INDUSTRIAL & SYSTEMS ENGINEERING
TEXAS A&M UNIVERSITY

FINAL COURSE PROJECT

Team 23
Dwarkanath Prabhu
Venkata Kartik Mutya

Table of Contents

EXECUTIVE SUMMARY

Our objective of this project was to identify the in-control and out-of-control samples. Due to high dimensionality, the noise components can add up to a great magnitude. As a result, the aggregated noise effect can overwhelm the signal effects thus making it harder to reject the null hypothesis. This phenomenon is known as *curse of dimensionality*. In this report since the number of dimensions are very high, we used principal component analysis (PCA) as the data reduction tool to reduce the data points and then used the Hotelling T^2 chart on the reduced data to isolate the in-control samples.

First, we calculated the mean vector and covariance matrix of the given data. Then, we calculated eigenvalues and eigenvectors of S to find the reduced dimension. These eigenvectors were used to form principal components from the original data. For the S matrix, we calculated the eigenvalues and arranged them in descending order. Thereafter, we plotted a graph and observed that the value of L for which MDL is minimum is 35. As 35 Principal Components (PC's) are not nearly small enough we then used scree plot i.e. the plot of eigenvalues against the number of principal components to further compress data and reduce the Principal Components (PC's). From the scree plot, we observed that there is a bend where the x-axis value is 4. Therefore, we chose only the first 4 principal components for our analysis.

For Principal Component Analysis (PCA), we calculated the vector y of principal components and then performed Phase I analysis on y. While performing the Phase I analysis of y, we approximated the upper control limit to 9.49 using a χ^2 distribution. We then plotted the Hotelling T^2 statistic for each sample. To isolate in-control data, we removed out-of-control samples and recalculated the T^2 statistic till we were left with only the in-control samples. i.e there were 461 in-control samples.

PROBLEM DESCRIPTION

We are provided a set of data collected from a manufacturing process, in which both the in-control and out-of-control data are present. We are asked to develop a method or a procedure to identify the data falling in the respective categories, i.e., which ones are in-control and which ones are out of control. This is amounted to a Phase I analysis, whose purpose is to isolate the in-control data for estimating the in-control distribution parameters so that a monitoring scheme can be set up for future missions.

Description of the dataset

The problem at hand has 552 samples, each with 209 data points.

- n = 552
- p = 209

This can be denoted as $\{x_j\}$, j = 1...552 and each x_j is a 209 x 1 vector. The task at hand is to identify in-control and out-of-control samples. The μ_0 and Σ_0 for this data are not known. Hence, this is a Phase I analysis with a sample size of 1. We will use $\bar{x}$ and S to estimate μ_0 and Σ_0.

Since the number of dimensions are very high, we will first reduce data using principal component analysis and then use the Hotelling chart to isolate in-control data.

APPROACH

Principal Component Analysis

To interpret the data in a more meaningful form, it is necessary to reduce the number of variables to a few, interpretable linear combinations of the data where each linear combination will correspond to a principal component. In order to achieve this reduction in the dimension of data we use Principal Component Analysis. PCA is a statistical procedure which converts the set of observations of possibly correlated variables into a set of values that are linearly uncorrelated variables using orthogonal transformation. The linearly correlated variables are said to be Principal Components.

After applying PCA to the original data set, we will have the same number of PCs as the number of elements in the original vector. In order to reduce the data dimension, we can only retain the first few principal components, corresponding to the largest values in eigenvalue.

The number of PC's to be kept can be decided using data reduction techniques such as:

Minimum description length (MDL) criterion

$$MDL\ (l) = n(p-l)\log\left(\frac{a_l}{g_l}\right) + l(2p-1)\log(n)\ /2$$

where n is the data sample size, p is the dimension of the covariance matrix. a_l and g_l are the arithmetic and geometric mean of the p-l smallest eigenvalues, respectively.

• MDL(l) is evaluated for l = 0, 1, ..., p − 1 and the number of PC to retain is chosen as the l that minimizes MDL(l).

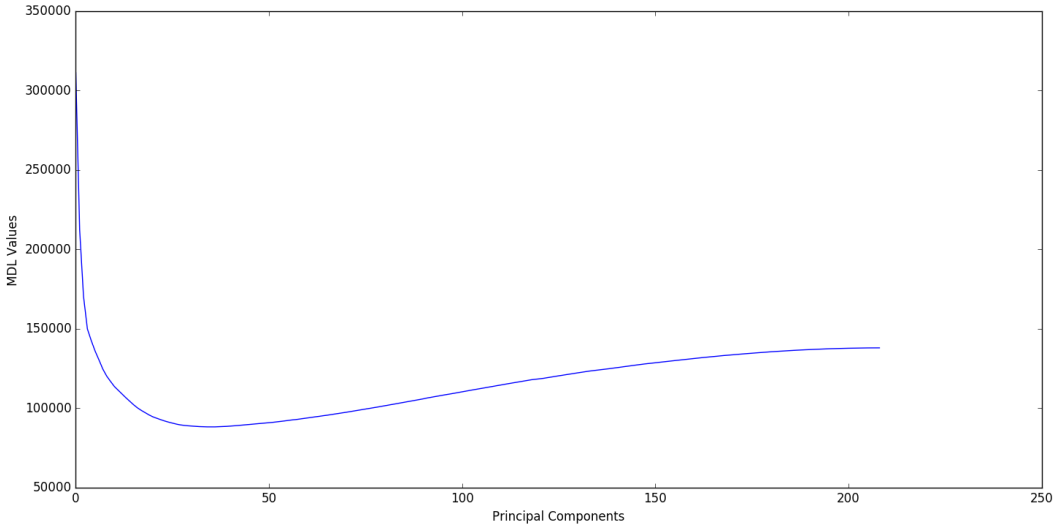
• Sometimes MDL(l) can retain too many eigenvalues. So it is a better practice to use MDL together with the scree plot.

For our problem we needed to estimate the covariance matrix (S). This was done using the equations:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We exported the data from excel using xlrd library in python. We then used the numpy library of python to calculate eigenvalues and eigenvectors of S. Using python, the MDL values were calculated for l = 0, 1, ..., 551 and plotted using matplotlib library.

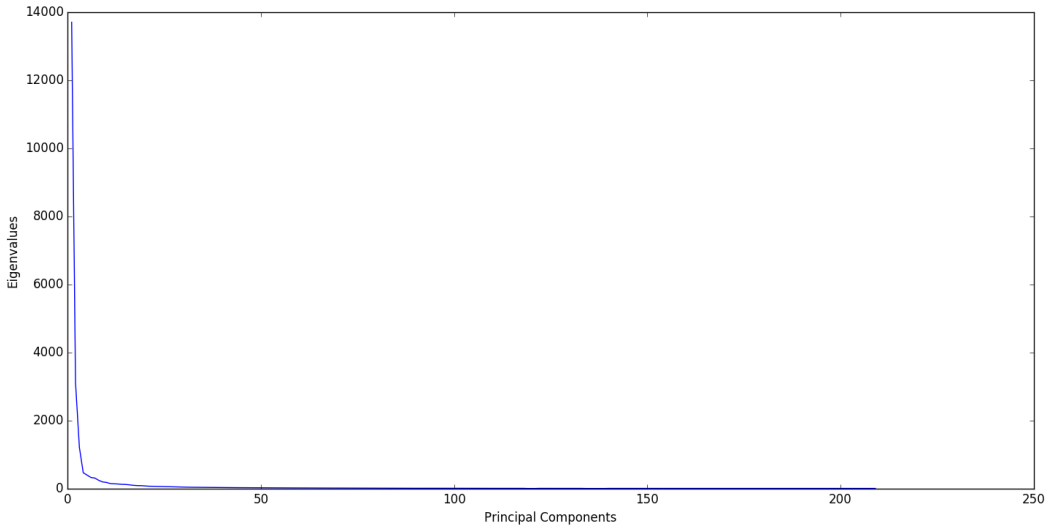


From the calculated MDL values, it was seen that the minimum was achieved when l = 35. However, 35 is also a large number of dimensions. Hence, we looked at the Scree plot.

Scree plot

With the eigenvalues ordered from largest to smallest, a scree plot is a plot of λ_i versus i. We look for an elbow (bend) in the plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

We plotted the scree plot using the matplotlib library in python.



From the graph, it can be seen that there is a bend at i = 4 i.e. the eigenvalues after this are very small and remain relatively constant. Hence only the first 4 principal components were chosen

Calculating principal components

The i^{th} principal component of data can be calculated using:

$$y_i = e_i^T x$$

Since, we are using only the first 4 principal components, i = 1,..4. Thus we can create a vector y of size 4x1. In the original data, we had 552 samples. So we get 552 such vectors. We conducted Phase I analysis on these vectors to isolate in-control samples.

Hotelling T^2 chart

This chart is used to isolate the in-control data.

Phase I: Phase I is used to identify the in-control training data (which are used to estimate the distribution parameters). Typically, we apply a chart to the training data to see if the training data are really in control. Then we remove all the out-of-control data and iterate until all training data are in control.

The formula for T^2 statistic for a sample size of 1, is given as

$$T^2 = (X_J - \overline{X})^T * s^{-1} * \left(X_J - \overline{X}\right) \;\; with \;\; UCL = \chi^2_{1-\alpha}(p)$$

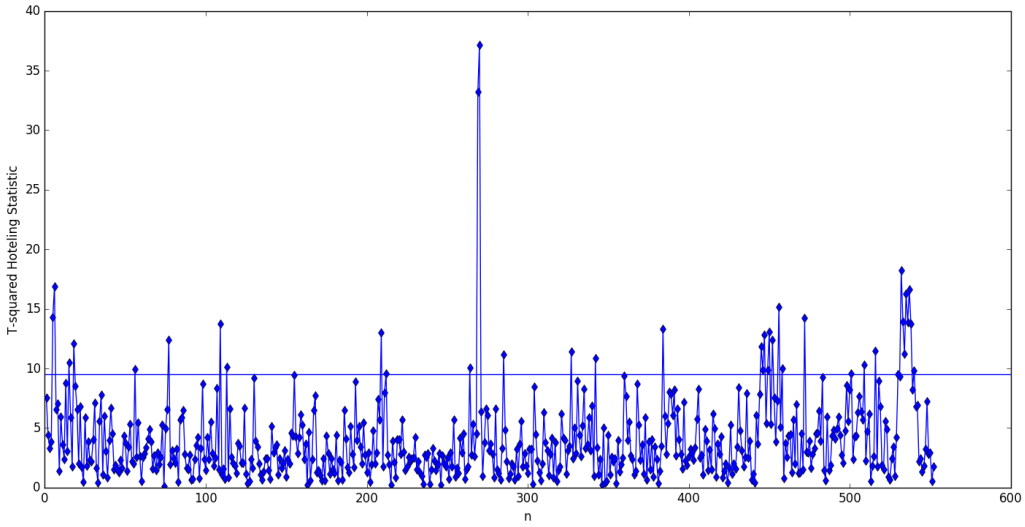
Here X_j is the j^{th} sample. In our question, the number of samples is 552 and p, dimension of the data, is 4. Hence there will be 552 such statistics.

If we choose $\alpha = 0.05$, we get:

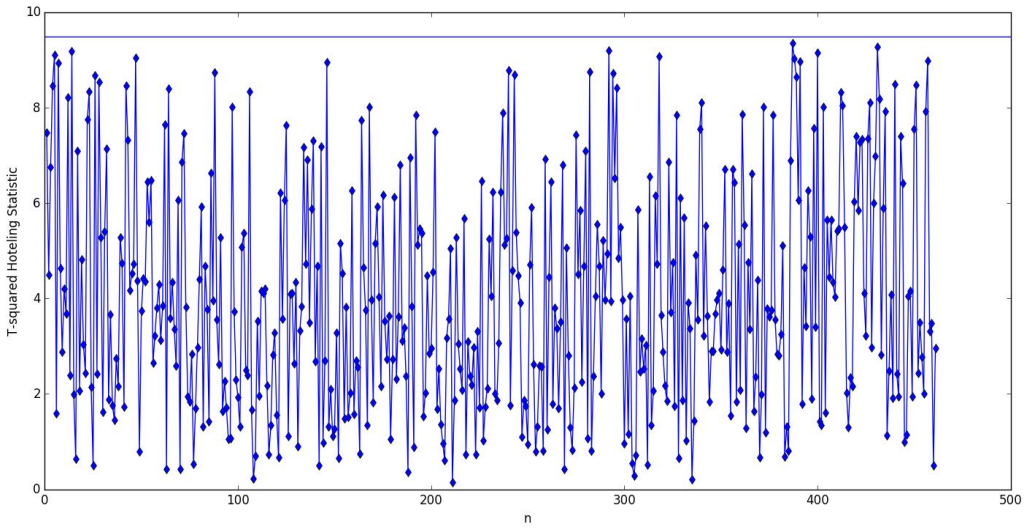
$$UCL = \chi^2_{0.95}(4) = 9.49$$

We calculated the T^2 statistic for the reduced data vector y and compared it with the above UCL. Then we plotted this statistic using matplotlib library in python.

The first iteration of the T^2 statistic showed several out of control samples.

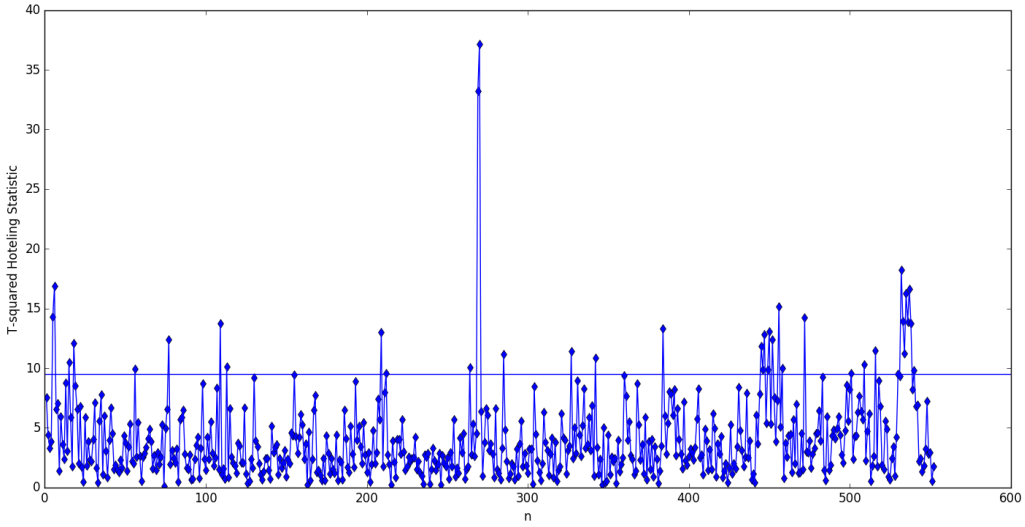


After a few iterations of removing out-of-control samples and recalculating the T^2 statistic, there were 461 in-control samples left.



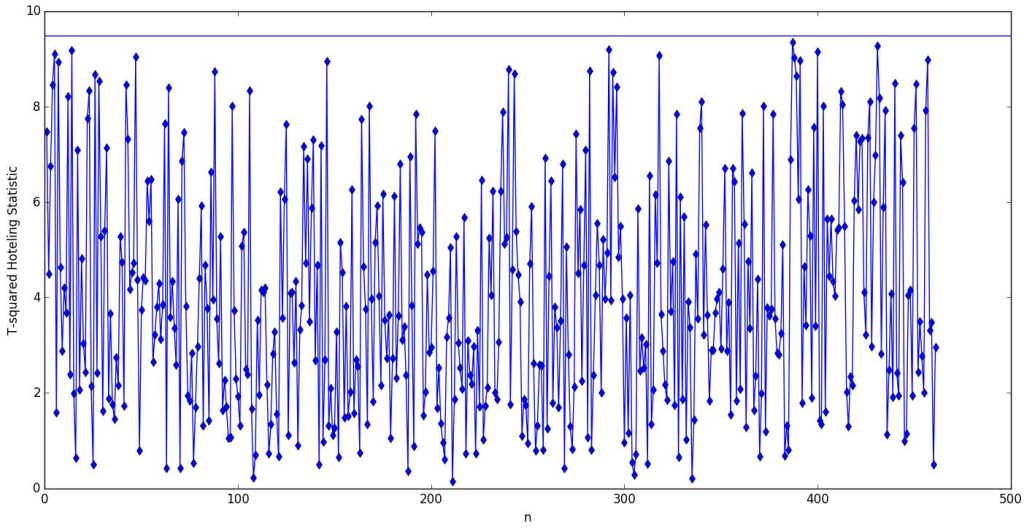
RESULTS

Post analysis we observe that a total of 91 data points are beyond our UCL of 9.49.



This graph indicated all the data points which are in and out of control

Therefore, we remove those 91 out-of-control data points and recalculate the T^2 statistic in order to prove that all the remaining 461 data samples are in-control.



This graph displays all the 461 data points which are in control

CONCLUSION

The purpose of this project was to enable teams to gain hands-on experience regarding developing methods for quality control and anomaly detection.
The project was based on a set of data collected from a manufacturing process, in which both in-control and out-of-control data are present. We developed a procedure to identify the data falling in the respective categories, i.e., which ones are in-control and which ones out of control. This is amounted to a Phase I analysis, whose purpose is to isolate the in-control data for estimating the in-control distribution parameters so that a monitoring scheme can be set up for future missions. Initially, to reduce the number of variables to a few i.e reduction in the dimension of data. We used Principal Component Analysis. After applying PCA to the original data set, we still had the same number of PCs as the number of elements in the original vector. So we decided the number of PC's to be kept using two data reduction techniques namely: MDL and Scree Plots.

First, we estimated the covariance matrix (S) using the equations:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We then exported the data from excel using xlrd library in python. Post which we used the numpy library of python to calculate eigenvalues and eigenvectors of S. Using python, the MDL values were calculated for l = 0, 1, ..., 551 and plotted using matplotlib library.
From the calculated MDL values, we observed that the minimum was achieved when l = 35. Since, 35 was also a large number of dimensions we then plotted the scree plot using the matplotlib library in python and observed a bend at i = 4 i.e. the eigenvalues after this were very small and remained relatively constant. Therefore, we only choose the first 4 principal components.
We then conducted Phase I analysis using the Hotelling T^2 chart to isolate in-control samples. The first iteration of the T^2 statistic showed 91 out of control samples. We then removed the out-of-control samples and recalculated the T^2 statistic to observe that there a total of 461 data points which are in-control.