# Homework 4 Question 2

Sahil Shah and Preey Shah

November 2017

The actual code that has been used to perform the necessary computations (matlab code) is in errors.m. In p2.sc, resides the code used to generate the boxplot. In scatterPlot.m, we write the code to generate the scatter plots that are needed. In the files data<i>.png, we have the scatter plots for sizes $10^i$. The required boxplots are in the files errorMean.png and errorsCovariance.png

## 1 Generating the Gaussian

Steps to generate multivariate Gaussian-

- Given the Covariance matrix $C$, calculate $A$, where $A$ is the matrix from which multivariate gaussian can be derived from independent gaussian decompositions (that is $A$ is such that $C = AA^T$). Take the covariance matrix as $C$. Given that $C = AA^T$. The eigen-decomposition of $C$ is given by $Q\lambda Q^{-1}$.Then $A$ is given by $Q\lambda^{1/2}Q^{-1}$ (can be verified by putting back the value). Using the *eig* function $Q$ and $\lambda$ are calculated. This way $A$ is calculated
- Using the given values of $C$ and $\mu$, $\lambda_1 = [0.5000, 0], \lambda_2 = [0, 5.0000]$. The value of $Q$ is $Q_1 = [-0.8660, -0.5000], Q_2 = [-0.5000, 0.8660]$. Using these values , $A_1 = [1.0893, -0.6621], A_2 = [-0.6621, 1.8538]$
- Now generate two independent gaussians. The transformation is given by $X = \mu + AW$. for every two values, apply the tranformation. This way you will get multivariate gaussian data.

## 2 Maximal estimates

The ML estimate for mean is given by $\mu_{MLE} = E(X)$ and for covariance matrix is given by $C_{MLE} = E((X - E(X)(X - E(X))^T))$. Now substituting the data.

As can be clearly seen the the value tends to the original estimate as $N$ increases

# 3 PCA

The principal mode of variation is along that vector on which the variance of projection is maximum. This vector is the eigenvector corresponding to which the eigenvalue of the covariance matrix is maximum. We the plot a line of length of that eigenvalue in the direction of this eigenvalue.
As can be seen from the diagram, this vector goes in the direction where there is maximum data. As sample points increases, it becomes a better representation of the data.