Homework 3

Please submit both your code and results. Please use R for this homework.

In this homework, you should use a new PBMC dataset with the given cell type labels stored as the Seurat object (v5) [*PBMC_w_labels.rds*]. This dataset has nine different cell types.

1. Data preparation. Perform the following steps in Seurat with default parameters:
   (1) log1pCP10k, (2) highly variable gene selection, (3) scaling data, and (4) PCA.

2. Implement the following four clustering algorithms on the PCA results (the first 50 PCs). Also, plot cells using their scores of the first two PCs and color them according to five sets of labels. Present these in five separate figures: four showing the cluster labels in the four clustering results and one showing the true cell type labels.
   (1) kmeans [Rfunction]: Set k to 9.
   (2) kmeans++ [Rfunction]: Set k to 9
   (3) Hierarchical clustering (complete linkage; Euclidean distance for the dissimilarity) [Rfunction]: Cut the dendrogram to obtain 9 clusters.
   (4) Seurat clustering: Choose an appropriate `resolution` parameter in FindCluster() to obtain 9 clusters.

3. Evaluate and compare the four clustering results using:
   (1) Silhouette score [Rfunction] (use the Euclidean distance).
   (2) Adjusted rand index (ARI) [Rfunction]. Please refer to this introduction to ARI and provide a brief explanation of it.

   Which clustering method has the best performance based on each metric? Can you explain the different conclusions?

4. Use the following nine combinations for parameters of the Seurat clustering algorithm. How does each parameter influence the number of clusters?

k.neighbor (FindNeighbors)	prune.SNN (FindNeighbors)	resolution (FindClusters)
10	default	default
20	default	default
50	default	default
default	1/5	default
default	1/10	default
default	1/15	default
default	default	0.5
default	default	1
default	default	1.5

5. Instead of setting k as nine in question 2, please calculate the Gap statistics for kmeans clustering, with k ranging from 1 to 15 [Rfunction]. What is the best k for kmeans based on Gap statistics?

6. Perform the tight clustering on the PCA results (50 PCs) [Rfunction]. Set `target`=9, and `k.min`=15. What are the meanings of these two parameters? Plot cells using their scores of the first two PCs, coloring them according to the cluster labels from tight clustering results (use the gray color for cells labeled -1). Please explain why some cells are labeled as -1.