Lab 3 - Parallelizing k-means

Sahil Saxena

October 26, 2021

1 Introduction

This report aims to assess the stability of k-means through a popular procedure outlined in Ben-Hur et al. [2001], which uses stability as a guide for picking k. The data used is the provided lingBinary.Rdata from Lab 2, which contains answers to various linguistic survey questions from $45,512$ respondents from across the United States. The report clusters the data based on survey responses using the stability method, through which different coding languages (R and C++) are experimented with for performance.

2 Performance of Similarity Computations

The correlation measure was used to compute similarity, as outlined by Fowlkes and Mallows in Ben-Hur et al. [2001]. It is worth noting that the asymptotic runtime for calculations in both languages is $O(n^2)$ for size n vectors, as every pair of observations $(x1[i], x2[j])$ is iterated through. Also, both methods do not explicitly store the q-by-q similarity matrix described, where q is the number of data points common to each subsample. Instead, each element of the matrix C_{ij} is computed by iterating over the original cluster assignment vector of size n, yielding a memory complexity of $O(n)$.

On randomly sampled membership vectors of size 5000 (for each of $k = 3, 4, and 5$ clusters), the similarity measure was calculated using both R and C++. As seen in the experiments, C++ is at least 23.35 times faster and up to 45.31 times faster than R. Actual running time does vary greatly between these languages; computational efficiency is a critical aspect to consider, especially when working with large datasets.

Type	k (# of clusters)	Elapsed Time (s)
R	3	4.390
C++	3	0.188
R	4	6.479
C++	4	0.143
R	5	3.958
C++	5	0.122

3 Stability Analysis of k-means

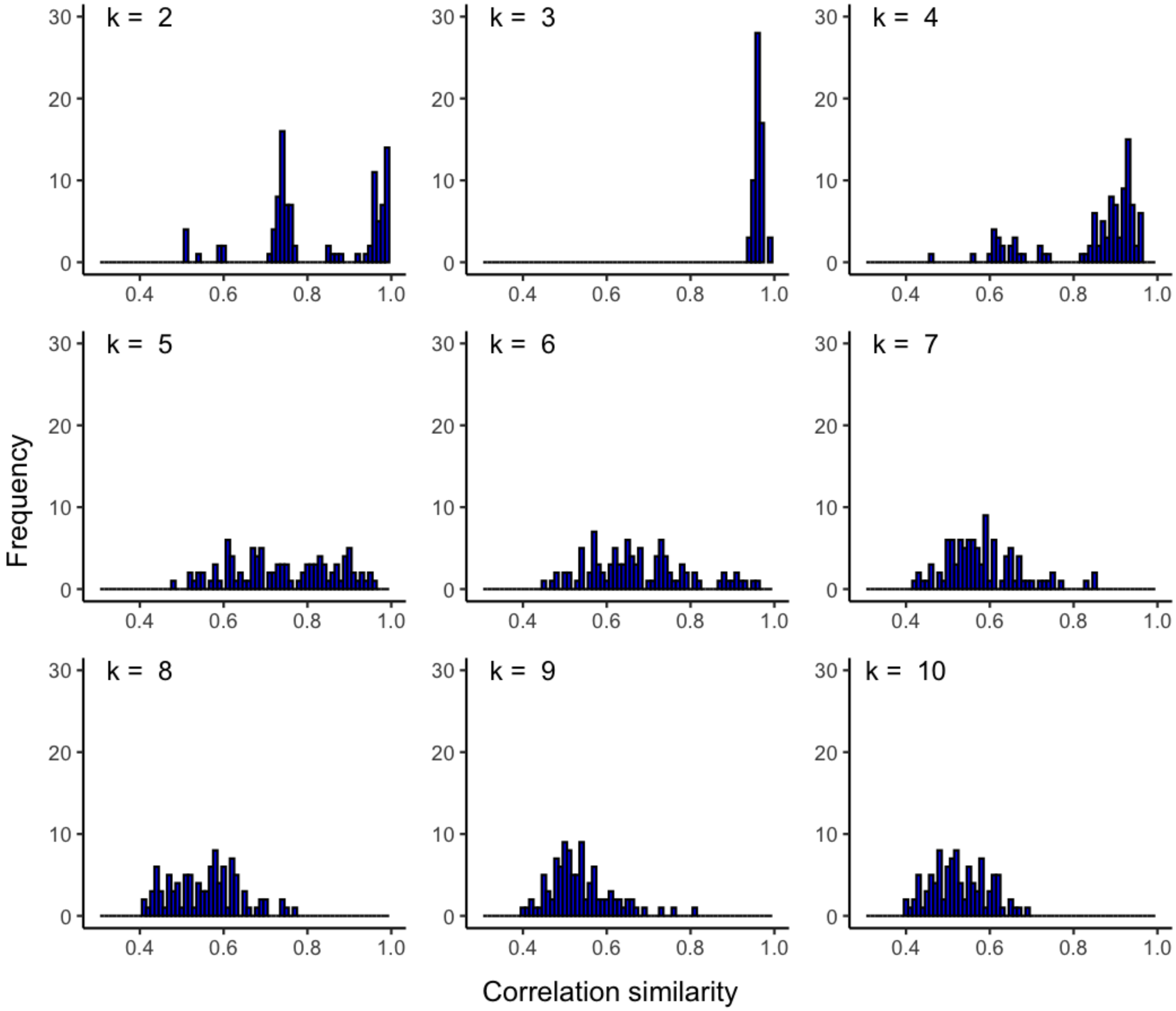

Figure 1: Distribution of the correlation similarity for different values of k.

Figure 1 shows the distribution of the correlation measure for different values of k. $k = 3$ produced the most stable results with high mean and low variance in the distribution of correlation similarity scores. As the number of clusters increases beyond 3, the spread of the distributions generally become greater and the mean decreases, indicating that k-means becomes unstable for more than 3 clusters.

Figure 2 reinforces this idea, where the cumulative density shows that more correlation values are near 1 for $k = 3$ than any other choice of k. Therefore, it is reasonable to choose $k = 3$ to cluster the linguistics dataset. $k = 3$ is trustworthy because in Lab 2, it was determined that 3 clusters produced the lowest average silhouette value. This suggests that using this stability method may be a reliable method for choosing an appropriate value of k, as it provides similar results to the silhouette test.

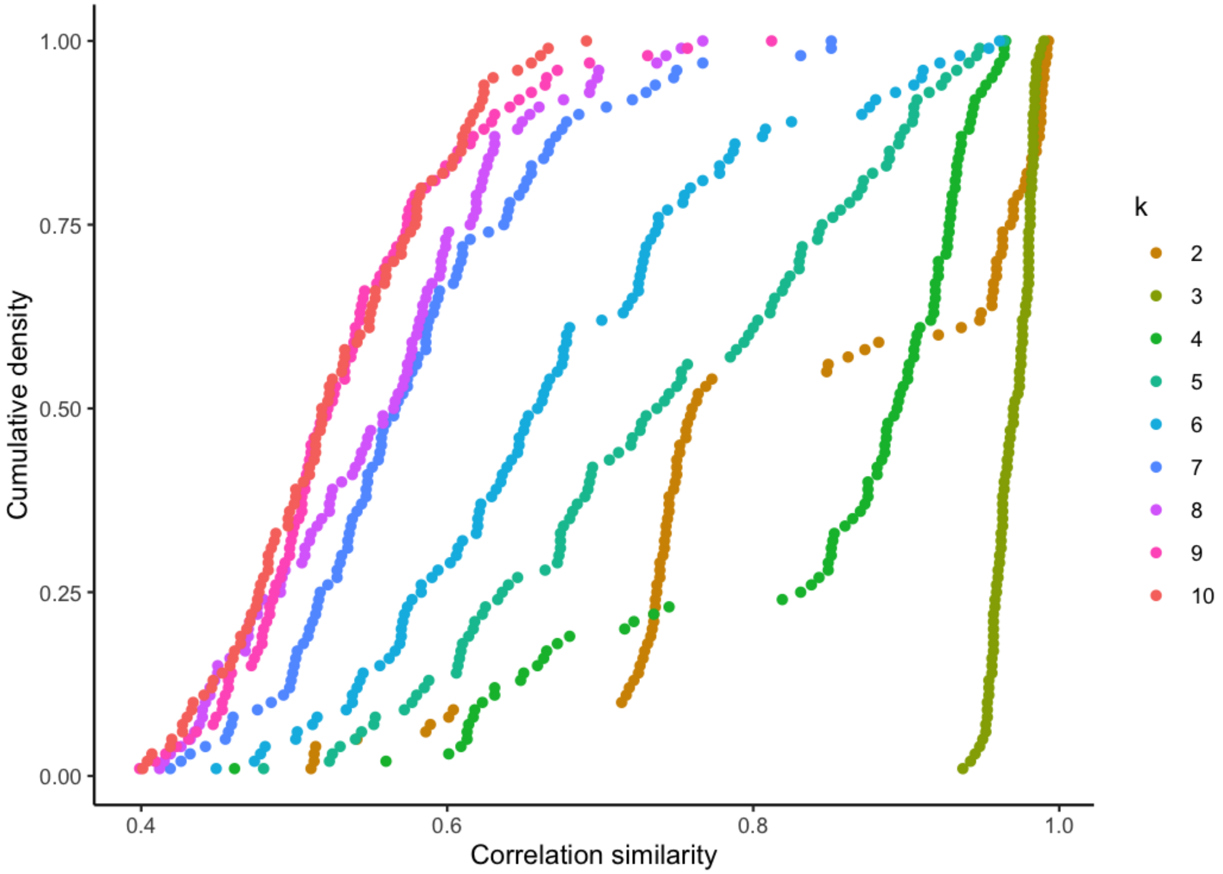

Figure 2: Cumulative distribution of the correlation similarity for different values of k.

4 Conclusion

It was determined that C++ runs much faster than R for the similarity calculation, which gives merit to the idea of using C++ rather than R when dealing with large datasets and lots of computation. Also, this stability method indicated that $k = 3$ clusters should be chosen for clustering the linguistics dataset from Lab 2, which also agrees with the choice of $k = 3$ clusters. It is fair to conclude that this stability method is a good way to test roboustness of clustering decisions in the future.

5 References

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In Pacific symposium on biocomputing, volume 7, pages 6–17, 2001.