Performance Assessment D212 – Data Mining II  
Task I

Doug Haunsperger

College of IT, Western Governors University

March 22, 2024

# Part I. Rese**arch Questi**on

## A1. Question Proposal

I propose to research the question, “Is doing a clustering analysis more effective when first reducing the dimensionality of the data using Principal Component Analysis (PCA)?”

## A2. Goals

The goal of this analysis would be to reduce the dimensionality of the data set because “the higher the number of features [in a data set], the more difficult it is to model them.” (Navot, 2022). The expected outcome would be a reduced-dimension data set ready to be submitted to a clustering algorithm.

# Part II. Technique Justification

## B1. Explanation of PCA

Principal component analysis is a method to determine a “lower-dimensional space by preserving the variance as measured in the high dimensional input space.” (Navot, 2022). The goal of PCA is to collapse the original input features to a smaller group of “weighted linear combinations of the original” features. (Bruce, Bruce, & Gedeck, 2020, p.285). These linear combinations, or components, are chosen to be orthogonal to each other in the high-dimensional feature space, and hence uncorrelated.

Building on the example given in Bruce, Bruce, & Gedeck, it can be shown that for *n* input variables, *X1, X2, … Xn,* one can calculate *n* components *Zi* (*i* = 1, 2, …, *n*) as:  
The weights () are the component loadings, which “transform the original variables into the principal components”. (2020, p.285). The first PC *Z1* is “usually the most important” and “accounts for greater variability among the predictors than any other component.” (Larose & Larose, 2019, p. 175). Likewise, additional PCs explain more and more of the variance of the original data set.

The loadings vectors <> are the eigenvectors of the correlation matrix ***R*** of the data set “corresponding to the eigenvalues λ*j* of ***R*** in descending order, ranging from highest to lowest.” (Steiger, 2015).

Typically, an analyst will choose a small number of PCs that explain the majority of the variance in their data set. Steiger relates three potential heuristics for choosing the number of PCs to retain:

1. Kaiser-Guttman rule: Components based on eigenvalues λ*j* > 1
2. Scree plot test: Examine the plot of eigenvalues for a “clear break” between a set of “meaningful components” and a long tail of “essentially meaningless” ones.
3. Arbitrary percentage of variance explained, such as 80%. (2015).

Applying this to our data set, where I have chosen seven input variables, I expect PCA will identify 2-4 components that explain the majority of the variance in the original seven.

## B2. Method Assumptions

Since PCA transformations are linear, finding the eigenvectors of the input variable matrix, one of the underlying assumptions for PCA to be effective is that the distribution of the input data set is generally linear (Navot, 2022). In Figure 1, the data points are roughly correlated to a straight line, making PCA meaningful. If the data was uniformly distributed, the calculated PCs would be meaningless.

Figure 1  
*Graph showing simple PCA on a two-dimensional dataset (Navot, 2022)*

*A diagram of a diagram

Description automatically generated*

# Part III. Data Preparation

## C1. Data Set Variables

I have chosen to do the clustering analysis on a set of 7 variables, each quantitative: Population, Age, Income, VitD\_levels, Initial\_days, TotalCharge, and Additional\_charges. Population and Income underwent a log transformation to make their distributions closer to normal. Technically, since both Population and Age take only integer values, they are properly categorized as discrete, but as they have a wide range of possible values, they are functionally continuous. The other five variables all take floating point values and are properly continuous.

## C2. Cleaned Data Set

To clean the data set, I removed records where Population = 0 in order to enable the log transformation. I did a log transform of Population and Income and renamed those columns to ‘logPop’ and ‘logIncome’. These variables were than scaled using the scikit-learn StandardScaler function. The cleaned and scaled data set (prior to dimension reduction) is attached as ‘clean\_medical\_data.csv’.

# Part IV. Analysis

## D1. Principal Components

With the variables chosen and transformed/scaled as described above, I calculated the principal component matrix as:

PC1 PC2 PC3 PC4 PC5 \

logPop 0.019302 0.088946 -0.013514 -0.001753 0.700725

Age -0.024496 0.701227 -0.014074 0.016304 -0.093891

logIncome -0.152723 -0.011103 -0.705311 0.691712 -0.002173

VitD\_levels -0.976272 -0.005772 0.001685 -0.214561 0.015377

Initial\_days 0.149964 0.005309 -0.708558 -0.689365 -0.009403

TotalCharge 0.010373 0.706749 0.009901 -0.001603 0.031395

Additional\_charges 0.000086 0.026281 0.000544 -0.001599 -0.706296

PC6 PC7

logPop 0.701849 0.089011

Age -0.083588 0.701010

logIncome -0.000921 -0.024879

VitD\_levels 0.014206 -0.019563

Initial\_days -0.010672 -0.000795

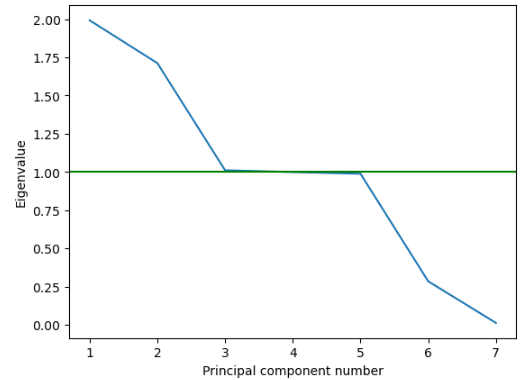
TotalCharge -0.031484 -0.705918

Additional\_charges 0.706480 -0.036596

## D2. Identification of Principal Components

Strictly applying the Kaiser criterion, one would determine that there are three meaningful principal components. The eigenvalues for PCs 3, 4, and 5 are all very close to 1, but PCs 4 & 5 are slightly below 1. Figure 2 below shows the scree plot.

Figure 2  
*Scree plot*



## D3/D4. Variance of Each Component and Total Variance

The percentage of explained variance for each PC is in the table below (both each component individually and cumulative), with the 3 chosen PCs highlighted in green:

|  |  |  |
| --- | --- | --- |
| PC # | Explained Variance % | Cumulative Explained % |
| 1 | 28.5 | 28.5 |
| 2 | 24.5 | 52.9 |
| 3 | 14.4 | 67.4 |
| 4 | 14.3 | 81.7 |
| 5 | 14.2 | 95.8 |
| 6 | 4.1 | 99.8 |
| 7 | 0.2 | 100 |

This data is re-presented visually in Figure 3, with the red line being the cumulative explained variance.

Figure 3  
*Explained variance chart*

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

## D5. Results

PCA of this data set provides 2 PCs that explain a majority of the variance, while 3 PCs explain rough 2/3 of the variance. I used this exact procedure in the first PA of D212, and selecting 2 PCs improved my clustering silhouette score from 0.25 for 2 clusters with the scaled data only, to over 0.5 for 2 clusters within the 2-component feature space. PCA can be an effective dimension reduction tool, if the underlying assumptions hold true.

# Part V. Attachments

## E. Third-party Code Sources

Middleton, K. (n.d.). *Getting Started with D206 | Principal Component Analysis.* Western Governors University. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95>

## F. References

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists : 50+ Essential Concepts Using R and Python*. O'Reilly Media, Inc.

Larose, C., & Larose, D. (2019). *Data Science Using Python and R.* Wiley.

Navot, S. (Apr. 6, 2022). DataLoop AI Blog. *The Curse of Dimensionality – Dimension Reduction*. <https://dataloop.ai/blog/the-curse-of-dimensionality-dimension-reduction/>

Steiger, J. (Feb. 16, 2015). Statpower. *Principal Components Analysis.* <https://www.statpower.net/Content/312/R%20Stuff/PCA.html>