Knowledge-Based Systems 196 (2020) 105792

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An online Bayesian approach to change-point detection for categorical data

Yiwei Fan, Xiaoling Lu *

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

ARTICLE INFO

Article history:
Received 24 March 2019
Received in revised form 3 December 2019
Accepted 17 March 2020
Available online 29 March 2020

Keywords:
Bayes factor
Change-point detection
Dirichlet-multinomial mixtures
Online strategy

ABSTRACT

Change-point detection for categorical data has wide applications in many fields. Existing methods either are distribution-free, not utilizing categorical information sufficiently, or have limited performance when there exists "rare events" (events that occur with low frequency). In this paper, we propose a Bayesian change-point detection model for categorical data based on Dirichlet-multinomial mixtures. Because of the introduction of prior information, our method performs well for the existence of "rare events". An online parameter estimation procedure and an online detection strategy are then designed to adapt to data streams. Monte Carlo simulations discuss the power of the proposed method and show advantages compared with existing algorithms. Applications in biomedical research, document analysis, health news case study and location monitoring indicate practical values of our method.

1. Introduction

Change-point detection (CPD) is becoming increasingly popular in several fields including quality control [1], biomedical research [2], economic [3,4], text mining [5], signal segmentation [6] and business analysis [7]. Change-points are referred to abrupt changes of the underlying distribution in data over time, which represent transitions that occur between states. Anomaly detection [8] and concept drift [9] are similar concepts, which are used in literature as well as CPD. There are two aims in CPD that are generally achieved simultaneously, detecting whether there exists a change-point and estimating the moment of occurrence if it exists.

Categorical data is widespread in practice such as health-care-related environment [10], recommendation system [11], text [12,13] and image [14,15]. Some nonparametric and distribution-free methods can be applied to CPD for categorical data, while they cannot utilize categorical information sufficiently. Existing methods for CPD in categorical data usually have limited performance when there exists "rare events" (events that occur with low frequency). Taking the classical method, Pearson's chi-squared test, as an example, it requires a minimum number of counts for each category to achieve a convincing result. In this paper, we focus on CPD for categorical data. Specifically, we assume each data point is a high-dimensional vector of counts, where the length of the vector is the number of categories. Furthermore, each coordinate of the vector represents the counts for each category, and the sum is the total number of trials. An online Bayesian approach is then proposed. Different from existing methods, we assume categorical data comes from Dirichlet-multinomial mixtures [12,14,16], where the prior Dirichlet distribution is introduced to reduce sensitiveness to "rare events". Under this assumption, we formulate the hypothesis testing problem for CPD and define the Bayes factor as the test statistic.

In modern big data era, data often comes in the form of streams. Large volumes of streaming data are infeasible to be accommodated in the machine's main memory. Hence, an online procedure of CPD is desired. To deal with data streams where data arrives continuously, we apply an incremental expectation–maximization (EM) to estimate parameters in our Bayesian model and design an online strategy for CPD. Power analyses and Monte Carlo simulations are presented to show the effectiveness of our method. Moreover, the proposed model is compared with some competitive competitors to show great improvements. Lastly, we apply it to various scenarios in real world including biomedical research, document analysis, health news case study and location monitoring.

The remainder of this paper is organized as follows. In Section 2, we review related work of CPD. In Section 3, a Bayesian approach for CPD in categorical data is presented. In Section 4, an online estimation procedure and an online detecting strategy are designed for data streams. In Sections 5 and 6, we do some simulations and applications. Section 7 concludes.

* Corresponding author at: School of Statistics, Renmin University of China, Beijing, China
E-mail address: xiaolinglu@ruc.edu.cn (X. Lu).

https://doi.org/10.1016/j.knosys.2020.105792
0950-7051/

2. Related work

2.1. Change-point detection

Existing methods of CPD can be divided into two classes, supervised and unsupervised. The supervised CPD can be regarded as a simple binary classification problem. The classification methods in machine learning are widely used such as nearest neighbor [17], decision trees [18], and Naive Bayes [19].

Compared with supervised methods, more commonly considered in CPD is unsupervised learning, that is, training samples are unlabeled. One of the popular ideas is to transform CPD into a hypothesis testing problem. Methods based on hypothesis testing, mainly include sequential analysis and monitoring distributions on two different time-windows. Wald [20] proposed the sequential probability ratio test. The idea is to divide the sequence into two subsets that are generated from two unknown distributions, and then measure whether the probability of observing certain subsequences under one distribution is expected to be significantly higher than that under another distribution. The cumulative sum [21] uses the similar principle as that of the sequential probability ratio test, where the input for the test is a residual from any predictor and the output is an alarm when the mean of the incoming data significantly deviates from zero. Monitoring distributions on two different time-windows uses statistical tests with the null hypothesis stating that the distributions are equal. Dries and Ruckert [22] presented three novel drift detection tests on the multivariate data in data streams. The first one is based on a rank statistic on density estimates for a binary representation of the data, the second compares average margins of a linear classifier induced by the 1-norm support vector machine, and the last one is based on the average zero–one, sigmoid or stepwise linear error rate of an SVM classifier. Ada and Berthold [23] employed different window positioning strategies, where the two windows can be of equal or progressive sizes. Another standard statistical technique is statistical process control [24,25], which considers learning as a process and monitors the evolution of this process by defining the system to be in one of the three states: in-control, out-of-control and warning. Besides, Harries et al. [26] presented a contextual approach, where a time-stamp of the examples is taken as an input feature to a batch classifier. Similar ideas are also considered in Bouchachia [27]. In addition, researchers developed subspace identification methods [28,29], clustering algorithms [30], and graph-based approaches [31] to CPD. A more detailed review can be referred to Gama et al. [9] and Aminikhanghahi and Cook [32].

One can see most literature assumes observations to be normal or at least continuous, thus these methods cannot be used to solve CPD in categorical data. Though there are some nonparametric methods that are distribution-free, they do not utilize categorical information sufficiently. Adams and MacKay [33] proposed a Bayesian online approach to compute the posterior distribution of the length of the time since the last change-point, given the data so far observed, and use a simple message-passing algorithm. This method is highly modular and can be applied to a variety of types of data. Matteson and James [34] considered nonparametric estimation of both the number of change-points and the positions at which they occur based on hierarchical clustering without any distribution assumptions.

Focusing on categorical data, a common approach to CPD is to design a test statistic to perform hypothesis testing on two consecutive sequences under multinomial assumptions. Wolfe and Chen [35] developed three test statistics, assuming there is at most one change-point. Horváth and Serbinowska [36] presented a likelihood ratio test statistic and a chi-square test statistic to detect multiple change-points in categorical data. Based on this work, Batsidis et al. [37] developed phi-divergence test statistics, taking the likelihood ratio and the chi-square test statistics as special cases. These methods usually have a requirement of a minimum number of counts for each category and are sensitive to "rare events". Besides hypothesis testing formulation, Höhle [38] proposed the likelihood ratio based cumulative sum for CPD in categorical data, constructing a beta-binomial or a Dirichlet-multinomial regression model. It performs well when there are seasonal patterns in sequences but may fail if the regression assumption is not true. A more robust and flexible CPD algorithm for categorical data with the existence of "rare events" is then desired.

To overcome the drawbacks mentioned above, we introduce the prior Dirichlet distribution and assume categorical data comes from Dirichlet-multinomial mixtures. Introducing a prior distribution of parameters has been considered in literature. Son and Kim [39] used a Bayesian method to detect changes of mean, covariance, or both in a sequence of independent multivariate normal observations by introducing noninformative priors and several versions. Gupta and Baker [40] implemented a Bayesian change-point model to detect change-points in seismicity rates, where the events are assumed to belong to a Poisson process. A Gamma distribution is introduced as a prior distribution to help seismic hazard calculation as it allows more rigorous accounting of uncertainties.

2.2. Offline and online strategies

Strategies of CPD algorithms can be classified as "offline" and "online". Traditionally in data mining, already collected data is processed in an offline mode. Offline strategies widely adopted in literature require the entire sequence, and do retrospective detection. Binary segmentation is one of the most popular offline strategies [35–37]. It is a sequential approach: first, one change-point is detected in the whole sequence, then the sequence is split around this change-point, then the operation is repeated on the two resulting subsequences. The computational complexity of binary segmentation is $\mathcal{O}(n \log n)$, where n is the length of the sequence. As n increases, the computational cost increases rapidly.

When data comes in the form of streams, the requirement of the whole sequence makes offline strategies useless. In recent literature, online strategies, which run concurrently with a data stream, processing each data item when it arrives and detecting a change-point as soon as possible after it occurs, are more popular. Online strategies are various and adaptive to specific CPD models. Adams and MacKay [33] computed the probability distribution of the length of the current "run" and updated it after a new data item arrived, which drops to zero when a change-point occurs. Some other researchers compared two sets of descriptors extracted online from the sequence at each moment: the immediate past set and the immediate future set, and decided whether the moment is a change-point [13,41]. Because of the saving of computational time and storage, online strategies receive more attention in today's big data era and are what we concern in this paper.

3. Mathematical model for Bayesian change-point detection

3.1. Change-point detection problem

Denote the data stream as $X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$, where $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iK})^\top$, $i = 1, 2, \ldots$ are independent high-dimensional variables with length K. Our goal is to detect the occurrence of a change-point in X. Before proceeding, we first give the definition of a change-point, which is same as that in most literature [9,11,13,32,37].

Definition 1 (*Change-point*)**.** A change-point represents a transition between different distributions that generate the data over time.

Denote the subsequence of X starting at time s with length m as $\mathscr{Y}(s,m)=\{\boldsymbol{X}_s,\boldsymbol{X}_{s+1},\ldots,\boldsymbol{X}_{s+m-1}\}$. All subsequences of length m in X can be obtained by moving a sliding window with width m, and the set is denoted as $Y=\{\mathscr{Y}(1,m),\mathscr{Y}(2,m),\ldots\}$. According to the work of Liu et al. [13], in order to sufficiently utilize the information contained in the consecutive time periods, the set Y is considered in the following, rather than the original observed stream X. Fig. 1 shows the illustration of notations.

To detect the occurrence of a change-point in X at $t, t=m+1,\ldots$, we consider the subsequences of two consecutive time periods, $\mathscr{Y}(t-m,m)$, $\mathscr{Y}(t,m)$. Then the CPD can be converted into a hypothesis testing, that is, whether $\mathscr{Y}(t-m,m)$, $\mathscr{Y}(t,m)$ come from a same distribution. As shown in literature [32,37], we have the following definition.

Definition 2 (*CPD*)**.** Denote the probability distribution of the subsequence $\mathscr{Y}(s,m)$ as $\mathbb{P}_{s,m}$, $s=1,\ldots$. The CPD problem at t can be defined as the following hypothesis testing problem with the null hypothesis being no change-point occurs and the alternative hypothesis being a change-point occurs,

$$\begin{aligned}\boldsymbol{H}_0:&\quad \mathbb{P}_{t-m,m}=\mathbb{P}_{t,m},\\ \boldsymbol{H}_1:&\quad \mathbb{P}_{t-m,m}\neq\mathbb{P}_{t,m}.\end{aligned}\tag{1}$$

3.2. Dirichlet-multinomial mixtures

If $\boldsymbol{X}_i=(x_{i1},x_{i2},\ldots,x_{iK})^\top$, $i=1,2,\ldots$ is categorical, where K is the number of categories, assume $\mathscr{Y}(s,m)$, $s=1,\ldots$ is from multinomial mixtures, consisting of J components, that is, $c_j, j=1,\ldots,J$. The corresponding hidden variable of $\mathscr{Y}(s,m)$ is $\boldsymbol{\gamma}_s=(\gamma_{s1},\ldots,\gamma_{sJ})^\top$, where

$$\gamma_{sj}=\begin{cases}1, & \mathscr{Y}(s,m)\text{ comes from }c_j,\\ 0, & \text{otherwise.}\end{cases}$$

Note that $\sum_{j=1}^{J}\gamma_{sj}=1$. Since $\boldsymbol{X}_i, i=s,\ldots,s+m-1$ are independent, the probability of $\mathscr{Y}(s,m)$ can be calculated as,

$$\begin{aligned}Pr(\mathscr{Y}(s,m)|\boldsymbol{p},\boldsymbol{\beta})&=\sum_{\boldsymbol{\gamma}_s}Pr(\boldsymbol{\gamma}_s|\boldsymbol{p})Pr(\mathscr{Y}(s,m)|\boldsymbol{\gamma}_s,\boldsymbol{\beta})\\&=\left(\prod_{i=s}^{s+m-1}\frac{\Gamma(n_i+1)}{\prod_{k=1}^{K}\Gamma(x_{ik}+1)}\right)\sum_{j=1}^{J}p_j\prod_{k=1}^{K}\beta_{kj}^{N_{s,m,k}},\end{aligned}\tag{2}$$

where $n_i=\sum_{k=1}^{K}x_{ik}$ is the number of trials in $\boldsymbol{X}_i$ for $i=s,\ldots,s+m-1$, $\boldsymbol{p}=(p_1,\ldots,p_J)^\top$ is a vector of length J with p_j representing the probability that $\gamma_{sj}=1$, $\boldsymbol{\beta}$ is a matrix of $K\times J$ with β_{kj} representing the probability that the kth category occurs in c_j, and $N_{s,m,k}=\sum_{i=s}^{s+m-1}x_{ik}$.

The maximum likelihood estimation of the parameters in (2) can be obtained by the expectation–maximization (EM) algorithm and its extensions. As described in Bouguila [14], there is a problem of sparsity of the estimators, which makes the detection result sensitive to "rare events". A common way to solve the sparsity is to smooth $\boldsymbol{\beta}$ by replacing β_{kj} with a linear combination, that is,

$$(1-\lambda)\beta_{kj}+\lambda\frac{\sum_s N_{s,m,k}}{\sum_s\sum_{k=1}^{K}N_{s,m,k}},$$

where λ represents the smoothing factor and is a parameter to be selected. A more flexible way to address the sparsity is to introduce the Dirichlet distribution with parameters $\boldsymbol{\alpha}_j=(\alpha_{1j},\ldots,\alpha_{Kj})^\top$ as a prior distribution for $\boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_j=(\beta_{1j},\ldots,\beta_{Kj})$ represents the jth column of $\boldsymbol{\beta}$. The probability density function is

$$P(\boldsymbol{\beta}_j|\boldsymbol{\alpha}_j)=\frac{\Gamma(\sum_{k=1}^{K}\alpha_{kj})}{\prod_{k=1}^{K}\Gamma(\alpha_{kj})}\prod_{k=1}^{K}\beta_{kj}^{\alpha_{kj}-1}.$$

From the Bayes formula, we have

$$\begin{aligned}Pr(\mathscr{Y}(s,m)|\boldsymbol{p},\boldsymbol{\alpha})&=\sum_{j=1}^{J}p_j\int Pr(\mathscr{Y}(s,m)|\boldsymbol{\gamma}_s,\boldsymbol{\beta}_j)P(\boldsymbol{\beta}_j|\boldsymbol{\alpha}_j)d\boldsymbol{\beta}_j\\&=\left(\prod_{i=s}^{s+m-1}\frac{\Gamma(n_i+1)}{\prod_{k=1}^{K}\Gamma(x_{ik}+1)}\right)\sum_{j=1}^{J}p_j\frac{\Gamma(\sum_{k=1}^{K}\alpha_{kj})}{\Gamma(\sum_{k=1}^{K}N_{s,m,k}+\sum_{k=1}^{K}\alpha_{kj})}\\&\quad\prod_{k=1}^{K}\frac{\Gamma(N_{s,m,k}+\alpha_{kj})}{\Gamma(\alpha_{kj})},\end{aligned}\tag{3}$$

where $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_J)$ is the matrix of prior parameters.

3.3. Bayes factor

The Bayes factor, proposed by Jeffreys [42] and developed by Kass and Raftery [43], is a method for the hypothesis testing based on Bayesian theories. Since the Bayes factor introduces the prior information, it is more robust than traditional methods that estimate probabilities directly by observational frequencies. The Bayes factor calculates the ratio of the likelihood when the alternative hypothesis and the null hypothesis are established respectively, measuring the relative support of the evidence. The Bayes factor of the hypothesis testing (1) is defined as follows,

$$\begin{aligned}B_{10}(t)&:=\frac{Pr(\mathscr{Y}(t-m,m),\mathscr{Y}(t,m)|\boldsymbol{H}_1)}{Pr(\mathscr{Y}(t-m,m),\mathscr{Y}(t,m)|\boldsymbol{H}_0)}\\&=\frac{Pr(\mathscr{Y}(t-m,m)|\boldsymbol{p},\boldsymbol{\alpha})Pr(\mathscr{Y}(t,m)|\boldsymbol{p},\boldsymbol{\alpha})}{Pr(\mathscr{Y}(t-m,m),\mathscr{Y}(t,m)|\boldsymbol{p},\boldsymbol{\alpha})},\end{aligned}\tag{4}$$

where $Pr(\mathscr{Y}(t-m,m)|\boldsymbol{p},\boldsymbol{\alpha})$, $Pr(\mathscr{Y}(t,m)|\boldsymbol{p},\boldsymbol{\alpha})$ and $Pr(\mathscr{Y}(t-m,m),\mathscr{Y}(t,m)|\boldsymbol{p},\boldsymbol{\alpha})$ are calculated by (3) accordingly.

It should be noted that

$$\begin{aligned}&\left(\prod_{i=t-m}^{t-1}\frac{\Gamma(n_i+1)}{\prod_{k=1}^{K}\Gamma(x_{ik}+1)}\right)\cdot\left(\prod_{i=t}^{t+m-1}\frac{\Gamma(n_i+1)}{\prod_{k=1}^{K}\Gamma(x_{ik}+1)}\right)\\&=\left(\prod_{i=t-m}^{t+m-1}\frac{\Gamma(n_i+1)}{\prod_{k=1}^{K}\Gamma(x_{ik}+1)}\right).\end{aligned}$$

So, (4) is equivalent to

$$B_{10}(t)=\frac{b_{t-m,m}b_{t,m}}{b_{t-m,2m}},\tag{5}$$

where

$$\begin{cases}b_{t-m,m}=\sum_{j=1}^{J}p_j\frac{\Gamma(\sum_{k=1}^{K}\alpha_{kj})}{\Gamma(\sum_{k=1}^{K}N_{t-m,m,k}+\sum_{k=1}^{K}\alpha_{kj})}\prod_{k=1}^{K}\frac{\Gamma(N_{t-m,m,k}+\alpha_{kj})}{\Gamma(\alpha_{kj})},\\ b_{t,m}=\sum_{j=1}^{J}p_j\frac{\Gamma(\sum_{k=1}^{K}\alpha_{kj})}{\Gamma(\sum_{k=1}^{K}N_{t,m,k}+\sum_{k=1}^{K}\alpha_{kj})}\prod_{k=1}^{K}\frac{\Gamma(N_{t,m,k}+\alpha_{kj})}{\Gamma(\alpha_{kj})},\\ b_{t-m,2m}=\sum_{j=1}^{J}p_j\frac{\Gamma(\sum_{k=1}^{K}\alpha_{kj})}{\Gamma(\sum_{k=1}^{K}N_{t-m,2m,k}+\sum_{k=1}^{K}\alpha_{kj})}\prod_{k=1}^{K}\frac{\Gamma(N_{t-m,2m,k}+\alpha_{kj})}{\Gamma(\alpha_{kj})}.\end{cases}$$

Similar to the likelihood ratio test statistic, we take the twice of the logarithm of the Bayes factor, that is,

$$2\ln(B_{10}(t))=2(LS_1(t)-LS_0(t)),\tag{6}$$

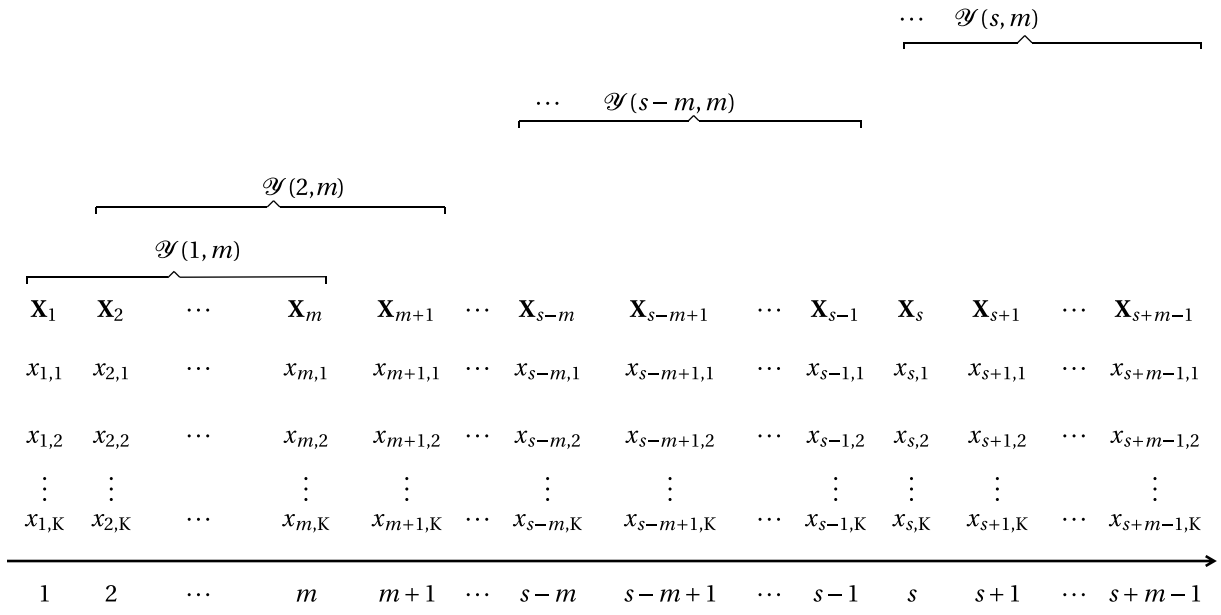

Fig. 1. Illustration of notations.

Table 1
Empirical scales of the Bayes factor.

$2\ln(B_{10}(t))$	$B_{10}(t)$	Evidence against $\boldsymbol{H}_0$
$0-2$	$1-3$	Not worth more than a bare mention
$2-6$	$3-20$	Positive
$6-10$	$20-150$	Strong
>10	>150	Very strong

where $LS_i(t) = \log Pr(\mathscr{Y}(t-m,m), \mathscr{Y}(t,m)|\boldsymbol{H}_i)$, $i = 0, 1$ denotes the log-likelihood of samples $\mathscr{Y}(t-m,m), \mathscr{Y}(t,m)$ under the assumption $\boldsymbol{H}_i$. Based on (6), the Bayes factor can be regarded as the difference between the log-likelihood of samples under the alternative hypothesis and the null hypothesis. As discussed in Kass and Raftery [43], when $2\ln(B_{10}(t))$ is sufficiently large, the log-likelihood of samples under the alternative hypothesis is much larger than that under the null hypothesis. We then conclude that the evidence is strong against $\boldsymbol{H}_0$, which means a change-point occurs. In general, the asymptotic distribution of the Bayes factor (5) under the null hypothesis does not have an explicit expression, thus the power of the testing problem (1) cannot be calculated exactly. Kass and Raftery [43] provided empirical scales of the Bayes factor to infer the hypothesis testing, as shown in Table 1.

3.3.1. Power analysis

In the traditional hypothesis testing, we generally discuss the extent to which the observational data rejects the null hypothesis and care about two types of errors, the first type of error $Pr(\text{reject } \boldsymbol{H}_0|\boldsymbol{H}_0)$ and the second type of error $Pr(\text{do not reject } \boldsymbol{H}_0|\boldsymbol{H}_1)$. The two types of errors indicate the probability that data rejects the null hypothesis when $\boldsymbol{H}_0$ established and does not reject the null hypothesis when the alternative hypothesis established. However, the situation is a bit different for the Bayes factor. The Bayes factor can provide the degree to which the data supports the null hypothesis [43] and it is not sufficient to just control the two errors as we show in the following. Before proceeding, we first define the following events [44]:

- Strong evidence: the data supports either the null hypothesis or the alternative hypothesis;
- Weak evidence: the data is not sufficient to support any of the hypotheses.

According to the discussion above, when $B_{10}(t)$ is sufficiently large, the data supports the alternative hypothesis; when $B_{10}(t)$ is sufficiently small, the data supports the null hypothesis. Denote η as a predetermined threshold, then the set of moments with strong evidence can be described as $D := \{t : 2\ln(B_{10}(t)) > \eta \text{ or } 2\ln(B_{10}(t)) < \eta^{-1}\}$. Similarly, the set of moments with weak evidence can be expressed as $W := \{t : \eta^{-1} < 2\ln(B_{10}(t)) < \eta\}$. If $t \in W$, the observational data cannot support any hypothesis, that is, the testing problem (1) has no solution. Therefore, in Bayesian hypothesis testing, besides controlling the probability of two types of errors, we also need to control the probability of weak evidence. It holds that

$$Pr(\boldsymbol{H}_0|\boldsymbol{H}_0) = 1 - Pr(\boldsymbol{H}_1|\boldsymbol{H}_0) - Pr(\eta^{-1} < 2\ln(B_{10}(t)) < \eta|\boldsymbol{H}_0), \quad (7)$$

$$Pr(\boldsymbol{H}_1|\boldsymbol{H}_1) = 1 - Pr(\boldsymbol{H}_0|\boldsymbol{H}_1) - Pr(\eta^{-1} < 2\ln(B_{10}(t)) < \eta|\boldsymbol{H}_1). \quad (8)$$

It can be seen that, unlike the traditional hypothesis testing, because of the existence of weak events, a small probability of the error does not lead to a high probability that the data supports the true hypothesis. Therefore, instead of controlling these two types of errors, we should directly control $Pr(\boldsymbol{H}_0|\boldsymbol{H}_0) = Pr(2\ln(B_{10}(t)) < \eta^{-1}|\boldsymbol{H}_0)$ and $Pr(\boldsymbol{H}_1|\boldsymbol{H}_1) = Pr(2\ln(B_{10}(t)) > \eta|\boldsymbol{H}_1)$. Assume we guarantee $Pr(\boldsymbol{H}_i|\boldsymbol{H}_i) \geq \xi_i, i = 0, 1$ where $0 \leq \xi_i \leq 1$, then by (7) and (8), we have

$$Pr(\boldsymbol{H}_j|\boldsymbol{H}_i) + Pr(\eta^{-1} < 2\ln(B_{10}(t)) < \eta|\boldsymbol{H}_i) \leq 1 - \xi_i,$$
$$i, j = 0, 1, j \neq i.$$

Therefore, it holds that $Pr(\boldsymbol{H}_1|\boldsymbol{H}_0) \leq 1 - \xi_0$ and $Pr(\boldsymbol{H}_0|\boldsymbol{H}_1) \leq 1 - \xi_1$, and the probability of weak evidence satisfies $Pr(\eta^{-1} < 2\ln(B_{10}(t)) < \eta|\boldsymbol{H}_i) \leq 1 - \xi_i, i = 0, 1$. When $\xi_i, i = 0, 1$ is sufficiently large, the two types of errors and the probability of weak evidence can be controlled. In Section 5, we will use the Monte Carlo simulation to further discuss the power of the Bayes factor in detail.

4. Online change-point detection

4.1. Online parameter estimation

Estimating the value of parameter $\boldsymbol{p}$ and $\boldsymbol{\alpha}$ in the Bayes factor (5) is a tricky problem. The optimization formula can be

expressed as follows,

$$\max \quad \ell(\boldsymbol{p}, \boldsymbol{\alpha}) = \sum_{s} \log Pr(\mathscr{Y}(s, m)|\boldsymbol{p}, \boldsymbol{\alpha}), \tag{9}$$

$$s.t. \quad \sum_{j=1}^{J} p_j = 1, p_j > 0, j = 1, \ldots, J, \tag{10}$$

$$\alpha_{kj} > 0, j = 1, \ldots, J, k = 1, \ldots, K. \tag{11}$$

The difficulty in solving the optimization problem (9)–(11) is the existence of latent variables $\boldsymbol{\gamma}_s, s = 1, 2, \ldots$. The expectation–maximization (EM) algorithm, widely used in missing data, censored observations and mixture distributions [45,46], is considered for our model. The EM algorithm has two main steps, E step and M step. E step is used to calculate the conditional expectation based on the current estimation. M step is used to maximize the expectation in the E step and update estimators. The estimators updated in the M step are used for the calculation in the next E step. The two steps are alternately performed until the predetermined convergence condition is satisfied. The batch EM algorithm for solving the optimization problem (9)–(11) is described as follows.

Step1. *Initialization.* Set $i = 0$ and give $\boldsymbol{p}^{(0)}, \boldsymbol{\alpha}^{(0)}$ in advance.

Step2. *Iteration.* Alternate the following E step and M step until the convergence condition satisfies.

E step. Denote $\boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}$ as the estimators in the ith iteration, then calculate the conditional expectation

$$E_{\boldsymbol{\gamma}}\left[\log Pr(Y, \boldsymbol{\gamma}|\boldsymbol{p}, \boldsymbol{\alpha})|Y, \boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}\right].$$

M step. Maximize the expectation with specific constraints, that is, solving the following optimization problem,

$$\begin{aligned} &\left(\boldsymbol{p}^{(i+1)}, \boldsymbol{\alpha}^{(i+1)}\right) \\ &= \underset{\boldsymbol{p},\boldsymbol{\alpha}}{\operatorname{argmax}} E_{\boldsymbol{\gamma}}\left[\log Pr(Y, \boldsymbol{\gamma}|\boldsymbol{p}, \boldsymbol{\alpha})|Y, \boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}\right], \\ s.t. \quad &\sum_{j=1}^{J} p_j = 1, p_j > 0, j = 1, \ldots, J, \\ &\alpha_{kj} > 0, j = 1, \ldots, J, k = 1, \ldots, K. \end{aligned}$$

Check whether the convergence condition satisfies. If not, set $i = i + 1$ and go back to E step.

Note that the batch EM algorithm requires the whole dataset to estimate parameters, which results in a large amount of calculation and is not proper for data streams. An online parameter estimation procedure is desired. Considering the key idea of the EM algorithm which is to guarantee the monotonic improvement of the log-likelihood, an incremental EM algorithm is conducted to our model, where estimators are updated when a new data item arrives [47]. The incremental EM is described as follows.

Step1. *Initialization.* Set $i = 0$ and give $a_{s,j}^{(0)}(j = 1, \ldots, J, s = 1, 2, \ldots), \boldsymbol{p}^{(0)}, \boldsymbol{\alpha}^{(0)}$ in advance.

Step2. *Iteration.* Alternate the following E step and M step until the convergence condition satisfies.

E step. Denote $a_{s,j}^{(i)}(j = 1, \ldots, J, s = 1, 2, \ldots), \boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}$ as the update value in the ith iteration, then calculate

$$\begin{aligned} a_{s,j}^{(i+1)} &= a_{s,j}^{(i)}, \quad s \neq i+1, \\ a_{i+1,j}^{(i+1)} &= Pr\left(\gamma_{i+1,j} = 1|\mathscr{Y}(i+1, m), \boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}\right) \\ &= \frac{p_j^{(i)} Pr(\mathscr{Y}(i+1, m)|\gamma_{i+1,j} = 1, \boldsymbol{\alpha}^{(i)})}{\sum_{l=1}^{J} p_l^{(i)} Pr(\mathscr{Y}(i+1, m)|\gamma_{i+1,l} = 1, \boldsymbol{\alpha}^{(i)})}. \end{aligned} \tag{12}$$

The conditional expectation is,

$$\begin{aligned} &E_{\boldsymbol{\gamma}}\left[\sum_{s} \log Pr(\mathscr{Y}(s, m), \boldsymbol{\gamma}_s|\boldsymbol{p}, \boldsymbol{\alpha})|Y, \boldsymbol{p}^{(i)}, \boldsymbol{\alpha}^{(i)}\right] \\ &\propto \sum_{s=1}^{i+1} \sum_{j=1}^{J} a_{s,j}^{(i+1)} \log p_j Pr(\mathscr{Y}(s, m)|\gamma_{sj} = 1, \boldsymbol{\alpha}_j). \end{aligned} \tag{13}$$

M step. Maximize (13) under constraints (10)(11). For $j = 1, \ldots, J, k = 1, \ldots, K$, we have

$$p_j^{(i+1)} = \frac{\sum_{s=1}^{i+1} a_{s,j}^{(i+1)}}{i+1} = \frac{i \cdot p_j^{(i)} + a_{i+1,j}^{(i+1)}}{i+1}.$$

Thus, the update of $\boldsymbol{p}^{(i+1)}$ only involves simple calculation. The solution of $\boldsymbol{\alpha}^{(i+1)}$ is relatively complicated, and has no analytical expression. We use the sequential least squares programming (SLSQP) algorithm to solve it, which is based on the Han–Powell quasi-Newton algorithm and can solve nonlinear optimization problems with constraints [48].

Check whether the convergence condition satisfies. If not, set $i = i + 1$ and go back to E step.

It can be seen that in the E step of each iteration, we only need to compute the expected value $a_{i+1,j}^{(i+1)}, j = 1, \ldots, J$ of the newly arriving data item. The updating of $\boldsymbol{p}$ is just a simple operation, though the updating of $\boldsymbol{\alpha}$ is a little more complicated.

In addition to parameters $\boldsymbol{p}, \boldsymbol{\alpha}$, the number of components J, is remained to be determined. The value of J can be given in advance based on some prior information. In the absence of any prior information, the BIC criterion can be used to select the optimal J, that is,

$$BIC(J) = -2\ell(\boldsymbol{p}, \boldsymbol{\alpha}) + |\Omega| \cdot \log(i), \tag{14}$$

where $|\Omega| = J(K+1) - 1$ represents the number of parameters to be estimated in the model and i is the number of data items used to estimate parameters in the incremental EM algorithm. The BIC criterion balances the log-likelihood and the model complexity, and can be used to select the optimal model. The smaller the BIC value is, the better the model achieves. Given the candidate set $\mathcal{J}$, the optimal J is the one that takes the smallest BIC, that is,

$$J^* = \underset{J \in \mathcal{J}}{\operatorname{argmin}} BIC(J). \tag{15}$$

The detailed steps of the estimation procedure are shown in Algorithm 1.

4.2. Online detection strategy

As we discussed in Section 2.2, online CPD strategies are popular nowadays. To determine whether there is a change-point, new data items are compared with old ones. Thus, perfect real time operation, that is, detecting change-points before the new data

Algorithm 1: Online parameter estimation using the incremental EM.

Input: $\mathcal{J}$, $a_{s,j}^{(0)}(j=1,\ldots,J, s=1,2,\ldots)$, $\boldsymbol{p}^{(0)}$, $\boldsymbol{\alpha}^{(0)}$, ϵ
Output: J^*, $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{\alpha}}$
for $J \in \mathcal{J}$ **do**
  1. Let $i=0$, $\boldsymbol{p}^{(-1)}=(+\infty,\cdots,+\infty)$, $\boldsymbol{\alpha}^{(-1)}=(+\infty)_{K\times J}$.
  while $|\boldsymbol{p}^{(i)}-\boldsymbol{p}^{(i-1)}|_\infty > \epsilon$ *or* $|\boldsymbol{\alpha}^{(i)}-\boldsymbol{\alpha}^{(i-1)}|_\infty > \epsilon$ **do**
    2. Update $a_{s,j}^{(i+1)}(j=1,\cdots,J, s=1,2,\cdots)$ according to (12).
    3. Calculate (13).
    4. Maximize (13) with constraints (10)(11), and update estimators $\boldsymbol{p}^{(i+1)}$, $\boldsymbol{\alpha}^{(i+1)}$.
  end
  5. Calculate $BIC(J)$ according to (14).
end
6. Based on (15), select the optimal J^*.
7. Output $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{\alpha}}$ corresponding to J^*.

item arrives, cannot be achieved by any change-point detection algorithm [32]. According to notations in Aminikhanghahi and Cook [32], an algorithm is defined as an θ real time algorithm if it needs at least θ data items to detect occurrence of change-points. An offline algorithm can be viewed as $+\infty$ real time. A smaller value of θ results in a more powerful and more responsive detection algorithm.

In Section 3.1, the incremental EM is used to do online parameter estimation until the convergence conditions are satisfied. Denote the number of steps achieved convergence as i. The parameter estimation process requires subsequences $\{\mathscr{Y}(1,m),\ldots,\mathscr{Y}(i+1,m)\}$, which is called a burn-in period [4]. After estimating parameters, an online detection strategy for our model is presented. For the moment $t \geq \max\{m+1, i+1\}$, to detect whether there is a change-point, we consider the immediate past subsequence $\mathscr{Y}(t-m,m)$ and the immediate future subsequence $\mathscr{Y}(t,m)$ to calculate $B_{10}(t)$ as (5). Then we solve the hypothesis testing problem (1) based on the empirical scale shown in Table 1. If the evidence does not against the null hypothesis $\boldsymbol{H}_0$, we conclude there is no change-point; otherwise, a change-point occurs. Note that the previous data is not required except for the last subsequence $\mathscr{Y}(t-m,m)$ and it needs m data items in $\mathscr{Y}(t,m)$ to detect the occurrence of a change-point at moment t. Thus our proposed strategy is a m real time online algorithm. We only need to store a subsequence of length $2m$ in the machine's main memory, saving computational resources significantly. Moreover, the change-point occurrence in the burn-in period can be detected using the retrospective strategy after estimating parameters. The whole procedure for our model is shown in Fig. 2.

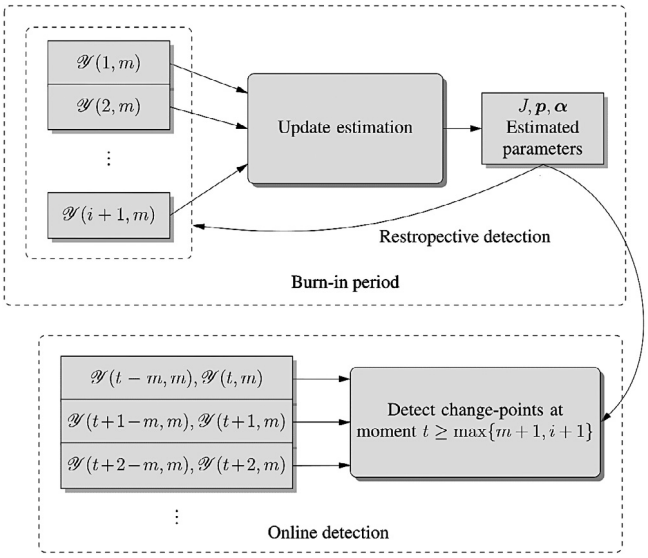

Fig. 2. Procedure of change-point detection.

5. Simulations

In this section, we first use Monte Carlo simulations to analyze the power of our proposed Bayes factor and then compare with other existing detection methods. In the following, we set $\eta = 2$ as the threshold value according to Table 1.

5.1. Monte Carlo power analysis

Generally, the asymptotic distribution of the Bayes factor $B_{10}(t)$ under the null hypothesis has no explicit expression. In this subsection, we do numerical simulations to approximate $Pr(\boldsymbol{H}_i|\boldsymbol{H}_i), i=0,1$ via Monte Carlo and give our suggestion of the choice of the sliding window width m.

(Data generating process) The simulation data is generated as follows. Consider a categorical sequence denoted as $X = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$. The corresponding latent variables are denoted as $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots\}$, where $\boldsymbol{\gamma}_i \sim Multi(1, \boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \ldots, p_J)$ and J being the number of components. Let

$$\begin{cases} p_j \sim Unif\left(0, \frac{1}{J-1}\right), j=1,\ldots,J-1, \\ p_J = 1 - \sum_{j=1}^{J-1} p_j. \end{cases}$$

Then $\boldsymbol{X}_i \sim Multi(n_i, \boldsymbol{\beta}\boldsymbol{\gamma}_i)$, where $\boldsymbol{\beta}$ is the $K \times J$ parameter matrix from Dirichlet priors according to our assumption, that is, $\boldsymbol{\beta} \sim Dirichlet(\boldsymbol{\alpha})$. For $j=1,\ldots,J$, let

$$\begin{cases} \alpha_{kj} \sim Unif(0,1), & k \neq (j \bmod K), \\ \alpha_{kj} = 5, & \text{otherwise}, \end{cases}$$

which means in the jth component, the $(j \bmod K)$th category is the dominant one and others are "rare events". The occurrence of a change-point at the moment t is marked as $\boldsymbol{\gamma}_{i-1} \neq \boldsymbol{\gamma}_i, i=1,2,\ldots$. Let $d_0 = 0, d_l \sim Poisson(20), l=1,2,\ldots$ and $\boldsymbol{\gamma}_{\sum_{i=1}^{l} d_{i-1}+1} = \boldsymbol{\gamma}_{\sum_{i=1}^{l} d_{i-1}+2} = \cdots = \boldsymbol{\gamma}_{\sum_{i=1}^{l+1} d_{i-1}}$.

To compute the power $Pr(\boldsymbol{H}_i|\boldsymbol{H}_i), i=0,1$, denote the set of true change-points as $\mathcal{T}$ and the set of non-change-points as $\mathcal{N}$. The power can be approximated as

$$Pr(\boldsymbol{H}_0|\boldsymbol{H}_0) = \frac{\sum_{s\in\mathcal{N}} 2\ln(B_{10}(s)) < \eta^{-1}}{|\mathcal{N}|},$$
$$Pr(\boldsymbol{H}_1|\boldsymbol{H}_1) = \frac{\sum_{s\in\mathcal{T}} 2\ln(B_{10}(s)) > \eta}{|\mathcal{T}|}.$$

Simulation 1. In this simulation, we show the power of the Bayes factor and explore the effect that the width of the sliding window has on the power. The data is generated as the data generating process. Set $K = 10, 20, 30, 40$, $m = 1,2,3,4,5,6,7,8$ and $n_i \sim Poisson(15), Poisson(20), Poisson(25), i=1,2,\ldots$. Thus, the sample size is small compared with the number of categories and there exists "rare events". Consider the first 1000 observations in the data stream. The average results over 100 replications are shown in Fig. 3.

The results show that $Pr(\boldsymbol{H}_0|\boldsymbol{H}_0)$ generally increases when K increases and decreases with n_i increasing. The trend is opposite for $Pr(\boldsymbol{H}_1|\boldsymbol{H}_1)$. It is consistent with the knowledge that the detection rate $Pr(\boldsymbol{H}_1|\boldsymbol{H}_1)$ will be lower for larger K and smaller sample size n_i. Considering the sliding window width, it can be seen that when the sliding window width m is small, the detection rate $Pr(\boldsymbol{H}_1|\boldsymbol{H}_1)$ is unsatisfying; when m is large, the detection accuracy

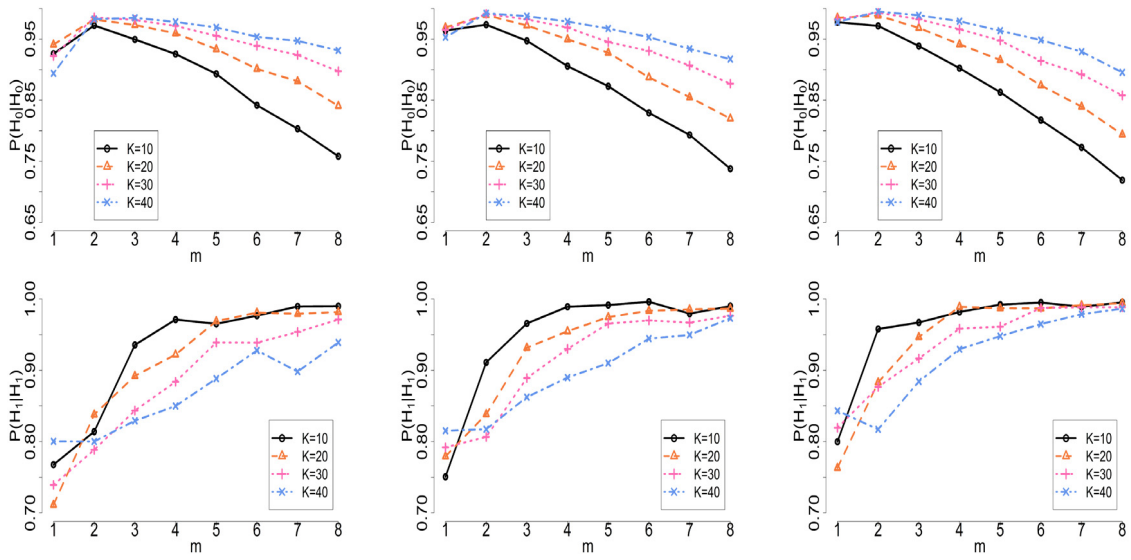

Fig. 3. Average results for Simulation 1 over 100 replications with $n_i \sim Poisson(15)$ (left panel), $n_i \sim Poisson(20)$ (middle panel), $n_i \sim Poisson(25)$ (right panel).

$Pr(\boldsymbol{H}_0|\boldsymbol{H}_0)$ is limited. We need to find a balance between these two powers. A basic rule is to set larger m for larger K and smaller n_i. In this case, we recommend $m = 4$ for $K = 10, 20$ and $m = 7$ for $m = 30, 40$.

5.2. Comparison

In this subsection, we compare the proposed method, Bayes factor (BF), with the following representative detection algorithms:

(1) the traditional chi-square test (CHI), where the significant level is set as 0.01 [37];
(2) a nonparametric multiple change-points detection model that is distribution free (ECP) [34], where the maximum number of change-points is set as 30 and the window size used to calculate between-segment distances is set as m;
(3) an online Bayesian change-points detector computing the probability of the length of the current "run" (BCPD) [33];
(4) a Bayesian change-point model that uses event occurrence data to indicate whether a change in event rates occurred, assuming that the event occurrences belong to a Poisson distribution (POI) [40]; (The method is designed for univariate time series with at most one change-point. We implement POI over K data streams for multi-dimensional data. Besides, a sliding window with size m is applied for multiple change-points. The union set of change-points over K runs is chosen as the final detection result.)
(5) the pruned exact linear time method (PELT) [49]; (The method is currently designed for univariate time series, thus PELT is implemented over K data streams for multi-dimensional data and the union set of change-points is chosen as the final detection result.)

In simulations, where the true change-points are known, the CPD can be seen as a binary classification problem. Common evaluation measures of the binary classification can be applied to it, such as precision (P), recall (R) and F-value (F) [4,13]. Specifically for CPD, we have the following definitions,

- P: the proportion of detections that are not false detections;
- R: the proportion change-points that are correctly detected.

Note that to avoid duplication, peaks of a change-point statistic are regarded as detection alarms. Denote $\hat{\mathcal{T}}$ as the set of detected change-points, then we have

$$\mathrm{P} = \frac{|\mathcal{T} \cap \hat{\mathcal{T}}|}{|\hat{\mathcal{T}}|}, \quad \mathrm{R} = \frac{|\mathcal{T} \cap \hat{\mathcal{T}}|}{|\mathcal{T}|}, \quad \mathrm{F} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}.$$

Simulation 2. In this simulation, we set $K = 10, 20, 30, 40$, $n_i \sim Poisson(15)$, $n_i \sim Poisson(20)$, $n_i \sim Poisson(25)$, $i = 1, 2, \ldots$, $m = 4$ for $K = 10, 20$ and $m = 7$ for $K = 30, 40$ as discussed in Section 4.1. The data is generated as the data generating process. Consider the first 1000 observations in the data stream. The average results as well as the standard errors over 100 replications are shown in Table 2.

From Table 2, in all situations, our method performs much better than other algorithms in precision and F-value, and is comparable in recall. As comparisons, the traditional method, CHI, performs limited since the requirement of the minimum sample size is not satisfied in the simulation. ECP, BCPD and PELT that are distribution free, cannot utilize the categorical information sufficiently, thus achieve unsatisfying values. Besides, though POI, designed for univariate data which follows a Poisson distribution, has a high value of recall, the precision value is quite low. We can conclude that the proposed method, BF, outperforms these existing detectors in our simulation.

For each evaluation metric, Fig. 4 shows the average values along with the confidence intervals (MEAN$\pm 2\times$SE) for the proposed method, BF. It can be seen that for our method, given K, all evaluation metrics are non-decreasing with n_i increasing. In general, the value increases significantly (except for the recall with $K = 10, 30$ and the F-value with $K = 30$) when n_i increases from 15 to 20 as there is no overlap between the confidence intervals. When n_i increases from 20 to 25, the value is non-decreasing and gradually close to 1 (the ideal value).

6. Applications

In this section, we apply our model to four datasets in the fields of biomedicine research, document analysis, health news case study and location monitoring, where the first two have been considered extensively in literature [36–38,50], the third one is public on the University of California at Irvine and the fourth one is collected by authors. In the following, we set $\eta = 2$.

Table 2
Average results as well as standard errors for Simulation 2 over 100 replications.

	$n_i \sim Poisson(15)$			$n_i \sim Poisson(20)$			$n_i \sim Poisson(25)$		
	P	R	F	P	R	F	P	R	F
	$K = 10$								
CHI	$0.139_{0.005}$	$0.586_{0.014}$	$0.222_{0.007}$	$0.147_{0.005}$	$0.650_{0.016}$	$0.240_{0.008}$	$0.145_{0.005}$	$0.650_{0.013}$	$0.235_{0.007}$
ECP	$0.559_{0.027}$	$0.306_{0.026}$	$0.373_{0.026}$	$0.554_{0.026}$	$0.310_{0.027}$	$0.369_{0.025}$	$0.537_{0.028}$	$0.268_{0.027}$	$0.329_{0.025}$
BCPD	$0.335_{0.015}$	$0.484_{0.017}$	$0.392_{0.016}$	$0.415_{0.016}$	$0.589_{0.016}$	$0.483_{0.016}$	$0.519_{0.017}$	$0.684_{0.014}$	$0.586_{0.016}$
POI	$0.345_{0.007}$	$\mathbf{0.994_{0.003}}$	$0.508_{0.008}$	$0.325_{0.007}$	$\mathbf{0.997_{0.001}}$	$0.487_{0.008}$	$0.322_{0.006}$	$\mathbf{0.999_{0.001}}$	$0.483_{0.008}$
PELT	$0.372_{0.014}$	$0.570_{0.026}$	$0.429_{0.012}$	$0.342_{0.015}$	$0.629_{0.025}$	$0.414_{0.013}$	$0.291_{0.011}$	$0.651_{0.025}$	$0.370_{0.012}$
BF	$\mathbf{0.943_{0.006}}$	$0.962_{0.011}$	$\mathbf{0.948_{0.009}}$	$\mathbf{0.981_{0.003}}$	$0.985_{0.005}$	$\mathbf{0.982_{0.003}}$	$\mathbf{0.992_{0.002}}$	$0.986_{0.006}$	$\mathbf{0.988_{0.005}}$
	$K = 20$								
CHI	$0.119_{0.004}$	$0.587_{0.013}$	$0.197_{0.006}$	$0.116_{0.004}$	$0.630_{0.012}$	$0.193_{0.006}$	$0.121_{0.004}$	$0.637_{0.015}$	$0.202_{0.006}$
ECP	$0.675_{0.023}$	$0.379_{0.019}$	$0.482_{0.02}$	$0.720_{0.021}$	$0.452_{0.02}$	$0.557_{0.019}$	$0.742_{0.023}$	$0.466_{0.02}$	$0.569_{0.02}$
BCPD	$0.155_{0.008}$	$0.250_{0.011}$	$0.190_{0.009}$	$0.287_{0.013}$	$0.426_{0.014}$	$0.339_{0.014}$	$0.398_{0.015}$	$0.555_{0.015}$	$0.459_{0.015}$
POI	$0.245_{0.004}$	$\mathbf{0.995_{0.002}}$	$0.391_{0.006}$	$0.218_{0.005}$	$\mathbf{0.999_{0.001}}$	$0.356_{0.007}$	$0.219_{0.004}$	$\mathbf{1.000_{0.000}}$	$0.357_{0.006}$
PELT	$0.483_{0.016}$	$0.530_{0.02}$	$0.499_{0.013}$	$0.404_{0.015}$	$0.550_{0.022}$	$0.434_{0.012}$	$0.365_{0.014}$	$0.568_{0.023}$	$0.417_{0.012}$
BF	$\mathbf{0.972_{0.004}}$	$0.912_{0.012}$	$\mathbf{0.937_{0.008}}$	$\mathbf{0.989_{0.002}}$	$0.964_{0.007}$	$\mathbf{0.975_{0.004}}$	$\mathbf{0.997_{0.001}}$	$0.982_{0.005}$	$\mathbf{0.989_{0.003}}$
	$K = 30$								
CHI	$0.117_{0.004}$	$0.442_{0.012}$	$0.183_{0.006}$	$0.126_{0.005}$	$0.456_{0.013}$	$0.194_{0.007}$	$0.118_{0.005}$	$0.411_{0.011}$	$0.181_{0.007}$
ECP	$0.601_{0.027}$	$0.323_{0.018}$	$0.415_{0.021}$	$0.639_{0.025}$	$0.366_{0.02}$	$0.467_{0.021}$	$0.716_{0.022}$	$0.421_{0.019}$	$0.527_{0.02}$
BCPD	$0.052_{0.005}$	$0.093_{0.008}$	$0.073_{0.006}$	$0.139_{0.009}$	$0.227_{0.013}$	$0.174_{0.01}$	$0.273_{0.011}$	$0.404_{0.013}$	$0.323_{0.012}$
POI	$0.379_{0.004}$	$\mathbf{0.997_{0.001}}$	$0.548_{0.004}$	$0.345_{0.004}$	$\mathbf{0.999_{0.001}}$	$0.512_{0.004}$	$0.337_{0.004}$	$\mathbf{1.000_{0.000}}$	$0.502_{0.005}$
PELT	$0.499_{0.016}$	$0.394_{0.021}$	$0.424_{0.018}$	$0.475_{0.018}$	$0.491_{0.022}$	$0.482_{0.013}$	$0.444_{0.015}$	$0.535_{0.023}$	$0.464_{0.014}$
BF	$\mathbf{0.983_{0.003}}$	$0.957_{0.006}$	$\mathbf{0.969_{0.004}}$	$\mathbf{0.994_{0.002}}$	$0.969_{0.008}$	$\mathbf{0.979_{0.006}}$	$\mathbf{0.993_{0.002}}$	$0.989_{0.002}$	$\mathbf{0.991_{0.002}}$
	$K = 40$								
CHI	$0.106_{0.003}$	$0.408_{0.011}$	$0.167_{0.005}$	$0.122_{0.005}$	$0.440_{0.014}$	$0.188_{0.007}$	$0.132_{0.004}$	$0.443_{0.011}$	$0.201_{0.006}$
ECP	$0.663_{0.028}$	$0.266_{0.016}$	$0.381_{0.019}$	$0.701_{0.026}$	$0.303_{0.016}$	$0.423_{0.019}$	$0.785_{0.021}$	$0.337_{0.016}$	$0.466_{0.018}$
BCPD	$0.015_{0.002}$	$0.031_{0.004}$	$0.034_{0.003}$	$0.060_{0.005}$	$0.112_{0.008}$	$0.089_{0.005}$	$0.158_{0.008}$	$0.250_{0.011}$	$0.192_{0.009}$
POI	$0.356_{0.004}$	$\mathbf{0.987_{0.003}}$	$0.522_{0.005}$	$0.321_{0.003}$	$\mathbf{0.998_{0.001}}$	$0.485_{0.004}$	$0.304_{0.003}$	$\mathbf{0.999_{0.001}}$	$0.465_{0.003}$
PELT	$0.454_{0.023}$	$0.330_{0.022}$	$0.418_{0.018}$	$0.475_{0.019}$	$0.387_{0.022}$	$0.435_{0.016}$	$0.491_{0.015}$	$0.476_{0.021}$	$0.458_{0.015}$
BF	$\mathbf{0.978_{0.003}}$	$0.896_{0.015}$	$\mathbf{0.927_{0.010}}$	$\mathbf{0.994_{0.001}}$	$0.954_{0.010}$	$\mathbf{0.971_{0.006}}$	$\mathbf{0.996_{0.001}}$	$0.972_{0.009}$	$\mathbf{0.981_{0.007}}$

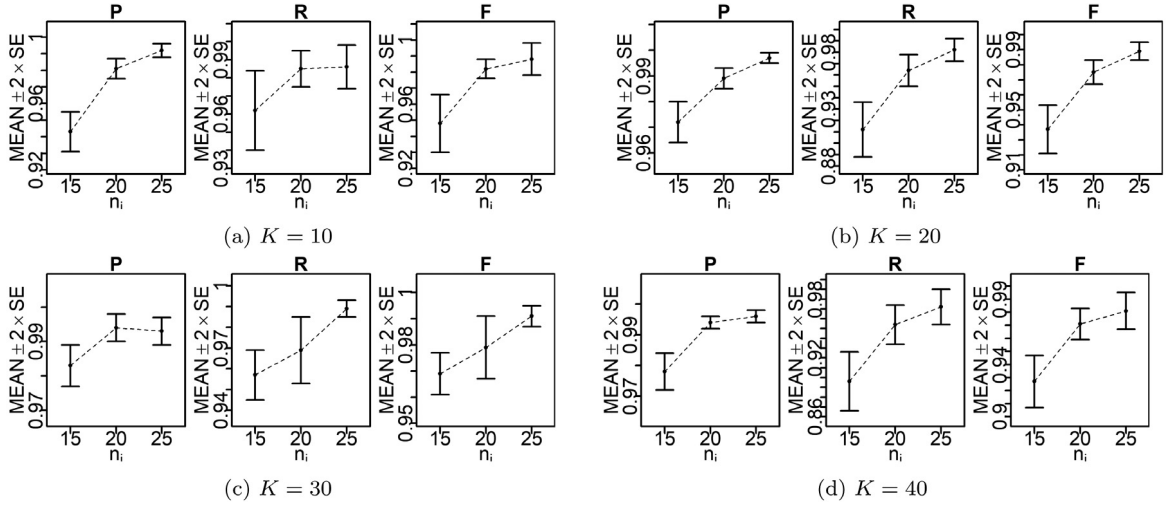

Fig. 4. Confidence intervals of BF for Simulation 2 over 100 replications.

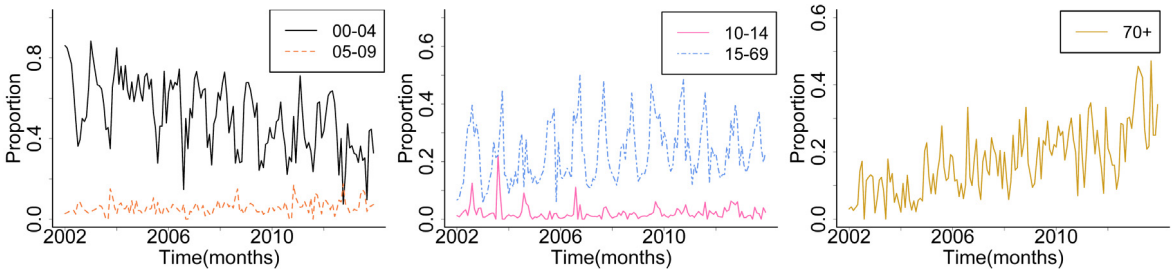

Fig. 5. Monthly proportions in five age-groups for the reported rotavirus cases in Brandenburg, Germany, 2002–2013.

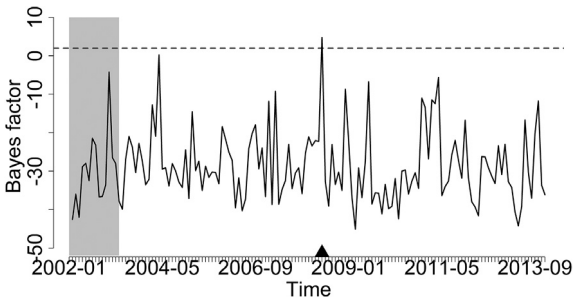

Fig. 6. Values of Bayes factor for varicella cases.

6.1. Agegroups of varicella cases

Varicella, is a highly contagious disease caused by the initial infection with varicella zoster virus. After the introduction of a vaccination recommendation, the disease incidence is effectively controlled. This dataset records the monthly number of rotavirus cases in the federal state Brandenburg during 2002–2013, where the population is divided into the five age-groups 00–04, 05–09, 10–14, 15–69, 70+ years. One point of research interest is to detect possible shifts in the age distribution of the cases after introducing vaccines in 2006, which are administered in children at the age of 4–6 months. Fig. 5 shows the time series of monthly proportions across the five age groups for the reported rotavirus cases. More specific description of this dataset can be found in Maëlle et al. [50].

From Fig. 5, it can be seen a shift in proportion away from the young group 00–04 to the old one 70+ around 2008–2009, after the introduction of vaccines. We now apply our proposed algorithm, BF, to detect possible change-points. Note that the average number of reported cases in each month is 289. In order to compare differences of monthly proportions, it is sufficient to set $m = 1$. Setting $K = 5$, Fig. 6 shows the value of Bayes factor where the shadow represents the burn-in period and the solid triangle represents the occurrence of a change-point, that is, May, 2008. The result indicates that after the introduction of vaccines, a shift in the age distribution of the cases occurs in May, 2008.

6.2. Lindisfarne scribes problem

The classical Lindisfarne Scribes problem is to decide the number of authors involved in translating the Lindisfarne Gospel, which was presumed to be a work of a monk named Eadfrith, Bishop of Lindisfarne in 698 and was translated into ancient English by one or more authors in the 10th century. Many statisticians have studied this problem [36,37,51,52]. Lindisfarne Gospel is divided into 64 chapters, where each chapter is assumed to be translated by one author and each author translated consecutive chapters. In ancient English, some word habits can be used to confirm the number of authors involved in translation. There are two datasets commonly used in the Lindisfarne Gospels [37].

- Dataset 1: $\boldsymbol{X}_i = (x_{i1}, x_{i2})^\top$ represents the number of times the third singular ends with $-s$ and $-\delta$ in the ith section, $i = 1, 2, \ldots, 64$, respectively.
- Dataset 2: $\boldsymbol{X}_i = (x_{i1}, x_{i2})^\top$ represents the number of times the second plural ends with $-s$ and $-\delta$ in the ith section, $i = 1, 2, \ldots, 64$, respectively.

The specific data can be found in Batsidis et al. [37]. Setting $K = 2$, we consider $m = 3, 4, 5$. Fig. 7 shows the results of our model for the two datasets, along with the results in literature. One can see that there are many kinds of segmentation. Our method and the method in Batsidis et al. [37] tend to detect less change-points than that in Horváth and Serbinowska [36]. For Dataset 1, all methods detect change-points around chapter 19 except for BF at $m = 3$. Horváth and Serbinowska [36] and BF detect change-pints at chapter 32, while Batsidis et al. [37] fails. For Dataset 2, all methods detect change-points at chapter 19. The first two statistics in Batsidis et al. [37] and BF at $m = 3, 4, 5$ only detect this change-point. By comparisons, we can conclude that our method is effective in detection change-points in Lindisfarne Scribes problem.

6.3. Twitter health news

In this subsection, we apply the proposed change-point detection method to the health news in Twitter dataset.[1] This dataset was collected using Twitter API, containing health news during September 30, 2013 to April 9, 2015 from more than 15 major health news agencies such as BBC, CNN, and NYT [53].

Here we focus on health news from BBC and track the degree of popularity of a given topic by monitoring the frequency of selected keywords [13]. Specifically, we focus on the Ashya King case. The Ashya King case concerned a boy with a brain tumor named Ashya King, whose parents took their son out of Southampton General Hospital (England) in August 2014 over a disagreement with doctors regarding his treatment to get proton therapy. The timeline of this case can be summarized as follows.

- On 28 August 2014, the parents took their son out of the hospital without telling the medical team and boarded a ferry to France.
- On 30 August 2014, King and his parents were found in Velez Malaga, Spain. King's parents were arrested and their son was sent to a local hospital for urgent treatment.
- On 5 September 2014, the High Court ruled that King could receive proton therapy in Prague.
- On 9 September 2014, King arrived at the Proton Therapy Center in Prague.
- On 15 September 2014, King began to receive post-operative radiotherapy at the Proton Therapy Center in Prague.
- On 24 October 2014, Ashya King returned to Spain after he received proton beam therapy.

We use the frequency of five keywords: "Ashya", "King", "proton", "beam", "therapy", which are shown in Fig. 8. Applying detection methods, Fig. 9 shows detected change-points. One can see POI and PELT detect no change-points. The detected change-points by CHI before August, 2014 are probably false positive, since the Ashya King case did not arise at that time. For our method, BF, when $m = 3, 5$, three change-points are detected, respectively around 2014-09-02, 2014-09-16 and 2014-10-24; when $m = 4$, two change-points, 2014-09-16 and 2014-10-24, are detected. As comparisons, ECP and BCPD also detect the change-point around 2014-09-02; CHI, ECP and BCPD detect the change-point around 2014-09-16; CHI ($m = 3$) and BCPD detect the change-point around 2014-10-24.

Considering the timeline, the Ashya King case happened on 28 August 2014. From 30 August 2014 to 15 September 2014, many important events happened including the arrest of King's parents, the High Court hearing, and the final judgment. During this period, the case was widely broadcast and got the nation talking. After 15 September 2014 when the final judgment was announced, King began to receive proton beam therapy in Prague. Discussions about this case were gradually declined. After King returned to Spain on 24 October 2014, discussions subsided. One can see that the change-points detected by our method are reasonable along the timeline.

[1] https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter.

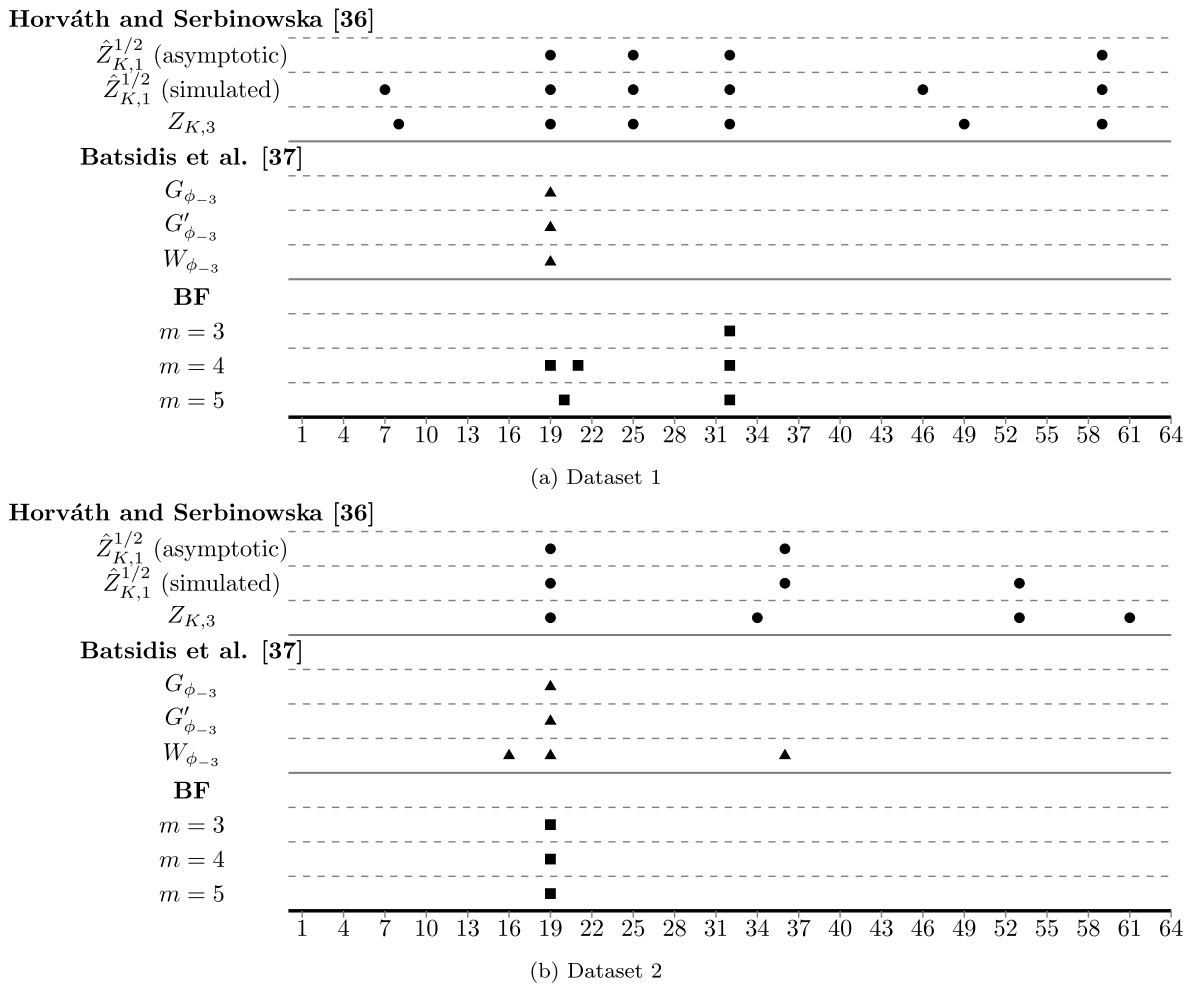

Fig. 7. Detected change-points in Lindisfarne Scribes problem.

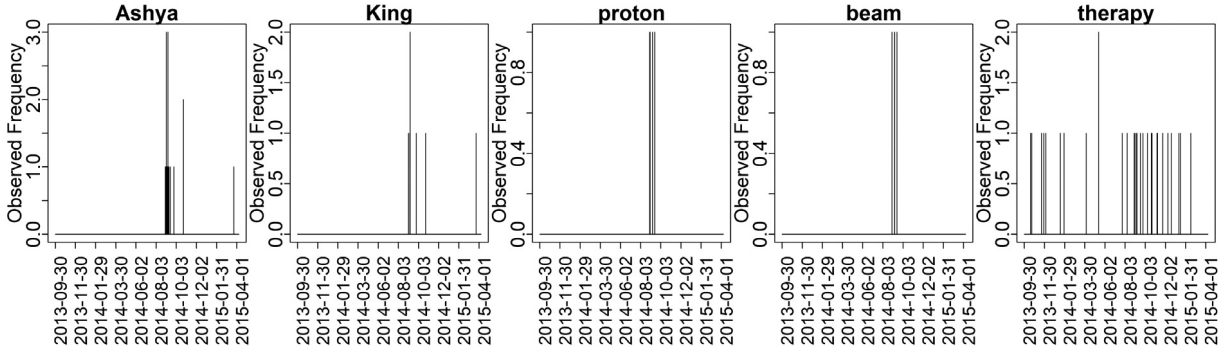

Fig. 8. Frequency of selected five keywords in Twitter health news from BBC.

6.4. Location monitoring

With the development of communication technology, we can obtain more and more users' location data via mobile devices such as smartphones and vehicle GPS trackers. Detecting changes in location data can help us mine valuable information such as patterns of users' behaviors to achieve highly commercial goals including precision marketing.

The dataset is collected from about 160,000 users by smartphones with more than 700 million records, from February 1, 2015 to May 17, 2015, a total of 106 days. Each record represents the location of the user at the time being recorded, mainly containing identification number, geographic information (longitude, latitude, country, province, city), and time. Note that the number of recording days and the number of records per day are quite different for each user. The user with the most records has more than 300,000 records, while the user with the least records has only one. From 160,000 users, we select those having more than 1000 records in more than 40 days. Moreover, in the preprocessing stage, we delete records with missing data and records that indicate users are moving, that is, users move a long distance in a short period. Besides, for each user, we extract one record every five minutes to obtain evenly recorded data. We analyze this dataset by cities and focus on Beijing, the capital of China. There are 17,476 users with 26,150,824 records. After

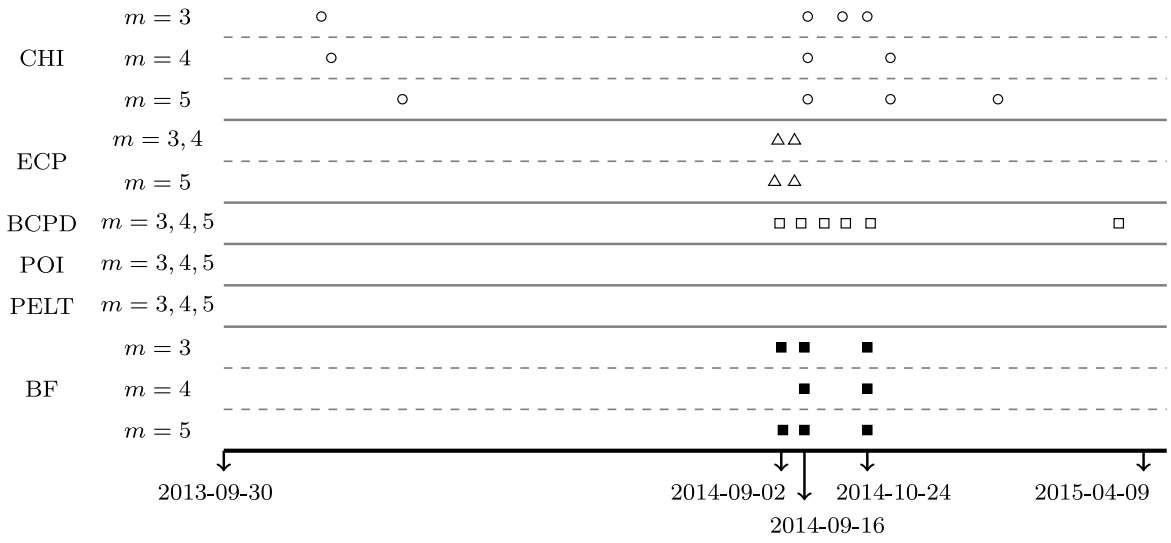

Fig. 9. Detected change-points in Twitter health news (BBC).

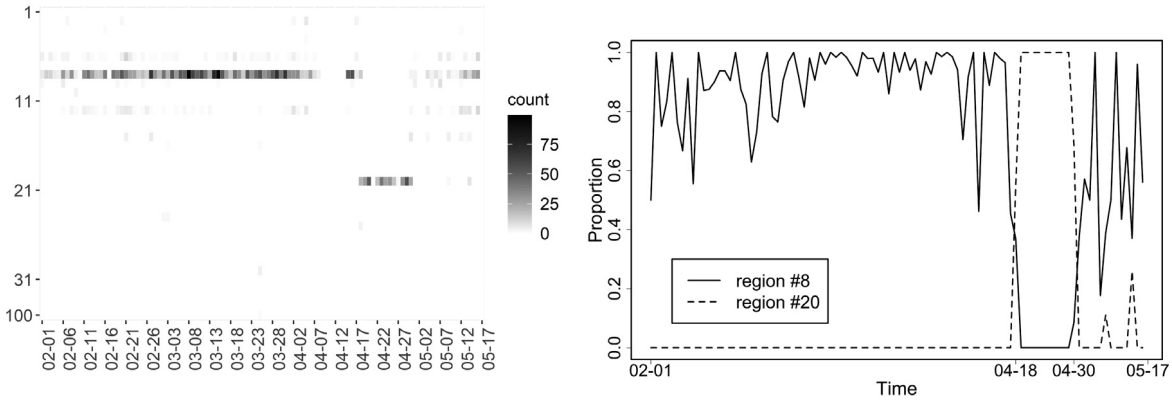

Fig. 10. The sequence (left) and the proportion of the first two most visited regions (right) of user #32.

preprocessing, there are 2256 users with 5283,386 records. Then we divide Beijing into 100 regions evenly by the grid method according to latitude and longitude. In the following, we detect change-points from the user-based perspectives.

We count the sequence $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$ for each user where $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iK})^\top$ with x_{ij} representing the frequency that the user visited region j in the ith day and $K = 100$ being the total number of regions.

Taking user #32 as an example, the sequence is shown in Fig. 10 (left), where each row represents a region and each column represents a day. As shown in the figure, the scales of the y-axis between 1–30 and 31–100 are different for a clearer display. Note that the number of visits of each user is limited in a day. There exists large amounts of "rare events", shown in light-colored block, making the dataset proper to apply our model. Since the number of categories is large, we consider $m = 7, 8, 9$. Applying detection methods, Fig. 11 shows the detected change-points. One can see that

(1) our method BF ($m = 7, 8, 9$) detects two change-points around 04–18 and 04–30, respectively;
(2) CHI ($m = 7, 8$), ECP ($m = 9$) and POI ($m = 7, 8, 9$) also detect these two change-points, along with many other change-points at earlier or later stage;
(3) CHI ($m = 9$) detects the change-point around 04–18 along with many other change-points at earlier stage, but fails to detect the change-point around 04–30;
(4) ECP ($m = 7, 8$) fails to detect these two change-points, while detects two other change-points at earlier stage;
(5) BCPD and PELT cannot detect any change-point.

Fig. 10 (right) shows the proportion of the first two most visited regions. Since the sum of the proportion of these two regions is quite large, other regions were rarely visited by the user. Before 04–18, the user almost only visited the region #8. During 04–18 to 04–30, the user turned to visit the region #20 and hardly went to the region #8. After 04–30, the user began to visit the region #8 again. Thus, we can see our method, BF, detects the two obvious change-points accurately. Regarding the number and moment of change-points detected, our method is competitive when there exists large amounts of "rare events". Moreover, considering all 2256 users, we find that there are 1438 users having change-points with our proposed detection method.

7. Conclusions

In this paper, we propose an online Bayesian approach to change-point detection for categorical data. Firstly, we formulate the change-point detection as a hypothesis testing problem. Secondly, we introduce the Dirichlet distribution as prior information and design the test statistic, Bayes factor. Thirdly, an online parameter estimation procedure and an online detection strategy are conducted to adapt to data streams. The proposed method is robust when some "rare events" exist. Simulations and

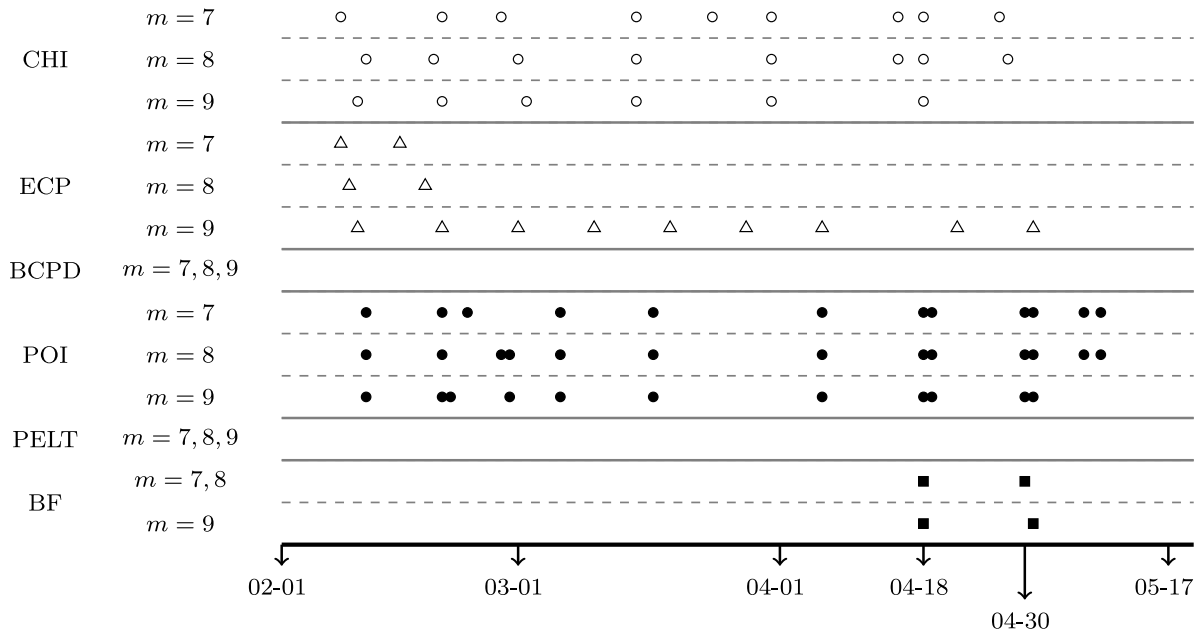

Fig. 11. Detected change-points for user #32.

applications show great advantages of our algorithm, compared with existing models.

A limitation of our method is the number of categories is fixed in this paper. The situation where it can change in data streams can be considered further. A dynamic change-point detection model is then desired for such case. Another concern is that though change-points are defined as abrupt changes of the underlying distribution in data over time, it is of interest to describe slow changes of events which may not directly reflect on the probability distribution. This problem may involve completely different perspectives such as the evolution of topics [54] and user preferences [55], which can be studied in the future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yiwei Fan: Methodology, Formal analysis, Software, Writing - original draft. **Xiaoling Lu:** Conceptualization, Methodology, Data curation, Supervision, Writing - review & editing.

Acknowledgments

The research was supported by the Ministry of Education Focus on Humanities and Social Science Research Base (Major Research Plan 17JJD910001) (China) and the fund for building world-class universities (disciplines) of Renmin University of China.

References

[1] X. Wan, Y. Wang, D. Zhao, Quality monitoring based on dynamic resistance and principal component analysis in small scale resistance spot welding process, Int. J. Adv. Manuf. Technol. 86 (9) (2016) 3443–3451.
[2] C.Y. Yau, Z. Zhao, Inference for multiple change points in time series via likelihood ratio scan statistics, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (4) (2016) 895–916.
[3] N. Chopin, Dynamic detection of change points in long time series, Ann. Inst. Statist. Math. 59 (2) (2007) 349–366.
[4] D.A. Bodenham, N.M. Adams, Continuous monitoring for changepoints in data streams using adaptive estimation, Stat. Comput. 27 (5) (2017) 1257–1270.
[5] Y. Zhang, H. Chen, J. Lu, G. Zhang, Detecting and predicting the topic change of knowledge-based systems: A topic-based bibliometric analysis from 1991 to 2016, Knowl.-Based Syst. 133 (2017) 255–268.
[6] N. Milosavljevic, A. Petrovic, ST segment change detection by means of wavelets, in: Neural Network Applications in Electrical Engineering, 2006, pp. 137–140.
[7] J.D. Sharpe, R.S. Hopkins, R.L. Cook, C.W. Striley, Evaluating google, twitter, and wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis, JMIR Public Health Surv. 2 (2) (2016).
[8] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (3) (2009) 15.
[9] J. Gama, I. liobait, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. 46 (4) (2014) 44.
[10] E. Topalidou, S. Psarakis, Review of multinomial and multiattribute quality control charts, Qual. Reliab. Eng. Int. 25 (7) (2009) 773–804.
[11] G. Wang, C. Zou, G. Yin, et al., Change-point detection in multinomial data with a large number of categories, Ann. Statist. 46 (5) (2018) 2020–2044.
[12] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal Machine Learning Research 3 (2003) 993–1022.
[13] S. Liu, M. Yamada, N. Collier, M. Sugiyama, Change-point detection in time-series data by relative density-ratio estimation, Neural Netw. 43 (2013) 72–83.
[14] N. Bouguila, Clustering of count data using generalized Dirichlet multinomial distributions, IEEE Trans. Knowl. Data Eng. 20 (4) (2008) 462–474.
[15] T. Masada, S. Kiyasu, S. Miyahara, Clustering images with multinomial mixture models, in: International Symposium on Advanced Intelligent Systems, 2007.
[16] P.D. Valpine, A.N. Harmon-Threatt, General models for resource use or other compositional count data using the dirichlet multinomial distribution, Ecology 94 (12) (2013) 2678–2687.
[17] L. Wei, E. Keogh, Semi-supervised time series classification, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, 2006, pp. 748–753.
[18] Y. Zheng, Y. Chen, Q. Li, X. Xie, W. Ma, Understanding transportation modes based on GPS data for web applications, ACM Trans. Web 4 (1) (2010) 1–36.
[19] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, M. Srivastava, Using mobile phones to determine transportation modes, ACM Trans. Sensor Netw. 6 (2) (2010) 1–27.
[20] A. Wald, Sequential tests of statistical hypotheses, Ann. Math. Stat. 16 (2) (1945) 117–186.
[21] S. Muthukrishnan, E. Den Berg, Y. Wu, Sequential change detection on data streams, 2007, pp. 551–556.

[22] A. Dries, U. Ruckert, Adaptive concept drift detection, Stat. Anal. Data Min. 2 (5) (2009) 311–327.
[23] I. Ada, M.R. Berthold, EVE: a framework for event detection, Evol. Syst. 4 (1) (2013) 61–70.
[24] J.B. Gomes, E. Menasalvas, P.A. Sousa, Learning recurring concepts from data streams with a context-aware ensemble, 2011, pp. 994–999.
[25] B. De Ketelaere, M. Hubert, E. Schmitt, Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data, J. Qual. Technol. 47 (4) (2015) 318–335.
[26] M.B. Harries, C. Sammut, K. Horn, Extracting hidden context, Mach. Learn. 32 (2) (1998) 101–126.
[27] A. Bouchachia, Fuzzy classification in dynamic environments, Soft Comput. 15 (5) (2011) 1009–1022.
[28] Y. Kawahara, T. Yairi, K. Machida, Change-point detection in time-series data based on subspace identification, in: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07, 2007, pp. 559–564.
[29] I. Naoki, J. Kurths, Change-point detection of climate time series by non-parametric method, in: Proceedings of the World Congress on Engineering and Computer Science, 2010, pp. 445–448.
[30] J. Zakaria, A. Mueen, E. Keogh, Clustering time series using unsupervised-shapelets, in: Data Mining (ICDM), 2012 IEEE 12th International Conference on, IEEE, 2012, pp. 785–794.
[31] H. Chen, N. Zhang, Graph-based change-point detection, Ann. Statist. 43 (1) (2015) 139–176.
[32] S. Aminikhanghahi, D.J. Cook, A survey of methods for time series change point detection, Knowl. Inf. Syst. 51 (2) (2017) 339–367.
[33] R.P. Adams, D.J. MacKay, Bayesian online changepoint detection, 2007, arXiv:0710.3742.
[34] D.S. Matteson, N.A. James, A nonparametric approach for multiple change point analysis of multivariate data, J. Amer. Statist. Assoc. 109 (505) (2014) 334–345.
[35] D.A. Wolfe, Y.-S. Chen, The changepoint problem in a multinomial sequence, Comm. Statist. Simulation Comput. 19 (2) (1990) 603–618.
[36] L. Horváth, M. Serbinowska, Testing for changes in multinomial observations: The lindisfarne scribes problem, Scand. J. Stat. 22 (3) (1995) 371–384.
[37] A. Batsidis, L. Horváth, N. Martín, L. Pardo, K. Zografos, Change-point detection in multinomial data using phi-divergence test statistics, J. Multivariate Anal. 118 (2013) 53–66.
[38] M. Höhle, Online change-point detection in categorical time series, in: Statistical Modelling and Regression Structures, Springer, 2010, pp. 377–397.
[39] Y.S. Son, S.W. Kim, Bayesian single change point detection in a sequence of multivariate normal observations, Statistics 39 (5) (2005) 373–387.
[40] A. Gupta, J.W. Baker, A Bayesian change point model to detect changes in event occurrence rates, with application to induced seismicity, in: 12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12, 2015.
[41] F. Desobry, M. Davy, C. Doncarli, An online kernel change detection algorithm, IEEE Trans. Signal Process. 53 (8) (2005) 2961–2974.
[42] H. Jeffreys, Some tests of significance, treated by the theory of probability, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 31, 1935, pp. 203–222.
[43] R.E. Kass, A.E. Raftery, Bayes factors, J. Amer. Statist. Assoc. 90 (430) (1995) 773–795.
[44] F.D. Santis, Statistical evidence and sample size determination for Bayesian hypothesis testing, J. Statist. Plann. Inference 124 (1) (2004) 121–144.
[45] O. Cappé, E. Moulines, On-line expectation-maximization algorithm for latent data models, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (3) (2009) 593–613.
[46] H. Wang, Y. Zhang, R. Nie, Y. Yang, B. Peng, T. Li, Bayesian image segmentation fusion, Knowl.-Based Syst. 71 (2014) 162–168.
[47] R.M. Neal, G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, Springer, 1998, pp. 355–368.
[48] D.P. Bertsekas, Nonlinear programming, J. Oper. Res. Soc. 48 (3) (1997) 334.
[49] R. Killick, P. Fearnhead, I.A. Eckley, Optimal detection of changepoints with a linear computational cost, J. Amer. Statist. Assoc. 107 (500) (2012) 1590–1598.
[50] S. Maëlle, S. Dirk, H. Michael, Monitoring count time series in R: Aberration detection in public health surveillance, 2014, arXiv:1411.1292.
[51] S.D. Silvey, The Lindisfarne scribes' problem, J. R. Stat. Soc. Ser. B Stat. Methodol. 20 (1) (1958) 93–101.
[52] A.F.M. Smith, Change-point problems: approaches and applications, Trabajos Estadistica Invest. Oper. 31 (1) (1980) 83–98.
[53] A. Karami, A. Gangopadhyay, B. Zhou, H. Kharrazi, Fuzzy approach topic discovery in health and medical corpora, Int. J. Fuzzy Syst. 20 (4) (2018) 1334–1345.
[54] H. Zhou, H. Yu, R. Hu, Topic evolution based on the probabilistic topic model: a review, Front. Comput. Sci. China 11 (5) (2017) 786–802.
[55] Q. Guo, L. Ji, J. Liu, J. Han, Evolution properties of online user preference diversity, Physica A 468 (2017) 698–713.