Phase I Analysis on multivariate data to estimate the in-control distribution parameters

SUBMITTED BY:

Team 18

Naga Ramesh Kamisetti (UIN: 724007881)

Tejaswi Petluri (UIN: 524004797)

Executive Summary

This report provides insight into the applied techniques and methods used in the Phase-I analysis of a large dimension multivariate data taken from a manufacturing process. The main objective is to identify and isolate the in-control data which will help estimate the in-control distribution parameters to design a monitoring scheme for future observations. Initial analysis proved to be difficult owing to the large dimensions of the data (552 x 209) and the resulting T-squared chart of the original data identifies very few out-of-control points due aggregated noise that leads to higher control limits and lower efficiency in detection. To mitigate the effects of aggregated noise, Principal Component Analysis is used to identify the vital few principal components that can explain 80% of the data. Pareto analysis, Scree plots and MDL plots are used in combination to identify the four principal components, of the centered data, capable of explaining 80% of the manufacturing data reducing the dimensions from 209 to 4. Multiple univariate charts, xbar charts, are plotted using the uncorrelated principal components to identify and eliminate out-of-control data. Since the multiple univariate charts are not capable of identifying trends and shifts, a multiple-CUSUM chart is plotted on the reduced data obtained from the multiple-univariate charts. The data is further reduced by plotting mCUSUM charts iteratively to remove outliers until the chart does not identify any more out-of-control data points. The final mCUSUM chart is in-control but it does not explain or signal the presence of spikes in data and hence, a T-squared chart is plotted to identify remaining out of control data. Multiple iterations of the T-squared chart result in an in-control T-squared chart that can now be used to monitor future observations. After multiple iterations of the univariate, mCUSUM and T-squared charts a total of 189 data points are eliminated. Multiple univariate charts are plotted on the reduced data to revalidate the in-control data points. This data can then be used to monitor future observations.

Abstract

Customers are the most important part of the supply chain and by use of quality control in manufacturing we can ensure that customers receive products that they are willing to pay for. The main objective is also to provide products free from defects that will in turn reduce the probability of a recall. Recall of products is very expensive for the company and hazardous to the customer based on the product. Improvement in quality also gives a company the much-needed competitive edge. Anomaly detection and testing in the manufacturing process is important and can be done using several statistical tools in use today to identify defective and or abnormalities in the product. The objective of the project is to analyze the multivariate high-dimension manufacturing data to identify and isolate in-control data from the given data. Phase-I, typically, is an iterative process to identify and isolate in-control data by removing out-of-control data observed after each iteration. This reduced data is then used to estimate the in-control design parameters which will be used to monitor future observations. This phase involves dealing with high dimension data that can be reduced to a vital few principal components that can explain the variation in the majority of the data. The outliers in the reduced data are identified by use of multiples univariate charts followed by mCUSUM charts. In case mCUSUM fails to explain the remaining out-of-control data, T-squared charts are plotted. After this iterative analysis, we obtain clean, in-control data which is then used to identify the in-control distribution parameters.

Introduction

A multivariate, high-dimension manufacturing data is given to us. Phase-I analysis on the dataset needs to done to establish the in-control design parameters. The original data consists of 552 rows and 209 attributes. These attributes need to be reduced in order to mitigate the effects of aggregated noise and simplify the analysis process. Even though the individual noise is relatively small, the aggregated noise overwhelms the signal effects and makes it harder to reject the null hypothesis i.e., identify in-control data. Upon plotting the T-squared chart on the original data using $ARL_0 = 370$ and $\alpha = 0.0027$, we observe a total of 8 out-of-control points out of the total 552 which is very less considering the high-dimensionality of the data.

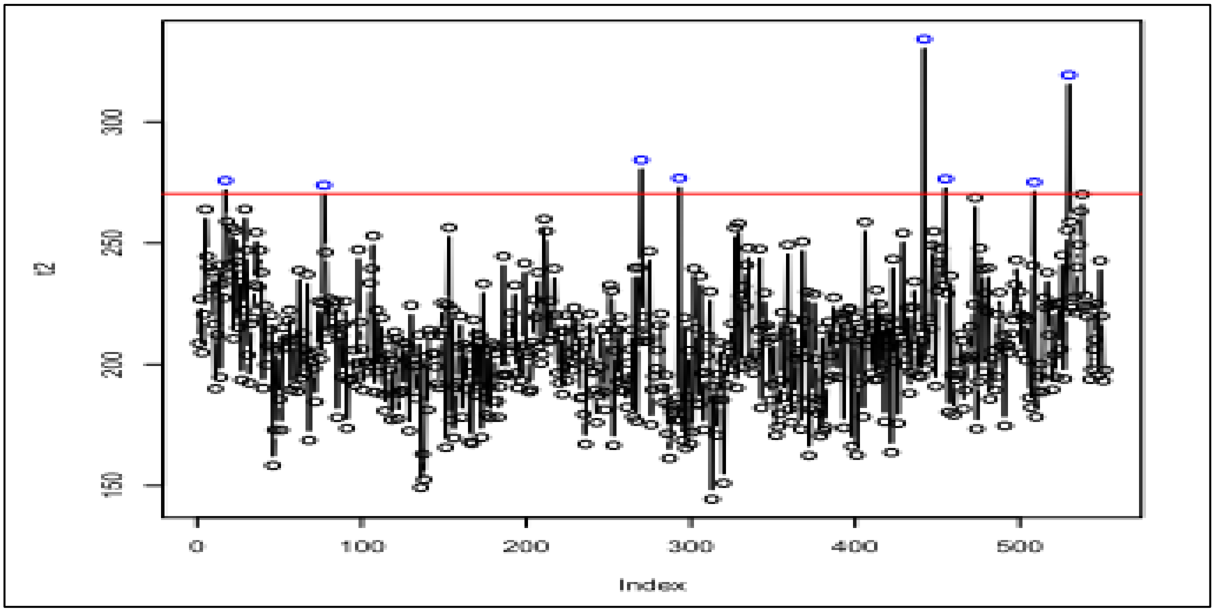


Fig 1: T^2 chart of the entire data with $ARL_0 = 370$ and $\alpha = 0.0027$

Principal Component Analysis

To overcome this curse of dimensionality, we use Principal Component Analysis. Using PCA, we identify the “vital few” principal components that are capable of explaining 80% of the variation in the entire data. Principal Component Analysis gives us 209 principal components. Each component explains the variation in a given direction. The first principal component is aligned along the direction of maximum variation and the following components are aligned in a particular way to make all the components uncorrelated in decreasing order of variation. Of the 209 principal components, we pick a few components that can explain the majority of the data, we have used four principal components that explain 80% of the data.

Pareto, Scree and MDL plots were plotted to help identify the major principal components. As we see in the MDL plot, the plot identifies 35 principal components which is a slightly larger number than what we prefer. We used Scree and MDL plots to identify the first four principal components as the major components.

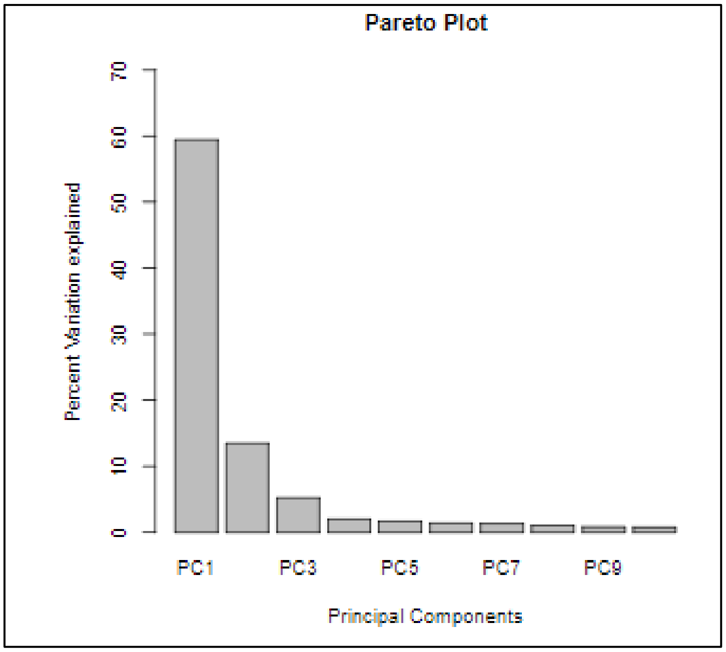


Fig 2: Pareto plot illustrating the percentage contributions of principal components.

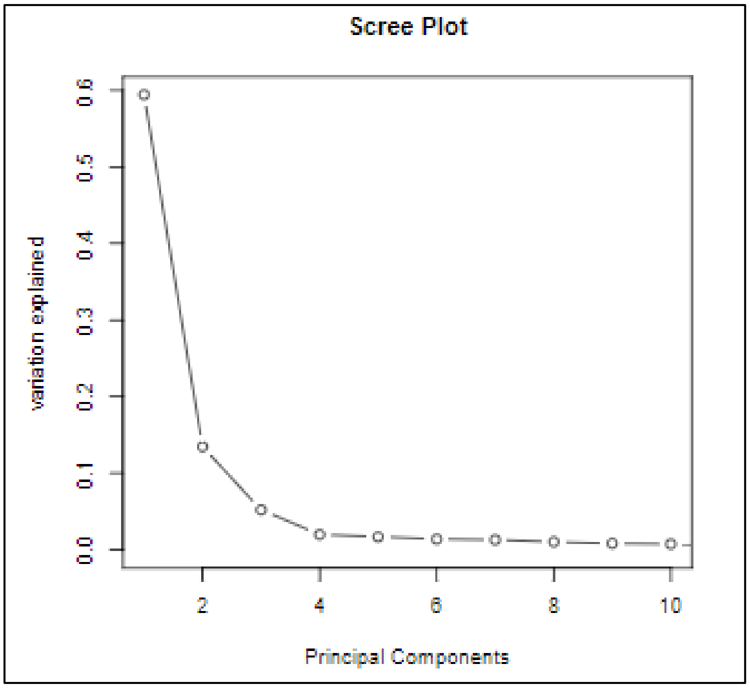


Fig 3: Scree plot indicating an elbow at PC=4.

MDL plot identifies 35 major principal components which is a large dimension data to work with. We use Scree and Pareto plots of the eigen values in combination with the MDL plots to identify the "vital four" components. We observe an elbow at PC=4 in the scree plot above after which the eigen values seem to be very small. The Pareto plot identifies the percentage of data that the principal components can explain and we notice that four PCs can explain up to 80% variation in the data.

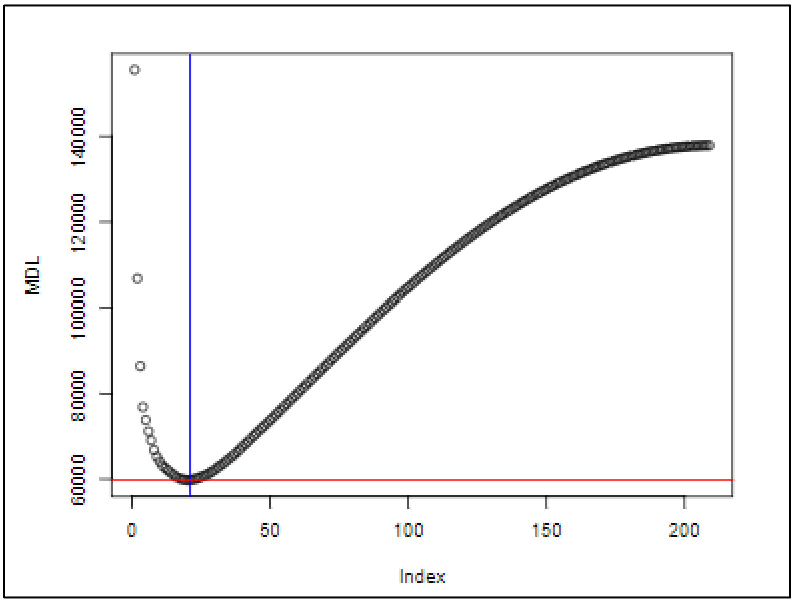


Fig 4: MDL plot illustrating the minimum value at 35.

After PCA, the dimensions of the reduced data were 552 rows and 4 attributes.

Data Reduction

After PCA, we identify four principal components of all the 209 attributes of the original data. These four **principal components are uncorrelated** and we can monitor these components using multiple univariate xbar charts. However, there will be an inflation in α and β errors and since we are using multivariate data, these values need to be adjusted so that the combined $ARL_0 = 370$, ARL_1 ~2, and the combined α=0.0027. Using these α and β values, multiple univariate charts for each of the principal components were plotted iteratively and the out-of-control data points were identified and removed.

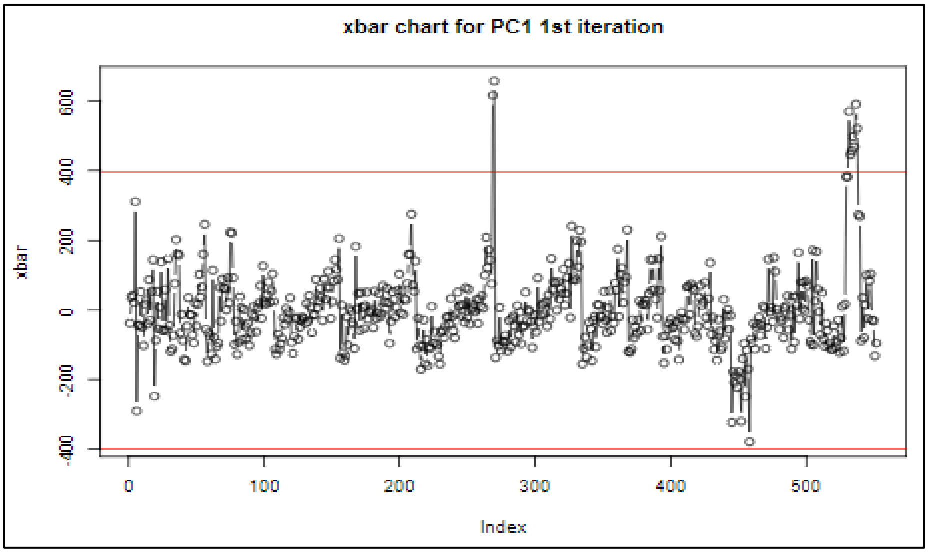


Fig 5.1.a: Univariate xbar chart for PC1 for the reduced data

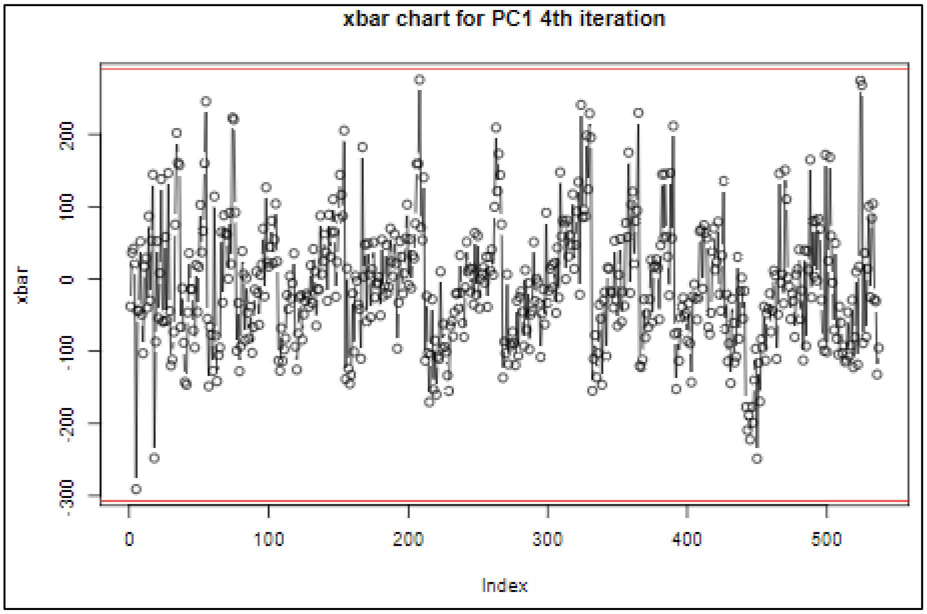


Fig 5.1.b: Univariate xbar chart for PC1 with in-control data

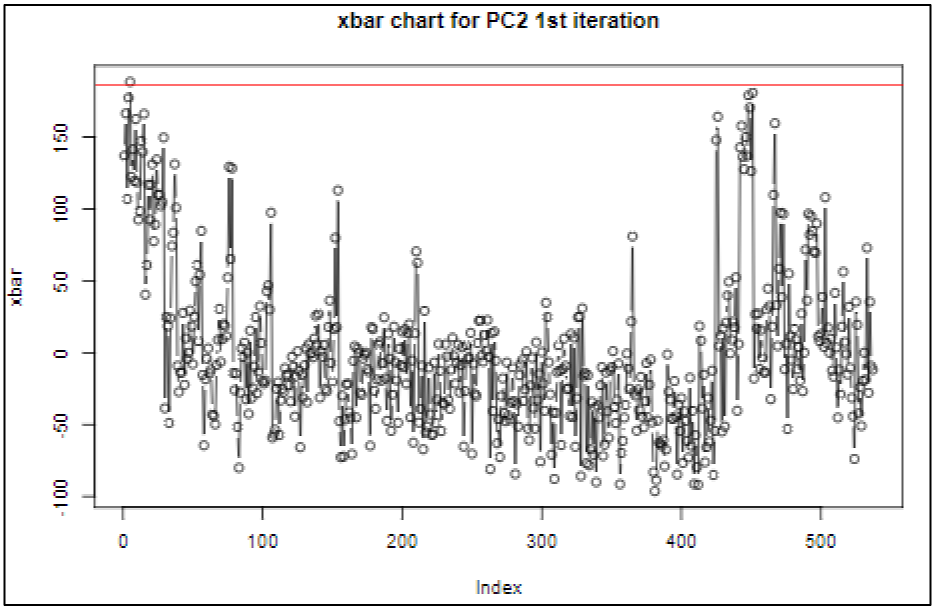


Fig 5.2.a: Univariate xbar chart for PC2 for the reduced data

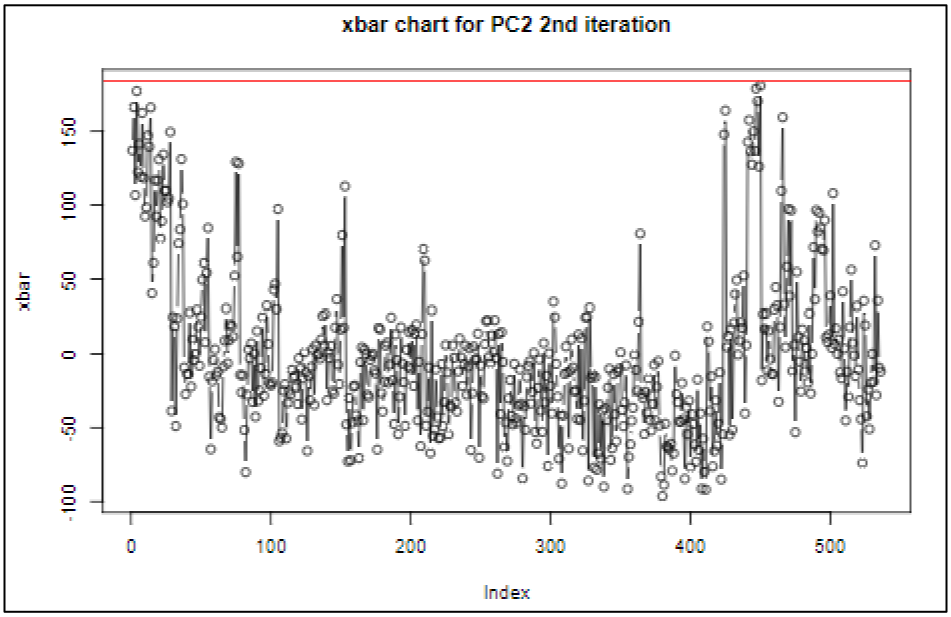


Fig 5.2.b: Univariate xbar chart for PC2 with in-control data

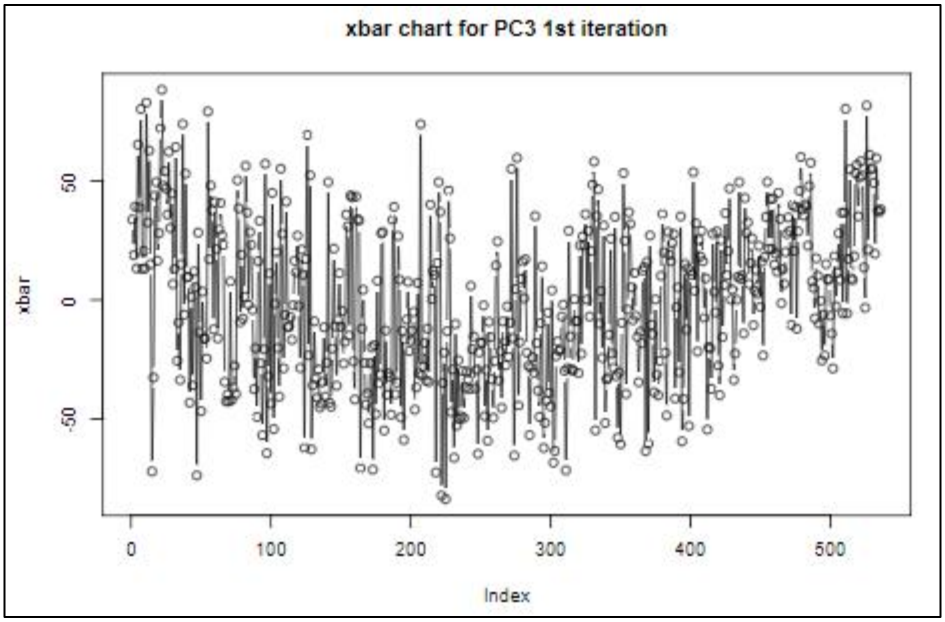


Fig 5.3: Univariate xbar chart for PC3 with in-control data in the first iteration

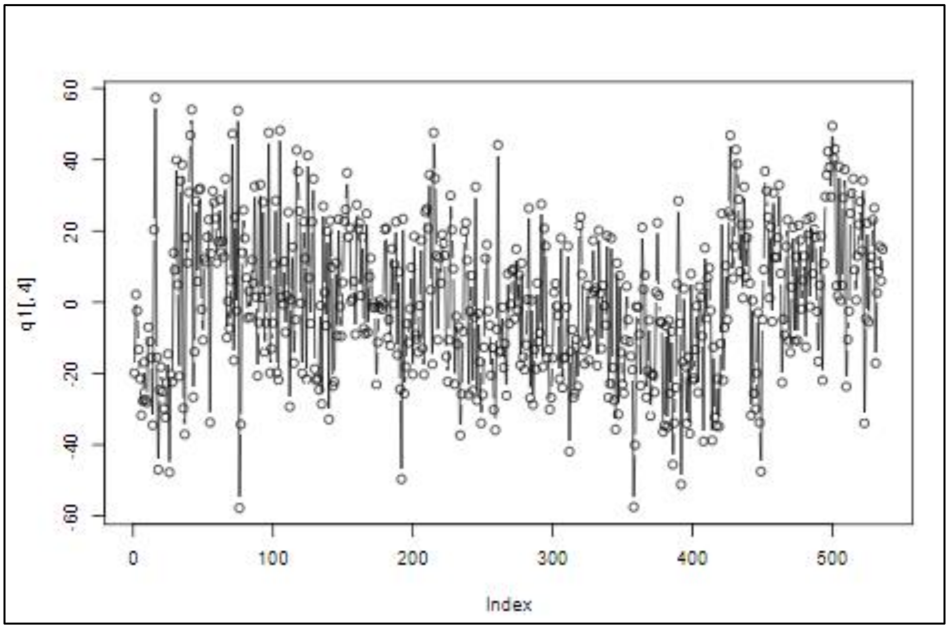
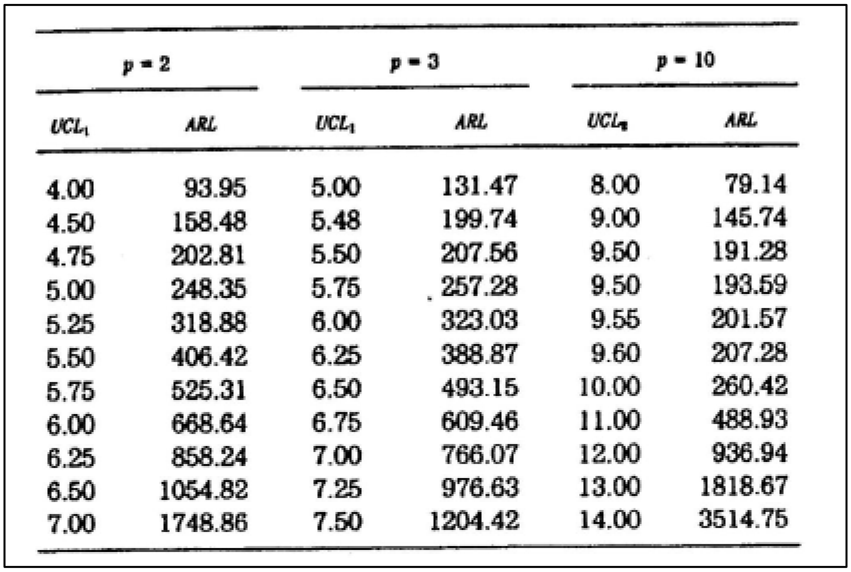


Fig 5.4: Univariate xbar chart for PC4 with in-control data in the first iteration

m-CUSUM

To improve the sensitivity of detection, and to identify any out-of-control points the univariate xbar charts failed to identify, we plot m-CUSUM charts on the reduced data. m-CUSUM chart has two parameters K and h. For this reduced data, we have used **k=1.5 and h=6.25**. "K=1.5" was arrived at by using half the statistical distance between μ_1 and μ_0 and there was a need to use a reasonable threshold to accommodate natural fluctuations in the data. "h" is calculated by interpolation of the existing h values in the literature.



$p = 2$		$p = 3$		$p = 10$	
UCL_1	ARL	UCL_1	ARL	UCL_1	ARL
4.00	93.95	5.00	131.47	8.00	79.14
4.50	158.48	5.48	199.74	9.00	145.74
4.75	202.81	5.50	207.56	9.50	191.28
5.00	248.35	5.75	257.28	9.50	193.59
5.25	318.88	6.00	323.03	9.55	201.57
5.50	406.42	6.25	388.87	9.60	207.28
5.75	525.31	6.50	493.15	10.00	260.42
6.00	668.64	6.75	609.46	11.00	488.93
6.25	858.24	7.00	766.07	12.00	936.94
6.50	1054.82	7.25	976.63	13.00	1818.67
7.00	1748.86	7.50	1204.42	14.00	3514.75

Fig 6: Literature results for h values for a corresponding p and ARL values.

m-CUSUM chart revealed several out-of-control data points which were removed by plotting the m-CUSUM charts iteratively removing the outliers after every iteration. The iterations are summarized in the table below.

Serial Number	Iteration number	Number of points removed	Samples removed	Number of data points remaining in the dataset
1	1st	74	5-49 444-472	462
2	2nd	35	3,4 50-58 77, 442, 443 473-478 492-506	427
3	3rd	28	2, 425, 482 490, 491 507, 509-511,	399

			515, 516, 518-524, 526, 529-536	
4	4th	12	76, 330, 483, 485-489, 508, 514, 517 and 525	387
5	5th	10	438, 440, 479, 480 481, 484 512, 513, 527 and 528	377
6	6th	3	436, 437 and 439	374
7	7th	1	435	373
8	8th	1	434	372
9	9th	1	433	371
10	10th	0	-	371

Table 1: Summary of out-of-control data points in m-CUSUM iteration

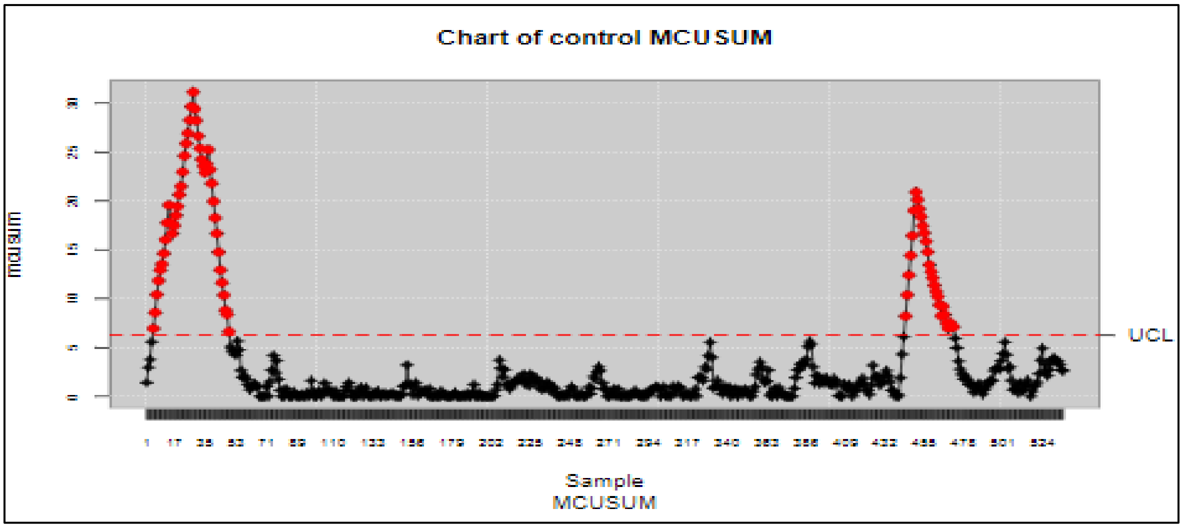


Fig 7a: m-CUSUM chart on the reduced data after PCA

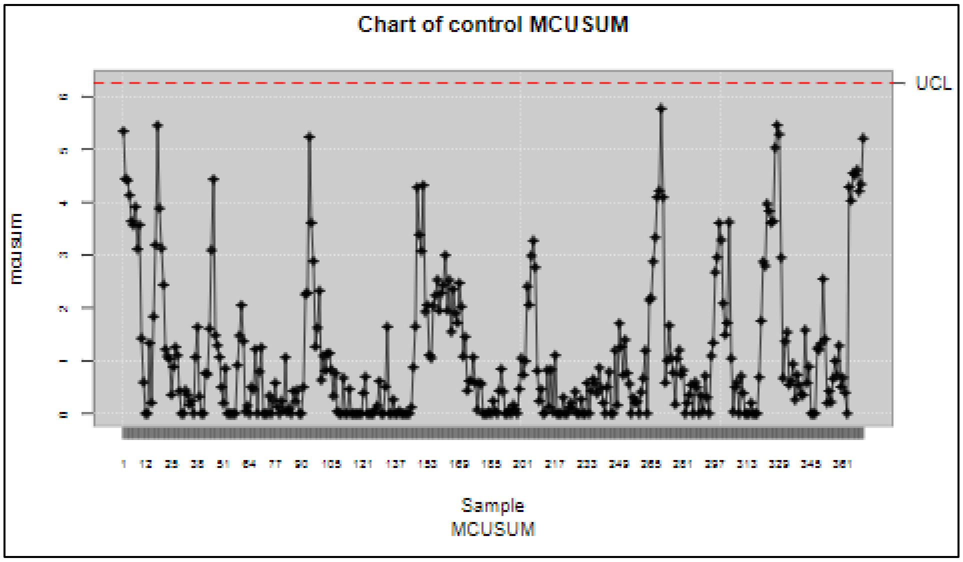


Fig 7b: m-CUSUM chart after the tenth iteration showing in-control data

Upon studying the m-CUSUM chart for in-control data, we notice the presence of large spikes which m-CUSUM does not recognize as out-of-control data. m-CUSUM fails to identify this because of the use of threshold values which it subtracts from the observation thus indicating it as in-control, making it less sensitive. To overcome this drawback, we use T^2 charts to identify the remaining out-of-control data.

T^2 chart

We use T^2 charts corresponding to Case – III(a) on the reduced data (371 x 4) using combined ARL_0 =370 and $\alpha_{combined}$ =0.0027. Multiple iterations of T^2 charts revealed several out-of-control data points were removed.

T^2 statistic used: $T^2 := (\mathbf{x}_j - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$

UCL used: $\chi^2_{1-\alpha}(p)$ where p=4, α = 0.0027

The out-of-control samples and the iterations are summarized in the table below.

Serial Number	Iteration number	Number of points removed	Samples removed	Number of data points remaining in the dataset
1	1st	5	1, 18, 94, 148 and 364	366
2	2nd	2	4, 300	364
3	3rd	1	89	363
4	4th	0	-	363

Table 2: Summary of out of control samples after T-squared charts' 4 iterations

The T^2 charts for the following iterations are below. At the end of the T^2 analysis, the entire data is in-control and the dimensions of the data are 363 rows and 4 columns.

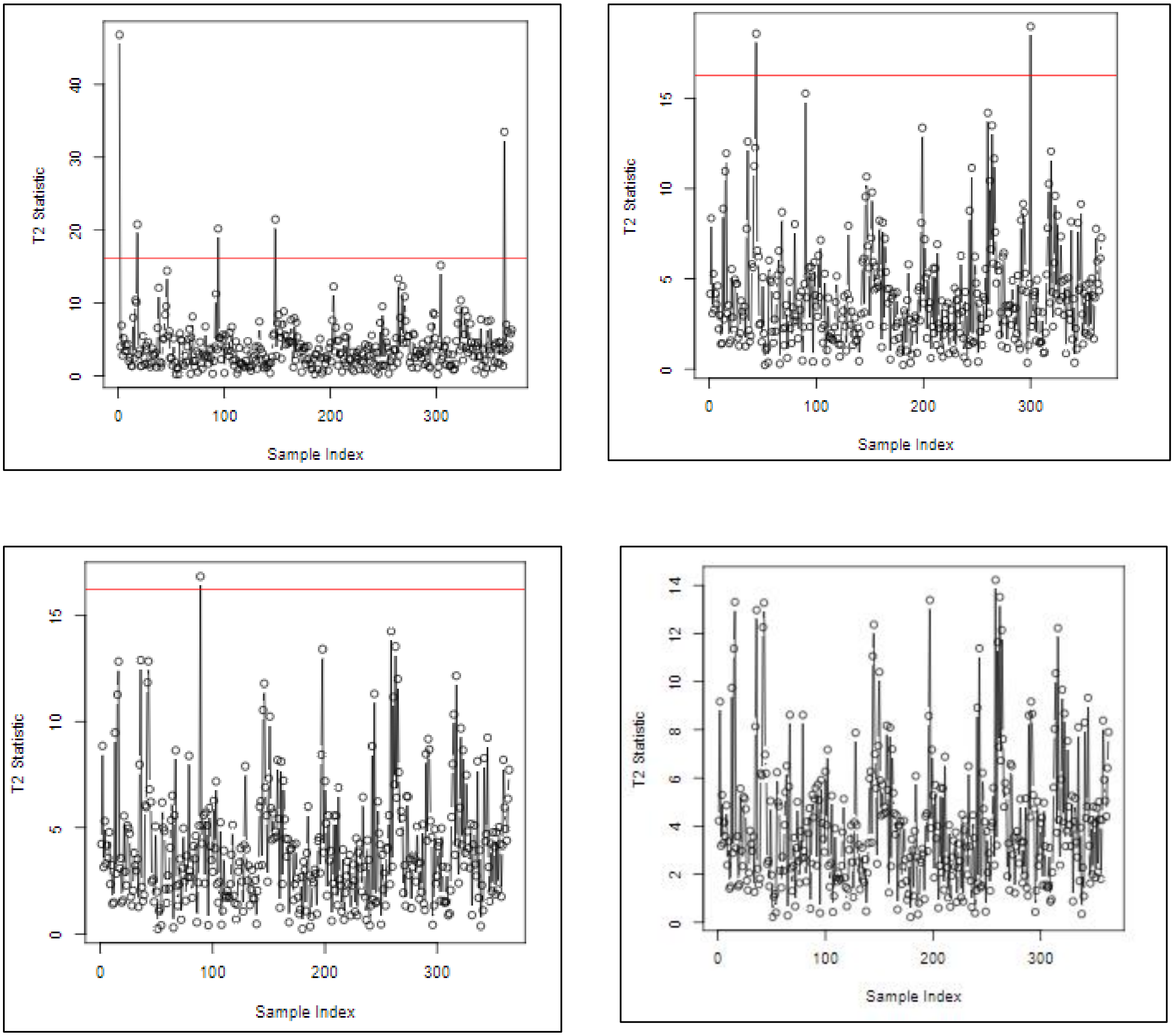


Fig 8: The T^2 charts for the four iterations of the reduced data.

The final data (363 x 4) is now used for estimating the in-control distribution parameters to design a monitoring scheme for future observations. Using this data, the mean vector and covariance matrix of the in-control data are calculated.

Conclusion

The given manufacturing data with 552 rows and 209 attributes was first analyzed for out-of-control points without data reduction. Data reduction was considered to mitigate the effects of the aggregated noise which overwhelmed the signal to indicate out-of-control data as in-control data.

Principal Component Analysis was used to identify the four principal components that can be used to explain 80% of the given data. Since the principal components are uncorrelated, multiple univariate xbar charts were plotted to initially identify the out-of-control data. We noticed trends and shifts in the univariate charts and m-CUSUM charts were plotted to identify and remove the out-of-control samples. The m-CUSUM charts resulted in reduced data but we realized that the data is not entirely in-control due to the presence of spikes in the data. T^2 charts were plotted to realize these out-of-control samples and they were subsequently removed. The T^2 charts still indicate small spikes but these are well inside the control limits.

This could have been avoided by removing out-of-control data initially by using T^2 charts but this would have led to removal of in-control data as well. M-CUSUM charts tend to remove the adjacent in-control points as a result of cumulative accumulation that identifies the next point also as out-of-control.

The final, reduced data now consists of 363 rows and 4 attributes and is in-control. This in-control data is now used to estimate the mean vector and co-variance matrix that can be used to monitor the future observations.

References

1. https://www.graphicproducts.com/articles/quality-control-in-manufacturing/
2. http://www.innovationservices.philips.com/news/managing-production-quality-control-and-10-steps-roadmap
3. https://cran.r-project.org/web/packages/qcc/index.html