Journal of Data Science 20 (1), 79–94
January 2022
DOI: 10.6339/22-JDS1036
Data Science in Action

Clustering US States by Time Series of COVID-19 New Case Counts in the Early Months with Non-Negative Matrix Factorization

Jianmin Chen[1] and Panpan Zhang[2,*]

[1]*Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A.*
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*

Abstract

The spreading pattern of COVID-19 in the early months of the pandemic differs a lot across the states in the US under different quarantine measures and reopening policies. We proposed to cluster the US states into distinct communities based on the daily new confirmed case counts from March 22 to July 25 via a nonnegative matrix factorization (NMF) followed by a k-means clustering procedure on the coefficients of the NMF basis. A cross-validation method was employed to select the rank of the NMF. The method clustered the 49 continental states (including the District of Columbia) into 7 groups, two of which contained a single state. To investigate the dynamics of the clustering results over time, the same method was successively applied to the time periods with an increment of one week, starting from the period of March 22 to March 28. The results suggested a change point in the clustering in the week starting on May 30, caused by a combined impact of both quarantine measures and reopening policies.

Keywords *change point; COVID-19; k-means clustering; non-negative matrix factorization*

1 Introduction

The COVID-19 pandemic has brought a great deal of challenges not only to the global public health system but also to the global economy. The whole world was hit hard by surprise in the early months of the pandemic. By July 25, 2020, there had been a total of 15,590,252 confirmed cases across the world, while the number of death cases had reached 630,294. The United States (US), one of the countries that have been most severely attacked by COVID-19, reported 4,009,808 confirmed cases and 143,663 death cases, respectively, by then (https://covid19.who.int). In the meantime, the economic crisis caused by the pandemic was unprecedented in its scale. The Missouri Economic Research and Information Center reported that the total value of exports of all 50 continental US states had fallen by 29.8%, from $414.95 billion in the second quarter of 2019 to $291.47 billion in the second quarter of 2020 (https://meric.mo.gov/data/us-state-exports). The unemployment rate from the US Department of Labor reached 14.7% in April, 2020, doubling that reported at the end of 2008 (7.2%) when the subprime mortgage crisis broke out. Although the unemployment rate declined to 10.2% in July, 2020, thanks to the stimulating measures by the government, it remained at a high level.

*Corresponding author. Email: panpan.zhang@pennmedicine.upenn.edu.

Received January 23, 2022; Accepted January 25, 2022

Updated statistics on the pandemic and government policies are of genuine and critical public concern. From the beginning, the Centers for Disease Control and Prevention (CDC) has been continuously updating the progress of COVID-19 and the status of public health safety in the US in their Morbidity and Mortality Weekly Report (https://www.cdc.gov/mmwr/index.html). A large body of literature has reported clinical characteristics of the disease (e.g., Fauci et al., 2020; Moghadas et al., 2020; Goyal et al., 2020). For the general public, however, the case and death counts at the nation and state levels are of direct concerns. Although there were high volumes of confirmed and death cases reported in the US in March and April, the spread of the disease effectively slowed down in late April and early May owing to the practice of social distancing, "mask wearing" recommendations, "stay-at-home" and anti-gathering orders, and many other quarantine measures (e.g., Tian et al., 2020). However, the number of new confirmed cases in some of the US states has surged since early June, two weeks after the resumption of business in a majority of the states. In terms of the dynamic pattern of the case counts, some states are more similar than others possibly due to difference in quarantine and reopening policies.

We clustered the US states by the patterns in their COVID-19 case count series in the early four months of the pandemic. This period covers the first two nationwide waves by summer 2020. The results may be of interest for the authorities to reflect on their decisions in response to the pandemic. Such times series clustering analysis had not yet been reported in the literature of the study for COVID-19 series on the US states. It was only recently that the clustering of COVID-19 case count series was reported for European countries. For instance, D'Urso et al. (2021a) considered fuzzy clustering methods, D'Urso et al. (2021b) proposed a k-medoids algorithm by accounting for cluster noises, and Vitale et al. (2021) developed a spatio-temporal Bayesian network model for clustering. The most popular time series clustering methods are *feature-based* algorithms where features are first generated from the raw data and then appropriate clustering algorithms are applied to form clusters (Liao, 2005). Many studies on the combination of feature generating procedure and clustering algorithms have been done. For example, an anytime algorithm based on the Haar wavelet decomposition followed by a modified k-means clustering was proposed by Lin et al. (2004); an agglomeration algorithm incorporated with principals component analysis was introduced by Shaw and King (1992). An appealing alternative to solving time series clustering problems is to exploit functional data analysis (FDA) techniques, such as the k-centers functional clustering algorithm (Chiou and Li, 2007) and an EM-based algorithm for Gaussian mixture models (Chen and Maitra, 2015), the latter of which has been in a COVID-19 application (Li et al., 2022; Tang et al., 2022). See Jacques and Preda (2014) for a concise survey for functional data clustering methods.

Our time series clustering was based on a *non-negative matrix factorization* (NMF) followed by a k-means clustering procedure. Both NMF and k-means are standard methods, but their integration gives a new prospective to clustering time series. The series of each state was properly approximated by a linear combination of a small number of NMF bases. The coefficients of the bases were used as features in a k-means clustering analysis. This NMF-based method is attractive for its robustness, stability, and scalability (Brunet et al., 2004; Devarajan, 2008). In particular, we considered a weighted NMF method allowing for missing values in the dataset. The number of bases was determined through a cross-validation method where, facilitated by a stratification on the level of the data, a randomly selected fold of data was treated as missing and predicted by the rest. Our clustering results provide a platform to further study the similarities of the states that are in the same cluster and the dissimilarities of those that are not. Further, we investigated the changes in the clustering results as additional data became available with

an increment of one week successively. The temporal changes between the consecutive analyses were explored via similarity measures to identify possible change point(s) in the clustering of the states. Our analysis result suggests that the policies varying among the states may have impacted the spreading or confinement of the pandemic. The implementation of the present method is publicly available at GitLab (https://gitlab.com/covid-19-analysis/covid19-nmf.git).

The primary novelty as well as the main contributions of the paper are reflected in the application of the integrated method of NMF and k-means clustering to the COVID-19 case count data in the early months in the US. The rest of the manuscript is organized as follows. In Section 2, we introduce the data source, basic descriptive statistics, and data preprocessing procedures. In Section 3, we elaborate the fundamental concept of NMF, and propose an NMF-based algorithm followed by k-means clustering. The proposed algorithm is applied to the preprocessed data and the clustering results for different study periods are presented in Section 4. The temporal dynamic of cluster structures is investigated in Section 4 as well. Lastly, we address some concluding remarks and carry out some discussions in Section 5.

2 Data

Daily new confirmed cases from each US state were retrieved from a public repository maintained by the Center for Systems Science and Engineering at the Johns Hopkins University (Dong et al., 2020). The start date of our study period was set to be Sunday (March 22, 2020). President Trump declared a national emergency concerning the COVID-19 outbreak on March 13, 2020, after which the number of tests increased substantially. On average, it took about a week to get the results of nasal and swab tests. Hence, few confirmed cases were reported in, for example, the District of Columbia and West Virginia, before March 22. The end date of our study period was Saturday, July 25, 2020, when the spread of the pandemic appeared to be slowing down in those states affected largely by the second wave. The entire study period has 126 days, about 4 early months of the pandemic in the US.

We considered 49 continental state-level entities (including the District of Columbia), hereafter referred to as 49 states for simplicity. Alaska and Hawaii were not included because the population movements between the two states and the continent were rather limited. Since the 49 states vary a lot in population size, the raw counts of cases are not comparable across them. Thus, we standardized the raw counts by the population estimates of the states at the end of 2019 from The US Census Bureau (https://www.census.gov). Specifically, the standardized number of daily new confirmed cases for each state was recorded in terms of "per million" over the study period.

Figure 1 displays the raw daily new confirmed cases in four states with drastically different patterns — New York, North Carolina, Louisianan and Massachusetts — over the study period. Periodicity was observed in the series with a period of 7 days. The vertical red dashed lines in Figure 1 are presented with 7-day gaps. The 7-day period was possibly caused by the routine of tests available on different business days and testing report delivery by the public health departments. In each panel, the thick blue curve shows the 7-day *moving average* (with boundary adjustment) of the new confirmed cases per million in the state. The moving averages are much smoother than the raw counts, allowing us to observe the long-term trend in the time series and to make comparisons among the states more easily. The moving average series are the input data for our clustering analysis.

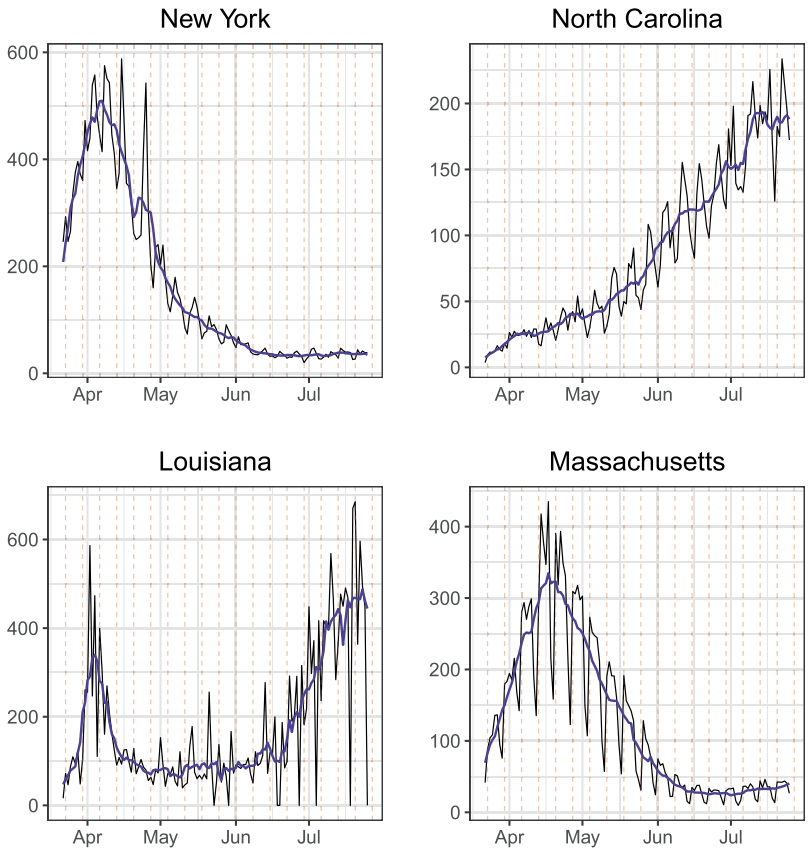

Figure 1: Daily confirmed cases in New York, North Carolina, Louisiana and Massachusetts over the study period, where the periodicity is reflected in vertical red dashed lines (7 days per gap).

3 Methods

Let (x_{ij}) be the input data matrix, $i = 1, 2, \ldots, 49$, $j = 1, 2, \ldots, 126$, where x_{ij} denotes the number of confirmed cases per million (7-day moving average) of state i on day j. With the starting date held on March 22, 2020, clustering can be done on each sub-period with incrementally extended ending date, in addition to the whole study period. The collected analyses on these sub-periods may help identifying an increment in time period that might have led to substantial changes in the clustering results. For a given time period, our clustering is carried out on the coefficients of NMF bases of the series. A potential change point in the cluster structure is suggested when the consistency measures of every two consecutive clustering outcomes are too low. The components of our method are presented next.

3.1 NMF

NMF is a class of matrix factorization methods where a high-dimensional non-negative matrix is approximated by the product of non-negative low-rank matrix factors. Consider an $n \times m$ non-negative matrix $\boldsymbol{X} := (x_{ij})$, where the rows and columns respectively represent observations and features. In our data analysis, we have $n = 49$ rows, each representing the time series of a state, and m columns, which is the number of days in the study period, each representing

one day of all the states in the time series. An NMF of $\boldsymbol{X}$ determines non-negative matrices $\boldsymbol{W} := (w_{ij})$ and $\boldsymbol{H} := (h_{ij})$ such that

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}, \tag{1}$$

where $\boldsymbol{W}$ and $\boldsymbol{H}$ are, respectively, $(n \times r)$ and $(r \times m)$ matrices with $r < m$ being the number of NMF bases of $\boldsymbol{X}$. In essence, each row of $\boldsymbol{H}$ is a basis vector of dimension m; each row of $\boldsymbol{X}$ is approximated by a linear combination of these bases, where the coefficients are stored in the rows of $\boldsymbol{W}$. A good approximation of $\boldsymbol{X}$ indicates that most of the variation in $\boldsymbol{X}$ is captured by the resultant linear combination of the bases. The number of bases r is usually much smaller than m. This dimension reduction procedure allows us to explore the $(n \times m)$ matrix $\boldsymbol{X}$ more efficiently through a low dimensional matrix $\boldsymbol{W}$. That is, the series of state i is represented by a coefficient vector of dimension r of the bases in the i-th row of $\boldsymbol{W}$.

The algorithm that we used specifically was the *weighted non-negative factorization* (WNMF) algorithm (Guillamet et al., 2001), which, as to be shown in the next subsection, facilitates the selection of the tuning parameter r. Let $\boldsymbol{V} := (v_{ij})$ be a weight matrix of the same dimension as $\boldsymbol{X}$, where v_{ij} is the weight of x_{ij}, reflecting its relative importance; specific definition of $\boldsymbol{V}$ is referred to Equation (6). The matrix approximation problem accordingly becomes

$$\boldsymbol{V} \odot \boldsymbol{X} \approx \boldsymbol{V} \odot (\boldsymbol{W}\boldsymbol{H}) \tag{2}$$

subject to the constraints of $w_{ij}, h_{ij} \geqslant 0$ for all i, j, where $\odot$ denotes the operator of element-wise product. When all the elements in $\boldsymbol{V}$ are ones, WNMF is reduced to the standard NMF.

Consider a squared error cost function

$$\|\boldsymbol{V} \odot (\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H})\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}\left(x_{ij} - \boldsymbol{w}_i^\top \boldsymbol{h}_j\right)^2, \tag{3}$$

where $\boldsymbol{w}_i^\top$ and $\boldsymbol{h}_j$ are respectively the i-th row of $\boldsymbol{W}$ and the j-th column of $\boldsymbol{H}$. Lee and Seung (2000) developed a multiplicative update rule that would minimize the cost function for NMF by iteratively updating $\boldsymbol{W}$ and $\boldsymbol{H}$. Wang et al. (2006) extended the multiplicative rule such that it would be applicable to WNMF. Let $t \in \mathbb{N}$ index the iteration process. At the t-th iteration, the updates of $\boldsymbol{W}^{(t+1)}$ and $\boldsymbol{H}^{(t+1)}$ from $\boldsymbol{W}^{(t)}$ and $\boldsymbol{H}^{(t)}$ are given by

$$\boldsymbol{H}^{(t+1)} \leftarrow \boldsymbol{H}^{(t)} \odot \frac{{\boldsymbol{W}^{(t)}}^\top (\boldsymbol{X} \odot \boldsymbol{V})}{{\boldsymbol{W}^{(t)}}^\top [(\boldsymbol{W}^{(t)}\boldsymbol{H}^{(t)}) \odot \boldsymbol{V}]}, \tag{4}$$

$$\boldsymbol{W}^{(t+1)} \leftarrow \boldsymbol{W}^{(t)} \odot \frac{(\boldsymbol{X} \odot \boldsymbol{V}){\boldsymbol{H}^{(t)}}^\top}{[(\boldsymbol{W}^{(t)}\boldsymbol{H}^{(t)}) \odot \boldsymbol{V}]{\boldsymbol{H}^{(t)}}^\top}, \tag{5}$$

where $\boldsymbol{X}^\top$ represents the transpose of $\boldsymbol{X}$. The alternating update goes on until the absolute difference between the maximum and the minimum costs over a number of successive iterations is below a preset tolerance. Equations (4) and (5) jointly guarantee that Equation (2) is non-increasing, and, consequently, that the algorithm at least converges to a local minimum (Lee and Seung, 2000). We used the WNMF algorithm implemented in R package *NMF* (Gaujoux and Seoighe, 2010).

Like any iterative algorithm, WNMF needs starting matrices $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$. We used the nonnegative double singular value decomposition (Boutsidis and Gallopoulos, 2008) to initialize the WNMF algorithm. In its basic form, this method selects the starting matrices that approximate the singular value decomposition of $\boldsymbol{X}$ by dropping the non-positive singular values and

replacing the negative entries in the unit-rank matrices formed by singular vectors with zeros. The algorithm does not grant randomization, which ensures reproducibility. It rapidly provides starting matrices with error almost as small as those from competing initialization approaches.

3.2 Rank Selection for NMF

It is critical to select an appropriate rank or the number of bases for NMF. We selected r through a cross-validation-based scheme extended from the method proposed by Kanagal and Sindhwani (2010). For an s-fold cross-validation, the entries in $\boldsymbol{X}$ are partitioned randomly into s sub-groups or folds. For a given r, each of fold is held out as testing data, while the remaining $(s-1)$ folds are used as training data with the held-out entries regarded as missing values to run a NMF. The mean squared prediction error (MSPE) of each held-out fold is computed by comparing the observed value with predictions using the NMF based on the training data. The total MSPE summed over all folds is used as the cross-validation criterion.

NMF for matrices with missing entries can be approached by a WNMF problem (Kim and Choi, 2009). Let the weight matrix $\boldsymbol{V}$ indicate the missing values in the held-out fold:

$$v_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ is observed;} \\ 0, & \text{if } x_{ij} \text{ is missing.} \end{cases} \tag{6}$$

The missing values can be predicted by the corresponding values in the approximation matrix $\boldsymbol{W}\boldsymbol{H}$ after implementing WNMF. Given a candidate set of $r \in \{2, 3, \ldots, 12\}$, where the maximum candidate value was about a quarter of the number of the states, the one with the minimum MSPE was selected for the value of r.

The right-skewness of the COVID-19 case counts in most states needs some care in WNMF. The data magnitudes in the right columns (later time) are much larger than those in the left (earlier time). A random partitioning of $\boldsymbol{X}$ may cause the data distribution in the testing fold significantly different from the counterparts in the training folds, potentially resulting in poor predictions. Having this in mind, we considered a stratified partitioning scheme. Specifically, all the entries of $\boldsymbol{X}$ were stratified into two strata based on the magnitudes, and a simple random sampling was executed within each stratum. The split of the two strata was set at the 75-th percentile of the observed data in the study period, giving the most stable results based on our experiments.

3.3 Clustering Procedure

Our clustering procedure based on WNMF has two steps, the first of which is dimension reduction with NMF. This step starts with rank selection of NMF as presented in the last subsection. For the selected rank r,

$$\boldsymbol{x}_i^\top \approx \boldsymbol{w}_i^\top \boldsymbol{H}, \tag{7}$$

where $\boldsymbol{x}_i^\top$ is defined analogously as $\boldsymbol{w}_i^\top$. Equation (7) implies that the (m-dimensional) features of the i-th subject in $\boldsymbol{X}$ are well approximated by the (r-dimensional) corresponding coefficients in $\boldsymbol{W}$, rendering a cluster analysis through $\boldsymbol{W}$.

The second step is to perform a clustering analysis using the basis coefficients $\boldsymbol{W}$ as input. A variety of well-developed clustering algorithms can be adopted, and our choice was the k-means clustering as available in R (Hartigan and Wong, 1979). We note that there are other available clustering methods (e.g., k-medoids and fuzzy clustering algorithms) that can be applied to $\boldsymbol{W}$.

Algorithm 1: Pseudo-algorithm of the NMF-based k-means clustering procedure for the daily COVID-19 case counts in 49 states.

Input: Raw COVID-19 daily COVID-19 case counts in 49 states
Output: Clustering results of the 49 states

```
Scale by population and apply 7-day moving average for each state to obtain X;
for r = 2 to 12 do
    Carry out cross-validation based on WNMF of X with rank r;
    Compute MSPE_r;
end
Select NMF rank r^(opt) = arg min_r{MSPE_r};
Obtain W based on NMF rank r^(opt);
for g = 2 to 10 do
    Carry out k-means for W with g clusters;
    Compute WSS for this g, WSS_g;
end
Select the number of clusters g^(opt) based on the elbow plot and practical
  interpretation;
return Result of k-means clustering for W with g^(opt) clusters;
```

However, the selection of clustering method is not the primary concern in the present study. Different combinations of basis coefficients represent different groups in both the pattern of the times series and their magnitudes. The algorithm depends on a random initialization. We used 500 random starting point and chose the result that gave the smallest within-cluster sum of squares (WSS). To select the number of clusters g, we used the elbow method which plots the WSS as a function of g, and chose g as the elbow of the curve. With 49 states, we considered candidate $g \in \{2, 3, \ldots, 10\}$. The determination of the elbow position may vary from case to case. Our selection considered both the curvature and the interpretation of the clustering results.

Algorithm 1 summarizes the major steps of the NMF-based k-means clustering algorithm for the daily COVID-19 new case counts from 49 states in a given time period.

3.4 Temporal Dynamics in Cluster Structure

Algorithm 1 can be applied to the new case count series of any time period, which facilitates an investigation of how the clustering results change over time as more data becomes available. We fixed the starting date at March 22, 2020 and considered a sequence of time periods with ending date starting from March 28, 2020, with an increment of one week until July 25, 2020. The incremental segment of 7 days is chosen according to the observed 7-day period in the data; see Figure 1. Such arrangement led to a total of 18 study periods. The clustering results over the 18 periods with incremental ending dates provided information about the temporal dynamics of the cluster structure.

At each ending date, we compared the obtained clustering results with or without the data from the most recent week. The agreement between the two clustering results was assessed by *adjusted rand index* (ARI, Hubert and Arabie, 1985). The ARI measurement was originally proposed to assess the accuracy of a clustering strategy by comparing its result with the ground truth. ARI ranges between -1 and 1, with higher values indicating that the clustering result is closer to the truth. We use it here to quantify the degree of agreement between a pair cluster

results. A small value of ARI suggests that the two clustering results are quite different from each other. A big drop after a sequence of high values may indicate a structural change in the clustering results.

4 Results

We first report the clustering results on the entire study period (from March 22 to July 25, 2020), then investigate the temporal dynamics in the clustering results, and finally compare the cluster structures before and after a possible change point.

4.1 Clustering Result for the Entire Study Period

Algorithm 1 was applied to the preprocessed time series data for the entire study period from March 22 to July 25, 2020. The NMF step yielded 12 bases as displayed in Figure 2. Most of the bases show distinctive features of peaks in terms of their timings and shapes. In chronological order, we see clearly a series of bases peaking with a size of about 10 or higher from early April to late July. The bases have different patterns in addition to their sizes. For example, basis 2 and 3 with early peaks are important for capturing high counts in early stage of the study period. Basis 2 is about a week later but goes down much more slowly than basis 3; basis 3 spans about 2 weeks while basis 2 spreads for about 5 weeks. Basis 1 and basis 11 with later peaks are important for capturing high counts in later stage of the study period, with basis 1 having an earlier climbing point and earlier peak than basis 11. Bases with less obvious interpretations and lower magnitude help properly adjust the ups and downs in describing the daily count curves of the 49 states.

The k-means step on the coefficients of the 12 bases grouped the 49 states into 7 clusters as summarized in Table 1, with a graphic visualization presented in the left panel of Figure 3. The means of the daily new case series per million (in population) for the 7 clusters are overlaid with the series themselves in each cluster in Figure 4. As expected, the curves in the same cluster appear similar in shape, but present different patterns or trends across the clusters. The states that are geographically close are likely to be in the same cluster due to the spillover effect. Other factors such as mobility, transport capacity, population density, and state-level social distancing measures may have been in effect, too, as discussed next.

In the sequel, we use the two-letter state abbreviations to represent the states for brevity in our discussion (https://www.ssa.gov/international/coc-docs/states.html).

There are two singleton clusters D and E, which contains AZ and LA, respectively, presenting completely different characteristics of the new case count curves as shown in Figure 4. The curve of AZ was flat early on, but started climbing at the beginning of June with a peak in the first week of July. Although a declining trend was observed since then, the counts had remained high until the end of the study period. The coefficient of basis 4 for AZ is outstanding, while the coefficient of basis 1 is relatively large as well, together precisely presenting the curve feature. On the other hand, the LA curve has two peaks, one in early April and the other in late July. It is the only bimodal curve in Figure 4. The coefficients of basis 1 and 3 for LA are both relatively greater than the counterparts of most of the other states, capturing the observed bimodal pattern.

The two states in Cluster A are NY and NJ whose new count case curves resemble each other throughout the study period. The early counts of daily new cases in NY and NJ were significantly higher than all of the other continental states. After peaking in early April, the

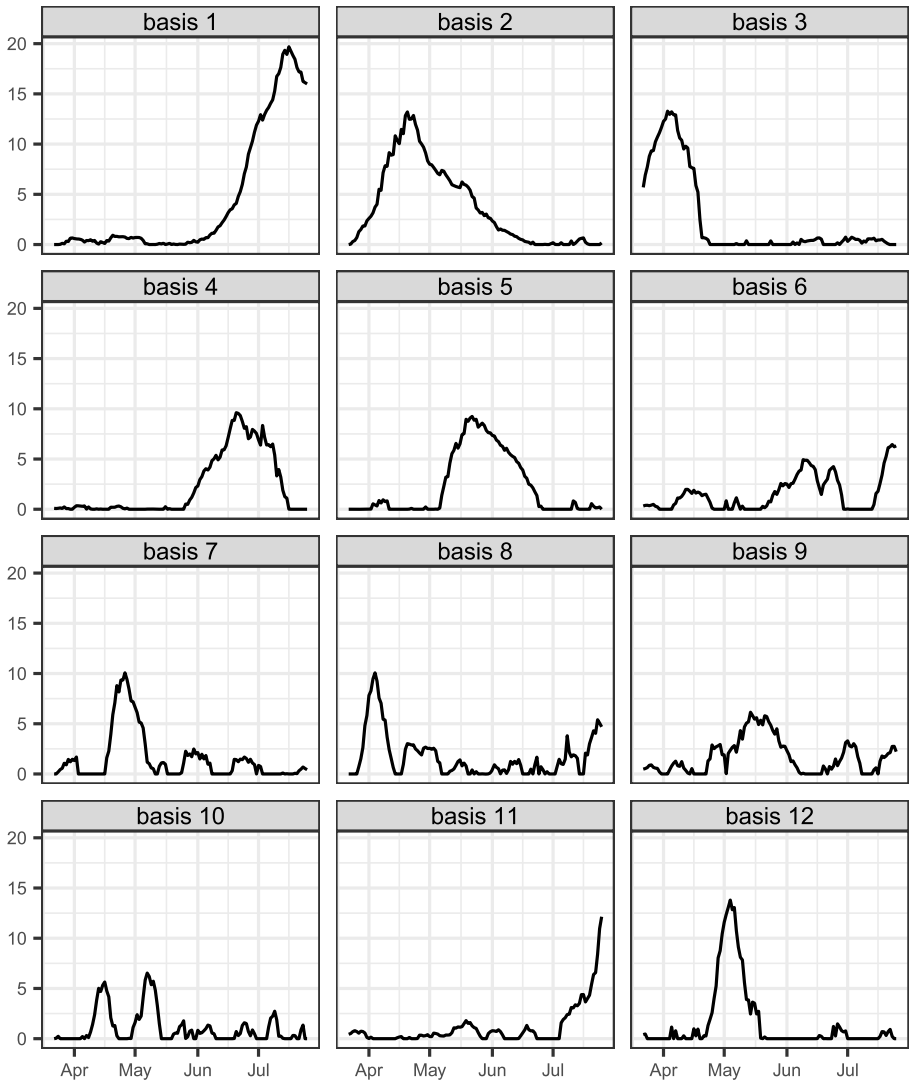

Figure 2: NMF bases for the proposed algorithm applied to the time series data from March 22 to July 25, 2020.

pandemic appeared to be well-controlled. The number of new cases went down and did not bounce back until July. Hence, the NY and NY curves are completely opposite to the AZ curve, reflected in relatively large coefficients of bases 2 and 3.

The four states in Cluster B, CT, DE, MA and RI, are all geographically close to NY and NJ. Their mean curve is similar to that of Cluster A in shape; both are unimodal and the peaks appear in the early times. The main difference among the curves in Clusters A and B is reflected in magnitude; the NY and NJ curves (in Cluster A) have been consistently higher over time. A minor difference is that the peaks in NY and NJ both occurred in early April, compared to the peaks of the four states occurring around late April. The four state curves did not show any tendency of rebound since May until the end of the study period, either.

Cluster C contains 7 states, DC, IA, IL, MD, MN, NE and VA. Their mean curve is flatter than those of Clusters A and B, and the values are smaller in magnitude in the early and mid

Table 1: Summary of the clustering results based on Algorithm 1 for the entire study period (March 22 to July 25, 2020).

Cluster	States
A	New Jersey, New York
B	Connecticut, Delaware, Massachusetts, Rhode Island
C	District of Columbia, Illinois, Iowa, Maryland, Minnesota, Nebraska, Virginia
D	Arizona
E	Louisiana
F	Alabama, Arkansas, California, Florida, Georgia, Idaho, Mississippi, Nevada, North Carolina, South Carolina, Tennessee, Texas, Utah
G	Colorado, Indiana, Kansas, Kentucky, Maine, Michigan, Missouri, Montana, New Hampshire, New Mexico, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, South Dakota, Vermont, Washington, West Virginia, Wisconsin, Wyoming

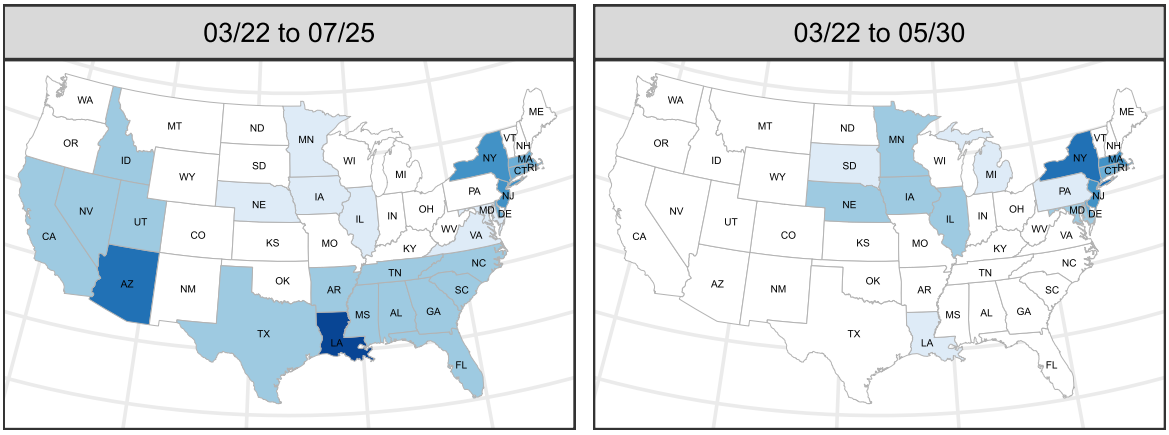

Figure 3: Clustering result for the entire study period (from March 22 to July 25, 2020) in the left panel; clustering result for the first half of the entire study period (from March 22 to May 30, 2020, where May 30, 2020 appears to be a potential change point of dynamics) in the right panel.

times. The peak of the average occurred in early May, followed by a decline until a second increasing period that started in July.

Cluster F contains 13 states with low counts of new cases from the start of the study period to June, followed by a continuing rise in June and July. The states form two big blocks respectively in the west (CA, NE, ID and UT) and south (AL, AR, FL, GA, MS, NC, SC, TN and TX). Among them, CA was one of the first states reporting confirmed cases, with several counties marked as hot spots. Most of the other states in this cluster keep the curves flat at the early and mid times. It is worthy of mentioning that the AZ curve is similar to the overall trend of the curves in Cluster F. However, the AZ curve is separated in a singleton group, possibly because it started increasing at the beginning of June, about two weeks prior to the climbing trend observed in the states in Cluster F.

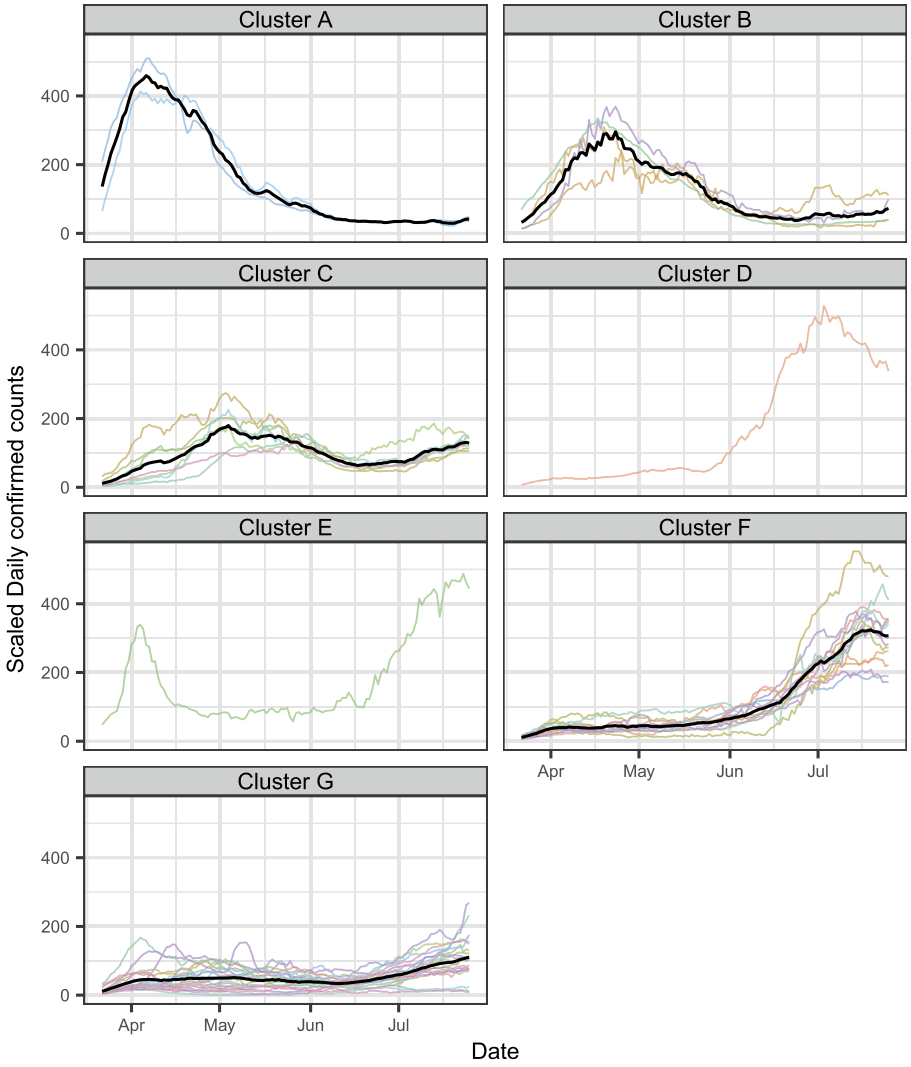

Figure 4: Daily confirmed case counts (after scaling and smoothing) of the states in each cluster; the mean curves are presented by black thick curves. Especially, each of the clusters D and E has one state member only.

The remaining 21 states are in Cluster G. Overall, these states have relatively low counts of new case every day, and their curves are flat throughout the study period. States in this cluster have either low population density or mandate strict quarantine measures.

4.2 Temporal Dynamic of Cluster Structure

In this section, we look into the temporal dynamics in cluster structure. The proposed clustering method was applied to 18 sub study periods (all starting from March 22) with end point successively incremented by one week, i.e., April 4, April 11, and so on. Let $T_1, T_2, \ldots, T_{18}$ indicate the 18 cumulatively incremental study periods in sequence. Table 2 summarizes the rank of NMF (denoted r) and the number of clusters (denoted g) of each study period. The number of bases r increases in general as the duration of the study extends. This is expected as longer

Table 2: The selected tuning parameters for 18 study periods with end point incremented by one week.

Study period	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
Rank of NMF (r)	2	5	4	6	6	8	7	8	8
Number of clusters (g)	5	5	5	6	5	5	6	6	6
Study period	T_{10}	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}	T_{16}	T_{17}	T_{18}
Rank of NMF (r)	7	8	8	12	11	12	11	12	12
Number of clusters (g)	6	7	7	8	6	7	7	7	7

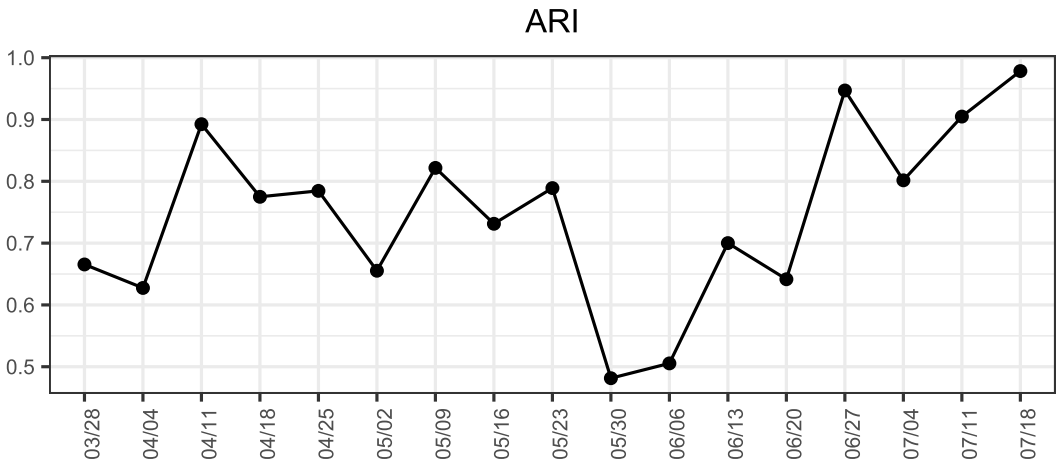

Figure 5: ARI for successive pairwise clustering results as the study period was extended by one week.

study period usually admits more complex curve patterns which require more bases to approximate. The number of clusters g started with 5 increased a little over time, and remained at 7 since T_{15}.

To identify possible structural changes in the clustering results, we checked the consistency in the resultant clusters as measured by ARI before and after each week (i.e., T_1 versus T_2, T_2 versus T_3 and so on). Figure 5 presents the ARI for each successive comparison between the clustering results from T_i versus those from T_{i+1}, $i = 1, 2, \ldots, 17$, where the horizontal axis represents the end point of the shorter time period in each pair. ARI has a moderate value at the very beginning, and vibrates in the early few weeks. A tremendous drop in T_{10} heralds a large degree of inconsistency. Hence, possible structural change in the clustering results may occur during the one-week periods before and after May 30. ARI remains relatively low until a rising in T_{12} as the week of June 20 is included. After a slight drop at T_{13}, ARI stays at a high level until the end of the study period.

The possible change in the clustering results as the study period up to the week of May 30 is appended echoes the findings in the clustering patterns in Figure 4. It was during the period from May 30 to June 20 that the rebound in many of the states in Clusters C, E, and G and the rise in the states in Clusters D and F occurred. Noticing the estimated mean of incubation period is around 5.34 days with 95% confidence interval [4.29, 6.40] (Zhang et al., 2020) and that

Table 3: Cluster details of the period from 03/22/2020 to 05/30/2020.

Cluster	States
A*	New York
B*	New Jersey, Massachusetts
C*	Connecticut, Delaware, Rhode Island
D*	District of Columbia, Illinois, Iowa, Maryland, Minnesota, Nebraska
E*	Louisiana, Michigan, Pennsylvania, South Dakota
F*	Alabama, Arizona, Arkansas, California, Colorado, Florida, Georgia, Idaho, Indiana, Kansas, Kentucky, Maine, Mississippi, Missouri, Montana, Nevada, New Hampshire, New Mexico, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, South Carolina, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming

the majority (97.5%) of the infected individuals developed symptoms within 11.5 days (Lauer et al., 2020), the actual change point took place in the week between May 17 and May 23.

4.3 Comparison with the Results from Data Before May 30

The study of the temporal dynamic of cluster structure in Section 4.2 induces a comparison between the clustering results including and excluding the data after May 30, respectively corresponding to T_{18} and T_{10} (March 22 to May 30, 2020). As shown in Table 2, there are 6 clusters for T_{10}, where the detailed clustering result is summarized in Table 3, followed by a graphic visualization given in the right panel of Figure 3.

Specifically, the ARI score between the clustering results of T_{18} and T_{10} is 0.36. The primary difference between the two clustering results is that Clusters D (AZ), F, and G (the majority) from T_{18} are grouped in one big cluster F* from T_{10}. As seen in Figure 4, the curves in Clusters D, F, and G are all flat over T_{10} in spite of several minor humps up to May 30. Secondly, LA, MI, PA, and SD are in one cluster (Cluster E*) from T_{10}, but from T_{18}, LA forms a singleton and MI, PA and SD belong to Cluster G. The four states had early peaks between April 7 and April 20, but the counts were significantly lower than that from NY, rendering them distinct from all the other states from T_{10}. Thirdly, NJ and NY are in the same cluster from T_{18}, but from T_{10}, NY itself forms a single-state cluster, and NJ and MA together form another cluster. Different from NY that increased from the beginning to the first week of April and then started declining right after, NJ had a long peak for almost two weeks in April. In addition, the magnitudes of the curves of NJ and MA were relatively close to each other throughout T_{10}, consistently smaller than that of NY. Cluster D* from T_{10} and Cluster C from T_{18} are similar, with the exception of VA in C but not in D*.

5 Discussion

Our clustering procedure is a combination of NMF and k-means clustering. NMF is used for dimension reduction and k-means is applied to the basis coefficients from the NMF. NMF itself is a clustering method in which the number of clusters is the rank and each data point is assigned to a cluster corresponding to the most highly expressed basis (Brunet et al., 2004). When the

basis is restricted to be orthogonal, NMF and k-means are equivalent in terms of the objective function (Ding et al., 2005). Our method is different from the traditional NMF clustering in that the clusters are formed by the similarities in basis coefficients. It has the advantage to capture patterns that are combinations of multiple bases, which may not be appropriate to be clustered to the groups represented by either one of the bases. Our method is also preferred to the standard k-means clustering because the k-means algorithm gives equal weights to all time points. However, data of early time points may have a much smaller magnitude. The difference in early time times may not contribute as much as the later times. As a result, states with distinctive patterns at early times such as NY, NJ and LA, cannot be successfully distinguished from other states by the standard k-means.

Our clustering of the 49 states appears to be reasonably reflecting the spread and control of the pandemic in the US up to July 25, 2020. The third wave since September 2020 as well as the differences in quarantine and business resumption measures are expected to reshuffle the clustering reported in this paper, which merits continuing research. Other factors that may be accounted include population mobility, vaccination information, virus mutation, spatial features of the states, among others. Some of the dynamic factors at the state-levels may be combined with the features constructed from the NMF in the clustering stage. Once the features are set, unsupervised learning approaches other than k-means could be fuller explored in a comparison study (Hirano et al., 2004; Gelbard et al., 2007; Madhulatha, 2011). In our state-level clustering, the spatial dependence among the states was discarded. At the county level, it would be less appropriate to do so. Incorporating spatial contiguity (Arumugadevi and Seenivasagam, 2015) in clustering county level data would be of great value in providing insights at a finer geographic resolution.

Supplementary Material

1. data_10_05.csv: This file contains the data from a public repository maintained by the Center for Systems Science and Engineering at the Johns Hopkins University (Dong et al., 2020). The data was retrieved on October 5, 2020. The case numbers may differ from those in the current version owing to possible modifications made after October 5, 2020.
2. nst-est2019-01.csv: This file contains the state-level population data, maintained by the US Census Bureau (https://www.census.gov). The data was released at the end of 2019.
3. pretreat.R: Codes for pre-processing the data (e.g., smoothing and scaling).
4. getnmfparameter.R: Codes for obtaining NMF ranks via the cross-validation method proposed in the paper.
5. model_fit.R: Codes for implementing the NMF method. The results of k-means clustering (including the selection of k) are given by running this file as well.
6. plotmaking.R: Codes for generating the figures in the paper.

References

Arumugadevi S, Seenivasagam V (2015). Comparison of clustering methods for segmenting color images. *Indian Journal of Science and Technology*, 8(7): 670.

Boutsidis C, Gallopoulos E (2008). SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4): 1350–1362.

Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12): 4164–4169.

Chen WC, Maitra R (2015). EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution. R Package. URL http://cran.r-project.org/package=EMCluster.

Chiou JM, Li PL (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(4): 679–699.

Devarajan K (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLOS Computational Biology*, 4(7): e1000029.

Ding C, He X, Simon HD (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In: *Proceedings of the 2005 SIAM International Conference on Data Mining* (H Kargupta, J Srivastava, C Kamath, A Goodman, eds.), 606–610. SIAM, Philadelphia, PA, USA.

Dong E, Du H, Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5): 533–534.

D'Urso P, De Giovanni L, Vitale V (2021a). Spatial robust fuzzy clustering of COVID-19 time series based on B-splines. *Spatial Statistics*, 100518. https://doi.org/10.1016/j.spasta.2021.100518.

D'Urso P, Mucciardi M, Otranto E, Vitale V (2021b). Community mobility in the European regions during COVID-19 pandemic: a partitioning around medoids with noise cluster based on space-time autoregressive models. *Spatial Statistics*, 100531. https://doi.org/10.1016/j.spasta.2021.100531.

Fauci AS, Lane HC, Redfield RR (2020). COVID-19—navigating the uncharted. *The New England Journal of Medicine*, 382(13): 1268–1269.

Gaujoux R, Seoighe C (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, 11: 367.

Gelbard R, Goldman O, Spiegler I (2007). Investigating diversity of clustering methods: an empirical comparison. *Data & Knowledge Engineering*, 63(1): 155–166.

Goyal P, Choi JJ, Pinheiro LC, Schenck EJ, Chen R, Jabri A, et al. (2020). Clinical characteristics of COVID-19 in New York City. *The New England Journal of Medicine*, 382(24): 2372–2374.

Guillamet D, Bressan M, Vitria J (2001). A weighted non-negative matrix factorization for local representations. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume I, 942–947. IEEE, Piscataway, NJ, USA.

Hartigan JA, Wong MA (1979). Algorithm as 136: a *k*-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108.

Hirano S, Sun X, Tsumoto S (2004). Comparison of clustering methods for clinical databases. *Information Sciences*, 159(3–4): 155–165.

Hubert L, Arabie P (1985). Comparing partitions. *Journal of Classification*, 2: 193–218.

Jacques J, Preda C (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8: 231–255.

Kanagal B, Sindhwani V (2010). Rank selection in low-rank matrix approximations: a study of cross-validation for NMFs. In: *Low-Rank Methods for Large-scale Machine Learning (Workshop in NIPS'10)* (A Gretton, M Mahoney, M Mohri, A Talwalkar, eds.). https://www.cs.umd.edu/~bhargav/nips2010.pdf.

Kim YD, Choi S (2009). Weighted nonnegative matrix factorization. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1541–1544. IEEE, Piscataway, NJ, USA.

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 172(9): 577–582.

Lee DD, Seung HS (2000). Algorithms for non-negative matrix factorization. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS'00)* (TK Leen, TG Dietterich, V Tresp, eds.), 535–541. MIT Press, Cambridge, MA, USA.

Li X, Zhang P, Feng Q (2022). Exploring COVID-19 in mainland China during the lockdown of Wuhan via functional data analysis. *Communications for Statistical Applications and Methods*. In press.

Liao TW (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11): 1857–1874.

Lin J, Vlachos M, Keogh E, Gunopulos D (2004). Iterative incremental clustering of time series. In: *Proceedings of the 9th International Conference on Extending Database Technology* (E Bertino, S Christodoulakis, D Plexousakis, V Christophides, M Koubarakis, K Böhm, E Ferrari, eds.), 106–122. Springer-Verlag, Berlin, Heidelberg, Germany.

Madhulatha TS (2011). Comparison between k-means and k-medoids clustering algorithms. In: *Advances in Computing and Information Technology* (DC Wyld, M Wozniak, N Chaki, N Meghanathan, D Nagamalai, eds.), 472–481. Springer, Berlin, Heidelberg, Germany.

Moghadas SM, Shoukat A, Fitzpatrick MC, Wells CR, Sah P, Pandey A, et al. (2020). Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 117(16): 9122–9126.

Shaw C, King G (1992). Using cluster analysis to classify time series. *Physica D: Nonlinear Phenomena*, 58(1–4): 288–298.

Tang C, Wang T, Zhang P (2022). Functional data analysis: an application to COVID-19 data in the United States. *Quantitative Biology*. In press. arXiv preprint: https://arxiv.org/abs/2009.08363.

Tian T, Tan J, Jiang Y, Wang X, Zhang H (2020). Evaluate the risk of resumption of business for the states of New York, New Jersey and Connecticut via a pre-symptomatic and asymptomatic transmission model of COVID-19. medRxiv preprint: https://doi.org/10.1101/2020.05.16.20103747.

Vitale V, D'Urso P, De Giovanni L (2021). Spatio-temporal object-oriented Bayesian network modelling of the COVID-19 Italian outbreak data. *Spatial Statistics*, 100529. https://doi.org/10.1016/j.spasta.2021.100529.

Wang G, Kossenkov AV, Ochs MF (2006). LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC bioinformatics*, 7: 175.

Zhang P, Wang T, Xie SX (2020). Meta-analysis of several epidemic characteristics of COVID-19. *Journal of Data Science*, 18(3): 536–549.