Exploring the space-time pattern of log-transformed infectious count of COVID-19: a clustering-segmented autoregressive sigmoid model

Xiaoping Shi*   Meiqian Chen and Yucheng Dong†

Abstract

At the end of April 20, 2020, there were only a few new COVID-19 cases remaining in China, whereas the rest of the world had shown increases in the number of new cases. It is of extreme importance to develop an efficient statistical model of COVID-19 spread, which could help in the global fight against the virus. We propose a clustering-segmented autoregressive sigmoid (CSAS) model to explore the space-time pattern of the log-transformed infectious count. Four key characteristics are included in this CSAS model, including unknown clusters, change points, stretched S-curves, and autoregressive terms, in order to understand how this outbreak is spreading in time and in space, to understand how the spread is affected by epidemic control strategies, and to apply the model to updated data from an extended period of time. We propose a nonparametric graph-based clustering method for discovering dissimilarity of the curve time series in space, which is justified with theoretical support to demonstrate how the model works under mild and easily verified conditions. We propose a very strict purity score that penalizes overestimation of clusters. Simulations show that our nonparametric graph-based clustering method is faster and more accurate than the parametric clustering method regardless of the size of data sets. We provide a Bayesian information criterion (BIC) to identify multiple change points and calculate a confidence interval for a mean response. By applying the CSAS model to the collected data, we can explain the differences between prevention and control policies in China and selected countries.

*Department of Mathematics and Statistics, Thompson Rivers University, email: xshi@tru.ca

†Center for Network Big Data and Decision-Making, Business School, Sichuan University, email: ycdong@scu.edu.cn

1 Introduction

During the COVID-19 outbreak, multiple complex factors resulted in the space-time pattern of spread. Fig. 1 shows the log-transformed infectious counts in each region in China, and in 33 selected countries at the end of April 20, 2020.

From Fig. 1, we can see two main characteristics of the spread: (i) the spread of COVID-19 has a space-time characteristic determined by different intervention policies, incomplete information, geographical locations, transport, climate, and so on; (ii) along the time, the log-transformed infectious counts presented different sigmoid (stretched S-shaped) curves. This phenomenon often happens in the life cycles of plants, animals, and viruses, which can rise and fall periodically. In each cycle, the sigmoid curve experiences three phases: slow rising, sharp rising, and slow falling.

Modeling the spread of COVID-19 in many regions over a long period of time is proven to be challenging. That is because many regions may not share the same spread pattern and different regions may exhibit various intervention policies that may cause instability in the models. The model for each region may have a large degree of noise, but a common cluster of all regions could have less noise by the law of large numbers. Thus, clustering is of importance to increase model fit. We may have to cluster all regions, even if the number of clusters is unknown. In addition, we should allow the model to incorporate unknown change points to further enhance the fitting performance. Ignoring the existence of change points may lead to poor model fitting and misleading model interpretation (Shi, Wang, Wei & Wu, 2016). Furthermore, it is often necessary to apply the model to updated data from an extended period of time. In the extended period, old clusters need to be updated and new change points may occur. Models with incorporated clusters and change points should be flexible and adaptive to the new data. In the next step, we shall consider the

2

nonlinear characteristics of the models.

Logarithmic transformation is often used for transforming count data, which includes zero values (Jin et al. , 2020) and grows exponentially over time. The simplest formula for exponential growth of a function y at the growth rate r, as time t goes on, is $y(t) = y(0)(1+r)^t$, which satisfies the linear differential equation $\frac{dy(t)}{dt} = \log(1+r)y(t)$. A nonlinear variation of this differential equation may lead $y(t)$ to a sigmoid function. For example, the solution of a nonlinear differential equation $\frac{dy(t)}{dt} = \log(1+r)y(t) - y^2(t)$ is the logistic function (Murray, p.308 , 1989; Liu & Stechlinski, p.84 , 2017). The exponential growth model has shown numerous applications in the modeling and controlling of complex systems. For example, the number of cells in a culture will increase exponentially until an essential nutrient is exhausted. A virus, for example SARS or COVID-19, has been found to spread exponentially (Katul et al. , 2020). The speed of spread slows down when an artificial immunization becomes available or intervention policies take effect. Other applications of the exponential growth model can be found in Physics (e.g., radioactive decay), Economics (e.g., a country's gross domestic product), Finance (e.g., investments), Computer science (e.g., computing growth and internet phenomena), and so on.

When systems have short-term memories and become more complex, it is extremely difficult to find a differential equation to describe the growth curve. In contrast, we may add some autoregressive terms in a regression function to adapt to the complex system. Kowsar et al. (2017) shows that an autoregressive logistic model was more accurate than a logistic model when it comes to predicting the behaviors of complex biological systems. The reason is that the added autoregressive terms, which behave like short-term memory, can make an appropriate adjustment to better fit the complex system. In the same spirit, we propose the clustering-segmented autoregressive sigmoid (CSAS) model with four key characteristics including unknown clusters, change points, stretched S-shaped curves, and

3

autoregressive terms. With the help of the CSAS model, we expect to understand how an outbreak is spreading in time and in space, to understand how the spread is affected by epidemic control strategies, and to apply the model to updated data from an extended period of time.

To identify this CSAS model, we first identify unknown clusters. There are many popular methods, such as K-means (Wang & Hartiganm , 1979) (implemented in the R function *kmeans*), Expectation-Maximization clustering for Gaussian Mixture Models (GMM-EM) (Akaho , 1995) (implemented in the R package *mclust*), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) (implemented in the R function *fpc::dbscan*), and Hierarchical clustering (Murtagh & Legendre , 2014) (implemented in the R function *hclust*). Except for GMM-EM, which can be considered to be parametric, all other methods need to predetermine the number of clusters or distance related parameters. To compare the dissimilarity of the curve time series, we need a nonparametric method that does not require predetermined parameters. Then, we can separate different regions from China and the selected 33 countries into clusters that share common patterns, segment the curve time series, and provide accurate fittings.

Our contributions include the following: (1) we propose the CSAS model to help understand how an outbreak is spreading in time and in space, to understand how the spread is affected by epidemic control strategies, and to apply the model to updated data from an extended period of time; (2) we provide a nonparametric graph-based clustering method with theoretical support, which furthermore proposes a very strict purity score that penalizes the overestimation of clusters. Simulations show that our method is fast and efficient for different sizes of data sets; (3) we give practical methods for segmentation and provide a confidence interval estimation for mean response; (4) we analyze the COVID-19 data in regions in China and selected countries, and explain the differences among the epidemic

4

prevention and control policies.

2   Main results

We assume the clustering-segmented autoregressive sigmoid (CSAS) model:

$$
Z_{i,t} = \sum_{m=1}^{M_i} \left\{ \beta_{1,i}^{(m)} + \beta_{2,i}^{(m)} \Phi(\beta_{3,i}^{(m)} + \beta_{4,i}^{(m)} t) \right.
$$
$$
\left. + \sum_{q=1}^{p} \beta_{q+4,i}^{(m)} Z_{i,t-q} + \varepsilon_{i,t}^{(m)} \right\} I(\tau_i^{(m-1)} < t \le \tau_i^{(m)}), \tag{1}
$$

where $Z_{i,t} = \log(1 + Y_{i,t})$; $Y_{i,t}$ is the number of confirmed cases for the ith $(1 \le i \le N)$ cluster and time $t \in [1, T]$; $i = \delta(j)$ for jth region with $1 \le j \le K$; $Z_{i,1-q} = 0$ for $q = 1, \ldots, p$; $I(A)$ is an indicator function taking 1 if A is true, 0 otherwise; $\tau_i^{(0)} = 0$, $\tau_i^{(M_i)} = T$, $\tau_i^{(m)}$ for $M_i > 1$ and $1 \le m \le M_i - 1$ are common change points for the ith cluster; $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$ is a cumulative distribution function (CDF) of the standard normal distribution representing the sigmoid curve; $\beta_{1,i}^{(m)}$'s and $\beta_{2,i}^{(m)}$'s are stretch location and scalar parameters, respectively; $\beta_{3,i}^{(m)}$'s and $\beta_{4,i}^{(m)}$'s are linear regression coefficients within the sigmoid curves; $\beta_{q+4,i}^{(m)}$'s are autoregressive regression coefficients; $\varepsilon_{i,t}^{(m)}$'s are independent random errors with a mean of zero and constant variance of $(\sigma_i^{(m)})^2$.

The CSAS model has four key characteristics: (1) it is implemented with unknown N different clusters among K regions. Due to the epidemic mechanism, human mobility and control strategy, the spread of epidemics displays a spatial propagation. We will propose a nonparametric method to cluster the regional data by applying the characteristic of sigmoid curve. This method does not introduce any factors and hence can be considered nonparametric; (2) the multiple S-shaped curves are described by change points. The

change points $\tau_i^{(m)}$ for $1 \leq m \leq M_i - 1$ are unknown and are related to the cluster (i). This is because different intervention policy releases such as lockdown, maintaining social distance, cancelling large events, closing schools, and so on, result in different segmented sigmoid curves among unknown clusters; (3) the regression function is mainly determined by the stretched S-curve $\beta_{1,i}^{(m)} + \beta_{2,i}^{(m)} \Phi(\beta_{3,i}^{(m)} + \beta_{4,i}^{(m)}t)$ and it allows a slight adjustment through the autoregressive terms, $\sum_{q=1}^{p} \beta_{q+4,i}^{(m)} Z_{i,t-q}$ which can be considered as a short-term memory for the response variable; and (4) after specifying both clusters and change points, we use the corresponding data sets to answer three questions. How do we estimate the regression coefficients? Are those coefficients significantly different from zero? How do we give a confidence interval for the mean response?

We give the following five remarks for the logarithmic transformation, the CDF function $\Phi(x)$, and the random error in the CSAS model.

Remark 1. $\log(1 + x)$ transformation is often used for transforming count data that include zero values (Jin et al. , 2020). When $Y_{i,t}$ is much smaller or larger than 1 in magnitude, $\log(1 + Y_{i,t}) \approx Y_{i,t}$ or $\log(1 + Y_{i,t}) \approx \log Y_{i,t}$ can be used. This transformation $\log(1 + Y_{i,t})$ of $Y_{i,t}$, which may grow exponentially over time, has two patterns, slow rises and slow falls, and hence can often be modeled by a stretched S-curve.

Remark 2. The nonlinear function $\Phi(x)$ is used to describe the stretched S-shaped curve. Other similar functions may be considered. For example, if we apply the approximation of $\Phi(x)$ by (Tocher , 1963), $\Phi(x) \approx \frac{1}{1+e^{-2\sqrt{2/\pi}x}}$ for all x, then we have

$$\beta_{1,i}^{(m)} + \beta_{2,i}^{(m)} \Phi(\beta_{3,i}^{(m)} + \beta_{4,i}^{(m)}t) \approx \beta_{1,i}^{(m)} + \frac{\beta_{2,i}^{(m)}}{1 + e^{-2\sqrt{2/\pi}(\beta_{3,i}^{(m)} + \beta_{4,i}^{(m)}t)}},$$

which is an extended logistic function of time t and is commonly used in logistic regression.

Remark 3. Mathematical modelling may provide an understanding of spread mecha-

nisms. The original mathematical model was proposed and solved by Daniel Bernoulli in 1760; see (Dietza & Heesterbeek , 2002). Recent developments and applications are mainly focused on the susceptible-infectious-recovered (SIR) model and its variants. The logistic function derived from a nonlinear differential equation may explain why we should apply the sigmoid curve to model the spread of disease (Murray, p.308 , 1989; Liu & Stechlinski, p.84 , 2017; Katul et al. , 2020).

Remark 4. The model for each region may have a large degree of noise, but a common cluster of all regions could have less noise because of the law of large numbers. So, we should consider an individual cluster in the CSAS model. From Fig. 7 C (Cluster 3), it can be seen that the noise is significantly smaller than that of Fig. 7 A (Province NM) or B (Province TJ). In addition, we should allow the model to incorporate unknown change points to further enhance the fitting performance. Fig. 8 F suggests that the residuals from the CSAS model without change points exhibit a clear trend. In contrast, the variance of noises in each segment should be constant; see Fig. 8 A-C. Models with incorporated clusters and change points should be flexible and adaptive to the new data; see the continued good performance of the CSAS model in the extended two-month data in the "Discussion and Conclusions" section.

Remark 5. With both autoregressive terms and CDF function in the CSAS model, the variance of random errors can be considered to be constant across segments. In Fig. 8 A, B and C, the model residuals are well-behaved across segments. The residuals in Fig. 8 D (the autoregressive terms are removed) and Fig. 8 E (the CDF function is removed as shown in the Long-Short-Term-Memory model in (Yang et al. , 2020)) suggest a time trend. This finding agrees with the fact that an autoregressive logistic model was more accurate than a logistic model as shown in (Kowsar et al. , 2017).

2.1 Clustering

To find clusters of all K regions, we consider the T dimensional series $\{\boldsymbol{Z}_j, 1 \leq j \leq K\}$, where its tth component is $Z_{j,t}$ for $1 \leq t \leq T$, and define the Euclidean distance between $\boldsymbol{Z}_{j_1}$ and $\boldsymbol{Z}_{j_2}$ as follows:

$$d(\boldsymbol{Z}_{j_1}, \boldsymbol{Z}_{j_2}) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (Z_{j_1,t} - Z_{j_2,t})^2}. \tag{2}$$

We construct an approximate shortest Hamiltonian path (SHP) based on a heuristic Kruska algorithm (HKA), which was proposed by (Biswas et al. , 2014) for a two-sample test. This was successfully applied into change point detection in (Shi, Wu and Rao, 2017, 2018). The HKA first sorts all edges in order of increasing distance defined in (2). First and foremost, the edge with a minimum distance must be selected. Then subsequent edges are chosen one-by-one from the remaining list of sorted edges according to the requirement of a path. If this current edge does not form a cycle with the previously selected edges, and every vertex connected by this current edge, or previously selected edges, has a degree not greater than 2, then this current edge must be selected. The HKA terminates when $K-1$ edges have been chosen. The approximate SHP is formed by chosen $K-1$ edges denoted as $\mathcal{P} = (j_1, \ldots, j_K)$. The next step is to find clusters based on $\mathcal{P}$. We define the edge set of $\mathcal{P}$ as $\mathcal{E}(\mathcal{P})$, and consider a subset of $\mathcal{E}(\mathcal{P})$:

$$\mathcal{E}^*(\mathcal{P}, \theta) = \{(j_s, j_{s+1}) \text{ for } s = 1, \ldots, K-1 \tag{3}$$
$$\text{such that } (j_s, j_{s+1}) \in \mathcal{E}(\mathcal{P}) \text{ and } d(\boldsymbol{Z}_{j_s}, \boldsymbol{Z}_{j_{s+1}}) \leq \theta \}.$$

We create a graph from the edge set $\mathcal{E}^*(\mathcal{P}, \theta)$ and define the connected components of this graph as a set of clusters $\mathcal{A} = \{\mathcal{A}_\ell, 1 \leq \ell \leq L\}$. We note that the R function

8

components in the R package *igraph* (Csardi & Nepusz , 2006) can calculate the connected components given the edge set.

Suppose that there is a set of classes $\mathcal{C} = \{\mathcal{C}_i, 1 \leq i \leq N\}$, where $\mathcal{C}_i = \{j|\delta(j) = i\}$. We need to measure how close the set of clusters $\mathcal{A}$ is to the predetermined set of classes $\mathcal{C}$. Purity (Manning, Raghavan and Schütze , 2008) is a measure of this extent defined as:

$$S(\mathcal{A}, \mathcal{C}) = \frac{1}{K} \sum_{\ell=1}^{L} \max_{1 \leq i \leq N} |\mathcal{A}_\ell \cap \mathcal{C}_i|. \tag{4}$$

In most cases, a bad clustering has a purity value close to 0 and a perfect clustering has a purity of 1. However, this measure may not give a realistic evaluation for overestimated clusters. For example, a purity score of 1 could happen by putting $\mathcal{A}_\ell = \ell$, $L = K$ and $N = 1$. In this case, one whole class is mis-clustered to K separate clusters with a purity score of 1.

We propose a very strict purity score to penalize overestimated clusters:

$$S^*(\mathcal{A}, \mathcal{C}) = \frac{1}{K} \sum_{\ell=1}^{L} \max_{1 \leq i \leq N} |\mathcal{A}_\ell \cap \mathcal{C}_i| - \frac{|L - N|}{\max(L, N)}. \tag{5}$$

Users may add additional weight on the second penalty term according to different requirements. Based on this very strict purity evaluation, a very bad clustering would have a purity value close to -1, and a perfect clustering will still have a purity of 1. If $\mathcal{A}_\ell = \ell$, $L = K$ and $N = 1$, then $S^*(\mathcal{A}, \mathcal{C}) = 1/L$, decreasing as L increases. Overestimated clusters may have a very strict purity score close to 0. A natural question comes: does our clustering have a very strict purity score of 1? To answer this question, we make the following assumptions.

Assumption 1. Let $\varepsilon_{i,t}$ be $Z_{i,t} - E(Z_{i,t})$. Assume that $\varepsilon_{i,t}$ is independent and identically

9

distributed (i.i.d.) satisfying $E(\varepsilon_{i,t}^4) < \infty$ for all $1 \le j \le K$ and $1 \le t \le T$.

Assumption 2. There exists a $\eta(T)$, satisfying that $\eta^2(T) > 2E(\varepsilon_{1,1}^2)$, $K << \{\eta^2(T) - 2E(\varepsilon_{1,1}^2)\}^2 T$ and $\min_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} d(E(\boldsymbol{Z}_{j_1}), E(\boldsymbol{Z}_{j_2})) > 2\eta(T)$.

In Assumption 1, if $\varepsilon_{i,t}$ is dependent, then we require the upper bound of

$$E \left| \sum_{t=1}^{T} (\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2 - E(\varepsilon_{j_1,t}^2) - E(\varepsilon_{j_2,t}^2) \right|^2 << T\eta^2(T)/K.$$

In Assumption 2, we require K to be quite small compared to T. Note that $d(E(\boldsymbol{Z}_{j_1}), E(\boldsymbol{Z}_{j_2}))$ is easy to evaluate because $d(E(\boldsymbol{Z}_{j_1}), E(\boldsymbol{Z}_{j_2})) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \{E(Z_{j_1,t}) - E(Z_{j_2,t})\}^2}$. We have the following Theorem 1.

Theorem 1. Suppose Assumptions 1-2 hold. Choose $\theta = \eta(T)$ as in (3). As $T \to \infty$, we have $P\{S^*(\mathcal{A}, \mathcal{C}) = 1\} \to 1$.

Proof of Theorem 1. We first prove that $P\{\max_{j_1 \neq j_2, \delta(j_1) = \delta(j_2)} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (Z_{j_1,t} - Z_{j_2,t})^2} > \eta(T)\} \to 0$. Because $\delta(j_1) = \delta(j_2)$, $Z_{j_1,t} - Z_{j_2,t} = \varepsilon_{j_1,t} - \varepsilon_{j_2,t}$. Then we have

$$
\begin{aligned}
P & \left\{ \max_{j_1 \neq j_2, \delta(j_1) = \delta(j_2)} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (Z_{j_1,t} - Z_{j_2,t})^2} > \eta(T) \right\} \\
& \le \sum_{j_1 \neq j_2} P \left\{ \sum_{t=1}^{T} (\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2 > \eta^2(T)T \right\} \\
& = \sum_{j_1 \neq j_2} P \left[\sum_{t=1}^{T} (\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2 - E\{(\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2\} \right. \\
& \left. \quad > \{\eta^2(T) - 2E(\varepsilon_{1,1}^2)\}T \right] \\
& \le \sum_{j_1 \neq j_2} P \left[\left| \sum_{t=1}^{T} (\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2 - E\{(\varepsilon_{j_1,t} - \varepsilon_{j_2,t})^2\} \right|^2 \right.
\end{aligned}
$$

10

$$> \{\eta^2(T) - 2E(\varepsilon_{1,1}^2)\}^2 T^2\Big]$$

$$\leq \frac{cK}{\{\eta^2(T) - 2E(\varepsilon_{1,1}^2)\}^2 T},\tag{6}$$

where c is a constant not related to either K or T. By Assumption 2, this upper bound converges to zero. Next, we prove that $P\{\min_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_{j_1,t} - Z_{j_2,t})^2} \leq \eta(T)\} \to 0$. By the Minkowski inequality and Assumption 2,

$$P\left\{ \min_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_{j_1,t} - Z_{j_2,t})^2} \leq \eta(T) \right\}$$

$$\leq P\left[\min_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} \sqrt{\frac{1}{T}\sum_{t=1}^{T}\{E(Z_{j_1,t}) - E(Z_{j_2,t})\}^2} \right.$$

$$\left. - \sqrt{\frac{1}{T}\sum_{t=1}^{T}\{\varepsilon_{j_1,t}) - \varepsilon_{j_2,t}\}^2} \leq \eta(T) \right]$$

$$\leq P\left[\min_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} 2\eta(T) - \sqrt{\frac{1}{T}\sum_{t=1}^{T}\{\varepsilon_{j_1,t}) - \varepsilon_{j_2,t}\}^2} \leq \eta(T) \right]$$

$$= P\left[\max_{j_1 \neq j_2, \delta(j_1) \neq \delta(j_2)} \sqrt{\frac{1}{T}\sum_{t=1}^{T}\{\varepsilon_{j_1,t}) - \varepsilon_{j_2,t}\}^2} \geq \eta(T) \right],$$

which converges to zero by (6).

By the HKA, for any $\mathcal{A}_\ell$, there exists $\mathcal{C}_i$ such that $\mathcal{C}_i = \mathcal{A}_\ell$ in probability, which implies that $\max_{1 \leq i \leq N} |\mathcal{A}_\ell \cap \mathcal{C}_i| = |\mathcal{A}_\ell|$ and $L = N$ hold in probability. So, $P(S^*(\mathcal{A},\mathcal{C}) = 1)$ converges to 1 as $T \to \infty$. The proof of Theorem 1 is finished.

To apply Theorem 1, we need to set the right value for θ. In real problems, θ could

be unknown. We shall propose a data driven method to select the threshold value of θ. A naive choice of θ based on outlier detection is

$$\hat{\theta} = \text{median}_{s=1,\dots,K-1}(x_s)$$
$$+ 2.5\left(1.483 \times \text{median}_{s=1,\dots,K-1}|x_s - \text{median}_{s=1,\dots,K-1}(x_s)|\right), \qquad (7)$$

where $x_s = d(\boldsymbol{Z}_{j_s}, \boldsymbol{Z}_{j_{s+1}})$ for $s = 1, \dots, K-1$, $\text{median}_{s=1,\dots,K-1}(x_s)$ and $1.483 \times \text{median}_{s=1,\dots,K-1}|x_s - \text{median}_{s=1,\dots,K-1}(x_s)|$ are robust estimates of mean and standard deviation of $\{x_s, s = 1, \dots, K-1\}$, respectively, and 2.5 is the cutoff value. It works well for relatively small N to K. The large values in the series of $\{x_s, s = 1, \dots, K-1\}$ would not affect the threshold value $\hat{\theta}$ and hence they could be successfully removed. However, if the distribution of x_s's, with the exception of outliers, is a mixture of two or more probability distributions which commonly occurs in multiple clusters, then $\hat{\theta}$ may not be consistent to θ. Therefore, we propose Algorithm 1 based on Bayesian information criterion (BIC) to choose θ.

2.2   Segmentation

Denote a set of change points as $C_i = \{\tau_i^{(1)}, \cdots, \tau_i^{(M_i-1)}\}$, where $i = 1, \dots, N$ and $M_i - 1$ is the number of change points. Since M_i is unknown in practice, we would need to estimate the change points. Consider the segment $[t_-, t^+]$ and define two residual sums of squares

$$S_{i,0}(t_-, t^+) = \min_{\boldsymbol{\beta}} \sum_{t=t_-}^{t^+} \{Z_{i,t} - f(t; \boldsymbol{\beta})\}^2, \qquad (8)$$

$$S_{i,1}(t_-, t_0, t^+) = \min_{\boldsymbol{\beta}} \sum_{t=t_-}^{t_0} \{Z_{i,t} - f(t; \boldsymbol{\beta})\}^2$$

12

$$+ \min_{\boldsymbol{\beta}} \sum_{t=t_0+1}^{t^+} \{Z_{i,t} - f(t;\boldsymbol{\beta})\}^2,\qquad(9)$$

where $1 \leq t_- < t_0 < t^+ \leq T$ and $f(t;\boldsymbol{\beta}) = \beta_1 + \beta_2\Phi(\beta_3 + \beta_4 t) + \beta_5 Z_{i,t-1} + \beta_6 Z_{i,t-2}$. Here, we consider two autoregressive terms. Then, the estimated change point is denoted as $\hat{t}_i(t_-, t^+)$:

$$\hat{t}_{i,t_-,t^+} = \arg\min_{t_-+\Delta/2 < t_0 < t^+ - \Delta/2} S_{i,1}(t_-, t_0, t^+).\qquad(10)$$

where Δ is the minimum distance between two adjacent change points and $t^+ - t_- > \Delta$.

In light of Bai & Perron (2003), we apply the BIC method for model comparison. Define

$$\mathrm{BIC}_{i,\nu}(t_-, t^+) = (t^+ - t_- + 1)\log\{\hat{\sigma}_{i,\nu}^2\} + 6(\nu + 1)\log(t^+ - t_-)\qquad(11)$$

where $\nu = 0$ or 1, $6(\nu+1)$ is the number of parameters and $\hat{\sigma}_{i,0}^2 = (t^+ - t_- + 1)^{-1} S_{i,0}(t_-, t^+)$ and $\hat{\sigma}_{i,1}^2 = (t^+ - t_- + 1)^{-1} S_{i,1}(t_-, \hat{t}_{i,t_-,t^+}, t^+)$. Combined with the Iterated Cumulative Sums of Squares Algorithm (ICSS) (Inclán & Tiao , 1994), we propose Algorithm 2 to estimate multiple change points.

In Algorithm 2, there are two main steps that include finding candidate change points and refining them. We set the minimum distance between two adjacent change points, Δ, to be 10 for real data analysis.

2.3 Fitting

First, we use the well-known *nls* function in the R package *stats* (R Core Team , 2020) to find the minimum value as shown in (8) and give t tests on regression coefficients, where initial values of parameters are given by grid search. Second, we give a confidence interval of regression function, denoted as $g_{i,t}(\boldsymbol{\beta}) = E(Z_{i,t}|Z_{i,t-1}, Z_{i,t-2})$ for $t \in [t_-, t^+]$, by the delta

13

method as follows. By first-order Taylor expansion at the solution $\hat{\boldsymbol{\beta}}$, we have

$$g_{i,t}(\boldsymbol{\beta}) \approx g_{i,t}(\hat{\boldsymbol{\beta}}) + \nabla g_{i,t}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

The approximate $(1 - \alpha)100\%$ confidence interval for $g_{i,t}(\boldsymbol{\beta})$ is

$$g_{i,t}(\hat{\boldsymbol{\beta}}) \pm t_{\alpha/2}^*(t^+ - t_- - 6)\sqrt{\nabla g_{i,t}(\hat{\boldsymbol{\beta}})\top \text{Var}(\hat{\boldsymbol{\beta}})\nabla g_{i,t}(\hat{\boldsymbol{\beta}})},$$

where $t_{\alpha/2}^*(t^+ - t_- - 6)$ is the $\alpha/2$ lower quantile of a t distribution with degrees of freedom $t^+ - t_- - 6$ and $\text{Var}(\hat{\boldsymbol{\beta}})$ can be estimated by the *nls* function in *R*. Here, we consider $\alpha = 0.05$.

3 Simulations

We consider three classes $N = 3$. Let $\mathcal{C}_1, \mathcal{C}_2$, and $\mathcal{C}_3$ be randomly seperated classes with $\cup_{i=1}^3 \mathcal{C}_i = \{1, \ldots, K\}$. Denote the number of elements in ith class as n_i with $\sum_{i=1}^3 n_i = K$. We produce $Z_{i,t}$ from the following model

$$Z_{i,t} = \sum_{m=1}^{M_i} \left\{ \beta_{1,i}^{(m)} + \beta_{2,i}^{(m)}\Phi(\beta_{3,i}^{(m)} + \beta_{4,i}^{(m)}t) \right\} I(\tau_i^{(m-1)} < t \leq \tau_i^{(m)})$$

$$+ \varepsilon_t, \tag{12}$$

where $i = 1, 2, 3$, $t = 1, \ldots, T$ and ε_t's are independent Normal errors with mean zero and variance σ^2.

For the first class, let $M_1 = 1$, $\beta_{1,1}^{(1)} = 0$, $\beta_{2,1}^{(1)} = 10$, $\beta_{3,1}^{(1)} = -4$, and $\beta_{4,1}^{(1)} = -0.05$. For the second class, let $M_2 = 2$, $\beta_{\ell,2}^{(1)} = 0$ for $\ell = 1, \ldots, 4$, $\tau_2^{(1)} = T/3$, $\beta_{1,2}^{(2)} = 0$, $\beta_{2,2}^{(2)} = 20$,

14

$\beta_{3,2}^{(2)} = -3$, and $\beta_{4,2}^{(2)} = 0.03$. For the third class, let $M_3 = 2$, $\beta_{\ell,3}^{(1)} = 0$ for $\ell = 1, \ldots, 64$, $\tau_3^{(1)} = 2T/3$, $\beta_{1,3}^{(2)} = 0$, $\beta_{2,3}^{(2)} = 5$, $\beta_{3,3}^{(2)} = -2$, and $\beta_{4,3}^{(2)} = 0.07$.

Fig. 2 plots different S-curves of $Z_{i,t}$ for $i = 1, 2, 3$ and $T = 150$. Next, we compare our graph-based clustering method as shown in Algorithm 1 with the model-based clustering method (Fratey & Raftery , 2002) implemented in the R function *Mclust* (Fraley et al. , 2020). Fig. 3 shows the averaged strict purity score as in (5) for estimated clusters based on these two methods for $\sigma = 0.1, 0.2, \ldots, 1$, different n_i's and 100 replications.

It can be seen from Fig. 3 that our method is very accurate for different σ and sizes of classes because its very strict purity scores are close to 1. This agrees with the conclusion in Theorem 1. In contrast, the performance of the model-based clustering method is affected by large σ and large sizes of classes. Specially, when $n_1 = 20$, $n_2 = 100$ and $n_3 = 200$, the computation is significantly slower compared with our method. The comparison of computing time is not presented here.

4   Real data analysis

We continue to use the data set of log-transformed infection counts from December 1, 2019 to April 20, 2020 from Chinese provinces/regions and the 33 countries, and present the clustering and S-shaped fitting with change points. Here, two autoregressive components $(p = 2)$ in (1) are suggested.

4.1   Clustering

Based on the graph-based clustering Algorithm, the clusters of COVID-19 in China and the rest of the world are presented in Fig. 4, where the optimal path is presented as a cycle with vertexes representing clusters in different colors and overextended curves. This way of

presentation is to transmit three aspects of information: (i) this analysis is for virus data, therefore, we should use the cycle and the sharp nodes to describe the structure of the virus; (ii) the optimal graph is a path connecting all nodes where nodes can be provinces/regions in China or countries in the world; and (iii) readers can quickly find the different clusters and where to separate them from the path.

From Fig. 4, we observe the following.

(1) As shown in COVID-19 cases in China, the 34 provinces/regions are clustered into 7 categories. Specifically, Hubei (HB), Xizang (XZ), Qinghai (QH), Macao (MO), Hong Kong (HK), and Taiwan (TW) are individually clustered into separate categories, and the remaining provinces/regions are all clustered into one category. This clustering result can be explained by the differences in epidemic control strategies among the provinces/regions: HB is the center of the COVID-19 breakout, with a large number of infection cases; underpopulated XZ and QH are both located on the Qinghai Tibet Plateau, with only a few infection cases; MO, HK, and TW are of self-governance: meaning their epidemic control strategies are different from all other regions in China. The model-based clustering method (Fratey & Raftery , 2002) suggests both HK and TW are to be in one cluster, which may not be correct.

(2) As shown in COVID-19 cases in the world, the 33 selected countries are clustered into 8 categories. Specifically, China (CN), Korea (KR), Japan (JP), Spain (ES), and Turkey (TR) are individually clustered into separate categories; Italy (IT) and Iran (IR) are clustered into one category; the United States of America (US), Germany (DE), France (FR), the United Kingdom of Great Britain (UK), Northern Ireland (GB), and Canada (CA) are clustered into one category; and the remaining countries are all clustered into one category. This clustering result is partly based on the timing of COVID-19 outbreaks in those countries. For example, the first large-scale outbreak was in CN, followed by KR

16

and JP. After that, infections in IR and IT experienced rapid growth, followed by the outbreaks in European countries and the US. Finally, the epidemic spread worldwide. In addition, the clustering is also based on the epidemic control strategies in each country. For example, in KR and JP, even while the epidemic broke out around the same time, the two countries had taken different strategies: JP adopted a "defensive strategy"' to ensure the health care system operated normally as usual, while KR used an "aggressive attack strategy" to comprehensively detect infections.

4.2   Segmentation and fitting

Based on the BIC-based ICSS Algorithm, we segment the curve time series and present the segmented fittings and confidence interval estimation for the log-transformed infection counts $Z_{i,t}(1 \leq i \leq N)$ of each cluster in China and the rest of the world; see Fig. 5 and 6, respectively.

We can obtain that all sigmoid curves share the form of multiple stages and multiple change points, with the exception of Cluster 7 (XZ) in China, with only one infection; the calculated change points of each cluster can still be explained by the differences in epidemic control strategies. See the details below.

(1) As shown in Fig. 5 A and Fig. 6 A, the sigmoid curves and change points are almost the same because HB province was the center of the COVID-19 outbreak in CN. In Fig. 6 A in CN, the first segment (19/12/01 to 19/12/13) was the germination period of the outbreak. In the second segment (19/12/13 to 20/01/16), COVID-19 seemed to have been controlled in CN. However, because many COVID-19 cases had not been found due to varied epidemic control strategies in the previous two stages, COVID-19 broke out in the third segment (20/01/16 to 20/01/26) and fourth segment (20/01/26 to 20/02/11) in CN. This coincided with Chunyun (the annual massive movement of people during

17

Chinese Lunar New Year), which particularly accelerated the outbreak. Finally, in the last two segments (20/02/11 to 20/02/27 and 20/02/27 to 20/04/20), COVID-19 was controlled and stabilized once the CN government implemented very strict epidemic control strategies, such as traffic control and home quarantine.

(2) As shown in Fig. 5 C, the sigmoid curves in HK and TW seem similar because they were both strongly affected by COVID-19 cases from mainland China. However, we find that the change points of COVID-19 in TW are about a week delayed compared to those in HK after COVID-19 started to break out in both regions. This is because TW responded in a timely manner to the COVID-19 outbreak and controlled it more quickly and effectively than HK, while the implementation of epidemic control strategies in HK lagged behind.

(3) As shown in Fig. 6, the number of new cases in China had tentatively stabilized since the last change point, 20/02/27, which was delayed by about one week in other clusters. In Fig. 6 A and B, the infections in CN and KR are mostly stable, but the epidemic situations in other countries have not been controlled effectively. Take the fifth cluster (Fig. 6 C) as an illustration, considering that this cluster had the fastest growth. The four segments can be explained as follows: (i) the infections in the first segment were mainly from oversea imports; (ii) in the second segment, COVID-19 seemed to have been controlled; (iii) COVID-19 broke out because of many unfound COVID-19 cases in previous segments; and (iv) in the last segment, COVID-19 began to come under control as governments declared states of emergency and started implementing strict measures to control the spread of the virus.

(4) As shown in Fig. 6 B, confidence intervals for KR tended to be quite narrow in width when the number of new cases had tentatively stabilized, resulting in more precise estimates of mean response, whereas confidence intervals for JP tended to be wide since JP

had adopted a "defensive strategy". In most of cases, confidence intervals produced precise results.

5   Discussion and Conclusions

A clustering-segmented autoregressive sigmoid model is developed to explore the space-time pattern of the log-transformed infectious count by the end of April 20, 2020. It performed well when it was applied to COVID-19 cases in both China and the 33 countries, and thus provides an efficient statistical model of COVID-19 spread to help fight against the virus. Currently, the infections in China are mostly stable, and the graph-based clustering algorithm is robust to the clusters from the 34 provinces/regions in China. When COVID-19 began to come under control, the clustering of the disease globally will become increasingly stable.

In fact, the CSAS model can adapt to an extended period of time when clusters have been updated and new change points have been identified. To do so, we use the last change points in time, 20/03/07 obtained from Fig. 5 or 2020/03/08 obtained from Fig. 6, as the start of the extended period at two-month intervals, from 20/03/07 to 20/05/07 or 20/03/08 to 20/05/08. In Fig 9, we show segmentations and fittings for log-transformed infection counts of each cluster in both China and the 33 countries during this extended period. We can see that the fittings continue to work well. We provide an R package, *GraphCpClust*, which can be accessed from https://github.com/Meiqian-Chen/GraphCpClust. From this R package, users can obtain the same results presented in this paper and can model data for another extended period of time. In addition, the data and code for another two papers (Shi, Wu and Rao, 2017, 2018) are included in this R package.

Regarding the dataset used in this article, Wuhan-2019-nCoV, we make the follow-

ing additional remarks: 1. Back in early March 2020, there were very few datasets on COVID-19, and especially few datasets containing timely epidemic data from each Chinese province. This dataset, Wuhan-2019-nCoV, collects national outbreak reports from WHO, as well as daily outbreak reports from provincial health and family planning commissions in China; 2. The Wuhan-2019-nCoV dataset is very timely updated and has been included in the "Open Source Wuhan" data resource. Therefore, we believe that the data quality of the Wuhan-2019-nCoV dataset is trustworthy. There are now more and more COVID-19 data resources available, such as WHO data (https://covid19.who.int/) and Our World in Data (https://ourworldindata.org/covid-data-switch-jhu). For these two datasets, we find that our model still works very well. Please see this webpage, http://graph-clustering-system.com/, for the three data analyses described above.

References

AKAHO, S. (1995). Mixture model for image understanding and the EM algorithm, *https://staff.aist.go.jp/s.akaho/papers/ETL-TR-95-13E.pdf*.

BAI, J. S. & PERRON, P. (2003). Computation and analysis of multiple structural change models. *J APPL ECONOM* **18(1)**, 1-22.

BISWAS, M., MUKHOPADHYAY, M. & GHOSH, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* **101**, 913-926.

CSARDI, G. & NEPUSZ, T.(2006). The igraph software package for complex network research. *InterJournal*, Complex Systems 1695. 2006. http://igraph.org. Accessed April 1, 2020.

DIETZA, K., HEESTERBEEK, J. A. P. (2002). Daniel Bernoulli's epidemiological model revisited. *MATH BIOSCI* **180**, 1-21

ESTER, M., KRIEGEL, H. P., SANDER, J. & XU, X.W.(1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96 Proceedings*, 226-231.

FRALEY, C.& RAFTERY, A.E.(2020) mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. R package version 2.2-5 . *Available at https://cran.r-project.org/web/packages/fpc/index.html.*

FRATEY, C.& RAFTERY, A.E. (1993). Model-based clustering, discriminant analysis and density estimation. *J AM STAT ASSOC* **97/458**, 611-631.

INCLÁN, C. & TIAO, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. Publications of the American Statistical Association. *J AM STAT ASSOC*,**89(427)**, 913-923.

JIN, B. S., WU, Y. H., RAO, C. R. & HOU, L.(2020). Estimation and model selection in general spatial dynamic panel data models. *Proc Natl Acad Sci* **117**, 5235-5241.

KATUL, G. G., MRAD, A., BONETTI, S., MANOLI, G., & PAROLARI, A. J. (2020). Global convergence of COVID-19 basic reproduction number and estimation from early-time SIR dynamics. *medRXiv.*

KOWSAR, R., KESHTEGAR., B., MAREY, M. A., & MIYAMOTO, A. (2017). An autoregressive logistic model to predict the reciprocal effects of oviductal fluid components on in vitro spermophagy by neutrophils in cattle. *SCI REP-UK* **7(1)**, 4482.

LIU, X.Z., & STECHLINSKI, P.(2017). Infectious Disease Modeling. *A Hybrid System Approch*, Springer, Heidelberg.

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H.(2008). Introduction to Information Retrieval, Cambridge University Press, Cambridge, England.

MURRAY, J. D.(1989) Mathematical Biology. Springer, Heidelberg.

MURTAGH, F. & LEGENDRE, P.(2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J CLASSIF* **31(3)**, 274-295.

R CORE TEAM(2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Accessed April 1, 2020.

SHI, X.P. , WANG, X.S., WEI, D.W. & WU, Y.H.(2016). A sequential multiple change-point detection procedure via VIF regression. *Comput Stat* **31**, 671-91.

SHI, X.P. , WU, Y.H. & RAO, C.R. (2017). Consistent and powerful graph-based change-point test for high-dimensional data. *Proc Natl Acad Sci* **114**, 3873-8.

SHI, X.P., WU, Y.H. & RAO, C.R. (2018). Consistent and powerful non-Euclidean graph-based change-point test with applications to segmenting random interfered video data. *Proc Natl Acad Sci* **115**, 5914-5919.

TOCHER, K. D. (1963). The Art of Simulation. *English University Press*, London.

WANG, J. A. & HARTIGANM, A.(1979). Algorithm as 136: a k-means clustering algorithm. *J ROY STAT SOC A STA* **28(1)**, 100-108.

YANG, ET AL. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J Thorac Dis* **12**, 165-174.
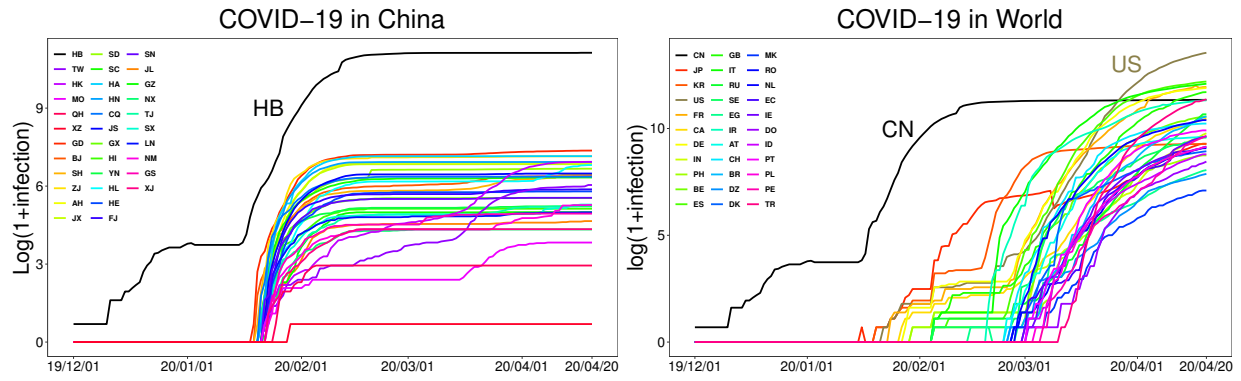
Figure 1: Plots of log-transformed infectious counts from December 1, 2019 to April 20, 2020 in China and in 33 selected countries. The data origin is from https://github.com/canghailan/Wuhan-2019-nCoV. The Alpha-2 codes applied here for China's provinces/regions and countries come from https://www.iso.org/obp.

Figure 2: Plots of S-curves for three classes.

24

Algorithm 1: Graph-based clustering Algorithm

Result: Output the optimal clusters $\mathcal{A}$.

1 Notations: $x(\theta) = \{x_s | x_s \leq \theta, s = 1, \ldots, K-1\}$ where x_s is defined in (7);

2 $\theta_{(s)}$ is the s'th largest element in set $\{x_s, s = 1, \ldots, K-1\}$;

3 $\hat{\sigma}^2(\theta)$ is the sample variance of $x(\theta)$;

4 $\mathrm{BIC}(\theta, \mathcal{A}) = (K-1)\log(\hat{\sigma}^2(\theta)) + 2L(\mathcal{A})\log(K-1)$, where $L(\mathcal{A})$ is the number of clusters in $\mathcal{A}$;

5 Initialize: Let $i = 1$, $L = 1$, and $\mathcal{A} = \{\mathcal{A}_1\}$ where $\mathcal{A}_1 = \{1, \ldots K\}$;

6 for $s = 2;\ s < K-1;\ s = s+1$ **do**

7   Let θ be $\theta_{(s)}$ and calculate the clusters based on $\mathcal{E}^*(\mathcal{P}, \theta)$ in (3) denoted as $\mathcal{A}_{\mathrm{temp}}$;

8   **if** $\mathrm{BIC}(\theta, \mathcal{A}_{temp}) < \mathrm{BIC}(\theta_{(s-1)}, \mathcal{A})$ **then**

9 $\mathcal{A} = \mathcal{A}_{\mathrm{temp}}$;

10   **else**

11   break;
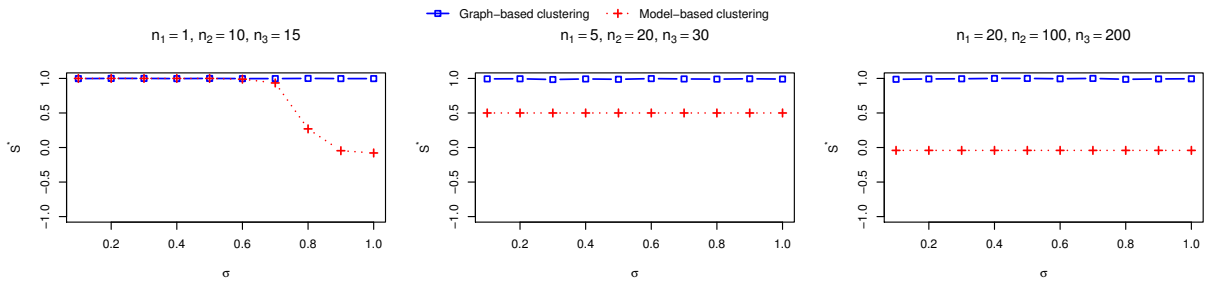
12   **end**

13 end

Figure 3: Comparisons of graph-based clustering method and model-based clustering method.

25

Algorithm 2: BIC-based ICSS Algorithm

    Result: Output the estimated change points $\hat{C}_i$.

1 **Notations:** $\hat{C}_{i,(s)}$ and $|\hat{C}_i|$ are the sth smallest element and number of elements of set $\hat{C}_i$, respectively;

2 **Initialization:** Let $t_- = 1$, $t_+ = T$, and $\hat{C}_i = \{0, T\}$;

3 **while** $t^+ - t_- > \Delta$ **do**

4 $t_{\text{first}} \leftarrow t^+$; $t_{\text{last}} \leftarrow t_-$;

5   **while** $\text{BIC}_{i,0}(t_-, t_{\text{first}}) \geq \text{BIC}_{i,1}(t_-, t_{\text{first}})$ **do**

6 $t_{\text{first}} \leftarrow \hat{t}_{i,t_-,t_{\text{first}}}$;

7   **end**

8   **while** $\text{BIC}_{i,0}(t_{\text{last}}, t^+) \geq \text{BIC}_{i,1}(t_{\text{last}}, t^+)$ **do**

9 $t_{\text{last}} \leftarrow \hat{t}_{i,t_{\text{last}},t^+}$;

10   **end**

11   **if** $t_{\text{first}} = t_{\text{last}}$ **then**

12 $\hat{C}_i \leftarrow \hat{C}_i \cup \{t_{\text{first}}\}$; break;

13   **else**

14 $\hat{C}_i \leftarrow \hat{C}_i \cup \{t_{\text{first}}, t_{\text{last}}\}$; $t_- \leftarrow t_{\text{first}}$; $t^+ \leftarrow t_{\text{last}}$;

15   **end**

16 **end**

17 **for** $s = 2$; $j < |\hat{C}_i|$; $s = s + 1$ **do**

18   **if** $\text{BIC}_{i,0}(\hat{C}_{i,(s-1)} + 1, \hat{C}_{i,(s+1)}) \leq \text{BIC}_{i,1}(\hat{C}_{i,(s-1)} + 1, \hat{C}_{i,(s+1)})$ **then**

19 $\hat{C}_i \leftarrow \hat{C}_i \setminus \{\hat{C}_{i,(s)}\}$;

20   **end**

21 **end**

22 $\hat{C}_i \leftarrow \hat{C}_i \setminus \{0, T\}$;
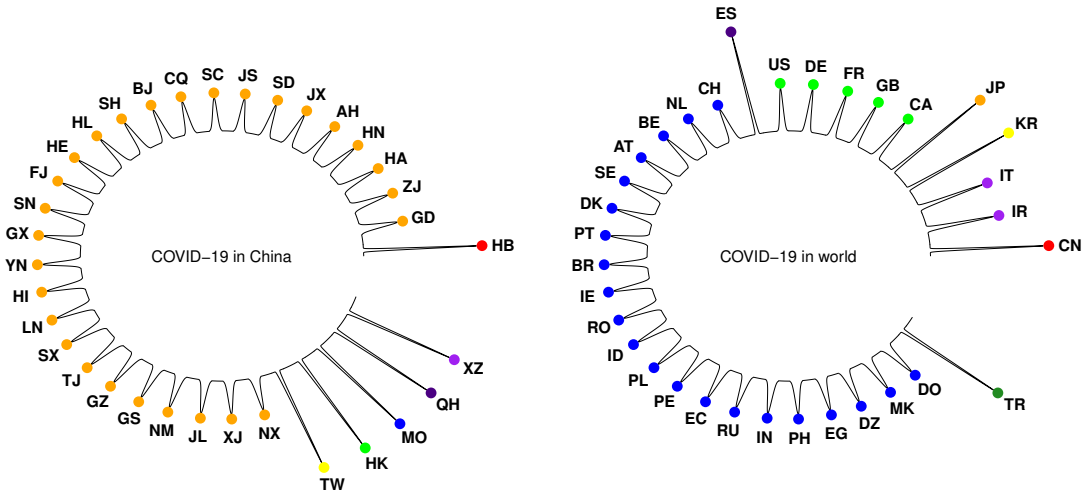

Figure 4: Plots of clusters in China and in 33 selected countries based on the log-transformed infection counts.
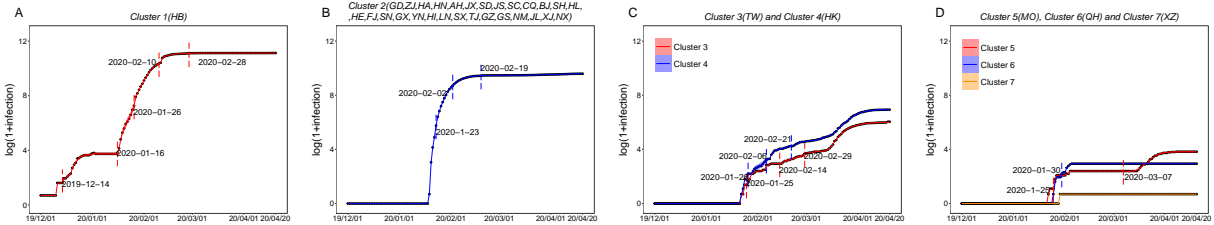

Figure 5: Plots of segmentations and fittings of provinces/regions in China based on the log-transformed infection counts.
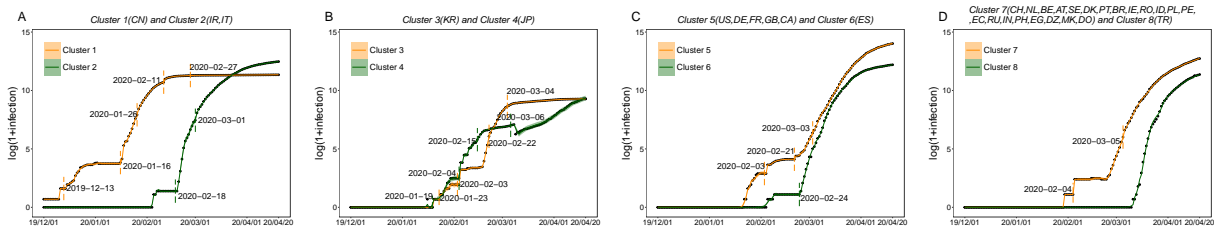
Figure 6: Plots of segmentations and fittings of 33 selected countries in the world based on the log-transformed infection counts.
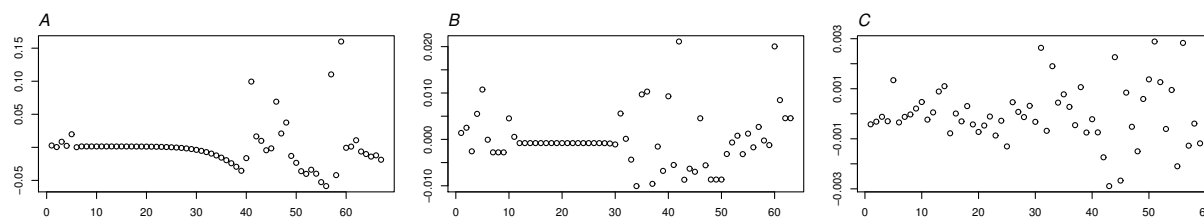

Figure 7: Plots of residuals of the last segment from the CSAS model for two separate provinces in China, NM (A) and TJ (B), and their common cluster 2 (C).
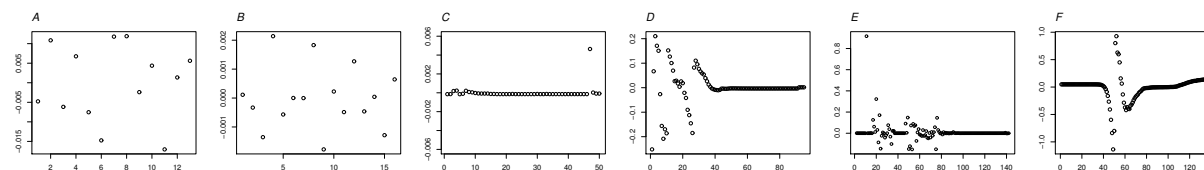

Figure 8: Plots of residuals for HB in China for the fourth segment (A), the fifth segment (B), and the last segment (C) from the CSAS model, for the last segment from the CSAS model without autoregressive terms (D), for the whole period from the autoregressive model of order 2 after taking the first difference (E), and for the whole period from the CSAS model without change points (F).
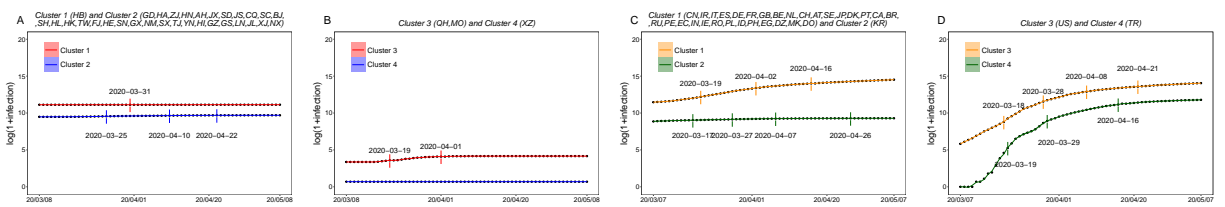
28

Figure 9: Plots of segmentations and fittings of each cluster in China (A-B) and in 33 selected countries (C-D) based on the log-transformed infection counts during the two-month extended period.