Quantitative Biology 2022, 10(2): 172–187
https://doi.org/10.15302/J-QB-022-0300

RESEARCH ARTICLE

Functional data analysis: An application to COVID-19 data in the United States in 2020

Chen Tang[1,†], Tiandong Wang[2,†], Panpan Zhang[3,†,*]

[1] Research School of Finance, Actuarial Studies & Statistics, The Australian National University, Canberra, ACT 2601, Australia
[2] Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[3] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA
* Correspondence: panpan.zhang@pennmedicine.upenn.edu

Received May 16, 2021; Revised June 28, 2021; Accepted July 11, 2021

***Background*: In this paper, we conduct an analysis of the COVID-19 data in the United States in 2020 via functional data analysis methods. Through this research, we investigate the effectiveness of the practice of public health measures, and assess the correlation between infections and deaths caused by the COVID-19. Additionally, we look into the relationship between COVID-19 spread and geographical locations, and propose a forecasting method to predict the total number of confirmed cases nationwide.**

***Methods*: The functional data analysis methods include functional principal analysis methods, functional canonical correlation analysis methods, an expectation-maximization (EM) based clustering algorithm and a functional time series model used for forecasting.**

***Results*: It is evident that the practice of public health measures helps to reduce the growth rate of the epidemic outbreak over the nation. We have observed a high canonical correlation between confirmed and death cases. States that are geographically close to the hot spots are likely to be clustered together, and population density appears to be a critical factor affecting the cluster structure. The proposed functional time series model gives more reliable and accurate predictions of the total number of confirmed cases than standard time series methods.**

***Conclusions*: The results obtained by applying the functional data analysis methods provide new insights into the COVID-19 data in the United States. With our results and recommendations, the health professionals can make better decisions to reduce the spread of the epidemic, and mitigate its negative effects to the national public health.**

Keywords: COVID-19; canonical correlation; cluster analysis; functional time series; forecasting; principal component analysis

Author summary: We study the COVID-19 time series data in the United States in 2020 via functional data analysis methods. We find that the practice of public health measures helps to reduce the spread of the epidemic, and there is a high canonical correlation between confirmed and death cases. The cluster structure at state level is closely related to the state geographical locations. Through a functional time series model, we are able to accurately predict the total number of infections in the nation relative to standard time series methods.

INTRODUCTION

Ever since December 2019 the coronavirus disease (COVID-19) has been spread to more than 180 countries over the world, leading to extremely negative impacts on global public health and economy. As of 08/25/2020, this virulent disease has caused a total of 23,721,008 confirmed and 815,029 death cases around the world, according to the data collected at Johns Hopkins University [1]. By the end of the study period (August

[†]These authors contributed equally to this work.

© The Author (s) 2022. Published by Higher Education Press

2020), the United States (US) is the most severely hit country with 5,759,147 confirmed and 177,873 death cases. According to the US Bureau of Labor Statistics, the unemployment rate of US has reached 14.7% in April 2020, a record-high over-the-month increase since January, 1948. Although the rate has declined to 8.4% in August 2020, it still remains at a relatively high level, compared to less than 4% before the pandemic.

A great amount of scientific efforts have been integrated to learn the progression of the disease, and to mitigate its negative impact on people's normal life. However, according to some recent research [2], it is evident that the COVID-19 has become mutating. As pointed out by Dr. Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases, on 06/24/2020, the vaccine for the COVID-19 would not likely to be available until 2021. On 07/27/2020, the National Institutes of Health has announced the phase 3 clinical trial of the investigational vaccine for the COVID-19. By now (August 2020) there is no medication that can directly eliminate the virus, many infected patients have to rely on their own immune systems to recover under the help of standard treatments. As more and more COVID-19 data become available, statisticians now commit to carrying out intensive data-driven analyses to precisely uncover the epidemic characteristics of the COVID-19.

In this paper, we exploit methods from the functional data analysis to analyze the COVID-19 data of both confirmed and death cases in the US from 01/21/2020 to 08/15/2020. In this study, our data are collected at state level (see Section of Preliminary Analysis), with the following research questions in mind:

(1) Does the practice of public health measures (*e.g.*, social distancing and mask wearing) help to mitigate the spread of the COVID-19? On the other hand, does the reopening of business exacerbate the spread of the disease?

(2) Is there any quantitative way to understand the correlation between infections and deaths caused by the COVID-19 in the US? Does the correlation vary from state to state?

(3) Is the spread of the COVID-19 related to the geographical locations of the infected regions or hot spots (at the state level) in the US?

(4) Is there a way to have some reasonable forecasts with regard to the total number of confirmed cases nationwide?

METHODS

In multivariate statistics, modes of variation are a set of vectors that are centered at mean, describing the variation in a population or sample. Typically, variation patterns are characterized via the standard eigenanalysis, *i.e.*, principal component analysis [3]. Analogously in the functional data analysis (FDA), modes of variation provide an efficient tool to visualize the variation of the functional curves around the mean function. Identifying modes of variation in functional data is usually done through the functional principal component analysis [4], providing new insights and precise interpretations of the functional data.

Functional principal component analysis

Consider a probability space $(\mathbb{R}_+, \mathcal{F}, \mathbf{P})$, and a compact interval $I \subset \mathbb{R}_+$. A stochastic process $X(\cdot)$ is called an L_2 process on I if

$$\mathbf{E}\int_I |X(t)|^2\, dt < \infty.$$

Let X_i, for $i = 1, 2, \ldots, n$, be independent realizations of the underlying L_2 process $X(\cdot)$. By convention in the FDA, functional data are given as

$$Y_i(t) = X_i(t) + \varepsilon_i(t), \tag{1}$$

for $t \in \mathcal{T}_i := \{t_{ij},\ j = 1, 2, \ldots, n_i\}$, a time schedule for subject i. The terms $\varepsilon_i(t_{ij})$'s are independent measurement errors with $\mathbf{E}(\varepsilon_i(t_{ij})) = 0$ and $\mathrm{Var}(\varepsilon_i(t_{ij})) = \sigma_{ij}^2$ for some constant σ_{ij}.

The functional principal component analysis (FPCA) is a powerful tool for dimension reduction in the FDA [4]. In essence, FPCA is an expansion of the realization $X_i(t)$ into functional bases consisting of the eigenfunctions of the variance-covariance structure of the process $X(\cdot)$, where the eigenfunctions are required to be orthogonal. Let $\{\phi_k : k \geqslant 1\}$ denotes the collection of orthogonal eigenfunctions. In addition, let $\mu(t) = \mathbf{E}(X(t))$ be the true mean function, and $\xi_{ik} := \int_I (X_i(t) - \mu(t))\phi_k(t)\, dt$ be the k-th functional principal component (FPC) of X_i (also called scores in the jargon). By the Karhunen-Loève Theorem [5, 6], we have

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t). \tag{2}$$

Due to the difficulties in estimating and interpreting the infinite sum in Eq. (2), a conventional treatment is to approximate it by a finite sum of K terms. In what follows, we set

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t) + \upsilon_i(t),$$

where $\upsilon_i(t)$ is the counterpart of $\varepsilon_i(t)$ in Eq. (1) owing to truncation. We refer interested readers to [7] for a comprehensive review of the fundamental theory of FPCA. One primary application of FPCA is to explore

© The Author (s) 2022. Published by Higher Education Press

modes of variation [8] for the functional data, reflecting the percentage of total variations contributed by each principal eigenfunctions.

When applying the FPCA method to the COVID-19 data of 50 states in the US, we notice that they are not identical in time schedules. The identification of the first COVID-19 case may differ by as long as 30 days across the country. Having observed the sparsity in the functional data during the early stage of the outbreak, we adopt the principal components analysis through conditional expectation (PACE) approach proposed in [9] for our analysis. The estimation of the FPC score of $\boldsymbol{Y}_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))^{\mathsf{T}}$, for $i = 1, 2, \ldots, n$, via PACE takes the following major steps:

(1) For each $\boldsymbol{Y}_i$, estimate the mean function $\boldsymbol{\mu}_i = (\mu(t_{i1}), \ldots, \mu(t_{in_i}))^{\mathsf{T}}$ and the covariance structure $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i} = \mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_i)$ by locally linear scatter and surface smoothers, respectively [10];

(2) Discretize the off-diagonal smoothed covariance to estimate the eigenfunctions $\boldsymbol{\phi}_{ik} = (\phi_k(t_{i1}), \ldots, \phi_k(t_{in_i}))^{\mathsf{Tc}}$, $k = 1, 2, \ldots, K$, and the corresponding eigenvalue λ_k [11];

(3) Adopt an Akaike information criterion (AIC) type criterion to select the number of eigenfunctions, *i.e.*, K, needed to approximate the process [12];

(4) Lastly, the estimation of the FPC score for the i-th subject is given through the conditional expectation:

$$\hat{\xi}_{ik} = \mathbf{E}(\xi_{ik} \,|\, \boldsymbol{Y}_i) = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^{\mathsf{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i), \tag{3}$$

where $\hat{\lambda}_k$, $\hat{\boldsymbol{\phi}}_{ik}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}$ and $\hat{\boldsymbol{\mu}}_i$ are respectively the estimates of λ_k, $\boldsymbol{\phi}_{ik}$, $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$ and $\boldsymbol{\mu}_i$.

In practice, the **R** package **fdapace** [13] allows us to apply the PACE method to the functional COVID-19 data directly. The computation results and corresponding discussions are given in the subsequent section.

Functional canonical correlations analysis

In statistics, canonical correlation analysis is a tool to make inference based on the cross-covariance matrix of multiple datasets. The analogue in FDA, named functional canonical correlation analysis (FCCA), aims to investigate the correlation shared by multiple functional datasets [14]. Specifically, here we use the FCCA method to quantitatively estimate the correlation between confirmed and death cases across the states.

We first outline the theoretical setup of the analysis method [15, 16], and refer the interested readers to [17] for a concise review. Let $X(\cdot)$ and $Y(\cdot)$ be two L_2 processes on two compact intervals, I_X and I_Y, respectively. In addition, let $\mathcal{H}_2(I)$ denote the Hilbert space of square-integrable functions on some compact interval I, with respect to Lebesgue measure. In addition, the notation $\langle \cdot, \cdot \rangle$ refers to the standard operation of inner product. By definition, the first canonical correlation is given by

$$\rho_1 = \sup_{u \in \mathcal{H}_2(I_X), v \in \mathcal{H}_2(I_Y)} \mathrm{Cov}(\langle u, X \rangle, \langle v, Y \rangle) = \mathrm{Cov}(\langle u_1, X \rangle, \langle v_1, Y \rangle),$$

subject to

$$\mathrm{Var}(\langle u, X \rangle) = 1 \qquad \text{and} \qquad \mathrm{Var}(\langle v, Y \rangle) = 1. \tag{4}$$

The pair of the optimal solutions (u_1, v_1) are called the canonical weight functions of ρ_1. For $k \geqslant 2$, the k-th canonical correlation coefficient, denoted ρ_k, is defined under the condition of orthogonality:

$$\mathrm{Cov}(\langle u_k, X \rangle, \langle u_j, X \rangle) = \mathrm{Cov}(\langle v_k, Y \rangle, \langle v_j, Y \rangle) = 0,$$
$$j = 1, 2, \ldots, k-1.$$

Set ρ_k, $k \geqslant 2$, and its associated weight functions (u_k, v_k) in an analogous way:

$$\rho_k = \sup_{u \in \mathcal{H}_2(I_X), v \in \mathcal{H}_2(I_Y)} \mathrm{Cov}(\langle u, X \rangle, \langle v, Y \rangle) = \mathrm{Cov}(\langle u_k, X \rangle, \langle v_k, Y \rangle),$$

subject to condition (4). In FDA, the inner products $\langle u, X \rangle$ and $\langle v, Y \rangle$ corresponding to weight functions u and v are called probe scores. Typically for $k \geqslant 1$, we refer to (u_k, v_k) as the canonical weight pair that optimizes the canonical criteria for ρ_k.

The FCCA approach, in essence, determines the projection of X in the direction of u_k as well as that of Y in the direction of v_k, such that their linear combinations are maximized for $k \geqslant 1$. Consider three kinds of variance and cross-covariance operators $\Sigma_{XX} : \mathcal{H}_2(X) \mapsto \mathcal{H}_2(X)$, $\Sigma_{YY} : \mathcal{H}_2(Y) \mapsto \mathcal{H}_2(Y)$ and $\Sigma_{YX} : \mathcal{H}_2(Y) \mapsto \mathcal{H}_2(X)$, which correspond to the variance structure of X, the variance structure of Y and the cross-covariance structure between X and Y, respectively. We then rewrite the definition expression of ρ_1 as follows:

$$\rho_1 = \sup_{u \in \mathcal{H}_2(I_X), v \in \mathcal{H}_2(I_Y)} \langle u, \Sigma_{YX} v \rangle, \tag{5}$$

subject to $u^{\mathsf{T}} \Sigma_{XX} u = v^{\mathsf{T}} \Sigma_{YY} v = 1$. Similar arguments can also be applied to $\rho_k, k \geqslant 2$, by accounting for the condition of orthogonality. A standard procedure for the change of basis in Eq. (5) implies that the problem is equivalent to an eigenanalysis of $\Sigma_{XX}^{-1/2} \Sigma_{YX} \Sigma_{YY}^{-1/2}$, *i.e.*, a maximization problem of Rayleigh quotient.

In this study, we use functions from the **R** package **fda** [18] to implement the FCCA method based on an integration of an exceedingly greedy procedure and an expansion of the functional basis. A variety of methods solving the optimization problem of functional canonical correlations have been developed in the literature; see for example, [15, 16, 19, 20].

Prior to estimating the functional canonical correlation between confirmed cases and death tolls in the US, some

© The Author (s) 2022. Published by Higher Education Press

additional pre-processing procedures to the data are necessary, as we observe that the date on which the first confirmed case is reported varies significantly across the states, and the number of death counts stays relatively low during the entire study period in several states.

Functional cluster analysis methods

Clustering is another common tool for data exploration in multivariate statistics, aiming at constructing homogeneous groups (called clusters) consisting of observed data that present some similar characteristics or patterns [21]. In contrast, observations from different groups are expected to be as dissimilar as possible. The functional cluster analysis (FCA) is an unsupervised learning process for functional data. In this section, we investigate the cluster structure in the US based on cumulative confirmed cases (after being scaled) across the states.

Two classical methods for the FCA are hierarchical clustering [22] and k-means clustering [23]. More recently, many clustering methods extended from them were proposed. See [24] for a summary.

Traditional k-means clustering methods for the FCA, *e.g.* [23], require a finite set of pre-specified basis functions in order to span a functional space, and assume the observed functional data to admit the basis expansion. The FCA is then completed by applying the standard k-means algorithm to the estimated basis coefficients. Alternatively, one may replace basis coefficients with FPC scores (*i.e.*, ξ_{ik}'s in Eq. (3)) for conducting the cluster analysis [25, 26, 27].

In modern FDA, there are two methods that are extensively popular for carrying out FCA. One approach is the EM-based algorithm [28], which assumes a finite (say $K \in \mathbb{N}$) mixture of Gaussian distributions. The model is given by

$$f(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \sum_{c=1}^{K} \pi_c \psi(\boldsymbol{x}\,|\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{6}$$

where $\boldsymbol{x}$ represents the data to be clustered, $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ is a collection of parameters; for each $c = 1, 2, \ldots, K$, $\pi_c \in (0, 1)$ is the mixing proportion such that $\sum_{c=1}^{K} \pi_c = 1$, and $\psi(\boldsymbol{x}\,|\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_c$ and variance $\boldsymbol{\Sigma}_c$. The log-likelihood of the model in Eq. (6) is optimized by the EM algorithm [29]. The setups of the E-step and the M-step are standard, available in a variety of articles, texts and tutorials, *e.g.* [30]. Given the number of clusters K, the EM algorithm partitions a set of n observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ into K clusters by solving

$$\arg\max_{c} \hat{\pi}_c \psi(\boldsymbol{x}_i\,|\,\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)$$

for each $i = 1, 2, \ldots, n$.

An alternative is the k-centers functional clustering (kCFC) algorithm proposed by [25], using the subspace spanned by the FPCs as cluster centers. This approach is popular as the distributional assumptions are relaxed.

Nonetheless, there are limitations in both algorithms. The kCFC algorithm assumes equal within-cluster variance, whereas the EM algorithm assumes a mixture of Gaussian distributions. Relevant discussions on the difference between the two algorithms are included in [24]. In the present analysis, it seems inappropriate to assume equal within-cluster variance, so we adopt the model-based EM algorithm, which is available in `R` package **EMCluster** [28]. Besides, the FDA objects created for FPCA (via the PACE algorithm) are ready to use. In other words, the inputs $\boldsymbol{X}$ in Eq. (6) usually take FPC scores. Since the clustering method is based on the EM algorithm, it is critical to select the initial values of the parameters appropriately. Specifically, we adopt a strategy developed in [31] for the initialization of the algorithm, with corresponding functions available in **EMCluster**.

Forecasting methods

When forecasting the COVID-19 data using functional data approaches, we need to preserve the temporal dynamics among curves to maintain the forecasting ability, which leads us to the context of functional time series [FTS, 32].

Converting data into FTS

An FTS consists of a set of random functions collected over time. So far in the literature, there are two common formations of an FTS object. One treats an FTS as a segmentation of an almost continuous time record into natural consecutive intervals, such as days, months, or quarters. This is a conventional treatment for financial data [33]. The other treats it as a collection of curves over a time period, where the curves are not functions of time; see [34] for example. The major difference between these two formations is whether the continuum of each function is a time variable. In this section, we start with the number of nationwide cumulative confirmed cases of the COVID-19 in the US, where we define and convert the FTS using the first convention outlined above. Based on several reports on the COVID-19 pandemic [35, 36], the average incubation period of the coronavirus is about 7 days, but the incubation period could last up to 14 days for most of the cases. Therefore, we segment the number of cumulative confirmed cases into curves with time intervals of 14

© The Author (s) 2022. Published by Higher Education Press

days. In line with the conventional treatment that segments a continuous curve into small intervals, we set the last value in one curve equal to the first value of the next.

Similar to the classical time series modeling, it is critical to appropriately handle the non-stationary data when modeling the FTS. To circumvent the issue of the non-stationarity in the COVID-19 data, we exploit the ideas from the literature dealing with FTS of share prices. In [37], the authors define the cumulative intraday returns (CIDR's), and later in [38], a formal test that justifies the stationarity of CIDR's curves has been developed. Here we convert the non-stationary cumulative confirmed case counts into stationary curves by calculating the daily growth rate in each curve:

$$r_{n,j} = 100 \times [\ln C_{n,j} - \ln C_{n,j-1}], \tag{7}$$

where $C_{n,j}$ denotes the number of confirmed cases on the j-th day of the n-th segment, for $j \in \{1,2,\ldots,14\}$, and $n \in \{1,2,\ldots,N\}$ with N being the total number of segments. Note that n is used as an index but not the total number of states in the rest of this section. Let $r_n(t)$ be the daily growth rates curve of the n-th segment, but values can only be observed at discrete time points j, such that $r_{n,j} = r_n(t_j)$.

Under the functional data framework, we assume the neighboring grid points are highly correlated, and smoothing serves as a tool for regularization, such that we borrow information from the neighboring grid points [17]. When applying to the daily growth rate for the cumulative confirmed cases, we assume that the underlying continuous and smooth function $Y_n(t)$ is observed at discrete points with smoothing error $\epsilon_n(t_j)$:

$$r_n(t_j) = Y_n(t_j) + \epsilon_n(t_j);$$

see [39] for details on smoothing. After applying the test from [38], we find that the p-value is equal to 0.989, suggesting the stationarity of the FTS with respect to the daily growth rates.

The rainbow plot proposed in [40] is effective in the visualization of the FTS. In a rainbow plot, functions that are ordered in time and colored with a spectrum of rainbow, such that functions from earlier times are colored in red, while the most recent ones are in violet. The rainbow plot captures the features of an FTS in two ways. Within each curve, the mode of variation reflects the pattern of the curve, and the ordering in color reveals the temporal dynamics over time.

In the rest of this section, we develop a forecasting scheme by considering a dynamic FPCA method that accounts for the temporal dynamics of the long-run covariance structure of the FTS. The forecasting is done through the dynamic FPCA scores. We use the root mean squared forecasting error and a nonparametric bootstrap approach to assess the accuracy of point and interval forecasts, respectively.

Dynamic FPCA

The major critics of classical FPCA stem from its incapability of making forecasts as it only aims to reduce the dimension of data by maximizing the variances explained by eigenfunctions. The dynamic FPCA, however, manages to reduce dimension towards the directions mostly reflecting temporal dynamics [41]. To better accommodate the need of capturing the temporal dynamics among the FTS and making better forecast, we here adopt the dynamic FPCA method [41]. The primary difference between classical and dynamic FPCA is that in performing the eigenanalysis, we replace the variance-covariance function with the long-run covariance function.

For an FTS, the long-run covariance function, which is the functional analogue of the long-run covariance matrix in standard time series analysis, plays an important role in accommodating the temporal dependence among functions. We now give the definition of long-run covariance function as follows. Let $\{Y_i(\tau), i \in \mathbb{Z}\}$ denote a sequence of stationary and ergodic functional time series, with τ being a bounded continuous variable. The long-run covariance function of $\{Y_i(\tau), i \in \mathbb{Z}\}$ is given by

$$C(s,t) := \sum_{i=-\infty}^{\infty} \gamma_i(s,t),$$

where $\gamma_i(s,t) = \mathrm{Cov}(Y_0(s), Y_i(t))$. In practice, we need to estimate $C(s,t)$ via a finite sample, *i.e.* $\{Y_i(\tau), i = 1,2,\ldots,n\}$ for some integer $n < \infty$. We use the lag-window estimator proposed by [44] to estimate the long-run covariance function $C(s,t)$:

$$\widehat{C}(s,t) := \sum_{i=-\infty}^{\infty} \mathcal{K}\left(\frac{i}{h}\right) \widehat{\gamma}_i(s,t), \tag{8}$$

where $\mathcal{K}(i/h)$ is a kernel function assigning different weights to the auto-covariance functions with different lags, and the parameter h is the bandwidth [43]. Typically, we assign more weights to the autocovariance functions of small lags and fewer to those of large lags. Here we use the "flat-top" type of kernels, as they provide smaller bias and faster rates of convergence [44]. For $k < 1$, the flat-top kernel is given by

$$K\left(\frac{i}{h}\right) = \begin{cases} 1, & 0 \leqslant |i/h| < k; \\ \dfrac{|i/h| - 1}{k - 1}, & k \leqslant |i/h| < 1; \\ 0, & |i/h| \geqslant 1. \end{cases}$$

© The Author (s) 2022. Published by Higher Education Press

The selection of the bandwidth crucially affects the accuracy of estimation. Here we adopt an adaptive bandwidth selection procedure [45] to get an approximately optimal bandwidth for the subsequent estimation of the long-run covariance of FTS. The estimator $\widehat{\gamma}_i(s,t)$ is given by

$$\widehat{\gamma}_i(s,t)=\begin{cases}\dfrac{1}{n-i}\displaystyle\sum_{r=1}^{n-i}Y_r(s)Y_{r+i}(t), & i\geqslant 0;\\ \dfrac{1}{n-i}\displaystyle\sum_{r=1-i}^{n}Y_r(s)Y_{r+i}(t), & i<0.\end{cases}$$

The dynamic FPCA is done through the Karhunen-Loève expansion exactly the same as Eq. (2), but the eigenvalues and eigenfunctions are estimated based on the eigenanalysis of $\widehat{C}(s,t)$.

Forecasting based on scores

We use m functional observations to get an ℓ-step-ahead forecast by the method developed in [46]. Note that applying a univariate time series forecasting method to the score vector $\widehat{\boldsymbol{\xi}}_k^{(m)}=\{\widehat{\xi}_{1k},\widehat{\xi}_{2k}\ldots,\widehat{\xi}_{mk}\}$ gives the estimated $\widehat{\xi}_{m+\ell|m,k}$, then the estimate of the ℓ-step-ahead forecast is given by

$$\widehat{Y}_{m+\ell|m}(t)=\widehat{\mu}(t)+\sum_{k=1}^{K}\widehat{\xi}_{m+\ell|m,k}\,\widehat{\phi}_k(t),$$

where $\widehat{\mu}(t)$ and $\{\widehat{\phi}_k(t),k=1,2,\ldots,K\}$ denote the estimated mean and FPCs, respectively.

For the application to the COVID-19 data, we forecast the daily number of the cumulative confirmed cases by the grid points on the forecasting curve. Without loss of generality, we consider the one-step-ahead forecast (*i.e.*, $\ell=1$). Note that this procedure generates daily forecasts up to 13 days, since by our assumption on the FTS conversion in section of converting data into FTS, the first grid point on the curve is the same as that on the last day of the study period. By implementing the FTS converting procedures from section of converting data into FTS, we obtain an FTS object consisting of 11 functional curves. We create an *expanding window analysis* framework, and apply the proposed forecasting method to get multiple one-step-ahead forecasts on the growth rate curve. The details are deferred to the following.

Forecast accuracy evaluation

In this section, we introduce the methods that assess the prediction accuracy for point and interval forecasts. To make forecast for the number of confirmed cases in the US in the next 13 days, we set $\ell=1$ throughout the section.

Point forecast For a point forecast, We use the root mean square forecast error (RMSFE) to evaluate its accuracy:

$$\mathrm{RMSFE}(j)=\sqrt{\frac{1}{N-n}\sum_{m=n}^{N-1}\left[C_{m+1}(j)-\widehat{C}_{m+1}(j)\right]^2},\qquad(9)$$

where N is the total number of segments in the FTS, m is the number of curves used in forecasting, $C_{m+1}(j)$ is the number of confirmed cases at the j-th point on the $(m+1)$-th curve, and $\widehat{C}_{m+1}(j)$ is the corresponding forecast value. The RMSFE measures the discrepancy between the forecast and the actual value.

Interval forecast To better capture the uncertainty of point forecasts, we construct the associated prediction intervals in this section. The authors in [46] suggest a curve prediction based on modeling FPC scores by scalar time series, while the authors in [47] develop a multivariate prediction scheme to utilize the valuable information hidden in the dependence structure of the data. Later on, a robust nonparametric bootstrap method is proposed in [48] to construct pointwise prediction intervals. In the present study, the difference between the results based on the methods from [46] and those from [47] is negligible as the FPCs obtained from dynamic FPCA are uncorrelated [41].

The construction of the prediction intervals is based on in-sample-forecast errors, *i.e.*, $\widehat{e}_{m+1|m}(j)=C_{m+1}(j)-\widehat{C}_{m+1}(j)$. Specifically, we generate a bootstrap sample (with replacement) of forecasting errors to obtain the upper and lower bounds, respectively denoted by $\eta^{\mathrm{ub}}(j)$ and $\eta^{\mathrm{lb}}(j)$. Then we choose a tuning parameter, δ_α, such that

$$\mathbf{P}\left\{\delta_\alpha\times\eta^{\mathrm{lb}}(j)\leqslant\widehat{e}_{m+1|m}\leqslant\delta_\alpha\times\eta^{\mathrm{ub}}(j)\right\}=(1-\alpha)\times 100\%.$$

The one-step-ahead (pointwise) prediction interval is given by

$$\widehat{C}_{m+1}(j)+\delta_\alpha\times\eta^{\mathrm{lb}}(j)\leqslant C_{m+1}(j)\leqslant\widehat{C}_{m+1}(j)+\delta_\alpha\times\eta^{\mathrm{ub}}(j).\quad(10)$$

We use the interval scoring rules [49] to evaluate the accuracy of the pointwise interval forecasts. The interval score for the pointwise interval forecast (c.f. Eq. (10)) at time point j is given by

$$\begin{aligned}&S_\alpha\left[\widehat{C}^{\mathrm{lb}}_{m+1}(j),\widehat{C}^{\mathrm{ub}}_{m+1}(j);C_{m+1}(j)\right]\\&=\left[\widehat{C}^{\mathrm{ub}}_{m+1}(j)-\widehat{C}^{\mathrm{lb}}_{m+1}(j)\right]\\&\quad+\frac{2}{\alpha}\left[\widehat{C}^{\mathrm{lb}}_{m+1}(j)-\widehat{C}_{m+1}(j)\right]\mathbf{1}\left\{\widehat{C}_{m+1}(j)<\widehat{C}^{\mathrm{lb}}_{m+1}(j)\right\}\\&\quad+\frac{2}{\alpha}\left[\widehat{C}_{m+1}(j)-\widehat{C}^{\mathrm{ub}}_{m+1}(j)\right]\mathbf{1}\left\{\widehat{C}_{m+1}(j)>\widehat{C}^{\mathrm{ub}}_{m+1}(j)\right\},\end{aligned}$$

© The Author (s) 2022. Published by Higher Education Press

where $\mathbf{1}\{\cdot\}$ represents a standard indicator function, and the level of significance, α, is conventionally set at 0.2. A lower interval score suggests that the interval forecast is more accurate. The ideal scenario is that the actual values of $C_{m+1}(j)$ values lie between $\widehat{C}^{\text{lb}}_{m+1}(j)$ and $\widehat{C}^{\text{ub}}_{m+1}(j)$ for all j. The mean interval score over $(N-n)$ one-step-ahead forecasts becomes

$$\bar{S}_{\alpha}(j) = \frac{1}{N-n}\sum_{m=n}^{N-1} S_{\alpha}\left[\widehat{C}^{\text{lb}}_{m+1}(j), \widehat{C}^{\text{ub}}_{m+1}(j); C_{m+1}(j)\right].$$

RESULTS

Preliminary analysis

We collected the number of confirmed and death cases of the COVID-19 from 50 states (49 continental states and Alaska) in the US between 01/21/2020 (the date of the first domestic confirmed case reported in the US) to 08/15/2020 (the weekend before school reopening in most of the states). The data of cumulative confirmed and death cases were collected at state level, from a publicly available repository released and updated by *The New York Times*. In what follows, the daily confirmed and death cases were obtained effortlessly.

The majority of our analyses is done at the state level. Noticing the significant differences in the population size for each state, we standardize the data, using the estimated population size for each state at the end of year 2019 from the US Census Bureau to scale all collected data (cumulative and daily cases), and save them in units of "per million". In Fig. 1, we plot the number of daily confirmed and death cases in the US. From early May to early June, the cumulative case-to-fatality rate (CFR) of the COVID-19 in the US has stayed consistently high around 6.01%, close to the estimate (6.1%) given in [50], which is slightly higher than the CFR in Wuhan, China (5.8%) reported on 02/01/2020 [51], and significantly higher than the global CFR (3.4%) according to WHO Director-General's opening remarks at the media briefing on 03/03/2020. In August, the CFR in the US has declined to 3.1%.

One important epidemic metric for the study of the infectious disease dynamics is the basic reproduction number, usually denoted as R_0. It refers to the average number of secondary cases per infectious case in a population where all the individuals thereof are susceptible to the infection. A variant called the instantaneous reproduction number, denoted as R_t, is the average number of cases generated by each infection at a given time t. We observe that 33 of 50 states in the US has had $R_t > 1$ on 05/29/2020, suggesting that the epidemic has not yet been fully contained by the end of May, 2020. In fact, this proportion climbs to 41 out of 50 in the following month. Upon the end of the study period, the proportion of the states with $R_t > 1$ has dropped to 16 out of 50, but the whole nation is still faced with the

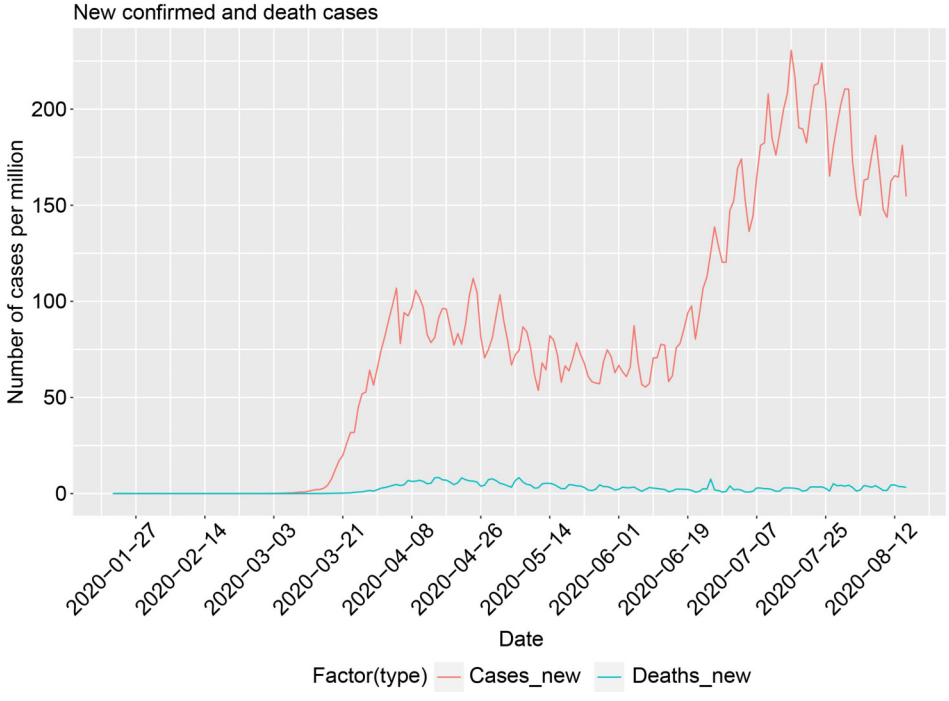

Figure 1. The number of daily confirmed and death cases (per million) in the US from 01/21/2020 to 08/15/2020.

© The Author (s) 2022. Published by Higher Education Press

potential risk of a massive spread owing to the reopening of schools in the fall. Some other results of critical epidemic characteristics and dynamics of the COVID-19 in the US have been reported in [52].

In this study, we treat the collected data as functional data, and adopt methods from the functional data analysis (FDA). Figure 2 shows the functional data of daily confirmed and death cases (per million) of the top five states in the US during the study period. As discussed in [53], the FDA methods are applicable to sparsely and irregularly spaced data in time, so are preferred to the standard time series methods in the analyses of the time series data. Besides, the FDA methods manage to capture the functional behavior of the underlying data which generate the process (see [19, 53, 54] for details), and have been widely adopted in a plethora of applications in public health and biomedical studies [55], but with limited applications to COVID-19 [56, 57]. In the next several sections, we list the adopted FDA methods, and present the analysis results when they are applied to the COVID-19 data in the US. It is worth mentioning that, throughout the manuscript, we assume that the functional data across the 50 states are mutually independent.

Modes of variation

To begin with, we plot the fitted mean curve (which estimates the trend over time), the fitted variance curve (which estimates the subject-specific variation) and the fitted covariance surface of daily confirmed cases across 50 states in Fig. 3.

The estimated mean is close to 0 in January and February, and starts climbing since early March until reaches a local maximum in late April. The overall trend in May is downwards, followed by a second wave that starts from June. The curve hits the peak at the end of the third week of July. The estimated variance curve looks similar to the fitted mean curve in shape, which stays low at the very early stage, and deviates from 0 at the end of the first week of March. The first local maximum emerges around 04/17/2020, and then starts

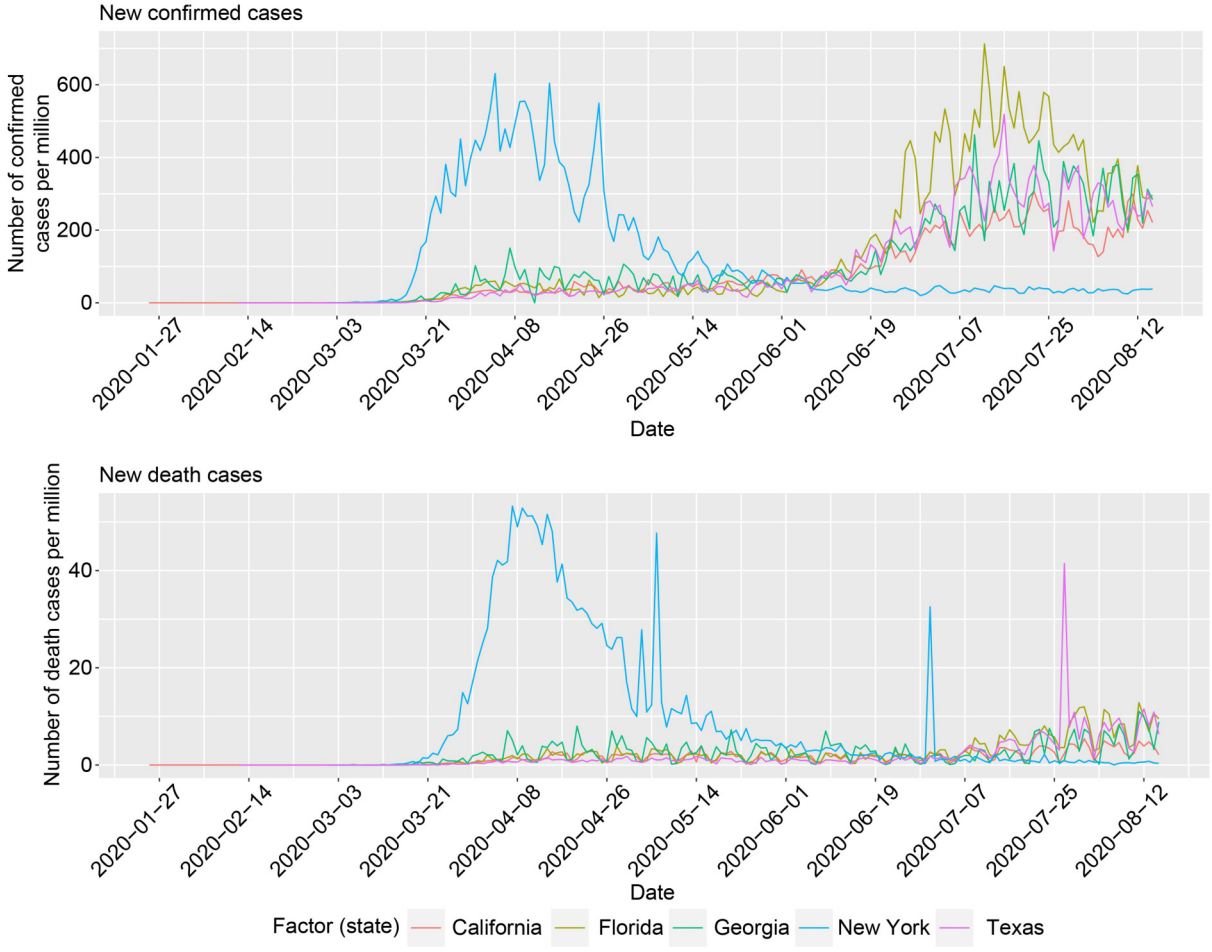

Figure 2. The number of daily confirmed and death cases (per million) of top five continental states in the US.

© The Author (s) 2022. Published by Higher Education Press

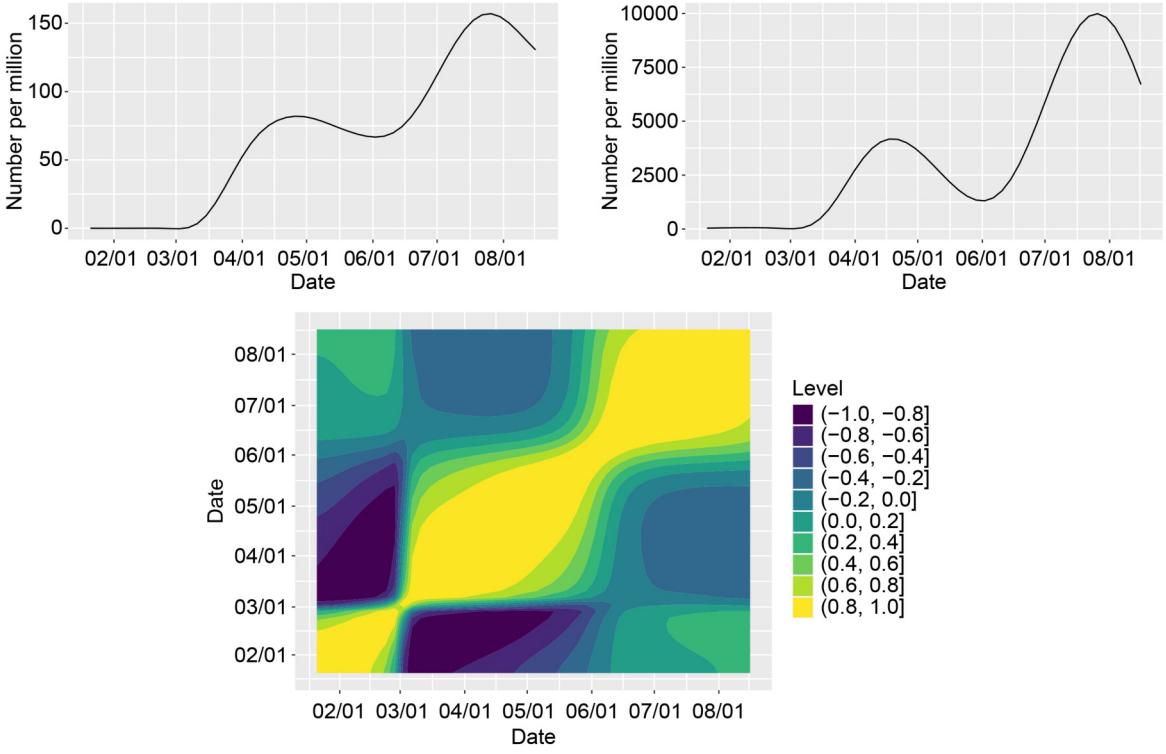

Figure 3. Fitted mean curve (top left panel), fitted variance curve (top right panel) and fitted correlation surface (bottom) of daily confirmed cases of COVID-19 in the US.

decreasing until 06/01/2020. The global maximum of the curve is observed in the third week of July, corresponding to the worst period of the attack of the COVID-19 to the country. The correlation surface is presented through a contour plot. Measurements at close time points appears to be highly correlated. The correlation between early and very late times are close to 0, whereas that between early and middle times tends to be negative. The correlation patterns among middle and later times are slightly negative. In particular, the correlation surface reveals that the increase in the number of confirmed cases follows three different stages, where the two break points are respectively 03/01 and 06/01. The strong positive correlation seems more persistent in the latter two stages.

Next we apply the PACE method to the COVID-19 data of daily confirmed cases in the US. The adaptive algorithm selects a total of six eigenfunctions (accounting for more than 99.99% of the total variation), where the first two together explains about 95% of the total variation of the data (see Fig. 4), indicating the remaining eigenfunctions are less important.

The first eigenfunction (accounting for 68.38% of the total variation) remains constant close to 0 during the first month of the study period, and starts decreasing afterwards. The value of the eigenfunction falls negative since 03/01/2020 until it reaches a local minimum around 04/17/2020. After that, the function value keeps increasing until it hits the peak around 07/23/2020. The first eigenfunction defines the most important mode of variation, corresponding to an overall mean effect. Besides, it reflects a contrast between middle and late times, where the break point emerges about one week after the announcement of business reopening in the majority of the states in the US.

The second eigenfunction presents a variation around a piecewise linear time trend. The value of the eigenfunction is negative at early times, and becomes positive since March, and hits a break point in the third week of April. Since then, it begins to decrease but stays positive during the rest of the study period. The second eigenfunction suggests that the second largest variation among the states is a scale difference along the direction of this functional curve.

Canonical correlations between confirmed and death cases

Now we inspect the functional canonical correlation between confirmed and death cases from the 50 states in the US. The first step is to pre-process the data. Noticing that diagnostic tests (*e.g.*, swab test) of the COVID-19 have not been widely implemented in most states until late March (for instance, the first drive-through

© The Author (s) 2022. Published by Higher Education Press

Figure 4. First two eigenfunctions for the US COVID-19 confirmed cases.

COVID-19 testing site in Philadelphia was not open to the public until 03/20/2020), we reschedule the starting date of study period to 04/01/2020 for the rest of the study in this section. Besides, the cumulative number of (scaled) confirmed and death cases (instead of scaled daily numbers) are used for the analysis so that no state is excluded due to sparsity, especially for those with few cases reported during April and early May.

In the literature, there are several different options of basis functions for the basis expansion in the FCCA. Having observed some periodic features in our functional data, we choose the Fourier basis here. The first canonical correlation is 0.985, which reveals a dominant pair of modes of variation that are highly positively correlated. The corresponding canonical weight functions are plotted in Fig. 5. The canonical weight function of confirmed cases resembles a sinusoidal function with a period of one month approximately, while the counterpart of death cases seemingly contrasts early and middle times to late times (primarily in July). A state has a negative score with a large absolute value on the canonical weight of confirmed cases if it has a large number of confirmed cases in April, late May and early June, but small counts in July and early August. Also, a state has a high positive score on the canonical weight of death cases if it has more death cases in July than any other time during the study period.

Next we plot the canonical variable scores for the death cases against confirmed cases in Fig. 6, where we

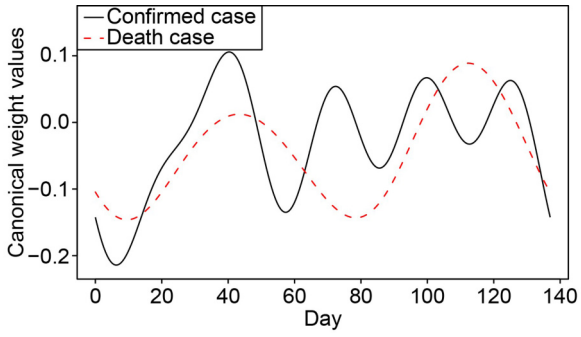

Figure 5. The first pair of canonical weight functions correlating confirmed and death cases of 50 states in the US.

see that the New York state lies on the bottom left corner. This is due to the fact that the New York state (especially New York City) has a lot of confirmed and death cases at the early stage of the pandemic, but later the transmission of the virus has been well controlled since July, thanks to the complete shutdown of the city as well as the rigorous adherence to public health measures, *e.g.* the practice of social distancing, mask wearing and the limit of the number of people allowed in essential businesses. Since July, less than 20 deaths has been reported daily in New York City, in contrast with an average of more than 700 deaths every day in April. Therefore, both canonical variable scores remain low in the state of New York.

Another "outlier" is New Jersey, which is

© The Author (s) 2022. Published by Higher Education Press

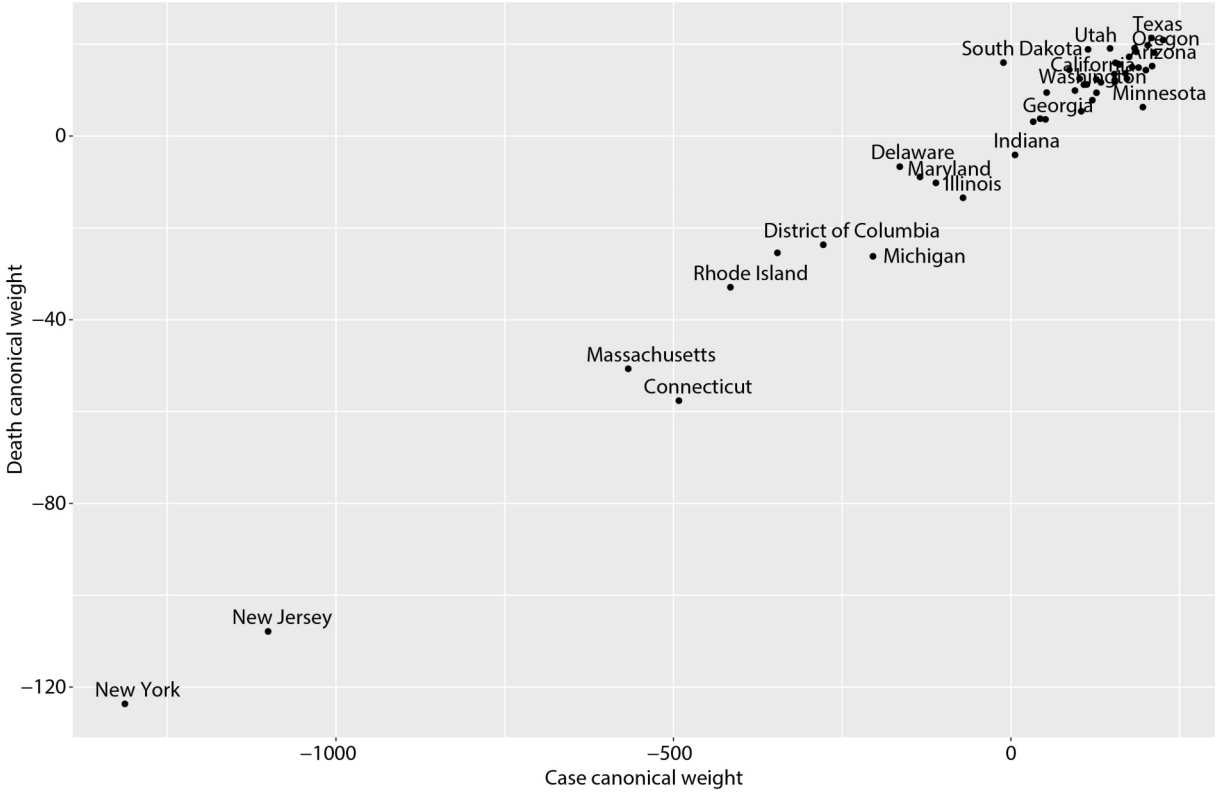

Figure 6. Canonical variable scores of death cases against confirmed cases of 50 states in the US.

geographically adjacent to New York state. The overall trends of confirmed and death cases in New Jersey are similar in shape to those in New York, but both are smaller in magnitude. Therefore, New Jersey sits close to New York in Fig. 6. Three states in the middle with negative scores in both canonical variables are Massachusetts, Connecticut and Rhode Island, all of which are geographically close to New York.

All other states are mostly scattered on the top right corner in Fig. 6. The one sits at the most top right is Texas, with positive scores in both canonical variables. The confirmed and death cases in Texas are not the largest in the first wave of outbreak (*i.e.*, April and May), but the state has seen a large surge in the daily confirmed and death cases since late June. Consequently, both of the canonical scores of Texas remain positive. In fact, Texas has controlled the spread of the disease well in April when a 14-day self-quarantine has been mandated, all nonessential businesses are closed, and ordered travel restrictions have been instituted between Texas and Louisiana (an outbreak occurs in New Orleans during that period). However, with the stay-at-home order lapsed in May, and the reopening of the economy started in June, Texas has experienced a huge increase in the COVID-19 cases, due to the increasing frequencies of indoor and outdoor gatherings in large groups of people without proper public health measures implemented.

Cluster analysis results

Similar to the FCCA part, we trim the study period to "04/01/2020 to 08/15/2020" as few (or even no) confirmed case has been observed in many states before April. The implementation of the EMCluster algorithm requires the predetermination of the number of clusters in prior. There are a couple of standard methods to obtain an optimal value of K. We select the value of K via the elbow method based on the sum of squares of the within-cluster variations. The computation reveals that the most appropriate value of K is 5, and the optimal clustering strategy associated with $K = 5$ is given in Table 1.

The first cluster includes New York, New Jersey and Illinois, the three states that have been most severely attacked by the COVID-19 in the first wave, as well as Florida, Texas and California, all of which have experienced a huge surge in both confirmed and death cases in the second wave. Overall, these six states are the worst hit by the COVID-19 throughout the study period.

The second cluster contains 24 states, including

© The Author (s) 2022. Published by Higher Education Press

Table 1 Clustering results based on the number of cumulative confirmed cases in 50 states in the US from 04/01/2020 to 08/15/2020

Cluster number	States
Cluster 1	CA FL IL NJ NY TX
Cluster 2	AL AZ CO CT GA IN IA LA MD MA MI MN MS MO NV NC OH PA RI SC TN VA WA WI
Cluster 3	AR UT
Cluster 4	KS KY NE OK
Cluster 5	AK DE DC ID ME MT NH NM ND OR SD VT WV WY

Georgia, North Carolina, Massachusetts, Pennsylvania and Washington, etc., most of which are geographically close (but not necessarily adjacent) to the hot-spot regions. Some states, *e.g*. Georgia, North Carolina and Colorado, actually have reached the peak of the COVID-19 cases in the second wave after the reopening of the business, when some mandatory protocols have been called off. Besides, with the relaxation of travel restrictions, an increase in population movements may also cause the cross infections among the residents in these states. The state of Washington, located at the northwest corner, also belongs to this cluster, where the first case in the US has been reported.

The third cluster only has two states: Utah and Arkansas. Utah is not directly adjacent those epidemic centers specified in the first cluster. The control of the COVID-19 remains reasonably well in Utah, and we speculate some possible reasons as follows. Though there is no evidence that youngsters are not likely to be infected, they are likely to have stronger immune systems (than elders). Utah has the lowest percentage of residents above 65 in the nation. Besides, Utah is among the healthiest states in the nation overall (https://www.americashealthrankings.org). The attack of the COVID-19 to the other state in the third cluster, Arkansas, has also been less serious during the second wave, although it is right next to Texas. The local government has maintained a series of policies that help reduce the spread of the disease, for example, the delay of school opening.

The fourth cluster contains Oklahoma, Kansas, Nebraska and Kentucky, where the major industries are agriculture, core mining, and aviation, etc. Besides, there are not too many tourist resorts in these states triggering large population movements. It is also necessary to note that the population densities are relatively small in these states. Hence, they are less affected by the COVID-19 compared to those in the first three clusters.

Lastly, a total of 14 states are classified in the fifth cluster, including Vermont, New Mexico, Alaska, Montana, and Wyoming, etc. These states are the least attacked by the virus. Although the data has been scaled, most of these states are large in their geographical sizes, leading to low population densities. In fact, among all 50 states, Alaska, Montana and Wyoming have the three lowest population densities in the nation.

For comparison, we also conduct an analogous analysis on a truncated study period from 04/01/2020 to 05/15/2020, around which business reopening starts over a large number of states. The clustering results are presented in Fig. 2, and the cluster structure appears to be greatly different from that given in Table 1. Essentially, the New York state and its neighbors are the epidemic centers for the period from 04/01/2020 to 05/15/2020, and the alterations in the cluster membership flag the significant impacts of reopening the business in accelerating the spread of the virus.

Forecasting results

Figure 7 presents the rainbow plot of both unsmoothed (left) and smoothed (right) growth rates of the cumulative confirmed cases of COVID-19 in the US from 04/04/2020 to 08/25/2020, leading to exactly 11 functional curves. From Fig. 7, we observe that all the curves share a similar temporal pattern, where the growth rate increases rapidly within the first four days, and then tapers off gradually. This coincides with a scientific finding that an infected person is most contagious early in the course of their illness [58]. Based on the color ordering in the rainbow plot, we see that the growth rate declines gradually from April to June, indicating the effectiveness of the practice of public health measures (including lock down policies implemented at the early stage of the outbreak), but it bounces back up again in July, flagging the effects of reopening.

Given that there are 11 functional curves in the study period, we start with the first $n = 8$ curves to generate one-step-ahead point forecasts, and repeat this forecasting procedure by adding one curve at a time until the first 10 curves are included, which gives us $(N - n = 3)$ one-step-ahead forecasts in total. In other words, for each time point j, there are three forecast values. We choose $n = 8$ as our starting point to ensure that an adequate number of curves are used for forecasting.

© The Author (s) 2022. Published by Higher Education Press

Table 2 Clustering results based on the number of cumulative confirmed cases in 50 states in the US from 04/01/2020 to 08/15/2020

Cluster number	States
Cluster 1	NJ NY
Cluster 2	CA CO CT FL GA IL IN LA MD MA MI NC OH PA TN TX VA WA
Cluster 3	AL AZ IA MN MS MO RI SC WI
Cluster 4	DE DC KS KY NE NV UT
Cluster 5	AK AR ID ME MT NH NM ND OK OR SD VT WV WY

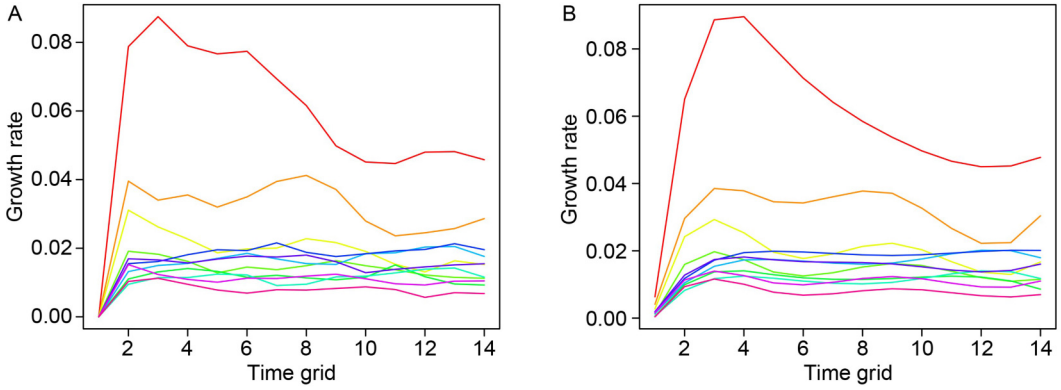

Figure 7. Functional time series for unsmoothed (left panel) and smoothed (right panel) growth rates of the cumulative confirmed cases of COVID-19 in the US.

We adopt a standard forecasting approach through the autoregressive integrated moving average (ARIMA) model as a competing method. The forecasting results are presented in Fig. 8, where the y-axis is the average of the RMSFE (scaled by 10) of the three one-step-ahead point forecasts. We observe that the mean values of the RMSFE of FTS forecasts are consistently smaller than the counterparts of standard ARIMA, suggesting that the proposed FTS method is preferred.

Figure 9 shows the mean interval scores (scaled by 10) of the interval forecasts obtained from the FTS model versus the standard ARIMA model. The computation results reveal that the proposed FTS method outperforms the standard ARIMA.

We apply the proposed method to forecast the number of cumulative confirmed cases in the next 13 days upon the last date of study period. Table 3 shows the predicted number and 80% confidence interval of nationwide cumulative confirmed cases (in thousands). In addition, the point forecasts and the associated 80% confidence bands for the cumulative confirmed cases in the US from 08/26/2020 to 09/07/2020 are depicted in Fig. 10, where we also compare the prediction results with those generated from an ARIMA model.

From Table 3, we see that our FTS forecasts tends to slightly underestimate the total counts, but the actual

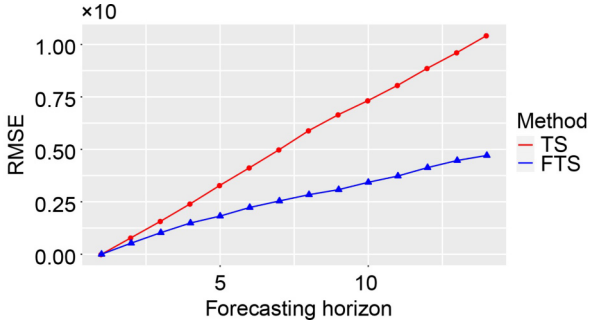

Figure 8. Average RMSFE values (scaled by 10) for the FTS and the standard ARIMA method for the next 13 days. The FTS method consistently outperforms the standard ARIMA. Note that the RMSFE curve starts from 0, corresponding to the assumption in section of converting data into FTS that the first day of the 14-day prediction segment coincides with the last day of the study period.

count still falls within the prediction intervals. Such deviation may be due to the reopening of the schools starting from mid-August, leading to the surge in the cumulative confirmed cases nationwide. When comparing the FTS forecasting results with the ARIMA model in Fig. 10, we see a narrower prediction interval for the FTS results, suggesting that it is preferred than the standard ARIMA model.

© The Author (s) 2022. Published by Higher Education Press

DISCUSSION

In this article, we conduct a functional data analysis of the time series data of COVID-19 in the US. Our answers to the four public health questions raised at the end of section of introduction are given as follows. Based on our results, it is evident that the practice of public health measures (*e.g.* the "stay-at-home" order and mask wearing) helps to reduce the growth rate of the epidemic outbreak over the nation. However, the implementation of the business reopening plans seems to have caused the rapid spread of the COVID-19 in some states, *e.g.* Texas and Florida.

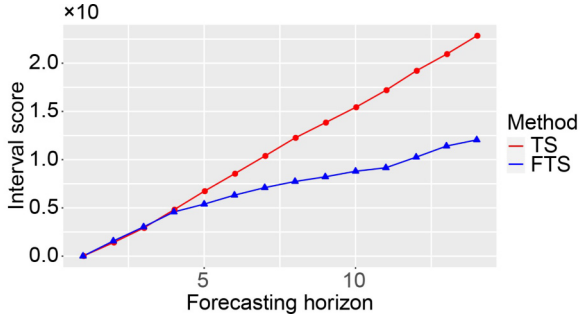

Figure 9. Average interval scores (scaled by 10) for the FTS method and the standard ARIMA model for the next 13 days. The FTS approach is preferred in that it produces smaller interval scores in most of the cases. Note that the RMSFE curve starts from 0, corresponding to the assumption in section of converting data into FTS that the first day of the 14-day prediction segment coincides with the last day of the study period.

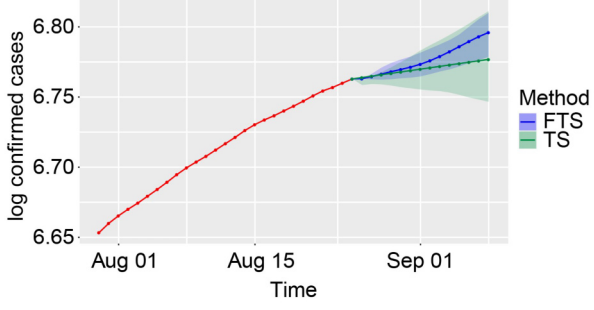

Figure 10. Point forecast and confidence bands for the number of total confirmed cases in the next 13 days (upon the last date of the study period).

We quantitatively assess the correlation between confirmed and death cases using the FCCA. Overall, we observe a high canonical correlation between confirmed and death cases, though the canonical variable scores vary from state to state. With the population size in each state carefully adjusted, we see that there is a substantial change in the cluster structure during early and late times, and 05/15/2020 (the average date of business reopening across the states) appears to be a potential change point. States that are geographically close to the hot spots are likely to be clustered together, and population density (at state level) appears to be a critical factor affecting the cluster structure.

In addition, we propose a forecasting scheme for the nationwide cumulative confirmed cases under the functional time series framework. The basis of forecasting is converting data to FTS. In the paper, we adopt a testing procedure proposed by [38] to test the stationarity of the converted FTS. However, we would like to point out that the test may experience power loss due to the difficulty of selecting the correct number of FPCs, especially when the growth rate of data has structural breaks [59]. Integrating information from the neighboring data points, the forecasting accuracy from the functional time series approach outperforms that from an ARIMA model. Forecasts are also made to the next 13 days of the study period, and comparisons with the actual counts are provided. Although our method tends to produce smaller estimated counts for the next 13 days, the actual values still fall within the prediction intervals. It is also worthwhile noticing that such underestimation may indicate the effects of school reopening in accelerating the spread of the virus over the country.

There are some limitations in the present study. The analysis is carried out at the relatively early stage of the COVID-19 outbreak in the US. Hence, the quality of data is an inevitable concern. The accuracy of the daily confirmed cases, including measurement errors and unreported cases, may be questionable. Besides, the exact date on which the first case emerged in each of the states might not be accurate, either. Second, when conducting functional data analysis across the states, we

Table 3 Number of cumulative confirmed cases (in thousands) in the US in the next 13 days (upon the last date of the study period)

Date	08/26	08/27	08/28	08/29	08/30	08/31	09/01	Date	09/02	09/03	09/04	09/05	09/06	09/07
Actual	5,839	5,884	5,931	5,975	6,009	6,045	6,089	Actual	6,122	6,168	6,220	6,263	6,293	6,318
Forecasts	5,815	5,840	5,863	5,884	5,907	5,935	5,970	Forecasts	6,012	6,060	6,110	6,161	6,210	6,252
Lower bound	5,788	5,786	5,790	5,799	5,805	5,817	5,839	Lower bound	5,865	5,885	5,913	5,946	5,976	6,006
Upper bound	5,843	5,902	5,948	5,981	6,010	6,049	6,090	Upper bound	6,131	6,176	6,233	6,312	6,390	6,457

© The Author (s) 2022. Published by Higher Education Press

have assumed mutual independence in the functional data due to the lack of satisfying methods. However, the spatial dependence factors, such as population mobility, should also be incorporated in a more comprehensive study, with focus on methodological development and then applications. Thirdly, upon the completion of the analysis, the vaccine had not been available in the US. While more and more residents are being vaccinated across the country, we are interested in assessing the impact of the vaccinated population via a change point analysis, and predicting the approximate time that human being lives would become normal again in the US. We will report our results elsewhere.

ACKNOWLEDGMENTS

The authors would like to thank two anonymous referees for their valuable comments on the manuscript.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Chen Tang, Tiandong Wang and Panpan Zhang declare that they have no conflict of interest or financial conflicts to disclose. All procedures performed in studies were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

OPEN ACCESS

This article is licensed by the CC By under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

REFERENCES

1. Dong, E., Du, H. and Gardner, L. (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis., 20, 533−534
2. Grubaugh, N. D., Hanage, W. P. and Rasmussen, A. L. (2020) Making sense of mutation: What D614G means for the COVID-19 pandemic remains unclear. Cell, 182, 794−795
3. Jollie, I. T. (2002) Principal Component Analysis. 2 edition. New York: Springer-Verlag
4. Dauxois, J., Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. J. Multivariate Anal., 12, 136−154
5. Karhunen, K. (1946) Zur spektraltheorie stochastischer prozesse. Annales Academiae scien-tiarum Fennicae. Series A. 1, Mathematica-physica, page 34
6. Loeve, M. (1995) Probability Theory: Foundations, Random Sequences. Princeton: D. Van Nostrand, Company
7. Shang, H. L. (2014) A survey of functional principal component analysis. AStA Adv. Stat. Anal., 98, 121−142
8. Jones, M. C. and Rice, J. A. (1992) Displaying the important features of large collections of similar curves. Am. Stat., 46, 140−145
9. Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. J. Am. Stat. Assoc., 100, 577−590
10. Fan, J. and Gijbels, I. (1996) Local Polynomial Modelling and its Application. London: Chapman & Hall
11. Capra, W. B. and Müller, H.-G. (1997) An accelerated-time model for response curves. J. Am. Stat. Assoc., 92, 72−83
12. Shibata, R. (1981) An optimal selection of regression variables. Biometrika, 68, 45−54
13. Carroll, C., Gajardo, A., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H. Zhu, C. and Lin, S.-C. (2020) fdapace: Functional data analysis and empirical dynamics. R package version 0.5.3, https://CRAN.R-project.org/package=fdapace
14. Leurgans, S. E., Moyeed, R. A. and Silverman, B. W. (1993) Canonical correlation analysis when the data are curves. J. R. Stat. Soc. B, 55, 725−740
15. He, G., Müller, H.-G. and Wang, J.-L. (2003) Functional canonical analysis for square integrable stochastic processes. J. Multivariate Anal., 85, 54−77
16. He, G., Müller, H.-G. and Wang, J.-L. (2004) Methods of canonical analysis for functional data. J. Stat. Plan. Inference, 122, 141−159
17. Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016) Functional data analysis. Annu. Rev. Stat. Appl., 3, 257−295
18. Ramsay, J. O., Graves, S. and Hooker, G. fda: Functional data analysis, 2020. R package version 5.1.4, https://CRAN.R-project.org/package=fda
19. Ramsay, J. O. and Silverman, B. W. (2002) Applied Functional Data Analysis: Methods and Case Studies. New York: Springer-Verlag
20. Yang, W., Müller, H.-G. and Stadtmüller, U. (2011) Functional singular component analysis. J. R. Stat. Soc. Series B Stat. Methodol., 73, 303−324
21. Kaufman, L. and Rousseeuw, P. J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. 1 edition. Hoboken: Wiley-Interscience
22. Ferreira, L. and Hitchcock, D. B. (2009) A comparison of hierarchical methods for clustering functional data. Commun. Stat. Simul. Comput., 38, 1925−1949
23. Abraham, C., Cornillon, P. A., Matzner-Lber, E. and Molinari, N. (2003) Unsupervised curve clustering using B-splines. Scand. J. Stat., 30, 581−595

© The Author (s) 2022. Published by Higher Education Press

24. Jacques, J. and Preda, C. (2013) Funclust: A curves clustering method using functional random variables density approximation. Neurocomputing, 112, 161−171
25. Chiou, J.-M. and Li, P.-L. (2007) Functional clustering and identifying substructures of longitudinal data. J. R. Stat. Soc. Series B Stat. Methodol., 69, 679−699
26. Chiou, J.-M. and Li, P.-L. (2008) Correlation-based functional clustering via subspace projection. J. Am. Stat. Assoc., 103, 1684−1692
27. Peng, J. and Müller, H.-G. (2008) Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. Ann. Appl. Stat., 2, 1056−1077
28. Chen, W.-C. and Maitra, R. (2015) EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution, 2015. R Package, http://cran.rproject.org/package=EMCluster
29. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B, 39, 1−22
30. Lee, G. and Scott, C. (2012) EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Comput. Stat. Data Anal., 56, 2816−2829
31. Biernacki, C., Celeux, G. and Govaert, G. (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Stat. Data Anal., 41, 561−575
32. Hörmann, S. and Kokoszka, P. (2010) Weakly dependent functional data. Ann. Stat., 38, 1845−1884
33. Hörmann, S. and Kokoszka, P. (2012) Weakly dependent functional data. In: Handbook of Statistics (Rao, T. S., Rao, S. S. and Rao, C. eds.,) volume 30, pp. 157–186. Amsterdam: Elsevier
34. Chiou, J.-M. and Müller, H.-G. (2009) Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. J. Am. Stat. Assoc., 104, 572−585
35. McAloon, C., Collins, Á., Hunt, K., Barber, A., Byrne, A. W., Butler, F., Casey, M., Griffin, J., Lane, E., McEvoy, D., *et al*. (2020) Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. BMJ Open, 10, e039652
36. Zhang, P., Wang, T. and Xie, S. X. (2021) Meta-analysis of several epidemic characteristics of COVID-19. J. Data Sci., 18, 536–549
37. Gabrys, R., Horvath, L. and Kokoszka, P. (2010) Tests for error correlation in the functional linear model. J. Am. Stat. Assoc., 105, 1113−1125
38. Horváth, L., Kokoszka, P. and Rice, G. (2014) Testing stationarity of functional time series. J. Econom., 179, 66−82
39. Hyndman, R. J. and Shahid Ullah, M. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Stat. Data Anal., 51, 4942−4956
40. Hyndman, R. J. and Shang, H. L. (2010) Rainbow plots, bag plots, and boxplots for functional data. J. Comput. Graph. Stat., 19, 29−45
41. Hörmann, S., Kidziński, Ł. and Hallin, M. (2015) Dynamic functional principal components. J. R. Stat. Soc. Series B Stat. Methodol., 77, 319−348
42. Panaretos, V. M. and Tavakoli, S. (2013) Fourier analysis of stationary time series in function space. Ann. Stat., 41, 568−603
43. Andrews, D. W. K. (1991) Heteroskedasticity and autocorrelation consistent covariant matrix estimation. Econometrica, 59, 817−858
44. Politis, D. N. and Romano, J. P. (1996) On at-top kernel spectral density estimators for homogeneous random fields. J. Stat. Plan. Inference, 51, 41−53
45. Rice, G. and Shang, H. L. (2017) A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. J. Time Ser. Anal., 38, 591−609
46. Hyndman, R. J. and Shang, H. L. (2009) Forecasting functional time series. J. Korean Stat. Soc., 38, 199−211
47. Aue, A., Norinho, D. D. and Hörmann, S. (2015) On the prediction of stationary functional time series. J. Am. Stat. Assoc., 110, 378−392
48. Shang, H. L. (2018) Bootstrap methods for stationary functional time series. Stat. Comput., 28, 1−10
49. Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc., 102, 359−378
50. Abdollahi, E., Champredon, D., Langley, J. M., Galvani, A. P. and Moghadas, S. M. (2020) Temporal estimates of case-fatality rate for COVID-19 outbreaks in Canada and the United States. CMAJ, 192, E666–E670
51. Omer, S. B., Malani, P. and Del Rio, C. (2020) The COVID-19 pandemic in the US: A clinical update. JAMA, 323, 1767–1768
52. Peirlinck, M., Linka, K., Sahli Costabal, F. and Kuhl, E. (2020) Outbreak dynamics of COVID-19 in China and the United States. Biomech. Model. Mechanobiol., 19, 2179−2193
53. Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis. J. R. Stat. Soc. B, 53, 539−561
54. Ramsay, J. O. (1982) When the data are functions. Psychometrika, 47, 379−396
55. Ullah, S. and Finch, C. F. (2013) Applications of functional data analysis: A systematic review. BMC Med. Res. Methodol., 13, 43
56. Boschi, T., Di Iorio, J., Testa, L., Cremona, M. A. and Chiaromonte, F. (2020) The shapes of an epidemic: Using functional data analysis to characterize COVID-19 in Italy. arXiv, 2008.04700v1
57. Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, Á., Zhou, X., Müller, H. G. and Wang, J. L. (2020) Time dynamics of COVID-19. Sci. Rep., 10, 21040
58. Siordia, J. A. Jr. (2020) Epidemiology and clinical features of COVID-19: A review of current literature. J. Clin. Virol., 127, 104357
59. Jiang, F., Zhao, Z. and Shao, X. (2020) Time series analysis of COVID-19 infection curve: A change-point perspective. J. Econom., doi: 10.1016/j.jeconom.2020.07.039

© The Author (s) 2022. Published by Higher Education Press