Analyses of daily COVID-19 cases across nations

Group11: Sibei Liu, Xue Jin, Yuchen Qi, Xinru Wang

05/01/2020

Introduction

COVID-19

Since it‘s first outbreak in January, the novel coronavirus (COVID-19) has been spreading rapidly through China and expanded to touch nearly every corner of the globe. Hundreds of thousands of people around the world have been sickened and over 200,000 have died. Efforts to contain the spread of the Covid-19 pandemic are now the top priority of governments. To make scientific decisions, such as quarantine, active monitoring, border controls, and lockdown, it is particularly crucial for policymakers to accurately predict how the spread of COVID-19 will change over time.

A logistic growth curve can be an effective way to capture the trajectory of cumulative cases of COVID-19. Characterized by an S-shaped curve, logistic growth model is approximately exponential at first, and growth rate accelerates as it approaches the midpoint of the curve but begins to decelerate as it approaches the model's upper bound, called the carrying capacity. In the COVID-19 case, The more people who have the virus, the more rapidly it spreads, and the growth will necessarily diminish when everybody is sick, which make the logistic model a good one to study the spread. In particular, this maximum limit would be the maximum number of cases a region can reach denoted by a. The t is the days since the first infection found. The b is the growth rate. And the c is the mid-point when the cumulative cases reach a/2.

$$f(t) = \frac{a}{1 + exp{-b(t-c)}}$$

Objectives

To help predict future spread of Covid-19 and to identify risk factors, our project aims to fit a logistic curve to the cumulative confirmed COVID-19 cases in each region of the world by developing an optimization algorithm and implement K-mean and Gaussian mixture model (with EM algorithm) to cluster these curves based on the fitted parameters.

Dataset

The dataset is a a subset of the open data, which contains the cumulative number of confirmed cases and death of COVID-19 from Jan 21 to March 24 from 163 countries/regions. Eight variables are recorded as following:

- Id: Record ID
- Province/State: The lcoal state/province of the record;
- Country/Region: The country/regiono of the record;

- Lat: Lattudiute of the record;
- Long: Longitude of the record;
- Date: Date of the record;
- ConfirmedCases: The number of confirmed case on that day;
- Fatalities: The number of death on that day;

We filter the countries that have confirmed cases greater 20 to fit the logistic curve. So in total only 116 coutries are used.

Statistical Methods

Adam Algorithm

Adam is A Method for Stochastic Optimization proposed in 2015 from Diederik P Kingma that only need the first-order gradient. The Stochastic Gradient descent (SGD) is often used when the objective function is typically non-convex (as in our case). The "Ada" is derived from "adaptive", meaning this method change the learning rate over time according to gradients before. The detailed proof and explanation can be found in Diederik's paper. Here we just extracted the fake code part from original paper to clarify.

Algorithm:

1. Reqiured: α: Stepsize
   $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
   $f(\boldsymbol{\theta})$: The objective function with parameter vetctor $\boldsymbol{\theta}$
   ϵ controls the converge
2. Required: $\boldsymbol{\theta}_0$: Initial guess of parameters
   $\mathbf{m}_0 \leftarrow \mathbf{0}$: Initialize the 1st moment vector as $\mathbf{0}$
   $\mathbf{v}_0 \leftarrow \mathbf{0}$: Initialize the 1st moment vector as $\mathbf{0}$
   $t \leftarrow 0$:Initialize time step =0
   while $\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} > \epsilon$ not converge, do
   $t \leftarrow t + 1$
   $\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$: Get gradients w.r.t objective function at timestep t
   $\boldsymbol{m}_t \leftarrow \beta_1 \cdot \boldsymbol{m}_{t-1} + (1 - \beta_1) \cdot \boldsymbol{g}_t$:Update biased first moment estimate
   $\boldsymbol{v}_t \leftarrow \beta_2 \cdot \boldsymbol{v}_{t-1} + (1 - \beta_2) \cdot \boldsymbol{g}_t^2$:Update biased second moment estimate
   $\hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t/(1 - \beta_1^t)$: Compute bias-corrected first moment estimate
   $\hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t/(1 - \beta_2^t)$: Compute bias-corrected second raw moment estimate
   $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \cdot \hat{\boldsymbol{m}}_t/(\sqrt{\hat{\boldsymbol{v}}_t + \epsilon}$: Update parameters
   End while Return $\boldsymbol{\theta}_t$ Result parameters
3. Defult setting:$\beta_1 = 0.9 \quad \beta_2 = 0.999 \quad \alpha = 0.001 \quad \epsilon = 10^{-8}$

Notes: g_t^t indicate the element-wise t power like $(g_t)^t$. Similarly, β_1^t and β_2^t also means the β_1 and β_2 to the power of t. In our case, we set the maximun time step $t = 10000$ to decrease the computation.

Loss function:

$$f = \sum_{i=1}^{n} (y_i - \frac{a}{1 + exp(-b(t - c))})^2$$

Gradient for parametrs a,b,c:

$$\nabla f(t, a) = \sum_{i=1}^{n} (\frac{2a}{(1 + e^{(-bt+bc)})^2} - \frac{2y}{1 + e^{(-bt+bc)}})$$

$$\nabla f(t,b) = -\sum_{i=1}^{n}(\frac{2a^2e^{(-bt+bc)}}{(1+e^{(-bt+bc)})^3} + \frac{2ae^{(-bt+bc)}(c-t)y}{(1+e^{(-bt+bc)})^2})$$

$$\nabla f(t,c) = -\sum_{i=1}^{n}(\frac{2a^2be^{(-bt+bc)}}{(1+e^{(-bt+bc)})^3} + \frac{2abe^{(-bt+bc)}}{(1+e^{(-bt+bc)})^2})$$

The initail guess of a,b,c in each country: a_0=two times the cumulative case in 24 March, b_0=0.3, c_0=40. For some special countries for example China and South Korea, the intial guess are adjusted for many times and the iteration also increases.

EM Algorithm

Cluster analysis is a method for finding clusters with similar characters within a dataset. And clustering methods can be divided into probability model-based approaches and nonparametric approaches[1]. The probability model-based approach contains Gussian Mixture Method, which assumes that the dataset follows a gussian mixture mixture distributions.

Given that $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \in \mathbb{R}^p$ be a collection of p dimensional data points. Assuming the following equation:

$$x_i \sim \begin{cases} N(\boldsymbol{\mu}_1, \Sigma_1), \text{with probability } p_1 \\ N(\boldsymbol{\mu}_2, \Sigma_2), \text{with probability } p_2 \\ \quad \vdots \quad\quad , \quad\quad \vdots \\ N(\boldsymbol{\mu}_k, \Sigma_k), \text{with probability } p_k \end{cases}$$

$\sum_{j=1}^{k} p_j = 1$

Let $\mathbf{r}_i = (r_{i,1}, ..., r_{i,k}) \in \mathbb{R}^k$ as the cluster indicator of $\mathbf{x}_i$, which takes form $(0, 0, ..., 0, 1, 0, 0)$ with $r_{i,j} = I\{\mathbf{x}_i \text{ belongs to cluster } j\}$. The cluster indicator $\mathbf{r}_i$ is a latent variable that cannot be observed. What is complete likelihood of $(\mathbf{x}_i, \mathbf{r}_i)$.

The distribution of $\mathbf{r}_i$ is

$$f(\mathbf{r}_i) = \prod_{j=1}^{k} p_j^{r_i,j}$$

The complete log-likelihood is

$$\ell(\theta; \mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{i,j}[\log p_i + \log f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)] = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{i,j}[\log p_i - 1/2\log|\Sigma| - 1/2(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top\Sigma(\mathbf{x}_i - \boldsymbol{\mu}_j)]$$

E-step Evaluate the responsibilities using the current parameter values

$$\gamma_{i,k}^{(t)} = P(r_{i,k} = 1|\mathbf{x}_i, \theta^{(t)}) = \frac{p_k^{(t)} f(\mathbf{x}_i|\boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} f(\mathbf{x}_i|\boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}$$

M-step

$\theta^{(t+1)} = \arg\max \ell(\mathbf{x}, \gamma^{(t)}, \theta)$.

Let $n_k = \sum_{i=1}^{n} \gamma_{i,k}$, we have

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{n_k}\sum_{i=1}^{n}\gamma_{i,k}\mathbf{x}_i$$

$$\Sigma_k^{(t+1)} = \frac{1}{n_k}\sum_{i=1}^{n}\gamma_{i,k}(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T$$

$$p_k^{(t+1)} = \frac{n_k}{n}$$

K-mean

The K-means algorithm partitions data into k clusters (k is predetermined). We denote $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k\}$ as the centers of the k (unknown) clusters, and denote $\mathbf{r}_i = (r_{i,1}, ..., r_{i,k}) \in \mathbb{R}^k$ as the "hard" cluster assignment of $\mathbf{x}_i$.

k-means finds cluster centers and cluster assignments that minimize the objective function

$$J(\mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{i,j}\|\mathbf{x}_i - \mu_k\|^2$$

K-means is a special case for Gussian Mixture. It is not required to consider small variances or the limit case of zero variances.

Method to select number of clusters

1. The Elbow Method

   Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish.

2. The Silhouette Method

   The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

3. Gap Statistic Method

   The idea of the Gap Statistic is to compare the within-cluster dispersion to its expectation under an appropriate null reference distribution.

Dunn Index

The Dunn index (DI) is a metric for evaluating clustering algorithms. It is an internal evaluation scheme, where the result is based on the clustered data itself. It aims to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. For a given assignment of clusters, a higher Dunn index indicates better clustering.

Result

Task 1

Task 1.1

After applying the Adam algorithm in 116 countries, we get the estimated a,b,c values for each country in **Table 1**. The maximum a value is 138340 from Italy.The b value ranges from 0.085(Singapore) to 3.857(Trinidad and Tobago). The c value changes from 70 (China, Taiwan) to 4(Uzbekistan).

country_region	a_value	b_value	c_value
Afghanistan	342	0.202	37
Albania	269	0.173	17
Algeria	723	0.258	30
Andorra	345	0.344	22
Argentina	970	0.315	23
Armenia	514	0.288	23
Australia	4072	0.293	58
Austria	10760	0.275	28
Azerbaijan	365	0.184	30
Bahrain	795	0.118	29
Bangladesh	99	0.244	18
Belarus	102	0.276	19
Belgium	8530	0.254	49
Bolivia	81	0.192	16
Bosnia and Herzegovina	352	0.292	19
Brazil	4507	0.380	27
Brunei	98	0.381	7
Bulgaria	459	0.253	16
Burkina Faso	252	0.363	14
Cambodia	168	0.317	56
Canada	5462	0.338	58
Chile	1862	0.318	21
Netherlands	11170	0.239	26
New Zealand	505	0.420	27
Nigeria	102	0.407	25
North Macedonia	309	0.325	27
Norway	5557	0.175	26
Oman	361	0.125	40
Pakistan	1774	0.326	26
Panama	715	0.321	14
Paraguay	74	0.195	19
Peru	678	0.322	16
Philippines	1091	0.240	54
Poland	1821	0.283	20
Portugal	4741	0.335	22
Qatar	889	0.175	19
Romania	1783	0.256	29
Russia	979	0.291	53
Rwanda	107	0.356	11
San Marino	230	0.191	19
Saudi Arabia	1551	0.288	23
Senegal	357	0.217	27
Serbia	627	0.286	18
Singapore	1262	0.085	67
Slovakia	254	0.332	13
Slovenia	805	0.200	16
South Africa	1303	0.343	20
Spain	79759	0.257	52
Sri Lanka	105	0.459	51
Sweden	4381	0.171	52
Switzerland	19766	0.261	28
Taiwan*	576	0.097	70
Thailand	1634	0.306	62
Trinidad and Tobago	53	3.857	6
Tunisia	419	0.242	24
Turkey	3770	0.537	13
Ukraine	212	0.395	21
United Arab Emirates	652	0.114	62
United Kingdom	16258	0.279	53
Uruguay	184	0.548	6
US	106991	0.389	29
Uzbekistan	50	0.729	4
Venezuela	95	0.426	5
Vietnam	418	0.102	69

country_region	a_value	b_value	c_value
China	78732	0.223	18
Colombia	777	0.335	18
Congo (Kinshasa)	115	0.360	14
Costa Rica	375	0.268	18
Cote d'Ivoire	342	0.857	15
Croatia	958	0.310	29
Cuba	122	0.363	13
Cyprus	272	0.234	15
Denmark	3258	0.170	24
Dominican Republic	640	0.498	23
Ecuador	2180	0.449	23
Egypt	806	0.193	39
Estonia	569	0.235	22
Finland	1570	0.216	55
France	39932	0.148	64
Georgia	151	0.140	29
Germany	65957	0.259	57
Ghana	300	0.332	15
Greece	1499	0.182	27
Guatemala	23	0.589	6
Honduras	32	0.549	8
Hungary	393	0.266	20
Iceland	1311	0.213	25
India	1060	0.253	54
Indonesia	1389	0.266	22
Iran	49441	0.131	33
Iraq	642	0.143	30
Ireland	2673	0.309	24
Israel	4055	0.304	33
Italy	138340	0.183	53
Jamaica	20	0.331	5
Japan	2195	0.094	60
Jordan	326	0.302	21
Kazakhstan	69	0.529	5
Kenya	237	0.320	18
Korea, South	8801	0.284	40
Kuwait	564	0.088	36
Kyrgyzstan	279	0.546	9
Latvia	411	0.270	22
Lebanon	829	0.169	35
Liechtenstein	55	0.500	15
Lithuania	432	0.451	25
Luxembourg	2213	0.354	24
Malaysia	3231	0.222	59
Malta	242	0.248	17
Martinique	135	0.251	18
Mauritius	115	0.492	7
Mexico	748	0.317	25
Moldova	273	0.285	16
Monaco	60	0.272	25
Montenegro	124	0.507	8
Morocco	357	0.291	22

Table 1. Estimated a,b,c values in each country

Untill 24 March, It is estimaed that there are 27 countries that pass the midpoint. They are : Belarus, Brunei, Cambodia, China, Denmark,Estonia, Guatemala, Honduras , Iran, Jamaica, Japan, Kazakhstan, Korea South, Liechtenstein, Norway, Pakistan, Peru, Qatar, San Marino, Slovakia, Slovenia, Sri Lanka, Sweden,Trinidad and Tobago, Uruguay,Uzbekistan, Venezuela.

If we define the cumulative cases at 24 March surpass the 80% of a value in corresponding country is "appraoching the end". Then there are 15 countries: Brunei, China, Guatemala, Honduras, Jamaica, Kazakhstan,Korea South, Liechtenstein, San Marino, Slovakia, Sri Lanka, Trinidad and Tobago, Uruguay, Uzbekistan, Venezuela.

Task 1.2

We select three kinds of countries to do the visualization: 1) In the very beginning stages of COVID-19 outbreak. Representatives: Afghanistan and Vietnam. 2) During the Outbreak stage. Representatives: UK and US. 3)Late stage of outbreak, which may produce a complete logistic curve. Representatives: China and South Korea. The a,b,c values of above 6 example countries are as follow:

country_region	a_value	b_value	c_value
Afghanistan	342	0.202	37
China	78732	0.223	18
Korea, South	8801	0.284	40
United Kingdom	16258	0.279	53
US	106991	0.389	29
Vietnam	418	0.102	69

Table 2. Estimated a,b,c values in 6 countries

The data from 25 March to 5 April (11 days) is used as test data to examine the predictivity of fitted model. The MSEs of training data(data before 24 March) and test data are as follow. Because the original data itself is relatively large, so the calculated MSE seems to be large.

Country	Train_error
Afghanistan	2.080206e+01
China	4.077602e+06
Korea_South	4.471121e+04
United_Kingdom	9.472004e+03
US	1.871744e+05
Vietnam	5.664849e+01

Table 3. MSE of train data

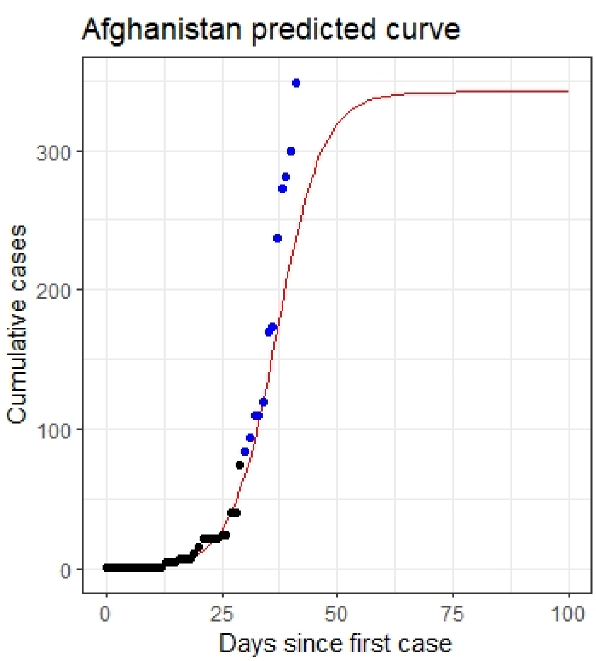
Country	test_error
Afghanistan	3.200053e+03
China	1.211702e+07
Korea_South	9.565317e+05
United_Kingdom	2.690240e+08
US	1.428445e+10
Vietnam	8.978671e+01

Table 4. MSE of test data

But if we visualize the model fitted value(red line) and observed values(train data is black and test data is blue). In the following plot, the fitted logistic curve fits the train data well, but deviations from test data in those two countries are different. The Afghanistan and Vietnam are both at the initial outbreak, so a dramatic increase of cases can be expected.

The maximum cases(a=342) is expected to be reached around the 60th day in Afghanistan. The deviation of test data before around 1 April is smaller than that after 1 April. But the data in April 5, apparently exceeds the estimaed a value, which denote the bias of our fitted model since we built the model only based on the data before 24 March.

For Vietnam, the The maximum cases(a=418) is expected to be reached around the 120th day. The fitteness of both train and test data is good.



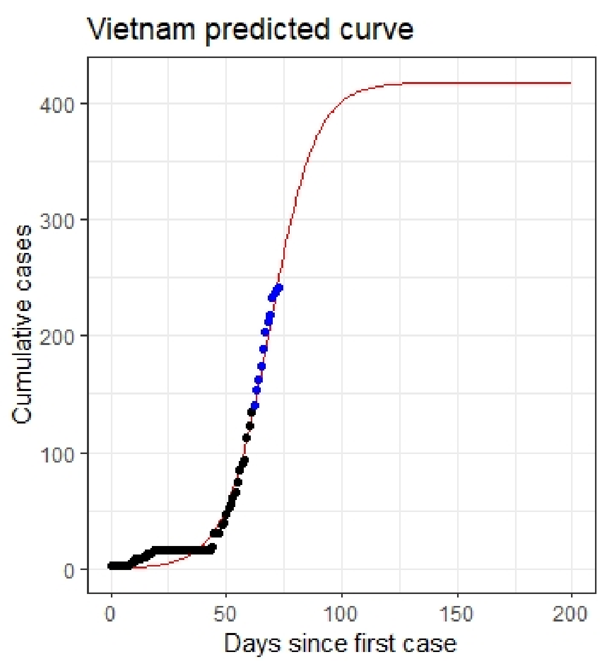


Figure 1. Afghanistan and Vietnam fitted and predicted values

In second kind country is as follow. The estimated a values are 16258 and 106991 for UK and US respectively.And the estimated stable stage when a is reach is 70th day and 50th day for UK and US respectively. For both of them, the red line fits black train data very well. But the increase of cases after 25 March is soaring, which is far away from the fitted line. To some extend, the **Figure 2** denotes the lack of predictivity beause the lack of data when we built the model.

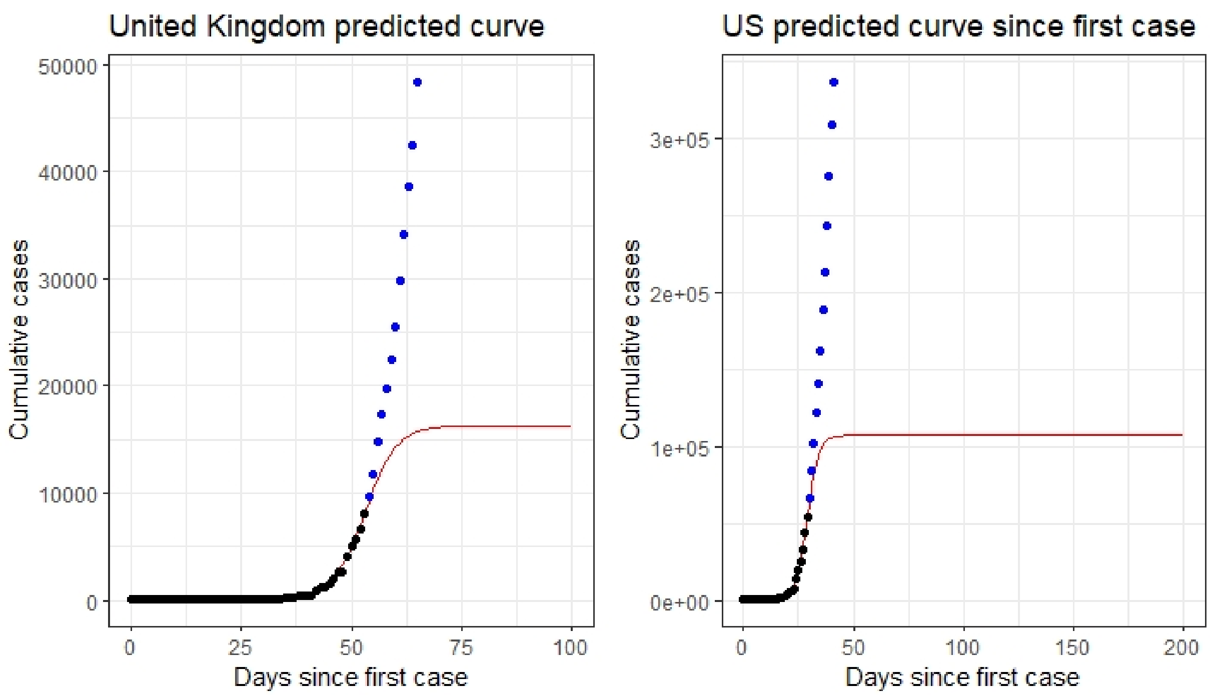


Figure 2. UK and US fitted and predicted values

In third kind country, who breakout reported at early Jan, their growths are very similar to each other. The problem of lack of predictivity re-appears that it estimates both of them already reached the end of spreading. But in fact both of them have increase cases after March 25. But the increase of cases is much slighter than UK and US. And the increase in China after 25 March is more flat given 1) it may already enters the stable part, which means the increase slows and 2) the interventions China takes may play an important role.

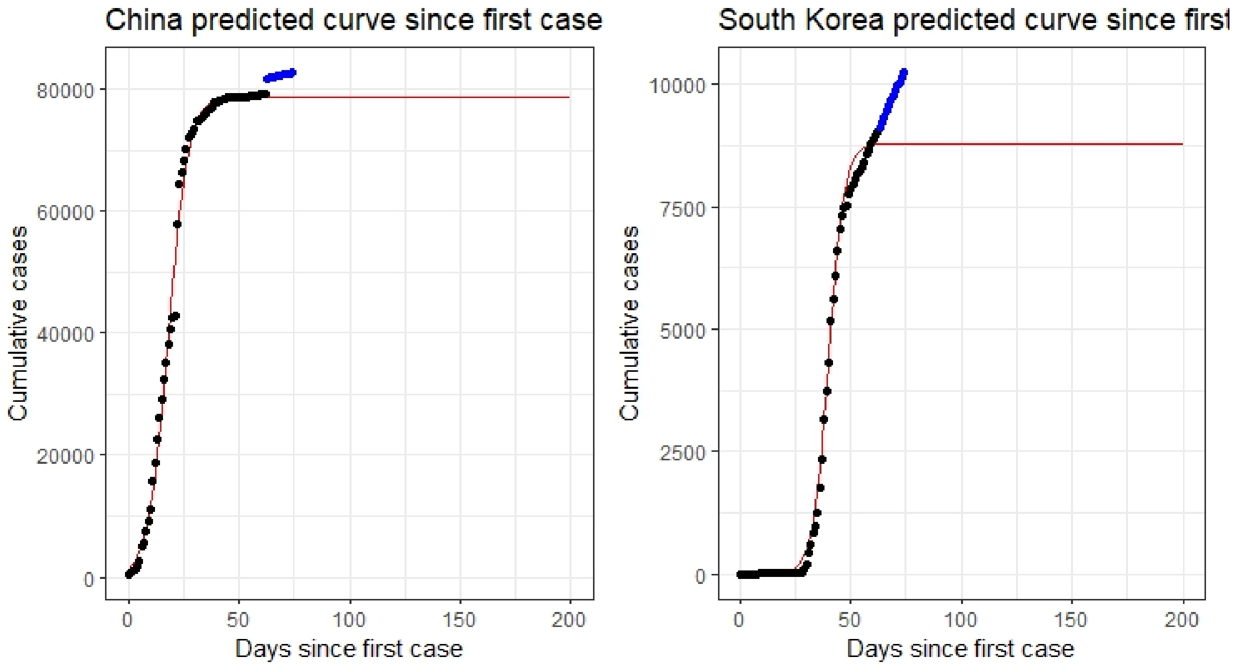


Figure 3. China and South Korea fitted and predicted values

Task 2

In order to choose the best clustering number, we use three different methods: The Elbow Method, The Silhouette Method and Gap Statistic Method. From the results (**Supplementay Fig. 1,2,3**), we finally choose three as our clustering number, given that when clustering number is five, there will be NA in GMM method.

method	a_value	b_value	c_value
GMM	355.1401	0.3757532	18.98756
GMM	16669.0556	0.2525830	41.13602
Kmeans	2472.2883	0.3293063	27.26126
Kmeans	109867.4000	0.2252000	47.80000

Table 5: Centering points of GMM and Kmeans

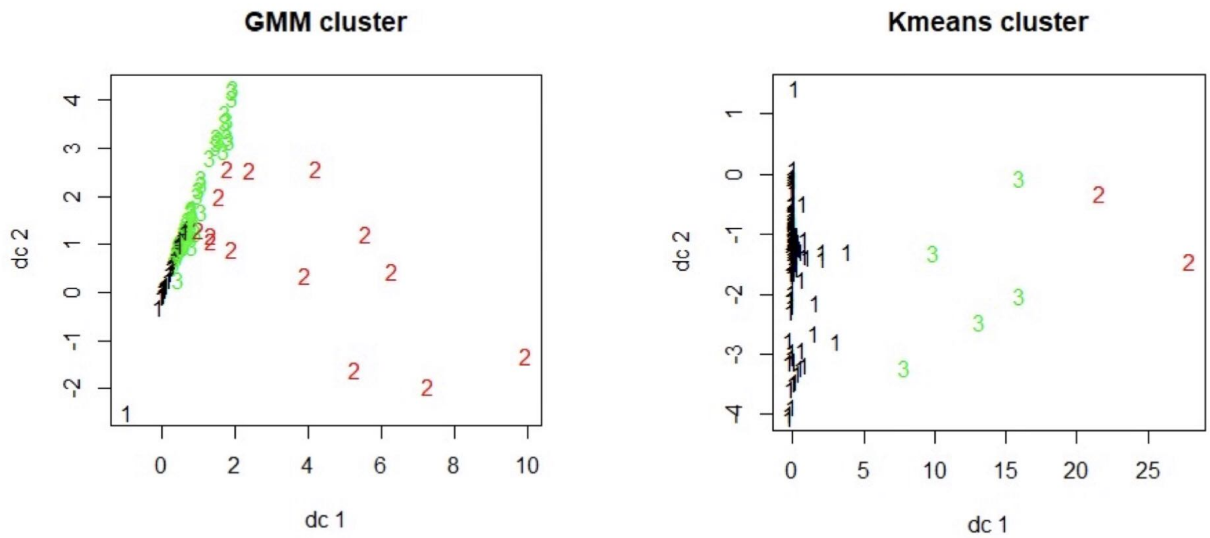
country	GMM_class	kmeans_class	a_value	b_value	c_value
Sri Lanka	3	1	105	0.459	51
Sweden	3	1	4381	0.171	52
Switzerland	2	1	19766	0.261	28
Taiwan*	3	1	576	0.097	70
Thailand	3	1	1634	0.306	62
Trinidad and Tobago	1	1	53	3.857	6
Tunisia	1	1	419	0.242	24
Turkey	3	1	3770	0.537	13
Ukraine	1	1	212	0.395	21
United Arab Emirates	3	1	652	0.114	62
United Kingdom	2	1	16258	0.279	53
Uruguay	1	1	184	0.548	6
US	2	2	106991	0.389	29
Uzbekistan	1	1	50	0.729	4
Venezuela	1	1	95	0.426	5
Vietnam	3	1	418	0.102	69

country	GMM_class	kmeans_class	a_value	b_value	c_value
Afghanistan	3	1	342.00	0.20200	37.00000
Albania	1	1	269.00	0.17300	17.00000
Algeria	3	1	723.00	0.25800	30.00000
Andorra	1	1	345.00	0.34400	22.00000
Argentina	3	1	970.00	0.31500	23.00000
Armenia	3	1	514.00	0.28800	23.00000
Australia	3	1	4072.00	0.29300	58.00000
Austria	2	1	10760.00	0.27500	28.00000
Azerbaijan	3	1	365.00	0.18400	30.00000
Bahrain	3	1	795.00	0.11800	29.00000
Bangladesh	3	1	99.00	0.24400	18.00000
Belarus	3	1	102.00	0.27600	19.00000
Belgium	2	1	8530.00	0.25400	49.00000
Bolivia	1	1	81.00	0.19200	16.00000
Bosnia and Herzegovina	1	1	352.00	0.29200	19.00000
Brazil	3	1	4507.00	0.38000	27.00000
Brunei	1	1	98.00	0.38100	7.00000
Bulgaria	3	1	459.00	0.25300	16.00000
Burkina Faso	1	1	252.00	0.36300	14.00000
Cambodia	3	1	168.00	0.31700	56.00000
Canada	3	1	5462.00	0.33800	58.00000
Chile	3	1	1862.00	0.31800	21.00000
China	2	3	78732.05	0.22511	17.91412
Colombia	3	1	777.00	0.33500	18.00000
Congo (Kinshasa)	1	1	115.00	0.36000	14.00000
Costa Rica	1	1	375.00	0.26800	18.00000
Cote d'Ivoire	1	1	342.00	0.85700	15.00000
Croatia	3	1	958.00	0.31000	29.00000
Cuba	1	1	122.00	0.36300	13.00000
Cyprus	1	1	272.00	0.23400	15.00000
Denmark	3	1	3258.00	0.17000	24.00000
Dominican Republic	3	1	640.00	0.49800	23.00000
Ecuador	3	1	2180.00	0.44900	23.00000
Egypt	3	1	806.00	0.19300	39.00000
Estonia	3	1	569.00	0.23500	22.00000
Finland	3	1	1570.00	0.21600	55.00000
France	2	3	39932.00	0.14800	64.00000
Georgia	3	1	151.00	0.14000	29.00000
Germany	2	3	65957.00	0.25900	57.00000
Ghana	1	1	300.00	0.33200	15.00000
Greece	3	1	1499.00	0.18200	27.00000
Guatemala	1	1	23.00	0.58900	6.00000
Honduras	1	1	32.00	0.54900	8.00000
Hungary	1	1	393.00	0.26600	20.00000
Iceland	3	1	1311.00	0.21300	25.00000
India	3	1	1060.00	0.25300	54.00000
Indonesia	3	1	1389.00	0.26600	22.00000
Iran	2	3	49441.00	0.13100	33.00000
Iraq	3	1	642.00	0.14300	30.00000
Ireland	3	1	2673.00	0.30900	24.00000

country	GMM_class	kmeans_class	a_value	b_value	c_value
Israel	3	1	4055.000	0.3040000	33.00000
Italy	2	2	138340.000	0.1830000	53.00000
Jamaica	1	1	20.000	0.3310000	5.00000
Japan	3	1	2195.000	0.0940000	60.00000
Jordan	1	1	326.000	0.3020000	21.00000
Kazakhstan	1	1	69.000	0.5290000	5.00000
Kenya	1	1	237.000	0.3200000	18.00000
Korea, South	2	1	8801.392	0.2836325	40.38837
Kuwait	3	1	564.000	0.0880000	36.00000
Kyrgyzstan	1	1	279.000	0.5460000	9.00000
Latvia	1	1	411.000	0.2700000	22.00000
Lebanon	3	1	829.000	0.1690000	35.00000
Liechtenstein	1	1	55.000	0.5000000	15.00000
Lithuania	1	1	432.000	0.4510000	25.00000
Luxembourg	3	1	2213.000	0.3540000	24.00000
Malaysia	3	1	3231.000	0.2220000	59.00000
Malta	1	1	242.000	0.2480000	17.00000
Martinique	1	1	135.000	0.2510000	18.00000
Mauritius	1	1	115.000	0.4920000	7.00000
Mexico	3	1	748.000	0.3170000	25.00000
Moldova	1	1	273.000	0.2850000	16.00000
Monaco	3	1	60.000	0.2720000	25.00000
Montenegro	1	1	124.000	0.5070000	8.00000
Morocco	1	1	357.000	0.2910000	22.00000
Netherlands	2	1	11170.000	0.2390000	26.00000
New Zealand	1	1	505.000	0.4200000	27.00000
Nigeria	3	1	102.000	0.4070000	25.00000
North Macedonia	3	1	309.000	0.3250000	27.00000
Norway	2	1	5557.000	0.1750000	26.00000
Oman	3	1	361.000	0.1250000	40.00000
Pakistan	3	1	1774.000	0.3260000	26.00000
Panama	3	1	715.000	0.3210000	14.00000
Paraguay	3	1	74.000	0.1950000	19.00000
Peru	3	1	678.000	0.3220000	16.00000
Philippines	3	1	1091.000	0.2400000	54.00000
Poland	3	1	1821.000	0.2830000	20.00000
Portugal	3	1	4741.000	0.3350000	22.00000
Qatar	3	1	889.000	0.1750000	19.00000
Romania	3	1	1783.000	0.2560000	29.00000
Russia	3	1	979.000	0.2910000	53.00000
Rwanda	1	1	107.000	0.3560000	11.00000
San Marino	1	1	230.000	0.1910000	19.00000
Saudi Arabia	3	1	1551.000	0.2880000	23.00000
Senegal	3	1	357.000	0.2170000	27.00000
Serbia	3	1	627.000	0.2860000	18.00000
Singapore	3	1	1262.000	0.0850000	67.00000
Slovakia	1	1	254.000	0.3320000	13.00000
Slovenia	3	1	805.000	0.2000000	16.00000
South Africa	3	1	1303.000	0.3430000	20.00000
Spain	2	3	79759.000	0.2570000	52.00000

Table 6: Cluster result of each country

The centering points of GMM and Kmeans method is shown in (**Table. 5**), and classification result of each country using these two method is shown in (**Table. 6**) and (**Fig. 4**). And the geographical distribution of countries in these classes using these two method can be seen in (**Fig. 5**), in which blue points are countries in class one, red points are countries in class two and yelloe points are countries in class three.



Figue 4: Visualized Cluster result of each country

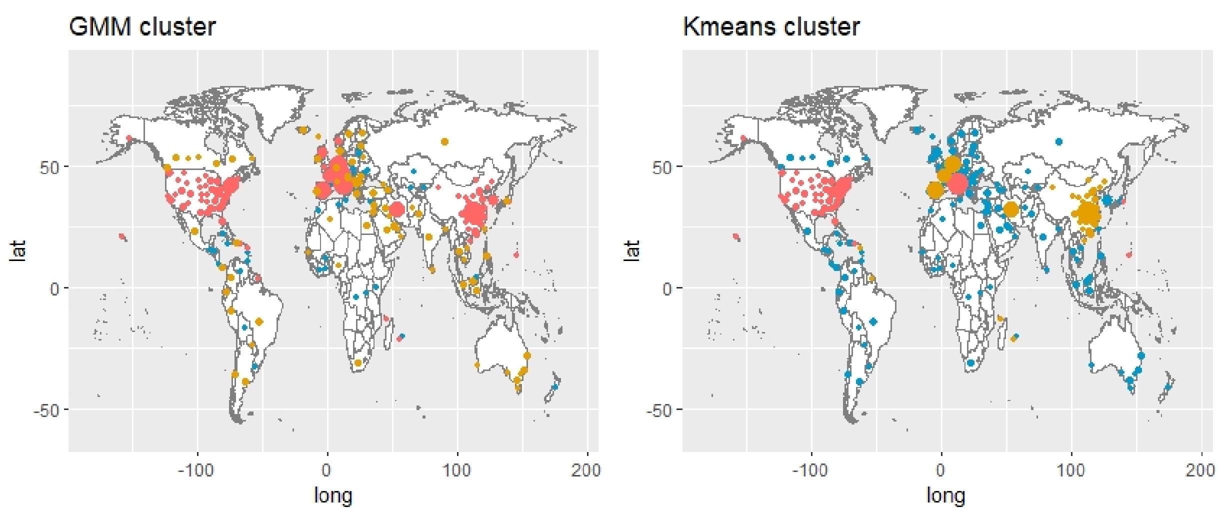


Figure 5: Clusters in map

To compare GMM and Kmeans method, we used Dunn Index method. From (**Table. 7**), we can see that the Dunn Index of Kmeans is higher than that of GMM. The reason may be that our data don't follow Gussian distribution. So we choose Kmeans to cluster our character value of each country. From (**Fig.5**) and (**Table. 6**), we can see that Italy and US fall into class two, and China, France, Germany, etc fall into class three. The reason may be that Italy and US have higher growth rate and larger maximum cases value according to the given dataset. There is two types of countries in class three: one is that they have already arrived maximum point and their start time is relatively earlier than other countries, such as China and South Korea, another is that they are still in early stage and still lack of detection of covid-19, so their data may not be accurate and will increase quickly later due to more and more test, such as Spain and France.

method	Dunn_index
Kmeans	0.1227
GMM	0.0013

Table 7: Dunn Index

Discussion

Task 1:

For most regions, the logistic curve is a reasonable model for fitting the cumulative cases, capturing the growth rate trend. However, when it comes to predicting future new cases, the logistic growth model has limitations, especially for this dataset. For example, China and Korea are predicted to have reached the upper bound, but the predicted number of cases after March 23 exceeds the estimated maximum. For China, one possible explanation is that the imported cases of novel coronavirus pneumonia account for this increasing trend, but our model fails to include the fluence outside a certain region, assuming each region is independent. As for Korea, a potential second wave of infection may be the result of "returning to normal life" and some citizens' ignoring social distancing. An alternative explanation is that the decreasing trend of growth rate based on the training data is attributed to Korea's rapidly responding to and mitigating the spread of this epidemic, but the maximum has not yet been reached. Piecewise functions may be suitable for such cases. In Afghanistan, UK and US, the growth rate after March 23 is much larger than the predicted one, which may be explained by the absence and inefficiency of intervention strategies. Generally, we cannot add the effect of factors such as public health interventions, newly developed treatments and vaccinations, and other regions' conditions outside a specific region, to this modeling process. Another factor needs to mention is that the data itself may not be accurate, that is the number of cases reported for a certain date may be smaller than the truth, as some cases may have not been tested or they may be tested falsely negative.

Although the limitations of a logistic curve may account for the discrepancy between the fitted curve and the recorded number for some regions, it is still useful for prediction when the date is not far away from the latest date in the training data in most circumstances based on our test result. When the recorded number of cases significantly exceeds the prediction, it may be necessary to consider whether social factors such as improper interventions exist, and use this to guide future strategies for controlling this disease.

Several optimization algorithms were implemented when fitting the curve. The Newton–Raphson method was considered for its fast convergence rate. As it was not easy to calculate the hessian matrix of the RSS for the original form of the parametric function, we transformed y and a to be the inverse. However, the starting values for this algorithm have a significant impact on the result, as the Newton method tends to find the local minimizer instead of the global one, which is especially a severe problem for non-linear least squares regression. To reduce the effect of this limitation, we considered adding the momentum when doing iteration. The final algorithm chosen is Adam, as it adopts an adaptive per-coordinate learning rate selection method and dampens oscillations. Adam may lead to the optimal solution, however it needs a large number of iterations which exceeds our computer capacity, thus our fitting may not minimize the RSS, resulting in an inaccurate prediction.

Task 2

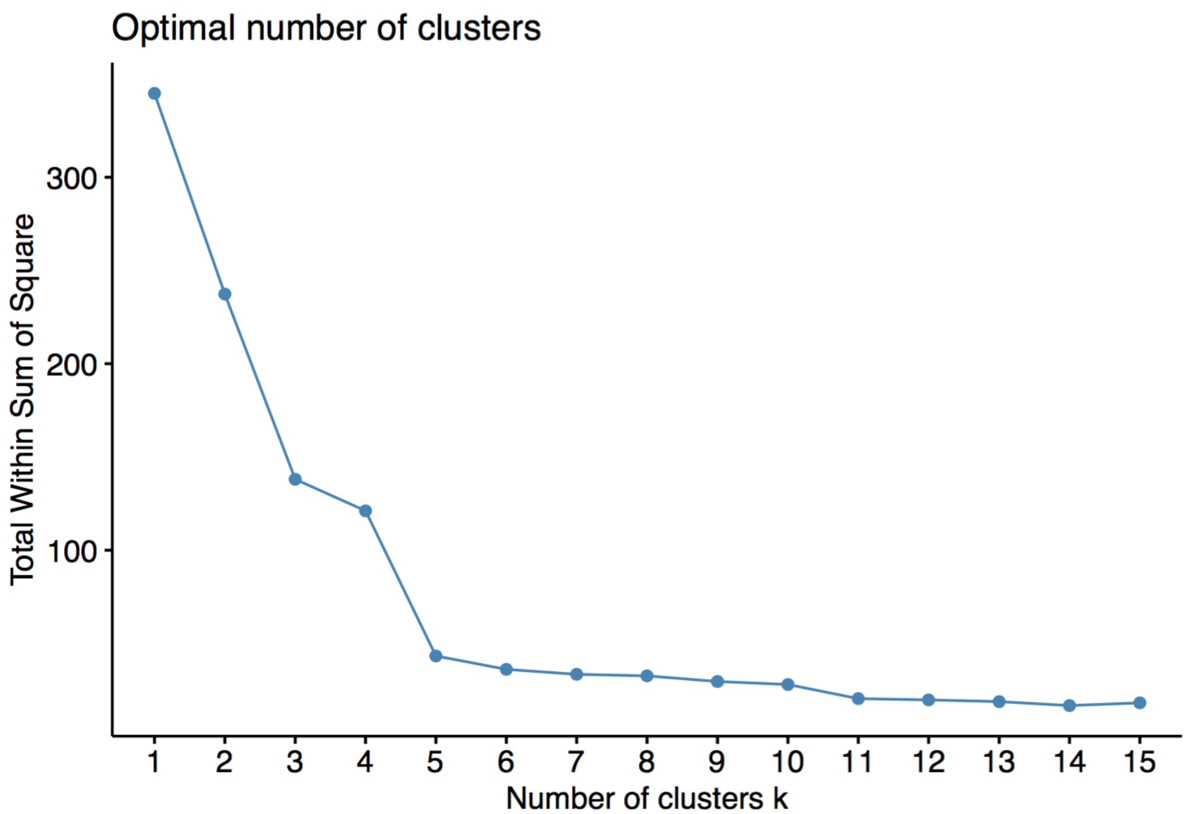
According to Kmeans CLassification, we have three clusters in these countries with different maximum cases, growth rate and mid-point. But due to the inaccurate data in early stage of some countries, we may get inaccurate estimate of a, b, c value, which leads to wrong classification of some countries, such as Spain and France. And Kmeans clustering also has some disadvantages, one of them is that this method assumes the clusters as spherical, so does not work efficiently with complex geometrical shaped data.

References

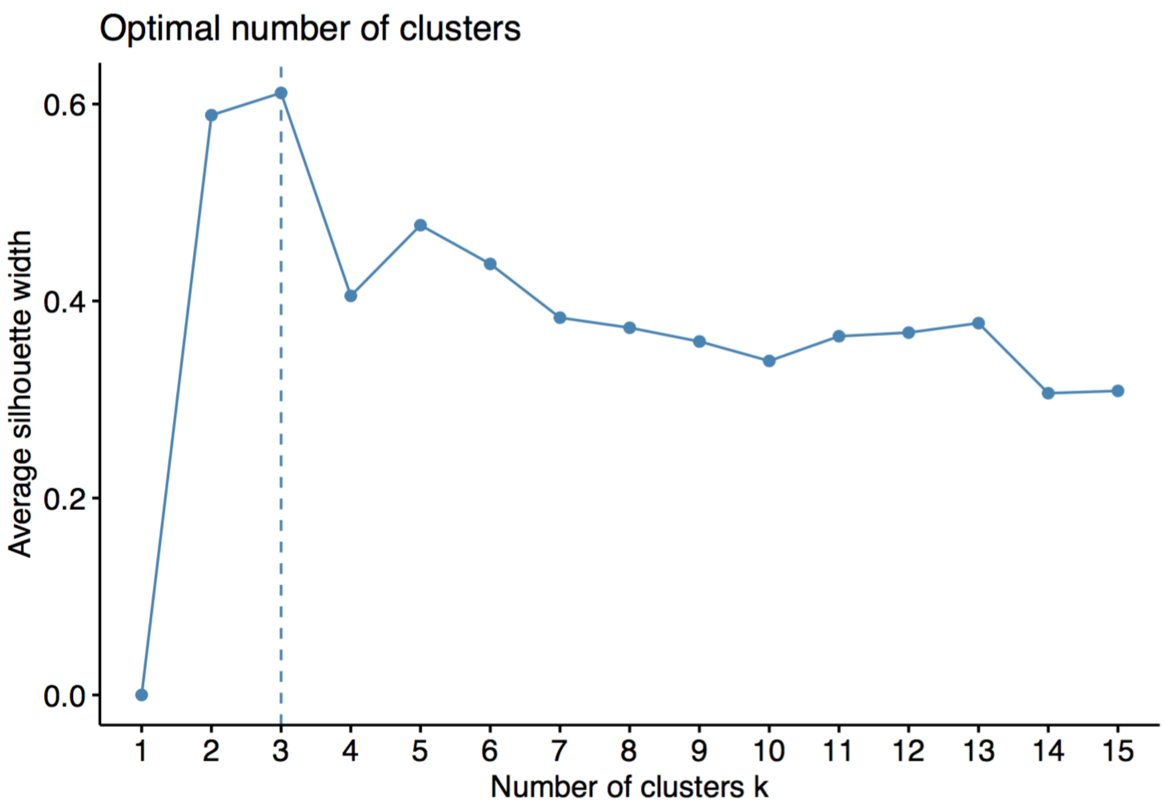
1 Miin-ShenYang, Chien-YoLai, Chih-YingLin. "A robust EM clustering algorithm for Gaussian mixture models." Pattern Recognition (2012).

2 Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. 2014. arXiv:1412.6980v9

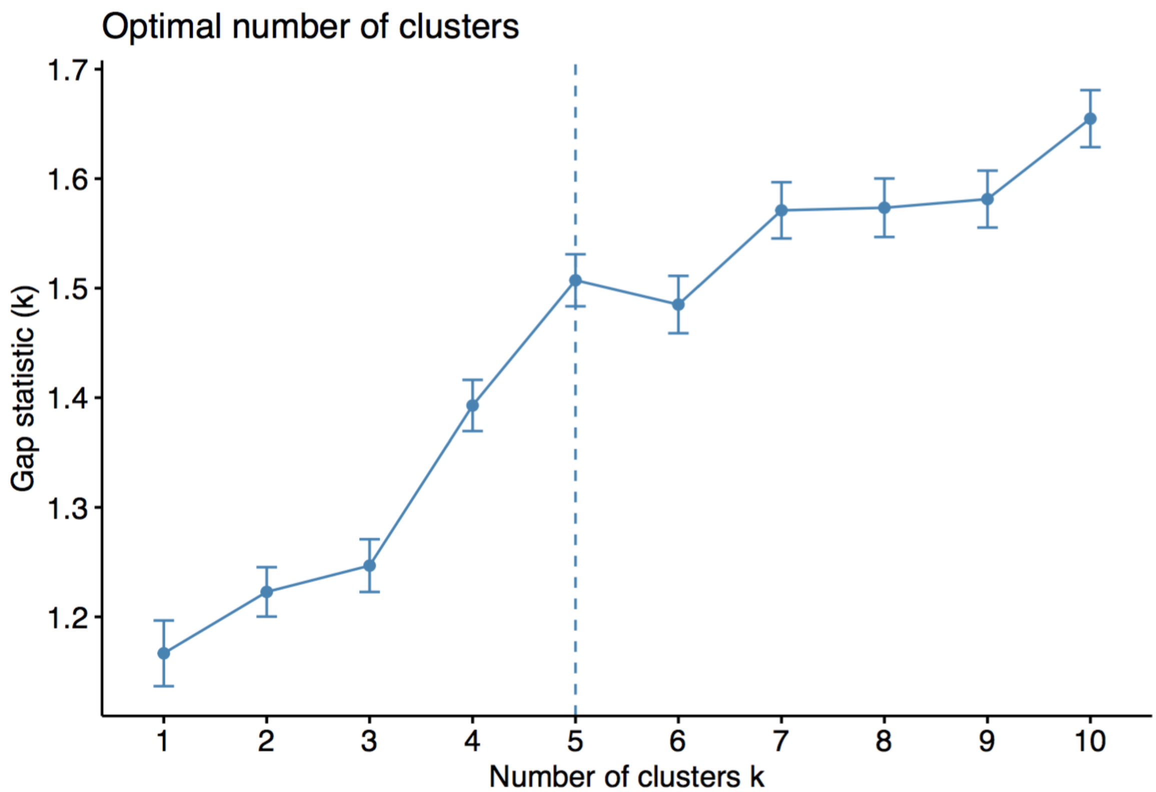
Supplementary



Supplementary Fig 1. WSS



Supplementary Fig 2. Silhouette Method



Supplementary Fig 3. Gap