USE OF UNSUPERVISED LEARNING ON GENETIC SEQUENCES OF COVID

- RAMGOPAL REDDY PUTTA

ABSTRACT:

This report demonstrates the use of unsupervised learning methods like Principal components analysis and K-Means clustering techniques. The data used in the analysis contains four types of genetic sequences (1, 2, 3 and 4). Scientists are exploring what new variants of the COVID virus are being created depending on the genetic sequence. The data is not labeled to identify the properties of the virus or to classify the type of virus variant depending on the genetic sequences recorded in the data. Hence, two types of unsupervised learning techniques are implemented to classify the data depending on the identical properties the virus possesses. They are Principal components analysis and K-means clustering. Using principal component analysis, the whole data with a large number of variables (high dimensional data) is reduced to just very few components (low dimensional data) and the analysis can be carried out over them for data classification. On these principal components, K -mean clustering technique is implemented, and data is clustered into distinct groups such that the observations within each group have similar properties. In this analysis, a high dimensional data with approximately 29000 variables is reduced to a low dimensional data with just 6 components. K means clustering is performed over these components and data is clustered into groups, where each group is considered as the properties of a COVID variant type.

UNSUPERVISED LEARNING:

Unsupervised learning occurs when the data is not labeled. Which means, the data has not been assigned with some labels identifying the properties or classifications. This kind of learning is used to find hidden patterns in large unlabeled datasets through some techniques like Principal components analysis, Clustering etc. In this method, there is no simple objective such as prediction of a response. These techniques for unsupervised learning are of growing importance in various fields like Marketing sector (grouping the shoppers characterized by their previous purchases or browsing history), Health sector (making subgroups of breast cancer patients grouped by their measurements of gene expression), grouping movies based on ratings assigned by the viewers, etc.

COMMON UNSUPERVISED LEARNING APPROACHES:

Two of the most widely used techniques are implemented in this analysis. They are Principal Components Analysis and K-Means Clustering.

Principal components analysis (PCA):

This is a tool used for data pre-processing or visualization. PCA produces a low-dimensional representation of the data. For instance, consider a data set with a large number of variables and many of them are correlated. That can be considered as a quite an unmanageable set. PCA helps in recognizing a sequence of linear combinations of the variables or features that have maximum variance and are mutually uncorrelated, by retaining the contribution of all the variables.

Consider a set of variables X_1, X_2, X_3…X_P The first component PC_1 is the linear combination of these variables and has the highest variance. This can be represented in the form of the below equation:

$$PC_1 = \alpha_{11}X_1 + \alpha_{21}X_2 + \alpha_{31}X_3 + ... + \alpha_{P1}X_P$$

The principal component is defined by a set of weights $\alpha_{1,}$ $\alpha_{2,}$ α_3 …… α_p These weights are referred to as loadings of the first component PC_1 The loadings are constrained so that their sum of squares is equal to one. If they are not constrained, they can become much bigger, which in turn makes variance much higher. The loading vector (α_1) defines a direction in feature space along which the data has got the highest variance.

Now, when the second principal component is considered, which also has a large variance but with a natural constraint that the second variable is uncorrelated with the first variable so that it can tell different information about the data. This can be indexed by 2 and represented in the form of the below equation:

$$PC_2 = \alpha_{12}X_1 + \alpha_{22}X_2 + \alpha_{31}X_3 + ... + \alpha_{P2}X_P$$

Where α_2 is the second principal component loading vector, with elements $\alpha_{12,}$ α_{22} …. α_{P2} Here, constraining PC_2 to be uncorrelated to PC_1 is equivalent to constraining the direction α_2 to be orthogonal to the direction of α_1

To better understand the strength of each principal component, the proportion of variance explained (PVE) is determined for each component. The PVE can be observed by looking at the variance of individual components relative to the sum of variances. So, PVEs sum up to one.

K-Means clustering:

This is implemented for discovering the unknown subgroups in the data. The data is segregated into a pre-specified number of distinct groups such that the observations within each group are similar to each other.

The main difference between the PCA and K- means is that the PCA looks for variation in the data. Whereas clustering looks at similarity among the observations. A ‘good clustering’ can be defined as the one in which the variation within the cluster is as small as possible.

The number of clusters is randomly assigned initially for the observations. For each cluster, the centroid (the average value for each feature for all the points in the cluster) is computed and then each observation is assigned to the closest centroid.

METHODOLOGY:

The data used in the analysis contains four types of genetic sequences (1, 2, 3, 4) and all the other entries are set to ‘Not available values’. Since the data has lot of empty fields, ‘Matrix completion’ method is implemented to fill the empty fields with most likely values. This means that if a feature is missing for a given observation, other observations are considered, which are similar to it in order to fill in the missing information. Once all the empty fields of the dataset are filled with likely values, the columns with the same values (with zero variance) are removed from the data. Before implementing PCA or clustering, the variables must be maintained with the same units, must be centered to zero and scaled to have the standard deviation one. Otherwise, one variable with a different unit or with large numbers, that variable will dominate the principal components or clustering.

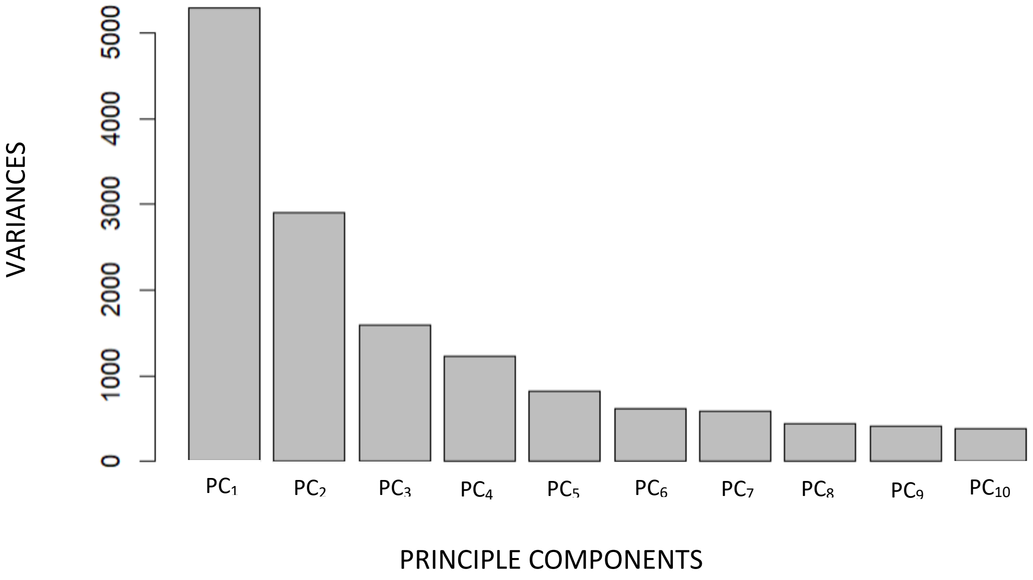
PCA analysis is now implemented over the data. The proportion of variance explained (PVE) is determined for each component to study the influential Principal components.

To this data, K-Means clustering is implemented and analyzed the data with respect to the subgroups formed. After randomly assigning the number of clusters, for each cluster, the centroid is calculated, and each data point is assigned to the cluster whose centroid is closest.

RESULTS AND DISCUSSION:

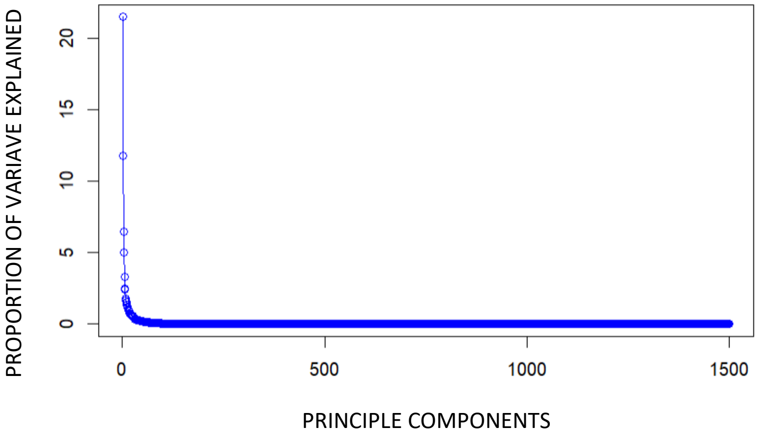
Principal components analysis (PCA):

PCA is implemented over the data and summary values (standard deviation, proportion of variances, cumulative proportion of variances) are obtained. For each principal component, standard deviation values are obtained, and it is found that the values gradually decrease from the first component. A bar plot is created to explain the variability of each component.

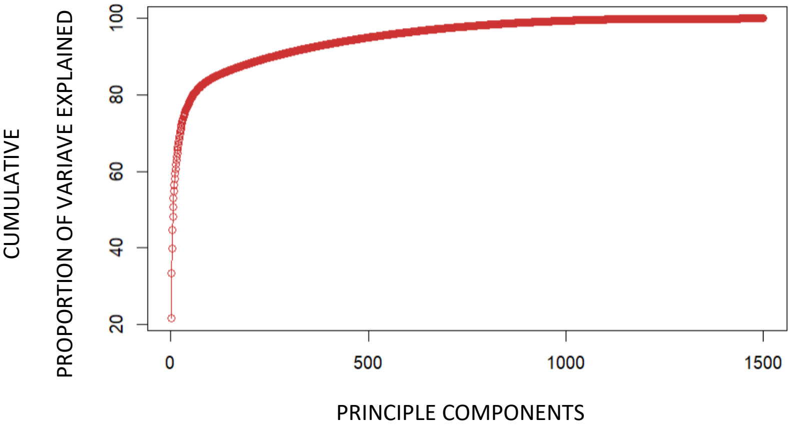


Plot 1. Bar graph to show principal components and their variability.

From the above bar plot, it is observed that the first principal component has got the highest variability and the variance decreased gradually for other principal components. The variance of the components, especially after sixth principal component (PC_6) are very low and can be excluded from consideration. Similarly, some plots were created to observe the proportion of variances in the data and the number of components to be considered in further analysis.



Plot 2. Proportion of variance explained by each of the principal component



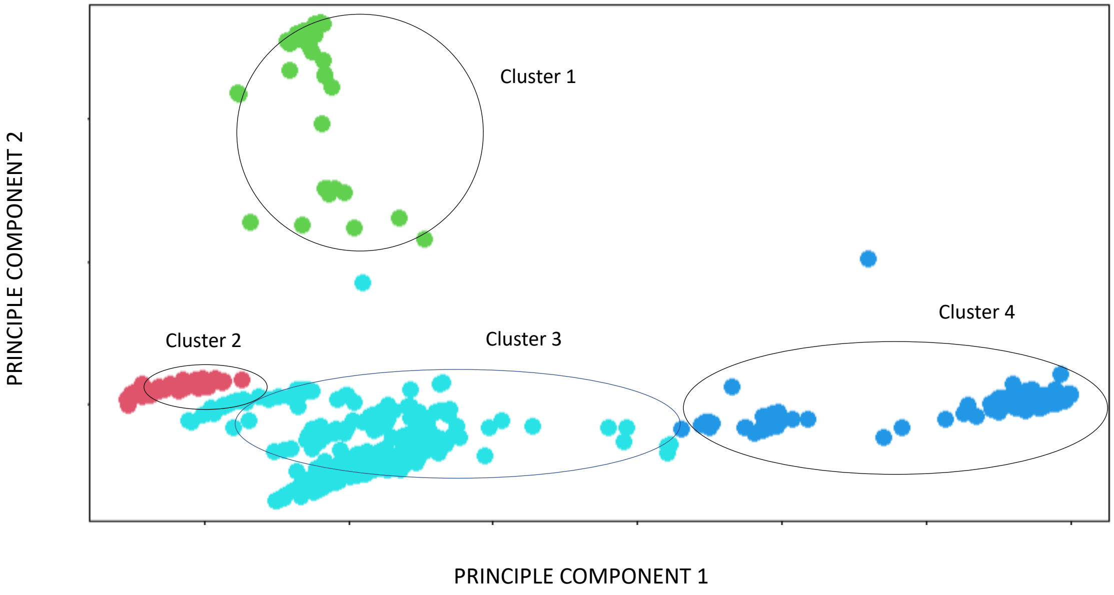
Plot 3. Cumulative Proportion of variance explained by each of the principal component

From plot 3, it is observed that the first '6' principal components explain around 50 percentage of the variance in the data. The same can be observed from plot 2 where each of the first 6 components explain a substantial amount of variance, and then there is an elbow in the plot which infers that there is a marked decrease in the variance explained by further components. Hence, these plots suggest that there will be little benefit in considering or examining more than 6 principal components.

Hence, the whole data with approximately 29000 variables (high dimensional data) is reduced to just 6 components (low dimensional data) and the analysis is performed to classify the data.

K-Means clustering:

The color of each observation indicated the cluster to which it belongs. Here, there is no ordering of the clusters. So, the coloring is arbitrary. Since the data has a large number of variables, the principal components are considered in clustering and plotted for the first two components.



Plot 4. K-Means clustering results with first two principal components

The data is grouped into four clusters, namely cluster 1, cluster 2, cluster 3 and cluster 4 with similar properties, and each cluster and can be considered as the properties of a specific type of COVID virus variant.

CONCLUSION:

Principal components analysis and K-means clustering techniques are applied over the unlabeled data and it is observed that using Principal components analysis, the high dimensional data (with 29000 variables) is brought down to a low dimensional data (with just 6 components). This number of principal components to be considered is obtained from the plots which were created by observing the proportion of variances in the data. After obtaining the influential principal components, K means clustering is performed over these components and data is clustered into four groups, where each group contains some distinct properties and is considered as a type of COVID variant.

APPENDIX (Source code):

Libraries imported:

```
library(ISLR2)
library(ggplot2)
library(dplyr)
library(RSpectra)
library(softImpute)

# Importing the data:
df_covid <- read.csv('C:/D_drive/DATA 5322/HW/HW_4/covid_vals.csv', header = TRUE, sep = ',')

# Imputing the data to fill the NA values:
Xtrue <- covid_1 %>% select(-date) %>% as.matrix()
corrupt <- Xtrue
set.seed(1)
data_1 <- softImpute(corrupt, rank=2, lambda=100, trace.it = TRUE)
d_frame  <- complete(corrupt,data_1)

# Converting matrix form to data frame
dataframe_data <- as.data.frame(d_frame)

dim (dataframe_data)

# Applying PCA:
pca.out <- prcomp(dataframe_data, scale = FALSE)
pca.out
names(pca.out)
summary(pca.out)
# Plotting the variance explained
plot (pca.out)
```

```
pve <- 100 * pca.out$sdev^2 / sum (pca.out$sdev^2)
pve


# Scree plots:
par(mfrow = c(1, 2))
plot (pve , type = "o", ylab = "PVE",
xlab = "Principal Component", col = "blue")

plot ( cumsum (pve), type = "o", ylab = "Cumulative PVE",
xlab = "Principal Component", col = "brown3")

# Kmeans clustering: (3 clusters)
km.out <- kmeans(pca.out$x, 3, nstart = 50)
km.out$cluster
par (mfrow = c(1, 1))

plot (pca.out$x[, 1:2], col = (km.out$cluster + 1), main = "K- Means Clustering",
xlab = "PC1", ylab = "PC2", pch = 20, cex = 2)


# Kmeans clustering (4 clusters):

km.out <- kmeans(pca.out$x, 4, nstart = 50)
km.out$cluster
par (mfrow = c(1, 1))
plot (pca.out$x[, 1:2], col = (km.out$cluster + 1), main = "K- Means Clustering",  xlab = "PC1", ylab =
"PC2", pch = 20, cex = 2)
```

REFERENCES:

[1] ISLRv2_website.pdf (su.domains)

(Textbook: An Introduction to Statistical Learning with Applications in R Second Edition)