AM 170B HW1

Simon Lee

April 5, 2022

1 I. Deriving PCA from the SVD

This weeks focus was on dimensionality reduction. In today's data driven world, data comes in the form of high dimensions and it becomes very hard to interpret when their are multiple features consisting of these high dimensions. Therefore we used techniques like the Principal component analysis (PCA) and the Independent component analysis (ICA) to due two widely different problems in this homework.

1.1 PCA from SVD

The first part of our assignment was to extract the first two principal components from the ovarian cancer dataset from HW1 and plot a variety of graphs. However the main objective of this portion of the assignment was to get the principal components from the single value decomposition (SVD). We derived the pricipal components using the following set of equations.

$$\bar{X} = U\Sigma V^T \tag{1}$$

By taking the dot product of the Singular values diagonals of the V vector by the mean centered data we obtain our principal components.

1.2 Plotting singular values and plot the fraction of variance corresponding to each singular value

Before we begin plotting the principal components, we were first asked to plot the singular values and plot the fraction of variance corresponding to each singular value.

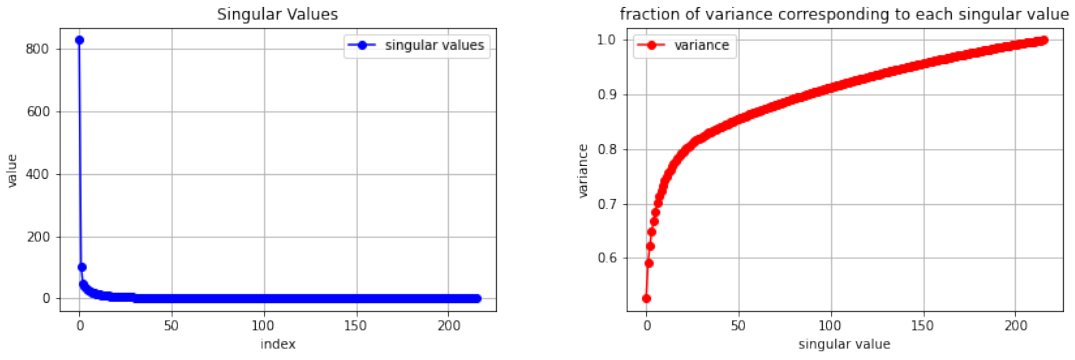

Figure 1: Plotting singular values and plot the fraction of variance corresponding to each singular value

The relationship of these two graphs are almost inverse of each other in the sense that the singular values are going in descending order while the fraction of variances are in ascending order.

1.3 Principal Component

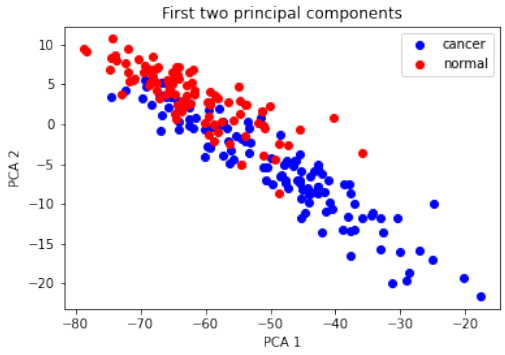

Figure 2: Plotting the first two principal components

In the final part of this problem we are asked to plot the first two principal components. In subsection 1.1 we explained how we derived this process from SVD. Therefore we were able to transform are very high dimensional dataset into a 2D representation. This representation can be seen in Figure 2.The data actually lines up really nicely and we can see that the cancer samples tend to spread out more vs. the normal samples.

2 II Signal decomposition.

In class we discussed another application of dimensionality reduction which was extracting unobserved signals from a multitude of incoming signals. This in my opinion is very cool and the objective of this assignment was to do this. However unlike problem 1, we have taken a different dimensionality reduction technique called the independent component analysis.

2.1 Plotting the original dataset

Lets first visualize our dataset by plotting the frequency of the signal over time. This can be seen below:

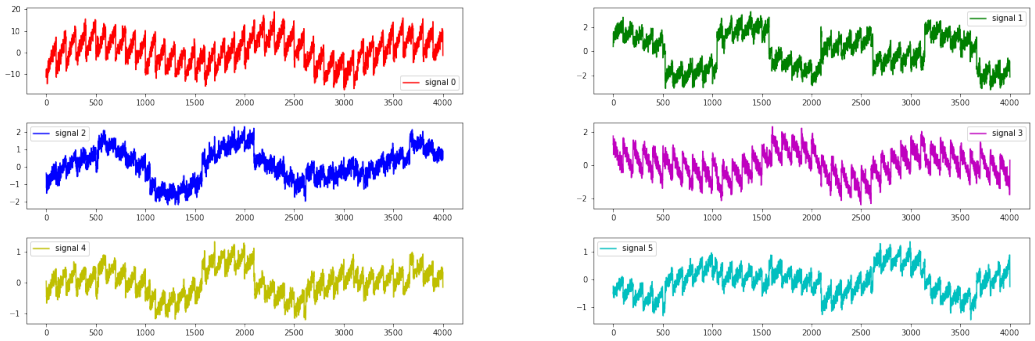

Figure 3: Plotting the mixed signals dataset over time

We can see that all these signals have different shapes and we did not going in on how many components gives us the cleanest unobserved signal. The key clue we were given was that they were well known mathematical functions.

2.2 ICA reduction

With the help of Professor Jonsson, we actually were told that the number of components was 4. Therefore we can see the the unobserved signals we have obtained by using ICA:

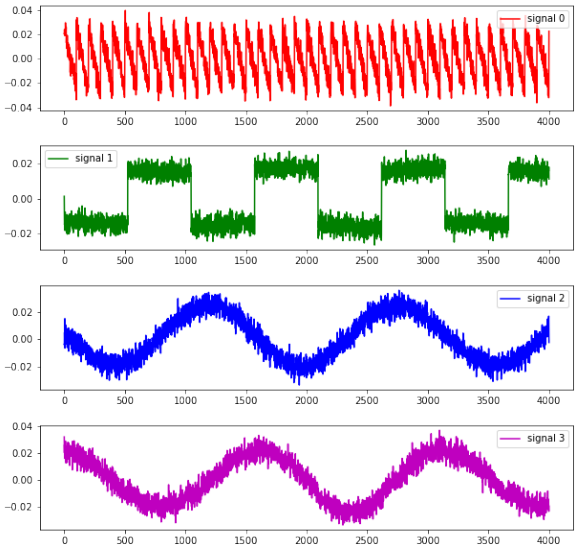

Figure 4: Plotting the unobserved signals using ICA

As expected we have obtained four signals that are very similar to well known mathematical functions. The first unobserved signal is what is called a sawtooth function. The sawtooth funciton takes the following piecewise function form:

$$x(t) = t - \lfloor t \rfloor \tag{2}$$

$$x(t) = t mod 1 \tag{3}$$

The second unobserved signal is a square wave which is commonly seen in Fourier analysis. Square waves take the mathematical form of a Fourier series and it can be generally defined by the following:

$$x(t) = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{\sin(2\pi(2k-1))}{2k-1} \tag{4}$$

And lastly unobserved signal 3 and 4 are our sinusoidal function sine and cosine. Because a transformation of some sort could be defined, we will consider both of them as a pair.

$$x(t) = sin(t) \tag{5}$$

$$x(t) = cos(t) \tag{6}$$

2.3 running PCA and its issues

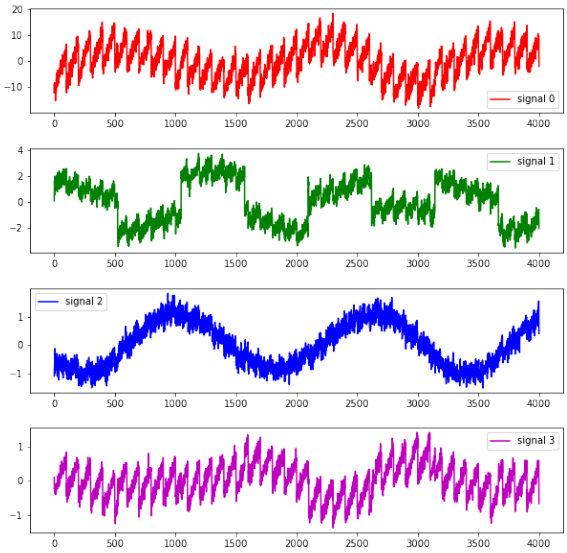

Figure 5: Plotting the unobserved signals using ICA

In Figure 5 we see the results from PCA. The biggest issue that we already see is that the unobserved signals are some mixture or conncatenation of mathematical functions. We see that in some cases we do get a nice signal like signal 2 but we can also see the rest of them are a mixture of the sawtooth function with a sinusoidal function or the square wave function. This is the biggest setback related to PCA in recovering unobserved signals. Therefore stemming from the results of both these techniques, we have determined that ICA is the better method for this type of work.

3 III Course Project

1. Describe at least one idea for your project (one paragraph).

In the field of research I am most particularly interested in disease onset. There are so many diseases that aren't totally understood and with the advances in data, I am hoping to run a problem in this field of image reconstruction, and computer vision. Alzheimer's disease is among one of those diseases that aren't very well understood, and I intend to look at a set of images within a 2-5 year frame and try to predict the brain structure. My co-adviser for this project, Razvan Marinescu ran a competition with this same objective and he has shown me a few code bases for which I can attempt to reconstruct the brain structures based on these predictions. Though this may not be the final project idea due to a list of personal caveats (my computer is slowly dying on me, my machine struggles to push stuff onto git, limited RAM/memory), I will definitely be in the exploring phases in the coming weeks on a final proposal.

2. List objective of the project in one or two sentences.

My tentative objective in my project is to reconstruct 3D brain structure based off temporal data given by ADNI.

3. List publicly available data sources to use. Describe the data source, data size, variables, and other aspects of that data that are relevant to your project. Justify its use to meet the objective you listed in 2.

ADNI is a public database ran by USC that holds MRI data of various human patients. This dataset fits the theme of the class and its dynamics due to the fact that it is a temporal dataset of images within a 2-5 year span. Though predicting Alzheimers is a relatively impossible task currently, we can predict whether a patient may develop this disease based on the analysis of the images which will help us predict their structure.