Econ 31703: Assignment 3

Due date: May 19, 2021

Exercise 1

(a) Write a function that calculates the K-means objective function for a given clustering:

$$\sum_{i=1}^{N} ||\mathbf{X}_i - \bar{\mathbf{X}}_{k_i}||_2^2 .$$

```
kmeans.objective <- function(K,data,clustering){
## data is a n × (p+1) matrix,
## where the first column denotes y and the rest denotes x's.
## clustering is a N × 1 vector,
## where each component takes a value from 1, ···, K.
⋮
return(a scalar K-means objective)
}
```

Here we discuss K-means clustering with X-property so only use regressors to compute the objective function.

(b) Write a function that updates means of K clusters when given a clustering: given $(k_1, \cdots, k_N)$,

$$\bar{\mathbf{X}}_k = \frac{1}{\sum_{i=1}^{N} \mathbf{1}_{\{k_i=k\}}} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{1}_{\{k_i=k\}}$$

for $k = 1, \cdots, K$.

```
kmeans.mean.update <- function(K,data,clustering){
⋮
return(a K × p matrix)
}
```

(c) Write a function that assigns each unit to a cluster when given means of K clusters: given $(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_K)$,

$$k_i = \arg\min_{1,\cdots,K} \left|\left|\mathbf{X}_i - \bar{\mathbf{X}}_k\right|\right|_2^2,$$

for $i = 1, \cdots, N$.

```
kmeans.clustering.update <- function(K,data,means){
⋮
return(a N × 1 vector)
}
```

(d) Write a wrapper function, which takes in an initial clustering and uses your `kmeans.mean.update` and `kmeans.clustering.update`. The wrapper function applies `kmeans.mean.update` to get cluster means and then updates the clustering with `kmeans.clustering.update`. The function keeps iterating between the two functions until a stopping criterion is met. For stopping criterion, stop the iteration when the number of iteration passes a set maximum `max`, or there is little update in the cluster means:

$$\max_k \left|\left|\bar{\mathbf{X}}_k^{(s)} - \bar{\mathbf{X}}_k^{(s-1)}\right|\right|_2^2 < \varepsilon.$$

```
kmeans <- function(K,data,initial.clustering,max=1000,eps=1-e6){
## loop using the two stopping criteria
⋮
ans <- list(clustering=, ## the final vector of clustering assignment
            means=, ## the final vector of cluster means
            objective=, ## the sequence of objectives updated
            status=, ## which stopping criterion is used?
return(ans)
}
```

During iteration, make sure that cluster means are always defined; if `kmeans.clustering.update` returns a clustering with an empty cluster, either use the cluster mean from the previous iteration or stop iteration and move on to the next initial value.

(e) Generate a dataset using the following DGP: with $\mathbf{X}_i \in \mathbb{R}^2$,

$$\begin{aligned} U_i &\overset{\text{iid}}{\sim} \text{uniform}[0,1], \\ \theta_i &= 1 + \mathbf{1}_{\{U_i \geq 0.2\}} + \mathbf{1}_{\{U_i \geq 0.5\}}, \\ (Y_i, \mathbf{X}_i) \,\big|\, (\theta_1, \cdots, \theta_N) &\overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_{\theta_i}, \Sigma_{\theta_i}\right). \end{aligned}$$

Note that $\theta_i \in \{1,2,3\}$; i.e. a finite mixture model. Let $N = 200$ and let

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0.1 & 0,1 \\ 0.1 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.2 \end{pmatrix}.$$

Report the true clustering based on θ_i, on the space of $\mathbf{X}_i$, as in Figure 1.

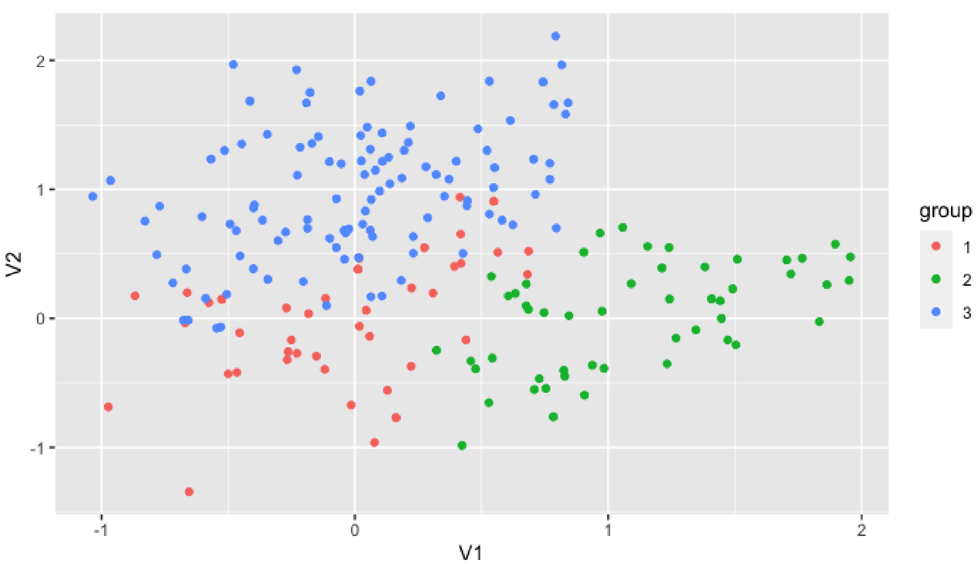

Figure 1: True clustering structure

Experiment with at least 10 values of initial clusterings and report the best K-means clustering on the space of $\mathbf{X}_i$. Discuss the result. Why or why not do you think K-means clustering algorithm captures the underlying DGP well?

(f) Simulate 1,000 datasets following the DGP described in **(e)**. Now, suppose K, the number of mixture components, is not known. Thus, you want to experiment with K. Let $\tilde{K}$ denote your choice of K. For each simulated dataset, experiment with $\tilde{K} = 2, 3, 4$ and 5 and compute

$$\begin{aligned} \text{bias}(\tilde{K}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\mu}_{k_i} - \mu_{\theta_i 1}\right)^2, \\ \text{variance}(\tilde{K}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\bar{U}_{k_i}\right)^2, \end{aligned}$$

for each $\tilde{K}$. k_i denotes the 'estimated' cluster for unit i; thus, $k_i \in \{1, \cdots, \tilde{K}\}$. $\bar{\mu}_k$ is the average of signal ($\mu_{\theta_i 1}$) for each estimated cluster and $\bar{U}_k$ is the average of noise ($Y_i - \mu_{\theta_i 1}$) for each estimated cluster: for each $k = 1, \cdots, \tilde{K}$,

$$\begin{aligned} \bar{\mu}_k &= \frac{1}{\sum_{i=1}^{N}\mathbf{1}_{\{k_i=k\}}}\sum_{i=1}^{N}\mu_{\theta_i 1}\mathbf{1}_{\{k_i=k\}}, \\ \bar{U}_k &= \frac{1}{\sum_{i=1}^{N}\mathbf{1}_{\{k_i=k\}}}\sum_{i=1}^{N}\left(Y_i - \mu_{\theta_i 1}\right)\mathbf{1}_{\{k_i=k\}}. \end{aligned}$$

Report the average of bias($\tilde{K}$) and variance($\tilde{K}$) across simulated datasets for each $\tilde{K}$. Discuss the result.

Table 1 is an example of an estimation result with $\tilde{K} = 2$, $N = 5$. θ_i is the true cluster and k_i is estimated cluster based on $\{X_{i1}, X_{i2}\}_{i=1}^{N}$. Note that $\theta_i \in \{1, 2, 3\}$ while $k_i \in \{1, 2\}$.

i	k_i	θ_i	$\mu_{\theta_i 1}$	$\mu_{\theta_i 2}$	$\mu_{\theta_i 3}$	Y_i
1	1	1	0	0	0	1.2
3	1	2	2	1	0	1.8
4	2	2	2	1	0	2.5
8	2	3	1	0	1	0.1
9	2	3	1	0	1	0.4

Table 1: An estimation result

In this example,

$$\begin{aligned}
\bar{\mu}_1 &= \frac{1}{2}(0+2) = 1, \\
\bar{\mu}_2 &= \frac{1}{3}(2+1+1) = \frac{4}{3}, \\
\bar{U}_1 &= \frac{1}{2}(1.2+1.8) - 1 = \frac{1}{2}, \\
\bar{U}_2 &= \frac{1}{3}(2.5+0.1+0.4) - \frac{4}{3} = -\frac{1}{3}
\end{aligned}$$

Then,

$$\begin{aligned}
\text{bias}(2) &= \frac{1}{5}\left[(1-0)^2 + (1-2)^2 + (4/3-2)^2 + (4/3-1)^2 + (4/3-1)^2\right] = \frac{8}{15} \\
\text{variance}(2) &= \frac{1}{5}\left[(1/2)^2 + (1/2)^2 + (-1/3)^2 + (-1/3)^2 + (-1/3)^2\right] = \frac{6}{25}
\end{aligned}$$