DS-6030 Homework Module 10

Tom Lever

07/31/2023

DS 6030 | Spring 2023 | University of Virginia

8. In Section 12.2.3, a formula for calculating Proportion of Variance Explained (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 the following two ways. 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).

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

```
the_prcomp <- prcomp(USArrests, scale = TRUE)
variances_of_principal_components <- the_prcomp$sdev^2
vector_of_PVEs <-
    variances_of_principal_components / sum(variances_of_principal_components)
vector_of_PVEs
```

```
# [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.

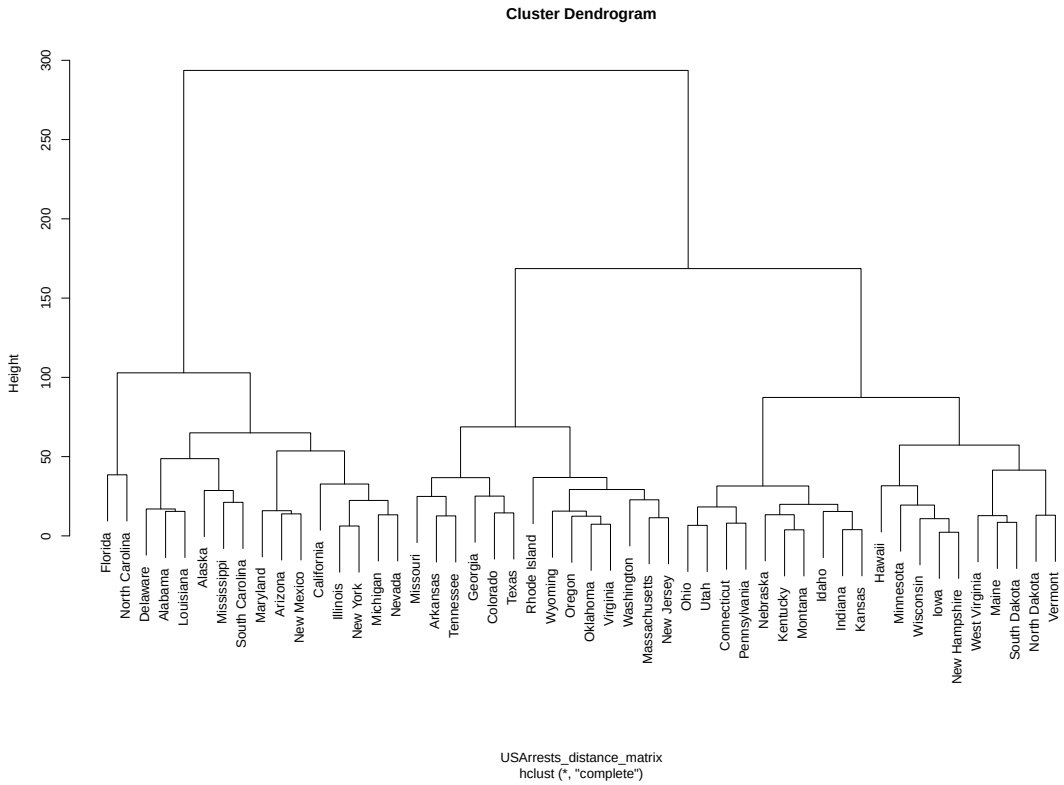
```
matrix_of_variable_loadings <- the_prcomp$rotation
centered_and_scaled_USArrests <- scale(USArrests)
centered_and_scaled_USArrests_matrix <-
    as.matrix(centered_and_scaled_USArrests)
squared_centered_and_scaled_USArrests_matrix <-
    centered_and_scaled_USArrests_matrix^2
sum_of_squared_centered_and_scaled_USArrests_matrix <-
    sum(squared_centered_and_scaled_USArrests_matrix)
product_of_centered_and_scaled_USArrests_matrix_and_matrix_of_variable_loadings <-
centered_and_scaled_USArrests_matrix %*% matrix_of_variable_loadings
product_of_centered_and_scaled_USArrests_matrix_and_matrix_of_variable_loadings_summed_on_colu
    apply(product_of_centered_and_scaled_USArrests_matrix_and_matrix_of_variable_loadings^2, 2
vector_of_PVEs <-
    product_of_centered_and_scaled_USArrests_matrix_and_matrix_of_variable_loadings_summed_on_
    sum_of_squared_centered_and_scaled_USArrests_matrix
vector_of_PVEs
```

```
#        PC1        PC2        PC3        PC4
# 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.

```
USArrests_distance_matrix <- dist(USArrests)
hclust_for_USArrests <- hclust(USArrests_distance_matrix, method = "complete")
plot(hclust_for_USArrests)
```



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

```
vector_with_group_memberships <- cutree(tree = hclust_for_USArrests, k = 3)
sorted_vector_with_group_memberships <- sort(vector_with_group_memberships)
sorted_vector_with_group_memberships
```

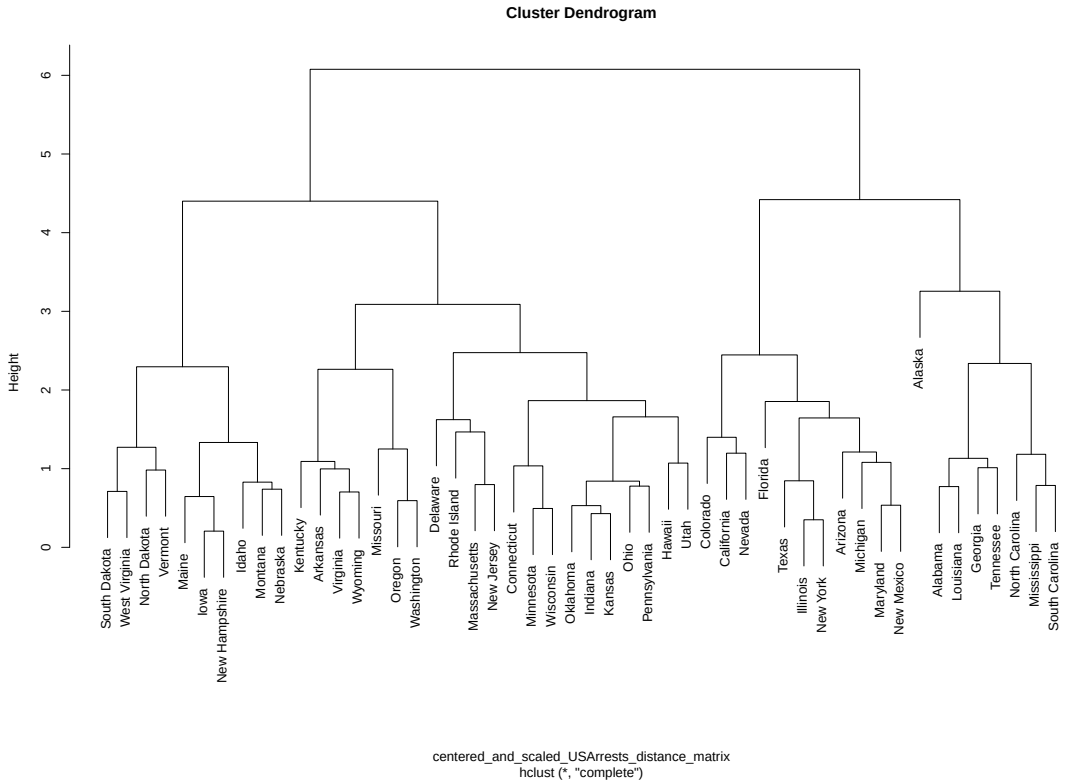
```
#        Alabama         Alaska        Arizona     California       Delaware
#              1              1              1              1              1
#        Florida       Illinois     Louisiana       Maryland       Michigan
#              1              1              1              1              1
#    Mississippi        Nevada     New Mexico       New York North Carolina
#              1              1              1              1              1
# South Carolina       Arkansas       Colorado        Georgia  Massachusetts
#              1              2              2              2              2
#       Missouri     New Jersey       Oklahoma         Oregon   Rhode Island
#              2              2              2              2              2
#      Tennessee          Texas       Virginia     Washington        Wyoming
#              2              2              2              2              2
#    Connecticut         Hawaii          Idaho        Indiana           Iowa
#              3              3              3              3              3
```

```
#          Kansas        Kentucky           Maine       Minnesota         Montana
#               3               3               3               3               3
#        Nebraska   New Hampshire    North Dakota            Ohio    Pennsylvania
#               3               3               3               3               3
#    South Dakota            Utah         Vermont   West Virginia       Wisconsin
#               3               3               3               3               3
```

As depicted above, in ascending order by group index, states Alabama through South Carolina belong to Group 1. States Arkansas through Wyoming belong to Group 2. States Connecticut through Wisconsin belong to Group 3.

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

```
centered_and_scaled_USArrests <- scale(USArrests)
centered_and_scaled_USArrests_distance_matrix <-
    dist(centered_and_scaled_USArrests)
hclust_for_centered_and_scaled_USArrests <-
    hclust(centered_and_scaled_USArrests_distance_matrix, method = "complete")
plot(hclust_for_centered_and_scaled_USArrests)
```



(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.

```
vector_with_group_memberships <- cutree(
    tree = hclust_for_centered_and_scaled_USArrests,
    k = 3
)
```

```
sorted_vector_with_group_memberships <- sort(vector_with_group_memberships)
sorted_vector_with_group_memberships
```

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

As depicted above, in ascending order by group index, states Alabama through Tennessee belong to Group 1. States Arizona through Texas belong to Group 2. States Arkansas through Wyoming belong to Group 3.

The states within each group are different. The numbers of states within in group are different. I believe the variables should be scaled before the inter-observation dissimilarities are computed.

According to When is centering and scaling needed before doing hierarchical clustering?, "For some types of well defined data, there may be no need to scale and center. A good example is geolocation data (longitudes and latitudes). If you were seeking to cluster towns, you wouldn't need to scale and center their locations.

For data that is of different physical measurements or units, its probably a good idea to scale and center. For example, when clustering vehicles, the data may contain attributes such as number of wheels, number of doors, miles per gallon, horsepower etc. In this case it may be a better idea to scale and center since you are unsure of the relationship between each attribute.

The intuition behind that is that since many clustering algorithms require some definition of distance, if you do not scale and center your data, you may give attributes which have larger magnitudes more importance."