Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive-Hazard Bayesian Online Change-Point Detection for Text Streams: A Dirichlet-Multinomial Formulation

Muhammad Ali Gunawan

guns12380@gmail.com

STIK Kendal https://orcid.org/0000-0002-0801-4939

Amalia Fitri

Universitas Pekalongan https://orcid.org/0009-0007-6361-1369

Research Article

Keywords: Adaptive hazard, Bayesian online change-point detection, Dirichlet-multinomial model, Text streams, Sequential statistical monitoring

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10299618/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Adaptive-Hazard Bayesian Online Change-Point Detection for Text Streams: A Dirichlet-Multinomial Formulation

Abstract

Bayesian online change-point detection provides a recursive filtering framework for sequential monitoring, yet standard implementations commonly assign the reset probability through a constant or duration-dependent hazard. This paper develops an adaptive-hazard formulation for sparse text streams using Dirichlet-multinomial predictive updating. Each document is represented as a high-dimensional word-count vector, and the hazard is modelled as a bounded logistic function of online Jensen–Shannon divergence between the current lexical distribution and a recent rolling history. The formulation preserves the standard run-length recursion while allowing lexical-distributional drift to modify the prior reset weight. The constant-hazard filter is recovered as a restricted case when the adaptive-hazard coefficient is zero. Numerical experiments were conducted on 7000 synthetic text streams and 700,000 sequential documents covering stationary, abrupt-shift, gradual-drift, recurring-topic, sparse-document, noisy-vocabulary, and weak-drift regimes. The adaptive-hazard filter maintained low false alarm rates under stationarity and remained comparable to the constant-hazard filter under abrupt and recurring topic shifts. Its clearest gains appeared under gradual lexical drift, where valid detection increased from 0.929 to 0.933 and mean delay decreased from 6.829 to 6.391, and under weak lexical drift, where valid detection increased from 0.135 to 0.169. Under sparse short documents, the method became more responsive but less conservative. An arXiv computational-linguistics application showed stable posterior filtering on a real weekly text stream without spurious boundary declaration.

Keywords Adaptive hazard · Bayesian online change-point detection · Dirichlet-multinomial model · Text streams · Sequential statistical monitoring

1 Introduction

Sequential change-point detection is a central problem in statistical monitoring, where the objective is to identify distributional changes as observations arrive over time. Classical procedures such as cumulative sum tests and related sequential schemes established foundational tools for detecting departures from a reference regime (Basseville & Nikiforov, 1993; Hinkley, 1970, 1971; Page, 1954; Tartakovsky et al., 2014). The field has since expanded through asymptotic theory, nonparametric procedures, multivariate detection, and broad algorithmic comparisons (Aminikhanghahi & Cook, 2017; Brodsky & Darkhovsky, 1993; Csörgő & Horváth, 1997; Matteson & James, 2014; van den Burg & Williams, 2020). Offline multiple change-point methods have further introduced penalized contrasts, optimal partitioning, pruning strategies, and scalable segmentation algorithms for large sequences (Fryzlewicz, 2014; Haynes et al., 2017; Jackson et al., 2005; Killick et al., 2012; Lavielle, 2005; Maidstone et al., 2017; Rigaill, 2015; Truong et al., 2018, 2020). Online detection remains a distinct problem because inference must be updated sequentially, using only the data available at the current time point.

Bayesian online change-point detection (BOCPD) provides a coherent filtering framework for this task by recursively updating the posterior distribution of the run length, defined as the elapsed time since the most recent change point (Adams & MacKay, 2007; Fearnhead & Liu, 2007). The framework separates the observation model from the hazard function. The observation model supplies predictive evidence for the incoming observation, whereas the hazard function specifies the prior probability of a segment reset. This modular structure has supported extensions to parameter estimation, model selection, nonparametric filtering, robust Bayesian updating, and scalable online inference (Agudelo-Espana et al., 2020; Altamirano et al., 2023; Caron et al., 2012; Detommaso et al., 2019; Knoblauch & Damoulas, 2018; Knoblauch & Lamperti, 2021; Ruggieri & Antonellis, 2016; Saatçi et al., 2010). In many implementations, however, the reset probability is specified through a constant or duration-dependent hazard, so it does not respond directly to distributional evidence carried by the incoming observation.

This restriction is consequential for text streams because observations are high-dimensional sparse count vectors, and changes often emerge as lexical-distributional movement rather than abrupt shifts in low-dimensional summaries. Standard information-retrieval representations treat documents as structured lexical objects, while multinomial and Dirichlet-multinomial models provide natural count-based representations for documents and word burstiness (Blei, 2012; Blei et al., 2003; Madsen et al., 2005; Manning et al., 2008). Related work on topic detection, document-based change-point detection, and online topic modelling has treated sequential textual variation through topic structure or document-level representations (Allan, 2002; Kim & Choi, 2015; Wang & Goutte, 2018). Information-theoretic divergences provide a complementary way to quantify movement between empirical lexical distributions (Cover & Thomas, 2006; Endres & Schindelin, 2003; Lin, 1991; Lu et al., 2020). A methodological gap remains in connecting sparse count-based predictive updating with a hazard function that responds explicitly to online lexical-drift evidence.

We propose an adaptive-hazard Dirichlet-multinomial BOCPD method for sparse text streams. Each document is represented as a word-count vector, and the one-step predictive distribution is obtained through Dirichlet-multinomial updating. The hazard is specified as a bounded logistic function of an online Jensen–Shannon divergence between the current document distribution and a recent rolling lexical history. The resulting filter preserves the standard BOCPD run-length recursion while allowing lexical-distributional drift to modify the prior reset probability. The constant-hazard model is recovered as a restricted case when the adaptive-hazard coefficient is zero.

The contribution of this article is threefold. First, it formulates a Dirichlet-multinomial BOCPD procedure for sparse text streams, with explicit posterior predictive updating for high-dimensional word-count observations. Second, it introduces a bounded logistic adaptive hazard that links online Jensen–Shannon drift to the prior reset probability, thereby separating boundary plausibility from segment-level predictive compatibility. Third, it provides a sparse computational implementation with posterior pruning and evaluates the method against constant-hazard Dirichlet-multinomial BOCPD, Jensen–Shannon cumulative sum monitoring, Jensen–Shannon Page–Hinkley monitoring, and Jensen–Shannon penalized segmentation across controlled regimes involving stationarity, abrupt shifts, gradual drift, recurring topics, sparse documents, noisy vocabulary, and weak lexical drift.

2 Standard BOCPD for text streams

Let $\mathbf{x}_{1:T} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ denote a sequence of text observations arriving over time. Each observation is represented as a sparse word-count vector

$$\mathbf{x}_t = (x_{t1}, \dots, x_{tV}) \in \mathbb{N}_0^V, \tag{1}$$

where $V$is the vocabulary size. The document length at time $t$is

$$m_t = \sum_{v=1}^{V} x_{tv}\,. \tag{2}$$

The objective of online change-point detection is to infer whether the distribution generating $\mathbf{x}_t$ has changed using only observations available up to time t. In Bayesian online change-point detection, this problem is expressed through a posterior distribution over the run length, which records the elapsed time since the most recent change point (Adams & MacKay, 2007; Fearnhead & Liu, 2007).

2.1 Run-length posterior

Let $r_t \in \{0,1, \dots, t\}$ denote the run length at time t. The event $r_t = 0$ indicates that a new segment starts at time t, while $r_t = r$ indicates that the current segment has continued for r time steps since the most recent reset. Bayesian online change-point detection recursively updates the joint distribution

$$p(r_t, \mathbf{x}_{1:t}), \tag{3}$$

and obtains the posterior run-length distribution by normalization:

$$p(r_t \mid \mathbf{x}_{1:t}) = \frac{p(r_t, \mathbf{x}_{1:t})}{\sum_{r_t'=0}^{t} p(r_t', \mathbf{x}_{1:t})}. \tag{4}$$

The recursion is defined by two mutually exclusive transitions. The first is a growth transition, in which the current segment continues and the run length increases by one. The second is a reset transition, in which a change point occurs and the run length is set to zero. Under a hazard function $H(r)$, the standard growth update is

$$p(r_t = r + 1, \mathbf{x}_{1:t}) = p(\mathbf{x}_t \mid r_{t-1} = r, \mathbf{x}_{1:t-1})\{1 - H(r)\}p(r_{t-1} = r, \mathbf{x}_{1:t-1}), \tag{5}$$

for $r = 0, \dots, t-1$. The reset update is

$$p(r_t = 0, \mathbf{x}_{1:t}) = \sum_{r=0}^{t-1} p(\mathbf{x}_t \mid r_t = 0, \mathbf{x}_{1:t-1})H(r)p(r_{t-1} = r, \mathbf{x}_{1:t-1}). \tag{6}$$

The predictive term $p(\mathbf{x}_t \mid r_{t-1} = r, \mathbf{x}_{1:t-1})$ is computed from the sufficient statistics associated with the candidate segment of run length r. Thus, the filtering recursion does not commit to a single segmentation path. Instead, it propagates a posterior distribution over

possible current segment lengths. This representation is useful for text streams because abrupt lexical changes, gradual drift, and transient lexical fluctuations can produce competing explanations at the same time point.

For numerical implementation, probabilities are stored on the logarithmic scale and normalized using a log-sum-exp operation. Without pruning, the number of candidate run lengths increases linearly with t, and the cumulative filtering cost over T observations is quadratic in T. This motivates the pruning strategy introduced in Sect. 3.4.

2.2 Constant-hazard limitation

A common BOCPD specification uses a constant hazard,

$$H(r) = H_0, 0 < H_0 < 1. \tag{7}$$

This assumption induces a geometric prior over segment lengths and remains computationally convenient when changes are abrupt and the expected duration is well specified. Its limitation is that the reset probability is fixed, or depends only on run length under a duration-dependent hazard $H(r)$, rather than on the discrepancy between $\mathbf{x}_t$ and recent observations. For online text streams, such discrepancy is informative: a document departing from recent lexical history signals segment renewal. Under a constant hazard, this evidence enters only through the one-step predictive likelihood, leaving the prior reset weight unchanged. The limitation is clearest under gradual lexical drift; Sect. 3 addresses it by conditioning the hazard on an online text-drift feature while preserving recursion.

2.3 Dirichlet-multinomial predictive model

For sparse text streams, a natural observation model treats each document as a multinomial count vector conditional on a segment-specific word-probability vector. Let

$$\boldsymbol{\theta}_s = (\theta_{s1}, \dots, \theta_{sV})$$

denote the word-probability vector for segment s, where $\theta_{sv} \geq 0$ and $\sum_{v=1}^{V} \theta_{sv} = 1$. Conditional on $\boldsymbol{\theta}_s$ and document length m_t, the document count vector follows

$$\mathbf{x}_t \mid \boldsymbol{\theta}_s, m_t \sim \text{Multinomial}\,(m_t, \boldsymbol{\theta}_s). \tag{8}$$

A Dirichlet prior is placed on the segment-specific word probabilities:

$$\boldsymbol{\theta}_s \sim \text{Dirichlet}\,(\boldsymbol{\eta}), \tag{9}$$

where

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_V), \eta_v > 0.$$

This prior is conjugate to the multinomial likelihood. If the candidate segment associated with run length $r$has accumulated word-count sufficient statistics

$$\mathbf{N}_{t-1}^{(r)} = (N_{t-1,1}^{(r)}, \dots, N_{t-1,V}^{(r)}), \tag{10}$$

then the posterior parameter vector before observing $\mathbf{x}_t$ is

$$\boldsymbol{\eta} + \mathbf{N}_{t-1}^{(r)}. \tag{11}$$

The resulting one-step posterior predictive distribution is Dirichlet-multinomial:

$$p(\mathbf{x}_t \mid r_{t-1} = r, \mathbf{x}_{1:t-1}) = \frac{m_t!}{\prod_{v=1}^{V} x_{tv}!} \frac{\Gamma(A_{t-1}^{(r)})}{\Gamma(A_{t-1}^{(r)} + m_t)} \prod_{v=1}^{V} \frac{\Gamma(\eta_v + N_{t-1,v}^{(r)} + x_{tv})}{\Gamma(\eta_v + N_{t-1,v}^{(r)})}, \tag{12}$$

where

$$A_{t-1}^{(r)} = \sum_{v=1}^{V} \left(\eta_v + N_{t-1,v}^{(r)}\right). \tag{13}$$

For posterior comparison across run-length states at a fixed time t, the multinomial coefficient

$$\frac{m_t!}{\prod_{v=1}^{V} x_{tv}!}$$

is invariant in $r$and can be omitted from normalized filtering probabilities. The log-predictive score used in the implementation is therefore

$$\log \pi_t^{(r)} = \log \Gamma(A_{t-1}^{(r)}) - \log \Gamma(A_{t-1}^{(r)} + m_t) + \sum_{v:x_{tv}>0} \left[\log \Gamma(\eta_v + N_{t-1,v}^{(r)} + x_{tv}) - \log \Gamma(\eta_v + N_{t-1,v}^{(r)})\right], \tag{14}$$

up to an additive constant independent of r. The summation is restricted to the non-zero entries of $\mathbf{x}_t$, which exploits sparsity in high-dimensional vocabularies.

The Dirichlet-multinomial predictive model is well suited to text streams because it provides a conjugate update, accommodates count-valued observations, and yields segment-specific sufficient statistics. It also provides a direct probabilistic basis for comparing whether an incoming document is more compatible with a continuing segment or a newly reset segment. The next section extends this formulation by replacing the constant hazard with a bounded adaptive hazard driven by online lexical-distributional drift.

3 Adaptive-hazard Dirichlet-multinomial BOCPD

The standard formulation in Sect. 2 assigns the reset probability through a hazard function that is either constant or dependent on the current run length. For sparse text streams, the incoming document also contains information about lexical-distributional departure from recent history. The proposed model introduces this information at the hazard level by conditioning the reset probability on an online text-drift feature, while retaining Dirichlet-multinomial predictive updating for the observation model.

Let $c_t \in \{0,1\}$ denote the change-point indicator at time t, where $c_t = 1$ indicates that a new segment starts at t. The adaptive-hazard model specifies

$$P(c_t = 1 \mid z_t) = H_t, \tag{15}$$

where z_t is a scalar drift feature computed from $\mathbf{x}_t$ and a recent rolling lexical history. This feature-conditioned hazard modifies the reset weight in the Bayesian online filtering recursion. The predictive likelihood remains Dirichlet-multinomial, so the model separates two sources of evidence: segment compatibility through the posterior predictive distribution and boundary plausibility through the adaptive hazard.

3.1 Online text-drift feature

The adaptive hazard is driven by a lexical-distributional discrepancy between the current document and the recent text stream. Let $\delta > 0$ be a smoothing constant. The smoothed empirical distribution of the current document is

$$\widehat{\boldsymbol{\phi}}_t = \frac{\mathbf{x}_t + \delta \mathbf{1}}{m_t + V\delta}, \tag{16}$$

where $\mathbf{1}$ is a V-dimensional vector of ones. Let w denote the rolling-window size. The smoothed empirical distribution of the recent lexical history is

$$\widehat{\boldsymbol{\phi}}_{t-w:t-1} = \frac{\sum_{j=\max(1,t-w)}^{t-1} \mathbf{x}_j + \delta \mathbf{1}}{\sum_{j=\max(1,t-w)}^{t-1} m_j + V\delta}. \tag{17}$$

For the first observation, where no previous document is available, the drift feature is set to zero.

The online text-drift feature is defined as the Jensen–Shannon divergence between the two smoothed distributions:

$$z_t = D_{\mathrm{JS}}\left(\widehat{\boldsymbol{\phi}}_t, \widehat{\boldsymbol{\phi}}_{t-w:t-1}\right). \tag{18}$$

For two discrete probability vectors **p**and **q**, the Jensen–Shannon divergence is

$$D_{\mathrm{JS}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} D_{\mathrm{KL}}(\mathbf{p} \parallel \mathbf{m}) + \frac{1}{2} D_{\mathrm{KL}}(\mathbf{q} \parallel \mathbf{m}), \mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q}), \tag{19}$$

where

$$D_{\mathrm{KL}}(\mathbf{p} \parallel \mathbf{m}) = \sum_{v=1}^{V} p_v \log \frac{p_v}{m_v}. \tag{20}$$

The smoothing constant ensures that all entries are strictly positive. Jensen–Shannon divergence is bounded and symmetric, making it a stable scalar feature for online lexical-distributional drift (Endres & Schindelin, 2003; Lin, 1991). In the sparse implementation, the

summation is evaluated over the union of non-zero terms in the current document and the rolling-window counts, with smoothing applied analytically.

3.2 Logistic adaptive hazard

The adaptive hazard maps $z_t$to a bounded reset probability:

$$H_t = h_{\min} + (h_{\max} - h_{\min})\sigma(\alpha + \beta z_t), \tag{21}$$

where

$$\sigma(u) = \frac{1}{1 + \exp(-u)}. \tag{22}$$

The constants $h_{\min}$ and $h_{\max}$ satisfy

$$0 < h_{\min} < h_{\max} < 1, \tag{23}$$

and prevent the reset probability from degenerating to zero or one. The coefficient βcontrols the sensitivity of the hazard to lexical-distributional drift. When $\beta = 0$, the hazard becomes

$$H_t = h_{\min} + (h_{\max} - h_{\min})\sigma(\alpha), \tag{24}$$

which is constant over time. Hence, the constant-hazard BOCPD model is nested as a restricted case.

To align the adaptive hazard with a target baseline hazard H_0, the intercept is set as

$$\alpha = \text{logit}\left(\frac{H_0 - h_{\min}}{h_{\max} - h_{\min}}\right), \tag{25}$$

provided that $h_{\min} < H_0 < h_{\max}$. This calibration ensures that the bounded logistic hazard coincides with the baseline reset probability when the drift feature is at its reference level. The functional form of the adaptive hazard is shown in Fig. 1.

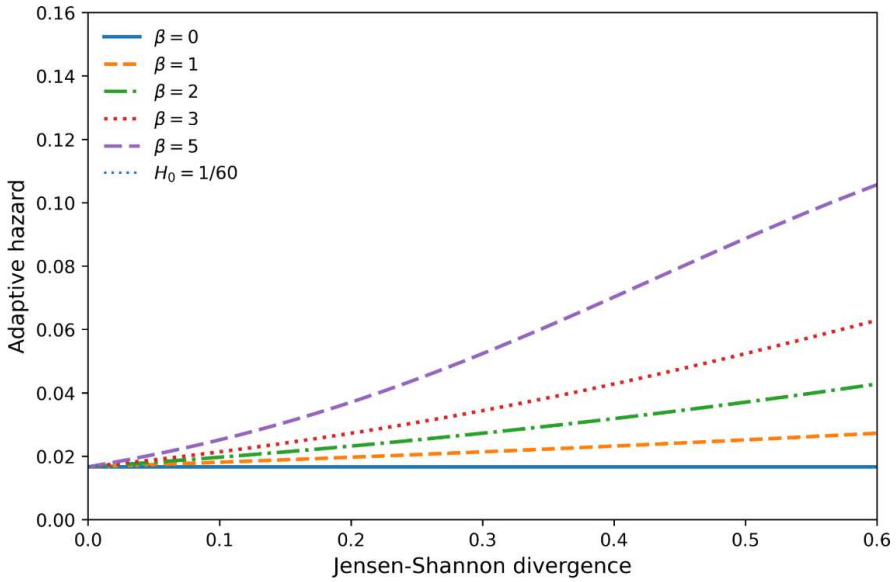


Fig. 1 Adaptive hazard as a bounded logistic function of Jensen–Shannon divergence

The bounded logistic specification has two roles. First, it translates lexical-distributional drift into a time-varying reset probability. Second, it regularizes the influence of the drift feature by imposing an upper bound on the hazard. This is important for sparse text streams, where divergence values can fluctuate because of document-length variation and sampling sparsity.

3.3 Posterior recursion

Let

$$\pi_t^{(r)} = p(\mathbf{x}_t \mid r_{t-1} = r, \mathbf{x}_{1:t-1}) \tag{26}$$

denote the Dirichlet-multinomial posterior predictive probability for candidate run length r. Let

$$p_{t-1}(r) = p(r_{t-1} = r \mid \mathbf{x}_{1:t-1}) \tag{27}$$

be the normalized run-length posterior at time $t-1$. The adaptive-hazard update forms unnormalized growth and reset weights over the active run-length set R_{t-1}.
For the growth transition,

$$q_t(r+1) = \pi_t^{(r)}(1 - H_t)p_{t-1}(r), r \in R_{t-1}. \tag{28}$$

For the reset transition,

$$q_t(0) = p_0(\mathbf{x}_t) \sum_{r \in R_{t-1}} H_t \, p_{t-1}(r), \tag{29}$$

where $p_0(\mathbf{x}_t)$ is the prior predictive probability for a new segment,

$$p_0(\mathbf{x}_t) = p(\mathbf{x}_t \mid \boldsymbol{\eta}). \tag{30}$$

Since H_t is common to all retained run-length states at time t, Eq. (29) emphasizes that the reset probability is modified by the current lexical-drift feature rather than by the previous run length.
The normalized posterior is

$$p_t(r) = p(r_t = r \mid \mathbf{x}_{1:t}) = \frac{q_t(r)}{\sum_{r' \in R_t^*} q_t\,(r')}, \tag{31}$$

where R_t^* denotes the candidate set obtained after applying the reset and growth transitions. This recursion differs from constant-hazard BOCPD only through the replacement of H_0 by H_t. The observation model and sufficient-statistic updates are unchanged. For each retained run-length state, the sufficient statistics are updated as

$$\mathbf{N}_t^{(r+1)} = \mathbf{N}_{t-1}^{(r)} + \mathbf{x}_t \tag{32}$$

for growth states, and

$$\mathbf{N}_t^{(0)} = \mathbf{x}_t \tag{33}$$

for the reset state. This representation preserves conjugacy and avoids refitting a separate model for each candidate segmentation.
The posterior change-point probability at time t is

$$p_t(0) = p(r_t = 0 \mid \mathbf{x}_{1:t}), \tag{34}$$

and the maximum a posteriori run length is

$$\hat{r}_t = \arg\max_{r \in R_t} p_t(r). \tag{35}$$

The posterior entropy,

$$E_t = -\sum_{r \in R_t} p_t\,(r) \log\, p_t(r), \tag{36}$$

is used as a measure of posterior uncertainty over the current segment length.

For reporting detection times in the numerical experiments, a candidate change point is declared when the MAP run length falls below a reset threshold after having previously exceeded a minimum run-length threshold. A minimum spacing rule is imposed between consecutive detections. This rule converts the posterior run-length path into an operational detection time while preserving the distinction between filtering probabilities and declared alarms.

3.4 Pruning and computational complexity

Without pruning, BOCPD maintains $t + 1$ possible run-length states at time t, so the posterior state space grows linearly in time and the cumulative filtering cost over T observations grows

quadratically in T. For high-dimensional text streams, this becomes inefficient even when each document is sparse. The proposed implementation uses probability pruning:

$$R_t = \{r \in R_t^* : p_t(r) > \varepsilon\}, \tag{37}$$

where $\varepsilon > 0$ is a pruning threshold. The reset state $r = 0$ and the MAP run-length state $\hat{r}_t$ are always retained. This prevents the filter from discarding the newly reset state or the most probable continuing state.

Let $\bar{R}$ denote the average number of retained run-length states, and let $\bar{s}$ denote the average number of non-zero entries in the document count vectors. The sparse implementation evaluates the Dirichlet-multinomial predictive score over non-zero terms and has practical filtering cost

$$O(T\bar{R}\bar{s}), \tag{38}$$

excluding the cost of maintaining rolling-window statistics for the drift feature. The online drift feature is updated from rolling counts and evaluated over sparse supports, so its computational burden remains modest in practice.

All filtering probabilities are computed on the logarithmic scale. For each candidate run length, the log-predictive score is

$$\log \pi_t^{(r)} = \log \Gamma(A_{t-1}^{(r)}) - \log \Gamma(A_{t-1}^{(r)} + m_t) + \sum_{v:x_{tv}>0} \left[\log \Gamma(\eta_v + N_{t-1,v}^{(r)} + x_{tv}) - \log \Gamma(\eta_v + N_{t-1,v}^{(r)})\right], \tag{39}$$

up to an additive constant independent of r. Posterior normalization is performed by log-sum-exp. This implementation avoids numerical underflow and supports long filtering sequences.
The adaptive hazard and pruning step play distinct roles. The adaptive hazard changes the statistical reset weighting by incorporating lexical-drift evidence. Pruning changes the computational representation of the posterior by controlling how many run-length hypotheses are retained.

3.5 Algorithm

The complete adaptive-hazard Dirichlet-multinomial BOCPD procedure is summarized in Algorithm 1.

Algorithm 1 Adaptive-hazard Dirichlet-multinomial BOCPD for sparse text streams
Input: text stream $\mathbf{x}_{1:T}$; vocabulary size V; Dirichlet prior $\boldsymbol{\eta}$; smoothing constant δ; rolling-window size w; adaptive-hazard parameters $h_{\min}, h_{\max}, \alpha, \beta$; pruning threshold ε.
Initialize: set $p_0(r_0 = 0) = 1$, $\mathbf{N}_0^{(0)} = \mathbf{0}$, and $R_0 = \{0\}$.
For $t = 1, \ldots, T$:

1. Compute the document length

$$m_t = \sum_{v=1}^{V} x_{tv}\,.$$

2. Compute the smoothed current-document distribution

$$\widehat{\boldsymbol{\phi}}_t = \frac{\mathbf{x}_t + \delta\mathbf{1}}{m_t + V\delta}.$$

3. Compute the smoothed rolling-history distribution

$$\widehat{\boldsymbol{\phi}}_{t-w:t-1} = \frac{\sum_{j=\max(1,t-w)}^{t-1} \mathbf{x}_j + \delta\mathbf{1}}{\sum_{j=\max(1,t-w)}^{t-1} m_j + V\delta}.$$

4. Compute the text-drift feature

$$z_t = D_{\mathrm{JS}}\left(\widehat{\boldsymbol{\phi}}_t, \widehat{\boldsymbol{\phi}}_{t-w:t-1}\right).$$

5. Compute the adaptive hazard

$$H_t = h_{\min} + (h_{\max} - h_{\min})\sigma(\alpha + \beta z_t).$$

6. For each $r \in R_{t-1}$, compute the Dirichlet-multinomial log-predictive score $\log \pi_t^{(r)}$.
7. Compute growth weights

$$q_t(r+1) = \pi_t^{(r)}(1 - H_t)p_{t-1}(r), r \in R_{t-1}.$$

8. Compute the reset weight

$$q_t(0) = p_0(\mathbf{x}_t) \sum_{r \in R_{t-1}} H_t\, p_{t-1}(r).$$

9. Normalize $q_t(r)$over all candidate states $R_t^*$to obtain

$$p_t(r) = p(r_t = r \mid \mathbf{x}_{1:t}).$$

10. Apply probability pruning:

$$R_t = \{r \in R_t^* : p_t(r) > \varepsilon\},$$

while always retaining $r = 0$ and

$$\hat{r}_t = \arg\max_{r \in R_t^*} p_t(r).$$

11. Update sufficient statistics:

$$\mathbf{N}_t^{(r+1)} = \mathbf{N}_{t-1}^{(r)} + \mathbf{x}_t, \mathbf{N}_t^{(0)} = \mathbf{x}_t.$$

12. Store $p_t(0)$, $\hat{r}_t$, posterior entropy E_t, and $\mid R_t \mid$.

Output: posterior run-length distributions, posterior change-point probabilities, MAP run-length path, posterior entropy, and retained-state counts.

4 Numerical experiments

We conducted numerical experiments to examine the operating characteristics of the adaptive-hazard filter under controlled forms of lexical nonstationarity. The experiments were designed to separate three statistical effects: the strength of the predictive evidence supplied by the Dirichlet-multinomial model, the contribution of the adaptive hazard when lexical drift accumulates over time, and the computational reduction obtained through posterior pruning. The data-generating mechanism was not intended to reproduce a specific corpus. Instead, it provides sparse count-valued streams with known regime paths, so that detection, localization, false alarm behavior, posterior uncertainty, and retained-state complexity can be evaluated directly.

All methods were applied to the same generated streams. The proposed method is denoted DM-BOCPD-AH. The primary Bayesian comparator is DM-BOCPD-CH, a constant-hazard Dirichlet-multinomial filter with the same observation model, sufficient-statistic updates, pruning rule, and detection rule. This paired comparison isolates the effect of replacing the fixed reset probability with the adaptive hazard. Three additional methods, JS-CUSUM, JS-PH, and JS-PELT, were included to compare Bayesian run-length filtering with direct monitoring or segmentation of the Jensen–Shannon drift sequence.

4.1 Data-generating mechanism

Each simulated stream consists of sparse word-count observations

$$\mathbf{x}_{i,1:T} = (\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,T}), \tag{40}$$

where $i = 1, \dots, N$ indexes independent streams and $t = 1, \dots, T$ indexes time. The baseline design used $N = 1000$ streams per scenario, $T = 100$ documents per stream, vocabulary size $V = 1000$, and $K = 6$ latent topics. Observations were generated from topic-specific multinomial distributions with document-level Dirichlet variation. This construction gives exact control over the latent regime path while retaining the sparse count structure of text streams.

Let $\boldsymbol{\psi}_k$ denote the word-probability vector for topic k. Topic-word distributions were generated from structured Dirichlet distributions,

$$\boldsymbol{\psi}_k \sim \text{Dirichlet}\,(\mathbf{a}_k), \tag{41}$$

with concentration parameters

$$a_{kv} = a_0 + a_1 I(v \in \mathcal{V}_k) + a_b I(v \in \mathcal{V}_b). \tag{42}$$

Here, $\mathcal{V}_k$ denotes the vocabulary subset emphasized by topic k, $\mathcal{V}_b$ denotes the background vocabulary, a_0 is the base concentration, a_1 controls topic separation, and a_b controls background-word prevalence. This specification produces topic distributions with shared background mass and topic-specific lexical concentration.

For stream i at time t, the scenario determines the latent regime $g_{i,t}$. Conditional on this regime, the document-specific word distribution is

$$\boldsymbol{\theta}_{i,t} \sim \text{Dirichlet}\,(\kappa\boldsymbol{\psi}_{g_{i,t}}), \tag{43}$$

where κcontrols within-topic document variability. Smaller values of κproduce noisier documents around the topic-level distribution. Document length is generated as

$$m_{i,t} = m_{\min} + Y_{i,t}, Y_{i,t} \sim \text{Poisson}\,(\lambda_m), \tag{44}$$

and the observed count vector follows

$$\mathbf{x}_{i,t} \sim \text{Multinomial}\,(m_{i,t}, \boldsymbol{\theta}_{i,t}). \tag{45}$$

The data-generating mechanism deliberately aligns with the Dirichlet-multinomial observation model while still creating nontrivial detection problems. Short documents introduce sampling variation, noisy vocabulary weakens topic evidence, gradual drift distributes boundary information over time, and weak lexical drift reduces separation between regimes. The principal parameter settings used in the final numerical experiments are reported in Table 1.

Table 1 Simulation parameter settings

Component	Setting
Number of streams per scenario	N=1000
Stream length	T=100
Vocabulary size	V=1000
Number of latent topics	k=6
Topic-specific vocabulary size	$\lfloor 0.08V \rfloor$
Background vocabulary size	$\lfloor 0.15V \rfloor$
Document length model	$m_{i,t} = m_{\min} + \text{Poisson}(\lambda_m)$
Baseline mean document length	approximately 50 tokens
Document-level variability	controlled by k
Baseline hazard	$H_0 = 1/60$
Adaptive-hazard lower bound	$h_{\min} = 0.001$
Adaptive-hazard upper bound	$h_{\max} = 0.15$
Adaptive-hazard coefficient	$\beta = 2.0$
Rolling-window size	$w = 10$
Smoothing constant	$\delta > 0$
Pruning threshold	$\varepsilon = 10^{-8}$

The adaptive-hazard intercept was calibrated so that the bounded logistic hazard coincides with the baseline hazard H_0 under the constant-hazard restriction. Thus, differences between DM-BOCPD-AH and DM-BOCPD-CH are attributable to the feature-dependent hazard response rather than to a different baseline reset level.

4.2 Topic-shift scenarios

Seven scenarios were constructed to separate distinct forms of lexical nonstationarity. Table 2 summarizes the regime structure and evaluation target for each scenario.

Table 2 Topic-shift scenarios used in the numerical experiments

Scenario	Description	Change-point structure	Evaluation target
S0 No-change	Stationary stream from one latent topic regime	No true change point	False alarm behavior
S1 Abrupt shift	Sudden transition between two topic regimes	$\tau = 60$	Abrupt boundary detection
S2 Gradual drift	Smooth transition between two lexical regimes	$\tau_0 = 50$, $\tau_1 = 70$, $midpoint\tau^* = 60$	Drift detection over a transition window
S3 Recurring topic	Topic changes and later returns to a previous regime	$\tau_1 = 40$, $\tau_2 = 75$	Detection of repeated regimes
S4 Sparse documents	Abrupt shift with short documents	$\tau = 60$	Robustness under high sampling variation
S5 Noisy vocabulary	Abrupt shift with stronger background contamination	$\tau = 60$	Robustness under weakened topic evidence
S6 Weak lexical drift	Low-separation transition between regimes	$midpoint\tau^* = 60$	Detection under weak boundary information

The first three scenarios provide basic calibration points. S0 measures the tendency to declare changes when the stream is stationary. S1 provides a clean abrupt boundary, where a well-specified predictive model should detect the transition rapidly. S2 spreads the transition over a finite interval, requiring the detector to respond to accumulating lexical movement rather than to a single discontinuity.

The remaining scenarios stress specific features of online text monitoring. S3 examines a recurring-regime case in which a topic that appeared earlier returns later in the stream. S4 shortens documents and therefore increases variability in empirical word distributions. S5 increases background vocabulary contamination, reducing the relative contribution of topic-specific terms. S6 reduces lexical separation between regimes and forms the weakest boundary signal in the experiment.

The scenario-mean Jensen–Shannon divergence trajectories are shown in Fig. 2. These trajectories display the drift feature before transformation by the bounded logistic hazard.

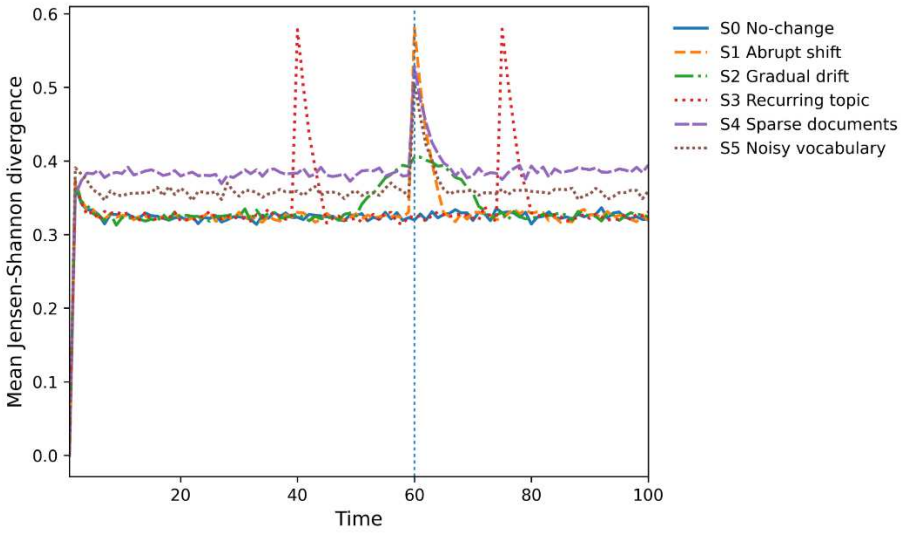


Fig. 2 Scenario-mean Jensen–Shannon divergence trajectories

The corresponding adaptive-hazard trajectories are shown in Fig. 3. This figure illustrates how lexical-distributional movement is converted into a time-varying prior reset probability.

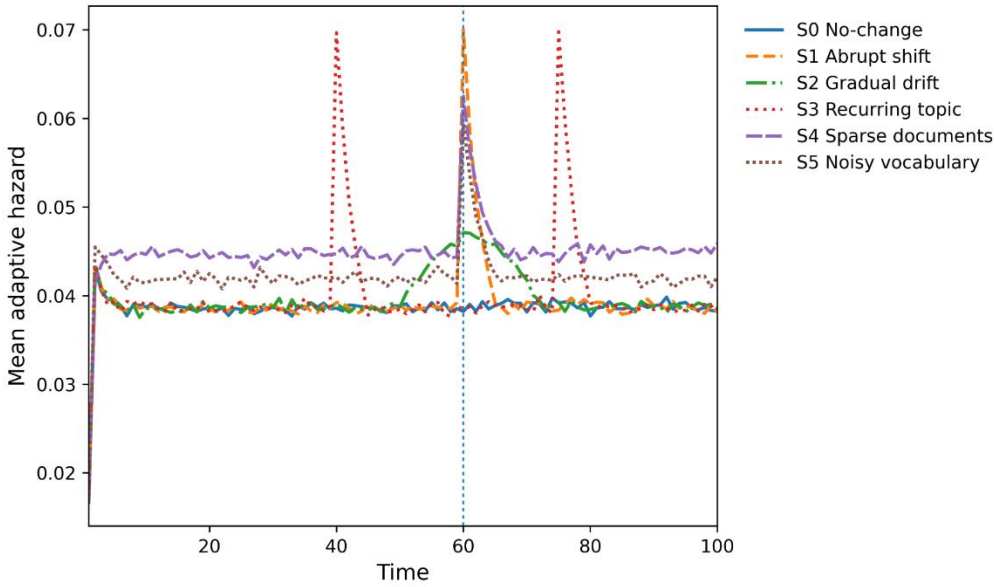


Fig. 3 Scenario-mean adaptive-hazard trajectories

The two trajectory figures clarify the role of z_t in the numerical design. Abrupt and recurring regimes create sharp local increases in the drift feature. Gradual drift produces a wider transition band. Weak drift produces a lower-amplitude signal, and sparse documents create more volatile short-term fluctuations. These distinctions are important because the adaptive hazard uses lexical-drift information to modify the reset prior; it does not replace the Dirichlet-multinomial predictive model.

4.3 Baseline methods

DM-BOCPD-CH was included as the primary Bayesian baseline. It uses the same Dirichlet-multinomial predictive distribution, the same pruning rule, and the same detection

rule as DM-BOCPD-AH, but replaces H_t with the fixed hazard $H_0 = 1/60$. This comparison gives a direct assessment of whether conditioning the reset prior on online lexical drift changes operating performance.

JS-CUSUM and JS-PH were used as sequential baselines operating directly on the Jensen–Shannon drift sequence. They represent a two-step strategy: reduce each document to a scalar lexical-distributional discrepancy and then apply a sequential monitoring rule. Their thresholds were calibrated using stationary streams, so their false alarm behavior was determined by the simulation design rather than by arbitrary tuning.

JS-PELT was included as an offline comparator. It applies penalized segmentation to the full Jensen–Shannon divergence sequence and therefore has access to future observations. It is not an online competitor, but it provides a reference for how direct segmentation of z_t behaves when the complete sequence is available. Including both online and offline baselines separates the effect of Bayesian run-length filtering from the effect of simply monitoring or segmenting the drift feature.

4.4 Evaluation metrics

The evaluation uses detection, localization, posterior-uncertainty, and computational summaries. Let $\hat{\tau}_i$ denote the first detected boundary for stream i, and let τ_i denote the reference boundary. For abrupt, sparse, and noisy scenarios, $\tau_i = 60$. For gradual and weak drift, $\tau_i = \tau^* = 60$, where τ^* is the midpoint of the transition. For recurring-topic streams, detections were evaluated against the corresponding regime transitions.

A detection was classified according to its position relative to the admissible region for the scenario. In the no-change scenario, every detection is a false alarm. In the gradual-drift scenario, detections inside the transition window $[50, 70]$ are treated as valid early detections, whereas detections before $\tau_0 = 50$ are false alarms. This rule avoids penalizing a method for detecting drift before the midpoint when the generating mechanism has already begun to change.

The valid detection rate is

$$\text{VDR} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{\tau}_i \text{ is valid}), \tag{46}$$

and the false alarm rate is

$$\text{FAR} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{\tau}_i \text{ is a false alarm}). \tag{47}$$

The missed detection rate is computed from the remaining streams whenever valid detections and false alarms form mutually exclusive categories.

For streams with valid detections, delay and localization error are defined as

$$d_i = \hat{\tau}_i - \tau_i, e_i = |\hat{\tau}_i - \tau_i|. \tag{48}$$

Mean and median summaries were computed over valid detections only. For BOCPD-based methods, posterior uncertainty at the detected time was measured through run-length entropy,

$$E_{\hat{\tau}_i} = -\sum_{r \in R_{\hat{\tau}_i}} p_{\hat{\tau}_i}(r) \log\, p_{\hat{\tau}_i}(r), \tag{49}$$

and maximum posterior mass,

$$M_{\hat{\tau}_i} = \max_{r \in R_{\hat{\tau}_i}} p_{\hat{\tau}_i}(r). \tag{50}$$

These quantities indicate whether a detected boundary is supported by a concentrated posterior distribution or by a diffuse set of competing run-length explanations.

Computational cost was summarized by runtime and retained run-length states after pruning. The average retained-state count records typical filtering complexity, while the maximum retained-state count records the largest posterior state space encountered in a stream. Monte Carlo uncertainty for a proportion $\hat{p}$was summarized by

$$\text{MCSE}\,(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \tag{51}$$

and Monte Carlo uncertainty for a mean $\bar{y}$was summarized by

$$\text{MCSE}\,(\bar{y}) = \frac{s_y}{\sqrt{n_y}}, \tag{52}$$

where s_y is the sample standard deviation over contributing streams and n_y is the number of streams used in the mean. These summaries treat the numerical experiments as a finite Monte Carlo study and quantify the simulation uncertainty associated with reported proportions and averages.

4.5 Detection performance

The final numerical experiment consisted of $N = 1000$ independent streams for each of the seven scenarios. Table 3 reports the main detection summaries for all methods, including valid detection rate, false alarm rate, mean detection delay, and mean localization error. The no-change scenario has no true boundary; performance in this scenario is therefore interpreted through false alarm behavior.

Table 3 Main detection performance across scenarios and methods

Scenario	Method	Valid detection	False alarm	Mean delay	Mean localization error
S0 No-change	DM-BOCPD-AH	0.000	0.005	NA	NA
S0 No-change	DM-BOCPD-CH	0.000	0.004	NA	NA

S0 No-change	JS-CUSUM	0.000	0.050	NA	NA
S0 No-change	JS-PH	0.000	0.050	NA	NA
S0 No-change	JS-PELT	0.000	0.050	NA	NA
S1 Abrupt shift	DM-BOCPD-AH	0.998	0.002	0.041	0.041
S1 Abrupt shift	DM-BOCPD-CH	1.000	0.000	0.048	0.048
S1 Abrupt shift	JS-CUSUM	0.982	0.014	1.173	1.173
S1 Abrupt shift	JS-PH	0.611	0.018	4.537	4.537
S1 Abrupt shift	JS-PELT	0.422	0.301	6.642	6.642
S2 Gradual drift	DM-BOCPD-AH	0.933	0.001	6.391	7.746
S2 Gradual drift	DM-BOCPD-CH	0.929	0.001	6.829	8.013
S2 Gradual drift	JS-CUSUM	0.567	0.004	3.951	4.857
S2 Gradual drift	JS-PH	0.666	0.000	5.179	6.587
S2 Gradual drift	JS-PELT	0.541	0.162	-0.397	13.765
S3 Recurring topic	DM-BOCPD-AH	1.000	0.000	0.036	0.036
S3 Recurring topic	DM-BOCPD-CH	1.000	0.000	0.041	0.041
S3 Recurring topic	JS-CUSUM	0.998	0.002	1.632	1.282
S3 Recurring topic	JS-PH	0.807	0.000	13.881	3.796
S3 Recurring topic	JS-PELT	0.527	0.329	16.698	1.476
S4 Sparse documents	DM-BOCPD-AH	0.919	0.077	0.551	0.551
S4 Sparse documents	DM-BOCPD-CH	0.943	0.050	0.614	0.614
S4 Sparse documents	JS-CUSUM	0.534	0.005	3.204	3.204
S4 Sparse documents	JS-PH	0.301	0.644	4.093	4.093
S4 Sparse documents	JS-PELT	0.164	0.574	9.860	9.860
S5 Noisy vocabulary	DM-BOCPD-AH	0.920	0.003	1.472	1.472
S5 Noisy vocabulary	DM-BOCPD-CH	0.917	0.003	1.552	1.552
S5 Noisy vocabulary	JS-CUSUM	0.304	0.014	4.095	4.095
S5 Noisy vocabulary	JS-PH	0.226	0.012	12.208	12.208
S5 Noisy vocabulary	JS-PELT	0.124	0.158	7.177	7.177
S6 Weak lexical drift	DM-BOCPD-AH	0.169	0.006	14.278	16.503
S6 Weak lexical drift	DM-BOCPD-CH	0.135	0.003	15.126	16.904
S6 Weak lexical drift	JS-CUSUM	0.062	0.001	28.177	28.177
S6 Weak lexical drift	JS-PH	0.783	0.006	16.309	17.831
S6 Weak lexical drift	JS-PELT	0.676	0.133	-2.368	9.182

Valid detection rates for the six nonstationary scenarios are displayed in Fig. 4. The no-change scenario is excluded from this figure because it contains no true boundary and is evaluated through false alarms.

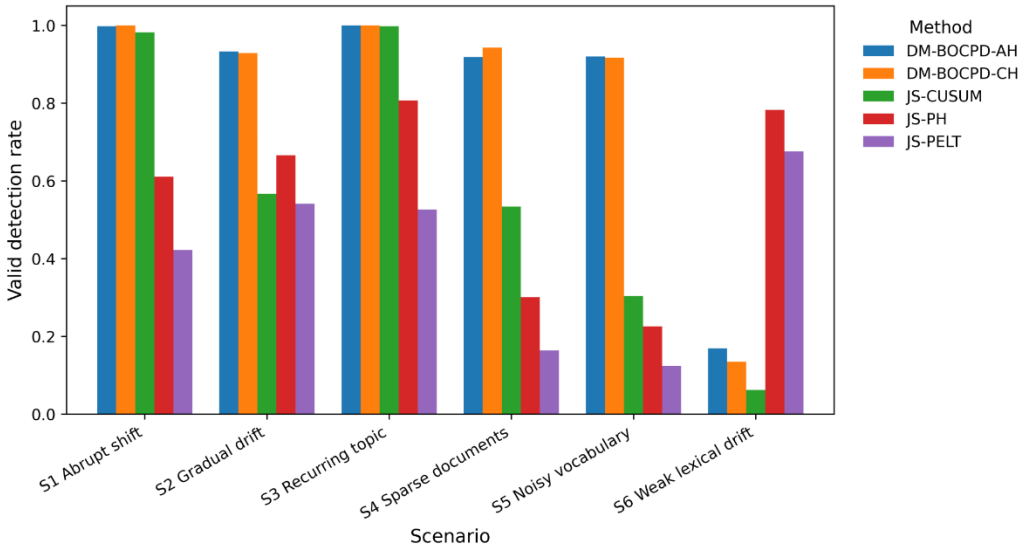


Fig. 4 Valid detection rate by scenario and detection method

Under abrupt and recurring-topic regimes, the two BOCPD filters behaved almost identically: DM-BOCPD-AH attained valid detection rates of 0.998 and 1.000, while DM-BOCPD-CH reached 1.000 in both cases, showing that the Dirichlet-multinomial predictive likelihood already supplies decisive reset evidence when lexical boundaries are sharp. Gradual drift and noisy vocabulary were more diagnostic of the adaptive hazard, with valid detection improving from 0.929 to 0.933 and from 0.917 to 0.920, respectively, without added false alarms. Under weak lexical drift, sensitivity rose from 0.135 to 0.169, although the problem remained difficult. Sparse documents reversed the pattern: responsiveness increased, but valid detection fell from 0.943 to 0.919 because short-document divergence carried greater sampling variation and reduced posterior conservativeness.

4.6 False alarm behavior

False alarm rates across scenarios and methods are shown in Fig. 5. This figure is central for interpreting detection performance because higher sensitivity is useful only when it does not arise from excessive premature signaling.

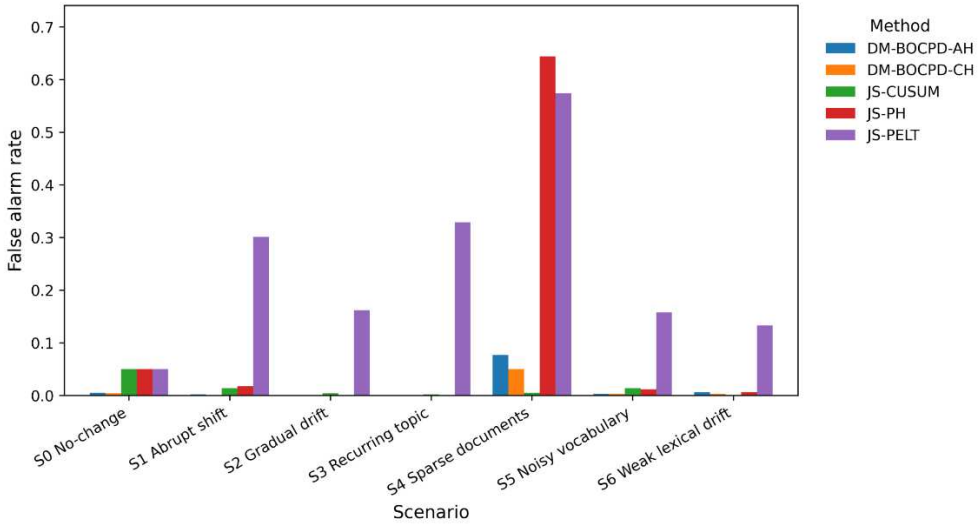


Fig. 5 False alarm rate by scenario and detection method

Under the stationary no-change regime, both BOCPD filters remained conservative: DM-BOCPD-AH and DM-BOCPD-CH yielded false alarm rates of 0.005 and 0.004, whereas the Jensen–Shannon benchmarks yielded 0.050. This contrast reflects the difference between posterior run-length filtering, which requires joint support from the predictive model and reset mechanism, and direct thresholding of a scalar drift score. The offline JS-PELT procedure further illustrated this distinction, producing elevated false alarms across nonstationary scenarios, including 0.301 under abrupt shift and 0.574 under sparse documents. The adaptive-hazard cost was concentrated in the sparse-document case, where false alarms rose from 0.050 to 0.077 because $z_t$became more sensitive to sampling variation under short lexical counts, despite bounded-hazard regularization in this regime.

4.7 Detection delay and localization

Mean detection delay for the two BOCPD filters is displayed in Fig. 6. The comparison is restricted to DM-BOCPD-AH and DM-BOCPD-CH because the two methods share the same predictive model and differ only in the hazard specification.

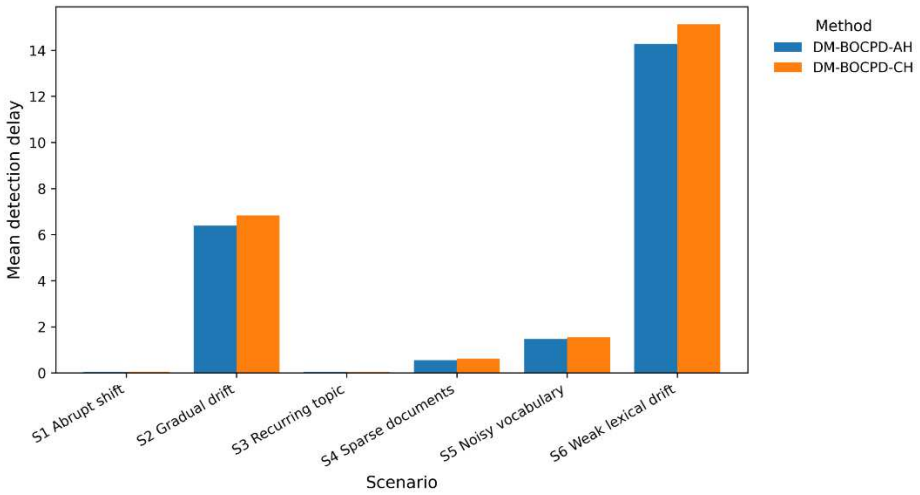


Fig. 6 Mean detection delay for adaptive-hazard and constant-hazard DM-BOCPD

The delay profile shows that adaptive hazard mainly changes responsiveness. Under abrupt and recurring-topic shifts, both filters detected boundaries almost immediately, with delays of 0.041 versus 0.048 and 0.036 versus 0.041 for DM-BOCPD-AH and DM-BOCPD-CH, because the predictive likelihood already supplied decisive reset evidence. Gradual drift showed the clearer timing gain: delay fell from 6.829 to 6.391 and localization error from 8.013 to 7.746. Noisy vocabulary and weak drift followed the same direction, reducing delay from 1.552 to 1.472 and from 15.126 to 14.278. Sparse documents confirmed the trade-off: faster detection, 0.551 versus 0.614, came with higher false alarms under severe sparsity in this setting.

4.8 Computational cost and posterior uncertainty

The computational and posterior summaries of the two BOCPD filters are reported in Table 4. The table gives posterior entropy, MAP posterior probability, average retained states, and maximum retained states. Retained-state counts are used as the principal computational

summary because the pruning rule directly controls the number of active run-length hypotheses.

Table 4 Posterior uncertainty and retained run-length states for BOCPD-based methods

Scenario	Method	Entropy at detection	MAP probability	Average retained states	Maximum retained states
S0 No-change	DM-BOCPD-AH	0.428	0.780	3.008	6
S0 No-change	DM-BOCPD-CH	0.409	0.745	3.006	6
S1 Abrupt shift	DM-BOCPD-AH	0.015	0.993	2.613	6
S1 Abrupt shift	DM-BOCPD-CH	0.012	0.995	2.609	6
S2 Gradual drift	DM-BOCPD-AH	0.262	0.890	2.963	8
S2 Gradual drift	DM-BOCPD-CH	0.262	0.891	2.953	8
S3 Recurring topic	DM-BOCPD-AH	0.013	0.995	2.419	6
S3 Recurring topic	DM-BOCPD-CH	0.015	0.994	2.414	6
S4 Sparse documents	DM-BOCPD-AH	0.253	0.895	4.309	15
S4 Sparse documents	DM-BOCPD-CH	0.261	0.893	4.199	15
S5 Noisy vocabulary	DM-BOCPD-AH	0.189	0.921	2.775	7
S5 Noisy vocabulary	DM-BOCPD-CH	0.183	0.924	2.773	7
S6 Weak lexical drift	DM-BOCPD-AH	0.340	0.842	3.066	7
S6 Weak lexical drift	DM-BOCPD-CH	0.311	0.859	3.053	7

Average retained run-length states are visualized in Fig. 7. The figure shows the posterior state space retained after pruning.

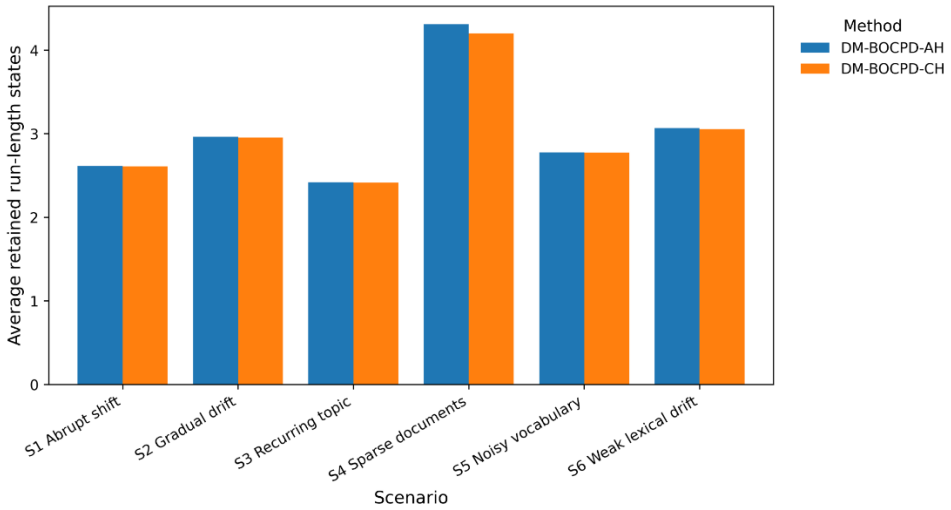


Fig. 7 Average retained run-length states for BOCPD-based methods

The retained-state counts were small across all scenarios. DM-BOCPD-AH retained between 2.419 and 4.309 states on average, and DM-BOCPD-CH retained between 2.414 and 4.199 states. The maximum retained-state count did not exceed 15. These values are far below the unpruned run-length space, which grows with time. The result confirms that probability pruning kept the online filter compact while preserving the main competing run-length hypotheses.

Sparse documents required the largest retained-state count. This is expected because short documents give less lexical evidence per time point, leaving more competing run-length explanations alive after normalization. The adaptive-hazard and constant-hazard filters had nearly identical retained-state profiles, indicating that the additional adaptive-hazard calculation did not materially increase posterior state complexity.

Posterior entropy at detection is shown in Fig. 8. Entropy is reported only for the two BOCPD filters because the JS-based baselines do not maintain a posterior distribution over run length.

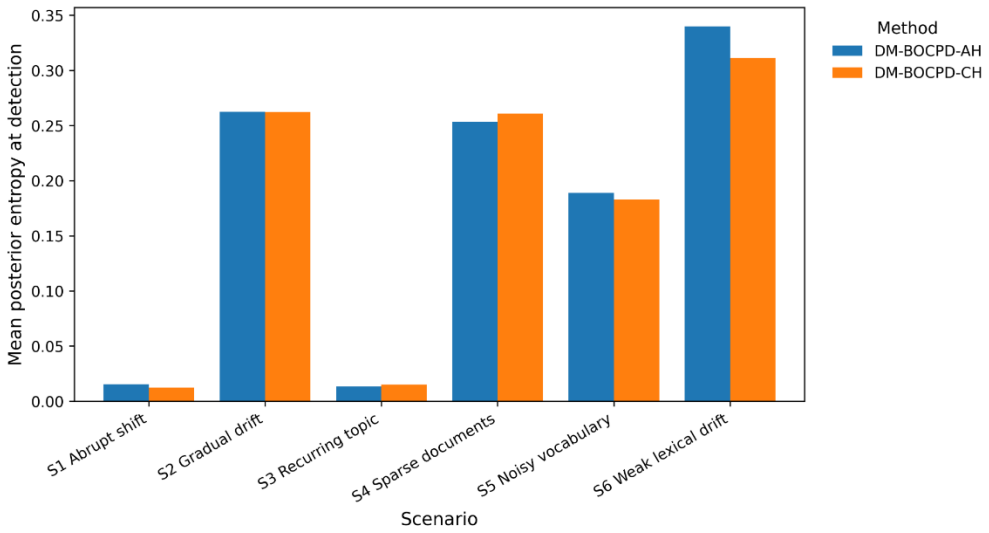


Fig. 8 Mean posterior entropy at detection for BOCPD-based methods

Posterior entropy tracked scenario difficulty: abrupt and recurring-topic shifts yielded concentrated run-length posteriors, with entropy 0.015/0.012 in S1 and 0.013/0.015 in S3 for DM-BOCPD-AH/CH, and MAP probabilities above 0.99. Gradual drift, sparse documents, noisy vocabulary, and weak lexical drift produced more diffuse posterior mass; entropy reached 0.262 in S2, 0.253/0.261 in S4, and 0.340/0.311 in S6. Thus, detection rates alone obscure posterior ambiguity. BOCPD makes this uncertainty explicit, showing that alarms under short documents or weak lexical contrast are statistically less concentrated than alarms generated by sharp topic changes. This distinction is central for sparse text monitoring in practice.

5 Application to arXiv cs.CL data

The proposed procedure was further examined on a real text stream constructed from arXiv metadata in the computational linguistics category, cs.CL. This application is intended as a stability check on a naturally varying sparse text stream rather than as an event-discovery benchmark with externally labelled change points. Titles and abstracts published between January 2023 and December 2024 were aggregated into weekly word-count vectors. After preprocessing and vocabulary truncation, the resulting stream contained $T = 106$ weekly periods, 5262 documents, and 1000 retained terms. The average number of documents per weekly period was 49.64.

The same filtering specification used in the numerical experiments was applied to the arXiv stream. DM-BOCPD-AH used the bounded logistic adaptive hazard driven by Jensen–Shannon divergence, while DM-BOCPD-CH used the same Dirichlet-multinomial predictive

model, pruning rule, vocabulary, and run-length recursion with a constant hazard. This paired design isolates the role of adaptive reset weighting because both filters operate on the same sparse count sequence and differ only in the hazard specification.

The arXiv cs.CL stream contained T=106 weekly periods, 5262 documents, and an average of 49.64 documents per period. The mean Jensen–Shannon divergence was 0.0769, with a maximum value of 0.2759. Neither filter selected a candidate boundary, and the maximum retained-state count was 3 for both methods. Table 5 therefore focuses on the quantities that distinguish the adaptive- and constant-hazard filters.

Table 5 Real-data summary for the arXiv cs.CL weekly text stream

Method	Mean hazard	Max hazard	Mean retained states	Candidate boundaries
DM-BOCPD-AH	0.0190	0.0263	2.99	0
DM-BOCPD-CH	0.0167	0.0167	2.99	0

The adaptive hazard increased the mean reset probability relative to the constant-hazard filter, but this increase did not produce a candidate boundary. Thus, the filter responded to lexical-distributional variation while retaining conservative posterior reset behaviour on the weekly arXiv stream

Figure 9 shows the online Jensen–Shannon divergence and the corresponding adaptive-hazard trajectory. The largest divergence occurred near the beginning of the stream, where the rolling lexical history was still short. Later fluctuations were more moderate. The adaptive hazard responded to these lexical movements, but the response remained bounded. This behavior follows directly from the model design: the drift feature changes the prior reset weight, whereas a posterior reset still requires support from the Dirichlet-multinomial predictive evidence.

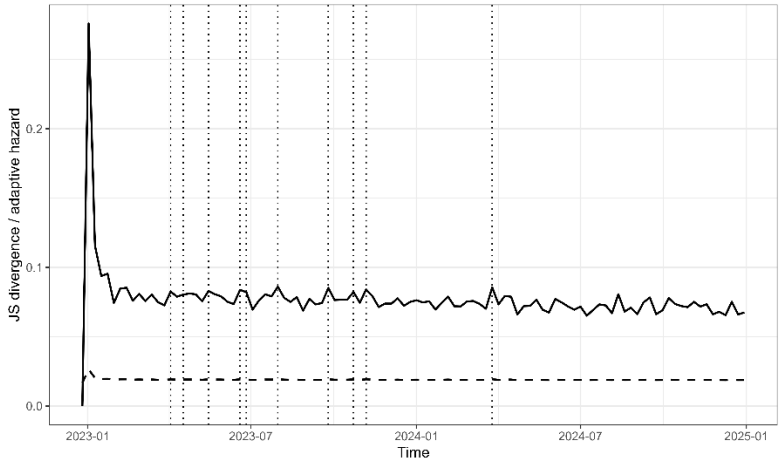


Fig. 9 Real-data Jensen–Shannon divergence and adaptive-hazard trajectory for the arXiv cs.CL weekly text stream

Figure 10 reports posterior reset probabilities under the adaptive- and constant-hazard filters. The posterior reset probability was 0.0167 at the initial time point, reflecting the baseline reset prior at the start of filtering. After the initial period, reset probabilities became numerically negligible. The largest post-initial reset probability was approximately 1.62×10^{-18} for DM-BOCPD-AH and 1.02×10^{-18} for DM-BOCPD-CH. Thus, the adaptive-hazard filter did not

convert moderate lexical variation into a false segment reset when the predictive evidence supported continuation.

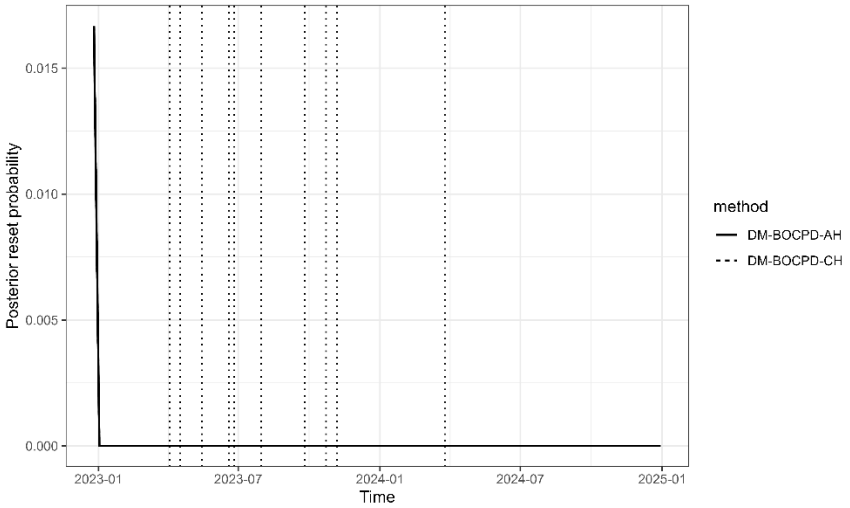


Fig. 10 Posterior reset probability for DM-BOCPD-AH and DM-BOCPD-CH on the arXiv cs.CL weekly text stream

Figure 11 shows the MAP run-length trajectories. Both filters produced an increasing MAP run length throughout the observation window, reaching 106 at the final weekly period. This trajectory indicates sustained posterior support for a continuing lexical regime rather than repeated segment resets. The close agreement between the adaptive- and constant-hazard trajectories also indicates that the adaptive prior did not override the predictive distribution when the incoming weekly documents remained compatible with the accumulated segment-level word distribution.

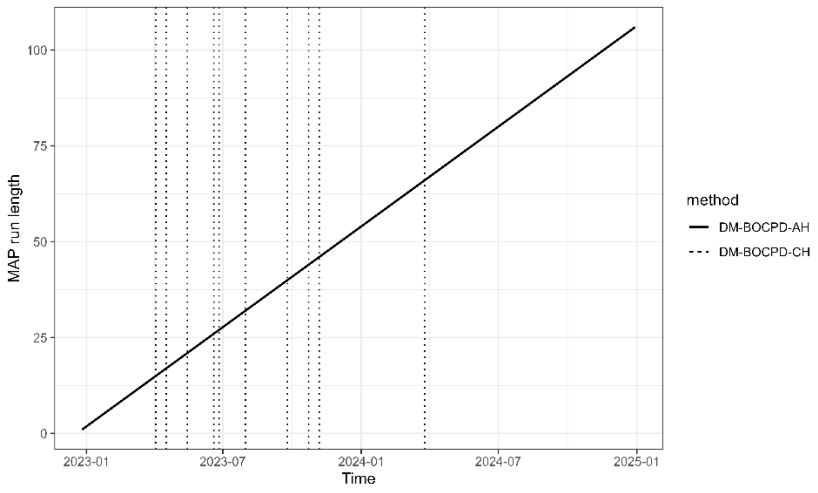


Fig. 11 MAP run-length trajectory for DM-BOCPD-AH and DM-BOCPD-CH on the arXiv cs.CL weekly text stream

Inspection of the weeks with the highest adaptive-hazard values supports the same interpretation. These weeks showed local vocabulary movements involving terms such as editing, encoding, autoregressive modelling, parsing, reward, community, adversarial evaluation, multimodal language models, and policy. Such movements indicate variation in

local lexical composition, but the Bayesian filter did not treat them as statistically supported segment boundaries. The application therefore complements the numerical experiments by showing conservative behavior on a sparse text stream with natural lexical variation. The adaptive hazard reacts to online drift evidence, while posterior reset remains governed by the joint effect of hazard weighting and Dirichlet-multinomial predictive support.

6 Concluding remarks

This article develops an adaptive-hazard Bayesian online change-point detection method for sparse text streams by combining Dirichlet-multinomial predictive updating with a bounded logistic hazard driven by online Jensen–Shannon divergence. The formulation preserves the recursive run-length filtering structure of Bayesian online change-point detection (Adams & MacKay, 2007; Fearnhead & Liu, 2007), while allowing the prior reset probability to respond to lexical-distributional movement in the incoming stream. The constant-hazard filter is recovered as the restricted case obtained when the adaptive-hazard coefficient is zero. The numerical experiments indicate that the method is best interpreted as a conditional extension of constant-hazard BOCPD, not as a uniformly dominant detector.

When lexical boundaries are abrupt or recurring, the Dirichlet-multinomial predictive likelihood already provides decisive reset evidence, and DM-BOCPD-AH and DM-BOCPD-CH therefore yield nearly identical performance, with valid detection rates close to one and negligible delays. The adaptive hazard is more informative when the change signal is distributed across time or weakly expressed in the one-step predictive likelihood. Under gradual lexical drift, valid detection increased from 0.929 to 0.933 and mean delay decreased from 6.829 to 6.391 without increasing false alarms. Under weak lexical drift, valid detection rose from 0.135 to 0.169 and mean delay decreased from 15.126 to 14.278. These gains are modest but statistically coherent because the hazard feature reacts to accumulated lexical movement, while the predictive distribution continues to determine whether segment-level evidence supports a reset.

The comparison with direct Jensen–Shannon monitoring clarifies the value of posterior run-length filtering. Classical cumulative sum procedures detect departures through accumulated evidence relative to a reference regime (Basseville & Nikiforov, 1993; Hinkley, 1970, 1971; Page, 1954), and penalized segmentation methods such as PELT estimate multiple change points when the full sequence is available (Killick et al., 2012; Truong et al., 2020). By contrast, DM-BOCPD-AH operates online and maintains a posterior distribution over the current segment length, thereby combining boundary plausibility with segment compatibility. The method also differs from text-stream approaches based on topic evolution or document-level topic structure (Kim & Choi, 2015; Wang & Goutte, 2018), since it monitors sparse word-count vectors directly through conjugate Dirichlet-multinomial filtering.

The main calibration trade-off appears under sparse documents: mean delay decreased from 0.614 to 0.551, but false alarms increased from 0.050 to 0.077 because short documents make Jensen–Shannon divergence more sensitive to sampling variation. The arXiv cs.CL application confirms conservative real-stream behavior: the adaptive hazard increased the mean hazard from 0.0167 to 0.0190, yet neither filter selected a candidate boundary, and posterior reset probabilities remained negligible after initialization. Computationally, probability pruning kept the filter compact, with fewer than 4.4 retained run-length states on average and at most 15 across simulations. The principal limitations are the use of a single divergence feature, controlled synthetic generation, and fixed hazard calibration; future

extensions should consider document length, vocabulary novelty, embedding-based semantic displacement, topic uncertainty, labeled event streams, and data-adaptive hazard sensitivity.

Acknowledgements
The author gratefully acknowledges the availability of open arXiv metadata used in the real-data application. The author also thanks the developers and maintainers of the open-source statistical software used for numerical computation, simulation, and graphical presentation.

Funding
No funding was received for conducting this study. No funds, grants, or other support were received during the preparation of this manuscript.

Statements and Declarations

Competing interests
The author declares no competing interests that are relevant to the content of this article.

Ethics approval
This study did not involve human participants, human data, animals, or clinical materials. Ethics approval was therefore not required.

Consent to participate
Not applicable.

Consent for publication
Not applicable.

Use of generative artificial intelligence
During the preparation of this manuscript, generative artificial intelligence tools were used for language editing, formatting assistance, and improvement of readability. The author reviewed, revised, and approved all content and takes full responsibility for the accuracy and integrity of the final manuscript.

Data and code availability
Synthetic text streams were generated from the article's data-generating mechanism, while the real-data application used public arXiv cs.CL metadata aggregated into weekly word-count vectors. R scripts for simulations, BOCPD filters, baselines, summaries, sensitivity analyses, and figures will be made available.

References

Adams, R. P., & MacKay, D. J. C. (2007). *Bayesian Online Changepoint Detection*. https://arxiv.org/abs/0710.3742

Agudelo-Espana, D., Gomez-Gonzalez, S., Bauer, S., Schölkopf, B., & Peters, J. (2020). Bayesian Online Prediction of Change Points. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, Proceedings of Machine Learning Research*, *124*, 320–329.

Allan, J. (2002). *Topic Detection and Tracking* (J. Allan, Ed.; Vol. 12). Springer US. https://doi.org/10.1007/978-1-4615-0933-2

Altamirano, M., Briol, F.-X., & Knoblauch, J. (2023). Robust and Scalable Bayesian Online Changepoint Detection. *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research*, *202*, 642–663.

Aminikhanghahi, S., & Cook, D. J. (2017). A Survey of Methods for Time Series Change Point Detection. *Knowledge and Information Systems*, *51*, 339–367. https://doi.org/10.1007/s10115-016-0987-z

Basseville, M., & Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brodsky, B. E., & Darkhovsky, B. S. (1993). *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers.

Caron, F., Doucet, A., & Gottardo, R. (2012). On-Line Changepoint Detection and Parameter Estimation with Application to Genomic Data. *Statistics and Computing*, *22*, 579–595. https://doi.org/10.1007/s11222-011-9248-x

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

Csörgő, M., & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley.

Detommaso, G., Hoitzing, H., Cui, T., & Alamir, A. (2019). *Stein Variational Online Changepoint Detection with Applications to Hawkes Processes and Neural Networks*.

Endres, D. M., & Schindelin, J. E. (2003). A New Metric for Probability Distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860. https://doi.org/10.1109/TIT.2003.813506

Fearnhead, P., & Liu, Z. (2007). On-Line Inference for Multiple Changepoint Problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 589–605. https://doi.org/10.1111/j.1467-9868.2007.00601.x

Fryzlewicz, P. (2014). Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics*, *42*(6), 2243–2281. https://doi.org/10.1214/14-AOS1245

Haynes, K., Eckley, I. A., & Fearnhead, P. (2017). Computationally Efficient Changepoint Detection for a Range of Penalties. *Journal of Computational and Graphical Statistics*, *26*(1), 134–143. https://doi.org/10.1080/10618600.2015.1116445

Hinkley, D. V. (1970). Inference about the Change-Point in a Sequence of Random Variables. *Biometrika*, *57*(1), 1–17. https://doi.org/10.1093/biomet/57.1.1

Hinkley, D. V. (1971). Inference about the Change-Point from Cumulative Sum Tests. *Biometrika*, *58*(3), 509–523. https://doi.org/10.1093/biomet/58.3.509

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., & Tsai, T. T. (2005). An Algorithm for Optimal Partitioning of Data on an Interval. *IEEE Signal Processing Letters*, *12*(2), 105–108.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal Detection of Changepoints with a Linear Computational Cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

Kim, T., & Choi, J. (2015). Reading Documents for Bayesian Online Change Point Detection. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1610–1619. https://doi.org/10.18653/v1/D15-1184

Knoblauch, J., & Damoulas, T. (2018). Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection. *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research*, *80*, 2718–2727.

Knoblauch, J., & Lamperti, F. (2021). Pruning algorithms for online changepoint detection. *International Conference on Artificial Intelligence and Statistics*, 1465–1473.

Lavielle, M. (2005). Using Penalized Contrasts for the Change-Point Problem. *Signal Processing*, *85*(8), 1501–1510. https://doi.org/10.1016/j.sigpro.2005.01.012

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Lu, J., Henchion, M., & Mac Namee, B. (2020). Diverging Divergences: Examining Variants of Jensen Shannon Divergence for Corpus Comparison Tasks. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6740–6744.

Madsen, R. E., Kauchak, D., & Elkan, C. (2005). Modeling Word Burstiness Using the Dirichlet Distribution. *Proceedings of the 22nd International Conference on Machine Learning*, 545–552. https://doi.org/10.1145/1102351.1102420

Maidstone, R., Hocking, T., Rigaill, G., & Fearnhead, P. (2017). On Optimal Multiple Changepoint Algorithms for Large Data. *Statistics and Computing*, *27*(2), 519–533. https://doi.org/10.1007/s11222-016-9636-3

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Matteson, D. S., & James, N. A. (2014). A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, *109*(505), 334–345. https://doi.org/10.1080/01621459.2013.849605

Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, *41*(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

Rigaill, G. (2015). A Pruned Dynamic Programming Algorithm to Recover the Best Segmentations with 1 to K_ Change-Points. *Journal de La Société Française de Statistique*, *156*(4), 180–205.

Ruggieri, E., & Antonellis, M. (2016). An Exact Approach to Bayesian Sequential Change Point Detection. *Computational Statistics & Data Analysis*, *97*, 71–86. https://doi.org/10.1016/j.csda.2015.11.010

Saatçi, Y., Turner, R. D., & Rasmussen, C. E. (2010). Gaussian Process Change Point Models. *Proceedings of the 27th International Conference on Machine Learning*.

Tartakovsky, A., Nikiforov, I., & Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press.

Truong, C., Oudre, L., & Vayatis, N. (2018). *ruptures: Change Point Detection in Python*.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective Review of Offline Change Point Detection Methods. *Signal Processing*, *167*, 107299. https://doi.org/10.1016/j.sigpro.2019.107299

van den Burg, G. J. J., & Williams, C. K. I. (2020). *An Evaluation of Change Point Detection Algorithms*. https://arxiv.org/abs/2003.06222

Wang, Y., & Goutte, C. (2018). Real-Time Change Point Detection Using On-Line Topic Models. *Proceedings of the 27th International Conference on Computational Linguistics*, 2505–2515.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx