Homework Set 4

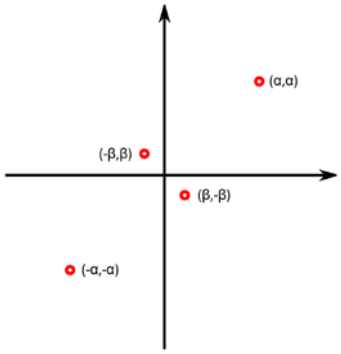
Problem 1 (Principal component analysis)
Joe's first job as a statistician was to analyze the countries that won medals in the 2012 London Olympic Games. He started his task by collecting data from 84 countries that had won at least one medal. The data's features (dimensions) for each country (sample) consisted of the number of medals won, the population size (in hundred thousands) and the GDP (in millions of US dollars) all taken in 2012. To start his analysis, Joe ran PCA on the data. The first eigenvector explained 99.9% of the data's variance. Joe's boss ran PCA by himself and he found that the first eigenvector explained only 74.0% of the data's variance. They compared what they both had done but the only difference they found was that Joe's boss used the GDP values in billions of dollars while Joe's GDP values were in millions of dollars.

Answer the following questions:

a) Why did they obtain different results?

b) How should Joe and his Boss have treated the data before the PCA?

Problem 2 (Principal component analysis)



a) Compute the covariance matrix $\mathrm{cov}(\boldsymbol{X})$ of the dataset composed of the four points shown in the figure [(α,α), $(-\alpha,-\alpha)$, $(-\beta,\beta)$, $(\beta,-\beta)$].

b) Compute the eigenvectors and eigenvalues of $\mathrm{cov}(\boldsymbol{X})$. Discuss how the eigenvectors and eigenvalues are affected by α and β.

Problem 3 (Principal component analysis)
PCA decomposition is a common approach for denoising. Suppose a signal $\boldsymbol{X}$ is measured. Assume the signal is additive. It can be decomposed into 2 matrices by using PCA, i.e., $\boldsymbol{X} = \boldsymbol{X_S} + \boldsymbol{X_N} = \boldsymbol{E_S}\boldsymbol{\Sigma_S}(\boldsymbol{F_S})^{\mathrm{T}} + \boldsymbol{E_N}\boldsymbol{\Sigma_N}(\boldsymbol{F_N})^{\mathrm{T}}$, where $\boldsymbol{X_S}$ and $\boldsymbol{X_N}$ respectively contain "information" and "noise" of the data. Typically noise is associated with a low level of variance in data. Hence, in the PCA decomposition, the first n eigenvectors that are associated with high variance are usually assumed to be the information. In contrast, the rest eigenvectors are assumed to be the noise. If one can identify the correct intrinsic dimensionality (e.g., correct n) of the data, he/she should expect to be able to filter out the noise while still capturing the important information in the data by reconstructing the signal $\boldsymbol{X_S}$ using only the first n eigenvectors.

This problem helps you practice the denoising using PCA. A dataset "04HW2_noisy.mat" consists of 1,965 gray-level images. These images are 20×28 in dimension and are distorted by adding Gaussian noises to each pixel with a standard deviation of 25. Import the dataset into MATLAB. To view the 10^{th} image, type:

```
>>colormap gray
>>imagesc(reshape(X(:, 10), 20, 28))'
```

Finish the following tasks:

a) Apply PCA to the noisy data. Suppose the correct intrinsic dimensionality of the data is 10 (i.e., only the first 10 eigenvalues and eigenvectors contain information; the rest eigenvalues and eigenvectors contain noise). Compute reconstructed images using the top 10 eigenvectors and plot the 10^{th}, 121^{st}, 225^{th}, 318^{th}, and 426^{th} original and reconstructed images.

b) Repeat part a), assuming the correct intrinsic dimensionality of the data is 2 and 30.

c) Determine the best intrinsic dimensionality of the dataset using the techniques learned from the class. You will need to check all the possible dimensionality. Explain the approaches you use and the reasons of choosing the best intrinsic dimensionality in details.