Performance Assessment: DIMENSIONALITY REDUCTION METHODS

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D212 – DIMENSIONALITY REDUCTION METHODS

Professor Middleton

August 22, 2024

Part I: Research Question

A. Describe the purpose of your data mining report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer by using PCA.

My practical research question is: "Can Principal Component Analysis (PCA) be used to identify key features for patients?" I will be utilizing the Medical Data set for this analysis.

2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

The primary goal of this data analysis is to identify the principal component within the Medical dataset. This goal is realistic and doable because the dataset has useful variables that can be analyzed with PCA. By finding these key variables, healthcare organizations can make better decisions, and improve patient care. Overall, the goal of PCA is to simplify the dataset by focusing on the most significant features while reducing less significant ones.

Part II: Method Justification

B. Explain the reasons for using PCA by doing the following:

1. Explain how PCA analyzes the selected data set. Include expected outcomes.

PCA converts the original correlated variables into new, uncorrelated variables known as principal components. This reduces the overall number of variables while keeping most important data. The first few components usually capture most of the data's variability, making

the analysis simpler but still useful. According to Jolliffe and Cadima (2016), PCA is widely used in data analysis for its ability to reduce data dimensionality, especially in cases where there is a large number of variables with potential correlations.

2. Summarize one assumption of PCA.

PCA assumes that the data is not heavily influenced by outliers. Since PCA aims to maximize variance, the presence of outliers can disproportionately affect the principal components, leading to misleading results. Therefore, it's important to preprocess the data and handle any outliers appropriately before applying PCA to ensure that the extracted components accurately reflect the underlying data structure (Jolliffe & Cadima, 2016).

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. Identify the continuous data set variables that you will need to answer the PCA question proposed in part A1.

The initial data set variables used for the analysis are ‘Age’ (continuous), ‘Income’ (continuous), ‘VitD_levels’ (continuous), and ‘TotalCharge’ (continuous). As you can see all the variables are continuous data which follows the requirements for PCA. As the technique depend on continuous data to capture the underlying variance structure and reduce dimensionality.

2. Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.

I standardized the continuous data variables identified in Part C1. The clean data set is called, 'PCA_medical.csv'.

Part IV: Analysis

D. Perform PCA by doing the following:

1. Determine the matrix of *all* the principal components.

The matrix of principal components was obtained by applying PCA to the standardized Medical Data set. This matrix includes new variables (principal components) that are linear combinations of the original variables, with each component accounting for a specific percentage of the total variance in the data. Most commonly, the first few components typically capture most of the variance, making them the most informative.

Below is a representation of the matrix showing the Principal Components:

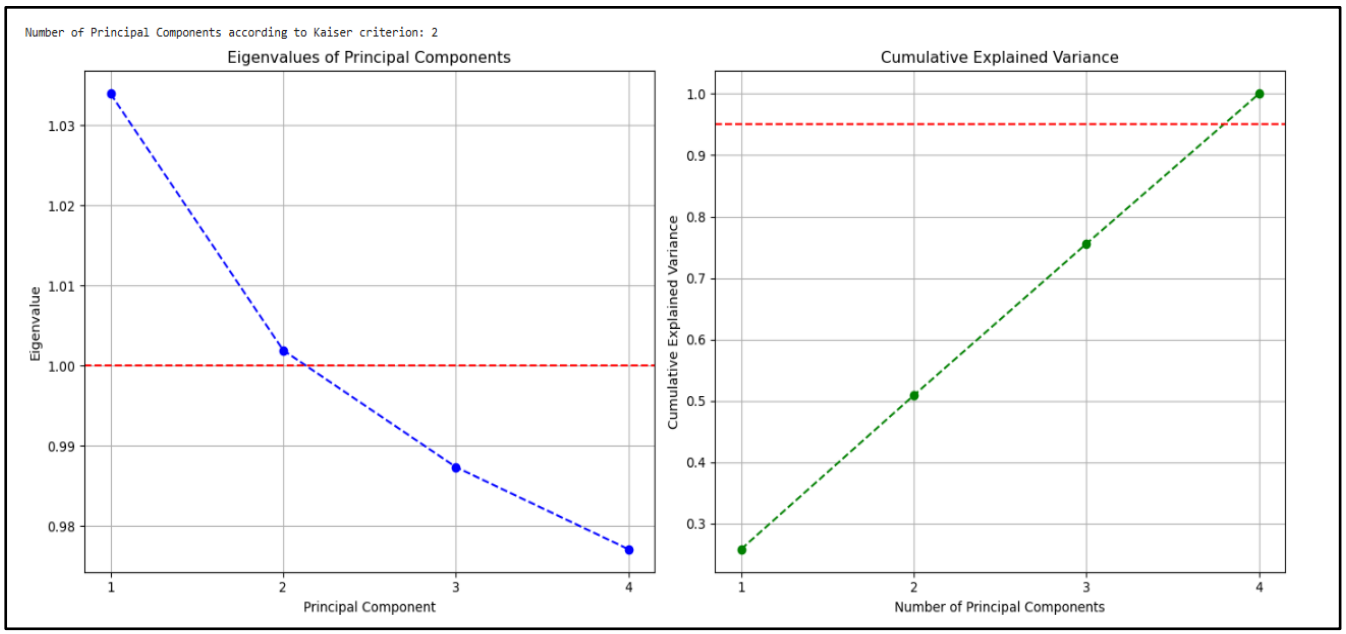
```
Matrix of Principal Components:
[[-0.55907815  0.55384824 -0.36412614 -0.49809227]
 [-0.12063183 -0.14175941  0.78245642 -0.59423411]
 [ 0.68796726  0.70151863  0.15375496 -0.10455654]
 [-0.44674449  0.42547386  0.48117929  0.62278234]]
```

2. Identify the *total* number of principal components, using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.

To figure out the total number of principal components, two general methods were used:

- Elbow Rule: This method suggests retaining the components before the point where the explained variance starts to level off (resembling an “elbow” in the plot). These components capture the most significant information.
- Kaiser Criterion: Corresponding to this criterion, I would retain components with eigenvalues greater than one. Eigenvalues represent the importance of each component.

A scree plot was generated to visualize how much variance each component explains. The optimal number of components was determined using the Elbow method. As shown in the graph below, the ideal number of components is two (indicated by the red line).



3. Identify the variance of *each* of the principal components identified in part D2.

After calculating and presenting the variance explained by each principal component, I identified a total of two principal components. This involved determining the percentage of total variance

captured by each component, helping us understand how much information each principal component retains from the original dataset.

```
Variance of Each Principal Component (%):
PC1: 25.85%
PC2: 25.05%
```

4. Identify the *total* variance captured by the principal components identified in part D2.

The analysis yielded two principal components, each capturing a specific percentage of the total variance in the dataset. The variances explained by each principal component are as follows:

Variance of Each Principal Component (%):

- PC1: 25.85%
- PC2: 25.05%

Together, these two principal components capture a total of 50.89% of the variance in the dataset. This indicates that the first two components are sufficient to represent a significant portion of the original data's variability.

5. Summarize the results of your data analysis.

Using Principal Component Analysis (PCA) simplified the dataset while retaining all essential information. By applying the elbow rule and Kaiser criterion, I determined the optimal number of principal components, which captured 50.89% of the total variance. This insight into key factors affecting patient outcomes is valuable for healthcare organizations.

The principal component matrix shows how the original variables (‘Age,’ ‘Income,’ ‘VitD_levels,’ and ‘TotalCharge’) combine to create new dimensions. Each value in the matrix represents how much a variable contributes to a principal component. These insights guide informed decision-making

Part V: Attachments

E. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

Data to Fish. (n.d.). How to Create a Covariance Matrix in Python. Retrieved from https://datatofish.com/covariance-matrix-python/

F. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. DOI: 10.1098/rsta.2015.0202

Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100. Available at: https://arxiv.org/abs/1404.1100