Analyses of daily COVID-19 cases across nations

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

05/01/2020

Objective

Statistical Methods

Adam Algorithm

Guassian mixture model (with 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:

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

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

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 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 (**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.

The centering points of GMM and Kmeans method is shown in (**Table. 1**), and classification result of each country using these two method is shown in (**Table. 2**) and (**Fig. 4,5**). And the geographical distribution of countries in these classes using these two method can be seen in (**Fig. 6**) and (**Fig. 7**), in which blue points are countries in class one, red points are countries in class two and yelloe points are countries in class three.

To compare GMM and Kmeans method, we used Dunn Index method. From (**Table. 3**), 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. 7**) and (**Table. 2**), 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.

Discussion

Task 1:

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.

Figures

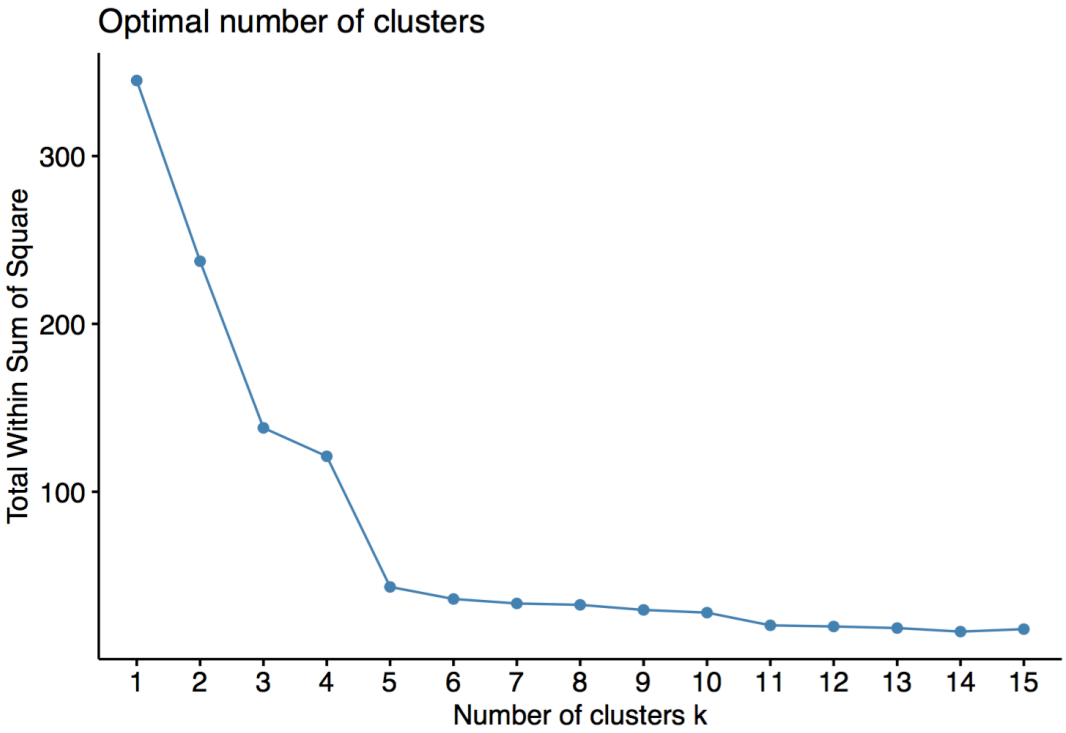

Figure 1: WSS results

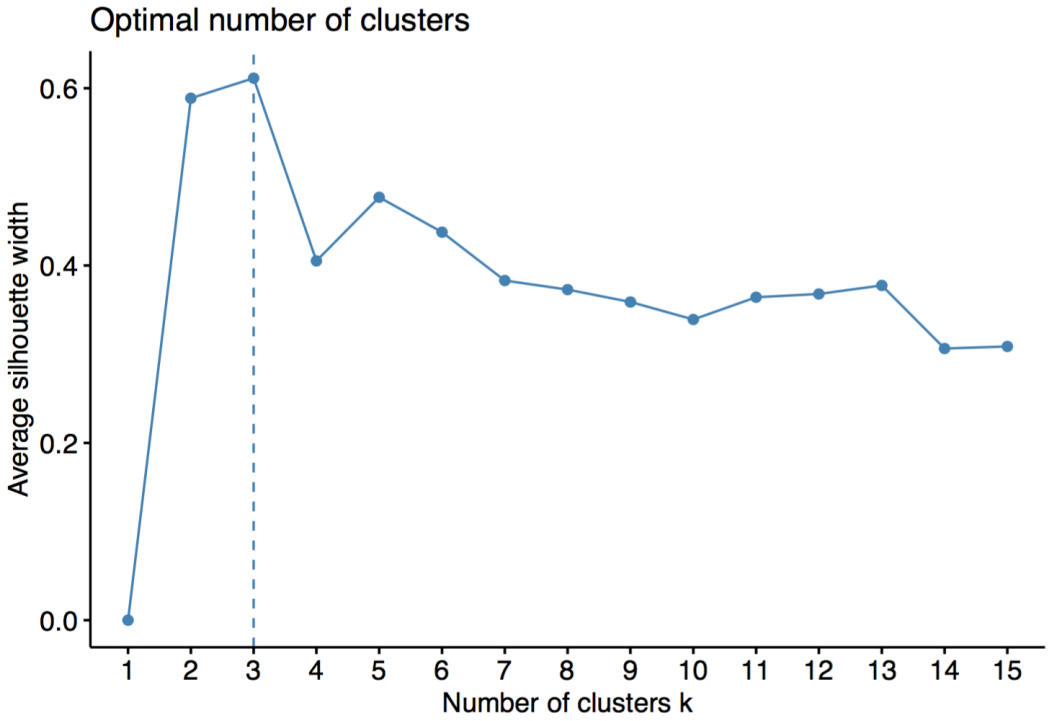

Figure 2: Result of Silhouette Method

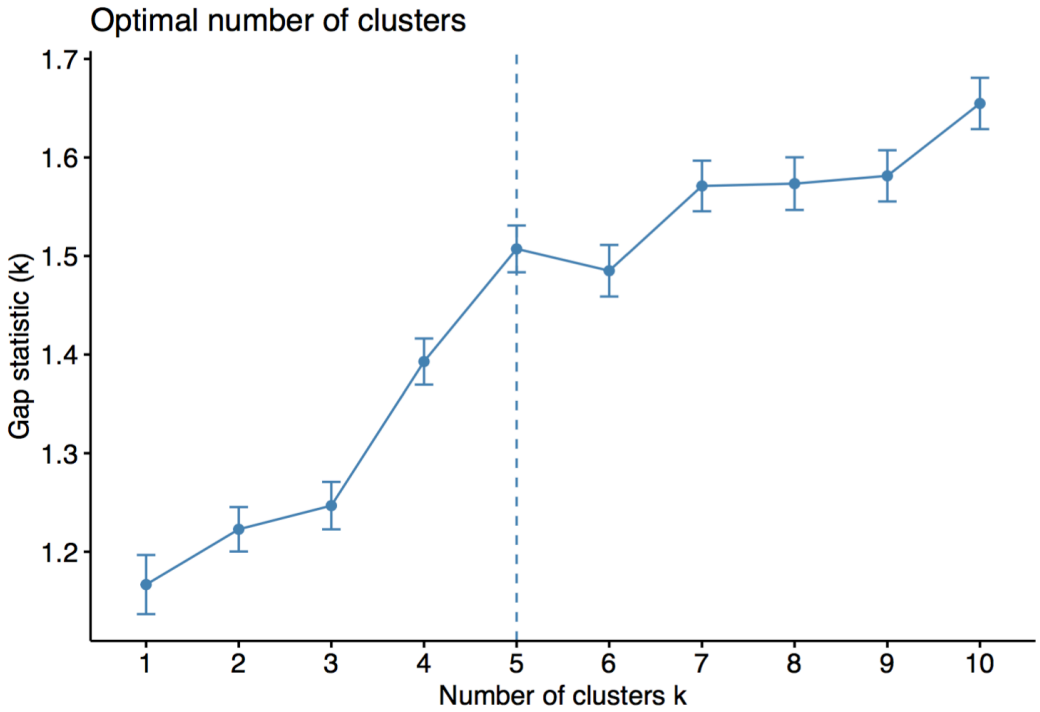

Figure 3: Result of Gap Statistic Method

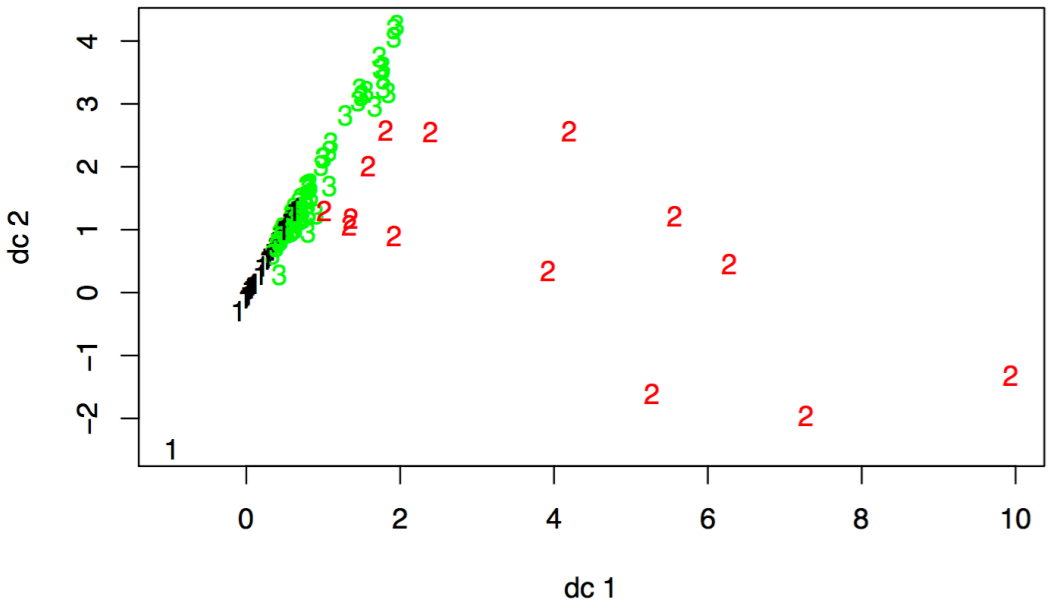

Figure 4: Classification of GMM

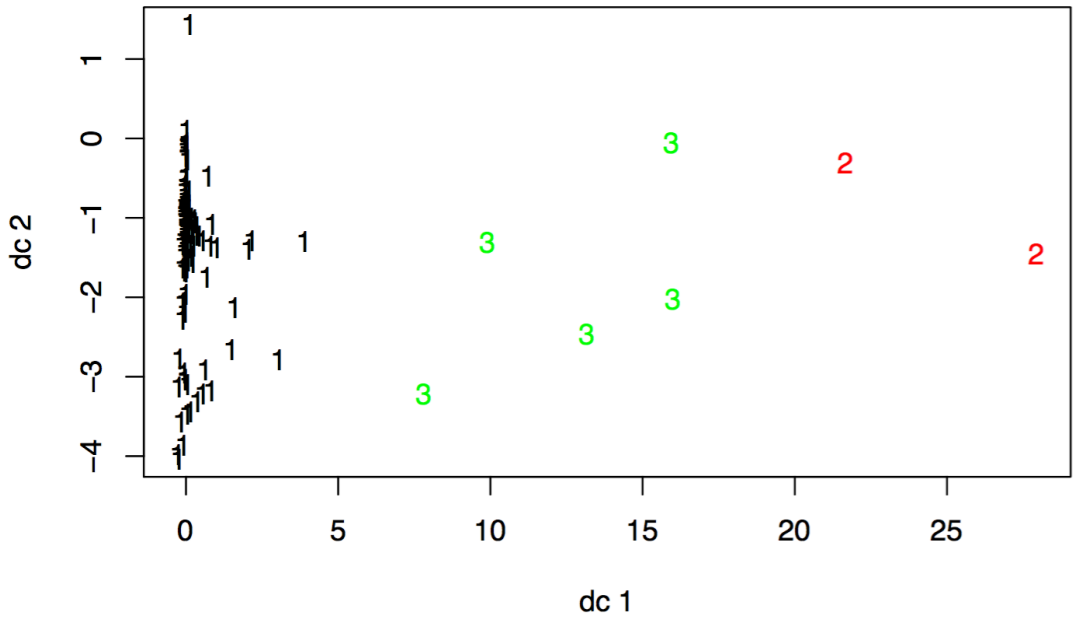

Figure 5: Classification of Kmeans

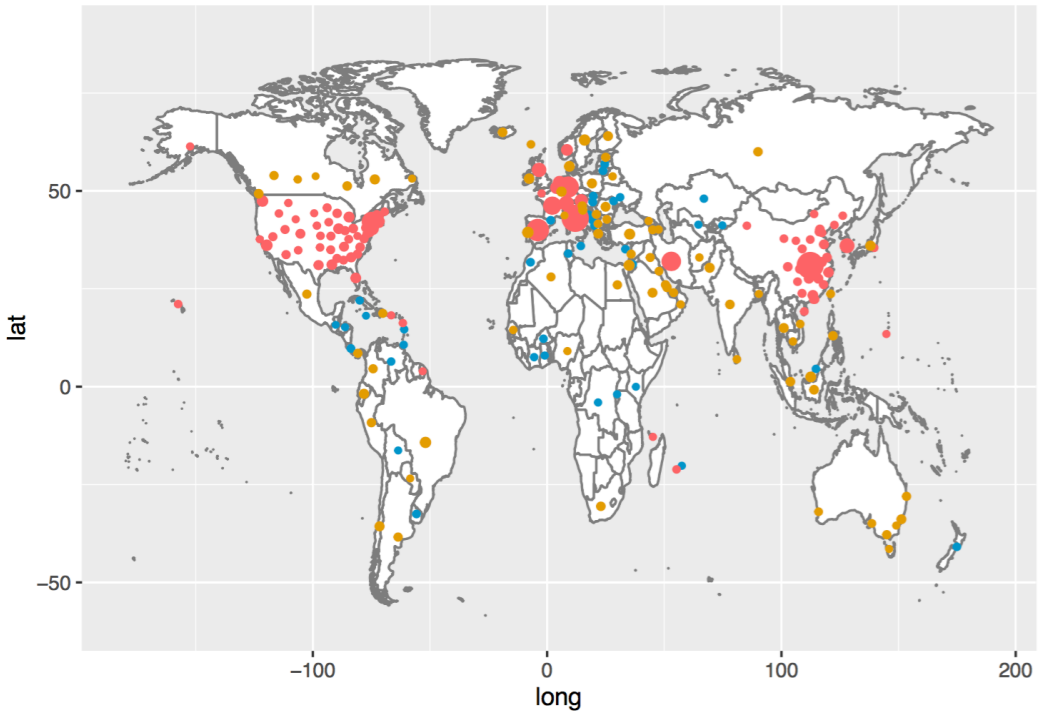

Figure 6: World map of classification using GMM method

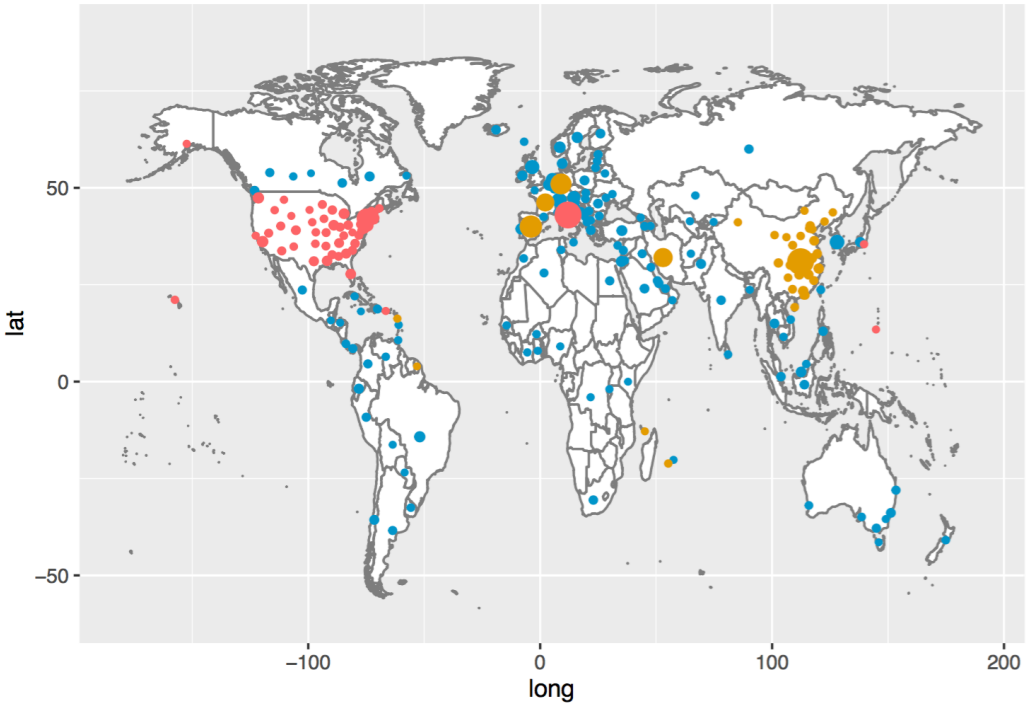

Figure 7: World map of classification using Kmeans method

Table

method	a_value	b_value	c_value
GMM	212.797	0.4922645	14.14861
GMM	43289.675	0.2424189	39.97001
GMM	1268.728	0.2631294	32.00442
Kmeans	1614.435	0.3345930	26.72833
Kmeans	122665.500	0.2860000	41.00000
Kmeans	62764.210	0.2040220	44.78282

Table 1. Centering point result

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

Table 2. Classification result

	country	GMM_class	kmeans_class	a_value	b_value	c_value
51	Israel	3	1	4055.000	0.3040000	33.00000
52	Italy	2	2	138340.000	0.1830000	53.00000
53	Jamaica	1	1	20.000	0.3310000	5.00000
54	Japan	3	1	2195.000	0.0940000	60.00000
55	Jordan	1	1	326.000	0.3020000	21.00000
56	Kazakhstan	1	1	69.000	0.5290000	5.00000
57	Kenya	1	1	237.000	0.3200000	18.00000
58	Korea, South	2	1	8801.392	0.2836325	40.38837
59	Kuwait	3	1	564.000	0.0880000	36.00000
60	Kyrgyzstan	1	1	279.000	0.5460000	9.00000
61	Latvia	1	1	411.000	0.2700000	22.00000
62	Lebanon	3	1	829.000	0.1690000	35.00000
63	Liechtenstein	1	1	55.000	0.5000000	15.00000
64	Lithuania	1	1	432.000	0.4510000	25.00000
65	Luxembourg	3	1	2213.000	0.3540000	24.00000
66	Malaysia	3	1	3231.000	0.2220000	59.00000
67	Malta	1	1	242.000	0.2480000	17.00000
68	Martinique	1	1	135.000	0.2510000	18.00000
69	Mauritius	1	1	115.000	0.4920000	7.00000
70	Mexico	3	1	748.000	0.3170000	25.00000
71	Moldova	1	1	273.000	0.2850000	16.00000
72	Monaco	3	1	60.000	0.2720000	25.00000
73	Montenegro	1	1	124.000	0.5070000	8.00000
74	Morocco	1	1	357.000	0.2910000	22.00000
75	Netherlands	2	1	11170.000	0.2390000	26.00000
76	New Zealand	1	1	505.000	0.4200000	27.00000
77	Nigeria	3	1	102.000	0.4070000	25.00000
78	North Macedonia	3	1	309.000	0.3250000	27.00000
79	Norway	2	1	5557.000	0.1750000	26.00000
80	Oman	3	1	361.000	0.1250000	40.00000
81	Pakistan	3	1	1774.000	0.3260000	26.00000
82	Panama	3	1	715.000	0.3210000	14.00000
83	Paraguay	3	1	74.000	0.1950000	19.00000
84	Peru	3	1	678.000	0.3220000	16.00000
85	Philippines	3	1	1091.000	0.2400000	54.00000
86	Poland	3	1	1821.000	0.2830000	20.00000
87	Portugal	3	1	4741.000	0.3350000	22.00000
88	Qatar	3	1	889.000	0.1750000	19.00000
89	Romania	3	1	1783.000	0.2560000	29.00000
90	Russia	3	1	979.000	0.2910000	53.00000
91	Rwanda	1	1	107.000	0.3560000	11.00000
92	San Marino	1	1	230.000	0.1910000	19.00000
93	Saudi Arabia	3	1	1551.000	0.2880000	23.00000
94	Senegal	3	1	357.000	0.2170000	27.00000
95	Serbia	3	1	627.000	0.2860000	18.00000
96	Singapore	3	1	1262.000	0.0850000	67.00000
97	Slovakia	1	1	254.000	0.3320000	13.00000
98	Slovenia	3	1	805.000	0.2000000	16.00000
99	South Africa	3	1	1303.000	0.3430000	20.00000
100	Spain	2	3	79759.000	0.2570000	52.00000

Table 2. Classification result

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

Table 2. Classification result

method	Dunn_index
Kmeans	0.1227
GMM	0.0013

Table 3. Dunn Index Result

References

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