**D212 Task 2: Dimensionality Reduction Methods**

**Joseph Duszynski**

**ID #011237236**

**A.1. Proposal of question**

Using the provided medical dataset (WGU, 2024 [1]), this project seeks to answer the following question: “Can principal component analysis (PCA) be used to reduce the continuous numerical dimensions of patient data to simplify later analysis, data collection, and data storage?” This reduction could aid in computational speed of future models, such as machine learning techniques aiming to predict patient readmission and occurrences of various health conditions.

**A.2. Defined goal**

This project aims to find the most relevant principal components worth retaining while proposing to discard the ones considered relatively unimportant. While some accuracy is lost in removing the least important principal components, the organization benefits from decreased data storage and faster production of models requiring fewer dimensions of data. This can aid hospital analysts with future modeling of patient readmission rates and predicting occurrences of certain health conditions.

**B.1. Explanation of PCA**

Principal component analysis is a dimensionality reduction method that uses a linear transformation from data points on *d* continuous variables (with the standard Euclidean axes) to a *d*-dimensional ellipsoid with (mutually orthogonal) axes defined by the eigenvalues and their eigenvectors, which capture variability within the data in decreasing order by the magnitude of those eigenvalues. The ellipsoid’s axes’ lengths are the eigenvalues of the transformation, so the least important principal components (i.e. smallest eigenvalues) can be discarded for capturing the least variability within the data.

Specifically, PCA first normalizes the data to a mean of 0 and variance of 1, as variables of larger scales will obfuscate important variance in variables of smaller scales. Next, it computes the covariance matrix of the normalized data, then finds the eigenvalues and associated eigenvectors of the covariance matrix. The resulting eigenvectors are sorted in descending order by eigenvalue. For a selected number *k* of components to retain, the first *k* of those eigenvectors (principal components) are kept while the others are discarded. The original data is then transformed to the principal component system by multiplying (on the right) the data by the matrix of the selected principal components (Jaadi, 2024 [1]).

PCA produces a set of principal components (or eigenvectors) with weights according to the initial selection of *d* variables to allow the original data to be transformed to the principal component coordinate system. The eigenvalues and covariance matrix can be used to calculate the explained variance for each of the components, aiding the user in determining a desired value of *k* components to retain.

**B.2. PCA assumption**

Calculating the covariance matrix naturally requires numerical data and implicitly assumes there are no null values present in the variables being used. Although highly unlikely, if the covariance matrix is of an even dimension, it may have no real eigenvalues (only complex ones), such as a matrix representing a rotation. Matrices with degenerate eigenvalues don’t necessarily have a set of *d* linearly independent eigenvectors to span the *d*-dimensional space of the covariance matrix. Additionally, outliers can distort the covariance matrix and have non-trivial influence on the determination of the eigenvalues and associated principal components.

PCA operates on the assumption that the variables have reasonably normal (gaussian) distributions with correlations between them. With no correlations, the off-diagonal elements of the covariance matrix would be 0 and the eigenvalues would all be 1 (and make no change to the underlying coordinate system). Non-normal distributions would be distorted by the standard normal scaling done during preprocessing (and provide unhelpful results).

As discussed in the previous section, PCA assumes all of its data has been standardized to the same scale (adjusting to mean of 0 and variance of 1) during preprocessing, as variables in different scales would have disproportionate influences on the procedure and invalidate the results. Although it could still calculate a covariance matrix and eigenvalues without proper scaling (standardizing), the results would not be applicable.

**C.1. Continuous data set variables**

The following are the 13 continuous (numeric) variables from the dataset to be used with PCA:

1. Lat
2. Lng
3. Population
4. Children
5. Age
6. Income
7. VitD\_levels
8. Doc\_visits
9. TotalCharge
10. vitD\_supp
11. Full\_meals\_eaten
12. Initial\_days
13. Additional\_charges

**C.2. Standardization of data set variables**

See the attached file “X\_pca\_scaled.csv” for the cleaned and scaled data.

The variables are scaled with StandardScaler() and the function scale\_data shown below:



**D.1. Principal components**

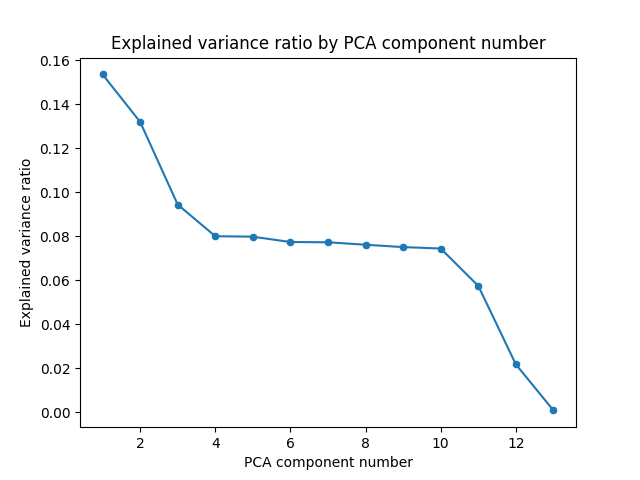
Below is the matrix of the 13 principal components and their corresponding (normalized) values with the 13 variables listed in section C.1.:

A screenshot of a computer screen

Description automatically generated

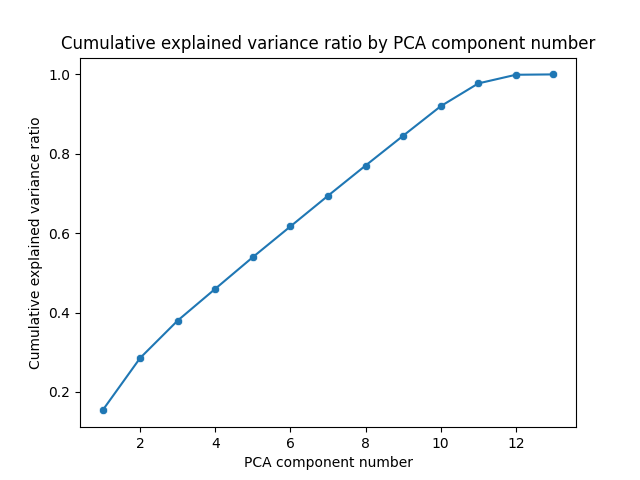
**D.2. Identification of the total number of components**

PCA produced the following scree plot:



A graph with a line and a red line

Description automatically generated



The Kaiser criterion states the most relevant principal components are those with an eigenvalue of 1.00. In the scree plot above, components 4 through 10 all have eigenvalues very close to 1.00. Looking over the cumulative explained variance ratio, retaining only 6 principal components (where the y = 1.00 line first intersects with the eigenvalue vs component number curve) would result in only 60% explained variance, a rather poor result. Since the tenth component has an eigenvalue of ~0.97 (very close to 1.00), and the explained variance is over 90% when 10 components are used (discussed in the following sections), the PCA process should retain 10 of the 13 components (i.e. ‘PC1’ through ‘PC10’).

**D.3. Variance of each component**

The ten principal components ‘PC1’ through ‘PC10’ have the following variances:

A screenshot of a computer

Description automatically generated

**D.4. Total variance captured by components**

The total variance captured by the ten principal components is shown in the last row of the image from the previous section – the cumulative explained variance ratio for ‘PC10’, which is 0.919877, or 92.0%

**D.5. Summary of data analysis**

With the 13 continuous (numeric) variables listed in section C.1., PCA initially computed 13 principal components, with its matrix shown in section D.1. Using the scree plot and Kaiser criterion discussed in section D.2., it was determined the ideal number of components would be 10. As displayed in section D.3., these ten components have explained variance ratios from 15.35% down to 7.44%, with a cumulative explained variance ratio of 91.99%. It is advised that this principal component system of 10 components be used to reduce the dimensionality of the original 13 variables.

**E. Sources for third-party code**

**1.** WGU. 2024. D212 Data Mining II “Data Sets and Associated Data Dictionaries”. Medical Data and Dictionary Files. Retrieved May 23, 2024, from <https://access.wgu.edu/ASP3/aap/content/jf8rcds032ldktfces9r.html>.

**2.** WGU. 2024. D206 Data Cleaning "Welcome to Getting Started With Principal Component Analysis". Retrieved May 25, 2024, from <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F7%2E%20D206%2DGettingStartedPCA%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>.

**F. Sources**

**1.** Jaadi, Zakaria. 2024. “A Step-by-Step Explanation of Principal Component Analysis (PCA)”. Retrieved May 25, 2024, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis.