PRINCIPAL COMPONENT ANALYSIS

Authors: G. H. Zerwes, *Universidade Federal do Espirito Santo, Vitória, Brazil*

Research Supervisor: L. Homri, *Arts et Métiers ParisTech, Metz, France*

1 Introduction

The Principal Component Analysis (PCA) is an algorithm used to solve problems related to dimensionality reduction, feature extraction, data compression, and data visualization.

Its basic working principle consists of projecting the data points on a space of fewer dimensions that still retains a large amount of the total information. The main choice to be made in this algorithm is the number of dimensions, or the number of principal components, for the data to be projected.

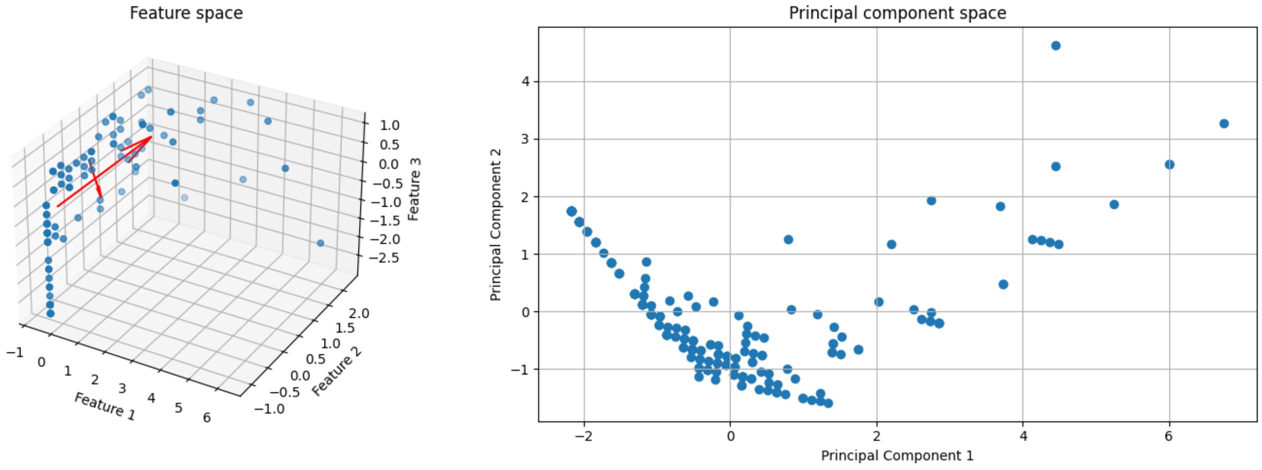

Figure 1: Example of PCA

An example of this algorithm can be seen in figure 1. The first figure shows the data in the three-dimensional feature space. Although in this case there are not many data points, nor many dimensions, it can be useful to reduce the dimensionality. For that, the PCA algorithm was applied and the principal components are represented by the red arrows on the left plot. The biggest arrow represents the first principal component, and the second arrow, the second. The data is then projected into the components and can be visualized in the second plot.

2 Pseudo-code

To use this algorithm on your data, the pseudo-code below shows a procedure that should be followed.

```
#Step 1: Gather the data.

X = [x_value_1, x_value_2, ...]

#Step 2: Fit and transform the data.
reduced_data = PCA.fit_transform(X)
```

3 Other recommendations

An important observation is that the model does not, by default, change the data to have unitary variance, it only transforms the data to have the mean value equal to zero.

In some occasions, it may be necessary to have unit variance in the dataset. This can be the case when the data will be posteriorly fed into a model that requires this condition. To achieve this, the 'whitening' technique may be employed.

Another point of relevance is that the PCA algorithm can only process data in batch processing, that is, the dataset must be fully loaded into the memory for the algorithm to be implemented.

A relevant metric to aid in the choice of the number of dimensions to consider is the explained variance percentage. This number estimates how much of the total variance of the original dataset can be explained by the reduced dataset after the transformation.

References

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*. Number 758. Springer.