Homework 2

1. (20 points)

(S, T) and (U, V) form an orthogonal and non-orthogonal coordinate system, respectively, for the same space. Let $\overrightarrow{e}_s = [S_x, S_y]$ and $\overrightarrow{e}_t = [T_x, T_y]$ as the orthonormal basis for system $[S, T]$. Let $\overrightarrow{e}_u = [U_x, U_y]$ and $\overrightarrow{e}_v = [V_x, V_y]$ be unit vectors and form a basis for system $[U, V]$. Figure 1 depicts the coordinate systems, with $X - Y$ forming the original coordinate system, $S - T$ forming the orthogonal one, and $U - V$ forming the non-orthogonal one.

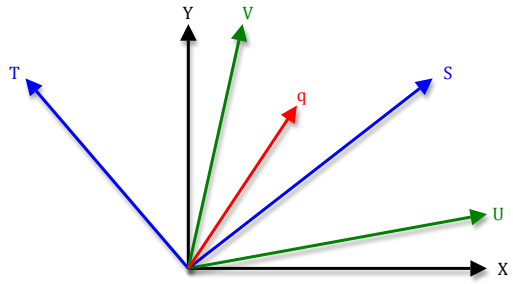

Figure 1: Different basis.

Let $\overrightarrow{q} = [q_x, q_y]$ be an arbitrary vector. Compute:
1). the coordinate of $\overrightarrow{q}$ in the space of $[S, T]$.
2). the coordinate of $\overrightarrow{q}$ in the space of $[U, V]$.
3). Can you explain why orthogonal basis makes things easier to handle?

2. (40 points)

Write your own class in Python to perform PCA analysis. Require that your input data follow the format where rows are samples and columns are variables. Your function should do the following:

(1). Do mean-centering.
(2). Compute the covariance matrix using `numpy.cov()`.
(3). Perform eigen-decomposition using `numpy.linalg.eig()`.
(4). Project the data onto the principal component axes.
(5). Return the variance and percent variance that each PC explains, all of the scores, and loadings.

3. (20 points)

Apply your own PCA function to the dataset "Homework_2_dataset_prob3.csv". In this dataset, columns correspond to variables.

- Plot the scores plot.
- Do you see a clear separation of the raw data?
- Can you still separate them after you project your raw data onto your first principal component?
- What message can you get from this observation?
- What is the variance of the projections on PC1 and PC2. What is the relationship between these variances and the eigenvalues of your covariance matrix?

4. (20 points)

Apply your own PCA function to the dataset "Homework_2_dataset_prob4.csv". In this dataset, rows correspond to different variables and columns correspond to different samples. You will need to transpose it before PCA analysis because your own PCA function requires columns to be variables. Plot the scree plot for PC1 and PC2 and indicate the percentage of variance that PC1 and PC2 explains respectively. Also plot the scores plot and loadings plot.