Homework 6

Cluster analysis

We perform hierarchical clustering on the states using the `USArrests` data in the `ISLR` package. For each of the 50 states in the United States, the dataset contains the number of arrests per 100,000 residents for each of three crimes: `Assault`, `Murder`, and `Rape`. The dataset also contains the percent of the population in each state living in urban areas, `UrbanPop`. The four variables will be used as features for clustering.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?

PCA

PCA can be used for image compression. Suppose we have a $n \times p$ data matrix $\boldsymbol{X}$. For $k < p$, define $V_k = (\phi_1, \ldots, \phi_k) \in \mathbb{R}^{p \times k}$, where ϕ_j is the jth PC direction. The Projection of $\boldsymbol{X}$ on to V_k is $\boldsymbol{X} V_k V_k^\top$, which can be thought of as an approximation to $\boldsymbol{X}$. It can be shown that $\boldsymbol{X} V_k V_k^\top$ is the best rank k approximation to $\boldsymbol{X}$.

In this question, we use the `jpeg` package to read and write the .jpeg files. We use a image of cat for illustration, and the sample codes are given in "image.R". Read the image using `img <- readJPEG('example.jpg')`. The image will be represented as three matrices as an array with each matrix corresponding to the RGB color value scheme and each element in a matrix corresponding to one pixel. Extract the individual color value matrices to perform PCA on each of them. Reconstruct the original image using the projections of the data with the first 20 PCs.

Now use your own .jpg image to perform image compression via PCA with different numbers of PCs (e.g., 50, 100, 200, ...).