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:

Conclusions

Figures

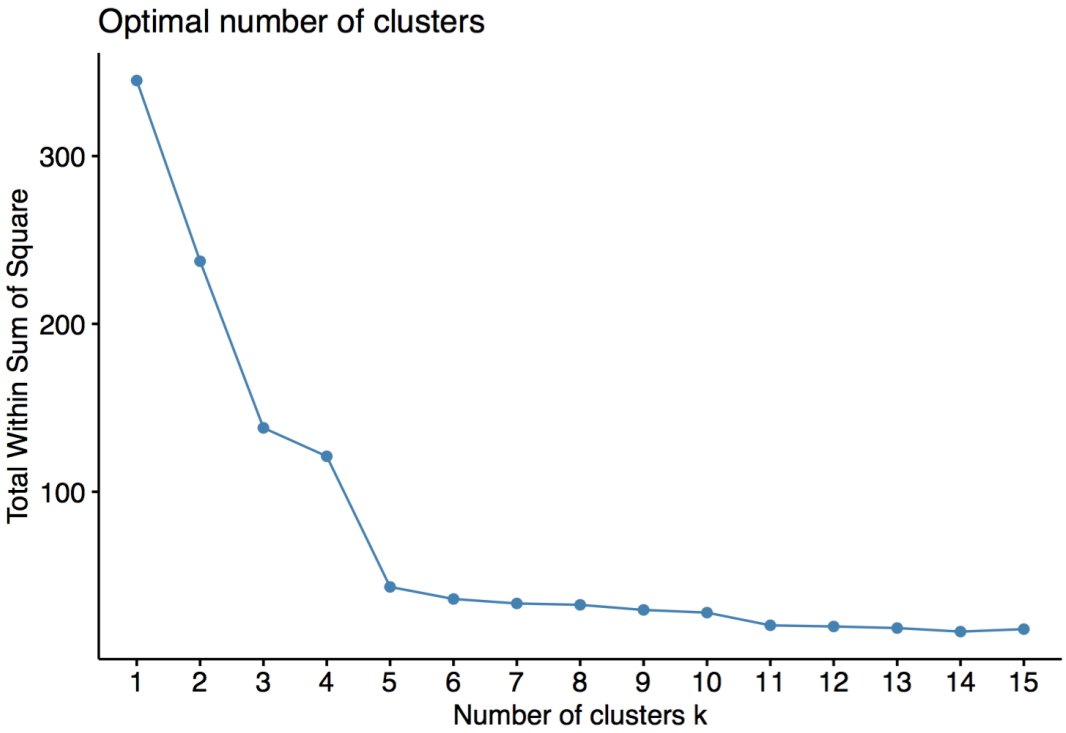

Figure 1: WSS results

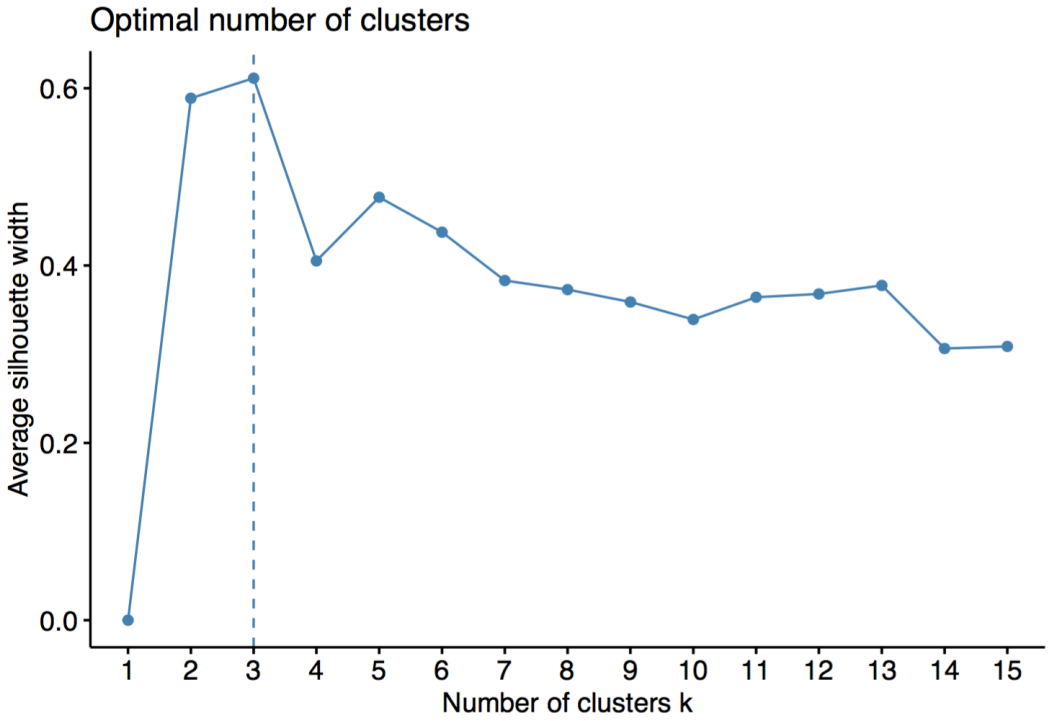

Figure 2: Result of Silhouette Method

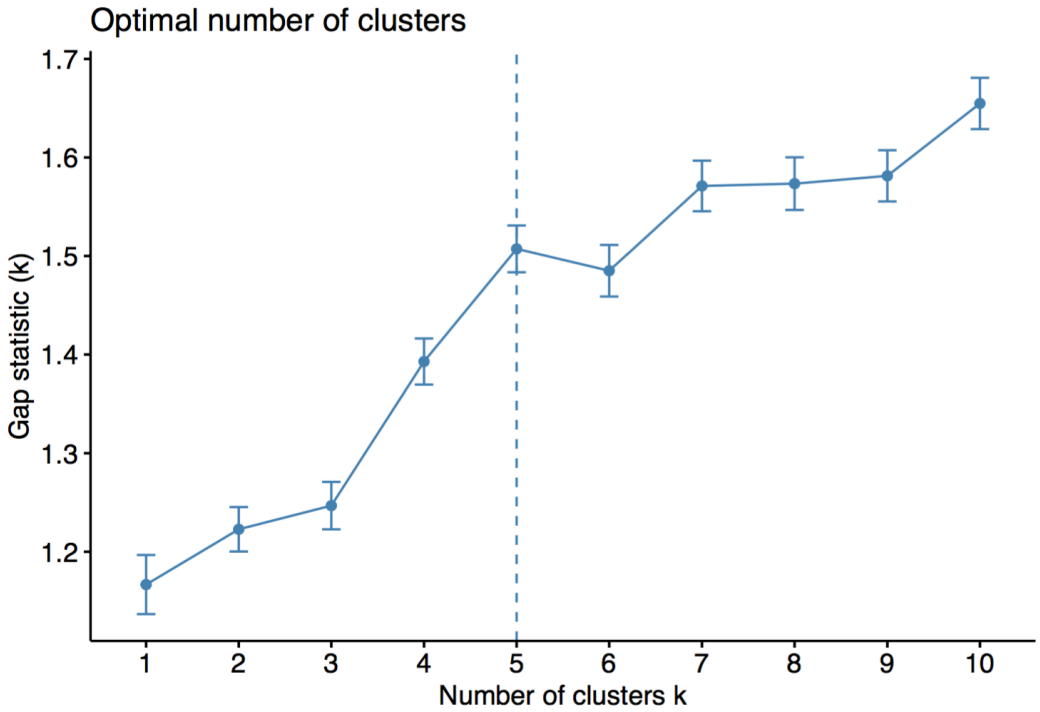

Figure 3: Result of Gap Statistic Method

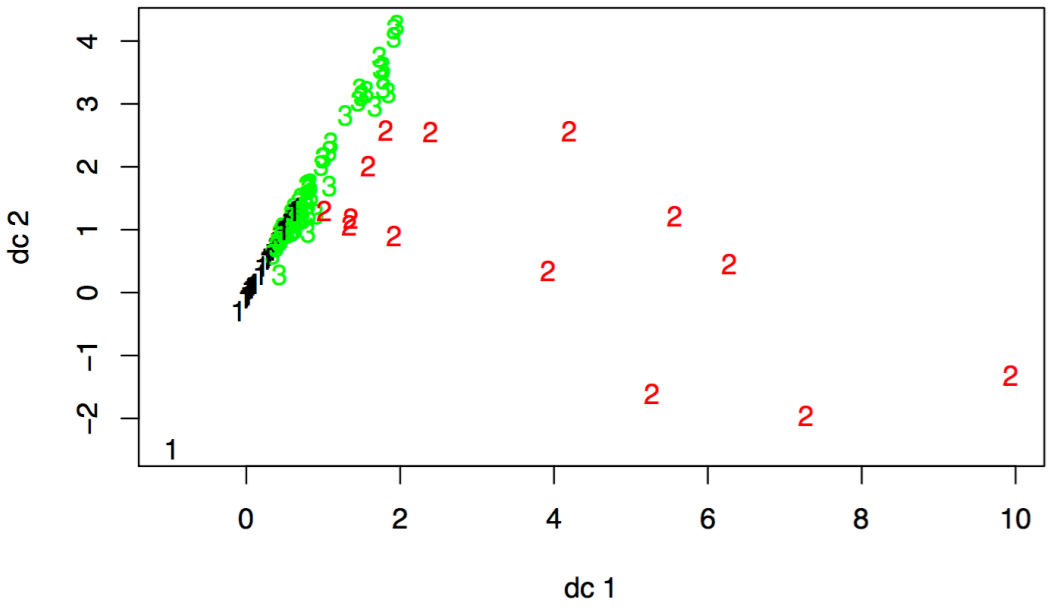

Figure 4: Classification of GMM

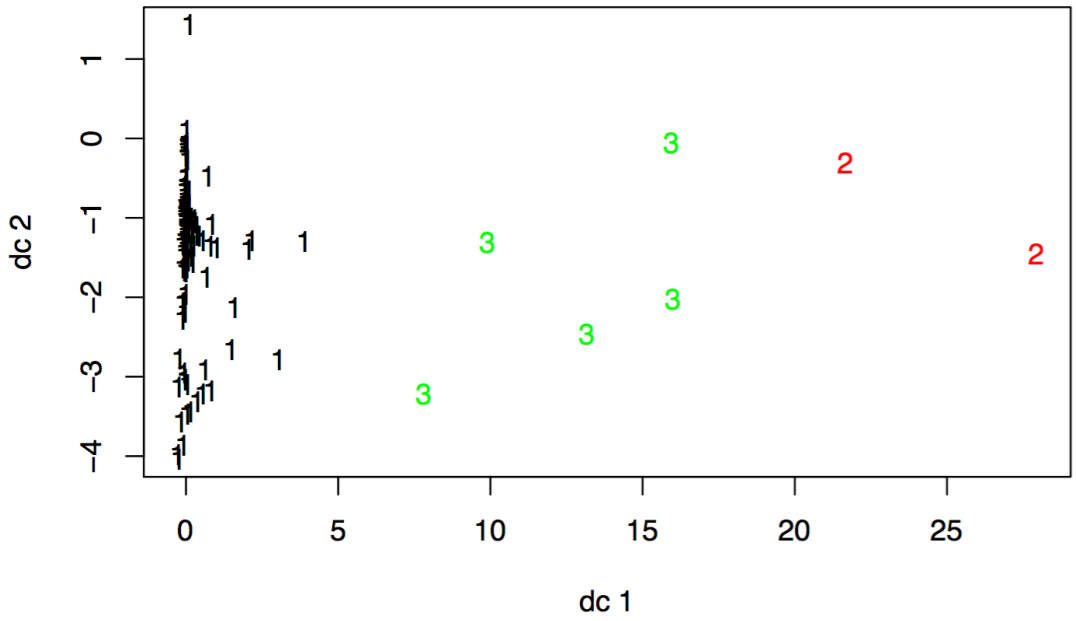

Figure 5: Classification of Kmeans

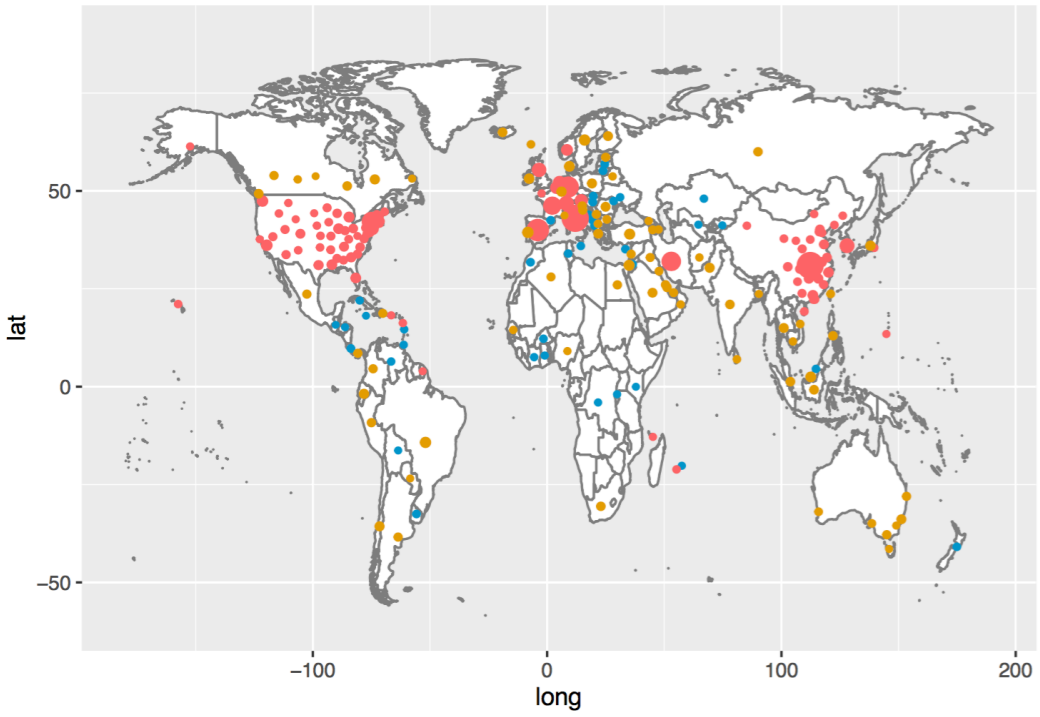

Figure 6: World map of classification using GMM method

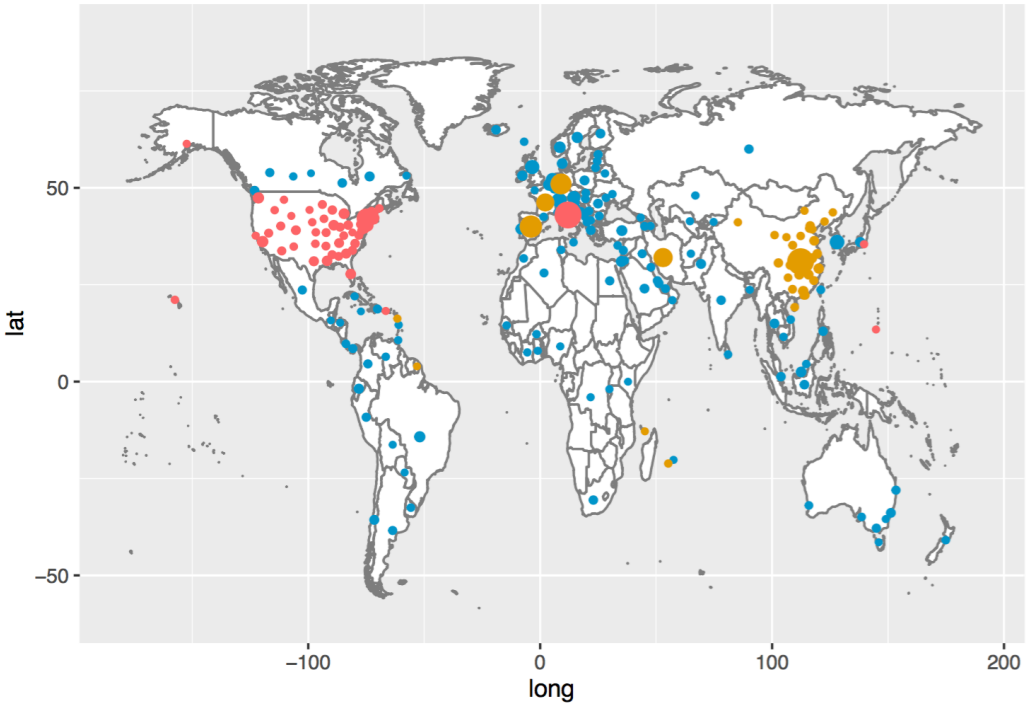

Figure 7: World map of classification using Kmeans method

References

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