H3: Principal Component Projection

(1) For the data set "8OX" introduced in class, there are $n = 45$ patterns from $k = 3$ categories, each pattern consists of $d = 8$ features. Each pattern can be denoted by $\mathbf{x}_i^{(k)},\ 1 \le i \le 15,\ 1 \le k \le 3$, where $\mathbf{x}_i^{(k)} \in R^d$.

(a) Compute the pooled $d \times d$ covariance matrix $C = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^{(k)} - \mathbf{u})(\mathbf{x}_i^{(k)} - \mathbf{u})^t$, where $\mathbf{u} = \frac{1}{n}\sum_{k=1}^{3}\sum_{i=1}^{15}\mathbf{x}_i^{(k)}$ is the mean vector.

(b) Report the eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$ of C.

(c) Report the percentage of $\gamma_j = \frac{\sum_{i=1}^{j}\lambda_i}{\sum_{i=1}^{d}\lambda_i},\ \ \forall\ 1 \le j \le d.$

(d) Plot n patterns using the first two principal components.

(e) Show the dendrogram (by complete linkage) of projected "8OX" data using only the first two principal components.

(f) Show the dendrogram (by complete linkage) of the original "8OX" data using the $d = 8$ features.

(2) For the data set "iris" introduced in class, there are $n = 150$ patterns from $k = 3$ categories, each pattern consists of $d = 4$ features. Repeat the same processes as required in problem **(1)**.

(3) For the data set "imox" introduced in class, there are $n = 192$ patterns from $k = 4$ categories, each pattern consists of $d = 8$ features. Repeat the same processes as required in problem **(1)**.