12.1.2 Example: Old Faithful Geyser Eruptions

The data for this example[1] is a set of 107 bivariate observations, that were taken from a study of the eruptions of Old Faithful Geyser in Yellowstone National Park, Wyoming (Weisberg, 1985, p. 231). A geyser is a hot spring which occasionally becomes unstable and erupts hot water and steam into the air. Old Faithful Geyser is the most famous of all geysers and is an extremely popular tourist attraction. The variables measured are duration of eruption (X_1) and waiting time until the next eruption (X_2), both recorded in minutes, for all eruptions of Old Faithful Geyser between 6 a.m. and midnight, 1–8 August 1978. Prior to clustering, one could argue that there are two or three possible clusters in the data.

Because the two variables are measured on very different scales (the standard deviations of X_1 and X_2 being approximately 1 and 13, respectively), the derived clusters (using any clustering algorithm) are completely determined by X_2, the interval between eruptions; the observations are divided into clusters by straight-line boundaries parallel to the horizontal axis. Without standardizing both variables, we cannot obtain a realistic partitioning of the data. So, for this example, we standardize the variables prior to clustering.

The results of this clustering study, where we set the number of clusters to be two or three for each method, are displayed in Figure 12.1. The most interesting result is that "perfect" clustering (according to our intuition) for both two and three clusters is accomplished only by the single-linkage, hierarchical agglomerative method (see first row of Figure 12.1). If we use the single-linkage results as the gold standard, we see that average-linkage and complete-linkage methods (second row), which produced the same results for two and three clusters, had one incorrect allocation for two clusters and three incorrect allocations for three clusters. Although both of the non-hierarchical clustering methods, `pam` and K-means (third row), had perfect clustering for two clusters, they performed poorly for three clusters, where they both had 45 incorrectly allocations.

[1] The data can be found in the file `geyser` on the book's website.

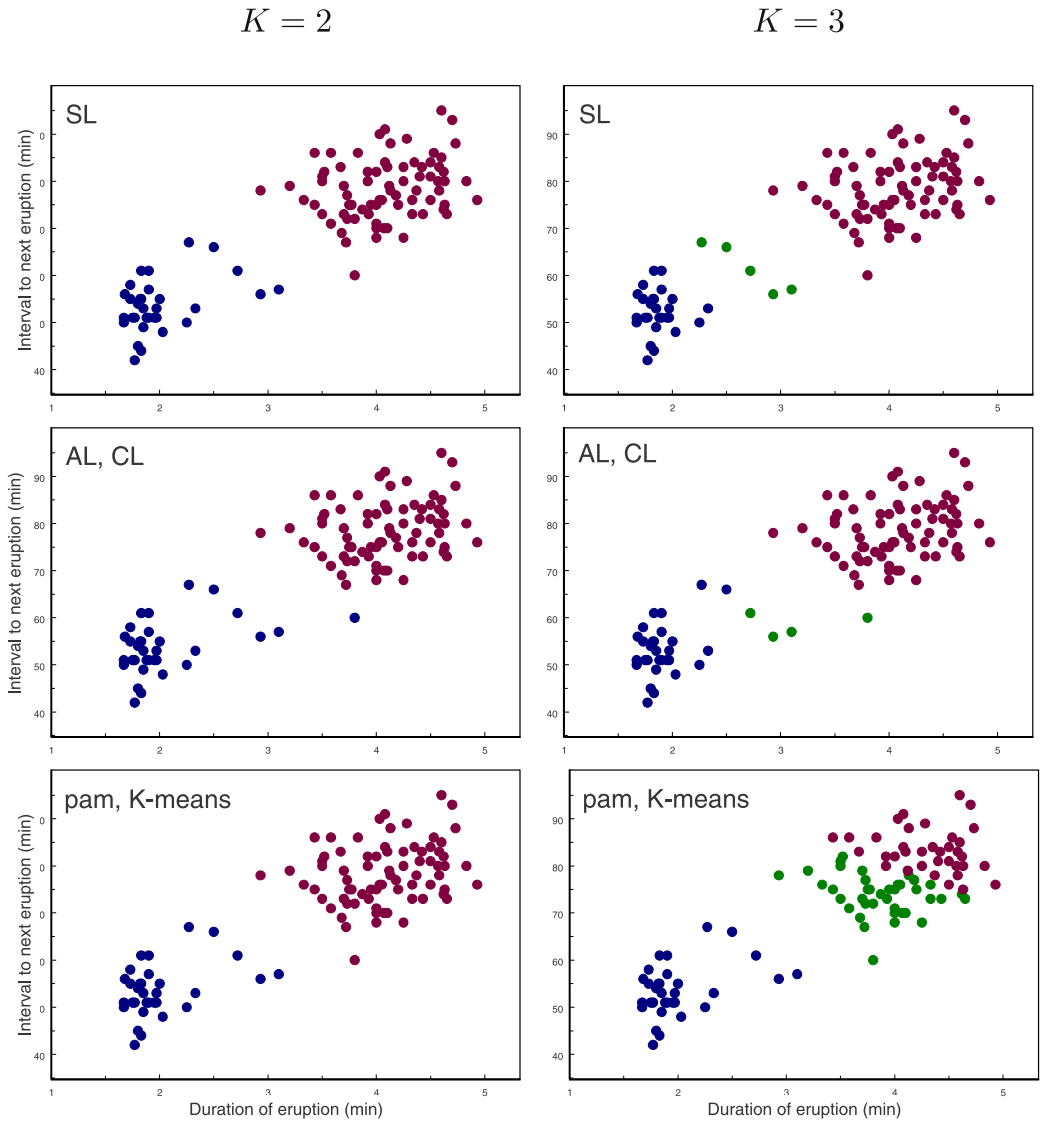

FIGURE 12.1. *Clustering results for Old Faithful Geyser data. The scatterplots in the left column panels are solutions for $K = 2$ classes, with red and blue as the two cluster colors. The scatterplots in the right column panels are solutions for $K = 3$ classes, with red, green, and blue as the three cluster colors. The first row is the single-linkage (SL) solutions, the second row is both average-linkage (AL) and complete-linkage (CL) solutions, the third row is both* `pam` *and K-means solutions.*