INDUSTRIAL & SYSTEMS ENGINEERING
TEXAS A&M UNIVERSITY

FINAL COURSE PROJECT
ISEN-614
SPRING 2022

Team 11
Rohit Rajesh Naik(732004183)
Madhusudhan Goplaraju(432007217)
Nitu Mahadev Munavalli(932008273)

Table of content

EXECUTIVE SUMMARY

Phase 1 analysis is the method of removing out-of-control Data from the raw data sets. And to give the parameters for phase 2 analysis. The main issue arises when we get high dimensionality data sets (in this case 408 dimensions) which results in the curse of dimensionality. And not all the information available in the high dimensional data is important. It is therefore preferred to reduce the dimensions to get VITAL HIGH data sets.

So, we did the principal component analysis using a correlation matrix. First, we found the eigenvalues and eigenvectors of the correlation matrix and then using those we found PCA. As we got 408 PCs, we applied ideas of scree plot, and Pareto plot to choose the best top 3 PCs. Using those PCs we did a multivariate CUSUM chart repeatedly to remove the data causing sustained mean shifts and then we used a multivariate Hotelling T^2 chart to identify and remove the data causing spike type changes, we did a few iterations to completely remove these types of data.

We then did a sanity check to see whether all data are in control, here we once again applied the multivariate CUSUM chart to check any mean shift. After that, we applied the hotelling t^2 chart and there was no out-of-control data. Finally, we got the 255 in-control data sets, using those we calculated the in-control sample parameters (i.e. sample covariance matrix and mean vector) data set for future monitoring i.e., phase 2 analysis. The key insights we got while doing this project were that although PCA reduces the dimensionality of the data and we used an 82% variance proportion, there might be chances that the other 20% of variances contain some important information. Both MCUSUM and hotelling t^2 charts are unable to detect the small spikes. After analyzing the raw data it was found that many variables are highly correlated up to 0.9.

PROBLEM DESCRIPTION

We have given data sets collected from the biomedical instruments in a healthcare process, which includes both in-control and out-of-control data. We must devise a system or procedure for identifying the data that fit into the categories, i.e., which are in control and which are out of control. This is a Phase I analysis, which intends to isolate the in-control data to estimate the in-control distribution parameters and set up a monitoring plan for phase 2 analysis.

Description of the dataset

(a) This dataset contains 462 data records and is connected to biometric signals.

(b) n=1, m=462, p=408.

The goal is to extract in-control data from the provided data collection. The values of μ_0 and Σ_0 for this data are unknown. As a result, this is a Phase I study with n=1. To estimate μ_0 and Σ_0, we'll utilize $\overline{x}$ and S from historic data.

APPROACH

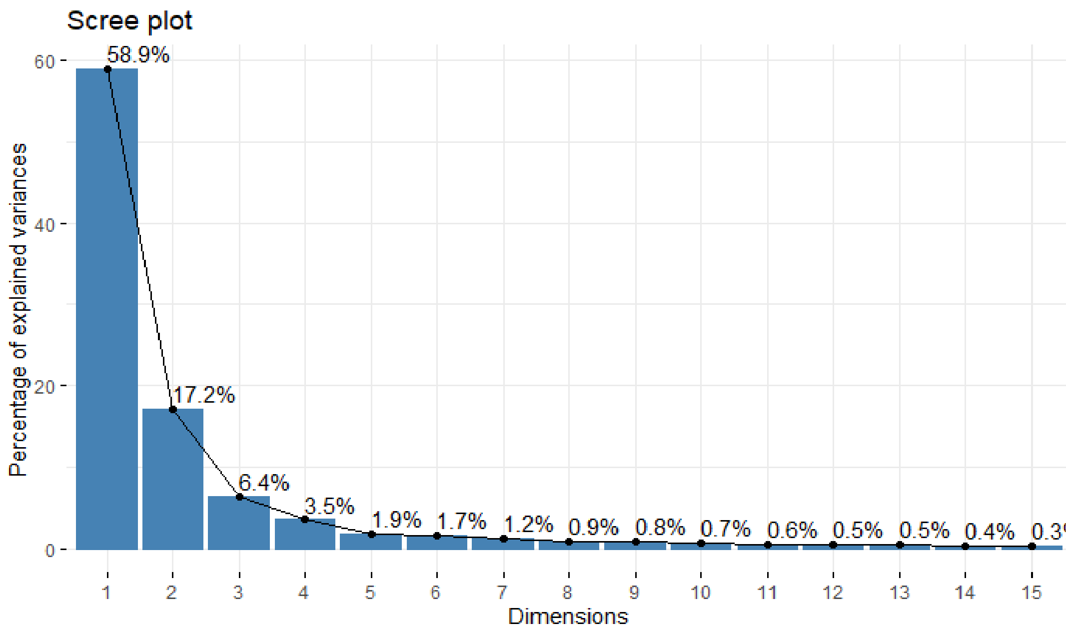
Due to the large dimension of the data set, we are using PCA to reduce the dimension. We used a multivariate CUSUM chart to detect small mean shits and then used the Hotelling T^2 chart to isolate in-control data from the large spike. Then a sanity check was made with the help of the M-CUSUM chart T^2 chart just to detect if any data is out of control and in the end just to ensure all final data are in control another multivariate CUSUM chart is used.

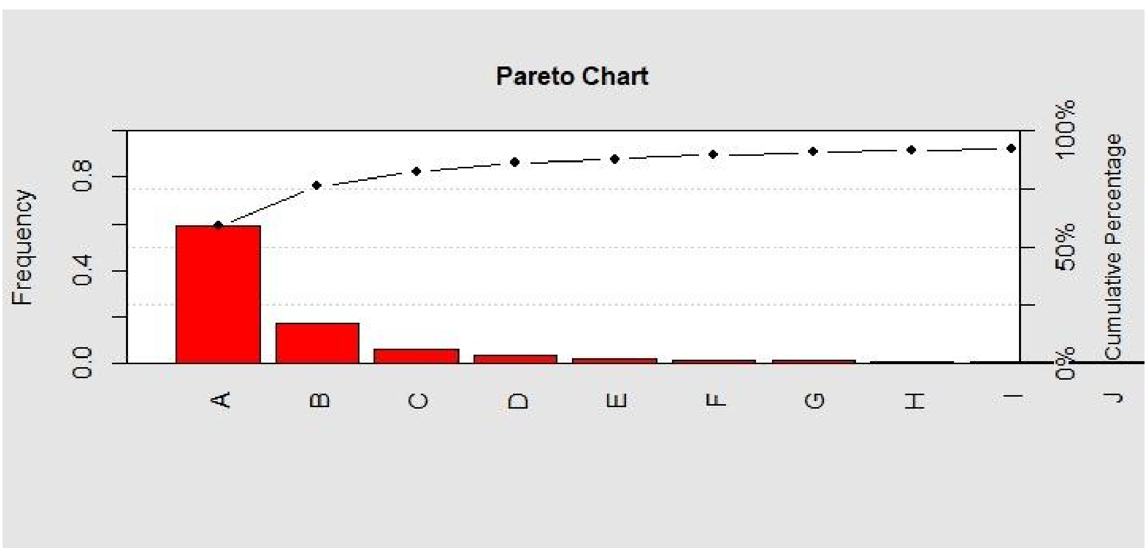
Dimensionality Reduction:

PCA

PCA is one of the tools to reduce the dimensions of data. We used correlation-based PCA. Doing the PCA will convert correlated data sets into linearly uncorrelated data using orthogonal transformation. The reasons for using correlation over covariance are that raw data has a widely varying range. And the relative magnitude of deviation is not important and as the physical meaning of the data is not given it is safer to use a correlation matrix and it captures most of the information.

Doing PCA alone will not reduce the dimensionality we need to choose the VITAL HIGH data sets. So, we used a scree plot and paret chart to choose PCs.





In the scree plot, we are getting elbow between 3rd PC and 4th PC, so it is always better to choose the small one. It is also verified that the first 3 PCs contain above 82% of the total variance proportion by pareto chart. So, we choose the first three PCs for our further analysis.

Control Charting

We have used multivariate charts rather than individual control charts because in individual PC charts α and β errors will increase.

And we have used the alpha value as 0.0027 but while doing the individual PC chart the alpha value will add up to 0.008 which will eventually reduce the ARL0 value and make the charts more sensitive. Which results in more false positives.

Multivariate CUSUM Chart:

It is one of the controlling charts which uses memory to detect the sustained mean swift in the data set. As we used 3 PCs it is relevant to use the h value as 5.5. We were interested in the mean shift of 2. That gives the value of k as 1.

$$Ci = \sum_{j=i-ni+1}^{i} (xj - \mu 0)$$

Namely, Ci = CUMSUM of previous n_i x's = $n_i(\overline{x} - \mu 0)$, where $\overline{x}$ is the average over previous n_i x's.

Then, define

$MC_i = \max\{0, (C^T_i \Sigma^{-1}_0 C_i)^{1/2} - k \cdot n_i\}$

We can use either m-EWMA or m-CUSUM analysis to remove the points that cause a sustained mean shift.

We had to do 8 iterations to completely remove out-of-control data. The table below gives detailed information about the number of samples before every iteration, UCL, and the number of out-of-control points found in each iteration.

M-CUSUM Chart			
Iteration	Number of Sample	UCL	Out of Control Points
1	462	5.5	18
2	444	5.5	48
3	396	5.5	60
4	336	5.5	30
5	306	5.5	29
6	277	5.5	12
7	265	5.5	2
8	263	5.5	0

Table 1: Details of Output from m-CUSUM chart

The following pictures are the 1st, 2nd and final iteration of the M-CUSUM chart.

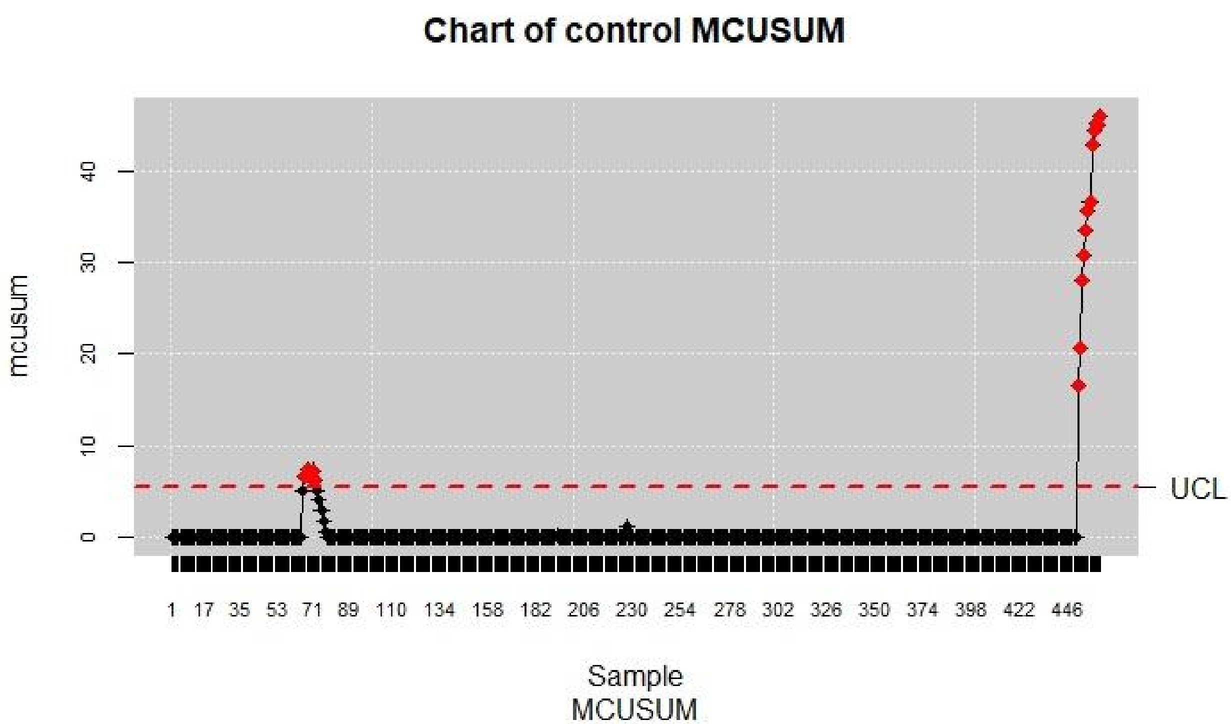


Figure 1: 1st Iteration of the m-CUSUM chart

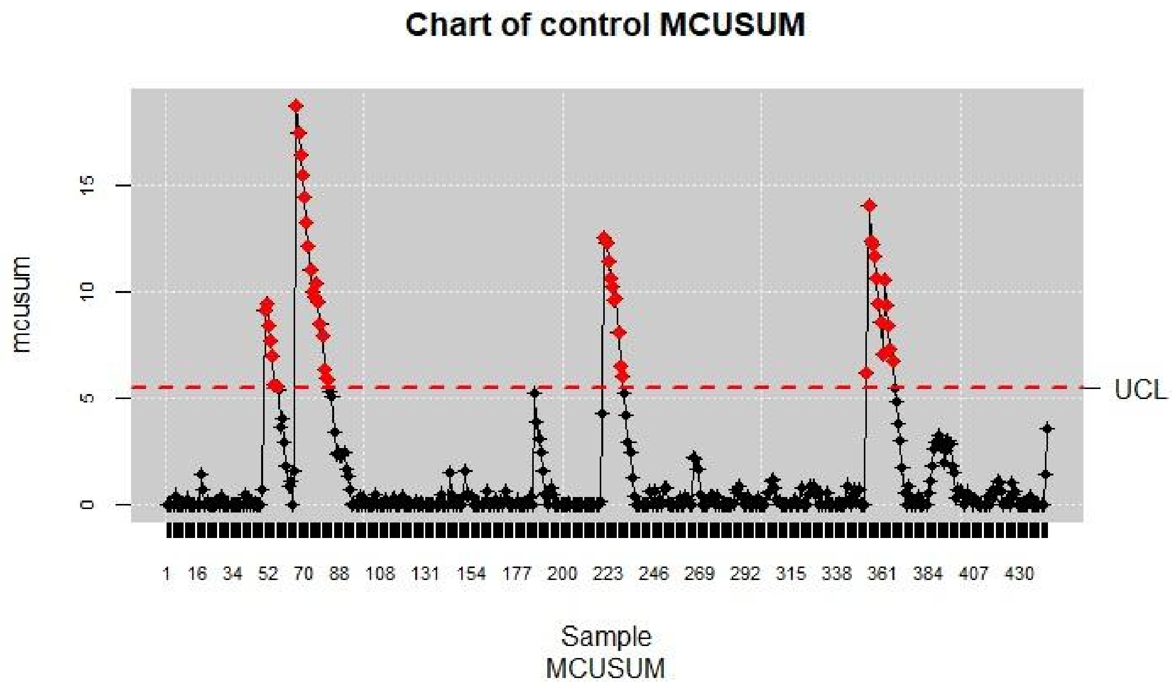


Figure 2: 2nd Iteration of the m-CUSUM chart

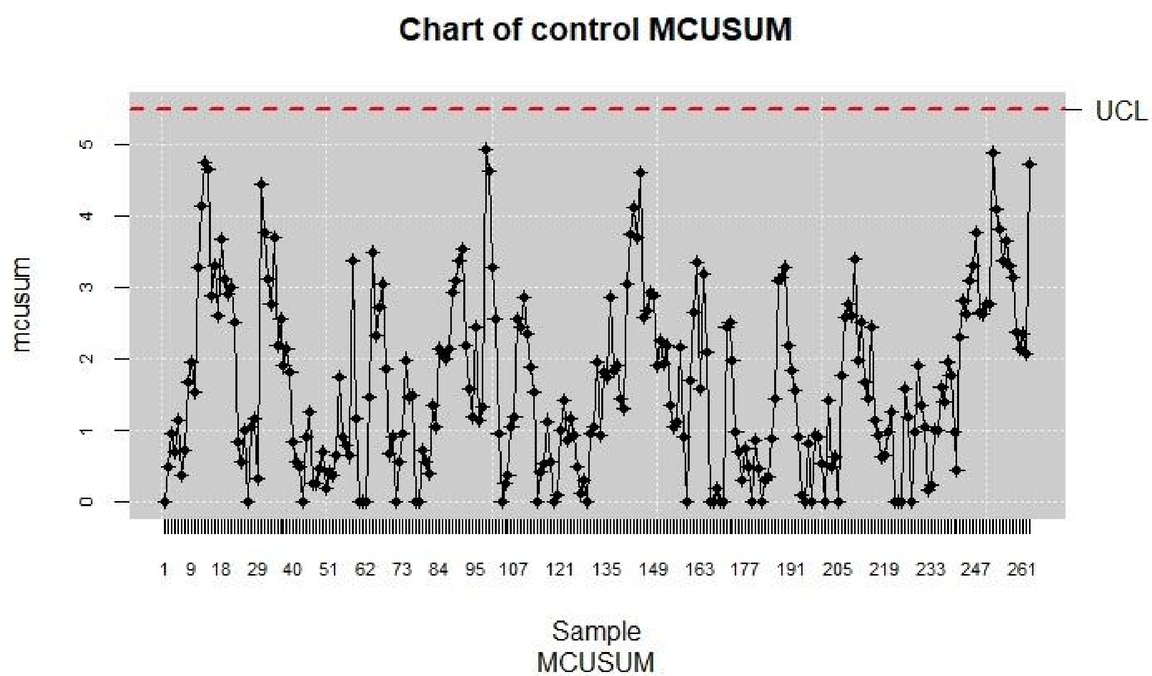


Figure 3: Final Iteration of the m-CUSUM chart

Charts of all the iterations can be found in Appendix.

Hotelling T^2 chart:

Typically, the Hotelling T^2 chart is used to detect the spike type changes, here we used it on the in-control data given by multivariate CUSUM to identify and remove the data sets causing spike type changes. We used M-CUSUM before Hotelling the t^2 chart to make sure that the data set does not contain any memory. This procedure has been iterated 3 times to ensure training data is in control. We have used α=0.0027. This gives the value of UCL as 13.85 using chi-square distribution and with a value of p=3.

We know that when n =1

$$T^2 = (x_j - \bar{x})^T S^{-1} (x_j - \bar{x})$$

$$UCL = \chi^2_{1-\alpha}(p)$$

Here

x_j= Mean of j sample

$\bar{x}$ = mean taken from historic data

S^{-1}= Covariance matrix

p = Dimension of the matrix

$\alpha = 0.0027$

T^2 Chart			
Iteration	**Number of Sample**	**UCL**	**Out of Control Points**
1	263	13.857	4
2	259	13.852	1
3	258	13.8512	0

Table 2: Details of Output from T^2 chart

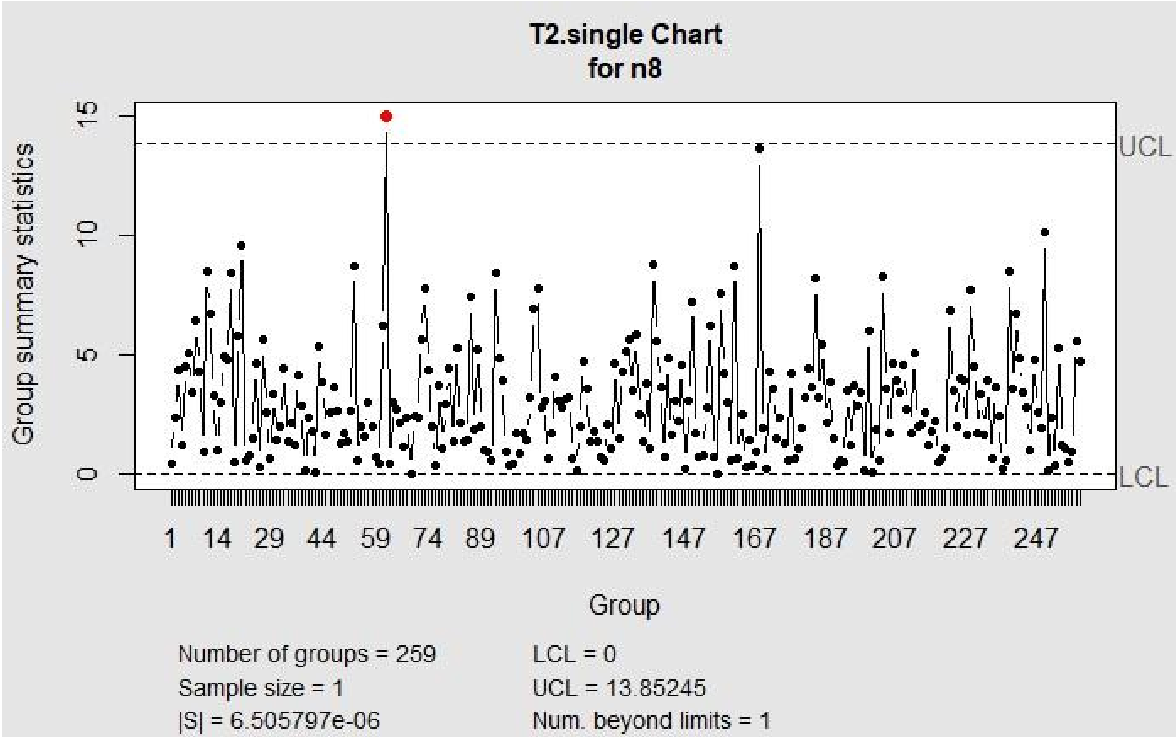


Figure 4: 1st Iteration of the T^2 chart

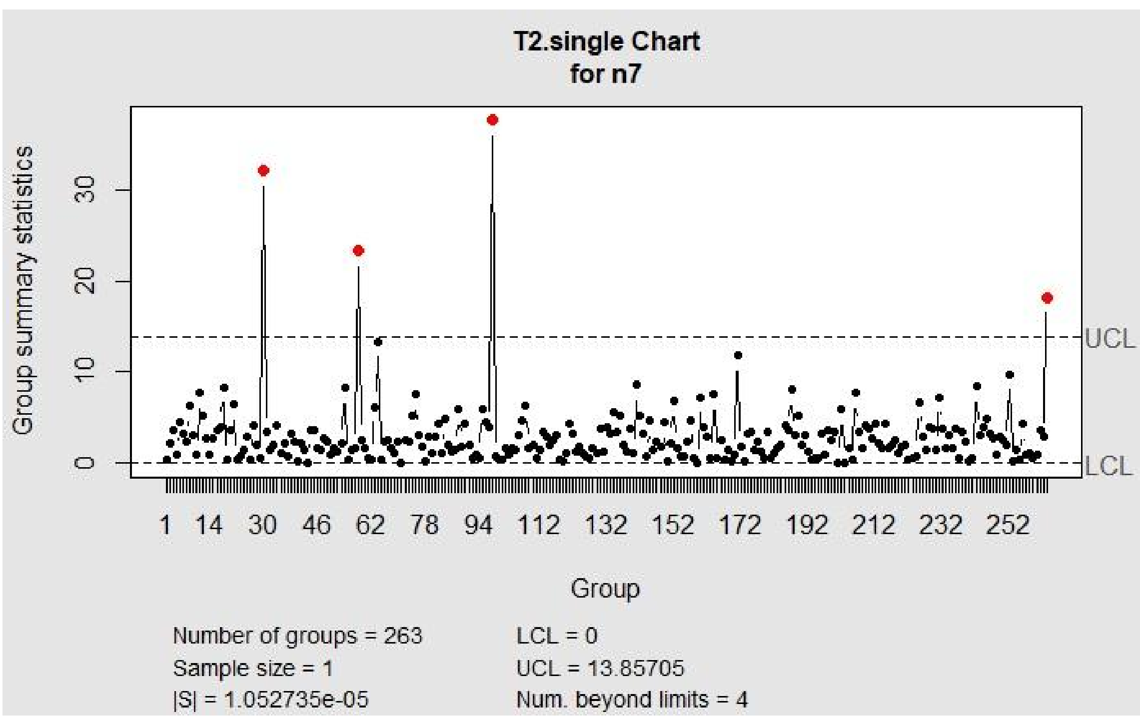


Figure 5: 2nd Iteration of the T^2 chart

Final iteration of t-square chart

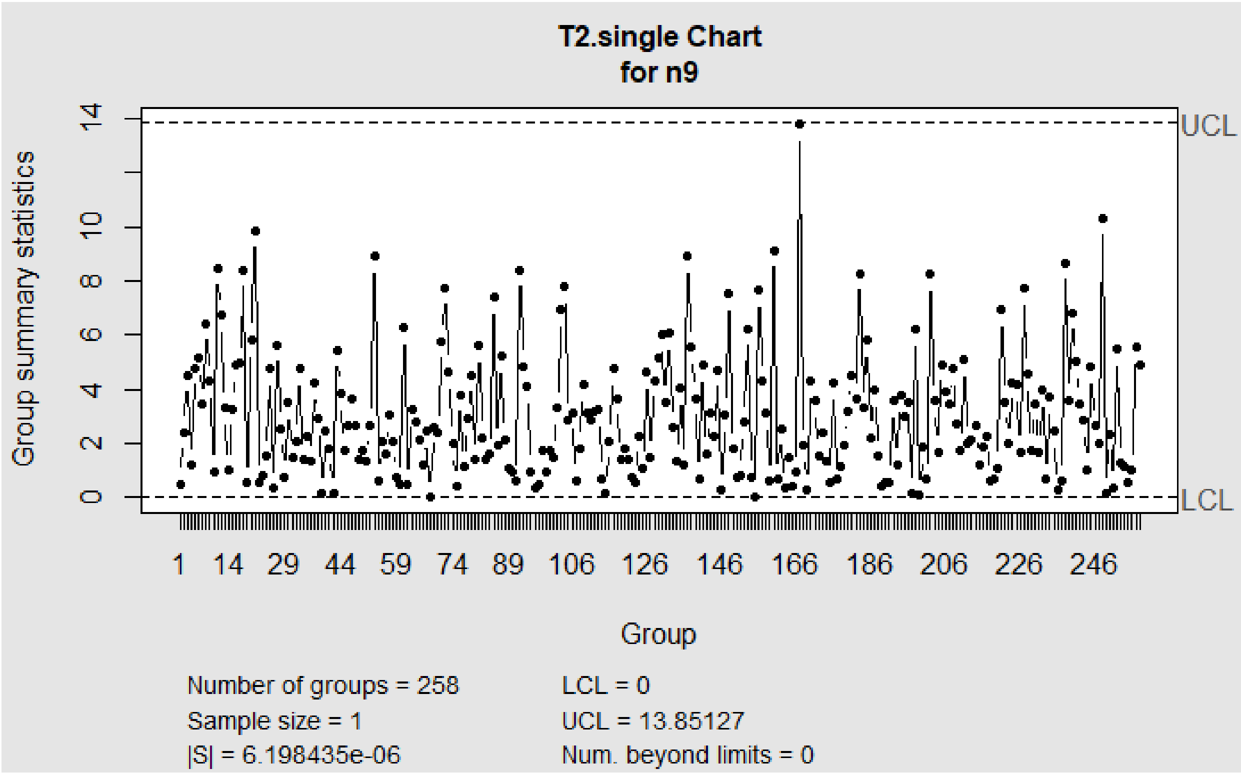


Figure 6: Final Iteration of the T^2 chart

Sanity Check:

It is necessary to check whether all in-control data are still in-control after going through MCUSUM and t^2 charts. So, again we applied MCUSUM and t^2 charts.
3 points come out of control insanity checking so, we removed those and applied t^2, where all data are in control.

m-CUSUM chart:

After removing out-of-control points using the T^2 chart, we use the M-CUSUM chart again to verify that all points are in control, thus it is also called a sanity check. Here we performed two iterations to get all in-control data points.

M-CUSUM Chart			
Iteration	Number of Sample	UCL	Out of Control Points
1	258	5.5	3
2	255	5.5	0

Table 3: Details of Output from m-CUSUM chart

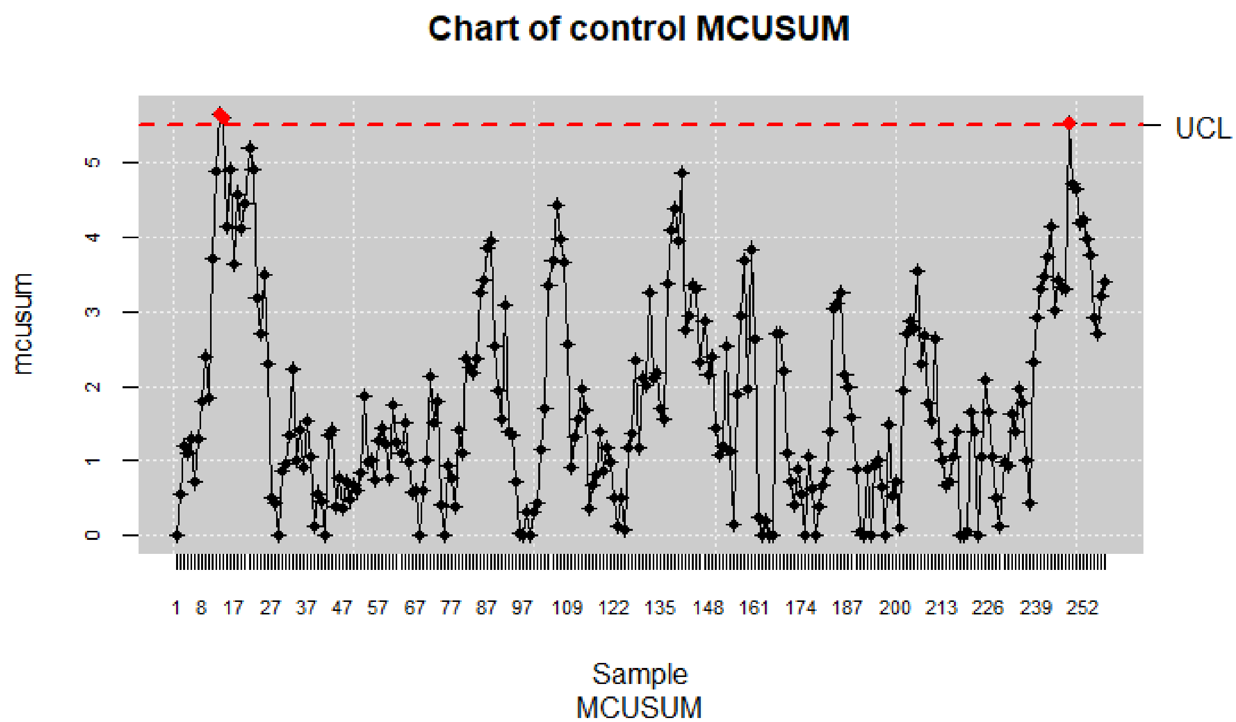


Figure 7: 1st Iteration of the m-CUSUM chart

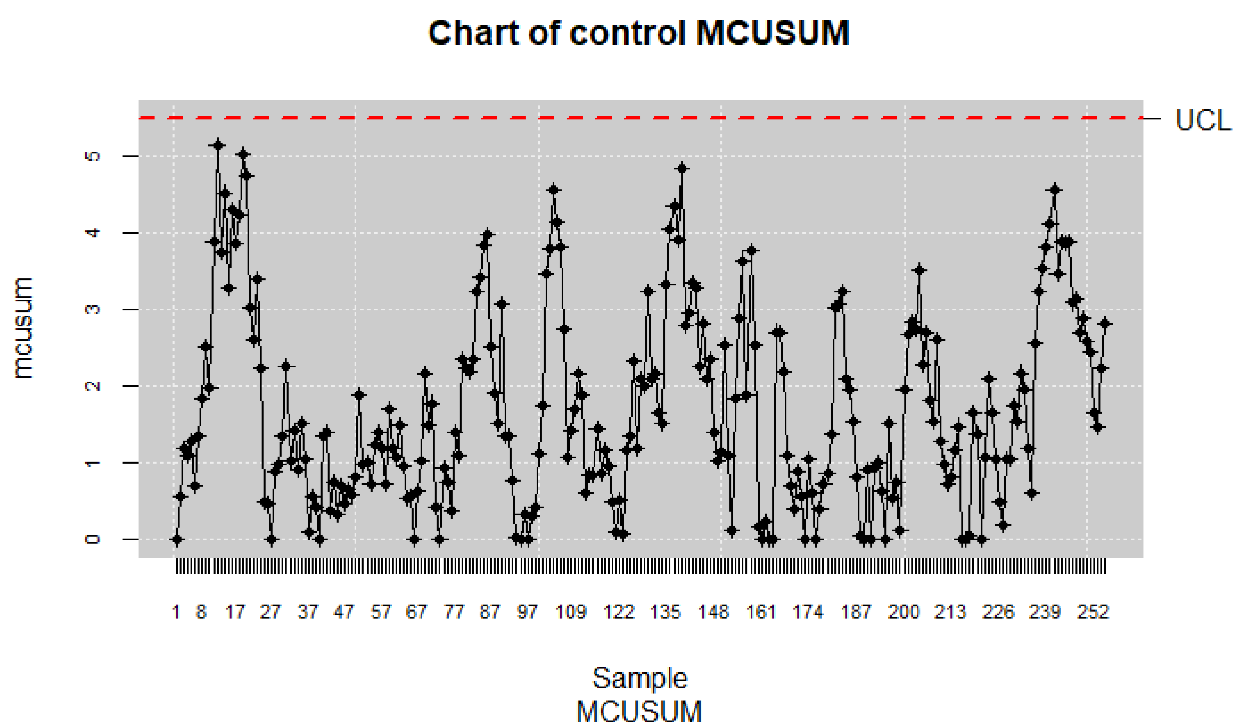


Figure 8: Final Iteration of the m-CUSUM chart

Hotelling T^2 chart after m-CUSUM:

Just to confirm spike changes are completely removed from our data set we performed the T^2 chart again and as expected we didn't find any spikes and all data sets were in control.

T			
Iteration	Number of Sample	UCL	Out of Control Points
1	255	13.847	0

Table 4: Details of Output from T^2 chart

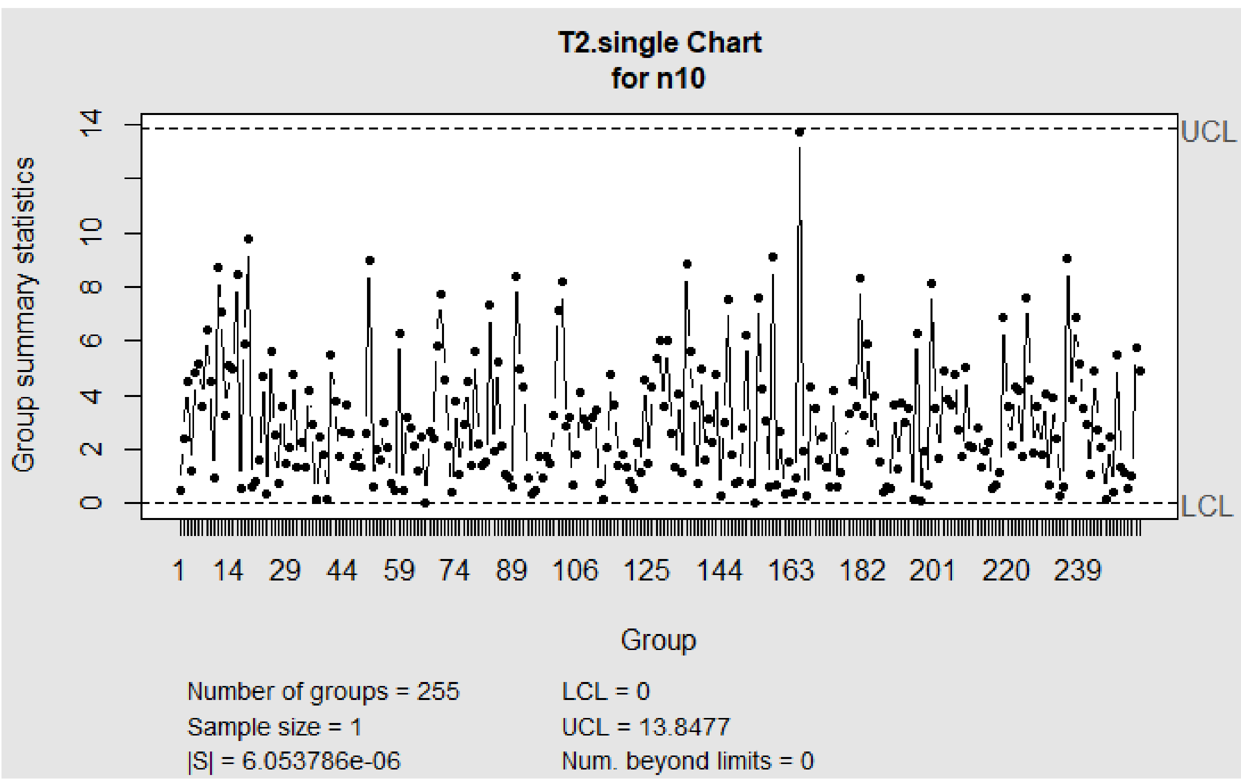


Figure 9: Final Iteration of the T^2 chart

RESULTS

After getting all in-control data sets we calculated the mean vector and sample covariance matrix for further analysis. Here we got 255 in-control batches out of 462.

Mean Vectors
-2.422
0.808
-0.110

Sample Covariance Matrix(3*3)		
0.027	-0.016	-0.001
-0.016	0.021	0.005
-0.001	0.005	0.021

Sample Correlation Matrix(3*3)		
1.000	-0.663	-0.035
-0.663	1.000	0.255
-0.035	0.255	1.000

In-control data sets(255*3)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X 1	-2.4823	-2.6346	-2.5816	-2.5743	-2.5776	-2.0559	-2.1472	-2.5689	-2.1157	-2.5116
X 2	0.8431	0.8890	1.0903	0.8984	0.7608	0.5418	0.7282	1.1143	0.7300	0.9429
X 3	-0.0291	-0.0155	-0.1306	-0.1815	0.0471	-0.1175	-0.0132	0.1547	-0.0176	-0.0393

	11	12	13	14	15	16	17	18	19	20
X 1	-2.2469	-2.0452	-2.4191	-2.0632	-2.5387	-2.4973	-2.3592	-2.0259	-2.0277	-2.3242
X 2	0.9221	0.7287	0.6340	0.6616	1.0178	1.0947	0.7946	0.5477	0.3705	0.7871
X 3	0.2340	-0.0057	-0.0602	-0.1189	-0.2752	0.2397	-0.0218	-0.0944	-0.1139	-0.0458

	21	22	23	24	25	26	27	28	29	30
X 1	-2.5390	-2.3117	-2.1124	-2.4386	-2.7135	-2.6750	-2.4141	-2.5818	-2.5208	-2.6021
X 2	0.8197	0.8461	0.5109	0.8576	0.9421	0.9981	0.7503	0.7284	0.7699	0.8142
X 3	-0.1447	-0.1451	-0.1331	-0.1531	-0.3273	-0.0704	-0.0365	-0.1296	-0.2250	-0.1107

	31	32	33	34	35	36	37	38	39	40
X 1	-2.4454	-2.5200	-2.5154	-2.3218	-2.1915	-2.6814	-2.4755	-2.4361	-2.6325	-2.4212
X 2	0.6192	0.9161	0.7268	0.7257	0.5156	0.8920	0.8545	0.8652	0.8863	0.7928
X 3	-0.0348	-0.2074	-0.0765	0.0073	-0.2326	-0.1620	-0.1122	-0.2946	-0.1073	-0.0734

	41	42	43	44	45	46	47	48	49	50
X 1	-2.1868	-2.6470	-2.4565	-2.2034	-2.6603	-2.2823	-2.3659	-2.4058	-2.2937	-2.5651
X 2	0.6743	1.0293	0.6902	0.6130	0.9283	0.5938	0.7037	0.6660	0.7017	0.8983
X 3	0.1388	0.0869	-0.2031	-0.0434	-0.2793	-0.2574	-0.0144	-0.2294	-0.2385	-0.2896

	51	52	53	54	55	56	57	58	59	60
X 1	-2.8019	-2.3580	-2.6189	-2.3754	-2.5096	-2.5737	-2.5465	-2.4179	-2.5032	-2.4510
X 2	0.9425	0.7041	0.9079	0.6848	0.8754	0.9862	0.8914	0.8390	0.6118	0.8263
X 3	0.1043	-0.1012	-0.2133	-0.2682	-0.3255	-0.1689	-0.1515	-0.1818	-0.3480	-0.1967

	61	62	63	64	65	66	67	68	69	70
X 1	-2.3096	-2.6067	-2.6075	-2.4404	-2.5863	-2.4205	-2.5288	-2.5073	-2.7575	-2.1754
X 2	0.5930	0.7804	0.9873	0.8897	0.8235	0.8150	0.7303	0.7077	0.9025	0.6103
X 3	-0.0474	-0.1522	-0.1646	-0.1993	-0.0071	-0.1007	-0.0622	-0.2351	-0.0158	-0.4540

	71	72	73	74	75	76	77	78	79	80
X 1	-2.2372	-2.3174	-2.3300	-2.5690	-2.4790	-2.5107	-2.6794	-2.3505	-2.4507	-2.6342
X 2	0.5339	0.8805	0.7587	0.9695	0.7775	0.9683	0.9222	0.7837	0.7053	0.8525
X 3	-0.0459	0.0038	-0.1517	-0.2896	-0.2427	0.1106	0.0756	0.0436	0.1280	-0.1116

	81	82	83	84	85	86	87	88	89	90
X 1	-2.5496	-2.4761	-2.8296	-2.5456	-2.6873	-2.5064	-2.2648	-2.5143	-2.4975	-2.8824
X 2	0.7832	0.7642	1.1431	0.7742	0.8481	0.7444	0.6954	0.9183	0.7858	1.0460
X 3	-0.1417	-0.0043	0.0294	-0.0585	-0.3189	-0.0312	-0.0759	-0.0045	-0.1421	-0.2107

	91	92	93	94	95	96	97	98	99	100
X 1	-2.1694	-2.3284	-2.2740	-2.4888	-2.3630	-2.5263	-2.4644	-2.5945	-2.5429	-2.5113
X 2	0.6727	0.9449	0.7033	0.8258	0.7088	0.9851	0.7600	0.9525	0.8004	0.8874
X 3	-0.3328	-0.1615	-0.1725	-0.1705	-0.1473	0.0018	-0.0495	0.0071	-0.0472	0.1368

	91	92	93	94	95	96	97	98	99	100
X 1	-2.1694	-2.3284	-2.2740	-2.4888	-2.3630	-2.5263	-2.4644	-2.5945	-2.5429	-2.5113
X 2	0.6727	0.9449	0.7033	0.8258	0.7088	0.9851	0.7600	0.9525	0.8004	0.8874
X 3	-0.3328	-0.1615	-0.1725	-0.1705	-0.1473	0.0018	-0.0495	0.0071	-0.0472	0.1368

	101	102	103	104	105	106	107	108	109	110
X 1	-2.0061	-2.0465	-2.2148	-2.2058	-2.3403	-2.4201	-2.6304	-2.2995	-2.2178	-2.3029
X 2	0.6339	0.7660	0.6367	0.7240	0.8328	0.7670	1.1023	0.5879	0.5635	0.8040
X 3	-0.0073	0.0326	0.0056	0.0593	-0.0845	0.0502	-0.0435	-0.2882	-0.1667	0.1218

	111	112	113	114	115	116	117	118	119	120
X 1	-2.2889	-2.2889	-2.4538	-2.2178	-2.6051	-2.4142	-2.5449	-2.5254	-2.6079	-2.4047
X 2	0.8952	0.7317	0.8306	0.6171	1.1110	0.8211	0.7848	0.8838	0.9477	0.8769
X 3	0.0472	-0.0784	-0.1501	-0.1016	0.0482	0.1570	-0.1039	0.0597	-0.0986	-0.0111

	121	122	123	124	125	126	127	128	129	130
X 1	-2.5349	-2.3556	-2.4020	-2.7174	-2.5980	-2.6896	-2.0717	-2.5505	-2.5090	-2.4083
X 2	0.8623	0.6541	0.8923	1.0911	0.9687	1.0010	0.6492	0.7387	1.0466	0.7892
X 3	-0.0741	-0.0237	0.0058	0.0229	-0.0595	0.0832	0.0099	0.0968	0.0445	0.2205

	131	132	133	134	135	136	137	138	139	140
X 1	-2.5404	-2.5220	-2.5321	-2.4684	-2.6454	-2.3847	-2.5995	-2.4369	-2.5130	-2.3050
X 2	1.0029	0.9647	0.6819	0.7396	0.9682	0.9535	1.0771	0.9083	1.0913	0.6919
X 3	-0.1718	-0.0144	-0.0815	-0.0707	0.2695	0.2018	0.0202	-0.0559	0.0476	-0.2642

	141	142	143	144	145	146	147	148	149	150
X 1	-2.6391	-2.5932	-2.6515	-2.4580	-2.7004	-2.3696	-2.5132	-2.3586	-2.3871	-2.1950
X 2	1.0392	1.0185	0.8714	0.7802	0.9990	0.4799	0.7295	0.7395	0.7260	0.6377
X 3	-0.1519	-0.1056	0.0794	-0.1311	-0.1318	-0.2276	-0.1481	-0.2197	-0.2249	-0.2535

	151	152	153	154	155	156	157	158	159	160
X 1	-2.0500	-2.5505	-2.4015	-2.3021	-2.3464	-2.3318	-2.4237	-2.2535	-2.4827	-2.4074
X 2	0.5083	0.8948	0.8001	0.4735	0.5513	0.5843	0.8932	0.4032	0.8677	0.9622
X 3	-0.2614	-0.1458	-0.0987	-0.3722	-0.2685	-0.2589	-0.0500	-0.1868	-0.1904	0.0405

	161	162	163	164	165	166	167	168	169	170
X 1	-2.4292	-2.3637	-2.3211	-2.4074	-2.2808	-2.5655	-2.4293	-2.6972	-2.4288	-2.5029
X 2	0.8065	0.9004	0.7578	0.7745	0.8254	0.9341	0.8371	0.9035	0.7396	0.9802
X 3	-0.1892	-0.0266	-0.1396	-0.2474	0.4050	0.0476	-0.0394	-0.2745	-0.3791	-0.0679

	171	172	173	174	175	176	177	178	179	180
X 1	-2.4723	-2.4760	-2.3726	-2.5983	-2.4882	-2.4632	-2.5541	-2.1530	-2.4546	-2.3876
X 2	1.0054	0.8604	0.8339	0.8694	0.7718	0.8318	1.0080	0.6770	0.8552	0.6265
X 3	-0.0392	0.0496	-0.1583	-0.3549	-0.1496	0.0315	-0.0358	-0.0021	-0.3837	-0.3368

	181	182	183	184	185	186	187	188	189	190
X 1	-2.5371	-2.4411	-2.4446	-2.4565	-2.5906	-2.2797	-2.4334	-2.3469	-2.4625	-2.6719
X 2	0.5953	0.8603	0.5765	0.9410	1.0596	0.6909	0.7490	0.8320	0.8034	0.8664
X 3	-0.3753	-0.3344	-0.0874	-0.2071	-0.1749	-0.2416	-0.1540	-0.1076	-0.0341	-0.0259

	191	192	193	194	195	196	197	198	199	200
X 1	-2.2576	-2.4149	-2.5124	-2.6389	-2.3779	-2.3283	-2.4555	-2.2327	-2.3379	-2.1672
X 2	0.7475	0.6393	0.9616	0.9057	0.8120	0.8691	0.8265	0.6975	0.7032	0.5003
X 3	-0.1597	-0.3414	-0.2505	0.0646	-0.0860	-0.3663	-0.1208	-0.2201	-0.0872	-0.4445

	201	202	203	204	205	206	207	208	209	210
X 1	-2.3819	-2.4538	-2.3842	-2.3702	-2.1745	-2.4035	-2.6019	-2.3506	-2.5548	-2.2461
X 2	0.7124	0.7820	0.5623	0.6818	0.7879	0.9125	1.0468	0.6728	0.8419	0.7997
X 3	-0.3791	-0.2923	-0.0911	-0.3881	-0.0231	-0.3251	-0.0609	-0.2716	-0.4044	-0.0097

	211	212	213	214	215	216	217	218	219	220
X 1	-2.4252	-2.3576	-2.2829	-2.3649	-2.4874	-2.4115	-2.5221	-2.2741	-2.3609	-2.7144
X 2	0.6617	0.9410	0.7943	0.8719	0.8960	0.7532	0.8213	0.6795	0.7451	0.9096
X 3	-0.2375	-0.1008	-0.1518	-0.2091	-0.2766	-0.0473	-0.1688	-0.1857	-0.4765	-0.1325

	221	222	223	224	225	226	227	228	229	230
X 1	-2.2371	-2.5428	-2.4892	-2.3543	-2.4968	-2.3051	-2.4088	-2.4909	-2.2724	-2.5366
X 2	0.7812	0.7443	0.6741	0.8285	0.7277	0.8160	0.9320	1.0415	0.7938	1.0764
X 3	-0.0096	0.0430	-0.0106	0.0655	-0.4977	0.1777	-0.1390	-0.0918	-0.1825	0.0210

	231	232	233	234	235	236	237	238	239	240
X 1	-2.3946	-2.1834	-2.3639	-2.3442	-2.4909	-2.1552	-2.1901	-2.0540	-2.4255	-2.2246
X 2	0.8781	0.8002	0.7549	0.7409	0.8902	0.8293	0.8210	0.5362	1.0549	0.7107
X 3	-0.1038	-0.1452	-0.3253	-0.1358	-0.1596	-0.3110	-0.0576	0.0295	-0.0581	-0.3062

	241	242	243	244	245	246	247	248	249	250
X 1	-2.3378	-2.4526	-2.0650	-2.4178	-2.3866	-2.4181	-2.1857	-2.3801	-2.5703	-2.2482
X 2	0.8473	0.8892	0.5688	0.9381	0.9383	0.8210	0.6935	0.8411	1.0385	0.7548
X 3	-0.2700	0.0285	-0.0807	-0.2078	-0.0064	-0.1473	-0.0281	-0.1194	-0.2802	-0.0711

	251	252	253	254	255
X 1	-2.4492	-2.4778	-2.3110	-2.0591	-2.1456
X 2	0.9027	0.8986	0.7829	0.6536	0.5583
X 3	-0.1829	-0.0373	-0.1882	0.0102	0.0086

CONCLUSION

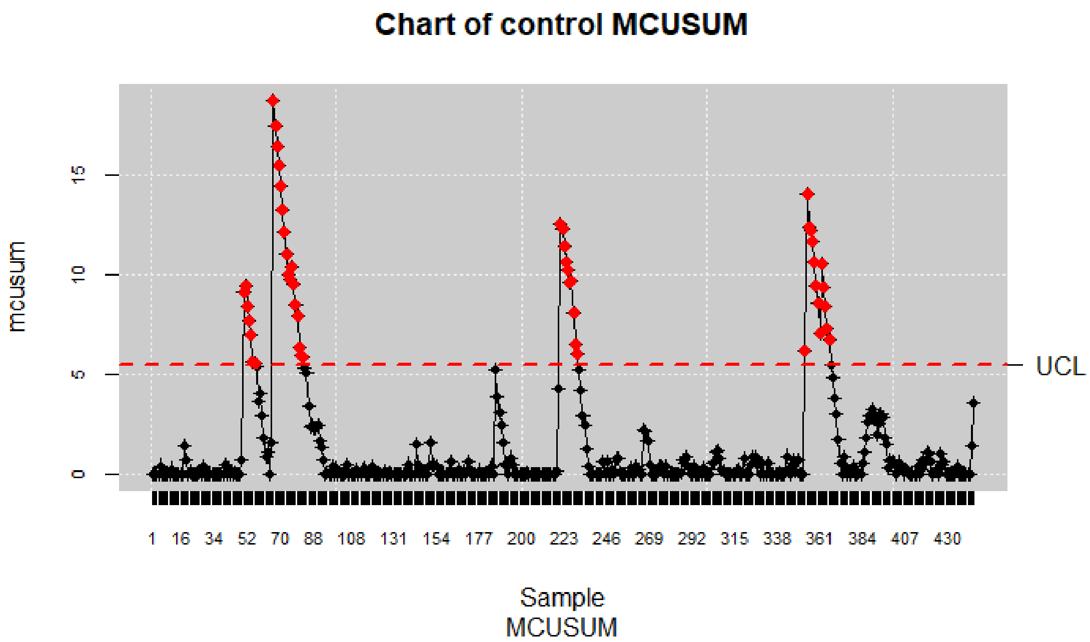
Phase 1 analysis, when the sample size is 1, is quite different from other types of cases because here X_j(variable) is not independent of S(sample covariance matrix). This results in t^2 following beta-like distribution but it is generalized to follow chi-square distribution. Having more dimensional data sets will add more complications to it. So, using PCA and relevant PCs selection procedure we are able to capture 82.2% of variances using 3 PCs. Reducing the dimensionality is the pre-processing procedure before we actually eliminate the out-of-control data. By repeatedly doing control charting we eliminated almost 207 out-of-control batches.

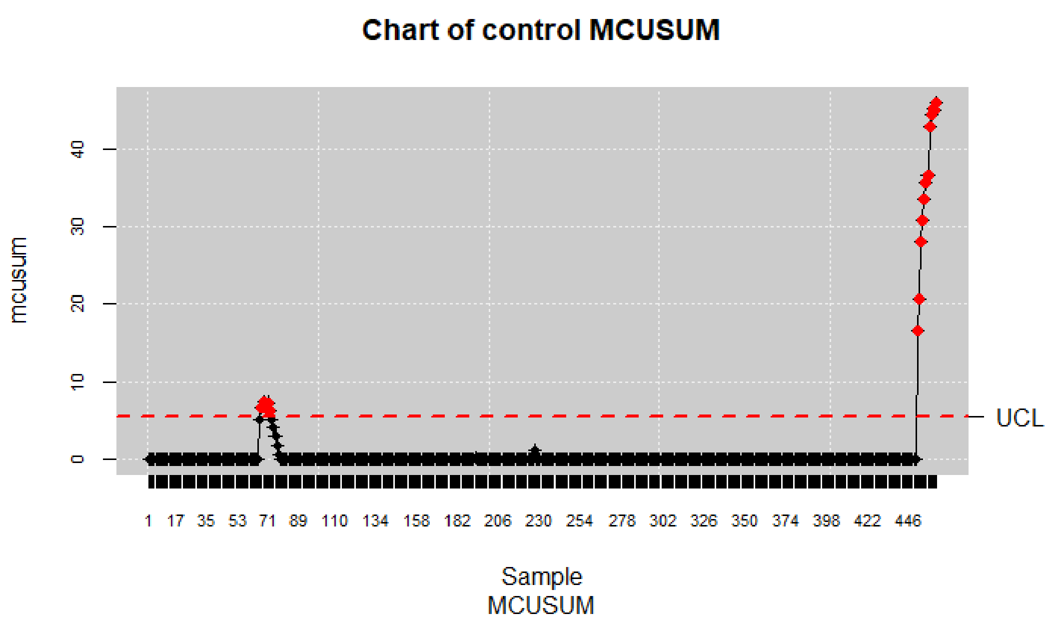
By doing MCUSUM alone we almost eliminated 202 batches. So, doing MCUSUM was enough to remove the out-of-control data but we used the t^2 chart to make sure that even spike-like changes in cursing data are also removed. As phase 2 analysis or other important procedures after phase 1 analysis are completely dependent on the output of phase 1 analysis it is vital to do a sanity check to make sure that we give the correct parameters for further analysis. We have 3 mean vectors and 3*3 sample covariance and correlation matrix along with in-control data for phase 2 analysis.

Appendix

Controlling Chart for M-CUSUM:

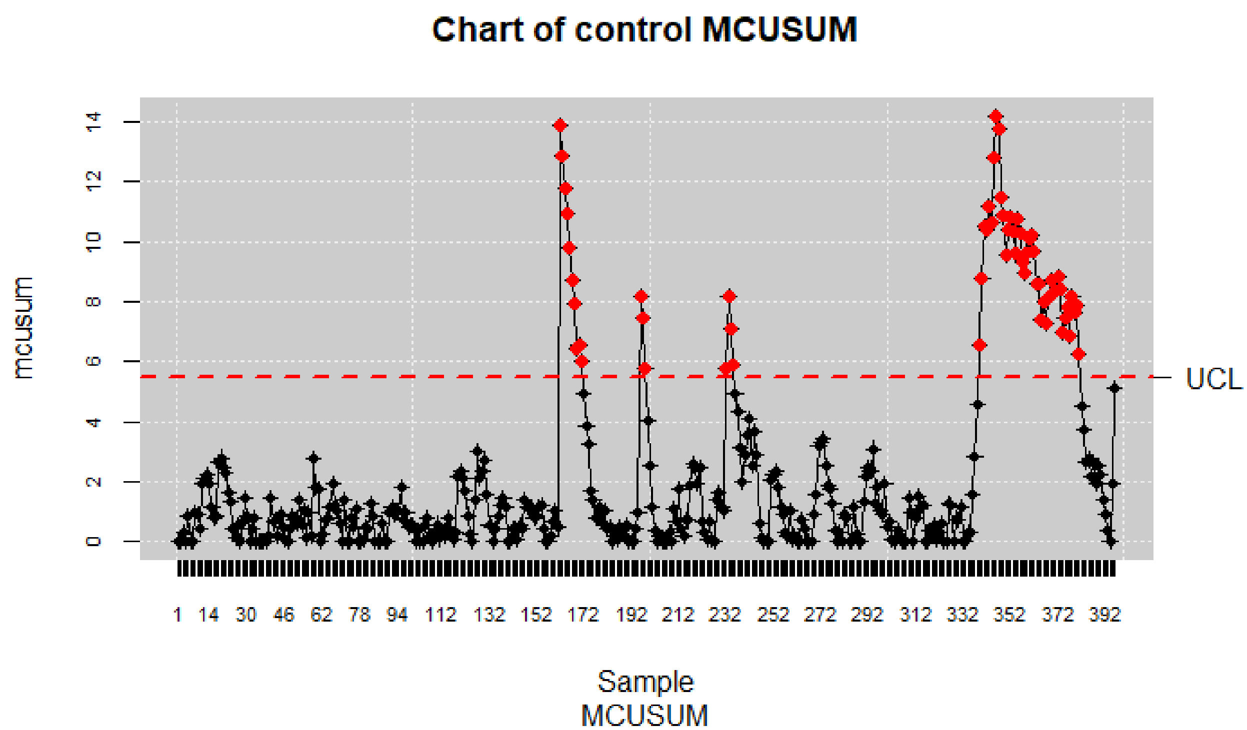
Iteration 1



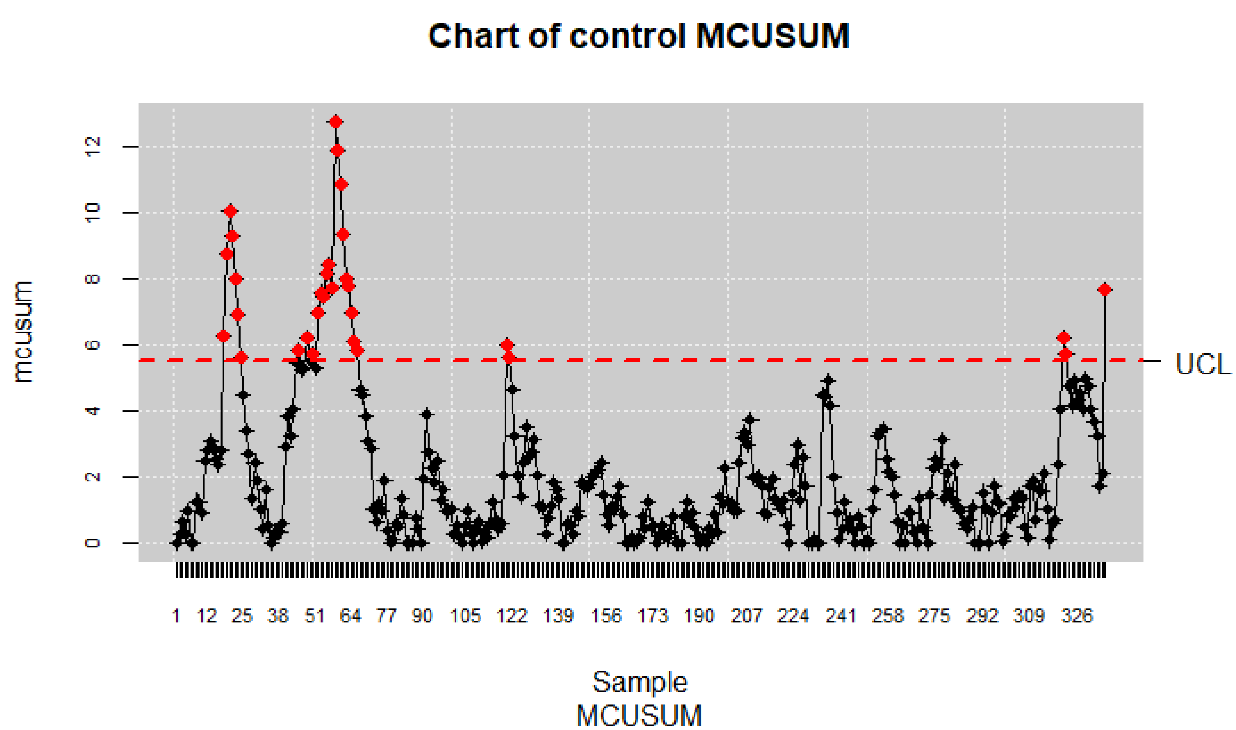


Iteration 2

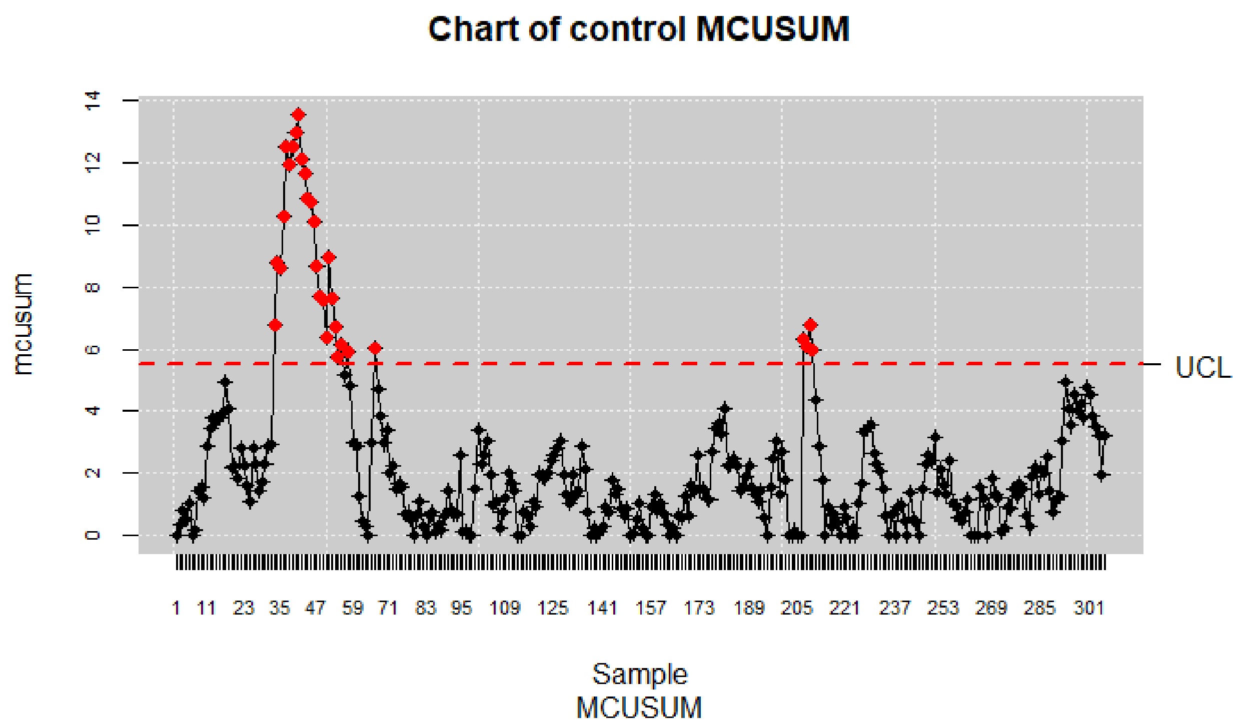
Iteration 3



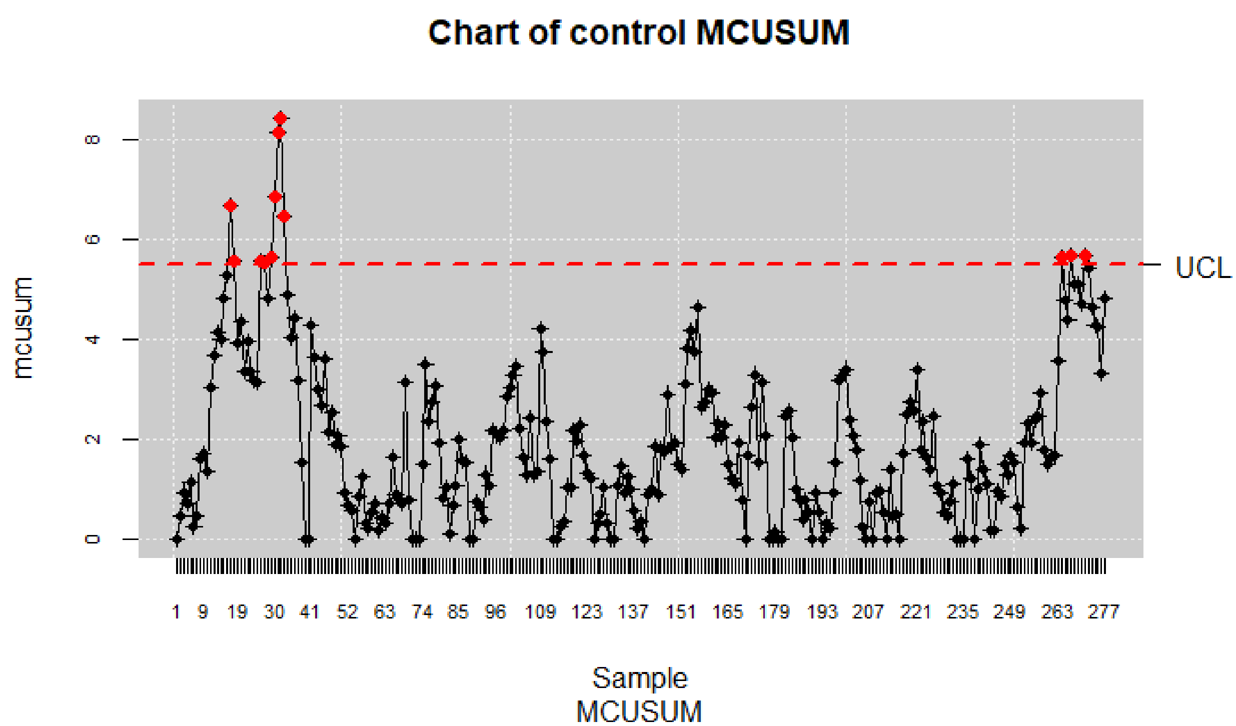
Iteration 4



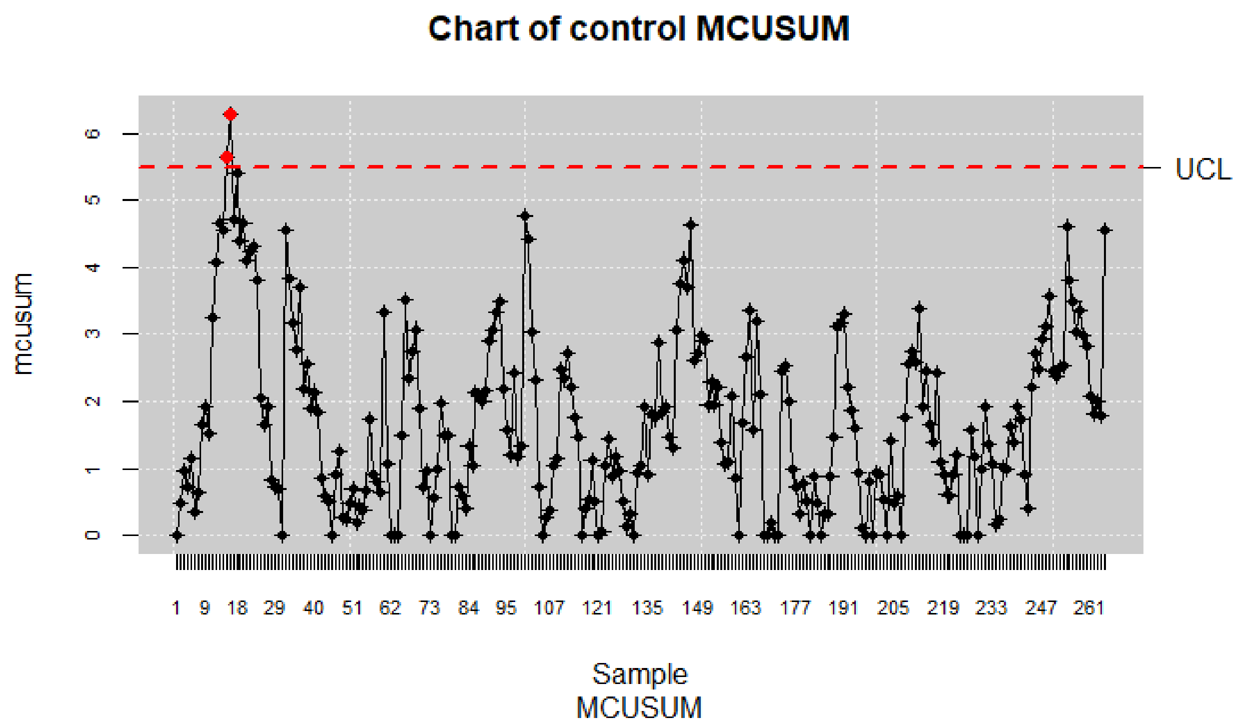
Iteration 5



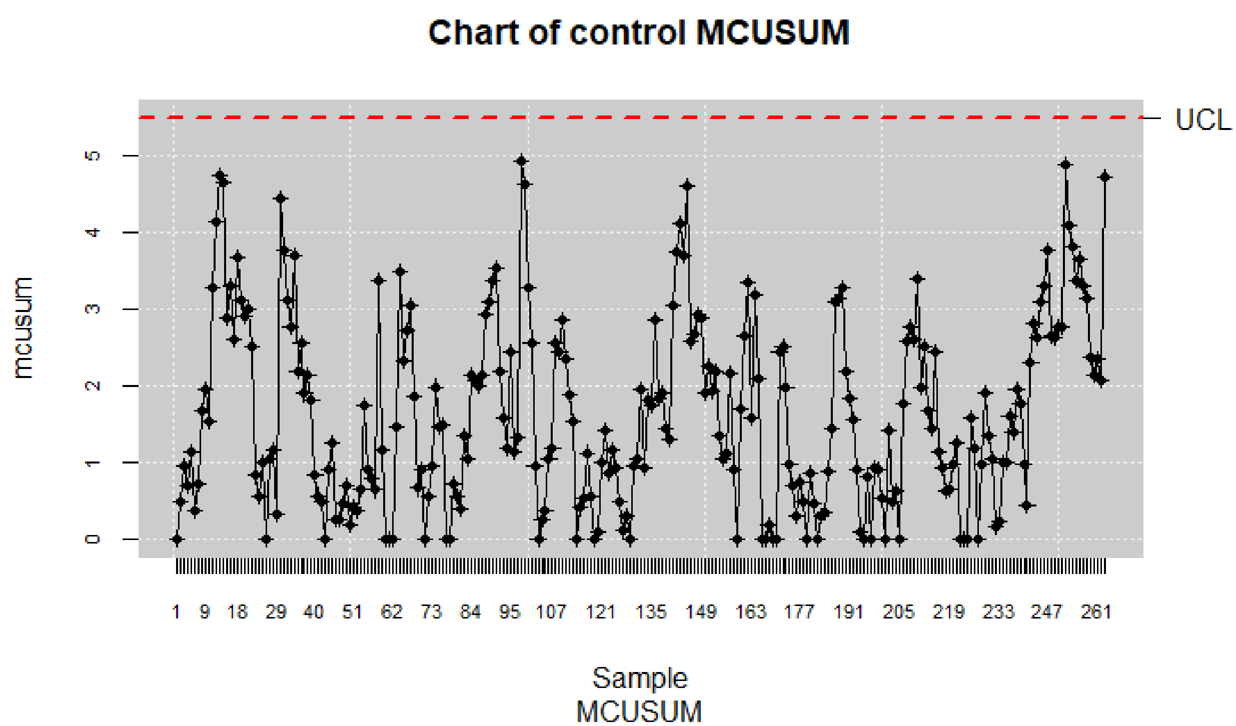
Iteration 6



Iteration 7

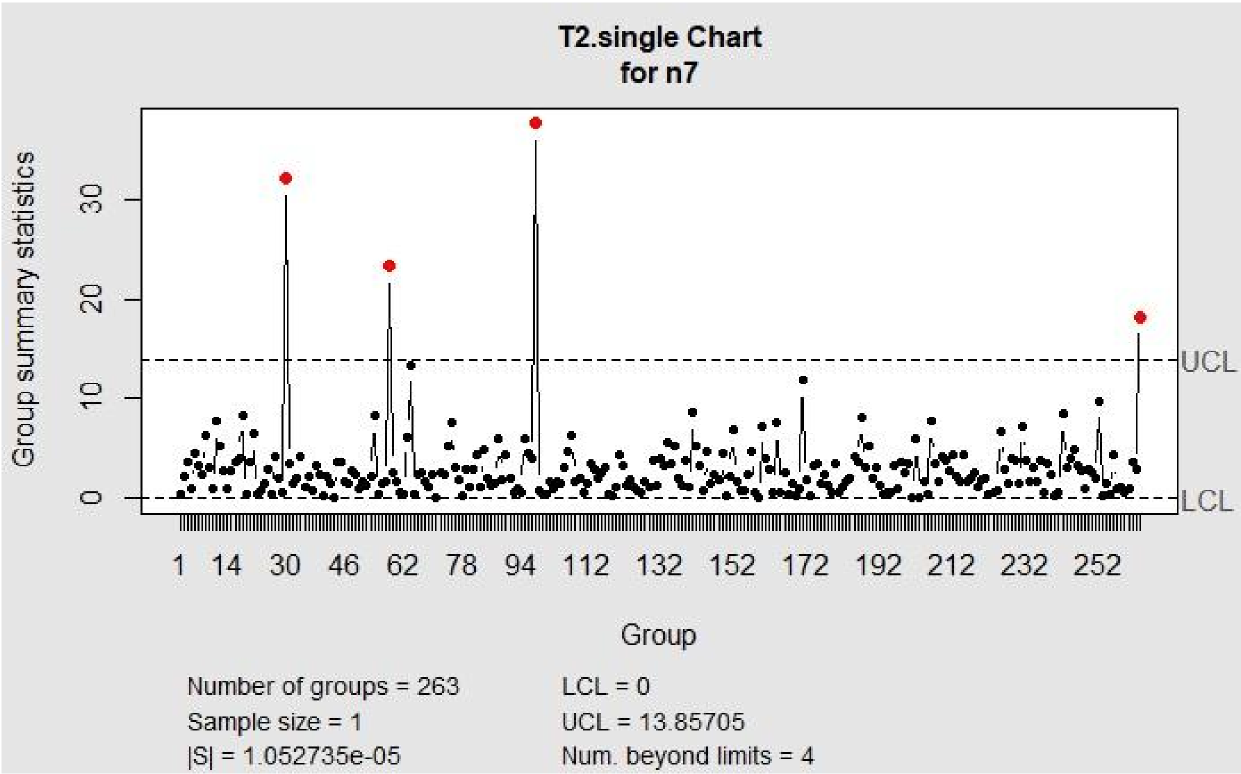


Iteration 8

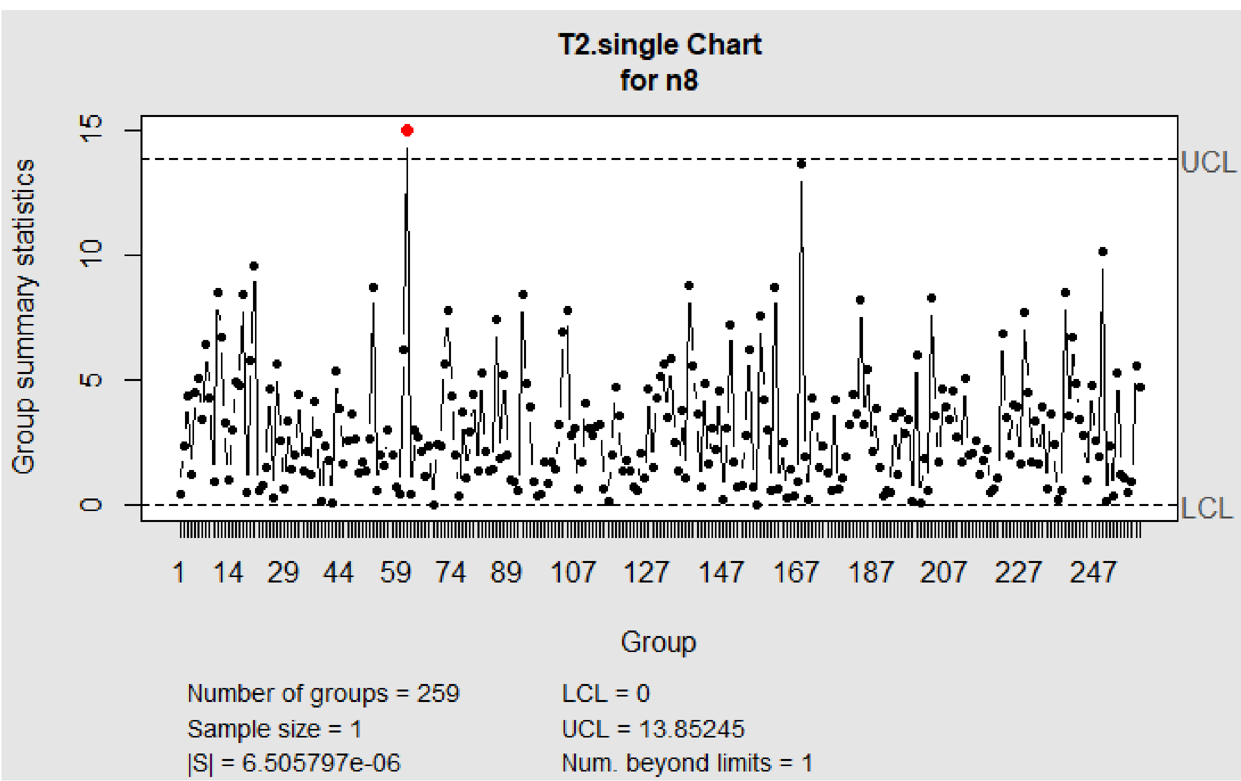


Controlling Chart for T^2 :

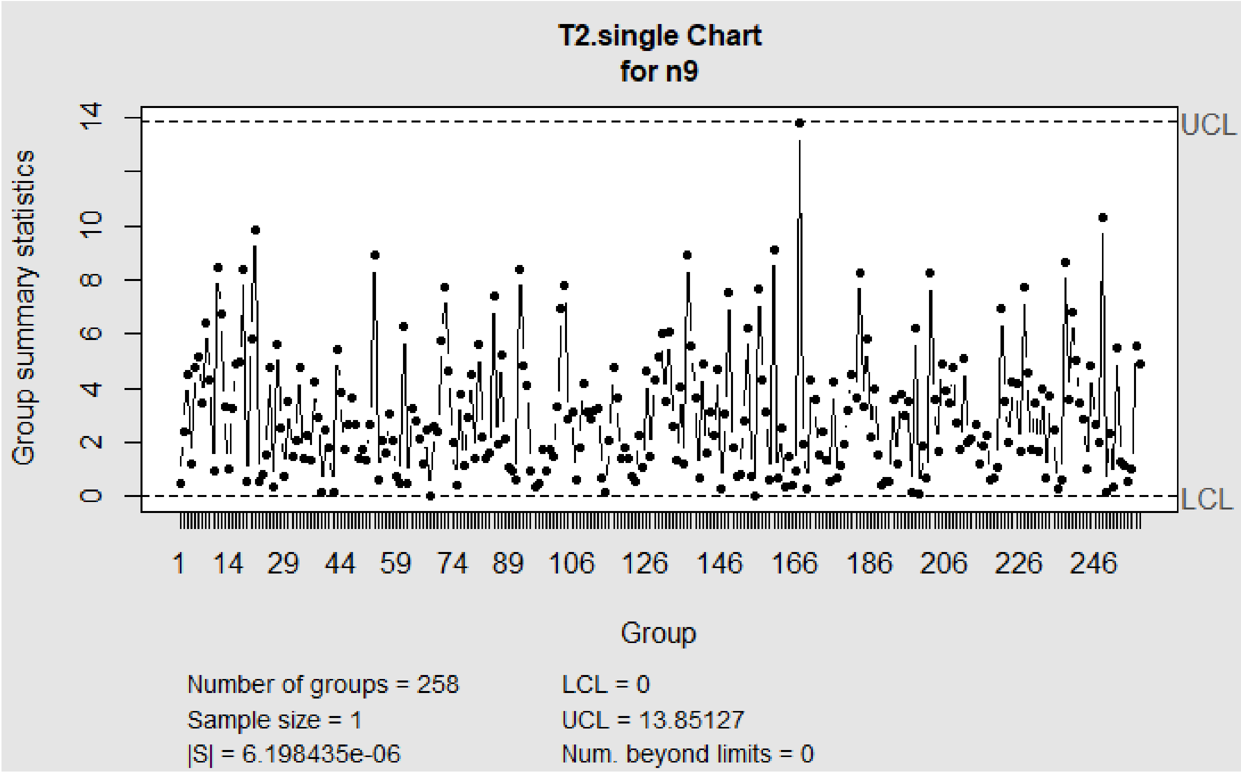
Iteration 1



Iteration 2



Iteration 3


T2.single Chart
for n9
Group summary statistics
0
2
4
6
8
10
14
UCL
LCL
1 14 29 44 59 74 89 106 126 146 166 186 206 226 246
Group
Number of groups = 258
Sample size = 1
|S| = 6.198435e-06
LCL = 0
UCL = 13.85127
Num. beyond limits = 0

REFERENCES

---

[1] A Multivariate CUSUM Control Chart for Monitoring Gumbel's Bivariate Exponential Data
July 2020/Quality and Reliability Engineering 37(1)/DOI:10.1002/qre.2717

[2] Principal Component Analysis via correlations or covariances? no better alternative?
November 2020/DOI:10.13140/RG.2.2.24735.18088

[3] Principal component analysis: a review and recent developments
Ian T. Jolliffe and Jorge Cadima/Published:
13 April 2016/https://doi.org/10.1098/rsta.2015.0202

[4] Principal Component Analysis in R; PCA of covariance or correlation matrix/Aedin Culhane/July 15, 2021

[5] Be careful with your principal components

Mats Björklund

First published: 21 August 2019 https://doi.org/10.1111/evo.13835Citations: 32

[6] Why principal component analysis is not an appropriate feature extraction method for hyperspectral data/publisher:IEEE

[7] ISEN 614 course pack 2022 spring