Functional Clustering of Covid-19 Countrywise Pandemic Performance

by

Xi Fang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

The Covid pandemic has lasted for over a year influencing everyone's physical and emotional well-beings. Our work is aimed at exploring the capability of various types of functional data clustering methods on the complex Covid data. We collect the Covid data from the Our World in Data website, where the data source is maintained by the John Hopkins University. In our study, we introduce the clustering methods that come from both non-parametric and model-based families. K-mean alignment method combines curve alignment and k-mean clustering, where there is no parametric assumptions of distribution. On the other hand, funHDDC and funFEM model the clustering on the Gaussian mixture distribution assumptions. funHDDC uses EM-like inference for parameters; funFEM is based on the Fisher EM algorithm, which combines Fisher method and EM algorithm in order to ensure the most discriminant group-specific subspace. We purposed the sequential clustering technique on the three stages of pandemic development. Model-based methods show good clustering stability on each stage compared to the non-parametric method in terms of Adjusted Rand Index (ARI). Through the mapping technique, we can conclude the clusters are very sensitive to the countries having either the most severe Covid cases or the fewest Covid cases in three algorithms. However, for countries that do not have the above extreme conditions, their clusters are unclear. The clustering algorithm, such as funFEM, would downgrade the number of clusters from three to two and others would show large variance in ARI indicating the reduction of the clustering stability.

Keywords: K-means; EM; Gaussian mixtures; Functional principal component analysis; Fisher EM algorithm (FEM); Sequential clustering

Acknowledgement

I would like to express my deepest appreciation to Professor Cristina Anton and Professor Adam Kashlak for always being supportive. Their expertise and valuable comments undoubtedly contribute a lot to this work. Also, I have special thanks for the senior master Tiantian Ma for her wonderful guidance at the very beginning, which helps me familiarize with the research topic much quicker. In addition, I would like to thank my parents and my sister for their suggestions and sympathetic ears. I could not complete the master study without their supports. In the end, thank you the University of Alberta and the department of Mathematical and Statistical Sciences. I really appreciate and will forever remember the experience of studying in this professional and loving family!

Contents

List of Figures

List of Tables

1 Introduction

The coronavirus pandemic broke out in Dec 2019. Until now, it has become a long-lasting and large-scale pandemic, which challenge everyone's physical, social and emotional well being. To prevent the spread of this pandemic, and to ensure the safety of public health, the governments all over the world have been managing to carry out various restrictions and lock down policy.

For example, by the end of May 2020, all the European countries have agreed to carry out mask policy, which requires people to wear a mask when showing up in public. Besides, other restrictions may include: restaurants, cafes and bars can not serve customers indoors; museums, cinemas, gyms and other non-essential shops are temporarily closed during the lock down period; grocery stores need to limit the indoor customer capacity in order to ensure the social distance of 6 feet; except for the medical workers and other essential workers, all the other employees should work remotely instead of working onsite.

However, along with different phase of this pandemic, countries may have different choices and attitude towards the restriction policies and this implies differences in the growth path of this pandemic among countries.

The reaction of the governments varies. Some countries may prefer to immediately implement a strict lock down policy and mask policy to reduce the risk of exposures and infections. Other countries may choose to relax the lock down restriction requirement because of the concern of the impact on social function.

During this hard time, our work is motivated by the will to monitor the growth of this pandemic within each country and to cluster the countries that have similar pandemic growth paths. Since Dec 2019, we've been collecting the Covid data of countries all over the world. In the end, we collected the data from 163 countries. The data records the daily growth of active Covid cases and death cases for each country.

In this thesis, we delve into the research on functional data clustering methods, including both non-parametric and model-based approaches. They are implemented on the discrete consecutive daily Covid case observations, which can be considered to have fine grid and have fulfilled the property of the functional data.

In section two, we introduce the source of Covid data and its components. Also, we introduce our work on the preprocessing of the Covid data. The data preprocessing, especially for the Covid data is quite challenging. Countries may have different timelines for the Covid case records. Besides, the quality of the Covid data for each country is different. We purposed several data cleaning and imputation techniques to solve this problem.

In section three, we introduce the related work of functional data clustering and the methodology of several clustering methods of our interest. Additionally, we bring up the idea of sequential functional data clustering, which could capture the stage by stage changes in the Covid data. In section four and five, we evaluate the performance of each clustering method and compare across methods with various clustering procedures.

2 Functional Data

2.1 Covid Data Collection

We collected the Covid-19 data from Our World in Data website, which is a database trusted by lots of research organizations and media (Science, Nature, BBC, CNN, etc). This dataset keeps tracking the growth of Covid-19 active cases and death cases of 209 countries per day since 31 December, 2019.

The Our World in Data website relies on the Covid data from John Hopkins University, which is maintained by their team of John Hopkins Center for Systems Science and Engineering (CSSE). John Hopkins University updates the Covid data each day by merging the newly collected data from governments and national organizations all over the world.

At hand the Covid data consists of (1) the cumulative total of Covid cases and death cases for each country; (2) the number of daily increase Covid cases and death cases for each country. (3) the number of (1) and (2) scaled by the country population per million.

2.2 Covid Data Pre-processing

We mainly work on the scaled version of daily increased Covid-19 cases. In this way we elicit the influence of various population scales across countries, which makes the daily increased Covid cases an ideal indicator to monitor the growth of Covid cases.

Within the raw data set, however, not all countries follow the same timeline and start recording the Covid case at the same time. Some countries begin late in the middle of March and others may start recording in the early January. In order to align the timelines, we impute zeros for all the missing days before a country has non-zero record since December 31, when the data set gets its first non-zero Covid case recorded.

Not every country was considered in our analysis. If a country starts with a big number of non-zero record, for example, 100 cases per million recorded at the first day when a country report finding active Covid cases, then the data is no longer reliable. We assume that any smoothing methods are not capable of relieving the great lack of record consistency. Therefore, we delete in total 37 countries with first record larger than 5 cases per million. Most of them are island countries with small populations.

In general, our Covid data records positive daily increase. However, this is not always the case. There are correction days which have negative values to reconcile the previous record mistakes.

To handle this problem, we replace the negative value with the average of its neighbour positives, and then subtracted the share of this change for all the days before the correction day, in order to: (1) maintain the monotonic growth of Covid data; (2) guarantee the magnitude of cumulative case still equivalent to the unrefined version; (3) implement the correction that allows the revised data to stay as close as possible to its true growth nature.

Nevertheless, a big correction number sometimes can be troublesome. For example, if one country has 200 cumulative cases by the end of March and then suddenly have a negative correction of 150 cases on April 1st, we would wonder whether the previous Covid records of this country is trustworthy, since this is a huge proportion of correction compared to the total cases. In the end, we make use of 50% as the threshold. If the absolute value of the negative correction is larger than that of the 50% of the previous cumulative total, then we would remove all of the records of this country. In total, we remove 9 countries and consider their Covid records are less trustworthy.

After following the previous steps, we then smooth and interpolate the data with a kernel smoother. The smoother applies the classic Gaussian kernel. We set the bandwidth equals to five, which is the minimum required bandwidth to smooth out all the missing values within our data set since December 31, 2019. Kernel smoother ensures positive interpolation of daily case increase and thus guarantees the monotonically increase of cumulative cases.

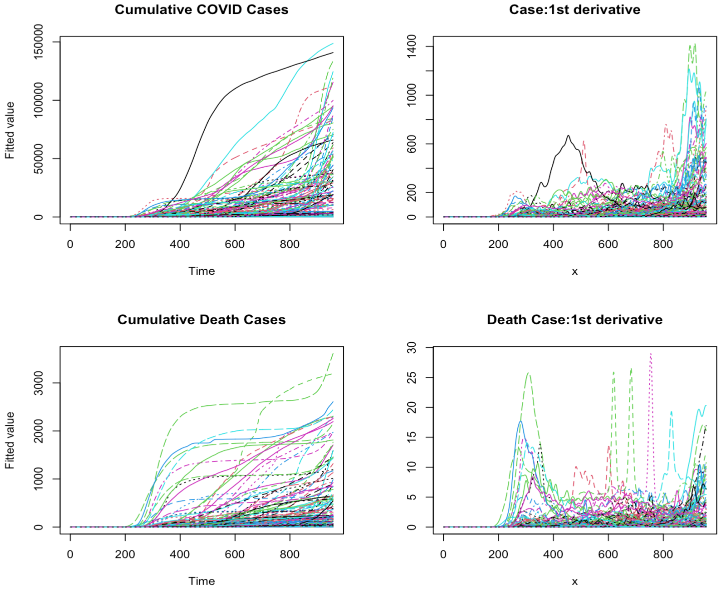

Figure 1: Covid Data from Dec 31th 2019 to Nov 13th 2020 after smoothing

2.3 Features of Functional Data

For a single phenomenon, one can have several observations at different time points in the range (t$_{min}$, t$_{max}$). For instance, the observation at the j$_{th}$ time point can be expressed by X(t$_j$), a discrete point taken its value at a multidimensional space.

However, observations can become more and more consecutive when the grid gets finer. Then a continuous family defined by $\mathcal{X}$={X(t); t:∈(t$_{min}$, t$_{max}$)} will be a good expression of observations considering the high-dimensional aspect of the data Ferraty and Vieu [2006].

The collected Covid data for each country is observed and recorded day by day. After data smoothing and interpolation, the Covid data has even finer grid. Hence, in this paper, we assume it as a functional data set, in which, theoretically, the variable becomes continuous and can take values in an infinite dimensional space.

3 The Clustering of Functional Data

Our research is aimed at finding a proper way to cluster countries through a broad exploration across the existing functional data clustering techniques. Ideally, a promising clustering technique should be robust, which means the clustering results will not have a big difference within a large number of experiments. Secondly, it should ensure result interpretability: countries within the same cluster will share some similarity in the growth feature of Covid cases; countries in different groups should show distinctive differences.

Unsupervised learning of Covid data is fascinating because of its uncertainty. There might be no best answer for unsupervised clustering. With the complex Covid data, the unsupervised clustering would be a more difficult task. In this case, the evaluation of the clustering results would be quite crucial and we choose the evaluation metrics from two perspectives, robustness and interpretability.

The robustness would evaluate the consistency of the clustering results in iterations and the interpretability would evaluate whether the clustering results have the practical meanings.

In this section, we introduce the methodology of several non-parametric and model-based functional data clustering techniques. Also, we introduce several evaluation metrics, including the Adjusted Rand Index(ARI) and the mapping technique, which evaluate the clustering robustness and interpretability respectively.

3.1 Related Work of Functional Data Clustering

Theoretically, functional data has infinite dimensional space, which is the main source of difficulty caused in the modeling process. To solve this problem, one of the most common solutions is to represent the discrete observations with several basis functions. The finite dimensional space spanned by the basis function could transform the functional data clustering problem into a regular model-based clustering method.

The functional principal component analysis (FPCA) can be an ideal candidate for the space representation on account of its interpretability and the superiority of data presentation Jacques and Preda [2014]. The work of Peng and Müller [2008] implements the k-means clustering on the principal component scores. They first assume the space can be spanned by some basis functions, then identify the basis function coefficients.

Another way is to treat the coefficients as random variables. The random variables can have group-specific probability distribution for clusters. funHDDC purposed by Bouveyron and Jacques [2011] builds up the model of probability distribution on FPCA scores, as an extension of the High-dimensional data clustering Bouveyron et al. [2007]. It is also worth to remark funHDDC have parsimonious assumptions on the principal components variances that offers different groups of submodels by applying various assumption restrictions. funHDDC computes its coefficients through a EM-like algorithm. While funFEM purposed by Bouveyron et al. [2015], inherits the idea of funHDDC, it keeps adding on the Fisher step before the EM steps to ensure the group-specific functional subspace is the most discriminant. funclust purposed by Jacques and Preda [2013] is based on the density approximation Delaigle and Hall [2010] of the functional variables. The parameters in the parametric mixtures models are also estimated by a EM-like algorithm.

For probabilistic model-based clustering methods, the likelihood-based selection criteria, such as BIC Schwarz [1978] and AIC Akaike [1974] are often used to evaluate the model fit and to select the optimal model. The funHDDC algorithm and funFEM algorithm all make use of BIC to choose the optimal submodel.

Unlike the previous works on model-based functional data clustering, the non-parametric clustering, such as k-means alignment clustering method purposed by Sangalli et al. [2010] is also of our interest.

This k-means alignment clustering method first considers to combine curve alignment with a functional k-means clustering. The curve alignment decouples the phase variability and amplitude variablility within curve, combined with the similarity evaluation across curves, it makes the k-means clustering more efficient.

Except for the k-means and the model-based clustering methods, the hierarchical clustering Ward [1963] is also very popular. The hierarchical clustering would start by treating each observation as a separate cluster, then iteratively merging the most similar clusters together. Considering the complexity of the implementation of the hierarchy clustering and the limited time, we do not include the hierarchical clustering method in this paper.

3.2 K-Means Clustering of Functional Data

The proper alignment of curves often plays a key role in the functional data clustering. Figure 1 shows the growth curves of Covid cases of 163 countries and their first derivatives. Looking at the derivative curves, there are growth spurts happened at different times. Some grow quite fast at an early stage and then quickly slow down; some take their time and do not get a growth spurt until the middle stage; others, however, start the spurt late but grows at a increasingly fast pace.

So the questions is: Do the above three groups represent three distinct curve shapes, or rather some of them follow the same growth path as long as we properly align them?

To figure out these questions, we take advantage of a k-means alignment clustering approach proposed by Sangalli et al. [2010].

3.2.1 Warping function and the similarity index

In this section, we first introduce the warping function and the similarity index defined in the work of Sangalli et al. [2010]. Assume that we have N curves $\mathbf{c}_i \in \mathcal{C}$, i = 1,....N, $\mathcal{C}$ is the set of curve.

In order to align $\mathbf{c}_i$ and $\mathbf{c}_j$, the following similarity index $\rho(\cdot,\cdot)$ is purposed:

$$\rho\left(\mathbf{c}_i,\mathbf{c}_j\right) = \frac{1}{d}\sum_{p=1}^{d} \frac{\int_R c'_{ip}(s)c'_{jp}(s)ds}{\sqrt{\int_R c'_{ip}(s)^2 ds}\sqrt{\int_R c'_{jp}(s)^2 ds}}, \tag{1}$$

where c_{ip} is the pth component of c_i and $\mathbf{c}_i = (c_{i1},\ldots,c_{id})$.

In this way, the similarity index averages the cosine values of the angle between two vectors: the derivative of c_i and the derivative of c_j. The maximal value of similarity index will be reached when two curves are identical. The similarity index is robust for any shifts and dilation, which means $\rho\left(\mathbf{c}_i,\mathbf{c}_j\right) = 1$ if $c_{ip} = A_p c_{jp} + B_p$.

The choice of warping function is also crucial and should be jointly considered with the choice of similarity index. The warping function can be defined as:

$$W = \{h : h(s) = ms + q\}, \tag{2}$$

where the dilation parameter $m \in R^+$ and the shift parameter $q \in R$ decouple the phase and amplitude variability within curves.

3.2.2 Domain of attraction and labelling function

Assume we have k template curves $\underline{\varphi} = \{\varphi_1,\ldots,\varphi_k\}$, where $\underline{\varphi} \subset \mathcal{C}$ and $\mathcal{C}$ represents the set of curves. Define the domain of attraction of a template curve φ_i as:

$$\Delta_i(\underline{\varphi}) = \left\{\mathbf{c} \in \mathcal{C} : \sup_{h\in W} \rho\left(\varphi_i, \mathbf{c}\circ h\right) \geq \sup_{h\in W} \rho\left(\varphi_j, \mathbf{c}\circ h\right), \forall j \neq i\right\} \tag{3}$$

Then along with the domain of attraction, define the labelling function as:

$$\lambda(\underline{\varphi}, \mathbf{c}) = \min\left\{j : \mathbf{c} \in \Delta_j(\underline{\varphi})\right\} \tag{4}$$

The value of the labelling function would indicate the choice of one template over other templates. The template chosen by the labeling function would have the largest similarity index with curve c compared to other templates. Then curve c would be aligned to this chosen template and be labelled.

3.2.3 Clustering and alignment steps

The optimization of clustering and alignment can be divided into two steps. First, find a set of k templates, and a set of warping functions h, such that:

$$\frac{1}{N}\sum_{i=1}^{N} \rho\left(\varphi_{\lambda\left(\underline{\varphi}, \mathbf{c}_i\right)}, \mathbf{c}_i \circ h_i\right) \geq \frac{1}{N}\sum_{i=1}^{N} \rho\left(\psi_{\lambda\left(\underline{\psi}, \mathbf{c}_i\right)}, \mathbf{c}_i \circ g_i\right), \tag{5}$$

where g is defined as a warping function of another set. This inequality should be satisfied for any other set of k templates and warping functions.

The next step is in charge of labelling. The curve c_i will be labeled by $\lambda\left(\underline{\varphi}, \mathbf{c}_i\right)$ and then the curve would be aligned to the template $\varphi_{\lambda\left(\underline{\varphi}, \mathbf{c}_i\right)}$ accordingly.

Ideally, the above two-step optimization steps should solve the problem. However, the problem in the first step is hard to solve. Hence, the k-means alignment algorithm iteratively runs two steps. The template identification step estimates the k templates identified in the previous assignment and alignment step, then comes back to the assignment and alignment step to align the curves with the previously estimated k templates.

Iteratively, the algorithm should be able to approach the optimal solution of k-means alignment-based clustering.

3.3 The Functional HDDC method

In this section, we introduce functional high-dimensional data clustering (funHDDC) method, which is a model-based clustering algorithm. This functional version of HDDC method takes advantage of the functional-specific latent mixture model to project the functional data into group-specific subspaces, and thus improves both the clustering performance and the result interpretability Bouveyron and Jacques [2011].

Unlike k-means clustering, the model-based method is based on the parametric Gaussian mixture model. Considering the infinite dimensional problem, model-based clustering methods usually first manage to reduce the dimensions. There are many ways to achieve this goal, such as the discretization of observed curves, representation of a group of basis or functional principal components (FPCA).

The discretization of observed curves along the time intervals is often regarded as the most straight-forward approach to solve the infinite dimensional problem. However, after the discretization, sometimes, we may still have more dimensions than the number of observations, which then turns out to be a high-dimensional problem. Hence, the idea of High-dimensional data clustering (HDDC) Bouveyron et al. [2007] is brought up in this case, which transforms the high-dimensional data into group-specific subspaces.

3.3.1 Data format transformation through basis expansion

The discrete observations for the i^{th} observed curves at j time points can be expressed as $x_{ij} = x_i\left(t_{ij}\right)$. However, in functional form, we assume the observed curves $\{x_1, \ldots, x_n\}$ are independent sample trials of a L_2 -continuous stochastic process $X = \{X(t)\}_{t\in[0,T]}$. Hence, the transformation from discrete observations to the continuous functional curves is necessary Bouveyron and Jacques [2011].

One way to achieve this is by representing the functional form with a basis expansion:

$$X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t), \tag{6}$$

where $\{\psi_1, \ldots, \psi_p\}$ is the basis and $\gamma = (\gamma_1(X), \ldots, \gamma_p(X))$ is a random vector in R^p. The basis expansion can be estimated through interpolation procedure. Coefficients, at the following steps, are fitted by the group-specific latent mixture models.

3.3.2 The group-specific functional latent mixture model

In this section, we introduce the group-specific functional latent mixture model purposed by Bouveyron and Jacques [2011].

First let us assume that in the kth cluster, we have n_k observed curves. Their coefficients are $\{\gamma_1, \ldots, \gamma_{n_k}\} \subset R^p$. We assume the coefficients are independent and all of them are of a random vector Γ. We then consider the stochastic process of the kth cluster can be represented with d_k dimensional latent subspace, where $d_k \leq p$.

Therefore, let the first d_k entries of the basis, $\{\varphi_{kj}\}_{j=1,\ldots,d_k}$ in $L_2[0,T]$ be the group-specific basis of the latent subspace $E_k[0,T]$. We obtain the group-specific basis through linear transformation:

$$\varphi_{kj} = \sum_{\ell=1}^{p} q_{k,j\ell}\psi_\ell, \tag{7}$$

where $q_{k,j\ell}$ are the elements in the orthogonal $p \times p$ matrix Q_k.

Let us split the matrix $Q_k = (q_{k,j\ell}) = [U_k, V_k]$ into U_k and V_k, where $U_k^t U_k = I_{d_k}$, $V_k^t V_k = I_{p-d_k}$ and $U_k^t V_k = 0$. The dimension of U_k is p by d_k and the dimension of V_k is p by $(p - d_k)$.

We assume the corresponding latent expansion coefficients of the group-specific basis $\{\varphi_{kj}\}_{j=1,\ldots,d_k}$ are also independent and are of a latent random vector $\Lambda \in \mathrm{R}^{d_k}$. The linear transformation between Γ of R^p and Λ of R^{d_k} therefore is:

$$\Gamma = U_k\Lambda + \varepsilon, \varepsilon \in R^p, \tag{8}$$

where ε is the random noise.

The assumptions on their distributions are:

$$\Lambda \sim \mathcal{N}(m_k, S_k) \tag{9}$$

$$\varepsilon \sim \mathcal{N}(0, \Xi_k) \tag{10}$$

$$\Gamma \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{11}$$

where Γ and ε all follow multivariate Gaussian density. m_k is mean of the kth group and $S_k = diag(a_{k1}, \ldots, a_{kd_k})$ is its corresponding covariance matrix. $\mu_k = U_k m_k$ and $\Sigma_k = U_k S_k U_k^t + \Xi_k$.

We assume that the Σ_k satisfies:

$$\Delta_k = Q_k^t \Sigma_k Q_k = diag(a_{k1}, \ldots, a_{kd_k}, b_k, \ldots, b_k), \tag{12}$$

where b_k models the noise term and the a_{kj} models the variance of the kth group. The advantage of this model is that it manages the clustering in a low-dimensional space through group-specific projection, however, the discriminative information is kept by the noise term b_k.

The density of γ hence follows a mixture of Gaussian distributions:

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; \mu_k, \Sigma_k) \tag{13}$$

where the prior probability is $\pi_k = P(Z_k = 1)$. Z_k is equal to 1 if the curve belongs to the kth group.

3.3.3 EM-based parameter estimation and MAP-based clustering

The parameters a_{kj}, b_k, Q_k and d_k are estimated through the iterative EM steps. The E step calculates the expectation first and the following M step updates the value of the parameters to maximize the likelihood. However, the hyper-parameter number of cluster K and the dimension of subspace d_k are left unknown. The hyper-parameters can not be estimated by maximizing the likelihood. Instead, the intrinsic dimension d_k should be chosen through the threshold of eigenvalues scree. Like the number of cluster K , the threshold can also be tuned by BIC.

By maximizing a posteriori (MAP) rule, γ_i would belong to the group that has the highest density $P\left(Z_{ik}=1 \mid \gamma_i\right)$, where Z_{ik} is an indicator for the kth cluster. If the ith curve belongs to the kth cluster, then $Z_{ik}=1$.

3.3.4 Parsimonious functional latent mixture models

The orginal functional latent mixture model can be referred as FLM$_{[a_{kj}b_kQ_kd_k]}$. Nevertheless the model can be more parsimonious by fixing several parameters to be common across different classes Bouveyron et al. [2007]. For example, if b_k is common, then the submodel becomes FLM $[a_{kj}bQ_kd_k]$. This submodel assumes there is no difference between the noise outside the group-specific subspaces.

In the experiments on Covid data, except for the original model, we select five submodels out of 28 options: FLM$_{[a_{kj}b_kQ_kd_k]}$, FLM$_{[a_{kj}bQ_kd_k]}$, FLM$_{[a_kb_kQ_kd_k]}$, FLM$_{[ab_kQ_kd_k]}$, FLM$_{[a_kbQ_kd_k]}$ and FLM$_{[abQ_kd_k]}$, considering their good performance in the clustering of Canadian temperature dataset Bouveyron and Jacques [2011].

The full model can be represented as FLM$_{[a_{kj}b_kQ_kd_k]}$. By assuming the noise outside groups are common, FLM$_{[a_{kj}bQ_kd_k]}$, FLM$_{[a_kbQ_kd_k]}$ and FLM$_{[abQ_kd_k]}$ are purposed; on the top of that if we assume the variance across groups are common, we have FLM$_{[abQ_k}$.

In the nested runs, despite initial randomization, we wrap a loop for all of the six models. Only the model that leads to the minimal BIC value will be chosen in the end.

3.4 The Discriminative Functional Mixture Model

In this section, we introduce funFEM algorithm. This functional data clustering algorithm is proposed by Bouveyron et al. [2015] on 2015. They adapt the work of Bouveyron and Brunet [2011] on Fisher discriminative subspace of functional data in multivariate case so that the funFEM is able to cluster in the discriminative functional context.

3.4.1 The discriminative functional mixture model

The functional data clustering is based on the discriminative function mixture model. At this step, funHDDC and funFEM, they share the similar latent mixture model assumptions. They all assume that the density of curve coefficients γ follows a mixture of Gaussian distributions and γ_i would belong to the k_{th} group that has the highest probability $P\left(Z_{ik}=1 \mid \gamma_i\right)$.

3.4.2 Three-step model inference

Unlike funHDDC, funFEM does not make use of the EM algorithm for model inference. Instead, it takes advantage of F (Fisher) step first to make sure that the functional subspace F is most discriminant, where $F[0,T]$ is a latent subspace, spanned by some basis functions in $L_2[0,T]$.

At the q-1 iteration of E step, the posterior probability is $t_{ik}^{(q)}=E\left[z_{ik} \mid \gamma_i, \theta^{(q-1)}\right]$. Conditioned on $t_{ik}^{(q)}$, following the idea of Fisher, this F step determines the orientation matrix U of the most discriminant functional group-specific subspace F. This discriminant subspace should have maximal between groups variance while minimal within group variance Fisher [1936].

After the F step, the M step follows. This step maximizes the log-likelihood $Q\left(\theta;\theta^{(q-1)}\right)=E\left[\ell\left(\theta;\mathbf{\Gamma},z_1,\ldots,z_n\right) \mid \mathbf{\Gamma},\theta^{(q-1)}\right]$ conditionally on the orientation matrix $U^{(q)}$ from F step. The maximization of log-likelihood yields updates for $\pi_k^{(q)}$, $\mu_k^{(q)}$, $\Sigma_k^{(q)}$ and $\beta_k^{(q)}$ at the $q^{(th)}$ iteration.

Finally, the E steps takes advantage of Bayes' theorem to calculate $t_{ik}^{(q)}=P\left(z_{ik}=1 \mid \gamma_i,\theta^{(q)}\right)$ when the ith curve belongs to the kth cluster. The posterior probabilities $t_{ik}^{(q)}$ can be calculated as:

$$t_{ik}^{(q)}=\frac{\pi_k^{(q)}\phi\left(\gamma_i,\theta_k^{(q)}\right)}{\sum_{l=1}^{K}\pi_l^{(q)}\phi\left(\gamma_i \mid \theta_l^{(q)}\right)}, \tag{14}$$

where $\theta_k^{(q)}=\left(\pi_k^{(q)},\mu_k^{(q)},\Sigma_k^{(q)},\beta^{(q)}\right)$ contains the parameters we updated from the previous M step.

3.4.3 Submodels of funFEM

By applying constraints on the parameters of Δ_k, we have up to 12 submodels of funFEM. The full model is $\mathrm{DFM}_{[\Sigma_k \beta_k]}$. If we assume that the covariance matrices Σ_1, ..., Σ_k are common across all the subgroups, then the DFM model can be simplified to $\mathrm{DFM}_{[\Sigma \beta_k]}$. Besides, the Σ_k can also be assumed spherical or diagonal. The spherical constraints can be applied to $\mathrm{DFM}_{[\alpha_k \beta_k]}$ while the diagonal constraints can be applied to $\mathrm{DFM}_{[\alpha_{kj} \beta_k]}$.

Moreover, the submodels may also apply more relaxed constraints on the noise parameter β_k by assuming that it is common across all the groups. For example, with common noise and diagonal covariance matrices, the funFEM submodel would be $\mathrm{DFM}_{[\alpha_{kj} \beta]}$.

In total, there are 12 submodels on account of all the possible constraints. In all of the following experiments, we will use "all" model mode, which means for each run of this clustering algorithm, we will give 12 nested runs, each nested run will test one specific submodel and returns its BIC value. The submodel with the smallest BIC value in the end will be chosen.

4 Application to Covid Data

In this section, we lay out the application of three clustering methods to the Covid dataset, including the evaluation of clustering stability through adjusted rand index (ARI), the mapping of clustering results and the sequential clustering techniques.

4.1 Application of k-means alignment method

In this section, we introduce the application of k-means alignment method. This algorithm is implemented in R package *fdakma*.

4.1.1 Parameter Tuning

There are several parameters that can be tuned in this algorithm: (1) number of clusters; (2) initial centers; (3) warping methods, of which options are curve shift and curve dilation; (4) similarity method, of which options are the measurements of functions' first derivatives and the measurements of distance between functions.

Unlike classification problems where a correct label for each class member is known, the best solution to a proper clustering for the Covid data set is left unknown. So in this section, we applied the clustering technique with a wide range of exploration on the parameter tuning, from which we start to approach the optimal.

Intuitively, we start with three clusters. Looking at the cumulative curves in Figure 1, their growth paths are of three levels: mild, moderate, aggressive. If we consider the distinctive spikes we find in the picture of first derivatives, the number of clusters could be more than three.

In order to align the time lines of curves without spoiling the difference among curve amplitudes, only the curve shift is allowed in the alignment. We choose the scaled L_2 distance between derivatives for similarity measures, in order to capture the trend of growth directly.

By default, this algorithm randomizes the initial cluster centers for each run. We run this clustering jobs 100 times and summarize the results. But the question arises: How do we estimate the consistency of clustering results in 100 runs?

4.1.2 Adjusted Rand Index

Adjusted Rand Index can be a good measure of clustering consistency. It is the corrected-for-chance version of the Rand index Vinh et al. [2010]. The Rand index is named after William M. Rand. He brought up this concept for evaluating the similarity between two clusters Rand [1971].

Given n elements in one set, at hand we compare two partitions into K subgroups: $A = \{A_1, \ldots, A_K\}$ and $B = \{B_1, \ldots, B_K\}$.

The Rand Index is thus defined as:

$$R = \frac{a+b}{a+b+c+d}$$

where a is the number of pairs in the same subset of A and B; b is the number of pairs in the different subset of A and B; c and d in turn count the pairs that are in the different subset of one partition way and the same subset of the other.

Partition	A	B
A	a	c
B	d	b

Table 1: The number of pairs in different subsets of A and B partitions

The numerator of Rand Index evaluates the consistency of clustering in partition A and partition B. Intuitively, we could get the idea: the larger the value of Rand Index, the higher the level of similarity between two clustering results. Adjusted Rand Index(ARI) employs a correction for chance in order to take into consideration the difference in modeling random clustering, but it shares the same idea of similarity measurement with Rand Index.

4.1.3 Clustering on the whole Covid data

For each run, we compare the clustering results in pairs. One pair is generated in each run. So in 100 runs, we have in total 100 pairs and 100 ARI calculating the similarity of clustering per pair.

Figure 2 uses boxplot to describe the distribution of ARI in 100 runs. The boxplot on the left shows ARIs computed for Covid case-based clustering. The median of box is close to 0.5 and the variance is large. Boxplot on the right shows the ARIs computed for the death case-based clustering. The ARI has a higher median close to 0.9 and the spread of boxplot gets smaller.

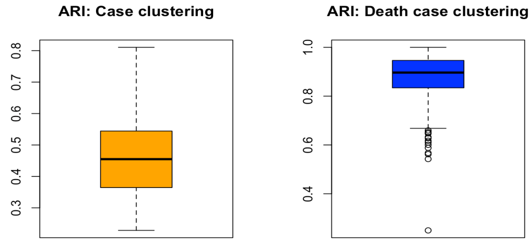

Figure 2: Boxplot of ARI in 100 runs

In this section, we explore the k-means clustering on the whole timeline of Covid data. Nevertheless, ARI shows the lack of consistency of this clustering method. Looking into the derivative curves in Figure 1, there are spikes hard to align with each other, which may cause difficulty in the curve registration step. We also did a broad exploration on the choice of cluster numbers. However, from two clusters to six clusters, neither upgrade or downgrade the cluster complexity improves the clustering consistency.

So in the next section, we bring up the idea of sequential clustering. Sequential clustering divides the whole timeline into three stages. It simplifies the registration of curves and gains more consistency of clustering at each stage.

4.1.4 Sequential Clustering

In this section, we split the Covid data into three subsets. The timelines for three stages are: stage one, from 2019-12-31 to 2020-04-30; stage two, from 2020-05-01 to 2020-08-31; stage three, from 2020-09-01 to 2020-11-13.

The partition criteria of stages follows closely to the turning points during the spread of this pandemic. The stage one ends by the time most of the countries started to take actions on epidemic prevention and control, such as the lock down policy, mask policy and the social distance requirement in public places.

The stage one takes four months. Similarly, the second stage lasts for another four months. It can be a representative of summertime, when the prevention and control policies have been implemented: people work from home and students have their summer vacations. The third stage can be considered as a activity recovery phase, where moderately easy control policies are implemented instead.

The box plots in figure 3 and figure 4 display the distribution of ARI in 100 runs at three stages, for Covid case-based clustering and death case-based clustering, respectively.

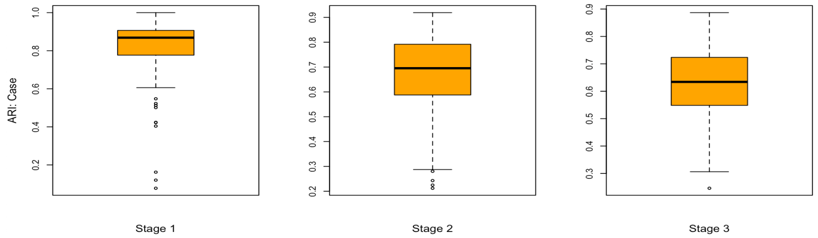

Figure 3: ARI of case-based k-means clustering at three stages

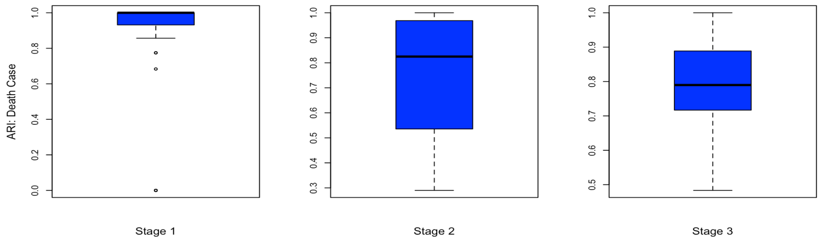

Figure 4: ARI of death case-based k-means clustering at three stages

The improvement of the case-based clustering consistency in 100 runs is very straightforward. Also, the sequential clustering allows us to stay tuned for the cluster change along the three time periods. We pick up the pairs with the highest ARI value at each stage, then mapping the cluster labels to a real world map. Figure 5 consists of three cluster maps on the results of the Covid case-based clustering and death-based clustering at three stages.

At the first stage, we downgrade the number of cluster from three to two. Less clusters, stronger clustering consistency. This is what we found when tuning the complexity of our cluster structures.

At the beginning of this pandemic, the situations of most countries are not complicated. A small group, which is the countries in red in the map of stage 1, begins this pandemic earlier than the rest of the world (countries in blue).

China on the map is in grey at the first stage, because we do not include China in the first clustering. China started this pandemic way ahead of the rest of the world which makes itself difficult to align with. The average ARI has increased 0.4 with two clusters, which has proved that two clusters can be a good choice for the clustering at stage one.

The pair with highest ARI shows good consistency in clustering between runs and can be considered as near optimal solution at each stage. The maps generated by such pairs also proves its interpret-ability.

In the sequential clustering of Covid cases, the maps are telling stories. At the first stage, the pandemic first breaks out in the red countries, mostly within North America and European countries, while countries in the blue cluster are the majority and are the safer places mostly without the impact of coronavirus.

At the second stage, the number of clusters have been upgraded from two to three, and blue cluster, used to be the majority, is losing its members. Russia, India, Argentina, Saudi Arabia and Iraq joined the red cluster; Brazil, even worse, got into the purple, which indicates this country is facing surprisingly bad situation with US. Red cluster is in the second place. Countries in red have better situation than countries in purple but worse than the blue ones. Many European countries have come back and joined the blue group at this stage, however, Spain and Sweden are left behind.

At the third stage, the blue cluster still dominates, but the cluster members keep switching their places. Argentina made its second move, from red to purple, while Canada, Libya and Iran jump into red. Spain falls into a worse situation, into purple, along with Peru, Brazil and US, which are the old purple members for two stages.

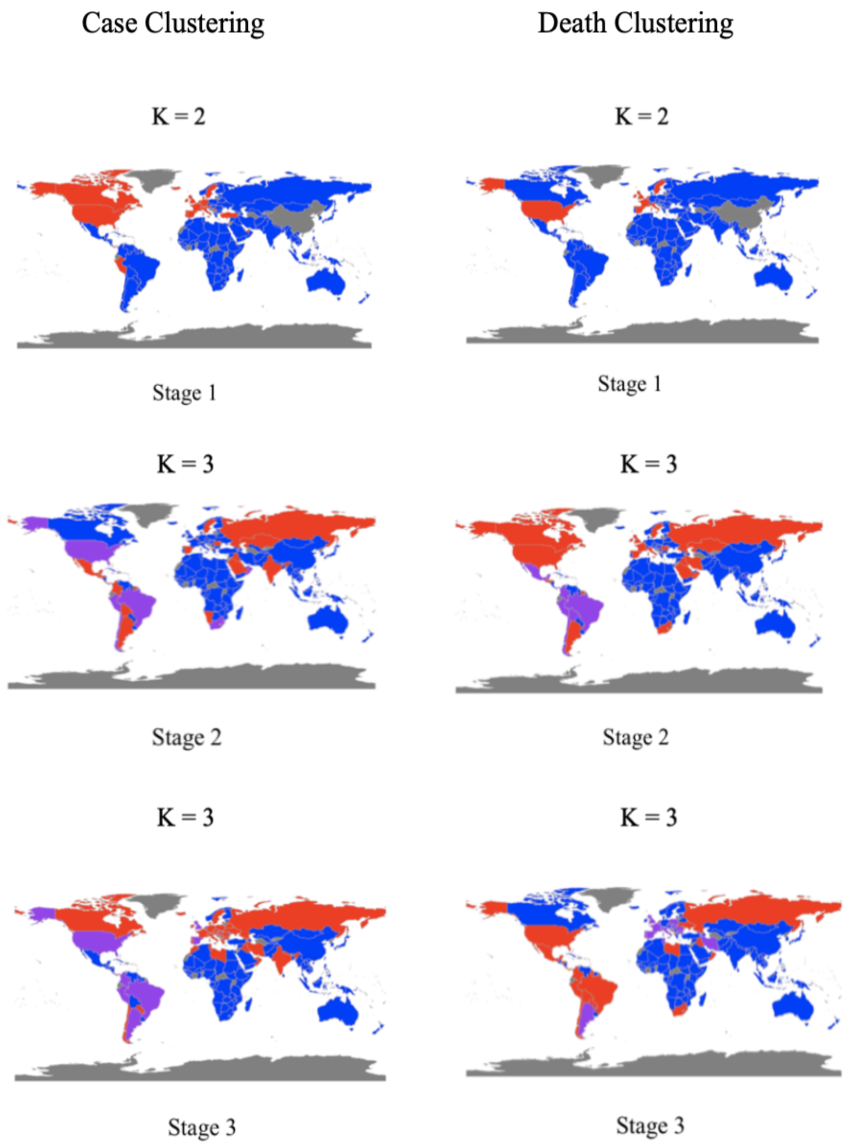

Figure 5: Maps of case-based and death-based k-means clustering at three stages

The maps in figure 6 tell a similar story on the growth of death cases in this pandemic. As the clustering of active cases and death cases share the same split of timeline, it is expected that country clusters will be mostly synchronous at three stages. Besides, intuitively, we would conclude that countries fall into the rather worse group in the case-based clustering will have a great chance to stay within the same level group of the clustering for death.

4.1.5 Confidence Interval of mean ARI

To further mathematically describe the distribution of ARI, we propose the use of confidence interval for mean ARI. The idea of confidence interval can be validated by the central limit theorem when the sample size is large enough, for the central limit theorem indicates that as the sample size gets larger, the sample mean should be getting approximately normally distributed.

Hence in this section, we first check out the distribution of mean ARI within 40,000 runs. We have three options for sample size: 50, 100 and 200. Out of 40,000 runs, with n = 50, we calculate the mean ARI of 50 samples; with n = 100, we calculate the mean ARI of 100 samples; with n = 200, we calculate the mean ARI of 100 samples. But it is worth mentioning that the samples are not entirely independent from each other.

We then visualize the distribution of the mean ARI with different sample size. The following histograms present the distribution of the mean ARI calculated for death clustering at three stages. The shapes of the histograms are approximately symmetric when sample size equals to 200, and thus the histograms prove that the distribution of the mean ARI satisfies the central limit theorem.

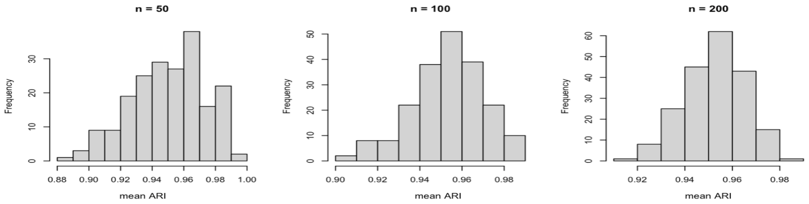

Figure 6: The distribution of mean ARI at the first stage

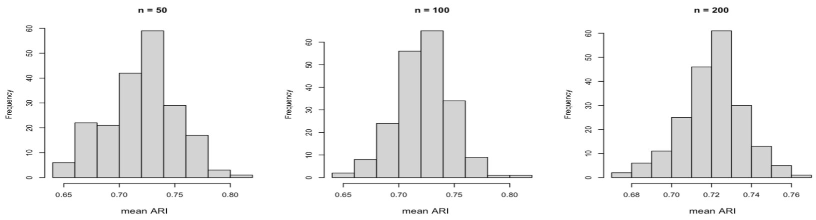

Figure 7: The distribution of mean ARI at the second stage

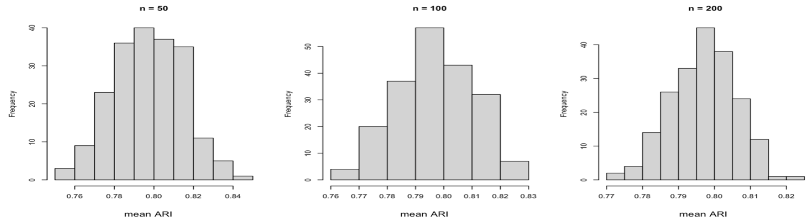

Figure 8: The distribution of mean ARI at the third stage

The following table calculates the confidence interval of mean ARI at three stages. They are not wide intervals as the variance has been sharply reduced by taking average of ARI values.

Sample size	Stage 1	Stage 2	Stage 3
50	(0.949, 0.956)	(0.718, 0.727)	(0.795, 0.799)
100	(0.950, 0.955)	(0.718, 0.724)	(0.796, 0.799)
200	(0.951, 0.954)	(0.719, 0.723)	(0.796, 0.798)

Table 2: Confidence interval of mean ARI at three stages

There are no overlaps between the intervals of first stage and other stages, for the lower bound of CI in stage one is higher than 0.9. All of the confidence intervals are over 0.5, which elaborates the decent clustering stability of the k-means alignment method.

4.2 Application of funHDDC to Covid data

In this section, we introduce the application of funHDDC algorithm, which is implemented in R package $funHDDC$ Schmutz and Bouveyron [2021].

4.2.1 Initial setups and parameter tuning of funHDDC

EM-based algorithms are easy to converge to the local optimal. In our experiment, within each run of this clustering algorithm, there are nested runs with random initialization of EM. Those nested runs will only keep the parameters resulting from the largest likelihood than others. The idea is purposed by Biernacki [2004] on 2004 to avoid the convergence to a local optimal.

There are three hyper-parameters that need tuning. The number of cluster K, the number of basis elements and the value of threshold. K can be 2 or 3 in the tuning process; the number of basis elements has the options of 25, 50 and 100; regarding the threshold, options are 0.1, 0.2 and 0.3. 0.1 is the default threshold value for $funHDDC$ algorithm.

At the beginning of our research, we carry out a much boarder exploration across various parameter combos rather than the current parameter tuning plan. For example, for the number of basis functions, with interval equals to 25, we have options of 25 number of basis, 50 number of basis, 75 number of basis, all the way to 200 number of basis. Beyond that we also have 0.05, 0.1, 0.15 ... and 0.3 for threshold tuning. However, this is extremely computational expensive and there is no big difference between the results with such fine parameter tuning. In the end, we keep 25, 50 and 100 as the number of basis and 0.1, 0.2 and 0.3 as the thresholds.

4.2.2 Clustering on the whole Covid data

Figure 9 describes the distribution of ARI values in 100 runs both for case-based clustering and death-based clustering. The number of basis and threshold are tuned through BIC for each run. In general, the boxplots show extremely high ARI values (close to 1) when the cluster design is less complex (number of cluster $K = 2$). If we increase the number of cluster from two to three, the median value of ARI is still close to 1, though the variance has increased due to the increase of clustering complexity.

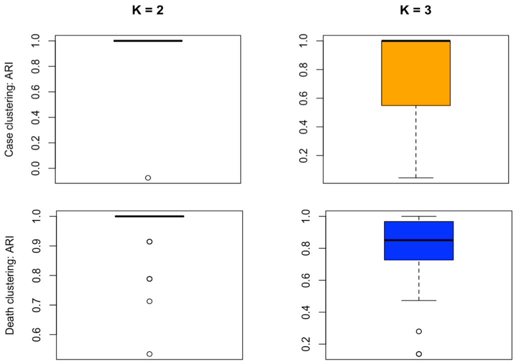

Figure 9: ARI values in 100 runs of different number of clusters

4.2.3 Sequential clustering at three stages

To simplify the task for clustering models, the next step remains to be the discretization of the time intervals into three sequential stages, on which we carry out the $funHDDC$ method. The timeline is divided into the same three stages as that of the k-means clustering, so that their results at each stage are comparable.

Figure 10 and Figure 11 describe the distribution of ARI values in 100 runs of different number of clusters. Both of them indicate that two clusters and three clusters are all able to stabilize the clustering model at the first two stages. They have almost equally satisfactory performance. However, at stage three, they are losing stability, though the median ARI is still close to 1. In this case, lower variance might result from higher number of clusters.

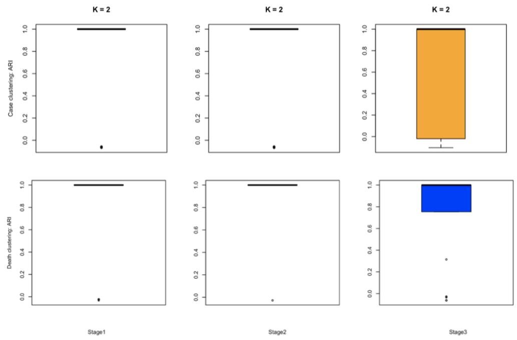

Figure 10: ARI values in 100 runs of two clusters

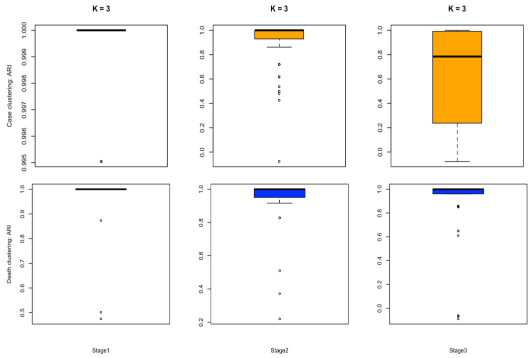

Figure 11: ARI values in 100 runs of three clusters

Overall, the sequential clustering contributes to the decomposition of clustering variance into three stages, also at the same time, increase the interpretability of clustering results. In the next section, we will look into the Covid world maps generated by the highest ARI value pairs at each stage, in order to approach the maximal clustering stability.

4.2.4 Maps of Covid clusters

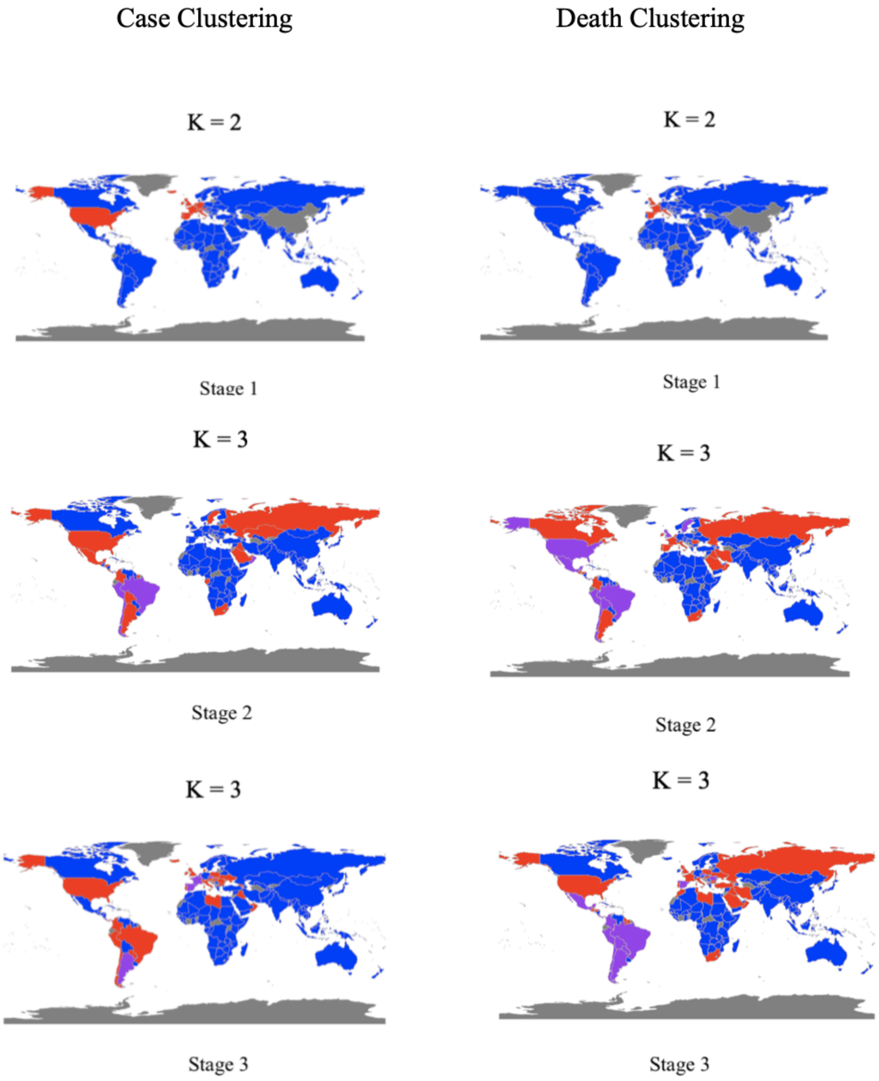

Figure 12: Maps of case-based and death-based funHDDC clustering at three stages

If we look into the maps of active case-based clustering at the three stages, we can tell that in general the maps are mirroring what is happening in real life. At stage one, the countries in red are European countries and the United States. They had a hard time at the beginning of the pandemic. At stage two, the pandemic is spreading quickly across the borders to Brazil, Argentina, Russia and India. Brazil is in the worst case, while other countries in blue, such as China is recovering fast from the attack of the coronavirus. At the third stage, Russian and many European countries are turning back to blue thanks to the Covid restriction policies, while US, Brazil and Argentina still struggle with the most severe Covid cases.

The flags of red and purple in those maps are very similar to the flags we got from the maps of k-mean clustering. Both of the clustering methods are very sensitive to the countries suffering with severe Covid cases. The clustering results of funHDDC show some extent of consistency of that of the k-means alignment method.

4.3 Application of funFEM to Covid data

Finally in this section, we introduce the application of funFEM algorithm tho the Covid data. This algorithm is implemented in R package $funFEM$ Bouveyron [2015].

4.3.1 Parameter tuning

There are two parameters need tuning. One is the number of clusters K, the other is the number of basis elements. To align with the previous experiments of funHDDC, options for K are two and three, while the options for the number of basis are 25, 50 and 100. By default, this algorithm tests 12 submodels in each run; the initiation of Fisher-EM inference algorithm is random; the maximum number of iterations before the stop of the Fisher-EM algorithm is 100.

Considering the randomness of clustering, all of the following experiments will run the clustering algorithm 100 times independently.

4.3.2 Clustering on the whole Covid data

Given a number of basis, the functional model will converge to the same optimum, so that the ARI always equals to 1 in 100 runs. The discriminative function mixture model has the best fit when the number of basis reaches 100.

Given two initial clusters and three initial clusters, the BIC values are on the same scale, though the BIC values of K = 2 are always smaller that the BIC values of K = 3.

The following maps of clusters presents the clustering results on the whole Covid data.

Clustering	K = 2	K = 3
Case	-160343	-153854
Death	-99737	-90243

Table 3: BIC values under different number of clusters

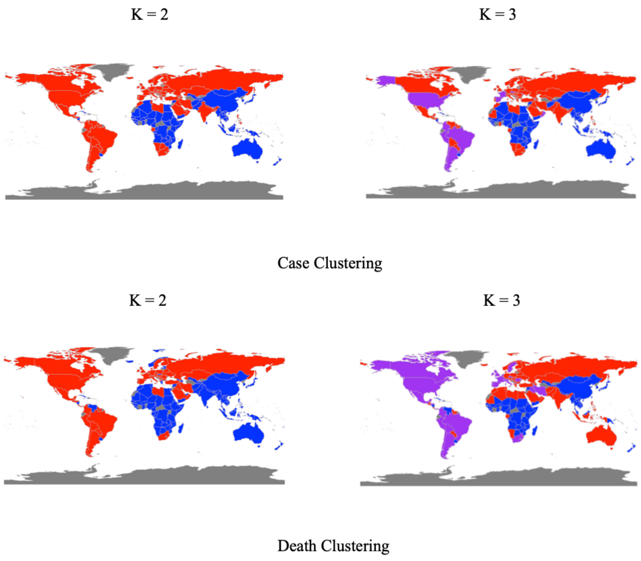

Figure 13: Maps of case-based clustering and death case-based clustering of the whole Covid data

The maps in Figure 13 show that China, Australia and most of the African countires are in blue, which can also be validated by the previous two clustering methods. North America and South America are still in the more severe case compared to the other regions.

4.3.3 Sequential clustering at three stages

The BIC values are comparable between different number of clusters at three stages. Therefore, on the same scale of goodness of fit, we choose the same number of clusters previous clustering methods used in order to prepare for the further comparison among three clustering approaches.

Given a number of basis functions and a fixed timeline, the functional model again successfully converges to the same optimum in 100 runs. With great clustering stability, clustering the data by separate chunks no longer decrease the variance. However, from the cluster maps, the sequential clustering still help to increase the interpretability across three different time points.

The BIC values at three stages are higher than the BIC values returned by the clustering on the whole Covid data, which indicates that in terms of goodness of fit, the sequential clustering no longer improve the fitness of modeling.

Clustering	Stage 1	Stage 2	Stage 3
Case	-109350	-143929	-143384
Death	-57853	-80776	-74900

Table 4: BIC values at three stages

The initial number of clusters set up for the three stages are two, three and three, respectively. In this way, the results of funFEM could align and compare with the clustering results of other methods. The maps are visualizing the clustering result with the highest ARI values out of 100 runs. But it is worth mentioning that the map in the middle of the second column has only two clusters, though its initial cluster setup is three. It warns us that sometimes the cluster downgrade happens in the funFEM and the clustering result does not always go with the initial setups.

But the following maps in Figure 14 still indicate that a good extent of interpretability remains in the sequential clustering process of funFEM. Asian countries and most of the African countries keep staying in the blue cluster, while India, European countries and South America countries, such as Argentina, are jumping across the red group and the purple group.

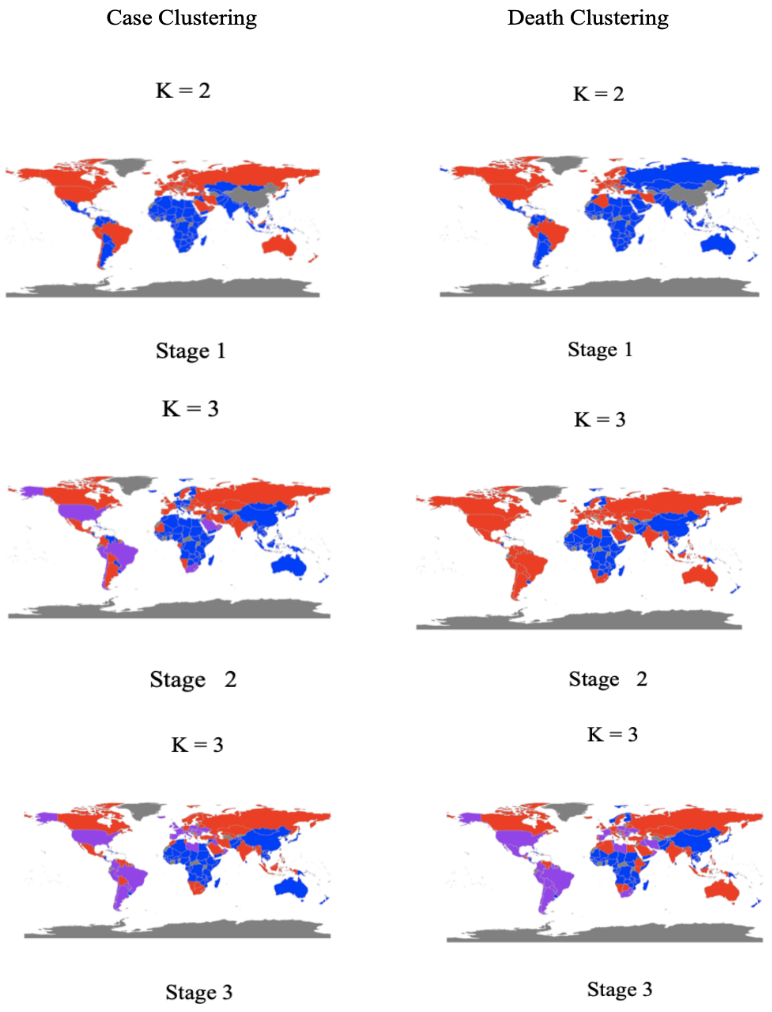

Figure 14: Maps of case-based and death-based funFEM clustering at three stages

The sequential maps generated by funFEM can be validated by the previous clustering approaches for it remains its sensitivity to the countries having most significant severe Covid cases, like Brazil, US and Spain.

However, from the visualization of clustering results, generally speaking, we could only speculate and observe their differences without a proper quantitative evaluation metric. Therefore, in the next section, we will take advantage of ARI to quantify changes among three clustering methods.

4.4 Comparison among three clustering methods

In this section, we compare the classification results of three clustering methods. We first collect the results returned by 100 runs of the application of each clustering method, then randomly draw a pair of the results from two different algorithms to calculate the adjusted rand index (ARI). In total, we draw 200 pairs for each comparison across two algorithms.

The following boxplots visualized the spread of ARI of comparison among three algorithms at three stages. Each row of the boxplot represents the comparison at one stage.

The plots indicate that the cross-algorithm calculated ARIs from the pair K-mean and funHDDC are higher than the other pairs, which means that K-mean and funHDDC have more similar clustering results than the other pairs with funFEM. This happens at three stages. At each stage, the ARI calculated across the clustering result of k-means alignment and funHDDC is over 0.6. So that we may conclude that overall, there is a moderate similarity between the clustering results of k-mean alignment and funHDDC methods.

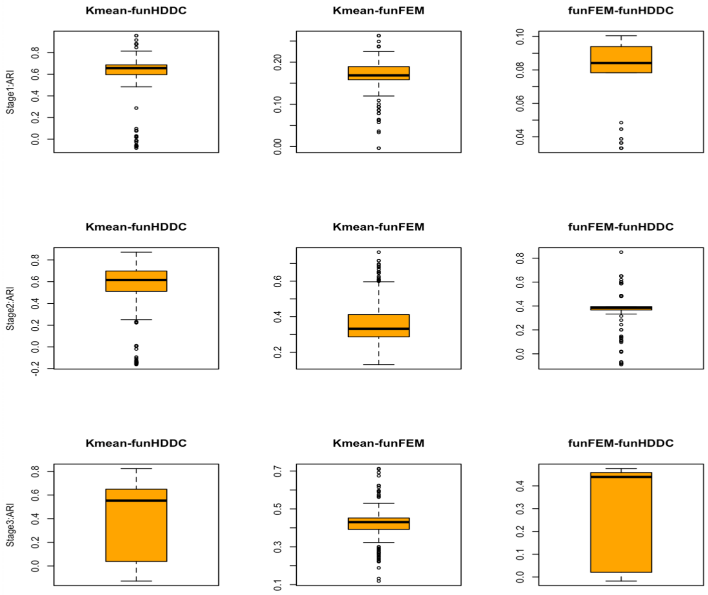

Figure 15: The comparison of case-based clustering results among three algorithms

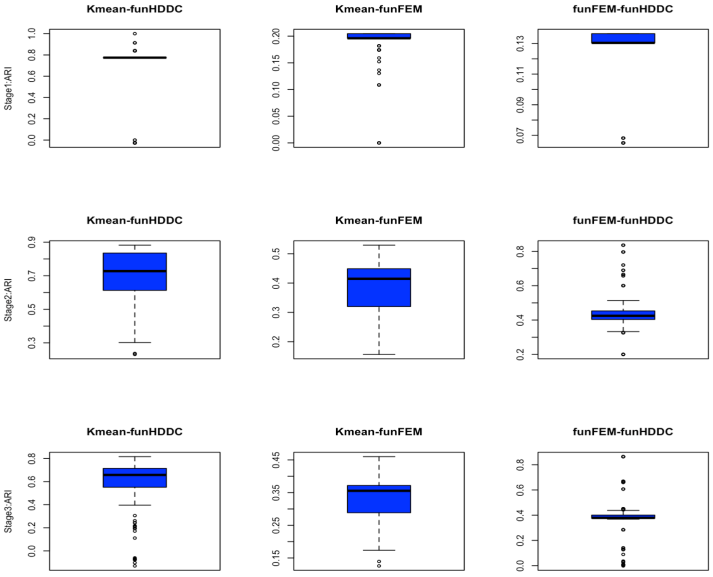

Figure 16: The comparison of death case-based clustering results among three algorithms

Considering there are randomness within the clustering results of 100 runs, at the next step, we manage to select the most stable clustering pair of each algorithm, which is the pair with the highest ARI value in each algorithm, then calculate the ARI across those pairs of different algorithms. In this way, the between group comparisons are with less variance.

The following tables of case-based clustering and death-based clustering validate what we have learned in the boxplots. Still, we find out the k-means and the funHDDC algorithm have the most similar clustering pairs, the ARI values calculated across the pairs of death clustering of k-means and funHDDC can be up to 0.86 at the second stage. However, the ARI values with the pairs of funFEM are still close to zero.

Stage	funFEM-funHDDC	K-mean-funHDDC	K-mean-funFEM
Stage 1	0.08	0.66	0.18
Stage 2	0.48	0.68	0.41
Stage 3	0.47	0.63	0.44

Table 5: Comparison of three methods of case-based clustering at three stages

Stage	funFEM-funHDDC	K-mean-funHDDC	K-mean-funFEM
Stage 1	0.13	0.77	0.2
Stage 2	0.47	0.86	0.45
Stage 3	0.38	0.67	0.37

Table 6: Comparison of three methods of death-based clustering at three stages

One explanation of the low ARI values across other two clusteirng method and funFEM can be the funFEM sometimes fails to keep three clusters. For example, at stage two, the clustering result of funFEM shown in the middle of the second column of figure 14 indicate there are only two clusters. This situation would make ARI quite low when compared with three-class clustering results.

5 Summary & Discussion

Functional data clustering has always been one of the most challenging topics compared to other clustering problems. The infinite dimension, various functional curve shapes and the curve shifts, all of which can become difficulties when researchers want to build up a functional data clustering algorithm.

In the past few years, non-parametric clustering method, such as k-means, and Gaussian mixture model-based clustering methods are quite popular. Researchers are trying to solve the difficulties of functional data clustering from different perspectives and they all show some strengths on the analysis of certain classic functional datasets: for example, the Canadian temperature data and the Berkeley growth study data Ramsay and Silverman [1997].

This year, we have an unprecedentedly large scale and long lasting pandemic. The Covid data of over 150 countries and the cumulative Covid curves having over three waves are making it one of the most difficult functional data for clustering.

In this thesis paper, our contribution is the board survey of different types of functional data clustering methods on the complex Covid data. It is also worth mentioning that unlike other labelled functional datasets, the clustering on the Covid data is unsupervised. The final answer for the clustering is unknown.

Therefore, the difficulties and the questions can be: (1) how to handle the preprocessing of the complex Covid data; (2) how to evaluate the clustering performance when there are no labels and no right answers; (3) with the complex Covid data, which type of algorithms would win the clustering game?

The main contribution of our study is trying to answer and solve the above questions. Firstly, we purposed a four-step Covid data cleaning process, which includes: timeline alignment, data consistency check, the data correction check and the data smoothing. The four steps of the Covid data preprocessing clean all the irregular Covid records and smooth out every negative Covid case correction.

Secondly, we purpose to evaluate the clustering results from two aspects in the unsupervised learning: the clustering stability and the clustering result interpretability. We introduce ARI as the metrics of the clustering stability.

We also introduced sequential clustering method and mapping to evaluate the clustering result interpretability. Sequential clustering technique monitors the changes of clusters along different development stage of this pandemic, and at the same time, reduces the difficulty of stage by stage clustering compared to the clustering on the whole timeline.

Thirdly, we select the functional data clustering methods that come from the very different clustering families. The non-parametric clustering method, k-means, does not have any parametric distribution assumptions, unlike the model-based clustering family: funHDDC and funFEM.

In terms of stability, the model-based method funHDDC and funFEM are better than the k-mean alignment method, though the k-mean method also have decent ARI values at three stages.

In terms of interpretability, all of the clustering methods have the capability to recognize countries with severe Covid cases condition, such as the United States and Spain at the beginning of this pandemic. In the middle stage, Brazil and Argentina have the red flags. The countries that three clustering methods find it hard to harmonize are in the fuzzy zone. Neither do they have the Covid condition as bad as the most severe countries, nor do they have the most clear Covid condition as China, Australia, and others.

The ARI can be very sensitive to the changes of countries in the fuzzy zones when comparing the clusteirng results of three algorithms across different stages. Also, the ARI is quite sensitive to the changes of the number of clusters. Though the initial number of clusters can be tuned and fixed in each algorithm, the downgrade of clusters could still happen from time to time in funFEM.

Overall, we find out both non-parametric and model-based clustering method have decent clustering stability. The clustering stability for death cases are generally better than that of the active cases. By reducing the number of clusters, the clustering stability will increase, which can be indicated by a smaller variance of ARI. All of the three clustering methods can be good indicators to the countries with red flags. But with more clusters, there will be more countries classified in the fuzzy zones. The clustering stability would drop down because of the uncertainty of finding the right cluster for those countries.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

C. Biernacki. Initializing EM using the properties of its trajectories in Gaussian mixtures. *Statistics and Computing*, 14(3):267–279, 2004. doi: 10.1023/B:STCO.0000035306.77434.31. URL https://doi.org/10.1023/B:STCO.0000035306.77434.31.

C. Bouveyron. *funFEM: Clustering in the Discriminative Functional Subspace*, 2015. URL https://CRAN.R-project.org/package=funFEM. R package version 1.1.

C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the fisher discriminative subspace. *Statistics and Computing*, 22(1): 301–324, Apr 2011. ISSN 1573-1375. doi: 10.1007/s11222-011-9249-9. URL http://dx.doi.org/10.1007/s11222-011-9249-9.

C. Bouveyron and J. Jacques. Model-based Clustering of Time Series in Group-specific Functional Subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300, 2011. doi: 10.1007/s11634-011-0095-6. URL https://hal.archives-ouvertes.fr/hal-00559561.

C. Bouveyron, S. Girard, and C. SCHMID. High-dimensional data clustering. *Computational Statistics Data Analysis*, 52:502–, 09 2007. doi: 10.1016/j.csda.2007.02.009.

C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4), Dec 2015. ISSN 1932-6157. doi: 10.1214/15-aoas861. URL http://dx.doi.org/10.1214/15-AOAS861.

A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171 – 1193, 2010. doi: 10.1214/09-AOS741. URL https://doi.org/10.1214/09-AOS741.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.

1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

J. Jacques and C. Preda. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171, 2013. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2012.11.042. URL https://www.sciencedirect.com/science/article/pii/S0925231213002233. Advances in artificial neural networks, machine learning, and computational intelligence.

J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014. doi: 10.1007/s11634-013-0158-y. URL https://doi.org/10.1007/s11634-013-0158-y.

J. Peng and H.-G. Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2(3), Sep 2008. ISSN 1932-6157. doi: 10.1214/08-aoas172. URL http://dx.doi.org/10.1214/08-AOAS172.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer series in statistics. Springer, 1997. ISBN 9780387949567. URL https://books.google.ca/books?id=vYXCsgEACAAJ.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. doi: 10.1080/01621459.1971.10482356. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356.

L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, May 2010. URL https://ideas.repec.org/a/eee/csdana/v54y2010i5p1219-1233.html.

A. Schmutz and J. J. . C. Bouveyron. *funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces*, 2021. URL https://CRAN.R-project.org/package=funHDDC. R package version 2.3.1.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. URL http://jmlr.org/papers/v11/vinh10a.html.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, mar 1963. doi: 10.1080/01621459.1963.10500845. URL https://doi.org/10.1080%2F01621459.1963.10500845.