Texas A&M University

ISEN 614 Course Project

Spring 2022

Team number: 13

Adhikari, Kiran. 731008005

Gawali, Aniket. 932002258

Table of Contents

1. Executive Summary

Data collected from a manufacturing process was provided, in which both in-control and out of control data were present. Our objective for this ISEN 614 project was to develop the best method to separate in-control and out-of-control data, a phase I analysis. We calculated the distribution parameter from control data which will be used for quality control monitoring of future unseen data. We took the ISEN 614 course reference and explored the literature related to the scope of our project to accomplish our objective. R programming language was used to do the analysis.

Due to the high dimensionality of the data, we were aware that the signal would get highly compromised, and detection would be complex. So we transformed our data into principle components. The covariance matrix was chosen to do so, and we have justified our preference in the correlation vs. covariance analysis section. After computing PCA, we found out that four principles captured around 80% variance of our data. We settled with 4 principle components based on scree plot results. We used four multiple univariate control charts for phase one analysis, and our principle components follow the standard normal distribution. We used the Industry-standard $\mp 3\sigma$ limit to set up the upper and lower control limits. Data points outside these control limits in each control chart were removed and updated data sets were plotted again on the control chart until no principle components data points were outside the control limit. This computation was carried out using a loop function.

After phase one analysis, we separated in-control data from the original dataset. The distribution parameter for future monitoring will be based on in-control data parameters. If we get new data, we have to transform it into four principle components and plot it in the univariate chart. We recommended MCUSUM and EWMA Chart alongside multiple univariate charts. The univariate chart is good at detecting a single large spike in data, whereas CUSUM and EWMA are good at detecting small mean-shift. If we follow both charts, we can get the advantages offered by both methods.

To decide whether our system is out of control or not, we explored four decision criteria. We then established the best decision rule based on Arlo and ARL1 values we get after calculation. Out of four decision rules, decision rule no 2 (if at least two univariate charts signal out of control, then the system is out of control) was acceptable and used for future monitoring.

2. Choice of method

For a given multivariate data, we have different methodologies available, including but not limited to Hotelling T^2, mCUSUM, mEWMA, and Principle Component Analysis. Given dataset has 209 variables. The accumulation of noise components from individual elements in such datasets will significantly increase. It could overwhelm the actual signal. This situation is termed as the curse of dimensionality[i]. Thus, we can not directly apply Hotelling T^2, mCUSUM, or mEWMA to the dataset. We will reduce the dimension to avoid this curse. Therefore we will employ PCA or Principle Component Analysis for data reduction. Another advantage of this method is it translates the variables into a set of uncorrelated variables with the same magnitude. The risk is we might lose some information that is only visible in higher dimensions. But this can be mitigated as well by using multiple charts for monitoring. Let us not concern ourselves with that and continue our analysis using PCA. We will try two methods for phase 1 analysis:

1. PCA with multiple univariate charts
2. PCA with mCUSUM

Number 1 is the apparent option to try since we do not have any correlation between the Principle Components. This method will also give the most flexibility in determining the performance i.e., $ARL_0 \; and \; ARL_1$. We can decide the signaling policy to adjust the performance. Woodall and Ncube (85) [ii]have shown that mCUSUM is often preferable to the Hotelling T^2 procedure. Hence a worthy method to try. We are using mCUSUM over EWMA because of its popularity. This method is not necessarily better or worse than the other, but historically, it's popular.

2.1. Correlation Vs. covariance analysis

Because a correlation matrix is standardized, the relative scale changes among elements so much that the eigenvalue/vectors are different[iii]. We observe the same phenomenon here. 1st PC explains 20% of the variability, which dropped from ~70% when using the covariance method.

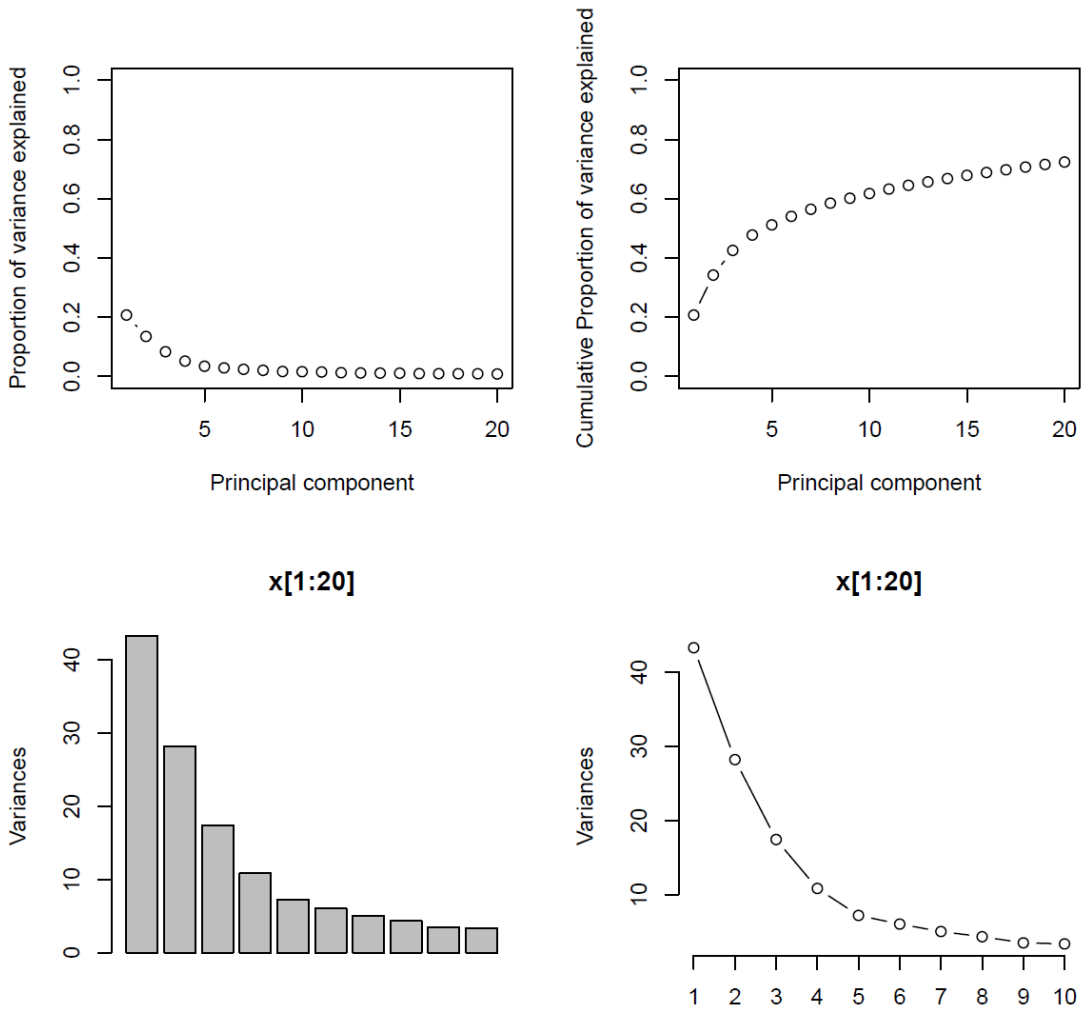

Figure 1 Screeplot for correlation

This drastic change could be due to the change in the scaling or difference in variances of variables. Again, we do not have any information on the application side. Units of the data are also unknown to us. So we can not tell for sure. This interpretation is by looking at the range of the dataset, which varies from -97 to 1442. The 1st variable ranges from -6 to 4, but the 100th variable ranges from 1168 to 1365. Borgognone, Bussi, and Hough (2001)[iv] show that the correlation matrix method can produce misleading results if the difference in the variances is significant. Here we will put our good faith in the capable hands of manufacturing engineers and assume this dataset has the same units, metric or imperial. It does not make sense to use different units for the same application. On the other hand, this dataset's variance is too large to ignore, and we might lose information by normalizing these variables. Hence, we justify the use of the covariance matrix method.

2.2. Choice of design parameters for PCA

Now that we have selected the PCA, we will need to decide how many principle components we should use, what will be our α (alpha) and should we use the Variance or Covariance matrix for the analysis.

2.2.1. Choice of α

This parameter will depend on the application. Accepted error for manufacturing Pacemaker will be completely different from manufacturing water bottles. We do not have any information on the application side for this dataset. All we know is it is the dataset for manufacturing. Thus, assuming industry-standard ∓3σ process control, our α will be 0.0027.

2.2.2. Number of PC selected

Although multiple methods are available to select the number of components, they are still very subjective. Results will vary depending on the number of variables and the data distribution. We have used the Cattell's Scree test, MDL, and Kaiser-Guttman (K-G) rule to identify the number of PCs. Results are summarized in Table 1.

Method	Covariance		Correlation	
	No. of PC	Variance explained	No. of PC	Variance explained
MDL	35	94.83%	35	80.95%
Scree test	4	80.10%	8	58.50%
K-G rule	15	89.96%	32	79.55%

Table 1 Comparison of methods for selecting the number of PC

2.2.2.1. MDL

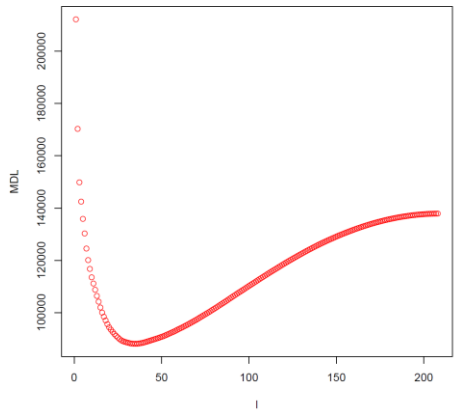

Figure 2 MDL plot for covariance method

MDL analysis gave the optimum number of PC as 35. The problem with MDL is it retains too many principle components than necessary[v]. Here keeping 15 components explain ~90% variability. It doesn't make sense to include 20 more PC for additional 5% variability.

2.2.2.2. Cattell's scree Test

Looking at the proportion of variance explained vs. pc graph, the "elbow" is at component 4. There is not much improvement in 4 to 5 or subsequent PC. But the first 4 PC only explain ~80% of the variance.

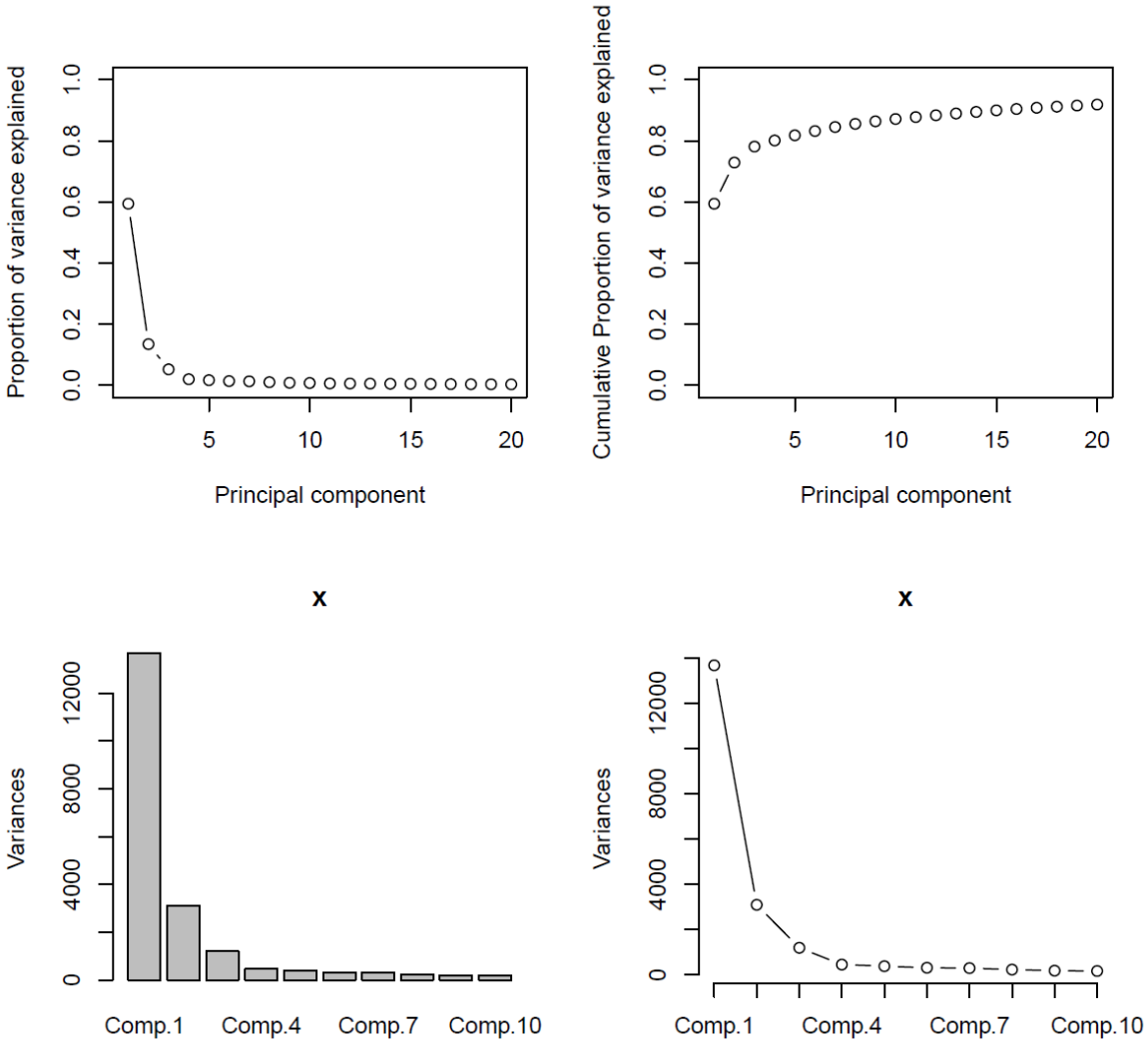

Figure 3 Screeplot using covariance method.

2.2.2.3. Kaiser-Guttman (K-G) rule

This rule says to retain the eigenvalues greater than the average. The average eigenvalue is 110 for the covariance method and 1 for the correlation method. As per the rule, we found that 15 and 32 eigenvalues were above average for covariance and correlation, respectively.

Having these estimates, choosing one is subjective. As Hakstian, Rogers, and Cattell (82)[vi] discussed, the method's performance varies given the number of variables, dataset size, variable commonalities, and proportion of major factors present. Our dataset does not resemble the formal or middle model considered in the paper for analysis. We have independent variables as formal and minor factors as in the

middle model. Here we will ignore the involvement of minor factors and assume we have the formal model. For a large number of variables (p>50), it is shown that the Scree test significantly outperformed the K-G rule. Hence, choosing 4 Principle Components for the analysis as per the Scree test.

3. Methodology

We have used prcomp() function in R to perform the analysis. This function returns PC values, rotation, variance etc.

1. Calculate the upper and lower control limits using the following formulas:

$$UCL = 3 \cdot \sigma$$

$$LCL = -3 \cdot \sigma$$

2. For each principle component used, compare the control limits with the value of scores/principle components. The scores lying above the UCL and below the LCL are the outliers. We will remove the outliers from our dataset.
3. Step 1 and 2 forms 1 iteration. After we remove the outliers, the data has essentially changed. This means variance will change, changing the control limits. Thus, we have performed Steps 1 and 2 in a loop until no outliers are left.

4. Sanity Check

Here we have verified that the output we got using prcomp() matches the theoretical method we have used in the coursework. We confirmed that eigenvalues, eigenvectors, and Principle component values remain the same.

Calculation of eigenvalues and eigenvectors is quite simple in R. We verified that the output of eigen(S)$values remains the same as prcomp(data)$sdev^2 where S = covariance matrix and data = given dataset. Similarly, eigen(S)$vectors also match with prcomp(data)$rotation.

To calculate components using the theoretical method, we have used the formula:

$$PC_i = e_i^T \cdot \left(x - \mu \right) \quad \text{for i = 1, 2, 3, 4.}$$

Where, e = eigenvector for i^{th} component.

x = given data of dimention n x p

μ = mean

From here, we verified that the output of our PC components using the above equation is the same as the output using the prcomp(data)$x.

We calculated the UCL and LCL from the above data for the first iteration and matched the outliers. Outliers from both the methods matched, and thus we can confirm that the way we used for calculating outliers is correct.

5. Performance of chart

We analyzed four decision rules to decide whether our system is out of control or not based ON the signal we received on our univariate charts. Decision rules are as follows:

1) If at least one chart is out of control, then the system is out of control
2) If at least two charts are out of control, then the system is out of control
3) If at least three charts are out of control, then the system is out of control
4) If all four charts are out of control, then the system is out of control

Decision rule	ARL_0	ARL_1
1	9	2
2	22945	8
3	1.27E+07	71
4	1.88E+10	1576

Table 2 Values for ARL for various decision rules

Our individual control charts follow a normal distribution, and they are not correlated. We considered $\mp 3\sigma$ mean shift for each univariate control chart and a 2σ mean detection for individual alpha and beta values. We calculated composite alpha and beta values for each decision rule and calculated ARL_0 and ARL_1 values. Out of for decision rule, decision rule no 2 has ARL_0 22945 and ARL_1 8, which is better than other decision rules. Hence, we will determine the system is out of control for future data if at least two univariate chart signals are out of control.

6. Conclusion

We have established the control charts for 4 PC as shown in Figure 4 and the control limits for 4 PCs are in Table 3.

Principle Component	UCL	LCL
PC1	244.97991	-244.97991
PC2	128.57560	-128.57560
PC3	103.20233	-103.20233
PC4	64.23331	-64.23331

Table 3 Control Limits for principle components

These control charts can be used for phase 2 analysis. Note that we are detecting the mean shift of 2σ.

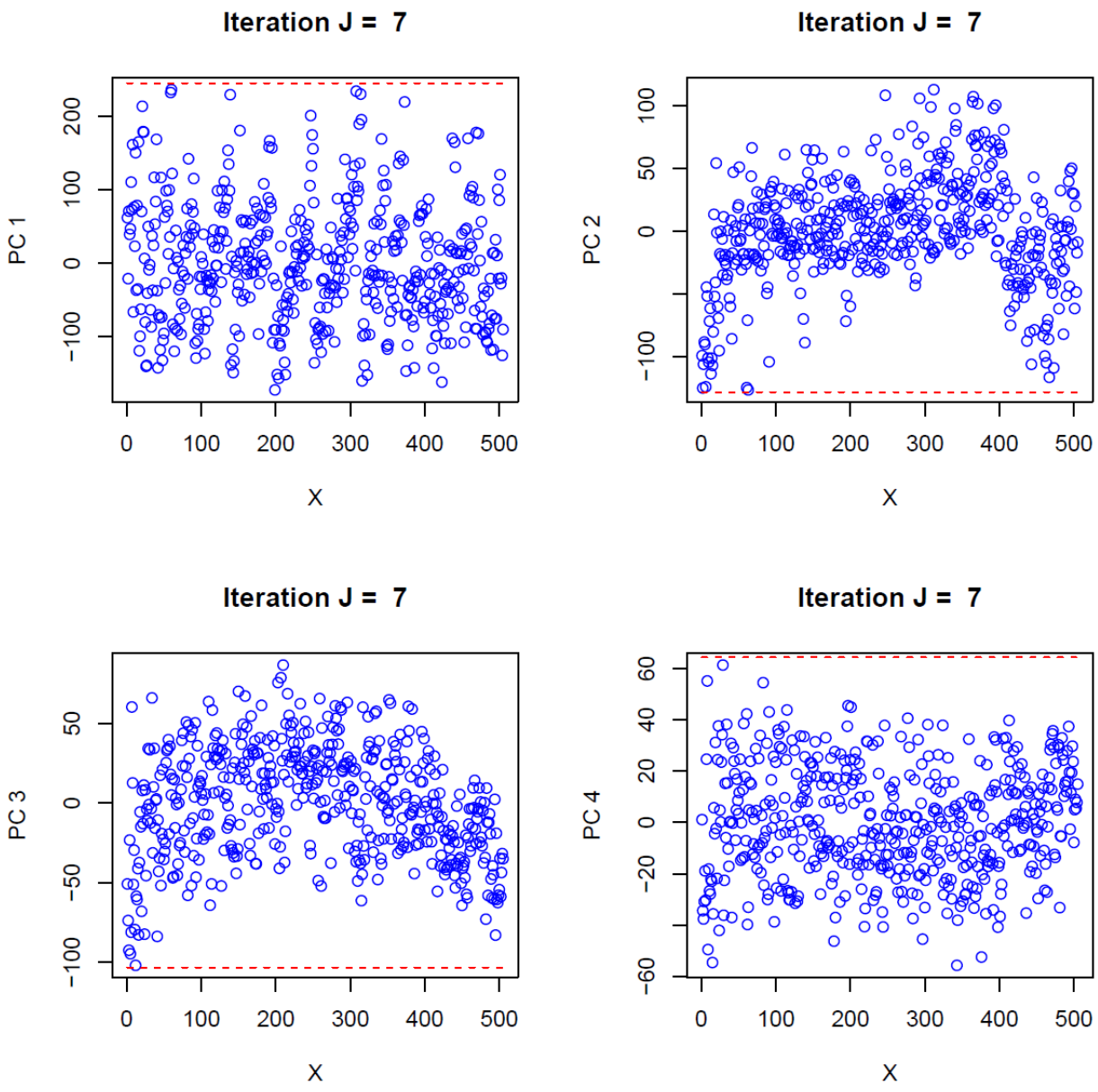

Figure 4 Control limits

7. References

[i] ISEN 614 Course Notes, Yu ding, Unit 41, Page 1.

[ii] William H. Woodall & Matoteng M. Ncube (1985) Multivariate CUSUM Quality-Control Procedures, Technometrics, 27:3, 285-292, DOI: 10.1080/00401706.1985.10488053

[iii] ISEN 614 Course Pack, Yu ding, Unit 42, Slide 450.

[iv] María G Borgognone, Javier Bussi, Guillermo Hough, Principle component analysis in sensory analysis: covariance or correlation matrix? Food Quality and Preference, Volume 12, Issues 5–7,2001, Pages 323-326, ISSN 0950-3293,

[v] ISEN 614 Course Notes, Yu ding, Unit 42, Page 5.

[vi] A. Ralph Hakstian , W. Todd Rogers & Raymond B. Cattell (1982) The Behavior Of Number-Of-Factors Rules With Simulated Data, Multivariate Behavioral Research, 17:2, 193-219, DOI: 10.1207/s15327906mbr1702_3