Analysis of the UN Economic Commission for Europe Dataset

Khang Ee Pang

University College Dublin

January 6, 2020

1 Introduction

The dataset of interest is sourced from the United Nations Economic Commission for Europe database, containing 14 variables:

Index	Variable
1	Population density, pers. per sq. km
2	Total population, male (%)
3	Total population, female (%)
4	Life expectancy at birth, women
5	Life expectancy at birth, men
6	Mean age of women at birth of first child
7	Computer use, 16-24, male
8	Computer use, 16-24, female
9	Unemployment rate
10	Youth unemployment rate
11	Exchange rate (XR), NCU per US$
12	GDP in agriculture (ISIC4 A): output approach, index, 2010=100
13	Persons killed in road accidents
14	Total length of motorways (km)

of 49 countries from the year 2013 to 2018. In this work, we explore the data using different dimension-reduction methods and its effect on the k-means clustering algorithm. We choose to focus on the year 2017.

2 Pre-processing

To address the missing values within the dataset, we choose to fill in the data in two stages. In the first stage, we assume for every country, the variables do not change rapidly across the years, and null values are replaced by the mean across years. For the values that remain missing, it is replaced by sampling randomly across other countries, the motivation is that the resulting distribution will closely mimic the starting distribution. Finally, the variables are standardized by dividing by its standard deviation, we label the standardized variable i by X_i.

3 PCA

A standard method for dimension-reduction is the principal component analysis (PCA). PCA fits a high-dimensional ellipsoid to the data and identifies the orientation of said ellipsoid. Because PCA only takes into account the second moments of the data, this method works best when all the variables are symmetric and unimodal, which in our case are not. After the orientation of the ellipsoid is identified, the data is projected onto the space spanned by the major principal axes of the ellipsoid.

Figure 1(a) shows the marginal distributions of the first three variables. This provides useful descriptions of the data between pairs of variables. However, if we are interested in the global structure of the data, we look at the data in the principal frame. Figure 1(b) shows the first three components of the data in the principal basis. The three distributions over the principal components provide a better global description of the data compared to the distribution over the original variables. Figure 1(c) shows the relative magnitude of the variance with respect to the corresponding principal component, around 60% of the variation of the data is contained in the first three principal components. Figure 1(d) shows the (unsigned) transformation matrix.

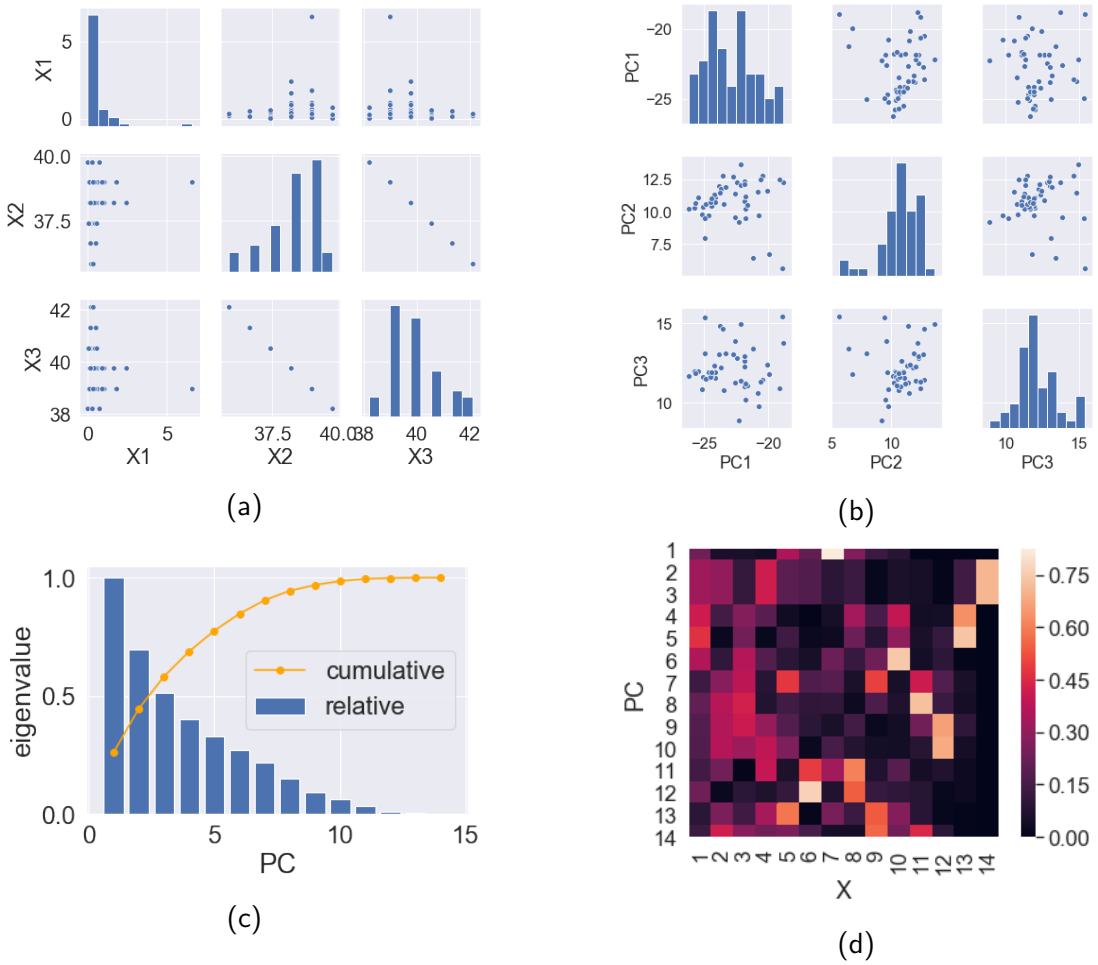


Figure 1: Marginal distribution of the first three (standardized) variables (a) compared to the transformed variables (b). (c) Scree plot displaying the relative magnitude of the corresponding eigenvalue of each principal component in blue and cumulative variance proportion in orange. (d) Contribution of each variable to the principal component.

4 UMAP

In this section, we investigate the uniform manifold approximation and projection (UMAP) method for dimension reduction. UMAP relies on the k-nearest neighbour algorithm to construct a weighted graph referred to as a fuzzy simplicial set, the fuzzy simplicial set is then represented in lower dimension using a graph layout algorithm [1]. UMAP does not assume the geometry of the data, hence is more robust compared to PCA. However, it is important to keep in mind that, unlike PCA, the embedding space of UMAP have no specific meaning.

Figure 2 shows a fuzzy simplicial set constructed with UMAP with $k = 7$ nearest neighbour. The embedding is optimized with respect to the cross entropy of the graph (a different embedding of the graph using a force-directed graph layout algorithm is included in the Appendix). The parameter k can be thought of as the degree of smoothing on the manifold, smaller k value better preserves the feature of the manifold but prone to noise. Motivated by the next section, we choose the largest k such that the resulting manifold has two connected components.

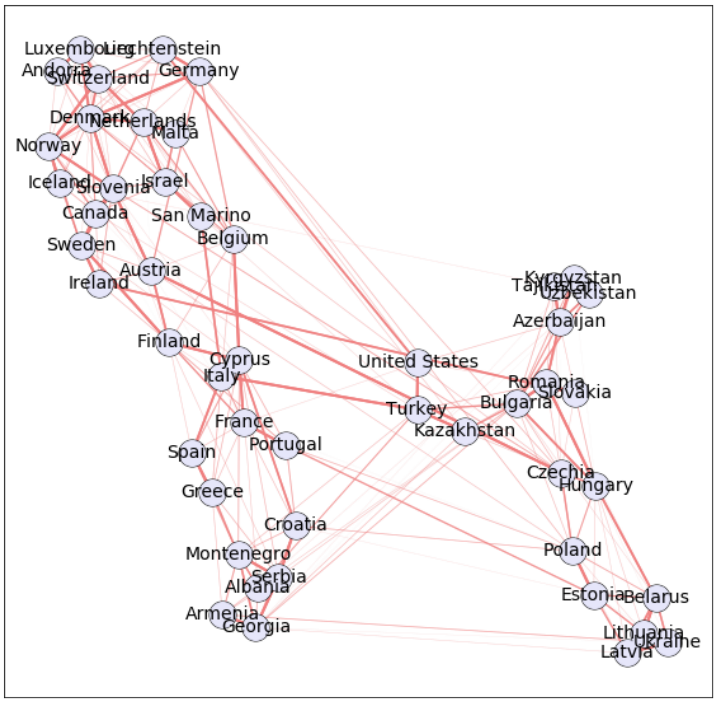


Figure 2: Graph of the fuzzy simplicial set generated by UMAP. The width of the edges corresponds to the membership strength between two nodes. The nodes are layout by minimizing the cross entropy of the graph. Here we use $k = 7$ with the Euclidean metric.

5 Clustering

We cluster the high-dimensional data into two groups using k-means, the result is shown in Figure 3(a) and (b) projected using PCA and UMAP respectively. As expected, k-means splits the data across the first principal component as the data are spread out the most along the first principal axis. We repeat the clustering using only the two-dimensional embeddings of the data obtained from PCA and UMAP, results are shown in Figure 3(c) and (d) respectively. Perhaps it is not surprising that the PCA embedding preserves the grouping, we expect the

angle between the hyperplane dividing the data and the plane spanned by the first two principal axes to be close to a right angle. It is however, surprising that UMAP embedding preserves the clustering since graph layout algorithms do not necessarily preserve the relative positions of the data.

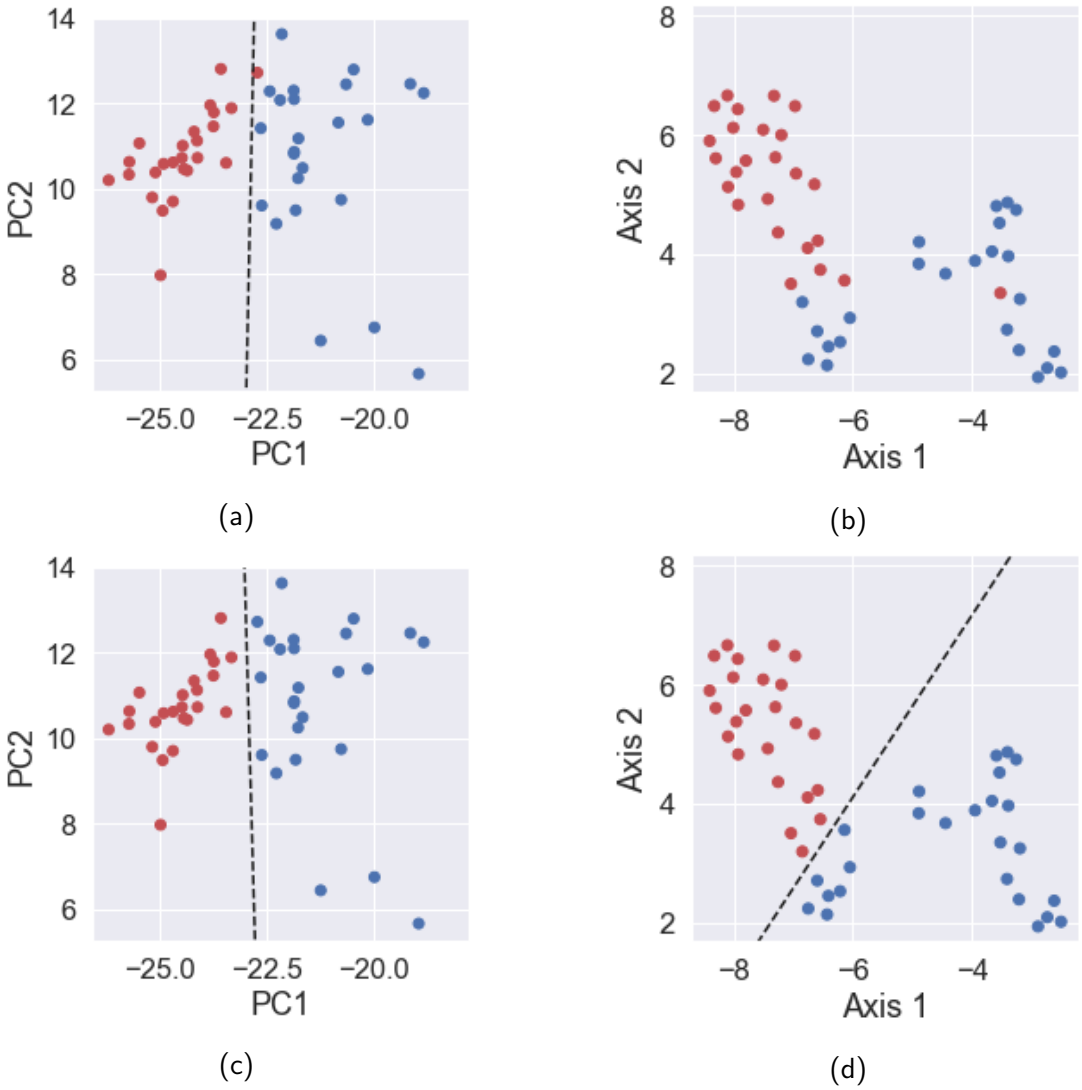
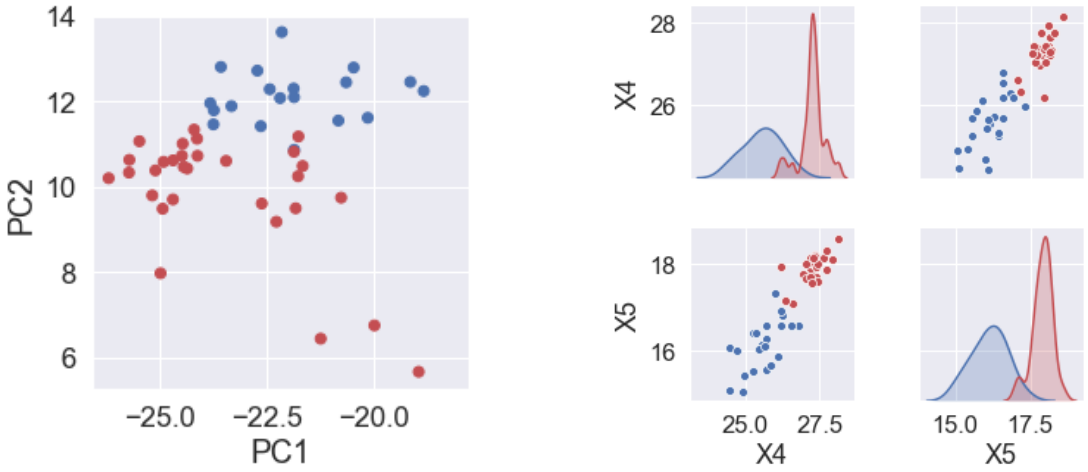


Figure 3: (a) PCA visualization of the clustering using high-dimensional data. (b) The same grouping as (a) but visualized using UMAP. (c) Clustering using only the first two principal components of the data. (d) Clustering using the two-dimensional UMAP embedding of the data. Dashed lines represent the decision boundary of the k-means clustering.

We found that removing X_4 and X_5 from the high dimensional data changes the clustering the most, the resulting clusters after the removal are shown in Figure 4(a). From Figure 4(b), we see why this is the case: the variable X_4 and X_5 both have a bimodal distribution, thus have a strong influence on the result of clustering. Removing the other 12 variables has minimal effect on the clustering, confirming our claim.



(a) (b)

Figure 4: (a) Result of clustering after removing X_4 and X_5 from the data. (b) Distribution of X_4 and X_5.

6 Conclusion

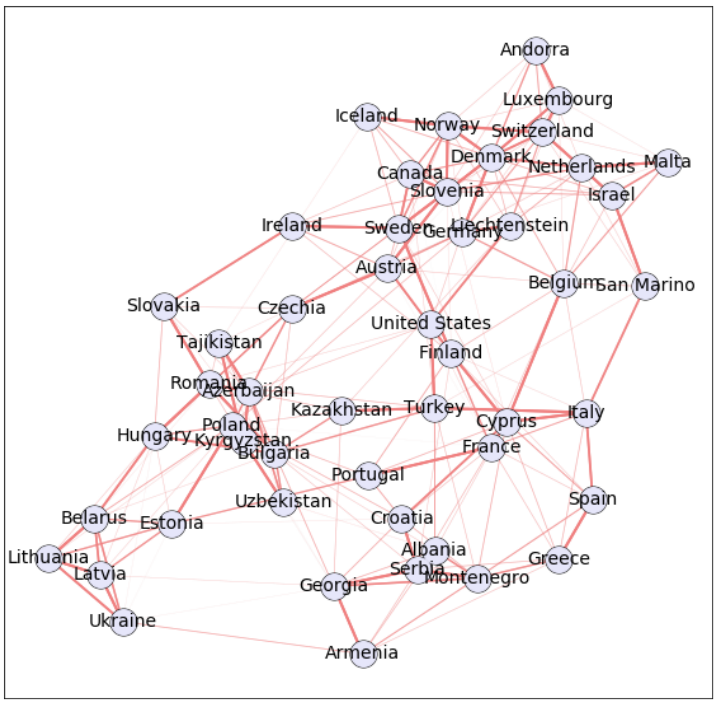
In general, UMAP is a better choice for data visualization compared to PCA when dealing with any non-multivariate-normally distributed dataset. However, because UMAP embedding does not transform distances linearly, UMAP may not be the optimal choice of dimension reduction technique for distance-based clustering methods such as k-means. In contrast, methods such as the density-based DBSCAN has high compatibility with UMAP. Using k-means, we saw that the countries are divided in to two groups according to the life expectency. To better investigate the difference in the effectiveness of PCA and UMAP as a visualization method a larger, labeled dataset is much preferable.

References

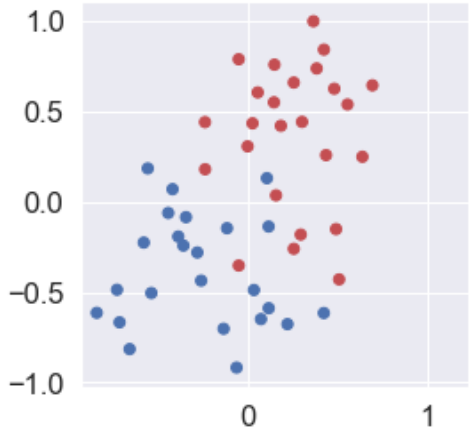
[1] L. McInnes, J. Healy, J. Melville. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Appendix

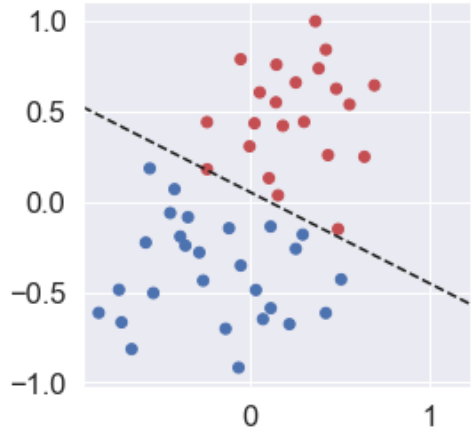
Since the author of [1] described two methods for graph embedding, we thought it may be interesting to compare the two. Figure 5(a) shows the same fuzzy simplicial set discussed in section 4 but embedded using the Fruchterman-Reingold force-directed graph layout algorithm. In Figure 5(b), we see a less distinct boundary between two clusters compared to the cross entropy minimization approach. This is due to the Fruchterman-Reingold algorithm does not preserve the relative orientations of the weakly linked nodes. A comparison with the k-means clustering using only the two-dimensional embedding of the data is shown in Figure 5(c).



(a)



(b)



(c)

Figure 5: (a) Graph of the fuzzy simplicial set embedded using a force-directed layout algorithm. The width of the edges corresponds to the membership strength between two nodes. $k = 7$, Euclidean metric. (b) k-means clustering using the high-dimensional data. (c) k-means clustering using only the two-dimensional embedding of the data. Dashed lines represent the decision boundary of the k-means clustering.