11.2: Cluster Analysis

In the previous section we examined the spectra of 24 samples at 635 wavelengths, displaying the data by plotting the absorbance as a function of wavelength. Another way to examine the data is to plot the absorbance of each sample at one wavelength against the absorbance of the same sample at a second wavelength, as we see in the following figure using wavelengths of 403.3 nm and 508.7 nm. Note that this plot suggests an underlying structure to our data as the 24 points occupy a triangular-shaped space. defined by the samples identified as 1, 2, and 3.

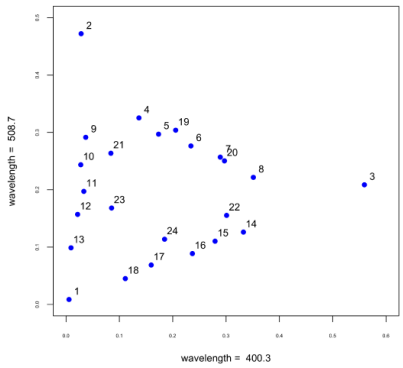

Figure 11.2.1: Plot showing the absorbance values for the 24 samples from Figure 11.1.1 at wavelengths of 400.3 nm and 508.7 nm. The numbers next to the points are index values for the samples.

We can extend this analysis to three wavelengths, as we see in the following figure, and, to as many as all 635 wavelengths (Of course we cannot examine a plot of this as it exists in 635-dimensional space!).

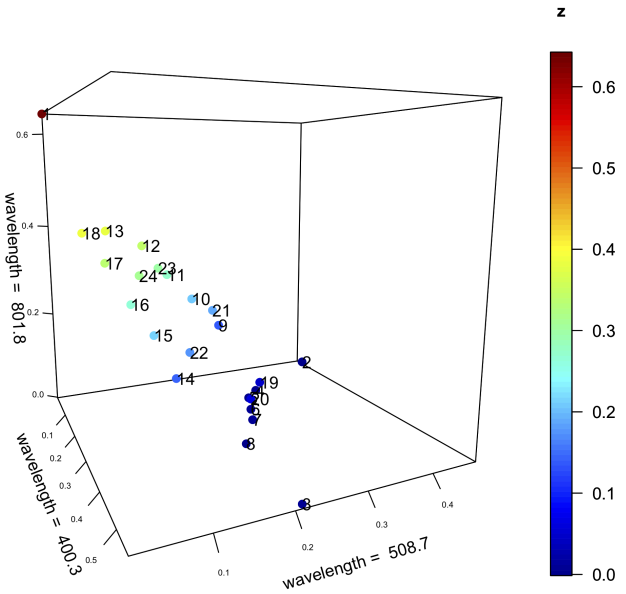

Figure 11.2.2: Plot showing the absorbance values for the 24 samples from Figure 11.1.1 at wavelengths of 400.3 nm, 508.7 nm, and 801.8 nm. The color of the points shows the absorbance at 801.8 nm, which is the *z*-axis. The numbers next to the points are index values for the samples. Note that the 24 points here also reside in a triangular-shaped space.

In both Figure 11.2.1 and Figure 11.2.2 (and the higher dimensional plots that we cannot display), some samples are closer to each other in space than are other points. For example, in Figure 11.2.1, samples 7 and 20 are closer to each other than any other pair of samples; samples 2 and 3, however, are further from each other than any other pair of samples.

How Does a Cluster Analysis Work?

A cluster analysis is a way to examine our data in terms of the similarity of the samples to each other. Figure 11.2.3 outlines the steps using a small set of six points defined by two variables, *a* and *b*. Panel (*a*) shows the six data points. The two points closest in distance are 3 and 4, which make the first cluster and which we replace with the red point midway between them, as seen in panel (*b*). The next two points closest in distance are 2 and 6, which make the second cluster and which we replace with the red point between them, as seen in panel (*c*). Continuing in this way yields the results in panel (*d*) where the third cluster brings together points 2, 3, 4, and 6, the fourth cluster brings together points 1, 2, 3, 4, and 6, and the final cluster brings together all six points.

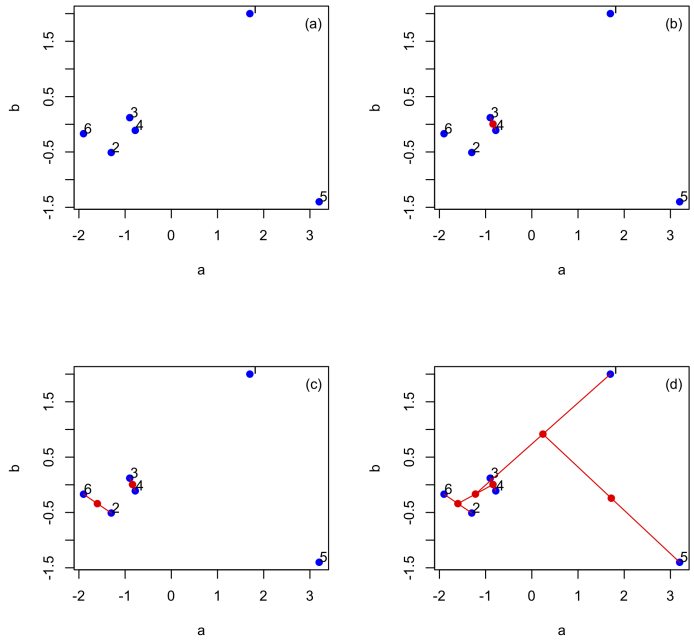

Figure 11.2.3: Example of how cluster analysis works. See the text for details.

To visualize the clusters, in terms of the identify of the points in the clusters, the order in which the clusters form, and the relative similarity of difference between points and clusters, we display the information in Figure 11.2.3*d* as the dendrogram shown in Figure 11.2.4, which shows, for example, that the clusters of points 3 and 4, and of 2 and 6 are more similar to each other than they are to point 1 and to point 6. The vertical scale, which is identified as Height, provides a measure of the distance of the individual points or clusters of points from each other.

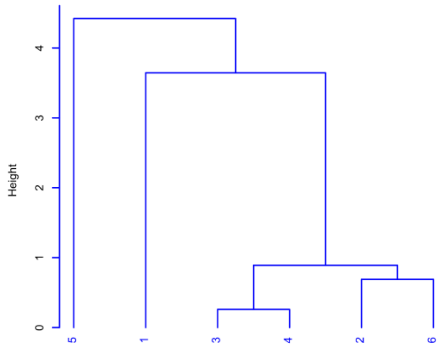

Figure 11.2.4: The results of the cluster analysis in Figure 11.2.3*d* displayed as a dendrogram.

How Do We Interpret the Results of a Cluster Analysis?

A cluster analysis of the 24 samples from Figure 11.1.1 is shown in Figure 11.2.5 using 40 equally-spaced wavelengths. There is much we can learn from this diagram about the structure of these samples, which we can divide into three distinct clusters of samples, as shown by the boxes. The samples within each cluster are more similar to each other than they are to samples in other clusters. One possible explanation for this structure is that the 24 samples are comprised of three analytes, where, for each cluster, one of the analytes is present at a higher concentration than the other two analytes.

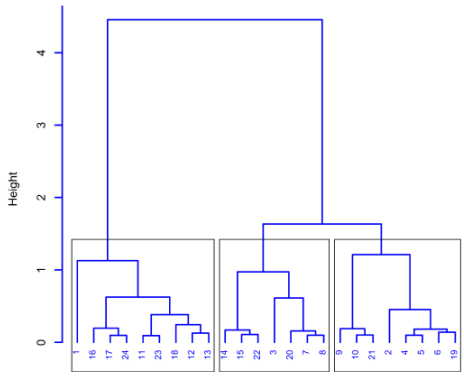

Figure 11.2.5: Cluster analysis of the 24 samples from Figure 11.1.1. The boxes divide the 24 samples into three distinct clusters.

This page titled 11.2: Cluster Analysis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by David Harvey.