DS-6030 Homework Module 10

Matt Scheffel

DS 6030 | Spring 2022 | University of Virginia

```
library(ISLR)
library(tidyverse)
```

8. In Section 12.2.3, a formula for calculating PVE was given in Equation 12.10.

We also saw that the PVE can be obtained using the `sdev` output of the `prcomp()` function.

On the `USArrests` data, calculate PVE in two ways:

(a) Using the `sdev` output of the `prcomp()` function, as was done in Section 12.2.3.

```
data(USArrests)

pca <- prcomp(USArrests, scale = TRUE)

# PVE
pve <- pca$sdev^2/sum(pca$sdev^2)

pve
```

```
#> [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

(b) By applying Equation 12.10 directly. That is, use the `prcomp()` function to compute the principal component loadings. Then, use those loadings in Equation 12.10 to obtain the PVE.

These two approaches should give the same results.

Hint: You will only obtain the same results in (a) and (b) if the same data is used in both cases. For instance, if in (a) you performed `prcomp()` using centered and scaled variables, then you must center and scale the variables before applying Equation 10.3 in (b).

```
data(USArrests)

# PCA
pca <- prcomp(USArrests, scale = TRUE)

# calculate principal component loadings
loadings <- pca$rotation

# center and scale variables
x <- scale(USArrests)

# calculate covariance matrix of x
cov_x <- cov(x)
```

```r
# calculate eigenvalues and eigenvectors of covariance matrix
eig <- eigen(cov_x)

# calculate total variance
total_var <- sum(eig$values)

# calculate PVE using Equation 12.10
pve <- eig$values/total_var

# print PVE
pve
```

```
#> [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

9. Consider the USArrests data.

We will now perform hierarchical clustering on the states.

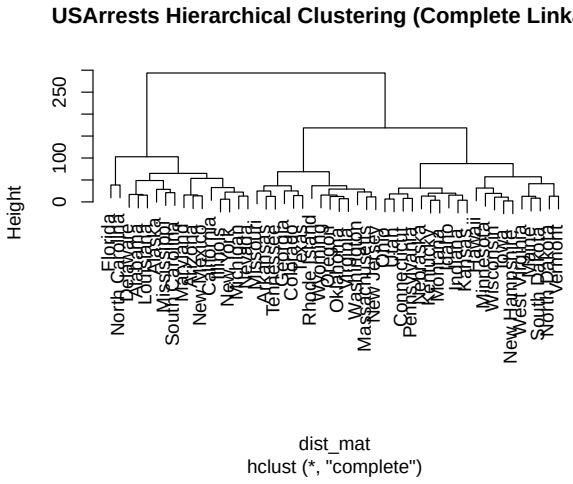
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

```r
data(USArrests)

# Euclidean distance matrix
dist_mat <- dist(USArrests)

# hierarchical clustering with complete linkage
hc_complete <- hclust(dist_mat, method = "complete")

# dendrogram
plot(hc_complete, main = "USArrests Hierarchical Clustering (Complete Linkage)")
```



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

```r
# cut the dendrogram at height 150 to obtain 3 clusters
clusters <- cutree(hc_complete, h = 150)

# states belonging to each cluster
cbind(State = rownames(USArrests), Cluster = clusters)
```

```
#>                State            Cluster
#> Alabama        "Alabama"        "1"
#> Alaska         "Alaska"         "1"
#> Arizona        "Arizona"        "1"
#> Arkansas       "Arkansas"       "2"
#> California     "California"     "1"
#> Colorado       "Colorado"       "2"
#> Connecticut    "Connecticut"    "3"
#> Delaware       "Delaware"       "1"
#> Florida        "Florida"        "1"
#> Georgia        "Georgia"        "2"
#> Hawaii         "Hawaii"         "3"
#> Idaho          "Idaho"          "3"
#> Illinois       "Illinois"       "1"
#> Indiana        "Indiana"        "3"
#> Iowa           "Iowa"           "3"
#> Kansas         "Kansas"         "3"
#> Kentucky       "Kentucky"       "3"
#> Louisiana      "Louisiana"      "1"
#> Maine          "Maine"          "3"
#> Maryland       "Maryland"       "1"
#> Massachusetts  "Massachusetts"  "2"
#> Michigan       "Michigan"       "1"
#> Minnesota      "Minnesota"      "3"
#> Mississippi    "Mississippi"    "1"
#> Missouri       "Missouri"       "2"
#> Montana        "Montana"        "3"
#> Nebraska       "Nebraska"       "3"
#> Nevada         "Nevada"         "1"
#> New Hampshire  "New Hampshire"  "3"
#> New Jersey     "New Jersey"     "2"
#> New Mexico     "New Mexico"     "1"
#> New York       "New York"       "1"
#> North Carolina "North Carolina" "1"
#> North Dakota   "North Dakota"   "3"
#> Ohio           "Ohio"           "3"
#> Oklahoma       "Oklahoma"       "2"
#> Oregon         "Oregon"         "2"
#> Pennsylvania   "Pennsylvania"   "3"
#> Rhode Island   "Rhode Island"   "2"
#> South Carolina "South Carolina" "1"
#> South Dakota   "South Dakota"   "3"
#> Tennessee      "Tennessee"      "2"
#> Texas          "Texas"          "2"
#> Utah           "Utah"           "3"
#> Vermont        "Vermont"        "3"
#> Virginia       "Virginia"       "2"
#> Washington     "Washington"     "2"
#> West Virginia  "West Virginia"  "3"
#> Wisconsin      "Wisconsin"      "3"
#> Wyoming        "Wyoming"        "2"
```

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

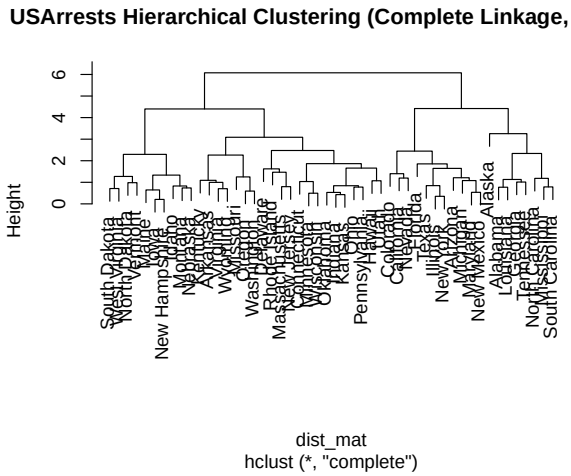
```r
data(USArrests)

# scale variables to have standard deviation one
scaled_data <- scale(USArrests)

# calculate Euclidean distance matrix
dist_mat <- dist(scaled_data)

# hierarchical clustering with complete linkage
hc_complete_scaled <- hclust(dist_mat, method = "complete")

# dendrogram
plot(hc_complete_scaled, main = "USArrests Hierarchical Clustering (Complete Linkage, Scaled)")
```



(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? Provide a justification for your answer.

Scaling the variables before computing inter-observation dissimilarities has the effect of giving equal weight to each variable in the clustering process. If variables are not scaled, variables with larger variances will dominate the clustering, and variables with smaller variances will be largely ignored. Scaling the variables can help to avoid biases in the clustering process.

Specific to the USArrests data, scaling the variables has the effect of making each variable directly comparable. Without scaling, variables such as Murder and Rape would have much larger variances than Assault and UrbanPop, and would therefore dominate the clustering process.

Scaling the variables helps to avoid biases in the clustering process and ensures that each variable is given equal weight in the clustering. If the variables are on different scales, it is difficult to make meaningful comparisons between them. However, if there is a strong reason not to scale the variables, such as domain knowledge or prior research suggesting that a specific variable should be given more weight, then it may be appropriate not to scale the variables.