

Mathematical Modelling and Decision Making - Assignment

Joseph Chotard

JPC844 – 1935844

- i. Briefly explain what Principal Components Analysis (PCA) is, how it is applied to a data set and what it can achieve [5 marks]**

PCA is an unsupervised (Kaggle & Encyclopedia Titanica)(Lever, Krzywinski, & Altman, 2017). This means PCA simplifies data with lots of dimensions into a smaller dimensional space while maintaining trends and patterns unlike feature elimination which just ignores certain dimensions. PCA extracts features by projecting them onto Principal Components (PCs). The goal is to find the smallest number of PCs that accurately represent the data. PCA allows dimensionality reduction without losing trends and patterns because PCs are essentially just a combination of all the previous features. Furthermore all PCs must be independent of one another (Brems, 2017) which means the maximum number of PCs is either the smallest of the number of samples or the number of features.

To apply PCA to a dataset:

1. Put the the data into a matrix
2. Centre it into the matrix A
3. Standardise the data
4. Calculate an estimate $A^T A$ of the covariance matrix of A
5. Compute the eigenvectors and their eigenvalues of $A^T A$
6. Take the eigenvectors with large enough eigenvalues and store them in descending order in a matrix B
7. Finally, compute AB which is the projection of all the data points onto the hyperplane defined by the chosen eigenvectors

- ii. Describe the dataset and explain the various values that it includes [5 marks]**

For this assignment, I am going to using Kaggle's Titanic dataset made by the Kaggle Team with the help of Encyclopaedia Titanica (Kaggle & Encyclopedia

Mathematical Modelling and Decision Making – Assignment

Titanica). This dataset has 891 values and 12 dimensions. The dimensions are as follow:

Table 1 Features and descriptions

Feature	Data Type	Description
Passenger ID	Integer	A unique ID for every passenger: is used for identification purposes.
Pclass	Integer	1 means 1 st class, 2 means 2 nd class and 3 means 3 rd class
Name	String	The passenger's name
Sex	String	Male or female
Age	Integer	The passenger's age
Sibsp	Integer	The number of siblings/spouses on board
Parch	Integer	The number of parents/children on board
Ticket	String	The ticket number
Fare	Rational Number	The price of the passenger's ticket
Cabin	String	Cabin number (77% missing data)
Embarked	Character	Port of embarkation: C = Cherbourg, France; Q = Queenstown, Ireland; S = Southampton, UK

Table 2 Labels and description

Label	Data Type	Description
Survival	Integer	1 means the passenger survived, 0 means they did not

- iii. Before conducting any investigation, you should have an expectation about what the results might be. Explain what you hope to obtain from applying PCA to this dataset. Which dimensions will you use to label the data? [15 marks]

By applying PCA and reducing the dimensionality of the data to 2 Principle

Mathematical Modelling and Decision Making – Assignment

Components, I expect to see a rather obvious separation between the passengers who survived and the passengers who didn't when I plot the data. From applying PCA, I hope to be able to apply a classification algorithm in order to predict if a passenger, given their information survives or not.

To classify this data, I'm going to use the survival label: this can be 0 or 1 which makes this a binary classification problem. I hope to see a clear separation between the passengers who survived when plotted onto 2 or, if necessary, 3 dimensions.

Intuitively, I can hypothesise that Gender and Passenger Class will probably have the most impact on the Principal Components because of their importance and the fact that I assume the gender distribution is even amongst the classes.

iv. What software will you use? Justify your choice. [5 marks]

For this project, I will be using `python 3.8.0` with the `scikit-learn (sklearn)` module for doing the data pre-processing and the PCA. I will also use the `matplotlib` module to visualise and plot the data. I've chosen Python over other software like MATLAB for multiple reasons. Firstly, I find Python syntax to be clearer and the code easier to write. I also prefer that Python is free and open source compared to the proprietary and expensive MATLAB which means I can work on my code from different machines very easily. Finally, Python also gives me more choice in the libraries/modules I want to use to write my code.

v. What pre-processing of the data will you do before applying PCA? [5 marks]

First, I separated the label from the features as we only one to apply PCA to the features. The rest of the pre-processing only applies to the features.

I first chose to remove the features that clearly have no value: this includes features that are unique and independent to every passenger like "PassengerID", "Name", "Ticket". I also chose to remove the Cabin feature as it is nearly unique to every passenger and we are missing more than 77% of the data.

Some features had missing data, so I replaced the missing values with the most common value from that feature.

Mathematical Modelling and Decision Making – Assignment

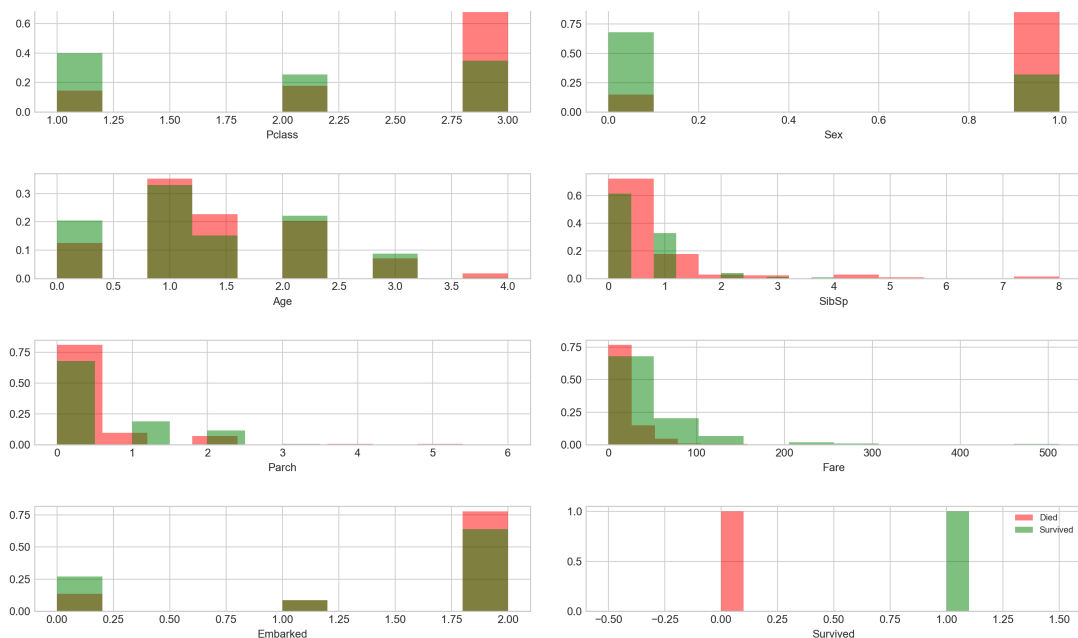
I then decided to categorise the age into 5 categories: $[-\infty; 18]$; $[18; 32]$; $[32; 48]$; $[48; 64]$; $[64; +\infty[$ which I noticed slightly increased the eigenvalues of the Principal Components. I tried one-hot encoding the Embarked and Sex feature as some studies suggested it could be helpful (Salkind, 2007) however I noticed that it ended up slightly reducing the variance of the Principal Components. Instead, I simply converted every nonnumeric value in each feature to a number unique to that value.

Finally, I standardise the data to make sure that features with more variance aren't favoured over features with lower variance.

vi. What insights into the structure of the data can you obtain by plotting projections of the data onto pairs of its existing dimensions? [10 marks]

Before plotting the data onto pairs of its existing dimensions let's plot the data onto each dimension individually to see which dimensions might be worth looking at it in more detail:

Figure 1 Histogram of passenger's who survived and those who didn't on the 7 features and the label



In this histogram, we divided every bar by the amount of people in that category to make sure that the fact that we have more people who died didn't affect the data. From a quick glance, we can see that Sex and Pclass, Fare and Sex seem to have the biggest variance in whether the passenger survived. Let's plot the data onto some of these dimensions.

Mathematical Modelling and Decision Making – Assignment

Figure 2 Data plotted onto Pclass and Fare

When we plot Pclass and Fare, we can see that there is quite a strong relation between both features. This, of course, makes a lot of sense as intuitively the better the class, the higher the price.

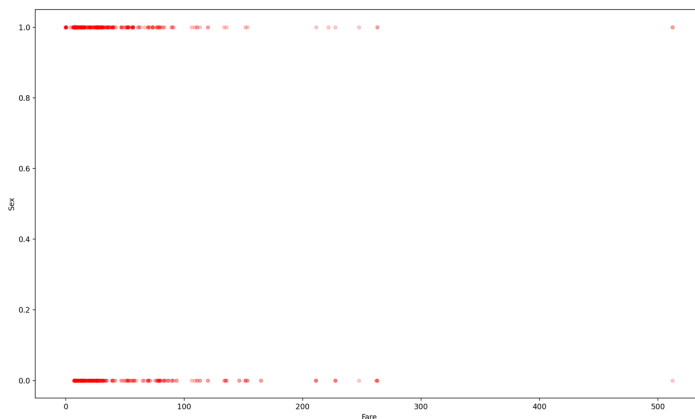
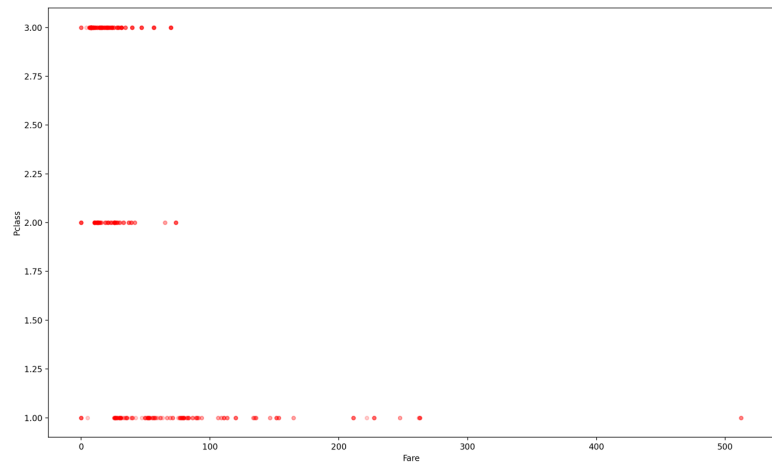
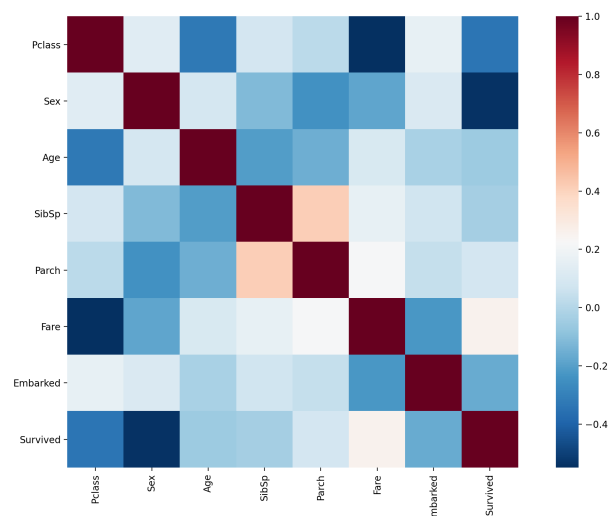


Figure 3 Data plotted onto Sex and Fare

However, if we look at Fare and Sex, we see that there is very little correlation which means that the gender distribution among the classes is very similar.

Figure 4 Correlation between features of original data

Instead of plotting every pair of features let's look at a heatmap of the correlation between features. We can see that there isn't much correlation between most features. This suggests that PCA might not be that efficient in reducing the dimensionality...



Mathematical Modelling and Decision Making – Assignment

vii. For each dimension of the data that you use to categorize the data:

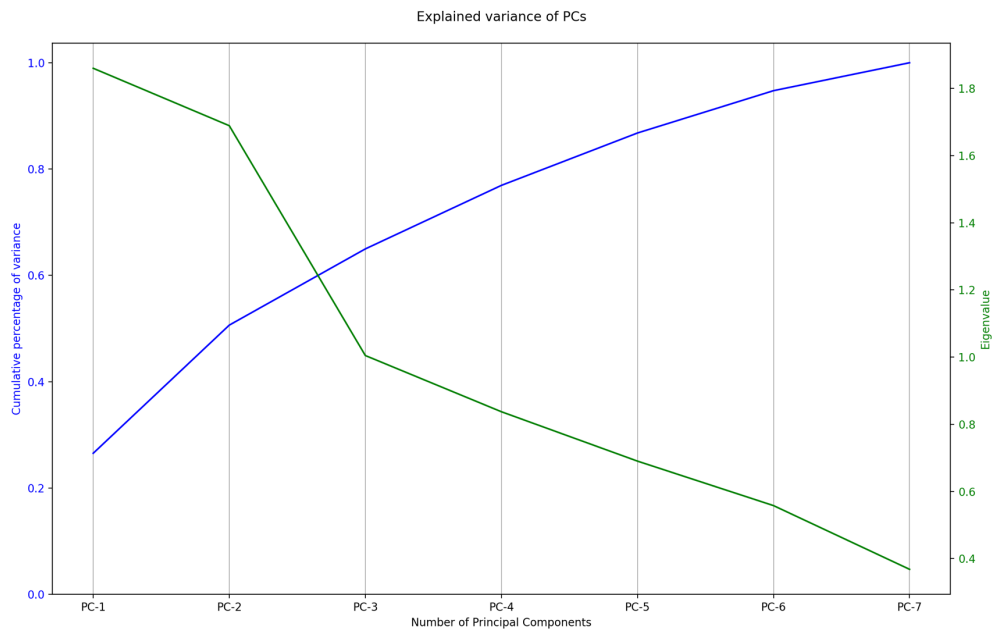


Figure 5 Explained variance of the Principal Components

Here are the eigenvalues for all 7 principal components (in green) along with the cumulative percentage of the variance of the principal components. The eigenvalues tell us that the principal components do not accurately reflect the original data as we need 6 out of the 7 PCs to represent at least 90% of the data (94.74%).

Here's the data from Figure 5:

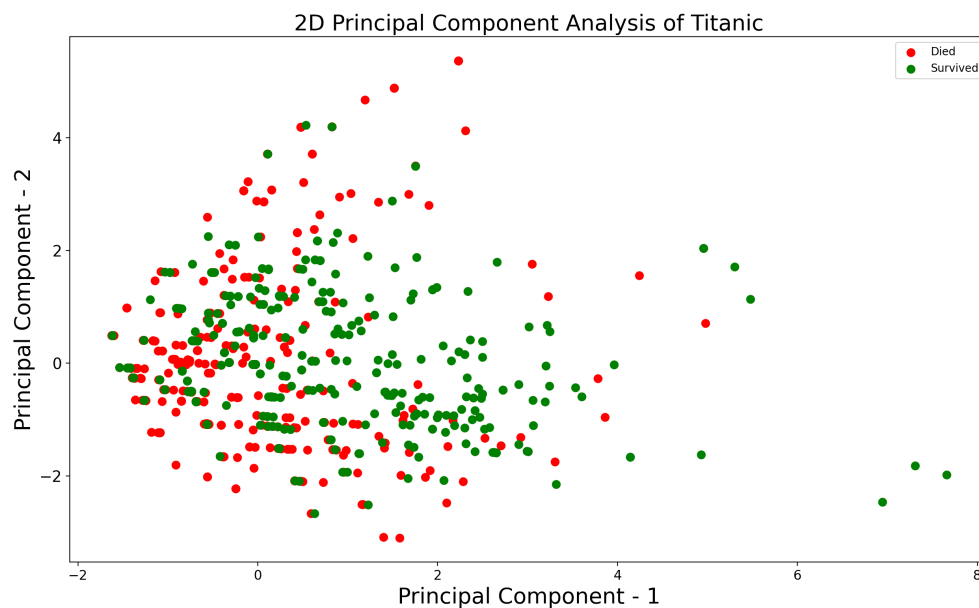
Table 3 Eigenvalues and explained variance

PC	Eigenvalue	Variance Ratio	Cumulative sum of Variance Ratio
1	1.86003984	0.26542175	0.26542175
2	1.68907441	0.24102553	0.50644728
3	1.00476078	0.14337616	0.64982344
4	0.83741472	0.11949641	0.76931985
5	0.69040919	0.09851919	0.86783904
6	0.55794861	0.07961749	0.94745652
7	0.36821761	0.05254348	1.00000000

Mathematical Modelling and Decision Making – Assignment

This suggests that the features in the original data that we selected are most likely already quite independent one of another. Let's project the data onto the first two principal components (Figure 6).

Figure 6 Data projected on first two PCs



As we can see, there isn't very much noticeable separation between the passengers who died and those who survived. This is understandable as Table 3 tells us that combined, the first two PCs only represent 50% of the data. If we plot the data onto the first three PCs, we'll have 65% of the data, let's see if it's clearer:

When we plot the data onto the first 3 PCs, we can start to see a separation between the two categories. This separation isn't perfect as there is a lot of noise. This is to be expected though as when we think about where the data came from, we can

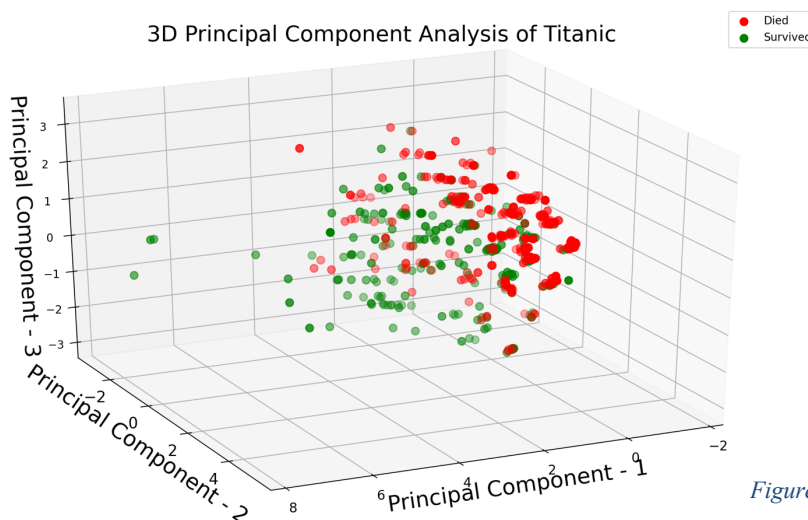


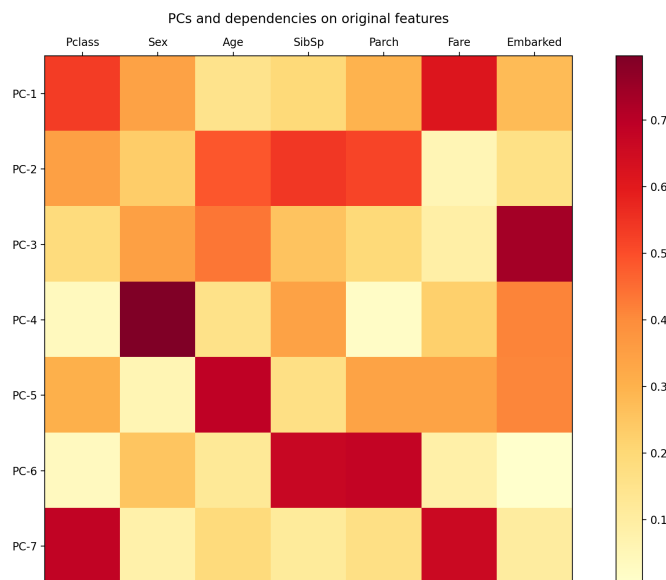
Figure 7 Data projected onto first 3 PCs

Mathematical Modelling and Decision Making – Assignment

intuitively understand that there was a part of chance in whether a passenger survived the sinking.

As I mentioned previously, this confirms the intuition I got from the data in Figure 4.

Figure 8 PCs and their dependencies on the original data



If we ignore the Survival row and column, we can see that the features do indeed have very little correlation between each other. When looking at the survival row/column, we see that the features that have the most correlation with the survival of a passenger are PClass and Sex. This means that when we look at the PCs and their dependencies on the

original components, we should find that PClass and Sex are quite important.

Here's that data plotted onto a heat map. We see that PC 1 is mostly a combination of PClass and Fare (which makes sense when you look at the correlation between Fare and PClass in Figure 4). And PC 4 is mostly composed of Sex. In this analysis of the Eigenvectors, we see that each PC mostly points towards one or sometimes two of the original dimensions which, once again, shows that the original dimensions are mostly orthogonal and independent.

viii. Present your conclusions about your investigation [15 marks]

Throughout this investigation, we found out that the two features that impacted survival the most were the Passenger Class/Fare as well as their Sex. We also discovered that to implement an accurate classifier, we would most likely need to use at least 6 out of 7 Principal Components (Figure 5).

Because we only have 7 features in the first place, it would be just as efficient to apply the classifier on the original features which means that PCA does not serve to

Mathematical Modelling and Decision Making – Assignment

much of a purpose for classifying this dataset.

We did learned that PClass and Fare along with SibSp and Parch are highly correlated (Figure 4), as such, if we decided to apply PCA on the data before training a classifier we could apply PCA selectively on PClass and Fare as well as on SibSp and Parch to reduce these 4 features into 2.

Once we do all this, I'm quite confident that we will be able to train a classifier to accurately identify whether a passenger survives or not based on the features from the training set.

ix. Bibliography

Brems, M. (2017, April 17). *A One-Stop Shop for Principal Component Analysis*.

Retrieved from www.towardsdatascience.com:

<https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>

Jolliffe, I., & Cadima, J. (2015). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Retrieved from The .

Kaggle, & Encyclopedia Titanica. (n.d.). *Titanic: Machine Learning from Disaster*. Retrieved from Kaggle: <https://www.kaggle.com/c/titanic/data>

Retrieved from Kaggle: <https://www.kaggle.com/c/titanic/data>

Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature methods*.

Powell, V. (2015, February 12). *Principal Component Analysis Explained Visually*.

Retrieved from setosa: <http://setosa.io/ev/principal-component-analysis/>

Salkind, N. J. (2007). *Encyclopedia of Measurement and Statistics*.