ROBUST ESTIMATION METHODS IN STATISTICAL QUALITY CONTROL

JAYASANKAR VATTATHOOR

Robust Estimation Methods in Statistical Quality Control

by

©Jayasankar Vattathoor

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirement for the Degree of

Master of Science in Statistics

Department of Mathematics and Statistics

Memorial University of Newfoundland

St. John's Newfoundland, Canada 21 September 2012

Abstract

Multivariate control charts are widely used in industry to monitor changes in the process mean and process variability. The classical estimators, sample mean and sample variance, used in control charts are highly sensitive to outliers in the data. In Phase-I monitoring, the control limits are set based on the historical data after the outliers have been identified and removed. The identification of the outliers in Phase-I is not straightforward. We propose robust control charts with high-breakdown robust estimators based on the re-weighted minimum covariance determinant (RMCD) and the re-weighted minimum volume ellipsoid (RMVE). These charts monitor the process mean and the process variability in the historical Phase-I data in the case of individual multivariate observations.

To monitor the process mean, we propose using Hotelling's T^2 control charts with RMCD and RMVE estimators of the mean and the covariance matrix. We set the control limits empirically based on a large number of Monte Carlo simulations. We

assessed the performance of these methods by considering different data scenarios and found that our methods improve on existing methods. We suggest using robust T^2 charts based on RMCD estimators for data with large samples and large dimensions and RMVE estimators for data with smaller samples and smaller dimensions. We also propose using robust versions of the MEWMS/MEWMV schemes to monitor process variability in Phase-I. The control limits of these robust control charts are set empirically, and the charts improve on existing methods. We also extended the concept of robust estimation in the context of generalized linear models.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Asokan Mulayath Variyath, for his guidance, advice, encouragement, and financial support. Working with Dr. Asokan was a rewarding experience because he shaped my overall understanding of statistics and helped me to appreciate the importance of the thesis topic. I also want to thank Dr. David Pike of the Department of Mathematics and Statistics and Mr. Nolan White of the Department of Computer Services for allowing me to use their computers for my time-consuming simulations. My sincere thanks to Dr. Jahrul Alam of the Department of Mathematics and Statistics for helping me to switch to Linux-based systems so I could access more computers and thus speed up the simulation study.

I am also grateful to the School of Graduate Studies and the Department of Mathematics and Statistics for their financial support in the form of a graduate fellowship and teaching assistantship. I would like to thank all the faculty and staff

of the Department of Mathematics and Statistics for the friendly atmosphere and for providing the necessary facilities.

I am very grateful to my parents, my wife Mini, and my son Rahul for their strong belief in my abilities and for their eternal love, support, and encouragement throughout my studies. I also thank the friends and well-wishers who directly or indirectly encouraged and helped me and contributed to this dissertation. Above all, I thank God for his grace in allowing me to successfully complete the program.

Contents

Abstract ii

Acknowledgements iv

List of Tables vii

List of Figures x

1 Introduction 1

1.1 Overview 1

1.2 Multivariate Process Monitoring 3

1.2.1 Monitoring Process Mean 4

1.2.2 Monitoring Process Variability 6

1.3 Background of Problem 9

2 Robust Estimators **15**

2.1 Desirable Properties of Robust Estimator 16

2.2 MVE and RMVE Estimators 20

2.3 MCD and RMCD Estimators 24

3 Robust Control Charts for Monitoring Process Mean **28**

3.1 Robust Control Charts 28

3.2 Computation of Control Limits 29

3.3 Performance Analysis 34

4 Robust Control Charts for Monitoring Process Variability **44**

4.1 MEWMS Control Charts 46

4.2 MEWMV Control Charts 48

4.3 Control Charts Based on L_e-norm Function 51

4.4 Robust Control Charts for Monitoring Variability 55

4.4.1 Performance Comparison for Phase-I Monitoring 57

5 Robust Regression **73**

5.1 Generalized Linear Model 75

5.1.1 Poisson Log Linear Model 76

5.1.2 Binary Logistic Regression Model 78

5.2 Robust Generalized Linear Regression 79
5.3 Simulation Studies 82
5.3.1 Poisson Regression Model 83
5.3.2 Binary Logistic Model 86
5.4 Comparison Study 90
5.4.1 Poisson Case 92
5.4.2 Binary Case 94

6 Conclusions and Future Work **97**

Bibliography **100**

List of Tables

3.1 Estimates of model parameters $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}, a_{3(p,1-\alpha)}$ for dimensions p = (2,...,10) and confidence levels α = (0.05, 0.01, 0.001) for T^2_{RMCD} control charts 33

3.2 Estimates of model parameters $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}, a_{3(p,1-\alpha)}$ for dimensions p = (2,...,10) and confidence levels 1-α = (95%, 99%, 99.9%) for T^2_{RMVE} control charts 33

4.1 Control limits for robust control charts with MEWMS scheme; p=2 and m=50 58

4.2 Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.10, p=2 and m=50 59

4.3 Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.20, p=2 and m=50 60

4.4 Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.30, p=2 and m=50 61

4.5 Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.40, p=2 and m=50 62

5.1 Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters under Poisson model with β = (3.0,3.5,0,0,2.0) in presence of single outlier 84

5.2 Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters under Poisson model with β = (3.0,3.5,0,0,2.0) in presence of two outliers 86

5.3 Simulated means (SM), standard errors (SSE), and the relative biases (RB) of estimates of the regression parameters under the binary model with β = (3.0,3.5,0,0,2.0) in the presence of single outlier 87

5.4 Simulated means (SM), standard errors (SSE), and the relative biases (RB) of estimates of the regression parameters under the binary model with β = (3.0,3.5,0,0,2.0) in the presence of two outliers 89

5.5 Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters for sample of size = 60 under Poisson model with β = (1.0,0.5) in presence of one or two outliers 93

5.6 Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters for sample of size = 60 under binary model with $\beta = (1.0,0.5)$ in presence of one or two outliers 96

List of Figures

3.1 Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 2 . 31

3.2 Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 6 . 31

3.3 Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 10 32

3.4 Probability of signal for RMCD/RMVE control limits for p= 2, m= 30 36

3.5 Probability of signal for RMCD/RMVE control limits for p= 2, m= 50 36

3.6 Probability of signal for RMCD/RMVE control limits for p= 2, m= 100 37

3.7 Probability of signal for RMCD/RMVE control limits for p= 2, m= 150 37

3.8 Probability of signal for RMCD/RMVE control limits for p= 6, m= 30 38

3.9 Probability of signal for RMCD/RMVE control limits for p= 6, m= 50 38

3.10 Probability of signal for RMCD/RMVE control limits for p= 6, m= 100 39

3.11 Probability of signal for RMCD/RMVE control limits for p= 6, m= 150 39

3.12 Probability of signal for RMCD/RMVE control limits for p= 10, m= 30 40

3.13 Probability of signal for RMCD/RMVE control limits for p= 10, m= 50 40

3.14 Probability of signal for RMCD/RMVE control limits for p= 10, m= 100 41

3.15 Probability of signal for RMCD/RMVE control limits for p= 10, m= 150 41

4.1 Probability of signal for robust $MEWMSL_1$ control chart for p= 2, m= 50, ω =0.30, $\mu_1 = \mu_2 = 0$ 65

4.2 Probability of signal for robust $MEWMSL_2$ control chart for p= 2, m= 50, ω =0.40, $\mu_1 = \mu_2 = 0$ 65

4.3 Probability of signal for robust $MEWMS$ control chart for p= 2, m= 50, ω =0.50, $\mu_1 = \mu_2 = 0$ 66

4.4 Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 0.50$ 66

4.5 Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 1.00$ 67

4.6 Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 2.00$ 67

4.7 Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 0$ 68

4.8 Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.20, $\mu_1 = \mu_2 = 0$ 68

4.9 Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.30, $\mu_1 = \mu_2 = 0$ 69

4.10 Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.40, $\mu_1 = \mu_2 = 0$ 69

4.11 Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.10, $\mu_1 = \mu_2 = 0$ 70

4.12 Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.20, $\mu_1 = \mu_2 = 0$ 70

4.13 Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.30, $\mu_1 = \mu_2 = 0$ 71

4.14 Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.40, $\mu_1 = \mu_2 = 0$ 71

Chapter 1

Introduction

1.1 Overview

Quality has become the basic consumer decision factor in a competitive market. Consumers who have long-standing relationships with the same suppliers may select alternatives when better-quality products or services are available. The quality of a product or service can be defined as the sum of the characteristics that impact its ability to satisfy the stated and implied needs of the customer. The manufacturing and service industries are placing more emphasis on the quality of their products and services as they realize that "the cost is long forgotten but the quality is remembered for ever." As the expectations of customers grow, businesses must continually

improve the quality of their products and services in order to remain competitive. High standards do not happen by chance; they evolve over time as a result of continuous improvement. Organizations can secure their future by engaging in continual improvement and adopting new processes for conformity assessment. A product or service should meet high standards in terms of both quality of the design and quality of conformance. The quality of the design reflects the customer requirements, and quality of conformance is achieved when the actual product or service is as close as possible to the design.

Statistical process control (SPC) is a set of statistical tools used to monitor and control a process to ensure that it produces a conforming product. SPC techniques help to identify the root causes of quality and productivity problems, so that appropriate corrective and preventative measures can be taken. SPC is usually applied to manufacturing processes, but it is suitable for any process with a measurable output. The use of SPC in industry has increased in recent years because of improvement in data collection and data-handling systems. The most widely used SPC technique is the control chart.

Control charts are important and effective SPC tools. They are used to identify and remove the assignable causes affecting a process, thereby ensuring that the process is in statistical control, i.e., it is affected by chance causes alone. Control charts are

graphical devices for detecting changes in the manufacturing conditions due to the presence of assignable causes by comparing the observed values with limits derived from the historical (Phase-I) data. The Phase-II data analysis consists of monitoring future observations based on the control limits found from the Phase-I estimates to determine whether or not the process continues to be in-control. The most commonly used variable-type control charts for univariate data are $\bar{X}$-R charts and $\bar{X}$-s charts, where the $\bar{X}$-chart is used to monitor the process mean, and the R-chart (or s-chart) is used to monitor the process variability.

1.2 Multivariate Process Monitoring

In many applications in industrial quality control, more than one quality characteristic is of interest, and hence multivariate control charts are more relevant than univariate charts. Individual charts for each quality characteristic can also be used. However, when the quality characteristics are correlated, multivariate control charts are more effective than multiple charts. The most commonly used multivariate chart to monitor the process mean is Hotelling's T^2 control chart (Hotelling, 1947). The S^2 chart, G-Chart, multivariate exponentially weighted mean square (MEWMS) chart, and multivariate exponentially weighted moving variance (MEWMV) chart are used to

monitor the process variance in the multivariate case. A brief description of these charts is given in the following sections.

1.2.1 Monitoring Process Mean

Hotelling's T^2 control chart monitors shifts in the process mean assuming that all the quality characteristics are normally distributed. In many situations, multivariate data are collected according to a rational-subgroup concept, i.e., sample data are collected at some time point in the process. Let $X_1, X_2, \cdots, X_p$ be the p quality characteristics of interest. We assume that X= $(X_1, X_2, \cdots, X_p)'$ is normally distributed with multivariate mean μ and covariance matrix Σ. We collect m samples (subgroups) of size n each at regular intervals. For the ith subgroup, we have n samples of p-dimensional observations: $(x_{i11}, x_{i12}, \cdots, x_{i1p})'$, $(x_{i21}, x_{i22}, \cdots, x_{i2p})'$, $\cdots$, $(x_{in1}, x_{in2}, \cdots, x_{inp})'$. The sample mean vector and sample covariance matrix for this subgroup are estimated as:

$$\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{ip})' = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{ij} \qquad and$$

$$S_i = \frac{1}{(n-1)}\sum_{j=1}^{n}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)', \qquad i = 1, 2, \cdots, m \tag{1.1}$$

where $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijp})'$ is the j-th ($j = 1, 2, \cdots, n$) p-dimensional observation from the i-th subgroup. The mean μ and covariance matrix Σ are estimated by

averaging the sample means and sample covariances over all m subgroups:

$$\bar{\bar{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{x}}_i$$
$$S = \frac{1}{m}\sum_{i=1}^{m}S_i. \tag{1.2}$$

Hotelling's T^2 statistic for the ith subgroup is

$$T_1^2(i) = n(\bar{\mathbf{x}}_i - \bar{\bar{\mathbf{x}}})' S^{-1}(\bar{\mathbf{x}}_i - \bar{\bar{\mathbf{x}}}). \tag{1.3}$$

The Phase-I and Phase-II control limits of the T_1^2 chart are found based on the F-distribution with (p, mn-m-p+1) degrees of freedom.

However, it is time-consuming and difficult to collect rational subgroups of size greater than one when the processing time is too large or the production rate is too slow. When the differences among repeated measurements are due to laboratory or analysis error, as in many chemical processes, it is not convenient to collect subgroups of size greater than one. Hence, individual multivariate observations are important. To monitor a multivariate process mean in this case, for the ith individual multivariate observation from a sample of size m, we calculate

$$T_2^2(i) = (\mathbf{x}_i - \bar{\mathbf{x}})' S_*^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) \tag{1.4}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$, $i = 1, 2, \cdots, m$, are the p-variate observations. The

sample mean $\bar{\mathbf{x}}$ and sample covariance matrix S_* are

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$$
$$S_* = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

As shown by Tracy, Young, and Mason (1992), the Phase-I control limits of the T_2^2 chart are found based on a beta distribution with (p/2, (m-p-1)/2) degrees of freedom, and the Phase-II control limits are based on an F-distribution with (p, m-p) degrees of freedom.

1.2.2 Monitoring Process Variability

Alt and Smith (1988) proposed multivariate control charts for monitoring process variation in the Phase-I data when the data are collected in subgroups where each p-dimensional data point follows a multivariate normal distribution with mean μ and covariance matrix Σ. They have extended the univariate s^2 chart to the multivariate case. The S^2 chart is based on the likelihood ratio test $H_0 : \Sigma = \Sigma_0$ vs $\Sigma \neq \Sigma_0$.
Given m subgroups each of size n from p-dimensional multivariate data, we define the statistic W_i for each subgroup $i = 1, 2, \cdots, m$:

$$W_i = -p(n-1) - (n-1)\ln(\mid S_i \mid) + (n-1)\ln(\mid \Sigma_0 \mid) + (n-1)tr(\Sigma_0^{-1}S_i) \quad (1.5)$$

where ln is the natural logarithm, *tr* is the trace function, $|\ .\ |$ is the determinant, and S_i is defined as in Eq. (1.1). When Σ_0 is known, the value of W_i is compared with the upper control limit (UCL)= $\chi^2_{[p(p+1)/2,(1-\alpha)]}$, where $1-\alpha$ is the confidence level. If the value of Σ_0 is not known, it can be estimated by $|\ S^*\ | = \frac{1}{m}\sum_{i=1}^{m} |\ S_i\ |$. Alt and Smith (1988) showed that $E(|\ S^*\ |) = b_1\ |\ \Sigma_0\ |$ where $b_1 = \frac{1}{(n-1)^p}\prod_{j=1}^{p}(n-j)$. Therefore, an unbiased estimate of Σ_0 is $\dfrac{|\ S^*\ |}{b_1}$ and we can construct the S^2 chart using the statistic W_i. The UCL is found empirically by Monte Carlo simulation such that the overall false alarm probability is α.

Alt and Smith (1988) introduced another chart known as the $|\ S\ |^{1/2}$ chart using the property that most of the probability distribution of $|\ S\ |^{1/2}$ is contained in the interval

$$E(|\ S\ |^{1/2}) \pm 3\sqrt{V(|\ S\ |^{1/2})} \tag{1.6}$$

where $E(|\ S\ |^{1/2}) = b_3\ |\ \Sigma_0\ |^{1/2}$ and $V(|\ S\ |^{1/2}) = (b_1 - b_3^2)\ |\ \Sigma_0\ |$ with $b_3 = \frac{(\frac{2}{n-1})^{p/2}}{\Gamma(\frac{n}{2})\Gamma(\frac{n-p}{2})}$. If Σ_0 is known, the UCL for the $|\ S\ |^{1/2}$ chart is

$$|\ \Sigma_0\ |^{1/2}\left(b_3 + 3\sqrt{b_1 - b_3^2}\right). \tag{1.7}$$

If Σ_0 is not known, $|\ \Sigma_0\ |^{1/2}$ is estimated using $\dfrac{|\ S^{**}\ |^{1/2}}{b_3}$, where $|\ S^{**}\ |^{1/2}$ is calculated by $\frac{1}{m}\sum_{i=1}^{m} |\ S_i\ |^{1/2}$, and the UCL is

$$|\ S^{**}\ |^{1/2}\left(1 + 3\frac{\sqrt{b_1 - b_3^2}}{b_3}\right). \tag{1.8}$$

Levinson, Holmes, and Mergen (2002) suggested the G chart, based on the comparison of the sample covariance matrix of each subgroup with an overall estimate of Σ_0. They calculated the weighted average of S_i in Eq. (1.1) and S in Eq. (1.2) as $S_2(i) = \dfrac{m(n-1)S + (n-1)S_i}{m(n-1) + (n-1)}$. For each subgroup, the statistic for the control chart is

$$G_i = \mathbf{k} \times (n-1)\{\ln(\mid S_2(i) \mid) - m \times \ln(\mid S \mid) - \ln(\mid S_i \mid)\} \tag{1.9}$$

where $\mathbf{k} = 1 - \{\dfrac{1.5}{n-1} \times \dfrac{2p^2+3p-1}{6(p+1)}\}$. The UCL of the G chart is $\chi^2_{[p(p+1)/2,(1-\alpha)]}$, where $1-\alpha$ is the confidence level. The process variability is monitored by comparing the value of G_i with the UCL, as for the other control charts, with LCL = 0.

As discussed earlier, obtaining samples with a subgroup size greater than one is difficult in many practical situations, and the monitoring of variability based on individual observations is preferred in such circumstances. Huwang, Yeh, and Wu (2007) proposed two control charts that use individual observations to monitor the process variability. They considered the following two situations:

- Changes in the process variability when there is no shift in the process mean;
- Changes in the process variability coupled with a shift in the process mean.

They introduced charts based on the MEWMS scheme for the first case and the MEWMV scheme for the second case, using the trace of the unbiased estimate of the

covariance matrix. They showed that these two charts perform better than multiple cumulative sum charts (MCUSUM) and multiple exponentially weighted moving average (MEWMA) charts for various scenarios. However, they could not explain the situation in which in-control and out-of-control covariance matrices have the same trace.

Memar and Niaki (2009) suggested new charts to overcome this deficiency. They modified the control charts of Huwang et al. (2007) by introducing the L_c norm function for any vector z= $(z_1, z_2, \cdots, z_p)$ of length p as:

$$\| z \|_c = \left(\sum_{i=1}^{p} | z |^c \right)^{1/c}. \tag{1.10}$$

Instead of the trace, they considered L_1 and L_2 functions (sum of absolute values or sum of squares) of the deviation of each diagonal element of the unbiased estimate of the covariance matrix from its target value. They showed that their $MEWMSL_1$, $MEWMSL_2$, $MEWMVL_1$, and $MEWMVL_2$ charts perform better than that of Huwang et al. (2007) under various scenarios.

1.3 Background of Problem

The historical Phase-I data is analyzed to determine whether the data indicates a stable (or in-control) process and to estimate the process parameters and construction

of control limits. The Phase-II data analysis consists of monitoring future observations based on control limits derived from the Phase-I estimates to determine whether the process continues to be in-control or not. But trends, step changes, outliers and other unusual data points in the Phase-I data can have an adverse effect on the estimation of parameters and the resulting control limits. ie. Any deviation from the main assumption (in our case, identically and independently distributed from multivariate normal distribution) may lead to out of control situation. So it becomes very important to identify and eliminate these data points prior to calculating the control limits. In this thesis, all these unusual data points are referred as "outliers". Care should be taken in the analysis of the Phase-I data, especially when outliers are present. Control limits based on data from unstable (or out-of-control) processes that contain outliers will be inaccurate, leading to ineffective Phase-II monitoring.

It is more difficult to detect outliers in multivariate data than in univariate data. Univariate outliers can be easily identified graphically but identification of multivariate outliers are often not possible in higher dimensions. More over, there are many ways that multivariate outliers can come from an out-of-control process such as:

a) a few or cluster of outliers due to changes of location in random directions;

b) multiple clusters of outliers in different directions;

c) data points with the same location as the good data but with more variability;

d) a shift in some of the elements of the location vector but not all of them.

e) multiple outliers are present and inflate the estimates in such a way that they mask each other so that it is difficult to detect.

Rocke and Woodruff (1996) stated that the most difficult multivariate outliers to detect are those that have the same variance-covariance matrix. These outliers are referred to as "shift outliers" because their center has been shifted from the center of the other data points. If shift outliers can be detected by robust estimation methods, then such methods will likely to work well for all other types of outliers.

The classical estimates, sample mean and sample covariance, are highly sensitive to outliers; we need estimation methods that are more robust. Sullivan and Woodall (1996) proposed an estimate of the covariance matrix based on successive differences of the multivariate observations to reduce the effect of shift outliers. This is equivalent to the use of the moving range to construct a Shewhart individual control chart in the univariate case.

Sullivan and Woodall (1996) defined the vector V_j to be

$$V_j = X_{j+1} - X_j, \quad j = 1, 2, \ldots, (m-1).$$

When the control chart is constructed, the unbiased estimator of the covariance matrix, $S_*(1) = \frac{1}{2(m-1)} \sum_{j=1}^{m-1} V_j V_j'$, replaces the covariance matrix S_* in Eq. (1.4). Successive-difference charts are effective for detecting sustained step changes but not

for detecting multiple multivariate outliers. Robust estimation methods are suitable for detecting multivariate outliers because of their high breakdown points, which ensure that the control limits are reasonably accurate.

Vargas (2003) introduced robust control charts that used two robust estimates of the location and scatter, namely the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE), to identify multivariate outliers. The exact distribution of T^2 with the robust estimators based on MVE and MCD was not available, so the control limits were obtained empirically. Jensen, Birch, and Woodall (2007) showed that the T^2_{MCD} and T^2_{MVE} control charts have better performance in the presence of outliers.

The MCD/MVE estimators have low statistical efficiency because they use only some of the data points. We propose control charts based on the re-weighted minimum covariance determinant (RMCD) and the re-weighted minimum volume ellipsoid (RMVE) to monitor the shift in the process mean and the shift in the variability. RMCD/RMVE estimators are statistically more efficient than MCD/MVE estimators and have a manageable asymptotic distribution. Chenouri, Steiner, and Variyath (2009) used RMCD estimators to monitor the Phase-II data when there is a shift only in the location. However, in many situations Phase-I control charts are necessary to assess performance and to identify outliers.

Vargas and Lagos (2007) proposed the robust G control chart (RG chart) to monitor the covariance matrix in the case of subgroup data. They modified the G chart suggested by Levinson et al. (2002) by using the MVE estimator of the covariance matrix of the full data, instead of the pooled covariance estimator S used by Levinson et al. (2002). They showed that the RG charts are able to detect changes in the variability. However, to date there are no robust control charts that monitor the covariance matrix for individual multivariate observations.

The problem of presence of outliers in the individual multivariate data can be viewed in three different perspectives:

- A shift in the mean vector of the process.
- A change in the covariance matrix process.
- A shift in mean vector together with a change in the covariance matrix.

The goal of this thesis is to address the outlier detection problem from these three perspectives with emphasis on robust estimators and to highlight the applications of the RMCD and RMVE in the areas of statistical quality control. We propose to use robust control charts based on RMCD/RMVE estimators and arrive the control limits empirically as the corresponding statistics do not have closed-form distributions. We fit a nonlinear regression model to find the control limits for a given sample size

in the case of RMCD/RMVE-based T^2 charts. Our simulation studies show that RMCD/RMVE-based charts perform well compared to existing charts in monitoring the shift in the process mean and the shift in the process variability. We also propose to use the outlier detection method with RMCD/RMVE estimators when estimating the parameters of a generalized linear regression model.

The remainder of this thesis is organized as follows. In Chapter 2, we discuss existing robust estimation methods, and we formally introduce the RMCD and the RMVE. In Chapter 3, we discuss the proposed robust charts for monitoring shifts in the mean vector for individual multivariate observations for Phase-I data. We compare the performance of the charts via simulation studies. In Chapter 4, we discuss multivariate control charts for individual observations to monitor process variability when the process exhibit shift in mean as well as variability. We compare the performance of the charts via simulation studies. In Chapter 5, we consider using the RMCD/RMVE estimators to identify and remove outliers in the covariate data and to find robust estimates of the regression parameters in the generalized linear model. In Chapter 6, we summarize our results and discuss directions for future research.

Chapter 2

Robust Estimators

To study a variable of interest and its properties, we need to know the parameters that characterize its distribution. In practical situations, the true parameter values are unknown and we must estimate them from the sample data. For example, suppose a p-variate quality characteristic follows a multivariate normal distribution characterized by mean vector μ and covariance matrix Σ. These parameters are often estimated by the sample mean and the sample covariance matrix, since they have most of the characteristics of good estimators. However, these estimators are highly sensitive to the presence of outliers. In contrast, robust estimators are not unduly affected by outliers. If outliers are present in the data, robust estimators are more appropriate. There are a number of such estimators available in the literature but

varying properties.

2.1 Desirable Properties of Robust Estimator

A good robust estimator has the following properties:

- Affine equivariance;
- High breakdown point;
- Statistical efficiency;
- Computational efficiency.

Affine equivariance : Consider a multivariate data set $X^m = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m)$ with m observations where $\mathbf{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jp})'$ represents the jth multivariate observation with dimension p, $j = 1, 2, \cdots, m$. Estimators $\mathbf{T}_m$ of the location parameter μ and $\mathbf{C}_m$ of the covariance matrix Σ are affine equivariant if for any nonsingular p×p matrix $\mathbf{A}$ and vector $\mathbf{b} \in \mathbb{R}^p$,

$$\begin{aligned} \mathbf{T}_m(\mathbf{AX} + \mathbf{b}) &= \mathbf{A}\mathbf{T}_m(\mathbf{X}) + \mathbf{b} \\ \mathbf{C}_m(\mathbf{AX} + \mathbf{b}) &= \mathbf{A}\mathbf{C}_m(\mathbf{X})\mathbf{A}'. \end{aligned} \tag{2.1}$$

Such estimators are unchanged or change in appropriate ways when the measurements and the parameters are transformed. Affine equivariance is important because it

makes the analysis independent of the measurement scale of the variables and of transformations or rotations of the data.

Breakdown point : The breakdown point concept introduced by Donoho and Huber (1983) is often used to assess robustness. The breakdown point is "the smallest proportion of the observations which can render an estimator meaningless." For example, let X^m be a random sample of m observations and $\mathbf{T}_m(\mathrm{X}^m)$ be the corresponding estimator of the parameter of interest. Consider replacing k points in X^m by arbitrary values and let the new data be represented by $\mathrm{X}^{m(k)}$. The finite-sample breakdown point of the location estimator $\mathbf{T}_m$ for the sample X^m is the smallest fraction $\frac{k}{m}$ of outliers that can carry the estimate over all bounds. It is given by

$$\epsilon(\mathbf{T}_m, \mathrm{X}^m) = min \left\{ \frac{k}{m}; \sup_{\mathrm{X}^{m(k)}} \|\mathbf{T}_m(\mathrm{X}^{m(k)}) - \mathbf{T}_m(\mathrm{X}^m)\| = \infty \right\} \tag{2.2}$$

where $\|.\|$ is the Euclidean norm.

If $\epsilon(\mathbf{T}_m, \mathrm{X}^m)$ is independent of the initial sample X^m, we say that the estimator $\mathbf{T}_m$ has the universal finite-sample breakdown point $\epsilon_m(\mathbf{T}_m)$. We can then calculate its limit $\epsilon = lim_{m\to\infty}\epsilon_m(\mathbf{T}_m)$, which is often called the asymptotic breakdown point or the breakdown point. A higher breakdown point implies a more robust estimator. The highest attainable breakdown point is $\frac{1}{2}$ in the case of the median in the univariate case. The breakdown point of a sample mean of size m is $1/m$, and hence for univariate data, the sample median is more robust than the sample mean.

It is difficult to find an affine-equivariant robust estimator since affine equivariance and high breakdown do not occur simultaneously. Lopuhaä and Rousseeuw (1991) and Donoho and Gasko (1992) pointed out that no affine-equivariant estimator can attain a finite-sample breakdown point of $\frac{(m-p+1)}{(2m-p+1)}$. The largest attainable finite-sample breakdown point of any affine-equivariant estimator of the location and scatter matrix is $\frac{(m-p+1)}{2m}$ (Davies, 1987). Relaxing the affine-equivariance condition to invariance under the orthogonal transformation makes it easy to find an estimator with the highest breakdown point of $\frac{1}{2}$.

Statistical efficiency: An estimator is said to be statistically efficient if it estimates the quantity of interest in the best possible manner. The definition of "best possible" depends on the choice of loss function, the function that quantifies the relative degree of undesirability of estimation errors of different magnitudes. The most common loss function is quadratic, resulting in the mean squared error (MSE) criterion of optimality. Hence, we consider an estimator to be efficient compared to some other estimator if its MSE is smaller for at least some values of the parameter. For example, for a sample of size m from the normal distribution with mean μ and standard deviation 1, the sample mean and sample median are unbiased estimators of μ and their MSEs are $1/m$ and $\pi/2m$ respectively. The mean is more efficient than the median since its MSE is smaller.

Computational efficiency : It should be possible to calculate the estimator in a reasonable amount of time. However, it is better to use an efficient method that takes a reasonable time but finds all the outliers than one that takes a lesser time and misses many of them.

The sample mean and the covariance matrix of the location and scatter parameters are affine equivariant but their sample breakdown point can be as low as $\frac{1}{m}$, where m is the sample size. Several multivariate robust estimators of μ and Σ have been proposed. These include the M-estimators (Maronna, 1976), the Stahel–Donoho estimators (Stahel, 1981; Donoho, 1982), the S-estimators (Rousseeuw and Yohai, 1984; Davies, 1987; Lopuhaä, 1989), and the MVE and MCD estimators (Rousseeuw, 1985). The M-estimators are computationally cheaper, but their breakdown point, under some general conditions, cannot exceed $\frac{1}{p+1}$ (Maronna, 1976; Huber, 1981), and the breakdown point reduces as the dimension increases. The Stahel–Donoho estimators are reasonably efficient and have the sample breakdown point $\frac{(m-2p+2)}{2m}$ (Donoho, 1982), but they are computationally expensive. The S-estimators can attain the sample breakdown point $\frac{(m-p+1)}{2m}$ but are also computationally expensive. The MCD and MVE estimators have the highest possible finite-sample breakdown point $\frac{(m-p+1)}{2m}$. The rate of convergence is $m^{-1/2}$ for MCD and $m^{-1/3}$ for MVE. However, these estimators have low asymptotic efficiency under normality. RMCD and RMVE

have better efficiency without compromising on the breakdown point and the rate of convergence. In the next two subsections, we discuss the MCD and MVE estimators and their re-weighted versions and the associated computational procedures.

2.2 MVE and RMVE Estimators

The MVE estimators of location and scatter of a distribution are determined by the ellipsoid of minimum volume that covers the subset of data points of size $h = \mathrm{m}^*\gamma$ where $(0.5 \leq \gamma \leq 1)$. Here $\epsilon = 1 - \gamma$ represents the breakdown point of the MVE estimators. The MVE location estimate is the geometrical center of the ellipsoid, and the MVE scatter estimate is the matrix that defines the ellipsoid, multiplied by an appropriate constant to ensure consistency (Rousseeuw and Van Zomeren, 1990; Woodruff and Rocke, 1994). Thus, the MVE estimator does not correspond to the sample mean and the sample covariance matrix of the data points that constitute the ellipsoid of minimum volume. The MVE estimator has its highest possible finite-sample breakdown point when $h = \frac{(m+p+1)}{2}$ (Davies, 1992; Loupuhaä and Rousseeuw, 1991). It has an $m^{-1/3}$ rate of convergence and a non-normal asymptotic distribution.

Calculating the exact MVE for a data set X^m would require examining all $\binom{m}{k}$ ellipsoids containing h observations of X^m to find the ellipsoid with the smallest

volume. While the MVE is interesting, finding the MVE estimator can be difficult in practice; it is essentially a two-step process. The first step is to find the best half-set consisting of h points. The second step involves finding the ellipsoid of minimum volume that covers the selected half-set. A given half-set is covered by many ellipsoids. Titterington (1975) found that the second step is equivalent to finding a D-optimal design for a design region where the points in the half-set are the design points. Thus, iterative algorithms that find D-optimal designs could be used to find the best covering ellipsoid. The first step is referred to as the subset problem, and the second step is referred to as the covering problem.

As the sample size m and the data dimension p increase, the computational effort required to find the half-sets increases exponentially. For example, if $m = 25$ and $p = 2$, so that $h = (25+2+1)/2 = 14$, then there are a total of $(28!)/(14!14!) = 40,116,600$ half-sets. When the best half-set has been found, additional calculations are needed to find the best covering ellipsoid.

Computing the MVE estimators is expensive or impossible for large sample sizes in high dimensions (Woodruff and Rocke, 1994). Rousseeuw and Leroy (1987) proposed an approximate sub-sampling algorithm to find these estimators. This algorithm considers a fixed number of random subsets, known as elemental subsets, each containing $p + 1$ points. For each elemental subset, the sample mean vector and sample

variance-covariance matrix are calculated, determining the shape of an ellipsoid. The size of this ellipsoid is then increased by multiplying by a constant until it covers at least h data points. The ellipsoid with the smallest volume is then used to obtain the MVE estimates. It has been shown that this sub-sampling algorithm retains the affine-equivariance property of the MVE estimator. Moreover, if all $\binom{m}{p+1}$ subsets of size $p+1$ are considered, then the solution of the algorithm has the same breakdown value as the exact MVE (Rousseeuw, 1985).

Croux and Haesbroeck (1997) modified the standard sub-sampling algorithm by taking the average of the solutions corresponding to several near-optimal subsets instead of considering only the optimal solution. They showed that their average solution maintains the breakdown value and has better finite-sample efficiency (Croux and Haesbroeck, 2002). Davies (1987) updated the center and scatter estimates corresponding to the best subset, using h observations in the MVE. Davies (1992) showed that the MVE estimators of location and scatter converge at rate $m^{-1/3}$ to a non-Gaussian distribution. This low rate of convergence implies that the asymptotic efficiency of the MVE estimators is 0%.

If robust multivariate estimators are to be of practical use in statistical inference they should offer reasonable efficiency under the normal model and a manageable

asymptotic distribution. A two-stage or re-weighted procedure provides both robustness and efficiency. A highly robust but perhaps inefficient estimator is first computed. This is used as a starting point to find a local solution for detecting outliers and computing the sample mean and covariance of the cleaned data set; see Rousseeuw and Van Zomeren (1990). This involves discarding those observations whose Mahalanobis distances exceed a fixed threshold.

The RMVE estimators are the weighted mean vector,

$$\bar{\mathbf{x}}_{RMVE} = \left(\sum_{i=1}^{m} w_i \mathbf{x}_i\right) / \left(\sum_{i=1}^{m} w_i\right), \tag{2.3}$$

and the weighted covariance matrix,

$$S_{RMVE} = c_{\alpha,p} * d_{\gamma,\alpha}^{m,p} * \sum_{i=1}^{m} w_i (\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE})(\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE})' / \sum_{i=1}^{m} w_i \tag{2.4}$$

where $c_{\alpha,p}$ and $d_{\gamma,\alpha}^{m,p}$ are the multiplication factor for consistency (Croux and Haesbroeck, 1999) and the finite-sample correction factor (Pison, Van Alest, and Willems, 2002). The weights are based on the robust distance :

$$RD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_{MVE})' S_{MVE}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{MVE})}. \tag{2.5}$$

The $RD(\mathbf{x}_i)$ is compared with $\sqrt{q_\alpha}$, where q_α is the (1-α)100% quantile of the chi-square distribution with p degrees of freedom, and weights are assigned to the corresponding observation as :

$$w_i = \begin{cases} 1 & \text{if } RD(\mathbf{x}_i) \leq \sqrt{q_\alpha} \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

It has been shown that the RMVE estimates do not improve on the convergence rate (and thus the 0% asymptotic efficiency) of the initial MVE estimator (Lopuhaä and Rousseeuw, 1991; Pison et al., 2002). As an alternative, a one-step M-estimator can be calculated with the MVE estimates as the initial solution (Croux and Haesbroeck, 1997; Woodruff and Rocke, 1990). This results in an estimator with the standard $m^{-1/2}$ convergence rate to a normal asymptotic distribution. This sub-sampling algorithm has been implemented in SPLUS, R, SAS, and MATLAB.

2.3 MCD and RMCD Estimators

An alternative high-breakdown estimator is based on the MCD; it was first proposed by Rousseeuw (1984). It is obtained by finding the half-set that gives the minimum value of the determinant of the variance-covariance matrix. The resulting estimator of the location is the sample mean vector of the points that are in the half-set. The estimator of the dispersion is the sample variance-covariance matrix of the points

multiplied by an appropriate constant to ensure consistency, as was done for the MVE. In contrast to the MVE, the MCD estimators correspond to the mean and covariance of a specific half-set. The MCD estimators are simple to calculate once the best half-set has been found; they do not require a solution to the covering problem.

The MCD estimators of the location and scatter of the distribution are determined by the subset of observations of size $h = \mathrm{m}^*\gamma$, where $(0.5 \leq \gamma \leq 1)$ whose covariance matrix has the smallest possible determinant. Here $\epsilon = 1-\gamma$ represents the breakdown point of the MCD estimators. The MCD location estimate $\bar{\mathbf{x}}_{MCD}$ is the average of this subset of h points. The MCD scatter estimate is given by $S_{MCD} = a_{\gamma,p} * b^n_{\gamma,p} * C_{MCD}$, where C_{MCD} is the covariance matrix of the subset, the constant $a_{\gamma,p}$ is the multiplication factor for consistency (Croux and Haesbroeck, 1999), and $b^n_{\gamma,p}$ is the finite-sample correction factor (Pison et al., 2002).

The MCD estimator has its highest possible finite-sample breakdown point when $h = \frac{(m+p+1)}{2}$. It has an $m^{-1/2}$ rate of convergence but low asymptotic efficiency under normality. Computing the exact MCD estimators ($\bar{X}_{MCD}$, S_{MCD}) is expensive or impossible for large sample sizes in high dimensions (Woodruff and Rocke, 1994) and so, as for the MVE, various approximate algorithms have been suggested. A fast algorithm was proposed independently by Hawkins and Olive (1999) and Rousseeuw

and Van Driessen (1999). The algorithm of Rousseeuw and Van Driessen, known as FAST-MCD, typically finds the exact MCD for small data sets and an approximate MCD for larger data sets. The FAST-MCD is implemented in SPLUS, R, SAS, and MATLAB.

As is the case for MVE estimators, MCD estimators are not efficient. Hence, a re-weighted version similar to that for MVE has been proposed by Rousseeuw and van Driessen (1999). This two-step procedure improves the efficiency while retaining the other properties of the MCD estimator. The asymptotic convergence rate of the MCD estimator is $m^{-1/2}$, and hence it is considered the best choice for the initial estimator of a two-step procedure. Based on the two-step approach, the RMCD estimators are

$$\bar{\mathbf{x}}_{RMCD} = \left(\sum_{i=1}^{m} w_i \mathbf{x}_i\right) / \left(\sum_{i=1}^{m} w_i\right) \tag{2.7}$$

$$S_{RMCD} = c_{\alpha,p} * d_{\gamma,\alpha}^{m,p} * \sum_{i=1}^{m} w_i (\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})' / \sum_{i=1}^{m} w_i \tag{2.8}$$

where $c_{\alpha,p}$ and $d_{\gamma,\alpha}^{m,p}$ are the multiplication factor for consistency and the finite-sample correction factor. The weights w_i= 0 or 1 are based on the robust distance as for RMVE:

$$RD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})' S_{MCD}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})}. \tag{2.9}$$

This re-weighting technique improves the efficiency of the initial MCD/MVE estimators while retaining (most of) its robustness. Hence, the RMCD/RMVE estimators

inherit the affine equivariance, robustness, and asymptotic normality properties of the MCD/MVE estimators with improved efficiency.

In this thesis, we propose using RMCD/RMVE estimators to construct robust control charts. In Chapter 3, we discuss the T^2 control chart for Phase-I with RMCD/RMVE estimators and in Chapter 4 we propose robust versions of the charts of Huwang et al. (2007) and Memar and Niaki (2009) for monitoring process variability.

Chapter 3

Robust Control Charts for Monitoring Process Mean

3.1 Robust Control Charts

As discussed in Chapter 1, outliers in the Phase-I sample may unduly influence the performance of the Hotelling's T^2 chart. The use of RMCD/RMVE estimators will make the standard T^2 chart robust. We propose using T^2 charts with robust estimators of the location and dispersion parameters to monitor changes in the mean vector when individual multivariate observations are considered. The RMCD/RMVE estimators inherit the properties of MCD/MVE estimators such as affine equivariance,

robustness, and asymptotic normality while achieving higher efficiency. We now define robust T^2 statistic using RMCD/RMVE estimators for the i-th multivariate observation $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ as

$$T^2_{RMCD}(\mathbf{x}_i) = (\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})' S^{-1}_{RMCD}(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})$$
$$T^2_{RMVE}(\mathbf{x}_i) = (\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE})' S^{-1}_{RMVE}(\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE}) \tag{3.1}$$

where $\bar{\mathbf{x}}_{RMCD}$ and $\bar{\mathbf{x}}_{RMVE}$ are the location estimators and S_{RMCD} and S_{RMVE} are the scatter estimators under the RMCD/RMVE methods based on m multivariate observations. The exact distribution of T^2 is not available, so the control limits for Phase-I data are obtained by inverting the empirical distribution of the T^2 values. In the next section we use Monte Carlo simulation to estimate the quantiles of the distributions of T^2_{RMCD} and T^2_{RMVE} for several sample sizes and dimensions. As will be seen shortly, the choice of T^2_{RMCD} or T^2_{RMVE} depends on the situation. For each dimension, we fit a smooth nonlinear model to find the control limits for a given sample size.

3.2 Computation of Control Limits

We performed a large number of Monte Carlo simulations to obtain the control limits. The limits are found by inverting the empirical distribution of T^2_{RMCD} and T^2_{RMVE}.

We generated n = 200,000 samples of size m from a standard multivariate normal distribution MVN(0, I_p) with dimension p. Because of the invariance of the T^2_{RMCD} and T^2_{RMVE} statistics, these limits are applicable for any values of μ and Σ.

Using the re-weighted MCD/MVE estimators $\bar{X}_{RMCD}$, S_{RMCD}, $\bar{X}_{RMVE}$, and S_{RMVE} with breakdown value γ=0.50, we calculated T^2 statistics for each observation in the data set using Eq. (3.1) and recorded the maximum value attained for each data set. We inverted the empirical distribution of the maximum of T^2_{RMCD} and T^2_{RMVE} to find the $(1-\alpha)100\%$ quantiles. We used the R function CovMcd() in the rrcov package to find the RMCD/RMVE estimates.

We found the quantiles of T^2 for m=(30, 31,..., 50, 55,..., 100, 110,..., 150, 175, 200) and p=(2, 3,...10) and derived the quantiles for α = (0.05, 0.01, 0.001). Scatter plots of the quantiles and the sample sizes for different dimensions suggest a family of nonlinear models of the form :

$$T^2_{(p,1-\alpha)} = a_{1(p,1-\alpha)} + \frac{a_{2(p,1-\alpha)}}{m^{a_{3(p,1-\alpha)}}}. \tag{3.2}$$

where $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}$ and $a_{3(p,1-\alpha)}$ are the model parameters which depends on the values of p and α. The parameters can be estimated for various values of p and α using the method of least squares.

The scatter plots of the actual and fitted values of the quantiles of T^2_{RMCD} and T^2_{RMVE} for p = 2, 6, and 10 and α =0.05 and 0.01 are given in Figs. 3.1, 3.2, and 3.3.

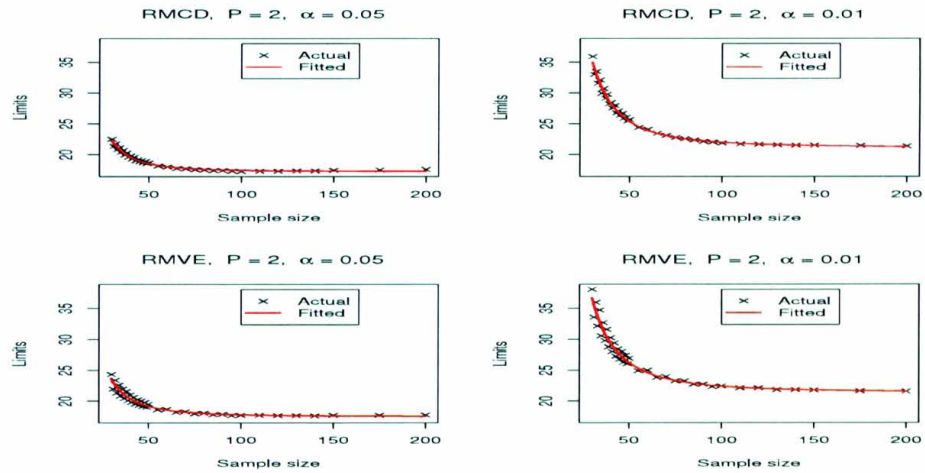

Figure 3.1: Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 2

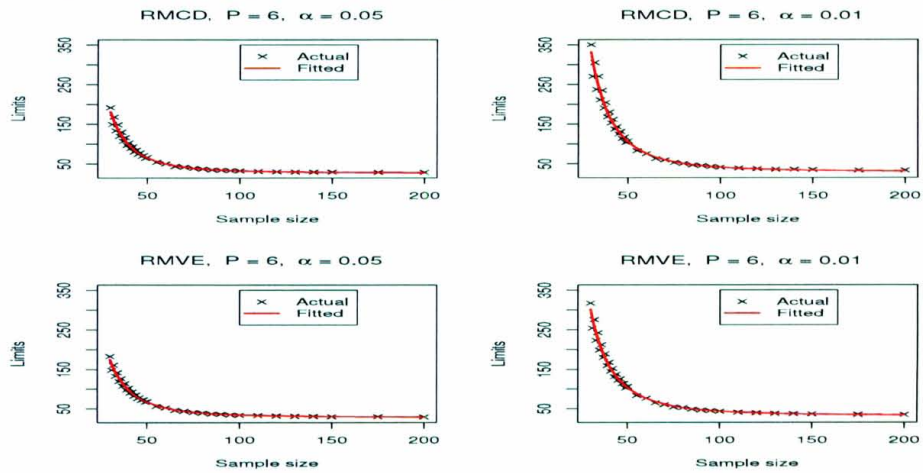

Figure 3.2: Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 6

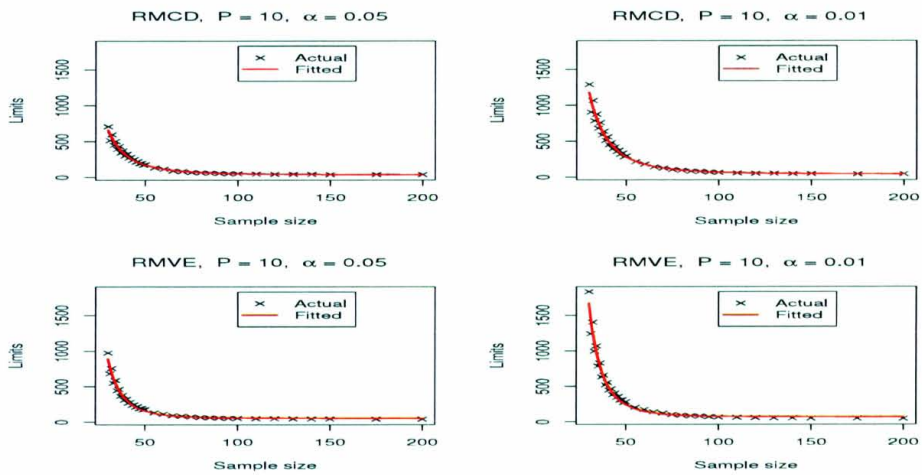

Figure 3.3: Scatter plot of T^2_{RMCD}/T^2_{RMVE} control limits and fitted curve for p = 10

The figures show that the non-linear fit is good, which helps us to find the T^2_{RMCD} and T^2_{RMVE} control limits for any given sample size and given values of p and α using Eq. (3.1) if the model parameters are available. The least-square estimates of the parameters $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}$ and $a_{3(p,1-\alpha)}$ for dimensions p=(2, 3,...,10) and for α = (0.05, 0.01, 0.001) corresponding to T^2_{RMCD} and T^2_{RMVE} charts respectively are given in Tables 3.1 and 3.2.

Table 3.1: Estimates of model parameters $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}, a_{3(p,1-\alpha)}$ for dimensions p = (2,...,10) and confidence levels α = (0.05, 0.01, 0.001) for T^2_{RMCD} control charts

	$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.001$		
p	$\hat{a}_{1,p,0.95}$	$\hat{a}_{2,p,0.95}$	$\hat{a}_{3,p,0.95}$	$\hat{a}_{1,p,0.99}$	$\hat{a}_{2,p,0.99}$	$\hat{a}_{3,p,0.99}$	$\hat{a}_{1,p,0.999}$	$\hat{a}_{2,p,0.999}$	$\hat{a}_{3,p,0.999}$
2	17.223	41102	2.647	21.134	38170	2.329	27.051	192909	2.508
3	20.134	35844	2.209	24.287	128924	2.344	31.350	1144947	2.718
4	23.152	269357	2.548	28.181	1272773	2.773	35.575	5989325	2.973
5	24.685	467949	2.524	28.437	1417059	2.632	31.013	2666196	2.593
6	26.962	1762051	2.746	29.654	3061216	2.711	31.662	5414248	2.669
7	24.892	1099128	2.493	22.882	1585224	2.416	19.058	3465278	2.444
8	27.236	2908821	2.667	27.245	4922576	2.644	28.326	12134778	2.710
9	23.974	2447649	2.534	21.420	4726835	2.554	18.772	14096595	2.676
10	31.894	12572909	2.914	37.085	34375654	3.033	56.573	172176786	3.301

Table 3.2: Estimates of model parameters $a_{1(p,1-\alpha)}, a_{2(p,1-\alpha)}, a_{3(p,1-\alpha)}$ for dimensions p = (2,...,10) and confidence levels 1-α = (95%, 99%, 99.9%) for T^2_{RMVE} control charts

	$\alpha = 0.05$			$\alpha = 0.01$			$\alpha = 0.001$		
p	$\hat{a}_{1,p,0.95}$	$\hat{a}_{2,p,0.95}$	$\hat{a}_{3,p,0.95}$	$\hat{a}_{1,p,0.99}$	$\hat{a}_{2,p,0.99}$	$\hat{a}_{3,p,0.99}$	$\hat{a}_{1,p,0.999}$	$\hat{a}_{2,p,0.999}$	$\hat{a}_{3,p,0.999}$
2	17.442	29553	2.494	21.365	31571	2.244	27.594	148747	2.434
3	20.286	22497	2.066	24.387	59096	2.130	31.326	338665	2.402
4	23.095	108855	2.286	27.549	291064	2.372	35.109	1255429	2.576
5	24.796	238966	2.334	28.302	508097	2.367	32.008	1063783	2.377
6	27.585	1041090	2.606	31.126	1882888	2.601	37.136	4714353	2.671
7	28.151	1541634	2.598	30.936	3183762	2.635	39.357	12199414	2.827
8	34.917	14798692	3.127	45.767	75616029	3.419	70.875	840512379	3.904
9	39.191	59094377	3.415	50.271	275604839	3.679	72.768	1960966919	4.039
10	50.733	950607720	4.099	68.154	4696452032	4.379	110.587	56398461817	4.881

3.3 Performance Analysis

We assess the performance of the proposed charts when outliers are present in the data due to the shift in the process mean. Jensen et al. (2007) concluded that MVE/MCD-based T^2 control charts perform well in terms of detecting outliers due to shift in process mean. Therefore, we compared the performance of our methods with MCD/MVE-based T^2 charts and standard T^2 charts. For each combination of p, m, and π, we generated a number of data sets. Of the m observations, $m \times \pi$ are random data points generated from the out-of-control distribution, and the remaining $m \times (1 - \pi)$ are generated from the in-control distribution so that the sample of m data points may contain some outliers. We set π to 0.20 to ensure that the sample contains a few outliers. Without loss of generality, we consider the in-control distribution to be N$(0,I_p)$. The out-of-control distribution is a multivariate normal with a small shift in the mean vector and with covariance matrix I_p. The mean shift is defined by a non-centrality parameter (δ), which is given by

$$\delta = (\mu_1 - \mu)' \Sigma^{-1} (\mu_1 - \mu) \tag{3.3}$$

where $(\mu_1 - \mu)$ is the shift in the mean vector. We calculated the proportion of data sets with at least one T^2_{RMCD} (or T^2_{RMVE}) statistic greater than the control limit; this is the estimated probability of signal for detecting outliers. We compared the

performance of these charts with T^2 charts with MCD and MVE estimators and the standard T^2 chart. We considered the probability of signal for different values of $\delta = (0, 5, 10, \cdots, 30)$, $m = (30,\ 50,\ 100,\ 150)$, $p = (2,\ 6,\ 10)$, and $\pi = (10\%, 20\%)$. We generated 50,000 data sets of size m for each combination of m, p, π, and δ and the probability of signal was estimated for $\alpha = 0.05, 0.01$, and 0.001. Figures 3.4 to 3.15 show the probability of signal for α =0.05 and 0.01 and different values of p and m.

Figures 3.4 to 3.15 show that when the value of the non-centrality parameter is zero or close to zero, the probability of a signal is close to α, as expected for an in-control process. As the value of the non-centrality parameter increases the probability of a signal also increases. Using this, we select the best method for identifying outliers. If the probability of a signal does not increase for an increase in the non-centrality parameter, then the estimator has broken down and is not capable of detecting outliers.

A careful examination of Figs. 3.4 to 3.15 shows that, for small values of p and m, T^2_{RMVE} performs well. As m and p increase, the T^2_{RMCD} chart is superior. For example, from Figs. 3.4, 3.5, 3.6, and 3.7 we see that T^2_{RMVE} has a slight advantage over T^2_{RMCD}. All the plots show that the T^2_{RMCD}/T^2_{RMVE} charts perform better than the T^2_{MCD}/T^2_{MVE} charts. When p is large, T^2_{RMCD} has a clear advantage over T^2_{RMVE};

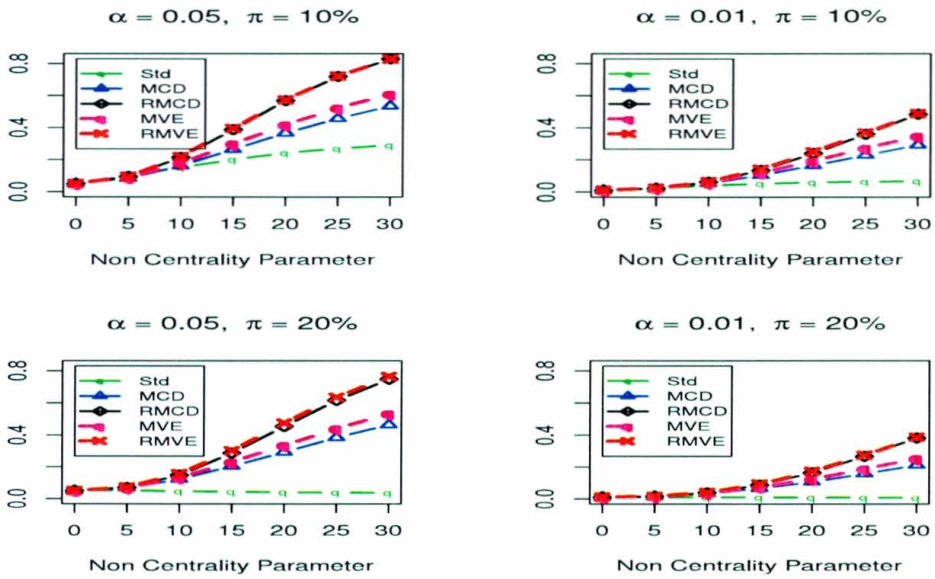

Figure 3.4: Probability of signal for RMCD/RMVE control limits for p= 2, m= 30

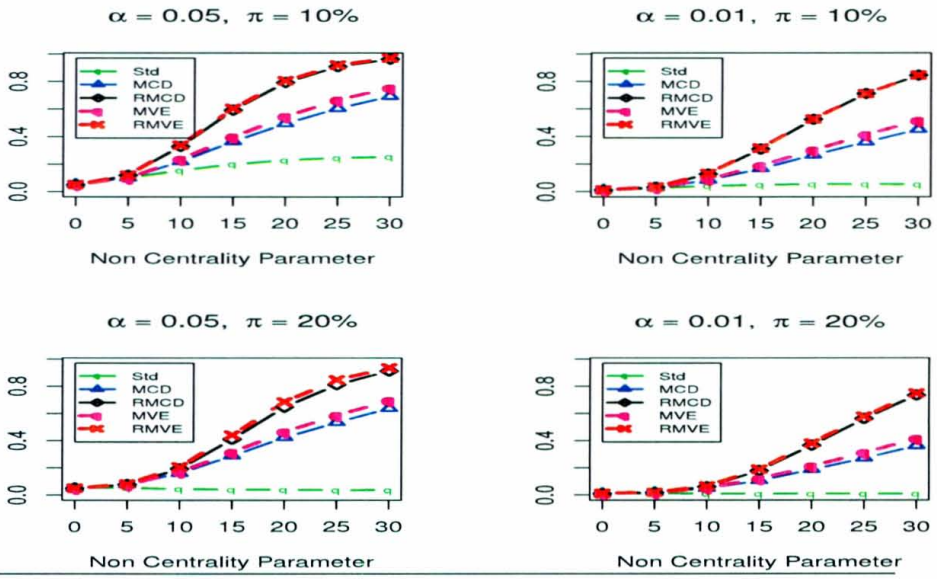

Figure 3.5: Probability of signal for RMCD/RMVE control limits for p= 2, m= 50

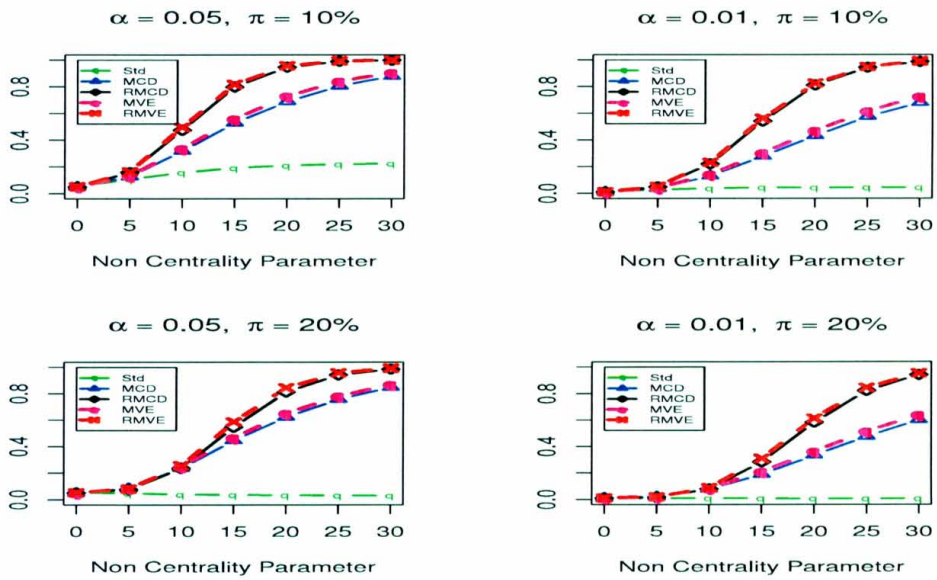

Figure 3.6: Probability of signal for RMCD/RMVE control limits for p= 2, m= 100

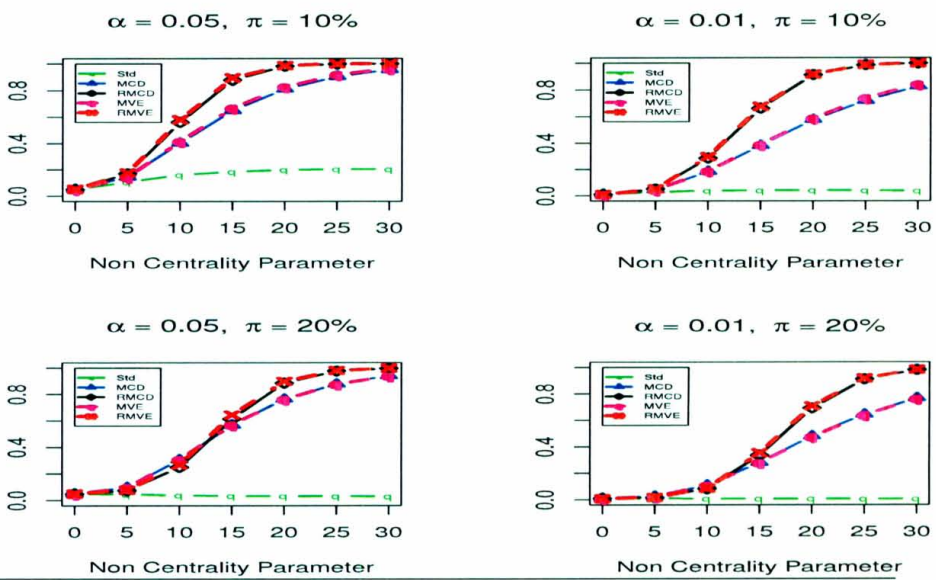

Figure 3.7: Probability of signal for RMCD/RMVE control limits for p= 2, m= 150

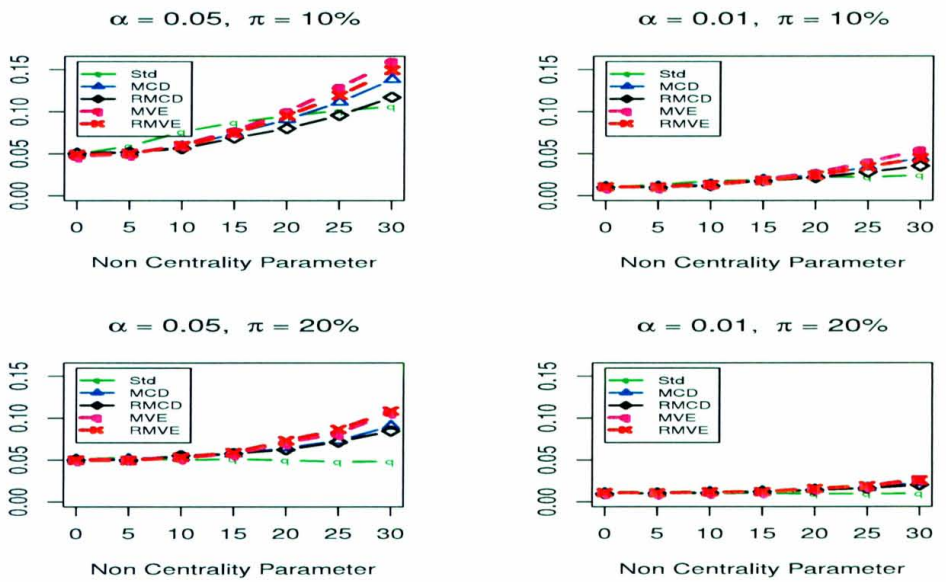

Figure 3.8: Probability of signal for RMCD/RMVE control limits for p= 6, m= 30

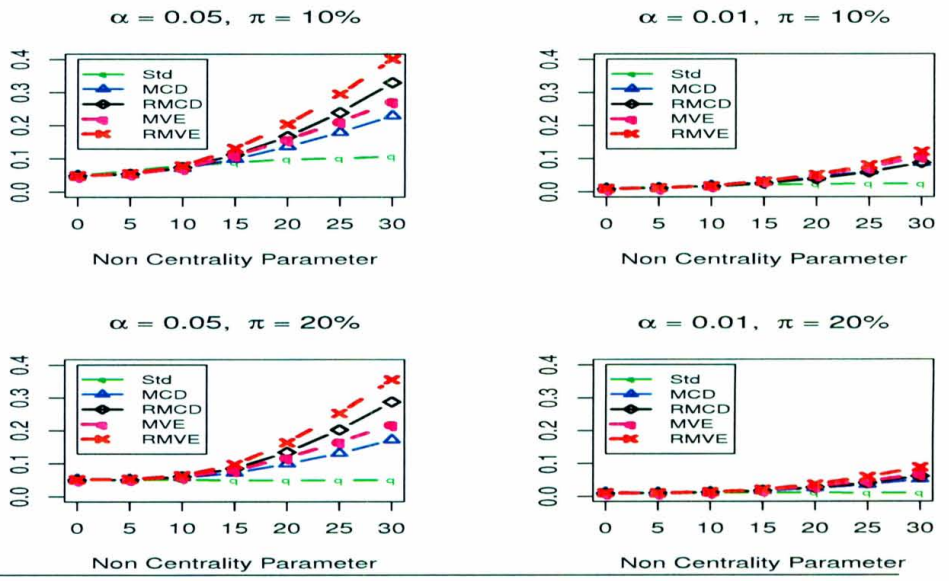

Figure 3.9: Probability of signal for RMCD/RMVE control limits for p= 6, m= 50

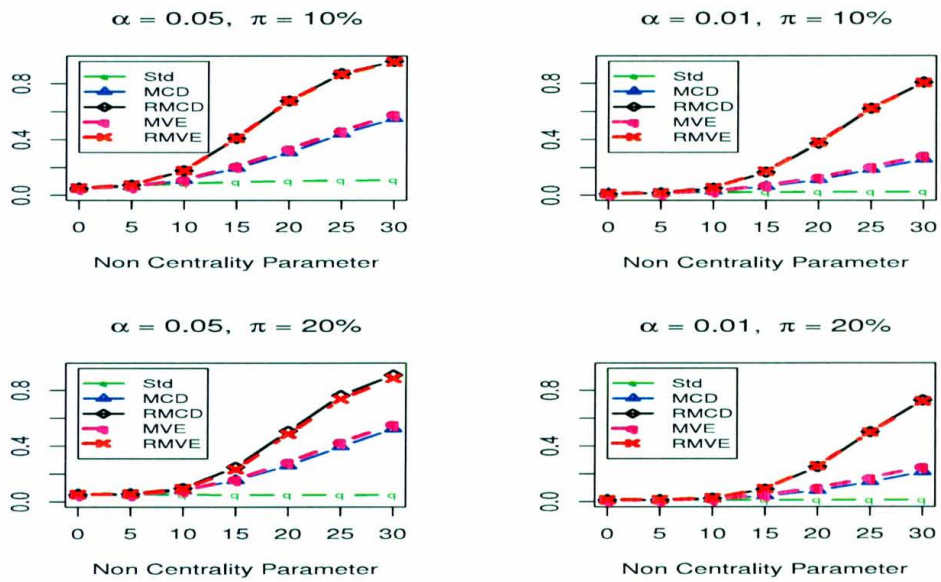

Figure 3.10: Probability of signal for RMCD/RMVE control limits for p= 6, m= 100

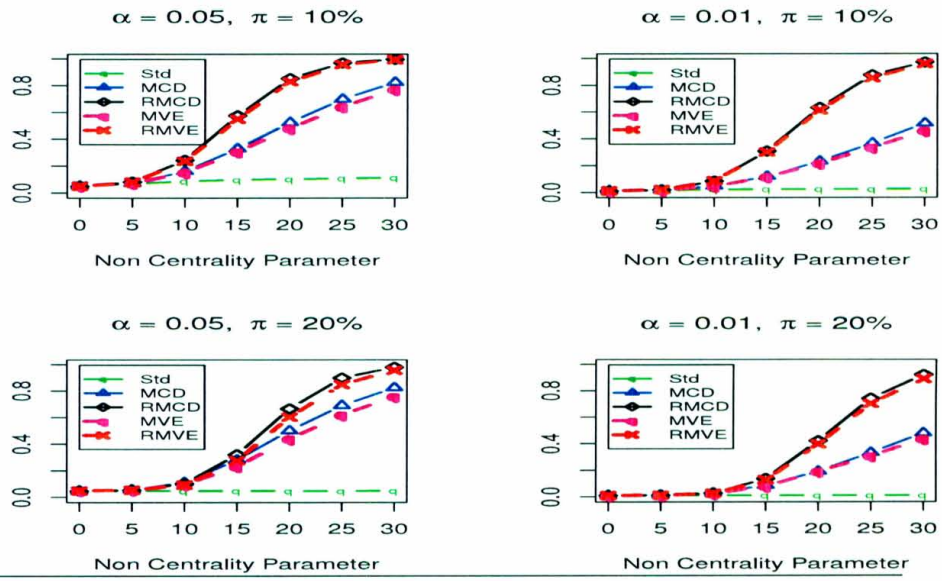

Figure 3.11: Probability of signal for RMCD/RMVE control limits for p= 6, m= 150

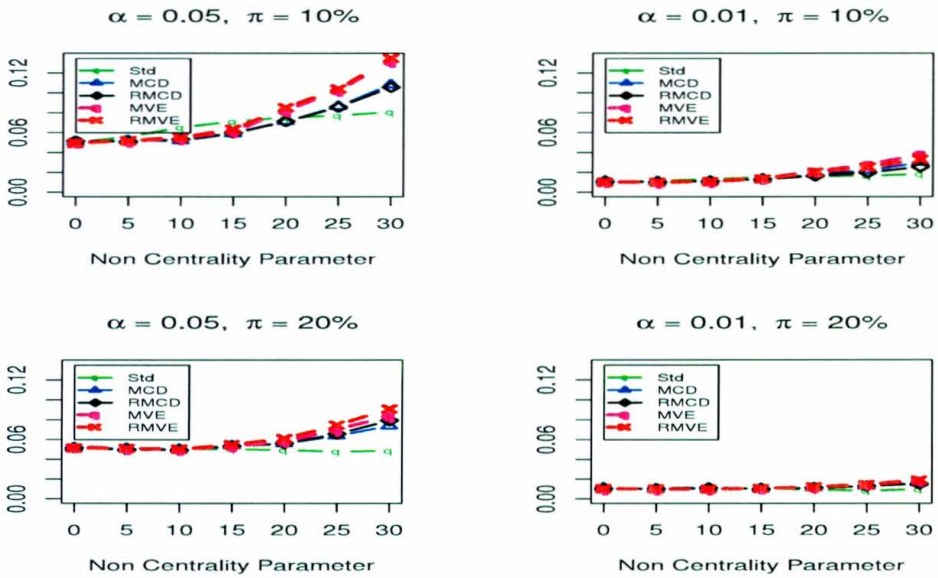

Figure 3.12: Probability of signal for RMCD/RMVE control limits for p= 10, m= 30

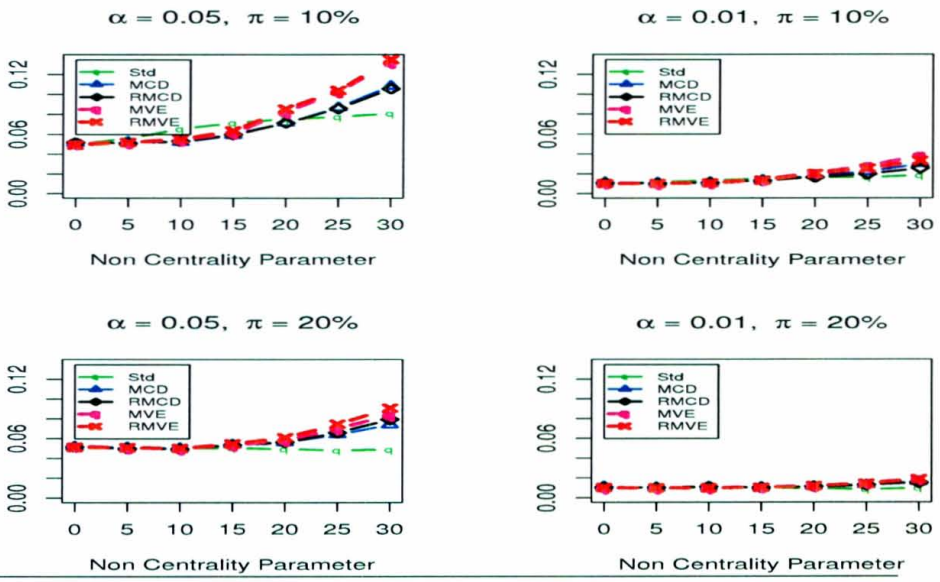

Figure 3.13: Probability of signal for RMCD/RMVE control limits for p= 10, m= 50

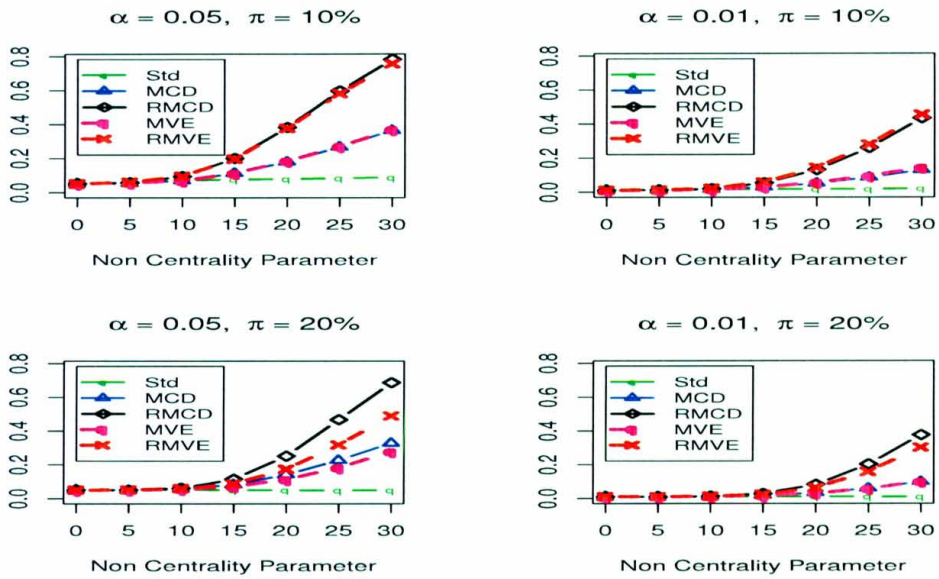

Figure 3.14: Probability of signal for RMCD/RMVE control limits for p= 10, m= 100

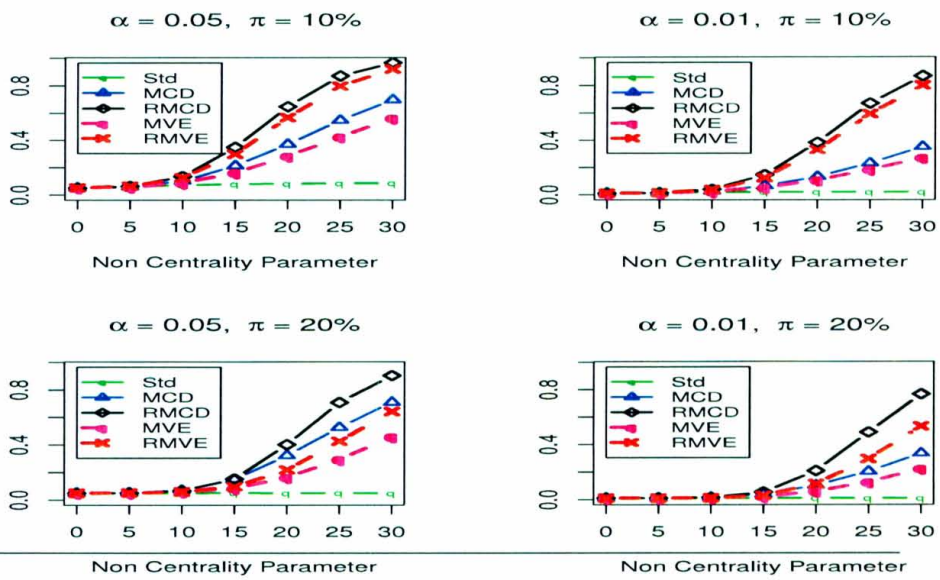

Figure 3.15: Probability of signal for RMCD/RMVE control limits for p= 10, m= 150

see Figs. 3.14 and 3.15. It is clear that the standard T^2 control chart has a limited ability to detect outliers, and T^2_{MCD} and T^2_{MVE} do not perform compared to the charts based on the re-weighted estimators.

As p increases for a fixed value of m, the breakdown points of RMCD and RMVE estimators decrease since the breakdown value is given by $\frac{(m-p+1)}{2m}$. This suggests that the larger the value of p, the larger m will need to be to maintain the breakdown point; this is demonstrated in Figs. 3.14 and 3.15. For dimensions 6 and 10, $m = 30$ or 50 is too small to detect outliers; see Figs. 3.8, 3.9, 3.12, and 3.13. In general, either RMCD or RMVE was superior for all the values of the non-centrality parameter, provided the proportion of outliers was not so high that the estimators broke down. This greatly simplifies the conclusions about when RMCD or RMVE estimators are preferred to MCD or MVE estimators.

When $m < 100$, T^2_{RMVE} is the best choice for small dimensions. When $m \geq 100$, T^2_{RMCD} is preferred. As p increases, the percentage of outliers that can be detected by T^2_{RMVE} decreases. For both charts, the higher the value of p, the lower the number of outliers that can be detected for smaller sample sizes. For Phase-I applications where the number of outliers is unknown, T^2_{RMVE} should be used only for smaller sample sizes, and it is also computationally feasible. T^2_{RMCD} should be used for larger sample sizes or when it is believed that the number of outliers is large. When the dimension

is large, larger sample sizes are needed to ensure that the estimator does not break down. Hence, for larger dimensions, T^2_{RMCD} is preferred with large sample sizes.

Chapter 4

Robust Control Charts for Monitoring Process Variability

We have seen that the robust versions of Hotelling's T^2 charts are good for monitoring the process mean for both multivariate observations with subgroup data and individual observations. To monitor multivariate process variability, control charts based on either the generalized variance (the determinant of the sample covariance matrix) or the likelihood ratio test for testing the equality of covariance matrices are generally used (Alt and Smith, 1988; Levinson et al., 2002). For these charts, the subgroup size should be larger than the number of quality characteristics of interest to ensure that the sample covariance matrix has full rank.

For individual observations, none of these procedures are applicable because the sample covariance matrix is not defined. The monitoring of multivariate process variability for individual observations has received little attention in the literature, although it is often more critical for improving the quality of manufacturing processes by reducing the variability rather than the detection of process mean shifts. However, based on the regression-adjusted variables, Hawkins (1981, 1991) developed control charts for univariate observations to monitor the process mean and extended them to monitor the process variability. Woodal and Ncube (1985) extended this to individual multivariate observations via multiple CUSUM and EWMA charts that combined p univariate charts. However, multiple charts are not effective if the quality characteristics are correlated. MacGregor and Harris (1993) developed exponentially weighted mean square error (EWMS) and exponentially weighted moving variance (EWMV) charts for individual univariate observations to detect changes in the process variability. This concept was extended by Huwang et al. (2007) to individual multivariate observations. They developed the MEWMS and MEWMV control charts. A brief review of these charts is given below.

4.1 MEWMS Control Charts

Let the random vector $g = (g_1, g_2, \cdots, g_p)'$ represent the process data with p quality characteristics, which follows a multivariate normal distribution with mean μ_g and covariance matrix Σ_g. It is assumed that the estimators of the parameters are either known or estimated from the Phase-I analysis of the in-control process with $\mu_g = \mu_0$ and $\Sigma_g = \Sigma_0$. Consider the transformation of the process variable $\mathbf{g}$ to $\mathbf{x}$, so that $\mathbf{x}$ follows a multivariate normal distribution with mean μ and covariance matrix Σ as defined below:

$$\begin{aligned} \mathbf{x} &= \Sigma_0^{-1/2}(g - \mu_0) \\ \mu &= \Sigma_0^{-1/2}(\mu_g - \mu_0) \\ \Sigma &= \Sigma_0^{-1/2}\Sigma_g\Sigma_0^{-1/2}. \end{aligned} \tag{4.1}$$

Obviously, for an in-control process the distribution of $\mathbf{x}$ is N(0, I_p), where I_p is a p×p identity matrix.

For individual observations, although the sample covariance matrix is not available, the matrix $\mathbf{x}\mathbf{x}'$ of each observation provides an unbiased estimator of Σ when the process mean does not shift (i.e., $\mu = 0$). However, $\mathbf{x}\mathbf{x}'$ is not a positive definite matrix. Hence, for the t-th individual observation $\mathbf{x}_t = (x_{t1}, x_{t2}, ...x_{tp})'$, Huwang et

al. (2007) defined the multivariate exponentially weighted moving average as

$$S_t = \omega \mathbf{x}_t \mathbf{x}_t^{'} + (1-\omega) S_{t-1} \ , \ t = 1, 2, 3... \tag{4.2}$$

where ω is a smoothing constant, $0<\omega<1$, and $S_0 = \mathbf{x}_1 \mathbf{x}_1^{'}$.

This can be simplified to

$$S_t = \sum_{i=1}^{t} c_i \mathbf{x}_i \mathbf{x}_i^{'} = X^{'} C X \tag{4.3}$$

where $c_1 = (1-\omega)^{t-1}$, $c_i = \omega(1-\omega)^{t-i}, i = 2, \cdots, t$, $\sum_{i=1}^{t} c_i = 1$, $X = (x_1, x_2, \cdots, x_t)^{'}$, and $C = diag(c_1, c_2, \cdots, c_t)$.

Huwang et al. (2007) showed that if the mean vector does not shift, S_t is positive definite for $t \geq p$ with probability 1 and $E(S_t) = \Sigma$. One way to measure the overall variability in a covariance matrix is to reduce the matrix to a single summary statistic. Two commonly used statistics are the determinant and the trace. Since the trace represents the total variation of the p quality characteristics of the covariance matrix, Huwang et al. (2007) proposed using the trace of S_t to monitor the changes in Σ. They showed that the trace of S_t can be written

$$tr(S_t) = \sum_{i=1}^{t} c_i tr(\mathbf{x}_i \mathbf{x}_i^{'}) = \sum_{i=1}^{t} c_i \left(\sum_{j=1}^{p} \mathbf{x}_{ij}^2 \right). \tag{4.4}$$

The mean and variance of $\mathrm{tr}(S_t)$ are p and $2p \sum_{i=1}^{t} c_i^2$ respectively, where $\sum_{i=1}^{t} c_i^2 = \frac{\omega}{(2-\omega)} + \frac{2-2\omega}{2-\omega}(1-\omega)^{2(t-1)}$ which will converge to $\frac{\omega}{(2-\omega)}$ as $\mathrm{t} \to \infty$. The control limits

for the MEWMS control chart are

$$E[tr(S_t)] \pm L\sqrt{Var[tr(S_t)]} = p \pm L\sqrt{2p\sum_{i=1}^{t} c_i^2} \tag{4.5}$$

where the value of L can be found by Monte Carlo simulation based on the in-control average run length (ARL_0). Huwang et al. (2007) found the value of L by simulation for p = 2,3, $\omega = 0.1, 0.2, .., 0.9$, and $ARL_0 = 370$.

4.2 MEWMV Control Charts

The MEWMS chart is designed to monitor the covariance matrix under the assumption that the process mean does not shift. However, the mean and the variability can vary simultaneously during the monitoring period. Therefore, it is desirable to construct a chart that can detect both changes in the process variability and shifts in the process mean.

Huwang et al. (2007) proposed the MEWMV chart based on the statistic V_t. The construction of V_t is similar to that of S_t except that the deviation of $\mathbf{x}_t$ is taken from $\mathbf{y}_t$, a predicted value of the mean shift at sampling point t. They defined V_t to be

$$V_t = \omega(\mathbf{x}_t - \mathbf{y}_t)(\mathbf{x}_t - \mathbf{y}_t)' + (1-\omega)V_{t-1},\ t = 1, 2, 3, ... \tag{4.6}$$

where ω is a smoothing constant, $0<\omega<1$, and $V_0 = (\mathbf{x}_1 - \mathbf{y}_1)(\mathbf{x}_1 - \mathbf{y}_1)'$.

The $\mathbf{y}_t$ values, the predicted mean shifts at sampling point t, are obtained by the multivariate exponentially weighted moving average of $\mathbf{x}_t$ proposed by Lowry et al. (1992). Huwang et al. (2007) defined $\mathbf{y}_t$ to be

$$\mathbf{y}_t = \lambda \mathbf{x}_t + (1-\lambda)\mathbf{y}_{t-1}, \ t = 1, 2, 3, ... \tag{4.7}$$

where $\mathbf{y}_0$=0, and λ is a smoothing constant (0<λ<1).

They showed that when $t \geq p$, V_t is a positive definite matrix with probability 1 and $E(V_t) \longrightarrow \frac{2(1-\lambda)^2}{(2-\lambda)}\Sigma$ as $t \to \infty$ so $\frac{2-\lambda}{2(1-\lambda)^2}V_t$ can be used to estimate Σ. Finding the mean and variance of V_t is not as easy as it was for S_t, and hence V_t is expressed in matrix form:

$$\begin{aligned} V_t &= (X-Y)'C(X-Y) \\ &= X'(I_t - M)'C(I_t - M)X \\ &= X'QX \end{aligned} \tag{4.8}$$

where I_t is a $t \times t$ identity matrix and X, Y, M, C, and Q are given by

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_t \end{pmatrix}_{t\times 1}, \quad Y = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdots \\ \mathbf{y}_t \end{pmatrix}_{t\times 1}, \quad M = \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ \lambda(1-\lambda) & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(1-\lambda)^{t-1} & \cdots & \lambda(1-\lambda) & \lambda \end{pmatrix}_{t\times t}$$

$C_{t\times t} = \text{diag}(c_1, c_2, \cdots, c_t)$, and $Q_{t\times t} = [q_{ij}] = (I_t - M)'C(I_t - M)$.

Therefore, the trace of V_t can be simplified to

$$tr(V_t) = tr(X'QX) = tr(QXX') = \sum_{i=1}^{t}\sum_{j=1}^{t} q_{ij} \sum_{k=1}^{p} x_{ik}x_{jk}. \tag{4.9}$$

The mean and variance of tr(V_t) are $p \times tr(Q)$ and $2p \times \sum_{i=1}^{t}\sum_{j=1}^{t} q_{ij}^2$ respectively. Thus, the control limits for the MEWMV chart are

$$E[tr(V_t)] \pm L\sqrt{Var[tr(V_t)]} = p \times tr(Q) \pm L\sqrt{2p\sum_{i=1}^{t}\sum_{j=1}^{t} q_{ij}^2} \tag{4.10}$$

where the value of L can be found using Monte Carlo simulation based on ARL_0. Huwang et al. (2007) found the values of L for p=2,3, $\omega = 0.1, 0.2, .., 0.9$, and $\lambda = 0.1, 0.2, .., 0.9$, and $ARL_0 = 370$.

They compared the performance of MEWMS and MEWMV charts with that of multiple CUSUM and EWMA charts (Hawkins, 1981, 1991). They used the regression adjusted variables method based on the out-of-control average run length (ARL_1); ARL_0 was set to be the same in every case. A bivariate normal process was considered with mean μ and covariance matrix Σ:

$$\mu = \begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The following shift scenarios were considered for an out-of-control situation:

- σ_1^2 =1.00 and σ_2^2 set to 1.00, 1.25, 1.50, 1.75, 2.00, 2.50, and 3.00.

- Correlation coefficient ρ set to 0, 0.25, 0.50, and 0.75.

- Shift in mean (μ_1 or μ_2) set to 0, 0.25, 0.50, 1.00, 2.00, and 3.00.

They found that the MEWMS chart outperforms the multiple CUSUM, multiple EWMA, and MEWMV charts when there is no location shift and when $\omega \leq 0.4$ in the cases where σ_1^2, σ_2^2, and ρ change. The MEWMV chart outperforms the multiple CUSUM and EWMA charts for $\omega \leq 0.2$ and $\lambda \leq 0.4$ and for smaller shifts in σ_2^2. If ρ changes while σ_1^2 and σ_2^2 are constant, the MEWMS and MEWMV charts outperform the multiple CUSUM and EWMA charts. However, if there is a location shift, only MEWMV charts can be used ($\omega, \lambda \leq 0.4$) since the MEWMS, multiple CUSUM, and EWMA charts are sensitive to location shifts.

4.3 Control Charts Based on L_c-norm Function

The trace of the estimator of the covariance matrix was used to derive the MEWMS and MEWMV charts. However, in many out-of-control instances some of the diagonal elements of the covariance matrix increase while others decrease. In these instances, the trace of the shifted covariance matrix will not have any considerable deviation from that of the in-control covariance matrix. The MEWMS and MEWMV methods can not detect such situations.

To overcome this problem, instead of trace, Memar and Niaki (2009) suggested using the sum of the absolute values or the sum of the squares of the deviation of each diagonal element from its target value. For $c \geq 1$, they defined the L_c-norm function for a vector $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p)'$of length p as $\| \mathbf{z} \|_c = (\sum_{i=1}^{p} | \mathbf{z} |^c)^{1/c}$. Using this L_c-norm function, Memar and Niaki (2009) modified the charts of Huwang et al. (2007) to overcome the problem of the in-control and out-of-control covariance matrices having the same trace. They proposed control charts named $MEWMSL_1$, $MEWMSL_2$, $MEWMVL_1$, and $MEWMVL_2$, based on the L_1-norm and L_2-norm, to improve the performance of the MEWMS and MEWMV charts.

They defined variables similar to those of Huwang et al. (2007) by transforming the process variable $\mathbf{g}$ to $\mathbf{x}$ so that $\mathbf{x} \sim \mathrm{N}(\mu, \Sigma)$ if the process is out of control and $\mathbf{x} \sim \mathrm{N}(0, I_p)$ if the process is in control. Let Σ_{ii} denote the ith diagonal element of the covariance matrix Σ of dimension $p \times p$. Then the vector of diagonal elements of Σ is $(\Sigma_{11}, \Sigma_{22}, \cdots \Sigma_{pp})'$ and the diagonal elements of I_p are $(1, 1, \cdots, 1)' = 1_p$. The L_1-norm and L_2-norm distances between the vector of diagonal elements of Σ and its expected value 1_p are labeled $D_1(\Sigma)$ and $D_2(\Sigma)$ respectively and are given by

$$D_1(\Sigma) = \| (\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{pp})' - 1_p \| = \sum_{i=1}^{p} | \Sigma_{ii} - 1 | \tag{4.11}$$

$$D_2(\Sigma) = \| (\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{pp})' - 1_p \|^2 = \sum_{i=1}^{p} (\Sigma_{ii} - 1)^2. \tag{4.12}$$

$D_1(\Sigma)$ and $D_2(\Sigma)$ are equal to zero when the process is in-control, and they have positive values when the process is out-of-control. This allows us to monitor the variability of individual observations. Since $E(S_t) = \Sigma$, Memar and Niaki (2009) introduced the $MEWMSL_1$ and $MEWMSL_2$ charts based on $D_1(S_t)$ and $D_2(S_t)$ with the MEWMS scheme:

$$D_1(S_t) = \| (S_{t(11)}, S_{t(22)}, \cdots, S_{t(pp)})' - 1_p \| = \sum_{i=1}^{p} | S_{t(ii)} - 1 | \tag{4.13}$$

$$D_2(S_t) = \| (S_{t(11)}, S_{t(22)}, \cdots, S_{t(pp)})' - 1_p \|^2 = \sum_{i=1}^{p} (S_{t(ii)} - 1)^2. \tag{4.14}$$

As for MEWMS charts, the control limits can be found using Monte Carlo simulation based on ARL_0. Since $D_1(S_t)$ and $D_2(S_t)$ are always non-negative, only upper control limits are considered. Similarly, $E(V_t) = \frac{2(1-\lambda)^2}{2-\lambda}\Sigma$ for large values of t and hence the $MEWMVL_1$ and $MEWMVL_2$ charts based on $D_1(V_t)$ and $D_2(V_t)$ with the MEWMV scheme are

$$\begin{aligned} D_1(V_t) &= \| (V_{t(11)}, V_{t(22)}, \cdots, V_{t(pp)})' - 2(1-\lambda)^2/(2-\lambda)1_p \| \\ &= \sum_{i=1}^{p} | V_{t(ii)} - 2(1-\lambda)^2/(2-\lambda) | \end{aligned} \tag{4.15}$$

$$\begin{aligned} D_2(V_t) &= \| (V_{t(11)}, V_{t(22)}, \cdots, V_{t(pp)})' - 2(1-\lambda)^2/(2-\lambda)1_p \|^2 \\ &= \sum_{i=1}^{p} [V_{t(ii)} - 2(1-\lambda)^2/(2-\lambda)]^2. \end{aligned} \tag{4.16}$$

V_t can be transformed to $V_t^* = \frac{2-\lambda}{2(1-\lambda)^2} V_t$ and the process variability can be monitored with respect to V_t^* as given below:

$$D_1(V_t^*) = \| \left(V_{t(11)}^*, V_{t(22)}^*, \cdots, V_{t(pp)}^*\right)' - 1_p \| = \sum_{i=1}^{p} \mid V_{t(ii)}^* - 1 \mid \tag{4.17}$$

$$D_2(V_t^*) = \| \left(V_{t(11)}^*, V_{t(22)}^*, \cdots, V_{t(pp)}^*\right)' - 1_p \|^2 = \sum_{i=1}^{p} (V_{t(ii)}^* - 1)^2. \tag{4.18}$$

Here too, upper control limits are found by Monte Carlo simulation based on p, ω, λ, and ARL_0. Memar and Niaki (2009) tabulated the UCLs for all four charts for $\omega = (0.1, 0.2, ..., 0.9)$, $\lambda = (0.1, 0.2, ..., 0.9)$, and p=2 and 3.

Memar and Niaki (2009) compared the performance of the $MEWMSL_1$, $MEWMSL_2$, $MEWMVL_1$, and $MEWMVL_2$ charts with the MEWMS and MEWMV charts in terms of the ARL criterion. All scenarios are considered in the bivariate case with $ARL_0 = 370$ and different values for the process mean vector μ and the covariance matrix Σ:

$$\mu = \begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The following scenarios were considered for an out-of-control situation:

- σ_1^2 =1.00 and σ_2^2 set to 1.00, 1.50, and 2.00.
- Correlation coefficient ρ set to 0, 0.10, and 0.90.
- Shift in mean (μ_1 or μ_2) set to 0, 0.5, 1, and 1.50.

Memar and Niaki (2009) set ω=(0.1, 0.2, 0.3, 0.4) for the $MEWMS$, $MEWMSL_1$, and $MEWMSL_2$ charts and ω=(0.1, 0.2, 0.3, 0.4) and λ=(0.1, 0.2, 0.3, 0.4) for the $MEWMV$, $MEWMVL_1$, and $MEWMVL_2$ charts. When the process standard deviation shifts, whether or not the correlation coefficient changes, the $MEWMSL_1$ and $MEWMSL_2$ charts perform better than the MEWMS chart, and the $MEWMVL_1$ and $MEWMVL_2$ charts perform better than the MEWMV chart. When there are shifts in the covariance matrix, $MEWMSL_1$ generally outperforms $MEWMSL_2$ and $MEWMVL_1$ generally outperforms $MEWMVL_2$. However, if only the correlation coefficient changes, the MEWMS and MEWMV charts outperform the $MEWMSL_1$, $MEWMSL_2$, $MEWMVL_1$, and $MEWMVL_2$ charts.

4.4 Robust Control Charts for Monitoring Variability

Huwang et al. (2007) and Memar and Niaki (2009) assumed that the in-control parameters μ_0 and Σ_0 are known when the control charts are constructed, and they used the in-control limits constructed under the assumption to monitor the Phase-II data. In practice, these parameters are not known and we have to estimate them from historical data or Phase-I data. The sample mean and sample covariance matrix are

unbiased and efficient estimators, but they are highly sensitive to the presence of outliers. It is therefore important to identify and eliminate outliers prior to calculating the control limits.

Since the RMCD/RMVE estimators are not unduly affected by outliers, we propose using the MEWMS, $MEWMSL_1$, $MEWMSL_2$, MEWMV, $MEWMVL_1$, and $MEWMVL_2$ charts with the RMCD/RMVE estimators. The process variable $\mathbf{g} = (g_1, g_2 ... g_p)'$ is considered to be from a multivariate normal distribution with mean μ_g and covariance matrix Σ_g. If the Phase-I data contain outliers, we have to detect and remove them before proceeding further. We use the robust estimators of the location and dispersion parameters based on RMCD/RMVE to construct charts for monitoring individual multivariate observations. These estimators inherit the properties of affine equivariance, robustness, and asymptotic normality while achieving higher efficiency in the transformed variable $\mathbf{x}$.

The new transformed variables are found by replacing the estimators in Eq. (4.1):

$$\mathbf{x}^{*} = S_{RMCD}^{-1/2}(g - \bar{\mathbf{x}}_{RMCD}) \tag{4.19}$$

$$\mathbf{x}^{**} = S_{RMVE}^{-1/2}(g - \bar{\mathbf{x}}_{RMVE})$$

where $\bar{\mathbf{x}}_{RMCD}$ and $\bar{\mathbf{x}}_{RMVE}$ are the mean vectors and S_{RMCD} and S_{RMVE} are the dispersion matrices under the RMCD/RMVE methods.

The new robust control charts are based on the transformed variables $\mathbf{x}^*$ and $\mathbf{x}^{**}$ with the MEWMS, $MEWMSL_1$, $MEWMSL_2$, MEWMV, $MEWMVL_1$, and $MEWMVL_2$ schemes for monitoring the process variability, with the mean vector constant or changing, as defined in Eqs. 4.4, 4.13, 4.14, 4.9, 4.17, and 4.18 respectively. Since the statistics considered in these equations are positive, we found upper control limits only. We performed 100,000 Monte Carlo simulations for various values of p, ω, and λ and confidence levels α =0.05, 0.01, and 0.0027. The control limits found for robust control charts under MEWMS scheme are given in Tables 4.1 for $\omega = (0.1, 0.2, \cdots, 0.9)$. The control limits found for robust control charts under MEWMV scheme are given in Tables 4.2 to 4.5 for $\omega = (0.1, 0.2, \cdots, 0.9)$, $\lambda = (0.1, 0.2, 0.3, 0.4)$.

4.4.1 Performance Comparison for Phase-I Monitoring

We are analysing the performance of the proposed charts when outliers are present due to a change in the process variance without a shift in process mean and a change in the process variance along with a shift in the process mean. The performance of the charts was assessed based on the probability of outlier detection. When the data come from an in-control process this probability should be close to a specified nominal value. When the data come from an out-of-control process, this probability should

Table 4.1: Control limits for robust control charts with MEWMS scheme; p=2 and m=50

		$MEWMSL_1$		$MEWMSL_2$		MEWMS	
Confidence Level	ω	RMCD	RMVE	RMCD	RMVE	RMCD	RMVE
95%	0.10	5.412	5.748	21.603	24.530	6.628	6.997
	0.20	5.832	6.213	25.737	29.214	7.236	7.623
	0.30	6.793	7.166	35.626	39.832	8.341	8.746
	0.40	8.035	8.486	50.974	56.596	9.628	10.092
	0.50	9.401	9.908	70.058	77.498	11.010	11.510
	6.00	10.853	11.440	93.967	104.456	12.441	13.026
	0.70	12.360	13.008	122.657	135.842	13.934	14.603
	0.80	13.852	14.627	153.937	171.414	15.465	16.217
	0.90	15.457	16.314	191.931	213.199	17.041	17.885
99%	0.10	9.872	10.444	76.409	85.406	11.317	11.867
	0.20	9.983	10.436	78.120	86.789	11.424	11.966
	0.30	10.556	11.100	89.793	97.747	12.104	12.634
	0.40	11.742	12.449	110.735	123.153	13.278	14.048
	0.50	13.388	14.209	144.763	160.265	14.994	15.830
	6.00	15.300	16.008	187.176	205.372	16.928	17.704
	0.70	17.344	18.146	241.363	264.818	19.047	19.862
	0.80	19.658	20.692	314.080	340.333	21.311	22.392
	0.90	21.911	23.109	390.889	423.006	23.543	24.671
99.73%	0.10	13.806	14.348	149.118	164.682	15.373	15.988
	0.20	13.816	14.831	153.692	173.276	15.315	16.281
	0.30	14.223	15.063	166.305	181.111	15.814	16.746
	0.40	15.664	16.550	192.689	222.040	17.194	18.117
	0.50	17.343	18.502	242.978	277.037	18.924	20.220
	6.00	19.618	20.672	306.466	344.296	21.349	22.290
	0.70	22.266	23.352	395.332	434.730	23.891	24.970
	0.80	25.557	26.345	512.444	560.157	27.238	28.037
	0.90	28.551	29.230	633.904	690.581	30.223	30.876

Table 4.2: Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.10, p=2 and m=50

λ =0.10		$MEWMVL_1$		$MEWMVL_2$		MEWMV	
Confidence Level	ω	RMCD	RMVE	RMCD	RMVE	RMCD	RMVE
95%	0.10	5.1051	5.4580	18.9569	22.1970	6.3490	6.6858
	0.20	5.5859	5.8724	23.6384	26.2945	6.9768	7.3236
	0.30	6.5007	6.8784	32.5779	36.6554	8.0317	8.4206
	0.40	7.6922	8.1441	46.1124	51.5411	9.2505	9.7644
	0.50	8.9646	9.4756	64.0539	71.5128	10.5578	11.0512
	6.00	10.3672	10.9375	85.8389	95.4815	11.9229	12.5050
	0.70	11.7332	12.2964	110.6252	121.2014	13.3535	13.9103
	0.80	13.2551	13.8833	139.9339	154.4089	14.7755	15.4457
	0.90	14.7619	15.4691	176.3471	190.5951	16.3208	17.0107
99%	0.10	9.5388	9.9933	69.7641	77.1772	10.9002	11.3466
	0.20	9.5884	10.0770	70.8158	79.5780	10.9416	11.5281
	0.30	10.1346	10.6259	80.6463	89.7291	11.6532	12.1176
	0.40	11.2927	11.9355	101.0684	111.9141	12.9636	13.5056
	0.50	12.7021	13.5692	130.1259	146.3130	14.3606	15.1600
	6.00	14.5995	15.3974	169.2676	190.3975	16.1671	17.1125
	0.70	16.5161	17.3952	220.6464	242.5054	18.2144	19.0374
	0.80	18.6475	19.4826	281.3431	310.4795	20.2393	21.1805
	0.90	20.7853	21.9090	345.4331	381.3439	22.3855	23.5929
99.73%	0.10	13.7291	14.2536	149.5252	159.6506	15.4335	15.7931
	0.20	13.5856	14.0653	145.5116	163.1156	15.1064	15.6325
	0.30	13.9353	14.3016	158.3426	165.0561	15.3163	15.7954
	0.40	15.0206	15.6274	178.7550	195.3429	16.5480	17.2081
	0.50	16.5865	17.4133	209.8106	241.3272	18.3472	19.2170
	6.00	18.4554	19.7040	272.8167	309.1900	20.1931	21.2886
	0.70	21.1065	22.0417	362.0970	400.4451	22.6739	23.5929
	0.80	23.9111	24.7855	472.1264	496.0729	25.5163	26.3991
	0.90	26.2912	27.5757	576.0018	629.0656	27.8740	29.2741

Table 4.3: Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.20, p=2 and m=50

λ =0.20		$MEWMVL_1$		$MEWMVL_2$		MEWMV	
Confidence Level	ω	RMCD	RMVE	RMCD	RMVE	RMCD	RMVE
95%	0.10	4.7694	5.0447	16.2966	18.6025	5.9698	6.2841
	0.20	5.3100	5.6954	21.2799	24.7984	6.7586	7.1228
	0.30	6.3551	6.6976	31.0610	34.8400	7.8946	8.2693
	0.40	7.5465	7.9912	44.6048	50.1130	9.1124	9.5664
	0.50	8.8206	9.3489	61.5216	68.9525	10.4496	10.9704
	6.00	10.2122	10.7855	82.8166	91.9596	11.7826	12.4060
	0.70	11.6226	12.1223	107.8605	118.7811	13.2036	13.7478
	0.80	13.0891	13.7081	135.9658	150.2081	14.6176	15.3054
	0.90	14.5519	15.2169	170.9465	186.3915	16.1452	16.8124
99%	0.10	8.6815	9.0476	58.6209	63.9531	10.0897	10.4833
	0.20	8.9625	9.4095	63.1119	70.7689	10.2991	10.9604
	0.30	9.7701	10.2321	74.3518	81.7940	11.3061	11.7709
	0.40	10.9581	11.5895	97.1529	106.7094	12.5011	13.1602
	0.50	12.6621	13.3729	128.9515	141.2902	14.2666	14.9489
	6.00	14.4285	15.2301	168.7584	185.1511	15.9831	16.8890
	0.70	15.9815	16.7943	206.3258	229.9903	17.5320	18.4221
	0.80	18.4227	19.2128	272.7901	299.6626	20.0221	20.8674
	0.90	20.0129	21.0024	325.0538	358.9017	21.5410	22.5706
99.73%	0.10	13.0980	13.8624	138.3264	153.2591	14.4417	15.2578
	0.20	12.8296	13.6818	131.3802	152.8525	14.2714	15.1862
	0.30	13.4524	14.1249	151.3848	166.9695	14.8736	15.5450
	0.40	14.2962	14.9945	165.6392	180.7477	15.8509	16.5292
	0.50	16.3207	17.3703	217.2912	239.2386	17.9022	18.9468
	6.00	18.3744	19.5418	273.4326	315.9891	20.1038	21.2763
	0.70	20.4233	21.2477	334.2219	367.7190	22.1160	22.9323
	0.80	23.1837	24.1086	451.2562	486.0798	24.9055	25.7932
	0.90	25.2601	26.6450	523.6656	577.8495	27.0784	28.3246

Table 4.4: Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.30, p=2 and m=50

λ =0.30		MEWMV		$MEWMVL_1$		$MEWMVL_2$	
Confidence Level	ω	RMCD	RMVE	RMCD	RMVE	RMCD	RMVE
95%	0.10	4.5552	4.9023	14.8746	17.4233	5.7436	6.1267
	0.20	5.1729	5.4960	20.0939	22.8512	6.5991	6.9462
	0.30	6.2440	6.5812	29.9073	33.7190	7.7884	8.1557
	0.40	7.5465	7.9912	44.6048	50.1130	9.1124	9.5664
	0.50	8.8020	9.2403	61.0219	67.5579	10.3746	10.8455
	6.00	10.1758	10.6877	82.5115	91.1071	11.7662	12.2677
	0.70	11.5283	12.1475	106.1085	117.3567	13.0872	13.7076
	0.80	12.9167	13.6758	134.5078	150.6758	14.4727	15.2587
	0.90	14.4485	15.1970	167.1397	183.3862	15.9619	16.7115
99%	0.10	8.1717	8.4880	51.1975	56.1086	9.6063	9.7838
	0.20	8.5368	9.0790	56.6789	64.9738	9.9021	10.3902
	0.30	9.3897	10.0101	70.9870	79.7028	10.8898	11.4966
	0.40	10.9581	11.5895	97.1529	106.7094	12.5011	13.1602
	0.50	12.3556	13.1490	122.9870	139.5134	13.9891	14.7934
	6.00	14.1516	14.9783	161.1694	180.2423	15.7632	16.6134
	0.70	16.1015	16.8706	209.2945	228.0611	17.7859	18.4972
	0.80	18.0493	19.0145	261.1538	291.7303	19.7076	20.6470
	0.90	20.0193	21.0335	325.5745	357.2021	21.6324	22.7225
99.73%	0.10	11.7549	12.1557	109.1773	120.8208	13.3243	13.7170
	0.20	11.9134	12.4614	112.0764	123.1893	13.3325	14.0248
	0.30	12.7302	13.5486	132.1648	151.3605	14.3185	14.9748
	0.40	14.2962	14.9945	165.6392	180.7477	15.8509	16.5292
	0.50	15.7494	16.6340	201.9270	223.7778	17.3570	18.1462
	6.00	18.1518	19.0668	274.3548	299.9406	19.7464	20.6298
	0.70	20.1266	20.9915	341.3798	366.4043	21.7606	22.8557
	0.80	23.1015	24.1432	432.3401	481.1007	24.7108	25.8285
	0.90	25.4964	26.1649	525.3491	553.3217	27.0467	28.0336

Table 4.5: Control limits for robust control charts with MEWMV scheme for various values of ω, λ =0.40, p=2 and m=50

λ =0.40		$MEWMVL_1$		$MEWMVL_2$		MEWMV	
Confidence Level	ω	RMCD	RMVE	RMCD	RMVE	RMCD	RMVE
95%	0.10	4.2394	4.5302	12.8501	14.8887	5.4383	5.7676
	0.20	5.0522	5.3596	19.2977	21.9912	6.4766	6.8318
	0.30	6.1807	6.5158	29.4252	32.7207	7.7240	8.0733
	0.40	7.4400	7.8874	43.5169	48.9075	9.0201	9.4506
	0.50	8.6455	9.1131	59.4628	66.2191	10.2364	10.6985
	6.00	9.9636	10.5018	79.5637	88.8396	11.5430	12.0840
	0.70	11.4370	12.0318	103.9654	114.4226	13.0319	13.6203
	0.80	12.8178	13.5835	132.2170	148.6244	14.3397	15.1573
	0.90	14.2879	15.0077	162.8353	178.8101	15.8570	16.6025
99%	0.10	7.6353	8.0855	45.5482	50.2086	8.9448	9.3942
	0.20	8.0785	8.5789	50.1714	57.7054	9.5158	9.9409
	0.30	9.1341	9.7908	67.5754	76.3891	10.6848	11.3067
	0.40	10.6288	11.2429	89.8766	100.7138	12.2619	12.8449
	0.50	12.1205	12.9184	118.7535	134.4263	13.8436	14.5348
	6.00	13.9323	14.7526	155.9947	177.4837	15.6153	16.4555
	0.70	15.8908	16.6762	205.8261	227.3480	17.5322	18.2816
	0.80	17.9238	18.9424	257.4751	291.3179	19.5502	20.6327
	0.90	19.7770	20.7617	320.3862	346.8489	21.4323	22.3480
99.73%	0.10	10.9849	11.4941	93.6568	101.7213	12.5972	12.9326
	0.20	11.3861	11.9649	107.2113	111.9252	12.8966	13.4966
	0.30	12.2103	12.8141	122.5920	138.0268	13.8004	14.3044
	0.40	13.7453	14.4432	155.2432	169.1351	15.3537	16.0699
	0.50	15.4734	16.3998	196.7133	219.6940	17.2050	18.0281
	6.00	17.7710	18.7498	256.4625	286.9265	19.4331	20.4683
	0.70	20.1427	20.8916	341.7016	356.0202	21.9232	22.7397
	0.80	22.9376	23.7725	420.1023	456.5187	24.5010	25.5212
	0.90	25.1089	25.9691	527.3190	561.8378	26.9420	27.7627

be large compared to the specified nominal value.

Following Huwang et al. (2007) and Memar and Niaki (2009), we consider a bivariate process with mean μ and covariance matrix Σ where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

For an in-control process, the parameters are set to $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$, and $\rho = 0$. If any one of these parameters is shifted, the process is out of control. We generated a number of data sets with $m = 50$ and $p = 2$. Of the m observations, $m \times \pi$ are random data points generated from the out-of-control distribution, and the remaining $m \times (1 - \pi)$ are generated from the in-control distribution. We set π to 0.20 to ensure that the sample contains a few outliers. The following shift scenarios were considered for an out-of-control situation:

- σ_1^2 =1.00 and σ_2^2 set to 1.00, 1.25, 1.50, 1.75, 2.00, 2.50, and 3.00.
- Correlation coefficient ρ set to 0, 0.25, 0.50, and 0.75.
- Shift in mean (μ_1 or μ_2) set to 0, 0.5, 1, and 2.

Following Huwang et al. (2007) and Memar and Niaki (2009), we considered smoothing parameters ω = (0.2,0.3,0.4,0.5) and λ = (0.1,0.2,0.3,0.4). For each chart, we consider the standard as well as robust versions based on MCD, RMCD, MVE,

and RMVE. The probability of a signal is estimated as the proportion of data sets with at least one data point greater than the control limit. We consider α = (0.05, 0.01, 0.0027). Figures 4.1 to 4.16 show the probability of a signal for $\alpha = 0.01$ and different values of p and m. The plots for $\alpha = 0.05$ and 0.0027 are omitted to save space. We show the probability of a signal for the standard chart, robust charts based on the MCD/MVE estimators and the proposed charts in each of the six methods. We can see that the proposed charts perform better than these MCD/MVE charts and standard chart in most of the scenarios considered. Each figure displays plots for four different values of ρ, showing the effect of changes in ρ. The performance of the proposed robust control charts are consistently better for all six charts. We have presented only a selected set of plots to save space.

From the plots, we see that the probability of a signal increases as the value of σ_2^2 increases for the proposed charts. In contrast, the charts based on the classical estimators break down and perform poorly compared to the proposed charts and the charts based on MCD/MVE. From Figs. 4.1 to 4.3 we see that the probability of a signal increases as ρ increases. This clearly indicates that as ρ increases, the proposed charts perform better and the performance of the charts based on the classical estimators deteriorates.

Figures 4.4, 4.5, and 4.6 show that under the MEWMV scheme, the chart detects

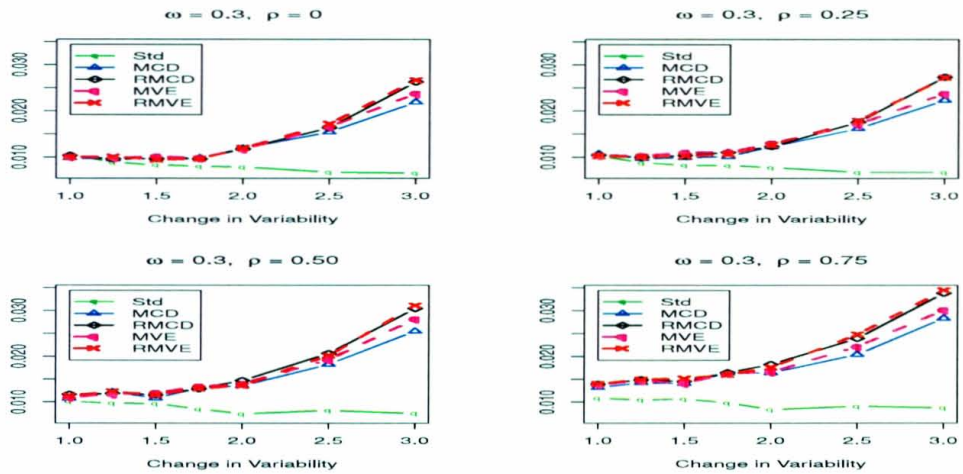

Figure 4.1: Probability of signal for robust $MEWMSL_1$ control chart for p= 2, m= 50, ω =0.30, $\mu_1 = \mu_2 = 0$

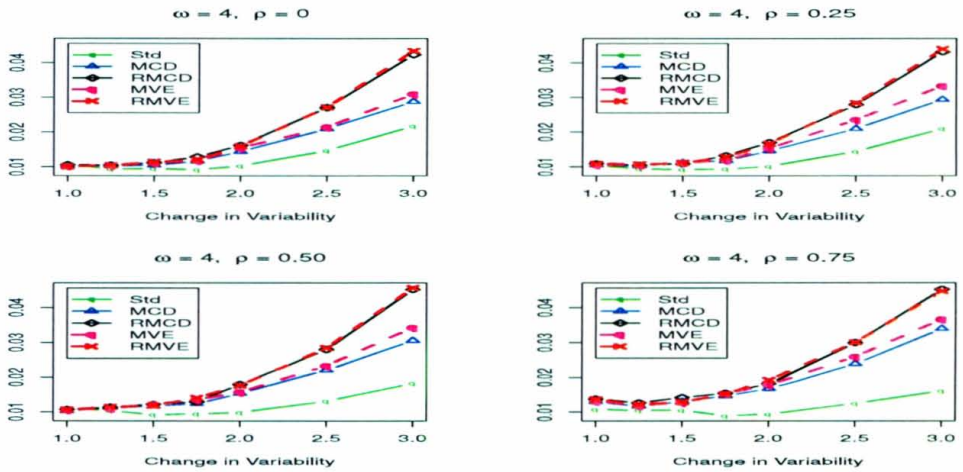

Figure 4.2: Probability of signal for robust $MEWMSL_2$ control chart for p= 2, m= 50, ω =0.40, $\mu_1 = \mu_2 = 0$

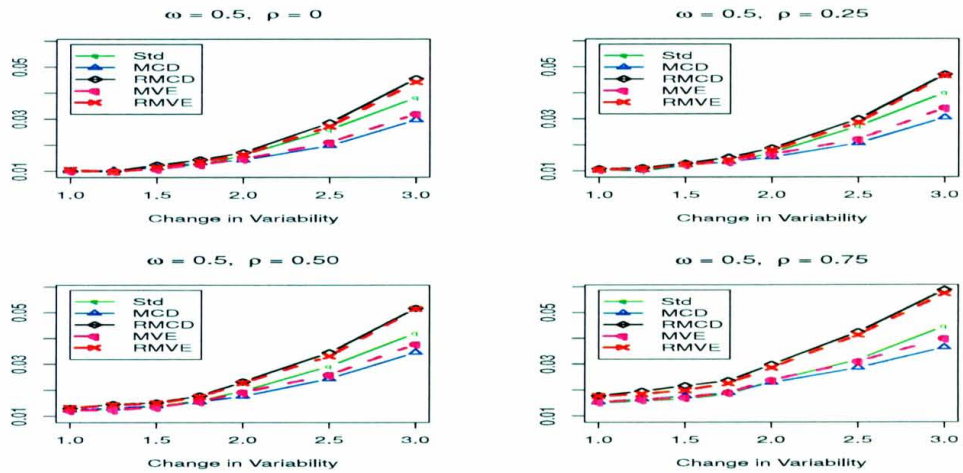

Figure 4.3: Probability of signal for robust $MEWMS$ control chart for p= 2, m= 50, ω =0.50, $\mu_1 = \mu_2 = 0$

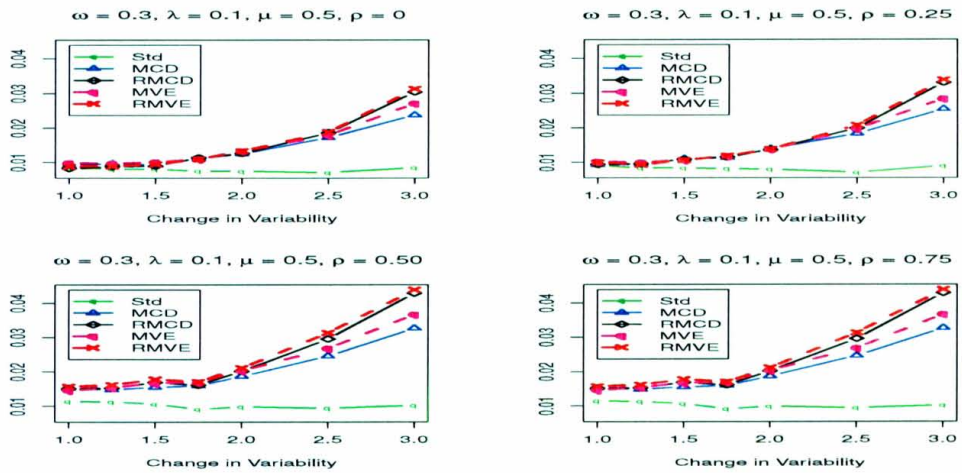

Figure 4.4: Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 0.50$

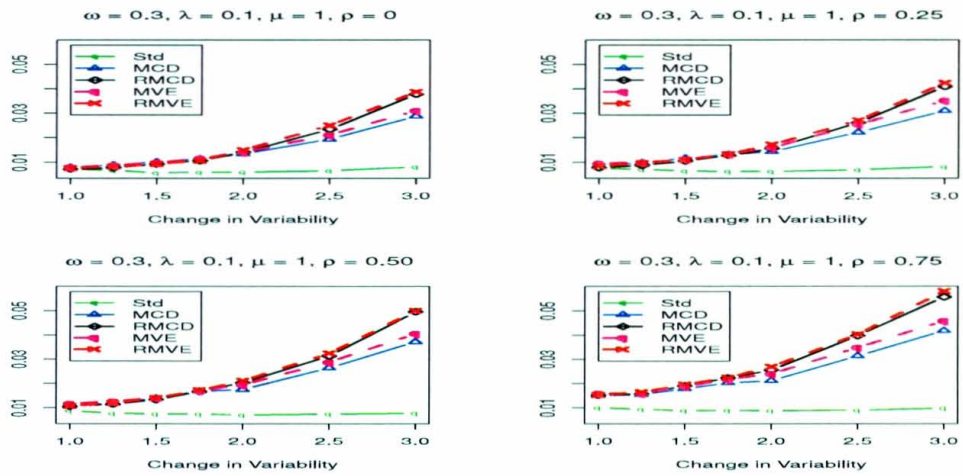

Figure 4.5: Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 1.00$

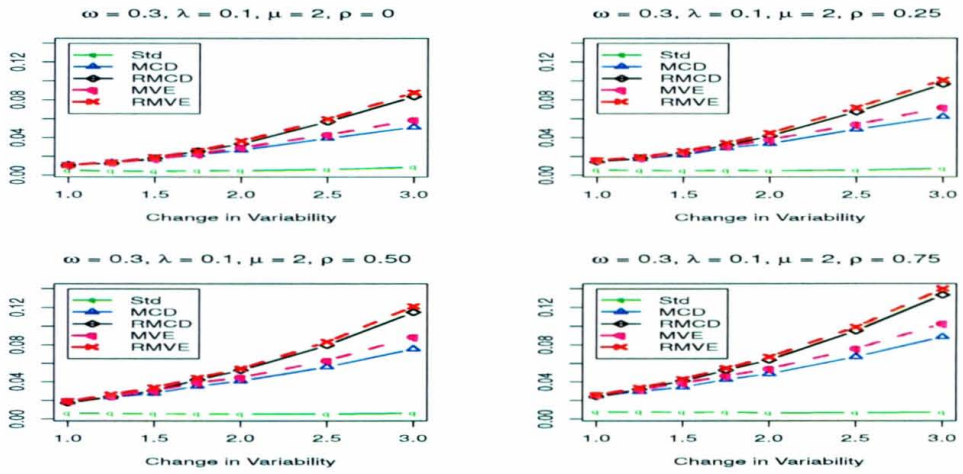

Figure 4.6: Probability of signal for robust $MEWMV$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 2.00$

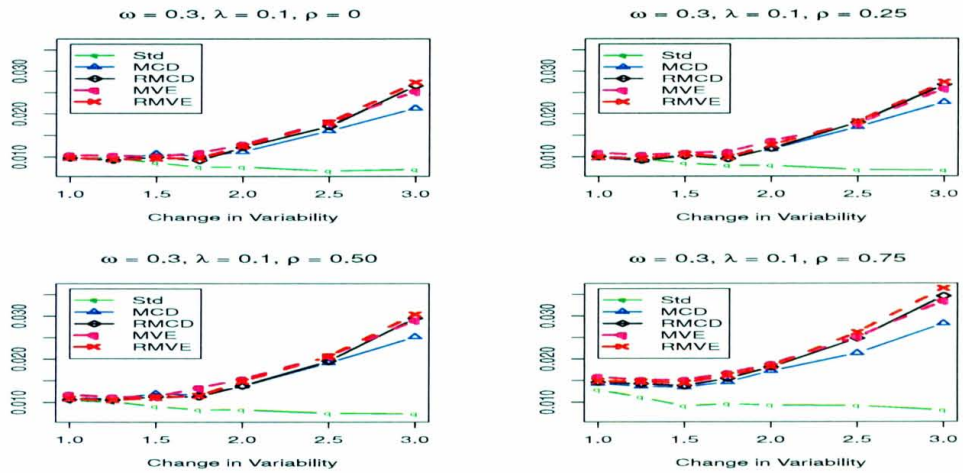

Figure 4.7: Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.10, $\mu_1 = \mu_2 = 0$

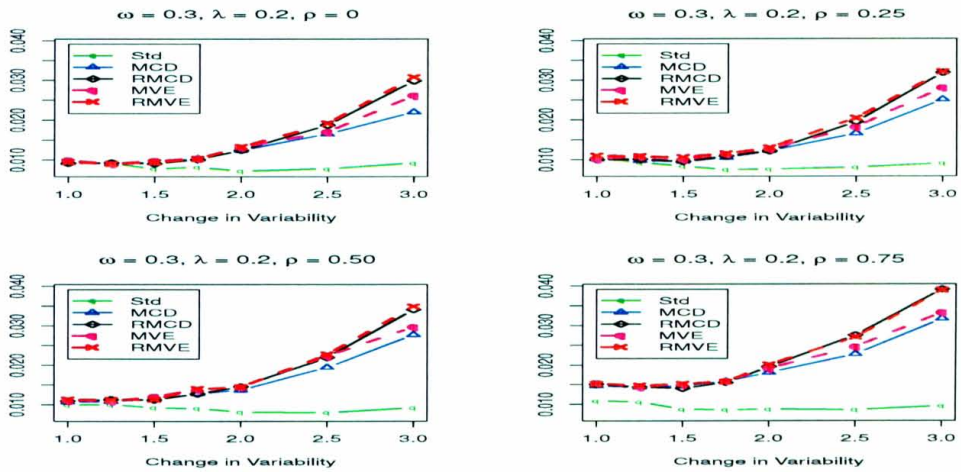

Figure 4.8: Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.20, $\mu_1 = \mu_2 = 0$

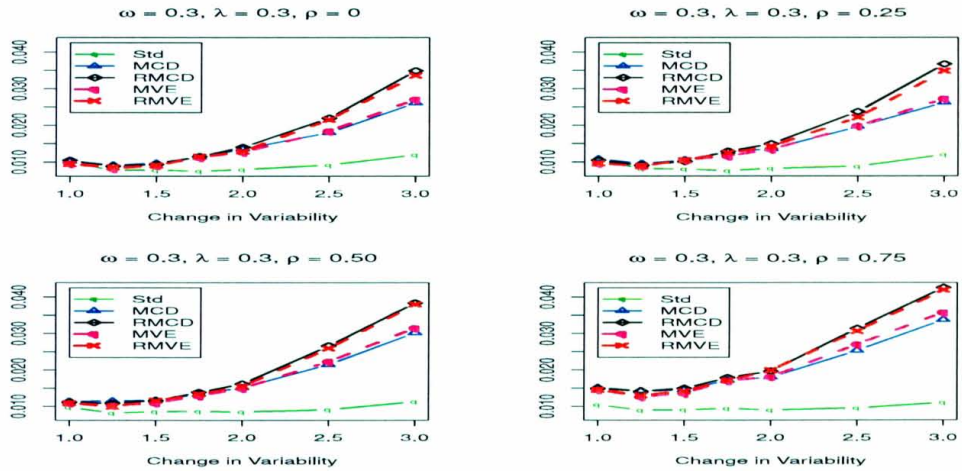

Figure 4.9: Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.30, $\mu_1 = \mu_2 = 0$

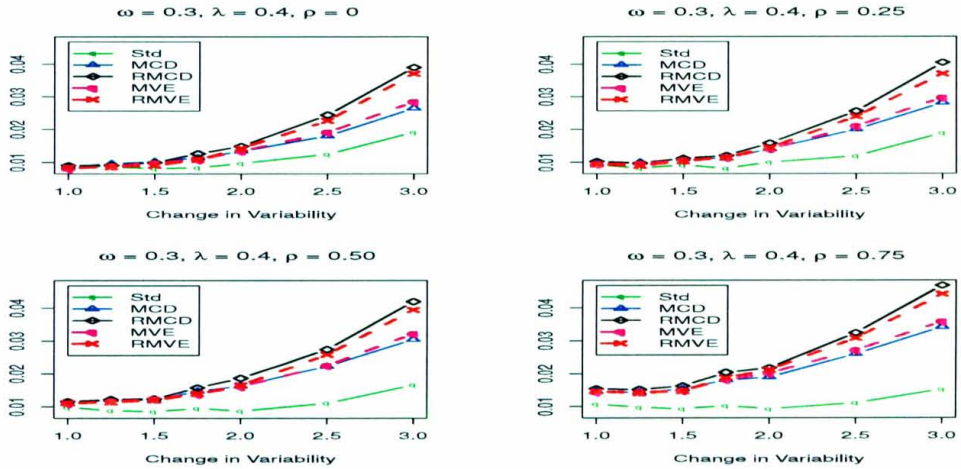

Figure 4.10: Probability of signal for robust $MEWMVL_1$ control chart for p= 2, m= 50, ω =0.30, λ =0.40, $\mu_1 = \mu_2 = 0$

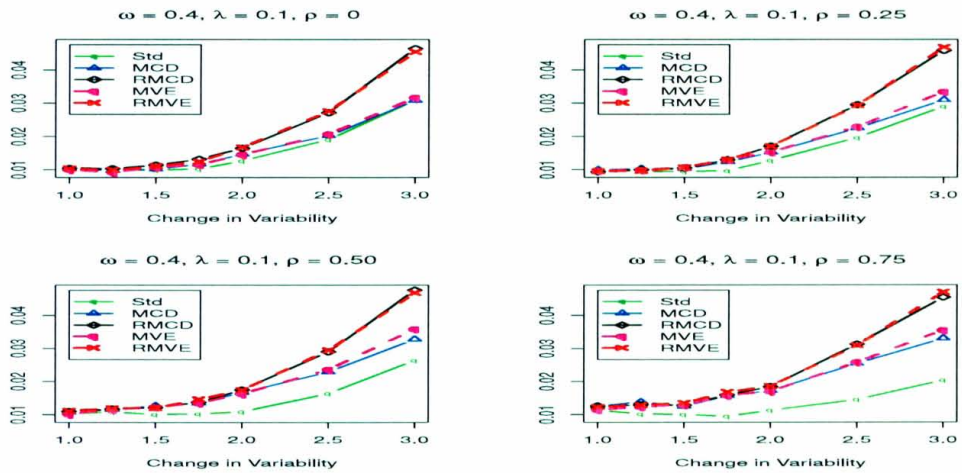

Figure 4.11: Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.10, $\mu_1 = \mu_2 = 0$

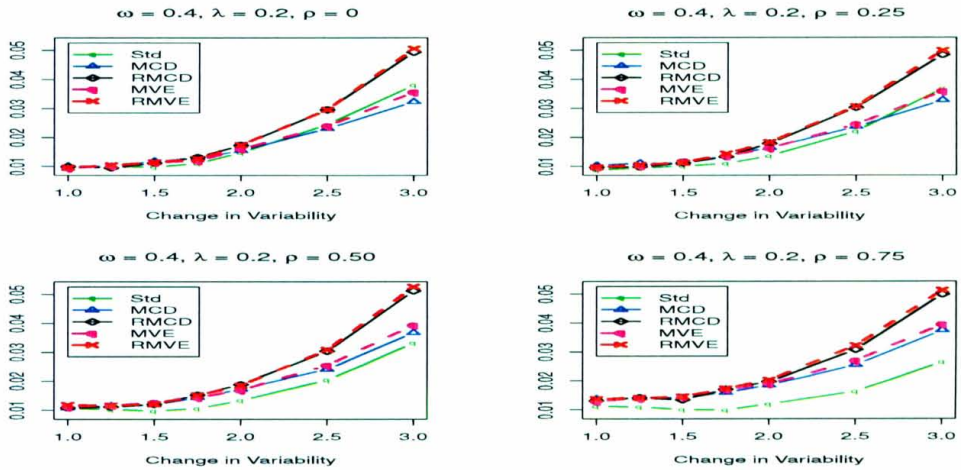

Figure 4.12: Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.20, $\mu_1 = \mu_2 = 0$

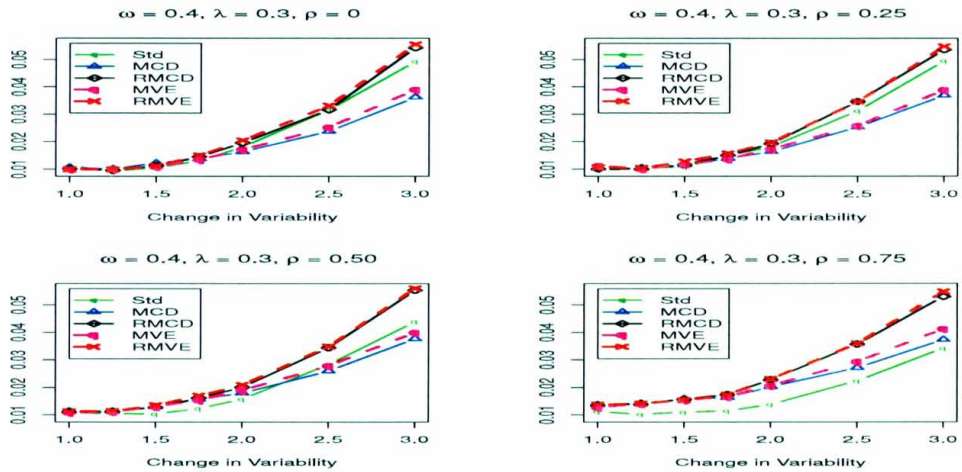

Figure 4.13: Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.30, $\mu_1 = \mu_2 = 0$

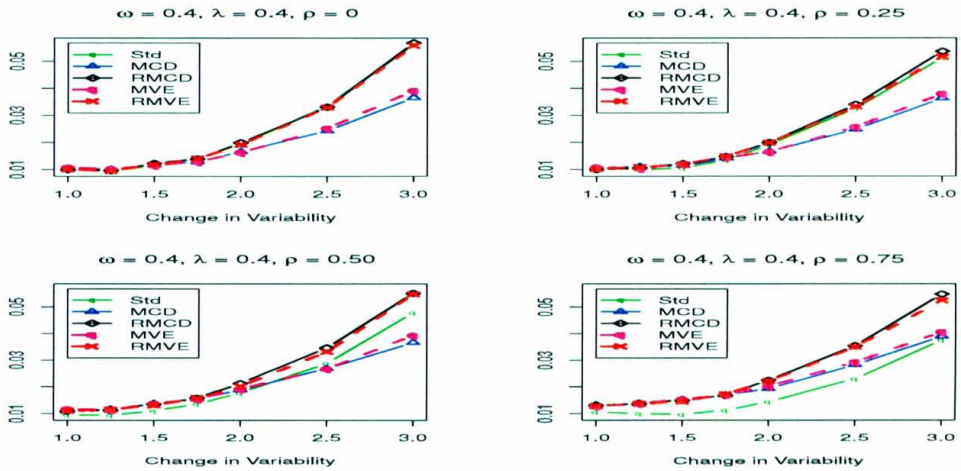

Figure 4.14: Probability of signal for robust $MEWMVL_2$ control chart for p= 2, m= 50, ω =0.40, λ =0.40, $\mu_1 = \mu_2 = 0$

the shift in location as well as the shift in variability. The probability of a signal increases as the shift in mean increases from 0.50 to 1.00 and increases further when it increases to 2.00. The MEWMS scheme fails to detect the shift in the mean especially when the magnitude of the shift is large.

Figures 4.7 to 4.10 and 4.11 to 4.14 show the effect of the changes in variability for various values of λ when ω= 0.3 and 0.4. Clearly, as σ_2^2 increases, the probability of a signal also increases. The changes in the value of σ_2^2 along with the changes in the value of ρ are also well detected by the proposed charts; see Figs. 4.7 to 4.14. All these plots clearly indicate that our robust charts with RMCD/RMVE estimates perform well compared to the other charts.

Chapter 5

Robust Regression

Regression analysis includes many techniques for modeling and analyzing several variables, when the focus is on establishing the relationship between a dependent variable (the response variable) and one or more independent variables (the covariates). Specifically, regression analysis helps to explain how the value of the response variable changes with changes in the covariates. Linear regression is a common regression technique with some basic assumptions of normality and independence for the response variable. The general form of a linear regression model is

$$y_i = \mathbf{x}_i^{'}\beta + \epsilon_i, \ i = 1, 2, \cdots, n \tag{5.1}$$

where y_i is the value of the ith response variable, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{'}$ is a vector of values of covariates corresponding to the ith response, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ is the

effects of the covariates $\mathbf{x}_i$ on y_i, and ϵ_i is the random error in the observed responses. The random errors are assumed to be independently and identically distributed as normal with zero mean and constant variance σ^2. We wish to estimate the regression parameter β from the observed responses and covariates.

The generalized linear model (GLM) is a flexible generalization of linear regression that allows response variables with non-normal distributions. The GLM allows the linear model to be related to the response variable via a link function and allows the variance of each measurement to be a function of its predicted value. Hence, GLM encompasses not only linear regression for normally distributed responses, but the logistic model for binary data, the log linear model for count data, and many other useful statistical models via its general formulation.

All these regression models work well when there are no outliers in the response and in the covariate data. Outliers, especially in the covariates, may unduly influence the estimation of the regression parameters. This causes bias and hence inconsistency in the estimators. Specifically, if there are no outliers, we can obtain a consistent estimate of the regression parameters. It is therefore important to identify outliers in the covariate data.

In the following section, we review the GLM, especially the Poisson and logistic

regression models. Then we introduce robust regression by identifying and down-weighting the outliers in the covariate data using the squared robust Mahalanobis distance and perform simulations to assess the performance of the proposed method.

5.1 Generalized Linear Model

The random component of a generalized linear model consists of a response variable Y with independent observations $(y_1, y_2. \cdots, y_n)$ from a distribution in the natural exponential family. This family has a probability density function or mass function of the form

$$f(y_i, \beta_i) = a(\beta_i)b(y_i)\exp[y_i Q(\beta_i]. \tag{5.2}$$

The term $Q(\beta_i)$ is called the natural parameter. The systematic component of a GLM relates a vector $(\eta_1, \eta_2, \cdots, \eta_n)$ to the explanatory variable through a linear model. Let $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$ be the values of the p-covariates for the ith case. Then

$$\eta_i = \mathbf{x}_i'\beta \tag{5.3}$$

where β is a $(p \times 1)$ vector of unknown parameters. The link connects the random and systematic components of the model.

Let $\mu_i = E(\eta_i),\ i = 1, 2, \cdots, n$. The model links μ_i to $\eta_i = g(\mu_i)$. Thus, g links

$E(\eta_i)$ to the explanatory variables through the formula

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\beta \quad : \quad i = 1, 2, \cdots, n. \tag{5.4}$$

The link function that transforms the mean to the natural parameter is called the canonical link. For this link, $g(\mu_i) = Q(\beta_i)$ where

$$Q(\beta_i) = \mathbf{x}_i'\beta \quad : \quad i = 1, 2, \cdots, n. \tag{5.5}$$

5.1.1 Poisson Log Linear Model

Let Y denote a count which follows a Poisson distribution, and let $\mu = E(Y)$. The Poisson probability mass function for Y is

$$f(y : \mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y.\log\mu). \tag{5.6}$$

The natural parameter is $\log\mu$, so the canonical link function is the log link, $\eta = \log\mu$. The model using this link is

$$\mu_i = e^{\mathbf{x}_i'\beta} \; ; \; i = 1, 2, \cdots, n \tag{5.7}$$

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is the vector of covariates for the i-th response and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ is the regression parameter.

Consider a data set containing count responses y_i for $i = 1, 2, \cdots, n$ and a $(p \times 1)$ vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ of covariates associated with the response. Let $\beta =$

$(\beta_1, \beta_2, \cdots, \beta_p)'$ be a $(p \times 1)$ vector of unknown regression parameters. Suppose the response y_i has the Poisson distribution with mean $m_i = e^{\mathbf{x}_i'\beta}$, then the probability mass function of y_i is given by

$$f(y_i) = \frac{e^{-m_i} m_i^{y_i}}{y_i!}. \tag{5.8}$$

The log-likelihood function is

$$\log L = c - \sum_{i=1}^{n} e^{\mathbf{x}_i'\beta} + \sum_{i=1}^{n} y_i \mathbf{x}_i'\beta \tag{5.9}$$

where c is a constant. The estimating equation of the parameter vector can be obtained by taking the partial derivative of the log-likelihood with respect to β; which is given by

$$\frac{\partial \log L}{\partial \beta} = R(\beta) = \sum_{i=1}^{n} (y_i - e^{\mathbf{x}_i'\beta})\mathbf{x}_i = \sum_{i=1}^{n} R_i(y_i, \mathbf{x}_i, \beta) = 0. \tag{5.10}$$

Since there is no closed-form solution to Eq. (5.10), we use the Newton–Raphson iterative method to estimate the regression parameter β:

$$\hat{\beta}^{t+1} = \hat{\beta}^t - \left[R'(\hat{\beta}^t)\right]^{-1} R(\hat{\beta}^t) \tag{5.11}$$

where $\hat{\beta}^t$ is the estimate of β in the tth iteration, $R(\beta) = \sum_{i=1}^{n} R_i(y_i, \mathbf{x}_i, \beta) = \sum_{i=1}^{n} (y_i - e^{\mathbf{x}_i'\beta})\mathbf{x}_i$, and $R'(\beta)$ is the first derivative of $R(\beta)$ with respect to β.

5.1.2 Binary Logistic Regression Model

Let Y be 1 or 0, representing the success or failure of a Bernoulli trial with specific probabilities P(Y=1)= π and P(Y=0)= 1-π, and E(Y)= π. This is a special case of the binomial distribution with $n = 1$, and the probability mass function for Y is

$$f(y:\pi) \;=\; \pi^{y}(1-\pi)^{1-y} \;=\; (1-\pi)exp\left(y\log\frac{\pi}{1-\pi}\right). \tag{5.12}$$

The natural parameter is $\log\dfrac{\pi}{1-\pi}$, so the canonical link function is the log link. We may write the link as

$$\pi(\mathbf{x}_i) \;=\; \frac{exp(\mathbf{x}_i'\beta)}{1+exp(\mathbf{x}_i'\beta)}\;;\; i=1,2,\cdots,n. \tag{5.13}$$

This is called the binary logistic model, where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$ contains the values of the p-covariates for the ith response and the $(p \times 1)$ parameter vector $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$.

Suppose the response y_i, $i = 1, 2, \cdots, n$, has a binary distribution with $\pi(\mathbf{x}_i) = \dfrac{exp(\mathbf{x}_i'\beta)}{1+exp(\mathbf{x}_i'\beta)}$, then the estimates of β can be obtained by solving the likelihood estimating equation using the Newton–Raphson method, as for the Poisson log-linear regression.

5.2 Robust Generalized Linear Regression

As discussed earlier, the regression models work well when there are no outliers in the response or in the covariate data. Outliers, especially in the covariates, may unduly influence the estimation of the regression parameters. There are many methods in the literature to down-weight these observations so that bias correction can be carried out. We propose identifying and down-weighting outliers in the covariate data so that outlier-free data can be used to fit the regression models. We use the squared robust distance based on the RMCD/RMVE estimators of the mean and covariance of the covariate data to identify outliers.

Consider the generalized linear model for discrete data (binary data or count data), where y_i, i=1,2...,n, is the discrete response collected from the ith individual. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ be the corresponding p-dimensional observed covariate vector corresponding to the response y_i and let $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ be the effects of the covariates $\mathbf{x}_i$ on the response y_i. We consider the situation where the data contain a covariate outlier corresponding to the jth observation y_j, i.e., $\mathbf{x}_j$ is contaminated. It is of primary interest to estimate $\beta = (\beta_1, \beta_2, \cdots, \beta_p)'$ based on the uncontaminated covariates $\tilde{\mathbf{x}}_i$. However, the observed $\tilde{\mathbf{x}}_i$'s include the contaminated $\mathbf{x}_j$. This causes bias and hence inconsistency in the estimators. If there were no outliers, we could

obtain a consistent estimate of β by solving the estimating equations. Our approach is to identify and down-weight the outliers in order to get a consistent estimate of β. In this thesis, we consider the situation where the covariate data has few contaminated data and the response data are free of outliers. We also assume that the covariates follow a normal distribution.

The Mahalanobis distance (Mahalanobis, 1936) and leverage are often used to detect outliers, especially in linear regression models. A data point that has a larger (squared) Mahalanobis distance than the rest of the sample is said to have higher leverage since it has a greater influence on the slope or coefficients of the regression equation. Note that the squared Mahalanobis distance for any sample data point $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ is similar to the Hotelling T^2 statistic for individual observations as given in Eq. (1.4) and reproduced below:

$$T^2(i) = (\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \; i = 1, 2, \cdots, n \tag{5.14}$$

where the sample mean $\bar{\mathbf{x}}$ and sample covariance matrix S are based on n sample points. The sample mean and sample covariance are highly sensitive to outliers, and hence robust estimation methods are preferred. The proposed RMCD/RMVE-based squared robust distance is used to identify and eliminate outliers in the covariate data.

The robust Hotelling T^2 statistic (Eq. 3.1) discussed in Chapter 3 is reproduced below:

$$T^2_{RMCD}(i) = (\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})' S^{-1}_{RMCD}(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD}) \tag{5.15}$$

$$T^2_{RMVE}(i) = (\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE})' S^{-1}_{RMVE}(\mathbf{x}_i - \bar{\mathbf{x}}_{RMVE})$$

where $\bar{\mathbf{x}}_{RMCD}$ and $\bar{\mathbf{x}}_{RMVE}$ are the location estimators and S_{RMCD} and S_{RMVE} are the scatter estimators under the RMCD/RMVE methods based on n covariate data. These values can be compared with the quantiles found via Eq. (3.2) and Tables 3.1 and 3.2 depending on the dimension and the confidence level. Observations with T^2_{RMCD}/T^2_{RMVE} values greater than the quantiles are considered outliers and need to be down-weighted. A step-by-step approach for estimating the robust regression parameters is as follows:

i) Compute the robust estimates of the mean and covariance of the covariate data.

ii) Compute the robust T^2 statistic for the covariate data for each response using Eq. (5.17).

iii) Find the critical values for the T^2 statistics for a given confidence level and dimension using Eq. (3.2).

iv) Identify the responses for which $T^2(\mathbf{x}_i) >$ the critical value; these are outliers.

v) Assign weight $w_i = 0$ to the response and to the covariates identified as outliers; otherwise assign weight $w_i = 1$.

vi) Estimate the regression parameters by solving the weighted score equation $\sum_{i=1}^{n} w_i R_i(y_i, \mathbf{x}_i, \beta) = 0$.

We conduct a simulation study for the Poisson log-linear model and the binary logistic regression model to study the effectiveness of our method.

5.3 Simulation Studies

We have conducted a large number of simulation studies to assess the performance of our method for the Poisson regression model and the binary logistic regression model. We examined the performance by estimating the regression parameters under models with one or two outliers. We repeated each simulation 10,000 times and computed the simulation means (SM), the standard errors (SSE), and the relative bias (RB) of these estimators. The relative bias for each regression parameter is $RB(\hat{\beta}_k) = \frac{|\hat{\beta}_k - \beta_k|}{SE(\hat{\beta}_k)} \times 100$. We used the R function glm() in the stats library to estimate the regression parameters.

5.3.1 Poisson Regression Model

We considered β =(3, 3.5, 0, 0, 2) with n=(150, 200, 250) for the Poisson model and n=(200, 250, 300) for the binary model. We generated the covariates $\tilde{\mathbf{x}}_i$ for the ith response by assuming that it follows a multivariate normal distribution with mean and covariance as given below and $\rho = 0.50$.

$$\Sigma = \begin{pmatrix} 1 & \sigma & \cdots & \sigma^{p-1} \\ \sigma & 1 & \cdots & \sigma^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{p-1} & \sigma^{p-2} & \cdots & 1 \end{pmatrix}_{p\times p}, \quad \mu = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{p\times 1}. \tag{5.16}$$

Data with a single outlier. To generate n count observations with one outlier, we first assume that outlier-free data $y_1, y_2, \cdots, y_n$ are generated following the Poisson density $P(Y_i = y_i) = \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!}$, with $\mu_i = e^{\tilde{\mathbf{x}}_i'\beta}$ where $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{ip})$ are the uncontaminated covariates. Now consider y_j as an outlying value among the n responses corresponding to the contaminated covariate $\mathbf{x}_j$. To get this $\mathbf{x}_j$, we shift the values of $\tilde{\mathbf{x}}_j$ by adding $\delta > 0$ to all p components of $\tilde{\mathbf{x}}_j$ and set $\mathbf{x}_i = \tilde{\mathbf{x}}_i$ for all $i \neq j$. We take δ =5.0 and thus $y_1, y_2, \cdots, y_n$ are a sample of n count observations with covariates corresponding to y_j as the single outlier. We estimated the regression estimates based on:

a) The contaminated data of size n.

b) The data excluding the contaminated covariate and the corresponding response.

c) The proposed method.

Table 5.1: Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters under Poisson model with β = (3.0,3.5,0,0,2.0) in presence of single outlier

# of Outliers: 1		m=150			m=200			m=250		
Sample size	Parameter	SM	SSE	RB	SM	SSE	RB	SM	SSE	RB
With Outlier	β_1	1.841	2.750	42	1.829	2.633	44	1.869	2.590	44
	β_2	3.674	3.032	6	3.775	2.881	10	3.843	2.775	12
	β_3	-1.243	3.194	39	-1.214	3.091	39	-1.248	3.006	42
	β_4	-1.230	3.181	39	-1.199	3.072	39	-1.241	2.968	42
	β_5	0.372	2.755	59	0.402	2.663	60	0.424	2.606	60
Without Outlier	β_1	3.000	0.002	1	3.000	0.001	1	3.000	0.001	1
	β_2	3.500	0.002	2	3.500	0.001	2	3.500	0.001	1
	β_3	0.000	0.002	0	0.000	0.002	2	0.000	0.001	2
	β_4	0.000	0.002	1	0.000	0.002	1	0.000	0.001	1
	β_5	2.000	0.002	1	2.000	0.001	0	2.000	0.001	0
Outliers	β_1	2.994	0.199	3	2.994	0.195	3	2.996	0.148	3
down-weighted	β_2	3.503	0.199	2	3.505	0.165	3	3.504	0.122	4
by RMCD	β_3	-0.008	0.246	3	-0.006	0.180	3	-0.005	0.168	3
	β_4	-0.004	0.224	2	-0.006	0.206	3	-0.006	0.174	3
	β_5	1.991	0.200	5	1.996	0.183	2	1.997	0.146	2
Outliers	β_1	2.997	0.176	2	2.995	0.180	3	2.997	0.146	2
down-weighted	β_2	3.500	0.196	0	3.503	0.180	2	3.501	0.122	1
by RMVE	β_3	-0.003	0.191	2	-0.006	0.221	3	-0.003	0.157	2
	β_4	-0.006	0.239	3	-0.005	0.216	2	-0.004	0.162	2
	β_5	1.992	0.226	4	1.994	0.175	4	1.995	0.142	3

Table 5.1 summarizes the results for the Poisson model with one outlier. We see that the regression estimates are biased by the outlier. However, the regression estimates based on the outlier-free data and the estimates based on our method are

close to the true regression parameters. The relative bias corresponding to these estimators is also very small. It is worth noting that for $m = 150$, 200, and 250 the RMCD method identifies the outliers in 99.36%, 99.45%, and 98.57% of the simulations, and the RMVE method identifies them in 99.42%, 99.44%, and 99.59% of the simulations. The method identified some outliers other than those generated, but this is negligible.

Data with two outliers. For the Poisson model with two outlying observations, the count responses are generated in a manner similar to that for a single outlier. After generating n count observations from a Poisson model with the covariate values, we create two covariate outliers, namely $\mathbf{x}_j$ and $\mathbf{x}_k$, $j \neq k$. The contaminated covariates $\mathbf{x}_j$ and $\mathbf{x}_k$ are obtained by adding $\delta > 0$ to all p components of $\tilde{\mathbf{x}}_j$ and subtracting δ from all p components of $\tilde{\mathbf{x}}_k$, with $\mathbf{x}_i = \tilde{\mathbf{x}}_i$ for all $i \neq j, k$, $i = 1, 2, \cdots, n$. We consider $\delta = 5.0$ for convenience and Table 5.2 summarizes the results.

We see that the regression estimates are more biased when there are two outliers. The estimates based on the outlier-free data and those based on our method are close to the true regression parameters, and the relative biases are also small. For $m = 150$, 200, and 250 the RMCD method identifies the outliers in 95.11%, 97.59%, and 98.30% of the simulations, and the RMVE method identifies them in 93.26%, 96.61%, and 97.64% of the simulations. The method again identified some outliers other than

Table 5.2: Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters under Poisson model with β = (3.0,3.5,0,0,2.0) in presence of two outliers

# of Outliers: 2		m=150			m=200			m=250		
Sample size	Parameter	SM	SSE	RB	SM	SSE	RB	SM	SSE	RB
With Outlier	β_1	1.801	2.746	44	1.860	2.671	43	1.828	2.592	45
	β_2	3.688	3.030	6	3.766	2.897	9	3.875	2.756	14
	β_3	-1.188	3.227	37	-1.203	3.069	39	-1.253	2.994	42
	β_4	-1.233	3.217	38	-1.282	3.040	42	-1.261	2.995	42
	β_5	0.306	2.749	62	0.387	2.661	61	0.446	2.578	60
Without Outlier	β_1	3.000	0.002	1	3.000	0.001	1	3.000	0.001	1
	β_2	3.500	0.002	1	3.500	0.001	1	3.500	0.001	1
	β_3	0.000	0.002	1	0.000	0.002	0	0.000	0.001	2
	β_4	0.000	0.002	0	0.000	0.002	0	0.000	0.001	1
	β_5	2.000	0.002	1	2.000	0.001	0	2.000	0.001	0
Outlier	β_1	2.940	0.668	9	2.963	0.513	7	2.977	0.393	6
down-weighted	β_2	3.517	0.711	2	3.507	0.471	2	3.507	0.383	2
by RMCD	β_3	-0.058	0.777	7	-0.038	0.551	7	-0.021	0.413	5
	β_4	-0.073	0.766	9	-0.039	0.585	7	-0.034	0.466	7
	β_5	1.896	0.743	14	1.945	0.570	10	1.976	0.385	6
Outlier	β_1	2.903	0.846	11	2.954	0.562	8	2.966	0.455	7
down-weighted	β_2	3.526	0.824	3	3.520	0.541	4	3.516	0.464	3
by RMCD	β_3	-0.094	0.936	10	-0.063	0.627	10	-0.036	0.493	7
	β_4	-0.108	0.937	12	-0.046	0.604	8	-0.037	0.530	7
	β_5	1.884	0.840	14	1.933	0.607	11	1.957	0.468	9

those generated, but these proportions are negligible.

5.3.2 Binary Logistic Model

Data with a single outlier. For the contaminated binary model with a single outlier, we first generate n binary responses, $y_1, y_2, \cdots, y_n$, assuming that they do not contain

any outliers. We generated n responses following the binary logistic model $P(Y_i = 1) = \frac{e^{\tilde{\mathbf{x}}_i'\beta}}{1+e^{\tilde{\mathbf{x}}_i'\beta}}$, with n covariates so that $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{ip})$ and $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$. The values of the covariates are chosen as for the Poisson model from MVN(μ, Σ) with p =5. To create an outlier covariate for the jth observation, we change the corresponding covariate values $\tilde{\mathbf{x}}_j$ as for the Poisson model by adding $\delta > 0$ to all p components to get $\mathbf{x}_j$. We again set $\delta = 5$. We retain $\mathbf{x}_i = \tilde{\mathbf{x}}_i$ for all $i \neq j$ as for the Poisson model. Table 5.3 summarizes the results.

Table 5.3: Simulated means (SM), standard errors (SSE), and the relative biases (RB) of estimates of the regression parameters under the binary model with β = (3.0,3.5,0,0,2.0) in the presence of single outlier

Pr. of selection =0.6		m=200			m=250			m=300		
Sample size	Parameter	SM	SSE	RB	SM	SSE	RB	SM	SSE	RB
With Outlier	β_1	2.485	1.035	50	2.531	0.880	53	2.548	0.780	58
	β_2	2.961	1.166	46	3.004	0.994	50	3.024	0.883	54
	β_3	-0.042	0.365	11	-0.039	0.314	12	-0.035	0.285	12
	β_4	-0.048	0.358	13	-0.037	0.317	12	-0.031	0.283	11
	β_5	1.629	0.755	49	1.670	0.640	52	1.678	0.569	57
Without Outlier	β_1	3.329	0.789	42	3.255	0.660	39	3.202	0.569	36
	β_2	3.888	0.923	42	3.798	0.762	39	3.743	0.660	37
	β_3	0.005	0.443	1	-0.002	0.376	0	0.000	0.334	0
	β_4	-0.006	0.437	1	0.000	0.379	0	0.002	0.332	1
	β_5	2.223	0.600	37	2.177	0.505	35	2.135	0.437	31
Outlier	β_1	3.281	0.819	34	3.226	0.678	33	3.182	0.585	31
down weighted	β_2	3.835	0.955	35	3.767	0.782	34	3.721	0.675	33
by RMCD	β_3	0.003	0.438	1	-0.003	0.373	1	-0.001	0.332	0
	β_4	-0.008	0.433	2	-0.002	0.377	0	0.001	0.330	0
	β_5	2.191	0.620	31	2.157	0.515	31	2.122	0.447	27
Outlier	β_1	3.288	0.852	34	3.231	0.671	34	3.181	0.588	31
down weighted	β_2	3.840	0.980	35	3.773	0.783	35	3.714	0.680	31
by RMVE	β_3	0.004	0.436	1	0.001	0.370	0	-0.004	0.335	1
	β_4	-0.004	0.438	1	-0.003	0.371	1	0.002	0.332	0
	β_5	2.188	0.641	29	2.158	0.518	31	2.118	0.449	26

We see that the estimators of the regression parameters are biased by the outliers. The estimates based on the outlier-free data and those based on our method are close to the true parameters, and the relative bias corresponding to these estimators is also small. For $m = 200$, 250, and 300 the RMCD method identifies the outliers in 97.00%, 97.90%, and 98.24% of the simulations, and the RMVE method identifies them in 96.62%, 97.74%, and 98.10% of the simulations. The method again identified some outliers other than those generated, but these proportions are negligible.

Data with two outliers. For the contaminated binary model with two outliers, we first generate n binary responses in a manner similar to that for a single outlier when the covariates are chosen from the MVN(μ, Σ) with dimension p. Suppose that two outlying covariates $\mathbf{x}_j$ and $\mathbf{x}_k$ arise as a result of a shift in the covariate values as for the Poisson model by adding $\delta > 0$ to all p components of $\tilde{\mathbf{x}}_j$ and subtract δ from all p components of $\tilde{\mathbf{x}}_k$, keeping the remaining values of the covariates unchanged. We again set δ to 5 and Table 5.4 summarizes the results. We see that the regression estimates are more biased by two outliers than by one outlier (Table 5.3). The regression estimates based on the outlier-free data and those based on our method are close to the true regression parameters, and the relative bias corresponding to these estimators is small. For $m = 200$, 250, and 300 the RMCD method identifies the outliers in 98.44%, 98.78%, and 99.05% of the simulations, and

Table 5.4: Simulated means (SM), standard errors (SSE), and the relative biases (RB) of estimates of the regression parameters under the binary model with $\beta = (3.0,3.5,0,0,2.0)$ in the presence of two outliers

# of Outlier 1		m=200			m=250			m=300		
Sample size	Parameter	SM	SSE	RB	SM	SSE	RB	SM	SSE	RB
With Outlier	β_1	1.904	0.971	113	1.986	0.849	120	2.071	0.780	119
	β_2	2.327	1.072	109	2.404	0.930	118	2.498	0.862	116
	β_3	-0.080	0.300	27	-0.071	0.269	26	-0.057	0.248	23
	β_4	-0.084	0.306	27	-0.072	0.267	27	-0.061	0.250	24
	β_5	1.215	0.703	112	1.273	0.611	119	1.342	0.568	116
Without Outlier	β_1	3.348	0.816	43	3.266	0.665	40	3.216	0.575	38
	β_2	3.902	0.922	44	3.809	0.759	41	3.753	0.664	38
	β_3	0.007	0.438	2	-0.001	0.375	0	0.004	0.336	1
	β_4	-0.005	0.443	1	-0.003	0.374	1	0.000	0.336	0
	β_5	2.235	0.621	38	2.172	0.499	34	2.147	0.444	33
Outlier	β_1	3.299	0.849	35	3.232	0.688	34	3.194	0.591	33
down weighted	β_2	3.848	0.959	36	3.772	0.784	35	3.729	0.681	34
by RMCD	β_3	0.005	0.434	1	-0.002	0.373	1	0.003	0.334	1
	β_4	-0.007	0.439	2	-0.005	0.372	1	-0.001	0.334	0
	β_5	2.202	0.641	32	2.149	0.514	29	2.131	0.453	29
Outlier	β_1	3.297	0.889	33	3.220	0.684	32	3.187	0.594	31
down weighted	β_2	3.848	1.025	34	3.760	0.787	33	3.721	0.684	32
by RMVE	β_3	-0.005	0.437	1	-0.004	0.378	1	0.000	0.332	0
	β_4	-0.005	0.446	1	0.002	0.374	1	-0.005	0.333	1
	β_5	2.199	0.659	30	2.145	0.523	28	2.124	0.458	27

the RMVE method identifies them in 98.20%, 98.66%, and 99.02% of the simulations. The method again identified some outliers other than those generated, but these proportions are negligible. This shows the effectiveness of our method for estimating regression parameters with a minimal effect from contaminated data.

5.4 Comparison Study

We compared the results of our method with the fully standardized Mallow's type quasi-likelihood (FSMQL) estimation approach of Bari and Sutradhar (2010) in Poisson and binary regression models. The FSMQL approach is a robust version of the quasi-likelihood estimation approach; brief details are given below. Quasi-likelihood estimation produces inconsistent estimates for the regression effects of β when outliers are present in the covariate data for GLMs for binary and count data. The quasi-likelihood estimating equation for estimating β in the GLM is

$$\sum_{i=1}^{n}\left[\frac{\partial\tilde{\mu_i}}{\partial\beta}V^{-1}(\tilde{\mu_i})(y_i-\tilde{\mu_i})\right]=0 \tag{5.17}$$

where $\tilde{\mu_i}=\mathrm{E}[Y_i]=\exp(\mathbf{x}_i'\beta)$ and $\mathrm{V}(\tilde{\mu_i})=\mathrm{var}[Y_i]=\tilde{\mu_i}$ for Poisson count data, and $\tilde{\mu_i}=\mathrm{E}[Y_i]=\frac{exp(\mathbf{x}_i'\beta)}{1+exp(\mathbf{x}_i'\beta)}$ and $\mathrm{V}(\tilde{\mu_i})=\mathrm{var}[Y_i]=\tilde{\mu_i}(1-\tilde{\mu_i})$ for binary data.

Cantoni and Ronchetti (2001) introduced a working Mallow's type quasi-likelihood (WMQL) approach. They suggested reducing the effect of outliers by introducing Huber's robust function for $r_i=\frac{(y_i-\mu_i)}{\sqrt{V(\mu_i)}}$ as

$$\psi_c(r_i)=\begin{cases} r_i & \text{if } \mid r_i \mid \leq c \\ c\ sign(r_i) & \text{otherwise} \end{cases} \tag{5.18}$$

where c is a tuning constant. The WMQL estimating equation is

$$\sum_{i=1}^{n}\left[w(\mathbf{x}_i)\frac{\partial\tilde{\mu_i}}{\partial\beta}V^{-1/2}(\tilde{\mu_i})\psi_c(r_i)-a(\beta)\right]=0 \tag{5.19}$$

where $\mathrm{a}(\beta) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i)\frac{\partial\tilde{\mu}_i}{\partial\beta}V^{-1}(\tilde{\mu}_i)E\left[\psi_c(r_i)\right]$, with $\tilde{\mu}_i = \mathrm{E}[Y_i]$, $\mathrm{V}(\tilde{\mu}_i) = \mathrm{var}[Y_i]$, and $w(x_i) = \sqrt{(1-h_i)}$ for both Poisson and binary data, where h_i is the ith diagonal element of the hat matrix $\mathrm{H} = X(X'X)^{-1}X'$ with $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$ being the $n \times p$ covariate matrix.

Bari and Sutradhar (2010) introduced FSMQL estimation approaches by modifying the robust weights and gradient functions to $\mathrm{var}(\psi_c(r_i))$ and $\frac{\partial\psi_c(r_i)}{\partial\beta}$ respectively. They demonstrated that the FSMQL approach produces almost unbiased and hence consistent estimates for the regression effect when outliers are present in the covariate data. The FSMQL estimating equation is

$$\sum_{i=1}^{n}\left[w(\mathbf{x}_i)\frac{\partial}{\partial\beta}\left\{\psi_c(r_i) - \frac{1}{n}\sum_{i}^{n}E(\psi_c(r_i))\right\}\{var(\psi_c(r_i))\}^{-1}\right. \\ \left.\times\left\{\psi_c(r_i) - \frac{1}{n}\sum_{i}^{n}E(\psi_c(r_i))\right\}\right] = 0. \tag{5.20}$$

They named this FSMQL$_1$. In FSMQL$_2$ they used the deviance $\psi_c(r_i) - E(\psi_c(r_i))$ instead of $\psi_c(r_i) - \frac{1}{n}\sum_{i}^{n}E(\psi_c(r_i))$ and the corresponding estimating equation is

$$\sum_{i=1}^{n}\left[w(\mathbf{x}_i)\frac{\partial}{\partial\beta}\left\{\psi_c(r_i) - E(\psi_c(r_i))\right\}\left\{var(\psi_c(r_i))\right\}^{-1}\times\left\{\psi_c(r_i) - E(\psi_c(r_i))\right\}\right] = 0$$

We examined the performance of our method by estimating the regression parameters β under both Poisson and binary models with one or two outliers. The simulation designs considered are similar to those of Bari and Sutradhar (2010) for

meaningful comparisons: n= 60 and p= 2 with $\beta = (\beta_1, \beta_2)$ =(1.0,0.5). We calculated the SM, SSE, and RB of these estimators based on 1000 simulations.

5.4.1 Poisson Case

Data with a single outlier : To generate n count observations with one outlier, first assume that in the absence of outliers, $y_1, y_2, \cdots, y_n$ are generated following the Poisson density $P(Y_i = y_i) = \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!}$, with $\mu_i = e^{\tilde{\mathbf{x}}_i'\beta}$ where $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2})$. The values of these two covariates are chosen from

$$\tilde{x}_{i1} \overset{iid}{\sim} N(0.5, 0.25) \text{ and } \tilde{x}_{i2} \overset{iid}{\sim} N(0.5, 0.5)$$

for $i = 1, 2, \cdots, n$. To make y_j the outlying response, shift the values of $\tilde{x}_{j1}$ and $\tilde{x}_{j2}$ as follows:

$$x_{j1} = \tilde{x}_{j1} + \delta \text{ and } x_{j2} = \tilde{x}_{j2} + \delta,\ \delta > 0$$

and set δ =2.0. Retain $x_{i1} = \tilde{x}_{i1}$ and $x_{i2} = \tilde{x}_{i2}$ for all $i \neq j$. Thus, $y_1, y_2, \cdots, y_n$ are a sample of n count observations with y_j as the single outlier.

Data with two outliers. For the Poisson model with two outlying observations, the count responses are generated in a manner similar to that for a single outlier. The two covariates $\tilde{x}_{i1}$ and $\tilde{x}_{i2}$ are chosen as

$$\tilde{x}_{i1} \overset{iid}{\sim} N(1.25, 0.25) \text{ and } \tilde{x}_{i2} \overset{iid}{\sim} N(2.25, 0.5).$$

After generating n count observations from a Poisson model with these covariate values, we create two outliers by shifting the covariate values $\tilde{x}_{j1}$, $\tilde{x}_{j2}$ and $\tilde{x}_{k1}$, $\tilde{x}_{k2}$ as follows:

$$\tilde{x}_{j1} = \tilde{x}_{j1} + \delta \text{ and } \tilde{x}_{j2} = \tilde{x}_{j2} + \delta,\ \delta > 0$$

$$\tilde{x}_{k1} = \tilde{x}_{k1} - \delta \text{ and } \tilde{x}_{k2} = \tilde{x}_{k2} - \delta,\ \delta > 0$$

and $\tilde{x}_{i1} = x_{i1}$ and $\tilde{x}_{i2} = x_{i2}$ for all $i \neq j, k$, $i = 1, 2, \cdots, n$. We again set $\delta = 2.0$. Table 5.5 summarizes the results for the Poisson model, here we reproduce results for FSMQL methods from Bari and Sutradhar (2010).

Table 5.5: Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters for sample of size = 60 under Poisson model with β = (1.0,0.5) in presence of one or two outliers

		$FSMQL_1$		$FSMQL_2$		RMVE method		Outlier-free data	
# of Outliers	Statistic	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1	SM	0.850	0.506	0.849	0.496	1.050	0.540	1.050	0.540
	SSE	0.328	0.229	0.322	0.225	0.541	0.531	0.541	0.531
	RB	46	3	47	2	9	8	9	8
2	SM	0.989	0.494	0.994	0.491	0.994	0.503	0.990	0.504
	SSE	0.168	0.089	0.164	0.087	0.110	0.060	0.122	0.064
	RB	6	8	4	10	5	5	8	6

From Table 5.5, we see that the regression parameter estimates using our method are close to the estimated values of the existing methods. They are close to the estimated values from the outlier-free data, and the relative bias is also close to that of the outlier-free data. Note that the method has identified 100% of the outliers in the one-outlier case and 99.35% in the two-outlier case.

5.4.2 Binary Case

Data with a single outlier. For the contaminated binary model with a single outlier, we first generate n binary responses, $y_1, y_2, \cdots, y_n$ assuming that they do not contain any outliers. We generated these n good responses following the binary logistic model $P(Y_i = 1) = \frac{e^{\tilde{\mathbf{x}}_i'\beta}}{1+e^{\tilde{\mathbf{x}}_i'\beta}}$, with two covariates so that $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2})$ and $\beta = (\beta_1, \beta_2)$. Suppose that the values of these two covariates are chosen from

$$\tilde{x}_{i1} \overset{iid}{\sim} N(-1.0, 0.25) \text{ and } \tilde{x}_{i2} \overset{iid}{\sim} N(-1.0, 0.5)$$

for $i = 1, 2, \cdots, n$. To create an outlier covariate $\mathbf{x}_j$, we change the corresponding covariate values $\tilde{x}_{j1}$ and $\tilde{x}_{j2}$:

$$x_{j1} = \tilde{x}_{j1} + \delta_1 \text{ and } x_{i2} = \tilde{x}_{i2} + \delta_2,\ \delta_1, \delta_2 > 0$$

with $x_{i1} = \tilde{x}_{i1}$ and $x_{i2} = \tilde{x}_{i2}$ for all $i \neq j$. We set δ_1 =3.0 and δ_1 =4.0. The remaining covariates are unchanged, i.e., $\tilde{x}_{i1} = x_{i1}$ and $\tilde{x}_{i2} = x_{i2}$ for $i \neq j, k$, $i = 1, 2, \cdots, n$.

The jth response y_j is replaced with a binary value corresponding to P(y_j =1)= π = 0.60 and 0.90.

Data with two outliers. For the contaminated binary model with two outliers, we first generate n binary responses in a manner similar to that for a single outlier with two covariates $\tilde{x}_{i1}$ and $\tilde{x}_{i2}$ chosen from the normal distribution as $\tilde{x}_{i1} \overset{iid}{\sim} N(0, 0.25)$ and $\tilde{x}_{i2} \overset{iid}{\sim} N(0, 0.5)$ for $i = 1, 2, \cdots, n$.

Suppose that two covariate outliers $\mathbf{x}_j$ and $\mathbf{x}_k$ arise as a result of a shift in the covariate values for $\tilde{x}_{j1}$, $\tilde{x}_{j2}$, $\tilde{x}_{k1}$, and $\tilde{x}_{k2}$:

$$\begin{aligned} x_{j1} &= \tilde{x}_{j1} + \delta_1 \quad , \quad x_{j2} = \tilde{x}_{j2} + \delta_2, \\ x_{k1} &= \tilde{x}_{k1} - \delta_1 \quad , \quad x_{k2} = \tilde{x}_{k2} - \delta_2. \end{aligned} \tag{5.21}$$

where $\delta_1, \delta_2 > 0$.

We retain $\tilde{x}_{i1} = x_{i1}$ and $\tilde{x}_{i2} = x_{i2}$ for all $i \neq j, k$. Consequently, for the large values of δ_1 =3.0 and δ_2 =4.0, the covariates corresponding to y_j and y_k become outliers. The j-th response y_j is replaced with a binary value corresponding to probability π = 0.60 and the kth response y_k is replaced with a binary value corresponding to probability π = 0.40. Table 5.6 summarizes the results for the binary model, here too we reproduce results for FSMQL methods from Bari and Sutradhar (2010).

Table 5.6: Simulated means (SM), standard errors (SSE), and relative biases (RB) of estimates of regression parameters for sample of size = 60 under binary model with β = (1.0,0.5) in presence of one or two outliers

		$FSMQL_1$		$FSMQL_2$		RMVE method		Outlier-free data	
# of Outliers	Statistic	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1,	SM	1.046	0.519	1.049	0.512	1.027	0.560	1.027	0.560
π=0.6	SSE	0.802	0.796	0.792	0.785	0.742	0.694	0.742	0.694
	RB	6	2	6	2	4	9	4	9
1,	SM	0.994	0.503	1.003	0.486	1.027	0.560	1.027	0.560
π=0.9	SSE	0.782	0.777	0.779	0.764	0.742	0.694	0.742	0.694
	RB	1	0	0	2	4	9	4	9
2,	SM	1.079	0.545	1.038	0.525	1.078	0.544	1.078	0.544
π=0.6,0.4	SSE	1.098	0.592	1.062	0.572	1.257	0.619	1.257	0.619
	RB	7	8	4	4	6	7	6	7

We see that the parameter estimates using our method are close to the true parameter values, and the relative bias is also small, as for the Poisson model. Note that the estimates based on our method are close to the estimates based on the outlier-free data, and the method has identified 100% of the outliers in the three cases considered.

Chapter 6

Conclusions and Future Work

There is much interest in control charts that monitor the process mean and process variance when individual multivariate observations are collected from an industrial process. The existing methods are influenced by outliers in the Phase-I data, which affect their efficiency in the Phase-II monitoring. Hence, it is important to develop methods that are not unduly influenced by outliers. In this thesis, we have proposed robust control charts using the high-breakdown robust estimation methods RMCD and RMVE to monitor the process mean and the process variance for individual multivariate observations. We have also discussed the use of robust estimation in generalized linear regression models.

We have proposed robust Hotelling's T^2 charts based on the RMCD/RMVE estimators for the Phase-I monitoring of the process mean, when individual multivariate observations are collected. The control limits for these charts are found empirically and a nonlinear regression model is used to find the control limits for any sample size. We studied the performance of our charts under various data scenarios using a large number of Monte Carlo simulations, and they performed better than the standard Hotelling's T^2 chart. We also compared our proposed charts with robust control charts based on MCD/MVE estimators using the concept of the probability of a signal. Our charts provided superior performance. Our simulation studies indicate that RMVE-based charts perform well for smaller sample sizes and smaller dimensions and RMCD-based charts perform well for larger sample sizes and larger dimensions in the case of robust T^2 charts.

We have proposed robust control charts using the MEWMS/MEWMV schemes based on RMCD/RMVE estimators for Phase-I monitoring of the process variance when individual multivariate observations are collected. We compared the performance of our charts under various data scenarios using a large number of Monte Carlo simulations. They perform better than existing charts, namely the MEWMS and MEWMV charts proposed by Huwang et al. (2007) and the $MEWMSL_1$, $MEWMSL_2$, $MEWMVL_1$, and $MEWMVL_2$ charts proposed by Memar and Niaki (2009). The

performance of the charts was studied for small values of the smoothing parameters ω and λ, and they were found to be better than the existing methods. We would like to extend the concept of robust control charts to the Phase-II monitoring of the process variance when individual multivariate observations are collected since detecting process variability changes is often more critical for improving quality than detecting process mean shifts.

Outliers in regression data, especially in the covariates, may unduly influence the estimates of the regression parameters. We have proposed a robust regression approach that identifies and down-weights these outliers using the squared robust Mahalanobis distance based on the RMCD/RMVE estimators of the covariate data. We assessed the performance of our method using a large number of Monte Carlo simulations. We showed that it is effective in freeing the GLM regression estimators from the effects of outlying covariates. We would like to extend the use of robust estimates of the multivariate mean and covariance matrix to regression models with correlated data.

Bibliography

[1] Alt, F.B., Smith, N.D. (1988). Multivariate Process Control. *Handbook of Statistics.* Krishnaiah, P.R., Rao, C.R. (Eds.), Elsevier Science Publishers: New York, 333-351.

[2] Bari, W., Sutradhar, B.C. (2010). On Bias Reduction in Robust Inference for Generalized Linear Models. *Scandinavian Journal of Statistics* **37**, 109-125.

[3] Cantoni, E., Ronchetti, E., (2001). Robust Inference for Generalized Linear Models. *Journal of American Statistical Association* **96**, 1022-1030.

[4] Chenouri, S., Steiner, S., and Variath, A.M. (2009). A Robust Multivariate Control Chart for Individual Observations. *Journal of Quality Technology* **41**, 259-271.

[5] Croux, C., Haesbroeck, G. (1997). An Easy Way to Increase the Finite Sample Efficiency of the Re-sampled Minimum Volume Ellipsoid Estimator. *Computational Statistics & Data Analysis* **25**, 125-141.

[6] Croux, C., Haesbroeck, G. (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis* **71**, 161-190.

[7] Croux, C., Haesbroeck, G. (2002). A Note on Finite-Sample Efficiencies of Estimators for the Minimum Volume Ellipsoid. *Journal of Statistical Computation and Simulation* **72**, 585-596.

[8] Davies, P.L. (1987). Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *Annals of Statistics* **15**, 1269-1292.

[9] Davies, P.L. (1992). The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. *Annals of Statistics* **20**, 1828-1843.

[10] Donoho D.L. (1982). *Breakdown Properties of Multivariate Location Estimators.* Ph.D. qualifying paper, Harvard University.

[11] Donoho D.L., Gasko, M. (1992). Breakdown Properties of Location Estimates Based on Half Space Depth and Projected Outlyingness. *Annals of Statistics* **20**, 1803-1827.

[12] Donoho, D.L., Huber, P.J. (1983). The Notion of Breakdown Point. In: *A Festschrift for Erich Lehmann*, Bickel P., Doksum K., Hodges J. (Eds.), pp. 157-184, Belmont, CA: Wadsworth.

[13] Hawkins, D.M. (1981). A CUSUM for a Scale Parameter. *Journal of Quality Technology* **13**, 228-231.

[14] Hawkins, D.M. (1991). Multivariate Quality Control Based on Regression-Adjusted Variables. *Technometrics* **33**, 61-75.

[15] Hawkins, D.M., Olive, D.J. (1999). Improved Feasible Solution Algorithm for High Breakdown Estimation. *Computational Statistics and Data Analysis* **30**, 1-11.

[16] Hotelling, H. (1947). In: *Techniques of Statistical Analysis*, McGraw Hill, New York, C. Eisenhart, H. Hastay, and W.A. Wallis (Eds.), pp. 111-184.

[17] Huber, P.J. (1981). *Robust Statistics.* John Wiley & Sons, New York.

[18] Huwang, L., Yeh, A.B., and Wu, C.V. (2007). Monitoring Multivariate Process Variability for Individual Observations. *Journal of Quality Technology* **39**, 258-278.

[19] Jensen, W.A., Birch J.B., and Woodall W.H. (2007). High Breakdown Estimation Methods for Phase-I Multivariate Control Charts. *Quality and Reliability Engineering International* **23**, 615-629.

[20] Levinson, W.A., Holmes, D.S., Mergen, E.A. (2002). Variation Charts for Multivariate Processes. *Quality Engineering*, **14(4)**, 539545.

[21] Lopuhaä, H.P. (1989). On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance. *Annals of Statistics* **17**, 1662-1683.

[22] Lopuhaä, H.P., Rousseeuw, P.J. (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Annals of Statistics* **19**, 229-248.

[23] Lowry, C. A., Woodal, W.H., Champ, C.W., Rigdon, W.E. (1992). A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics* **34**, 46-53

[24] MacGregor, J.F., Harris, T.J. (1993). The Exponentially Weighted Moving Variance. *Journal of Quality Technology* **25**, 106-18.

[25] Mahalanobis, P.C. (1936). On the Generalised Distance in Statistics. *Proceedings of the National Institute of Sciences of India* **2**, 49-55.

[26] Maronna, R.A. (1976). Robust M-Estimators of Multivariate Location and Scatter. *Annals of Statistics* **4**, 51-67.

[27] Memar, A.O., Niaki, S.T.A. (2009). New Control Charts for Monitoring Covariance Matrix with Individual Observations. *Quality and Reliability Engineering International* **25**, 821-838.

[28] Pison, G., Van Alest, S., and Willems, G. (2002). Small Sample Corrections for LTS and MCD. *Metrika* **55**, 111-123.

[29] Rocke, D.M., Woodruff, D.L. (1996). Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association* **91**, 1047-1061.

[30] Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. In: *Mathematical Statistics and Applications B*, W. Grossmann, G. Pflug, I. Vincze, and W. Werz (Eds.), Reidel, 283-297.

[31] Rousseeuw, P.J., Leroy, A.M. (1987). Robust Regression and Outlier Detection, John Wiley & Sons, New York, NY.

[32] Rousseeuw, P.J., Van Driessen K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* **41**, 212-223.

[33] Rousseeuw P.J., Van Zomeren B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of American Statistical Association* **85**, 633-639.

[34] Rousseeuw, P.J., Yohai, V. (1984). Robust Regression by Means of S-Estimators. In: *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Hardle, and R.D. Martin (Eds.), Lecture Notes in Statistics **26**, Springer, New York, 256-272.

[35] Stahel, W.A. (1981). Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators, Ph.D. Thesis, ETH, Zurich.

[36] Sullivan, J.H., Woodall, W.H. (1996). A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* **28**, 398-408.

[37] Titterington, D.M. (1975). Optimal Design: Some Geometrical Aspects of D optimality. *Biometrika* **62**, 313-319.

[38] Tracy, N.D., Young, J.C., and Mason, R.L. (1992). Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* **24**, 88-95.

[39] Vargas, J.A. (2003). Robust Estimation in Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* **35**, 367-376.

[40] Vargas, J.A., Lagos, C.J. (2007). Comparison of Multivariate Control Charts for Process Dispersion. *Quality Engineering* **19:3**, 191-196.

[41] Woodal, W.H., Ncube, M.M. (1985). Multivariate CUSUM Quality Control Charts. *Technometrics* **27**, 285-292.

[42] Woodruff, D.L., Rocke, D.M. (1994). Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators. *Journal of American Statistical Association* **89**, 888-896.