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.

   (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? Provide a justification for your answer.