PCA Literature Review

When faced with a large set of correlated variables, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the set. Principal component analysis (PCA) refers to the process by which principal components are computed and the subsequent use of these components in understanding the data [1]. PCA is an unsupervised learning method and is also used a tool for data visualization. PCA finds a low dimensional representation of a data that contains the most possible amount of variance of the dataset. Each of the dimensions found by PCA is a linear combination of the p features present in the dataset.

The first principal component of a set of features is the normalized linear combination of the features

(1)

Which has the largest variance. By normalized, we mean that . We refer to the elements as the loadings of the first principal components. Together, the loadings make up the principal component loading vector.

To compute the first principal component, we must solve the optimization problem,

Maximize subject to (2)

The objective that we are maximizing in (2) is just the sample variance of the n values of We refer to as the scores of the first principal component. The loading vector defines the direction in the feature space along which the data varies the most. If we were to project the n data points onto this direction, the projected values are the principal component scores themselves.

The second principal is the linear combination of that has maximal variance out of all linear combinations that are uncorrelated with . Putting this constraint on the second principal component is equivalent to constraining its direction to be orthogonal to the direction of the first principal component.

Diagram

Description automatically generated

Fig. 1 [2]

Based on fig.1, the pseudocode to calculate principal components is as follows:

1. Create the data matrix from the given dataset.
2. Subtract from mean value to center the data to have mean = 0.
3. Calculate its correlation/covariance matrix.
4. Compute the eigenvectors and eigenvalues of the covariance matrix.
5. Based on understanding of the domain of the problem and the data, find new attributes to work with
6. Use the new data set for further analysis.

Hierarchical Clustering Literature Review

Clustering refers to techniques for finding subgroups, or clusters in a dataset. We seek to find groups such that the observations within each group are similar to each other and are different from observations in other groups.

In hierarchical clustering, we end up with a tree like visual representation of the observations, called a dendrogram, that allows us to view at once the clustering obtained for each possible number of clusters, from 1 to n.

In hierarchical clustering we look into dissimilarity between clusters and we use amongst 4 different kind of dissimilarity linkages to compute it:

1. Min-Linkage – chooses minimum distance between data points to be dissimilarity between 2 clusters
2. Max-Linkage – chooses maximum distance between data points to be dissimilarity between 2 clusters
3. Average Linkage – chooses average of all the distances between data points to be the dissimilarity between 2 clusters.
4. Ward Linkage – minimizes the total in-cluster variance.

Diagram

Description automatically generated

Fig. 2

Based on fig 2, the hierarchical clustering algorithm is as follows:

1. Start by considering each point as its own cluster
2. Choose appropriate dissimilarity measure and create a distance matrix.
3. Put the two points with the minimum dissimilarity into the same cluster.
4. Choose appropriate linkage and create an updated distance matrix.
5. Repeat steps 3 and 4 until all points have been assigned a cluster and a dendrogram is created.
6. Use gap statistic to find optimal number of clusters and cut dendrogram at that point.

Result of PCA

On implementing principal component analysis on this dataset, the loading vectors formed were as given in fig. 11

Text

Description automatically generated

Fig. 11

Based on the results, we interpreted the components as below:

1. PC1 placed most load on child mortality, life expectancy and total fertility. So, we interpreted it as giving information on Birth Rate/Life Expectancy.
2. PC2 placed most load on imports and exports and we interpreted it as Trade.
3. PC3 placed most load on health and inflation and since both were related to inflation, we thought of it as inflation.
4. Finally, PC4 placed most load and income and GDP, and we interpreted it as telling us about the economy of a nation.

The choice of number of principal components to perform further clustering analysis on was based on the proportion of variance explained. 3 PC explain 76% of the total variance in the data. 4 and 5 PC’s explain 87% and 94% of the total variance. To find the best one of the three options, we implemented K-Means and Hierarchical clustering using all the three cases and compared them using silhouette score and the geospatial plots.

Results of both

Chart, line chart

Description automatically generated Chart

Description automatically generated

Fig. 3 Fig. 4

On performing hierarchical clustering with the complete dataset, the optimal number of clusters were found to be 3 using the gap statistic, fig. 3. This was the chosen metric to find the optimal number of clusters for hierarchical clustering. For the same number of clusters, we calculated its silhouette score and found that to be 0.25.

Also, to make sense of the cluster, we computed averages of each attribute for each cluster and the result can be seen in fig. 5.

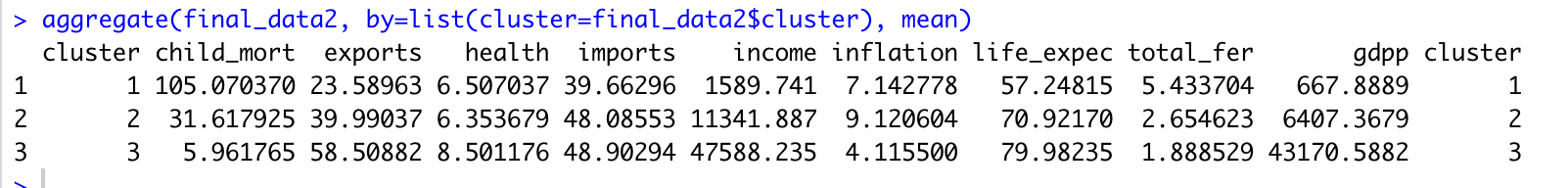


Fig. 5

Based on these averages, we concluded that countries in immediate need of help would belong to cluster 1 and the countries which were at low risk and do not need money were part of cluster 3. Among other averages, the average of child mortality for cluster 1 was 105 and was the highest while it had the lowest average for GDP. It also had high inflation and low life expectancy. Based on the geospatial plot in fig.6 we see that the clustering makes sense based on real world understanding. Western countries in yellow have been found to be at low or no risk and hence do not need any help. The Asian and Middle Eastern countries in green were found to be relatively stable and do not need help. Finally, the countries in blue were found to be at high risk and hence needed urgent help in the form of economic relief. These were found to be some African countries Afghanistan.

Map

Description automatically generated

Fig. 6

After performing PCA, and getting the new attributes, we performed hierarchical clustering on the principal components. To find the best number of principal components, we performed clustering on 3, 4 and 5 PC’s. The criterions were the silhouette score and the validity of the geospatial obtained based on real world understanding.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **3 PC** | **4 PC** | **5 PC** |
| % of variance explained | 76.13 | 87.2 | 94.53 |
| # of clusters | 3 | 4 | 3 |
| Gap Statistic | 0.749 | 0.774 | 0.786 |
| Silhouette Score | 0.24 | 0.32 | 0.23 |

Fig. 7

The number of clusters in each case was 3,4 and 3 respectively for 3, 4 and 5 PC. While the gap statistic score was highest for 3 clusters of the 5 PC case, we used silhouette score to compare between the three models and found that the 4 clusters 0 4 PC gave a relatively high silhouette score and decided to use that case moving forward.

Map

Description automatically generatedMap

Description automatically generatedMap

Description automatically generated

Fig.8 Fig.9 Fig.10

Further on inspecting the accompanying geospatial plots, we can notice differences in the clustering of the three different PC’s. Most differences were observed in the clustering of South American, East Asian and some African countries. For e.g. Argentina was placed in the high-risk cluster with 5 PC whereas it was placed in the medium risk cluster with other two cases. Similarly, Brazil was seen to have no risk with 3 PC whereas it was placed in the medium risk cluster with both the other implementations. Again, based on some real-world understanding, we chose the 4 clusters given by 4 principal components.

Results comparing K means and Hierarchical

Finally, we compare the results of both K-Means clustering and Hierarchical clustering with the complete data and after performing PCA as well.

|  | K Means (whole data) | K Means (4 PC) | Hierarchical (Whole data) | Hierarchical (4 PC) |
| --- | --- | --- | --- | --- |
| # Clusters | 3 | 3 | 3 | 4 |
| Silhouette Score | 0.28 | 0.33 | 0.25 | 0.32 |

We found better clusters after performing PCA on the dataset. The silhouette score for both K Means and Hierarchical using the complete dataset was found to be 0.28 and 0.25 respectively. The score for both after PCA was 0.33 for K Means and 0.32 for Hierarchical.

Map

Description automatically generatedMap

Description automatically generated

Fig.11 Fig. 12

On comparing implementations using PCA using geospatial plots, we concluded that both the clustering algorithms created similar clusters. Both placed Western countries along with Australia and Japan in the low-risk cluster, Most of the Asian and South American countries were placed in the medium risk cluster and African and East Asian countries were placed in the high-risk cluster needing urgent help in the form of economic relief.

Conclusion

In this project, K Means and agglomerative Hierarchical clustering was performed on a dataset containing socio-economic features of various countries to find groups of countries in need of economic aid. After performing the clustering methods on the complete dataset, Principal Component Analysis was performed on the data to find a low dimensional representation of the features. Finally, clustering was performed on the principal components and results were compared using silhouette scores and geospatial plots of country clusters formed. Both the algorithms clustered countries in a similar manner that made sense based on real world understanding of the socio-economic situation.

An important part to highlight in this project is the complete project

References

[1] James, Witten, Hastie and Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed,

[2] Tabassum and Afrin, “Comparative performance of using pca with k-means and fuzzy c means clustering for customer segmentation”, 2015

[3]