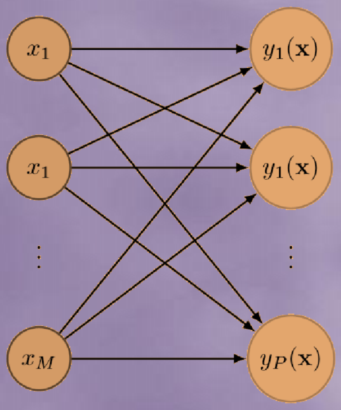

Linear Models

Daniel Khashabi[1]
KHASHAB2@ILLINOIS.EDU

0.1 K-means and Silhouette Statistic:

The k-means clustering method is described in the previous section. So I don't bring the explanations here again. The silhouette statistic of a measurement shows how well it fits in its own cluster versus how well it fits in its next closest cluster. The silhouette value for i-th measurement could be calculated by

$$s(i) = \frac{b(i) - a(i)}{\max\left\{a(i) - b(i)\right\}}$$

Based on this formula, $-1 \leq s(i) \leq 1$, while generally $s(i) \geq$ and slightly around zero. As $s(i)$ gets closer to one, the better clustering is done. The "Silhouette" test uses the silhouette statistic to test how well data is clustered. The silhouette test subdivides the data into successively more clusters looking for the first minimum of the silhouette statistic. The silhouettes for k-means with different k values in Fig. 1. Based on the this figure, we choose $k = 2$.

The silhouette diagram for clustering for $k = 2$ is shown in Fig. 2. While $k = 2$ is the best value for clustering, the data is so intertwined and not so appropriate for fully clustering. That's why the silhouette for a big proportion of data is negative. In order to visualize the results of the clustering we use PCA to decompose the data into various dimension. The Fig 3 shows the proportion of eigenvalues(variances). In demonstrations we use 8 first PCs. The result of pairwise visualization is shown in Fig. **??** Note that we use this pairwise visualization because the data in different dimensions is intertwined. So we need to see the clustering from different directions.

The clustering figure in the pairwise in Fig. **??** also show how the data is intertwined

[1] This is part of my notes; to see the complete list of notes check web.engr.illinois.edu/ khashab2/learn.html. This work is licensed under a Creative Commons Attribution-NonCommercial 3 License.

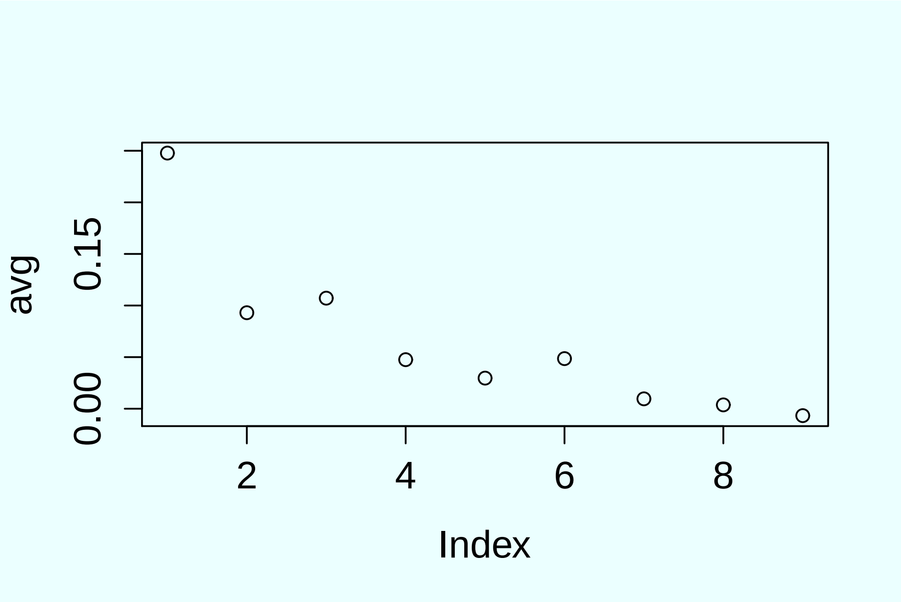

Figure 1: The silhouettes for k-means with different k values.

0.1.1 Hierarchical Clustering:

In hierarchical clustering we make use of mutual dissimilarity measure to build a hierarchical clustering structure. It this method we only need pairwise dissimilarity. To create this hierarchical clustering structure, we can follow different paths. One possible way is to do bottom-up clustering, in which we assume the data points to be in different clusters and we gradually merge them together until all observations are in one unit cluster. In *top-down clustering* we start with one cluster and gradually split them. In order to define a dissimilarity between two clusters each with more than one elements, we could use different approaches. For example *Single-linkage* assumes the nearest neighbours in each groups. In *complete-linkage* we use furthest neighbours from each group. Another reasonable way is to assume *group average*. There also some other methods for doing this comparison.

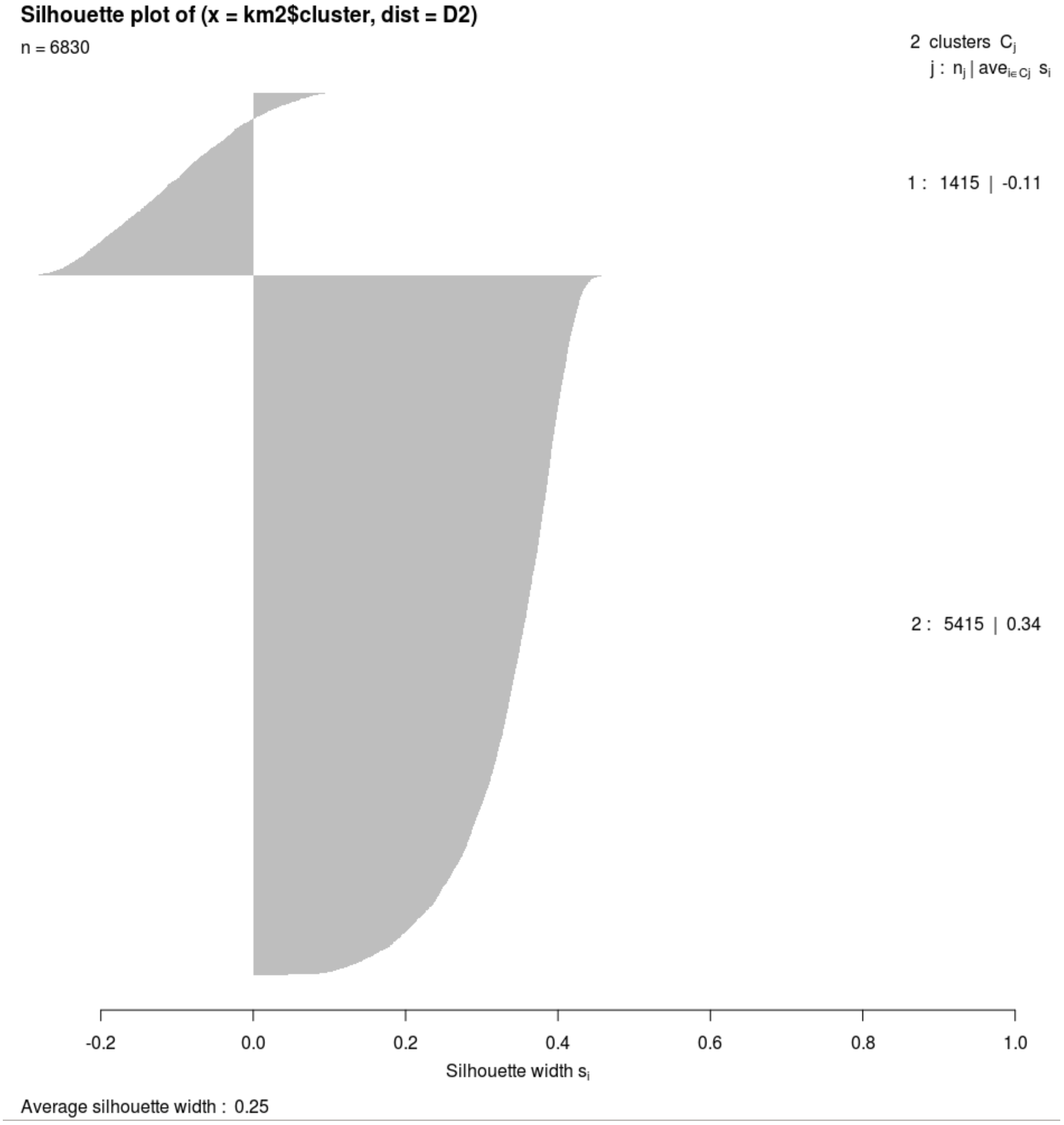

Figure 2: The silhouette diagram of k-means clustering for $k = 2$.

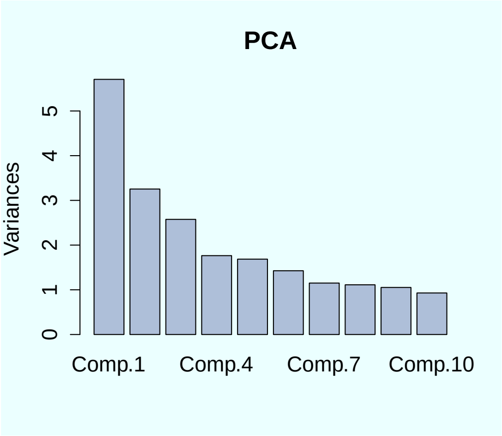

Figure 3: Variance portion of all eigenvalues of major directions.

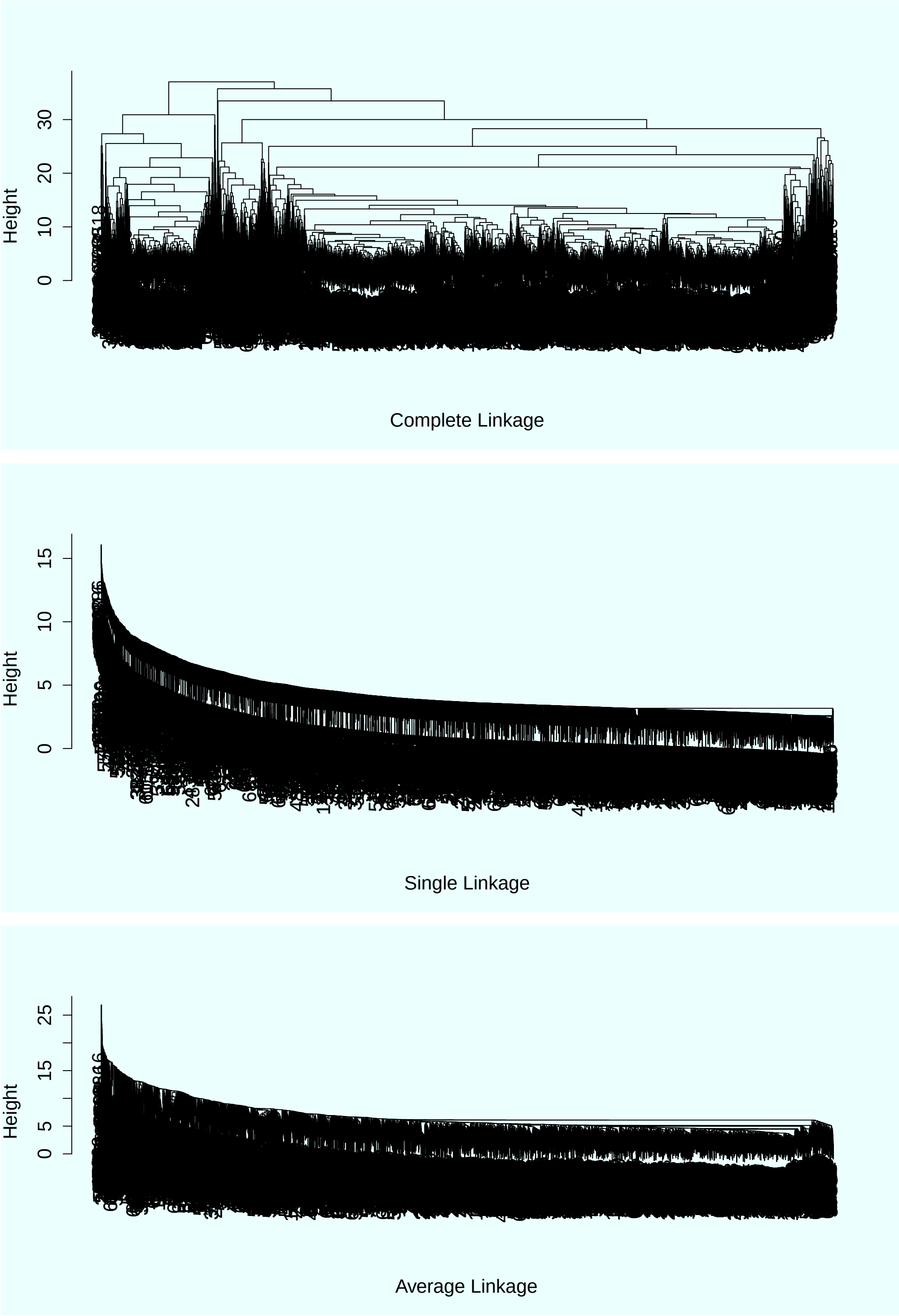

Figure 4: The various clustering methods in hierarchical clustering.