UNIVERSITY OF BIRMINGHAM

ON SOME MULTIVARIATE PROCESS CONTROL CHARTS

by

FADHIL ABBAS HASAN ALFARAG

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Mathematics
College of Engineering and Physical Science
The University of Birmingham
June 2016

Abstract

To maintain the quality of a product or to improve the reliability of a process, all industries need to monitor several parameters about their production process. Control charts are some visualization tools for monitoring processes statistically. They have been in use in the manufacturing processes for a quite long time, but all of them were based on either a single characteristic of the process or they used several different charts for different characteristics ignoring the dependence between the characteristics. With the ease of computing power and advances in technology, it is now easier to monitor several characteristics at the same time and to include their interdependencies as well. In this work, we propose a few control charting schemes to monitor several characteristics of a process at the same time and to detect when it goes out of control. Our objective is to reduce the false alarms (the scheme detects a problem when actually there is none) as well as to quickly detect the correct out-of-control situation. The novelty of the proposed schemes are that they do not depend on commonly assumed Normal distribution of the process variables and is applicable for a much wider range of data distributions.

At first, we make a detailed literature review of some univariate and multivariate control charts. We perform a comparison study of the commonly used multivariate control charts when the underlying distribution is not normal and show that they perform poorly giving a very high false alarm rate. Next we propose some nonparametric multivariate control charts based on the lengths of the multivariate rank vectors. The ideas are sim-

ilar to the ones proposed by Liu (1995), however, we show that our proposed methods are computationally simpler in any dimension and we study their performance through simulations and real data examples.

We propose some more multivariate versions of Shewhert type, CUSUM and EWMA control charts based on spatial sign vectors and signed rank vectors. We briefly discuss the issue of affine invariance and study the performance of the proposed charts through simulations. We also discuss several design parameters in the construction of these charts. None of the proposed charts depend on the assumption of underlying distribution or estimation of distributional parameters.

Acknowledgements

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Statistical Quality Control

Today maintaining quality of a product or service is paramount to any organisation providing that service or making that product and to achieve that goal statistical process control procedures have a very important role to play. Some of these procedures continuously monitor a process characteristic and produces some out of control signal when the variability in the process goes beyond the acceptable limits. One of the most widely used procedure is a control chart, which was first proposed by Dr. Walter Shewhart in 1924. Control charts gained popularity during the World War II to maintain the quality of ammunitions and other war related products. After that, its uses declined for a while. But these procedures again gained popularity in Japan and other countries in developing new technologies, and by the new millennium, everyone realised the importance of controlling the quality of processes using statistical techniques.

There are two main causes of variation in a process. The natural factors in the process which occur randomly are referred to as the common causes of variation. The common causes are considered as chance causes and are unavoidable. The special cause of variation considered as an assignable and identified cause of variation. The special causes

1

of variation should be discarded. The process that is operating with only common causes of variation present is said to be in statistical control. The special or abnormal cause of variability may sometimes be present in the output of a process. Some of the sources of this abnormal variation in the process are: improperly adjusted machines, operator errors and defective raw materials Kruger and Xie (2012). Control charts are visual tools which aid to detect when the processes are out-of-control. We may describe here an example from Montgomery (2009), which deals with the manufacturing of automotive engine piston rings, where the outside diameter of the ring is an important quality characteristic. The process can be controlled at the mean ring diameter of 74 mm, and standard deviation of 0.01 mm. A basic control chart for average ring diameter provides a central line at 74mm and the upper and lower control limits at 74.0135 and 73.9865 respectively. At every half hour a random sample of 5 rings is taken. The average ring diameter of the sample, $\bar{X}$ is computed and plotted on the chart. If all of the points fall within the control limits, we can say that the process is in statistical control. However, if an observation falls outside the control limits, we say that the process is out of control.

The univariate statistical process control charts can be defined as a monitor for one variable or quality characteristic of a process. It is important to detect the shifts in the location parameter of the univariate quality characteristic as quickly as possible. Some of the most popular univariate control charts for monitoring means of the process variable are univariate Shewhart $\bar{X}$-charts, the exponentially weighted moving average (EWMA) charts and the cumulative sum (CUSUM) control charts Ryan (1989). The Shewhart chart can be quick to detect large shifts in the process mean. However, the EWMA and CUSUM control charts are more effective in detecting small shifts in the process mean than Shewhart control chart.

The main feature of any control chart is that, it contains a central line that represent the target value and two other critical values, namely, the upper control limit (UCL) and

the lower control limit(LCL). These control limits are chosen to determine if all of the sample points are within the lower and upper control limits and the process will be in-control. Figure 1.1 is a simple example of a control chart where the process is in-control. On the other hand, if a few of the sample points are plotted outside the UCL or LCL, it means that the process is out of control and there may be some special cause of variation. As an example, the control chart in Figure 1.2 shows that the process is moving away from its in-control state near time point 15 but then going back to the in-control state near time point 16. This suggests that we need to investigate and find a corrective action.
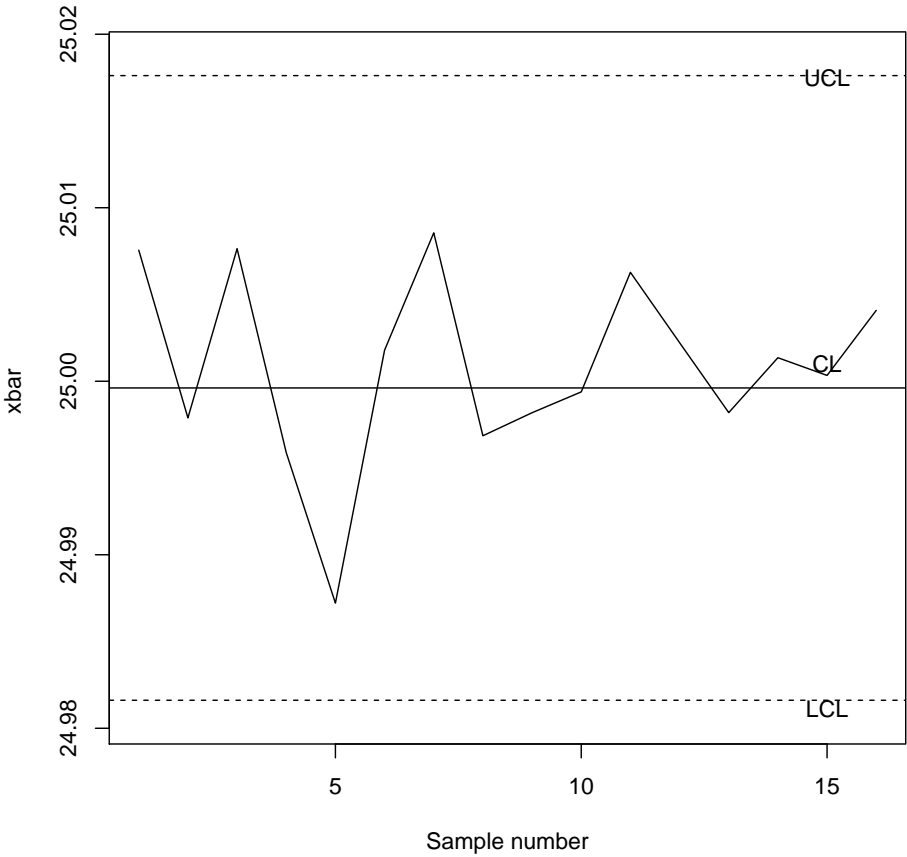
Figure 1.1: Control chart – the process is in control
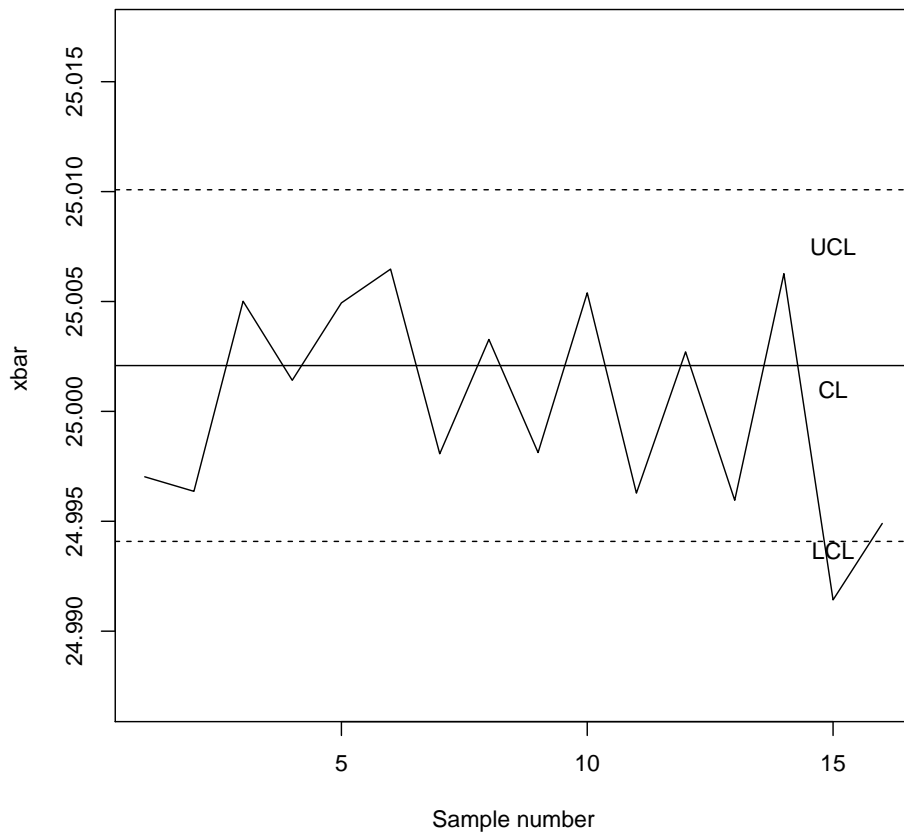
Figure 1.2: Control chart – the process is out of control

1.1.1 Some Definitions

Here we define some of the terminologies which will used through out this thesis.

Definition 1.1.1 *A Process*: *A process is a way to take input(materials) from a provider, transform them and obtain output (products) and offer this to a customer. The process is involved in production and quality control relevant areas.*

Definition 1.1.2 *Signal*: *The signal indicate the processes have most likely changed and there is a shift in the process characteristic which is being measured.*

4

Definition 1.1.3 *In control: The process is said to be in control if the variation in the process characteristic is only due to common chance causes.*

Definition 1.1.4 *Out-of-control: The process is said to be out of control if it is not producing according to specifications and the variations are due to special or assignable causes.*

Definition 1.1.5 *False alarm: The control procedure indicates an out-of-control signal when the process is in-control.*

Definition 1.1.6 *Phase I: Phase I of the monitoring scheme is used to check retrospectively whether the reference or historical data is in-control or not.*

Definition 1.1.7 *Phase II: Phase II of the monitoring scheme is used to check whether the future observations are still in-control or not with control limits obtained from in-control Phase I data.*

Definition 1.1.8 *Spherical distribution: A random vector X has a distribution spherical symmetric about μ if rotation of X a round μ dose not change the distribution $X - \mu = A(X - \mu)$, where A is a orthogonal $d \times d$ matrix.*

Definition 1.1.9 *Elliptical distribution: A random vector X has a distribution elliptical symmetric with parameter μ and Σ if it is an afinally equivariant to that of a spherically symmetric random vector X, $Y = A^T X$, where $\Sigma = A^T A$, with $rank(\Sigma) = k \leq d$.*

Definition 1.1.10 *Run length: The number of samples taken from the process before a control chart produces an out-of-control signal.*

Definition 1.1.11 *Average Run Length (ARL): The expected number of samples required for the control chart to produce a signal that the process is out of control, is called*

the Average Run Length (ARL). It is a major criterion for measuring the performance of a control chart Chang and Wu (2011). A control chart is said to be good when it has a large ARL with the process in-control. On the other hand, it will have a small ARL when the process is out-of-control, ARL=1/P(A point fall outside the control limits).

1.2 Review of Some Control Charts

1.2.1 Shewhart Charts

Most of the theoretical work for Shewhart control charts were developed in 1920s (Oakland, 1996). The basic idea is based on central limit theorem. If we take a sample of size n, from a population with a mean μ and a standard deviation σ, then as n increases, the distribution of sample mean, $\bar{x}$ approaches a normal distribution with a mean μ and a standard deviation of $\sigma/\sqrt{n}$. In Shewhart control chart we plot the sample mean against the sample number. The control limits will decide whether the process is in-control or out-of-control. Suppose X is a sample statistic, and that the mean of X is μ_x and the standard deviation of X is σ_x. Then the control limits for univariate Shewhart control charts are given as follows:

$$UCL = \mu_x + K\sigma_x \quad CL = \mu_x \quad \text{and} \quad LCL = \mu_x - K\sigma_x$$

where K is the distance between the control limits and the central line, defined in standard deviation units. For the control statistic, sample mean, $\bar{x}$, the standard deviation of the sample mean depends on sample of size n: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. When the process is in-control, we know that with probability $(1 - \alpha)$, the sample mean $\bar{x}$ will lie between $\mu - Z_{\alpha/2}\sigma_{\bar{x}}$ and $\mu + Z_{\alpha/2}\sigma_{\bar{x}}$, where $Z_{\alpha/2}$ is the $1 - \alpha/2$-th quantile of the standard normal distribution. Thus, the constant K in the control limits is determined by α. These models are called *K-sigma control limits*. Determination of the control limits are equivalent to setting up

the critical region in a test of hypothesis problem, where α is the level of significance or the probability of type I error.

In Figure 1.3, we present an univariate Shewhart chart for simulated observations from univariate normal distribution with mean equal to 74 and standard deviation equal to 0.01 (Montgomery, 2009). The sample sizes are $n = 5$ and thus the standard deviation of the sample mean, $\bar{X}$, is $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.01/\sqrt{n} = 0.0045$. Therefore, when the process is in control with $\mu = 74$, we may assume that $\bar{X}$ is distributed approximately as normal. Consequently, the upper and lower control limits will be given by

$$UCL = 74 + 3(0.0045) = 74.0135 \quad \text{and} \quad LCL = 74 - 3(0.0045) = 73.9865.$$

In this example, we have used $K = 3$ to construct the 3-sigma control limits and that gives $\alpha = 0.0027$. Clearly, Figure 1.3 shows that all points fall within the control limits, which mean that the process is in control.

The choice of sample size is important for such control charts and the convergence to the normal distribution depends on the process. If we have m samples, each containing n observations for the characteristic, that means the sample averages are $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$. Therefore, we can estimate μ, from the process average $\bar{X} = (\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m)/m$. The advantage of the Shewhart's technique lies in its ability to separate between special causes of variation and common cause of variation. This lead to a possible diagnosis and correction of many production problems. It often brings substantial improvements in product quality and reduction of variation and rework Leavenworth and Grant (1976).

1.2.2 Cumulative Sum (CUSUM) Control Charts

The univariate cumulative sum control scheme (CUSUM) were first proposed by Page (1954), and have been studied by many authors including Ewan (1963); Page (1963); Johnson (1961); Brook and Evans (1972); Lucas (1985); Chatterjee and Qiu (2009); Yang
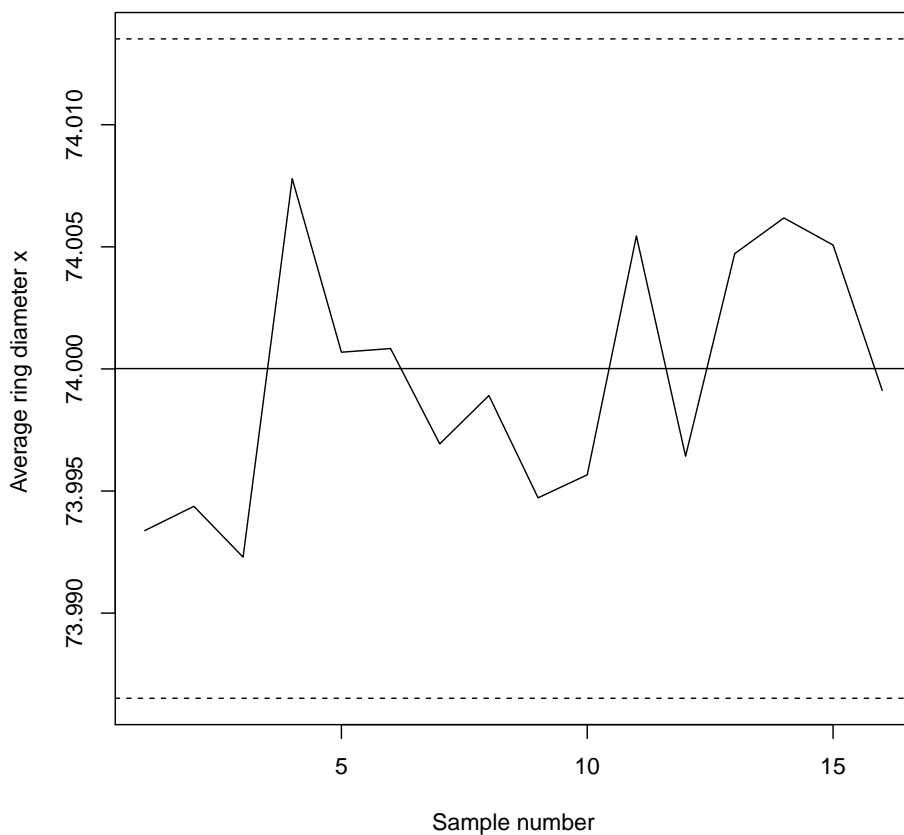
Figure 1.3: Univariate Shewhart control chart for a simulated data from univariate normal distribution.

and Cheng (2011). The biggest drawback of the Shewhart control chart is its inability to detect small shifts in the location parameter of the process characteristic. On the contrary, the cumultaivr sum control charts are quite sensitive to small shifts in location and detects them quite quickly. Another advantage with CUSUM charts are that they detect a systematic shift in the process mean over a period of time. There are two major disadvantages of the CUSUM control charts. Firstly, they are quite slow to discover large process shift. Secondly, they are not efficient to test historical data. The cumulative sum control chart directly incorporates current and previous information of sample values of

the process by plotting the cumulative sums of the variation of the sample values from a target value (Montgomery, 2009). For individual sample size $n \geq 1$, the cumulative sum control chart plots the points

$$S_m = \sum_{i=1}^{m} (\bar{x}_i - \mu_0), \tag{1.1}$$

against m where $\bar{x}_i$ represent the sample mean of the i-th sample, and μ_0 represents the target value for the process mean.

If the process remains in control at the target value μ_0, the cumulative sum S_m is a random walk with mean zero and if the mean shifts to some other value $\mu_1 > \mu_0$, a positive drift will increase the cumulative sum. If the mean shifts downward to some $\mu_2 < \mu_0$ then a negative drift in the cumulative sum will improve. If we observe change in the trend of the CUSUM values, it indicates that the process mean is drifting away from the target value (Oakland, 1996).

Runger and Testik (2004) proposed extensions to one-sided univariate CUSUM procedures based on Page (1954) using a sequential probability ratio test argument. Consider the testing of hypothesis problem $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$. Assuming that the data is coming from a normal distribution, the sequential probability ratio test at time m reject H_0 when the log-likelihood ratio is greater than a constant c_1, and fails to reject H_0 if it is less than a constant c_0. The log-likelihood ratio is given by

$$
\begin{aligned}
\ell &= \sum_{i=1}^{m} \log \frac{f(x_i, \mu_1)}{f(x_i, \mu_0)} = \sum_{i=1}^{m} \log \frac{f(x_i, \mu_0 + \delta)}{f(x_i, \mu_0)} = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu_0 - \delta)^2 - (x_i - \mu_0)^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{m} ((x_i - \mu_0) - \delta)^2 - (x_i - \mu_0)^2 = \frac{\delta}{\sigma} \sum_{i=1}^{m} \sigma^{-1}(x_i - \mu_0 - \frac{\delta}{2}).
\end{aligned}
$$

Then, the CUSUM procedure is to produce an out-of-control signal if

$$\frac{\delta}{\sigma} \sum_{i=1}^{m} \sigma^{-1}(x_i - \mu_0 - \frac{\delta}{2}) > c_1 \tag{1.2}$$

9

Then, an one-sided procedure will produce an out-of control signal if

$$S_m = max \frac{\delta}{\sigma}[S_{m-1} + \sigma^{-1}(x_i - \mu_0 - \frac{\delta}{2}), 0] > c \tag{1.3}$$

(for details see Runger and Testik (2004)). The two-sided procedure can be developed as to produce an out-of-control signal when

$$S_m^+ = max(S_{m-1}^+ + \sigma^{-1}(x_i - \mu_0 - \frac{\delta}{2}), 0) > H \tag{1.4}$$

or

$$S_m^- = max(S_{m-1}^- + \sigma^{-1}(\mu_0 - x_i - \frac{\delta}{2}), 0) > H \tag{1.5}$$

where $S_0^+ = S_0^- = 0$, and H is a suitably chosen cut-off. The value of H depends on δ and the k-sigma limit we would like to construct.
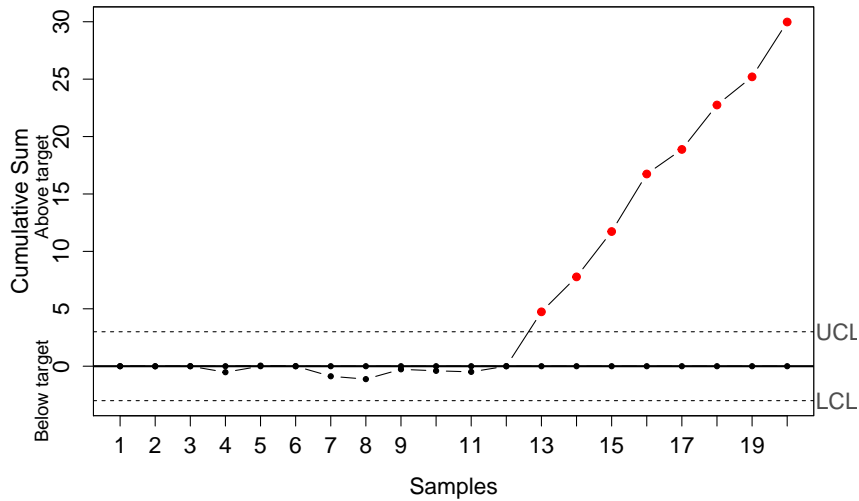
Figure 1.4: An example CUSUM control chart

For illustration we consider the following example with 20 data points

$$324.925, 324.675, 324.725, 324.350, 325.350, 325.225, 324.125, 324.525, 325.225, 324.600,$$

$$324.625, 325.150, 328.325, 327.250, 327.825, 328.500, 326.675, 327.775, 326.875, 328.350$$

which are each the average of samples of size 4 taken from a process that has an estimated mean of 325. Based on process data, the process standard deviation is 1.27 and therefore the sample means have a standard deviation of $1.27/\sqrt{4} = 0.635$. Figure 1.4 shows a CUSUM control chart for this data with $\delta = 1.0$ and a 3-sigma limit is considered. We observe that the last few points are beyond the control limits and an increasing trend in the chart indicating a systematic positive drift in the process.

1.2.3 EWMA control Charts

There is another approach to detect smaller shifts in the process variable mean remedying the problem of the Shewhart's chart. Consider samples with size n, we can define at the m-th stage a moving average of lag t as

$$S_m = (\bar{X}_m + \bar{X}_{m-1} + \cdots + \bar{X}_{m-t+1})/t. \tag{1.6}$$

This is an unbiased estimator of μ and the variance is given by

$$V(S_m) = \frac{1}{t^2} \sum_{i=m-t+1}^{m} V(\bar{X}_i) = \frac{\sigma^2}{nt} \tag{1.7}$$

which is smaller than the variance of $\bar{X}_m$ used in the Shewhart's chart. If we construct a control chart based on the moving averages S_m, the 3-sigma control limits will be given by

$$UCL = \mu + \frac{3\sigma}{\sqrt{nt}} \quad \text{and} \quad LCL = \mu - \frac{3\sigma}{\sqrt{nt}}. \tag{1.8}$$

Extending the idea, Roberts (1959) proposed an exponentially weighted moving average (EWMA)

$$Z_m = r\bar{X}_m + (1 - r)Z_{m-1} \tag{1.9}$$

where Z_m represents a weighted average of all previous sample means, $0 < r \leq 1$ is a tuning constant, and the starting value is $Z_0 = \mu$. As a result, the upper and lower control limits are

$$UCL = \mu + 3\sigma\sqrt{\frac{r}{(2 - r)n}} \quad \text{and} \quad LCL = \mu - 3\sigma\sqrt{\frac{r}{(2 - r)n}}. \tag{1.10}$$

The control chart using Z_m as the plotting statistic is known as the exponentially weighted moving average (EWMA) charts. The choice of the weighting factor r makes the chart sensitive to a small or gradual drift in the process. The value of r determines how the older data affects the control statistic. If $r = 1$, it reduces to the Shewhart's control chart and thus a large value of r gives more weight to the recent samples and a small value of r gives more weight to past samples. The value of r is usually taken to be in the interval $(0.2, 0.3)$ (Hunter, 1986). However, this choice is arbitrary and Lucas and Saccucci (1990) provided some tables to help users to choose the value of r.

Consider the following data consisting of 20 points.

$$52.0, 47.0, 53.0, 49.3, 50.1, 47.0, 51.0, 50.1, 51.2, 50.5,$$

$$49.6, 47.6, 49.9, 51.3, 47.8, 51.2, 52.6, 52.4, 53.6, 52.1$$

Here the target mean is 50 and the process standard deviation from the historical data is 2.0539. We plot an EWMA control chart for this data in Figure 1.5 with $r = 0.2$. The crosses in the plot indicates the raw data. We observe that the process is in control whereas raw data points are outside the control limits.
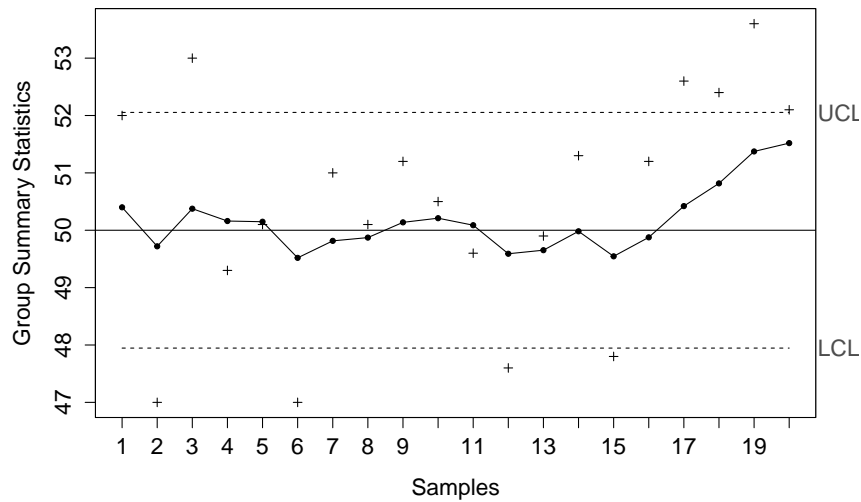
Figure 1.5: An example EWMA control chart

1.2.4 Nonparametric Control Charts

All previous control charting schemes often require the assumption of normality of the data, which may be inflexible to validate in application. Such charts are optimal and most efficient when the underlying distribution is normal. However, it may not be easy to detect a change in the quality of a manufactured product if the underlying distribution deviates slightly from normal distribution. The performance of such charts are affected significantly. To overcome this problem, many people tried to construct control charts that do not require any specific parametric distribution assumption including Liu (1995); Chakraborti and van de Wiel (2008); Bakir (2012); Ross and Adams (2012); Zou and Tsung (2011). These control chart are generally known as nonparametric or distribution free control charts.

The nonparametric methods have a few assumptions, and does not require large samples. For this reason the nonparametric methods have great utility in quality control, because in most cases there would be a lack of information and the underlying process

13

distribution is not known, as well as data measured on a non normal distribution.

Chakraborti and van de Wiel (2008) proposed a phase II Shewhart type control chart based on reference data from phase I and the well known Mann-Whitney statistic. Let $X = (X_1, ..., X_m)$ be the historical reference data obtained from an in-control process. Let $Y = (Y_1, ..., Y_n)$ denotes an arbitrary test sample of size n. To test whether the distribution of the test data Y is stochastically greater than the reference in-control data X, we can write the Mann-Whitney statistic as

$$M_{xy} = \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j) \tag{1.11}$$

where $I(X_i < Y_j)$ is the indicator function for $\{X_i < Y_j\}$. The large values of the statistic M_{xy} indicates a positive drift in the process. In contrast, the small values indicate a negative drift. A control chart based on this Mann-Whitney statistic produces an out-of-control signal for the h-th test sample if

$$M_{XY}^h < L_{mn} \quad \text{or} \quad M_{XY}^h > U_{mn} \tag{1.12}$$

where M_{XY}^h is the statistic for the h-th sample and L_{mn} and U_{mn} are the lower and upper control limits, respectively, obtained from the distribution of Mann-Whitney test statistic.

This is a Shewhart type control chart as the control statistic computed at the h-th sample depends on the h-th sample only. Bakir (2012) proposed a similar Shewhart type control chart based on Kolmogorov-Smirnov statistic as well.

Ross and Adams (2012) presented two nonparametric control charts based on Kolmogorov-Smirnov and Cramer Von Mises test statistics. These can detect any changes in the process distribution during Phase II monitoring. The proposed charts are based on adapting the change-point model. The main idea is to detect a change point in a fixed number of observations. It divides the observations into two parts by using the test hypothesis for

continuous distributions. The null hypothesis is that there is no change point and all observations are having the same distribution. The alternative hypothesis is, that there exists a single change point in the second part of the observations. We can write theses hypothesis as

$$H_0 : X_i \sim F_0 \quad \text{for} \quad i = 1, \dots, t \quad \text{against} \quad H_1 : X_1, \dots, X_k \sim F_0, \quad X_{k+1}, \dots, X_t \sim F_1. \tag{1.13}$$

The first of the two proposed charts uses the Cramer-von-Mises statistic and the second one is based on the Kolmogorov-Smirnov statistic. Both of them depend on the empirical distribution functions of the observation before and after the change point. Define $S_1 = \{X_1, \dots, X_k\}$ and $S_2 = \{X_{k+1}, \dots, X_t\}$ and their empirical distribution functions as

$$\hat{F}_{S_1}(x) = \frac{1}{k} \sum_{i=1}^{k} I(X_i \leq x) \tag{1.14}$$

and

$$\hat{F}_{S_2}(x) = \frac{1}{t-k} \sum_{i=k+1}^{t} I(X_i \leq x) \tag{1.15}$$

Now we can define the Kolmogorov-Smirnov statistic as

$$D_{k,t} = \sup_{x} |\hat{F}_{S_1}(x) - \hat{F}_{S_2}(x)|. \tag{1.16}$$

We reject H_0 when $D_{k,t} > h_{k,t}$, which means that there is no change in the distribution function.

We can also define the Cramer-Von-Mises statistic as

$$W_{k,t} = \sum_{i=1}^{t} |\hat{F}_{S_1}(X_i) - \hat{F}_{S_2}(X_i)|^2. \tag{1.17}$$

15

We reject the null hypothesis if $W_{k,t} > h_{k,t}$ for a suitable cut-off $h_{k,t}$. Observe that

$$E(W_{k,t}) = \frac{t+1}{6t}$$

and

$$Var(W_{k,t}) = \frac{(t+1)[(1-3/4k)t^2 + (1-k)t - k]}{45t^2(t-k)}$$

Then we can define a standardized statistic as

$$W_t = \max_k \frac{W_{k,t} - E(W_{k,t})}{\sqrt{Var(W_{k,t})}}, \quad 1 < k < t. \tag{1.18}$$

Also, if $W_t > h_t$ for a suitable h_t, we reject the null hypothesis and conclude that some point of data had a change point. In addition, if $W_t < h_t$, in this case we do not reject the null hypothesis, which means, no change occurred. The value of h_t will depend on the choice of process design parameters.

1.3   Multivariate Control Charts

With the advancement of technology, it became easier to record several characteristics or variables of a process and it is becoming increasingly important to monitor several process variables together to control the quality characteristic of a process. The applications of statistical process control methods moved beyond the manufacturing into engineering, medicine, epidemiology, environmental science, biology, genetics, finance and even law enforcements and athletics (Stoumbos et al., 2000). Although one can use separate univariate control charts to monitor the process characteristics individually, they do not capture the dependence structure of the variables among themselves. At statistical procedure which captures this dependence structure is likely to be more sensitive to the deviation of the process from the target parameters. Hawkins (1991) suggested

using a single control chart takes care of the correlations among the process variables. The first attempt to construct a truly multivariate control chart is by Hotelling (1947), who constructed a Shewhart type control chart based on multivariate T^2 statistic with applications in air samples from the bomb sights. Alt (1985) discussed the application of Hotelling's T^2 in control charts in detail and the survey article by Jackson (1985)) mentioned several modifications of the distribution theory of this control statistic for Phase II control charts. Alwan (1986) proposed a CUSUM control chart based on the T^2 statistic.

Crosier (1988) suggested a CUSUM of T (COT) instead of T^2. He also proposed another multivariate CUSUM chart based on the idea of the T^2 statistic for Phase I control charts suing the observations directly to construct the CUSUM statistic instead of T^2. We will review these proposed methods later in detail. In a different approach, Woodall and Ncube (1985) proposed a multivariate CUSUM control chart based on individual CUSUM control chart for the process variables. It produces a out-of-control signal whenever one of the univariate charts are out-of-control. The biggest advantage of this chart is its interpretability and one can identify the variable causing the out of control signal quite easily. However, it suffers from the drawback that it does not take into account the correlation structure of the variables and perform poorly when there is high correlation among the process variables. Although they discussed their control chart in the context of location shift only, it can be extended easily to scale shifts as discussed by Hawkins (1981). Exponentially weighted moving average control charts were extended to the multivariate case by Lowry et al. (1992) for phase I and phase II process control. Recently, Zhang et al. (2010) proposed a multivariate control chart extending the exponentially weighted moving average with the likelihood ratios.

All of the above multivariate control chart procedures are heavily dependent on the assumption of multivariate normal distribution for the underlying distribution of the process variable. However, the performance of this charts suffers heavily if the distribution shifts

a little from normality. Liu (1995) proposed a few distribution free control charts based on the notions of multivariate data depths. While the proposals are quite attractive, they suffer heavily due to computational complexities of the depth functions beyond dimension 2 and that makes the proposed scheme a little impractical. Liu and Tang (1996) discussed the use of simulation and bootstrap methods in determining the process control limits. Abu-Shawiesh and Abdullah (2001) introduced a new robust Shewhart-type control chart for monitoring the location of a bivariate process using Hodges–Lehmann estimators.

Qiu and Hawkins (2001, 2003) introduced a nonparametric CUSUM procedure for multivariate process control based on the ranks within the measurement components and also on the ranking between the measurement components and their in-control location parameter. Thissen et al. (2005) used mixture modeling for process monitoring in case of non-normality. Chakraborti et al. (2001) noted that, although multivariate process control problems are important, multivariate non-parametric statistical process techniques are not sufficiently well developed. Bersimis et al. (2007) presented a recent literature review on multivariate process control techniques.

In this PhD thesis, our objective is to fill the void in the literature by constructing a fee multivariate control chart procedures using notions of multivariate sign and rank functions. While the proposed methods do not depend on the assumption of multivariate normality, they are also computationally simple, making the proposed schemes attractive. We discuss the performance of the proposed methods using average run length for in-control and out-of-control processes.

1.4   Outline of The Thesis

In Chapter 2, we present a detailed review of Shewhart type multivariate control charts based on Hotelling's T^2 statistic, multivariate CUSUM control charts as proposed by Crosier (1988) and distribution free control charts proposed by Liu (1995). Apart from

some theoretical details of these charts, we present performance studies with simulations.

In Chapter 3, we use the notion of spatial signs and and spatial ranks to extend the multivariate control charts proposed by Liu (1995). We present theoretical results to determine the control limits and perform a finite sample simulation to study the average run lengths of the proposed schemes.

In Chapter 4, we proposed Shewhart type multivariate control charts based on spatial sign and signed rank statistic. It is well known that spatial sign statistic is not invariant under affine transformations of the data and that makes the proposed schemes to perform poorly in the presence of high correlations among the process variables. We briefly discuss the issue of affine invariance and proposed some affine invariant modifications of those control charts. Again we present the performance of the proposed control charts through finite sample simulations.

In Chapter 5, some extensions of multivariate CUSUM control charts are proposed using multivariate sign and signed rank statistics. We derived the control limits of the proposed charts using simulations. The average run lengths of the proposed schemes under different distributional assumptions are studied using simulated data.

In Chapter 6, we propose a multivariate extension of the exponentially weighted moving average (EWMA) control charts based on some multivariate notions of sign vector and signed rank vectors and in Chapter 7 we present some concluding remarks and directions for further research.

CHAPTER 2

REVIEW OF SOME MULTIVARIATE

CONTROL CHARTS

2.1 Hotelling's T^2 Chart

Assume that, the vector of process variables, $\boldsymbol{X}$ follows a d-dimensional normal distribution with mean vector $\boldsymbol{\mu}_0$ and the covariance matrix Σ, when the process is in-control and when the process is out-of-control the mean vector is given by $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. We assume for the time being that there is no change in the covariance matrix Σ. Hotelling (1947) proposed the first multivariate control chart based on T^2 statistic. An observation $\boldsymbol{X}_n$, at the n-th stage, is said to be out of control if

$$T_n = \sqrt{(\boldsymbol{X}_n - \boldsymbol{\mu}_0)^\top \Sigma^{-1} (\boldsymbol{X}_n - \boldsymbol{\mu}_0)} > SCL$$

where SCL is the control limit.

It is easy to observe that when the process is in-control, the statistic T_n^2 has a χ^2 distribution with d degrees of freedom, and when the process is out-of-control, its distribution

is a non-central χ^2 with d degrees of freedom and non-centrality parameter

$$\delta = \sqrt{(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)}.$$

We can decide on the control limit SCL by fixing our in-control average run length (ARL) of the chart, that is, the expected number of sample required to produce an out-of-control signal, when the process is actually in-control. This is also known as signal for the *false alarm*. We would like the in-control ARL to be as large as possible and out-of-control ARL to be as low as possible. But similar to the hypothesis testing problems with Type I and Type II errors, we cannot achieve both the goals togethers. If we take SCL^2 to be $\chi^2_{d,1-\alpha}$, the $100(1-\alpha)$-th percentile of the χ^2 distribution with d degrees of freedom, then the in-control ARL would be $1/\alpha$.

In Figure 2.1 we present the ARL curves for the multivariate T^2-charts as a function of δ for $d = 2, 5, 10$ and 20, with ARL in the vertical axes and non-centrality parameter in the horizontal axis. To obtain the ARL for $\delta > 0$, we simulate observations from d-variate normal distribution with mean vector $\boldsymbol{\mu}$ having the first element as δ and all other elements as 0. The covariance matrix Σ is taken to be the d-dimensional identity matrix $\mathbf{I}_d$. We report the average run lengths in 1000 simulations before an out of control signal is produced. The in-control ARLs in 2.1(a) and (b) are 200 and 500 respectively. Thus, the SCLs are the square roots of the 99.5 and 99.8 percentiles of the χ^2 distribution with d degrees of freedom. In both of these cases it can be seen that these charts are quite good in detecting large shifts in the location vector $\boldsymbol{\mu}$. However, for smaller shifts, even with $\delta = 1$, the average run lengths are quite high.

One practical problem with multivariate T^2 statistic based control charts is their lack of robustness. If the underlying distribution deviates slightly from the normal distribution, the performance of these charts are affected significantly. For instance, we can observe

Figure 2.1: ARL curves for multivariate T^2-charts when the underlying distribution is multivariate normal for in-control ARL (a) 200 and (b) 500 respectively for dimension $d = 2$, 5, 10 and 20 and non-centrality parameter δ.

their behaviour for multivariate Laplace and t distribution, where they are defined as follows:

Multivariate Laplace Distribution: The probability density function is given by

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto \frac{1}{|\Sigma|^{1/2}} \exp\{-\sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}\}, \tag{2.1}$$

Multivariate t Distribution: The probability density function is given by,

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2}) \nu^{d/2} \pi^{d/2} |\Sigma|^{1/2} [1 + \frac{1}{d}(\boldsymbol{x} - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]^{(\nu+d)/2}} \tag{2.2}$$

where $\boldsymbol{\mu}$ is a d vector, and Σ is a $d \times d$ scale matrix, and ν is the degree of freedom,

Figure 2.2: ARL curves for multivariate T^2-charts when the underlying distribution is multivariate Laplace with the same upper control limits as in Figure 2.1(a) and (b) respectively for dimension d and non-centrality parameter δ.

$\boldsymbol{x} \in \mathbb{R}^d$.

To illustrate, we present ARL curves of the multivariate T^2-charts when the underlying distribution is multivariate Laplace in Figure 2.2 and multivariate t distribution with 3 degrees of freedom in Figure 2.3 with location vector $\boldsymbol{\mu}$ and scale matrix Σ as described in the normal case. It can be observed that in both of these cases they fail to maintain the in-control ARL as it is much smaller than $1/\alpha$ with the same control limit and the false alarm rate will be quite high.

If at every stage of sampling, if we select a sample of size n, instead of a single

23

Figure 2.3: ARL curves for multivariate T^2-charts when the underlying distribution is multivariate Student's t distribution with 3 degrees of freedom with the same upper control limits as in Figure 2.1(a) and (b) respectively for dimension d and non-centrality parameter δ.

observation $\boldsymbol{X}_m$, then the T^2 statistic changes to

$$T_m^2 = m(\bar{\boldsymbol{X}}_m - \boldsymbol{\mu}_0)^{\top} \Sigma^{-1} (\bar{\boldsymbol{X}}_m - \boldsymbol{\mu}_0)$$

where $\bar{\boldsymbol{X}}_m$ is the sample mean vector of the m-th sample. Since the distribution of T_m^2 is χ^2 with d degrees of freedom when the process is in control, we can use the same control limit. Even if the underlying distribution is not multivariate normal, by multivariate central limit theorem , the distribution of T_m^2 will be approximately χ^2 distribution with d degrees of freedom for large values of m and the control charts perform better.

In Phase II control charts, let $\bar{\boldsymbol{X}}_m$ and $\mathbf{S}_m$ represents the sample mean vector and

sample covariance matrix of the m-th sample of size m, and also let $\bar{\bar{\boldsymbol{X}}}$ and $\bar{\mathbf{S}}$ denotes the average of $\bar{\boldsymbol{X}}_m$ and $\mathbf{S}_m$ respectively, then we define

$$T^2 = m(\bar{\boldsymbol{X}}_m - \bar{\bar{\boldsymbol{X}}})^\top \bar{\mathbf{S}}^{-1}(\bar{\boldsymbol{X}}_m - \bar{\bar{\boldsymbol{X}}})$$

which has a $d(m+1)(n+1)/(mn-m-d+1)F_{d,mn-m-d+1}$ distribution if the process is in-control. The control limit SCL is calculated accordingly.

Note that, this multivariate control chart does not have a lower control limit as $T^2 \geq 0$ and values close to 0 show conformity with the process specification.

2.2 Multivariate CUSUM Charts

In the context of univariate process control, Tukey (1986) commented that detecting a shift of 5 standard deviations is nearly trivial, whereas detecting a shift of 0.05 standard deviations is nearly impossible. Cumulative sum (CUSUM) control charts are developed for the univariate schemes to detect smaller shifts and the CUSUM charts designed to detect 1 standard deviation shift are widely used in practice. To extend the univariate CUSUM procedures, Woodall and Ncube (1985) proposed using d univariate CUSUM charts simultaneously and evaluate their performance together. However this proposal ignores the dependence structure of the process variables and the average run length of the control scheme heavily depends on the direction of shift if the process variables do not have identical standard deviations. While using principal components instead of original variables resolves the dependence structure problem, it alleviates the dependence of ARL on the direction of the shift. So there is a need to consider some CUSUM schemes which are truly multivariate in nature.

Crosier (1988) proposed the most immediate way of extending a multivariate Shewhart type T^2-chart to a CUSUM procedure, which is to form a cumulative sum of the statistics

for T_n, $n \geq 1$. We will refer this as *cumulative sum of T* (COT) chart. Define $S_0 = 0$ and

$$S_n = \max\{0, S_{n-1} + T_n - k\} \tag{2.3}$$

for some $k \geq 0$ where T_m^2 is the Hotelling T^2 statistic for the n-th sample. It will produce an out of control signal whenever $S_n > h$, for some control limit h. The constants k and h are determined to achieve a fixed in-control ARL and to make the chart most efficient to detect a specified shift in δ, the non-centrality parameter defined in the previous section. We can define h as the decision interval or the specific control limits and k as the reference number. Crosier (1988) discussed in detail on the choice of k and h for different values of δ and also discussed about setting some different initial value to S_0 from 0.
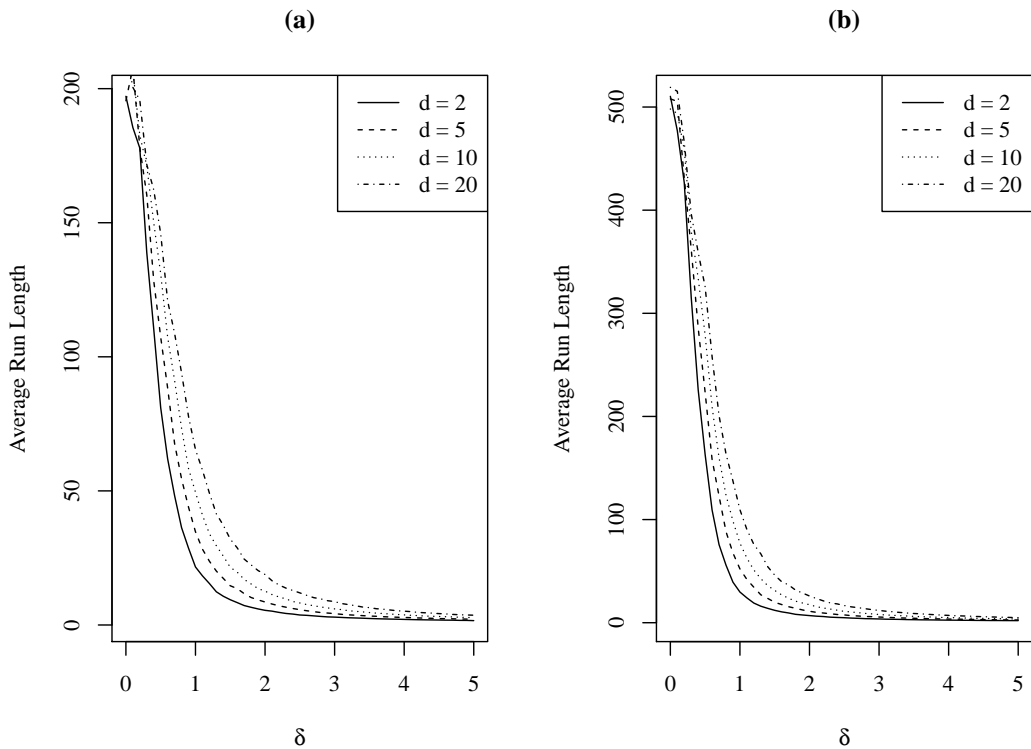
Figure 2.4: ARL curves for multivariate COT control charts when the underlying distribution is multivariate normal for in-control ARL (a) 200 and (b) 500 respectively for dimension d and non-centrality parameter δ.

In Figure 2.4, we present the ARL curves when the underlying distribution is multivariate normal for $d = 2, 5, 10$ and 20. Here the values of h and k are as given in Crosier (1988) to detect any shift in the mean vector producing $\delta = 1$ and in-control ARL of 200 and 500 respectively. The ARL decreases when delta increases. These curves indicate that COT control charts produce better detection of small shifts compared to multivariate Shewhart-type T^2-charts.

(a) **(b)**
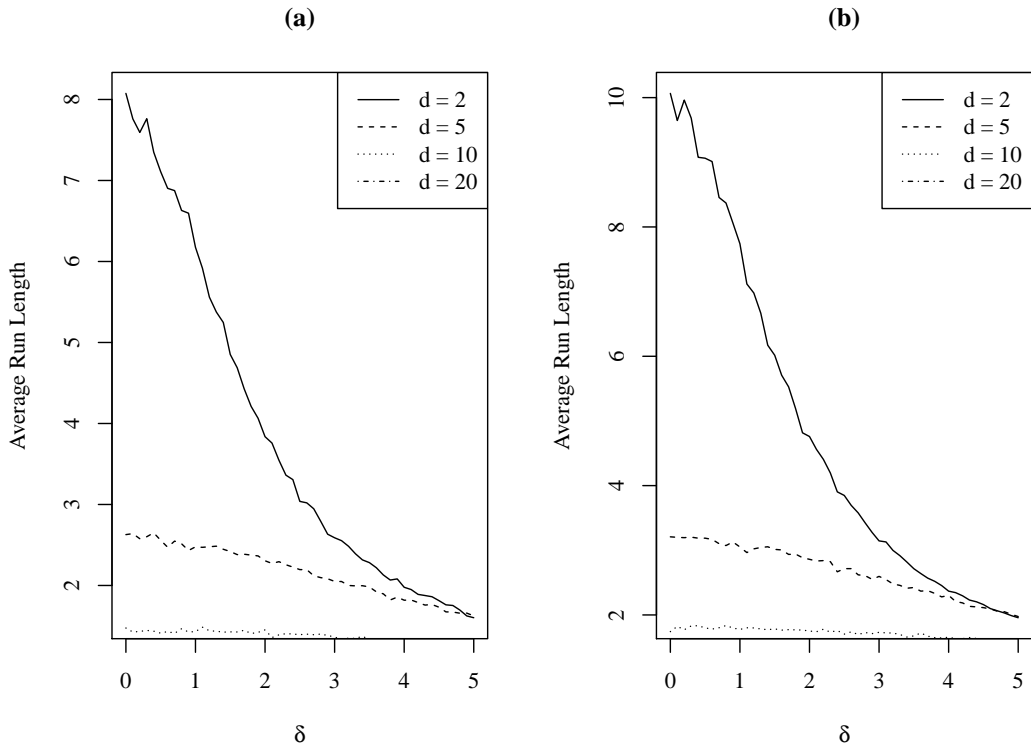
Figure 2.5: ARL curves for multivariate COT control charts when the underlying distribution is multivariate Laplace with the same upper control limit as in Figure 2.4(a) and (b) respectively for dimension d and non-centrality parameter δ.

The ARL curves for the multivariate COT scheme are presented in Figures 2.5 and 2.6, when the underlying distributions are multivariate Laplace and Student's t distribution with 3 degrees of freedom, respectively. The same phenomenon is observed, as before, that they fail to even attain the intended in-control average run lengths and thus producing a

(a)   **(b)**
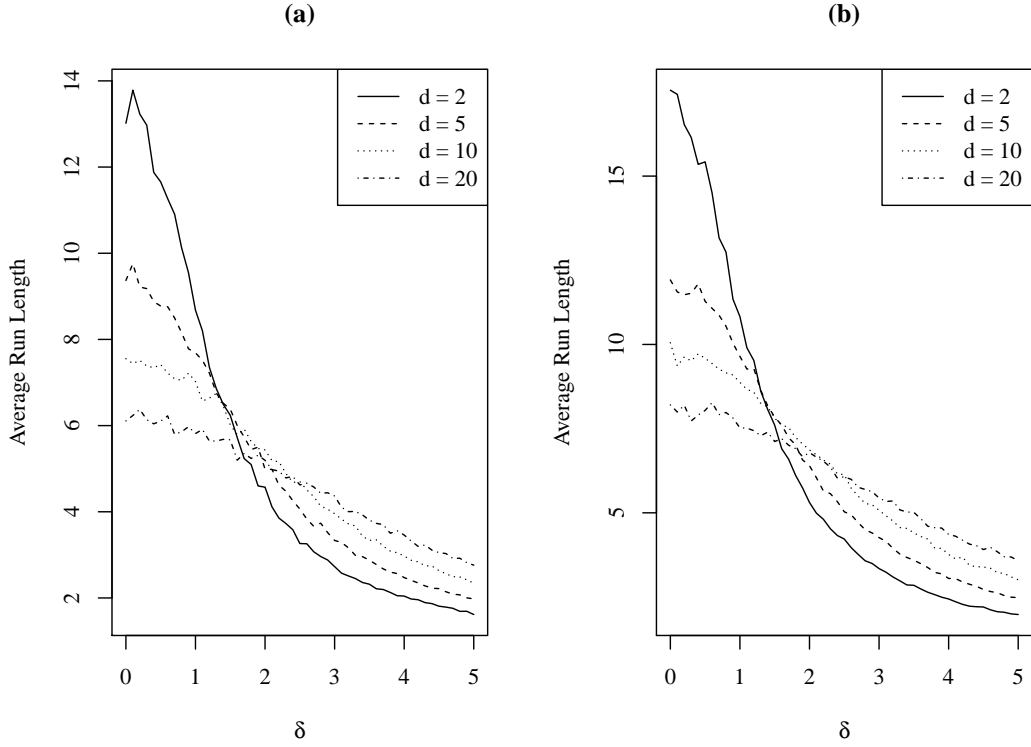
Figure 2.6: ARL curves for multivariate COT control charts when the underlying distribution is multivariate Student's t distribution with 3 degrees of freedom with the same upper control limit as in Figure 2.4(a) and (b) respectively for dimension d and non-centrality parameter δ.

large false alarm rates.

Instead of working with a CUSUM of T-statistics, Crosier (1988) proposed a fully vector-valued multivariate CUSUM scheme. Define

$$C_n = [(\boldsymbol{S}_{n-1} + \boldsymbol{X}_n - \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\boldsymbol{S}_{n-1} + \boldsymbol{X}_n - \boldsymbol{\mu}_0)]^{1/2},$$

and

$$\boldsymbol{S}_n = \begin{cases} 0 & \text{if } C_n \leq k \\ (\boldsymbol{S}_{n-1} + \boldsymbol{X}_n - \boldsymbol{\mu}_0)(1 - k/C_n) & \text{if } C_n > k, \end{cases} \quad (2.4)$$

28

where $\boldsymbol{S}_0 = \boldsymbol{0}$ and $k > 0$. Let

$$Y_n = \{\boldsymbol{S}_n^\top \Sigma^{-1} \boldsymbol{S}_n\}^{1/2}. \tag{2.5}$$

An out of control signal is produced when $Y_n > h$, for some control limit h. We have seen earlier that the average run lengths of the multivariate T^2-chart and COT chart depend on the location vector $\boldsymbol{\mu}$ and the scale matrix Σ only through the non-centrality parameter δ as they are solely based on the Hotelling's T^2 statistics. However, it is not obvious that the distribution of Y_n defined above depends on μ and Σ only through δ. Healy (1987) provided a detailed proof on this dependence.
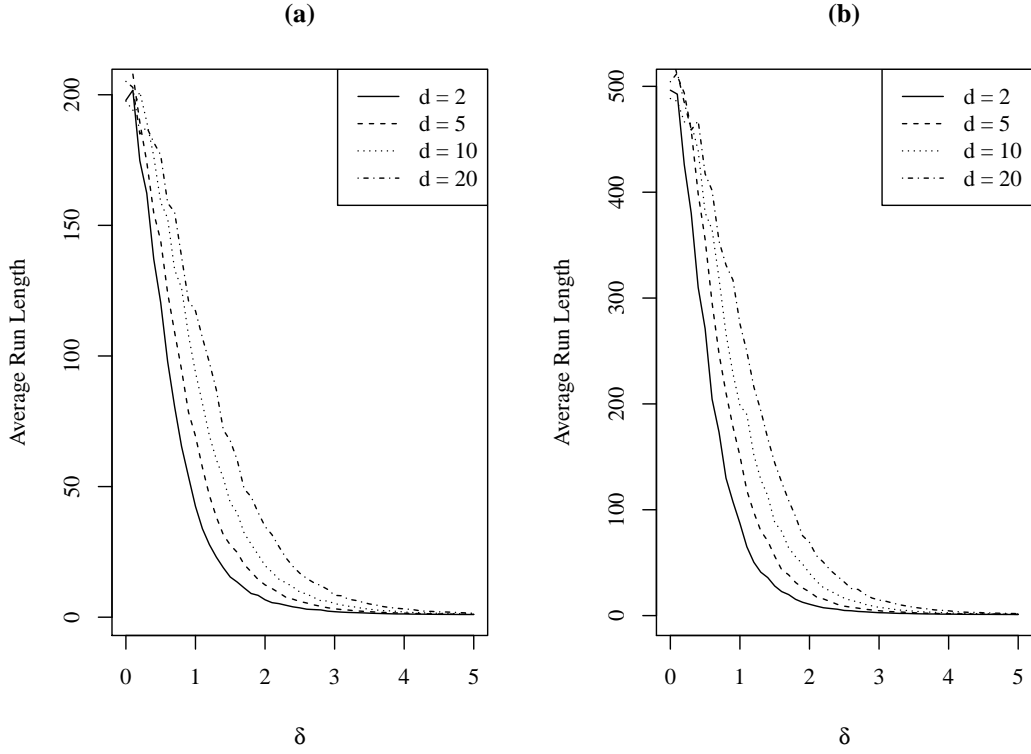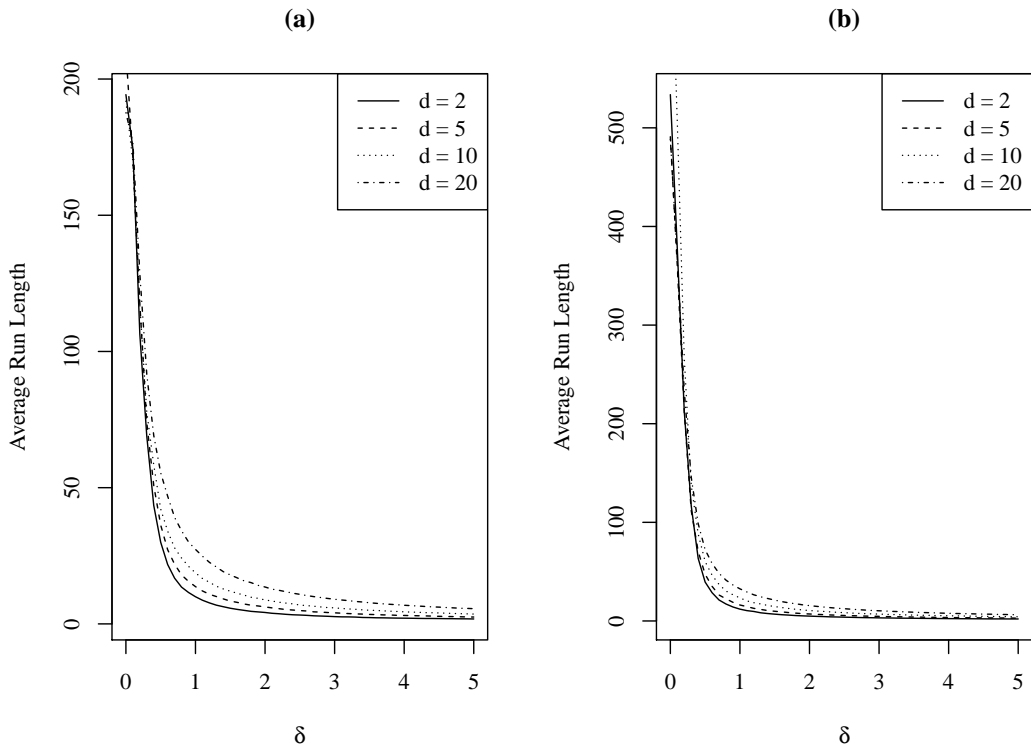

Figure 2.7: ARL curves for multivariate CUSUM control charts when the underlying distribution is multivariate normal for in-control ARL (a) 200 and (b) 500 respectively for dimension d and non-centrality parameter δ.

In Figure 2.7, we present the ARL curves for the proposed multivariate CUSUM charts

for underlying normally distributed process variables. They indicate a quicker detection of smaller shifts compared to both multivariate T^2-charts and COT charts. Determination of the constants h and k is straightforward. To detect a shift of the noncentrality parameter δ, one can choose $k = \delta/2$. In our examples, we design the charts to detect a shift of $\delta = 1$ most efficiently and hence we choose $k = 0.5$. The values of h to achieve in-control ARLs of 200 and 500 are given in Crosier (1988) for this particular choice of k.
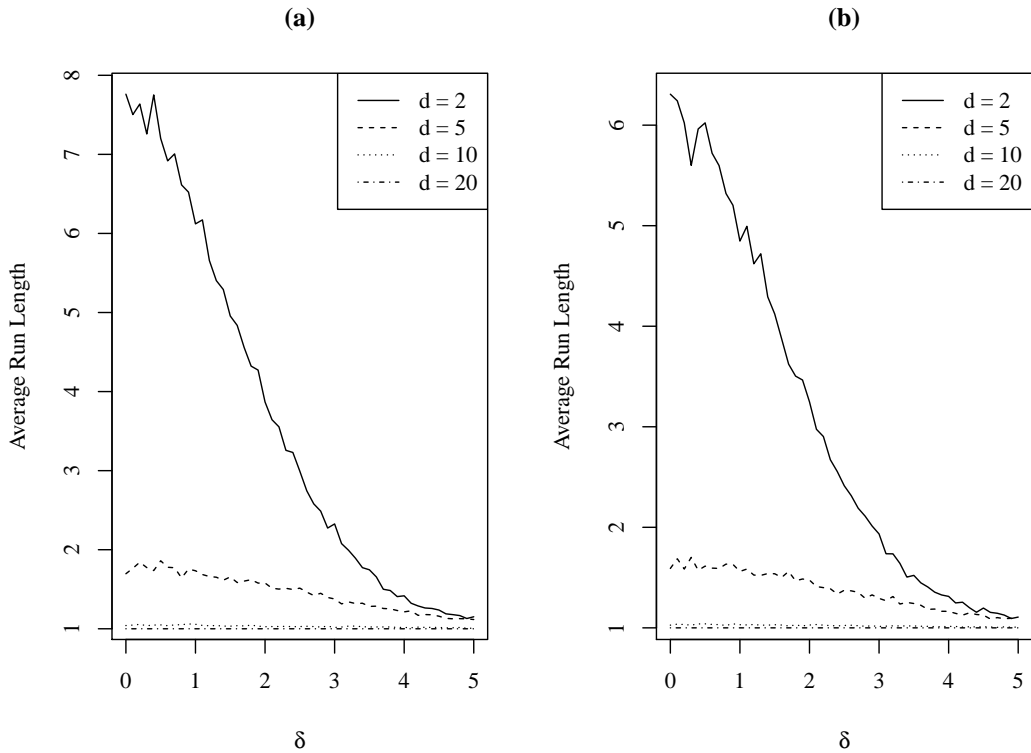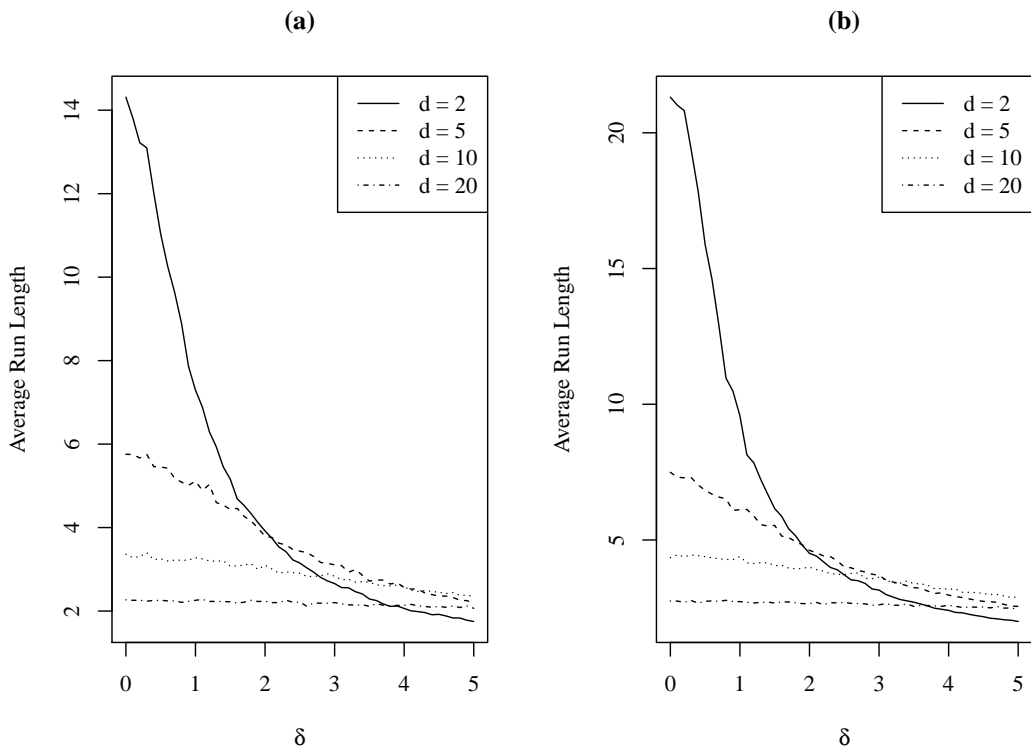

Figure 2.8: ARL curves for multivariate CUSUM control charts when the underlying distribution is multivariate Laplace with the same upper control limit as in Figure 2.7(a) and (b) respectively for dimension d and non-centrality parameter δ.

In Figures 2.8 and 2.9, we present the ARL curves for the above CUSUM control charting scheme when the underlying distributions are multivariate Laplace and Student's t distribution with 3 degrees of freedom. One can observe the same phenomenon as before that they fail to even attain the intended in-control average run lengths and thus producing
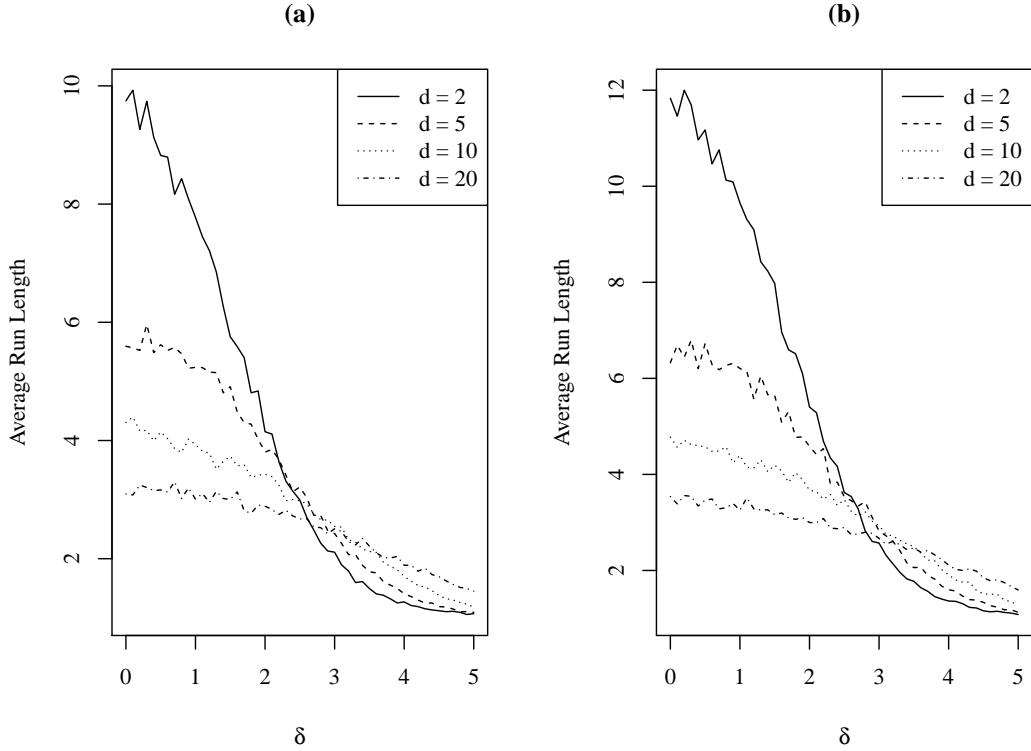
30
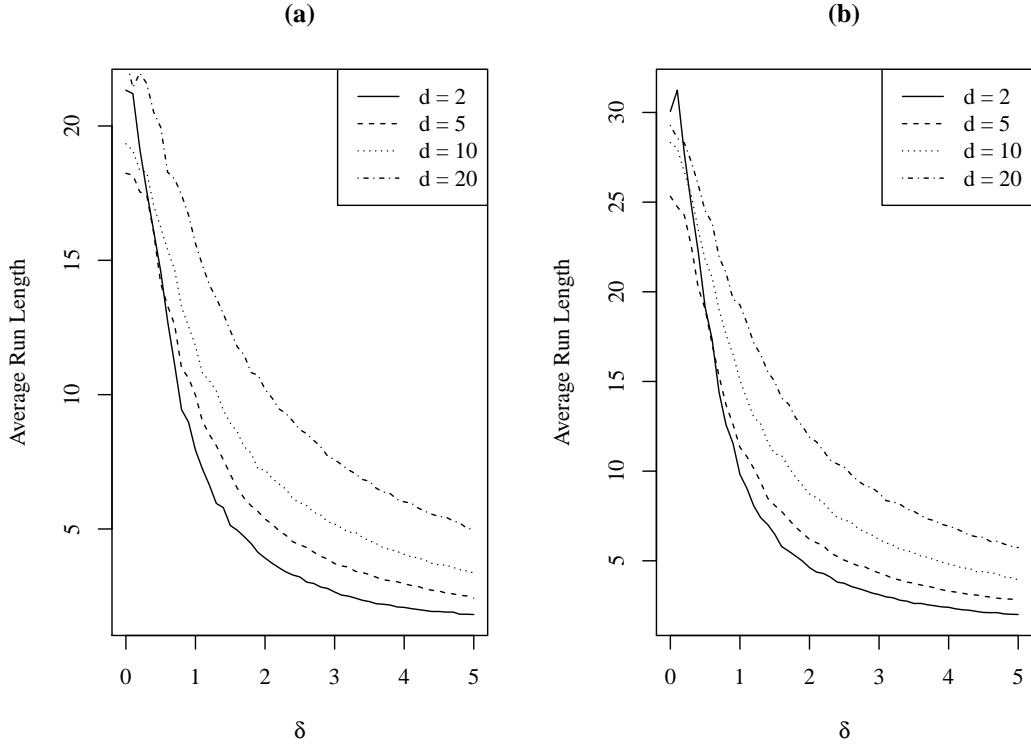
Figure 2.9: ARL curves for multivariate CUSUM control charts when the underlying distribution is multivariate Student's t distribution with 3 degrees of freedom with the same upper control limit as in Figure 2.7(a) and (b) respectively for dimension d and non-centrality parameter δ.

a large false alarm rates. This shows that the design of these schemes depend heavily on the assumption of multivariate normality of the underlying process variables.

In a different development, Hawkins (1991) tried to find test statistics that are more powerful than T^2 for non-normal distributions and have better interpretation. This procedure depends on the residuals from the regression equation of each variable on all others. Suppose $\boldsymbol{X} = (X_1, ..., X_d)^\top \sim N_d(\boldsymbol{\mu}, \Sigma)$, and the process is in control when $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\Sigma = \Sigma_0$. Let Z_j represent the residual obtained from the linear regression of the j-th variable, X_j on other variables $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d$. Suppose, this scaler Z_j has been rescaled to have unit variance. Then the regression residual vector is given by $\boldsymbol{Z} = \mathbf{A}(\boldsymbol{X} - \boldsymbol{\mu}_0) = [\operatorname{diag}(\Sigma_0^{-1})]^{-1/2}\Sigma_0^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_0)$, where $\mathbf{A} = [\operatorname{diag}(\Sigma_0^{-1})]^{-1/2}\Sigma_0^{-1}$. Us-

ing these residuals, he proposed two overall group diagnostics to detect out-of-control processes. The first proposal is :

$$MCZ = \max(\max(L_{ni}^+, -L_{ni}^-))$$

where $L_{ni}^+ = \max(0, L_{n-1,i}^+ + Z_{ni} - K)$, $L_{ni}^- = \min(0, L_{n-1,i}^- + Z_{ni} + K)$, and $L_{0,i}^+ = L_{0,i}^-$, for $i = 1, \ldots, d$, Z_{ni} is the i-th component of $\boldsymbol{Z}_n$, the residual vector at the n-th stage and k is a constant.

The second proposal is:

$$ZNO = \sum_{i=1}^{d} (L_{ni}^+ + L_{ni}^-)^2.$$

The form of ZNO represents the Euclidean norm of the resultant vector of the CUSUM for upward and for downward shift in mean. If the CUSUMs exceed the decision interval h, it produces an out of control signal. The value of h is found from univariate CUSUM chart because it is dependent on a signal random variable that follows the standard normal distribution.

2.3 Control Charts Based on Data Depth

Over the last two decades, the data depth functions have become useful tools to construct non-parametric methods for multivariate data. Zuo and Serfling (2000) discussed several notions of data depth, which can be used to construct outward ordering of the sample points. Liu (1995) proposed a few control charts based on these notions of data depths, which do not depend on the assumption of multivariate normality for the underlying process variables. There are several types of data depth functions. For example, Mahalanobis's depth, Tukey depth, Simplicial depth and Majority depth, which are defined as follows:

Mahalanobis depth (MD_F): Mahalanobis (1936) proposed the Mahalanobis distance

between a vector $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ and the mean vector $\boldsymbol{\mu}_F = (\mu_1, \ldots, \mu_d)^\top$ as

$$d^2(\boldsymbol{x}, \boldsymbol{\mu}_F) = (\boldsymbol{x} - \boldsymbol{\mu}_F)^\top \Sigma_F^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_F), \tag{2.6}$$

where Σ_F^{-1} is the covariance matrix. Liu and Singh (1993) defined Mahalanobis depth as

$$MD_F(\boldsymbol{x}) = [1 + d^2(\boldsymbol{x}, \boldsymbol{\mu}_F)]^{-1}. \tag{2.7}$$

The sample version of MD_F is obtained by replacing $\boldsymbol{\mu}_F$ and Σ_F by their estimates, the sample mean $\bar{\boldsymbol{X}}$ and sample covariance matrix $\mathbf{S}$

$$MD_F(\boldsymbol{x}) = [1 + (\boldsymbol{x} - \bar{\boldsymbol{X}})^\top \mathbf{S}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{X}})]^{-1}. \tag{2.8}$$

Simplicial depth (SD): (Liu, 1988) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}$ be identical and independently distributed with distribution function F. The Simplicial depth at point $\boldsymbol{x} \in \mathbb{R}^d$ is defined as

$$SD(F; \boldsymbol{x}) = P_F(\boldsymbol{x} \in S\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}\}) \tag{2.9}$$

where $S\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}\}$ is a d-dimensional simplex. The sample version is defined as

$$SD(F_n; \boldsymbol{x}) = \binom{n}{d+1} \sum_{1 \le i_1 \le \cdots \le i_{d+1} \le n} I(\boldsymbol{x} \in S\{\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d+1}}\}) \tag{2.10}$$

where

$$I(\boldsymbol{x} \in S\{\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d+1}}\}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in S\{\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_{d+1}}\} \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a random sample from F.

Majority depth ($M_j D$): (Liu and Singh, 1993) Suppose $\boldsymbol{X}_1, \ldots, X_d$ is a random sample

from F, and $H(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ denotes a unique hyperplane passing through these points. A point $\boldsymbol{x} \in \mathbb{R}^d$ is in the major side if the half-space containing $\boldsymbol{x}$ has probability greater than or equal to $1/2$, and a depth function can be defined as

$$M_j D(F, \boldsymbol{x}) = P_F\{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d) : \boldsymbol{x} \text{ is in a major side}\}. \tag{2.12}$$

The sample version of $M_j D(F, \boldsymbol{x})$ is $M_j D(F_n, \boldsymbol{x})$ by replacing F by F_n. For $d = 1$, we have

$$M_j D(F, x) = \frac{1}{2} + \min\{F(x), 1 - F(x-)\}.$$

Tukey's depth (TD): (Tukey, 1975) The Tukey's depth or the *halfspace depth* of a point $\boldsymbol{x} \in \mathbb{R}^d$ is defined as

$$TD(F; \boldsymbol{x}) = \inf_{\psi} F(\psi) \tag{2.13}$$

where ψ is a closed half space containing $\boldsymbol{x}$. The sample version of Tukey's depth is defined by changing F to F_n. For $d = 1$, it equals $TD(F; x) = \min\{F(x), 1 - F(-x)\}$.

Liu (1995) proposed a few control charts based on these notions of data depths, which do not depend on the assumption of multivariate normality for the underlying process variable. Let us consider G be the distribution of the process vector when the process is in-control and $D_G(\boldsymbol{y})$ denotes the depth of a point $\boldsymbol{y} \in \mathbb{R}^d$ with respect to G. Define

$$r_G(\boldsymbol{y}) = P\{D_G(\boldsymbol{Y}) \leq D_G(\boldsymbol{y}) | \boldsymbol{Y} \sim G\}. \tag{2.14}$$

If the depth function D_G is unavailable, we can use the historical data $\boldsymbol{Y}_1, \boldsymbol{Y}]_2, \ldots, \boldsymbol{Y}_m$ to construct the empirical c.d.f. G_m and the empirical version of the depth function $D_{G_m}(\cdot)$ and define

$$r_{G_m}(\boldsymbol{y}) = \#\{\boldsymbol{Y}_j | D_{G_m}(\boldsymbol{Y}_j) \leq D_{G_m}(\boldsymbol{y}), \quad j = 1, \ldots, m\}/m. \tag{2.15}$$

34

The r chart proposed by Liu (1995) is a plot of $r_{G_m}(\boldsymbol{X}_i)$ against the index i. It can be shown that the distribution of $r_{G_m}(\boldsymbol{X}_i)$ converges to a $U(0,1)$ distribution when the process is in-control and hence we can take the central line as $CL = 0.5$ and set a lower control limit of α to achieve an in-control ARL of $1/\alpha$.

Figure 2.10: A r chart based on data depth for a sample from multivariate normal distribution

To illustrate how it works, we simulate a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ with $m = 1000$ from a standard bivariate normal distribution to create our historical in-control data set. We simulate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{50}$ also from the same bivariate standard normal distribution, however, we simulate $\boldsymbol{X}_{51}, \ldots, \boldsymbol{X}_{100}$ from a bivariate normal distribution with mean vector $(2, 2)$ and the covariance matrix

$$\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}.$$

In Figure 2.10, we make a r chart for this example data and observe that it detects the shift in location and the scale quite effectively. We have considered $\alpha = 0.005$ here. In

this example, we have used the simplicial depth function introduced by Liu (1988).

To improve on the performance, Liu (1995) proposed another control chart called Q-chart. Let F_n denote the empirical distribution of the sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. Then define

$$Q(G_m, F_n) = \frac{1}{n} \sum_{i=1}^{n} r_{G_m}(\boldsymbol{X}_i). \tag{2.16}$$

We consider the data from the process in batches or subsamples. Let us assume that each subsample has a size n. A Q chart plots $Q(G_m, F_n^j)$ against the index j, where F_n^j is the empirical distribution of the $\boldsymbol{X}_i$s in the j-th subsample. Liu (1995) also showed that the Q chart should have a central line, $CL = 0.5$ and a lower control limit of

$$LCL = (n!\alpha)^{1/n}/n$$

for small values of α giving an in-control ARL of $1/\alpha$.
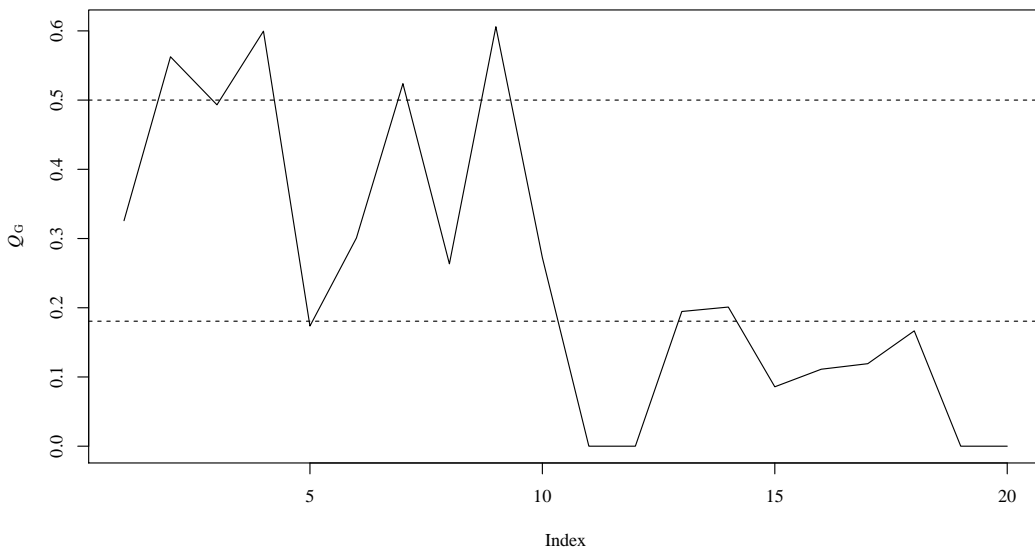
Figure 2.11: A Q chart based on data depth for a sample from multivariate normal distribution

36

In Figure 2.11, we present a Q chart for the same example data used in Figure 2.10 with a subsample size of $n = 5$. We can see that it prevents early out-of-control signals when the data is in-control.
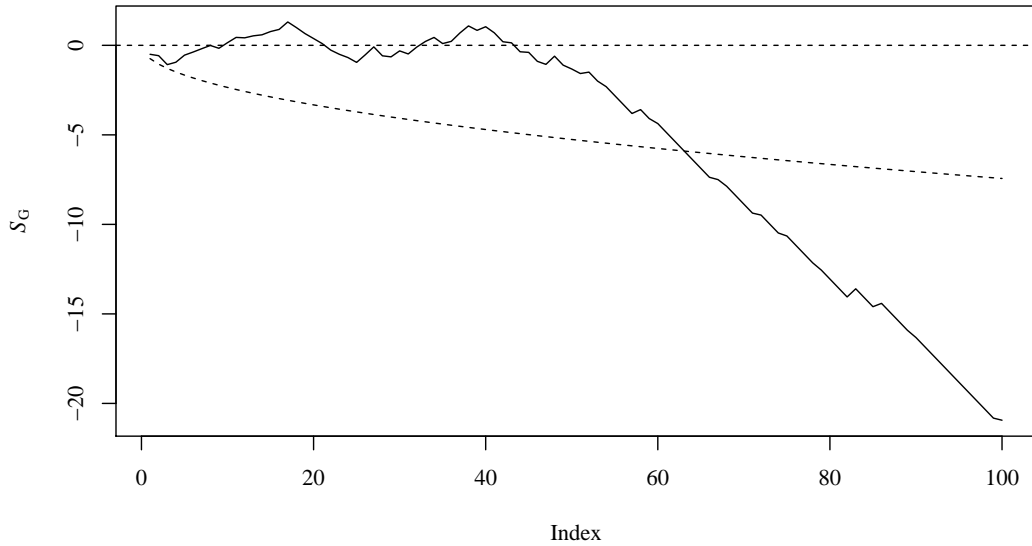
Figure 2.12: A S-chart based on data depth for a sample from multivariate normal distribution.

Extending the idea of an univariate CUSUM control charts, Liu (1995) proposed another kind of control chart called S-chart. Define

$$S_n(G) = \sum_{i=1}^{n}[r_G(X_i) - \frac{1}{2}] \tag{2.17}$$

and

$$S_n(G_m) = \sum_{i=1}^{n}[r_{G_m}(X_i) - \frac{1}{2}] \tag{2.18}$$

If we know the distribution G, a S-chart plots $S_n(G)$ against n. By a simple application of central limit theorem, it can be shown that this S-chart has the lower control limit $-(z_\alpha(n/12)^{1/2})$ where $z_{\alpha/2}$ is the $1 - \alpha/2$-th quantile of the standard normal distribution.

37

When the distribution G is unknown, we plot $S_n(G_m)$ against n and the lower control limit is given by $-[z_\alpha \sqrt{n^2[(\frac{1}{m} + \frac{1}{n})]/12}]$. The central line in both of this charts are 0.

In Figure 2.12, we present a S-chart for the same example data used in Figure 2.10. We can see that it prevents early out-of-control signals when the data is in-control.

Though the construction of the r, Q and S charts based on data depth do not depend on multivariate normality, computation of depth function is usually computationally very intensive for $d > 2$ and for that reason it has a very limited use in practice.

CHAPTER 3

MULTIVARIATE RANK BASED CONTROL

CHARTS

3.1   Definition and Basic Properties:

We generalize the univariate sign function $\text{sign}(x) = x/|x|$, for $x \neq 0$ to a multivariate notion of sign vector by defining

$$
\text{Sign}(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{for } \boldsymbol{x} \neq \boldsymbol{0} \\[2mm] \boldsymbol{0} & \text{for } \boldsymbol{x} = \boldsymbol{0} \end{cases}
$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_d^2}$ for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. Note that this multivariate notion of sign vector is nothing but the unit direction vector of $\boldsymbol{x}$ and was used in the literature to construct various statistics based on signs (for detail see, (Oja, 1999)).

Similarly, we define a multivariate version of the rank vector by

$$
Rank(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} Sign(\boldsymbol{x} - \boldsymbol{X}_i)
$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is a random sample with a common distribution

39

function F. Guha (2012) observed that

1. $\|Rank(\boldsymbol{x})\| < 1$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

2. $Rank(\boldsymbol{x}) = \boldsymbol{0}$ implies that $\boldsymbol{x}$ is the spatial median of the d-dimensional data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

3. Smaller values of $\|Rank(\boldsymbol{x})\|$ implies that $\boldsymbol{x}$ is more central to the data cloud and larger values of $\|Rank(\boldsymbol{x})\|$ indicates that $\boldsymbol{x}$ is more extreme. Moreover the direction of the vector $Rank(\boldsymbol{x})$ suggests the direction in which $\boldsymbol{x}$ is extreme compared to the data cloud.

4. Following Koltchinskii (1997), $E_F(Rank(\boldsymbol{x}))$ is a 1-1 function of the multivariate distribution function F.

We can also note that $Rank(\boldsymbol{x})$ is the inverse function of the multivariate geometric quantile function (Chaudhuri, 1996) $Q(\boldsymbol{u})$ in the sense that $Rank(\boldsymbol{x}) = \boldsymbol{u}$ implies that $Q(\boldsymbol{u}) = \boldsymbol{x}$ and vice-versa. If we define a measure of outlyingness by $\|Rank(\boldsymbol{x}\|$, then it is easy to verify that this measure of outlyingness is invariant under orthogonal and homogeneous scale transformations (see Serfling (2004), for some related discussion). A population version of that outlyingness function can be defined as

$$R_G(\boldsymbol{x}) = \|E_G(Rank(\boldsymbol{x})\| = \left\| E_G \left(\frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|} \right) \right\|$$

where the vector $\boldsymbol{X}$ has a distribution G.

In this section, we propose some control charts following the ideas of Liu (1995) based on multivariate ranks.

3.2 The r-Charts

We propose a new Shewhart type multivariate control chart using this definition of multivariate rank function and the r chart proposed by Liu (1995). Given a historical data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ when the process is in-control, we define

$$R_{G_m}(\boldsymbol{x}) = \|Rank_{G_m}(\boldsymbol{x})\| = \left\| \frac{1}{m} \sum_{i=1}^{m} \frac{x - \boldsymbol{Y}_i}{\|\boldsymbol{x} - \boldsymbol{Y}_i\|} \right\| \tag{3.1}$$

and

$$r_{G_m}(\boldsymbol{x}) = \#\{\boldsymbol{Y}_j | R_{G_m}(\boldsymbol{Y}_j) \leq R_{G_m}(\boldsymbol{x}), \ \ j = 1, \ldots, m\}/m. \tag{3.2}$$

We state the following result on the distribution of $r_G(\boldsymbol{X})$ when the process is in-control.

Theorem 3.2.1 *Assume that a d-dimensional random vector $\boldsymbol{X}$ has a conntinuous distribution function G. Let $U(0,1)$ denotes an uniform distribution on (0,1). Let $r_G(\boldsymbol{x}) = P(R_G(\boldsymbol{X}) \leq R_G(\boldsymbol{x}))$. If the multivariate rank function of $\boldsymbol{x}$ $(R_G(x))$ has a continuous distribution, then*

(i) $r_G(\boldsymbol{X})$ is distributed uniformly on (0,1).

(ii) $r_{G_m}(\boldsymbol{X})$ converges in distribution to $U(0,1)$ as $m \to \infty$.

Proof. The proof of (i) is trivial as G is assumed to have a continuous multivariate distribution, $R_G(\boldsymbol{X})$ will also have a univariate continuous distribution and $r_G(\boldsymbol{x})$ is the distribution function of $R_G(\boldsymbol{X})$. To prove (ii), note that

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |r_{G_m}(\boldsymbol{x}) - r_G(\boldsymbol{x})| \to 0 \quad \text{a.s.} \quad \text{as } m \to \infty$$

by Lemma 2 of Makinde and Chakraborty (2015). $\qquad\square$

We plot $r_{G_m}(\boldsymbol{X}_i)$ against the index i to construct an r chart. We can use the central line to be $CL = 0.5$ and the chart will signal a process to be out-of-control when $r_{G_m}(\boldsymbol{X}_i) > 1 - \alpha$ so that the in-control ARL is $1/\alpha$.
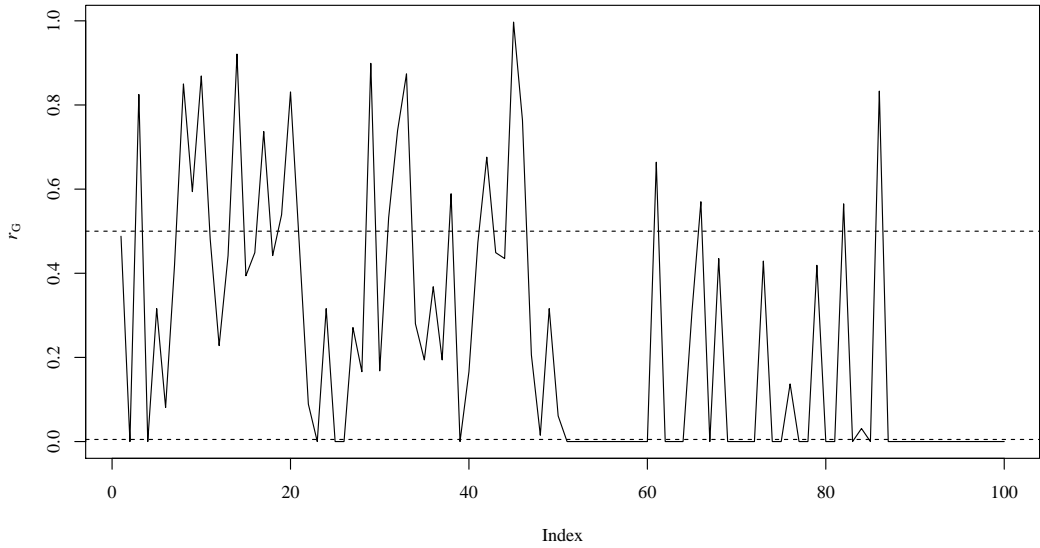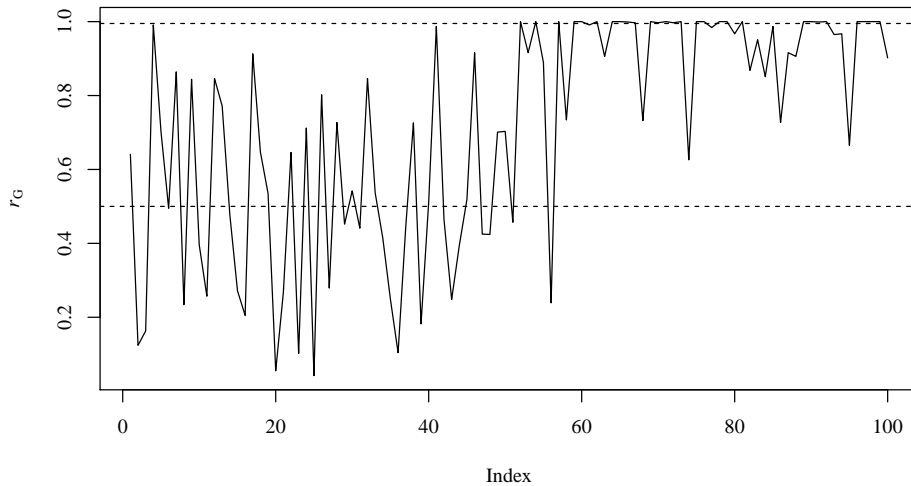

Figure 3.1: A r chart based on multivariate ranks for a sample from multivariate normal distribution

To illustrate how it works, we simulate a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ with $m = 1000$ from a standard bivariate normal distribution to create our historical in-control data set. We simulate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{50}$ also from the same bivariate standard normal distribution, however, we simulate $\boldsymbol{X}_{51}, \ldots, \boldsymbol{X}_{100}$ from a bivariate normal distribution with mean vector $(2, 2)$ and the covariance matrix

$$
\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}.
$$

We have considered $\alpha = 0.005$ here. A r chart based on multivariate ranks is presented in Figure 3.1. It shows a similar behaviour as 2.10.

42

3.3 The Q-Charts

If we have a sample of size n in every run instead of a single observation, we can modify the proposed r chart. Let us define

$$Q(G, F_n) = \frac{1}{n} \sum_{i=1}^{n} r_G(\boldsymbol{X}_i) \tag{3.3}$$

and

$$Q(G_m, F_n) = \frac{1}{n} \sum_{i=1}^{n} r_{G_m}(\boldsymbol{X}_i) \tag{3.4}$$

where $r_{G_m}(\boldsymbol{x})$ is as defined before.

Theorem 3.3.1 *Assume that the conditions in Theorem 3.2.1 hold. Then*

(i) $\sqrt{n}[Q(G, F_n) - \frac{1}{2}] = \sqrt{n}[\frac{1}{n} \sum_{i=1}^{n} r_G(X_i) - \frac{1}{2}]$ *converges in distribution to* $N(0, 1/(12))$ *as* $n \to \infty$.

(ii) $[Q(G_m, F_n) - \frac{1}{2}] = [\frac{1}{n} \sum_{i=1}^{n} r_{G_m}(X_i) - \frac{1}{2}]$ *converges in distribution to* $N\{0, 1/(12)\}$, *as* $\min(m, n) \to \infty$.

Proof. Proof of (i) is a straightforward application of central limit theorem. For (ii), observe that $\sqrt{m}(Q(G_m, F_n) - Q(G, F_n))$ converges in distribution to $N(0, 1/12)$ and $\sqrt{n}(Q(G, F_n - 1/2)$ converges in distribution to $N(0, 1/12)$ by (i) as $\min(m, n) \to \infty$. Also observe that $(Q(G_m, F_n) - 1/2) = (Q(G_m, F_n) - Q(G, F_n)) + (Q(G, F_n) - 1/2)$. $\square$

Let the j-th sample be $(\boldsymbol{X}_1^j, \ldots, \boldsymbol{X}_n^j)$ and its distribution function is denoted by F_n^j, then we propose a Q-chart by plotting $Q(G, F_n^j)$ against j if the in-control distribution G is known and $Q(G_m, F_n^j)$ against the index j if G is unknown and it is estimated by the historical data with empirical distribution function G_m. For both of them, the central line is $CL = 0.5$. For large values of m and n, the upper control limit, UCL, for the first
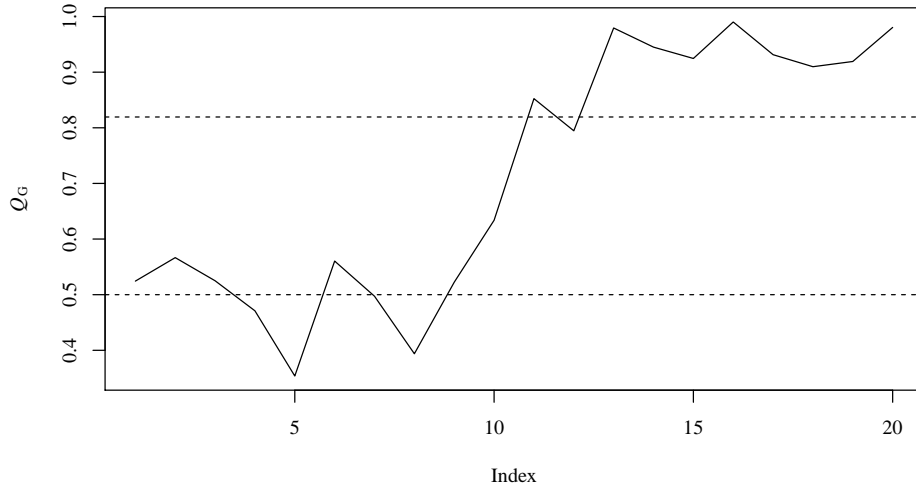
Figure 3.2: A Q chart based on multivariate ranks for a sample from multivariate normal distribution

one is given by $0.5 + z_\alpha/\sqrt{12n}$ and for the second one it should be

$$UCL = 0.5 + z_\alpha \sqrt{\frac{1}{12}\left(\frac{1}{m} + \frac{1}{n}\right)}$$

where z_α is the $100(1-\alpha)$-th percentile of the standard normal distribution.

For small values of n, we consider the exact distribution of $Q(G, F_n)$.

Theorem 3.3.2 *The upper control limit of the control chart of $Q(G, F_n^j)$ against j is given by $1 - (n!\alpha)^{1/n}/n$ for $\alpha \le 1/n!$*

Before we prove the theorem, we state a result from Feller (1971) on the distribution function of sum of uniform random variables.

Proposition 3.3.1 *Let $U_1, \ldots, U_n$ be n i.i.d. uniformly distributed random variables over the interval $(0,1)$. Let $H_n(t)$ denotes the distribution function of $\sum_{i=1}^n U_i$, that is $H_n(t) =$*

$P(\sum_{i=1}^{n} U_i \le t)$. *Then for each* $n = 0, 1, 2, \ldots$, $H_n(t) = 0$ *for* $t \le 0$ *and*

$$H_n(t) = \frac{1}{n!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}(t-k)_+^n$$

where

$$(x)_+^n = \begin{cases} 0 & \text{if } x \le 0 \\ x^n & \text{otherwise.} \end{cases}$$

Proof of Theorem 3.3.2. Note that, $1 - Q(G, F_n)$ can be viewed as an average of $U(0,1)$ random variables and we need to find a number ℓ_α such that $P(1 - Q(G, F_n) \le \ell_\alpha) = \alpha$ and consequently $H_n(n\ell_\alpha) = \alpha$. Therefore, for $\alpha \le 1/n!$,

$$\frac{1}{n!}(n\ell_\alpha)^n = \alpha$$

gives $\ell_\alpha = (n!\alpha)^{1/n}/n$. Hence the UCL for the chart based on $Q(G, F_n)$ is given by

$$UCL = 1 - \frac{(n!\alpha)^{1/n}}{n}.$$

As an illustration of the proposed Q-chart, we use the same simulated data from Figure 3.1 with a sample size of $n = 5$ and use an upper control limit of $1 - (n!\alpha)^{1/n}/n$ for $\alpha \le 1/n!$ with $\alpha = 0.005$. The in-control ARL for such a chart will be $1/\alpha$. A Q chart is presented in Figure 3.2. It shows a similar behaviour as 2.11.

3.4 The S-Chart

A natural extension of the univariate CUSUM control chart would be to plot $S_n(G)$ or $S_n(G_m)$ defined by

$$S_n(G) = \sum_{i=1}^{n}[r_G(\boldsymbol{X}_i) - \frac{1}{2}]$$

45

and

$$S_n(G_m) = \sum_{i=1}^{n} [r_{G_m}(\boldsymbol{X}_i) - \frac{1}{2}].$$

Note that $S_n(G) = n(Q(G, F_n - 0.5)$ and $S_n(G_m, F_n) = n(Q(G_m, F_n) - 0.5)$ and we can almost immediately derive the following result from Theorem 3.3.1.

Theorem 3.4.1 *Under the assumptions of Theorem 3.3.1, we have the following:*

(i) $S_n(G)$ converges in distribution to $N(0, n/12)$ as $n \to \infty$.

(ii) $S_n(G_m)$ converges in distribution to $N(0, n^2(1/m + 1/n)/12)$ as $\min(m, n) \to \infty$.
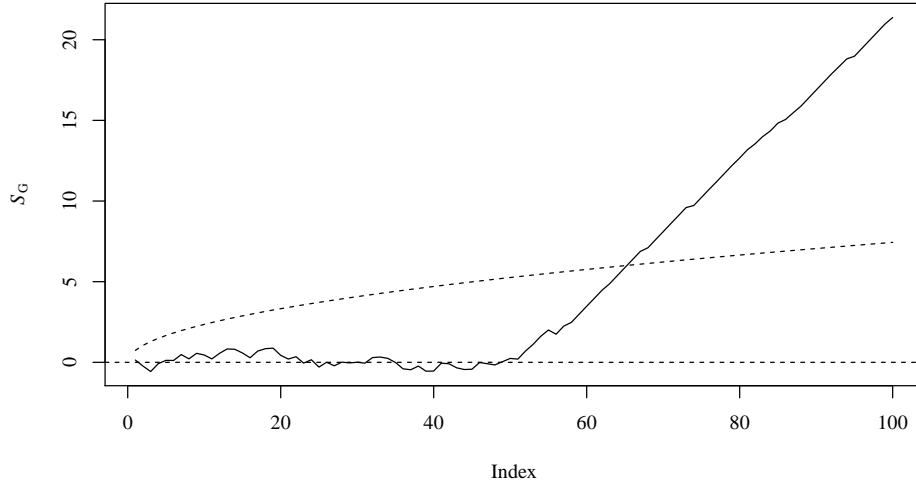
Figure 3.3: A S chart based on multivariate ranks for a sample from multivariate normal distribution

The above Theorem implies that to attain an average run length of $1/\alpha$, the upper control limit for the chart based on $S_n(G)$ should be $z_\alpha \sqrt{n/12}$ and that for the chart based on $S_n(G_m)$ should be $z_\alpha \sqrt{n^2[(1/m) + (1/n)]/12}$ for sufficiently large values of m and n. Observe that the upper control limit is not a straight line rather a curve of order $\sqrt{n}$. We illustrate this chart in Figure 3.3 with the same simulated data as in Figure 3.1.

46

Now note that as the chart goes up very fast, we may like to look into a standardised version of this S-chart for plotting by defining

$$S_n^*(G) = S_n(G)/\sqrt{n/12}$$

or

$$S_n^*(G_m) = S_n(G_m)/\sqrt{n^2[(1/m) + (1/n)]/12}$$

and then we can use the upper control limit of $UCL = z_\alpha$ and a central line $CL = 0$.

The standardized version of the S-chart is illustrated in Figure 3.4.
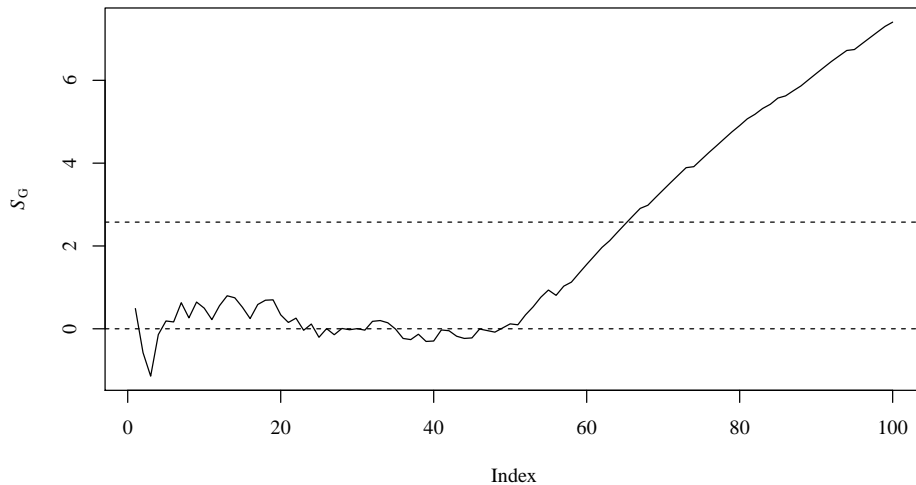
Figure 3.4: A S^* chart based on multivariate ranks for a sample from multivariate normal distribution

3.5 Average Run Lengths of the Proposed Charts

In this section we study the performance of the proposed r-chart based on multivariate ranks through average run lengths. We consider $\alpha = 0.005$ so that theoretical average run lengths of the proposed scheme should be 200 when the process is in-control. We

simulate from multivariate normal, Laplace and t distribution with 3 degrees of freedom as defined before for dimensions $d = 2, 3, 5,$ and 10. The location vector $\boldsymbol{\mu}$ is taken as a d-dimensional vector with the first element to be δ and all other elements to be 0. The scale matrix $\Sigma = \mathbf{I}_d$ the $d \times d$ identity matrix. In Figure 3.5, we present the average run lengths obtained from 1000 simulations against δ.

We observe that the performance of the r-chart is the best when the underlying distribution is multivariate normal. However, for multivariate Laplace and t distribution, they fail to detect small shifts in the location for higher dimensional data efficiently. Though their performance is still better than the T^2-chart discussed earlier.

Next we study the performance of the proposed Q-chart based on multivariate ranks through average run lengths. We consider $\alpha = 0.005$ so that the theoretical average run lengths of the proposed scheme should be 200 when the process is in-control. We simulate from multivariate normal, Laplace and t distribution with 3 degrees of freedom as defined before for dimensions $d = 2, 5, 10$ and 20. The location vector $\boldsymbol{\mu}$ is taken as a d-dimensional vector with the first element to be δ and all other elements to be 0. The scale matrix $\Sigma = \mathbf{I}_d$, the $d \times d$ identity matrix. The sub-sample size is $n = 5$. In Figure 3.6, we present the average run lengths obtained from 1000 simulations against δ.

We observe a similar behaviour as in Figure 3.5.

In Figure 3.7, we present ARL curves for S-charts when $\alpha = 0.005$ thus the in control ARL is 200 for multivariate normal distribution with the same simulation setting as above for dimension $d = 2, 5, 10$ and 20. In Figure 3.8, we present the ARL curve for multivariate Laplace distribution with dimension $d = 2$ only.

Next we present a small simulation study on the average run length of the proposed r and Q-charts for bivariate normal distribution, when the scale matrix is not an identity
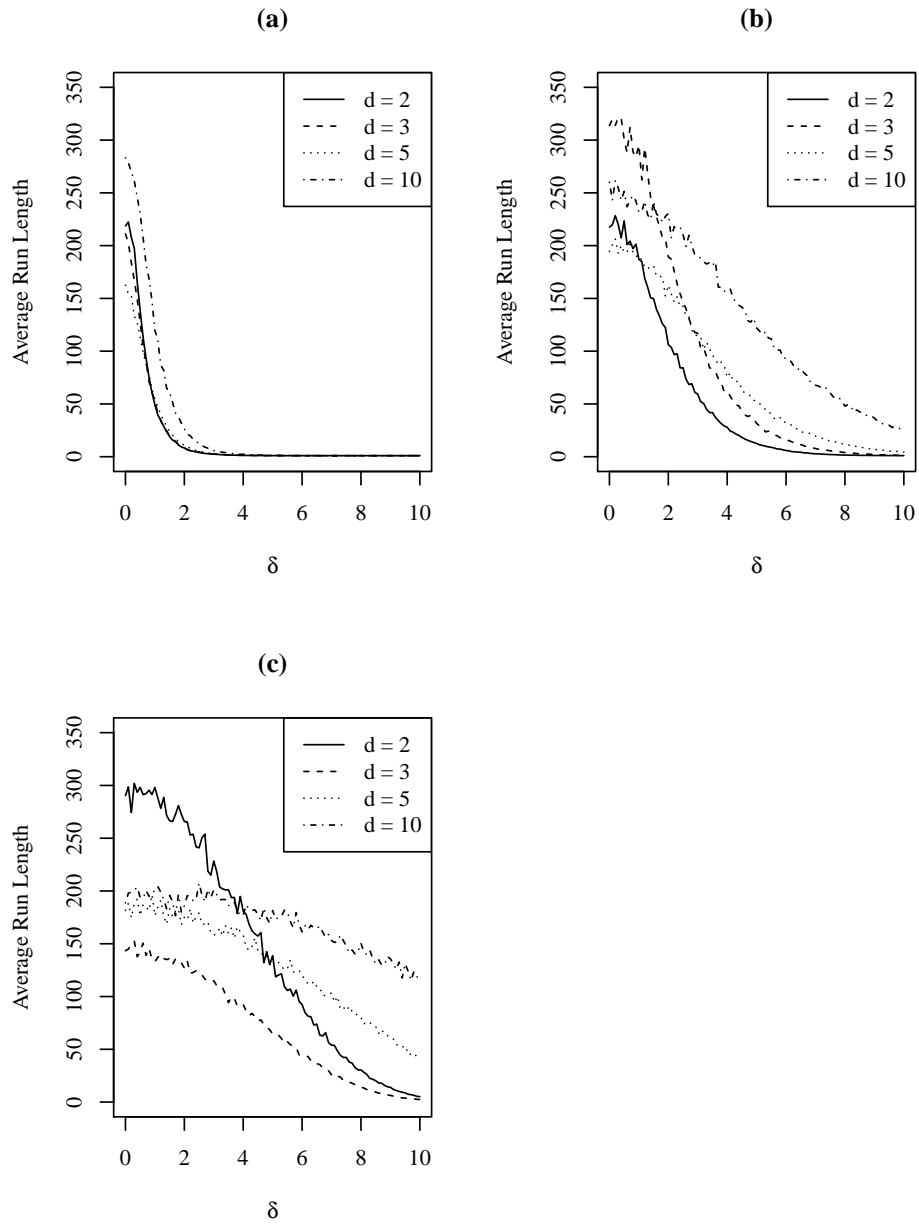
Figure 3.5: ARL curves for r-charts based on multivariate ranks when the distribution of the process variables are (a) normal , (b) Laplace, and (c) t distributions with in-control ARL 200 for dimensions $d = 2, 3, 5$ and 10.
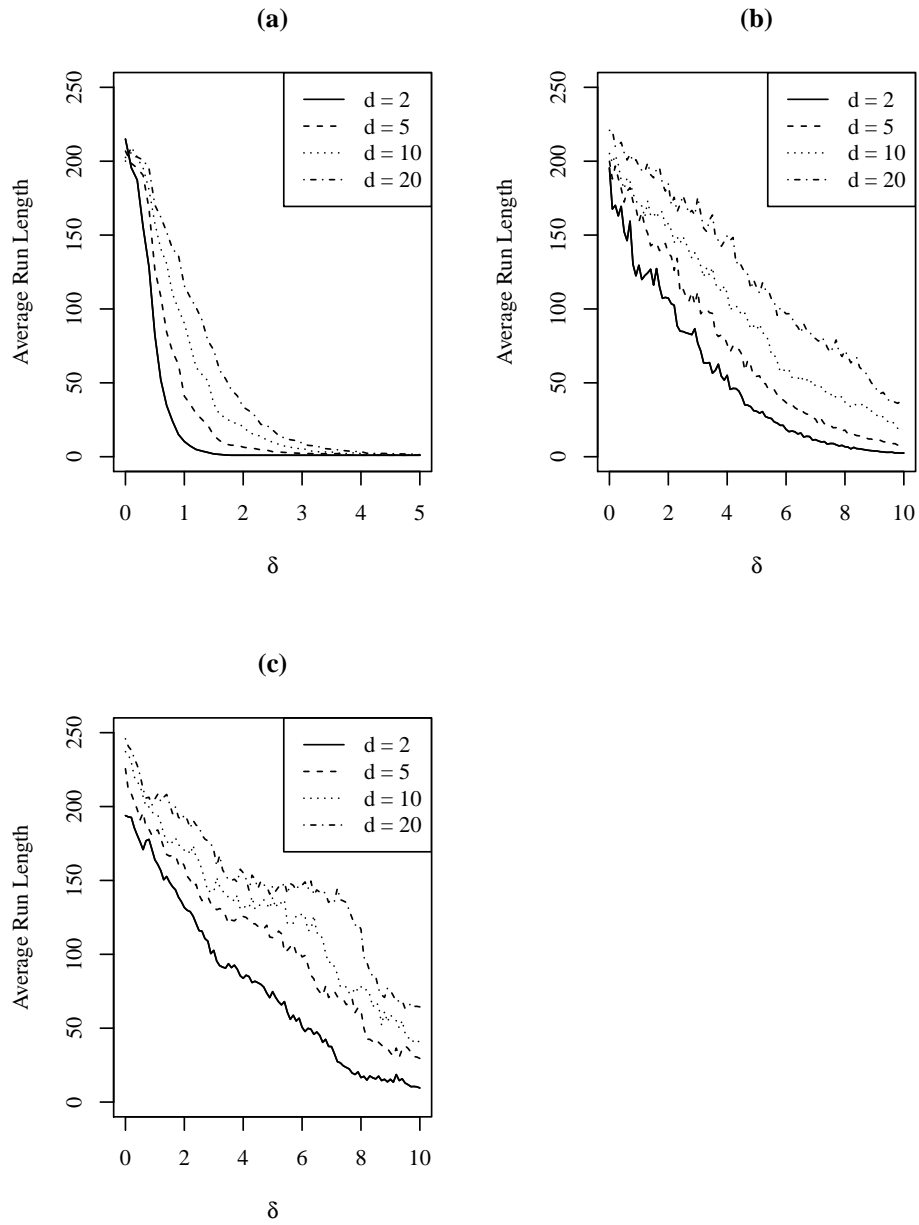
Figure 3.6: ARL curves for multivariate Q-charts based on multivariate ranks when the distribution of the process variable are (a) normal , (b) Laplace, and (c) t distributions for in-control ARL 200 for dimensions $d = 2, 5, 10, 20$.
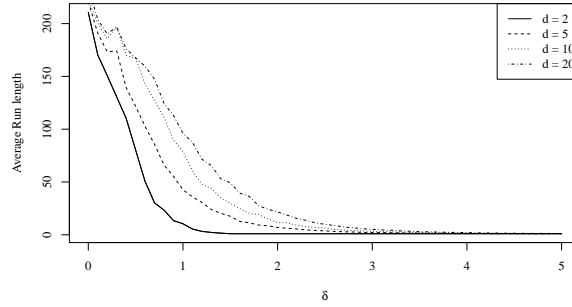
Figure 3.7: ARL curve for multivariate S-charts based on ranks when the distribution of the process variable is a normal distribution for in-control ARL 200 for dimension $d = 2, 5, 10$ and 20 and non-centrality parameter δ.

matrix. We consider

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

for $\rho = 0.0, 0.5, 0.8$ and 0.9. The ARL curve for the r-chart is presented in Figure 3.9 and that for the Q-chart is presented in Figure 3.10. We observe that for different values of ρ, the ARL curves are different. This is due to the fact that multivariate rank functions defined in this chapter are not invariant under affine transformations. We discuss more about affine invariance and invariant versions of multivariate rank functions in the next Chapter.

Observe that, the r, Q and S-charts can be computed for any dimension unlike the multivariate control charts proposed by Liu (1995). The computational ease of the charts makes these proposals attractive. However, we need to investigate further to decide on the control limits for smaller values of n as we have seen that the proposed charts do not perform very well with the upper control limit based on asymptotic results, especially for non-normal distributions.

51
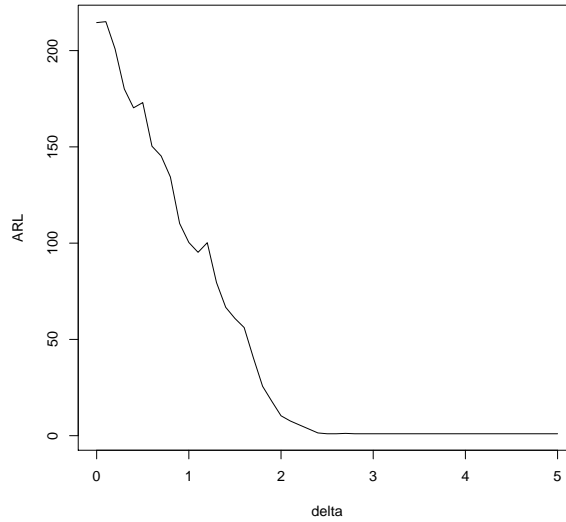
Figure 3.8: ARL curve for multivariate S-charts based on ranks when the distribution of the process variable is a Laplace distribution for in-control ARL 200 for dimension $d = 2$.

3.6 Real Data Examples

In this example we considered a real data which originated from a process capability study for turning aluminium pins. This data was obtained from a manufacturing process which recorded 6 measurements for each of the 70 observations (Fuchs and Kenett, 1998). The manufacturing process of turning aluminium pins is often monitored by a quality engineer. The first three measurements are diameter measurements on three different locations on the main part of the pin. The fourth measurement represented a diameter measurement at the cap. The last two are the lengths measurements, with the cap and without it. We considered the first 30 observations as the historical in-control data, which is presented in Table 3.1. To illustrate our proposed control charts, we have used the following 40 observations in Table 3.2.

In Figure 3.11, we have made an r-chart with these 40 new observations.We observe that 9 observations are beyond the control limit, namely, the observation numbers 10,
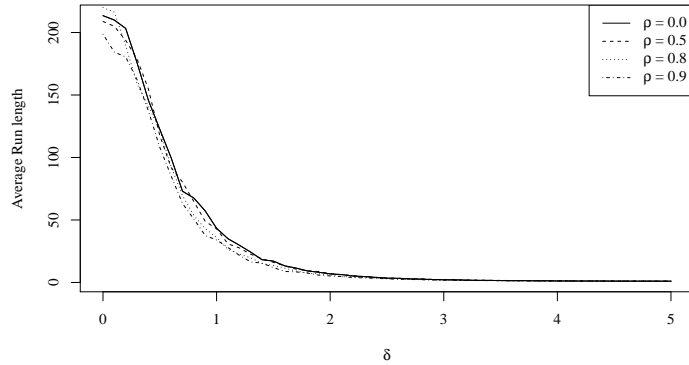
Figure 3.9: ARL curves for multivariate r-charts based on multivariate ranks when the distribution of the process variable is bivariate normal distributions with correlation ρ for in-control ARL 200.

17, 18, 19, 22, 23, 25, 31, and 36. Here we have taken the upper control limit to be $UCL = 0.995$. In Figure 3.12, we produce a Q-chart by taking subsamples of size $n = 4$. It shows that the subsample number 3, 4, 5 and 6 are beyond the control limit. The upper control limit for the Q chart is $1 - (4!0.005)^{0.25}/4 = 0.8529$. In Figure 3.13, we produce the S-chart with upper control limit given by $z_{0.005}\sqrt{k/12}$ for the k the observation. We observe that at 10th observation, it crosses the upper control limit and stays above as univariate CUSUM charts usually do.

3.7 Concluding Remarks

In practice, the assumption of normality of the data may be often required for more traditional techniques that may be hard to justify in practice (Chakraborty and Chaudhuri, 1999). In many applications, the assumption of normality (or some other specified distribution) of data may be violated that affects significantly on the performance of the control chart. Liu (1995) was among the first few to propose a control chart which does not depend on the assumption of normality. Though the properties of the proposed control charts are quite attractive due to their distribution-free nature, they did not provide
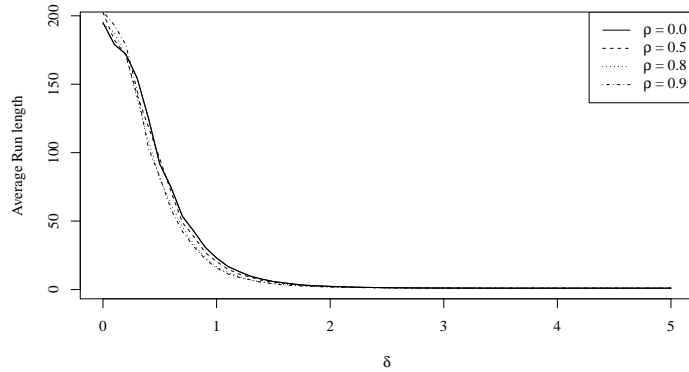
Figure 3.10: ARL curves for multivariate Q-charts based on multivariate ranks when the distribution of the process variable is bivariate normal distributions with correlation ρ for in-control ARL 200.

any theoretical or simulation results on average run lengths of the proposed methods for their out-of-control behaviours. Most of the depth functions (e.g. simplicial, half-space or majority depth) are computationally intensive and nearly impossible to compute exactly for dimensions greater than 2. Therefore, these control charts based on data depths have limited use and infeasible for processes with more than 2 process variables.

In this chapter, we considered a notion of multivariate rank function, popularly known as spatial ranks, which retain some important features of the univariate rank function and computationally very simple with time complexity of $O(n)$ for any dimension $d \geq 2$. We propose some control charts following the idea of Liu (1995) using the distribution of the lengths of the multivariate rank vectors. It can be shown that the r-chart proposed in here is equivalent to the Hotelling's T^2 chart when the distribution is multivariate normal and also it is equivalent to the T^2 chart with optimal control limits if the underlying distribution is spherically symmetric. Therefore, this is an optimal control chart in terms of out-of-control ARL whenever the distribution is spherically symmetric and the shift occurs only in the location vector. We have also discussed some extensions of the r-chart to Q-chart and S-chart, where S-charts are analogous to CUSUM charts and detects small
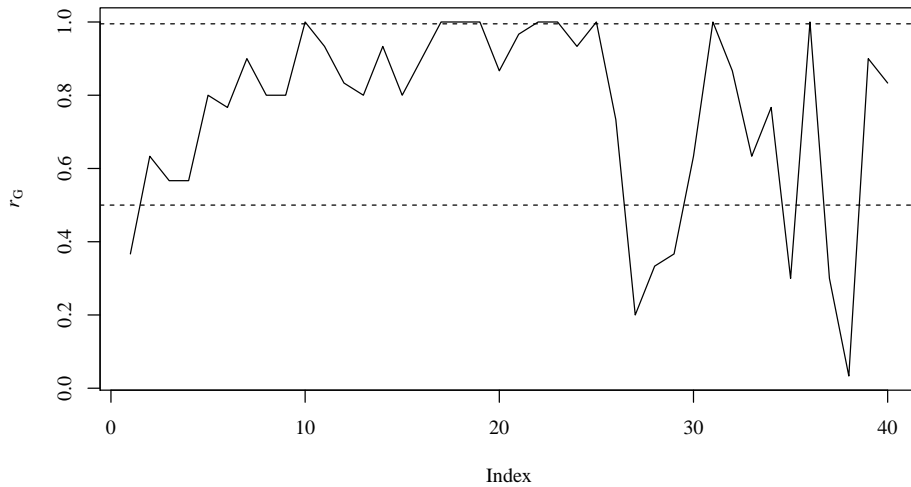
Figure 3.11: r-chart for the Aluminium pin data.

shifts better than r-charts.

A major inadequacy of the proposed multivariate rank function is that it is invariant under orthogonal transformations but not under general nonsingular transformations, which makes the proposed control charts optimal only for spherically symmetric distributions, that is, they are optimal only when the scatter matrix (or, the covariance matrix) Σ associated with the distribution is λI_d for some constant λ. To resolve this issue, we can use affine invariant multivariate ranks as discussed in Chakraborty (2001). The optimal behaviour of the proposed control charts as discussed above then extends to the elliptically symmetric distributions or when Σ is a general non-singular positive definite matrix. It still retains the computational simplicity with only added complexity of computing an estimate of the scatter matrix Σ, if that is unknown. All of our theoretical results and simulations show that the proposed procedures are very promising in terms of computational simplicity in high dimensions as well as performance in detecting out-of-control signals.
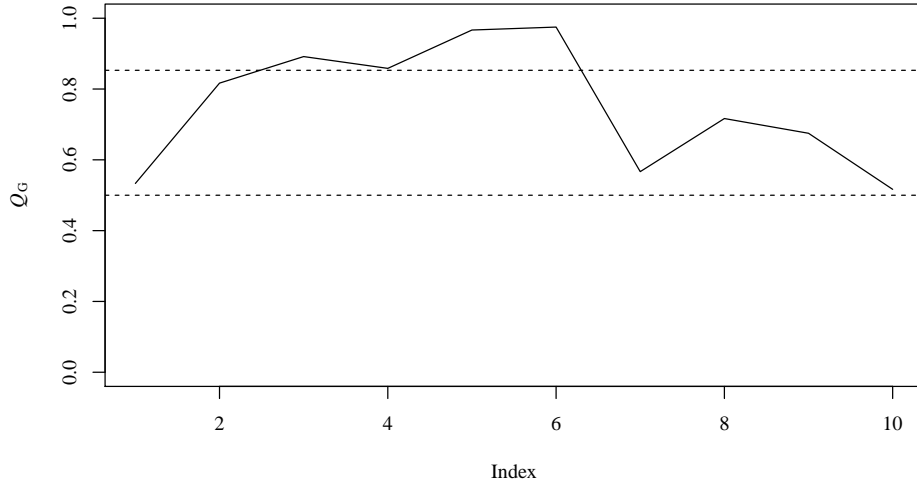
Figure 3.12: Q-chart for the Aluminium pin data.
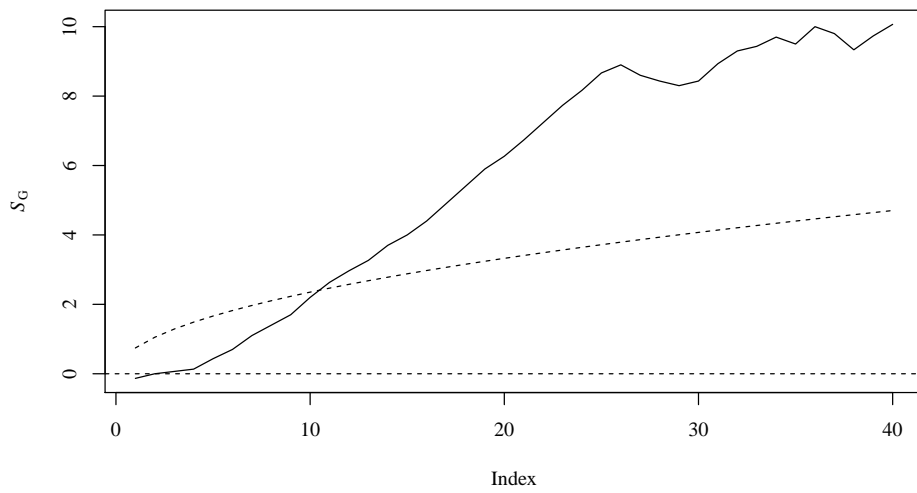
Figure 3.13: S-chart for the Aluminium pin data.

56

Table 3.1: Historical Data for Aluminium Pins

	1	2	3	4	5	6
1	9.99	9.97	9.96	14.97	49.89	60.02
2	9.96	9.96	9.95	14.94	49.84	60.02
3	9.97	9.96	9.95	14.95	49.85	60.00
4	10.00	9.99	9.99	14.99	49.89	60.06
5	10.00	9.99	9.99	14.99	49.91	60.09
6	9.99	9.99	9.98	14.99	49.91	60.08
7	10.00	9.99	9.99	14.98	49.91	60.08
8	10.00	9.99	9.99	14.99	49.89	60.09
9	9.96	9.95	9.95	14.95	50.00	60.15
10	9.99	9.98	9.98	14.99	49.86	60.06
11	10.00	9.99	9.98	14.99	49.94	60.08
12	10.00	9.99	9.99	14.99	49.92	60.05
13	9.97	9.96	9.96	14.96	49.90	60.02
14	9.97	9.96	9.96	14.96	49.91	60.02
15	9.97	9.97	9.96	14.97	49.90	60.01
16	9.97	9.97	9.96	14.97	49.89	60.04
17	9.98	9.97	9.96	14.96	50.01	60.13
18	9.98	9.97	9.97	14.96	49.93	60.06
19	9.98	9.98	9.97	14.98	49.93	60.02
20	9.98	9.97	9.97	14.97	49.94	60.06
21	9.98	9.97	9.97	14.97	49.93	60.06
22	9.98	9.97	9.97	14.97	49.91	60.02
23	9.98	9.97	9.96	14.98	49.92	60.06
24	10.00	9.99	9.98	14.98	49.88	60.00
25	9.99	9.99	9.99	14.98	49.91	60.04
26	10.00	9.99	9.99	14.99	49.85	60.01
27	10.00	10.00	9.99	14.99	49.91	60.05
28	10.00	9.99	9.99	15.00	49.92	60.04
29	10.00	9.99	9.99	14.99	49.89	60.01
30	10.00	10.00	9.99	14.99	49.88	60.00

Table 3.2: New Data on Aluminium Pins

	1	2	3	4	5	6
1	10.00	9.99	9.99	14.99	49.92	60.03
2	10.00	9.99	9.99	15.00	49.93	60.03
3	10.00	10.00	9.99	14.99	49.91	60.02
4	10.00	9.99	9.99	14.99	49.92	60.02
5	10.00	9.99	9.99	14.99	49.92	60.00
6	10.00	10.00	9.99	15.00	49.94	60.05
7	10.00	9.99	9.99	15.00	49.89	59.98
8	10.00	10.00	9.99	14.99	49.93	60.01
9	10.00	10.00	9.99	14.99	49.94	60.02
10	10.00	10.00	9.99	15.00	49.86	59.96
11	10.00	9.99	9.99	14.99	49.90	59.97
12	10.00	10.00	10.00	14.99	49.92	60.00
13	10.00	10.00	9.99	14.98	49.91	60.00
14	10.00	10.00	10.00	15.00	49.93	59.98
15	10.00	9.99	9.98	14.98	49.90	59.99
16	9.99	9.99	9.99	14.99	49.88	59.98
17	10.01	10.01	10.01	15.01	49.87	59.97
18	10.00	10.00	9.99	14.99	49.81	59.91
19	10.01	10.00	10.00	15.01	50.07	60.13
20	10.01	10.00	10.00	15.00	49.93	60.00
21	10.00	10.00	10.00	14.99	49.90	59.96
22	10.01	10.01	10.01	15.00	49.85	59.93
23	10.00	9.99	9.99	15.00	49.83	59.98
24	10.01	10.01	10.00	14.99	49.90	59.98
25	10.01	10.01	10.00	15.00	49.87	59.96
26	10.00	9.99	9.99	15.00	49.87	60.02
27	9.99	9.99	9.99	14.98	49.92	60.03
28	9.99	9.98	9.98	14.99	49.93	60.03
29	9.99	9.99	9.98	14.99	49.89	60.01
30	10.00	10.00	9.99	14.99	49.89	60.01
31	9.99	9.99	9.99	15.00	50.04	60.15
32	10.00	10.00	10.00	14.99	49.84	60.03
33	10.00	10.00	9.99	14.99	49.89	60.01
34	10.00	9.99	9.99	15.00	49.88	60.01
35	10.00	10.00	9.99	14.99	49.90	60.04
36	9.90	9.89	9.91	14.88	49.99	60.14
37	10.00	9.99	9.99	15.00	49.91	60.04
38	9.99	9.99	9.99	14.98	49.92	60.04
39	10.01	10.01	10.00	15.00	49.88	60.00
40	10.00	9.99	9.99	14.99	49.95	60.01

CHAPTER 4

SHEWHART TYPE CHARTS BASED ON

MULTIVARIATE SIGNS

4.1   Introduction

Consider a process with d process variables, which have a joint d-dimensional distribution F with location parameter $\boldsymbol{\mu} \in \mathbb{R}^d$ and the $d \times d$ scale matrix Σ. If the second moments exist, Σ may be a scalar multiple of the covariance matrix. Let us assume that the in-control values of the parameters are $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\Sigma = \Sigma_0$, where both are known in Phase-I control charts. Then as we discussed earlier, a Shewhart type multivariate control chart can be defined using the Hotelling's T^2 statistic

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

where $\bar{\boldsymbol{X}}$ is the sample mean based on a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n from the distribution F. If F is multivariate normal and the process is in-control, T^2 has an χ^2 distribution with d degrees of freedom and if the process mean shifts to $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$, but Σ_0 remains the covariance matrix, then T^2 will have a non-central χ^2 distribution with d degrees of freedom and non-centrality parameter $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. In most applications,

however, Σ_0 will be unknown and needs to be estimated by the sample covariance matrix $\mathbf{S}$ and the statistic will be defined by

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0).$$

Under multivariate normality, the statistic $(n-d)T^2/(d(n-1))$ will have a F distribution with d and $n-d$ degrees of freedom. However, when the underlying distribution of the process variables is not multivariate normal, we may not have a simple exact distribution for the T^2 statistic. For large sample size n, T^2 will have a χ^2 distribution with d degrees of freedom, if we assume that the second moments of F exist. But we have seen in Chapter 2 that the performance of control charts based on T^2 for non-normal distributions is poor. Moreover the assumption of existence of second moments may not be valid for the process variables. In this chapter, we consider some control charts based on multivariate sign vectors defined earlier.

4.2   Multivariate Signs

The multivariate sign vector was defined in Chapter 3 as

$$\text{Sign}(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{for } \boldsymbol{x} \neq \mathbf{0} \\ \mathbf{0} & \text{for } \boldsymbol{x} = \mathbf{0} \end{cases}$$

where $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_d^2}$ for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. Note that this is nothing but the direction vector of $\boldsymbol{x}$. Now for the target location vector $\boldsymbol{\mu}_0$, we may define the average direction a sample of size n compared to $\boldsymbol{\mu}_0$ as

$$\boldsymbol{S}_n = \frac{1}{n} \sum_{i=1}^{n} \text{Sign}(\boldsymbol{X}_i - \boldsymbol{\mu}_0).$$

Let

$$\mathbf{C}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Sign}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\{\mathrm{Sign}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\}^\top.$$

Then we define a statistic similar to the T^2 statistics as

$$W_n = n\boldsymbol{S}_n^\top \mathbf{C}_n^{-1} \boldsymbol{S}_n.$$

It is easy to observe that W_n will have a χ^2 distribution with d degrees of freedom as $n \to \infty$ (Chaudhuri, 1992). If we define the upper control limit to be $\chi^2_{d,\alpha}$, the $(1 - \alpha)$-th quantile of the Chi-squared distribution with d degrees of freedom, then the control chart based on W_n will have an average run length of $1/\alpha$ for large values of n.
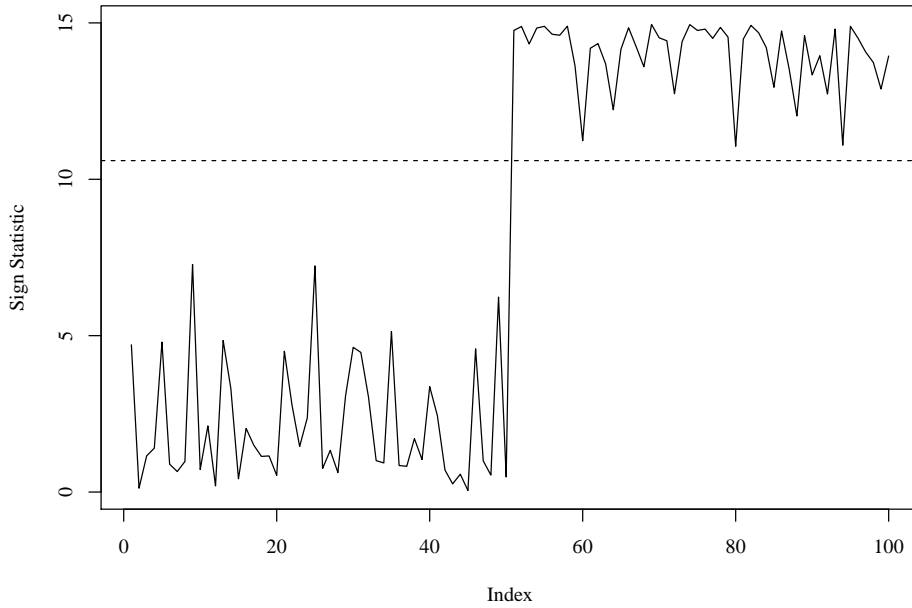
Figure 4.1: A Shewhart type control chart based on multivariate sign statistic.

To illustrate the proposed control chart based on the statistic W_n, we present an example with 100 simulated sample of size $n = 15$ each from the bivariate normal distribution with covariance matrix $\Sigma = \mathbf{I}_2$. The first 50 samples are with mean $(0, 0)^\top$ and the last 50

samples are with mean $(2, 0)$. Figure 4.1 shows that the proposed chart detects the shift in location quite easily for this example. The upper control limit $UCL = \chi^2_{2,0.005} = 10.5966$.
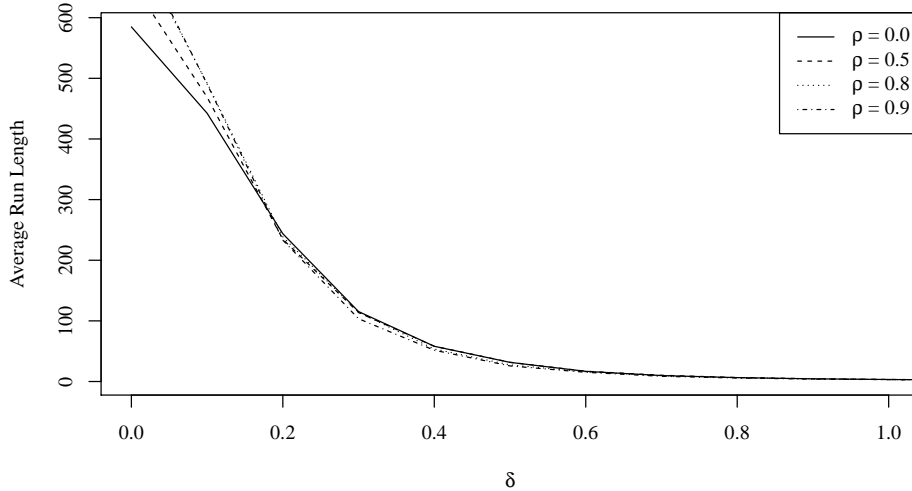
Figure 4.2: Average Runlengths of the control chart based on multivariate sign statistic for bivariate normal distribution with different values of the correlation coefficient ρ.

In Figure 4.2 we present average run lengths of the proposed control chart when the underlying distribution is bivariate normal with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and covariance matrix Σ such that

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{and} \quad \delta^2 = \boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu}.$$

We choose $\rho = 0.0$, 0.5, 0.8, and 0.9. Here the sample size $n = 15$ and the simulation size is 1000 for computing the average run lengths. Observe that for small values of δ, the average run lengths are different for different values of ρ, which is due to the fact that the sign statistic is not invariant under non-singular transformations of the data. We would to like to note here that the average sign vector $\boldsymbol{S_n}$ is equivariant under orthogonal

transformation, that is,

$$\boldsymbol{S}_n(\mathbf{A}\boldsymbol{X}_1, \ldots, \mathbf{A}\boldsymbol{X}_n) = \mathbf{A}\boldsymbol{S}_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$$

where $\mathbf{A}$ is any orthogonal matrix. Hence the statistic W_n is invariant under orthogonal transformations, that is

$$W_n(\mathbf{A}\boldsymbol{X}_1, \ldots, \mathbf{A}\boldsymbol{X}_n) = W_n(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

However, $\boldsymbol{S}_n$ is not equivariant under arbitrary non-singular transformation, that is the above relation does not hold when $\mathbf{A}$ is any arbitrary non-singular matrix. This is a major drawback behind using this sign statistic to construct a control chart.

4.3 Affine Invariant Multivariate Signs

Following Chakraborty et al. (1998), we define an affine invariant version of the multivariate sign statistic. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent and identically distributed d-variate random vectors. Let $\beta = \{i_0, i_1, \ldots, i_d\}$ be a subset of $d+1$ indices of $\{1, 2, .., n\}$ and define the $d \times d$ matrix $\mathbf{X}(\beta)$ with columns $\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{i_d} - \boldsymbol{X}_{i_0}$. Define multivariate sign statistic as

$$\boldsymbol{S}_n^{(\beta)} = \frac{1}{n} \sum_{i=1}^n \frac{\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)}{\|\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\|}. \tag{4.1}$$

This sign statistic actually calculates the average direction of the data vectors in a transformed space determined by coordinate axes $\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_0}, \ldots, \boldsymbol{X}_{i_d} - \boldsymbol{X}_{i_0}$. Note that for any nonsingular $d \times d$ matrix $\mathbf{A}$ and a d-dimensional vector $\boldsymbol{b}$, if we transform the observations $\boldsymbol{X}_i$ to $\boldsymbol{Y}_i = \mathbf{A}\boldsymbol{X}_i + \boldsymbol{b}$, the location parameter $\boldsymbol{\mu}_0$ transforms to $\mathbf{A}\boldsymbol{\mu}_0 + \boldsymbol{b}$. The transformation matrix $\mathbf{X}(\beta)$ is transformed to $\mathbf{A}\mathbf{X}(\beta)$ and $\boldsymbol{S}_n^{(\beta)}$ remains invariant,

that is

$$\boldsymbol{S}_n^{(\beta)}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) = \boldsymbol{S}_n^{(\beta)}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).$$

Let us now define

$$\mathbf{C}_n^{(\beta)} = \frac{1}{n} \sum_{i=1}^n \frac{\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top \{\mathbf{X}(\beta)^\top\}^{-1}}{\|\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\|^2}.$$

Then define the control statistic

$$W_n^{(\beta)} = n\{\boldsymbol{S}_n^{(\beta)}\}^\top \{\mathbf{C}_n^{(\beta)}\}^{-1} \{\boldsymbol{S}_n^{(\beta)}\}.$$

Now it is easy to observe that the statistic is invariant under affine transformations. Chakraborty et al. (1998) used the statistic $W_n^{(\beta)}$ to construct a test for the location vector in a multivariate one-sample problem. If the process is in-control, that is the underlying distribution of the process variables has location vector $\boldsymbol{\mu}_0$, the large sample distribution of the statistic $W_n^{(\beta)}$ is chi-squared with d degrees of freedom (Theorem 2.2 Chakraborty et al., 1998). So to achieve an in-control ARL of $1/\alpha$, we set the upper control limit as $\chi^2_{\alpha,d}$, $(1 - \alpha)$-th quantile of the chi-squared distribution with d degrees of freedom.

4.4 Selection of β

We observe that the construction of the control statistic depends on the choice of the transformation matrix $\mathbf{X}(\beta)$ or the set of indices β. The small sample performance as well as the behaviour of the statistic when the process is out of control depends heavily on the choice of the transformation matrix. So a natural question arises at this point is how to select β. We will try to provide a solution to that next.

Let $g(\boldsymbol{x})$ denotes the elliptically symmetric density $\{\det(\Sigma)\}^{-1/2} f(\boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{x})$, where

Σ is a $d \times d$ positive definite matrix and $f(\boldsymbol{x}^\top \boldsymbol{x})$ is a continuous spherically symmetric density around the origin $\mathbb{R}^d$. The $\boldsymbol{X}_i$'s are assumed to be i.i.d. observations with common elliptically symmetric density $g(\boldsymbol{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the location of elliptic symmetry of the distribution.

Theorem 4.4.1 *If the process is in-control, that is, $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, the large sample distribution of $W_n^{(\beta)}$ is chi-squared distribution with d degrees of freedom. If the process is out-of-control, that is, $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, we further assume that $\log f$ is twice differentiable almost everywhere (w.r.t. Lebesgue measure) on $\mathbb{R}^d$, then the conditional limiting distribution of $\sqrt{n} \boldsymbol{S}_n^{(\beta)}$ is normal with mean vector $\Lambda_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\beta))$ that depends on f, $\Sigma^{-1/2}\boldsymbol{\delta}$ and $\Sigma^{-1/2}\mathbf{X}(\beta)$, where $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{\mu}_0$. The limiting covariance matrix is $\Psi_1(\Sigma^{-1/2}\mathbf{X}(\beta))$ that depends on $\Sigma^{-1/2}\mathbf{X}(\beta)$. Also, the limiting conditional distribution of the control statistic $W_n^{(\beta)}$ is a non-central chi-squared distribution with d degrees of freedom and the limiting average run length monotonically decreases with the non-centrality parameter*

$$\phi_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\beta)) =$$

$$[\Lambda_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\beta))]^\top [\Psi_1(\Sigma^{-1/2}\mathbf{X}(\beta))]^{-1} [\Lambda_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\beta))],$$

where ϕ_1 is such that for any f, $\boldsymbol{\delta}$ and Σ and any two invertible matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\phi_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{A}) \geq \phi_1(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{B})$ whenever $\mathbf{B}^\top \Sigma^{-1}\mathbf{B} = \lambda \mathbf{I}_d$ for some $\lambda > 0$.

Theorem 4.4.1 follows immediately from Theorem 2.2 of Chakraborty et al. (1998). The main implication of the above theorem is that if we wish to minimize the average run length of the proposed chart for $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, the optimal choice of $\mathbf{X}(\beta)$ is obtained if $\{\mathbf{X}(\beta)\}^\top \Sigma^{-1}\mathbf{X}(\beta) = \lambda \mathbf{I}_d$ for some $\lambda > 0$. For this optimal choice of the transformation matrix, $\mathbf{X}(\beta)$, the non-centrality parameter becomes $c_1(f)\boldsymbol{\delta}^\top \Sigma^{-1}\boldsymbol{\delta}$, where the constant $c_1(f)$ depends on the spherically symmetric density f only.

In practice, Σ is unknown and we need to have a consistent estimator, $\hat{\Sigma}$ of Σ and we make the matrix $\{\mathbf{X}(\beta)\}^{\top}\hat{\Sigma}^{-1}\mathbf{X}(\beta)$ as close as possible to a diagonal matrix with all diagonal entries equal. Define

$$v(\beta) = \frac{\text{trace}[\{\mathbf{X}(\beta)\}^{\top}\Sigma^{-1}\mathbf{X}(\beta)]/d}{\{\det[\{\mathbf{X}(\beta)\}^{\top}\Sigma^{-1}\mathbf{X}(\beta)]\}^{1/d}}.$$

Note that $v(\beta)$ is the ration of the arithmetic and geometric mean of the eigenvalues of the matrix $\{\mathbf{X}(\beta)\}^{\top}\Sigma^{-1}\mathbf{X}(\beta)$ and hence

$$v(\beta) \geq 1.$$

$v(\beta) = 1$ implies all eigenvalues are equal and the matrix is a diagonal matrix with all diagonal entries equal. Thus to choose the optimal subset β, we minimize $v(\beta)$ over all subset of observations of size $d+1$. For practical implementation, we may stop once $v(\beta)$ is sufficiently close to 1. The consistent estimator $\hat{\Sigma}$ can be taken to be the sample variance-covariance matrix if the second moments exist or the MCD estimate of Σ as proposed by Rousseeuw and Leroy (1987). If we denote the choice of β by $\hat{\beta}$ which is obtained by minimizing $v(\beta)$, then by Theorem 3.3 of Chakraborty (2001), $v(\hat{\beta})$ converges in probability to 1 as $n \to \infty$ and the matrix $\mathbf{X}(\hat{\beta})\{\mathbf{X}(\hat{\beta})\}^{\top}$ converges in probability to a scalar multiple of Σ. For the rest of the thesis, we will omit the notation $\hat{\beta}$ to avoid the notational complexities and β will denote the optimal choice as discussed in this section.

4.5 Multivariate Signed Ranks

Generalizing univariate Wilcoxon's signed rank statistic to an affine invariant version of multivariate signed rank, we define the spatial signed-rank function as

$$\boldsymbol{Q}_n(\boldsymbol{x}) = \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{x} - \boldsymbol{X}_i)}{\|\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{x} - \boldsymbol{X}_i)\|} + \frac{\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{x} + \boldsymbol{X}_i - 2\boldsymbol{\mu}_0)}{\|\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{x} + \boldsymbol{X}_i - 2\boldsymbol{\mu}_0)\|} \right].$$

Note that $\boldsymbol{Q}_n(\boldsymbol{x})$ is the affine invariant spatial rank of $(\boldsymbol{x} - \boldsymbol{\mu}_0)$ with respect to $\boldsymbol{X}_1 - \boldsymbol{\mu}_0, \ldots, \boldsymbol{X}_n - \boldsymbol{\mu}_0$ and their reflections $\boldsymbol{\mu}_0 - \boldsymbol{X}_1, \ldots, \boldsymbol{\mu}_0 - \boldsymbol{X}_n$ (Möttönen and Oja, 1995). Define the statistic based on spatial signed rank function as

$$\boldsymbol{R}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Q}_n(\boldsymbol{X}_i).$$

Note that, it can be written in the form of a V-statistic [Serfling (1980), Chapter 5.1.2] as follows:

$$\boldsymbol{R}_n = \frac{1}{2n^2} \sum_i \sum_j \frac{\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i + \boldsymbol{X}_j - 2\boldsymbol{\mu}_0)}{\|\{\mathbf{X}(\beta)\}^{-1}(\boldsymbol{X}_i + \boldsymbol{X}_j - 2\boldsymbol{\mu}_0)\|},$$

and thus when the process is in-control, that is, the location vector $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\sqrt{n}\boldsymbol{R}_n$ is asymptotically normally distributed with mean vector 0 and a variance covariance matrix depending on the distribution F. To construct a distribution free control chart, define

$$\mathbf{D}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Q}_n(\boldsymbol{X}_i)\{\boldsymbol{Q}_n(\boldsymbol{X}_i)\}^{\top}$$

and the control statistic

$$Y_n = n\boldsymbol{R}_n^{\top}\mathbf{D}_n^{-1}\boldsymbol{R}_n.$$

Following Möttönen et al. (1997), the asymptotic distribution of the statistic Y_n is a chi-squared distribution with d degrees of freedom when $\boldsymbol{\mu} = \boldsymbol{\mu}_0$.

To illustrate the proposed control chart based on the statistic Y_n, we present an example with 100 simulated sample of size $n = 15$ each from the bivariate normal distribution with covariance matrix $\Sigma = \mathbf{I}_2$. The first 50 samples are with mean $(0, 0)^{\top}$ and the last 50 samples are with mean $(2, 0)$. Figure 4.3 shows that the proposed chart detects the shift in location quite easily for this example. The upper control limit $UCL = \chi^2_{2,0.005} = 10.5966$.

If the process is out-of-control, that is $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, the statistic Y_n will have a large sample distribution of non-central chi-squared with d degrees of freedom and non-centrality

67
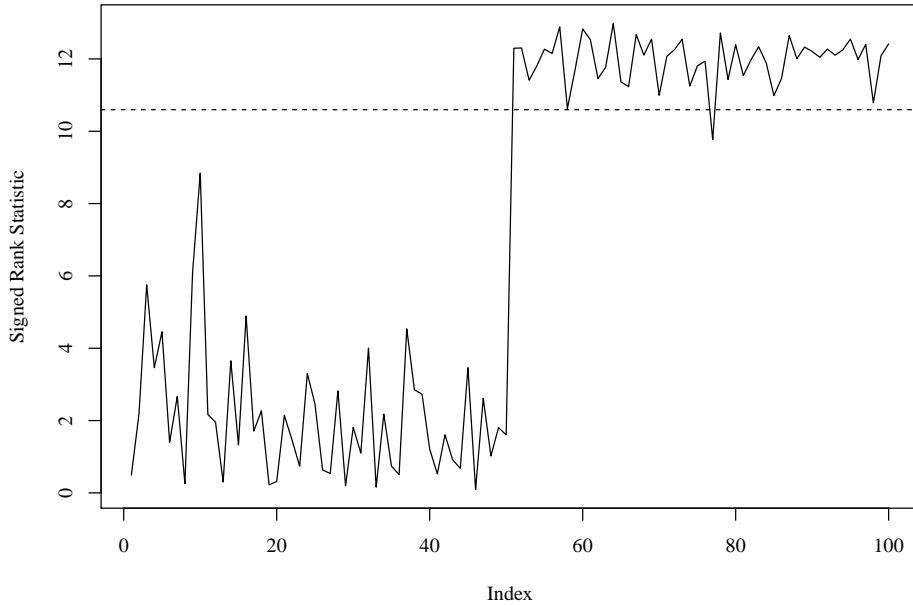
Figure 4.3: A Shewhart type control chart based on multivariate signed rank statistic.

parameter $\phi_2(f, \Sigma^{-1/2}\boldsymbol{\delta}, \Sigma^{-1/2}\mathbf{X}(\beta))$, which depends on the spherically symmetric density f, the scale matrix Σ, the shift $\boldsymbol{\delta} = \boldsymbol{\mu} - \boldsymbol{\mu}_0$ and the transformation matrix $\mathbf{X}(\beta)$ through $\Sigma^{-1/2}\boldsymbol{\delta}$ and $\Sigma^{-1/2}\mathbf{X}(\beta)$. If we select the transformation matrix $\mathbf{X}(\beta)$ in the way described in the previous section, the non-centrality parameter becomes $c_2(f)\boldsymbol{\delta}^\top \Sigma^{-1}\boldsymbol{\delta}$, for the optimal choice of β and $c_2(f)$ is a constant, which depends on the spherically symmetric density f only.

4.6 Performance Study Through Simulations

In this Section, we study the average run lengths of the proposed control charts through simulations. We have considered multivariate normal, Laplace and t distribution with 3 degrees of freedom. For all of these distributions, the location vector is taken to be $\boldsymbol{\mu}$, whose first element is δ and all other elements are 0, and the scale matrix Σ is taken to be the d dimensional identity matrix $\mathbf{I}_d$. When the process is in-control, the location vector

68

$\boldsymbol{\mu}$ is the vector of all zeros that is, $\delta = 0$. The simulation size is 1000.

Table 4.1: Average run length comparisons of the control charts based on multivariate signs for different sample sizes for normal distribution.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	10	6506.841	4268.670	2198.432	421.762	43.854
	20	394.581	251.673	127.921	13.284	1.732
	30	294.998	177.699	62.760	5.529	1.109
	50	259.290	111.367	34.787	2.150	1.003
	100	216.338	55.936	10.530	1.125	1.000
3	10	10806.069	8918.407	4889.837	806.761	88.846
	20	589.596	438.372	230.449	23.692	2.446
	30	372.276	234.015	93.671	7.143	1.186
	50	279.326	131.079	40.723	2.653	1.002
	100	224.678	69.784	12.948	1.164	1.000
5	10	14868.694	11360.185	6078.859	772.952	50.214
	20	3376.237	2306.215	1034.472	119.680	7.531
	30	692.013	501.990	216.290	16.642	1.472
	50	359.373	216.412	67.665	3.647	1.017
	100	265.213	97.691	19.229	1.252	1.000

First we consider the control chart based on multivariate sign statistic. We should note that the upper control limit of the control chart is based on the asymptotic distribution of the control statistic. To study the performance of the proposed chart for finite samples, we present some simulation studies in Tables 4.1, 4.2 and 4.3, which present the average run lengths of the control charts for different sample sizes and dimension. The large sample in-control average run length of these charts are 200, that is, the upper control limit is taken to be $\chi^2_{d,0.005}$. We observe that for small sample sizes, the in-control ARL is very large compared to 200, however, that slowly converges to 200 as the sample size increases. The performances are comparable for all three distributions indicating the distribution free nature of the multivariate sign based control statistic.

Next, we present ARL curves for the proposed control chart based on multivariate
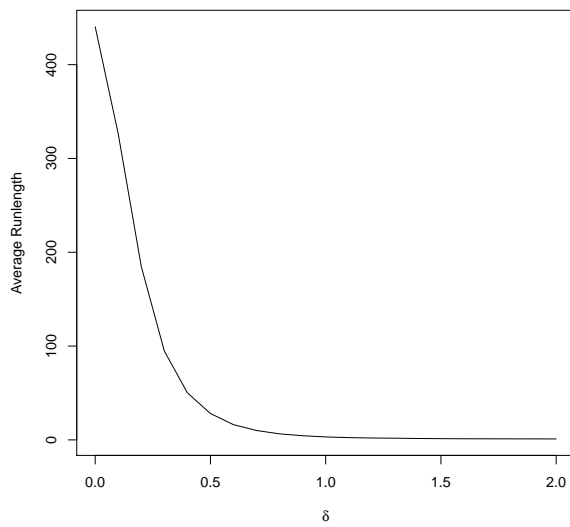
(a) Normal
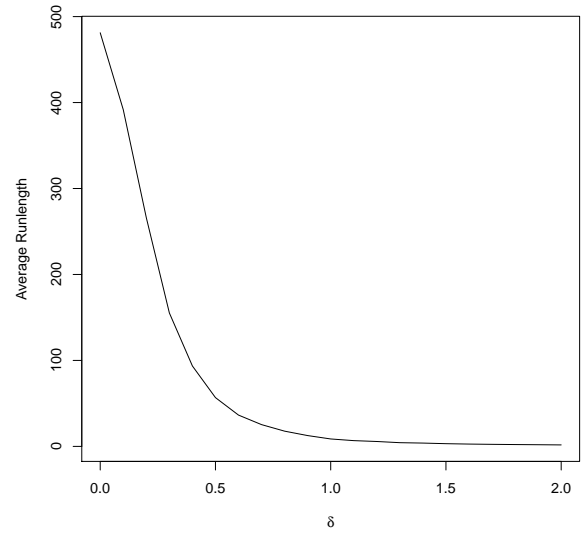
(b) Laplace
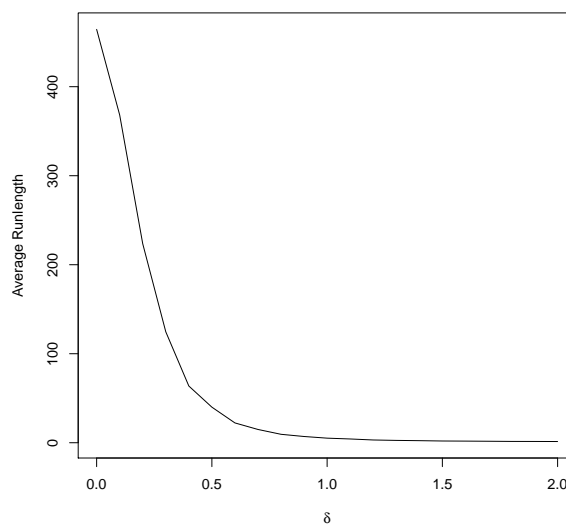
(c) t with 3 d.f.

Figure 4.4: ARL curves for control chart based on multivariate signs when the distribution of the process variables are (a) normal , (b) Laplace, and (c) t distributions with in-control ARL 200 for dimension $d = 2$, sample size $n = 15$.
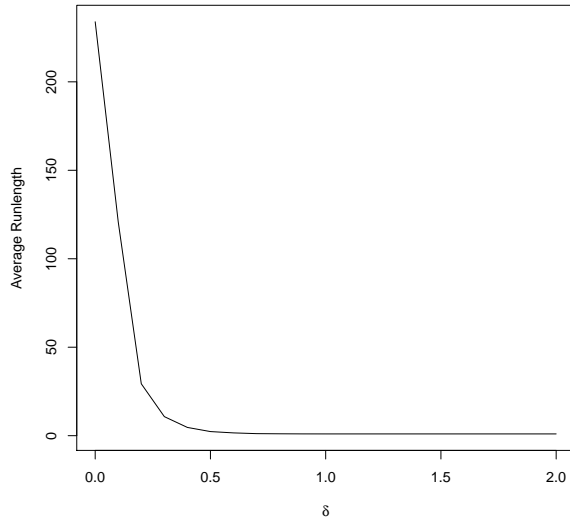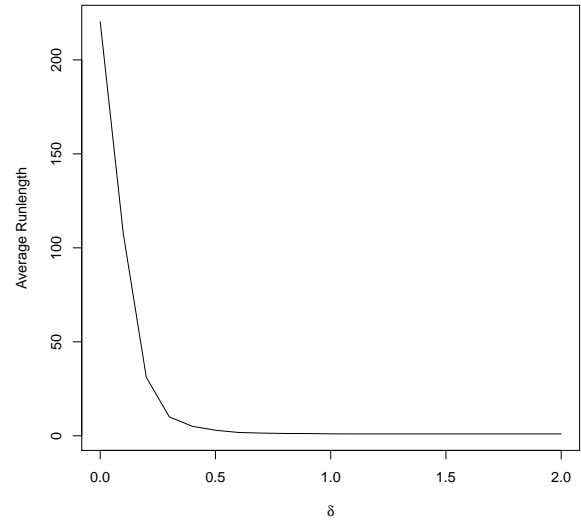
(a) Normal

(b) Laplace

(c) t with 3 d.f.

Figure 4.5: ARL curves for control chart based on multivariate signs when the distribution of the process variables are (a) normal , (b) Laplace, and (c) t distributions with in-control ARL 200 for dimension $d = 2$, sample size $n = 50$.

Table 4.2: Average run length comparisons of the control charts based on multivariate signs for different sample sizes for Laplace distribution.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	10	6321.941	5250.247	2671.932	552.867	84.670
	20	417.180	303.755	178.154	30.321	4.237
	30	281.740	204.406	100.676	12.848	1.991
	50	240.584	140.747	52.823	5.172	1.158
	100	221.350	86.398	20.409	1.817	1.002
3	10	10656.036	9190.211	7484.067	2450.889	476.826
	20	604.912	546.422	379.290	99.594	13.268
	30	361.631	322.274	205.148	34.255	4.194
	50	273.061	188.867	103.142	11.952	1.641
	100	221.482	130.415	44.182	3.261	1.030
5	10	15015.211	14175.920	11491.703	4668.530	974.680
	20	3263.290	2894.315	2481.027	899.455	148.573
	30	684.103	643.854	513.994	160.042	21.535
	50	385.757	316.153	216.626	46.741	4.962
	100	256.928	201.439	107.582	11.445	1.358

signs in Figure 4.4 and 4.5 with in-control ARL 200 and dimension $d = 2$ for sample sizes $n = 15$ and 50. We observe that for all three distributions, the proposed chart is good in detecting large shifts and its behaviour is very similar to the multivariate Shewhart type control chart based on T^2. As we have observed earlier that they fail to attain the in-control ARL of 200 due to the small sample size.

Now we consider the control chart based on multivariate signed rank statistic. We should note that the upper control limit of the control chart is based on the asymptotic distribution of the control statistic. To study the performance of the proposed chart for finite samples, we present some simulation studies in Tables 4.4, 4.5 and 4.6, which present the average run lengths of the control charts for different sample sizes and dimension. The large sample in-control average run length of these charts are 200, that is, the upper control limit is taken to be $\chi^2_{d,0.005}$. We observe that for small sample sizes, the in-control

Table 4.3: Average run length comparisons of the control charts based on multivariate signs for different sample sizes for t distribution with 3 degrees of freedom.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	10	6671.345	4672.954	2367.570	240.127	34.189
	20	402.280	279.389	143.506	17.688	2.505
	30	282.559	175.377	83.291	7.478	1.338
	30	282.559	175.377	83.291	7.478	1.338
	50	236.062	124.305	38.793	2.980	1.024
	100	219.345	66.180	13.863	1.245	1.000
3	10	10934.092	8633.842	5337.985	1037.985	117.232
	20	591.546	435.390	258.884	34.864	3.665
	30	366.525	253.831	114.498	10.103	1.551
	50	279.133	149.588	48.628	3.539	1.047
	100	238.345	79.189	17.669	1.331	1.000
5	10	15078.739	11085.290	6878.782	972.865	80.316
	20	3093.862	2496.209	1284.969	171.754	13.601
	30	727.364	504.221	253.292	24.169	2.241
	50	364.768	227.059	84.109	5.167	1.076
	100	271.926	107.970	24.291	1.566	1.000

ARL is very large compared to 200, however, that slowly converges to 200 as the sample size increases. The performances are comparable for all three distributions indicating the distribution free nature of the multivariate signed rank based control statistic.

In Figures 4.6 and 4.7, we present ARL curves for the proposed control chart based on multivariate signed ranks with in-control ARL 200 and dimension $d = 2$ for sample sizes $n = 15$ and 50. We observe that for all three distributions, the proposed chart is good in detecting large shifts and its behaviour is very similar to the multivariate Shewhart type control chart based on T^2. As we have observed earlier that they fail to attain the in-control ARL of 200 due to the small sample size.

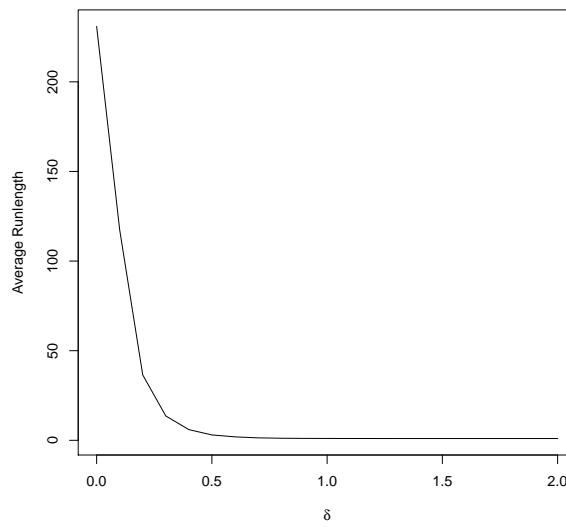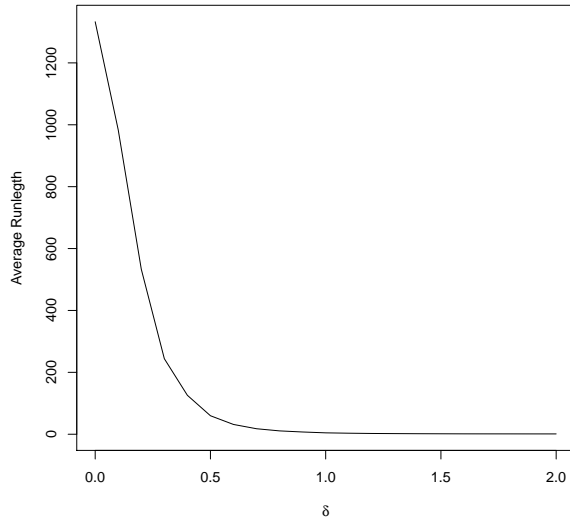(a) Normal
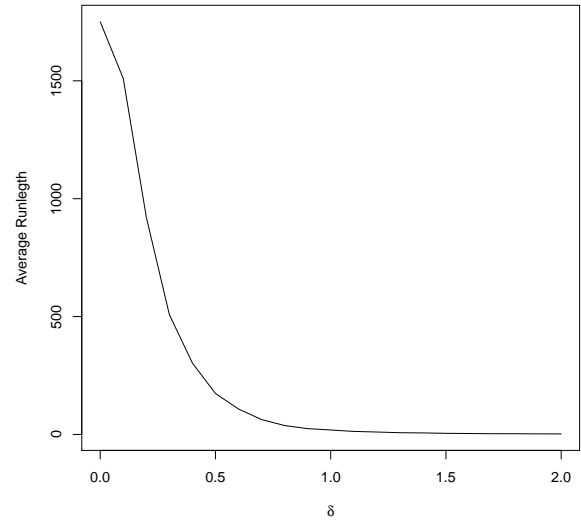
(b) Laplace
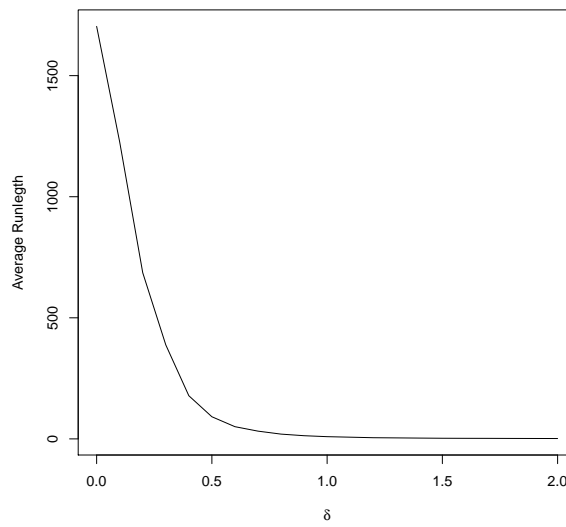
(c) t with 3 d.f.

Figure 4.6: ARL curves for control chart based on multivariate signed ranks when the distribution of the process variables are (a) normal , (b) Laplace, and (c) t distributions with in-control ARL 200 for dimension $d = 2$, sample size $n = 15$.
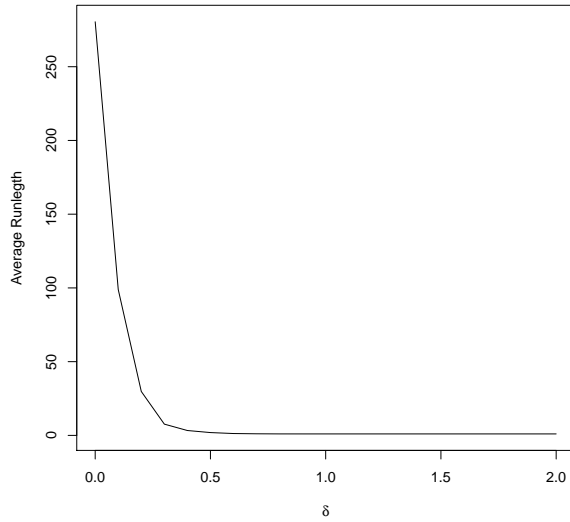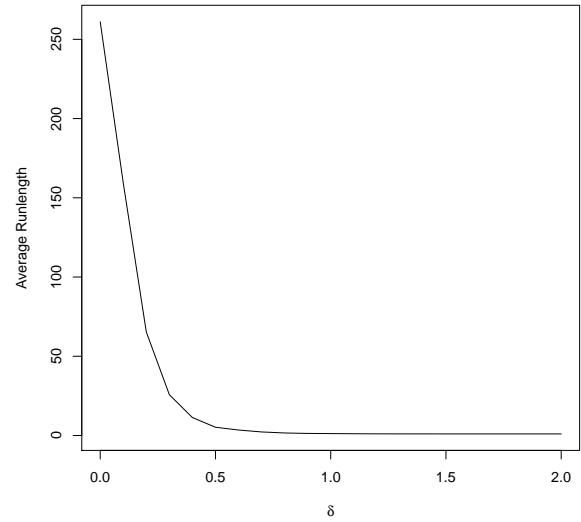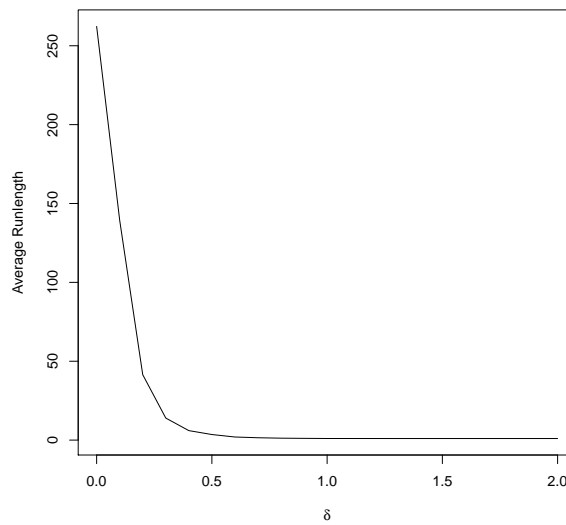
(a) Normal

(b) Laplace

(c) t with 3 d.f.

Figure 4.7: ARL curves for control chart based on multivariate signs when the distribution of the process variables are (a) normal , (b) Laplace, and (c) t distributions with in-control ARL 200 for dimension $d = 2$, sample size $n = 50$.

Table 4.4: Average run length comparisons of the control charts based on multivariate signed ranks for different sample sizes for normal distribution.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	20	561.646	345.323	154.593	13.898	1.611
	30	325.862	187.705	62.518	4.662	1.069
	50	282.118	111.254	26.224	1.800	1.000
	100	225.409	49.196	8.149	1.065	1.000
3	20	1027.371	691.027	340.631	31.885	2.533
	30	457.962	271.159	108.029	7.393	1.142
	50	301.106	146.573	37.929	2.283	1.001
	100	234.632	66.845	11.879	1.090	1.000
5	20	6941.433	5020.804	2165.903	236.959	15.432
	30	834.313	605.514	257.055	17.806	1.572
	50	403.584	233.175	67.178	3.264	1.008
	100	280.856	93.924	16.656	1.179	1.000

4.7 Concluding Remarks

In this Chapter, we have proposed two multivariate Shewhart type control charts based on multivariate sign vectors and signed ranks. Note that, for phase I control charts, we assume that the target location vector $\boldsymbol{\mu}_0$ is known and the scale matrix Σ is also known. In such a situation, we do not need to worry about affine invariance. If $\boldsymbol{X}$ is elliptically symmetric with scale (or, scatter) matrix Σ about $\boldsymbol{\mu}_0$, then $\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}_0)$ is spherically symmetric about $\boldsymbol{0}$. Thus if we know Σ and $\boldsymbol{\mu}_0$, we can transform the observations accordingly to have spherical symmetry and the non-invariant versions of multivariate sign and signed ranks will work with less computational complexity.

We also note that if the random vector $\boldsymbol{X}$ is spherically symmetric about $\boldsymbol{\mu}_0$, the sign vector $(\boldsymbol{X} - \boldsymbol{\mu}_0)/\|\boldsymbol{X} - \boldsymbol{\mu}_0\|$ is uniformly distributed over the d-dimensional unit sphere. Thus, the sign statistic $\boldsymbol{S}_n$, which is an average of such uniformly distributed random vectors, becomes distribution free when the process is in-control. We do not need

Table 4.5: Average run length comparisons of the control charts based on multivariate signed ranks for different sample sizes for Laplace distribution.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	20	630.941	454.212	267.688	45.096	5.248
	30	350.548	261.282	131.642	15.844	2.121
	50	257.197	164.291	59.562	5.866	1.177
	100	235.423	103.541	25.989	2.014	1.004
3	20	1156.574	1020.460	703.892	171.548	19.659
	30	469.069	397.777	250.303	44.154	4.780
	50	301.235	232.561	120.022	14.884	1.776
	100	247.797	140.854	50.334	3.910	1.041
5	20	7858.754	7625.387	6080.594	2143.015	379.533
	30	988.461	912.023	705.649	215.363	28.746
	50	432.388	385.723	258.028	53.416	5.559
	100	279.847	215.802	118.096	12.533	1.484

any large sample approximation for the distribution free nature of our proposed control statistic and that makes it quite useful for a large number of distributions. Similarly, signed ranks are also distribution free when the process is in-control. In practice, it may be useful to obtain the upper control limits of the proposed charts for finite small samples using simulations or other numerical methods and we do not have to rely on asymptotic distributions.

The affine invariant procedures suggested in this Chapter are useful for Phase II control charts or even for Phase I control charts where we know the target location vector $\boldsymbol{\mu}_0$, but the scale matrix Σ is unknown. We must admit that finding out the optimal transformation matrix $\mathbf{X}(\beta)$ is computationally intensive and for higher dimensions it may take a huge amount of time. For faster computational algorithms, one may see Chakraborty and Chaudhuri (2008). However, for an affine invariant version of the control statistic, there is an easy alternative. We may consider any consistent estimate $\hat{\Sigma}$ of Σ. For example, a sample variance covariance matrix if the second moments exist or minimum covariance

Table 4.6: Average run length comparisons of the control charts based on multivariate signed ranks for different sample sizes for t distribution with 3 degrees of freedom.

d	n	δ				
		0.0	0.1	0.2	0.5	1.0
2	20	612.789	415.654	207.848	24.137	3.036
	30	349.023	219.758	91.742	8.627	1.469
	50	275.530	135.984	41.056	3.073	1.037
	100	235.981	70.055	15.354	1.354	1.000
3	20	1129.102	916.146	467.576	56.020	5.145
	30	469.421	329.589	158.369	13.782	1.704
	50	313.192	179.256	60.029	4.275	1.084
	100	248.034	88.947	19.048	1.415	1.000
5	20	9127.768	6073.142	3359.564	458.021	34.343
	30	1098.144	774.117	397.974	35.967	2.919
	50	438.816	277.704	110.215	6.246	1.152
	100	275.436	127.288	30.188	1.717	1.001

determinant estimator (Rousseeuw and Van-Driessen, 1999) can be used and observations can be transformed using $\hat{\Sigma}^{-1/2}$. In this approach, the individual sign vectors or the statistics $\boldsymbol{S}_n$ and $\boldsymbol{R}_n$ are not invariant under affine transformations, but the control statistics W_n and Y_n are affine invariant. The transformation approach described in this Chapter are quite useful for general purpose as it produces affine invariant versions of multivariate signs and multivariate ranks.

The proposed control chart methods utilizes the idea behind the tests of location for a multivariate one sample problem. Their construction is similar to the Shewhart chart, which uses only the current sample and no past history of the process. As expected these procedures are good in detecting large shifts in the location, however, they cannot detect small shifts efficiently.

CHAPTER 5

MULTIVARIATE CUSUM CONTROL

CHARTS

5.1   Introduction

We have already mentioned that Shewhart type control charts as presented in Chapter 4, are not efficient in detecting small shifts in the location parameter. We know that univariate cumulative sum (CUSUM) control charts efficiently detects small shifts in the location parameter with a target in-control ARL. We can also design an univariate CUSUM control chart which is optimal in detecting a specified amount of shift. For univariate CUSUM control charts to detect shifts in the process means, the control statistics after the j-th sample is drawn is given by $T_j = \max(0, T_{j-1} + (X_j - \mu_0) - k\sigma)$, where μ_0 is the mean of the process variable when the process is in-control, σ^2 is the variance of the process variable X_j, $T_0 = 0$ and $k > 0$. The idea is to shrink the cumulative sum $T_{j-1} + (X_j - \mu_0)$ towards the origin by k standard deviations. If we wish to extend the similar definition to the multivariate case, there are some problems in choosing a vector valued $\boldsymbol{k}$ and also maximum is not uniquely defined. Crosier (1988) defined a multivariate CUSUM scheme by shrinking the vector $\boldsymbol{T}_{j-1} + (\boldsymbol{X}_j - \boldsymbol{\mu}_0)$ towards the origin in the same direction and

the shrinking is determined by covariance matrix Σ and a scalar parameter k. We have discussed the multivariate CUSUM scheme proposed by Crosier (1988) in detail in Chapter 2. We have observed that while the proposed scheme performs quite well in detecting small shifts in the process mean when the underlying distribution is multivariate normal, it behaves poorly when the underlying distribution deviate from normality. In this chapter, we propose two CUSUM control charts based on the notions of multivariate sign and signed ranks introduced in the previous chapter.

5.2 Multivariate CUSUM Charts Based on Sign Vectors

In the monitoring of sample processes, let us assume that every sample has size n. We consider a Phase I control chart, where the in-control location parameter $\boldsymbol{\mu}_0$ is known. Let the j-th sample be $\boldsymbol{X}_{1,j}, \ldots, \boldsymbol{X}_{n,j}$. Then the sign statistic for the j-th sample is given by

$$\boldsymbol{S}_{j,n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|}.$$

Let

$$\Sigma_1 = E\left[\left\{ \frac{\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|} \right\} \left\{ \frac{\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|} \right\}^{\top} \right].$$

Then define a multivariate CUSUM control scheme as follows:

- Let $\boldsymbol{T}_0 = \boldsymbol{0}$ and $k > 0$ is a predetermined constant.

- At the j-th stage, define $C_j = \sqrt{n(\boldsymbol{T}_{j-1} + \boldsymbol{S}_{j,n})^{\top} \Sigma_1^{-1} (\boldsymbol{T}_{j-1} + \boldsymbol{S}_{j,n})}$.

- Define
$$\boldsymbol{T}_j = \begin{cases} \boldsymbol{0} & \text{if } C_j \leq k \\ (\boldsymbol{T}_{j-1} + \boldsymbol{S}_{j,n})(1 - k/C_j) & \text{if } C_j > k. \end{cases}$$

- Define $Y_j = n\boldsymbol{T}_j^\top \Sigma_1^{-1} \boldsymbol{T}_j$. The process is said to be out of control if $Y_j > H$ for some $H > 0$.

Note that, C_j is the length of the cumulative sum of the sign vectors standardised by the scale matrix and when it is less than a threshold parameter k, we renew the cumulative sum to the zero vector and $Y_j = (C_j - k)^2$ after the last renewal. So k acts as an shrinkage parameter as well which shirks the the standardised length of the cusum vector towards zero. Larger values of k will shrink more and that will not be efficient to detect small shifts. On the other hand, small values of k will not shrink much and that will help to detect small shifts in the location efficiently.

To illustrate the proposed multivariate CUSUM control chart based on the statistic Y_j, we present an example with 100 simulated sample of size $n = 15$ each from the bivariate normal distribution with covariance matrix $\Sigma = \mathbf{I}_2$. The first 50 samples are with mean $(0, 0)^\top$ and the last 50 samples are with mean $(1.0, 0)^\top$. We have taken $k = 0.5$. Figure 5.1 shows that the proposed control statistic increases sharply when the process becomes out of control.

If the underlying distribution of the process vector is spherically symmetric around $\boldsymbol{\mu}_0$, the sign vectors $(\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0)/\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|$ are uniformly distributed over the unit sphere and do not depend on the actual distribution F. The covariance matrix of the sign vector is given by $\Sigma_1 = (1/d)\mathbf{I}_d$. If the underlying distribution is elliptically symmetric around $\boldsymbol{\mu}_0$ with a known scale matrix Σ, we can transform the vector of process variables by $\Sigma^{-1/2}$ and then the underlying distribution of the transformed vector becomes spherically symmetric and we can use the above procedure with $\Sigma_1 = (1/d)\mathbf{I}_d$.

An important task in designing this multivariate CUSUM control chart is determining the control limit H, which depends on the parameter k and the required in-control average run length of the proposed chart. As we have noted earlier that proposed scheme is distribution-free for elliptically symmetric distributions, we perform a small simulation
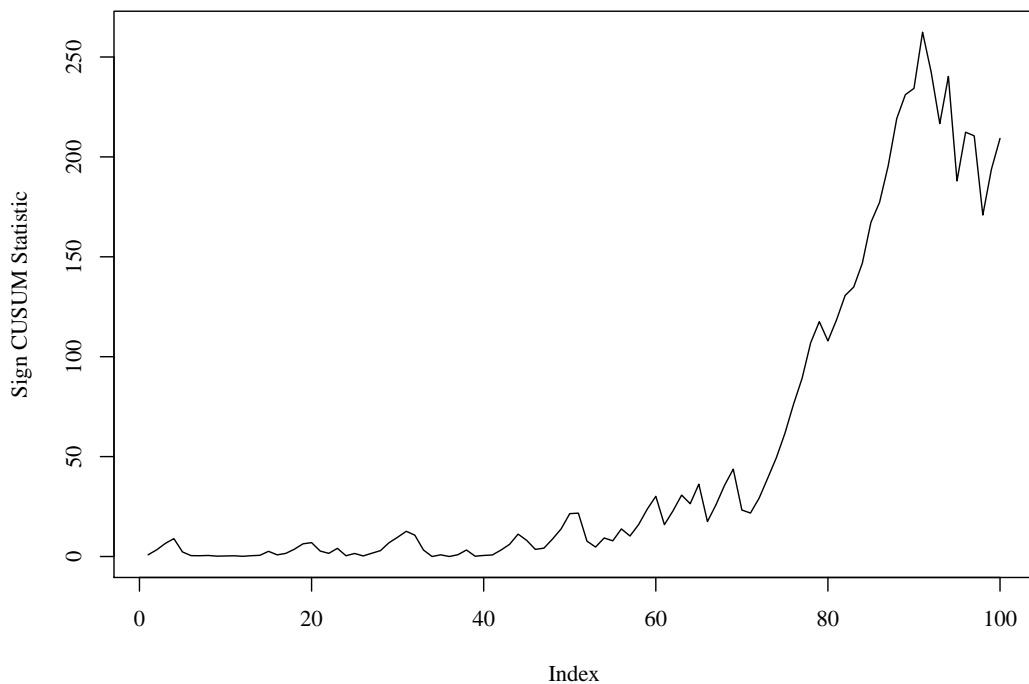
81

Figure 5.1: An example CUSUM control chart based on multivariate sign statistic.

study to determine the value of H for different dimensions d and parameter k. Figure 5.2 shows the average run lengths against control limit H, when the data is simulated from the multivariate normal distribution with in-control mean of the process as the zero vector and the covariance matrix $\mathbf{I}_d$. We observe that the value of the control limits decreases with increase in the parameter k.

From Figure 5.2, we summarise the control limits of the proposed CUSUM control chart in Table 5.1 when the average run lengths are 200 and 500. Using these control limits, we present a simulation study of average run lengths of the proposed CUSUM control chart for multivariate normal, Laplace and t distribution with 3 degrees of freedom in Figures 5.3, 5.4, 5.5. For all of these distributions, the location vector is taken to be $\boldsymbol{\mu}$, whose first element is δ and all other elements are 0, and the scale matrix Σ is taken to be

82

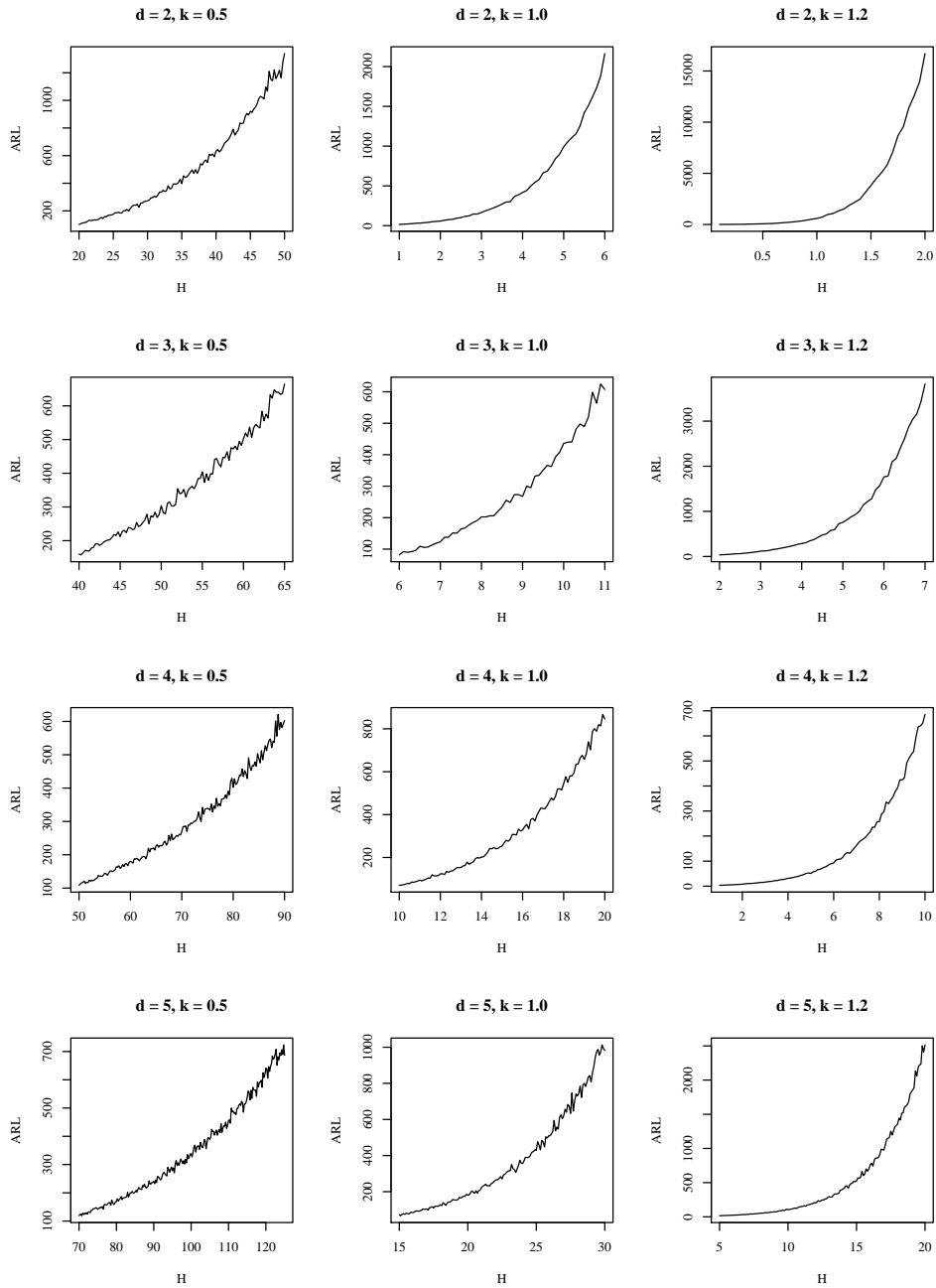Figure 5.2: Simulated average run length (ARL) of the proposed chart when the process is in-control against control limit H for different dimensions d and the parameter k for elliptically symmetric distributions.

the d dimensional identity matrix $\mathbf{I}_d$. When the process is in-control, we have $\delta = 0$. The simulation size is 1000. The control limits, H for different dimensions and different values

Table 5.1: Simulated values of the cutoff H for multivariate sign CUSUM charts with average run lengths 200 and 500 for different dimensions

|	ARL = 200			ARL = 500		
	k			k		
d	0.5	1.0	1.2	0.5	1.0	1.2
2	26.25	3.2	0.75	37.0	4.2	0.95
3	43.25	8.0	3.6	60.0	10.6	4.6
4	63.25	13.8	7.4	86.0	17.6	9.2
5	83.75	20.8	12.0	112.5	25.7	14.7

of k are shown in the Figures. We observe that the same control limit maintains the in-control ARL of 200 for all of these spherically symmetric distributions, thus showing the robust nature of the proposed procedure. We observe that for smaller values of k, the charts detect the small shifts in location quickly and for larger values they detect larger shifts efficiently. We also observe from the plots that the behaviour of the average run length for different dimensions are very similar for multivariate normal and t distribution when the process is out of control. However, the average run lengths for multivariate Laplace distribution are slightly higher for higher dimensions when the process is out of control.

Note that, if the scale matrix Σ is unknown when the underlying distribution is elliptically symmetric, we may replace Σ by any consistent estimator. For example, if the second moment exists, we may use the sample variance covariance matrix $\mathbf{S}$ as an estimator of Σ or we may use minimum covariance determinant estimator proposed by Rousseeuw and Van-Driessen (1999), which is a robust estimator of Σ. If the distribution is not elliptically symmetric, we do not usually have a simplified form for the matrix Σ_1 and that needs to be estimated either from the in-control data, or we may replace Σ_1 in the above scheme
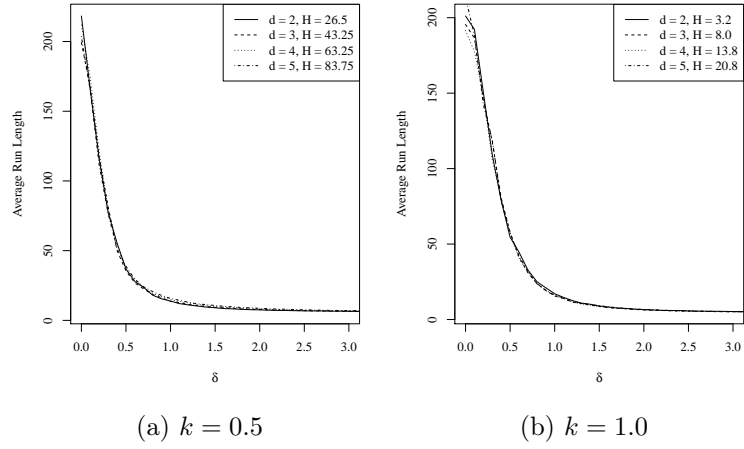
(a) $k = 0.5$ (b) $k = 1.0$
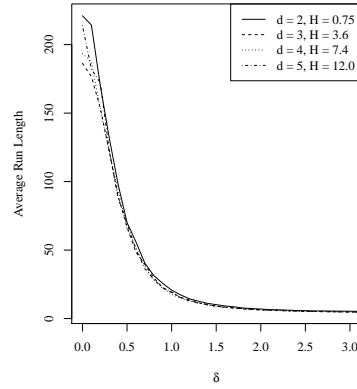
(c) $k = 1.2$

Figure 5.3: ARL curves for CUSUM control charts based on multivariate signs when the distribution of the process variables are multivariate normal with in-control ARL 200.

by its consistent estimator for the j-th sample

$$\mathbf{V}_j = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|} \right\} \left\{ \frac{\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} - \boldsymbol{\mu}_0\|} \right\}^{\top}.$$

For small n, the performance of the proposed scheme may be affected due to this estimation.
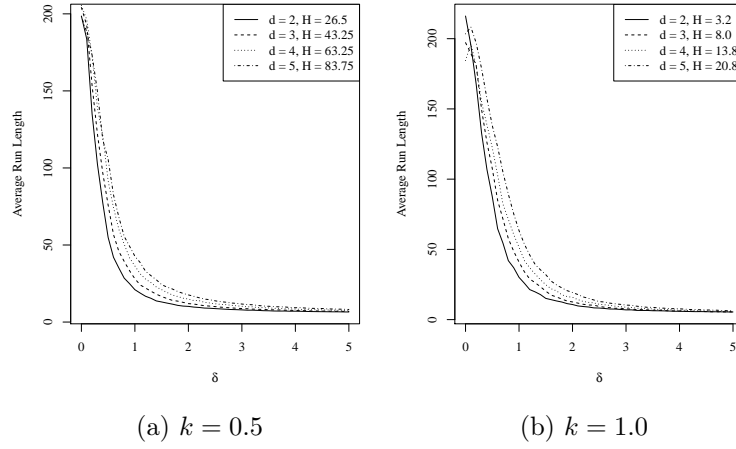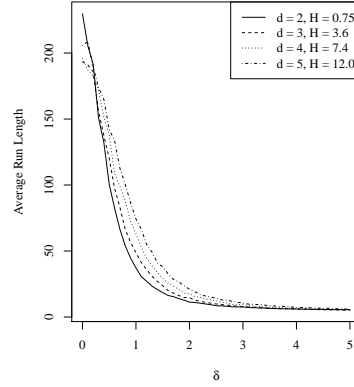
(a) $k = 0.5$

(b) $k = 1.0$

(c) $k = 1.2$

Figure 5.4: ARL curves for CUSUM control charts based on multivariate signs when the distribution of the process variables are multivariate Laplace with in-control ARL 200.

5.3   Multivariate CUSUM Control Charts Based on Signed Rank

Let the j-th sample be denoted by $\boldsymbol{X}_{1,j} \dots, \boldsymbol{X}_{n,j}$. Then we define the multivariate signed rank of a vector $\boldsymbol{x} \in \mathbb{R}^d$ with respect to this sample is

$$\boldsymbol{Q}_{j,n}(\boldsymbol{x}) = \frac{1}{2n} \sum_{i=1}^{n} \left\{ \frac{\boldsymbol{x} - \boldsymbol{X}_{i,j}}{\|\boldsymbol{x} - \boldsymbol{X}_{i,j}\|} + \frac{\boldsymbol{x} + \boldsymbol{X}_{i,j} - 2\boldsymbol{\mu}_0}{\|\boldsymbol{x} + \boldsymbol{X}_{i,j} - 2\boldsymbol{\mu}_0\|} \right\}$$
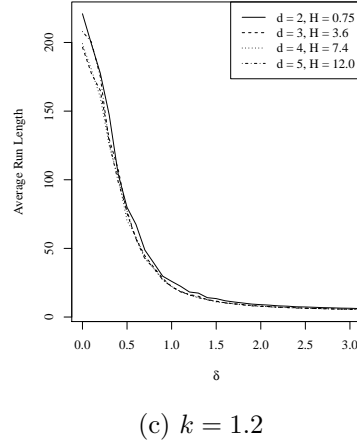
(a) $k = 0.5$

(b) $k = 1.0$

(c) $k = 1.2$
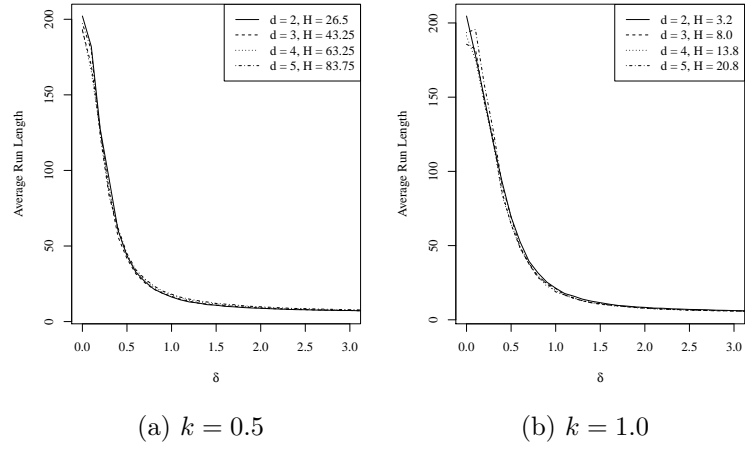
Figure 5.5: ARL curves for CUSUM control charts based on multivariate signs when the distribution of the process variables are multivariate t with 3 degrees of freedom with in-control ARL 200.

where $\boldsymbol{\mu}_0$ is the location vector when the process is in-control. We define the average signed rank of the j-th sample as

$$\boldsymbol{R}_{j,n} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Q}_{j,n}(\boldsymbol{X}_{i,j}) = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{l=1}^{n} \frac{\boldsymbol{X}_{i,j} + \boldsymbol{X}_{l,j} - 2\boldsymbol{\mu}_0}{\|\boldsymbol{X}_{i,j} + \boldsymbol{X}_{l,j} - 2\boldsymbol{\mu}_0\|}.$$

87

Now based on the signed rank statistics, we define a multivariate CUSUM control charting scheme as follows:

- Let $\boldsymbol{T}_0 = \boldsymbol{0}$ and $k > 0$ is a predetermined constant.

- Define a $d \times d$ matrix $\mathbf{B}_j$ as

$$\mathbf{B}_j = \frac{1}{nj} \sum_{l=1}^{j} \sum_{i=1}^{n} \boldsymbol{Q}_{l,n}(\boldsymbol{X}_{i,l})\{\boldsymbol{Q}_{l,n}(\boldsymbol{X}_{i,l})\}^{\top}.$$

- At the j-th stage, define $C_j = \sqrt{n(\boldsymbol{T}_{j-1} + \boldsymbol{R}_{j,n})^{\top} \mathbf{B}_j^{-1} (\boldsymbol{T}_{j-1} + \boldsymbol{R}_{j,n})}$.

- Define

$$\boldsymbol{T}_j = \begin{cases} \boldsymbol{0} & \text{if } C_j \le k \\ (\boldsymbol{T}_{j-1} + \boldsymbol{R}_{j,n})(1 - k/C_j) & \text{if } C_j > k. \end{cases}$$

- Define $Z_j = n\boldsymbol{T}_j^{\top} \mathbf{B}_j^{-1} \boldsymbol{T}_j$. The process is said to be out of control if $Z_j > H$ for some $H > 0$.

To determine the control limit H for the above multivariate control chart based on signed rank vectors, we run a small simulation study with for $n = 15$ and when the underlying in-control distribution of the process variables vector is multivariate normal with mean as the zero vector and the covariance matrix $\Sigma = \mathbf{I}_d$. Figure 5.6 presents the average run lengths against the control limit H for different dimension d and the parameter k. We summarise the control limits H for different dimensions to obtain average run lengths of 200 and 500 when the process is in control in Table 5.2. We again observe that the control limits increase with the dimension, but decrease with increasing values of the parameter k.

With the control limits obtained in Table 5.2, we present a simulation study of average run lengths of the proposed CUSUM control chart with multivariate signed rank vectors
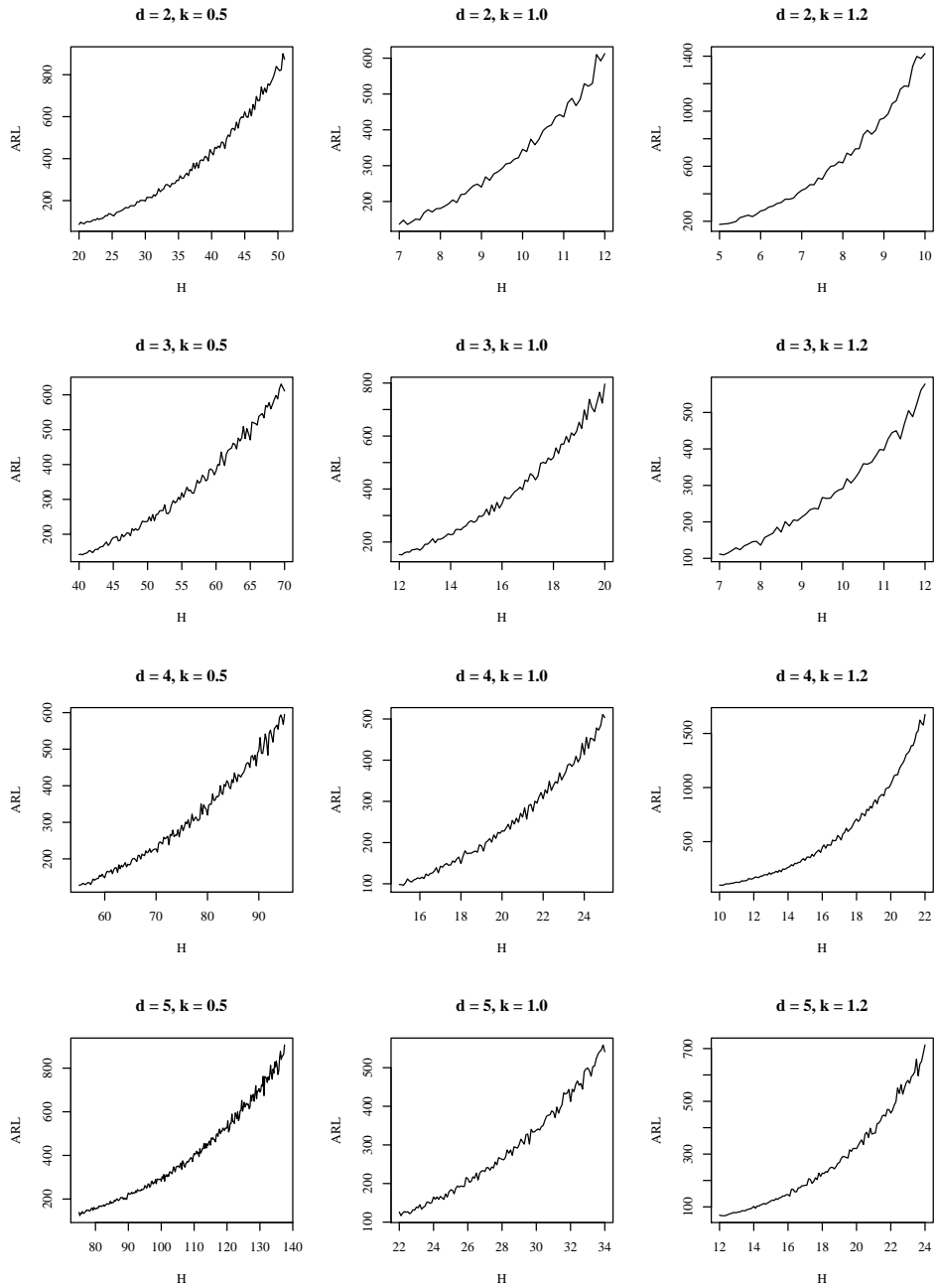
Figure 5.6: Simulated average run length (ARL) of the proposed chart when the process is in-control against control limit H for different dimensions d and the parameter k for elliptically symmetric distributions.

for multivariate normal, Laplace and t distribution with 3 degrees of freedom in Figures 5.7, 5.8, 5.9. For all of these distributions, the location vector is taken to be $\boldsymbol{\mu}$, whose

89

Table 5.2: Simulated values of the cutoff H for multivariate signed rank CUSUM charts with average run lengths 200 and 500 for different dimensions

| | ARL = 200 | | | ARL = 500 | | |
| | k | | | k | | |
d	0.5	1.0	1.2	0.5	1.0	1.2
2	29.5	8.3	5.4	42.25	11.45	7.4
3	47.0	13.2	8.6	64.5	17.6	11.6
4	65.5	19.2	13.0	90.0	24.9	16.6
5	86.5	25.9	17.2	117.25	33.3	22.3

first element is δ and all other elements are 0, and the scale matrix Σ is taken to be the d dimensional identity matrix $\mathbf{I}_d$. When the process is in-control, we have $\delta = 0$. The simulation size is 1000 and the sample size $n = 15$. The control limits, H for different dimensions and different values of k are shown in the Figures. We observe that the same control limit maintains the in-control ARL of 200 for all of these spherically symmetric distributions, thus showing the robust nature of the proposed procedure. We observe that for smaller values of k, the charts detect the small shifts in location quickly and for larger values they detect larger shifts efficiently. We also observe from the plots that the behaviour of the average run length for different dimensions are very similar for multivariate normal and t distribution when the process is out of control. However, the average run lengths for multivariate Laplace distribution are slightly higher for higher dimensions when the process is out of control.

5.4 Concluding Remarks

We have proposed two multivariate CUSUM control charts based on multivariate sign vectors and signed rank vectors. From the simulations studies of the average run lengths, we observe that both of these charts are quite efficient in detecting a small shift in location. The parameter k shrinks the multivariate CUSUM statistic towards the origin and hence
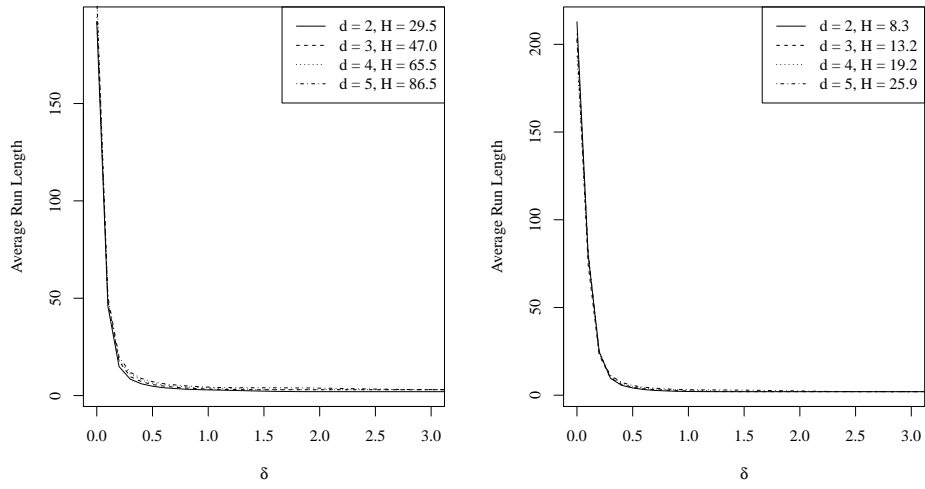
(a) $k = 0.5$

(b) $k = 1.0$
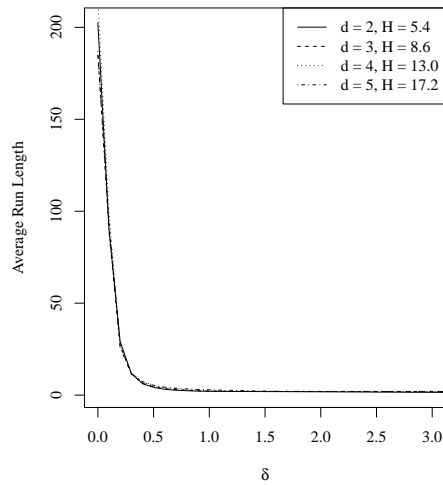
(c) $k = 1.2$

Figure 5.7: ARL curves for CUSUM control charts based on multivariate signed ranks when the distribution of the process variables are multivariate normal with in-control ARL 200.

determines the shift at which the chart is optimal. As a further study, we need to look into the theoretical relation between the parameter k and the location shift δ. That
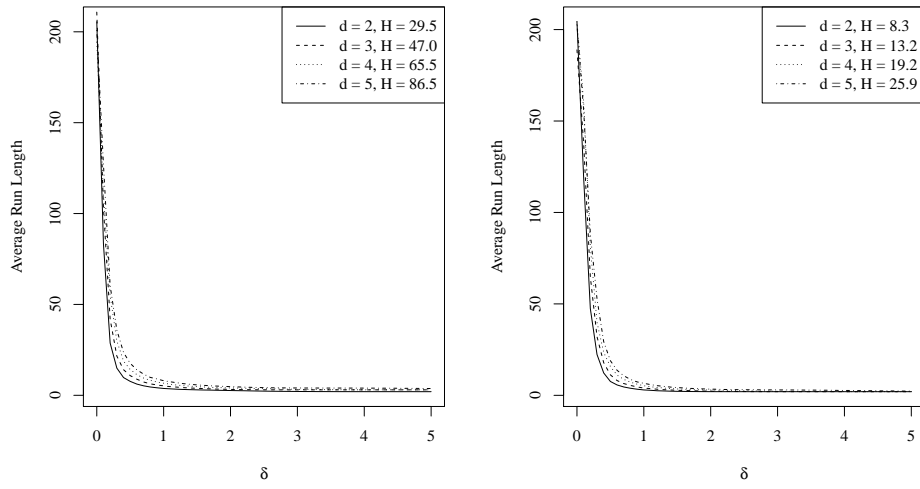
(a) $k = 0.5$

(b) $k = 1.0$

(c) $k = 1.2$

Figure 5.8: ARL curves for CUSUM control charts based on multivariate signed ranks when the distribution of the process variables are multivariate Laplace with in-control ARL 200.

result will provide a better understanding on the selection of k. The determination of the control limit H is based on simulations in our study. Though we observe that the control
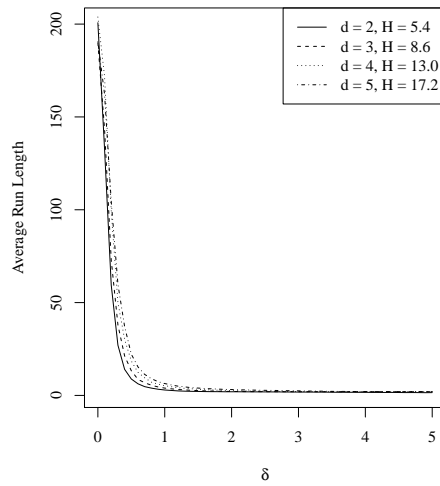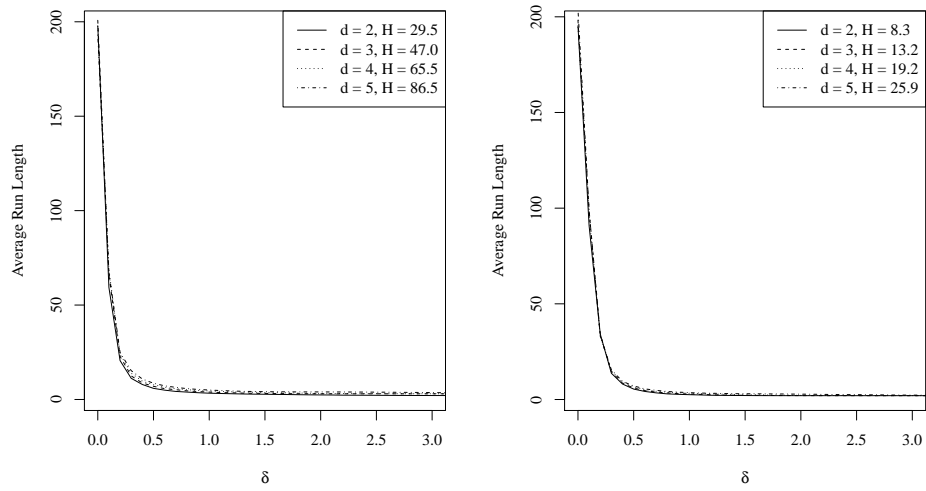
(a) $k = 0.5$
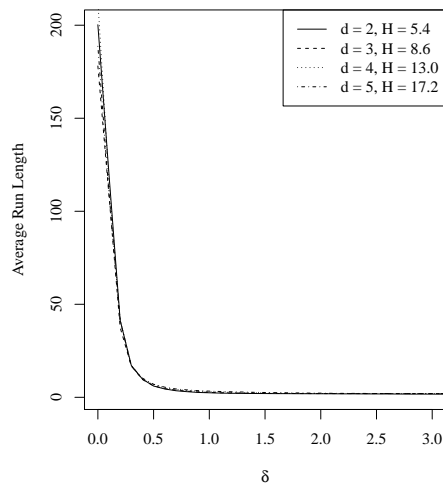
(b) $k = 1.0$

(c) $k = 1.2$

Figure 5.9: ARL curves for CUSUM control charts based on multivariate signed ranks when the distribution of the process variables are multivariate t with 3 degrees of freedom with in-control ARL 200.

limit thus obtained provides adequate results in our simulation studies, we still need to establish some theoretical results in determining the value of H.

In univariate CUSUM control charts, V-masks are used to make the charts more efficient in determining the small shifts quickly. We can construct similar V-masks for multivariate CUSUM control charts too, but we omit that construction in the present discussion and limit ourselves to the basic construction of a multivariate CUSUM chart. As we have seen before, multivariate CUSUM charts are not uniquely defined and there is a scope to explore other extensions of multivariate CUSUM charts as well.

CHAPTER 6

EXPONENTIALLY WEIGHTED MOVING AVERAGE CONTROL CHARTS

6.1 Introduction

The control chart based on exponentially weighted moving averages (EWMA) was defined at first by Roberts (1959) as $Z_0 = \theta_0$,

$$Z_k = r\bar{X}_k + (1-r)Z_{k-1}, \quad k > 0 \tag{6.1}$$

where θ_0 is the target location of the process variable when the process is in-control. Z_k represents a weighted average of all previous sample means and is located a fraction of smoothing parameter $0 < r < 1$ of the distance from Z_{k-1} to $\bar{X}_k$ on their connecting straight line, $0 < r < 1$ is some constant, and $\bar{X}_k$ is the average of the k-th sample for $k = 1, 2, \ldots$. It can be written as

$$
\begin{aligned}
Z_k &= (1-r)^2 Z_{k-2} + r(1-r)\bar{X}_{k-1} + r\bar{X}_k \\
&= (1-r)^k \theta_0 + \sum_{i=1}^{k} r(1-r)^{k-i}\bar{X}_i
\end{aligned}
$$

or, equivalently,

$$(Z_t - \theta_0) = \sum_{i=1}^{k} r(1-r)^{k-i}(\bar{X}_i - \theta_0).$$

It can be seen easily that the expected value of Z_k equals θ_0, while the expected values of the $\bar{X}_k$'s are all same as θ_0 if the process is in-control.

Following the above representation we can write the standard deviation of Z_k as

$$\sigma_{Z_k} = \sqrt{\frac{r}{2-r}[1 - (1-r)^{2k}]}\sigma_{\bar{X}}, \qquad (6.2)$$

where $\sigma_{\bar{X}}$ is standard deviation of the sample averages $\bar{X}$'s. As $k \to \infty$, the limiting value of the above standard deviation can be written as

$$\sigma_Z = \sqrt{\frac{r}{2-r}}\sigma_{\bar{X}}.$$

Roberts (1959) suggested control limits of $\mu_0 \pm 3\sigma_Z$.

Lowry et al. (1992) proposed a multivariate version of the EWMA control chart as a generalization of the univariate EWMA chart, which is sensitive in detecting a small shift in process mean. Let $\bar{X}_k$ be the sample mean of the d-dimensional k-th sample, $\{X_{k1}, \ldots, X_{kn}\}$, then define a d-dimensional vector

$$v_k = (1-r)v_{k-1} + r\bar{X}_k.$$

Let $\Sigma_{\bar{X}}$ be the covariance matrix of the k-th sample mean $\bar{X}_k$, then define the control statistic

$$T_k^2 = \frac{2-r}{r}v_k^T \Sigma_{\bar{X}}^{-1} v_k.$$

Here $r \in (0,1)$ is a smoothing parameter and the process is said to be out-of-control if $T_k^2 > H$, for some H, which is determined by the in-control ARL of the process.

Table 6.1: ARL Values for MEWMA Charts for Bivariate Normal Distribution

		r			
		0.2	0.4	0.6	0.8
δ	H	9.65	10.29	10.53	10.58
0.0		205.94	191.82	193.29	195.81
0.1		177.17	182.42	185.05	183.41
0.5		33.22	50.83	72.59	96.01
1.0		10.25	13.42	19.16	26.98
1.5		5.49	5.62	7.30	10.07
2.0		3.78	3.57	3.81	4.90

We can observe that T_k^2 converges in distribution to a chi-squared distribution with d degrees of freedom as $k \to \infty$. We can use the control limit H to be $\chi_{d,\alpha}^2$, the $(1 - \alpha)$-th quantile of the chi-squared distribution with d degrees of freedom. However, the performance of the proposed chart may not be optimal for small values of k. Lowry et al. (1992) obtained some values of H using simulations to attain an average run length of 200 when the process is in-control and has a bivariate normal distribution. In Table 6.1, we present some ARL values of the proposed chart using simulation size of 1000 from bivariate normal distribution with in-control mean vector to be the zero vector and the covariance matrix Σ is the identity matrix. When the process is out-of-control, the mean vector shifts to a vector whose first element is δ and rest of the elements are zero. We sample only one observation at every time point. Figure 6.1 shows detailed ARL curves for different values of r for the bivariate normal distribution. We can observe that this chart is quite efficient in detecting small shifts in the location vector. We also observe that smaller values of r are useful in detecting very small shifts whereas a bit larger values are better for detecting larger shifts. The choice of the smoothing parameter r depends on at what shift we want to design our control chart to be optimal. For a good discussion and some optimal values of r, see Lowry et al. (1992).

The control limits H were obtained for multivariate normal distributions, however,
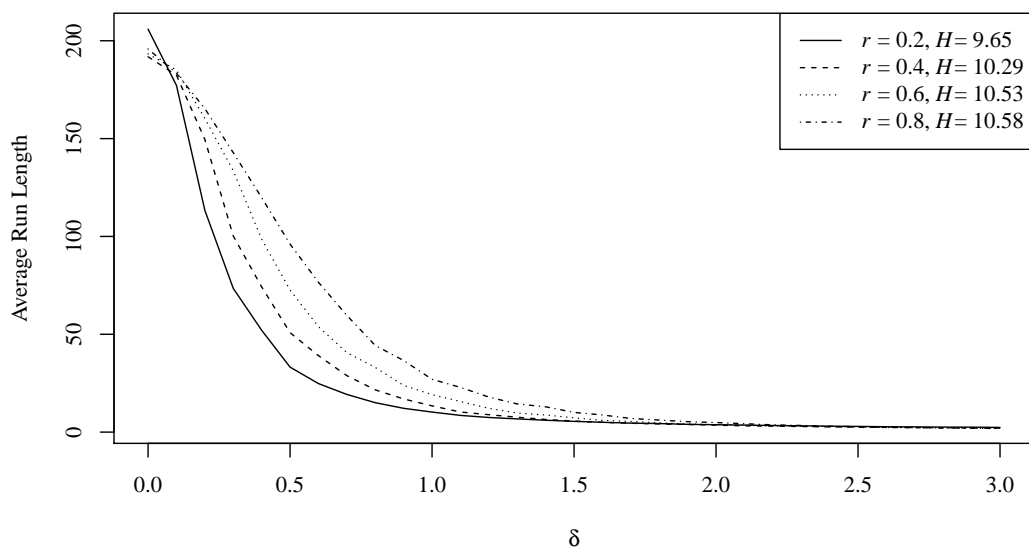
Figure 6.1: Average run lengths of multivariate EWMA control charts for bivariate normal distribution.

as with other procedures discussed before, the multivariate EWMA control charts break down when the in-control distribution departs from multivariate normality. In Figure 6.2, we present ARL curves in the similar setting with $r = 0.1$ for multivariate normal, Laplace and t distribution with 3 degrees of freedom in differen dimensions. The control limits, H, are determined using simulations for the multivariate normal distributions to attain an in-control ARL of 200. We observe that the ARL curves for the multivariate normal distributions are quite nice and as expected. However, the control charts with the same control limits even fail to attain the in-control ARL of 200 for multivariate Laplace and t distributions. This illustrates a lack of robustness of the multivariate EWMA chart based on sample averages. In this Chapter, we proposed two multivariate control charts based multivariate sign vectors and signed rank vectors and illustrate their distribution-free nature when the process variables have an elliptically symmetric distribution.
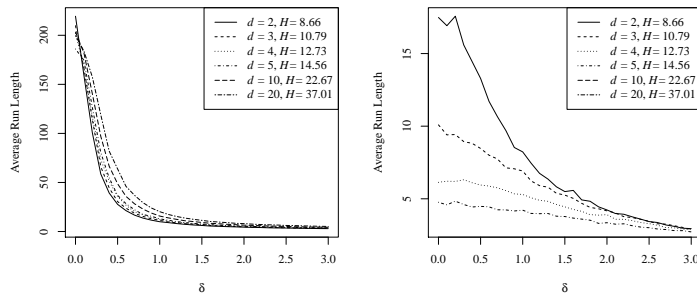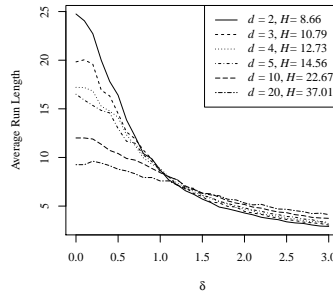
Figure 6.2: ARL curves for multivariate EWMA chart when the distribution of the process variable are (a) normal (b) Laplace, and (c) t distributions for dimensions $d = 2, 3, 4, 5, 10,$ and 20.

6.2 EWMA Charts Based on Sign Vectors

In this section we define a new exponentially weighted moving average control chart based on multivariate sign vectors defined earlier. Suppose that in-control location vector $\boldsymbol{\theta}_0$ is known. At each time point, we sample n observations and the k-th sample is denoted by $\{\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn}\}$. Then we propose a multivariate EWMA chart as follows:

- Define a d-dimensional vector $\boldsymbol{S}_0 = \boldsymbol{0}$.

- After the k-th sample $\{\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn}\}$ is observed, define

$$
\begin{aligned}
\boldsymbol{S}_k &= (1-r)\boldsymbol{S}_{k-1} + r\frac{1}{n}\sum_{i=1}^{n} Sign(\boldsymbol{X}_{ki} - \boldsymbol{\theta}_0) \\
&= (1-r)\boldsymbol{S}_{k-1} + \frac{r}{n}\sum_{i=1}^{n} \frac{\boldsymbol{X}_{ki} - \boldsymbol{\theta}_0}{\|\boldsymbol{X}_{ki} - \boldsymbol{\theta}_0\|}
\end{aligned}
$$

  where $r \in (0, 1)$ is a smoothing parameter.

- Define a $d \times d$ matrix as

$$
\begin{aligned}
\mathbf{C}_k &= \frac{1}{nk}\sum_{j=1}^{k}\sum_{i=1}^{n} Sign(\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0)\{Sign(\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0)\}^\top \\
&= \frac{1}{nk}\sum_{j=1}^{k}\sum_{i=1}^{n} \frac{(\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0)(\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0)^\top}{\|\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0\|^2}.
\end{aligned}
$$

- The control statistic is defined as $Y_k = [(2-r)n/r]\boldsymbol{S}_k^\top \mathbf{C}_k^{-1}\boldsymbol{S}_k$.

- The process is said to be out-of-control if $Y_k > H$, where H is a predetermined constant to attain an acceptable in-control ARL.

Theorem 6.2.1 *Let $\boldsymbol{X}_{ji}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ are independent and identically distributed with a d-dimensional distribution F. Then the variance-covariance matrix of the vector $\boldsymbol{S}_k$, defined above, is given by*

$$
\frac{r(1 - (1-r)^{2k})}{(2-r)n}\Sigma_1
$$

where

$$
\Sigma_1 = E\left[\frac{(\boldsymbol{X} - \boldsymbol{\theta}_0)(\boldsymbol{X} - \boldsymbol{\theta}_0)^\top}{\|\boldsymbol{X} - \boldsymbol{\theta}_0\|^2}\right]
$$

$\boldsymbol{X}$ having the distribution F. As $k \to \infty$, $Var(\boldsymbol{S}_k)$ converges to $\{r/((2-r)n)\}\Sigma_1$.

100

Proof. Let $\boldsymbol{Z}_k$ be the average sign vector for the k-th sample, that is,

$$\boldsymbol{Z}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_{ki} - \boldsymbol{\theta}_0}{\|\boldsymbol{X}_{ki} - \boldsymbol{\theta}_0\|}.$$

Then $\boldsymbol{S}_k$ can be written as

$$\begin{aligned}
\boldsymbol{S}_k &= (1-r)\boldsymbol{S}_{k=1} + r\boldsymbol{Z}_k \\
&= (1-r)^2 \boldsymbol{S}_{k-2} + r(1-r)\boldsymbol{Z}_{k-1} + r\boldsymbol{Z}_k \\
&= \sum_{j=1}^{k} r(1-r)^{k-j} \boldsymbol{Z}_j
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{S}_k) &= \sum_{j=1}^{k} r^2 (1-r)^{2(k-j)} \mathrm{Var}(\boldsymbol{Z}_j) \\
&= \frac{r(1-(1-r)^{2k})}{2-r} \mathrm{Var}(\boldsymbol{Z}_1) \\
&= \frac{r(1-(1-r)^{2k})}{(2-r)n} \Sigma_1.
\end{aligned}$$

As $k \to \infty$, $\mathrm{Var}(\boldsymbol{S}_k)$ converges to $\{r/((2-r)n)\}\Sigma_1$. $\qquad\square$

Note that if F is a spherically symmetric distribution around $\boldsymbol{\theta}_0$, then $\Sigma_1 = (1/d)\mathbf{I}_d$ and the variance covariance matrix of $\boldsymbol{S}_k$ is given by

$$\frac{r(1-(1-r)^{2k})}{(2-r)nd}\mathbf{I}_d \tag{6.3}$$

and asymptotically,

$$\frac{r}{(2-r)nd}\mathbf{I}_d. \tag{6.4}$$

Therefore, when the in-control distribution of the process variables are spherically symmetric around $\boldsymbol{\theta}_0$, we can define a simplified EWMA control chart statistic as

$$Y_k^* = \frac{(2-r)nd}{r}\boldsymbol{S}_k^\top \boldsymbol{S}_k,$$

which will have a chi-squared distribution with d degrees of freedom as $k \to \infty$. Note that, the distribution of $(\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0)/\|\boldsymbol{X}_{ji} - \boldsymbol{\theta}_0\|$ is uniform over the d-dimensional unit sphere and does not depend on the distribution of $\boldsymbol{X}_{ji}$. Thus, the statistic Y_k or Y_k^* are distribution-free even in finite samples.

If the observations are from elliptically symmetric distributions with a known scale matrix Σ, we can transform the vector of process variables by $\Sigma^{-1/2}$ to have spherically symmetric distributions and again use the control statistic Y_k^* as discussed above with the transformed variables. However, if the scale matrix Σ is unknown or the underlying distribution is not elliptically symmetric, Σ_1 may not be known or may not have a simple form as above. In such a situation, we plug in a consistent estimator, $\mathbf{C}_k$, of Σ_1 based on all $n \times k$ observations obtained until the k-th sample to construct the control statistic Y_k. Thus in the construction of our proposed EWMA chart, we do not need to make any specific distributional assumptions other than some regularity conditions for the large sample results to hold. It is easy to observe that, as $k \to \infty$, Y_k converges to a chi-squared distribution with d degrees of freedom and we can use a large sample upper control limit of $UCL = \chi_{d,\alpha}^2$ to have an in-control ARL of $1/\alpha$.

To illustrate the proposed control chart based on the statistic Y_k, we present an example with 100 simulated sample of size $n = 15$ each from the bivariate normal distribution with covariance matrix $\Sigma = \mathbf{I}_2$. The first 50 samples are with mean $(0,0)^\top$ and the last 50 samples are with mean $(0.5,0)^\top$. We have taken $r = 0.2$. Figure 6.3 shows that the proposed chart detects this small shift in location quite easily for this example. The upper

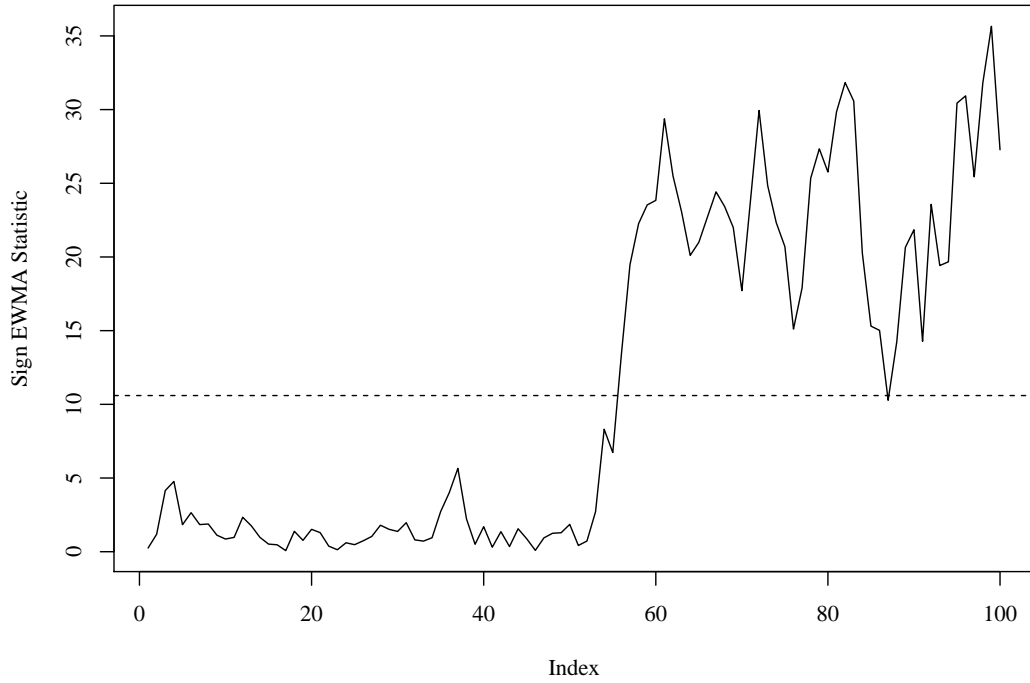control limit is taken as $UCL = \chi^2_{2,0.005} = 10.5966$.
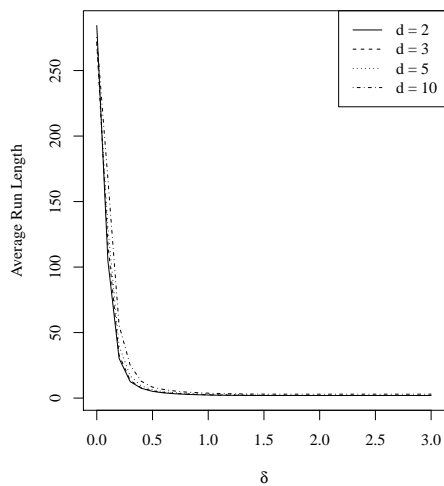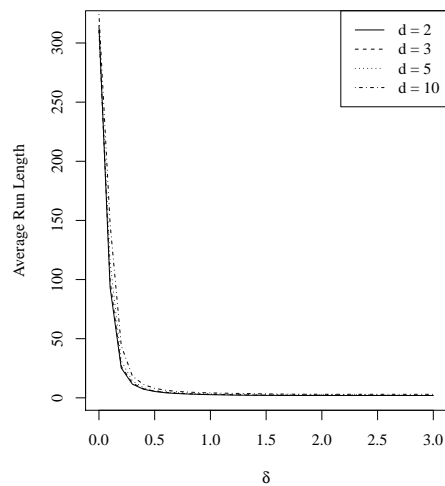

Figure 6.3: An example EWMA control chart based on multivariate sign statistic.

Now we present a study on the average run lengths of the proposed control charts through simulations. We have considered multivariate normal, Laplace and t distribution with 3 degrees of freedom. For all of these distributions, the location vector is taken to be $\boldsymbol{\theta}$, whose first element is δ and all other elements are 0, and the scale matrix Σ is taken to be the d dimensional identity matrix $\mathbf{I}_d$. When the process is in-control, the location vector $\boldsymbol{\theta}$ is the vector of all zeros that is, $\delta = 0$. The simulation size is 1000. The upper control limit is taken to be $\chi^2_{d,0.005}$ to have an asymptotic in-control ARL of 200. The ARL curves are presented in Figures 6.4, 6.5, 6.6 for different dimensions d and the smoothing parameter r. We observe that behaviour of the average run lengths with the shift in location are very similar for these 3 spherically symmetric distributions,

(a) $r = 0.3$

(b) $r = 0.2$
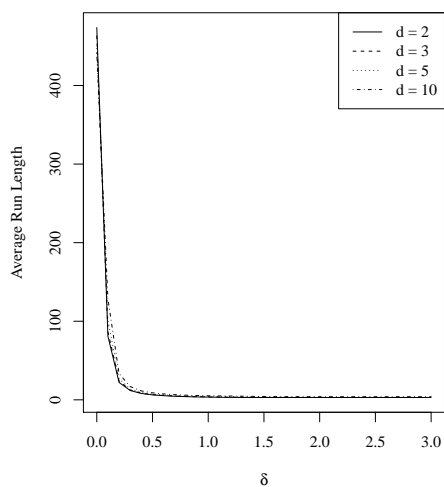
(c) $r = 0.1$

(d) $r = 0.05$

Figure 6.4: ARL curves for EWMA control charts based on multivariate signs when the distribution of the process variables are multivariate normal with in-control ARL 200 and sample size $n = 15$.
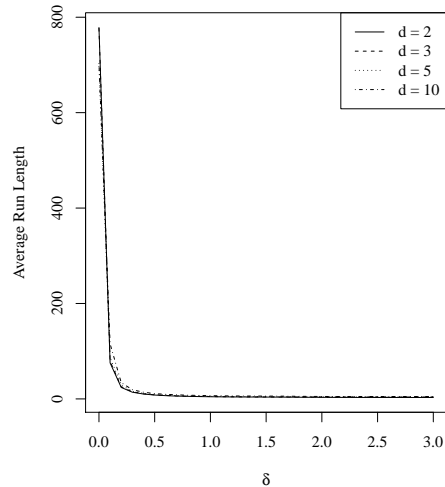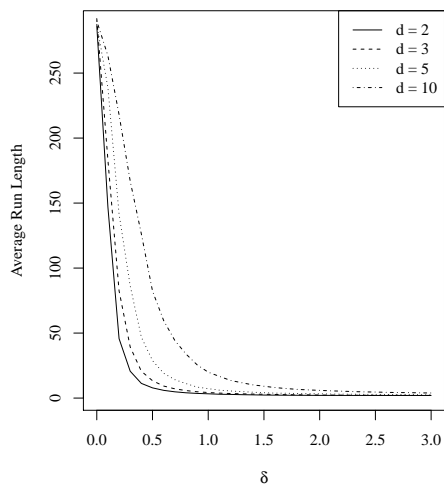
(a) $r = 0.3$
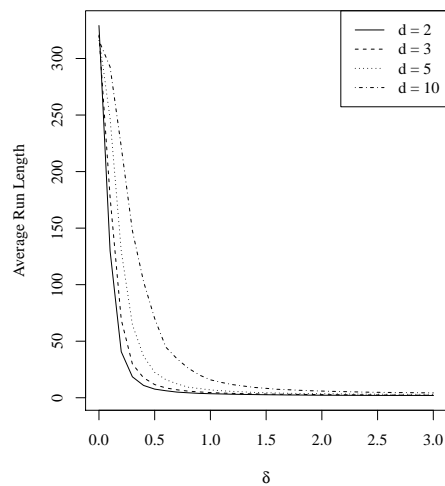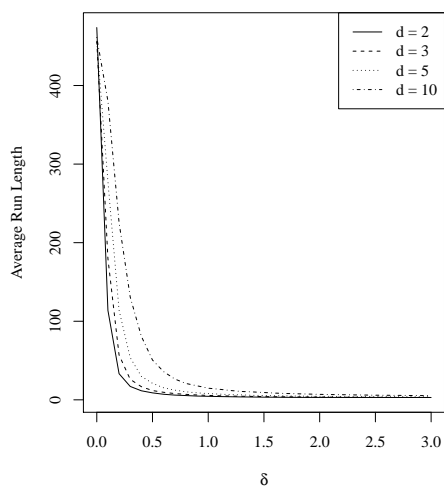
(b) $r = 0.2$

(c) $r = 0.1$

(d) $r = 0.05$

Figure 6.5: ARL curves for EWMA control charts based on multivariate signs when the distribution of the process variables are multivariate Laplace with in-control ARL 200 and sample size $n = 15$.
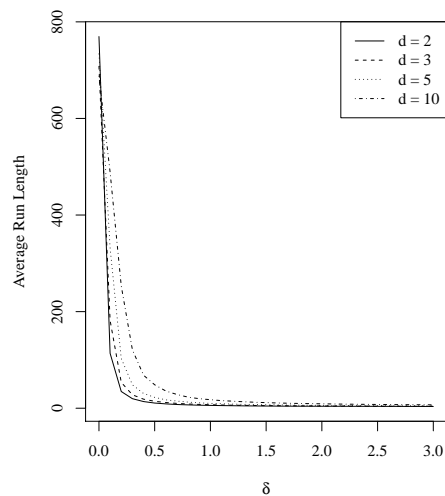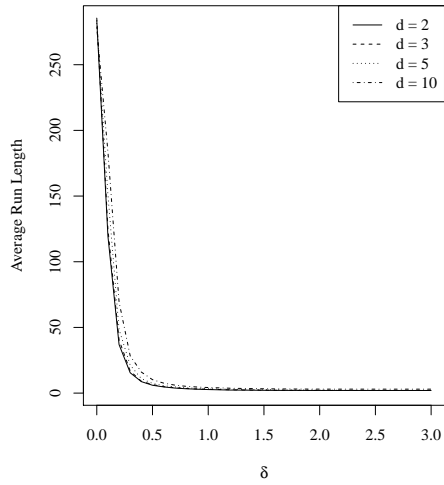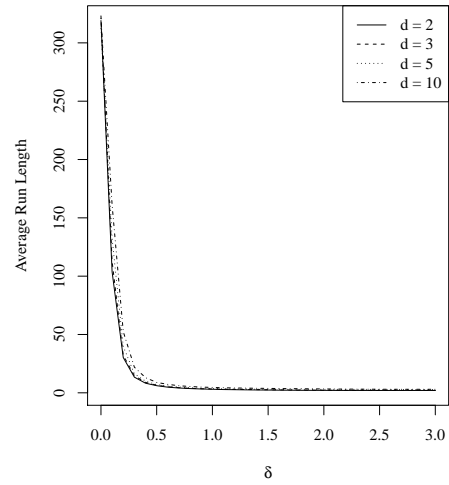
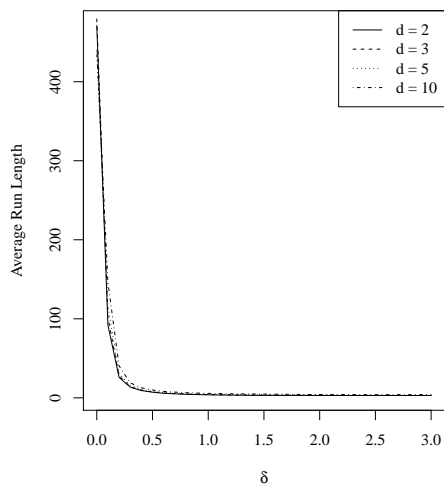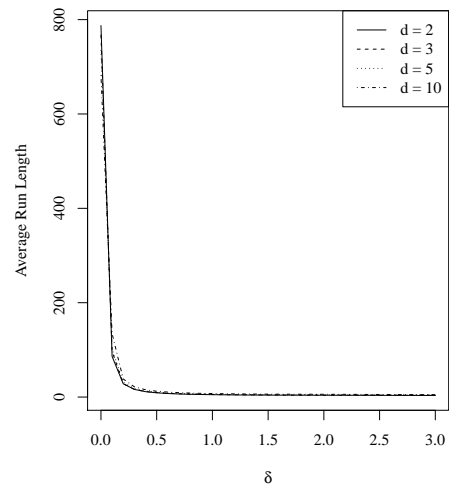Figure 6.6: ARL curves for EWMA control charts based on multivariate signs when the distribution of the process variables are multivariate t with 3 degrees of freedom with in-control ARL 200 and sample size $n = 15$.

which illustrates the distribution-free nature of the proposed method. We observe that the asymptotic upper control limit works best with $r = 0.3$ to attain the in-control ARL of 200. As r decreases, the in-control ARLs increase, however, all of these figures suggest that the proposed control charts are quite efficient in detecting small shifts in location.

6.3 EWMA Charts Based on Signed Rank Vectors

Now we consider the signed rank vectors to propose another multivariate EWMA control chart. Suppose that in-control location vector $\boldsymbol{\theta}_0$ is known. At each time point, we sample n observations and the k-th sample is denoted by $\{\boldsymbol{X}_{1,k}, \ldots, \boldsymbol{X}_{1,n}\}$. Define the signed rank of a vector $\boldsymbol{x} \in \mathbb{R}^d$ with respect to the k-th sample as

$$
\boldsymbol{Q}_n(\boldsymbol{x}) = \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{\boldsymbol{x} - \boldsymbol{X}_{i,k}}{\|\boldsymbol{x} - \boldsymbol{X}_{i,k}\|} + \frac{\boldsymbol{x} + \boldsymbol{X}_{i,k} - 2\boldsymbol{\theta}_0}{\|\boldsymbol{x} + \boldsymbol{X}_{i,k} - 2\boldsymbol{\theta}_0\|} \right].
$$

Then we propose a multivariate EWMA chart as follows:

- Define a d-dimensional vector $\boldsymbol{R}_0 = \boldsymbol{0}$.

- After the k-th sample $\{\boldsymbol{X}_{1,k}, \ldots, \boldsymbol{X}_{n,k}\}$ is observed, define

$$
\begin{aligned}
\boldsymbol{R}_k &= (1-r)\boldsymbol{R}_{k-1} + r\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_n(\boldsymbol{X}_{i,k}) \\
&= (1-r)\boldsymbol{R}_{k-1} + \frac{r}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\boldsymbol{X}_{i,k} + \boldsymbol{X}_{j,k} - 2\boldsymbol{\theta}_0}{\|\boldsymbol{X}_{i,k} + \boldsymbol{X}_{j,k} - 2\boldsymbol{\theta}_0\|}
\end{aligned}
$$

where $r \in (0,1)$ is a smoothing parameter.

- Define a $d \times d$ matrix as

$$
\mathbf{B}_k = \frac{1}{nk}\sum_{j=1}^{k}\sum_{i=1}^{n} \boldsymbol{Q}_n(\boldsymbol{X}_{i,j})\{\boldsymbol{Q}_n(\boldsymbol{X}_{i,j})\}^{\top}.
$$

107

- The control statistic is defined as $W_k = [(2-r)n/r]\boldsymbol{R}_k^\top \mathbf{B}_k^{-1} \boldsymbol{R}_k$.

- The process is said to be out-of-control if $W_k > H$, where H is a predetermined constant to attain an acceptable in-control ARL.

Let us define the conditional expectation,

$$\boldsymbol{U}^{(2)}(\boldsymbol{X}_1) = E\left[\frac{\boldsymbol{X}_1 + \boldsymbol{X}_2 - 2\boldsymbol{\theta}_0}{\|\boldsymbol{X} + \boldsymbol{X}_2 - 2\boldsymbol{\theta}_0\|}|\boldsymbol{X}_1\right].$$

Define

$$\Sigma_2 = E(\boldsymbol{U}^{(2)}(\boldsymbol{X}_1)\{\boldsymbol{U}^{(2)}(\boldsymbol{X}_1)\}^\top).$$

Then we have the following Theorem on the variance covariance matrix of the vector $\boldsymbol{R}_k$ defined above.

Theorem 6.3.1 *Let $\boldsymbol{X}_{i,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ are independent and identically distributed with a d-dimensional distribution F. Then the variance-covariance matrix of the vector $\boldsymbol{R}_k$, defined above, is given by*

$$\frac{r(1-(1-r)^{2k})}{(2-r)n}\Sigma_2.$$

As $k \to \infty$, $Var(\boldsymbol{R}_k)$ converges to $\{r/((2-r)n)\}\Sigma_2$.

The derivation of Σ_2 follows from Chaudhuri (1992) and the proof of the Theorem is almost identical to the proof of Theorem 6.2.1. Some examples and simplified versions of Σ_2 are derived in Chaudhuri (1992), however, except for multivariate normal distributions, derivation of Σ_2 is quite complicated and we use a consistent estimator $\mathbf{B}_k$ of Σ_2 to construct the control statistic W_k. Following Möttönen et al. (1997), it is easy to observe that W_k has a limiting chi-squared distribution with d degrees of freedom as $k \to \infty$,

when the process is in-control. Thus we can use a large sample upper control limit of $UCL = \chi^2_{d,\alpha}$ to have an in-control ARL of $1/\alpha$.

To illustrate the proposed control chart based on the statistic W_k, we present an example with 100 simulated sample of size $n = 15$ each from the bivariate normal distribution with covariance matrix $\Sigma = \mathbf{I}_2$. The first 50 samples are with mean $(0,0)^\top$ and the last 50 samples are with mean $(0.5, 0)^\top$. Figure 6.7 shows that the proposed chart detects this small shift in location quite easily for this example. The upper control limit is taken as $UCL = \chi^2_{2,0.005} = 10.5966$.
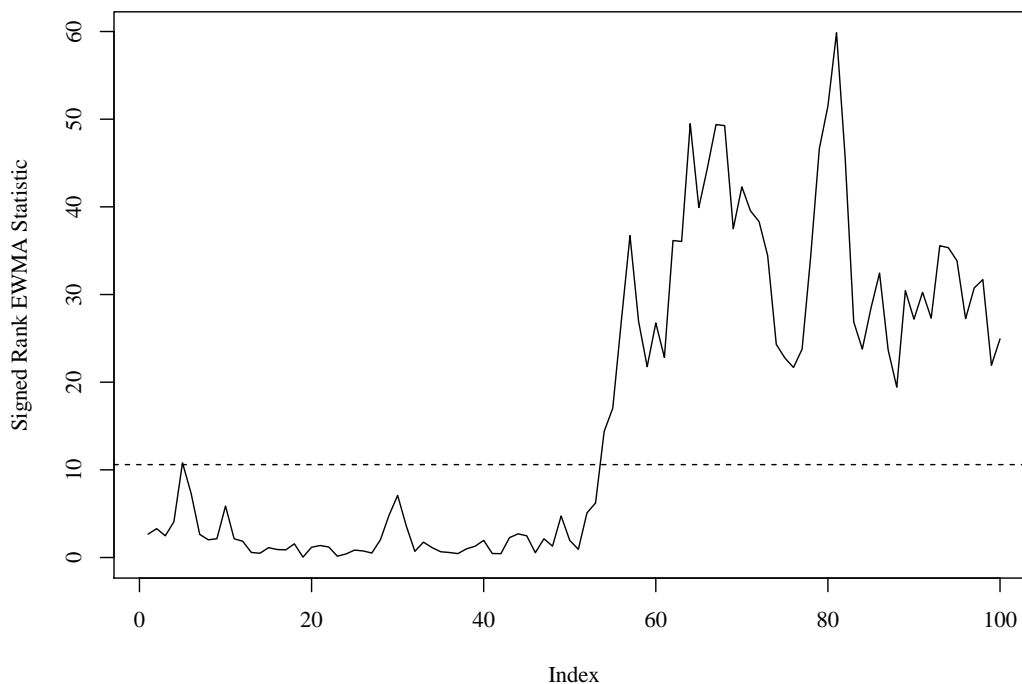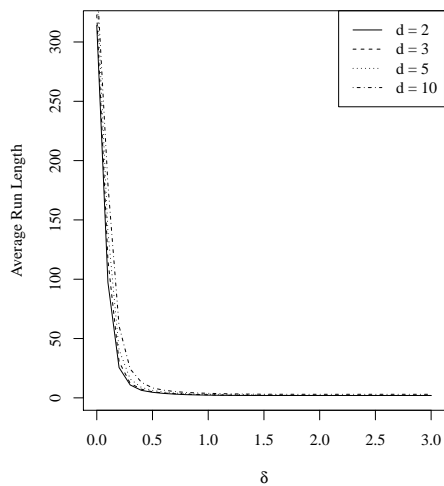

Figure 6.7: An example EWMA control chart based on multivariate signed rank statistic.

Now we present a study on the average run lengths of the proposed control charts through simulations. We have considered multivariate normal, Laplace and t distribution with 3 degrees of freedom. For all of these distributions, the location vector is taken to

be $\boldsymbol{\theta}$, whose first element is δ and all other elements are 0, and the scale matrix Σ is taken to be the d dimensional identity matrix $\mathbf{I}_d$. When the process is in-control, the location vector $\boldsymbol{\theta}$ is the vector of all zeros that is, $\delta = 0$. The simulation size is 1000. The upper control limit is taken to be $\chi^2_{d,0.005}$ to have an asymptotic in-control ARL of 200. The ARL curves are presented in Figures 6.8, 6.9, 6.10 for different dimensions d and the smoothing parameter r. We observe that the behaviour of the average run lengths with the shift in location are very similar for these 3 spherically symmetric distributions, which illustrates the distribution-free nature of the proposed method. We observe that the asymptotic upper control limit works best with $r = 0.3$ to attain the in-control ARL of 200. As r decreases, the in-control ARLs increase, however, all of these figures suggest that the proposed control charts are quite efficient in detecting small shifts in location.

6.4   Concluding Remarks

In this chapter, we have proposed two multivariate EWMA control charts based on multivariate signs and signed rank vectors. We have already noted that if the vector of process variables, $\boldsymbol{X}$, is spherically symmetric about $\boldsymbol{\theta}_0$, the sign vector $(\boldsymbol{X}-\boldsymbol{\theta}_0)/\|\boldsymbol{X}-\boldsymbol{\theta}_0\|$ is uniformly distributed over the d-dimensional unit sphere. Thus, the EWMA statistic based on signed vectors, $\boldsymbol{S}_k$, becomes distribution free when the process is in-control. We do not need any large sample approximation for the distribution free nature of our proposed control statistic and that makes it quite useful for a large number of distributions. Similarly, signed ranks are also distribution free when the process is in-control. In practice, it may be useful to obtain the upper control limits of the proposed charts for finite small samples using simulations or other numerical methods and we do not have to rely on asymptotic distributions. For phase I control charts, we assumed that the target location vector $\boldsymbol{\theta}_0$ is known and the scale matrix Σ is also known. If $\boldsymbol{X}$ is elliptically symmetric with scale (or, scatter) matrix Σ about $\boldsymbol{\theta}_0$, then $\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\theta}_0)$ is spherically symmetric
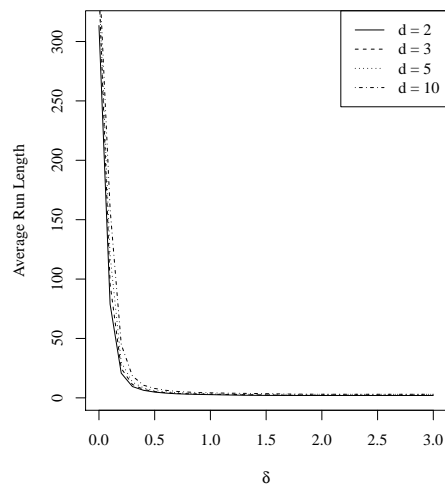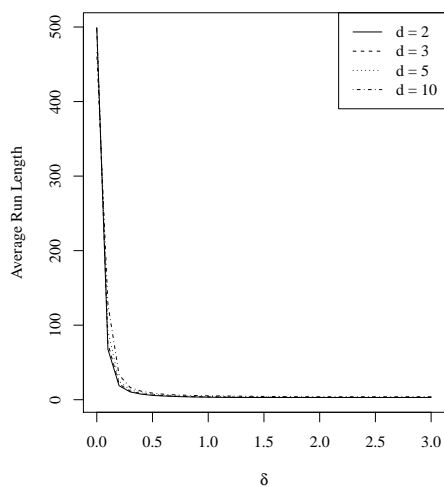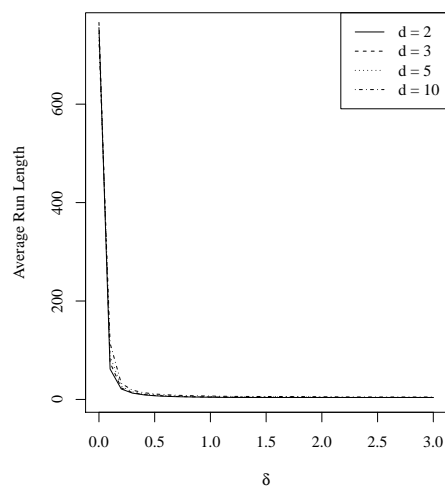
Figure 6.8: ARL curves for EWMA control charts based on multivariate signed ranks when the distribution of the process variables are multivariate normal with in-control ARL 200 and sample size $n = 15$.
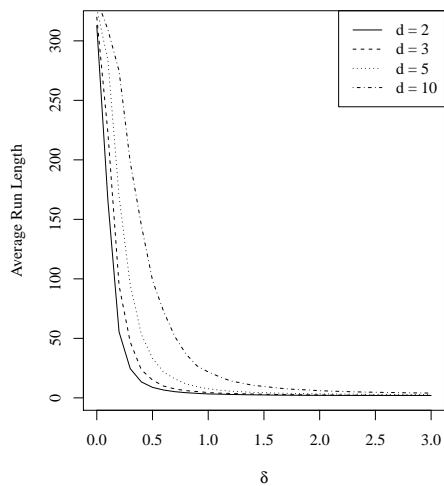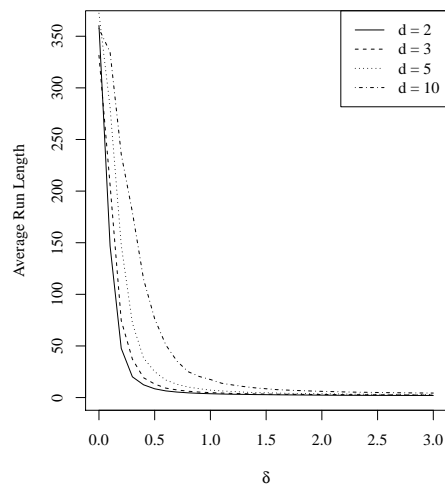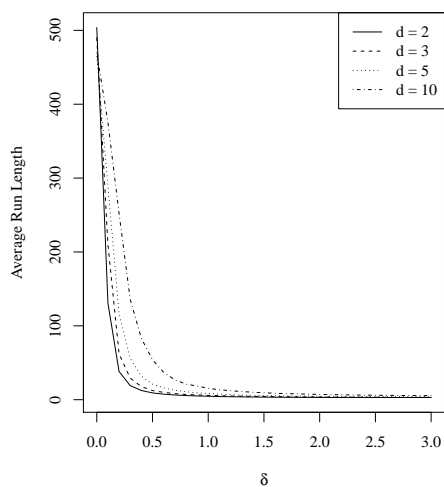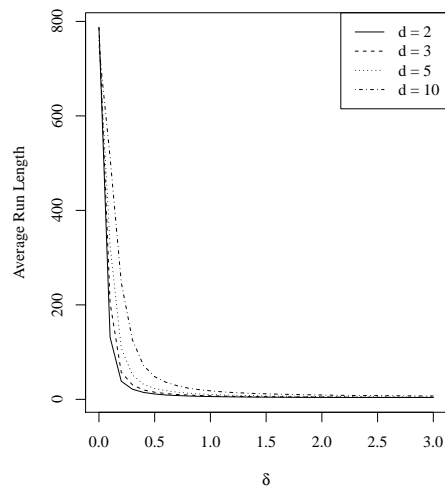
(a) $r = 0.3$

(b) $r = 0.2$

(c) $r = 0.1$

(d) $r = 0.05$

Figure 6.9: ARL curves for EWMA control charts based on multivariate signed ranks when the distribution of the process variables are multivariate Laplace with in-control ARL 200 and sample size $n = 15$.

Figure 6.10: ARL curves for EWMA control charts based on multivariate signed ranks when the distribution of the process variables are multivariate t with 3 degrees of freedom with in-control ARL 200 and sample size $n = 15$.

113

about $\mathbf{0}$. Thus if we know Σ and $\boldsymbol{\theta}_0$, we can transform the observations accordingly to have spherical symmetry and the distribution-free nature of the statistics are preserved.

Note that the proposed multivariate sign vectors and the signed rank vectors are invariant under orthogonal transformations, but they are not invariant under general affine transformations of the data. For Phase I control chart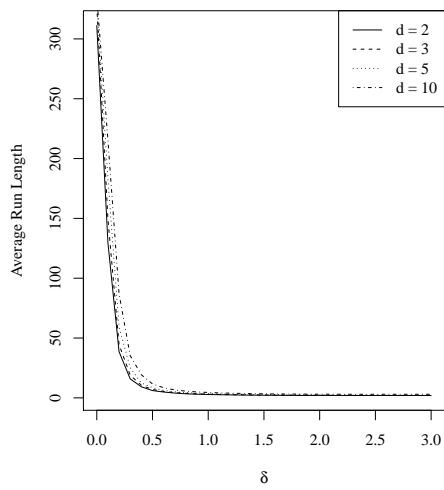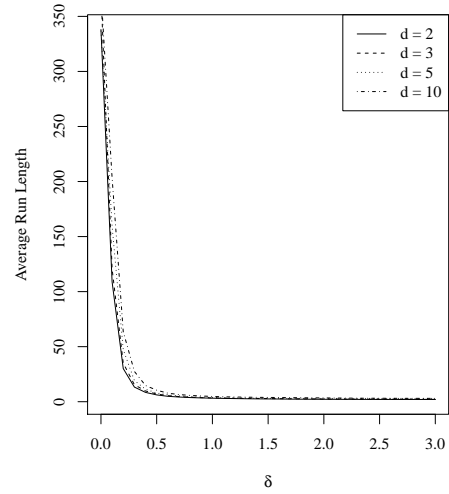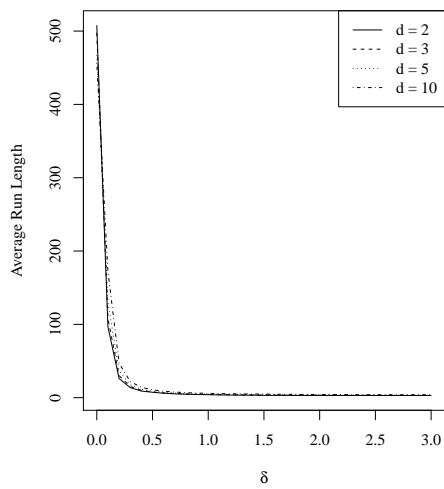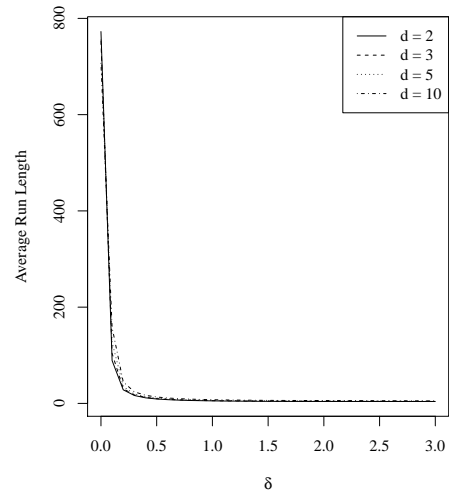s, when $\boldsymbol{\theta}_0$ and Σ are known, this lack of affine invariance does not pose any problem as we can transform the vector of process variables by $\Sigma^{-1/2}$ as described above. However, for Phase II control charts or when the scale matrix Σ is unknown we need to consider affine invariant procedure to avoid the lack of efficiency in the presence of high correlations among the process variables. Zou and Tsung (2011) considered an affine invariant version of multivariate ranks based on Hettmansperger and Randles (2002). However, their procedure is computationally quite complex. We may consider the transformation retransformation approach introduced in Chapter 4. However, we can use some simple techniques to attain affine invariance. Consider any consistent estimate $\hat{\Sigma}$ of Σ. For example, a sample variance covariance matrix if the second moments exist or minimum covariance determinant estimator (Rousseeuw and Van-Driessen, 1999) can be used and observations can be transformed using $\hat{\Sigma}^{-1/2}$. In this approach, the individual sign vectors or the statistics $\boldsymbol{S}_k$ and $\boldsymbol{R}_k$ are not invariant under affine transformations, but the control statistics Y_k and W_k are affine invariant.

In this Chapter, we have used the upper control limit of the proposed charts based on the limiting distributions. However, we need to consider the design aspect of these charts in more detail and derive the control limits for different values of r for small to finite samples. The performance of the proposed methods also depend on the sample size n as the convergence of the estimated covariance matrices $\mathbf{C}_k$ and $\mathbf{B}_k$ depends on n. We propose to discuss the effect of n on the choice of control limits in a more detailed future study.

114

Chapter 7

Conclusion

7.1 Concluding Remarks

Univariate control charts are simple but immensely useful statistical tool in industrial process monitoring and quality improvement. With technological advancements, collection and storing of large databases with several process variables have become easier over the years, and process engineers are increasingly feeling the need to exploit the dependence structure of the process variables to monitor the process better. For a long time, only multivariate control chart available for this purpose was Hotelling's T^2 chart based on T^2 statistic. In the late 1980s and 1990s, there were some developments in multivariate CUSUM and EWMA charts, which we have discussed in detail in Chapters 1 and 2. However, all of them were based on the assumption of multivariate normality for the underlying distribution of the process variables, which is usually very difficult to validate in practice and there are many instances when one knows at the outset that the underlying distribution differs from normal significantly. This gap in the literature creates a demand for developing control charts which do not depend on a specific distributional assumption. They are quite often referred as nonparametric control charts. We have also discussed some of the nonparametric control charts in the univariate set up in Chapter 1.

In Chapter 2, we have discussed some popular multivariate control charts in detail and compared their performances with respect to average run lengths (ARL) for different dimensions and multivariate distributions. In particular, we have discussed the CUSUM control charts proposed by Crosier (1988) and EWMA control charts proposed by Lowry et al. (1992). We have observed that they perform as intended when the distribution is multivariate normal, but they fail even to attain the in-control ARL when the distributions are multivariate Laplace or t, and thus they produce a large false-alarm rate. We have chosen these multivariate distributions because their shape is very close to the multivariate normal distribution but t distribution has a heavier tail and Laplace distribution has a sharper peak. Failure to perform in these very normal-like distributions makes these procedures very non-robust. Liu (1995) was among the first few to propose a control chart, which does not depend on normality. Her proposal was based on depth functions, simplicial depth (Liu, 1995) or half-space depths (Tukey, 1975). Though the properties of the proposed control charts are quite attractive due to their distribution-free nature, they did not provide any theoretical or simulation results on average run lengths of the proposed methods for their out-of-control behaviours. Most of the depth functions (e.g. simplicial, half-space or majority depth) are computationally intensive and nearly impossible to compute exactly for dimensions greater than 2. Therefore, these control charts based on data depths have limited use and infeasible for processes with more than 2 process variables.

In this work, we consider a notion of multivariate spatial rank function (Oja, 1999), which retains some important features of the univariate rank function and also it is computationally very simple with time complexity of $O(n)$ for any dimension $d \geq 2$. We propose a control chart following the idea of Liu (1995) using the univariate ranks of the lengths of the multivariate rank vectors. It can be shown that the r-chart proposed in Chapter 3 is equivalent to the Hotelling's T^2 chart when the distribution is multivariate

116

normal and also it is equivalent to the T^2 chart with optimal control limits if the underlying distribution is spherically symmetric. Therefore, this is an optimal control chart in terms of out-of-control ARL whenever the distribution is spherically symmetric and the shift occurs only in the location vector. We have also discussed some extensions of the r-chart to Q-chart and S-chart, where S-charts are analogous to CUSUM charts and detects small shifts better than r-charts. All of our theoretical results and simulations show that the proposed procedures are very promising in terms of computational simplicity in high dimensions as well as performance in detecting out-of-control signals.

A major inadequacy of the proposed multivariate rank function is that it is invariant under orthogonal transformations but not invariant under general nonsingular transformations, which makes the proposed control charts optimal only for spherically symmetric distributions, that is, they are optimal only when the scatter matrix (or, the covariance matrix) Σ associated with the distribution is λI_d for some constant λ. To resolve this issue, one can use affine invariant multivariate ranks as discussed in Chakraborty (2001). The optimal behaviour of the proposed control charts as discussed above then extends to the elliptically symmetric distributions or when Σ is a general non-singular positive definite matrix. It still retains the computational simplicity with only added complexity of computing an estimate of the scatter matrix Σ, if that is unknown.

On a completely different notion, we proposed some Shewhert type multivariate control charts based on multivariate sign and signed rank vectors, which are similar to the Hotelling's T^2 charts. We have noted that these control charts are distribution-free when the underlying distribution of the proposed variables are spherically symmetric. Even when the the distribution is elliptically symmetric with known scale matrix Σ, we can transform the observations with $\Sigma^{-1/2}$ to make them spherically symmetric and apply our proposed method without any modification. However, when the scale matrix Σ is not known, we construct some affine invariant versions of multivariate signs and and signed

ranks based on transformation re-transformation methodology proposed by Chakraborty (2001). We have also discussed some procedures to choose the optimal data-driven transformation matrix. These charts possess computational simplicity as well as large sample properties similar to the multivariate sign and signed rank tests of location. They may not be optimal when the distribution is normal, but they may provide very robust control charting techniques for deviations in distributional assumptions.

Two extensions of multivariate CUSUM control charts were proposed in Chapter 5 extending the definition of a multivariate CUSUM chart as proposed by Crosier (1988). We suggested the control limits of the proposed charts based on simulation studies and then studied their performances for different distributions. We have observed that the proposed charts are very robust and perform similarly across the distributions. They detect the small shifts in location quite efficiently.

In Chapter 6, we have discussed in detail the multivariate EWMA control chart proposed by Lowry et al. (1992). We have observed that while they are efficient in detecting a small shift in location when process variables are distributed as multivariate normal, their performance breaks down if the underlying distribution deviates from normality. We propose two EWMA control charts based on multivariate sign and signed rank vectors. We observe through simulation studies their robust behaviour under distributional assumptions. The proposed procedures are again computationally very simple. We have also discussed the issue of affine invariance in this context.

Overall, our objective in this work was to propose some multivariate control charts which exploits the dependence structure of the underlying process variables but do not depend on the assumption of multivariate normality. While the proposed charts are not fully distribution-free, they do not depend on the specific distribution as long as the underlying distribution is elliptically symmetric. We have constructed Shewhert-type, CUSUM and EWMA control charts.

7.2 Further Research

In the current work, we need to investigate in detail either theoretically or using simulations on the effect of sample sizes in determination of small sample control limits of the proposed charts based on multivariate sign and signed rank vectors. We also need to investigate further the effect of the smoothing parameter r on the construction of the EWMA charts. As a simple extension, one can use different values of r for different process variable (or, for different coordinates of the vector $\boldsymbol{X}$). While the theoretical details are not so difficult to work out, the practical implementation of such a choice might be difficult.

We would also like to consider theoretical determination of the average run lengths and their relation with the parameter k in the construction of the multivariate versions of CUSUM control charts based on multivariate sign and signed rank vectors. To attain better performance, V-masks are used in the construction of univariate CUSUM control charts, and we would like to consider those options as well. The issue of affine invariance can be dealt with similarly as we have done in Chapter 4 or 6.

We have mentioned in Chapter 1 that Kolmogorov-Smirnov test statistic and/or Cramer-von-Mises test statistic can be used to construct nonparametric control charts for the univariate processes. Dhar et al. (2014) proposed multivariate versions of Kolmogorov-Smirnov and Cramer-von-Mises statistics for comparing multivariate distributions based on multivariate ranks and quantile vectors. We can use those definitions to construct multivariate control charts, which will be truly nonparametric in nature. Our earlier proposals might be good enough to detect location shifts, but control charts based on these statistics would be able to detect any kind of distribution shift, eg. location, scale or both.

List of References

Abu-Shawiesh, M. O. and Abdullah, M. B. (2001). A new robust bivariate control chart for location. *Communications in Statistics – Simulation and Computation*, 28:2671–2686.

Alt, F. B. (1985). Multivariate quality control. In Johnson, N. L., Kotz, S., and Read, C. R., editors, *Encyclopedia of the Statistical Sciences*, volume 6, pages 111–122. John Wiley.

Alwan, L. C. (1986). Cusum quality control – multivariate approach. *Communications in Statistics - Theory and Methods*, 15:3531–3543.

Bakir, S. T. (2012). A nonparametric shewhart-type quality control chart for monitoring broad changes in a process distribution. *International Journal of Quality, Statistics, and Reliability*, vol. 2012:10 pages.

Bersimis, S., Psarakis, S., and Panaretos, J. (2007). Multivariate statistical process control charts: an overview. *Quality and Reliability Engineering International*, 23(5):517–543.

Brook, D. and Evans, D. A. (1972). An approach to the probability distribution of cusum run length. *Biometrika*, 59(3):539–549.

Chakraborti, S. and van de Wiel, M. A. (2008). A nonparametric control chart based on the mann-whitney statistic. In Balakrishnan, N., Peña, E. A., and Silvapulle, M. J.,

editors, *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1 of *Collections*, pages 156–172. Institute of Mathematical Statistics.

Chakraborti, S., van der Laan, P., and Bakir, S. T. (2001). Nonparametric control charts: An overview and some results. *Journal of Quality Technology*, 33:304–315.

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics*, 53(2):380–403.

Chakraborty, B. and Chaudhuri, P. (1999). On affine invariant sign and rank tests in one-and two-sample multivariate problems. In Ghosh, S., editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 499–522. Marcel Dekker Inc.

Chakraborty, B. and Chaudhuri, P. (2008). On an optimization problem in robust statistics. *Journal of Computational and Graphical Statistics*, 17(3):683–702.

Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8(3):767–784.

Chang, Y. M. and Wu, T. L. (2011). On average run lengths of control charts for autocorrelated processes. *Methodology and Computing in Applied Probability*, 13(2):419–431.

Chatterjee, S. and Qiu, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *The Annals of Applied Statistics*, 3:349–369.

Chaudhuri, P. (1992). Multivariate location estimation using extension of r-estimates through u-statistics type approach. *The Annals of Statistics*, 20(2):897–916.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3):291–303.

Dhar, S. S., Chakraborty, B., and Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli*, 20(3):1484–1506.

Ewan, W. D. (1963). When and how to use cu-sum charts. *Technometrics*, 5(1):1–22.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2*. Wiley, New York, NY, second edition.

Fuchs, C. and Kenett, R. S. (1998). *Multivariate Quality Control: Theory and Applications*. CRC Press.

Guha, P. (2012). *On scale-scale curves for multivariate data based on rank regions*. PhD thesis, University of Birmingham.

Hawkins, D. M. (1981). A cusum for a scale parameter. *Journal of Quality Technology*, 13:228–231.

Hawkins, D. M. (1991). Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33(1):61–75.

Healy, J. D. (1987). A note on multivariate cusum procedures. *Technometrics*, 29(4):409–412.

Hettmansperger, T. P. and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika*, 89(4):851–860.

Hotelling, H. (1947). Multivariate quality control illustrated by the air testing of sample bomb sights. In Eisenhart, C., Hastay, M. W., and Wallis, W. A., editors, *Selected Techniques of Statistical Analysis*. McGraw-Hill, New York.

Hunter, J. S. (1986). The exponentially weighted moving average. *Journal of Quality Technology*, 18:203–210.

Jackson, J. E. (1985). Multivariate quality control. *Communications in Statistics – Theory and Methods*, 14:2657–2688.

Johnson, N. L. (1961). A simple theoretical approach to cumulative sum control charts. *Journal of the American Statistical Association*, 56(296):835–840.

Koltchinskii, V. (1997). M-estimation, convexity and quantiles. *The annals of Statistics*, pages 435–477.

Kruger, U. and Xie, L. (2012). *Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control*. John Wiley & Sons, Ltd., Chichester.

Leavenworth, R. S. and Grant, E. L. (1976). *Statistical Quality Control*. Tata McGraw-Hill Education.

Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Sciences*, 85(6):1732–1734.

Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, R. Y. and Tang, J. (1996). Control charts for dependent and independent measurements based on bootstrap methods. *Journal of the American Statistical Association*, 91:1694–âĂŞ1700.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1):46–53.

123

Lucas, J. M. (1985). Counted data cusum's. *Technometrics*, 27(2):129–144.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, 32:1–29.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Makinde, O. S. and Chakraborty, B. (2015). On some nonparametric classifiers based on distribution functions of multivariate ranks. In Nordhausen, K. and Taskinen, S., editors, *Modern Nonparametric, Robust and Multivariate Methods*, pages 249–264. Springer International Publishing.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*. Wiley, New York, sixth edition.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5(2):201–213.

Möttönen, J., Oja, H., and Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics*, 25(2):542–552.

Oakland, J. S. (1996). *Statistical Process Control*. Routledge.

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: a review. *Scandinavian Journal of Statistics*, 26(3):319–343.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41:100–115.

Page, E. S. (1963). Controlling the standard deviation by cusums and warning lines. *Technometrics*, 5(3):307–315.

Qiu, P. and Hawkins, D. M. (2001). A rank-based multivariate cusum procedure. *Technometrics*, 43(2).

Qiu, P. and Hawkins, D. M. (2003). A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *Journal of the Royal Statistical Society, Series D*, 52:151–164.

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2):102–116.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

Rousseeuw, P. J. and Van-Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Runger, G. C. and Testik, M. C. (2004). Multivariate extensions to cumulative sum control charts. *Quality and Reliability Engineering International*, 20(6):587–606.

Ryan, T. P. (1989). *Statistical Methods for Quality Improvement*. John Wiley & Sons.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference*, 123(2):259–278.

Stoumbos, Z. G., Reynolds Jr, M. R., Ryan, T. P., and Woodall, W. H. (2000). The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, 95(451):992–998.

Thissen, U., Swierenga, H., Wehrens, R., Melssen, W. J., and Buydens, M. C. (2005). Multivariate statistical process control using mixture modelling. *Journal of Chemometrics*, 19:23–âĂŞ31.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, 40(1):72–76.

Woodall, W. H. and Ncube, M. M. (1985). Multivariate cusum quality-control procedures. *Technometrics*, 27(3):285–292.

Yang, S.-F. and Cheng, S. W. (2011). A new non-parametric cusum mean chart. *Quality and Reliability Engineering International*, 27(7):867–875.

Zhang, J., Li, Z., and Wang, Z. (2010). A multivariate control chart for simultaneously monitoring process mean and variability. *Computational Statistics & Data Analysis*, 54(10):2244–2252.

Zou, C. and Tsung, F. (2011). A multivariate sign ewma control chart. *Technometrics*, 53(1):84–97.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.