Dimensionality Reduction & ClusteringTechniques

Aya Ben Hriz

November 22, 2021

1 Introduction

In machine learning, to catch useful indicators and obtain a more accurate result, we tend to add as many features as possible at first. However, after a certain point, the performance of the model will decrease with the increasing number of elements. This phenomenon is often referred to as "The Curse of Dimensionality." So how could we overcome the curse of dimensionality and avoid overfitting especially when we have many features and comparatively few training samples? One popular approach is dimensionality reduction.

We will talk also about clustering, the method of identifying similar groups of data in a dataset. Clustering generally depends on some sort of distance measure. But in high dimensional spaces, distance measures do not work very well. You reduce the number of dimensions first so that your distance metric will make sense.

We will use dimensionality reduction and clustering techniques to explore an artificial data set containing data on user profiles and data on conversations from an imaginary dating app.

2 Exploring our data set

Our dataset consists of 3000 observations and of 16 variables.

	userid <int>	date.crea <chr>	score <dbl>	n.matches <int>	n.updates.photo <int>	n.photos <int>	last.connex <chr>
1	1	2011-09-17	1.495834	11	5	6	2011-10-07
2	2	2017-01-17	8.946863	56	2	6	2017-01-31
3	3	2019-05-14	2.496199	13	3	4	2019-06-17
4	4	2015-11-27	2.823579	32	5	2	2016-01-15
5	5	2014-11-28	2.117433	21	1	4	2015-01-15
6	6	2017-06-05	1.700014	14	2	6	2017-07-03

3 Identifying correlation in the variables

3.1 Correlation Matrix:

Correlations between variables play an important role in a descriptive analysis. A correlation measures the relationship between two variables, that is, how they are linked to each other.

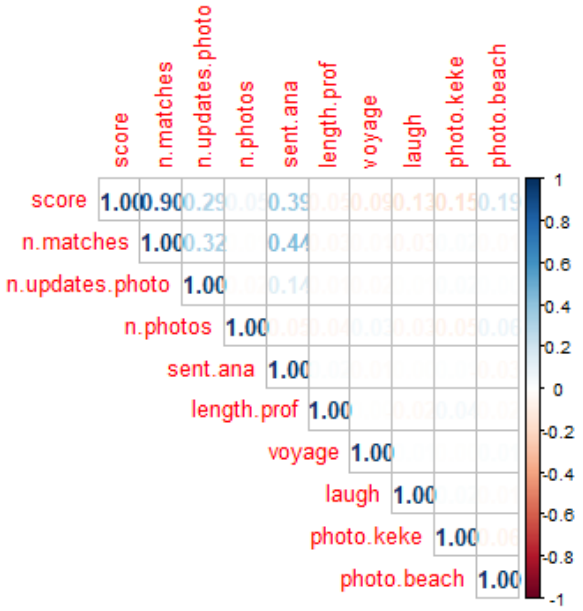

Figure 1: Correlation matrix

First of all, correlation ranges from -1 to 1. It gives us an indication of two things:

1- The direction of the relationship between the 2 variables.

2- The strength of the relationship between the 2 variables.

Regarding the direction of the relationship: On the one hand, a negative correlation implies that the two variables under consideration vary in opposite directions, that is, if a variable increases the other decreases and vice versa. On the other hand, a positive correlation implies that the two variables under consideration vary in the same direction, i.e., if a variable increases the other one increases and if one decreases the other one decreases as well.

Regarding the strength of the relationship: The more extreme the correlation coefficient (the closer to -1 or 1), the stronger the relationship. This also means that a correlation close to 0 indicates that the two variables are independent, that is, as one variable increases, there is no tendency in the other variable to either decrease or increase.

As an illustration, in Figure 1 the Pearson correlation between photo.keke and score found above is -0.15, meaning that the 2 variables vary in the opposite direction. On the contrary, from the correlation matrix, we see that the correlation between n.matches and score is 0.90 meaning that people with more matches tend to have higher scores. This again make sense.

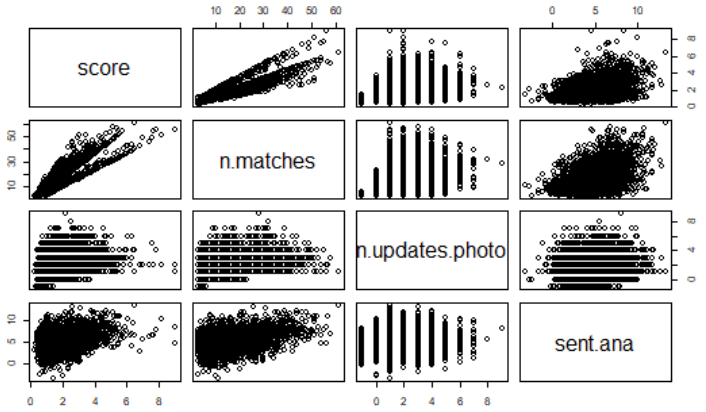

Figure 2: Scatterplot for several pairs of variables

Figure 2 indicates that score is positively correlated with n.matches, n.updates.photo and sent.ana.

3.2 Checking normality

Our next step now is checking if our distribution is normal or not to use the right correlation test. We make use of ggqqplot over score and n.matches to do that. The theoretical quantile-quantile plot is a tool to explore how a batch of numbers deviates from a theoretical distribution and to visually assess whether the difference is significant for the purpose of the analysis
From Figures 3 and 4 we can conclude that it's not a normal distribution since the figures don't reflect a pattern we would expect from a normally distributed batch of values. It is not a straight line so we use Spearman's test.

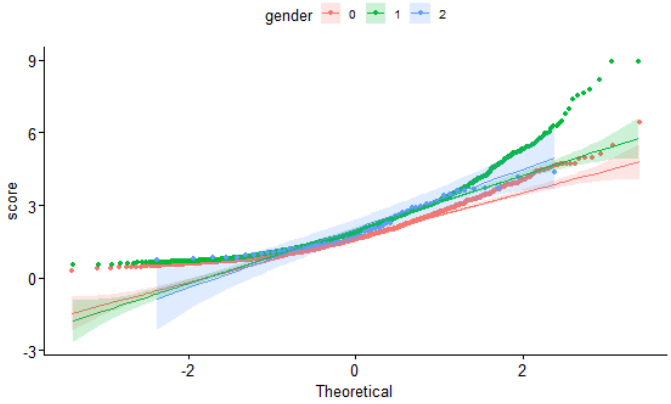

Figure 3: Theoretical QQ plot of score.

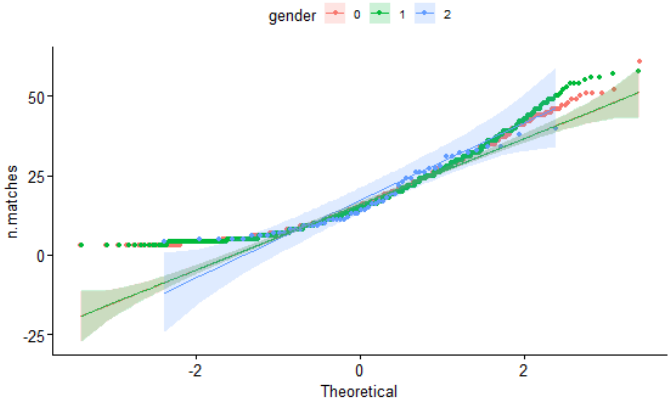

Figure 4: Theoretical QQ plot of n.matches

3.3 Spearman's test

As a reminder, Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. And its interpretation is similar to that of Pearsons, e.g. the closer is to the stronger the monotonic relationship. Here, rho is the Spearman's correlation coefficient, so the correlation coefficient between score and n.matches in Figure 5, is 0.9248536 and the p-value is less than 2.2e-16. Our correlation coefficient is close to 1 and 1 indicates a strong positive correlation : this means that score increases with n.matches.

4 Dimensionality Reduction

4.1 PCA

4.1.1 What is PCA and how is it used?

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed. Here are the setps involved in the PCA to have a better understaing of the algorithm.

Step 1: Standardize the dataset.

Step 2: Calculate the covariance matrix for the features in the dataset.

Step 3: Calculate the eigenvalues and eigenvectors for the covariance matrix.

Step 4: Sort eigenvalues and their corresponding eigenvectors.

Step 5: Pick k eigenvalues and form a matrix of eigenvectors.

Step 6: Transform the original matrix.

4.1.2 PCA - Individual map

In Figure 6, the first principal component explains 58.6% of the variability in the data while the second one explains 21.9% of it. The labeled individuals are those with the higher contribution to the plane construction. Further the observation is far from the origin for the x axis the highest the contribution to the first principal component. On the other side, further the observation is far from the origin for the y axis the highest the contribution to the second principal component.

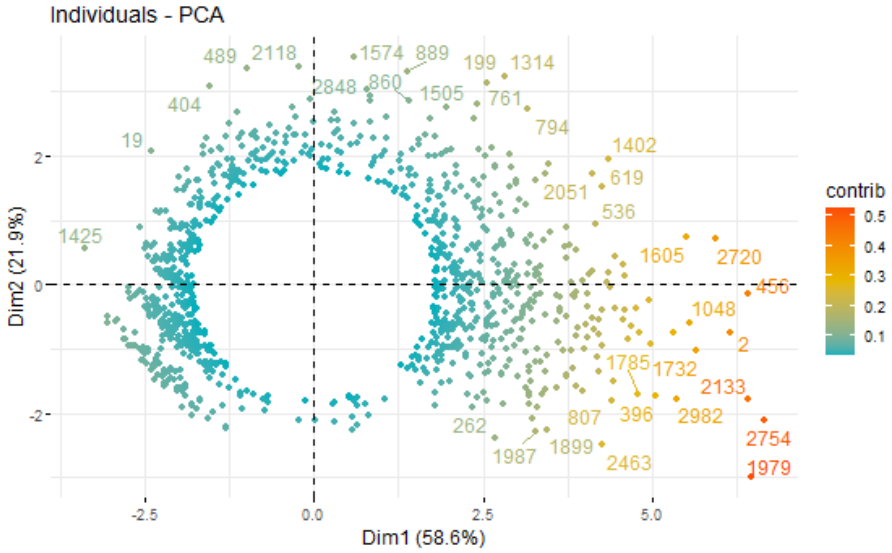

Figure 5: Individual map with a sample of individuals.

4.1.3 PCA - Biplot

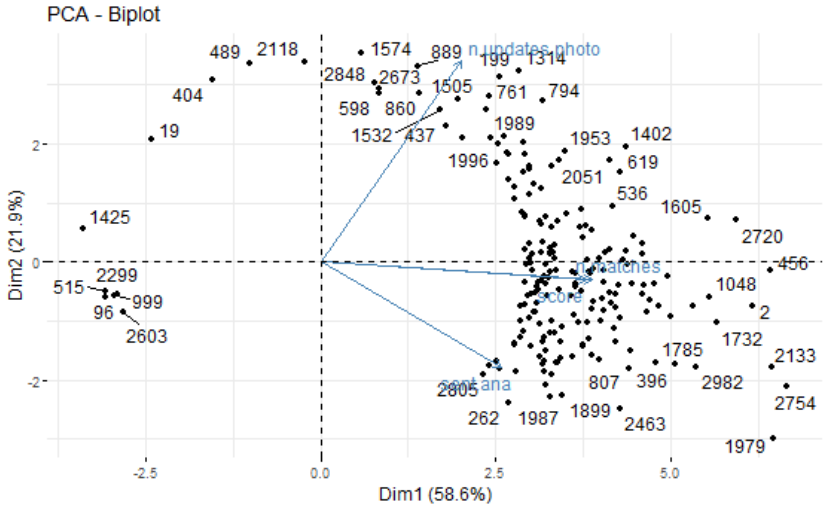

Figure 6: Individual map with a sample of individuals.

In Figure 7, all variables sent.ana, score and n.matches have positive values in the PC1 axis, while n.updates.photo is positive in PC2's and sent.ana,

score and n.matches are negative.
Since all the variables are positive in PC1, those which constrain the system the most are n.updates.photo and (then) n.matches and score (in PC1 axis). In PC2 axis, the two set of variables have opposite effects on the system.

4.1.4 Loadings of each Principal Component

Importance of components	PC1	PC2	PC3	PC4
Standard deviation	1.5307	0.9358	0.8280	0.30937
Proportion of variance	0.5857	0.2189	0.1714	0.02393
Cumulative proportion	0.5857	0.8047	0.9761	1.00000

Figure 7: Loadings of each Principal Component.

We can use the proportion to determine which principal components explain most of the variability in the data. The higher the proportion, the more variability that the principal component explains. The size of the proportion can help us decide whether the principal component is important enough to retain. In Figure 8, PC1 with a proportion of 0.5857 explains 58.5% of the variability in the data. Therefore, this component is important to include. PC4 has a proportion of 0.02393, and thus explains only 0.2% of the variability in the data. This component may not be important enough to include. We notice from Figure 8 that PC1 along with PC2 contribute to explaining 80% of the variability in the data.

4.2 MCA

4.2.1 What is MCA and how is it used?

The Multiple correspondence analysis (MCA) is for summarizing and visualizing a data table containing more than two categorical variables. It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative.

4.2.2 MCA - Scree Plot

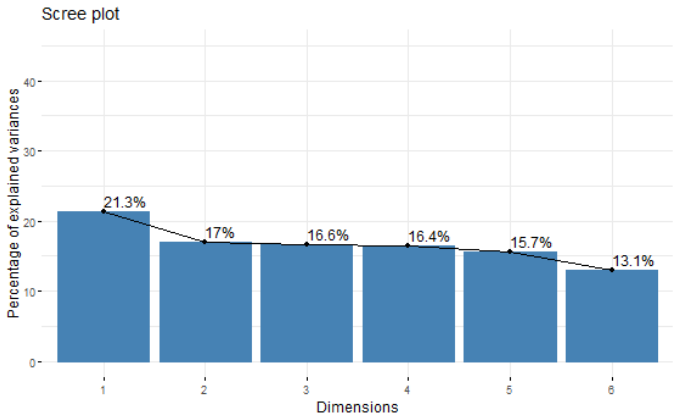

Figure 8: MCA Scree Plot.

In Figure 9, the scree plot shows that the eigenvalues start to form a straight line after the first principal component. Therefore, the remaining principal components account for a very small proportion of the variability and are probably unimportant. he two dimensions 1 and 2 are sufficient to retain 28.3% of the total inertia (variation) contained in the data.

4.2.3 MCA - Variable categories

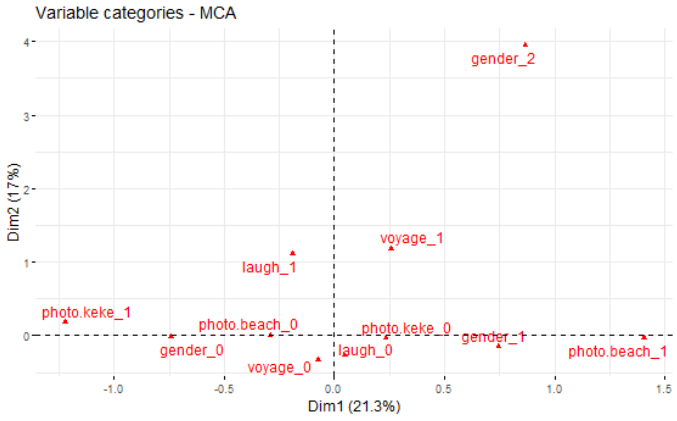

Figure 9: MCA - Variable categories.

The plot above gives an idea of what pole of the dimensions the categories are actually contributing to. It is evident that the categories photo.beach_1 and gender_1 have an important contribution to the positive pole of the first dimension, while the categories photo.keke_1 and gender_0 have a major contribution to the negative pole of the first dimension.

4.2.4 MCA - Biplot

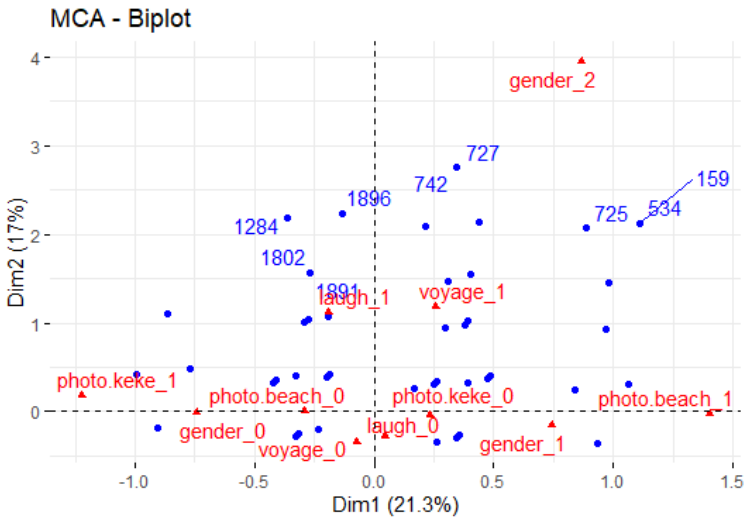

Figure 10: MCA - Biplot.

Figure 11 shows a global pattern within the data. Rows (individuals) are represented by blue points and columns (variable categories) by red triangles. The distance between any row points or column points gives a measure of their similarity (or dissimilarity). Row points with similar profile are closed on the factor map. The same holds true for column points.

5 K-means

5.1 Finding the Clusters

K Means clustering is one of the simplest yet efficient unsupervised algorithms. Our objective is to find K number of groups or "clusters" which

are similar to each other. Each cluster is associated with a centroid which is unique to each cluster. This algorithm iterates until the centroids do not change its position.
We found earlier the optimal number of components which capture the greatest amount of variance in the data. In this step, we will use k-means clustering to view the top 2 PCA components. In order to do this, we will first fit these principal components to the k-means algorithm and determine the best number of clusters. Determining the ideal number of clusters for our k-means model can be done by measuring the sum of the squared distances to the nearest cluster center aka inertia. Much like the scree plot in fig. 9 for MCA, the k-means scree plot below indicates the percentage of variance explained, but in slightly different terms, as a function of the number of clusters.

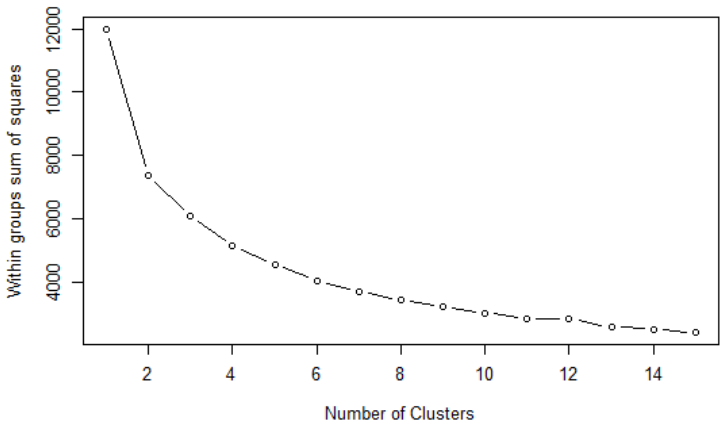

Figure 11: k-means scree plot

Fig. 12 shows that after 4 clusters at (the elbow) the change in the value of inertia is no longer significant and most likely, neither is the variance of the rest of the data after the elbow point. Therefore we can discard everything after k=4 and proceed to the last step in the process.

5.2 Visualising and Interpreting the Clusters

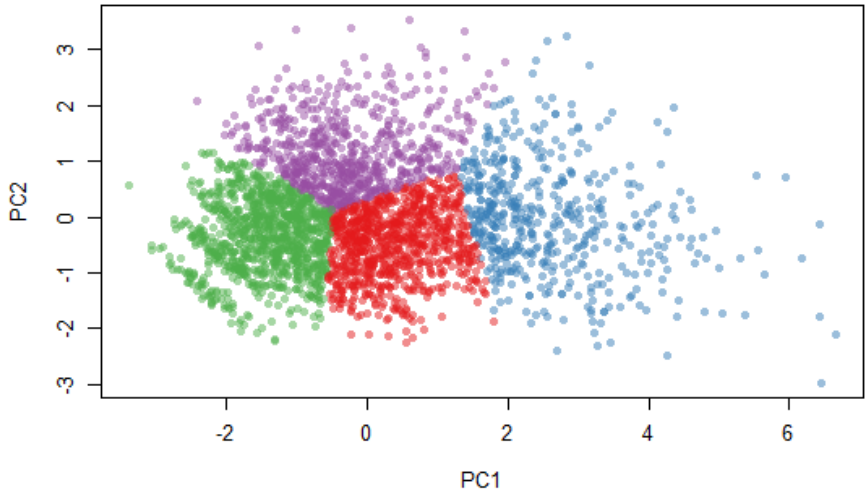

Figure 12: Clustering scatter plot.

Figure 13 shows some clearly defined clusters in the data reducing it to two principal components and using K-means . Now that we know how many clusters there are in our data, we have a better sense of how many groups we can label the population with.

6 Hierarchical Clustering

6.1 How HC works

The algorithm of the HCPC method, as implemented in the FactoMineR package, can be summarized as follow:

Compute principal component methods: PCA, (M)CA or MFA depending on the types of variables in the data set and the structure of the data set. At this step, you can choose the number of dimensions to be retained in the output by specifying the argument ncp.

Compute hierarchical clustering: Hierarchical clustering is performed using the Ward's criterion on the selected principal components. Ward criterion is used in the hierarchical clustering because it is based on the multidimensional variance like principal component analysis.

Choose the number of clusters based on the hierarchical tree: An initial partitioning is performed by cutting the hierarchical tree.

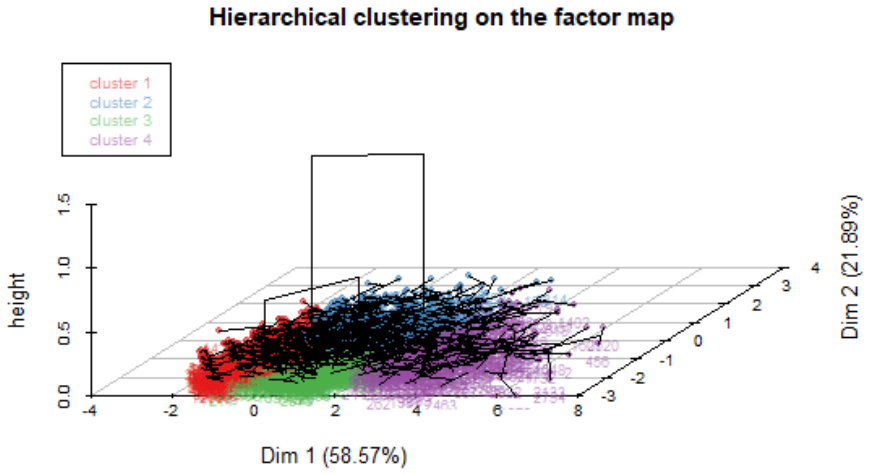

Figure 13: Hierarchical clustering on the factor map.

6.2 HC vs K-means

To recap, k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.

Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

K-means is less computationally intensive compared to HC and are suited with very large dataset. Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow. Also, Hierarchical clustering don't work as well as, k means when the shape of the clusters is hyper spherical.

Moreover, the pros of HC compared with k-means are:
1 .Ease of handling of any forms of similarity or distance.
2. Consequently, applicability to any attributes types.

On the other hand, in hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram. While it is not the case of K Means clustering that needs advance knowledge of K i.e. no. of clusters.

7 Conclusion

We can pull some conclusions regarding our dataset based on the previous cluster and principle component analysis:

1- We can reduce our dimensions from 16 features into just 2 dimensions and still retain more than 80% of the variances using PCA. The dimensionality reduction can be useful if we apply the new PCA for machine learning applications.

2- We can separate our data into at least 4 clusters based on all of the numerical features, with more than 80.5% of the total sum of squares come from the distance of observations between clusters.