# Phase I Analysis of High Dimensional Data Using Multivariate Control Statistics

ISEN 614: Advanced Quality Control Project
Team no. 2

Teja Gambhir (UIN: 525001619)

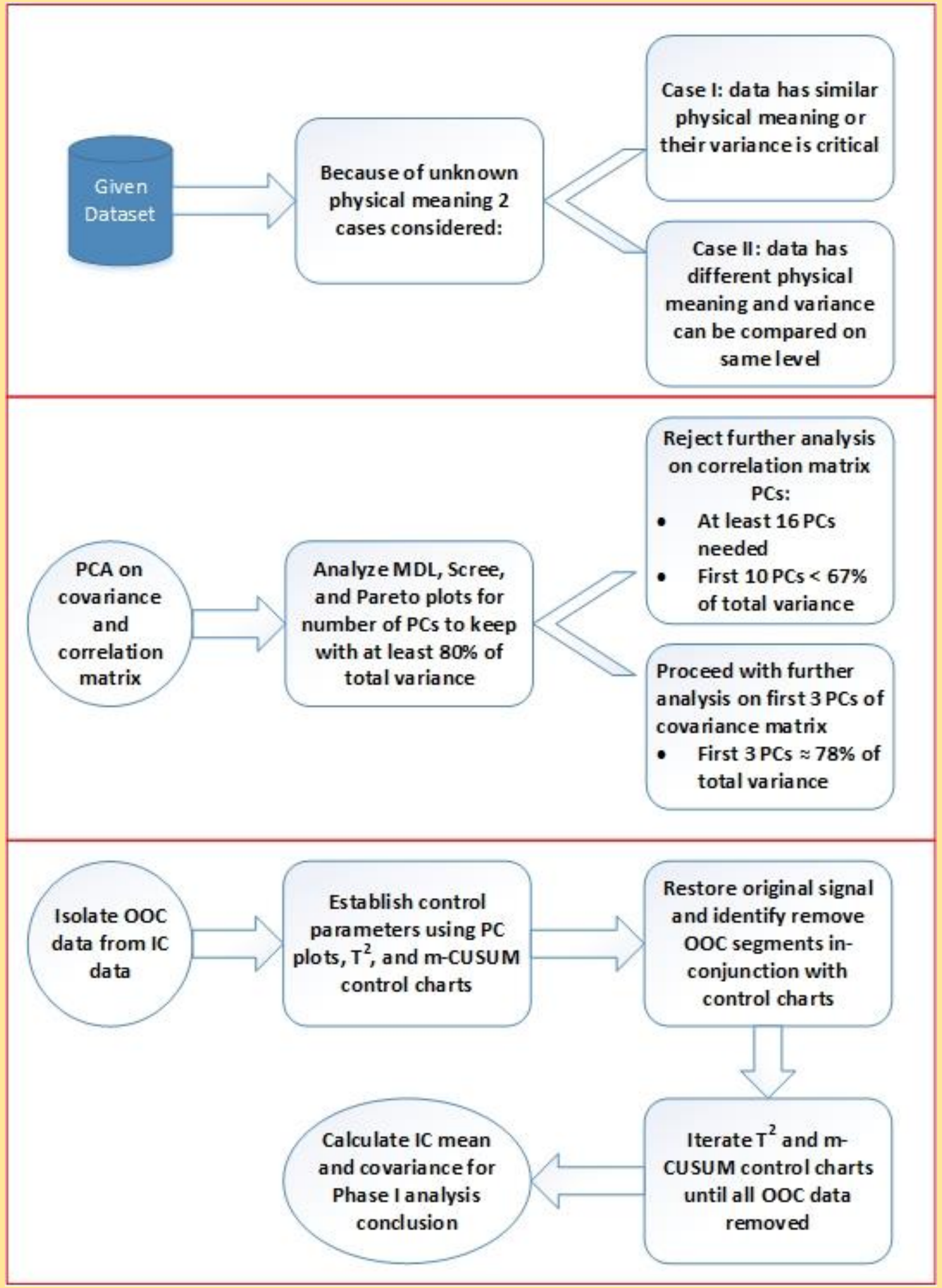
Pulkit Jain (UIN: 625005181)

# Purpose

- The purpose of this project is to conduct a Phase I analysis on a dataset with 209 characteristics and 552 observations with sample size of 1
- The end goal is to establish control charts with in-control mean and covariance so that future observations can be monitored in Phase II analysis
- To do so, multivariate statistical methods were used to isolate all out-of-control (OOC) data from the in-control (IC) data

# Approach

- Because the physical meaning of the dataset was unknown, PCA was performed on the covariance and the correlation matrix
- The number of PCs to retain had to account for close to 80% of the total variation for substantial representation of data
- Since the total number of observations equal 552, the $ARL_0$ had to be substantially less in order to observe sufficient OOC points
- Multiple iterations of the $T^2$ and m-CUSUM control charts were needed in-conjunction in order to ensure both control charts had all in-control data



Given Dataset

Because of unknown physical meaning 2 cases considered:

Case I: data has similar physical meaning or their variance is critical

Case II: data has different physical meaning and variance can be compared on same level

PCA on covariance and correlation matrix

Analyze MDL, Scree, and Pareto plots for number of PCs to keep with at least 80% of total variance

Reject further analysis on correlation matrix PCs:
- At least 16 PCs needed
- First 10 PCs < 67% of total variance

Proceed with further analysis on first 3 PCs of covariance matrix
- First 3 PCs ≈ 78% of total variance

Isolate OOC data from IC data

Establish control parameters using PC plots, $T^2$, and m-CUSUM control charts

Restore original signal and identify remove OOC segments in-conjunction with control charts

Iterate $T^2$ and m-CUSUM control charts until all OOC data removed

Calculate IC mean and covariance for Phase I analysis conclusion

**Figure 1: Pictorial Summary of Approach**

# Principle Component Analysis of Correlation Matrix

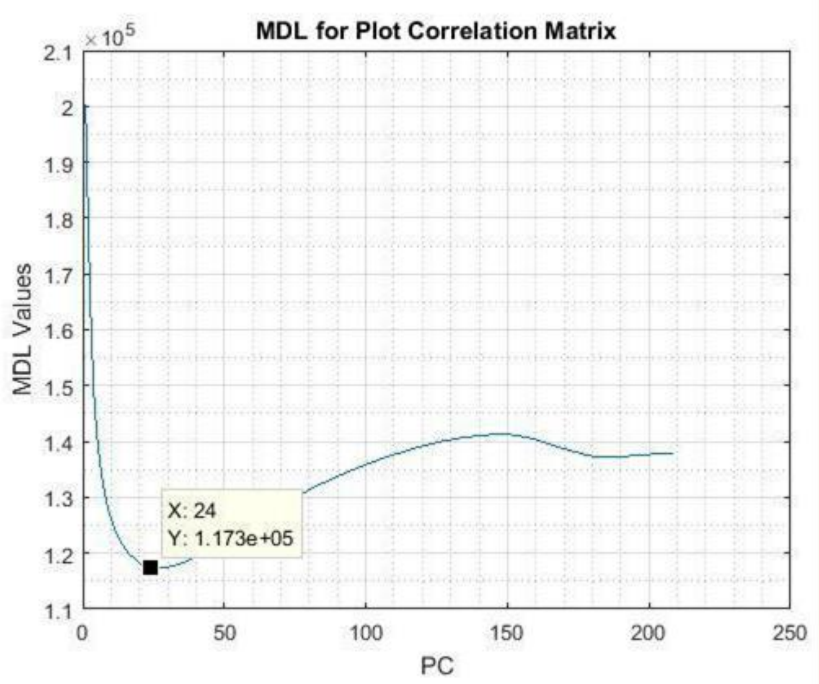


**Figure 2: MDL Plot for PCA on Correlation Matrix**

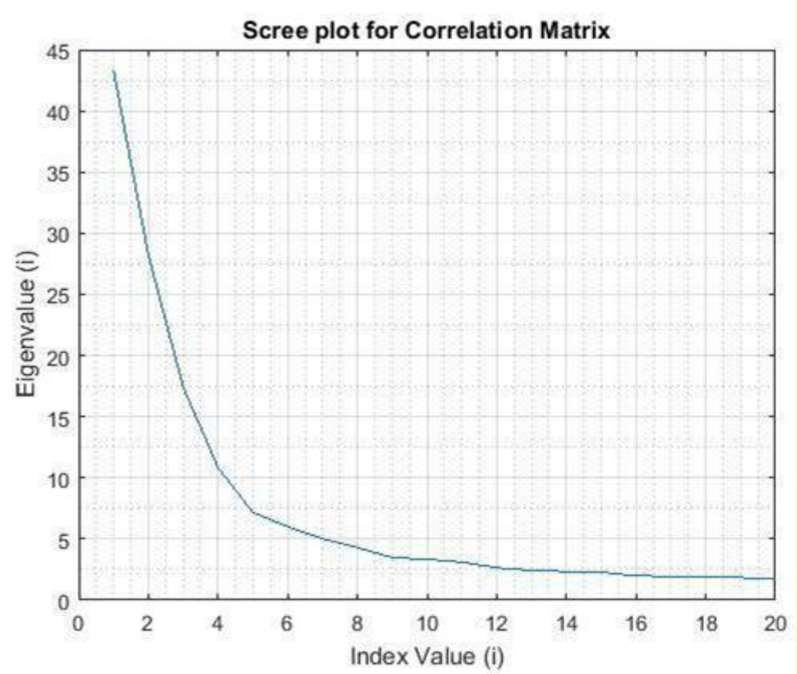


**Figure 3: Scree Plot for PCA on Correlation Matrix**

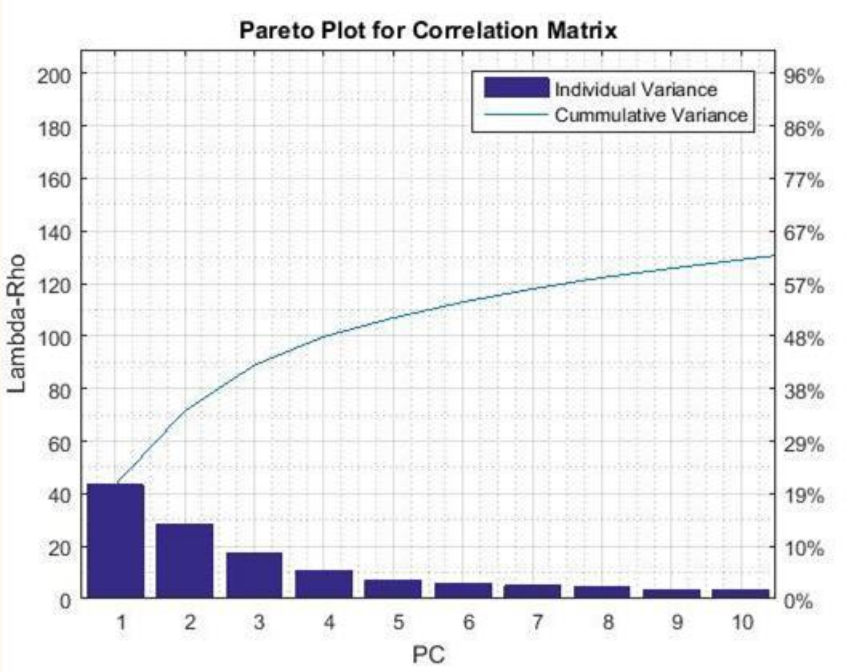


**Figure 4: Pareto Plot for PCA on Correlation Matrix**

- MDL plot retains 24 PCs, which is too many
- Scree plot begins to level only after 16 PCs, but retaining 16 PCs is still too many
- Pareto plot shows 10 PCs contribute less than 67% of the total variance
- Because a total variance of around 80% is not reached with a small number of PCs, further analysis with the PCs for the correlation matrix was forgone

# Principle Component Analysis of Covariance Matrix

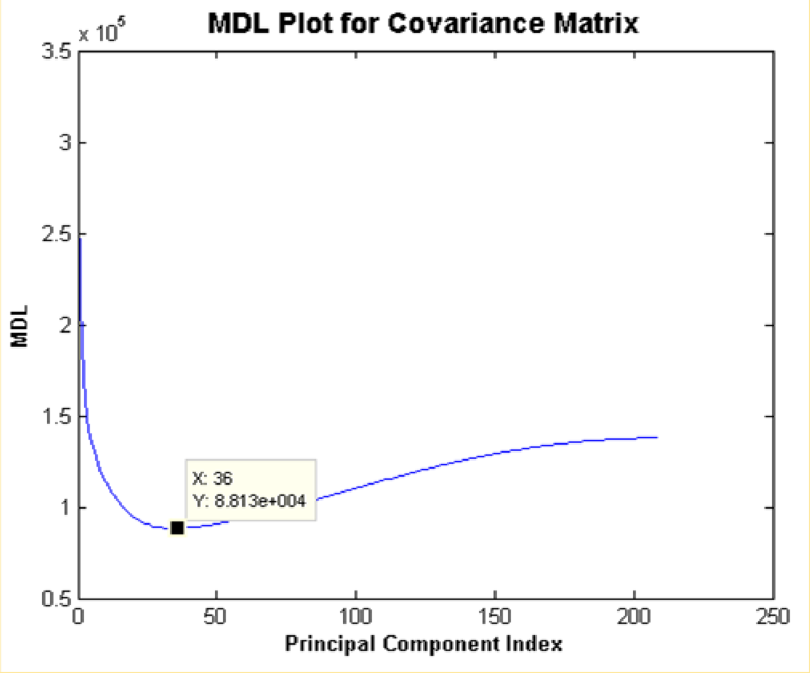


**Figure 5: MDL Plot for PCA on Covariance Matrix**

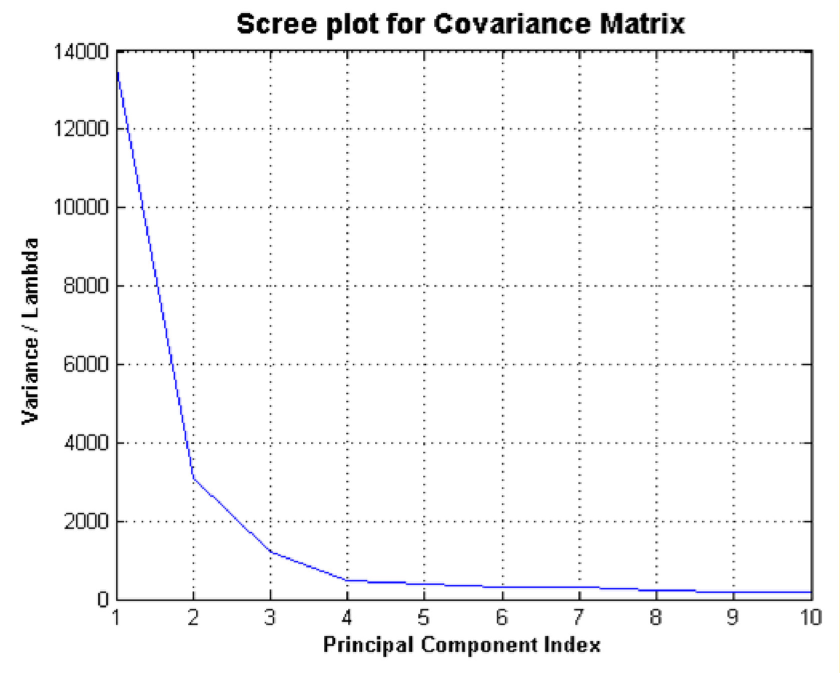


**Figure 6: Scree Plot for PCA on Covariance Matrix**

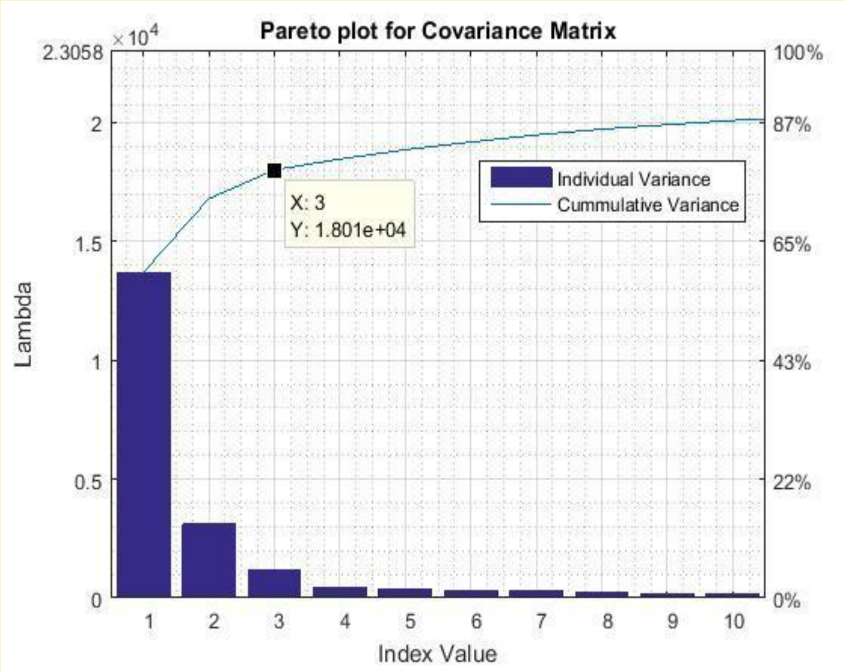


**Figure 7: Pareto Plot for PCA on Covariance Matrix**

- MDL plot retains 36 PCs, which is too many
- Scree and Pareto plots show 3 PCs have a total variance of 78%, which approximately meets our criteria of a total variance of 80%
- The first 3 PCs of the covariance matrix were chosen for further analysis

# Establishing Control Chart Parameters: m-CUSUM

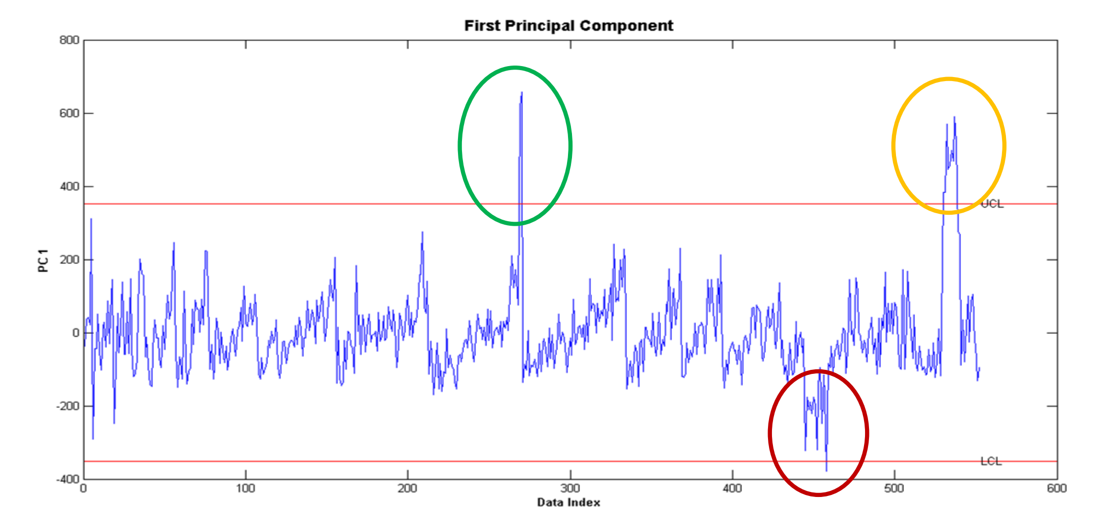


**Figure 8: PC1 for PCA on Covariance Matrix**

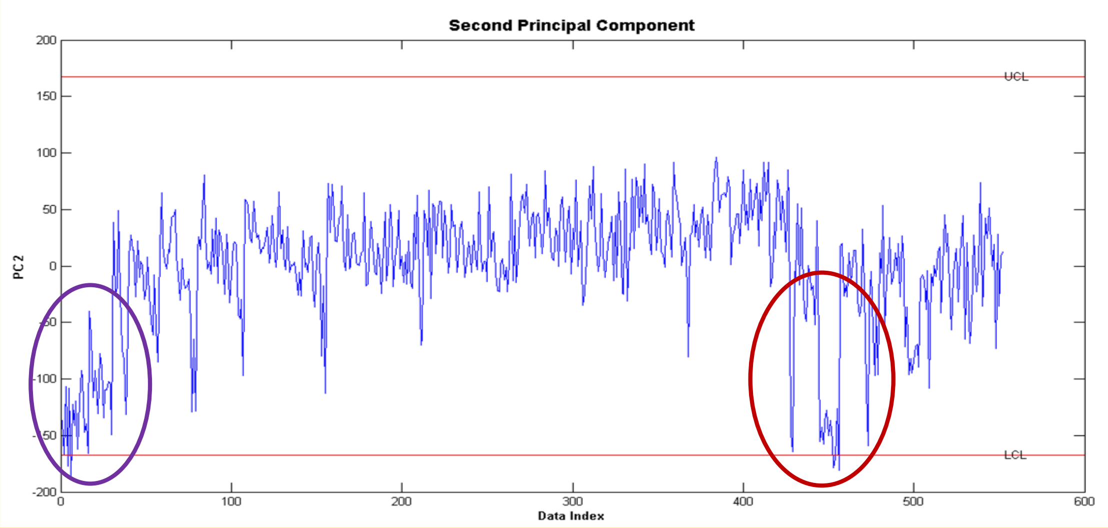


**Figure 9: PC2 for PCA on Covariance Matrix**

- OOC data was observed on PC1 and PC2, but not on PC3
- The first 38 points observed were OOC on PC2 and on the m-CUSUM for the 3-sigma mean shift
- All three charts signaled for data points from 500 to 552
- Because the m-CUSUM closely identifies the same OOC points as the PCs, a **mean shift of 3-sigma** was chosen instead of a mean shift of 2-sigma
- Chosen parameters: **$\mathbf{ARL_0}$ = 200** and **UCL = 5.48**

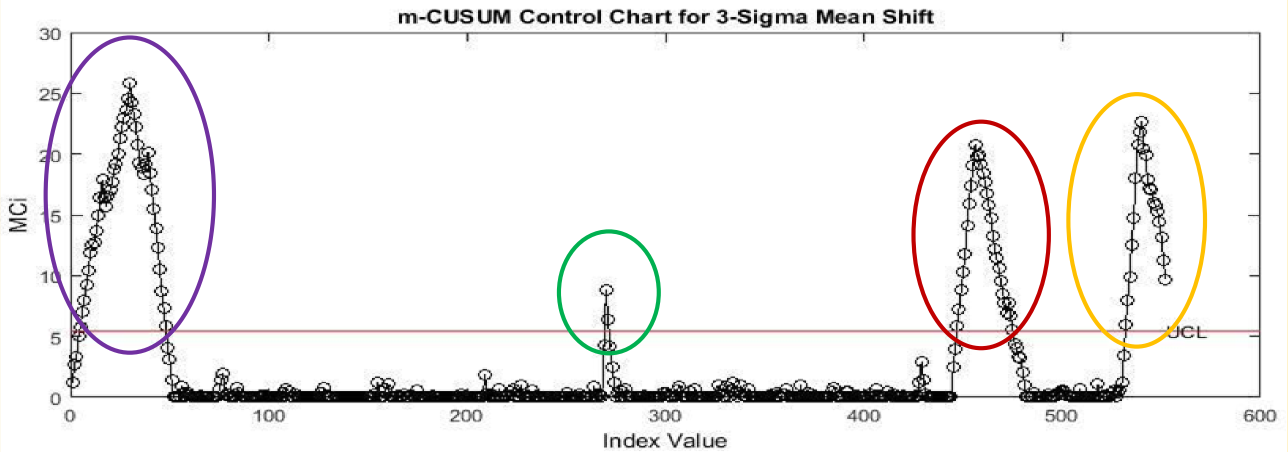


**Figure 10: m-CUSUM for 3-Sigma Mean Shift**

# Establishing Control Chart Parameters: $T^2$

- The $T^2$ chart was able to detect a similar trend because of its sensitivity to large, sustained mean shifts
- An **alpha of 0.005** was used corresponding with an $ARL_0$ of 200

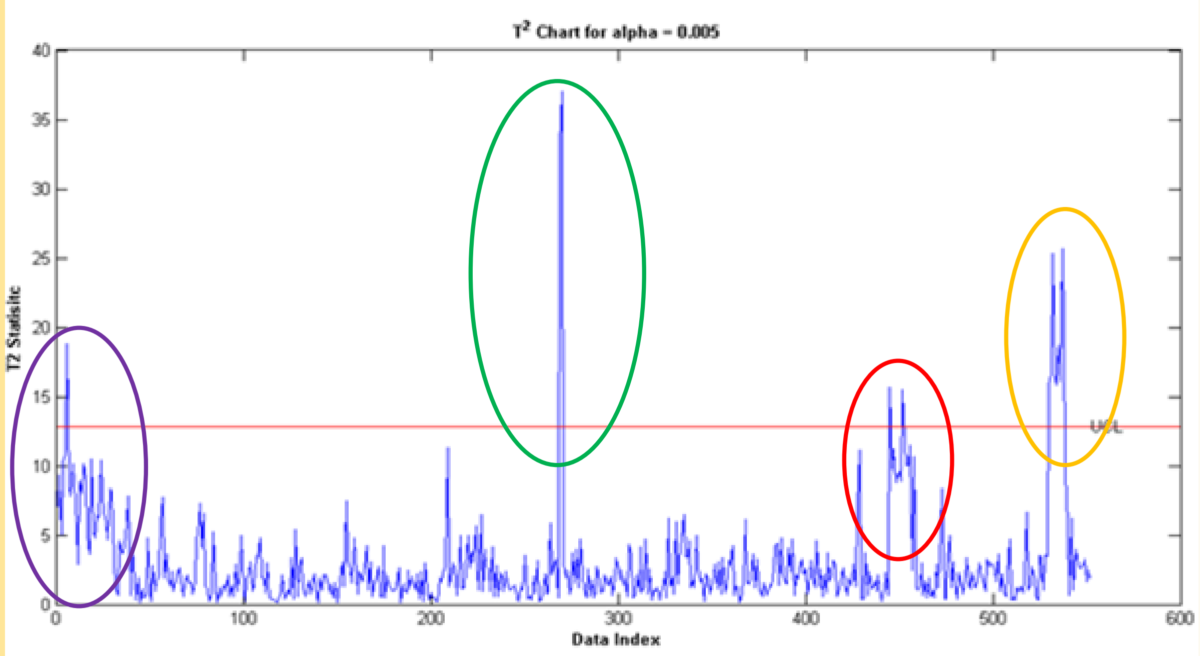


**Figure 11: T2 Control Chart for 3 PCs**

# Restoration of Original Signal Profiles

- Plot shows original profile signal for 3 OOC segments
- The 4th segment shows the profile for the IC data points
- Segment 3 and segment 2 deviate the most from the IC segment

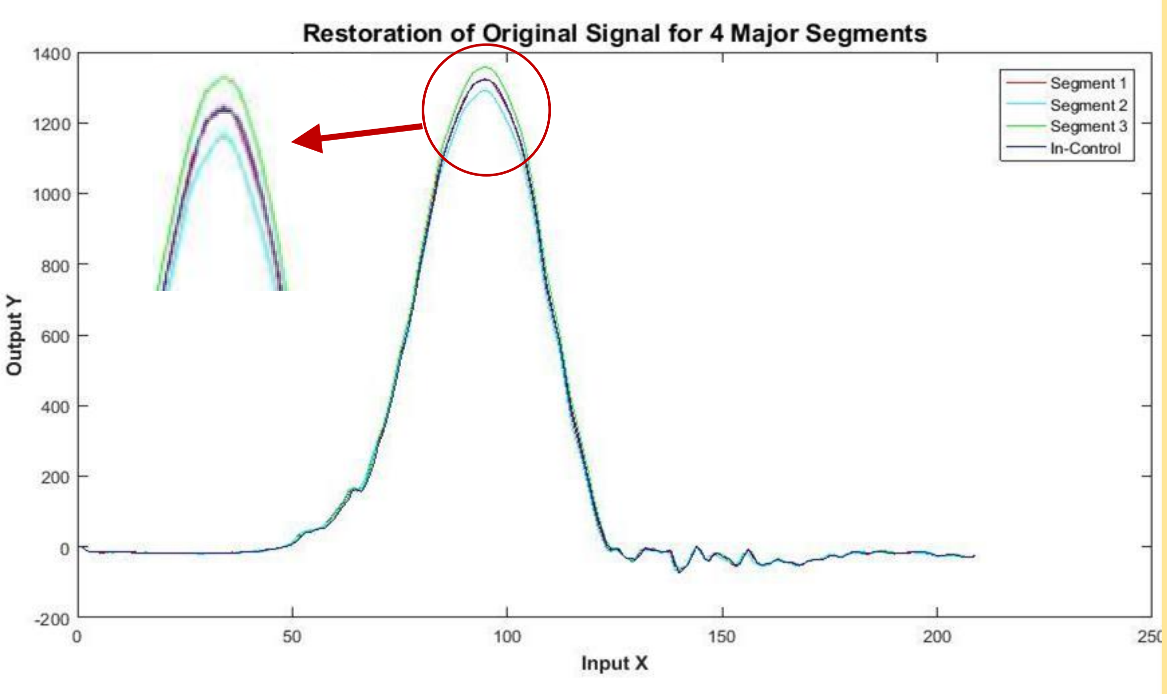


**Figure 12: Restoration of Original Signal Profiles for 4 Major Segments**

# Isolate OOC Data from IC Data: T2 and m-CUSUM Iterations

- The 1st round of $T^2$ was followed by the 1st round of m-CUSUM to establish IC points on both charts
- The 2nd round of $T^2$ showed OOC points, so multiple rounds were needed for both charts to remove all OOC data points

| $T^2$ and m-CUSUM Control Chart Out-of-Control Observations | | | | |
|---|---|---|---|---|
| **Round** | **Control Chart** | **Iteration** | **OOC Points** | **Remaining Observations** |
| One | $T^2$ | 1 | 9 | 458 |
| | | 2 | 7 | 451 |
| | | 3 | 1 | 450 |
| | mCUSUM | 1 | 5 | 445 |
| | | 2 | 1 | 444 |
| Two | $T^2$ | 1 | 3 | 441 |
| | | 2 | 1 | 440 |
| | -CUSUM | 1 | 1 | 439 |
| | | 2 | 1 | 438 |
| Three | $T^2$ | 1 | **0** | **438** |

**Table 1: OOC Points for Control Chart Iterations**

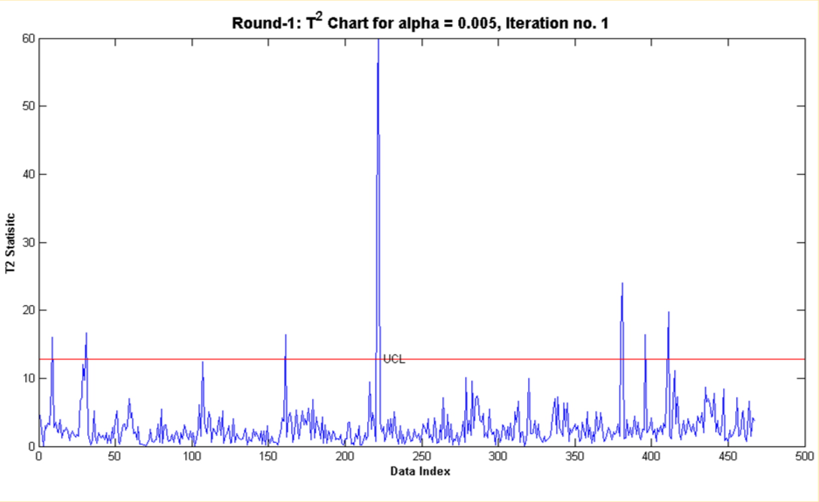


**Figure 13: $T^2$ Round 1 Iteration 1**

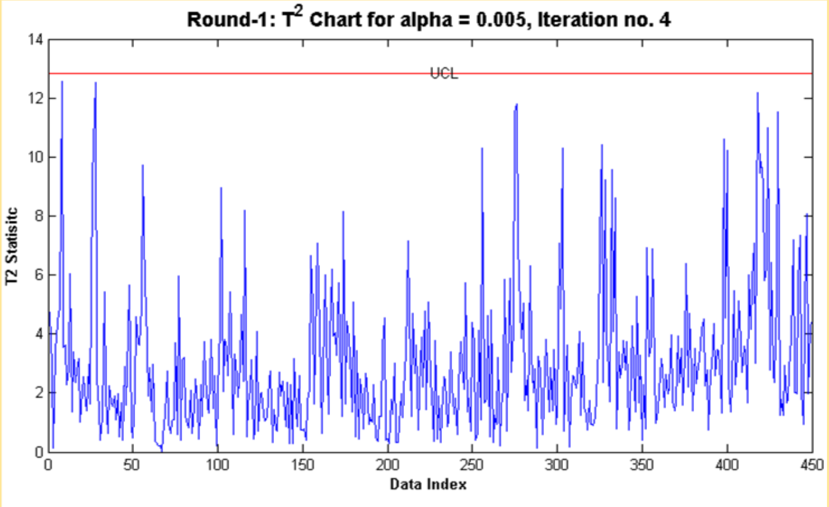


**Figure 14: $T^2$ Round 1 Iteration 4**

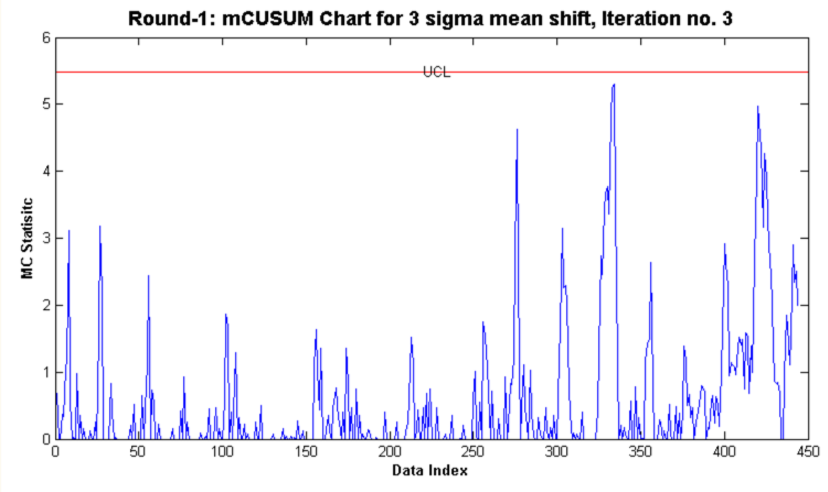


**Figure 15: m-CUSUM Round 1 Iteration 3**

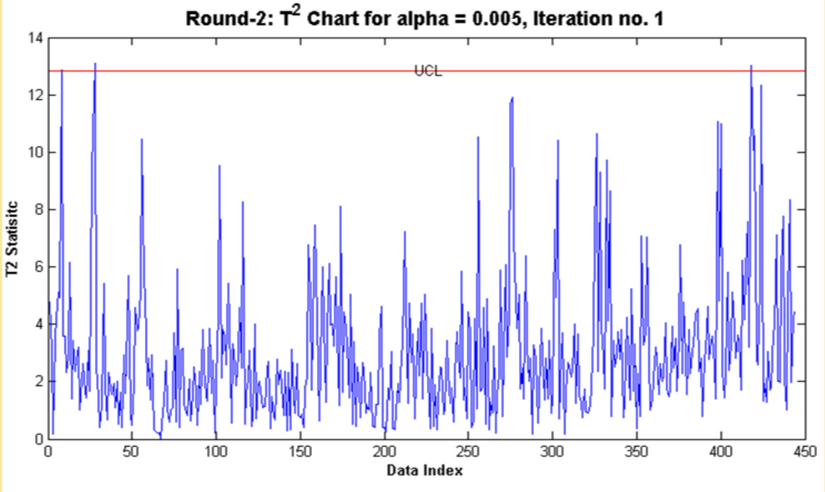


**Figure 16: m-CUSUM Round 1 Iteration 4**

# Results

- After 3 rounds of iterations, all OOC data points were removed

- A total of 438 observation out of the original 552 were found to be in control

- In-control mean: $\begin{bmatrix} 11.27567 \\ 16.49829 \\ 5.62782 \end{bmatrix}$

- In-control covariance: $\begin{bmatrix} 5724.7 & 519.95 & -456.43 \\ 519.95 & 1180.8 & -354.18 \\ -456.43 & -354.18 & 1084.2 \end{bmatrix}$

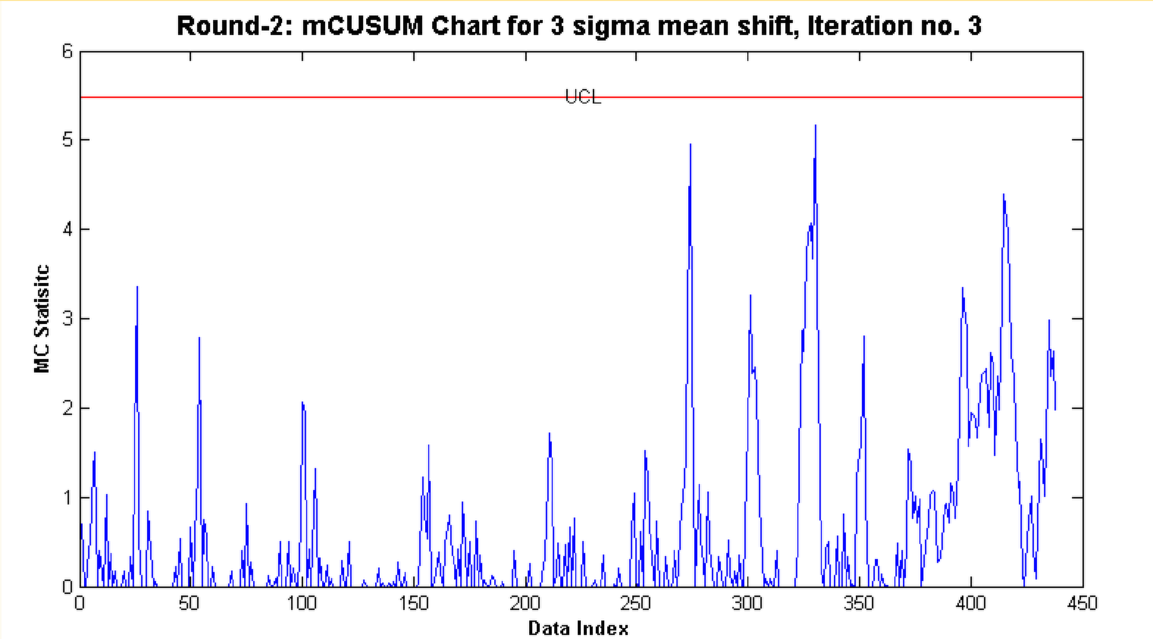


**Figure 17: m-CUSUM Round 2 Iteration 3**

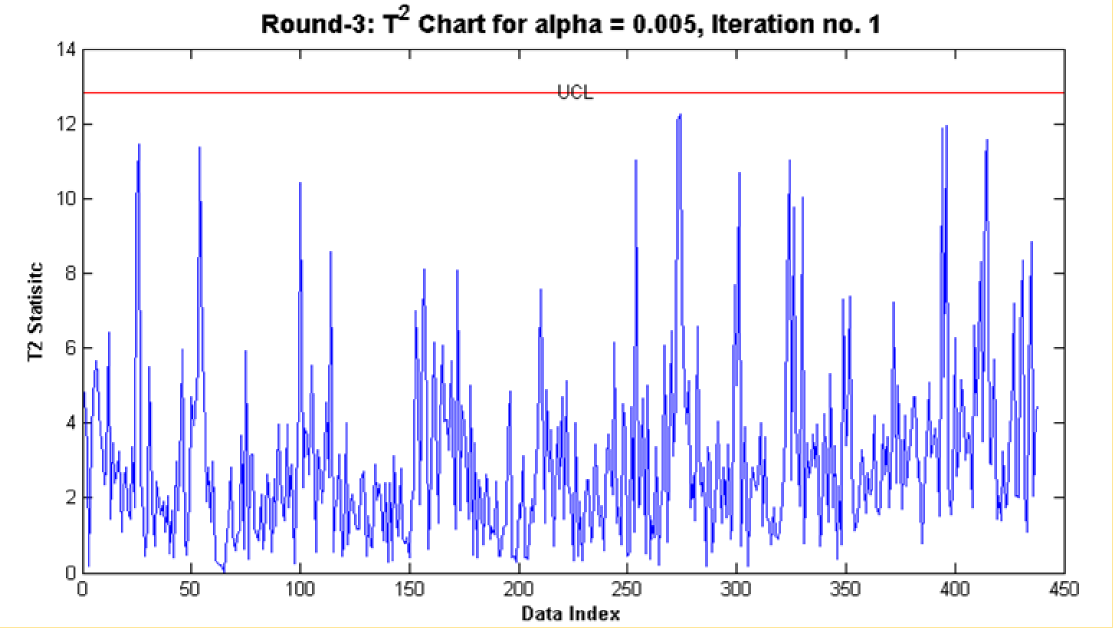


**Figure 18: T$^2$ Round 3 Iteration 1**

# Conclusion

- Isolating the OOC data from the IC data by doing multiple iterations of the $T^2$ and m-CUSUM control charts helped identify 6 more data points during the 2nd round of iterations that were OOC
- Despite performing the PCA, the non-diagonal elements of the covariance matrix are non-zero, which implies the values are correlated and thus dependent
- This may be attributed to noise, which made it difficult to achieve zero covariance
- With the conclusion of Phase I analysis, the control charts established with the IC mean and covariance can be used to monitor future observations
- This project forced us to rely on our own judgment and to rationalize all methods used, similar to a possible industry scenario
- Having done the project we now have a better understanding of the relative advantages in utilizing one method over another