Pattern Recognition Letters 157 (2022) 83–89

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Multiple change point clustering of count processes with application to California COVID data

Shuchismita Sarkar[a,*], Xuwen Zhu[b]

[a] Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, OH, USA
[b] Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL, USA

ARTICLE INFO

Article history:
Received 9 April 2021
Revised 22 February 2022
Accepted 28 March 2022
Available online 30 March 2022

Edited by Maria De Marsico

Keywords:
Finite mixture modeling
Count process
Multiple change point estimation
EM algorithm

ABSTRACT

In this paper, a model-based clustering algorithm relying on a finite mixture of negative binomial Lévy processes is proposed. The algorithm models heterogeneous stochastic count process data and automatically estimates multiple change points upon fitting the mixture model. Such change point estimation identifies time points when deviation from the standard process has occurred and serves as an important diagnostic tool for analyzing temporal data. The proposed model is applied to the COVID-positive ICU cases in the state of California with very interesting results.

1. Introduction

Change point estimation aims to identify the time points when the underlying probability distribution deviates from the standard process. The literature of change point estimation has a long history [1,2], but most of the papers study the existence of a single change point [3–5]. Multiple change point estimation was first proposed by [6] and can be found in more recent literature. Such an important development increases the flexibility and modeling capability of the change point methods, especially when the process is distributed over a long period of time. Specifically, there is currently a scarcity of work on multiple change point estimation in count process data. In our paper, a stochastic negative binomial non-homogeneous Lévy process [7] is considered.

The objective of cluster analysis is to partition heterogeneous data points into multiple groups, known as clusters, such that each group has observations similar in data features. A finite mixture model [8] is a convex combination of several probability distributions known as mixture components. In model-based clustering [9], finite mixture models are used for finding clusters in data. This approach assumes a one-to-one relationship between a mixture component and an underlying data group. In recent years, model-based clustering has emerged as a powerful tool for modeling complex data in areas such as medicine [10], sociology [11], trade network [12], life event sequences [13], and many more [14–17] have analyzed time series data under mixture model setup. For count data modeling, Poisson mixture models have been extensively used and studied by [18,19], and [20].

Specifically, [21] proposed a novel extension of the Poisson mixture model for clustering count process data using intensity functions. Other interesting works include negative binomial process count and mixture modeling [22] and model-based clustering for stochastic process data in its functional form [23].

To the authors' knowledge, change point estimation and model-based clustering methodologies have only been studied in the common framework by [24]. The method relies on the matrix-variate mixture and exhaustively searches for a shift in the mean or variance. In this paper, our focus is on long-time count process data that can have multiple change points. An exhaustive search, in this case, would be computationally infeasible. The authors take an alternative approach and search for an optimal interval at each step of the algorithm which contains a change point most likely. Then the change point is exhaustively tested in the interval. We assume a mixed type of change point where there is a gap between two logit transformed segments.

In the following manuscript, Section 2 introduces the negative binomial change point process algorithm with a preliminary introduction to the finite mixture models and model-based clustering. Section 3 constructs various simulation settings for validating the proposed algorithm. The method is then applied to COVID-19 California hospital patient data reported by the California Department

* Corresponding author.
E-mail addresses: ssarkar@bgsu.edu (S. Sarkar), xzhu20@cba.ua.edu (X. Zhu).

https://doi.org/10.1016/j.patrec.2022.03.025
0167-8655/

of Public Health in Section 4. Section 5 concludes the paper with a discussion of paper contributions.

2. Methodology

2.1. Finite Mixture Modeling and EM algorithm

Let $\underline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ be an observed random sample from a mixture model with probability distribution function

$$g(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{y}; \boldsymbol{\Psi}_k), \tag{2.1}$$

where K represents the number of components, and $f_k(.; \boldsymbol{\Psi}_k)$ is the k^{th} mixture component with parameter $\boldsymbol{\Psi}_k$. τ_k is known as mixing proportion, *i.e.*, the prior probability that $\boldsymbol{y}$ originates from the k^{th} component. It has constraints $0 < \tau_k \leq 1$ and $\sum_{k=1}^{K} \tau_k = 1$. The maximum likelihood estimation of $\boldsymbol{\Psi}^\top = \{(\tau_k, \boldsymbol{\Psi}_k^\top), k = 1, 2, \ldots, K\}$ is obtained by employing the expectation-maximization (EM) algorithm [25] which is an iterative procedure known for its convenience in finding maximum likelihood estimates (MLE) in the presence of missing or incomplete data.

Let Z_i represent the unknown membership label of $\boldsymbol{y}_i$. Then, the complete-data i^{th} observation can be written as $\{\boldsymbol{y}_i, Z_i\}$ which yields a complete-data likelihood function of the following form:

$$\mathcal{L}_c(\boldsymbol{\Psi}; \underline{\boldsymbol{y}}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \Big(\tau_k f_k(\boldsymbol{y}_i; \boldsymbol{\Psi}_k)\Big)^{I(Z_i=k)}, \tag{2.2}$$

with $I(.)$ being the indicator function. At the expectation (E) step of the EM algorithm, the conditional expectation of the complete-data log-likelihood function given observed data (known as the Q-function) is found. This involves the calculation of posterior probabilities $\pi_{ik} = P(Z_i = k|\boldsymbol{y}_i; \boldsymbol{\Psi}_k)$. At the maximization (M) step, the Q-function is maximized with respect to $\boldsymbol{\Psi}_k$. The E- and M-steps are iterated until a specified convergence is met. If the number of components (K) is unknown, the optimal mixture order is usually selected based on the Bayesian information criterion (BIC) [26], or integrated completed likelihood (ICL) [27], or entropy criterion [21]. This process is commonly known as model-based clustering and it assumes that each mixture component is responsible for modeling a specific data group. This one-to-one correspondence establishes the classification rule for estimating membership as $\hat{z}_i = \underset{k}{\operatorname{argmax}}\{\hat{\pi}_{ik}\}$.

2.2. Lévy processes

A stochastic process $\boldsymbol{X} = \{X_t, t \geq 0\}$ is called a Lévy process [7,28], if it has right continuous left limited path with independent, stationary increments. A random vector $\boldsymbol{W} \in \mathbb{R}^T$ is infinitely divisible if, for every positive integer n, there exists i.i.d. random variables $W_{1,n}, W_{2,n}, \ldots, W_{n,n}$ such that $\boldsymbol{W}$ has the same distribution as $W_{1,n} + W_{2,n} + \cdots + W_{n,n}$. There is a close correspondence between a Lévy process and an infinitely divisible distribution according to which each Lévy process can be associated with an infinitely divisible distribution. The Poisson process is the most popular example of a Lévy process used for count data. The Poisson distribution is characterized by equality of mean and variance, and hence, is not appropriate for overdispersed data where sample variance is larger than the sample mean. A common approach to address the overdispersion issue is to employ a compound probability distribution of Poisson processes where the Poisson rate (λ) follows a gamma distribution with shape parameter α and rate parameter β. This results in a negative binomial (NB) process with parameters $r = \alpha$ and $p = 1/(\beta + 1)$, where p can be considered as the success probability that some event happened. Usually, r is viewed as the number of failures until the experiment is stopped, but the framework can be extended such that r takes non-integer values. A compound Poisson process is a Lévy process. The NB process, being a compound Poisson process, has the property of a Lévy process [29].

A generalization of Lévy process offering additional flexibility is time-inhomogeneous Lévy process [30,31] where the stationarity assumption can be relaxed while satisfying the other requirements of a Lévy process. A negative binomial time-inhomogeneous process whose increments are independent, but not stationary has been considered in this paper.

2.3. Finite mixture model of negative binomial process

Consider time interval $(t_0, t_T]$ split into small intervals of equal length t as $\{(t_0, t_1], \ldots, (t_{T-1}, t_T]\}$. Let $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{T}$ denote a count process associated to subject i where y_{ij} is the number of events happening in interval $(t_{j-1}, t_j]$. Given observed random sample of n count processes, we assume that they are generated from Eq. (2.1) where

$$f_k(\boldsymbol{y}; \boldsymbol{\Psi}_k) = \prod_{j=1}^{T} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_k)^{r_k} p_k^{y_{ij}} \tag{2.3}$$

is a negative binomial (NB) process associated to the k^{th} component with parameters $r_k (\geq 0)$ and p_k $(0 \leq p_k \leq 1)$. When $p_k \to 0$ and $r_k p_k \to \lambda_k$, the NB(r_k, p_k) process behaves like a Poisson(λ_k) process. Eq. (2.3) leads to the complete-data likelihood function

$$\mathcal{L}_c(\boldsymbol{\Psi}; \underline{\boldsymbol{y}}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\tau_k \prod_{j=1}^{T} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_k)^{r_k} p_k^{y_{ij}}\right)^{I(Z_i=k)}, \tag{2.4}$$

where Z_i still denotes the unknown membership of the i^{th} count process.

2.4. Finite mixture model of negative binomial change point process

Now consider that the success probability (p_k) associated to the k^{th} component changes m_k times $(0 \leq m_k < T)$ during the observed time interval $(t_0, t_T]$. Without loss of generality we assume that the shift occurs at the beginning of the m_k sub intervals of the form $(t_{j-1}, t_j]$; $j = 1, 2, \ldots, T-1$. With the change points denoted as $\zeta_{k1}, \ldots, \zeta_{km_k}$ and two end points $t_0 = \zeta_{k0}, t_T = \zeta_{k(m_k+1)}$, we define the following change in the logit of $(1 - p_k)$ as

$$\log\left(\frac{1 - p_k}{p_k}\right) = \begin{cases} \eta_{k0} + \nu_{k0} j, & \text{if } \zeta_{k0} < j \leq \zeta_{k1} \\ \eta_{k1} + \nu_{k1} j, & \text{if } \zeta_{k1} < j \leq \zeta_{k2} \\ \vdots & \\ \eta_{km} + \nu_{km} j, & \text{if } \zeta_{km} < j \leq \zeta_{k(m_k+1)}. \end{cases} \tag{2.5}$$

Since NB(r_k, p_k) results from a continuous mixture of Poisson(λ_k) distribution where λ_k follows a gamma distribution with shape parameter $\alpha_k = r_k$ and rate parameter $\beta_k = \frac{1-p_k}{p_k}$, the change in the logit of $(1 - p_k)$ indicates a change in $\log(\beta_k)$. Let p_{kl} denotes the success probability in the interval $(\zeta_{kl}, \zeta_{k(l+1)}]$. The formulations according to Eqs. (2.1), (2.3) and (2.5) lead to

$$\mathcal{L}_c(\boldsymbol{\Psi}; \underline{\boldsymbol{y}}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\tau_k \prod_{l=0}^{m_k} \prod_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_{kl})^{r_k} p_{kl}^{y_{ij}}\right)^{I(Z_i=k)}. \tag{2.6}$$

The corresponding Q-function has the form

$$Q(\boldsymbol{\Psi};\dot{\boldsymbol{\Psi}},\underline{\boldsymbol{y}}) = \sum_{i=1}^{n}\ddot{\pi}_{ik}\Bigg(\sum_{j=1}^{T}\log\binom{r_k+y_{ij}-1}{y_{ij}} + \sum_{l=0}^{m_k}\sum_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}}(r_k\log(1-p_{kl})+y_{ij}\log(p_{kl}))\Bigg). \quad (2.7)$$

The posterior probabilities are given by

$$\ddot{\pi}_{ik} = \frac{\dot{\tau}_k f_k(\boldsymbol{y}_i;\dot{\boldsymbol{\Psi}}_k)}{\sum_{k'=1}^{K}\dot{\tau}_{k'}f_{k'}(\boldsymbol{y}_i;\dot{\boldsymbol{\Psi}}_{k'})}. \quad (2.8)$$

From here, the mixing proportion $\ddot{\tau}_k$ is obtained as $\frac{1}{n}\sum_{i=1}^{n}\ddot{\pi}_{ik}$. The maximum likelihood estimates of η_{kl} and ν_{kl} can be found by maximizing

$$Q_{kl} = \sum_{i=1}^{n}\ddot{\pi}_{ik}\sum_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}}(\dot{r}_k\log(1-\dot{p}_{kl})+y_{ij}\log(\dot{p}_{kl})). \quad (2.9)$$

For estimating r_k

$$Q_k = \sum_{i=1}^{n}\ddot{\pi}_{ik}\Bigg(\sum_{j=1}^{T}\log\binom{\dot{r}_k+y_{ij}-1}{y_{ij}} + \sum_{l=0}^{m_k}\sum_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}}\dot{r}_k\log(1-\ddot{p}_{kl})\Bigg) \quad (2.10)$$

needs to be maximized. In the above expressions, the parameters with a single dot on top represent the estimates from the previous iteration and the ones with double dot refer to the estimates of the current iteration of the EM algorithm. If there is no change point associated to the k^{th} component *i.e.* if $m_k = 0$, Eq. (2.6) becomes equivalent to Eq. (2.4).

2.5. Algorithm for finding multiple change points

A common approach for change point estimation is to consider this as a model selection problem [1]. The idea involves fitting several competing models and selecting the best model based on BIC. If a change point exists in the observed time interval $(t_0, t_T]$, there are $T-1$ possible choices $\{t_1,\ldots,t_{T-1}\}$. Even for a time series with moderate length, an exhaustive search at each time point becomes computationally infeasible for m_k change points associated with K mixture components. Employing a sequential approach for multiple change point estimation has been suggested by the studies of [32,33]. Application of segmentation algorithm that involves partitioning the entire time series into smaller pockets [34–36] is particularly useful in this context. The algorithm used in this paper (Algorithm 1) is a combination of segmentation and exhaustive search approach.

It is assumed that there are at least h time points between two consecutive change points. In the first step of this algorithm, the list of change points $\mathcal{CP}list$ is set to *NULL* and a null model $\mathcal{M}_{null}$ with no change point is fitted with K mixture components according to Eq. (2.4). In the next step, the observed interval is partitioned into smaller intervals of length h producing break points $\{b_0,\ldots,b_d\}$ where $b_0 = t_0$, $b_d = t_T$ and $d = \lfloor T/h\rfloor + 1$. Here, the $\lfloor.\rfloor$ operator returns the integer part of a real number. For each mixture component, $d-2$ models according to Eq. (2.6) are fitted with the assumption that there exist change points at$\{b_i, b_{i+1}\}\cup\mathcal{CP}list$ where $1\le i\le d-2$. The best c models (c is user specified) among these $d-2$ candidates are chosen based on BIC value. The intervals associated to these c models are combined to get a favorable interval ($\mathcal{I}_{fav}$). An exhaustive search is performed for each time point within the interval which involves fitting $c(h+1)$ models with the assumption that change points exist at $\{cp\}\cup\mathcal{CP}list$ where $cp\in\mathcal{I}_{fav}$. If the BIC value for a fitted model is less than the BIC value of $\mathcal{M}_{null}$, the associated time point $\mathcal{CP}_{est}$ is an estimated change point for the related mixture component and the corresponding model is denoted as $\mathcal{M}_{current}$. $\mathcal{CP}list$ is updated to include the estimated change point. This process is repeated for each component. Then a search for the next set of change points is conducted in the same fashion *i.e.* identification of an optimal interval followed by an exhaustive search. The algorithm stops when all searches yield higher BIC.

Once the algorithm stops due to no BIC improvement, a final validation check is done on the estimated change point matrix to find out whether the change points identified earlier are still valid. In this process, one change point at a time is removed and the resulting BIC is compared with $BIC_{\mathcal{M}_{current}}$. A change point is dropped if the removal of it improves the BIC value. $\mathcal{M}_{current}$ is also updated accordingly. An illustration of the change point estimation process applied to simulated data is available in Fig. 1.

Algorithm 1: Algorithm for model based clustering of count process with multiple change points.

Given: count processes: $Y_{n\times T}$, mixture order: K, interval between two consecutive change points: h, number of favorable intervals to consider: c

Initialization: Construct null model $\mathcal{M}_{null}$ with no change point (as in Eq~) ;

$\mathcal{M}_{current} \leftarrow \mathcal{M}_{null}$;

$\mathcal{BIC}_{current} \leftarrow \mathcal{BIC}_{\mathcal{M}_{null}}$;

$\mathcal{BIC}_{old} \leftarrow \infty$;

$\mathcal{CP}list \leftarrow NULL$;

while $\mathcal{BIC}_{current} < \mathcal{BIC}_{old}$ **do**

  $\mathcal{BIC}_{old} \leftarrow \mathcal{BIC}_{current}$;

  for *k in 1:K* **do**

    Create break points $\{b_0,\ldots,b_d\}$ where $b_0 = t_0, b_d = t_T, b_{i+1} - b_i = h$ and $d = \lfloor T/h\rfloor + 1$;

    for *i in 1:(d-2)* **do**

      $interval_i \leftarrow \{b_i, b_{i+1}\}$;

      Construct model $\mathcal{M}_{interval_i}$ with change points at $\{b_i, b_{i+1}\}\cup\mathcal{CP}list$ (as in Eq~2.6);

      $\mathcal{BIC}_{interval_i} \leftarrow \mathcal{BIC}_{\mathcal{M}_{interval_i}}$;

    end

    $\mathcal{I}_{fav} \leftarrow \underset{\{interval_i\}}{\text{argmin}}\{\mathcal{BIC}_{interval}\}\cup\underset{\{interval_i\}}{\text{argmin}}\{\mathcal{BIC}_{interval}\setminus \mathcal{BIC}_{interval_{(1)}}\}\cdots\cup\underset{\{interval_i\}}{\text{argmin}}\{\mathcal{BIC}_{interval}\setminus\{\mathcal{BIC}_{interval_{(1)}},\ldots,\mathcal{BIC}_{interval_{(c-1)}}\}\}$;

    for *cp in* $\mathcal{I}_{fav}$ **do**

      Construct model $\mathcal{M}_{exhaustive_{cp}}$ with change points at $cp\cup\mathcal{CP}list$ (as in Eq~2.6);

      $\mathcal{BIC}_{exhaustive_{cp}} \leftarrow \mathcal{BIC}_{\mathcal{M}_{exhaustive_{cp}}}$;

    end

    $\mathcal{CP}_{est} \leftarrow \underset{\{cp\}}{\text{argmin}}\{\mathcal{BIC}_{exhaustive_{cp}}\}$;

    if $\mathcal{BIC}_{exhaustive_{\mathcal{CP}_{est}}} < \mathcal{BIC}_{current}$ **then**

      $\mathcal{BIC}_{current} \leftarrow \mathcal{BIC}_{exhaustive_{\mathcal{CP}_{est}}}$;

      $\mathcal{M}_{current} \leftarrow \mathcal{M}_{exhaustive_{\mathcal{CP}_{est}}}$;

      $\mathcal{CP}list \leftarrow \mathcal{CP}list\cup\mathcal{CP}_{est}$;

    end

  end

end

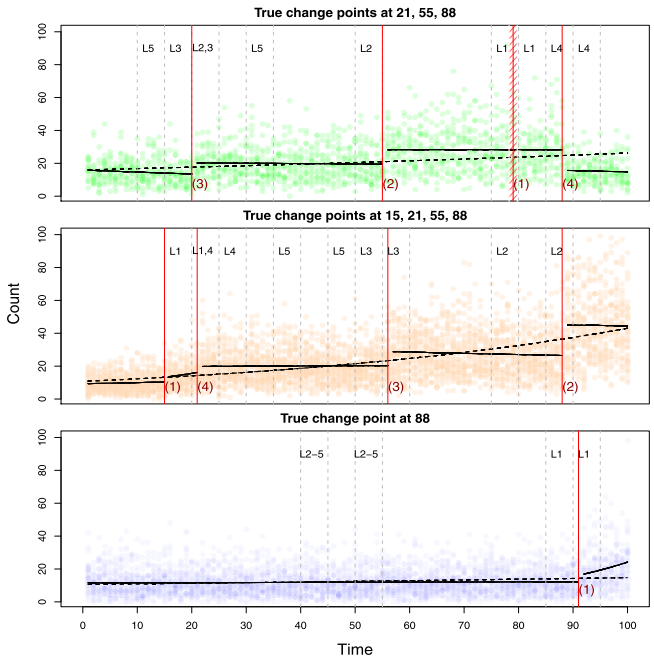


Fig. 1. Illustration of change point estimation.

Table 1
Change points associated to the experimental setups.

Component	$T = 100$				$T = 200$			
Green		21	55	88		42	110	176
Orange	15	21	55	88	30	42	110	176
Blue				88				176

3. Experimental validation

In this section, we illustrate the application of the proposed model under 8 different setups. In each case, 50 3-component NB process mixtures with multiple change points have been simulated. The experimental setups consider the cases when $n = 100, 400$ and $T = 100, 200$ where n denotes the number of data points simulated from the mixture and T represents the length of the time series. All possible combinations of n and T give rise to 4 frameworks. Within each framework, two cases are considered. The first case is the one where mixtures are generated with equal mixing proportions. In the second case, unequal mixing proportions with $\boldsymbol{\tau} = (0.4, 0.45, 0.15)$ have been used.

For each mixture, there are $3, 4$, and 1 change points associated with the first (green), second (orange), and third (blue) components respectively.

The change points associated with each time series length can be found in Table 1. Change point 88 is common to all components for the $T = 100$ case. The same can be said about change point 176 for $T = 200$. For both values of T, the green and orange components share two more change points at $21, 55$ and $42, 100$ respectively. The orange component has an additional change point at 15 for $T = 100$ case and at 30 for $T = 200$ case. It can be noted that this change point is quite close to the second change point of the orange component which makes the change point estimation task rather challenging for this component.

Fig. 1 presents an illustration of the change point estimation process described in Algorithm 1 for a mixture generated with equal mixing proportion and hundred data points for $T = 100$ case. Here we decided to inspect the best two intervals as $\mathcal{I}_{fav}$ during the interval search process. In the given example, the algorithm undergoes five loops before BIC improvement stops. Fig. 1 has three panels, each representing one mixture component. The true points associated with the mixtures are also indicated in the figure. The dashed vertical lines in grey color portray the two best intervals selected in each loop with the loop numbers displayed on the upper right-hand side indicating the order of selection.

The red vertical lines stand for the detected change points. The numbers presented in parenthesis on the lower right-hand side of the red lines indicate the sequence of estimation. If a chosen interval does not contain a red line, that implies that no point in that interval was identified as a change point. For the third component, intervals $(40, 45)$ and $(50, 55)$ were selected as the most favorable intervals in loops 2 to 5. However, no change point was detected in those intervals.

Finally, a validation check is applied to verify whether the previously found change points are still valid. Model is updated if dropping some change points improves BIC value. Such change point removals are shown with red strike marks in Fig. 1. In the given example, 79 was detected as a change point for the green component in the first loop, but it was removed during the validation check.

Table 2 displays the result of the conducted experiments. The ARI column has the average adjusted Rand index [37] for the associated setups which measures the similarity between two classification vectors with 1 indicating a perfect match. The numbers within the parentheses indicate the standard deviation of ARI values. The mean and standard deviations are calculated from the ARI values resulting from the models fitted on the 50 data sets generated under the associated experimental setup. It is to be noted that, for each combination of n, T and mixing proportion, 50 data sets are randomly generated using the same parameters. For each of the 50 datasets, the algorithm uses 10 random starts for detecting membership assignments using the null model. After choosing the best null model (random seed) based on BIC value, the interval search for change point begins followed by a thorough search on two best chosen intervals.

Although Fig. 1 suggests considerable overlap between the components, all experimental scenarios yield an almost perfect agreement. This remarkable performance can be attributed to the tail of the time series. The next two columns report the total number of estimated change points and the total number of correctly estimated change points for each component. An estimated change point is considered to be correct if it falls within ± 3 time points of the true change point. Since 50 data sets were simulated for each setup, there are a total of $150, 200$, and 50 change points associated with the three components. The last column reports the mean and standard deviation of the number of correctly estimated change points for each component. The summary statistics are calculated over the 50 simulated data sets.

The change points associated with the green and orange components are usually correctly estimated. The number of correct estimations increases with n and T. The change point associated with the blue component, on the other hand, is not easy to estimate with precision. The composition of each of the mixture components (Fig. 1) can help understand the reason behind this behavior. All change points in green and orange components indicate an abrupt change which is easier to estimate. The change in the blue component is gradual. Although the algorithm always finds a change point for this component, in some cases it fails to identify the exact location of change within ± 3 time points.

Fig. 1 also displays the estimated mean profile for each component. The solid black lines represent the estimated mean profile for each segment fitted by the model with change points. The mean profile estimated by the null model is given by dashed black lines. The mean profile plot suggests that the change point model is particularly useful for the green and blue components. The green

Table 2
Change point detection results for the experiments.

n	T	mixing proportion	Component	Average (Std. dev.) ARI	Total # est. CP	Total # correctly est. CP	Average (Std. dev.) # correctly est. CP
100	100	Equal	Green	1.000 (0.000)	146	146	2.920 (0.274)
			Orange		193	185	3.700 (0.544)
			Blue		50	41	0.820 (0.388)
		Unequal	Green	1.000 (0.000)	151	149	2.980 (0.141)
			Orange		199	195	3.900 (0.303)
			Blue		50	27	0.540 (0.503)
	200	Equal	Green	0.993 (0.051)	151	149	2.980 (0.141)
			Orange		198	193	3.860 (0.405)
			Blue		50	44	0.880 (0.328)
		Unequal	Green	1.000 (0.000)	151	150	3.000 (0.000)
			Orange		200	200	4.000 (0.000)
			Blue		51	35	0.700 (0.463)
400	100	Equal	Green	1.000 (0.000)	150	150	3.000 (0.000)
			Orange		201	200	4.000 (0.000)
			Blue		50	48	0.960 (0.198)
		Unequal	Green	1.000 (0.000)	150	150	3.000 (0.000)
			Orange		200	200	4.000 (0.000)
			Blue		51	43	0.860 (0.351)
	200	Equal	Green	1.000 (0.000)	150	150	3.000 (0.000)
			Orange		200	200	4.000 (0.000)
			Blue		52	46	0.920 (0.274)
		Unequal	Green	1.000 (0.000)	151	150	3.000 (0.000)
			Orange		200	200	4.000 (0.000)
			Blue		53	43	0.860 (0.351)

component undergoes an increase in counts with a peak observed in the $(55, 88)$ segment followed by a decline. The solid black lines associated with the change point model accurately capture this rise and fall phenomenon. For the blue component, after the true change point at 88, the probability of success has a steep increase. In this case, the change point model is successfully capturing the change in trend as well. The orange component, on the other hand, only experiences an increase in the success probability, but the slope is not as sharp as the blue component. In this case, in addition to the efficacy of the change point model, the null model also yields a decent fit. For the green and blue components, the null model fails to follow the trend of the data points.

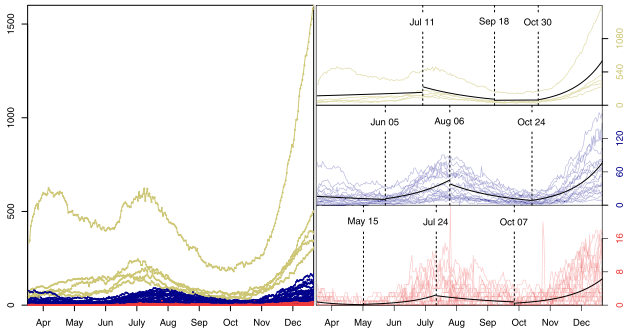
4. Application

In this section, county-level COVID-19 data from 1st April to 31st December, 2020 released by California Department of Public Health and The Los Angeles Times has been analyzed.

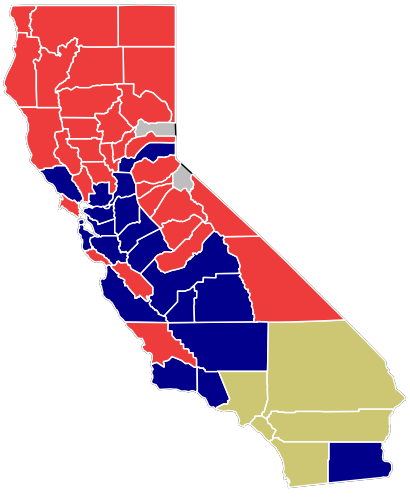
This dataset consists of daily ICU positive patients for 56 counties in California state for 275 days. Models are fitted for $K = 1, 2, 3, 4, 5$ components and the entropy values are used for selecting the valid number of clusters, where entropy of a model-based clustering result is given by $Ent(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n}\pi_{ik}\log(\pi_{ik})$. The lower the entropy criterion is, the more distinct mixture clusters are. The optimal entropy value of $\mathbf{1.27e^{-46}}$ is achieved by $K = 3$ components.

Fig. 2 a demonstrates the grouped curves by color. On the right, the dashed lines mark the change points estimated by the algorithm. The solid black lines represent mean profiles fitted by the model. It can be seen that the algorithm provides a reasonable partition to the curves. The curves in each group share similar characteristics and change points, whereas the magnitude of counts is rather different from group to group. The khaki group including counties San Bernardino, San Diego, Los Angeles, Orange, and Riverside has the highest daily ICU positive patients over the entire series. The blue and red groups, on the other hand, have relatively low ICU positive counts overall.

In the late spring, the numbers gradually dipped due to a number of regulations and close-downs implemented as well as stay-



(a) Grouped daily ICU positive patients curves for 56 counties in California



(b) 3-cluster solution presented on California map

Fig. 2. Analysis of California data.

at-home orders from the state governor. The second wave comes during the summer months May – July after the state slowly re-opened beaches, schools, restaurants, and bars in an attempt to salvage a deeply distressed economy.

During the end of June and the beginning of July, several close-down orders were re-imposed. Most indoor businesses were ordered to close, especially in khaki counties, which results in a decline in numbers starting from July 11 for the khaki group. On August 28, the state governor released the "Blueprint for a Safer Economy" guideline which classified all counties into 4 tiers based on the number of daily cases and current test positivity rate. The assignment can move to a lower or higher tier if recent numbers change. Re-openings will be permitted according to the assigned category. This smart movement effectively controlled the numbers during September – October. Interestingly, khaki counties experience a flat segment during Sep 18 – Oct 30. We can clearly see an increasing trend for all three groups towards the end of the year starting from October, which is probably driven by the holiday season and the presidential election.

California map is constructed in Fig. 2b according to the partition. The distribution is highly dependent on the geographical location of the county. The khaki counties are in South California. The blue group is along the west coast including the Bay area. The majority of red group counties are in the North California area.

5. Discussion

In this paper, a model-based clustering algorithm for negative binomial count process data is proposed with the functionality of multiple change point estimation. For a moderately long time series, an exhaustive search for multiple change points is a computationally challenging task. The computational difficulty has been successfully handled by adopting a search algorithm that first identifies a few optimal intervals likely holding a change point and then performs an exhaustive search on those small optimal intervals. Although the algorithm operates like a greedy search, there is a validation step at the end which verifies whether a previously detected change point is still valid after including the new ones. The proposed method is thus capable of partitioning the time series observations into homogeneous groups or clusters and then finding breakpoints along time within each cluster. The authors are aware of only one work [24] having similar functionality, but that model considers a single change point only. It is also worth noting that, the proposed model works for both abrupt and gradual change scenarios, which makes it suitable for a wide range of applications. The model has been tested on eight experimental setups with varying mixing proportions, sample size, and count process length. The performance of the model was assessed by its capability to recover the true classification and precision in finding change points. It was noted that the model works better on abrupt change points. For the gradual change point, the model can sense the existence of a change point, but sometimes it fails to identify the exact location of the change. As a future work, the precision of estimation of gradual change points can be improved by implementing a combination of interval search and multivariate adaptive regression splines (MARS) [38].

The model proposed in this paper was applied to the California county COVID-19 data which returned geographically meaningful partition of the counties and a reasonable estimation of multiple change points observed along time. Other possible future works in this direction include modeling heterogeneous multivariate count process data with change point estimation as well as using control chart test statistics for detection of abrupt change points.

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.

References

[1] J. Chen, A.K. Gupta, Parametric statistical change point analysis: with applications to genetics, medicine, and finance, Springer Science & Business Media, 2011.
[2] P.R. Krishnaiah, B. Miao, Review about estimation of change points, in: Handbook of Statistics, volume 7, Elsevier B.V., 1988, pp. 375–402.
[3] E.S. Page, On problem in which a change in parameter occurs at an unknown points, Biometrika 42 (1957) 248–252.
[4] K.J. Worsley, On the likelihood ratio test for a shift in location of normal populations, Journal of the American Statistical Association 74 (1979) 365–367.
[5] J. Chen, A. Gupta, A bayesian approach to the statistical analysis of a smooth-abrupt change point model, Advances and Applications in Statistics 7(1) (2007) 115–126.
[6] L.C. Zhao, P.R. Krishnaiah, Z.D. Bai, On detection of the number of signals in presence of white noise, Journal of Multivariate Analysis 20 (1986) 1–25.
[7] O. Kallenberg, Foundations of modern probability, Springer Science & Business Media, 2006.
[8] G. McLachlan, D. Peel, Finite mixture models, John Wiley and Sons, Inc., New York, 2000.
[9] V. Melnykov, Challenges in model-based clustering, WIREs: Computational Statistics 5 (2013) 135–148.
[10] P. Schlattmann, Medical applications of finite mixture models, Springer, 2009.
[11] I.C. Gormley, T.B. Murphy, A mixture of experts latent position cluster model for social network data, Statistical Methodology 7 (2010) 385–405.
[12] V. Melnykov, S. Sarkar, Y. Melnykov, On finite mixture modeling and model-based clustering of directed weighted multilayer networks, Pattern Recognition 112 (2021) 107641.
[13] Y. Zhang, V. Melnykov, X. Zhu, Model-based clustering of time-dependent categorical sequences with application to the analysis of major life event patterns, Statistical Analysis and Data Mining: The ASA Data Science Journal 14 (3) (2021) 230–240.
[14] V. Melnykov, X. Zhu, Studying crime trends in the usa over the years 2000–2012, Advances in Data Analysis and Classification 13 (1) (2019) 325–341.
[15] T. Roick, D. Karlis, P.D. McNicholas, Clustering discrete-valued time series, Advances in Data Analysis and Classification 15 (1) (2021) 209–229.
[16] S. Sarkar, V. Melnykov, X. Zhu, Tensor-variate finite mixture modeling for the analysis of university professor remuneration, The Annals of Applied Statistics 15 (2) (2021) 1017–1036.
[17] S.D. Tomarchio, A. Punzo, A. Maruotti, Parsimonious hidden markov models for matrix-variate longitudinal data, arXiv preprint arXiv:2107.04330 (2021).
[18] D. Karlis, Em algorithm for mixed poisson and other discrete distributions, ASTIN Bulletin: The Journal of the IAA 35 (1) (2005) 3–24.
[19] D. Karlis, E. Xekalaki, Mixed poisson distributions, International Statistical Review/Revue Internationale de Statistique (2005) 35–58.
[20] C. Etienne, O. Latifa, Model-based count series clustering for bike sharing system usage mining: a case study with the vélibsystem of paris, ACM Transactions on Intelligent Systems and Technology (TIST) 5 (3) (2014) 1–21.
[21] J. Ng, T.B. Murphy, Model-based clustering of count processes, Journal of Classification (2020), doi:10.1007/s00357-020-09363-4.
[22] M. Zhou, L. Carin, Negative binomial process count and mixture modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2) (2013) 307–320.
[23] F. Chamroukhi, H. Nguyen, Model-based clustering and classification of functional data, WIREs 9 (2019).
[24] X. Zhu, Y. Melnykov, On finite mixture modeling of change-point processes, Journal of Classification 39 (1) (2022) 3–22.
[25] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood for incomplete data via the EM algorithm (with discussion), Jounal of the Royal Statistical Society, Series B 39 (1977) 1–38.
[26] G. Schwarz, Estimating the dimensions of a model, Annals of Statistics 6 (1978) 461–464.
[27] C. Biernacki, G. Celeux, E.M. Gold, Assessing a mixture model for clustering with the integrated completed likelihood, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 719–725.
[28] S. Ken-Iti, Lévy processes and infinitely divisible distributions, Cambridge university press, 1999.
[29] T.J. Kozubowski, K. Podgórski, Distributional properties of the negative binomial Lévy process, Citeseer, 2008.
[30] W. Kluge, Time-inhomogeneous Lévy processes in interest rate and credit risk models, Verlag nicht ermittelbar, 2005 Ph.D. thesis.
[31] N. Koval, Time-inhomogeneous Lévy processes in cross-currency market models, Verlag nicht ermittelbar, 2005 Ph.D. thesis.
[32] J. Bai, Estimating multiple breaks one at a time, Econometric theory (1997) 315–352.
[33] Y.S. Niu, N. Hao, H. Zhang, Multiple change-point detection: A selective overview, Statistical Science (2016) 611–623.

[34] L.Y. Vostrikova, Detecting ædisorderg in multidimensional random processes, in: Doklady Akademii Nauk, volume 259, Russian Academy of Sciences, 1981, pp. 270–274.
[35] A.B. Olshen, E. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based dna copy number data, Biostatistics 5 (4) (2004) 557–572.
[36] P. Fryzlewicz, Wild binary segmentation for multiple change-point detection, Annals of Statistics 42 (6) (2014) 2243–2281.
[37] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.
[38] J.H. Friedman, Multivariate adaptive regression splines, The annals of statistics 19 (1) (1991) 1–67.