Assignment 6 part 2

Patrick, Martin
Nicholas

April 2022

For part 1, see the notebook

PCA - Principal Component Analysis Is can be broken down to five steps, in understanding what **PCA** is.

1. Standardize the range of continuous initial variables
2. Compute the covariance matrix to identify correlations
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
4. Create a feature vector to decide which principal components to keep
5. Recast the data along the principal components axes

To understand what **Principal Component Analysis** is, we first need to look at the basics. **PCA** is a dimensionality-reduction method that is used to reduce the dimensionality of large data sets. This is done by by transforming large sets of variables into smaller one, while still keeping the essential data to conduct the analysis.

For instance we can take a look at the ***Divorce Predictors*** dataset used in Assignment 7, which contains 55 columns where in 54 is questions.

In [13]:
```
dataset.head()
```

Out[13]:

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54	Status
0	2	2	4	1	0	0	0	0	0	0	...	2	1	3	3	3	2	3	2	1	Divorced
1	4	4	4	4	4	0	0	4	4	4	...	2	2	3	4	4	4	4	2	2	Divorced
2	2	2	2	2	1	3	2	1	1	2	...	3	2	3	1	1	1	2	2	2	Divorced
3	3	2	3	2	3	3	3	3	3	3	...	2	2	3	3	3	3	2	2	2	Divorced
4	2	2	1	1	1	1	0	0	0	0	...	2	1	2	3	2	2	2	1	0	Divorced

5 rows × 55 columns

In [14]:
```
dataset.tail()
```

Out[14]:

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54	Status
165	0	0	0	0	0	0	0	0	0	0	...	1	0	4	1	1	4	2	2	2	Married
166	0	0	0	0	0	0	0	0	0	0	...	4	1	2	2	2	2	3	2	2	Married
167	1	1	0	0	0	0	0	0	0	1	...	3	0	2	0	1	1	3	0	0	Married
168	0	0	0	0	0	0	0	0	0	0	...	3	3	2	2	3	2	4	3	1	Married
169	0	0	0	0	0	0	0	1	0	0	...	3	4	4	0	1	3	3	3	1	Married

5 rows × 55 columns

Figure 1: Divorce data

As it can be seen in 1, the many questions can be reduced to smaller *Principal Components*, where instead of having the 54 questions, we might reduce it to merely 10 or 20 questions.

0.1 Standardization

When standardize the range of the initial variable so that each one contribute equally, it is important to make sure no variable will dominate others. For instance a variable that ranges between o and 100 will dominate over a variable that ranges between 0 and 1.
Mathematically, this can be done with the following:

$$z = \frac{value - mean}{Standard\ Deviation}$$

0.2 Covariance Matrix Computation

This is a step we take to see if there are any relationship between our variables. A coveriance matrix is a $p\ x\ p$ symmetric matrix where p is the dimensions.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

0.3 Compute the eightvectors and eightvalues

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the *principal components* of the data.
The eigenvector/value numbers is equal to the number of dimensions of the data. For example, for a 3-dimensional data set, there are 3 variables, therefore there are 3 eigenvectors with 3 corresponding eigenvalues.

0.4 Feature vector

In this part we look at our features in the data set, and determine whether to keep all the components, or discard those of lesser significance (Which is the lower eigenvalue)

0.5 Recast the data along the principal component axes

Here we use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data.