UNIVERSITY OF
PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

BIG DATA MANAGEMENT

1st Academic Year Assignment 2023-2024, due 3/11/2023

Collaborators
Georgakis Giorgos, 1052995
Benekos Konstantinos, 1067754
Servos Dimitrios, 1067563
Petroulis Veliassarios, 1074579

Abstract

"The purpose of this work is to become familiar with R and Python environments in order to execute the Principal Component Analysis (PCA) method, as well as to implement distance and similarity measures useful in data processing."The project consists of 6 tasks. Task 1 involves an online quiz with 15 questions on R. Task 2 concerns summarizing 2 papers. Task 3 involves creating a Principal Component Analysis (PCA) algorithm in R and Python for a dataset (Cars93). In Task 4, we developed a function in R and Python that calculates the Euclidean distance for pairs of vectors. Task 5 focuses on how to create a function in the R and Python environments that takes two vectors as input and calculates the cosine similarity between them. Finally, in Task 6, we created a function in both R and Python that takes two vectors as input and computes their distance."

Task 2

Paper 1

This article discusses the application of dimensionality reduction, in particular P.C.A. (Principal Component Analysis), to three benchmark input-output tables of the U.S.A.(for the years 2002, 2007, 2012) as well as for the year 2019 using data given by the U.S. Bureau of Economic Analysis (BEA). The dimensions of each table were reduced to 70x70 matrices eliminating the variables "housing industry" and "fictitious household industry". Dimensionality reduction is a way to reduce the complexity of a model, its origins can be traced back to the physiocrats (tableau economique) and similar techniques (the Schur, SVD) have been identified in plethora of other works since then.

The article describes/applies P.C.A. as a dimensionality reduction technique that transforms data (points in space) to lower dimension through linear combinations of the original data. Basically, it transforms correlated variables into fewer uncorrelated variables and then projects the original data into the reduced P.C.A. space using the eigenvectors/eigenvalues of the variance/covariance matrix (Principal Components). The resulting data captures most of the variance of the original data set.

Even though P.C.A. has not been used extensively in input-output analysis, it holds a key advantage in the sense that it is able to identify the relative importance of the industries operating in the economy. The author goes on to describe how a P.C.A. is applied to input-output resulting in the computation of the Principal Components and their transformation into the PCA space using the variance/covariance matrix.

$$(\mathrm{PC}) = \frac{(\overline{H'} \cdot \overline{H})}{70-1} \cdot \mathrm{PC} \tag{1}$$

- $\overline{\mathrm{H'}} \cdot \overline{\mathrm{H}}$: variance/covariance matrix (on its diagonal, the eigenvalues are ranked from max to min).

- (70-1): degrees of freedom.

From the resulting $\overline{PC}$, we keep the eigenvectors corresponding to the top two eigenvalues, hence reducing the dimensionality to a 70×2 matrix.

The text goes on to discuss the application of P.C.A. in the context of identifying Key industries. It then goes on to compare P.C.A with the Leontief inverse as well as the estimates for forward and backwards linkage (FL, BL respectively). Computing the Pearson correlation efficient, a strong and positive relationship between the first PC and the unweight BL of each industry is revealed. Still FL remains important for understanding industry interrelationships and economic structural changes. Using both the first and second PCs (to account for FL) is essential for a comprehensive analysis.

Next the 70 industries are clustered based on similarity/homogeneity. First the k-means criterion (divides n observation into k clusters) is used to part our data then we choose the n of clusters based on the highest Silhouette Score (how well the data fits in its assigned cluster) which in our case is three for every year. K-means is based on the principal of minimizing the Euclidean distance (similarity measure) withing in each cluster, meaning

from each point to the "center" of the cluster. Then the industries are classified into the three clusters based on their Euclidean distance to each centroid.

To test the effectiveness of P.C.A. the author goes on to compare its findings with those of total BL and FL. Whilst neither of them taken individually produces anything like the results in ranking as the P.C.A., both of them taken together produce a similar ranking. This supports our results as meaningful.

In effect it is concluded that the P.C.A. stands out as a valuable tool for analyzing input-output data, refining industry rankings and revealing cluster structures since it is undoubtedly better at capturing the variance associated with each of the industries, (because it allows for said variance to be measured as the distance of a "data point " from zero) than traditional method. As well as giving us the analytical advantage by organizing the data into well-defined clusters and dendrograms(number of branches based on the k-mean criteria) and yet having similar raking as the BL,FL approach. The author encourages for further research on various levels.

Paper 2

This research describes the process of creating a Socioeconomic Status (SES) Index in the absence of data on income or consumption, using Principal Component Analysis (PCA) on variables related to the ownership of durable assets, access to infrastructure and services, and housing characteristics.

One of the reasons of using PCA is that collecting precise income data is labor-intensive and requires extensive resources for household surveys. PCA (Principal Component Analysis) is a statistical technique for multiple variables that is used to reduce the number of dimensions (variables) while retaining as much variance as possible for our data. As a result, it allows us to visualize and find correlations among the respective variables. It is particularly useful when analyzing composite variables, i.e., a large number of potential variables that could be collected, some of which may provide similar information. Therefore, a natural approach is to use methods such as Principal Component Analysis (PCA) to try to organize the data. Data from the Demographic and Health Survey (DHS) were used for the above analysis, and the following variables were used: ownership of durable assets, access to infrastructure and services (e.g., sanitation facilities and water source), and housing characteristics (e.g., number of sleeping rooms and construction material), with 27355 observations (framedata dimensions, [observations * variables]).

Additionally, descriptive analyses were conducted for all variables. It is worth mentioning that data in categorical form are not suitable for PCA, as their numbering has no meaning. Therefore, qualitative categorical variables must be recoded into binary variables. Another issue is missing values, which pose a problem both in the application of PCA and in the results. If we replace the mean in each missing value, it reduces the variance, which is a problem, so we must be very careful about how we handle the missing values. From the results of the PCA table in this study, emphasis was placed on the factor score. It includes the coefficients' scores for each variable. These scores indicate how each variable is related to the principal component. Usually, a variable with a positive score is associated with a higher Socioeconomic Status (SES), while a variable with a negative score is associated

with a lower SES. Interpreting the weights from our example, a household in urban Brazil with more goods, water from a cable reaching the residence, sanitation leading to a conduit, integrated floor investments, and a high number of sleeping rooms will achieve a higher SES score.

The finding is similar for rural Brazil, unless it includes any sanitation facilities and a well. In urban Ethiopia, a household with more goods and water from the conduit to the compound will achieve a higher SES score. In rural Ethiopia, owning any goods or having access to infrastructure such as water or sanitation will lead to a higher SES score. Next, another statistical method is used, cluster analysis, which allows the separation of observations into groups based on certain characteristics. This particular study/publication introduces a new methodology for how to use an SES Index from PCA.

It highlights the main advantage, which is the avoidance of measurement problems associated with income- and consumption-based methods, such as seasonality. Furthermore, compared to other statistics, PCA is simple to compute. Despite the problems, recent work on SES indices based on PCA suggests that they can be validated and are robust. However, it is the user's responsibility to bear in mind that PCA should primarily be considered as a brief empirical method.

Task 3

I)

Table 1: PCA Results-R

	Eigenvalues	Variance Explained
pc1	21659304372	88.398
pc2	2698659158	11.014

Comment: The above results were generated using the statistical technique PCA (Principal Component Analysis).

II)

Table 2: PCA Results

eigenvalues	explained variance ratio
1931.513	0.91
168.453	0.079

Comment: The above results were generated using the statistical technique PCA (Principal Component Analysis).

In the tables above, we have extracted the eigenvalues corresponding with the top two principal components, as well as the percentages of variance those explain. This is done since, as referred to in the above mentioned papers, those two top principal components explain most of the variance in the given sample. So by just selecting those two we can still explain most of the variance whilst reducing the dimensionality of our data. Indeed when observing our results in table 1 1 the first two pc's, with respective eigenvalues, explain together 99.4% of the total variance in our sample. The same is true for table 2 2 where the two pc's, and their respective eigenvalues, explain in total 98.9% of the total variance of the sample.As a result, it assists us in further analyzing and visually representing our data.

Task 4

I)

- $x_1 = (1, 2, 3, 4, 5, 6)$
- $y_1 = (1, 2, 3, 4, 5, 6)$
- $x_2 = (-0.5, 1, 7.3, 7, 9.4, -8.2, 9, -6, -6.3)$
- $y_2 = (0.5, -1, -7.3, -7, -9.4, 8.2, -9, 6, 6.3)$
- $x_3 = (-0.5, 1, 7.3, 7, 9.4, -8.2)$
- $y_3 = (1.25, 9.02, -7.3, -7, 5, 1.3)$
- $x_4 = (0, 0, 0.2)$
- $y_4 = (0.2, 0.2, 0)$

For each pair of the 4 vectors, we calculate the Euclidean distance using the following formula.

- $d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$

Distance

- Euclidean distance$d(x_1, y_1) = 0$
- Euclidean distance$d(x_2, y_2) = 40.78382$
- Euclidean distance$d(x_3, y_3) = 24.21059$
- Euclidean distance$d(x_4, y_4) = 0.3464102$

II)

Figure 1: Data

ΑΑ χρήστη	Διάρκεια ομιλίας (σε λεπτά)	Πλήθος SMS που έχουν σταλεί	Χρήση Internet (σε MB)
1	25000	14	7
2	42000	17	9
3	55000	22	5
4	27000	13	11
5	58000	21	13

Comment: We are searching for the profile of the user that most closely resembles the profile of the user with code 5.Below, we will examine the transformation we applied using the Euclidean distance to convert it into a similarity metric in order to identify which profile is more similar to that of user 5

Similarity measure

- $s = 1/(1 + d)$

Using the function described above, we discovered that the profile associated with user 3 is the one that most closely resembles that of user 5. User 3 also had the highest similarity index ($s_3 = 1/(1 + d3) = 0.0003332211$).

Question:Which values have the greatest impact on the Euclidean distance that arises in each case?

The variables with the highest variance because if a variable had low variance, all of its values would be close to the mean. As a result, taking any pair of values and placing them in two vectors wouldn't significantly affect the Euclidean distance measure between them.

Task 5

- $x_1 = (9.32, -8.3, 0.2)$
- $y_1 = (-5.3, 8.2, 7)$
- $x_2 = (6.5, 1.3, 0.3, 16, 2.4, -5.2, 2, -6, -6.3)$
- $y_2 = (0.5, -1, -7.3, -7, -9.4, 8.2, -9, 6, 6.3)$
- $x_3 = (-0.5, 1, 7.3, 7, 9.4, -8.2)$
- $y_3 = (1.25, 9.02, -7.3, -7, 15, 12.3)$
- $x_4 = (2, 8, 5.2)$
- $y_4 = (2, 8, 5.2)$

For each pair of vectors, we calculate the cosine Similarity using the following formula.

- $\cos(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$

Cosine similarity

- cosine Similarity$d(x_1, y_1) = -0.773$
- cosine Similarity$d(x_2, y_2) = -0.654$
- cosine Similarity$d(x_3, y_3) = -0.140$
- cosine Similarity$d(x_4, y_4) = 1$

Cosine similarity, briefly explained

Cosine similarity is a similarity measure that ranges from -1 to 1. When it's 0, it indicates no similarity between the vectors (a 90-degree angle). A value of 1 represents perfect similarity, while -1 signifies that the vectors point in completely opposite directions

Task 6

- x_1=("Green", "Potato", "Ford")
- y_1=("Tyrian purple", "Pasta", "Opel")
- x_2=("Eagle", "Ronaldo", "Real madrid", "Prussian blue", "Michael Bay")
- y_2=("Eagle","Ronaldo", "Real madrid", "Prussian blue", "Michael Bay")
- x_3=("Werner Herzog", "Aquirre, the wrath of God", "Audi", "Spanish red")
- y_3=("Martin Scorsese", "Taxi driver", "Toyota", "Spanish red")

For each pair of vectors, we calculate the nominal distance using the following formula.

$$\text{nominalDistance}(\text{x1}, \text{y1}) = \frac{\sum(\text{x1} = \text{y1})}{\text{length}(\text{x1})} \tag{2}$$

- nominalDistance(xi,yi)=1 Perfect similarity.
- nominalDistance(xi,yi)=0 No similarity.

Distance

- nominalDistance$d(x_1, y_1) = 0$
- NominalDistance$d(x_2, y_2) = 1$
- NominalDistance$d(x_3, y_3) = 0.25$