DD2343 Assignment 1

jacmalm@kth.se

November 10, 2021

1 Principal Component Analysis

1.1 Explain why data-centering is required before performing PCA. What might happen if we perform PCA on non-centered data?

PCA attempts to find new axes that are linear combinations of the old axes, such that these new axes, or principal components, minimize the mean squared distance between the original data points and their projection onto the principal axes [1]. One key thing to note is that these principal components are limited by the need to cross through the origin. This limitation means that data that is shifted cannot be approximated as well. A simple example is to consider the points $\{-1, 1\}$, $\{0, 0, \}$, $\{1, -1\}$. These points all lie perfectly along a 1 dimensional manifold, the line with slope -1 going through the origin. The mean of this dataset is 0. Thus, we expect the first principle component to consist of this line, and we expect all of the variance to be explained by the first principle component, as the PCA model is perfectly respected.When we run the PCA code (with the lines responsible for centering the data commented out in the SKLearn library file) given in appendix 1, this is exactly what we see. Great! PCA works.

However, if we shift the location of this manifold, while keeping its shape intact, this will change. If we now consider the points $\{49, 51\}$, $\{50, 50\}$, $\{51, 49\}$, it is easy to see that they also perfectly lie along a line, indeed a line with the same slope as the previous example. However, this line does not pass through the origin, and will thus not be found by PCA. Instead,

what we expect to happen, is that the first principle component will be a line that goes through the origin and minimizes the distance from the data points and their projection onto the first principle axis. As the data is symmetric around their mean, we expect this to be a line going through the mean point,$\{50, 50\}$, and the origin.This is a line with slope 1. In fact, this line is perpendicular to the line that the data points lie along. When looking at the explained variance, we can also see that this line does not explain all of the variance, necessitating us to use both principal components to perfectly be able to get back all of our data, meaning that PCA fails to detect the inherent 1-dimensionality of the dataset. Again, these conclusions are verifiable with provided code in Appendix 1, and visually demonstrated in figure 1 and figure 2.

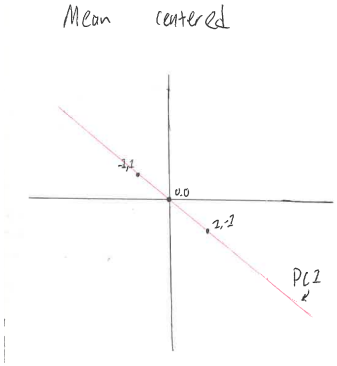


Figure 1.

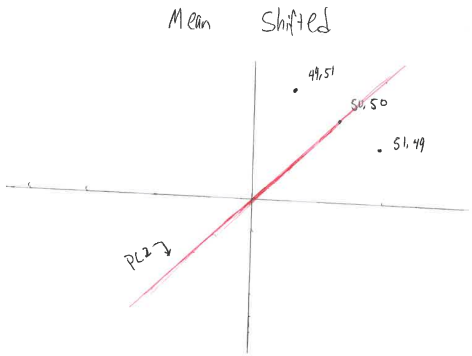


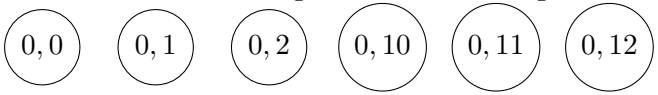
Figure 2.

2 Multidimensional Scaling and Isomap

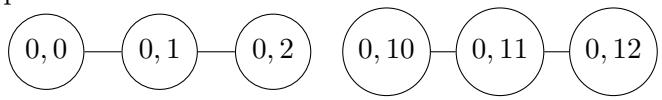
2.1 Argue that the process to obtain the neighbourhood graph G in the Isomap method may yield a disconnected graph. Provide an example. Explain why this is problematic.

The Isomap algorithm works similarly to MDS in that it relies on computing a Gram matrix S from a distance matrix D.However,dissimilarly to MDS,this distance matrix is not made up of euclidean distances. Instead, the distances are approximations of the shortest path from point A to point B where the path is restricted to the theoretical manifold our data distribution makes up. Computing the shape of this manifold as well as the length of the path through the manifold from point A to B is infeasible. An approximation of the manifold can be created by drawing a neighbourhood graph G. This neighbourhood graph G is created one of two ways. Either by joining each point to its K nearest neighbours, where a point A is considered closer to point B than to C if the euclidean distance between A-B is smaller than A-C. Alternatively the neighbourhood graph can be constructed by choosing to join a point A to all of its neighbours that lie within a certain distance d.

To demonstrate how this process can lead to a disconnected graph, assume we have the following nodes located at given coordinates.



If we go by the k-nearest neighbour approach and set k=2, or go by max euclidean distance between nodes and set d to smaller than 8, our resulting graph will be disconnected.



According to the Isomap algorithm, when we compute the distance matrix, the entry between point A and point B will be the summation of the distances between all nodes that lie along the shortest path between A and B through the neighbourhood graph G constructed in the previous step. If this is a disconnected graph, the distance between any point A to any point B where A and B are members of different disjoint subsets in G will be infinite, or undefined. In either of these cases the spectral decomposition of the Gram matrix corresponding to the distance matrix will be undefined and the Isomap algorithm will fail.

References

[1] *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2013.

3 Appendices

3.1 Appendix 1

```
import numpy as np
from sklearn.decomposition import PCA

X = np.array([[49, 51], [50, 50], [51, 49]])
X_mean_centered = np.array([[-1, 1], [0, 0], [1, -1]])
pca = PCA()
pca.fit(X)
print(pca.components_)
print(pca.singular_values_)
print(pca.explained_variance_ratio_)

pca.fit(X_mean_centered)
print(pca.components_)
print(pca.singular_values_)
print(pca.explained_variance_ratio_)
```