EFFICIENT AND HIGHLY ROBUST HOTELLING T^2 CONTROL CHARTS USING REWEIGHTED MINIMUM VECTOR VARIANCE

HAZLINA BINTI ALI

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2013

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to :

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Abstrak

Carta kawalan Hotelling T^2 adalah alat yang berkesan bagi kawalan proses berstatistik untuk persekitaran multivariat. Walau bagaimanapun, prestasi carta kawalan Hotelling T^2 tradisional yang menggunakan penganggar lokasi dan serakan klasik biasanya dicemari oleh kesan pelitupan dan *swamping*. Bagi mengurangkan masalah ini, penganggar teguh telah disyorkan. Penganggar teguh yang paling popular dan digunakan secara meluas dalam carta kawalan Hotelling T^2 adalah penentu kovarians minimum (MCD). Terkini, penganggar yang lebih baik dikenali sebagai varians vektor minimum (MVV) telah diperkenalkan. Penganggar ini mempunyai titik kerosakan yang tinggi, varians samaan affin dan pengiraan yang lebih cekap. Oleh kerana cirinya yang baik, kajian ini mencadangkan untuk mengganti penganggar klasik dengan penganggar lokasi dan serakan MVV dalam pembinaan carta kawalan Hotelling T^2 bagi cerapan individu pada analisis Fasa II. Walau bagaimanapun, penganggar MVV didapati mempunyai beberapa kelemahan seperti tidak tekal pada taburan normal, tidak saksama untuk sampel bersaiz kecil dan kurang cekap pada titik kerosakan yang tinggi. Bagi meningkatkan ketekalan dan kesaksamaan MVV, penganggar tersebut telah didarabkan masing-masing dengan faktor ketekalan dan faktor pembetulan. Bagi mengekalkan titik kerosakan di samping mempunyai kecekapan statistik yang tinggi, penganggar MVV berpemberat semula (RMVV) telah dicadangkan. Seterusnya, penganggar RMVV tersebut digunakan dalam pembinaan carta kawalan Hotelling T^2. Carta teguh Hotelling T^2 yang baharu ini menghasilkan kesan positif dalam mengesan titik terpencil dan pada masa yang sama mampu mengawal kadar penggera palsu. Di samping analisis terhadap data simulasi, analisis ke atas data sebenar juga mendapati carta teguh Hotelling T^2 yang baharu ini dapat mengesan cerapan luar kawalan dengan lebih baik berbanding carta lain yang diselidik dalam kajian ini. Berdasarkan prestasi yang baik terhadap analisis data simulasi dan sebenar, carta teguh Hotelling T^2 yang baharu ini adalah merupakan alternatif yang baik bagi carta Hotelling T^2 yang sedia ada.

Kata kunci: Penganggar Cekap, Kawalan Proses Berstatistik Multivariat, Varians Vektor Minimum Berpemberat Semula, Carta Hotelling T^2 Teguh, Penganggar Multivariat Teguh

Abstract

Hotelling T^2 control chart is an effective tool in statistical process control for multivariate environment. However, the performance of traditional Hotelling T^2 control chart using classical location and scatter estimators is usually marred by the masking and swamping effects. In order to alleviate the problem, robust estimators are recommended. The most popular and widely used robust estimator in the Hotelling T^2 control chart is the minimum covariance determinant (MCD). Recently, a new robust estimator known as minimum vector variance (MVV) was introduced. This estimator possesses high breakdown point, affine equivariance and is superior in terms of computational efficiency. Due to these nice properties, this study proposed to replace the classical estimators with the MVV location and scatter estimators in the construction of Hotelling T^2 control chart for individual observations in Phase II analysis. Nevertheless, some drawbacks such as inconsistency under normal distribution, biased for small sample size and low efficiency under high breakdown point were discovered. To improve the MVV estimators in terms of consistency and unbiasedness, the MVV scatter estimator was multiplied by consistency and correction factors respectively. To maintain the high breakdown point while having high statistical efficiency, a reweighted version of MVV estimator (RMVV) was proposed. Subsequently, the RMVV estimators were applied in the construction of Hotelling T^2 control chart. The new robust Hotelling T^2 chart produced positive impact in detecting outliers while simultaneously controlling false alarm rates. Apart from analysis of simulated data, analysis of real data also found that the new robust Hotelling T^2 chart was able to detect out of control observations better than the other charts investigated in this study. Based on the good performance on both simulated and real data analysis, the new robust Hotelling T^2 chart is a good alternative to the existing Hotelling T^2 charts.

Keywords: Efficient Estimators, Multivariate Statistical Process Control, Reweighted Minimum Vector Variance, Robust Hotelling T^2 Chart, Robust Multivariate Estimator

Acknowledgement

I wish to express my sincere appreciation to those who have contributed to this thesis and supported me in one way or the other during this amazing journey.

Firstly, my sincere appreciations to my supervisor Associate Professor Dr. Sharipah Soaad Syed Yahaya without whose guidance, support, patience and encouragement, this study could not have materialized. I am indeed deeply indebted to her. My sincere thanks also to my co-supervisor, Professor Dr. Zurni Omar for his encouragement and support throughout this study. I would also like to thank Universiti Utara Malaysia (UUM) for sponsoring my study.

Thanks to Professor Dr. Maman A. Djauhari, for his guidance and all the useful discussions and brainstorming sessions, especially during the conceptual development stage. To all of my friends who had directly or indirectly lend me their friendship, moral support and endless encouragement during my study, thank you from the bottom of my heart.

I am deeply grateful to my husband Zainuddin Mohamad for his personal support and for being the good listener I could ever wish for and above all is his great patience at all time. Words cannot express the feelings I have for my parents (Ali Salim and Halimah Akob) and my siblings for emotionally constant support. Finally, to my beloved children Nurqistina, Muhammad Aqil and Nurqaisara that have been a constant source of strength and inspiration. I would not have been here if it is not for you all.

Table of Contents

Permission to Use i

Abstrak ii

Abstract iii

Acknowledgement iv

Table of Contents v

List of Tables xi

List of Figures xiv

List of Appendices xvi

List of Abbreviations xvii

Declaration Associated with this Thesis xix

CHAPTER ONE INTRODUCTION 1

1.1 Introduction 1

1.1.1 Phase I vs. Phase II 4

1.1.2 Hotelling T^2 Control Chart 6

1.2 Problem Statement 12

1.3 Objective 17

1.4 Significance of the Study 18

CHAPTER TWO LITERATURE REVIEW 20

2.1 Introduction 20

2.2 Multivariate Outliers 21

2.3 Robust Estimators 26

2.3.1 Properties of Robust Estimators 26

2.3.2 Types of Robust Estimators 30

2.3.3 MinimumVolume Ellipsoid (MVE) Estimator 30

2.3.4 Minimum Covariance Determinant (MCD) Estimator 31

2.3.5 Reweighted MCD Estimator 35

2.4 Minimum Vector Variance (MVV) 37

2.4.1 Vector Variance 39

2.5 Multivariate Control Chart for Individual Observations. 42

2.6 Robust T^2 Chart 45

CHAPTER THREE METHODOLOGY 53

3.1 Introduction 53

3.2 Minimum Vector Variance (MVV) Estimators 55

3.2.1 MVV Algorithm..57

3.2.2 Computational Efficiency 60

3.3 Robust Hotelling T^2 Control Charts Based On MVV Estimators ($\mathrm{T}^2_{\mathrm{MVV}}$) 62

3.3.1 Estimation of Control Limits 64

3.3.2 Implementation Procedures..65

3.4 Performance Evaluation..66

3.4.1 Choice of Sample Size and Number of Quality Characteristics 67

3.4.2 Types of Contamination and Process of Evaluation 68

3.5 Consistency and Unbiasedness 70

3.6 Reweighted Minimum Vector Variance (RMVV)...71

3.7 Robust Hotelling T^2 Control Charts Based On RMVV Estimators ($\mathrm{T}^2_{\mathrm{RMVV}}$) 72

CHAPTER FOUR ROBUST HOTELLING T^2 CHART WITH MINIMUM VECTOR VARIANCE 73

4.1 Introduction...73

4.2 Simulated Data Analysis...74

4.2.1 Estimation of Control Limits 75

4.2.2 Probability of Detection of Outliers 77

4.2.3 False Alarm Rates 78

4.3 Real Data Analysis 92

4.4 Discussion 99

CHAPTER FIVE THE EFFECT OF CONSISTENT MINIMUM VECTOR VARIANCE ESTIMATORS ON HOTELLING T^2 CONTROL LIMITS 102

5.1 Introduction 102

5.2 Consistency Factor 103

5.3 Correction Factor 105

5.4 Investigation through Simulation Experiment 107

5.5 Estimation of Control Limit 113

5.6 Performance of $\mathrm{T}^2_{\mathrm{MVV}}$ Control Chart 116

5.7 Real Data Analysis 122

5.8 Discussion 124

CHAPTER SIX A ROBUST AND EFFICIENT REWEIGHTED ESTIMATOR OF MULTIVARIATE LOCATION AND SCATTER 127

6.1 Introduction 127

6.2 The Statistical Efficiency of MVV Estimators for Different Breakdown Points 128

6.3 Reweighted Minimum Vector Variance (RMVV) Estimator 130

6.3.1 The Distribution of Robust MSD Based On MVV Estimators................131

6.3.2 The Estimator...135

6.3.3 Algorithm 137

6.4 Efficiency 140

6.5 Finite-Sample Robustness...142

6.6 Discussion 148

CHAPTER SEVEN ROBUST HOTELLING T^2 CHART BASED ON REWEIGHTED VERSION OF MVV ESTIMATOR.......................................150

7.1 Introduction...150

7.2 RMVV Hotelling T^2 Control Chart 151

7.3 Estimation of Control Limits for RMVV Hotelling T^2 Control Chart 152

7.4 Performance Evaluation. 153

7.5 Simulation Results 154

7.5.1 Probability of Detecting Outliers 155

7.5.2 False Alarm Rates 161

7.6 Real Data Analysis 174

7.7 Conclusion 179

CHAPTER EIGHT CONCLUSION AND AREA OF FURTHER RESEARCH 181

8.1 Conclusion 181

8.2 Implications 185

8.3 Limitation 186

8.4 Areas for Further Research 186

REFERENCES 188

List of Tables

Table 3.1: The number of operations 61

Table 3.2: Average number of iterations to compute robust MSD 61

Table 3.3: The values of *n* and *p* 68

Table 4.1: Control limits 76

Table 4.2: False alarm rates for dimension, $p = 2$ 87

Table 4.3: False alarm rates for dimension, $p = 5$ 88

Table 4.4: False alarm rates for dimension, $p = 10$ 89

Table 4.5: False alarm rates for dimension, $p = 15$ 90

Table 4.6: False alarm rates for dimension, $p = 20$ 91

Table 4.7: Historical data set (Phase I data) 95

Table 4.8: Estimates of location vector, covariance matrix and UCL. 96

Table 4.9: The Hotelling T^2 values for the future data (Phase II) 97

Table 5.1: Values of $\vartheta_{\mathrm{p,n}}^{\alpha}$ for $\mathrm{c(h)}S_{\mathrm{MVV}}$ 107

Table 5.2: The slope and R-square for $\mathrm{T}_{\mathrm{MVV(o)}}^{2}$ and $\mathrm{T}_{\mathrm{MVV(I)}}^{2}$ 112

Table 5.3: Control limits 115

Table 5.4: Probability of detection and false alarm rates of the corresponding control charts with dimension, p = 2 117

Table 5.5: Probability of detection and false alarm rates for independent case with dimension, p = 5 118

Table 5.6: Probability of detection and false alarm rates for dimension, p = 10 119

Table 5.7: Probability of detection and false alarm rate for dimension, p = 15 120

Table 5.8: Probability of detection and false alarm rate for dimension, p = 20 121

Table 5.9: Estimates of location vector, covariance matrix and UCL. 123

Table 5.10: Hotelling T^2 values for future data (Phase II) 124

Table 6.1: Asymptotic relative efficiency of MVV scatter matrix estimator w.r.t classical covariance estimator for normal model. 130

Table 6.2: Slope and R-square for MSD_{MCD} and MSD_{MVV} 132

Table 6.3: Asymptotic relative efficiency of the scatter matrix for RMVV estimator with different breakdown point (BP=0.25 and 0.5) w.r.t MVV estimator with (BP=0.5) for normal model. 141

Table 6.4: Asymptotic relative efficiency of the scatter matrix for MVV and RMVV estimator with BP=0.25 with relative to MVV and RMVV estimator with BP=0.5 respectively. 141

Table 6.5: Location estimator: 10% outliers with mean shift 3 (mild contamination) 145

Table 6.6: Location estimator: 10% outliers with mean shift 5 (moderate contamination) 146

Table 6.7: Location estimator: 20% outliers with mean shift 3 (moderate contamination) . 147

Table 6.8: Location estimator: 20% outliers with mean shift 5 (extreme contamination) ... 148

Table 7.1: Probability of detection for the corresponding control charts with dimension, $p = 2$ 156

Table 7.2: Probability of detection for the corresponding control charts with dimension, $p = 5$ 157

Table 7.3: Probability of detection for the corresponding control charts with dimension, $p = 10$ 158

Table 7.4: Probability of detection for the corresponding control charts with dimension, $p = 15$ 159

Table 7.5: Probability of detection for the corresponding control charts with dimension, $p = 20$ 160

Table 7.6: False alarm rate for the corresponding control charts with dimension, $p = 2$ 164

Table 7.7: False alarm rate for the corresponding control charts with dimension, $p = 5$ 166

Table 7.8: False alarm rate for the corresponding control charts with dimension, $p = 10$ ... 168

Table 7.9: False alarm rate for the corresponding control charts with dimension, $p = 15$ ... 170

Table 7.10: False alarm rate for the corresponding control charts with dimension, $p = 20$. 172

Table 7.11. Estimates of location vector, covariance matrix and UCL. 175

Table 7.12. Hotelling T^2 values for future data (Phase II) 176

List of Figures

Figure 4.1: Probability of signal when p = 2.......... 80

Figure 4.2: Probability of signal when p = 5.......... 81

Figure 4.3: Probability of signal when p = 10.......... 82

Figure 4.4: Probability of signal when p = 15.......... 83

Figure 4.5: Probability of signal when p = 20.......... 84

Figure 4.6: Hotelling T^2 control charts.......... 98

Figure 5.1: Determinant of $\vartheta_{p,n}^{\alpha} c(h) S_{MVV}$ when n → ∞ and p = 10 108

Figure 5.2: QQ plot between χ_p^2 distribution versus simulated $T_{MVV(o)}^2$ for n = 10,000.......... 110

Figure 5.3: QQ plot between χ_p^2 distribution versus simulated $T_{MVV(I)}^2$ for n = 10,000 111

Figure 6.1: QQ plot between χ_p^2 distribution versus simulated MSD_{MCD} for n = 10,000.......... 133

Figure 6.2: QQ plot between χ_p^2 distribution versus simulated MSD_{MVV} for n = 10,000.......... 134

Figure 7.1: False alarm when p = 2.......... 165

Figure 7.2: False alarm when p = 5.......... 167

Figure 7.3: False alarm when p = 10.......... 169

Figure 7.4: False alarm when p = 15 171

Figure 7.5: False alarm when p = 20 173

Figure 7.6: Hotelling T^2 control charts 177

Figure 7.7: Hotelling T^2 control charts 178

List of Appendices

Appendix A FLOW CHART FOR PROCESS OF CALCULATING $\mathrm{T}_{\mathrm{MVV}}^{2}$ 196

Appendix B PROGRAM FOR MVV ESTIMATOR 198

Appendix C PROGRAM FOR HOTELLING T^2... 199

List of Abbreviations

ARE	Asymptotic relative efficiency
BP	Breakdown point
CD	Covariance determinant
Fast MCD	Fast minimum covariance determinant
HDS	Historical data set
MCD	Minimum covariance determinant
MLE	Maximum likelihood estimators
MSPC	Multivariate statistical process control
MSD	Mahalanobis squared distances
MSD_{MCD}	Mahalanobis squared distances based on MCD estimators
MSD_{MVV}	Mahalanobis squared distances based on MVV estimators
MSE	Mean squared error
MVN	Multivariate normal distribution
MVE	Minimum volume ellipsoid
MVV	Minimum vector variance
$\mathrm{MVV}_{0.25}$	MVV estimators with breakdown point of 0.25
$\mathrm{MVV}_{0.5}$	MVV estimators with breakdown point of 0.5
n	Sample size
p	Number of dimension
PDS	Positive definite and symmetric matrix
RMCD	Reweighted minimum covariance determinant
RMVV	Reweighted minimum vector variance
$\mathrm{RMVV}_{0.25}$	RMVV estimators with breakdown point of 0.25
$\mathrm{RMVV}_{0.5}$	RMVV estimators with breakdown point of 0.5
SPC	Statistical process control
T_0^2	Traditional Hotelling T^2 chart without cleaning the outliers

T_S^2	Traditional Hotelling T^2 chart with standard approach, cleans the outliers once
T_{MCD}^2	Hotelling T^2 chart based on MCD estimators
T_{RMCD}^2	Hotelling T^2 chart based on RMCD estimators
T_{MVV}^2	Hotelling T^2 based on MVV estimators
$\mathrm{T}_{\mathrm{MVV(o)}}^2$	Hotelling T^2 based on the original T_{MVV}^2
$\mathrm{T}_{\mathrm{MVV(I)}}^2$	Hotelling T^2 based on the improved MVV estimators in terms of consistency and unbiased
$T_{RMVV_{0.25}}^2$	Hotelling T^2 chart based on RMVV estimators with breakdown point of 0.25
$T_{RMVV_{0.5}}^2$	Hotelling T^2 chart based on RMVV estimators with breakdown point of 0.5
UCL	Upper control limit
VV	Vector variance

Declaration Associated with this Thesis

Ali, H., Djauhari, M. A., & Syed-Yahaya, S. S. (2008). *On the distribution of FMCD-based robust mahalanobis distance*. Publish in proceedings of the 3rd International Conference on Mathematics and Statistics (ICoMS-3), Institut Pertanian Bogor, Indonesia. Paper no: 134 -1506.

Ali, H., Djauhari, M. A., & Syed-Yahaya, S. S. (2009). *On Robust Mahalanobis Distance Issued From Fast and MVV*. Publish in book of abstracts, The International Conference of Robust Statistics (ICORS 2009), Universita degli Studi di Parma Facolta di Economia, 14 - 19 June 2009, Parma, Italy. Paper no: 2.

Ali, H., Syed-Yahaya, S. S., & Omar, Z. (2010). *Comparison of Hotelling T^2 Control Chart Based on MVV Robust Estimators for Bivariate Case*. Publish in proceedings of the Conference on Quantitative Sciences and its Aplications, Penang, Malaysia. Paper no: 167.

Ali, H., Syed-Yahaya, S. S., & Omar, Z. (2011). *Efficient and Highly Robust Hotelling T^2 Control Chart for Individual Observations*. Publish in book of abstracts, The International Conference of Robust Statistics (ICORS 2011), Universidad de Valladolid, 27 June- 1 July 2011, Valladolid, Spain. Paper no: 7

Syed-Yahaya, S. S., Ali, H. & Omar, Z. (2011). An Alternative Hotelling T^2 Control Chart Based on Minimum Vector Variance (MVV). *Modern Applied Science*,5(4). Doi:10.5539/mas.v5n4p132.

Ali, H., Syed-Yahaya, S. S. & Omar, Z. (2012). *The Application of Consistent minimum Vector Variance (MVV) Estimators on Hotelling T^2 Control Chart.* Publish in proceedings of Int. Conf. Sci. Tech. & Soc. Sciences under Springer. (In press)

Ali, H. & Syed-Yahaya, S. S. (2013). On Robust Mahalanobis Distance Issued from Minimum Vector Variance. *Far East Journal of Mathematical Sciences (FJMS)*, Volume 74 No. 2, pp. 249-268.

Ali, H., Syed-Yahaya, S. S. & Omar, Z. (2013). Robust Hotelling T^2 Control Chart with Consistence minimum Vector Variance. *Journal of Mathematical Problems in Engineering*. (Accepted)

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Success of a firm very much depends on the quality of its product. Be it goods or services, the firm has little chance of success if its core product is of inferior quality (Ferrel & Hartline, 2008). To ensure that the quality of a product is always up to a certain level, the process behavior needs be monitored and the quality of the process has to be improved. This will consequently lead to business success, growth and enhanced competitiveness. To better meet customers' expectations, many manufacturing industries have reviewed their processes and improve specifications with acceptable standards by reducing variability in the process and product, which substantially will improve performance. Thus, identifying the cause of variation to reduce variability in a process is vital in monitoring quality.

There are two distinct causes of variations in a process namely the common and special cause variations. While common cause variation can be reduced by management intervention, the special cause is hard to gauge as this variation affects the process in unpredictable ways. However, special cause can be detected by some statistical techniques. It can be eliminated from the process by the worker or process control team in charge of the particular segment of the process, which is referred as local action. When all the special-cause variation is eliminated, the process is said to be in-statistical control. The second type of variation, known as common-cause variation, is inherent in the process. This type of variation is predictable probabilistically and randomly distributed. It is the natural variation in a process and typically requires more skill in reducing, which is usually in charge by engineering department. The variation that originates from special causes is generally much greater than it is for common causes.

To examine the source of variations for special causes, manufacturers turn to statistical process control. Statistical Process Control (SPC) is a broad field of research and applications devoted to the improvement of products and processes. The basic procedure of SPC consists of the following steps:

1. the development of a statistical model from historical data collected when the process runs under normal operating conditions;
2. the determination of control limits for the statistical model; and
3. the detection of process faults when on-line data exceeds the control limits, followed by the diagnoses of the cause of the faults.

One of the main challenges faced by SPC is to simultaneously monitor product with multiple quality characteristics especially when the number of characteristics is large. In monitoring the quality of a product or process, quite often more than one quality characteristic are measured on each manufactured item, thus producing a multivariate response. These quality measurements are usually correlated with each other. Multivariate SPC (MSPC) methods are designed by taking into account the correlations among the variables and the ability to simultaneously monitor the variables through time. Like other statistical detection problems, MSPC is concerned with Type I and Type II error. The former which is also known as false alarm, occurs when good data is classified as defective, while the latter occurs when the test fails to detect defective data point. A good and reliable method should be able to control these two errors.

One of the main techniques employed in SPC is process control charts, where the purpose is to achieve and maintain statistical control and capability (Montgomery, 2005). Control charts are known to be effective tools for monitoring the quality of processes in MSPC and are applied in many industries. Data occur sequentially in time and are often reduced to a statistic(s) which represent the current state of the process. The statistics are then plotted on a chart with a process limit identified as the upper and lower control limits (UCL and LCL). The control limits are the common features of the chart, and this chart is specifically known as control chart. A process is deemed stable or in control if all the points (statistics) fall within the limits. Otherwise, the process is signaled as out of control and corrective action on the process may be needed.

The first original study of multivariate control chart was introduced by Hotelling (1947). Since then, as to provide for a wider spectrum, the study on multivariate quality control charts continues to expand. Currently, three of the most frequently considered multivariate control charts are Hotelling's T^2, the MEWMA (Multivariate

Exponentially Weighted Moving Average) and the MCUSUM (Multivariate Cumulative Sum). However, the Hotelling's T^2 is still the most frequently selected tool for multivariate charting procedure due to the fact that T^2 chart possesses almost all the desirable characteristics for a multivariate control chart such as ease of application, flexibility, sensitivity to small process changes, and the availability of software for application (Mason & Young, 2002). Moreover T^2 chart is widely accepted by quality engineers and operators because of its similarity in appearance to the univariate (Shewhart) chart (Prins & Mader, 1997). However, it is not a panacea, as it is not free of limitations.

In the construction of a control chart for monitoring the variability of a univariate or multivariate process, Alt (1985) has defined two phases of the process as phase I and phase II. It is useful to distinguish between methods and applications of the two phases. Although the two phases are both dedicated to identify out-of-control situations, each phase has a unique objective. These phases are also called retrospective and prospective analysis respectively (see Montgomery, 2005).

1.1.1 Phase I vs. Phase II

The purpose of a control chart is to ensure that a process is in control by achieving and maintaining statistical control at each phase of the process. In phase I, a preliminary data set is analyzed to determine whether the process is in control, by establishing the initial control limits and estimating the in-control parameters of the process. The goals of Phase I as stated by Woodall, Spitzner, Montgomery and Gupta (2004) are:

- To understand the variation in a process over time
- To evaluate the process stability
- To model the in-control process performance

This phase also involves the process of detecting outliers that cause the process to become unstable. Thus, in this phase, one needs to identify and remove atypical observations in preliminary data set before the in-control parameters are estimated and the initial control limit is computed. A typical observation located at an extreme distance from the main part of the sample data is considered as a variation due to special-cause. While in phase II, control charts are used with future observations for detecting possible departures from parameters estimated in phase I. The reason we seek to remove the presence of special causes of variation from preliminary data set is due to the fact that their inclusion can result in biased sample estimates of the population mean vector and covariance matrix. The existence of special cause variation would consequently lead to the inflation of control limits and reduction of power to detect process changes in phase II. Therefore a successful phase II analysis very much depends on a successful phase I analysis in estimating in-control mean, variance, and covariance parameters.

1.1.2 Hotelling T^2 Control Chart

Several statistical tests have been presented for identifying the presence of special causes of variation, and one of the most frequently used is the Hotelling T^2 statistics for the reasons mentioned earlier. With the T^2 statistic, the corresponding control chart has only a UCL since all the generated values are positive. The computation of UCL offers some differences between Phase I and Phase II, due to the distinction of the T^2 statistic probability distribution. The main purpose of the Hotelling T^2 statistics in Phase I control chart is to clean the preliminary data set from multivariate outliers and other distributional deviations. The preliminary data set collected in retrospective analysis involves either initial subgroups or individual observations. In many situations, data are collected according to the rational subgroups concept. Nevertheless, sometimes data come in the form of individual observations especially when the production rate is too slow to conveniently collect subgroup size greater than one. For individual multivariate observations, the parameter estimates for the mean vector and covariance matrix in Phase I is based on pooling all the observations (Jackson, 1985; Tracy, Young & Mason, 1992; Wierda, 1994; Lowry & Montgomery, 1995).

Phase I begins with the cleansing process by first selecting a value for α, i.e. the probability of making a Type I error. The choice of α will directly determine the size of the control region, $1-\alpha$. A type I error is made if an observation is declared as an outlier when in fact it is not. Suppose that the preliminary data set in Phase I consist of $\boldsymbol{x_i} = \{\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{ip}}\}$ where i =1,2,…,n and $\boldsymbol{x_i}$ represents p-dimensional of time ordered vectors that are independent of each other. Hotelling T^2 statistic for individual observations which is similar to Mahalanobis distance, is given as

$$T_i^2 = (\boldsymbol{x_i} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) \tag{1.1}$$

This statistic (T_i^2) is used to monitor the process via a T^2 chart. A T^2 value which exceeds the UCL limit signifies that the corresponding observation is an outlier and should be deleted from the preliminary data set. If $\boldsymbol{x_i}$ is assumed to come from multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and the parameters are known, the UCL for the cleansing process is determined using the chi square (χ^2) distribution as follows,

$$UCL = \chi^2_{(\alpha,p)} \tag{1.2}$$

where $\chi^2_{(\alpha,p)}$ is the upper αth quantile of a chi-square distribution having p degrees of freedom. However, when $\boldsymbol{\mu}$ and Σ are unknown, we estimate these parameters from a historical data set using sample mean vector ($\overline{\boldsymbol{x}}$) and the sample covariance matrix ($\boldsymbol{S}$). Thus, if we consider a sample $\boldsymbol{x_i} = \{\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{ip}}\}$ of p-variate observations, the T^2 statistic for $\boldsymbol{x_i}$ can be constructed in the following manner:

$$T_i^2 = (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^t \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) \tag{1.3}$$

where $\boldsymbol{S} = S_{uv}$and u, v = 1, ... , p. The sample mean and covariance matrix are estimated as;

$$\bar{\boldsymbol{x}}_{\boldsymbol{u}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_{\boldsymbol{iu}} \quad \text{and} \quad \boldsymbol{S} = (n-1)^{-1}\sum_{i=1}^{n}{}' \,(\boldsymbol{x}_{\boldsymbol{iu}} - \bar{x}_u)(\boldsymbol{x}_{\boldsymbol{iv}} - \bar{\boldsymbol{x}}_{\boldsymbol{v}})^t \quad (1.4)$$

The control limit is determined by

$$UCL = \left[\frac{(n-1)^2}{n}\right] B_{\left(\alpha;\frac{p}{2},\frac{n-p-1}{2}\right)} \quad (1.5)$$

where $B_{\left(\alpha;\frac{p}{2},\frac{n-p-1}{2}\right)}$ is the upper αth quantile of a Beta distribution (Mason & Young, 2002). The observations with T_i^2 values greater that UCL, will be deleted from the preliminary data. This signals the possible presence of outlier observations in the process which exist due to special-cause variation. With the remaining observations (preliminary data minus outliers), we calculate new estimates for the mean vector and covariance matrix. Then we calculate UCL using the remaining observations. Again, we remove any outliers identified and repeat the process until a homogeneous set of observations is obtained. The final set of data is the historical data set (HDS). This repeated process is known as iterative re-estimation procedure (Mason & Young, 2002). Once a homogeneous reference HDS is obtained and the common estimates $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are computed using Equation 1.4, the next step is to construct a T^2 control chart for Phase II. Assume the process is being monitored by observing a new single observation vector,$\boldsymbol{x_g} = \{\boldsymbol{x_{g1}}, \boldsymbol{x_{g2}}, \dots, \boldsymbol{x_{gp}}\}$, on p variables at each time point. The T^2 value associated with $\boldsymbol{x_g}$ is given by

$$T_g^2 = (\boldsymbol{x_g} - \overline{\boldsymbol{x}})^t \boldsymbol{S}^{-1} (\boldsymbol{x_g} - \overline{\boldsymbol{x}}) \tag{1.6}$$

At this phase, the control limit is determined by

$$UCL = \left[\frac{p(n+1)(n-1)}{n(n-p)}\right] F_{(\alpha;p,n-p)}(1.7)$$

where n is the size of the HDS and $F_{(\alpha;p,n-p)}$ is the αth quantile of F distribution. In this phase, F distribution is used because the calculation of T^2 statistic involves different data from those used to estimate the parameters.

The traditional T^2 chart works well when number of process variable is not too large i.e. $p < 10$. As the number of variables grows, the efficiency of T^2 chart in detecting shift will depreciate (Mason, Champ, Tracy, Wierda & Young, 1997). In addition, the estimators are easily affected by multivariate outliers. Three major types of multivariate outliers are always discussed in the T^2 control chart i.e. shifts in the mean vector, a departure from the in-control covariance structure (counter-relationship) or combinations of the two situations (Ye, Emran, Chen & Vilbert, 2002). A shift in the mean vector occurs when one or more of the p variables is out of control then causing the mean vector, $\boldsymbol{\mu_0}$, to change to some new vector, $\boldsymbol{\mu_1}$. The situation of counter-relationship on the other hand, occurs when a correlation structure between two or more p variables changed from the variable relationship established in the covariance matrix. Although the T^2 control chart can detect both mean shift and counter-relationship, nevertheless the T^2 control chart is more responsive to counter-relationship than mean shifts because the T^2 control chart relies largely on the correlated structure of variables (covariance matrix) for signal detection. An example is illustrated in Ryan (1989) with two variables and a high positive correlation between the two variables while they are in control. For this example, the T^2 control chart signals an observation with a counter-relationship, but does not signal an observation with an out-of-control mean shift on one variable because both variables shift in the same direction and thus still maintain their relationship of a positive correlation. This implies that detecting outliers via mean shifts is more difficult compared to counter-relationship. Any control chart that can circumvent the shift in the mean often can perform well for other types of changes (Jensen, Birch & Woodall, 2007). In addition, other research has shown that the use of the classical sample covariance matrix, with all the individual observations pooled, impairs the detection of a sustained step shift in the mean vector (Williams, Woodall, Birch & Sullivan, 2006).

Sullivan and Woodall (1996,1998) revealed that the T_i^2 chart constructed based on Equation 1.3 using the covariance matrix calculated from Equation 1.4 is not effective in detecting a shift or trend in the mean vector because the variance estimates inflate when special-cause of variations are present (outliers) in Phase I.

This approach is only effective in detecting a small number of very extreme observations, but failed to detect more moderate outliers (Vargas, 2003). Since the purpose of multivariate control chart is to monitor the stability of a multivariate process, the stability should be achieved when the estimates of means, variances, and covariance of the process variables remain stable. For that reason, observations used for the computation of Hotelling T^2 statistics require the assumption of a multivariate normal distribution. Violation of this assumption can lead to incorrect control limits and reduction of the probability of detection in Phase I, which consequently will cause the probability of the Type I error or false alarm rate to be out of control and the power to detect changes (probability of detection) will be reduced in Phase II process (Chang & Bai, 2004; Ramaker, van Sprang, Westerhuis, & Smilde, 2004)

When working with high dimension multivariate data, there is a high probability that outliers are present in the dataset. The existence of outliers is usually the main cause of the violation of normality assumption. The T_i^2 is the squared distance from the *i*-th data point, $\boldsymbol{x_i}$, to the centre described by the sample mean, $\bar{\boldsymbol{x}}$. Once multiple individuals or clusters of data points are separated from a main group, the sample mean vector, $\bar{\boldsymbol{x}}$, thought to represent the data centre, will likely be pulled away from the middle of the larger group of points. Then, the classical sample mean and sample covariance matrix from Equation 1.4 will be distorted. If that is the case, the UCL given in Equation 1.5 and 1.7 will not be effective in detecting outliers anymore. These effects of outliers or groups of outliers on the sample mean and covariance matrix are typically referred to as *masking* or *swamping* effects. Masking effect exists when the UCL fails to detect the outliers (false negative) while swamping effect which is also known as false positive occurs when observations are incorrectly declared as outliers.

Desirable results could be obtained when the estimator is robust even though the data set contains outliers (Rousseeuw & van Driessen, 1999; Hubert, Rousseeuw & Branden, 2005; Vargas, 2003; Jensen et al., 2007). In contrast, the iterative re estimation procedure fails to deal with the problem of masking and swamping because of its nature in identifying outlier's one point at a time (Chenouri, Steiner and Mulayath, 2009). To address this problem, it is necessary to have a procedure that locates all the outliers simultaneously. Nowadays researchers are focusing on the development of robust multivariate statistical process control methods to handle the problem of outliers. These methods are not entirely distribution free but are less sensitive to the assumption of normality than the usual parametric methods. Robust techniques are specifically designed to be relatively insensitive to outliers (Huber, 1977). Another alternative avenue is to consider statistical methods that are distribution free.

1.2 Problem Statement

Robust estimators are known to be more effective in detecting the deviation of data, or outliers as compared to the classical estimators (Hampel, Ronchetti, Rousseeuw &

Stahel, 1986). There are two approaches to deal with outliers when using robust methods. The first approach is to identify and remove outliers before using the remaining good data points to calculate the classical estimators. The second approach is to use the robust estimators in place of classical estimators (Beckman & Cook, 1983). A wide range of robust estimators of multivariate location and scatter are available; see Maronna and Zamar (2002); Maronna, Martin and Yohai, (2006) for a review. Nonetheless, the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators introduced by Rousseeuw (1985) have received considerable attention by scientific community and widely used in practice. The advantage of using MVE estimators is that, they have high breakdown point of approximately 50% and they are also affine equivariant (Lopuhaa & Rousseeuw, 1991, p. 236). However the computation of MVE estimator is very expensive and it may not even be computationally feasible (Hadi, 1992, p. 762). In addition, there is no fast algorithm known to compute the estimator. This is due to the fact that MVE has poor rate of convergence (Lopuhaa & Rousseeuw, 1991, p. 237) and fails to cope with large sample of more than 30 (Rousseeuw & van Driessen, 1999, p. 213). To alleviate the complexity of MVE, Rousseeuw (1985) also introduced the minimum covariance determinant (MCD) method. MVE and MCD estimators have the same characteristics with respect to affine equivariance, high breakdown value and bounded influence function properties (Rousseeuw & Leroy, 1987). The difference is only in the criteria they used where MVE uses minimizing the volume of the ellipsoid on $h = (n + p - 1)/2$data, while MCD uses minimizing the determinant of the covariance matrix based on the h data. The MCD estimator is more attractive than MVE because it has a better convergence rate of $n^{-1/2}$ compared to $n^{-1/3}$ of MVE (Butler, Davies & Jhun, 1993; Croux & Haesbroeck, 1999) and MCD gives the exact solution (Hadi, 1992; Hubert et al., 2005).

Lopuhaa and Rousseeuw (1991) discovered that the efficiency of high breakdown estimators were quite low, and proposed the reweighted version of MCD (RMCD) to alleviate the problem. Croux and Haesbroeck (1999) employed RMCD and noticed that this approach maintains the breakdown point of the initial MCD estimators, while attaining better efficiency. To compute the initial MCD estimator and its reweighted, various algorithms have been suggested. Most of the algorithms attempt to increase the computational efficiency because to obtain approximate values of these estimators is not only expensive, but could be impossible for large sample sizes with large number of quality characteristics (dimensions). Nevertheless, the main contribution in this domain is the Fast MCD algorithm proposed by Rousseeuw and van Driessen (1999) and improved by Hubert et al. (2005) which is available in many computer packages such as Matlab, R, SAS, and S-Plus. However, Fast MCD is not without limitation. For example, the use of minimum covariance determinant as the objective function in data concentration process can be computationally laborious especially when the data set is of high dimension. On the other hand, as Angiulli and Pizzuti (2005) have pointed out, the computational efficiency is as important as effectiveness. Furthermore, as noted by Fauconnier and Haesbroeck (2009), Fast MCD algorithm may return different results when used repeatedly in the same or in different statistical packages and could be more critical when n/p is small (*np*> 5). To overcome the weaknesses of Fast MCD algorithm, Herwindiati (2006) proposed minimum vector variance (MVV) as an alternative measure of multivariate data concentration. Herwindiati, Djauhari and Mashuri (2007) revealed that MVV was successfully used as an objective function in Fast MCD algorithm to substitute the MCD criterion. The findings showed that MVV is computationally more efficient than Fast MCD and as effective as Fast MCD in labeling outliers. A detail explanation about this method is discussed in Chapter 2 and 3.

The study on the significant role of MVE, MCD and RMCD estimators in scientific application can be easily found in the literature specifically in the construction of robust Hotelling T^2 chart. Vargas (2003) and Jensen et al. (2007) introduced robust control charts based on MVE and MCD estimators for multivariate individual observations. They applied these estimators using the first approach i.e. to identify and remove outliers in Phase I analysis and then calculate the classical estimators using the remaining good data points for Phase II analysis. Through this approach, the computability and breakdown point of the estimators become more important, but statistical efficiency is not as crucial because the highly robust estimators will eventually be replaced by classical estimators in Phase II analysis (Jensen et al., 2007). Nonetheless, they noticed some drawbacks when MVE and MCD were used in Phase I. The T^2 issued from MVE failed to perform under large sample size. Conversely, T^2 issued from MCD needed a larger sample size when large numbers of outliers were present to ensure that MCD estimator did not breakdown and lost its ability especially when monitoring with more variables (p).

To abate the problems, Chenouri et al. (2009) proposed robust Hotelling T^2 chart based on RMCD estimator. Besides possessing the nice properties of MCD estimator, this estimator is not unduly influenced by outliers and is more efficient than MCD. Thus, they used RMCD estimator in place of classical estimators in constructing Hotelling T^2 chart for Phase II data. Using the same approach as Chenouri et al. (2009), Alfaro and Ortega (2009) made a comparison study for the performance of Hotelling T^2 control chart based on robust estimators of MCD, MVE, RMCD, and trimmed estimator. They concluded their work by recommending the use of T^2based on trimmed estimator and RMCD when there are few outliers in the production process because of their ability in controlling false alarm rates. However, in the manufacturing of products which emphasizes more on outliers detection than the false alarms generated (Alfaro & Ortega, 2009), the T^2 based on MCD can be considered as better alternatives. This is due to the fact that the Hotelling T^2 control charts based on MCD performed well in terms of probability of detecting outliers. Theoretically, if the percentage of outliers' detection increases, the chart should also be able to control the overall false alarm rate, α (Jensen et al., 2007). However the finding in Alfaro and Ortega (2009) showed a conflict between the probability of detecting outliers and the ability of robust control chart in controlling the overall false alarm rate when robust charts were used under certain conditions.

To alleviate this conflict, we proposed robust Hotelling T^2 control charts based on recently introduced robust estimator known as minimum vector variance (MVV). MVV estimators possess the nice properties of MCD such as breakdown point and affine equivariant properties. In addition, the estimators have better computational efficiency compared to MCD (Herwindiati et al., 2007; Djauhari, 2007). Due to the nice properties of MVV, we were inspired to investigate on the performance of the estimators by integrating them in the Hotelling T^2 control chart on Phase II data. Since these estimators were used directly in Phase II analysis without any screening process in Phase I, they must always be reliable. Thus, for a more rounded and reliable estimators, we further investigated on other properties which were not discussed before such as consistency, biasness and efficiency. Based on the result of the investigation, the MVV estimators were further improved and used in the Hotelling's T^2 chart.

1.3 Objective

The ultimate goal of this research is to find an alternative Hotelling T^2 control chart which can improve the performance of the existing charts in terms of false alarm rate and probability of outliers detection specifically for individual observations. In achieving this goal, the following objectives need to be accomplished.

1. To investigate and compare the performance of Hotelling T^2 control chart using MVV estimators with the traditional Hotelling T^2 control charts and the robust Hotelling T^2control chart based on MCD estimators.
2. To improve the MVV estimators by adding the proportionality constants to ensure that the estimators are consistent at normal model and unbiased at small samples.
3. To investigate on the performance of Hotelling T^2 control charts using the improved estimators in (2).
4. To develop a new robust estimator known as Reweighted MVV (RMVV), based on MVV algorithm.
5. To investigate and compare the performance of the new robust Hotelling T^2control charts using RMVV in (4) with the Hotelling T^2charts using improved MVV estimators in (2), MCD estimators and Reweighted MCD estimators.
6. To evaluate the performance of the improved and the new robust Hotelling T^2 control charts using real industrial data.

1.4 Significance of the Study

This study contributes towards knowledge development in robust estimation and Statistical Process Control (SPC) especially in the construction of control charts. With regards to robust estimation, some improvements were made on MVV estimators in terms of consistency and biasness, followed by the development of a new robust estimator known as RMVV and its algorithm. The new estimator offers high statistical efficiency, consistent and unbiased. The new estimators when used in the Hotelling T^2 chart can improve the performance of the control chart in monitoring the quality of a product even when dealing with a product of high dimensional quality characteristics. This will consequently reduce the operational cost of the company. Additionally, the researchers in industries will not be constrained with the normality assumption as required by the traditional Hotelling T^2.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Monitoring production process quality using control charts is an area which is actively investigated. Concerted efforts by groups of researchers from the areas of statistical quality control and detection of outliers contribute to the continuous works in improving the existing multivariate control charts so that the product performance is always at its best. Hotelling T^2 statistic was the first statistic known to be used in multivariate control chart. The control chart was then referred to as Hotelling T^2 control chart. The purpose of using this statistic is to monitor the stability of a multivariate process in Phase I and II. Analysis in Phase I seek to identify a stable historical data set (HDS). From this dataset, the in-control mean vector and variance-covariance matrix are estimated, which later will be used in the Phase II analysis. A successful process monitoring in Phase II totally depends on the estimates of the parameters obtained from a stable HDS. However, the classical estimators are easily affected by outliers. The shift in the mean vector is the most difficult types of multivariate normal outliers to be detected when using distance-based method like Hotelling T^2 (Rocke & Woodruff, 1996). The existence of outliers can violate the normality assumption. This violation may lead to the inflation of control limits and reduction of the probability of detection in Phase I, which consequently will cause the probability of the Type I error or false alarm to be out of control and the power to detect changes will be reduced in Phase II process. False alarm rate is the probability of out-of-control signal when a process is in control. The value becomes large if the process is unstable due to increase in variability. Inflated false alarm rate can lead to unnecessary process adjustments and loss of confidence in the control chart as a monitoring tool (Chang & Bai, 2004). However, a control chart with small false alarm also has its downside such that the chart tends to be less sensitive to process fault, and it may result in large detection delay (Chen, 2010). Hence, a method which can control the false alarm rate to the desired level is necessary.

2.2 Multivariate Outliers

The study of outliers is as important for multivariate data as it is for univariate samples (Barnett & Lewis, 1994). Nevertheless, it is more difficult to detect outliers in multivariate than univariate data. There are various definitions given for outliers. An exact definition of an outlier often depends on the hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnet and Lewis (1994) indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs, similarly, Johnson (1992) defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. Meanwhile, Beckman and Cook (1983) interpret an outlier as a collective to refer to either a discordant or contaminate observation. Discordant observation is any observation that appears surprising or discrepant to investigator. Contaminant is any observation that is not a realization from the target distribution. Rousseeuw and van Zomeren (1990) also define an outlier to be contaminating data. No matter how it is defined, in general, outliers refer to a point/s that is surprisingly different from the rest of the data. An immediate consequence of the presence of outliers is that they may cause apparent non-normality.

As the dimensions of the data increase, the presence of outliers in the datasets will also increase. Beckman and Cook (1983) had discussed in detail on the need to study on outliers and their effect on linear models. They stated that the existing outliers in the data will affect the estimation of a population parameter, hence, causing the inability of the model to provide an adequate fit or statistical explanation. The presence of outliers can hardly be detected using naked eyes when the dimension is more than 2. This is the risk the researchers have to be cautious about when working with large datasets of high dimensions. Thus, a reliable method is needed to identify outliers especially for this sort of datasets. In their attempt to transform random vectors to be random variables so that outliers could be seen more clearly, Beckman and Cook (1983) suggested the most popular transformation that is via Mahanalobis distance. The works on Mahalanobis distance could be found in almost any literature on multivariate analysis, including outliers' studies.

Mahanalobis square distance (MSD) is a prominent method for outlier detection using the classical mean vector, $\overline{\boldsymbol{x}}$, and covariance matrix, $\boldsymbol{S}$, by assuming this estimation is close to the true values of location vector $\boldsymbol{\mu}$ and shape matrix Σ, and is formulated as follows,

$$d_i^2(\boldsymbol{x_i}, \overline{\boldsymbol{x}}) = (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^t \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})$$

For multivariate normally distributed data, MSDs are approximately chi-square distributed with p degrees of freedom (χ_p). An outlier would then be defined as an observation having larger distance value than the critical value i.e. $\chi_{p,\alpha}$ (Mardia, Kent & Bibby, 2000; Serfling, 1980). Since this study is based on individual observations, the formula for Hotelling T^2 chart which is similar to Mahalanobis distance is given as

$$T_i^2(\boldsymbol{x_i}, \overline{\boldsymbol{x}}) = (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^t \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) \tag{2.1}$$

The T^2 statistic uses the statistical distance that incorporates the multivariate variance-covariance matrix to measure the distance of an observation from the multivariate mean vector of a population. However, the T^2 statistic is sensitive to outliers. This statistic works well with single outliers but is not suitable for applications where multiple outliers are possible due to the effect of outliers on the classical estimates (Alfaro & Ortega, 2009).

As mentioned in the previous chapter, the three types of multivariate outliers which always appear in T^2 control chart are shifts in the mean vector, departure from the in-control covariance structure (counter-relationship) or the combinations of the two situations. However, the T^2 control chart is less effective in detecting mean shift as compared to detecting counter relationship. The performance of T^2statistic is also influenced by masking and swamping effect due to the non-robustness (sensitiveness) of the classical estimators to outliers. These estimators are sensitive to outliers and will be greatly influenced by their presence. The effect of masking and swamping defined by Barnett and Lewis (1994) and Davies and Gather (1993) are as follows,

> ***Masking effect***. It is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small. Therefore we fail to detect the outliers (false negative).

> ***Swamping effect***. It is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers. Therefore the observations are incorrectly declared as outliers (false positive).

Problems of masking and swamping can be resolved by using robust estimates of scatter and location, where they are less affected by outliers. Beckman and Cook (1983) considered robust estimation to be one of the best ways to accommodate outliers in the estimation problems and encouraged the routine use of the estimators. Rousseeuw and Leroy (1987) also mentioned that the use of robust estimates of the multidimensional distribution parameters can often improve the performance of the detection procedures in the presence of outliers. Thus, the development of robust estimation is indeed needed to prevent these errors from influencing the statistical model.

2.3 Robust Estimators

Since the assumption of normality required in the classical estimation methods is frequently not satisfied, Huber (1964) suggested the robust estimators. One of the goals using robust estimator, as stated by Hampel (1985), is to identify the deviation of data, or outliers. Compared to the classical methods, robust statistics will give a clearer variability description between an outlier and 'good data', whereas the classical statistics will vaguely show the difference. Robust estimation methods can be used in two different approaches. The first approach is to identify and remove outliers, then use classical estimators on the remaining good data points. In the second approach, the robust estimators are used in place of classical estimators (Beckman & Cook, 1983). In searching for highly robust estimators of location and scatter, there are several qualities that the estimators should possess.

2.3.1 Properties of Robust Estimators

There are four major measures or properties that can be used to determine the usefulness of a robust multivariate estimator (Jensen et al., 2007). The first is the breakdown point (BP), where it was introduced by Hampel (1971, 1974) as a measure for the robustness of an estimator against outliers. The breakdown point is defined as the smallest percentage of outliers that can cause an estimator to take arbitrary large values. Finite sample breakdown point (Donoho & Huber, 1983) is a very popular global measure of robustness. It is the smallest amount of contamination necessary to upset an estimator entirely. Let $X^{(0)} = \{x_1, \dots, x_n\}$ be a random sample of n observations and $t_n(X^{(0)})$ the corresponding estimator if we replace m arbitrary points in $X^{(0)}$ by arbitrary values if the new data is now $X^{(m)}$. Then the finite sample breakdown point for estimator t_n for sample $X^{(0)}$ is

$$\varepsilon_n^*\left(t_n, X^{(0)}\right) = min\left\{\frac{m}{n}; sup_{X^{(m)}}\left|t_n(X^{(m)}) - t_n(X^{(0)})\right|\right\}$$

If $\varepsilon_n^*\left(t_n, X^{(0)}\right)$ is independent of the initial sample $X^{(0)}$, we say the estimator t_n has the universal finite sample breakdown point $\varepsilon_n^*(t_n)$. Therefore $\varepsilon_n^* = \lim_{n\to\infty} \varepsilon_n^*(t_n)$. A higher breakdown point implies more robust estimator. In the univariate case, the usual mean has very low BP which is equals to ${}^{1}/_{n}$, while median possess the maximum possible value with BP = 50%. The higher the BP, the more resistant the estimator is to bad data. In other words, the less susceptible it is to the masking effect. Some literatures say that for realistic applications, a BP greater than or equal 20% is usually satisfactory (Zuo, 2006).

The second property to consider is affine equivariance, which is an important and often desirable property of statistical estimates. When an estimator is affine equivariant, changing the measurement scale or affine transformations should not affect the properties of the estimator. Suppose a random sample $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$ from a *p*-variate normal $MVN_p(\mu, \Sigma)$, we want to estimate μ and Σ then it is desirable that the estimator are independent of the choice of coordinate system. Formally if the estimator for μ and Σ are t_n and c_n respectively then t_n and c_n are called affine equivariant if for any nonsingular $p \times p$ matrix **A** and vector $b \in R^p$

$$t_n(\boldsymbol{A}x_1 + b, \boldsymbol{A}x_2 + b, \dots, \boldsymbol{A}x_n + b) = \boldsymbol{A}t_n(x_1, x_2, \dots, x_n) + b \tag{2.2}$$

$$c_n(\boldsymbol{A}x_1 + b, \boldsymbol{A}x_2 + b, \dots, \boldsymbol{A}x_n + b) = \boldsymbol{A}c_n(x_1, x_2, \dots, x_n)\boldsymbol{A'} \tag{2.3}$$

When an estimator possesses the affine equivariant property, it will not get influenced by an affine transformation. This is an important property that needs to be considered when searching for robust statistics. The estimators of location and dispersion that are considered in this study are all affine equivariant.

The third property is statistical efficiency of the estimator. This property concerns on how well the estimator makes use of all the good data available. Efficiency is always a very important performance measure for any statistical procedure (Zuo, 2006, p.7). In his seminal paper, Huber (1964) took into account both the robustness and the efficiency issues in the famous “minimax” (minimizing worst-case asymptotic variance) approach. Robust estimators are commonly not very efficient. The univariate median serves as a perfect example. It is the most robust affine equivariant location estimator with the best breakdown point and the lowest maximum bias at symmetric distributions (see Huber 1964). Yet for its excellent robustness, it has to pay the price of low efficiency with relative to the mean at normal and other light-tailed models.

Finally, a further important and desirable feature of an estimator is the computational efficiency for easy and fast computation. It is common to measure data in terabytes or megabytes, and some real time applications require the outliers to be detected within seconds or at most a few minutes. Estimating robust estimators of location and scatter with large dimension is one of the primary problems encountered in multivariate settings such as in the area of quality control. Many industries for example healthcare, machinery, agriculture, information, and financial will directly be affected as these industries deal with products of multi-dimensional specifications. The computational time and cost of analyzing the product (data) will escalate as the dimension gets larger, and the probability that outliers will be present in the data sets will increase. With the existence of outliers in the dataset, the application of classical statistical methods such as in quality control will no longer be precise and reliable. Pena and Prieto (2001) stated that it is entirely appropriate to develop special methods to handle special cases. For higher dimension and large multivariate data sets, computational speed seems to be one of the most difficult goals to achieve. Additionally, Angiulli and Pizzuti (2005) have pointed out, the computational efficiency is as important as effectiveness in detecting outliers.

2.3.2 Types of Robust Estimators

A wide range of robust estimators of multivariate location and scatter are available. Some of them are based on the minimization of a robust scale of Mahalanobis distances such as M-estimator (Campbell, 1980), minimum volume ellipsoid (MVE), minimum covariance determinant (MCD) estimates (Rousseeuw 1984, 1985), *S* estimates (Davies, 1987), and τestimates (Lopuhaä & Rousseeuw, 1991). Others are based on projections, for example, the Stahel–Donoho estimate (SDE), *P* estimates (Maronna, Stahel, & Yohai, 1992) and Kurtosis1 (Pena & Prieto, 2001). Nonetheless, the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimator introduced by Rousseeuw (1984;1985) has received a considerable attention by scientific community and widely used in practice.

2.3.3 MinimumVolume Ellipsoid (MVE) Estimator

MVE method uses $h=(n+p-1)/2$ data to construct robust location and scatter estimator (Rousseeuw & van Zomeren, 1990), which give the minimum volume of ellipsoid among all possible subsets of *h*. This estimator is then used to generate the robust Mahanalobis distance. The advantage of using MVE estimators is that they have high breakdown point of approximately 50% and also affine equivariant (Lopuhaä & Rousseeuw, 1991). However the computation of MVE estimators is very expensive and it may not even be computationally feasible (Hadi, 1992). In addition, there is no fast algorithm known to compute the MVE estimators. This is due to the fact that MVE has poor rate of convergence (Lopuhaä & Rousseeuw, 1991) and fail to cope with large sample of more than 30 (Rousseeuw & van Driessen, 1999).

2.3.4 Minimum Covariance Determinant (MCD) Estimator

To alleviate the complexity of MVE, Rousseeuw (1984; 1985) also introduced the minimum covariance determinant (MCD) method. MVE and MCD estimators have the same characteristics with respect to breakdown point and affine equivariant properties (Rousseeuw & Leroy, 1987). The only difference is in the criteria used such that MVE minimizes the volume of the ellipsoid on $h=(n+p-1)/2$ data, while MCD minimizes the determinant of the covariance matrix based on the h data. The MCD estimator is more attractive than MVE because it has a better convergence rate of $n^{-1/2}$ compared to $n^{-1/3}$ of MVE (Butler et al., 1993).

However, computing the exact MCD estimators is very expensive or even impossible for large sample sizes in high dimensions (Woodruff & Rocke, 1994). Various algorithms have been suggested to obtain an approximate value for this estimator. Most of them are to increase the computational efficiency. For example, feasible solution algorithm (FSA) in Hawkins (1994) and Hawkins and Olive (1999), MULTOUT in Woodruff and Rocke (1994), Fast MCD algorithm in Rousseeuw and van Driessen (1999), block adaptive computationally-efficient outlier nominators (BACON) in Billor, Hadi, and Vellemen (2000), improved Fast MCD algorithm in

Hubert et al. (2005). The most recent work on improving the algorithm was proposed by Herwindiati (2006) using variance vector instead of covariance determinant in Fast MCD algorithm. However, the main contribution in this domain is the Fast MCD algorithm which has been available in many computer packages such as Matlab, R, SAS, and S-Plus. Furthermore, its applications can be found in a very wide spectrum area, for example, multivariate statistical process control, multivariate process capability analysis, information sciences, data depth, data mining and etc. Thus, this proves that Fast MCD is very well accepted as an algorithm for MCD robust estimators.

For a finite sample of observation $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ in $\mathbb{R}^p$ the MCD is determined using the Fast MCD algorithm by selecting the subset X =$\{x_{i1}, x_{i2}, \dots, x_{in}\}$ of size h yielding the maximum possible breakdown point, i.e.

$$h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$$

where $\lfloor \, . \, \rfloor$ denotes the greatest integer part. A larger value of h would result in more efficient estimates, but at the expense of a reduced breakdown value (Croux & Haesbroeck, 1999). The MCD estimators are estimated with $1 \le h \le n,$ which minimizes the determinant of covariance matrix i.e. $|S|$ among all possible subsets of size h. The main method used in the estimation of MCD is Mahalanobis squared distances (MSD). The squared distances for the sample are defined as

$$d_i^2 = (\boldsymbol{x_i} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) i = 1,2, \dots n \tag{2.4}$$

The Fast MCD algorithm had been developed by Rousseeuw and van Driessen (1999) and consists of the following concentration steps. Let H_{old} be an arbitrary subset containing h data points.

1. Take a subset from X as H_{old} containing $h = \left[\frac{n+p+1}{2}\right]$ data points. Compute the mean vector $\boldsymbol{\overline{x}_{H_{old}}}$ and covariance matrix $\boldsymbol{S_{H_{old}}}$of all observations belonging to H_{old}
2. Compute the MSDs $d_{H_{old}}^2(i)$ for $i = 1, \dots, n$.
3. Sort these MSDs in ascending order. This ordering defines a permutation π on the index set.
$$d_{H_{old}}^2(\pi(1)) \le d_{H_{old}}^2(\pi(2)) \le \dots \le d_{H_{old}}^2(\pi(n))$$
4. Let H_{new} be the new subset of h elements indexed by $\pi(1), \pi(2), \dots, \pi(h)$
5. Calculate $\boldsymbol{\overline{x}_{H_{new}}}$, $\boldsymbol{S_{H_{new}}}$and $d_{H_{new}}^2$. If det($\boldsymbol{S_{H_{new}}}$) = det($\boldsymbol{S_{H_{old}}}$) the process is stopped, else, when det($\boldsymbol{S_{H_{new}}}$) < det($\boldsymbol{S_{H_{old}}}$) the process is continued and go to step 2. If det($\boldsymbol{S_{H_{new}}}$)=0, repeat step 1-5.

Let det($\boldsymbol{S_{H_k}}$) be generated from the k-th iteration. Thus, $\det(\boldsymbol{S_{H_1}}) \ge \det(\boldsymbol{S_{H_2}}) \ge \dots \ge \det(\boldsymbol{S_{H_k}}) = \det(\boldsymbol{S_{H_{k+1}}})$. From the k-th iteration, this algorithm gives $\boldsymbol{\overline{x}_{H_k}} = \boldsymbol{\overline{x}_{MCD}}$, $\boldsymbol{S_{H_k}} = \boldsymbol{S_{MCD}}$.

The location estimator is then defined as

$$\boldsymbol{\overline{x}_{MCD}} = \frac{1}{h} \sum_{i=1}^{h} \boldsymbol{x_i} \tag{2.5}$$

and the estimator of scatter by

$$\boldsymbol{S_{MCD}} = \frac{c(h)s(h,n,p)}{h-1}\sum_{i=1}^{h}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{MCD}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{MCD}})^t \tag{2.6}$$

The proportionality constant, $c(h)$, known as the coefficient of consistent factor makes $\boldsymbol{S_{MCD}}$ Fisher consistent when the distribution of $\boldsymbol{x}$ is elliptically symmetric and unimodal with $\boldsymbol{\mu}$ and dispersion matrix Σ (Butler et al., 1993; Croux & Haesbroeck, 1999). Fisher consistencyis a standard concept in robust statistics which denotes that the functionals evaluated at the model distribution return the true parameter value, Σ (Croux & Rousseeuw, 1992). Fauconnier and Haesbroeck (2009, p.6) had presented two approaches in defining the coefficients of consistency factor for MCD scatter matrix which are theoretical and empirical approach. Theoretical consistency factor was derived by Butler et al. (1993) and further discussed in Croux and Haesbroeck (1999) based on the functional form of the MCD estimator. If $\boldsymbol{x} \sim N(\mu, \Sigma)$, theoretical consistency factor (c_1) is defined as

$$c_1 = \frac{h/n}{P(\chi^2_{p+2} < \chi^2_{p,1-h/n})} \tag{2.7}$$

where χ^2_{p+2} denotes the α cut-off point of the χ^2_p distribution which leaves α of the values at its right. While empirical consistency factor (c_2) given by Rousseeuw and van Driessen (1999, p.218) depends on n, p and h (via the estimators $\overline{\boldsymbol{x}}_{\boldsymbol{MCD}}$ and

$\boldsymbol{S_{MCD}}$) and more generally on the data at hand. The empirical consistency factor is defined as

$$c_2 = \frac{Med_i d^2_{(\bar{\boldsymbol{x}}_{\boldsymbol{MCD}}, \boldsymbol{S_{MCD}})}(i)}{\chi^2_{p;0.5}} \text{where} i = 1,2,\ldots,n \qquad (2.8)$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{MCD}}$ and $\boldsymbol{S_{MCD}}$ are the MCD estimators computed from the optimal subset of data. Fauconnier and Haesbroeck (2009) stated that, the consistency factor c_2 is frequently referred to in literature as a scaling factor. This factor allows one to improve the distribution of robust distances computed on non-normal data and is used when the exact form of the consistency factor is not known.

The second proportionality constant, $\boldsymbol{S_{MCD}}(h, n, p)$, known as a finite sample correction factor serves the purpose of reducing the small sample bias of $\boldsymbol{S_{MCD}}$. The actual value of this factor depends also on n and p. It was obtain by Pison, van Alest and Willems (2002) through a combination of Monte Carlo simulation and parametric interpolation, under the assumption that $\boldsymbol{S_{MCD}}(h, n, p) \to 1$ as $n \to \infty$ for fixed p.

2.3.5 Reweighted MCD Estimator

Besides high resistance to outliers, if robust multivariate estimators are to be of practical use in statistical inference they should offer a reasonable efficiency under the normal model and a manageable asymptotic distribution (Rousseeuw & van

Zomeren, 1990). Croux and Haesbroeck (1999) verified that MCD estimators are not very efficient at normal models. They showed a conflict between efficiency and breakdown point, where the efficiency of MCD estimators decreases when the breakdown point increases, especially when the number of dimension becomes higher. Since the efficiency of high breakdown methods can be quite low, Rousseeuw and van Zomeren (1990) proposed Reweighted MCD (RMCD) estimator and Lopuhaä and Rousseeuw (1991); Lopuhaä (1999); Croux and Haesbroeck (1999) employed the reweighted version.

The basic concept of one-step reweighted proposed by Rousseeuw and van Zomeren (1990) is to skip those outlying observations and compute the sample mean and covariance matrix of the rest of the data. The RMCD estimators $\bar{\boldsymbol{x}}_{\boldsymbol{RMCD}}$ and $\boldsymbol{S}_{\boldsymbol{RMCD}}$ (shown below) are computed using Fast MCD algorithm in Section 2.3.4 by giving weight $w_i = 0$ to observations for $d^2_{MCD}(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{\boldsymbol{MCD}}) > \chi^2_{p,0.025}$, and $w_i = 1$ otherwise, and $m = \sum_{i=1}^{n} w_i$. The $\alpha = 0.025$ cut-off point of the χ^2_p distribution is suggested by Rousseeuw and van Driessen 1999, Croux and Haesbroeck 1999 and Pison and van Aelst 2004. The formula for RMCD estimators of location and scatter are as follows:

$$\bar{\boldsymbol{x}}_{\boldsymbol{RMCD}} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{m} \tag{2.9}$$

$$\boldsymbol{S}_{\boldsymbol{RMCD}} = c(m)s(m,n,p)\frac{\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\boldsymbol{RMCD}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{\boldsymbol{RMCD}})^t}{m-1} \tag{2.10}$$

The factors $c(m)$ and $s(m,n,p)$ guarantee consistency of the reweighted scatter estimator and improve its small sample behavior, like the corresponding factor in Equation 2.6. The finding from the simulation study by Croux and Haesbroeck (1999) noticed that RMCD maintains the breakdown point of the initial estimators, while attaining better efficiency.

However, Croux and Haesbroeck (1999) also emphasized that, the positive breakdown point is not a guarantee for robustness, since the corresponding bias may become extremely large but still remain bounded. Moreover the gains in efficiency come at the price of a larger bias, as Rousseeuw (1994) well pointed out. The reason is that all these methods are non-adaptive, and higher efficiency can only be obtained by tuning the parameters, which in turn affects the bias under contamination. Through simulation study on finite-sample robustness, Croux and Haesbroeck (1999) have shown that the RMCD with breakdown point of 0.25 is more precise and outperforms RMCD with breakdown point 0.5 under contamination. For that reason, RMCD with breakdown point 0.25 is more acceptable and has been used in the LIBRA package under MATLAB 7.8.0 (R2009a).

2.4 Minimum Vector Variance (MVV)

Although Fast MCD algorithm is well accepted, nevertheless, it is not without limitation. For example, the use of minimum covariance determinant as the objective function in data concentration process can be computationally laborious especially when the data set is of high dimension. On the other hand, as Angiulli and Pizzuti (2005) have pointed out, the computational efficiency is as important as effectiveness. Furthermore, as noted by Fauconnier and Haesbroeck (2009), Fast MCD algorithm may return different results when used repeatedly in the same or in different statistical packages and could be more critical when n/p small. To overcome the weaknesses of Fast MCD algorithm, Herwindiati (2006) proposed minimum vector variance (MVV) as an alternative measure of multivariate data concentration.

Minimum vector variance algorithm was introduced by Herwindiati (2006) for the purpose of increasing the computational efficiency of Fast MCD. Under higher dimensions, the determinant is more complicated to compute. As an alternative measure to the long and tedious computation of covariance determinant in data concentration, minimum vector variance (MVV) was proposed as an alternative measure for multivariate data concentration. The use of vector variance in place of covariance determinant as the objective function of the stopping rule will be discussed in the next section. Herwindianti (2006) and Djauhari, Mashuri, and Herwindiati (2008) have shown that MVV has met three of the four major properties of a good robust estimator namely high breakdown point, affine equivariance and computational efficiency, as discussed in Section 2.3.1. Herwindiati et al. (2007) revealed that the MVV and Fast MCD algorithms have the same structures and only differ in their objective functions. If the objective function of Fast MCD is minimizing the covariance determinant, the MVV is minimizing the vector variance. Their findings showed that MVV is computationally more efficient than Fast MCD and as effective as Fast MCD in labeling outliers. A detailed explanation about this method is discussed in Chapter 3 (Methodology).

2.4.1 Vector Variance

The two popular measures of dispersions in the study of multivariate analysis are the total variance (TV) and the covariance determinant (CD). If $\boldsymbol{x}$ is a random vector of$p$ dimension with Σ as its $(p \times p)$ covariance matrix, then TV = $Tr(\Sigma) = \lambda_1 + \lambda_2 + \dots + \lambda_p$ and CD = $|\Sigma| = \lambda_1 \lambda_2 \dots \lambda_p$. Pena and Rodriguez (2003) gave a very comprehensive discussion for the role of TV and CD in measuring the spread of multivariate data. CD has a much more general use than TV in every literature on multivariate analysis. This is because of the unstable TV's role where it only deals with the variance without the involvement of the whole structure of covariance matrices. Although CD has wider applications than TV, however, it has several drawbacks. The main drawback lies in the property of having the covariance determinant zero, $|\Sigma|$= 0. This occurs when there is a variable with variance 0 or when there is a variable which is a linear combination of any other variables (Herwindiati, 2006). The matrix of this condition is known as singular matrix and has no inverse.Because of this drawback, Herwindiati (2006) and Djauhari (2007) proposed another measure of multivariate dispersion based on TV, which is known as vector variance (VV).

Djauhari (2007) introduced and demonstrated how VV represents the degree of variation of multivariate distribution. Consider $\boldsymbol{x}$ and $\boldsymbol{y}$, two random vectors of p and q dimensions where p and q are not necessarily equal, having a covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{11} and Σ_{22} are variance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\Sigma_{12} = \Sigma_{21}^{t}$ is the covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$. To measure the linear relationship between the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, Cleroux and Ducharme (1989) defined the correlation coefficient between them as

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{Tr(\Sigma_{12}\Sigma_{21})}{\sqrt{Tr(\Sigma_{11}^{2})Tr(\Sigma_{22}^{2})}}$$

and called this measure ‘vector correlation’. According to this measure Herwindiati (2006) and Djauhari (2007) call $Tr(\Sigma_{12}\Sigma_{21})$ as the vector covariance between $\boldsymbol{x}$ and $\boldsymbol{y}$ and $Tr(\Sigma_{11}^{2})$ and $Tr(\Sigma_{22}^{2})$ as the vector variance (VV) of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

By definition, VV is the sum of square of all elements on the diagonal of covariance matrix. If $\boldsymbol{x}$ is a random vector of p dimension with Σ as its covariance matrix, then

VV of $\boldsymbol{x}$ is $Tr(\Sigma^2)$. Its value indicates the degree of how multivariate distribution is scattered. The larger the value of VV the more scattered the distribution around its mean vector in a subspace of dimension $q \leq p$. It is equal to zero if and only if the distribution degenerates at the mean vector.

The used of VV instead of covariance determinant as multivariate data concentration measures have several advantages. This matter was discussed by Djauhari (2007). First, its computation is very efficient even for covariance matrix of large size because VV is of quadratic form while CD is of multilinear form. Thus, the number of operations of VV is smaller than CD such that VV is of order $O(p^2)$ and CD is of order $O(p^3)$. Second, VV does not need the condition that the covariance matrix is non-singular, unlike CD. The singularity problem usually arises when the number of variable p is larger than the number of sample size n.

Another advantage of VV was illustrated by Djauhari (2007) via comparison of the power (probability of detection) of vector variance-based test with covariance determinant-based test. In general, both tests have similar performance when p is small such as $p = 2$. However, the power of VV is greater than CD to a larger shift of covariance structure when p and n are large. Djauhari (2007) and Djauhari et al. (2008) showed the derivation of the asymptotic distributional properties and the convergence performance of sample VV. They proved that the distribution of sample VV is sufficient to be approximated by the multivariate normal distribution.

When applied on multivariate process variability monitoring, Djauhari et al. (2008) revealed that VV showed better power in detecting the difference between the two covariance structures while CD failed to detect the difference. Another interesting characteristic of VV is its geometric property which is similar to the univariate case such that its value is small if all variables have small variances, and becomes large if at least one variable has a large variance. In the special case where $p = q = 1$, VV equals the square of the classical variance. If we compare the properties of VV with those of CD, we find that CD is a "strong" measurement in the sense that $|\Sigma| = 0$ if at least one variable has zero variance or if there is at least one variable that is a linear combination of the others (Djauhari et al., 2008). On the other hand, VV is "weak" in the sense that $Tr(\Sigma^2) = 0$ if and only if all variables have zero variance. A small value of VV means that all diagonal elements of Σ (variances) are small, however, a large value of VV does not mean that all variables have large variances.

2.5 Multivariate Control Chart for Individual Observations.

A T_i^2 statistic in Equation 2.1 that is based on the classical estimators is equivalent to the Mahalanobis squared distance (MSD). However, there is a problem when using T^2 statistic (or MSD) to detect outliers. The classical estimators are known to be sensitive to the presence of even one outlier. When these estimators are used in T^2 statistic to detect outlier, the process might suffer from the masking and swamping effect. This is because its breakdown point of $1/n$ goes to 0 as the sample size increases (Jensen et al., 2007). That is, a single arbitrarily large outlier can render the T^2 statistic ineffective.

In many situations, multivariate data are collected according to the rational subgroup concept. A rational subgroup represents a sample of data taken at some point in the process. However, sometimes it may not be possible to collect rational subgroup of size larger than one. Dealing with individual observations in the construction of control chart can be more challenging. Jackson (1985), Tracy et al. (1992), Wierda (1994), and Lowry and Montgomery (1995) suggest pooling all the data to estimate the mean vector and covariance matrix. Then, the Hotelling T^2 in Equation 2.1 is calculated for each observation. However Sullivan and Woodall (1996) have shown that by taking the sample covariance matrix from the pooling of HDS lead to poor properties in detecting mean shifts in the mean vector. Moreover Prins and Mader (1997) and Mason, Champ, Tracy, Wierda and Young (1997) had mentioned two weaknesses using this approach. First is the difficulty in obtaining the control limit (UCL) due to the restriction of the multivariate normal distribution assumption. A sufficiently large preliminary data set is needed to obtain a reasonably accurate control limits if the violation occurs. However to increase sample size for example is often impractical or too expensive. Second, the pooling of data may include out-of-control samples in the historical data set which consequently may cause an adverse effect on the phase II control limits.

Xu (2003) investigated on the effect of the violation of normality assumption on the performance of T^2 chart. The study revealed that in case of false alarm rate of $\alpha =$ 0.05, 0.01 and 0.001, the UCL which was obtained from beta distribution (Equation 1.5) was overestimated when the real data come from uniform distribution (0,1). On the other hand when the data came from exponential distribution with mean 2, the UCL was under estimated. Because of the differences between the desired false alarm rate and the observed rates, the author concluded that these UCLs may not be appropriate to represent the upper control limit in Phase II if the actual distribution is not normal.

To circumvent these problems, one has to be able to identify and eliminate apparent outliers from the data. Once a homogeneous HDS is obtained, one can then perform pooling and use the mean vector and covariance matrix for future data sets consisting of individual observations (Prins & Mader, 1997). One natural approach to overcome these effects is to substitute into Equation 2.1 with estimators of the mean vector and covariance matrix that are not affected by outliers or groups of outliers.

One suggested approach is to use a covariance matrix estimator based on successive differences which is robust to a sustained shift in the mean vector. This was proven by Sullivan and Woodall (1996) when their proposed SW1 technique showed that the T^2 chart using the sample covariance matrix failed in detecting shifts in the mean. To rectify the problem, they suggested that the covariance matrix should be estimated using the vector difference between two successive observations $\boldsymbol{v_i} = x_{i+1} - x_i$, where i = 1, 2,..,n-1. Nonetheless, the approach was effective in detecting small number of outliers but otherwise for large number of outliers. Instigated by the weaknesses, Sullivan and Woodall (1998) proposed another approach based on Atkinson and Mulira (1993) stalactite chart known as SW2. This idea is based on Mahalanobis distance with multiple-step method. It started by randomly selecting $(p + 1)$ observations to calculate the mean and covariance matrix, and then used in Mahalanobis distance. Next, the $(p + 2)$ observations with the smallest distance were selected to calculate new estimates. The process continued by adding one by one observation until all the observations are included. Therefore at each step, outliers were removed until the final subset included all the observations except the outliers. Again, this technique remains vulnerable to data that contains large number of outliers and also depends on the robustness of its initial random sample (Vargas, 2003). Another approach is to use robust estimators of the process parameters.

2.6 Robust T^2 Chart

Robust estimation has been a useful approach in the area of statistics due to the good properties shown under some deviations of distributional assumptions. In MSPC, this type of estimation is widely used and investigated by researchers. Thus, searching for reliable estimators become the main research topic for those in the area of MSPC. It is necessary to introduce robust estimators in T^2 chart, but this has to be done with caution. The robust estimator must have nice properties such as affine equivariance, high breakdown point, asymptotic normality, high in statistical and computational efficiency for the chart to be reliable. From literatures, various types of robust estimators were suggested for control charts, but the most popular estimators are MVE and MCD.

Estimation of MVE and MCD was introduced in T^2 chart in two different approaches. The first approach is to use these robust estimators to identify and remove outliers in Phase I analysis and then use the classical estimators on the remaining good data points for Phase II analysis. Using this approach, the computability and breakdown point of the estimator become more important, but statistical efficiency is not as crucial because the highly robust estimators will eventually be replaced by classical estimators in Phase II analysis (Jensen et al., 2007). The second approach is using these robust estimators which are calculated at Phase I and then used directly in Phase II analysis. This approach does not have to go through the process of outliers cleaning in Phase I by assuming that these robust estimators are not influenced by outliers. However the robust estimators should possess higher statistical efficiency (Chenouri et al., 2009). Researchers working on the construction of robust Hotelling T^2 chart, incessantly are trying to introduce various types of robust estimators to improve the performance of the process by using either one of these approaches.

The MVE and MCD robust estimators were first introduced in the construction of T^2 control chart, based on the first approach. Vargas (2003) proposed three robust estimators in constructing T^2 control chart for identifying multiple outliers and a step shift in the mean vector for multivariate individual observations in Phase I. They suggested minimum volume ellipsoid (MVE), minimum covariance determinant (MCD) and a trimmed type estimator (trimming of extreme values is determined by using Mahalanobis distance). The three robust T^2control charts were compared with the traditional T^2control chart and two more alternative T^2control charts i.e. the successive difference estimator of covariance matrix (SW1) and the outlier detection algorithm (SW2). The study also tested all the charts using real data from Quesenberry (2001). The MVE and MCD estimators were obtained using sub-sampling and Fast MCD algorithm respectively. Vargas (2003) consider $ncp = (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^t\Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$ as the non-centrality parameter that measures the severity of a shift from the in-control mean vector ($\boldsymbol{\mu_0} = 0$ and $\Sigma = \boldsymbol{I_p}$) to the out-of-control mean vector ($\boldsymbol{\mu_1} = 5, 15, 25$) with dimension p = 3, 5, 10 and sample size of n = 30, 50, 100. The k outliers was generated with different values for each n, where for *n:30, k=2, 4, 6* ; *n:50, k=2, 5, 10* ; *n:100, k=5, 10, 20* observations. Control limit were set based on 5000 simulation and all methods had an overall false alarm probability of $\alpha = 0.05$. Performance evaluation was based on the detection of outliers and false detection probability (false alarm rate) for different *ncp* values. Based on the simulation result, the study recommended using T^2 chart based on

MVE estimators for detecting multiple outliers and T^2 chart based on SW1 to detect step shifts in the mean vector. Both estimators were compared for the case of p=2 and n=30, and the result demonstrated that MVE showed better performance in terms of probability of outliers' detection. However the robust procedure (MVE and MCD estimators) under their study are not sensitive to step shift in the mean vector.

Jensen et al. (2007) detected some disadvantages in using sub-sampling algorithm by Vargas (2003) in calculating the MVE estimator. This algorithm would generate the different estimates value depending on the number of subsamples used. They compared MVE estimators based on sub-sampling algorithm with MCD based on Fast MCD algorithm, but with more combinations of p, n, and k. The performance evaluation considered dimensions p = 2, 3, 5, 7, 10 with sample size of n = 20,…,100 and k = *0, 2, 4,…, 48* random data points generated from the out-of-control distribution and the other $n - k$ observations were generated from the in-control distribution with significance level of $\alpha = 0.05$. The in-control distribution was a multivariate normal where it could be assumed that $\boldsymbol{\mu} = 0$and $\Sigma = \boldsymbol{I_p}$ without loss of generality. The out-of-control distribution was a multivariate normal with the same variance-covariance matrix but the mean vector had been shifted by some amount. Since the finite sample distribution of MCD and MVE estimators were unknown, the UCLs were determined based on the generation of 200,000 data sets from $MVN_p(0, \boldsymbol{I_p})$ for each combination of n and p. The T^2 statistic for each observation in the data set was calculated and the maximum value attained for each data set was recorded. The 95th percentile of this generated empirical distribution was the simulated control limit. Performance evaluation of the robust control charts was based on the probability of a signal for out-of-control data (probability of detection) and the probability of a signal for in-control data (false alarm rate). When the value of the non-centrality parameter is small (towards in-control process), the probability of a signal is close to $\alpha = 0.05$. As the value of the non-centrality increase (towards out-of-control process) the probability of a signal will increase. From the simulation result, they concluded that classical estimator should be used if only one outlier is expected. When $n \leq 50$, the MVE will be the best estimator, unless the percentage of outliers is greater than 25% or 30%. On the other hand, when n>50, the MCD is preferred as long as the percentage of outliers is less than 40%. They noticed that there are some drawbacks when MVE and MCD are used in Phase I. First, the ability of MVE and MCD estimator in detecting outliers decreases in the case of high dimensions. Second, although MVE performs well in detecting small or large number of outliers, it is only computationally feasible when the sample size is small. Meanwhile, the MCD has limited ability in detecting outliers which need larger sample size if the data is suspected of having large number of outliers. Jensen et al. (2007) also noted that when monitoring with more variables (p), larger sample sizes are needed to ensure that the MCD estimator does not breakdown and lose its ability to detect any outliers.

To abate the problems, Chenouri et al. (2009) proposed alternative estimator known as reweighted MCD (RMCD) introduced by Lopuhaä and Rousseeuw (1991) and Willems, Pison, Rousseeuw and van Alest (2002). Besides inheriting the nice properties of MCD estimator, this estimator is not unduly influenced by outliers and has high statistical efficiency than MCD. For that reason they introduced RMCD in T^2 chart using the second approach i.e. they proposed robust control charts for Phase II data based on the RMCD estimates of location and scatter parameters from Phase I. In identifying control limit, Chenouri et al. (2009) applied the Slutsky theorem when the finite sample distribution of the MCD and RMCD estimators was unknown with large sample size (n > 200) where T^2_{RMCD} has an asymptotic χ^2_p distribution. However, for small sample size ($n \leq 200$), they estimated appropriate quantiles (99% and 99.5%) of T^2_{RMCD} with sample size n, dimension p and breakdown point $1-\alpha$. The performance of robust T^2 control chart was judged based on the probability of detecting changes in the process behavior of the Phase II data, which was different from the data structure of Phase I. The change in the process was based on the shift in the process mean vector $ncp = (\boldsymbol{\mu} - \boldsymbol{\mu_0})^t\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0})$ assuming that there was no change in covariance structure. They made a comparison study for the performance of Hotelling T^2 control chart based on classical and robust estimators of MVE, MCD and RMCD. The simulation results showed that the robust control chart based on RMCD performed better than others methods.

Using the same approach as Chenouri et al. (2009), Alfaro and Ortega (2009) conducted a comparison study on the performance of Hotelling T^2 control chart based on robust estimators of MCD, MVE, RMCD, and trimmed estimator. They analyzed and compared the performance of the classical T^2 with the robust control charts based on different sample sizes n = 25, 50, 100 and 1000 with dimensions of p = 2, 3 and 5 and the overall false alarm probability of α = 0.05 and 0.01. Since the distributions of the alternative statistics were unknown, the reference control limits were determined by simulation technique similar to Vargas (2003) and Jensen et al. (2007). They assessed the performance from two perspectives, namely false alarm rates and probability of detection. Alfaro and Ortega (2009) concluded their work by recommending the use of T^2 based on trimmed estimator and RMCD when there are few outliers in the production process due to the charts' good control of false alarm rate. However, in product manufacturing which emphasizes more on outliers detection as compared to the false alarms generated, then T^2 based on MCD were better alternatives since the charts performed well in terms probability of outliers detection. In theory, if the percentage of outliers' detection increases, the chart should also have the ability to control the overall false alarm rate, α (Jensen et al., 2007). With regards to the aforementioned problems, Vargas (2003), Jensen et al., (2007), Alfaro and Ortega (2009) and Chenouri et al., (2009) tried incessantly to improve the performance of the control chart by using good robust estimators. However, their works were restricted on small and medium dimensions only due to the complexity of the high dimension data. Maintaining good performance on high dimensional data has its advantageous and also disadvantageous. While trying to preserve the identity of the original variable without reducing the dimension, researchers have to compromise with computational efficiency. The computational time and cost of analyzing the product (data) will escalate as the dimension gets larger, furthermore the probability of the presence of outliers will also increase. Computational efficiency is one of the main problems that need to be addressed in multivariate settings especially in the area of quality control (Mason and Young, 2002, p.9). MVV estimators was proven to be computational efficient. In addition, this estimator has high breakdown point, affine equivariance and more importantly the Σ does not need to be positive definite. Furthermore, the vector variance is not limited to low dimension and can be used efficiently for high dimension data set as well as on non-singular and singular covariance matrix. Due to the nice characteristics of MVV, thus, this study uses MVV estimators in the construction of Hotelling T^2 control chart.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

The purpose of this study is to develop a better robust estimator to be used in Hotelling T^2 control chart in Phase II. The estimator should be suitable for either lower or higher dimension, possess high computational efficiency, has good control of the false alarm rate and effective in detecting changes. One of the latest offerings in the study of robust estimators in the multivariate data is the minimum variance vector (MVV) proposed by Herwindiati (2006). Apart from being on par with the popular minimum covariance determinant (MCD) (Rousseeuw and van Driessen, 1999) for its robustness, this estimator has the edge over MCD in terms of computational efficiency. Industries would prefer procedures with high computational efficiency especially when dealing with high dimensional quality characteristics.

At the beginning of this chapter, we will discuss briefly on the characteristic of MVV which motivate us to propose these estimators to be used in Hotelling T^2 in place of the usual mean vector and covariance matrix. Then we formally introduce robust control chart based on the MVV estimators (T^2_{MVV}) on Phase II data. In order to assess the performance of T^2_{MVV} control charts, various conditions were created by manipulating the number of observations (n), number of dimensions or quality characteristics (p), proportion of outliers (ε) and mean shifts (non-centrality) values (μ_1). The performance of T^2_{MVV} control chart was evaluated based on the assumption that there are no changes in the covariance structure. Performance evaluation measured the effectiveness in terms of the probability of outlier(s) detection and false alarm rate (type I error) on Phase II data. It is worthwhile to investigate on the performance using both measures because they are closely related (Ramaker van Sprang, Westerhuis & Smilde, 2004). When the data comes from an in-control process the false alarm rate should be close to a nominal value, α. In this study, α was set to be equal to 0.05 by referring to Vargas 2003, Jensen et al. (2007), Chenouri et al. (2009), Alfaro and Ortega (2009). When data comes from an out-of-control process then the probability of detection should be large enough to ensure that the chart is able to monitor on-line data and quickly detect shifts in the process of Phase II.

The MVV estimators are expected to perform well in Hotelling T^2 control chart in terms of controlling false alarm rate, improving the probability in detecting outliers and simultaneously increasing the computational efficiency. Since these estimators are used directly in Phase II without the process of outliers cleaning, they must be statistically efficient (refer to Chapter 2 Section 2.5). For better efficiency, we then proposed the reweighted version of MVV after making MVV estimators consistent and unbiased. The subject on consistency and unbiasedness is discussed in detail in Chapter 5 while the issue on efficiency is discussed in Chapter 6. In Chapter 7,the application of reweighted version of MVV on the Hotelling T^2 chart is thoroughly explained.

3.2 Minimum Vector Variance (MVV) Estimators

Herwindianti (2006) and Herwindiati et al. (2007) had proved that MVV estimators possess three major properties of a good robust estimator i.e. high breakdown point (BP=0.5), affine equivariance and computational efficiency. Interestingly, MVV estimator has the same characteristics as MCD with respect to breakdown point and affine equivariance property. Like MCD, the main method used in the estimation of MVV is the Mahalanobis squared distances (MSD) which is defined as in Equation (2.4). Let $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ be a set of *p*-variate observations. Denote the MVV estimators for the location parameter and scatter by $\boldsymbol{m_{MVV}}$ and $\boldsymbol{S_{MVV}}$ respectively. Now let $H \subseteq X$, the $\boldsymbol{m_{MVV}}$ and $\boldsymbol{S_{MVV}}$ are determined based on the set *H* consisting of $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ data that produces $\boldsymbol{S_{MVV}}$ with minimum $Tr(\boldsymbol{S^2_{MVV}})$ among all possible sets of h data. The MVV algorithm that had been discussed in Herwindiati et al. (2007) is akin to the Fast MCD algorithm proposed by Rousseeuw and van Driesen (1999), except for some changes in the concentration step (C-step) where the computation of covariance determinant is replaced by the vector variance. The basic theorem of Fast MCD algorithm as introduced by Rousseeuw and van Driessen (1999, p. 214) is stated in Theorem 3.1 below,

Theorem 3.1: *Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ be a sequence of i.i.d random vectors of p dimension where the second moment exist. Let H_1 be a subset of $\Omega = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$ of h elements, and $\boldsymbol{\bar{x}_1} = \frac{1}{h}\sum_{x_i \in H_i} \boldsymbol{x_i}$ and $\boldsymbol{S_1} = \frac{1}{h}\sum_{x_i \in H_i}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_1})(\boldsymbol{x_i} - \boldsymbol{\bar{x}_1})^t$ be its mean vector and covariance matrix. If $|\boldsymbol{S_1}| \neq 0$, let $d_i^2 = (\boldsymbol{x_i} - \boldsymbol{\bar{x}_1})^t \boldsymbol{S_1}^{-1}(\boldsymbol{x_i} - \boldsymbol{\bar{x}_1})$ for all $i = 1,2, \ldots, n$ and $H_2 = \{\boldsymbol{x_{\sigma(1)}}, \boldsymbol{x_{\sigma(2)}}, \ldots, \boldsymbol{x_{\sigma(h)}}\}$ where σ is a permutation on the index set such that $d_{\sigma(1)}^2 \leq d_{\sigma(2)}^2 \leq \cdots \leq d_{\sigma(n)}^2$. If $\boldsymbol{\bar{x}_2}$ and $\boldsymbol{S_2}$ are respectively the mean vector and covariance matrix associated to H_2, then $|\boldsymbol{S_2}| \leq |\boldsymbol{S_1}|$ with equality if and only if $\boldsymbol{\bar{x}_2} = \boldsymbol{\bar{x}_1}$ and $\boldsymbol{S_2} = \boldsymbol{S_1}$.*

Based on this theorem, H_2 is more concentrated than H_1 in the sense that the volume of the smallest ellipsoid that covers H_2 is less than or equal to that of the smallest ellipsoid that covers H_1. Equivalently, the determinant of the covariance matrix $\boldsymbol{S_2}$ of all vectors belonging to H_2 is smaller than or equal to that of the covariance matrix $\boldsymbol{S_1}$ of all vectors in H_1. This necessary condition for H_2 to be more concentrated than H_1 is then used by Rousseeuw and van Driessen (1999) in developing Fast MCD. They define “minimizing covariance determinant (CD)" as the objective function in their C-step. We know that determinant operator is in multilinear form. Thus, when the number of variables *p* gets larger, the computational efficiency of CD dwindles rapidly. Specifically, the number of operations in the computation of CD is of order $O(p^3)$. The works of Herwindiati (2006), Herwindiati et al. (2007), and Djauhari

(2007) lead to another notion of data concentration which is demonstrated in Theorem 3.2.(Djauhari, Adnan, Lee & Ali, unpublished manuscript)

Theorem 3.2: *Let $\boldsymbol{x_1},\boldsymbol{x_2},\dots,\boldsymbol{x_n}$ be a sequence of i.i.d random vectors of p dimension where the second moment exist. Let H_1^* be a subset of $\Omega = \{\boldsymbol{x_1},\boldsymbol{x_2},\dots,\boldsymbol{x_n}\}$ of h elements, and $\overline{\boldsymbol{x}}_{\boldsymbol{1}}^* = \frac{1}{h}\sum_{x_i \in H_i^*} \boldsymbol{x_i}$ and $\boldsymbol{S}_{\boldsymbol{1}}^* = \frac{1}{h}\sum_{x_i \in H_i^*}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\boldsymbol{1}}^*)(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{1}}^*)^t$ be its mean vector and covariance matrix. If $|\boldsymbol{S}_{\boldsymbol{1}}^*| \neq 0$, let $\delta_i^2 = (\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{1}}^*)^t {\boldsymbol{S}_{\boldsymbol{1}}^*}^{-1}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{1}}^*)$ for all $i = 1,2,\dots,n$ and $H_2^* = \{\boldsymbol{x}_{\boldsymbol{\sigma}^*(\boldsymbol{1})},\boldsymbol{x}_{\boldsymbol{\sigma}^*(\boldsymbol{2})},\dots,\boldsymbol{x}_{\boldsymbol{\sigma}^*(\boldsymbol{h})}\}$ where σ^* is a permutation on the index set such that $\delta_{\sigma^*(1)}^2 \leq \delta_{\sigma^*(2)}^2 \leq \cdots \leq \delta_{\sigma^*(n)}^2$. If $\overline{\boldsymbol{x}}_{\boldsymbol{2}}^*$ and $\boldsymbol{S}_{\boldsymbol{2}}^*$ are respectively the mean vector and covariance matrix associated to H_2^*, then $Tr|(\boldsymbol{S}_{\boldsymbol{2}}^*)^2| \leq Tr|(\boldsymbol{S}_{\boldsymbol{1}}^*)^2|$ with equality if and only if $\overline{\boldsymbol{x}}_{\boldsymbol{2}}^* = \overline{\boldsymbol{x}}_{\boldsymbol{1}}^*$ and $\boldsymbol{S}_{\boldsymbol{2}}^* = \boldsymbol{S}_{\boldsymbol{1}}^*$.*

In Theorem 3.2, the role of CD as a multivariate dispersions measure is replaced by the sum of squares of all elements of covariance matrix, which is the vector variance (VV) or $Tr({\boldsymbol{S}_{\boldsymbol{MVV}}}^2)$.

3.2.1 MVV Algorithm

The search for a minimum $Tr({\boldsymbol{S}_{\boldsymbol{MVV}}}^2)$ for each *H* subset requires a finite number of steps to achieve convergence. However, that is no guarantee that the final value $Tr({\boldsymbol{S}_{\boldsymbol{MVV}}}^2)$ of the iteration process is the global minimum of the MVV objective function. Therefore, an approximate MVV solution can be obtained by taking many initial choices of H_1 subsets, applying C-step for each subset and later choose a specific number of subsets (e.g. 10) that produce the lowest vector variance. For ease of understanding, the MVV algorithm is partitioned into two stages. The first stage involves creating initial subsets, while the second stage is the concentration steps. Let $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$ be a *p*-variate random sample of size *n*.

Stage 1: Creating Initial Subsets.

This stage is repeated 500 times

1. Draw a random subset (H_0) with number of observations, $h = p + 1$. Compute the mean vector $\boldsymbol{\bar{x}_{Ho}}$ and covariance matrix $\boldsymbol{S_{Ho}}$.

   $$\boldsymbol{\bar{x}_{Ho}} = average(H_0) \quad \text{and} \quad \boldsymbol{S_{Ho}} = cov(H_0)$$

2. Compute the MSDs $d_0^2(i) = (\boldsymbol{x_i} - \boldsymbol{\bar{x}_{Ho}})^t \boldsymbol{S_o}^{\boldsymbol{-1}} (\boldsymbol{x_i} - \boldsymbol{\bar{x}_{Ho}})$ for $i = 1, \ldots, n$.
3. Sort these MSDs in ascending order, $d_0^2(\pi(1)) \leq d_0^2(\pi(2)) \leq \ldots \leq d_0^2(\pi(n))$. This ordering defines a permutation π on the index set.
4. Take a new subset $H_1 = \{\pi(1), \ldots, \pi(h)\}$ where $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$, then calculate $\boldsymbol{\bar{x}_{H_1}}$, $\boldsymbol{S_{H_1}}$, $Tr(\boldsymbol{S_{H_1}}^2)$and compute MSD, where

   $$d_1^2(i) = (\boldsymbol{x_i} - \boldsymbol{\bar{x}_{H_1}})^t \boldsymbol{S_{H_1}}^{-1} (\boldsymbol{x_i} - \boldsymbol{\bar{x}_{H_1}}) \quad \text{for } i = 1, \ldots, n.$$

5. Repeat step 3 and 4 for H_2
6. Sort the 500 values of $Tr(\boldsymbol{S_{H_2}}^2)$ in ascending order, then select 10 subsets of H_2 which have the lowest $Tr(\boldsymbol{S_{H_2}}^2)$. These subsets are treated as the initial

subsets and their mean vectors, $\bar{\boldsymbol{x}}_{\boldsymbol{H_2}}$ and covariance matrices, $\boldsymbol{S}_{\boldsymbol{H_2}}$ will be used in Stage 2.

Stage 2: Concentration Steps (C-step)

This process will be repeated until convergence for each of the 10 subsets. Convergence means $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) = Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$, where *k* is number of iterations.

1. Compute the MSDs by using $\bar{\boldsymbol{x}}_{\boldsymbol{H_2}}$ and $\boldsymbol{S}_{\boldsymbol{H_2}}$, where

   $d_2^2(i) = (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_2}})^t {\boldsymbol{S}_{\boldsymbol{H_2}}}^{-1} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_2}})$ for $i = 1, \dots, n$.

2. Repeat step 3 and 4 in Stage 1 until $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) = Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$. If $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) > Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$ the process is continued. This process will be repeated until convergence is achieved.

3. When convergence is achieved for all the 10 subsets, choose the subset (*H**) that generates the lowest $Tr(\boldsymbol{S}_{\boldsymbol{H_k}}^2)$. From *H**, calculate $\bar{\boldsymbol{x}}_{\boldsymbol{H^*}} = \boldsymbol{m}_{\boldsymbol{MVV}}$ and $\boldsymbol{S}_{\boldsymbol{H^*}} = \boldsymbol{S}_{\boldsymbol{MVV}}$ as the location and scatter estimators for *MVV* respectively.

From the *k*-th iteration, this algorithm gives $\bar{\boldsymbol{x}}_{\boldsymbol{H^*}} = \boldsymbol{m}_{\boldsymbol{MVV}}$, $\boldsymbol{S}_{\boldsymbol{H^*}} = \boldsymbol{S}_{\boldsymbol{MVV}}$ as the location and scatter estimators for MVV respectively. The location estimator is defined as

$$\boldsymbol{m}_{\boldsymbol{MVV}} = \frac{1}{h}\sum_{i=1}^{h} \boldsymbol{x_i} \tag{3.1}$$

and the scatter estimator by

$$\boldsymbol{S}_{\boldsymbol{MVV}} = \frac{1}{h}\sum_{i=1}^{h} (\boldsymbol{x_i} - \boldsymbol{m}_{\boldsymbol{MVV}})(\boldsymbol{x_i} - \boldsymbol{m}_{\boldsymbol{MVV}})^t \tag{3.2}$$

3.2.2 Computational Efficiency

Theoretically, it is clear that the objective function of minimizing VV is computationally more efficient than the initial objective function of minimizing CD because VV is in quadratic form while CD is of multilinear form. In terms of the number of operations, VV is of $O(p^2)$ while CD as mentioned before is of $O(p^3)$ (Herwindiati et al., 2007; Djauhari et al.,2008). To verify this statement, we carried out an investigation to compare the number of operations in the computation of VV and CD for several values of p. Results for the number of operations is shown in Table 3.1. Our finding discovers that, the number of operations of CD tends to be equal to $\frac{2}{3}p$ times more than VV when p gets larger. For example, for p = 75, the number of operations of CD is approximately 50 times more than VV.

To illustrate on the computational efficiency of this algorithm as compared to Fast MCD algorithm, we presented a simulation study focusing on the number of iterations necessary for robust MSD issued from MVV estimators as well as on Fast MCD in the concentration steps (C-steps). The MVV algorithm was executed using MATLAB 7.8.0 (R2009a), while Fast MCD algorithm using *mcdcov.m* in the LIBRA package under MATLAB 7.8.0 (R2009a). Random data were generated from p-variate standard normal distribution $N_p(0, I)$ for several values of p with a constant n = 100 based on 100 replications. Table 3.2 displays the result of the average number of iterations. We find that the speed of convergence of MVV is higher than

Fast MCD. This certainly reduces the time consumption. Moreover, unlike Fast MCD, MVV algorithm is still working even though h is equal p as shown in the last row of Table 3.2 where $h = 100$ and $p = 100$. This clearly illustrates that MVV algorithm is more flexible to be employed on singular or non-singular covariance matrices as Σ does not need to be positive definite.

Table 3.1: The number of operations

p	**Number of operations**	
	VV	**CD**
10	128	826
25	698	11376
50	2648	87126
75	5848	289751
100	10298	681751
150	22948	2283876
200	40598	5393501
250	63248	10510626
300	90898	18135251

Table 3.2: Average number of iterations to compute robust MSD

p	**MVV**	**Fast MCD**
2	5.14	5.22
3	5.02	5.21
4	5.23	5.43
5	4.91	5.20
10	4.31	4.64
15	3.74	4.14
20	3.35	3.87
25	3.46	3.83
30	3.04	3.45
40	2.18	2.70
50	1.59	2.00
75	1.46	2.00
100	2.95	-

3.3 Robust Hotelling T^2 Control Charts Based On MVV Estimators ($\boldsymbol{T^2_{MVV}}$)

As stated before in Chapter 1, the construction of control chart is divided into two phases. In Phase 1, a historical data set is analyzed to determine whether the process is in-control by establishing the initial control limits and estimating the in-control parameters of the process. While in Phase II, the control chart is used with future observations for detecting possible departures from parameters estimated in Phase 1. This study introduced MVV estimators in T^2 chart using the second approach (refer to Section 2.6) i.e. construct robust control chart for Phase II data based on the MVV estimates of location and scatter parameters from Phase I. For such approach, different observations will be used in the two phases.

Suppose that $\boldsymbol{x_i} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ is the *p*-variate random sample of *n* observations of preliminary data set in Phase I. Assume that $\boldsymbol{x_i}$ are independent and follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. If $\boldsymbol{\mu}$ and Σ are unknown then we need to estimate them using an in-control data set. The process of identifying the in-control data set from $\boldsymbol{x_i}$ is referred to as Phase I operation. From the preliminary data set, compute $\boldsymbol{\bar{x}}$ and $\boldsymbol{S}$. Using this estimates, compute $T^2(i)$ for *i* = 1,..., *n*, using Equation (1.3).To get in-control data set, first we need to identify outliers by using UCL based on Beta distribution as follows,

$$UCL_1 \sim \left[\frac{(n-1)^2}{n}\right] B_{\left(\frac{p}{2}, \frac{n-p-1}{2}\right)} \tag{3.3}$$

Observations with $T^2(i) > UCL_1$ are considered as outliers and should be removed. The sample with the outliers removed (n_c) is then used to compute the new estimates, $\boldsymbol{\overline{x}_N}$ and $\boldsymbol{S_N}$. Using these estimates, compute $T^2(g)$ statistic for Phase II observation, where $\boldsymbol{x_g} \notin \boldsymbol{x_i}$, such that

$$T^2(g) = (\boldsymbol{x_g} - \boldsymbol{\overline{x}_N})\boldsymbol{S_N}^{-1}(\boldsymbol{x_g} - \boldsymbol{\overline{x}_N})^t \tag{3.4}$$

By using the suggested α, p and n_c values, compute the upper control limit using F distribution such that

$$UCL \sim \left[\frac{p(n_c+1)(n_c-1)}{n_c(n_c-p)}\right] F_{(p,n_c-p)} \tag{3.5}$$

However, this standard approach is only effective in eliminating extreme outliers in small sample sizes, but it fails to detect moderate outliers especially when the number of variables increases (Vargas, 2003; Williamset al., 2006; Jensen et al., 2007; Chenouri et al., 2009). To alleviate the problem, we proposed using MVV estimator in Phase I data, $\boldsymbol{x_i}$. Since the estimator is known to be free from outliers due to its estimation process, they could be readily used as in-control estimators in Phase II. Let $\boldsymbol{x_g} = \{\boldsymbol{x_{n+1}}, \boldsymbol{x_{n+2}}, \ldots\}$ where $\boldsymbol{x_g} \notin \boldsymbol{x_i}$ and $\boldsymbol{m_{MVV}}$ and $\boldsymbol{S_{MVV}}$ represent the MVV mean vector and covariance matrix estimators, respectively. We define a robust Hotelling's T^2 for Phase II data, $\boldsymbol{x_g}$, based on these MVV estimates as

$$T_{MVV}^2(g) = (\boldsymbol{x_g} - \boldsymbol{m_{MVV}})\boldsymbol{S_{MVV}}^{-1}(\boldsymbol{x_g} - \boldsymbol{m_{MVV}})^t \tag{3.6}$$

3.3.1 Estimation of Control Limits

The application of robust estimators in place of the mean and covariance structure in T^2chart in Equation 3.4 will cause the distributional properties of the traditional T^2 (Equation 3.5) to change (William et al., 2006). To demonstrate the performance of $T_{MVV}^2(g)$ in Equation 3.6, we need a better understanding about its distribution in order to obtain appropriate control limits i.e. UCL. Since the distribution of T_{MVV}^2 is unknown, we apply Monte Carlo simulation method to estimate the quantiles of the $T_{MVV}^2(g)$, for several combinations of sample sizes (n) and dimensions (p) discussed in Section 3.4.1. The 95% quantile of $T_{MVV}^2(g)$ for the chosen sample size n and dimension p in Phase I is estimated by generating K = 5000 samples of size n from a standard multivariate normal distribution $MVN_p(0, I_p)$. For each data set of size n, we compute the MVV mean vector ($\boldsymbol{m_{MVV}}(k)$) and covariance matrix ($\boldsymbol{S_{MVV}}(k)$) such that $k = 1, \dots, K$. In addition, for each data set, we randomly generate a new observation $x_{g,k}$ treated as a Phase II observation from $MVN_p(0, \boldsymbol{I_p})$ and calculate the corresponding $T_{MVV}^2(g, k)$ values as given by Equation 3.6. The empirical distribution function of $T_{MVV}^2(g)$ is based on the simulated values

$$T_{MVV}^2(g, 1), T_{MVV}^2(g, 2), \dots, T_{MVV}^2(g, K) \tag{3.7}$$

We sort $T^2_{MVV}(g,k)$ values in ascending order, and the UCL for the control chart is the 95% quantile of the 5000 statistics.

3.3.2 Implementation Procedures

A step by step approach for the construction of a T^2_{MVV} control chart is given as follows;

Phase I

1. Decide on the sample size *n*, number of dimensions *p* and the overall false alarm probability of α.
2. Simulate or collect the Phase I data $\boldsymbol{x_i} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$.
3. Use $\boldsymbol{x_i}$ to compute the MVV estimates, $\boldsymbol{m_{MVV}}$ and $\boldsymbol{S_{MVV}}$.
4. Based on the chosen α, *n* and *p* values, compute the control limit using quantile based on the simulated values as Equation 3.7.

Phase II

5. Compute T^2_{MVV} for each of the Phase II data (new observation) as per Equation 3.6 and plot it on a control chart with the estimated limit in Phase I (step 4).
6. Interpret and evaluate the performance of this chart by identifying the out-of-control points or patterns.

3.4 Performance Evaluation

In order to analyze the performance of the T^2_{MVV} control chart when instability process is present, we conduct a simulation study that has been designed to encompass several different scenarios, which are assumed to influence the strength and weaknesses of the T^2_{MVV} control chart. These scenarios or conditions will be discussed in Section 3.4.2 later. The performance of the proposed chart was later compared with the popular existing robust method, the robust T^2 chart based on MCD, and also the traditional Hotelling T^2 control charts. For the traditional chart we employed two approaches; first approach denoted as T^2_0 is without cleaning the outliers as being adopted by Alfaro and Ortega (2009) and the second approach, which is known as the standard approach, cleans the outliers once (T^2_S) (refer Equations 3.4 and 3.5). The performance of all Hotelling T^2 charts were evaluated based on the probability of detection and the false alarm rate in the process behavior based on the phase II data using estimated mean vector and covariance matrix from different estimators in Phase I.

The programs and simulations were run using MATLAB 7.8.0 (R2009a).The algorithm of MVV was executed using the MATLAB 7.8.0 (R2009a), while Fast MCD algorithm using *mcdcov.m* in the LIBRA package under MATLAB 7.8.0 (R2009a).

3.4.1 Choice of Sample Size and Number of Quality Characteristics

Sample size determination for multivariate problems has always been somewhat subjective depending on the statistical tool being used. In general, it is expected that a large n produces better estimation results, since larger sample sizes increase the precision of the estimators (Chou, Mason & Young, 2001). Rousseeuw and van Zomeren (1990, p.649) stated that "any outlier method can get into trouble" if n/p is relatively small and, as a rule of thumb, they recommended applying robust multivariate methods only when $n/p > 5$. However, to determine how large the sample size should be taken depending on the number of quality characteristics involved in the monitoring process (Mason & Young, 2002). Correspondingly, if there are more quality characteristics that need to be monitored in a multivariate process, then there are more parameters to be estimated, hence, more number of samples needs to be taken.

This study focused on small (2 and 5), medium (10) and slightly high (15 and 20) number of quality characteristics (dimensions) with reasonable values of sample sizes. Based on most recent works such as Vargas (2003), Jensen et al. (2007), Chenouri et al. (2009) and Alfaro and Ortega (2009), the choices of values for *p* and *n* are in the range of values listed in Table 3.3. All these values were covered in this study.

Table 3.3: The values of n and p

$\boldsymbol{p}$	$\boldsymbol{n}$
2	10, 25, 50,100, 200, 500
5	30, 50, 100, 200, 500
10	50, 100, 200, 500
15	80, 100, 200, 500
20	100, 200, 300, 500

3.4.2 Types of Contamination and Process of Evaluation

A successful process monitoring in Phase II totally depends on the estimates of the parameters obtained from a stable HDS. However, the estimators are easily affected by outliers. Thus, the data in Phase I and II were contaminated with certain values of shift in the mean vector ($\boldsymbol{\mu_1}$) and also certain proportions of outliers (ε).We simulate 1000 datasets of various conditions created by manipulating the number of observations, dimensions and levels of contamination. To examine the effect of contamination on the charts' performance, we have considered a contaminated model by using a mixture of normal

$$(1-\varepsilon)N_p(\boldsymbol{\mu_0}, \Sigma_0) + \varepsilon N_p(\boldsymbol{\mu_1}, \Sigma_1) \tag{3.8}$$

where ε is the proportion of outliers, $\boldsymbol{\mu_0}$ and Σ_0 are the in-control parameters while $\boldsymbol{\mu_1}$ and Σ_1 are the out-of-control parameters. In this study we assume contamination with shift in the mean but no changes in covariance structure, therefore, the covariance matrix Σ_0 and Σ_1 in Equation (3.8) represent the identity matrix of p dimensions ($\boldsymbol{I_p}$). To check on these conditions, we consider ε to be 0, 0.1 or 0.2. While for the probability of detecting a change which depends on the shift in the mean vector, we set $\boldsymbol{\mu_1}$ to be a vector of size p with value of 0 (when there is no change), 3 or 5. Manipulation on the mean shifts and percentage of outliers generate 5 different types of contaminated distributions categorized as ideal, mildly contaminated, moderately contaminated and extremely contaminated as follows,

1) $N_p(0, \boldsymbol{I_p})$ -Ideal (no contamination)
2) $(0.9)N_p(0, \boldsymbol{I_p}) + (0.1)\mathrm{N_p}(3, \boldsymbol{I_p})$ - Mild contamination
3) $(0.8)N_p(0, \boldsymbol{I_p}) + (0.2)\mathrm{N_p}(3, \boldsymbol{I_p})$ -Moderate contamination
4) $(0.9)N_p(0, \boldsymbol{I_p}) + (0.1)\mathrm{N_p}(5, \boldsymbol{I_p})$ - Moderate contamination
5) $(0.8)N_p(0, \boldsymbol{I_p}) + (0.2)\mathrm{N_p}(5, \boldsymbol{I_p})$ - Extreme contamination

Each of these model was paired with different combinations of sample sizes, *n*, and number of dimensions, p (refer to Table 3.3) to create various conditions which are capable of highlighting the strengths and weaknesses of the charts (Alfaro & Ortega, 2009).

Next, in Phase II, we simulate data from multivariate normal distribution $MVN_p(\boldsymbol{\mu_1}, \boldsymbol{I_p})$, where $\boldsymbol{\mu_1}$ is the shift in the mean vector with values similarly assigned to Phase I (i.e. 0, 3, and 5). Each of these charts was tested on 5 types of contaminations with 23 combinations of n and p which totaled up to 115 conditions. For each condition, the false alarm rates and probability of detection were determined. Thus, for Phase II observations, we simulate 1000 new datasets of different sample sizes (n) and dimensions (p) in Table 3.3. To determine the false alarm rate and probability of detection, we randomly generate a Phase II observation with in-control and out-of-control parameters respectively, and calculate the proposed robust Hotelling T^2 statistics. The false alarm rate or probability of detection was estimated as the proportion of statistic values above the control limits of 1000 replications. The flowchart for the process of calculating the $\boldsymbol{T^2_{MVV}}$ is presented in Appendix A.

3.5 Consistency and Unbiasedness

The properties of MVV estimators will be discussed in detail in Chapter 5. This chapter also demonstrates the attempt to improve the MVV estimators in achieving consistency at normal model. Nonetheless, in practice we always deal with finite samples, therefore the issue of bias in a finite sample will exist and should also be considered. The advantage of having an unbiased estimator for a finite sample is that this estimator remains unbiased even though the sample size becomes larger (Pison et al., 2002). Due to the aforementioned issues, the following analysis seeks to improve the performance of MVV by making it unbiased for finite samples which in consequence will improve the performance of Hotelling T^2_{MVV} chart in general.

3.6 Reweighted Minimum Vector Variance (RMVV)

In Section 3.3, we introduced MVV estimator in the Hotelling T^2 chart for Phase II analysis, where these estimators were calculated at Phase I and then used directly in Phase II. Since this approach does not have to go through the process of outliers cleaning in Phase I, thus, higher statistical efficiency is vital because the highly robust estimators should not be unduly influenced by outliers in the Phase I data. Nevertheless, there is a conflict between the statistical efficiency and breakdown point where the efficiency of a robust estimator decreases when the breakdown point increases, especially when the number of dimension becomes higher (Rousseeuw & van Zomeren, 1990; Croux & Haesbroeck, 1999).

To check whether the conflict exists in MVV estimators, this study continued with the investigation on statistical efficiency of MVV estimators for different breakdown point. Two commonly chosen breakdown points are BP = 0.5 with $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ and BP = 0.25 with h = *(0.75)n*. To illustrate on how the efficiencies of MVV estimators vary with different breakdown points (BP) and dimensions (p) under normal model, we computed the asymptotic relative efficiency (ARE). When the conflict between efficiency and high breakdown value exist, we then proposed the reweighted version of MVV. Reweighted approach can maintain the breakdown point of the initial MVV estimators, while attaining a better efficiency. The detail about this analysis is discussed in Chapter 6.

3.7 Robust Hotelling T^2 Control Charts Based On RMVV Estimators (T^2_{RMVV})

Based on the results in Chapter 6, we proposed another robust control chart for Phase II data based on the RMVV location and scatter parameters estimated from Phase I. A detailed analysis is discussed in Chapter 7. The distribution of the Hotelling T^2 statistic based on RMVV (T^2_{RMVV}) differs from the T^2_{MVV}. We constructed an approximate distribution using ideas that were similar to the construction of the T^2_{MVV} distribution in Section 3.3.1. Then, the investigation on the performance of T^2_{RMVV} was conducted similar to the approach used in T^2_{MVV}.

CHAPTER FOUR
ROBUST HOTELLING T^2 CHART WITH MINIMUM VECTOR VARIANCE

4.1 Introduction

The organization of this chapter is determined by the types of data being analyzed. Two types of data namely the simulated and the real data are used for the investigation. The chapter starts with the presentation of the result for the simulated data analysis followed by the real data analysis.

This study integrates the MVV estimators in the Hotelling T^2 control chart for Phase II data using the same approach as Chenouri et al. (2009) and Alfaro and Ortega (2009) for monitoring the multivariate observations. Even though RMCD was observed to be better than MCD in controlling the false alarm rate (the discussion in Chapter 2, Section 2.6), in this chapter, comparisons are made based on the initial MCD since the algorithm for the proposed method follows the algorithm of the initial MCD. We want to compare the algorithm in its original state and diagnosing problems that might arise using the proposed method (MVV) in constructing Hotelling T^2 control chart. If there is a need to improve the method, this will be continued in the next chapter.

4.2 Simulated Data Analysis

The presentation of this subchapter is sectioned by the two measurements of the performance namely probability of outliers' detection and false alarm rates. The performance of the proposed chart is then compared with robust Hotelling T^2 chart using MCD (T^2_{MCD}) and the traditional Hotelling T^2 control charts. For the traditional chart we employed two approaches; first approach denoted as T^2_0 is without cleaning the outliers as being adopted by Alfaro and Ortega (2009) and the second approach, which is known as the standard approach, cleans the outliers once (T^2_S). Each of T^2_O, T^2_S , T^2_{MCD} and T^2_{MVV} charts was tested on 5 types of contaminations on 23 combinations of *n* and *p* which totaled up to 115 conditions. For each condition, the probability of detection and false alarm rates were determined. Each of these control charts is exposed to various scenarios which are known to highlight the strengths and weaknesses of the chart, such as number of dimensions (*p*), sample sizes (*n*), percentage of outliers (ε), and the mean shift (μ_1). This study only focused on independent case. The results for the false alarm rates and probability of outliers' detection are presented in tables and figures respectively. The programs and simulations were run using MATLAB 7.8.0 (R2009a). The algorithm of MVV was executed using the MATLAB 7.8.0 (R2009a), while Fast MCD algorithm using *mcdcov.m* in the LIBRA package under MATLAB 7.8.0 (R2009a).

4.2.1 Estimation of Control Limits

In this section, we present the control limit of the $T^2_{MVV}(g)$ control chart (Equation 3.8) by using simulated data with different combinations of sample sizes, *n*, and number of dimensions, *p*. The control limit of the $T^2_{MVV}(g)$ is then compared with robust Hotelling T^2 chart using MCD (T^2_{MCD}) and the traditional Hotelling T^2 control charts. We use quantile in estimating the distribution of $T^2_{MVV}(g)$ and $T^2_{MCD}(g)$ obtained via Monte Carlo method. This study focuses on multiple dimensions with reasonable values of sample size *n*. In order to estimate the 95% quantile of $T^2_{MVV}(g)$ and $T^2_{MCD}(g)$ for a given Phase I of sample size *n* and dimension *p*, we generate *K* = 5000 samples of size n from a standard multivariate normal distribution, $MVN_p(0, \boldsymbol{I_p})$. For each data set of size *n*, we compute the MVV and MCD mean vector and the modified covariance matrix estimates. In addition, for each data set, we randomly generate a new observation $\boldsymbol{x_{g,k}}$ treated as a Phase II observation from $MVN_p(0, \boldsymbol{I_p})$ and calculate the corresponding $T^2_{MVV}(g, k)$ and $T^2_{MCD}(g, k)$ values. We sort $T^2_{MVV}(g, k)$ and $T^2_{MCD}(g, k)$ values in ascending order, and the UCL is the 95% quantile of the 5000 statistics. The control limits for T^2_{MVV} and T^2_{MCD} calculated using the Monte Carlo method and the control limits for T^2_O and T^2_S based on Beta and *F* distributions respectively are presented in Table 4.1. From this table, we see that the UCL values for T^2_{MVV} are large as compared to traditional methods (T^2_O and T^2_S) and also T^2_{MCD}.

Table 4.1: Control limits

p	n	T_O^2	T_S^2	T_{MCD}^2	T_{MVV}^2
2	10	11.0360	13.1050	30.6250	76.0122
	25	7.4275	7.7676	12.1798	32.4008
	50	6.6447	6.7431	8.2762	28.1107
	100	6.3039	6.3444	7.4463	24.6037
	200	6.1443	6.1598	6.4127	21.7088
	500	6.0518	6.0566	6.1558	20.4264
5	30	15.6006	16.9160	27.6404	41.9567
	50	13.4506	13.9253	18.3456	33.5214
	100	12.1579	12.3055	14.7736	28.4822
	200	11.5915	11.6454	12.5765	25.6204
	500	11.2738	11.2904	11.4941	22.5296
10	50	25.9552	27.7721	39.8024	62.9323
	100	21.5264	21.9969	26.0646	43.1889
	200	19.7975	19.9570	20.9145	34.8509
	500	18.8777	18.9275	19.5618	31.1418
15	80	33.6517	35.1547	42.7942	69.0937
	100	31.5083	32.5354	37.4367	61.0544
	200	27.9034	28.2452	29.4809	45.9981
	500	26.0882	26.1820	26.6016	39.7181
20	100	42.5747	44.6890	52.7273	83.5238
	200	36.2033	36.8213	38.8163	57.2303
	300	34.4609	34.7659	36.0595	52.3438
	500	33.1766	33.3434	33.6574	47.7808

4.2.2 Probability of Detection of Outliers

The graphs illustrating the performance of the four charts in terms of probability of detection are exhibited in Figure 4.1 to 4.5. Each figure represents different dimension (p). For each condition, the performance of the control chart is regarded as better in detecting changes when the value of the probability is closer to 1. Under bivariate case (p = 2) as presented in Figure 4.1, initially T_S^2 showed better detection than other charts at mild and moderate contamination. However, the good performance of T_S^2 only sustain at n = 10, 25. Once the value of n and p increased, which can be clearly observed in Figure 4.2 – 4.5, the line representing T_{MVV}^2 is consistently at the highest location in the graphs with the probability value of approximately 1, and overlapping with T_{MCD}^2 line under most of the conditions. There are instances when the T_{MCD}^2 line started with lower values creating gaps between the two lines but merged later on when the n values increased. This situation occurs when the sample size is small with 20% outliers and mean shift 3.

Overall, the T_{MVV}^2 and T_{MCD}^2 control charts consistently achieved high probability in detecting outliers. One can observe that, the lines representing T_O^2 and T_S^2 charts are always at the lowest and second lowest respectively, creating a very wide gap between the other two lines (T_{MVV}^2 and T_{MCD}^2). This pattern repeats even within the same dimension for p > 5. Result on the T_S^2 chart reveals that the chart perform so well when the number of outliers is small (small p and low percentage of contamination), but underperform when the number of outliers gets larger (large *p* and high percentage of contamination). This weakness can be mitigated by the use of robust Hotelling T^2 chart.

4.2.3 False Alarm Rates

The performance of a chart is not only judged by its ability in detecting outliers, but also in controlling the level of false alarm rate. False alarm rate is the probability of out-of-control signal when a process is in control or also known as probability of Type I error. The value becomes large if the process is unstable due to increase in variability. Inflated false alarm rate can lead to unnecessary process adjustments and loss of confidence in the control chart as a monitoring tool (Chang & Bai, 2004). Hence, a method which can control the false alarm rate to the desired level is necessary.

The control chart is considered to be in control of its false alarm if the empirical value is close to the nominal value, α. For the purpose of comparison and checking on the level of robustness, we consider using the Bradley's liberal criterion of robustness as a reference. Bradley (1978) specified three criteria for robustness namely stringent, moderate and liberal which are respectively defined as $\alpha \pm 0.1\alpha$, $\alpha \pm 0.2\alpha$, and $\alpha \pm 0.5\alpha$. A statistic is considered robust if its empirical Type I error (false alarm) rates lie in one of the ranges. Nevertheless, the closer the value to α, the more robust is the statistic or in other words the procedure considered robust and has better ability in controlling false alarm. However, Guo and Luh (2000) considered a test to be robust if its empirical Type I error rate does not exceed 0.075 for the 5% level significance used. This implies that in the context of robustness, it is acceptable for a test to be conservative (< 0.025) than liberal (> 0.075). Taking into consideration Bradley's liberal criterion for robustness (1978) and Guo and Luh's (2000) justification, and also keeping in mind that inflated false alarm rate could mislead the ability of a chart as a monitoring tool, we proposed an interval between 0.025 and 0.055 to determine a chart's ability in controlling its false alarm rate. Thus in the tables, the values that are closest to the nominal value and within the 0.025 and 0.055 are highlighted.

Table 4.2 to 4.6 which recorded the false alarm rates for each condition are arranged based on the ascending number of dimensions (variables) namely $p = 2, 5, 10$,15 and 20 with α = 0.05. The first column in each table displays the number of sample sizes, followed by the percentage of outliers and non centrality values respectively in the second and third column. The last four columns record the false alarm rates of the control charts investigated in this study; namely T_O^2, T_S^2, T_{MCD}^2 and T_{MVV}^2.

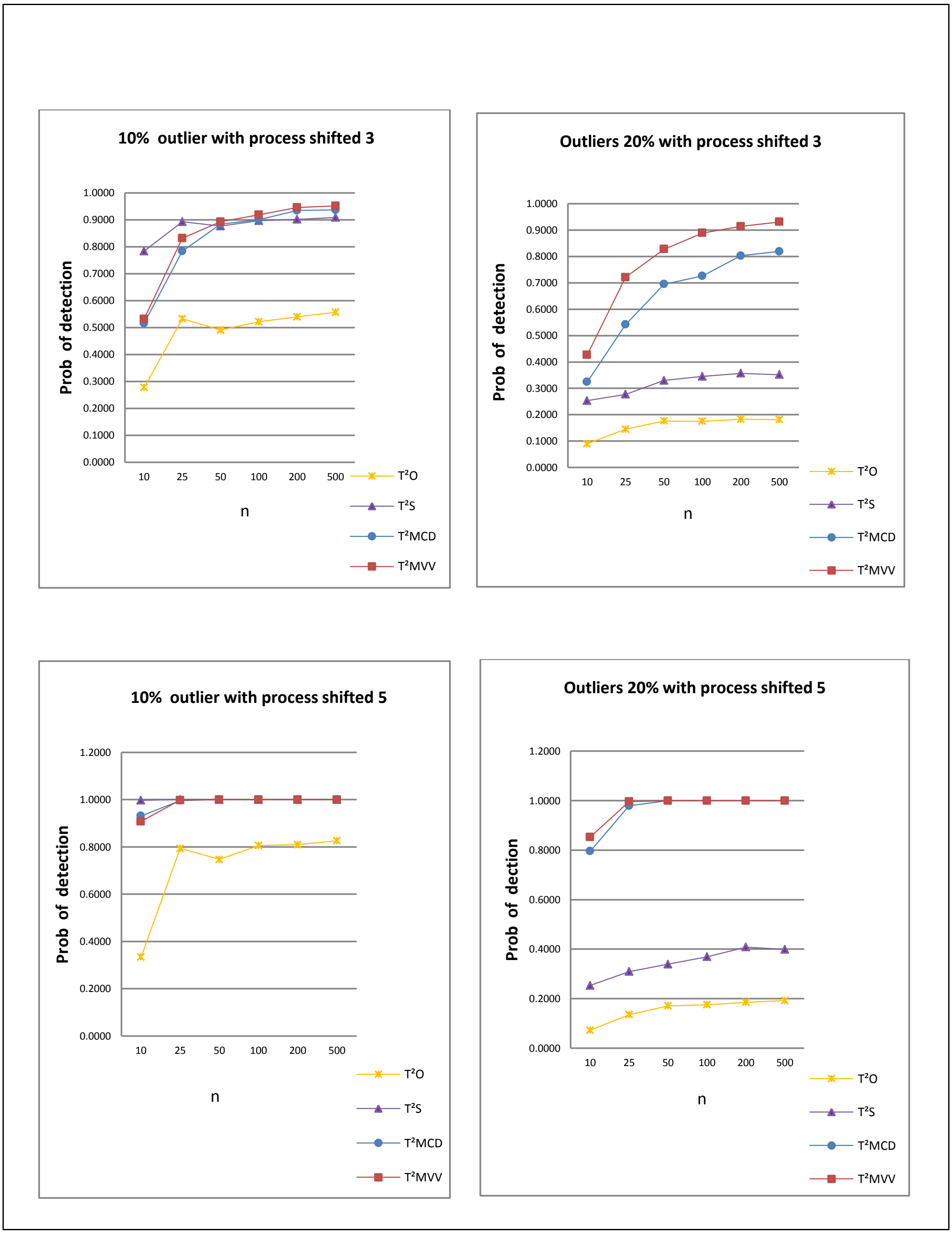


Figure 4.1: Probability of signal when p=2.

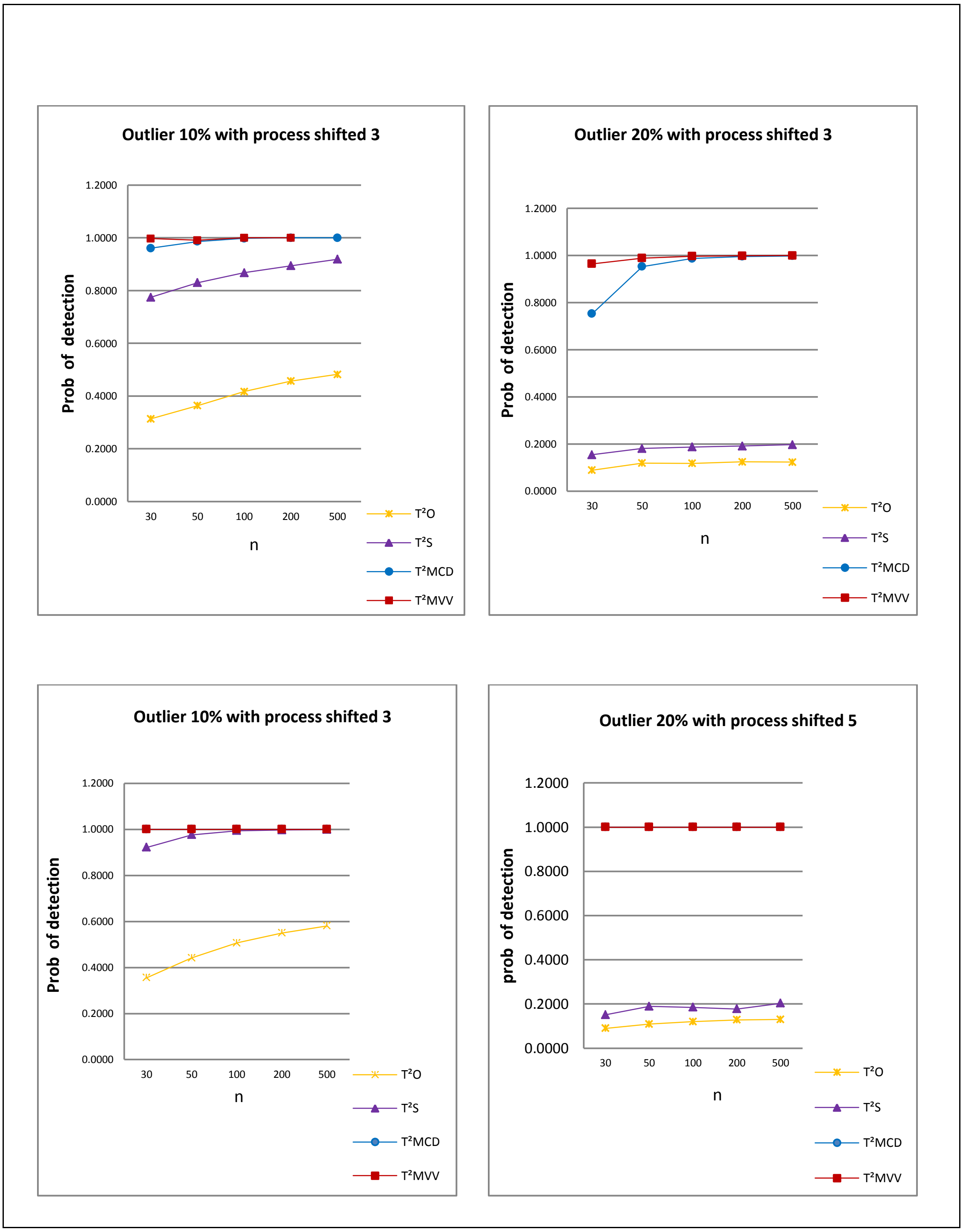


Figure 4.2: Probability of signal when p=5.

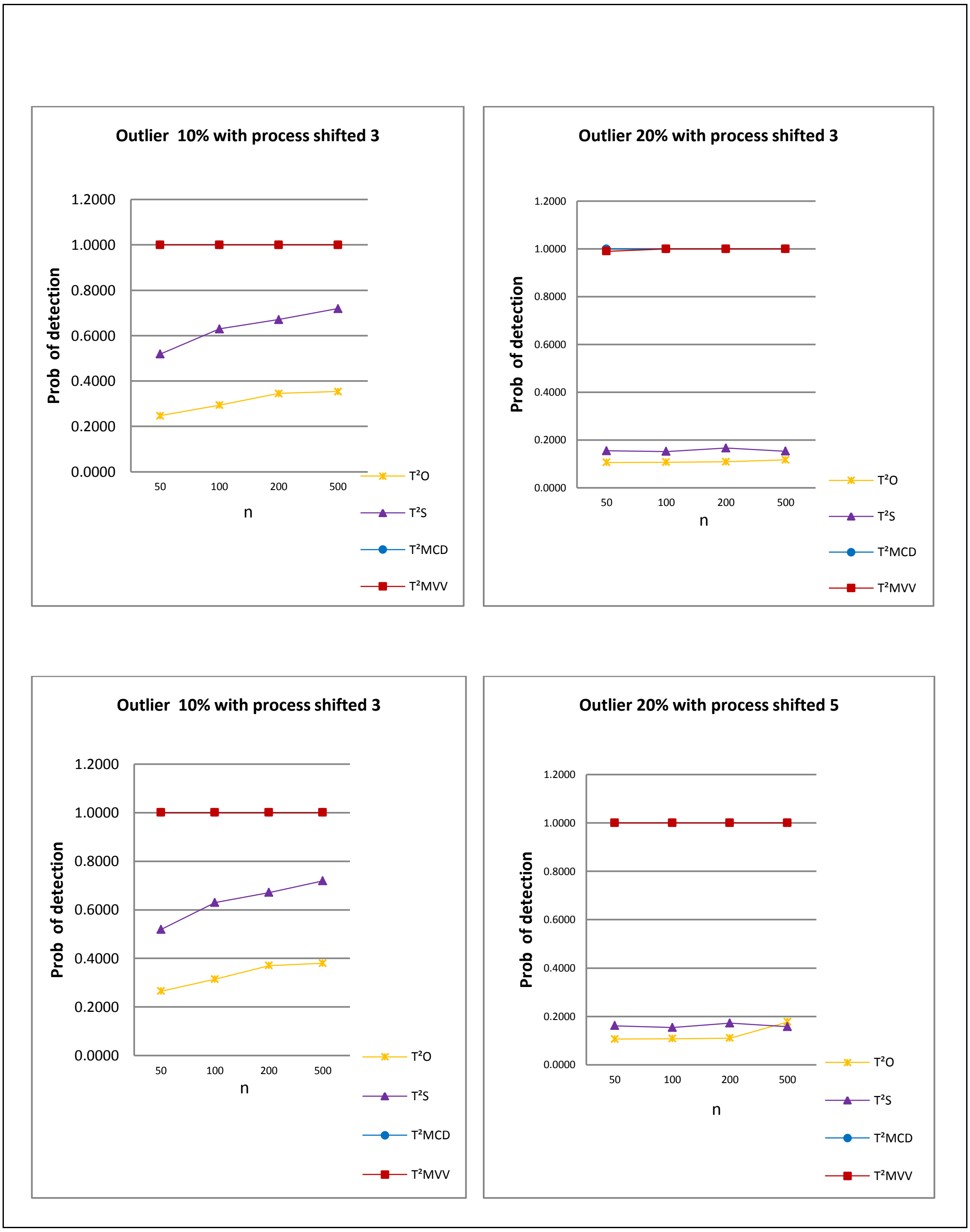


Figure 4.3: Probability of signal when p=10.

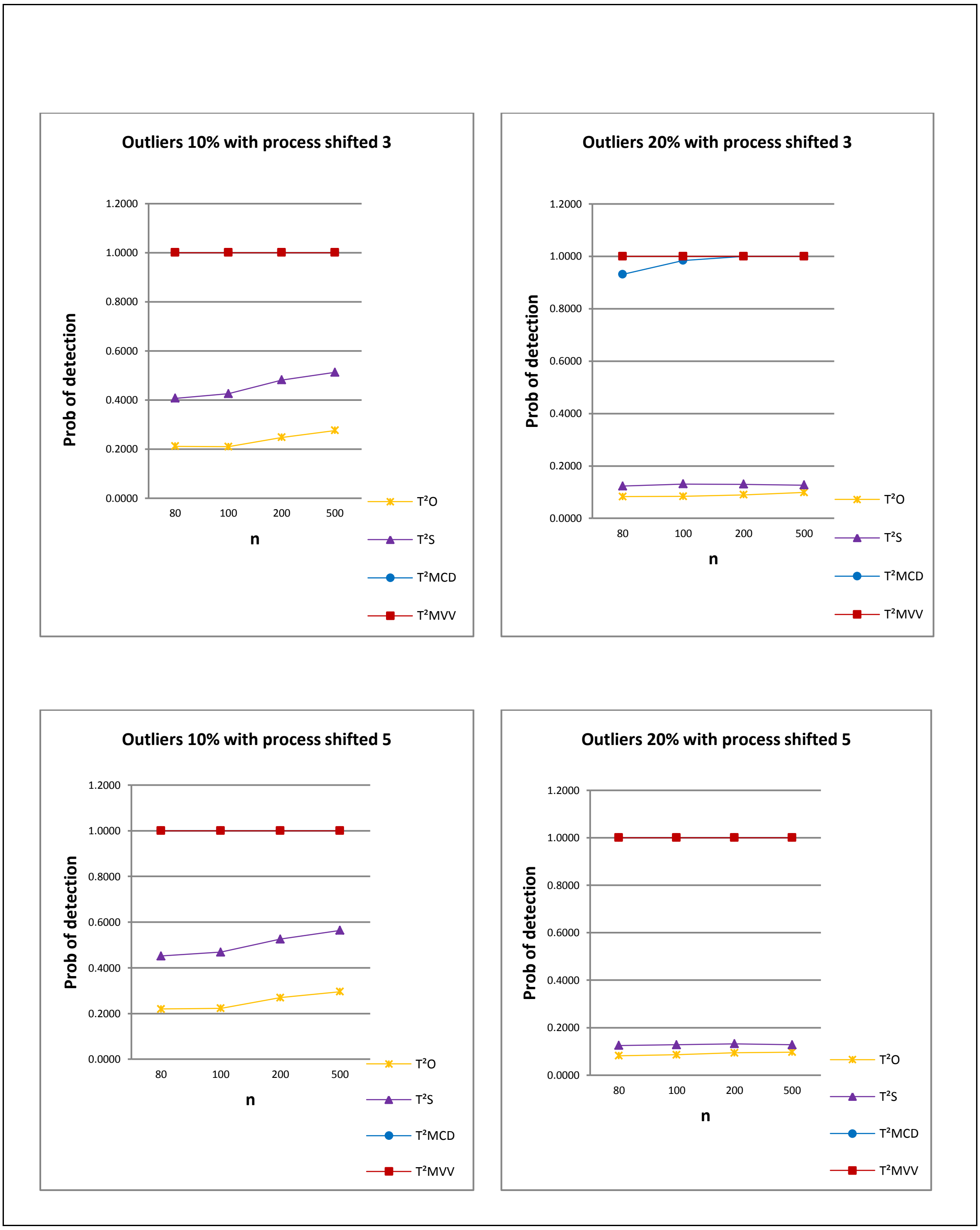


Figure 4.4: Probability of signal when p=15.

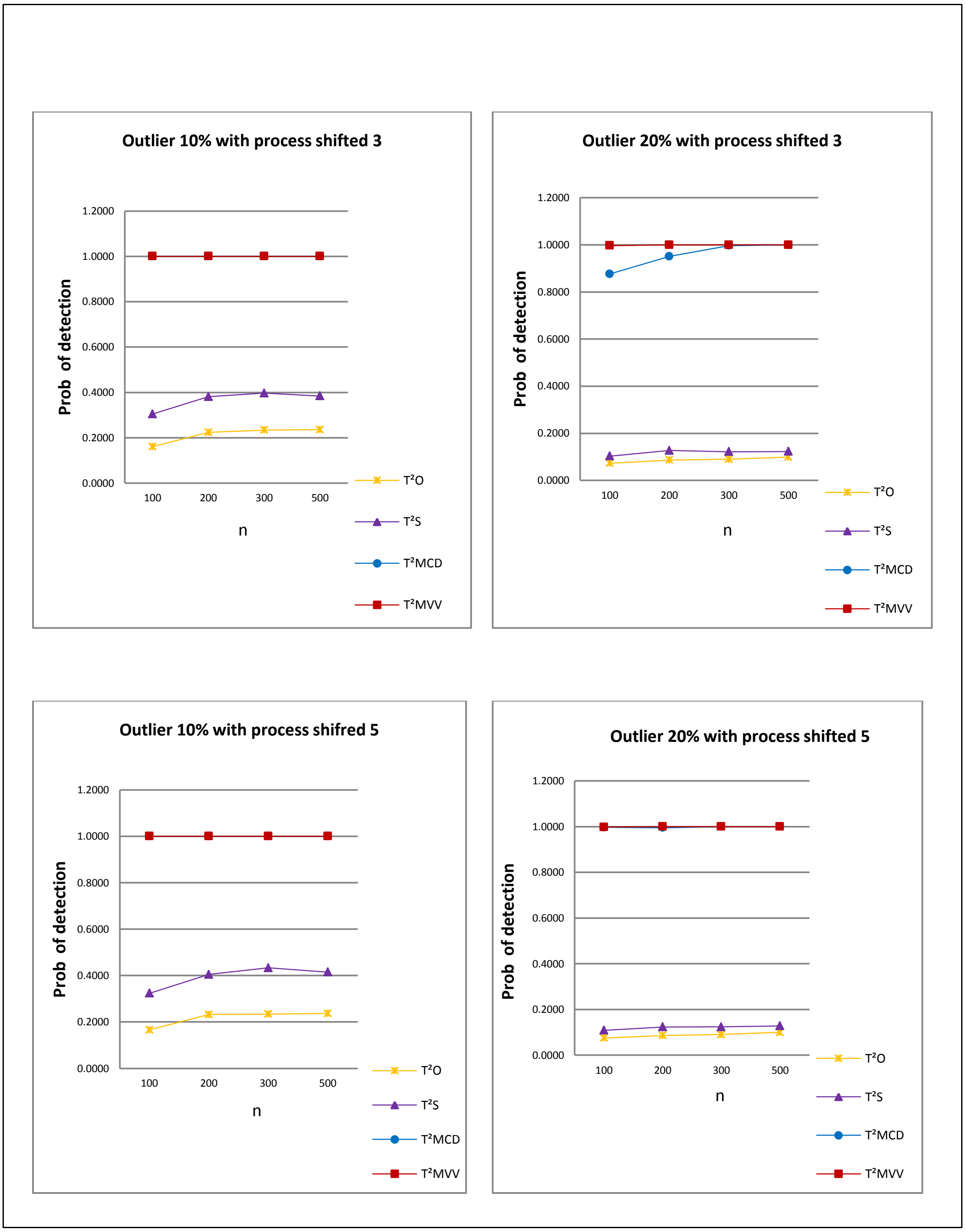


Figure 4.5: Probability of signal when p=20.

For the bivariate (p = 2) case presented in Table 4.2, the overall results on false alarm rates show that T_S^2 outperforms the other control charts, followed by T_{MVV}^2. Even though the results for T_S^2 under most conditions are well controlled, however under ideal condition (no contamination) the chart failed to control the false alarm, causing the rate to inflate to 0.1000. The T_{MCD}^2 and T_O^2 control charts are badly affected when the sample size is very small, which are verified by the rates of false alarm which are far below the nominal value except for ideal condition. When the percentage of outliers increased to 20%, we observed that the rates for T_{MVV}^2, T_{MCD}^2 and T_O^2 charts dwindle as the sample size increased, but the T_S^2 chart is still in control of its false alarm rate. The performance of the robust T_{MVV}^2 chart is much better than the T_{MCD}^2. The T_{MVV}^2 chart performs well in controlling false alarm rates except when the percentage of outliers is large.

When the dimension increased to p = 5, T_S^2 still show the best performance in controlling false alarm rate compared to other charts (refer to Table 4.3). Nevertheless, the rates for T_S^2 chart under ideal condition are still high (very far above the nominal value, $\alpha = 0.05$). We also notice improvements in the robust T_{MVV}^2 charts especially when the percentage of outliers is large, but the chart's performance is still below T_S^2 and T_O^2. In contrast, the false alarm rates for T_{MCD}^2 chart worsen with values as small as 0.0020.

Table 4.4 displays the false alarm rates for the case of p = 10. T_S^2 maintains to be the best performer in controlling false alarm rates but the chart seems to be deviating from the nominal value ($\alpha = 0.05$) when n = 50. There are also noticeable improvements in most of the conditions for T_O^2, T_{MVV}^2 and T_{MCD}^2 charts. Even though T_{MVV}^2 chart is not robust under the influence of extreme contamination, the false alarm rates for the chart are just slightly below the 0.025 level (no less than 0.022). However, the rates for T_{MCD}^2 chart are still far below the nominal level despite showing some improvement in the performance.

Under the case of p = 15, as can be clearly observed in Table 4.5, all the charts show better results than the previous case. Great improvement could be detected in T_S^2 chart under ideal condition and T_{MVV}^2 under extreme contamination, but T_{MCD}^2 chart is still unable to control its false alarm rates under the latter condition. As we scrutinized the false alarm rates for p = 20 in Table 4.6, we discover sporadic improvements under different conditions. There is no obvious improvement in the pattern could be observed. However, we can clearly observe that T_{MCD}^2 chart perform badly in controlling false alarm rate in all conditions.

Table 4.2: False alarm rates for dimension, p=2

Sample Size (n)	**% outliers(ε)**	**Mean shift (μ₁)**	**Control Charts**			
			$\boldsymbol{T_O^2}$	$\boldsymbol{T_S^2}$	$\boldsymbol{T_{MVV}^2}$	$\boldsymbol{T_{MCD}^2}$
10	0	0	0.0530	0.1000	0.0520	0.0520
	10%	3	0.0170	0.0650	0.0450	0.0290
		5	0.0160	0.0630	0.0450	0.0250
	20%	3	0.0180	0.0540	0.0330	0.0210
		5	0.0180	0.0480	0.0330	0.0110
25	0	0	0.0590	0.0980	0.0530	0.0480
	10%	3	0.0290	0.0600	0.0390	0.0280
		5	0.0230	0.0670	0.0390	0.0290
	20%	3	0.0280	0.0470	0.0190	0.0090
		5	0.0240	0.0390	0.0190	0.0050
50	0	0	0.0560	0.0920	0.0540	0.0580
	10%	3	0.0200	0.0480	0.0350	0.0230
		5	0.0160	0.0490	0.0340	0.0230
	20%	3	0.0210	0.0370	0.0180	0.0080
		5	0.0160	0.0340	0.0170	0.0060
100	0	0	0.0550	0.0930	0.0490	0.0460
	10%	3	0.0210	0.0470	0.0300	0.0200
		5	0.0160	0.0490	0.0290	0.0200
	20%	3	0.0210	0.0350	0.0150	0.0050
		5	0.0160	0.0350	0.0150	0.0040
200	0	0	0.0580	0.0950	0.0690	0.0600
	10%	3	0.0210	0.0510	0.0490	0.0310
		5	0.0180	0.0470	0.0500	0.0310
	20%	3	0.0200	0.0410	0.0280	0.0050
		5	0.0180	0.0360	0.0280	0.0020
500	0	0	0.0500	0.0880	0.0630	0.0520
	10%	3	0.0190	0.0480	0.0490	0.0270
		5	0.0160	0.0390	0.0480	0.0260
	20%	3	0.0170	0.0370	0.0230	0.0040
		5	0.0160	0.0350	0.0230	0.0040
Total highlighted			1	16	11	2

Table 4.3: False alarm rates for dimension, p = 5

Sample Size (n)	**% outliers (ε)**	**Mean shift (μ₁)**	**Control Charts**			
			T_O^2	T_S^2	T_{MVV}^2	T_{MCD}^2
30	0	0	0.0460	0.0790	0.0500	0.0430
	10%	3	0.0280	0.0510	0.0300	0.0100
		5	0.0260	0.0530	0.0330	0.0100
	20%	3	0.0300	0.0620	0.0210	0.0050
		5	0.0320	0.0620	0.0200	0.0000
50	0	0	0.0530	0.0790	0.0490	0.0650
	10%	3	0.0270	0.0480	0.0350	0.0130
		5	0.0260	0.0460	0.0370	0.0130
	20%	3	0.0260	0.0560	0.0220	0.0040
		5	0.0250	0.0520	0.0230	0.0020
100	0	0	0.0540	0.0740	0.0380	0.0320
	10%	3	0.0290	0.0510	0.0300	0.0140
		5	0.0280	0.0420	0.0320	0.0140
	20%	3	0.0300	0.0520	0.0170	0.0020
		5	0.0290	0.0490	0.0190	0.0020
200	0	0	0.0430	0.0740	0.0390	0.0410
	10%	3	0.0250	0.0450	0.0350	0.0200
		5	0.0240	0.0420	0.0350	0.0200
	20%	3	0.0270	0.0460	0.0220	0.0010
		5	0.0270	0.0440	0.0220	0.0010
500	0	0	0.0390	0.0620	0.0430	0.0420
	10%	3	0.0200	0.0410	0.0360	0.0160
		5	0.0190	0.0350	0.0370	0.0170
	20%	3	0.0210	0.0430	0.0190	0.0030
		5	0.0200	0.0420	0.0190	0.0030
Total highlighted			5	16	4	0

Table 4.4: False alarm rates for dimension, p = 10

Sample Size (n)	**% outliers (ε)**	**Mean shift (μ_1)**	**Control Charts**			
			T^2_O	T^2_S	T^2_{MVV}	T^2_{MCD}
50	0	0	0.0570	0.0920	0.0520	0.0530
	10%	3	0.0410	0.0690	0.0370	0.0210
		5	0.0380	0.0670	0.0380	0.0210
	20%	3	0.0420	0.0720	0.0250	0.0080
		5	0.0410	0.0720	0.0220	0.0020
100	0	0	0.0550	0.0780	0.0450	0.0420
	10%	3	0.0330	0.0570	0.0390	0.0190
		5	0.0340	0.0550	0.0350	0.0200
	20%	3	0.0350	0.0560	0.0240	0.0030
		5	0.0340	0.0520	0.0230	0.0030
200	0	0	0.0430	0.0730	0.0520	0.0540
	10%	3	0.0330	0.0530	0.0390	0.0200
		5	0.0320	0.0520	0.0420	0.0200
	20%	3	0.0340	0.0500	0.0250	0.0020
		5	0.0340	0.0490	0.0240	0.0020
500	0	0	0.0510	0.0750	0.0540	0.0490
	10%	3	0.0330	0.0540	0.0390	0.0220
		5	0.0330	0.0520	0.0390	0.0230
	20%	3	0.0340	0.0580	0.0260	0.0040
		5	0.0340	0.0550	0.0230	0.0040
Total highlighted			6	9	5	1

Table 4.5: False alarm rates for dimension, p = 15

Sample Size (n)	**% outliers (ε)**	**Mean shift (μ_1)**	**Control Charts**			
			$\boldsymbol{T_O^2}$	$\boldsymbol{T_S^2}$	$\boldsymbol{T_{MVV}^2}$	$\boldsymbol{T_{MCD}^2}$
80	0	0	0.0450	0.0650	0.0560	0.0580
	10%	3	0.0330	0.0550	0.0470	0.0260
		5	0.0330	0.0560	0.0430	0.0230
	20%	3	0.0350	0.0560	0.0270	0.0060
		5	0.0360	0.0540	0.0320	0.0030
100	0	0	0.0430	0.0680	0.0520	0.0490
	10%	3	0.0330	0.0590	0.0450	0.0240
		5	0.0330	0.0610	0.0430	0.0240
	20%	3	0.0330	0.0540	0.0250	0.0030
		5	0.0330	0.0560	0.0220	0.0020
200	0	0	0.0440	0.0620	0.0470	0.0540
	10%	3	0.0290	0.0520	0.0420	0.0330
		5	0.0280	0.0520	0.0410	0.0310
	20%	3	0.0310	0.0560	0.0200	0.0040
		5	0.0300	0.0550	0.0240	0.0040
500	0	0	0.0530	0.0690	0.0470	0.0460
	10%	3	0.0370	0.0540	0.0390	0.0270
		5	0.0370	0.0530	0.0390	0.0260
	20%	3	0.0390	0.0530	0.0260	0.0060
		5	0.0380	0.0520	0.0290	0.0060
Total highlighted			6	8	6	1

Table 4.6: False alarm rates for dimension, p = 20

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts			
			T_O^2	T_S^2	T_{MVV}^2	T_{MCD}^2
100	0	0	0.0410	0.0720	0.0530	0.0500
	10%	3	0.0340	0.0550	0.0400	0.0210
		5	0.0330	0.0560	0.0420	0.0240
	20%	3	0.0310	0.0570	0.0300	0.0060
		5	0.0300	0.0570	0.0240	0.0000
200	0	0	0.0450	0.0660	0.0510	0.0490
	10%	3	0.0320	0.0510	0.0370	0.0220
		5	0.0340	0.0520	0.0380	0.0220
	20%	3	0.0360	0.0550	0.0310	0.0050
		5	0.0380	0.0550	0.0250	0.0020
300	0	0	0.0430	0.0680	0.0440	0.0390
	10%	3	0.0400	0.0480	0.0350	0.0210
		5	0.0390	0.0490	0.0340	0.0210
	20%	3	0.0410	0.0470	0.0240	0.0060
		5	0.0420	0.0470	0.0220	0.0050
500	0	0	0.0520	0.0690	0.0530	0.0560
	10%	3	0.0390	0.0560	0.0400	0.0280
		5	0.0380	0.0570	0.0400	0.0280
	20%	3	0.0390	0.0550	0.0350	0.0030
		5	0.0390	0.0570	0.0320	0.0040
Total highlighted			4	10	5	2

4.3 Real Data Analysis

The investigation of Hotelling T^2 issued from MVV is continued with the application on real data. Real data were furnished to us by Asian Composites Manufacturing Sdn. Bhd. (ACM). ACM is a Joint Venture Company based in Bukit Kayu Hitam, Kedah, Malaysia, owned by Boeing & Hexcel. This company is involved in the production of advanced composite panels for the aircraft industry. ACM produces flat and contoured primary (Aileron Skins, Spoilers and Spars) and secondary (Flat Panels, Leading Edges and MISC: Components) structure composite bond assemblies and sub-assemblies for aerospace industries. It was awarded with the AS 9100 rev C Certification (the highest level of qualification for aerospace manufacturers) after the British Standards Institution (BSI), a member of the International Aerospace Quality Group(IAOG). This certifies that ACM has met its standards and requirements for quality management systems.

Spoilers are vital devices in an airplane. Their function is to increase lifts when the airplane is flying. They are plates fitted to the top surface of the wings which can be extended upward into the airflow and spoil it. By doing so, the spoiler creates a carefully controlled stallover the portion of the wing behind it, greatly reducing the lift of that wing section (http://en.wikipedia.org/wiki/Spoiler_aeronautics). The products are used in civilian, defense, and space applications, which cannot compromise any mistakes, albeit a minor one. Thus, careful monitoring is required to ensure that no variation occur in the process. Any slight mistake could jeopardize a human life.

For the purpose of this research, ACM has provided us the real data on spoilers which consists of several features such as trim edge (X_1), trim edge spar (X_2), and drill hole (X_3). A sample of 47 products (n = 47) was furnished to us by ACM. Out of the total, 21 products were collected from 2009, while the rest were from 2010. Hence, we decided to use the 2009 products as phase I historical data, and considered the products from 2010 as future data in this study. The details of the historical and future data are displayed in Tables 4.7 and 4.9 respectively. The products consist of 3 quality variables (dimensions) namely trim edge, trim edge spar, and drill hole. These variables were used to compare the three methods used to construct control charts. Estimates for the location vector ($\bar{\boldsymbol{x}}$) and scatter matrix ($\boldsymbol{S}$) are presented in Table 4.8. The calculation of upper control limit (UCL) based on this estimates are presented in the last column of the table. The values of the T^2 statistics based on the above estimators appear in the last four columns of Table 4.9. The graphical presentation of the corresponding control charts are put on view in Figure 4.6.

When comparing the values of the T^2 statistics in Table 4.9 with the corresponding control limits in Table 4.8, we observe that the three statistics T^2_{MVV}, T^2_{MCD} and T^2_S signal observations 20, 22 and 25 as out-of-control but T^2_O only signals 20 and 25 as out-of-control observations and fails to signal observation 22. The result for T_O^2 is as expected since the analysis on the probability of detection using simulated data showed that T_O^2 was not as effective as the other charts in detecting outliers. Chart (a), (b), (c) and (d) in Figures 4.6 represent the control chart for T_O^2, T_S^2,T_{MCD}^2 and T_{MVV}^2 respectively. Even though the performance of T_S^2 chart in this example is on par with the proposed T_{MVV}^2 chart and also T_{MCD}^2 chart, but the outcome could be due to the small number of quality characteristics (dimension) of the product. As revealed in the simulation study, T_S^2 performed well in detecting outliers under low dimension (not more than 5) only, but underperformed when the dimension increased to above 5.

Table 4.7: Historical data set (Phase I data)

Product No.	Trim edge (x1)	Trim edge spar (x2)	Drill hole (x3)
1	-0. 0011	0.0003	0.0128
2	0.0011	0.0021	0.0246
3	0.0252	0.0308	0.0378
4	-0. 0017	0.0109	0.0177
5	-0. 0005	-0. 0010	0.0106
6	0.0016	-0.0059	0.0128
7	0.0004	0.0001	0.0062
8	0.0078	0.0003	0.0159
9	0.0076	0.0089	0.0097
10	0.0020	0.0005	0.0071
11	0.0108	0.0011	0.0092
12	0.0039	0.0034	0.0425
13	0.0060	-0.0033	0.0160
14	0.0066	0.0100	0.0056
15	0.0045	-0.0067	0.0147
16	0.0110	-0.0207	0.0337
17	0.0047	0.0059	0.0065
18	0.0077	0.0003	0.0191
19	0.0015	0.0123	0.0124
20	0.0011	0.0038	0.0104
21	0.0056	0.0065	0.0063

Table 4.8: Estimates of location vector, covariance matrix and UCL.

Types of Control Chart	Location Vector ($\bar{x}$)	Scatter Matrix (S)	Upper Control Limit (UCL)
T_O^2	$[0.00504 \quad 0.00284 \quad 0.01579]$	$\begin{bmatrix} 0.00004 & 0.00002 & 0.00003 \\ 0.00002 & 0.00009 & 0.00001 \\ 0.00003 & 0.00001 & 0.00011 \end{bmatrix}$	11.035
T_S^2	$[0.00365 \quad 0.00256 \quad 0.01209]$	$\begin{bmatrix} 0.00001 & 0.00000 & 0.00000 \\ 0.00000 & 0.00003 & -0.00001 \\ 0.00000 & -0.00001 & 0.00003 \end{bmatrix}$	11.798
T_{MCD}^2	$[0.00414 \quad 0.00207 \quad 0.01096]$	$\begin{bmatrix} 0.00002 & 0.00000 & 0.00000 \\ 0.00002 & 0.00009 & -0.00002 \\ 0.00000 & -0.00002 & 0.00003 \end{bmatrix}$	21.946
T_{MVV}^2	$0.00336 \quad 0.00354 \quad 0.00913$	$\begin{bmatrix} 0.00001 & 0.00001 & 0.00000 \\ 0.00001 & 0.00003 & 0.00000 \\ 0.00000 & 0.00000 & 0.00001 \end{bmatrix}$	41.298

Table 4.9: The Hotelling T^2 values for the future data (Phase II)

Product No.	**x1**	**x2**	**x3**	T^2_O	T^2_S	T^2_{MCD}	T^2_{MVV}
1	0.0041	0.0087	0.0129	0.5582	1.4242	1.7659	4.3908
2	0.0047	0.0109	0.0124	0.9003	2.5492	2.4694	5.1695
3	0.0031	0.0057	0.0096	0.4992	0.4936	0.3437	0.2992
4	0.0035	-0.0020	0.0101	0.5463	1.0157	0.5456	1.5064
5	0.0040	-0.0028	0.0125	0.4592	0.9588	0.4580	3.7869
6	0.0031	0.0008	0.0061	0.9013	1.7480	1.2527	2.2421
7	-0.0019	0.0101	0.0112	3.0933	4.1372	4.4404	6.5361
8	0.0009	0.0039	0.0082	0.8061	1.2884	0.6837	1.0556
9	-0.0052	0.0090	0.0203	7.3602	9.6843	14.9766	26.0499
10	-0.0008	0.0110	0.0184	3.6198	5.8035	9.7417	19.1760
11	-0.0021	0.0139	0.0170	5.3839	8.0897	11.8717	19.6313
12	-0.0017	0.0092	0.0061	2.7387	4.7949	2.9788	8.1388
13	-0.0010	0.0133	0.0138	3.8058	5.6890	7.4040	11.3895
14	-0.0030	0.0002	0.0053	2.0548	6.3468	3.3086	9.1498
15	0.0016	0.0134	0.0151	2.5073	5.0227	6.8054	12.3881
16	0.0027	0.0086	0.0070	1.1976	1.8980	1.0679	2.0563
17	0.0004	0.0086	0.0087	1.5798	2.2630	1.7597	2.8765
18	-0.0036	0.0136	0.0129	5.7910	7.9657	9.2817	13.9293
19	-0.0028	0.0003	0.0078	1.8304	4.7003	2.4178	4.8791
20	0.0120	0.0123	0.0768	**38.1397**	**190.2969**	**214.9233**	**894.5184**
21	-0.0015	0.0004	0.0115	1.2651	2.3301	1.5486	2.0641
22	0.0009	0.0232	0.0202	8.4181	**19.7720**	**24.6552**	**45.2462**
23	-0.0035	0.0088	0.0107	3.7588	5.1645	4.8793	7.5328
24	0.0016	0.0061	0.0066	1.0602	1.7564	0.9320	2.23575
25	-0.0228	-0.0466	0.0231	**42.8447**	**134.6222**	**68.6307**	**116.02933**
26	0.0037	-0.0038	0.0147	0.4832	1.3946	0.7796	7.32655

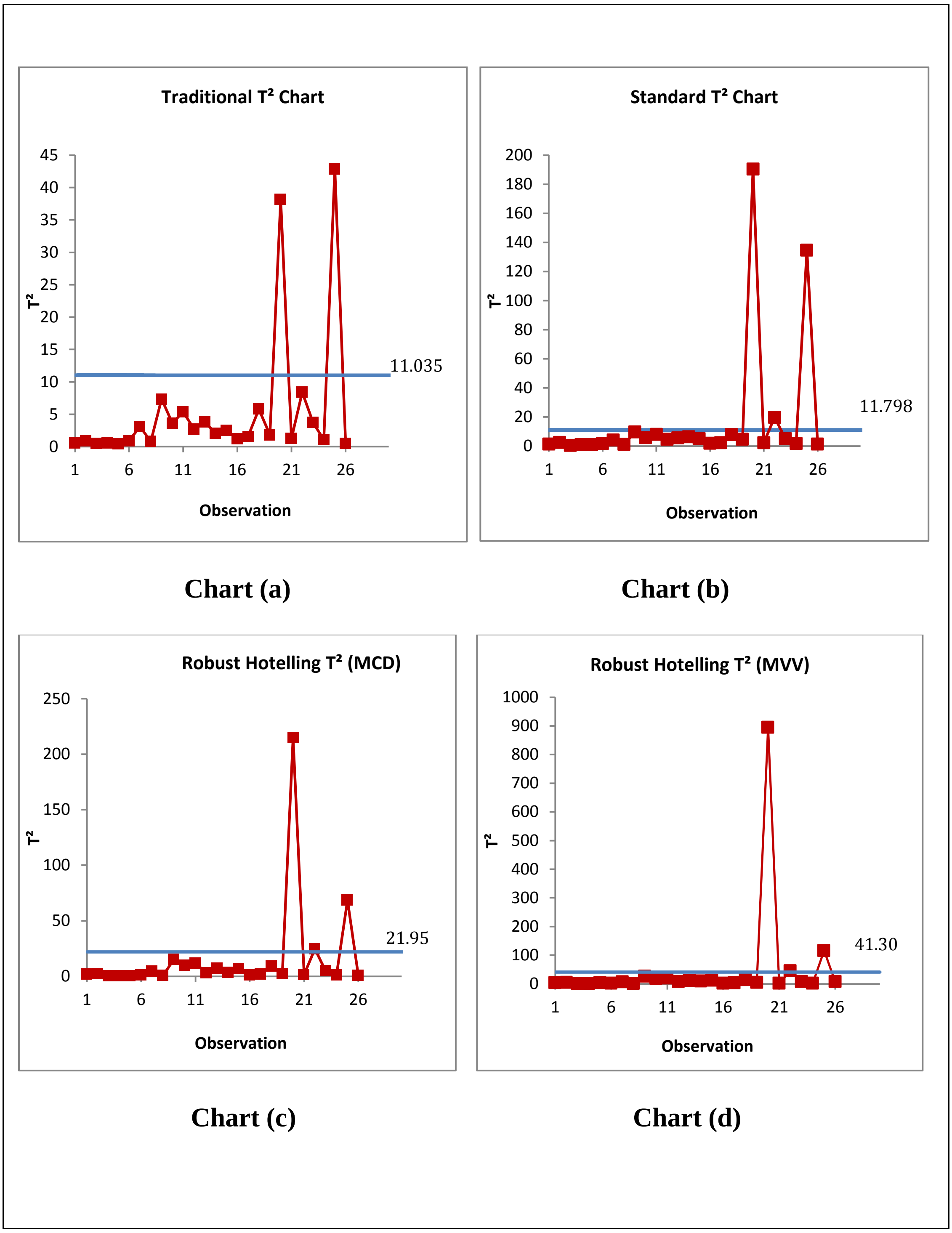


Chart (a)

Chart (b)

Chart (c)

Chart (d)

Figure 4.6: Hotelling T^2 control charts

4.4 Discussion

Hotelling T^2 chart is well accepted as a reliable method to monitor production; however, under conditions of non-normality, this chart is known to be underperformed. Alternative on the Hotelling T^2 statistic particularly on the location and scatter measures are recommended in order to produce a reliable chart regardless of the conditions. This study proposed an alternative to the the Hotelling T^2 chart by using a robust estimator known as minimum variance vector (MVV) for its location and scatter measures. MVV not only has all the properties of the well-known minimum covariance determinant (MCD) such as high breakdown point and affine equivariant, but also has better computational efficiency. The performance of our proposed robust Hotelling T^2 chart using MVV in terms of false alarm rate and probability of detection were compared with the robust Hotelling T^2 chart using MCD and the traditional Hotelling T^2 chart.

Investigation on the T_S^2 and T_{MCD}^2 by Alfaro and Ortega (2009) showed a conflicting result between the percentage of outliers detection and the overall false alarm rate such that when the probability of detection increased, the false alarm rates inflate away from the nominal value. However, our proposed chart, T_{MVV}^2 performed so well in terms of detecting outliers and also in controlling false alarm rates. Even though the traditional Hotelling T_S^2 chart performed so well in terms of controlling false alarm rates, but this chart fail to achieve good probability of detection especially when the number of quality characteristics is large. On contrary, the Hotelling T^2_{MCD} chart performs wonderfully in detecting outliers, however the chart fails terribly in controlling false alarm rates. With its good performance in terms of detecting outliers and controlling false alarm rates, plus the good properties of its statistics, Hotelling T^2_{MVV} chart is indeed a good alternative to the multivariate control chart.

When the investigation continued with the real industrial data, the results concurred with the earlier results obtained from simulation study which support both robust MVV and MCD estimators in detecting outliers. However, given the poor performance of MCD estimators in controlling false alarm rates, MVV estimators should be the one to choose for when searching for a good robust estimator. Nevertheless the values of Hotelling T^2 statistics and UCL estimates using MVV are large as shown in the real data analysis result (refer to Table 4.8 and 4.9). The result concurs with the simulated UCL values produced by T^2_{MVV} shown in Table 4.1. From the Table 4.1, we could observe the obvious differences between the simulated control limits for T^2_{MVV} with the usual control limits (T^2_O and T^2_S) and the simulated control limits for T^2_{MCD}.

MVV estimators have the same characteristics as the MCD estimators with respect to breakdown point and affine equivariance property, and their algorithms also display the same structures, but only differ in their objective function (MCD uses $|\Sigma|$ while MVV uses $Tr(\Sigma^2)$. Thus, by following the steps and procedures applied on MCD to

MVV, this study will attempt to improve the MVV estimators to achieve consistency at normal model. Nonetheless, in practice we always deal with finite samples, therefore the issue of bias in a finite sample will exist and should be considered. The advantage of having an unbiased estimator for a finite sample is that this estimator remains unbiased impartial even though the sample size becomes larger (Pison et al., 2002). Due to the aforementioned issues, the next analysis seeks to improve the performance of MVV by making it unbiased for finite samples and consistent at normal model, which in consequence will improve the UCL and the performance of Hotelling T^2 chart in general. Analysis for these improvements will be discussed in Chapter 5.

CHAPTER FIVE
THE EFFECT OF CONSISTENT MINIMUM VECTOR VARIANCE ESTIMATORS ON HOTELLING T^2 CONTROL LIMITS

5.1 Introduction

The simulated and the real data investigation on the robustness of the T^2_{MVV} had been done in Chapter 4. The result show that T^2_{MVV} performs well in terms of detecting outliers and also in controlling false alarm rates. However, in this chapter we developed a theory that leads to an improvement in the properties of MVV estimators when they are used in Hotelling T^2 chart. A main aspect of our viewpoint on this improvement is inspired by the MVV characteristic that is affine equivariant, where that measurement scale changes or other linear transformations do not alter the behaviour of analysis in Hotelling T^2 chart. Looking at the performance of T^2_{MVV} chart in Chapter 4 by comparing with the T^2_O, T^2_S and T^2_{MCD} charts, there have room of improvement for MVV estimators. This deficiency can be seen from the value of the T^2 statistics and estimated UCL for T^2_{MVV} is a very large to be consistent at normal model. Hence we are inspired to study the asymptotic properties of MVV estimators. The asymptotic properties of estimators are their properties as the number of observations in a sample becomes very large and tends to infinity. So we will pay attention to the concepts of consistency, unbiasedness and efficiency. Nevertheless, in this chapter we are focusing on the adjustments to the consistency and unbiasedness of the MVV estimators. We conducted simulation experiments to show the need for improvement. The efficiency of the MVV estimators will be investigated in Chapter 6.

The organization of the remaining part of this chapter is as follows. Section 5.2 and 5.3 discusses the adjustment done on the MVV scatter estimator to ensure that it is consistent and unbiased. Investigation through simulation experiment to illustrate the consistency and unbiased of MVV estimators at multivariate normal data is discussed in Section 5.4. In the following Section 5.5, we estimate the control limits of the improved Hotelling T^2_{MVV} charts by simulation. In section 5.6, we investigate on the improved MVV estimator through simulation study. A real data analysis from aircraft industry is presented to illustrate the applicability of the proposed charts in section 5.7. Finally, result and discussion are given in the last section.

5.2 Consistency Factor

The aim of Hotelling T^2 chart in Phase I is to estimate the in-control parameters of location, $\boldsymbol{\mu}$ and scatter, Σ. The usual estimators for these parameters are the normal maximum likelihood estimators (MLE). The estimation of these parameters is based on the data set $\boldsymbol{x}$={$\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$} from multivariate normal distributionwith density

$$f_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{\left(-\frac{(x-\mu)^t\Sigma^{-1}(x-\mu)}{2}\right)} \qquad (5.1)$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathbb{Z}^+$. However the distribution of (5.1) is only an approximation because a portion of the data may be contaminated by outliers (Croux & Rousseeuw, 1992). With the existence of outliers, MLE which are known to be sensitive to outliers will not be able to precisely estimate the parameters. To address this problem, we propose MVV estimators i.e. robust estimators with highest breakdown point (50%) proposed by Herwindianti (2006) to replace the MLE. We compute the MVV estimators in the Phase I data sets, with location and scatter estimators as defined in Equation 3.1 and 3.2 respectively. The MVV estimator that was described in Herwindiati et al. (2007) has a fixed integer h such that;

$$\left\lfloor \frac{n+p+1}{2} \right\rfloor \leq h < n$$

The preferred choice of h for outlier detection is its lower bound, which yields the breakdown value, BP $= \frac{n-2(p-1)}{2n}$. Let $\boldsymbol{m_{MVV}}$ and $\boldsymbol{S_{MVV}}$ be the mean and the scatter matrix calculated from the h observations out of $\boldsymbol{x_i}$, whose classical scatter matrix has the lowest vector variance resulting from h smallest MSD. The $\boldsymbol{S_{MVV}}$ is a scatter $p \times p$ matrix which is positive definite, symmetric (PDS) and affine equivariant (Herwindianti, 2006). Robust scatter estimator is typically calibrated to be consistent at normal model (a.k.a Fisher consistency). In order to achieve consistency under the normal model, $\boldsymbol{S_{MVV}}$ (Equation 3.2) is multiplied by a consistency factor, $c(h)$, as follows,

$$c(h)\boldsymbol{S_{MVV}} = \frac{c(h)}{h}\sum_{i=1}^{h}(\boldsymbol{x_i} - \boldsymbol{m_{MVV}})(\boldsymbol{x_i} - \boldsymbol{m_{MVV}})^t \tag{5.2}$$

The approximation of consistency factor can be obtained from elliptical truncation in the multivariate normal distribution based on squared distance. If $\boldsymbol{x_i} \sim N(\boldsymbol{\mu}, \Sigma)$, $c(h)$ is defined as

$$c(h) = \frac{h/n}{P(\chi^2_{p+2} < \chi^2_{p,h/n})} \tag{5.3}$$

where $\chi^2_{p,h/n}$ is the h/n-quantile of χ^2_p distribution. This formula is derived by Butler et al. (1993) and further discussed in Croux and Haesbroeck (1999) based on the functional form of the MCD estimator. Since MVV have the same functional form with the MCD estimator, we used Equation 5.3 as the consistency factor for $\boldsymbol{S_{MVV}}$. Albeit guaranteed consistency under normality distribution Pison et al. (2002) showed that MCD estimators were biased for small sample sizes. Thus, the consistency factor in Equation 5.3 might not be sufficient to make MVV estimator unbiased for small sample sizes. For that reason, we include the computation of correction factor at any sample size *n* and dimension *p*.

5.3 Correction Factor

A simulation study on the effect of correction factor on the MVV estimator is carried out for several sample sizes *n* and dimension *p=2,5,10,15,20*. We generated data sets

$X^{(j)} \epsilon\ \mathbb{R}^{n \times p}$ from standard multivariate normal distribution. It suffices to consider the standard multivariate normal distribution since the MVV is affine equivariant. For each data set $X^{(j)}, j = 1, \dots, m$ we then determine the $\boldsymbol{c(h)S_{MVV}^{(j)}}$ in Equation 5.2. If the estimator is unbiased, $E[\boldsymbol{c(h)S_{MVV}}] = \boldsymbol{I_p}$, therefore the *p*-th root of the determinant of $\boldsymbol{c(h)S_{MVV}}$ equals 1(Pison et al., 2002). The mean of the *p*-th root of the determinant is given as

$$mean(|\boldsymbol{c(h)S_{MVV}}|) = \frac{1}{m}\sum_{\mathrm{j}=1}^{m}(\left|\boldsymbol{c(h)S_{MVV}^{(j)}}\right|)^{1/p}$$

To determine the correction factor, we performed $m = 1000$ simulations for different sample sizes *n* and dimensions *p*, with $\alpha = 0.05$ such that

$$\vartheta_{p,n}^{\alpha} = \frac{1}{mean(|\boldsymbol{c(h)S_{MVV}}|)} \tag{5.4}$$

The computed values are displayed in Table 5.1. Then, using $\vartheta_{p,n}^{\alpha}$ in Equation 5.4 as the correction factor for $\boldsymbol{c(h)S_{MVV}}$, we obtain

$$\boldsymbol{\vartheta_{p,n}^{\alpha}c(h)S_{MVV}} = \frac{\vartheta_{p,n}^{\alpha}c(h)}{h}\sum_{i=1}^{h}(\boldsymbol{x_i} - \boldsymbol{m_{MVV}})(\boldsymbol{x_i} - \boldsymbol{m_{MVV}})^t \tag{5.5}$$

For $\boldsymbol{\vartheta_{p,n}^{\alpha}c(h)S_{MVV}}$ can be considered consistent and unbiased, the determinant of $\boldsymbol{\vartheta_{p,n}^{\alpha}c(h)S_{MVV}}$ should approach 1.

Table 5.1: Values of $\vartheta_{p,n}^{\alpha}$ for $\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}}$

***p*=2**		***p*=5**		***p*=10**		***p*=15**		***p*=20**	
$\boldsymbol{n}$	$\boldsymbol{\vartheta_{p,n}^{\alpha}}$	$\boldsymbol{n}$	$\boldsymbol{\vartheta_{p,n}^{\alpha}}$	$\boldsymbol{n}$	$\boldsymbol{\vartheta_{p,n}^{\alpha}}$	$\boldsymbol{n}$	$\boldsymbol{\vartheta_{p,n}^{\alpha}}$	$\boldsymbol{n}$	$\boldsymbol{\vartheta_{p,n}^{\alpha}}$
10	5.8276	30	2.9304	50	2.3127	80	1.9801	100	1.8503
25	5.4200	50	2.9598	100	2.2098	100	1.9528	200	1.7762
50	5.7715	100	2.9309	200	2.1524	200	1.8937	300	1.7490
100	5.7206	200	2.8912	500	2.1045	500	1.8476	500	1.7180
200	5.6679	500	2.8579						
500	5.5842								

5.4 Investigation through Simulation Experiment

Garther and Becker (1997) have emphasized that robust estimators to be used in the method of outliers detection should have sufficient rate of convergence to some true underlying model parameter for consistency and unbiasedness. A sequence of asymptotically unbiased estimators for parameter θ is called consistent if $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$. To illustrate the analysis on the consistency of MVV estimator at multivariate normal, data are randomly generated from $N(0, \boldsymbol{I_p})$. The experiment is carried out for several values of sample sizes *n* until convergent for a fixed moderate dimension such that, *p=10*. Figure 5.1 shows the determinants of $\boldsymbol{\vartheta_{p,n}^{\alpha} c(h) S_{MVV}}$ corresponding to the sample size, *n*. As the value of *n* increases, we can observe that the determinant approaches 1 which implies that the $\boldsymbol{\vartheta_{p,n}^{\alpha} c(h) S_{MVV}}$ is consistent.

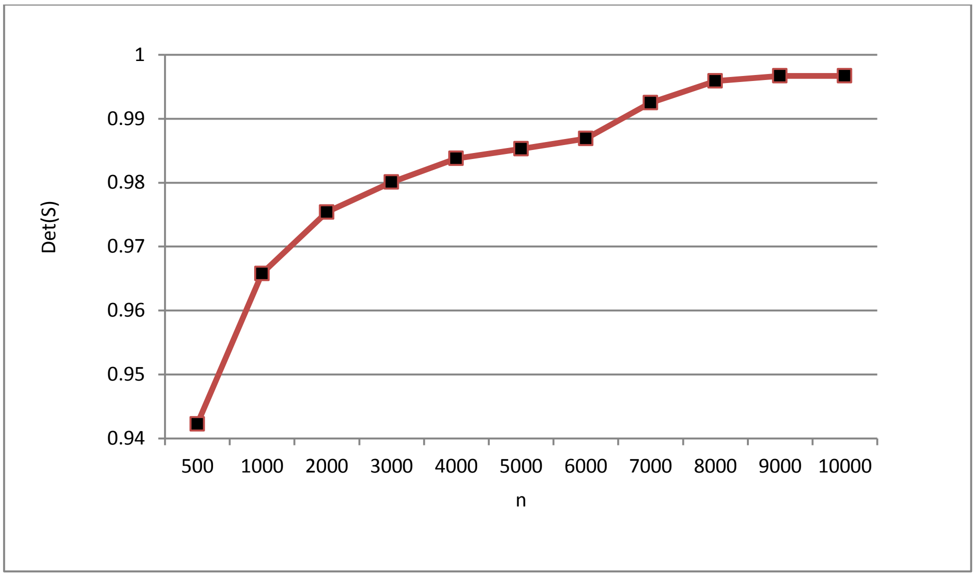


Figure 5.1: Determinant of $\boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}}$ *when n → ∞ and p = 10*

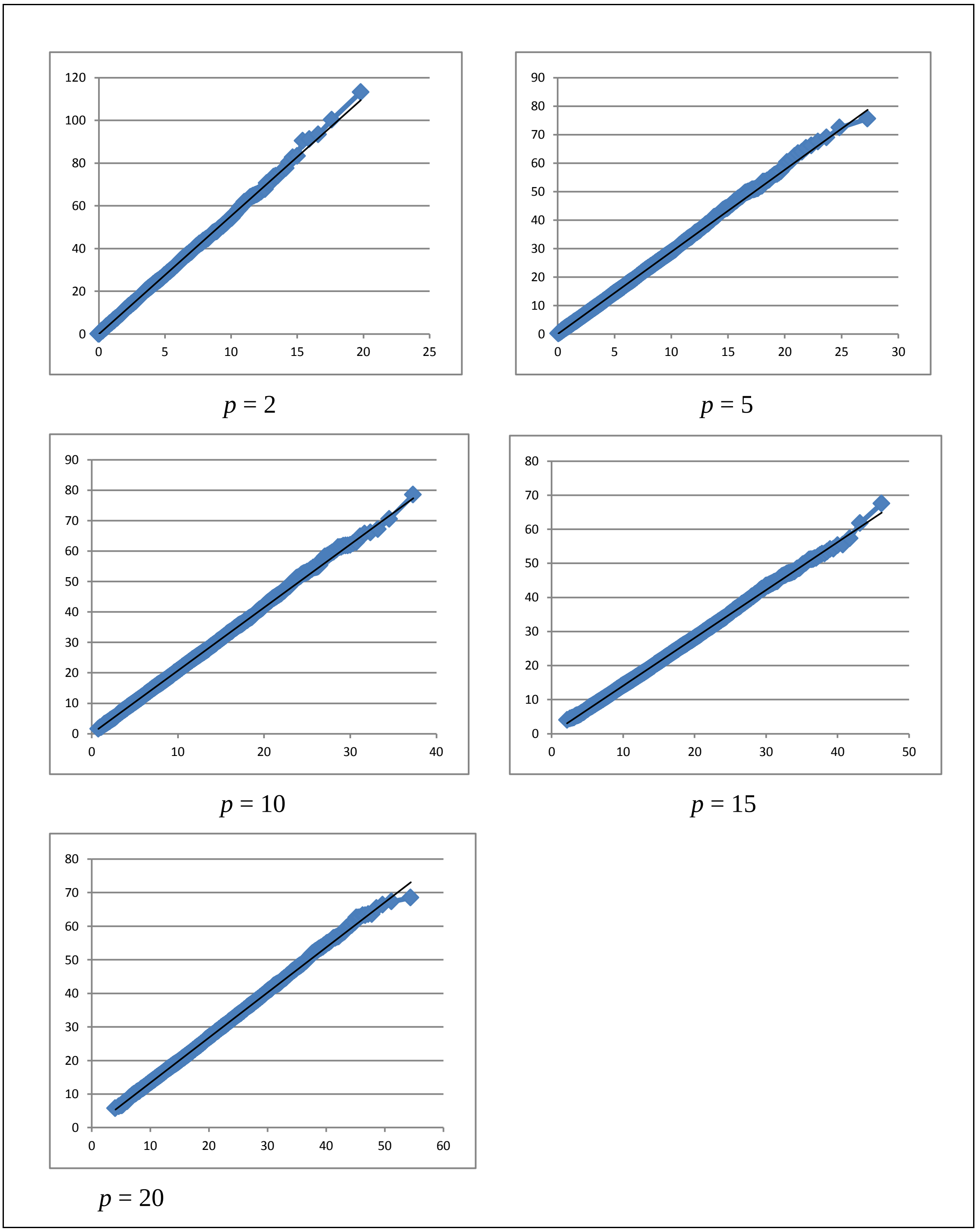
Next, the investigation using simulation experiment continues to show that $\boldsymbol{m}_{\boldsymbol{MVV}}$ and $\boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}}$ which replaced the MLE estimators, $\boldsymbol{\mu}$ and Σ, in Hotelling T^2 are consistent and unbiased. The squared distances of any affine-equivariant robust location and scatter estimators which are consistent and unbiased under normal model is asymptotically χ^2 distributed (Grubel & Rocke, 1990; Rocke & Woodruff, 1996; Garther & Becker, 1997).Therefore if $\boldsymbol{m}_{\boldsymbol{MVV}}$ and $\boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}}$ are consistent and unbiased estimators for $\boldsymbol{\mu}$ and Σ, then with observations $\boldsymbol{x}_{\boldsymbol{i}}$ i.i.d in $\mathbb{R}^p \sim N_p(\boldsymbol{\mu}, \Sigma)$, it follows that

$$d_i^2 = (\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})\boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}}{}^{-1}(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})^t$$

is asymptotically χ_p^2 distributed. Since d_i^2 is similar to Hotelling T^2, the asymptotic distribution of the improved Hotelling T_{MVV}^2 when $n \to \infty$ should also follow χ_p^2 distribution if the estimators are consistent. If we consider a sample of p quality characteristics such that $\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \dots, x_{ip}\}$ where i=1,2, ...,n as a phase I data set, then the improved T_{MVV}^2 statistic for $\boldsymbol{x_i}$ can be constructed in the following manner:

$$T_{MVV(I)}^2(i) = (\boldsymbol{x_i} - \boldsymbol{m_{MVV}})\boldsymbol{\vartheta_{p,n}^{\alpha} c(h) S_{MVV}}^{-1}(\boldsymbol{x_i} - \boldsymbol{m_{MVV}})^t \tag{5.6}$$

To check on the distributions of the improved T_{MVV}^2, we employed the QQ plots as done by Garrett (1989)and the results are shown in Figure 5.2 and 5.3. Each figure represents the QQ plot of χ_p^2 distribution versus the original T_{MVV}^2 ($T_{MVV(o)}^2$) and improved T_{MVV}^2 ($T_{MVV(I)}^2$) respectively. Random data were generated from multivariate standard normal distribution $MVN(0, I_p)$. This study is carried out for the sample size of *n = 10,000* with dimensions of *p = 2, 5, 10, 15, 20.* Based on the plots, it is seemingly reasonable to claim that the distribution of $T_{MVV(o)}^2$ and $T_{MVV(I)}^2$ is asymptotically equal to χ_p^2 distribution. To further clarify the situation, the goodness of fit on those plots is evaluated based on the slope and the *R*-square of the straight line in accordance to the data plot, as shown in Table 5.2.



$p = 2$

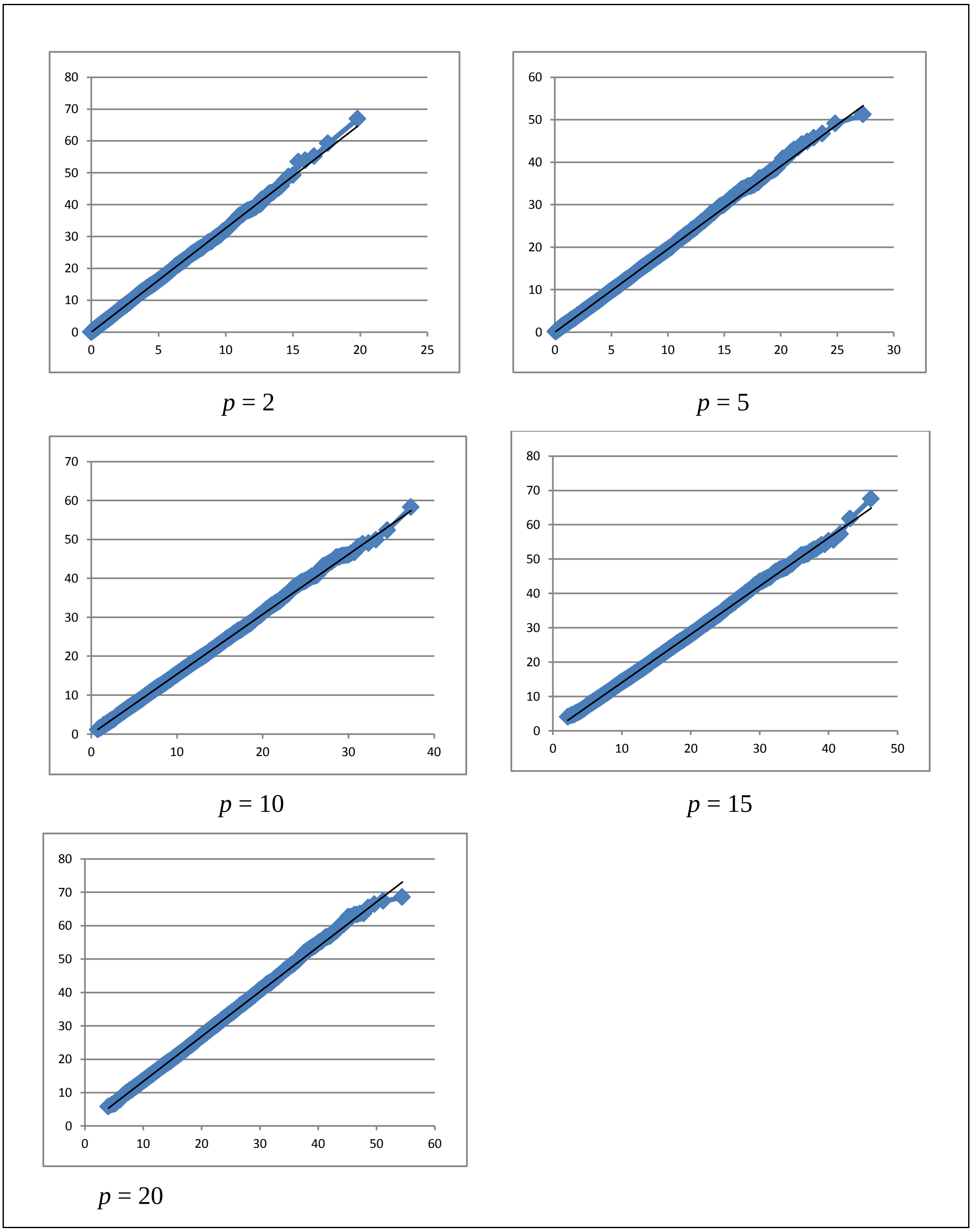
$p = 5$

$p = 10$

$p = 15$

$p = 20$

Figure 5.2: QQ plot between χ^2_p distribution versus simulated $T^2_{MVV(o)}$ for n=10,000



p = 2

p = 5

p = 10

p = 15

p = 20

Figure 5.3: QQ plot between χ_p^2 distribution versus simulated $T^2_{MVV(I)}$ for n=10,000

Table 5.2: The slope and R-square for $T^2_{MVV(o)}$ and $T^2_{MVV(I)}$

n = 10 000		$\boldsymbol{T^2_{MVV(o)}}$	$\boldsymbol{T^2_{MVV(I)}}$
p=2	R^2	0.999	0.999
	slope	3.264	1.001
p=5	R^2	0.999	0.999
	slope	1.952	1.020
p=10	R^2	0.999	0.999
	slope	1.538	1.003
p=15	R^2	0.999	0.999
	slope	1.404	1.001
p=20	R^2	0.999	0.999
	slope	1.341	1.005

The hypothetical distribution represents the χ^2_p without error if all points are in a straight line with slope equals 1 and *R*-square also equals 1 (Ali, Djauhari & Syed-Yahaya, 2008). From this table we observe that the *R*-square values for all *p*'s are 0.999. With regards to the slopes, we can see a considerable difference in the values between the original $T^2_{MVV(o)}$ and $T^2_{MVV(I)}$ especially when p = 2. The slopes for $T^2_{MVV(I)}$ are consistent and approximately equals to 1 regardless of the dimensions (p). In contrast, the slopes for $T^2_{MVV(o)}$ are quite a distance away from 1 even though the pattern shows a declining in values towards 1 as *p* increases. We observe that the values for the two measurements (R^2 and slopes) are very close to the ideal value, which signify that the χ_p^2 distribution fits well with the simulated $T_{MVV(I)}^2$ values. The result implies that the constant $\vartheta_{p,n}^{\alpha}c(h)$ fulfills the condition of the multiplicative factors to makes the $\boldsymbol{S_{MVV}}$ estimators consistent and unbiased for Σ.

5.5 Estimation of Control Limit

Let $\boldsymbol{x_i} = \{\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{ip}}\}$ be the *p*-variate random sample of *n* observations of preliminary data set in Phase I. Calculate the $\boldsymbol{m_{MVV}}$ and $\boldsymbol{\vartheta_{p,n}^{\alpha}c(h)S_{MVV}}$ estimators. Since the estimators are known to be free from outliers due to their estimation process, they could be readily used as in-control estimators in Phase II. By using these estimates, $T_{MVV(I)}^2(g)$ statistic is computed for Phase II observation, $\boldsymbol{x_g} = \{\boldsymbol{x_{n+1}}, \boldsymbol{x_{n+2}}, \dots\}$ where $\boldsymbol{x_g} \notin \boldsymbol{x_i}$, such that

$$T_{MVV(I)}^2(g) = (\boldsymbol{x_g} - \boldsymbol{m_{MVV}})\boldsymbol{\vartheta_{p,n}^{\alpha}c(h)S_{MVV}}^{-1}(\boldsymbol{x_g} - \boldsymbol{m_{MVV}})^t \qquad (5.7)$$

We present the control limit of the improved $T_{MVV(I)}^2(g)$ control chart by using simulated data with different combinations of sample sizes, *n*, and number of dimensions, p. The control limit of $T_{MVV(I)}^2(g)$ chart is then compared with the control limit of $T_{MVV(o)}^2(g)$ chart, robust Hotelling T^2 chart using MCD (T_{MCD}^2) and the traditional Hotelling T^2 charts (T_0^2 and T_S^2). We apply Monte Carlo method to estimate the quantiles of the $T_{MVV(I)}^2(g)$, for several combinations of sample sizes and dimensions. In order to estimate the 95% quantile of $T_{MVV(I)}^2(g)$ for a given

Phase I of sample size *n* and dimension *p*, we generate *K* = 5000 samples of size *n* from a standard multivariate normal distribution, $MVN_p(0, I_p)$. For each data set of size *n*, we compute the MVV mean vector and the modified covariance matrix estimates, $\boldsymbol{m_{MVV}}(k)$ and $\boldsymbol{\vartheta^{\alpha}_{p,n}c(h)S_{MVV}}(k)$ respectively from $k = 1, \dots, K$. In addition, for each data set, we randomly generate a new observation $\boldsymbol{x_{g,k}}$ treated as a Phase II observation from $MVN_p(0, I_p)$ and calculate the corresponding $T^2_{MVV(I)}(g, k)$ values. The empirical distribution function of $T^2_{MVV(I)}(g)$ is based on the simulated values

$$T^2_{MVV(I)}(g, 1), T^2_{MVV}(g, 2), \dots, T^2_{MVV}(g, K) \qquad (5.8)$$

We sort $T^2_{MVV(I)}(g, k)$ values in ascending order, and the UCL is the 95% quantile of the 5000 statistics. The results of the investigation are presented in Table 5.3. We observe that the estimated UCLs for $T^2_{MVV(o)}(g)$ are large as compared to the traditional control charts (T^2_O and T^2_S) and MCD chart (T^2_{MCD}). However, after making the MVV scatter estimator consistent and unbiased as shown in Equation 5.7, the results improved immensely. As we can see here, the UCLs are closer to the traditional UCLs (Table 5.3).

Table 5.3: Control limits

P	n	T_{O}^{2}	T_{S}^{2}	T_{MCD}^{2}	$T_{MVV(o)}^{2}$	$T_{MVV(I)}^{2}$
2	10	11.0360	13.1050	30.6250	76.0122	19.7067
	25	7.4275	7.7676	12.1798	32.4008	9.4443
	50	6.6447	6.7431	8.2762	28.1107	8.0556
	100	6.3039	6.3444	7.4463	24.6037	7.1969
	200	6.1443	6.1598	6.4127	21.7088	6.4469
	500	6.0518	6.0566	6.1558	20.4264	6.1788
5	30	15.6006	16.9160	27.6404	41.9567	19.5315
	50	13.4506	13.9253	18.3456	33.5214	15.9398
	100	12.1579	12.3055	14.7736	28.4822	14.0082
	200	11.5915	11.6454	12.5765	25.6204	12.9297
	500	11.2738	11.2904	11.4941	22.5296	11.5868
10	50	25.9552	27.7721	39.8024	62.9323	34.5417
	100	21.5264	21.9969	26.0646	43.1889	25.5450
	200	19.7975	19.9570	20.9145	34.8509	21.4812
	500	18.8777	18.9275	19.5618	31.1418	19.8112
15	80	33.6517	35.1547	42.7942	69.0937	42.6982
	100	31.5083	32.5354	37.4367	61.0544	38.6310
	200	27.9034	28.2452	29.4809	45.9981	30.6107
	500	26.0882	26.1820	26.6016	39.7181	26.1456
20	100	42.5747	44.6890	52.7273	83.5238	53.9627
	200	36.2033	36.8213	38.8163	57.2303	39.3577
	300	34.4609	34.7659	36.0595	52.3438	36.8230
	500	33.1766	33.3434	33.6574	47.7808	34.4221

5.6 Performance of $\boldsymbol{T^2_{MVV}}$ Control Chart

Next, our investigation continues with the performance of the T^2_{MVV} control charts before and after making the MVV scatter estimator consistent and unbiased. The performance is measured by the probability of detection and false alarm rates. To determine the false alarm rate and probability of detection, we randomly generate a Phase II observation with in-control and out of control parameters respectively from Phase 1 and calculate $T^2_{MVV(o)}$ and $T^2_{MVV(I)}$ statistics. The false alarm rate or probability of detection is estimated as the proportion of statistic values that are above the control limits of 1000 replications. Data for Phase I are simulated based on the various conditions created for this study. To examine the effect of contamination on the charts' performance, we have considered a contaminated model discussed in Section 3.4.2. The results of the investigation are presented in Table 5.4 – 5.8. Each table represents each dimension arranged in ascending order i.e. p = 2, 5, 10, 15 and 20 with α = 0.05. The first column in each table displays the number of sample sizes, followed by the percentage of outliers and non-centrality values respectively in the second and third column. As shown in Tables 5.4 – 5.8, the performance of the control chart for $T^2_{MVV(o)}$ and $T^2_{MVV(I)}$ in terms of probability detection and false alarm rate for each condition remain the same despite the changes in UCL. This indicates that the consistent and unbiased MVV scatter estimator is able to improve the UCL value while simultaneously maintains the good performance of the Hotelling T^2 control chart.

Table 5.4: Probability of detection and false alarm rates of the corresponding control charts with dimension, p = 2

Sample Size(n)	% outliers(ε)	Mean shift(μ $_1$)	$T^2_{MVV(o)}$		$T^2_{MVV(I)}$	
			Probability Detection	**False alarm**	**Probability Detection**	**False alarm**
	0	0	0.0520		0.0520	
	10%	3	0.5320	0.0450	0.5320	0.0450
		5	0.9080	0.0450	0.9080	0.0450
	20%	3	0.4270	0.0330	0.4270	0.0330
		5	0.8530	0.0330	0.8530	0.0330
25	0	0	0.0530		0.0530	
	10%	3	0.8320	0.0390	0.8320	0.0390
		5	0.9980	0.0390	0.9980	0.0390
	20%	3	0.7210	0.0190	0.7210	0.0190
		5	0.9960	0.0190	0.9960	0.0190
50	0	0	0.0540		0.0540	
	10%	3	0.8930	0.0350	0.8930	0.0350
		5	1	0.0340	1	0.0340
	20%	3	0.8280	0.0180	0.8280	0.0180
		5	1	0.0170	1	0.0170
100	0	0	0.0490		0.0490	
	10%	3	0.9190	0.0300	0.9190	0.0300
		5	1	0.0290	1	0.0290
	20%	3	0.8890	0.0150	0.8890	0.0150
		5	1	0.0150	1	0.0150
200	0	0	0.0690		0.0690	
	10%	3	0.9460	0.0490	0.9460	0.0490
		5	1	0.0500	1	0.0500
	20%	3	0.9140	0.0280	0.9140	0.0280
		5	1	0.0280	1	0.0280
500	0	0	0.0630		0.0630	
	10%	3	0.9520	0.0490	0.9520	0.0490
		5	1	0.0480	1	0.0480
	20%	3	0.9310	0.0230	0.9310	0.0230
		5	1	0.0230	1	0.0230

Table 5.5: Probability of detection and false alarm rates for independent case with dimension, p = 5

Sample Size (n)	% outliers (ε)	Mean shift (μ1)	Control Charts			
			$T^2_{MVV(o)}$		$T^2_{MVV(I)}$	
			Probability Detection	False alarm	Probability Detection	False alarm
30	0	0	0.0500		0.0500	
	10%	3	0.9770	0.0300	0.9770	0.0300
		5	1	0.0330	1	0.0330
	20%	3	0.9650	0.0210	0.9650	0.0210
		5	1	0.0200	1	0.0200
50	0	0	0.0490		0.0490	
	10%	3	0.9910	0.0350	0.9910	0.0350
		5	1	0.0370	1	0.0370
	20%	3	0.9890	0.0220	0.9890	0.0220
		5	1	0.0230	1	0.0230
100	0	0	0.0380		0.0380	
	10%	3	1	0.0300	1	0.0300
		5	1	0.0320	1	0.0320
	20%	3	0.9970	0.0170	0.9970	0.0170
		5	1	0.0190	1	0.0190
200	0	0	0.0390		0.0390	
	10%	3	1	0.0350	1	0.0350
		5	1	0.0350	1	0.0350
	20%	3	0.9990	0.0220	0.9990	0.0220
		5	1	0.0220	1	0.0220
500	0	0	0.0430		0.0430	
	10%	3	1	0.0360	1	0.0360
		5	1	0.0370	1	0.0370
	20%	3	1	0.0190	1	0.0190
		5	1	0.0190	1	0.0190

Table 5.6:Probability of detection and false alarm rates for dimension, p = 10

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts			
			$T^2_{MVV(o)}$		$T^2_{MVV(I)}$	
			Probability Detection	False alarm	Probability Detection	False alarm
50	0	0	0.0520		0.0520	
	10%	3	1	0.0370	1	0.0370
		5	1	0.0380	1	0.0380
	20%	3	0.9990	0.0250	0.9990	0.0250
		5	1	0.0220	1	0.0220
100	0	0	0.0450		0.0450	
	10%	3	1	0.0390	1	0.0390
		5	1	0.0350	1	0.0350
	20%	3	1	0.0240	1	0.0240
		5	1	0.0230	1	0.0230
200	0	0	0.0520		0.0520	
	10%	3	1	0.0390	1	0.0390
		5	1	0.0420	1	0.0420
	20%	3	1	0.0250	1	0.0250
		5	1	0.0240	1	0.0240
500	0	0	0.0540		0.0540	
	10%	3	1	0.0390	1	0.0390
		5	1	0.0390	1	0.0390
	20%	3	1	0.0260	1	0.0260
		5	1	0.0230	1	0.0230

Table 5.7:Probability of detection and false alarm rate for dimension, p = 15

Sample Size (n)	% outliers (ε)	Mean shift (μ$_1$)	Control Charts			
			$T^2_{MVV(o)}$		$T^2_{MVV(I)}$	
			Probability Detection	False alarm	Probability Detection	False alarm
80	0	0	0.0560		0.0560	
	10%	3	1	0.0470	1	0.0470
		5	1	0.0430	1	0.0430
	20%	3	1	0.0270	1	0.0270
		5	1	0.0320	1	0.0320
100	0	0	0.0520		0.0520	
	10%	3	1	0.0450	1	0.0450
		5	1	0.0430	1	0.0430
	20%	3	1	0.0250	1	0.0250
		5	1	0.0220	1	0.0220
200	0	0	0.0470		0.0470	
	10%	3	1	0.0420	1	0.0420
		5	1	0.0410	1	0.0410
	20%	3	1	0.0200	1	0.0200
		5	1	0.0240	1	0.0240
500	0	0	0.0470		0.0470	
	10%	3	1	0.0390	1	0.0390
		5	1	0.0390	1	0.0390
	20%	3	1	0.0260	1	0.0260
		5	1	0.0290	1	0.0290

Table 5.8:Probability of detection and false alarm rate for dimension, p = 20

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts			
			$T^2_{MVV(o)}$		$T^2_{MVV(I)}$	
			Probability Detection	False alarm	Probability Detection	False alarm
100	0	0	0.0530		0.0530	
	10%	3	1	0.0400	1	0.0400
		5	1	0.0420	1	0.0420
	20%	3	0.9970	0.0300	0.9970	0.0300
		5	0.9980	0.0240	0.9980	0.0240
200	0	0	0.0510		0.0510	
	10%	3	1	0.0370	1	0.0370
		5	1	0.0380	1	0.0380
	20%	3	1	0.0310	1	0.0310
		5	1	0.0250	1	0.0250
300	0	0	0.0440		0.0440	
	10%	3	1	0.0350	1	0.0350
		5	1	0.0340	1	0.0340
	20%	3	1	0.0240	1	0.0240
		5	1	0.0220	1	0.0220
500	0	0	0.0530		0.0530	
	10%	3	1	0.0400	1	0.0400
		5	1	0.0400	1	0.0400
	20%	3	1	0.0350	1	0.0350
		5	1	0.0320	1	0.0320

5.7 Real Data Analysis

The application of the improved method $T^2_{MVV(I)}$ on real data is illustrated by using data in Chapter 4 in Section 4.3 which involves the production of advanced composite panels for the aircraft industry. The product consists of 3 quality variables (dimensions) namely trim edge (x1), trim edge spar (x2), and drill hole (x3). The details of the 21 spoilers were collected as Phase I data and 23 spoilers as Phase II data displayed in Table 4.7 and 4.9 respectively.

Estimates for the location vector ($\bar{\boldsymbol{x}}$) and scatter matrix ($\boldsymbol{S}$) calculated using Phase I data are presented in Table 5.9. The upper control limits (UCLs) based on the estimates are displayed in the last column of the table. The values of the various types of T^2 statistics for Phase II data used in this study are shown in Table 5.10. As could be observed in Table 5.9, the UCL for $T^2_{MVV(I)}$ is smaller than the rest of the values except T^2_O. There is only a small different between the $T^2_{MVV(I)}$ (11.5513) and the original T^2_O (11.035). When compared with the original T^2_{MVV} ($T^2_{MVV(o)}$), we observe a large disparity between the two UCL values ($T^2_{MVV(o)}$= 41.298 and $T^2_{MVV(I)}$= 11.5513). Nevertheless, the ability of $T^2_{MVV(o)}$ and $T^2_{MVV(I)}$ in detecting the out of control data (highlighted) still remain the same as we can see in Table 5.10. Four statistics, namely $T^2_{MVV(o)}$, $T^2_{MVV(I)}$, T^2_{MCD} and T^2_S signal observations 20, 22 and 25 as out-of-control but T^2_O only signals 20 and 25 as out-of-control observations and fails to signal observation 22. Even though with low UCL value, T_O^2 is unable to detect the out of control data unlike $T_{MVV(I)}^2$.

Table 5.9: Estimates of location vector, covariance matrix and UCL.

Types of Control Chart	**Location Vector ($\overline{\boldsymbol{x}}$)**	**Scatter Matrix ($\boldsymbol{S}$)**	**Upper Control Limit (UCL)**
T_O^2	$[0.00504 \quad 0.00284 \quad 0.01579]$	$\begin{bmatrix} 0.00004 & 0.00002 & 0.00003 \\ 0.00002 & 0.00009 & 0.00001 \\ 0.00003 & 0.00001 & 0.00011 \end{bmatrix}$	11.035
T_S^2	$[0.00365 \quad 0.00256 \quad 0.01209]$	$\begin{bmatrix} 0.00001 & 0.00000 & 0.00000 \\ 0.00000 & 0.00003 & -0.00001 \\ 0.00000 & -0.00001 & 0.00003 \end{bmatrix}$	11.798
T_{MCD}^2	$[0.00414 \quad 0.00207 \quad 0.01096]$	$\begin{bmatrix} 0.00002 & 0.00000 & 0.00000 \\ 0.00002 & 0.00009 & -0.00002 \\ 0.00000 & -0.00002 & 0.00003 \end{bmatrix}$	21.946
$T_{MVV(o)}^2$	$[0.00336 \quad 0.00354 \quad 0.00913]$	$\begin{bmatrix} 0.00001 & 0.00001 & 0.00000 \\ 0.00001 & 0.00003 & 0.00000 \\ 0.00000 & 0.00000 & 0.00001 \end{bmatrix}$	41.298
$T_{MVV(I)}^2$	$[0.00336 \quad 0.00354 \quad 0.00913]$	$\begin{bmatrix} 0.00003 & 0.00002 & -0.00001 \\ 0.00002 & 0.00007 & -0.00001 \\ -0.00001 & -0.00001 & 0.00002 \end{bmatrix}$	11.5513

Table 5.10:Hotelling T^2 values for future data (Phase II)

No.	$\boldsymbol{T_O^2}$	$\boldsymbol{T_S^2}$	$\boldsymbol{T_{MCD}^2}$	$\boldsymbol{T_{MVV(o)}^2}$	$\boldsymbol{T_{MVV(I)}^2}$
1	0.5582	1.4242	1.7659	4.3908	1.5661
2	0.9003	2.5492	2.4694	5.1695	1.8438
3	0.4992	0.4936	0.3437	0.2992	0.1067
4	0.5463	1.0157	0.5456	1.5064	0.5373
5	0.4592	0.9588	0.4580	3.7869	1.3507
6	0.9013	1.7480	1.2527	2.2421	0.7997
7	3.0933	4.1372	4.4404	6.5361	2.3313
8	0.8061	1.2884	0.6837	1.0556	0.3765
9	7.3602	9.6843	14.9766	26.0499	9.2913
10	3.6198	5.8035	9.7417	19.1760	6.8396
11	5.3839	8.0897	11.8717	19.6313	7.0019
12	2.7387	4.7949	2.9788	8.1388	2.9029
13	3.8058	5.6890	7.4040	11.3895	4.0623
14	2.0548	6.3468	3.3086	9.1498	3.2635
15	2.5073	5.0227	6.8054	12.3881	4.4185
16	1.1976	1.8980	1.0679	2.0563	0.7334
17	1.5798	2.2630	1.7597	2.8765	1.0260
18	5.7910	7.9657	9.2817	13.9293	4.9682
19	1.8304	4.7003	2.4178	4.8791	1.7402
20	**38.1397**	**190.2969**	**214.9233**	**894.5184**	**319.0497**
21	1.2651	2.3301	1.5486	2.0641	0.7362
22	8.4181	**19.7720**	**24.6552**	**45.2462**	**16.1381**
23	3.7588	5.1645	4.8793	7.5328	2.6867
24	1.0602	1.7564	0.9320	2.23575	0.7974
25	**42.8447**	**134.6222**	**68.6307**	**116.02933**	**41.3844**
26	0.4832	1.3946	0.7796	7.32655	2.6132

5.8 Discussion

The UCL value for the Hotelling T^2 control chart using consistent and unbiased MVV estimators seemed to improve significantly from the Hotelling T^2 control chart using the original MVV estimators. The improved control chart ($T_{MVV(I)}^2$) was put to test on simulated and real industrial data. The finding showed that the improved T_{MVV}^2 performed well in detecting out of control data with a more stringent UCL value as compared to the original T_{MVV}^2 (unimproved, $T_{MVV(o)}^2$). With good properties and performance, this improved MVV estimators should be considered as alternative estimators to replace the usual mean and variance vector in the construction of the robust Hotelling T^2 control chart as well as other multivariate statistical procedures.

Even though Herwindianti (2006) and Herwindiati et al., (2007) had proved that MVV estimators possess three major properties of a good robust estimators i.e. high breakdown point (BP=0.5), affine equivariance and computational efficiency, the statistical efficiency of MVV estimators has never been shown before. The statistical efficiency is always a very important performance measure for any statistical procedure (Zuo, 2006, p.7). If robust multivariate estimators are to be of practical use in statistical inference they should offer a reasonable efficiency under the normal model and a manageable asymptotic distribution. Nonetheless, robust estimators are commonly not very efficient. Minimum covariance determinant (MCD) estimators introduced by Rousseeuw (1985) served as a perfect example. It has good theoretical properties i.e. affine equivariance, high breakdown value, bounded influence function and also has better convergence rate (Butler et al., 1993; Croux & Haesbroeck, 1999). However the estimators are not efficient at normal models and this is especially true at high breakdown point; see Croux and Haesbroeck (1999). To overcome the low efficiency drawback of the MCD estimators, thus Rousseeuw and van Zomeren (1990) suggested reweighted version to attain high statistical efficiency. Croux and Haesbroeck (1999) employed the reweighted version and noticed that this approach maintains the breakdown point of the initial MCD estimators, while attaining a better efficiency.

Taking into consideration the above problem, in next the chapter we propose an improvement over the algorithm as suggested by Herwindiati (2006) in the context of statistical efficiency. It consists of a one-step reweighted for MVV estimators. The reweighting scheme will be able to maintain the outlier resistance of the initial estimator and at the same time attains 100% efficiency at the normal distribution (Croux & Haesbroeck, 1999).

CHAPTER SIX
A ROBUST AND EFFICIENT REWEIGHTED ESTIMATOR OF MULTIVARIATE LOCATION AND SCATTER

6.1 Introduction

In contrast with the traditional method, when the estimators in Hotelling T^2 control chart are calculated in Phase I and directly used in Phase II analysis, the estimators should possess high statistical efficiency (Jensen et al., 2007; Chenouri et al., 2009). However, as we are already aware, there is a conflict between breakdown point and statistical efficiency as demonstrated in MCD estimators (Croux & Haesbroeck, 1999). Is the issue faced by the MCD estimators similar to MVV estimators? In addressing the issue, we proceed with further analysis to examine the properties of the estimators of MVV from the perspective of statistical efficiency.

The organization of this chapter is as follows. Section 6.2 is the investigation on statistical efficiency of MVV estimators for different breakdown points. Section 6.3 proposes reweighted version of MVV estimators and describes the algorithm to approximately calculate the estimates. Section 6.4 studies the asymptotic efficiency of the proposed estimates while section 6.5 shows results of the investigation on the finite-sample behavior of the estimator using simulation technique. Finally, conclusion result and discussion are given in the last section.

6.2 The Statistical Efficiency of MVV Estimators for Different Breakdown Points

MVV estimators were shown to be computationally efficient in Chapter 3. However, in the selection of robust estimators for any statistical estimation problems, besides high breakdown point, they should also offer a reasonable efficiency under the normal model and a manageable asymptotic distribution (Rousseeuw & van Zomeren, 1990). Nevertheless, there is a conflict between statistical efficiency and breakdown point where the efficiency of high breakdown estimators decreases when the breakdown point increases, especially when the number of dimension becomes higher (Rousseeuw & van Zomeren, 1990; Croux & Haesbroeck, 1999).

To check whether the conflict exists in MVV estimators, this study continues with the investigation on statistical efficiency of MVV estimators for different breakdown points. Two commonly chosen breakdown points are BP = 0.5 with $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ and BP = 0.25 with h = (0.75)n. The former yields highest breakdown while the latter gives a better compromise between efficiency and breakdown point. To illustrate how the efficiencies of MVV estimators vary with different breakdown points (BP) and quality characteristics (p) under normal model, we compute the asymptotic relative efficiency (ARE). The computation of asymptotic relative efficiency (ARE) is based on the definition given by Serfling (1980). For any parameter $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{p}}$, and two estimators $\widehat{\boldsymbol{\theta}}^{(j)}$ which are p-variate normal with mean $\boldsymbol{\theta}$ and non-singular covariance matrices $\Sigma_j(F)/n$ where for j=1,2 and F is the corresponding distribution, the ARE of $\widehat{\boldsymbol{\theta}}^{(\mathbf{2})}$ to $\widehat{\boldsymbol{\theta}}^{(\mathbf{1})}$ is

$$\text{ARE}(\widehat{\boldsymbol{\theta}}^{(\mathbf{2})}, \widehat{\boldsymbol{\theta}}^{(\mathbf{1})}, F) = \left(\frac{|\Sigma_1(F)|}{|\Sigma_2(F)|}\right)^{1/p} \tag{6.1}$$

Table 6.1, shows the ARE for the MVV scatter matrix with relative to the classical covariance estimator at normal model for several values of p with BP = 0.25 and 0.5. For each p, we observe a decrease in the efficiency value when BP changes from 0.25 to 0.5. As p increases in size, the efficiency value decreases regardless of the BPs. We could observe that, choosing the highest possible breakdown point (BP = 0.5) results in the loss of efficiency. Due to the conflict between efficiency and high breakdown value, Croux and Haesbroeck (1999) suggested taking BP = 0.25 as a compromise between efficiency and robustness, where the corresponding estimator can still cope with realistic amount of contamination in the data, but is much more precise when no outliers are present than the usual choice of BP = 0.5. However, we want to gain efficiency while retaining the highest breakdown point. This can be achieved by computing the reweighted version of the robust estimator (Rousseeuw & Leroy 1987; Rousseeuw & van Zomeren,1990; Rousseeuw & van Driessen 1999; Lopuhaä, 1999, Pison & van Aelst 2004).

Table 6.1: Asymptotic relative efficiency of MVV scatter matrix estimator w.r.t classical covariance estimator for normal model.

	p = 2	p = 5	p = 10	p = 15	p = 20
BP = 0.25	1.4176	1.3225	1.2411	1.2000	1.1740
BP = 0.5	1.0073	1.0000	0.9978	0.9980	0.9983

6.3 Reweighted Minimum Vector Variance (RMVV) Estimator

To increase the efficiency of robust estimators, reweighted version of the estimators is often used in practice. Rousseeuw and van Zomeren (1990) proposed a one-step reweighted version of MCD estimators by giving weight 0 to observations for which the robust Mahalanobis squared distance (MSD) statistics, $d^2_{MCD(i)}$ exceeds a threshold value. The determination of the threshold value very much depends on the exact distribution of robust distances. Nevertheless, an unsolved problem is that the exact distribution of robust distances is unknown for finite sample sizes. The common approach is to compare the squared distances with the percentage points of their asymptotic χ^2_p distribution (Cerioli, Riani & Atkinson, 2008). The usual suggestion for the threshold (e.g. Rousseeuw & Leroy 1987, p. 260; Rousseeuw & van Driessen, 1999, p. 218; Pison & van Aelst, 2004, p. 312) is to take the 0.025% cut-off point of the χ^2_p distribution.

Based on the aforementioned references, the MVV estimators are reweighted in order to improve their efficiency. Thus, in this study, the observations with $d^2_{(MVV)i} > \chi^2_{p,0.025}$, which can reasonably be suspected as outliers are given 0 weight.

The outliers will be removed and the sample mean and covariance matrix are then computed using the rest of the data. These estimators are known as the reweighted mean and covariance matrix. To ensure that the $d^2_{(MVV)i}$ statistic is asymptotically χ^2_p distribution, we investigate on the asymptotic distribution of the robust MSD through simulation.

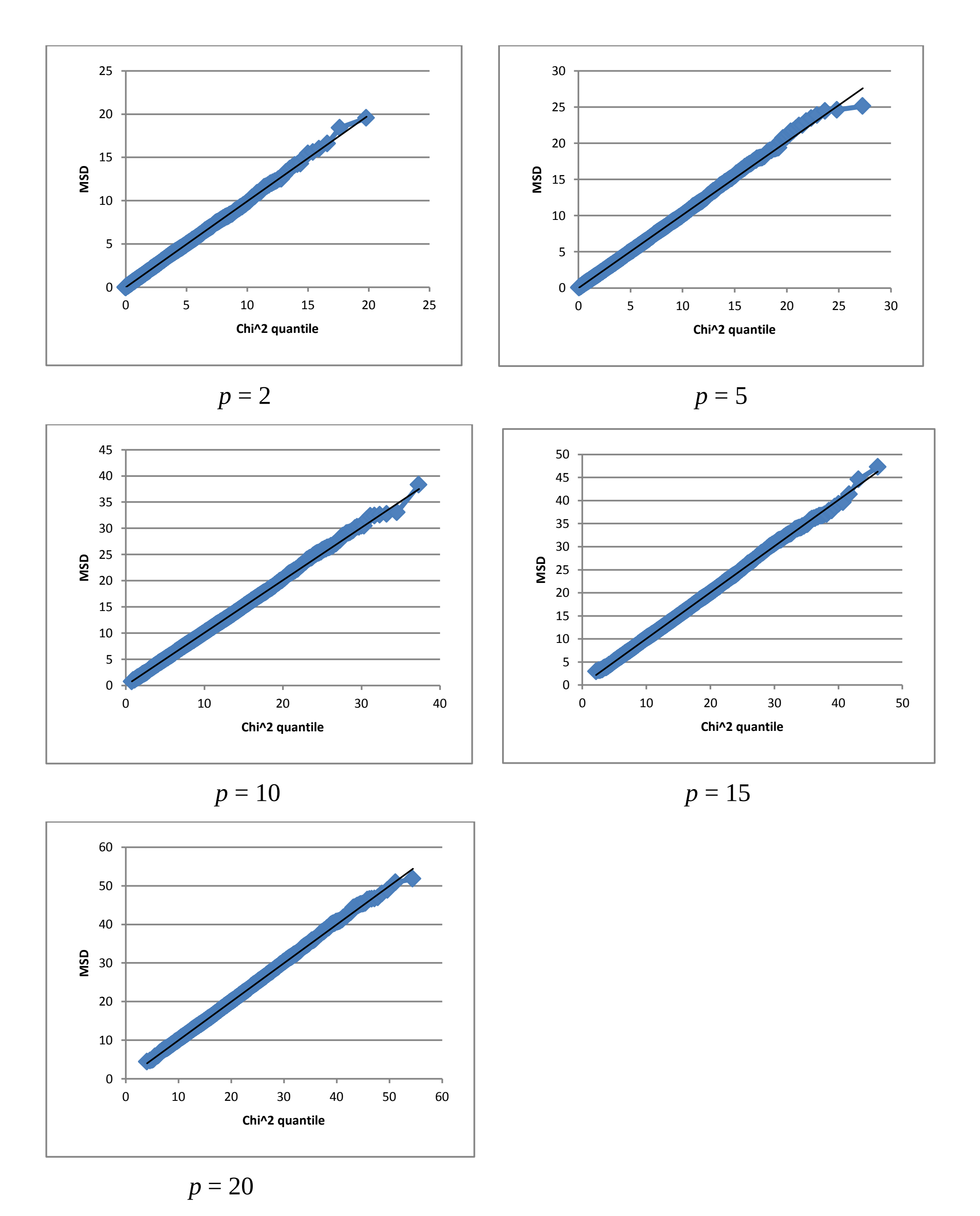
6.3.1 The Distribution of Robust MSD Based On MVV Estimators

To investigate on the asymptotic distribution of robust MSD based on MVV estimators and compare it with robust MSD based on MCD, we apply simulation experiments to show that the distributions fit the χ^2_p distribution as done by Garrett (1989). Random data of n = 10000 were generated from multivariate standard normal distribution $MVN(0, I_p)$ for several dimensions, p = 2, 5, 10, 15 and 20. Shown in Figure 6.1 and 6.2 are the QQ plots for the quantile of χ^2_p distribution on the horizontal axis versus the quantile of simulated MSD_{MCD} and MSD_{MVV} on the vertical axis respectively. Based on both figures, it is seemingly reasonable to claim that the distribution of MSD_{MCD} and MSD_{MVV} are asymptotically equal to χ^2_p distribution. To further clarify the situation, the goodness of fit of those plots is evaluated using the slope and the R-square of the straight line in accordance to the data plot. The hypothetical distribution represents the χ^2_p without error if all points are in a straight line with slope equals 1 and R-square also equals 1 (Ali *et al*., 2008).

Table 6.2 presents the values of the slope and R-square for the MSD_{MCD} and MSD_{MVV}. From this table we observe that both of them have the R-square values equal to 0.999 for all p's. With regards to the slopes, we can see a small difference in the values between MSD_{MCD} and MSD_{MVV}. The slopes for MSD_{MVV} is consistent and approximately equals to 1 regardless of the dimensions (p). From this result we can see that the values for the two measurements (R^2 and slopes) are very close to the ideal value, which signify that the simulated MSD_{MVV} and MSD_{MCD} values fit well with the χ_p^2 distribution.

Table 6.2: Slope and R-square for MSD_{MCD} and MSD_{MVV}

$\boldsymbol{n}$ **= 10 000**		$\boldsymbol{MSD_{MCD}}$	$\boldsymbol{MSD_{MVV}}$
p=2	R^2	0.999	0.999
	slope	0.995	1.001
p=5	R^2	0.999	0.999
	slope	1.010	1.020
p=10	R^2	0.999	0.999
	slope	1.005	1.003
p=15	R^2	0.999	0.999
	slope	1.002	1.001
p=20	R^2	0.999	0.999
	slope	0.999	1.005



$p = 2$

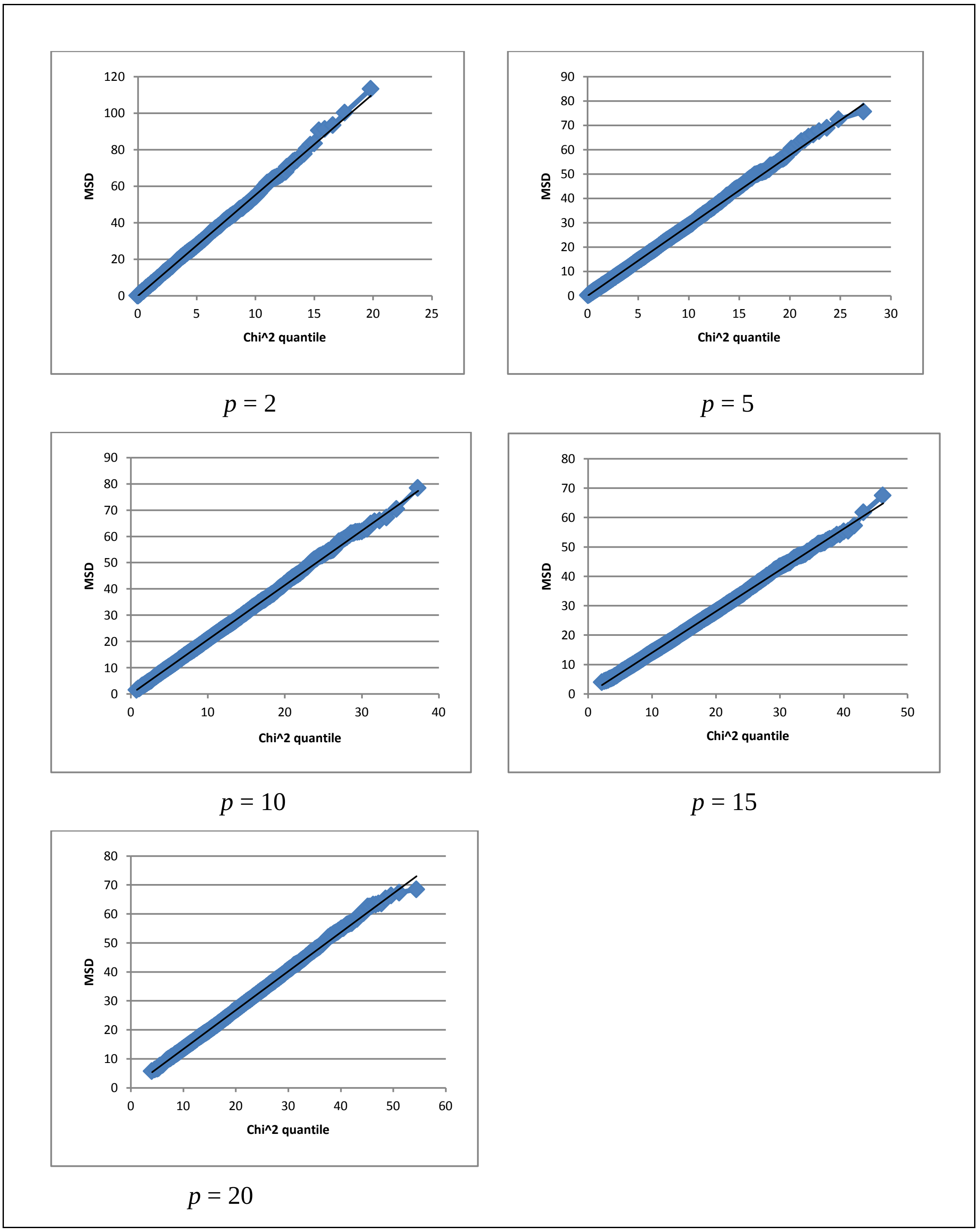
$p = 5$

$p = 10$

$p = 15$

$p = 20$

Figure 6.1: QQ plot between χ_p^2 distribution versus simulated MSD_{MCD} for n=10,000



$p = 2$

$p = 5$

$p = 10$

$p = 15$

$p = 20$

Figure 6.2: QQ plot between χ^2_p distribution versus simulated MSD_{MVV} for n=10,000

6.3.2 The Estimator

As discussed in the Section 6.3.1, the distribution of robust MSD based on MVV and MCD estimators are asymptotically equal to χ_p^2 distribution. Thus we come to a decision to use the usual suggestion for the choice of cut-off value for reweighted MCD i.e. $\chi^2_{p,0.025}$ (Rousseeuw & Leroy, 1987; Rousseeuw & van Driessen, 1999; Lopuhaä, 1999; Croux & Haesbroeck, 1999). The raw RMVV estimators of location and scatter are determined as follows:

$$\boldsymbol{m}_{\boldsymbol{RMVV}}^{\boldsymbol{raw}} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x_i}}{m} \tag{6.2}$$

$$\boldsymbol{S}_{\boldsymbol{RMVV}}^{\boldsymbol{raw}} = \frac{\sum_{i=1}^{n} w_i(\boldsymbol{x_i} - \boldsymbol{m}_{\boldsymbol{RMVV}}^{\boldsymbol{raw}})(\boldsymbol{x_i} - \boldsymbol{m}_{\boldsymbol{RMVV}}^{\boldsymbol{raw}})^t}{m} \tag{6.3}$$

where $w_i = 0$ if $d_{i_{MVV}}^2(\boldsymbol{x_i}, \boldsymbol{m_{MVV}}) > \chi^2_{p,0.025}$ and $w_i = 1$ otherwise. Therefore m represent number of observations with $d^2_{(MVV)i} \leq \chi^2_{p,0.025}$. Scatter estimators are typically calibrated to be consistent for the normal distribution, thus the consistency and correction factors are needed to guarantee Fisher consistency for the reweighted estimator and improve its biasness for small sample behavior. We take consistency factor, $c^*(m)$ as in Equation (5.3) such that

$$c^*(m) = \frac{m/n}{P(\chi^2_{p+2} < \chi^2_{p,m/n})} \tag{6.4}$$

Albeit this process guarantees consistency under normal distribution, this consistency factor is not sufficient to make the RMVV estimator unbiased for small sample sizes. To overcome the insufficiency issue, we compute the correction factor, $\vartheta^{*\alpha}_{m,n,p}$, via simulation approach for several sample sizes n and dimension p. We generated data sets $X^{(j)} \epsilon \mathbb{R}^{n\times p}$ from standard normal distribution, $N_p(0, I)$. For each data set $X^{(j)}$, $j = 1, \ldots, r$ we then determine the RMVV estimators of location and scatter as in Equation (6.2) and (6.3) followed by $\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw(j)}}_{\boldsymbol{RMVV}}$. If the estimator is unbiased, we should have $E[\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}] = \boldsymbol{I_p}$. Thus, we expect the *p*-th root of the determinant of $\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}$ equals 1 and the mean of the *p*-th root of the determinant is given by

$$r(|\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}|) = \frac{1}{r}\sum\nolimits_{\mathrm{j}=1}^{r}(\left|\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw(j)}}_{\boldsymbol{RMVV}}\right|)^{1/p},$$

where $|\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}|$ denotes the determinant of a square matrix $\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}$. We perform $r = 1000$ simulations for different sample sizes *n* and dimensions *p*, with value of $\alpha = 0.05$. The correction factor for $\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}$ is given as;

$$\vartheta^{*\alpha}_{m,n,p} \coloneqq \frac{1}{r(|\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}^{\boldsymbol{raw}}_{\boldsymbol{RMVV}}|)} \tag{6.5}$$

Next, we determine the RMVV location and scatter as follows,

$$\boldsymbol{m_{RMVV}} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x_i}}{m} \tag{6.6}$$

$$\boldsymbol{S_{RMVV}} = \vartheta^{*\alpha}_{m,n,p} c^*(m) \frac{\sum_{i=1}^{n} w_i(\boldsymbol{x_i} - \boldsymbol{m_{RMVV}})(\boldsymbol{x_i} - \boldsymbol{m_{RMVV}})^t}{m} \tag{6.7}$$

Finally, the squared robust Mahalanobis distances become

$$d_{i_{RMVV}}^{2}(\boldsymbol{x_i}, \boldsymbol{m_{RMVV}}, \boldsymbol{S_{RMVV}}) = (\boldsymbol{x_i} - \boldsymbol{m_{RMVV}})^t \boldsymbol{S}_{\boldsymbol{RMVV}}^{-1}(\boldsymbol{x_i} - \boldsymbol{m_{RMVV}}) \tag{6.8}$$

6.3.3 Algorithm

We now develop an algorithm to calculate an approximate RMVV solution, where the basis of our algorithm follows the generalization of MVV algorithm in Chapter 3 in Section 3.2. Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$ be a *p*-variate random sample of size *n*. We consider two typical choices of breakdown point (BP), namely BP=0.5 with $h = \lfloor (n + p + 1)/2 \rfloor$ and BP=0.25 with h = (0.75)*n*. For that reason, we use two different algorithms with different formula in determining the h subset. The difference between the two algorithms occurs in step-4 of Stage 1. Below is the complete algorithm to calculate an approximate RMVV solution.

Stage 1: Creating Initial Subsets.

This stage is repeated 500 times

1. Draw a random subset H_0 with number of observations, $h = p + 1$. Compute the mean vector $\boldsymbol{\bar{x}_{H_0}}$ and covariance matrix $\boldsymbol{S_{H_0}}$.

   $\boldsymbol{\bar{x}_{H_0}} = average(H_0)$ and $\boldsymbol{S_{H_0}} = cov(H_0)$

2. Compute the MSDs $d_0^2(i) = \left(\boldsymbol{x_i} - \boldsymbol{\bar{x}_{H_0}}\right)^t \boldsymbol{S_{H_0}}^{-1}\left(\boldsymbol{x_i} - \boldsymbol{\bar{x}_{H_0}}\right)$ for $i = 1, \dots, n$.

3. Sort these MSDs in ascending order, $d_0^2(\pi(1)) \leq d_0^2(\pi(2)) \leq \dots \leq d_0^2(\pi(n))$. This ordering defines a permutation π on the index set.
4. Take a new subset $H_1 = \{\pi(1), \dots, \pi(h)\}$ where

   i. $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ for BP=0.5 or

   ii. h = (0.75)n for BP=0.25

   then calculate $\bar{\boldsymbol{x}}_{\boldsymbol{H_1}}$, $\boldsymbol{S}_{\boldsymbol{H_1}}$, $Tr\left({\boldsymbol{S}_{\boldsymbol{H_1}}}^2\right)$ and compute MSD,

   $d_1^2(i) = (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_1}})^t {\boldsymbol{S}_{\boldsymbol{H_1}}}^{-1}\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_1}}\right)$ for $i = 1, \dots, n$.
5. Repeat step 3 and 4 for H_2
6. Sort the 500 values of $Tr\left({\boldsymbol{S}_{\boldsymbol{H_2}}}^2\right)$ in ascending order, then select 10 subsets of H_2 which have the lowest $Tr\left({\boldsymbol{S}_{\boldsymbol{H_2}}}^2\right)$. These subsets are treated as the initial subsets and their mean vectors, $\bar{\boldsymbol{X}}_{\boldsymbol{H_2}}$ and covariance matrices, $\boldsymbol{S}_{\boldsymbol{H_2}}$ will be used in Stage 2.

Stage 2: Concentration Steps (C-step)

This process will be repeated until convergence is achieved for each of the 10 subsets. Convergence means $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) = Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$, where k is the number of iterations.

1. Compute the MSDs by using $\bar{\boldsymbol{x}}_{\boldsymbol{H_2}}$ and $\boldsymbol{S}_{\boldsymbol{H_2}}$, where

   $d_2^2(i) = (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_2}})^t {\boldsymbol{S}_{\boldsymbol{H_2}}}^{-1}\left(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_{\boldsymbol{H_2}}\right)$ for $i = 1, \dots, n$.

2. Repeat step 3 and 4 in Stage 1 until $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) = Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$. If $Tr(\boldsymbol{S}_{\boldsymbol{k-1}}^2) > Tr(\boldsymbol{S}_{\boldsymbol{k}}^2)$ the process is continued. This process will be repeated until convergence is achieved.
3. When convergence is achieved for all the 10 subsets, choose the subset (*H**) that generates the lowest $Tr(\boldsymbol{S}_{\boldsymbol{H_k}}^2)$. From *H**, calculate $\overline{\boldsymbol{x}}_{\boldsymbol{H^*}} = \boldsymbol{m}_{\boldsymbol{MVV}}$ and $\boldsymbol{S}_{\boldsymbol{H^*}} = \boldsymbol{S}_{\boldsymbol{MVV}}$ as the location and scatter estimators for MVV respectively.

$$\boldsymbol{m}_{\boldsymbol{MVV}} = \frac{1}{h}\sum_{i=1}^{h} \boldsymbol{x}_{\boldsymbol{i}} \tag{6.9}$$

$$\boldsymbol{S}_{\boldsymbol{MVV}} = \frac{1}{h}\sum_{i=1}^{h}(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})^t \tag{6.10}$$

Reweighted Steps: Equations (6.9) and (6.10) respectively define the MVV estimates of location and scatter. By using these estimates,

4. Compute the Mahalanobis squared distances for all observations $\boldsymbol{x}_{\boldsymbol{1}}, \boldsymbol{x}_{\boldsymbol{2}}, \dots, \boldsymbol{x}_{\boldsymbol{n}}$ such that $d_{i_{MVV}}^2(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{m}_{\boldsymbol{MVV}}) = (\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})^t \boldsymbol{S}_{\boldsymbol{MVV}}^{-1}(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{m}_{\boldsymbol{MVV}})$ where *i*=1,2,...,*n*.
5. The raw RMVV estimators in Equation (6.2) and (6.3) are computed by giving weight $w_i = 0$ to observations with $d_{i_{MVV}}^2(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{m}_{\boldsymbol{MVV}}) > \chi_{p,0.025}^2$, and $w_i = 1$ otherwise.
6. Compute the consistency factor using Equation (6.4) for the raw RMVV covariance matrix.
7. Compute the correction factor for $\boldsymbol{c}^*(\boldsymbol{m})\boldsymbol{S}_{\boldsymbol{RMVV}}$ by using Equation (6.5).

8. Compute the reweighted MVV estimators of location and scatter using Equation (6.6) and (6.7) respectively.

6.4 Efficiency

To gain more insight in the RMVV estimators and observe how reweighting affects their performance, we compute the asymptotic relative efficiency where it may give some indication of how good the estimators are. We compute the asymptotic relative efficiency (ARE) based on Equation (6.1).

Table 6.3 shows the asymptotic relative efficiency (ARE) of the RMVV scatter estimators with different breakdown point of 0.25 and 0.5 denoted respectively as $\text{RMVV}_{0.25}$ and $\text{RMVV}_{0.5}$, with relative to the MVV estimator with breakdown point of 0.5 ($\text{MVV}_{0.5}$) at normal model computed using the following equation,

$$\text{ARE}(\boldsymbol{S}_{\boldsymbol{RMVV}}, \boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}(\boldsymbol{BP}=\boldsymbol{0.5})}) = \left(\boldsymbol{\vartheta}_{\boldsymbol{p,n}}^{\boldsymbol{\alpha}}\boldsymbol{c}(\boldsymbol{h})\boldsymbol{S}_{\boldsymbol{MVV}(\boldsymbol{BP}=\boldsymbol{0.5})}/\boldsymbol{S}_{\boldsymbol{RMVV}}\right)^{1/p} \quad (6.11)$$

Note that the ARE for $\text{RMVV}_{0.25}$ is less efficient than $\text{MVV}_{0.5}$ for all *p*’s. However, the ARE values improve as *p* increases. When the BP of RMVV estimator is increased to 0.5, we observe that the efficiency of the estimator increases considerably. For *p* = 2 and 10, the ARE’s are above 1 while for other dimensions, the values ranging from 0.9975 to 0.9987 are almost equal to 1. From Table 6.3 we can deduce that by reweighting MVV, we can achieve high efficiency while simultaneously maintain highest breakdown point. To show the effect of BP on efficiency, let us refer to Table 6.4. The first row records the efficiency of MVV estimator while the second row records the efficiency of the reweighted version of the estimator i.e. RMVV. Apparently, the MVV estimator is more efficient when BP = 0.25, however after reweighting the estimator (RMVV), the efficiency at BP = 0.5 improves and outdo when BP = 0.25.

Table 6.3: Asymptotic relative efficiency of the scatter matrix for RMVV estimator with different breakdown point (BP=0.25 and 0.5) w.r.t MVV estimator with (BP=0.5) for normal model.

	***p* = 2**	***p* = 5**	***p* = 10**	***p* = 15**	***p* = 20**
RMVV$_{0.25}$ w.r.t MVV$_{0.5}$	0.6984	0.7490	0.8012	0.8293	0.8489
RMVV$_{0.5}$ w.r.t MVV$_{0.5}$	1.0217	0.9984	1.0015	0.9975	0.9987

Table 6.4:Asymptotic relative efficiency of the scatter matrix for MVV and RMVV estimator with BP=0.25 with relative to MVV and RMVV estimator with BP=0.5 respectively.

	***p* = 2**	***p* = 5**	***p* = 10**	***p* = 15**	***p* = 20**
MVV$_{0.25}$ w.r.t MVV$_{0.5}$	1.4073	1.3225	1.2439	1.2024	1.1760
RMVV$_{0.25}$ w.r.t RMVV$_{0.5}$	0.9620	0.9922	0.9951	0.9997	0.9996

Thus, the RMVV$_{0.5}$ estimators possess both high efficiency and high breakdown point, hence, making these estimators more appealing. Nevertheless we should be aware that the gains in efficiency come at the price of a larger bias under contamination. The reason is that higher efficiency can only be obtained by increasing tuning parameters, which in turn affects the bias under contamination (Rousseeuw, 1994).

Our study then continued with the investigation on finite-sample robustness of RMVV estimators to support the above ARE results. For that purpose, a simulation study was conducted and discussed in the following section. Since this study focus on Hotelling T^2 for which the shift in the mean vector is of main concern, it would be more apt to focus on the RMVV location estimator.

6.5 Finite-Sample Robustness

To study on the finite-sample robustness of the RMVV location estimator, we performed simulations on contaminated data sets. In each simulation we generated L=1000 data sets of $N(0, I_p)$ with p = 2, 5, 10 and 20 representing small, medium and slightly high number of quality characteristics (dimensions) with reasonable values of sample sizes n = 50, 100, 200 and 500. Refer to section 3.4.2 of Chapter 3 for the generation of contaminated data sets.

To measure the robustness, we used the bias and the mean squared error (MSE) as suggested by Rousseeuw, van Driesen, van Aelst and Agullo (2004). For each simulation we compute the mean squared error and bias of the mean (location) vectors, $\hat{\boldsymbol{\mu}}^{\boldsymbol{l}}_{\boldsymbol{RMVV}}$, as in Roelant, van Aelst and William (2009),

$$MSE(\hat{\boldsymbol{\mu}}_{\boldsymbol{RMVV}}) = n\left[\frac{\sum_{j=1}^{p}\sum_{l=1}^{L}\left\{(\hat{\boldsymbol{\mu}}_{\boldsymbol{RMVV}})_j^{(l)}\right\}^2}{pL}\right] \tag{6.12}$$

$$bias(\hat{\boldsymbol{\mu}}_{\boldsymbol{RMVV}}) = \left[\frac{1}{p}\sum_{j=1}^{p}\left\{\frac{\sum_{l=1}^{L}(\hat{\boldsymbol{\mu}}_{\boldsymbol{RMVV}})_j^{(l)}}{L}\right\}^2\right]^{1/2} \tag{6.13}$$

where*l* = 1, …,*L*; *j* = 1, …, *p*

Tables 6.5, 6.6, 6.7 and 6.8 show the MSE and bias from mild, moderate and extreme contamination for RMVV$_{0.5}$, RMVV$_{0.25}$ and MVV$_{0.5}$ when p = 2, 5, 10 and 20 respectively. In general, across the type of contaminations, there is a diminution in the value of MSE when *p* increases except for moderate (Table 6.7) and extreme contamination (Table 6.8) when *p* = 20, *n* = 50. For most conditions, the RMVV$_{0.25}$ location estimator yields the lowest value of MSE, followed by RMVV$_{0.5}$ and then MVV$_{0.5}$. For larger sample sizes, the bias values for all estimators reduce closer to zero.

Although the RMVV$_{0.25}$ estimator produces the smallest MSE value, however if we scrutinize each table, we could observe inconsistency in the generation of the smallest bias values. As shown in Table 6.5, under mild contamination, RMVV$_{0.25}$ produce the smallest bias value when *p* = 2 and *n* = 50, but when *n* increases, RMVV$_{0.5}$ estimator outperforms RMVV$_{0.25}$ in the number of smallest bias values. Nonetheless, when *p* increases to 5 and 10, RMVV$_{0.25}$ reverts back to be the better performer. Meanwhile as p increases to 20, RMVV$_{0.5}$ produces the smallest bias except for n = 500.

For moderate contamination with mean shift 5 as shown in Table 6.6, RMVV$_{0.25}$ generates the smallest bias value for almost all combinations of p and n, except for small sample, n = 50. For the other moderate contamination (20% with mean shift 3) Table 6.7 shows that RMVV$_{0.5}$ is more dominant in generating the smallest value of bias especially when p = 20. Under the condition of extreme contamination as presented in Table 6.8, RMVV$_{0.25}$ outperforms RMVV$_{0.5}$ when p = 2 and 5, but when p increases to 10 and 20, RMVV$_{0.5}$ is better in generating small bias values.

Nevertheless, overall, RMVV$_{0.25}$ is the better performer as compared to RMVV$_{0.5}$ and MVV$_{0.5}$ because the estimator is not easily influenced by outliers (resulting in small MSE). Although the RMVV$_{0.25}$ has lower efficiency compared to RMVV$_{0.5}$ for normal data, but with regards to contamination, RMVV$_{0.25}$ on the whole is able to produce lower values of MSE and bias.

Table 6.5: Location estimator: 10% outliers with mean shift 3 (mild contamination)

n	50		100		200		500	
p= 2	MSE	bias	MSE	bias	MSE	bias	MSE	bias
$RMVV_{0.5}$	3.0032	0.0114	3.1907	0.0032	3.4245	0.0034	3.6031	0.0017
$RMVV_{0.25}$	1.7446	0.0066	1.6572	0.0067	1.6788	0.0052	1.6964	0.0029
$MVV_{0.5}$	3.4706	0.0080	4.0558	0.0047	4.5698	0.0043	5.0813	0.0018
p= 5								
$RMVV_{0.5}$	2.0571	0.0056	2.0776	0.0038	2.0825	0.0098	2.0380	0.0044
$RMVV_{0.25}$	1.4971	0.0066	1.4342	0.0037	1.4038	0.0078	1.3780	0.0031
$MVV_{0.5}$	2.1349	0.0127	2.4213	0.0092	2.7794	0.0114	3.1466	0.0048
p= 10								
$RMVV_{0.5}$	1.8022	0.0126	1.8966	0.0151	1.8359	0.0136	1.7480	0.0089
$RMVV_{0.25}$	1.4394	0.0092	1.3617	0.0120	1.3560	0.0117	1.3441	0.0079
$MVV_{0.5}$	1.8042	0.0107	1.9474	0.0148	2.0003	0.0154	2.4890	0.0105
p= 20								
$RMVV_{0.5}$	1.4990	0.0219	1.7181	0.0188	1.8215	0.0170	1.7075	0.0093
$RMVV_{0.25}$	1.4176	0.0226	1.3657	0.0192	1.3404	0.0173	1.3157	0.0087
$MVV_{0.5}$	1.4984	0.0222	1.7278	0.0196	2.0224	0.0175	2.1374	0.0118

Table 6.6: Location estimator: 10% outliers with mean shift 5 (moderate contamination)

n	50		100		200		500	
p= 2	MSE	bias	MSE	bias	MSE	bias	MSE	bias
RMVV$_{0.5}$	2.9739	0.0102	3.1905	0.0031	3.4192	0.0037	3.5983	0.0018
RMVV$_{0.25}$	1.7294	0.0024	1.6315	0.0033	1.6520	0.0017	1.6837	0.0009
MVV$_{0.5}$	3.4553	0.0072	4.0614	0.0044	4.5754	0.0043	5.0873	0.0019
p= 5								
RMVV$_{0.5}$	2.0643	0.0057	2.0525	0.0043	2.0807	0.0100	2.0394	0.0045
RMVV$_{0.25}$	1.4904	0.0065	1.4372	0.0042	1.4036	0.0077	1.3773	0.0031
MVV$_{0.5}$	2.1370	0.0133	2.4064	0.0073	2.7612	0.0114	2.9932	0.0047
p= 10								
RMVV$_{0.5}$	1.7936	0.0117	1.8934	0.0157	1.8708	0.0121	1.7473	0.0090
RMVV$_{0.25}$	1.4463	0.0120	1.3674	0.0130	1.3545	0.0112	1.3464	0.0080
MVV$_{0.5}$	1.7902	0.0107	1.9474	0.0148	2.1854	0.0129	2.4737	0.0117
p= 20								
RMVV$_{0.5}$	1.5196	0.0193	1.7361	0.0190	1.8213	0.0153	1.7012	0.0098
RMVV$_{0.25}$	1.4310	0.0219	1.3555	0.0186	1.3334	0.0175	1.3161	0.0090
MVV$_{0.5}$	1.5189	0.0195	1.7487	0.0189	2.0339	0.0158	2.1368	0.0118

Table 6.7: Location estimator: 20% outliers with mean shift 3 (moderate contamination)

n	50		100		200		500	
p= 2	MSE	bias	MSE	bias	MSE	bias	MSE	bias
$\boldsymbol{RMVV_{0.5}}$	2.7968	0.0109	2.9460	0.0034	3.1887	0.0043	3.1918	0.0037
$\boldsymbol{RMVV_{0.25}}$	1.6785	0.0168	1.6481	0.0188	1.7205	0.0177	1.7133	0.0145
$\boldsymbol{MVV_{0.5}}$	3.1571	0.0142	3.6666	0.0011	4.2234	0.0043	4.4614	0.0044
p= 5								
$\boldsymbol{RMVV_{0.5}}$	2.0241	0.0111	2.0430	0.0033	2.0979	0.0110	2.0288	0.0053
$\boldsymbol{RMVV_{0.25}}$	1.4396	0.0058	1.3793	0.0030	1.3890	0.0071	1.4244	0.0036
$\boldsymbol{MVV_{0.5}}$	2.0840	0.0109	2.3109	0.0060	2.6152	0.0143	2.8112	0.0147
p= 10								
$\boldsymbol{RMVV_{0.5}}$	1.7607	0.0128	1.9035	0.0154	1.9249	0.0103	1.8271	0.0093
$\boldsymbol{RMVV_{0.25}}$	1.4070	0.0114	1.3522	0.0132	1.3428	0.0111	1.4040	0.0095
$\boldsymbol{MVV_{0.5}}$	1.7682	0.0140	1.9238	0.0166	2.1292	0.0120	2.4336	0.0129
p= 20								
$\boldsymbol{RMVV_{0.5}}$	1.9817	0.3067	1.7059	0.0183	1.8512	0.0154	1.7339	0.0102
$\boldsymbol{RMVV_{0.25}}$	2.1080	0.3707	1.5289	0.0224	1.6660	0.0257	2.1956	0.0154
$\boldsymbol{MVV_{0.5}}$	1.9839	0.3072	1.7151	0.0184	2.0105	0.0158	2.1117	0.0103

Table 6.8: Location estimator: 20% outliers with mean shift 5 (extreme contamination)

n	50		100		200		500	
p= 2	MSE	bias	MSE	bias	MSE	bias	MSE	bias
RMVV$_{0.5}$	2.7651	0.0056	2.9251	0.0017	3.1455	0.0046	3.1678	0.0037
RMVV$_{0.25}$	1.5730	0.0021	1.5113	0.0023	1.5761	0.0017	1.5600	0.0022
MVV$_{0.5}$	3.1161	0.0110	3.6765	0.0019	3.8672	0.0049	4.4345	0.0038
p= 5								
RMVV$_{0.5}$	2.0273	0.0129	2.0449	0.0026	2.0951	0.0108	2.1688	0.0037
RMVV$_{0.25}$	1.4407	0.0060	1.3797	0.0032	1.3915	0.0071	1.4221	0.0037
MVV$_{0.5}$	2.0922	0.0117	2.3082	0.0048	2.5998	0.0143	3.2375	0.0039
p= 10								
RMVV$_{0.5}$	1.7522	0.0104	1.9073	0.0151	1.9257	0.0108	1.8272	0.0093
RMVV$_{0.25}$	1.4154	0.0108	1.3397	0.0153	1.3439	0.0111	1.2739	0.0093
MVV$_{0.5}$	1.7605	0.0116	1.9352	0.0149	2.2967	0.0137	2.4251	0.0132
p= 20								
RMVV$_{0.5}$	2.6433	0.4052	1.7082	0.0189	1.8547	0.0155	1.7954	0.0103
RMVV$_{0.25}$	2.9294	0.4261	1.8758	0.0633	1.8166	0.0199	1.6256	0.0084
MVV$_{0.5}$	2.6465	0.4055	1.7144	0.0191	2.2006	0.0158	2.1027	0.0116

6.6 Discussion

The result of the investigation on the statistical efficiency of MVV estimators for different breakdown point showed that the conflict between efficiency and breakdown point occurred in MVV estimators. Hence, to maintain the highest breakdown value and simultaneously achieving high efficiency, this study developed a one-step reweighted version of minimum vector variance estimator (RMVV). The development and availability of fast algorithm for computing the RMVV has brought renewed interest to this estimator. The finding proved that reweighting leads to an important gain in efficiency at the same time maintaining the highest breakdown value. Thus, the RMVV$_{0.5}$ estimator possesses both high efficiency and high breakdown point, making these estimators more appealing. However, if the data is suspected to be contaminated by outliers, we recommend using RMVV$_{0.25}$ estimators because it has a smaller MSE and bias.

CHAPTER SEVEN
ROBUST HOTELLING T^2 CHART BASED ON REWEIGHTED VERSION OF MVV ESTIMATOR

7.1 Introduction

In the previous chapter, we have introduced the reweighted version of MVV known as RMVV. Through simulation study, we have shown that these estimators are consistent, unbiased and attained high efficiency. In this chapter, we will investigate on the performance of these estimators in constructing the Hotelling T^2 chart. Since efficiency and breakdown point are inversely related (Croux & Haesbroeck, 1999), and the efficiency value changes with respect to breakdown point, thus this chapter will demonstrate on the construction of the robust Hotelling T^2 chart using the RMVV estimators taking into consideration the two breakdown points used in the previous chapter namely 0.5 and 0.25. The respective robust Hotelling T^2 charts are denoted as $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$. The investigation will then continue with the comparison of $T^2_{RMVV_{0.5}}$, $T^2_{RMVV_{0.25}}$, Hotelling T^2 control charts with MVV (T^2_{MVV}), the improved T^2_{MVV} ($T^2_{MVV(I)}$) which was proposed in Chapter 5, MCD estimators (T^2_{MCD}) and Reweighted MCD estimators with breakdown point 0.25 (T^2_{RMCD}).

The outline of this chapter is as follows. In Section 7.2, we formally introduce two robust control charts with different breakdown point based on RMVV estimators. The approximations of the control limit for $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ are discussed in

Section 7.3 followed by the simulation study on the performance of $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ with contaminated data in Section 7.4. The following Section 7.5 gives an example on real sample data, and finally, Section 7.6 concludes the chapter.

7.2 RMVV Hotelling T^2 Control Chart

Suppose that $\boldsymbol{x_i} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$ is the *p*-variate random sample of *n* observations of preliminary data set in Phase I such that $\boldsymbol{x_i}$ are independent with unknown $\boldsymbol{\mu}$ and Σ. To estimate the in-control parameters, we need to use in-control data set. The process of identifying the in-control data set from $\boldsymbol{x_i}$ is referred to as Phase I process. Since the RMVV estimators are known to be free from outliers due to its estimation process, they could be readily used as in-control estimators in Phase II process where the phase II observations are $\boldsymbol{x_g} = \{\boldsymbol{x_{n+1}}, \boldsymbol{x_{n+2}}, \dots\}$, $\boldsymbol{x_g} \notin \boldsymbol{x_i}$.

From the preliminary data set, $\boldsymbol{x_i}$, the RMVV mean vector and covariance matrix estimators ($\boldsymbol{m_{RMVV}}$ and $\boldsymbol{S_{RMVV}}$) are determined by using Equations (6.6) and (6.7) presented in Chapter 6. Since we are investigating RMVV Hotelling T^2 chart with two typical choices of breakdown point, namely BP=0.5 with $h = \lfloor (n + p + 1)/2 \rfloor$ and BP=0.25 with h = (0.75)*n*, thus, we need to calculate RMVV estimators using two different algorithms with different formula in determining the h subset. The algorithm was discussed in Section 6.3.3. Subsequently, by using these two types of

RMVV estimators, we define the two robust Hotelling's T^2 for Phase II data, $\boldsymbol{x_g}$, as follows,

$$T^2_{RMVV_{0.5}}(g) = (\boldsymbol{x_g} - \boldsymbol{m_{RMVV_{0.5}}})\boldsymbol{S^{-1}_{RMVV_{0.5}}}(\boldsymbol{x_g} - \boldsymbol{m_{RMVV_{0.5}}})^t \qquad (7.1)$$

$$T^2_{RMVV_{0.25}}(g) = (\boldsymbol{x_g} - \boldsymbol{m_{RMVV_{0.25}}})\boldsymbol{S^{-1}_{RMVV_{0.25}}}(\boldsymbol{x_g} - \boldsymbol{m_{RMVV_{0.25}}})^t \qquad (7.2)$$

7.3 Estimation of Control Limits for RMVV Hotelling T^2 Control Chart

To demonstrate the performance of $T^2_{RMVV_{0.5}}$ in Equation 7.1 and $T^2_{RMVV_{0.25}}$ in Equation 7.2, we need a better understanding about its distribution to obtain appropriate control limits (UCL). Since the exact distributions of the finite sample for RMVV estimators are unknown, we approximate control limit by adopting the same Monte Carlo simulation method used for the construction of the T^2_{MVV} control limit as discussed in Section 3.3. The quantiles of the $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ are estimated for several combinations of sample sizes and dimensions. In order to estimate the 95% quantile of $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ for a given sample size *n* and dimension *p* in Phase I, we first generate *K* = 5000 samples of size *n* from a standard multivariate normal distribution $MVN_p(0, \boldsymbol{I_p})$. Then, for each data set of size *n*, we compute RMVV mean vector and covariance matrix estimates which respectively are denoted as $\boldsymbol{m_{RMVV}}(k)$ and $\boldsymbol{m_{RMVV}}(k)$ where $k = 1, \dots, K$. In addition, for each data set, we randomly generate a new observation $\boldsymbol{x_{g,k}}$ treated as a Phase II observation from $MVN_p(0, \boldsymbol{I_p})$ and calculate the corresponding $T^2_{RMVV_{0.5}}(g,k)$ and $T^2_{RMVV_{0.25}}(g,k)$ values as given by Equation 7.1 and 7.2. The empirical distribution functions of $T^2_{RMVV_{0.5}}(g,k)$ and $T^2_{RMVV_{0.25}}(g,k)$ are based on the simulated values

$$T^2_{RMVV}(g,1), T^2_{RMVV}(g,2), \dots, T^2_{RMVV}(g,K) \tag{7.3}$$

Next, we sort $T^2_{RMVV_{0.5}}(g,k)$ and $T^2_{RMVV_{0.25}}(g,k)$ values in ascending order and the UCL is the 95% quantile of the 5000 statistics.

7.4 Performance Evaluation.

The performance of the $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ control charts were put to test under various conditions. For such purpose, a simulation study had been designed to encompass several different scenarios (conditions), which were assumed to influence the strength and weaknesses of the proposed control charts. These various conditions were created by manipulating the number of observations (*n*), the number of dimensions (*p*), and the level of contamination by using different proportion of outliers (ε) and several mean shifts values ($\boldsymbol{\mu_1}$) as discussed in Section 3.4 of Chapter 3.

The performance of $T^2_{RMVV_{0.5}}(g,k)$ and $T^2_{RMVV_{0.25}}(g,k)$ charts was judged based on the false alarm rate and probability of detection of the process of the Phase II data. Finally, the new charts were compared with the existing robust Hotelling T^2 charts using MVV (T^2_{MVV}), the improved T^2_{MVV} ($T^2_{MVV(I)}$), MCD (T^2_{MCD}) and RMCD (T^2_{RMCD}). Each of these charts was tested on 5 types of contaminations on 23 combinations of *n* and *p* which totaled up to 115 conditions. For each condition, the false alarm rates and probability of detection were determined. The programs and simulations were run using MATLAB 7.8.0 (R2009a). The algorithm of MVV and RMVV were executed using the MATLAB 7.8.0 (R2009a), while Fast MCD algorithm to compute MCD and RMCD estimators used *mcdcov.m* in the LIBRA package under MATLAB 7.8.0 (R2009a). The computation of the RMCD estimator using *mcdcov.m* algorithm was based on breakdown point 0.25 as discussed in Section 2.3.5 of Chapter 2.

7.5 Simulation Results

Since the performance of false alarm rate and probability of detection for T^2_{MVV} and $T^2_{MVV(I)}$ are the same, and only differ in the control limits (UCL's) as discussed in the result of Chapter 5, hereafter, T^2_{MVV} will represent both robust Hotelling T^2 chart using MVV(T^2_{MVV}) and improved T^2_{MVV} ($T^2_{MVV(I)}$). Basically, this section compares the performance of T^2_{MCD}, T^2_{MVV}, T^2_{RMCD}, $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ control chart in terms of probability of detection and false alarm rate. The presentation of the results for the probability of detection and false alarm rate in this chapter differs from the presentation in Chapter 4 because of the increase in the number of robust methods. The results of the investigation are presented in tables for the probability of detection. For ease of comparison, we will refer to the tables and graphical presentation for the false alarm rates.

7.5.1 Probability of Detecting Outliers

Tables 7.1-7.5 which recorded the probability of detection for each condition are arranged based on the ascending number of dimensions (variables) namely p = 2, 5, 10, 15 and 20 with α = 0.05. The first column in each table displays the number of sample sizes, followed by the percentage of outliers and non-centrality values (mean shifts) respectively in the second and third column. The last four columns record the probability of detection of the control charts investigated in this study namely $T^2_{MCD}, T^2_{MVV},\ T^2_{RMCD}, T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$.

For each condition, the performance of the control chart is regarded as better in detecting changes when the value of the probability is closer to 1. Table 7.1 presents the probability of detection for the bivariate case (p = 2). Under most conditions, $T^2_{RMVV_{0.25}}$ shows better detection than other charts, especially when the percentage of outliers is large (20%).

Table 7.1: Probability of detection for the corresponding control charts with dimension, p = 2

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
10	10%	3	0.5160	0.5320	0.5120	0.5080	0.5940
25			0.7840	0.8320	0.8420	0.7730	0.8270
50			0.8840	0.8930	0.9210	0.8590	0.9180
100			0.9010	0.9190	0.9290	0.9040	0.9370
200			0.9350	0.9460	0.9530	0.9460	0.9580
500			0.9370	0.9520	0.9510	0.9620	0.9580
10	10%	5	0.9310	0.9080	0.9240	0.8800	0.9380
25			0.7840	0.9980	1.0000	0.9930	0.9990
50			0.8840	1.0000	1.0000	1.0000	1.0000
100			0.9010	1.0000	1.0000	1.0000	1.0000
200			0.9350	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
10	20%	3	0.3240	0.4270	0.3550	0.4070	0.4870
25			0.5420	0.7210	0.6290	0.6900	0.7400
50			0.6950	0.8280	0.7620	0.8010	0.8850
100			0.7260	0.8890	0.7710	0.8900	0.9150
200			0.8030	0.9140	0.8200	0.9390	0.9430
500			0.8190	0.9310	0.8320	0.9490	0.9510
10	20%	5	0.7960	0.8530	0.8590	0.8340	0.9120
25			0.9790	0.9960	1.0000	0.9900	0.9980
50			1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	0.9990	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000

Table 7.2: Probability of detection for the corresponding control charts with dimension, p = 5

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
30	10%	3	0.9610	0.9770	0.9690	0.9710	0.9900
50			0.9860	0.9910	0.9970	0.9860	0.9990
100			0.9980	1.0000	1.0000	0.9970	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
30	10%	5	1.0000	1.0000	1.0000	1.0000	1.0000
50			1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
30	20%	3	0.7530	0.9650	0.8350	0.9600	0.9840
50			0.9530	0.9890	0.9800	0.9850	0.9980
100			0.9870	0.9970	0.9990	0.9970	1.0000
200			0.9960	0.9990	1.0000	1.0000	1.0000
500			0.9980	1.0000	1.0000	1.0000	1.0000
30	20%	5	1.0000	1.0000	1.0000	1.0000	1.0000
50			1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000

Table 7.3: Probability of detection for the corresponding control charts with dimension, p = 10

Sample Size (n)	% outliers (ε)	Mean shift (μ$_1$)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
50	10%	3	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
50	10%	5	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
50	20%	3	0.8840	0.9990	0.8870	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
50	20%	5	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000

Table 7.4: Probability of detection for the corresponding control charts with dimension, p = 15

Sample Size (n)	% outliers (ε)	Mean shift (μ$_1$)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
80	10%	3	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
80	10%	5	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
80	20%	3	0.9310	1.0000	0.9930	1.0000	1.0000
100			0.9840	1.0000	0.9840	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
80	20%	5	1.0000	1.0000	1.0000	1.0000	1.0000
100			1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000

Table 7.5: Probability of detection for the corresponding control charts with dimension, p = 20

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
100	10%	3	1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
300			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
100	10%	5	1.0000	1.0000	1.0000	1.0000	1.0000
200			1.0000	1.0000	1.0000	1.0000	1.0000
300			1.0000	1.0000	1.0000	1.0000	1.0000
500			1.0000	1.0000	1.0000	1.0000	1.0000
100	20%	3	0.8760	0.9970	0.8760	0.9990	0.9470
200			0.9510	1.0000	0.9000	1.0000	0.9840
300			0.9990	1.0000	0.9970	1.0000	0.9890
500			1.0000	1.0000	1.0000	1.0000	0.9920
100	20%	5	0.9970	0.9980	0.9970	0.9980	0.9670
200			0.9950	1.0000	0.9500	1.0000	0.9940
300			1.0000	1.0000	1.0000	1.0000	0.9990
500			1.0000	1.0000	1.0000	1.0000	0.9990

As we go across Tables 7.2, 7.3 and 7.4 which represent *p* = 5, 10 and 15 respectively, we could observe that $T^2_{RMVV_{0.25}}$ always record the highest percentage of detection with most of the values achieving the perfect 100% detection. However, as *p* reaches 20, the chart's performance slightly drops. However, this situation only occurs when the percentage of outliers is large as exhibited in Table 7.5. At this point, better ability of detection could be observed from T^2_{MVV} and $T^2_{RMVV_{0.5}}$.

7.5.2 False Alarm Rates

Tables 7.6 – 7.10 record the false alarm rates for all charts. We will alternately refer to the visual presentation in Figure 7.1 to 7.5 and the numerical values in Table 7.6 – 7.10. Each figure represents different type of contaminated distributions as discussed in Chapter 3 in Section 3.4.2, categorized as ideal, mildly contaminated, moderately contaminated and extremely contaminated. For each condition, the performance of the control chart is regarded as better in controlling false alarm rates when the empirical rate is closer to the nominal value, α = 0.05. In the tables, the values that are closest to the nominal value but not less than 0.025 and exceeding 0.055 are highlighted. This value was chosen based on Bradley's interval of robustness (discussed in Section 4.2.3 in Chapter 4).

Under bivariate case (p = 2) as presented in Table 7.6 and Figure 7.1, overall, T^2_{MVV} shows better performance in controlling false alarm rate since it has the highest number of highlighted values followed by $T^2_{RMVV_{0.5}}$. T^2_{MVV} shows better control of false alarm rate for ideal, mildly contaminated, and moderately (10% with mean shift 5) contaminated distributions. However, under moderate (20% with mean shift 3) and extreme contamination, the false alarm rates fall below the 0.025 level. The result implies that under bivariate case, T^2_{MVV} perform poorly when the percentage of contamination is high.

Table 7.7 and Figure 7.2 exhibits the false alarm rates when p = 5. For this case in general, $T^2_{RMVV_{0.25}}$ chart has better ability in controlling false alarm rates. Trailing behind is $T^2_{RMVV_{0.5}}$. Under small and moderate sample sizes, $T^2_{RMVV_{0.25}}$ surpasses the performance of the other charts. However, for the largest sample size i.e. 500, $T^2_{RMVV_{0.5}}$ is able to control false alarm rates better than $T^2_{RMVV_{0.25}}$.

The performance of the charts in terms of false alarm rates for the case of p = 10 is displayed in Table 7.8 and Figure 7.3. The overall results on false alarm rates show that $T^2_{RMVV_{0.5}}$ clearly outperforms the other control charts, except for ideal condition where the rate seem to be deviating from the nominal value for n = 100 and 500. Meanwhile, when the sample size is very small under 10% contamination with 3 and 5 shifts in the mean vector, the $T^2_{RMVV_{0.25}}$ performs better.

Under the case of p = 15, as can be clearly observed in Table 7.9 and Figure 7.4, the performance of the robust $T^2_{RMVV_{0.5}}$ chart is also much better than the other charts especially for moderate contamination. Nonetheless under ideal condition, the rate of $T^2_{RMVV_{0.5}}$ slightly diverges from the nominal value, while T^2_{MVV} and $T^2_{RMVV_{0.25}}$ seem to have better control of false alarm rate. For 10% contamination, T^2_{MVV} and $T^2_{RMVV_{0.25}}$

display relatively good performance, but their performance declines when the contamination increases to 20%.

The evaluation on false alarm rate for p = 20 in Table 7.10 clearly shows that in general, $T^2_{RMVV_{0.5}}$ has more ability in controlling the rate. Under ideal condition, T^2_{MCD} appears to have better control of false alarm, but for other conditions, $T^2_{RMVV_{0.5}}$ seems to outperform all the charts. Even T^2_{RMCD} could not compete well with any of the charts using MVV, be it T^2_{MVV}, $T^2_{RMVV_{0.5}}$ or $T^2_{RMVV_{0.25}}$. Both charts using MCD namely T^2_{RMCD} and T^2_{MCD} produce false alarm rates far below the nominal level for all cases.

Table 7.6: False alarm rate for the corresponding control charts with dimension, p = 2

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			$\boldsymbol{T^2_{MCD}}$	$\boldsymbol{T^2_{MVV}}$	$\boldsymbol{T^2_{RMCD}}$	$\boldsymbol{T^2_{RMVV_{0.5}}}$	$\boldsymbol{T^2_{RMVV_{0.25}}}$
10	0%	0	0.0520	0.0520	0.0510	0.0510	0.0600
25			0.0480	0.0530	0.0470	0.0490	0.0480
50			0.0580	0.0540	0.0570	0.0530	0.0490
100			0.0460	0.0490	0.0430	0.0460	0.0450
200			0.0600	0.0690	0.0450	0.0700	0.0600
500			0.0520	0.0630	0.0470	0.0590	0.0520
10	10%	3	0.0290	0.0450	0.0310	0.0450	0.0370
25			0.0280	0.0390	0.0280	0.0390	0.0340
50			0.0230	0.0350	0.0350	0.0450	0.0360
100			0.0200	0.0300	0.0340	0.0330	0.0320
200			0.0310	0.0490	0.036	0.0580	0.0540
500			0.0270	0.0490	0.0360	0.0470	0.0460
10	10%	5	0.0250	0.0450	0.0260	0.0450	0.0370
25			0.0290	0.0390	0.0280	0.0390	0.0340
50			0.0230	0.0340	0.0360	0.0430	0.0360
100			0.0200	0.0290	0.0350	0.0330	0.0320
200			0.0310	0.0500	0.0350	0.0590	0.0540
500			0.0260	0.0480	0.0360	0.0470	0.0460
10	20%	3	0.0210	0.0330	0.0230	0.0330	0.0250
25			0.0090	0.0190	0.0200	0.0200	0.0150
50			0.0080	0.0180	0.0230	0.0240	0.0200
100			0.0050	0.0150	0.0190	0.0220	0.0240
200			0.0050	0.0280	0.0190	0.0370	0.0400
500			0.0040	0.0230	0.0180	0.0310	0.0360
10	20%	5	0.0110	0.0330	0.0110	0.0320	0.0250
25			0.0050	0.0190	0.0170	0.0200	0.0150
50			0.0060	0.0170	0.0300	0.0240	0.0200
100			0.0040	0.0150	0.0270	0.0230	0.0230
200			0.0020	0.0280	0.0300	0.0360	0.0410
500			0.0040	0.0230	0.0340	0.0310	0.0380
Total highlighted			0	12	7	9	4

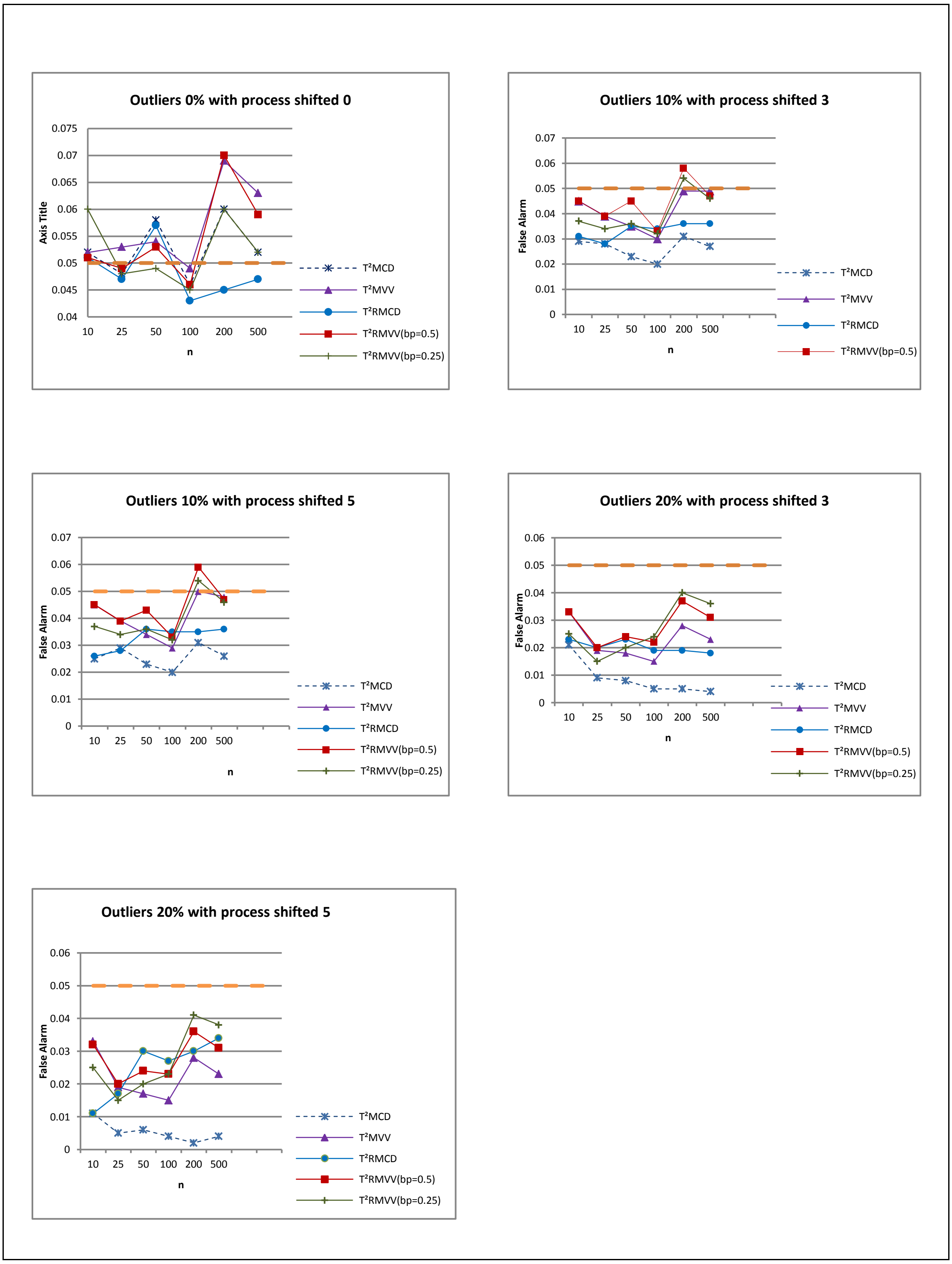


Figure 7.1: False alarm when p=2.

Table 7.7: False alarm rate for the corresponding control charts with dimension, $p = 5$

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
30	0%	0	0.0430	0.0500	0.0430	0.0500	0.0440
50			0.0650	0.0490	0.0620	0.0540	0.0640
100			0.0320	0.0380	0.0400	0.0400	0.0400
200			0.0410	0.0390	0.0450	0.0420	0.0460
500			0.0420	0.0430	0.0430	0.0480	0.0440
30	10%	3	0.0100	0.0300	0.0120	0.0290	0.0330
50			0.0130	0.0350	0.0270	0.0400	0.0410
100			0.0140	0.0300	0.0320	0.0350	0.0370
200			0.0200	0.0350	0.0380	0.0390	0.0460
500			0.0160	0.0360	0.0350	0.0460	0.0350
30	10%	5	0.0100	0.0330	0.0110	0.0340	0.0300
50			0.0130	0.0370	0.0270	0.0420	0.0420
100			0.0140	0.0320	0.0320	0.0340	0.0370
200			0.0200	0.0350	0.0380	0.0390	0.0460
500			0.0170	0.0370	0.0350	0.0460	0.0340
30	20%	3	0.0050	0.0210	0.0070	0.0200	0.0200
50			0.0040	0.0220	0.0160	0.0270	0.0260
100			0.0020	0.0170	0.0280	0.0270	0.0320
200			0.0010	0.0220	0.0340	0.0330	0.0420
500			0.0030	0.0190	0.0330	0.0370	0.0320
30	20%	5	0.0000	0.0200	0.0020	0.0190	0.0200
50			0.0020	0.0230	0.0130	0.0270	0.0240
100			0.0020	0.0190	0.0280	0.0290	0.0320
200			0.0010	0.0220	0.0340	0.0330	0.0420
500			0.0030	0.0190	0.0320	0.0370	0.0320
Total highlighted			0	2	1	11	13

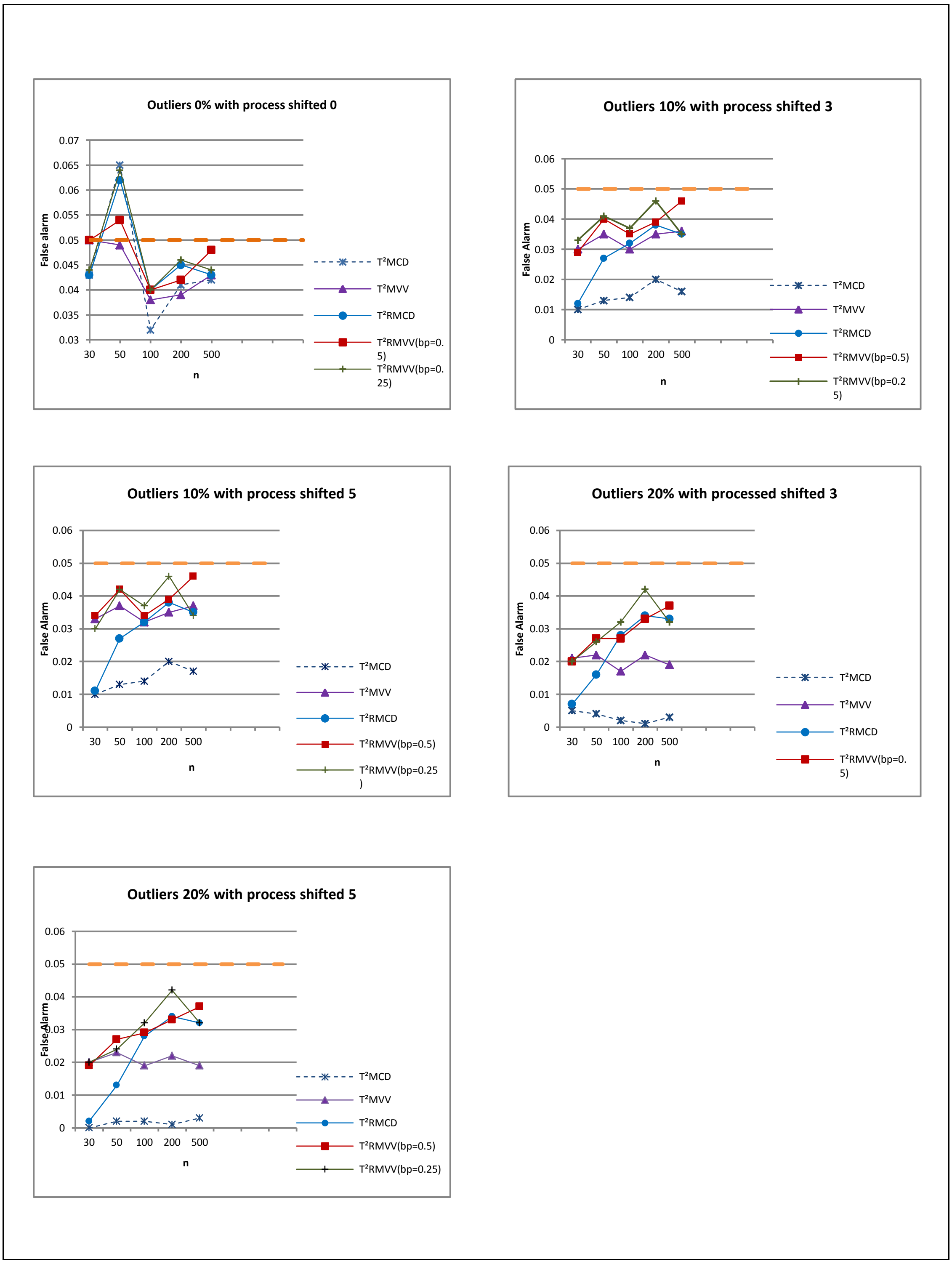


Figure 7.2: False alarm when p=5

Table 7.8: False alarm rate for the corresponding control charts with dimension, p = 10

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
50	0%	0	0.0530	0.0520	0.0570	0.0490	0.0570
100			0.0420	0.0450	0.0490	0.0440	0.0490
200			0.0540	0.0520	0.0440	0.0490	0.0590
500			0.0490	0.0540	0.0480	0.0550	0.0510
50	10%	3	0.0210	0.0370	0.0230	0.0360	0.0470
100			0.0190	0.0390	0.0400	0.0450	0.0340
200			0.0200	0.0390	0.0440	0.0480	0.0510
500			0.0220	0.0390	0.0450	0.0490	0.0530
50	10%	5	0.0210	0.0380	0.0230	0.0370	0.0390
100			0.0200	0.0350	0.0390	0.0440	0.0340
200			0.0200	0.0420	0.0440	0.0480	0.0510
500			0.0230	0.0390	0.0450	0.0510	0.0540
50	20%	3	0.0080	0.0250	0.0100	0.0280	0.0230
100			0.0030	0.0240	0.0260	0.0280	0.0260
200			0.0020	0.0250	0.0280	0.0430	0.0340
500			0.0040	0.0260	0.0420	0.0450	0.0560
50	20%	5	0.0020	0.0220	0.0040	0.0230	0.0230
100			0.0030	0.0230	0.0260	0.0260	0.0260
200			0.0020	0.0240	0.0280	0.0450	0.0340
500			0.0040	0.0230	0.0420	0.0450	0.0560
Total highlighted			1	0	2	13	6

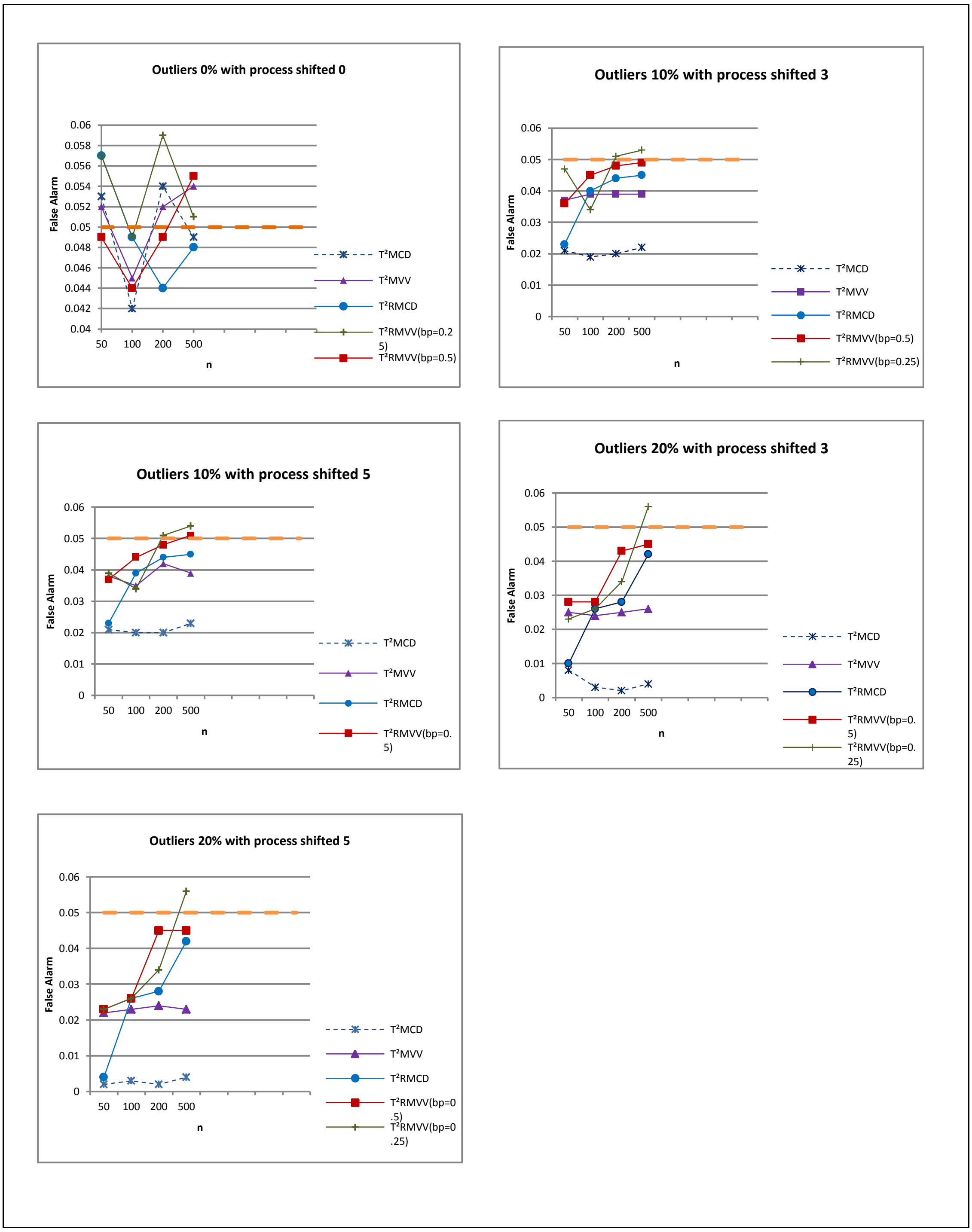


Figure 7.3: False alarm when p=10

Table 7.9: False alarm rate for the corresponding control charts with dimension, p = 15

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
80	0%	0	0.0580	0.0560	0.0580	0.0570	0.0600
100			0.0490	0.0520	0.0520	0.0490	0.0490
200			0.0540	0.0470	0.0540	0.0450	0.0610
500			0.0460	0.0470	0.0510	0.0480	0.0510
80	10%	3	0.0260	0.0470	0.0330	0.0440	0.0430
100			0.0240	0.0450	0.0400	0.0350	0.0390
200			0.0330	0.0420	0.0460	0.0530	0.0470
500			0.0270	0.0390	0.0450	0.0500	0.0530
80	10%	5	0.0230	0.0430	0.0300	0.0450	0.0390
100			0.0240	0.0430	0.0380	0.0380	0.0430
200			0.0310	0.0410	0.0460	0.0520	0.0460
500			0.0260	0.0390	0.0450	0.0460	0.0520
80	20%	3	0.0060	0.0270	0.0120	0.0290	0.0270
100			0.0030	0.0250	0.0150	0.0330	0.0240
200			0.0040	0.0200	0.0350	0.0430	0.0330
500			0.0060	0.0260	0.0430	0.0500	0.0570
80	20%	5	0.0030	0.0320	0.0100	0.0290	0.0290
100			0.0020	0.0220	0.0190	0.0270	0.0220
200			0.0040	0.0240	0.0340	0.0460	0.0340
500			0.0060	0.0290	0.0430	0.0490	0.0560
Total highlighted			1	4	1	12	4

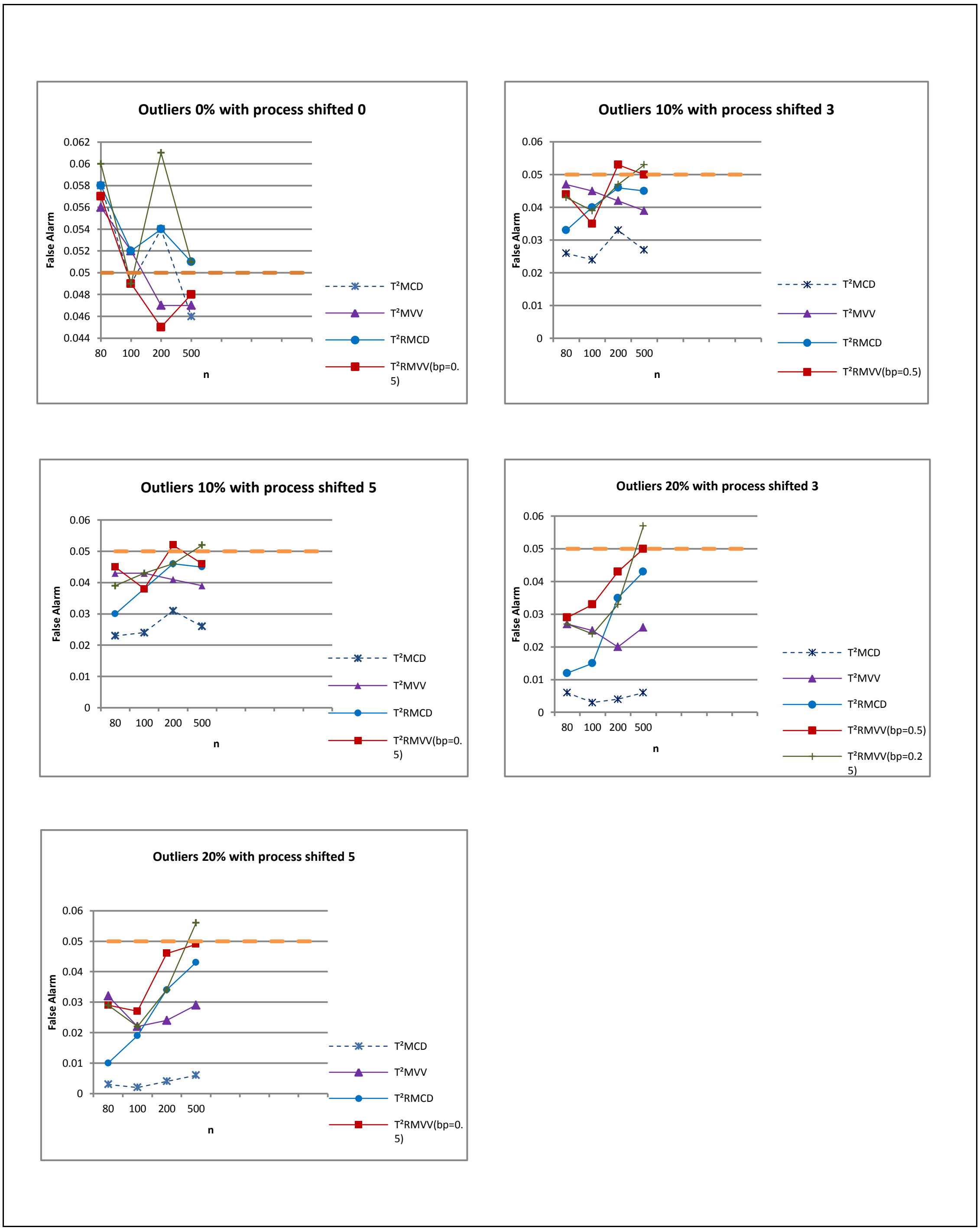


Figure 7.4: False alarm when p=15

Table 7.10: False alarm rate for the corresponding control charts with dimension, p = 20

Sample Size (n)	% outliers (ε)	Mean shift (μ_1)	Control Charts				
			T^2_{MCD}	T^2_{MVV}	T^2_{RMCD}	$T^2_{RMVV_{0.5}}$	$T^2_{RMVV_{0.25}}$
100	0%	0	0.0500	0.0530	0.0510	0.0530	0.0490
200			0.0490	0.0510	0.0490	0.0470	0.0520
300			0.0390	0.0440	0.0440	0.0440	0.0460
500			0.0560	0.0530	0.0530	0.0490	0.0510
100	10%	3	0.0290	0.0450	0.0310	0.0450	0.0370
200			0.0280	0.0390	0.0280	0.0390	0.0340
300			0.0230	0.0350	0.0350	0.0450	0.0360
500			0.0200	0.0300	0.0340	0.0330	0.0320
100	10%	5	0.0240	0.0420	0.0340	0.0420	0.0390
200			0.0220	0.0380	0.0350	0.0420	0.0410
300			0.0210	0.0340	0.0360	0.0390	0.0400
500			0.0280	0.0400	0.0460	0.0490	0.0560
100	20%	3	0.0060	0.0300	0.0100	0.0330	0.0270
200			0.0050	0.0310	0.0310	0.0370	0.0260
300			0.0060	0.0240	0.0360	0.0390	0.0320
500			0.0030	0.0350	0.0440	0.0560	0.0520
100	20%	5	0.0000	0.0240	0.0050	0.0270	0.0230
200			0.0020	0.0250	0.0280	0.0300	0.0210
300			0.0050	0.0220	0.0360	0.0360	0.0310
500			0.0040	0.0320	0.0040	0.0470	0.054
Total highlighted			2	3	3	14	3

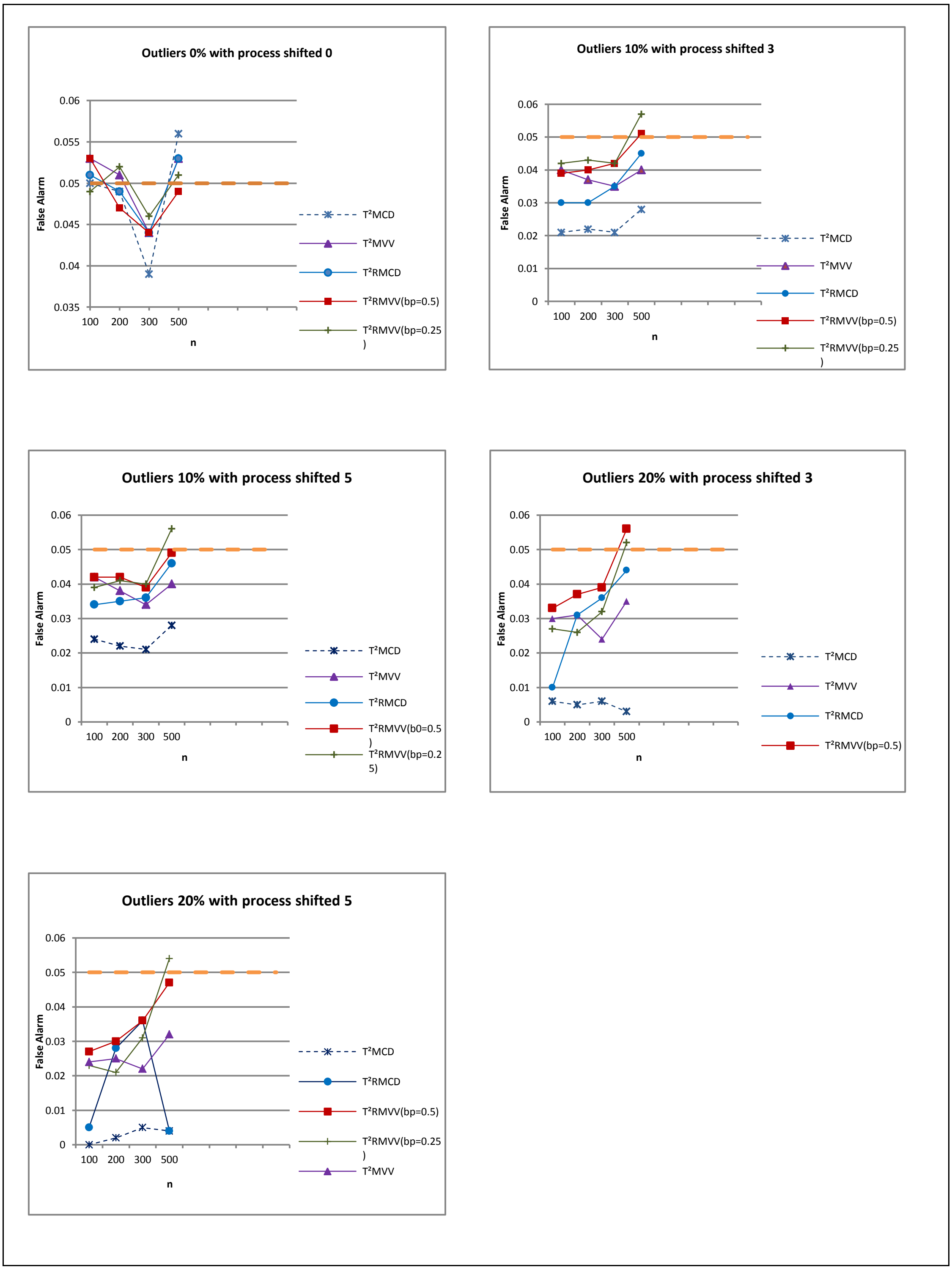


Figure 7.5: False alarm when p=20

7.6 Real Data Analysis

The investigation of $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ continues with the application on real data. Real data were furnished to us by Asian Composite Manufacturing Sdn. Bhd. (ACM) as discussed in Chapter 4 in Section 4.3. The historical data set (Phase I) and the future data set (Phase II) are shown in Table 4.7 and 4.9 respectively. The product consisted of 3 quality variables (dimensions) namely trim edge, trim edge spar, and drill hole. The performance of the proposed charts ($T^2_{MVV(o)}$, $T^2_{MVV(I)}$, $T^2_{RMVV_{0.5}}$, $T^2_{RMVV_{0.25}}$) are then compared with robust Hotelling T^2 chart using *MCD* (T^2_{MCD}) and RMCD (T^2_{RMCD}), and also the traditional Hotelling T^2 control charts where T^2_0 is without cleaning the outliers and T^2_S is the standard approach which cleans the outliers once.

Estimates for the location vector ($\bar{\boldsymbol{x}}$) and scatter matrix ($\boldsymbol{S}$) are presented in Table 7.11. The calculation of the upper control limits (UCLs) based on the estimates are presented in the last column of the table. The values of the T^2 statistics based on the above estimators appear in the Table 7.12. The graphical presentation of the corresponding control charts are put on view in Figure 7.6 and 7.7. Charts (a), (b), (c), (d) in Figure 7.6 and (e), (f), (g), (h) in Figure 7.7 represent the control chart for traditional T^2 chart (T^2_O), standard T^2 chart (T^2_S), T^2_{MCD}, T^2_{MVV}, $T^2_{MVV(I)}$, T^2_{RMCD}, $T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$ respectively.

Table 7.11. Estimates of location vector, covariance matrix and UCL.

Types of Control Chart	Location Vector($\bar{x}$)	Scatter Matrix(S)	UCL
$\boldsymbol{T_O^2}$	$[0.00504\quad 0.00284\quad 0.01579]$	$\begin{bmatrix} 0.00004 & 0.00002 & 0.00003 \\ 0.00002 & 0.00009 & 0.00001 \\ 0.00003 & 0.00001 & 0.00011 \end{bmatrix}$	11.035
$\boldsymbol{T_S^2}$	$[0.00365\quad 0.00256\quad 0.01209]$	$\begin{bmatrix} 0.00001 & 0.00000 & 0.00000 \\ 0.00000 & 0.00003 & -0.00001 \\ 0.00000 & -0.00001 & 0.00003 \end{bmatrix}$	11.798
$\boldsymbol{T_{MCD}^2}$	$[0.00414\quad 0.00207\quad 0.01096]$	$\begin{bmatrix} 0.00002 & 0.00000 & 0.00000 \\ 0.00002 & 0.00009 & -0.00002 \\ 0.00000 & -0.00002 & 0.00003 \end{bmatrix}$	21.946
$\boldsymbol{T_{MVV(o)}^2}$	$[0.00336\quad 0.00354\quad 0.00913]$	$\begin{bmatrix} 0.00001 & 0.00001 & 0.00000 \\ 0.00001 & 0.00003 & 0.00000 \\ 0.00000 & 0.00000 & 0.00001 \end{bmatrix}$	41.298
$\boldsymbol{T_{MVV(I)}^2}$	$[0.00336\quad 0.00354\quad 0.00913]$	$\begin{bmatrix} 0.00003 & 0.00002 & -0.00001 \\ 0.00002 & 0.00007 & -0.00001 \\ -0.00001 & -0.00001 & 0.00002 \end{bmatrix}$	11.551
$\boldsymbol{T_{RMCD}^2}$	$[0.00414\quad 0.00207\quad 0.01096]$	$\begin{bmatrix} 0.00001 & 0.00000 & 0.00000 \\ 0.00000 & 0.00003 & -0.00001 \\ 0.00000 & -0.00001 & 0.00002 \end{bmatrix}$	24.427
$\boldsymbol{T_{RMVV_{0.5}}^2}$	$[0.00336\quad 0.00354\quad 0.00913]$	$\begin{bmatrix} 0.00003 & 0.00002 & -0.00001 \\ 0.00002 & 0.00006 & -0.00001 \\ -0.00001 & -0.00001 & 0.00002 \end{bmatrix}$	16.503
$\boldsymbol{T_{RMVV_{0.25}}^2}$	$[0.00414\quad 0.00207\quad 0.01096]$	$\begin{bmatrix} 0.00002 & 0.00000 & 0.00000 \\ 0.00000 & 0.00005 & -0.00002 \\ 0.00000 & -0.00002 & 0.00003 \end{bmatrix}$	13.680

Table 7.12. Hotelling T^2 values for future data (Phase II)

n	T_O^2	T_S^2	T_{MCD}^2	$\boldsymbol{T_{MVV(o)}^2}$	$\boldsymbol{T_{MVV(I)}^2}$	$\boldsymbol{T_{RMCD}^2}$	$\boldsymbol{T_{RMVV_{0.5}}^2}$	$\boldsymbol{T_{RMVV_{0.25}}^2}$
1	0.5582	1.4242	1.7659	4.3908	1.5661	3.1188	2.7991	1.7757
2	0.9003	2.5492	2.4694	5.1695	1.8438	4.3613	3.1113	2.4832
3	0.4992	0.4936	0.3437	0.2992	0.1067	0.6070	0.4903	0.3456
4	0.5463	1.0157	0.5456	1.5064	0.5373	0.9636	1.6756	0.5487
5	0.4592	0.9588	0.4580	3.7869	1.3507	0.8088	3.1742	0.4605
6	0.9013	1.7480	1.2527	2.2421	0.7997	2.2125	1.1559	1.2597
7	3.0933	4.1372	4.4404	6.5361	2.3313	7.8423	1.5581	4.4651
8	0.8061	1.2884	0.6837	1.0556	0.3765	1.2075	0.0625	0.6875
9	7.3602	9.6843	14.9766	26.0499	9.2913	**26.4505**	10.1632	**15.0599**
10	3.6198	5.8035	9.7417	19.1760	6.8396	17.2050	8.2001	9.7958
11	5.3839	8.0897	11.8717	19.6313	7.0019	20.9668	7.6269	11.9376
12	2.7387	4.7949	2.9788	8.1388	2.9029	5.2610	1.6758	2.9954
13	3.8058	5.6890	7.4040	11.3895	4.0623	13.0763	4.0550	7.4451
14	2.0548	6.3468	3.3086	9.1498	3.2635	5.8434	2.1624	3.3270
15	2.5073	5.0227	6.8054	12.3881	4.4185	12.0191	5.4472	6.8432
16	1.1976	1.8980	1.0679	2.0563	0.7334	1.8860	0.5881	1.0738
17	1.5798	2.2630	1.7597	2.8765	1.0260	3.1078	0.4603	1.7694
18	5.7910	7.9657	9.2817	13.9293	4.9682	16.3925	4.2017	9.3333
19	1.8304	4.7003	2.4178	4.8791	1.7402	4.2700	0.7299	2.4312
20	**38.1397**	**190.2969**	**214.9233**	**894.5184**	**319.0497**	**379.5799**	**393.5026**	**216.1176**
21	1.2651	2.3301	1.5486	2.0641	0.7362	2.7351	0.4172	1.5572
22	8.4181	**19.7720**	**24.6552**	**45.2462**	**16.1381**	**43.5439**	**19.3300**	**24.7922**
23	3.7588	5.1645	4.8793	7.5328	2.6867	8.6175	1.5065	4.9065
24	1.0602	1.7564	0.9320	2.23575	0.7974	1.6460	0.4294	0.9372
25	**42.8447**	**134.6222**	**68.6307**	**116.02933**	**41.3844**	**121.2098**	**47.0107**	**69.0120**
26	0.4832	1.3946	0.7796	7.32655	2.6132	1.3768	4.9503	0.7839

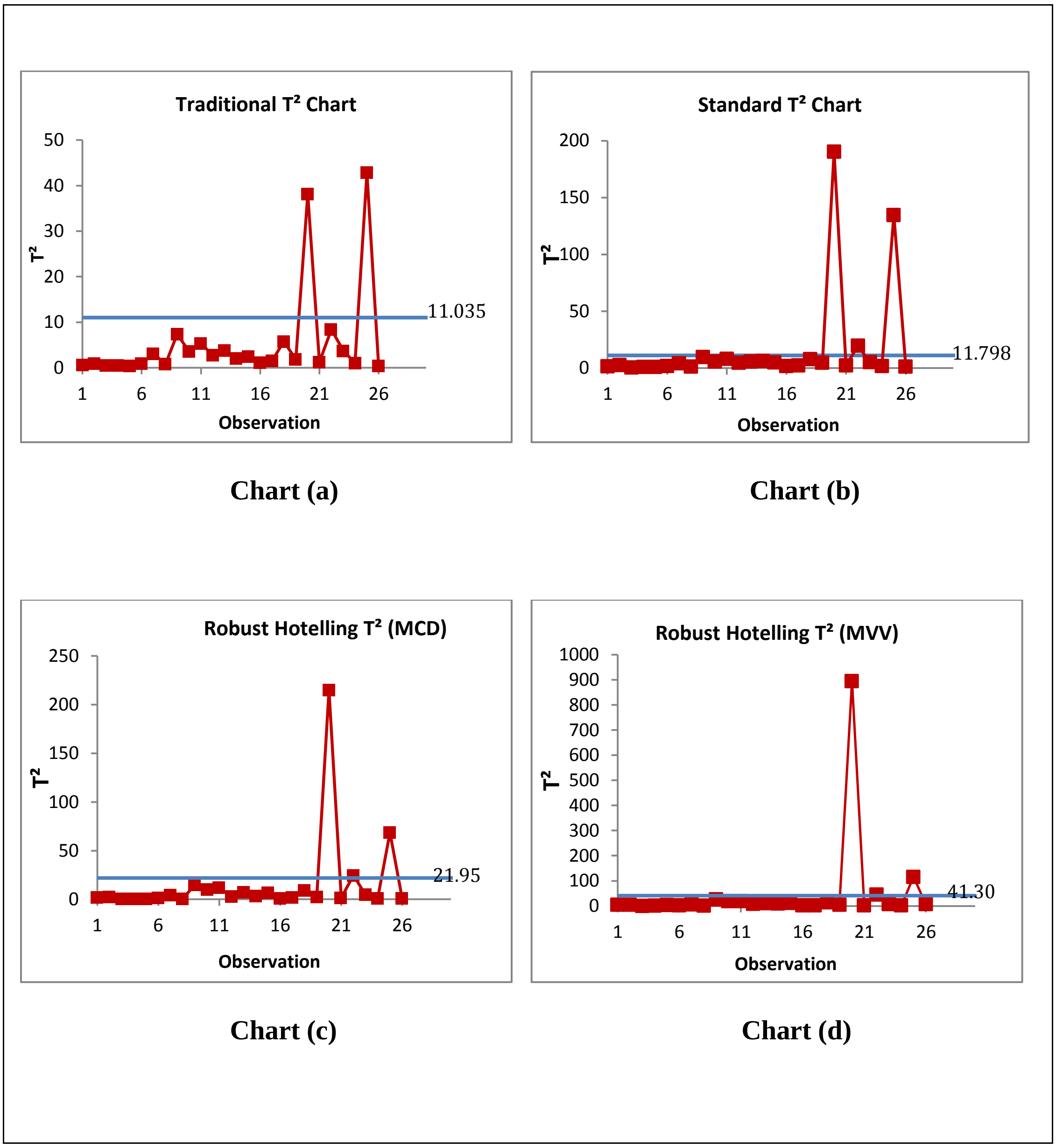


Chart (a)

Chart (b)

Chart (c)

Chart (d)

Figure 7.6: Hotelling T^2 control charts

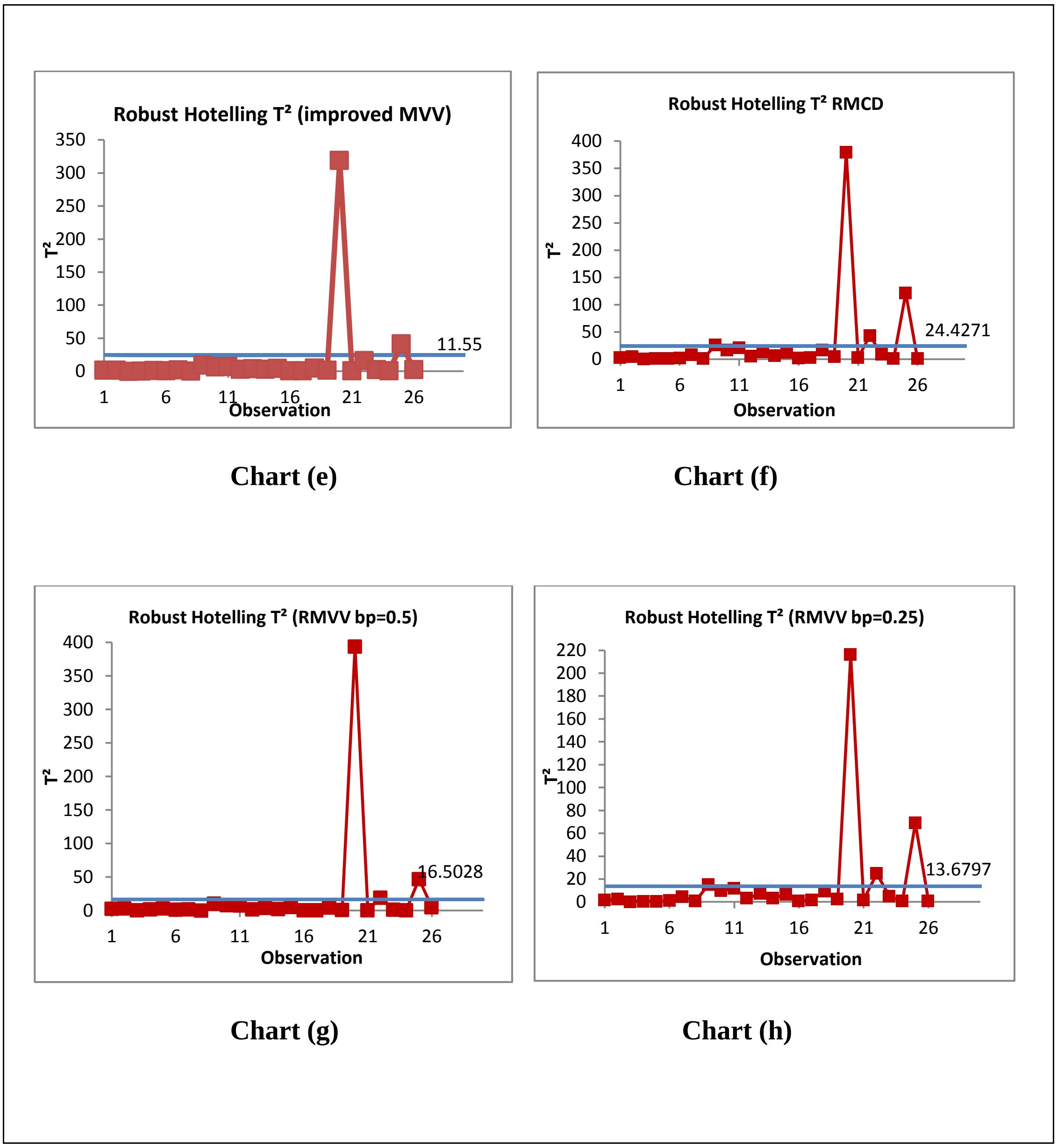


Chart (e) **Chart (f)**

Chart (g) **Chart (h)**

Figure 7.7: Hotelling T^2 control charts

As we can observe, the UCLs in Table 7.11 for $T^2_{MVV(I)}$ become smaller and are closer to the UCLs of T^2_O and T^2_S which used exact distribution as discussed in Section 5.7 of Chapter 5. When comparing the values of the T^2 statistics with their corresponding UCLs, we observe that $T^2_{MVV(O)}$, $T^2_{MVV(I)}$, $T^2_{RMVV_{0.5}}$, $T^2_{RMVV_{0.25}}$, T^2_{MCD}, T^2_{RMCD} and T^2_S signal observations 20, 22 and 25 as out-of-control, but T^2_O only signals 20 and 25 as out-of-control observations and fails to signal observation 22. Interestingly $T^2_{RMVV_{0.25}}$ and T^2_{RMCD} also signal observation 9 as out-of-control, which indicates that reweighted versions of MCD and MVV estimator with breakdown point 0.25 are more efficient in detecting out-of-control signal than the other charts. The performance is also graphically presented in Figure 7.7.

7.7 Conclusion

In this chapter, we proposed another alternative to the Hotelling T^2 chart by using robust estimator known as reweighted minimum variance vector (RMVV) for its location and scatter measures with two different breakdown points. Even though MVV estimators possess the good properties such as affine equivariant, high breakdown point and has better computational efficiency, this estimator is low in statistical efficiency. Thus, MVV was later improved in terms of its statistical efficiency in detecting outliers via reweighted scheme. The performance of the proposed robust Hotelling T^2 chart using RMVV with breakdown point 0.5 and 0.25 performed so well in terms of detecting outliers and also in controlling false alarm rates, but their ability differed on certain conditions.

The $T^2_{RMVV_{0.25}}$ control chart consistently achieved high probability in detecting outliers for low and moderate number of dimensions (p) with small sample size such that the range of p is from 2 to 10 and $n \leq 100$. However, the performance of $T^2_{RMVV_{0.25}}$ in detecting outliers dwindled when p increased to 20 and $T^2_{RMVV_{0.5}}$ showed better ability in handling this situation. In the context of false alarm rates, on the whole, the $T^2_{RMVV_{0.5}}$ control chart is the best performer especially for large sample size with high dimensions. Under low dimensions, $T^2_{RMVV_{0.5}}$ control chart was outdone by T^2_{MVV} and $T^2_{RMVV_{0.25}}$ control chart when $p = 2$ and 5 respectively.

Generally, $T^2_{RMVV_{0.5}}$ demonstrates the best performance compared to the other charts especially for high dimension. The chart is more outstanding with relative to T^2_{MVV} and $T^2_{RMVV_{0.25}}$ in terms of controlling false alarm rate, but the performance of the other two charts cannot be undermined. In the case of low dimension, $T^2_{RMVV_{0.25}}$ is more recommended because it appeared to be more efficient in detecting outliers as proven in the simulated and real data analysis. In real data analysis, $T^2_{RMVV_{0.25}}$ chart and T^2_{RMCD} chart were able to signal observation 9 as out-of-control but other charts failed to do so. Despite the good performance in the real data analysis, T^2_{RMCD} chart showed conflicting performance between false alarm rates and probability of detection such that increasing the probability of detection will increase the false alarm rates away from the nominal value and vice versa. This phenomenon also occurs in T^2_{MCD}.

CHAPTER EIGHT
CONCLUSION AND AREA OF FURTHER RESEARCH

8.1 Conclusion

The ultimate goal of this research is to search for alternative Hotelling T^2 control chart which can improve the performance of the existing charts (traditional Hotelling T^2 and robust Hotelling T^2 issued from MCD and RMCD) in terms of false alarm rate and probability of detection specifically for individual observations. In achieving this goal, firstly we proposed a robust Hotelling T^2 control chart based on minimum vector variance (MVV) estimators by using the second approach where these robust estimators calculated at Phase I are then used directly in Phase II analysis. This second approach does not have to go through the process of outliers cleaning in Phase I because these robust estimators are resistant and not influenced by outliers. MVV is a new robust estimator which possesses the good properties as MCD i.e. affine equivariance and high breakdown point; moreover it has a better computational efficiency as compared to MCD.

In statistical quality control, control limit is an essential element that depends on the distribution of the statistic used. Since the statistical distributions for the robust statistics in this study are unknown, the reference control limits were determined by Monte Carlo simulation method. The evaluations on the performance of the proposed charts were based on the probability of detection and false alarm rates. These charts were then compared with the performance of the traditional charts and the chart issued from MCD estimators. In general, the result showed that T_{MVV}^2 charts were able to detect out of control signals and simultaneously control false alarm rates even with large number of quality characteristics (dimensions). In contrast, the MCD charts performed well in detecting out of control signals but failed in controlling false alarm rates. The traditional chart (T_S^2), however was able to control false alarm rates but not effective in detecting out of control signals.

Investigation on the proposed charts continued with the real industrial data from Asian Composite Manufacturing Sdn. Bhd. (ACM). This company is involved in the production of advanced composite panels for the aircraft industry. ACM has provided us the real data on spoilers which consisted of several features such as trim edge (X_1), trim edge spar (X_2), and drill hole (X_3). The results on real data concurred with the results obtained from the simulation study which support both robust MVV and MCD estimators in detecting outliers. Nonetheless, in this case, performance of T_S^2 chart was on par with the T_{MVV}^2 chart and also T_{MCD}^2 chart. The outcome could be due to the small number of quality characteristics (dimension) of the product. As revealed in the simulation study, T_S^2 performed well in detecting outliers under low dimension (not more than 5) only, but underperformed when the dimension increased to above 5.

Despite the good performance of T^2_{MVV}, the estimated UCLs for Hotelling T^2 chart issued from MVV estimators were large as compared to the traditional and MCD charts. We then took the task to improve the MVV scatter (S_{MVV}) estimator in terms of consistency and biasedness. Investigation through simulation experiment were done to illustrate the consistency and unbiasedness of the MVV estimator at multivariate normal data. The inclusion of consistency and unbiased factor made the S_{MVV} estimator consistent and unbiased at normal model. When put to test on the simulated and real data, the improved control chart, $T^2_{MVV(I)}$, showed great improvement in the control limit values while maintaining its good performance in terms of false alarm and probability of detection.

Since the MVV estimators were directly used in Phase II analysis, they should possess higher statistical efficiency in order to reduce the influence of outlying observations. However, the highly robust affine equivariant estimators with the best breakdown point commonly have to compensate with low statistical efficiency. To mitigate the problem, first we investigated on the asymptotic relative efficiency (ARE) of MVV estimators. The AREs were computed for two different breakdown points such that BP = 0.5 with $h = \left\lfloor\frac{n+p+1}{2}\right\rfloor$ and BP = 0.25 with h = (0.75)n. For each p, we found a decrease in the efficiency value when BP changes from 0.25 to 0.5. Hence, to increase the efficiency while retaining the highest breakdown point, we proceeded to improve the minimum vector variance (MVV) estimators in the context of statistical efficiency via reweighted version (RMVV).

We developed an algorithm to calculate an approximate RMVV solution, where the basis of our algorithm followed a generalization of the MVV algorithm. Interestingly, the reweighted scheme was able to maintain the breakdown point of 0.5 and attain higher efficiency at the normal distribution. But the gain in efficiency increased the bias under contamination. Thus, for a balance between breakdown point and statistical efficiency when the data is suspected to be contaminated by outliers, RMVV with BP = 0.25 is recommended.

Since the ability of RMVV differed with respect to different breakdown points, the investigation on RMVV in Hotelling T^2 in terms of probability of detection and false alarm rates were later conducted on both breakdown points. Both the RMVV charts ($T^2_{RMVV_{0.5}}$ and $T^2_{RMVV_{0.25}}$) were found to be more effective in detecting multiple outliers and controlling false alarm rate compared to the other charts. However, each of them had its advantage over the other charts depending on the combinations of sample size and the proportion of outliers present. The $T^2_{RMVV_{0.25}}$ chart performed better for small sample sizes with low dimensions. In contrast, the $T^2_{RMVV_{0.5}}$ chart was better for large sample sizes of high dimensions. The analysis of ACM spoilers data clarified the situation whereby under small dimension (p = 3) and small sample size (n = 26) the $T^2_{RMVV_{0.25}}$ chart was more capable of detecting out of control data.

8.2 Implications

As we know the performance of traditional Hotelling T^2 control chart using classical estimators in Phase I suffer from masking and swamping effect. Although previous researches have introduced several robust control charts which are capable of addressing the problem of masking and swamping, but there are some disadvantages, particularly in their ability in controlling the false alarm rates. Therefore our goal was to propose alternative Hotelling T^2 control charts which can perform well in detecting outliers while simultaneously controlling false alarm rates.

In this final chapter, we would like to share some of the advances that emerged from this study. In its original state, the MVV estimators when applied in Hotelling T^2 chart had already shown positive impact in detecting outliers and controlling false alarm. While, its counterpart, the MCD estimators showed conflicting ability between the two measurements. After reweighting the MVV estimators, the efficiency of the estimators increased and the reweighted MVV (RMVV) further improved the performance of Hotelling T^2 chart and outperformed the Hotelling T^2 chart issued from reweighted MCD estimators (RMCD)

As a conclusion, the presence of outliers might alter the supposed normal distribution to be non-normal, which consequently will inflate false alarm, suffer loss of power, and will cause spurious detection of out of control process. The RMVV charts may serve as alternative to some other control charts which are unable to handle the problem of non-normality. The proposed robust Hotelling T^2 control charts hold some advantages such that they can handle low, medium and high dimensional quality characteristics and are also able to reduce the computational time. For that reason, our proposed charts are deemed more suitable to be applied in various real life situations especially those related to production process control.

8.3 Limitation

As with any study, the restricted selection of the robust estimators like MCD and RMCD in the context of comparison may limit the generalization of the findings. However, this limitation is necessary because our proposed methods were based on the Mahanalobis distance, moreover these estimators are the most popular and well accepted currently. The choices of number of dimensions, sample sizes and mixed normal models for the generation of the data set, surely does not completely reflect the intricacies of real data sets.

8.4 Areas for Further Research

In the short term, the results of this dissertation are expected to provide additional explanations and approaches in process monitoring and control. However, there are always rooms for improvement. The complexity in estimating MVV could be made simpler so that the proposed charts are more adaptive to industries.

We made improvement on the properties of MVV estimators in terms of consistency and unbiasedness and we also introduced reweighted version of MVV estimator to attain better efficiency. However, the analyses were demonstrated via simulation alone with no mathematical proof. Thus, to support the finding, mathematical proof could be suggested for future research.

There are still some unanswered questions related to high breakdown estimation methods for multivariate control charts. In this study the asymptotic distribution of the T^2_{MCD}, T^2_{RMCD}, T^2_{MVV} and T^2_{RMVV} statistics is considered as χ^2_p. It would be better to study on the exact distribution of MVV and RMVV estimators. Through the exact distribution, the use of approximate control limits is much simpler to obtain than via simulation.

In this study, we only considered high breakdown estimators that are resistant to shifts in the mean vector. However, less study were conducted on the effect of changes in the variance-covariance matrix as done by Levinson, Holmes, and Mergen (2002) and Khoo and Quah (2003, 2004). Thus it could be suggested for future research.

REFERENCES

Alfaro, J. L., & Ortega, J. F. (2009). A comparison of robust alternatives to Hotelling's T^2 control chart. *Journal of Applied Statistics*, 36(12),1385-1396.

Ali, H., Djauhari, M. A., & Syed-Yahaya, S. S. (2008). *On the distribution of FMCD-based robust mahalanobis distance*. Publish in proceeding of the 3rd International Conference on Mathematics and Statistics (ICoMS-3), Institut Pertanian Bogor, Indonesia, 134-1506.

Alt, F. B. (1985). Multivariate quality control. *Encyclopedia of Statistical Sciences* (Vol. 6, Kotz S., Johnson N. L., eds.). New York: Wiley.

Angiulli, F., & Pizzuti, C. (2005). Outlier mining and large high-dimensional data sets.*IEEE Transaction Knowledge Data Engineering*, 17(2), 203–215.

Atkinson, A. C., & Mulira, H. M. (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing*,3, 27-35.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data.* New York: John Wiley.

Beckman, R., & Cook, R. (1983). Outlier.......s. *Technometrics*, 25, 119–149.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34, 279-298.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psyschology*, 31, 144-152.

Butler, R. W., Davies, P. L., & Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics,* 21, 1385-1400.

Campbell, N. A. (1980). Robust procedure in multivariate analysis I robust covariance estimation. *Applied Statistics,* 29, 231-237.

Cerioli, A., Riani, M., & Atkinson, A. C. (2008). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*. Doi:10.1007/s11222-008-9096-5

Chang, S. C., & Bai, D. S. (2004). A multivariate T^2 control chartc for skewed populations using weighted standard deviations. *Qual. Reliab. Engng. Int.*, 20**,** 31-46.

Chen, T. (2010). On reducing false alarms in multivariate statistical process control. *Chemical Engineering Research and Design*, 88 (4). 430 - 436. ISSN 0263-8762.

Chenouri, S., Steiner, S. H., & Mulayath, A. (2009). A multivariate robust control chart for individual observations, *Journal of Quality Technology*, 41(3), 259-271.

Chou, Y. M., Mason, R. L. & Young, J. C.(2001). The control chart for individual observations from a multivariate non-normal distribution. *Communications in Statistics-Theory and Methods*, 30(8), 1937-1949.

Cleroux, R., & Ducharme, G. R. (1989). Vector correlation for elliptical distributions. *Commun.Statist. Theor. Meth.* 18(4):1441-1454.

Croux, C., & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scattermatrix estimator, *Journal of Multivariate Analysis* 71, 161-190.

Croux, C., & Rousseeuw, P. J.(1992). A class of high-breakdown scale estimators based on subranges. *Communication statistics – Theory meth.*, 21(7), 1935-1951.

Davies, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15, 1269-1292.

Davies, P. L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88, 782-792.

Djauhari, M. A., Mashuri, M., & Herwindiati, D. E. (2008). Multivariate process variability monitoring, *Communications in Statistics - Theory and Methods*, 37(11), 1742-1754.

Djauhari, M. A. (2007). A measure of multivariate data concentration. *Journal of Applied Probability and Statistics* 2, 139-155.

Djauhari, M. A., Adnan, R., Lee, M. H., & Ali, H. *An Equivalent Objective Function of Fast MCD*. unpublished manuscript.

Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology, 6(4), 363-379.*

Ferrel, O. C. & Hartline, M. D. (2008). *Marketing Strategy* (4rd ed.).South Western: Edition Thomson Learning Inc.

Garrett, R. G. (1989). The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32, 319-341.

Gather, U., & Becker, C. (1997). Convergence Rates in Multivariate Robust Outlier Identification. *In: Mathematics* 34, 101-107.

Grubel, R., & Rocke, D. M. (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal of Multivariate Analysis* 35, 203- 222. doi:10.1016/0047-259X(90)90025-D

Guo, Jiin-Huarng and Luh, Wei-Ming, (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistic & Probability letters*. 49: 1-7.

Hadi, A. S. (1992). Identifying multivariate outlier in multivariate data. *Journal of Royal Statistical Society B*,53, 761-771.

Hample, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27, 95-107.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of MathematicsStatistics,* 42, 1887-1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of theAmerican Statistical Association,* 69, 382-393.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hawkins, D.M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17, 197-210.

Hawkins, D. M., & Olive, D. J. (1999). Improved feasible solution algorithm for high breakdown estimation. *Computational Statistics and Data Analysis,* 30, 1-11.

Herwindiati, D. E. (2006). *A new criterion in robust estimation for location and covariance matrix, and its application for outlier labeling*. Unpublished Ph.D thesis, Institut Teknologi Bandung.

Herwindiati, D. E., Djauhari, M. A., & Mashuri, M. (2007). Robust multivariate outlier labeling. *Communication in Statistics-Computation and Simulation*, 36: 1287-1294.

Hotelling, H. (1947). Techniques of Statistical Analysis. InC. Eisenhart, M.W. Hastay, & W.A. Wallis, *Multivariate quality control*(pp. 111-184). New York: McGraw-Hill.

Hubert, M., Rousseeuw, P., & Branden, K. V. (2005). Robpca: A new approach to robust principal components analysis. *Technometrics,* 47, 64-79.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math.Statist*, 35, 73-101.

Huber, P. J. (1977). Robust covariances. *In Statistical Decision Theory and Related Topics* (Eds. S. S. Gupta and D. S. Moore). Academic Press, NewYork. 165-191.

Jackson, J. E. (1985). Multivariate quality control. *Communications in statistics: Theory and Methods*,11, 2657-2688.

Jensen,W. A., Birch, J. B. & Woodall, W. H. (2007). High breakdown estimation methods for Phase I multivariate control charts, *Qual. Reliab. Eng. Int.* 23, 615–629.

Johnson, R. (1992). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Khoo, M. B. C. & Quah, S. H. (2003). Multivariate control chart for process dispersion based on individual observations. *Quality Engineering*, 15, 639-642.

Khoo, M. B. C. & Quah, S. H. (2004). Alternatives to multivariate control chart for process dispersion. *Quality Engineering*,16, 423-435.

Lopuhaä, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariance estimators of multivariate location and covariance matrices. *Annal of Statistics*, 19, 229-248.

Lopuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Statist.*, 27, 1638-1665.

Lowry, C. A. & Montgomery, D. C. (1995). A review of multivariate control charts. *IIE Transactions*, 27, 800-810.

Levinson, W. A., Holmes, D. S., & Mergen, A. E. (2002). Variation charts for multivariate processes. *Quality Engineering*, 14, 539-545.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (2000). *Multivariate Analysis* (7rd ed.). London: Academic Press.

Maronna, R. A., Stahel, W. A., & Yohai, V. J. (1992). Bias-Robust Estimators of Multivariate Scatter Based on Projections. *Journal of Multivariate Analysis*, 42, 141-161.

Maronna, R. A., & Zamar, R. (2002). Robust estimation of location and dispersion for high-dimensional datasets. *Technometrics*, 44, 307-317.

Maronna, R.A., Martin, R. D, & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York, NY: John Wiley & Sons.

Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J. & Young, J. C. (1997) Assessment of multivariate process control techniques. *Journal of Quality Technology*, 29 (2), 140-143.

Mason, R. L., & Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: ASA-SIAM.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control* (5rd ed.). New York: Wiley.

Pena, D., & Prieto, J. F. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 3, 286-322.

Pena, D., & Rodriguez, J. (2003). Descriptive measures of multivariate scatter and linear dependence. *Journal of multivariate analysis*, 85, 361-374.

Pison, G., van Aelst, S., & Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, 55, 111-123.

Pison, G., & van Aelst, S. (2004). Diagnostic plots for robust multivariate methods. *J. Comput. Graph. Stat*., 13, 310–329.

Prins, J., & Mader, D. (1997). Multivariate control charts for grouped and individual observations. *Qual. Eng*., 10 (1), 49-57.

Quesenberry, C. P. (2001). The multivariate short-run snapshot *q* chart. *Quality Engineering,* 13, 679-683.

Ramaker, H., van Sprang, E. N. M., Westerhuis, J. A., & Smilde, A. K. (2004).The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 73, 181-187.

Rocke, D., and Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91, 1047-1061.

Roelant, E., van Aelst, S., and Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika,* 70(2), 177-204. doi: 10.1007/500184-008-0186-3

Rousseeuw, P. J., van Driessen, K., van Aelst, S., & Agullo, J. (2004). Robustmultivariateregression.Technometrics, 46(3).doi:10.1198/004017004000000329

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J. (1985). Mathematical Statistics and Applications, B. In W. Grossman, G. Pflug, I. Vincze, & W. Wertz,*Multivariate estimation with high breakdown point*. (pp. 283-297). D. Reidel Publishing Company.

Rousseeuw, P.J. (1994). Unconventional features of positive-breakdown estimators, *Statist. Probab. Lett*., 19, 417-431.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley.

Rousseeuw, P. J., & van Driessen, K. (1999). A Fast algorithm for the minimum covariance determinant estimator.*Technometrics,* 41, 212-223.

Rousseeuw, P. J., and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association,* 85(411), 633-639.

Ryan, T. P. (1989). *Statistical methods for quality improvement*. New York: John Wiley & Sons

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Sullivan, J. H., & Woodall, W. H. (1996). A comparison of multivariate control charts for individual measurements. *J. Qual. Technol,* 28 (4), 398-408.

Sullivan, J. H., & Woodall, W. H. (1998). Adapting control charts for the preliminary analysis of multivariate observations, *Commun. Stat. Simulation Comput*. 27, 953-979.

Tracy, N. D., Young, J. C. & Mason, R. L. (1992). Multivariate control charts for individual observations. *Journal of Quality Technology,* 24, 88-95.

Vargas, J. A. (2003). Robust estimation in multivariate control charts for individual observation, *J. Qual. Technol*. 35, 367-376.

Wierda, S. J. (1994). Multivariate statistical process control-recent results and directions for future research, *Statistica Neerlandica,* 48,147-168.

Willems, G., Pison, G., Rousseeuw, P. J. & van Alest, S. (2002). A robust Hotelling test. *Metrika,* 55, 125-138.

Williams, J. D., Woodall, W. H., Birch, J. B. & Sullivan, J. H. (2006). Distribution of Hotelling's T^2 statistic based on the successive differences estimator, *Journal of Quality Technology,* 38(3), 217-229.

Woodruff, D. L., Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association,* 89, 888-896

Woodall, W. H., Spitzner, D. J., Montgomery, D. C., & Gupta, S. (2004). Using control charts to monitor process and product profiles. *Journal of Quality Technology*, 36, 309-320.

Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical processcontrol. *Journal of Quality Technology*, 31(4), 376.

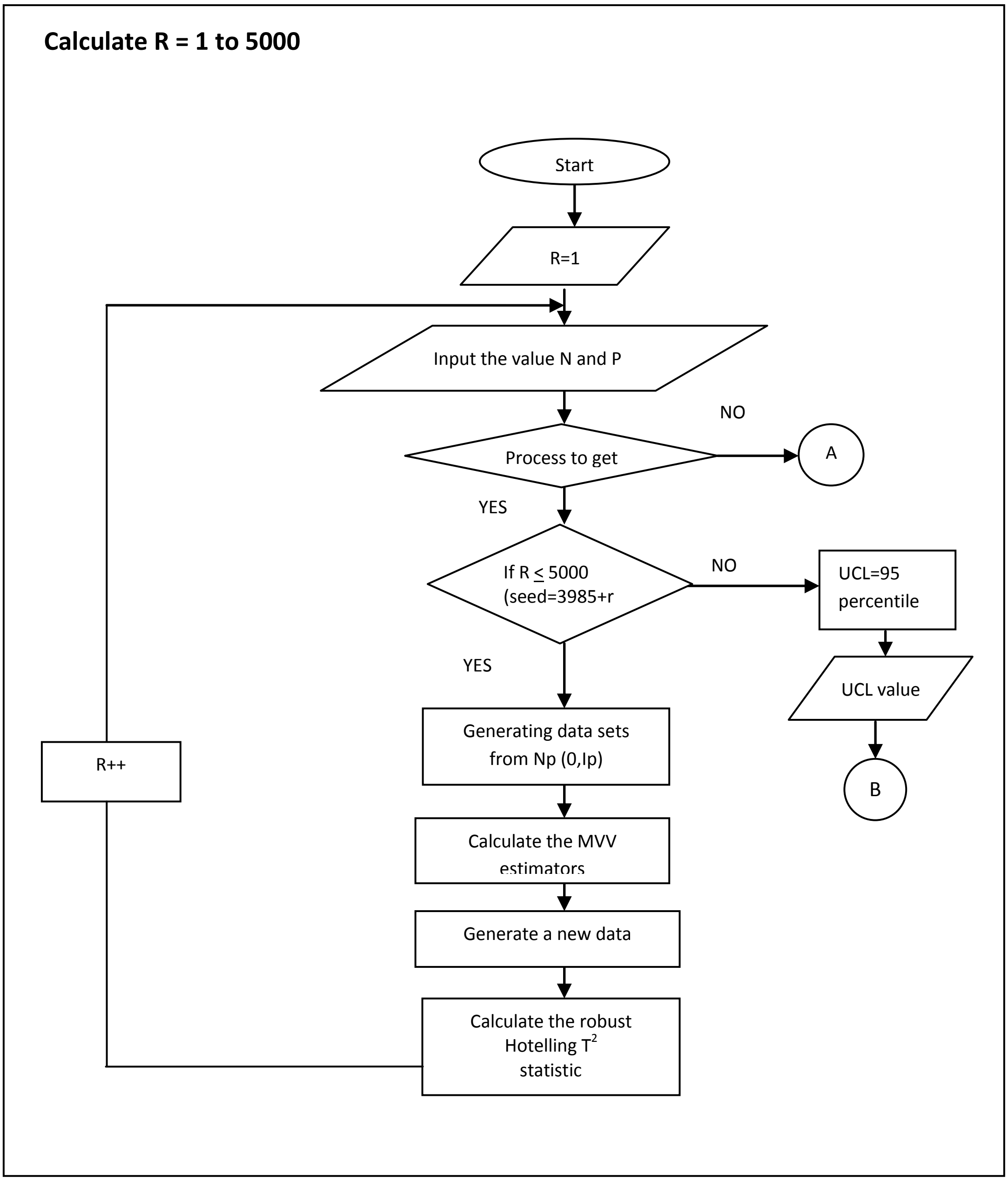
Xu, J. (2003). *Multivariate Outlier detection and process monitoring*. Unpublished Ph.D dissertation, University of Waterloo, Canada.

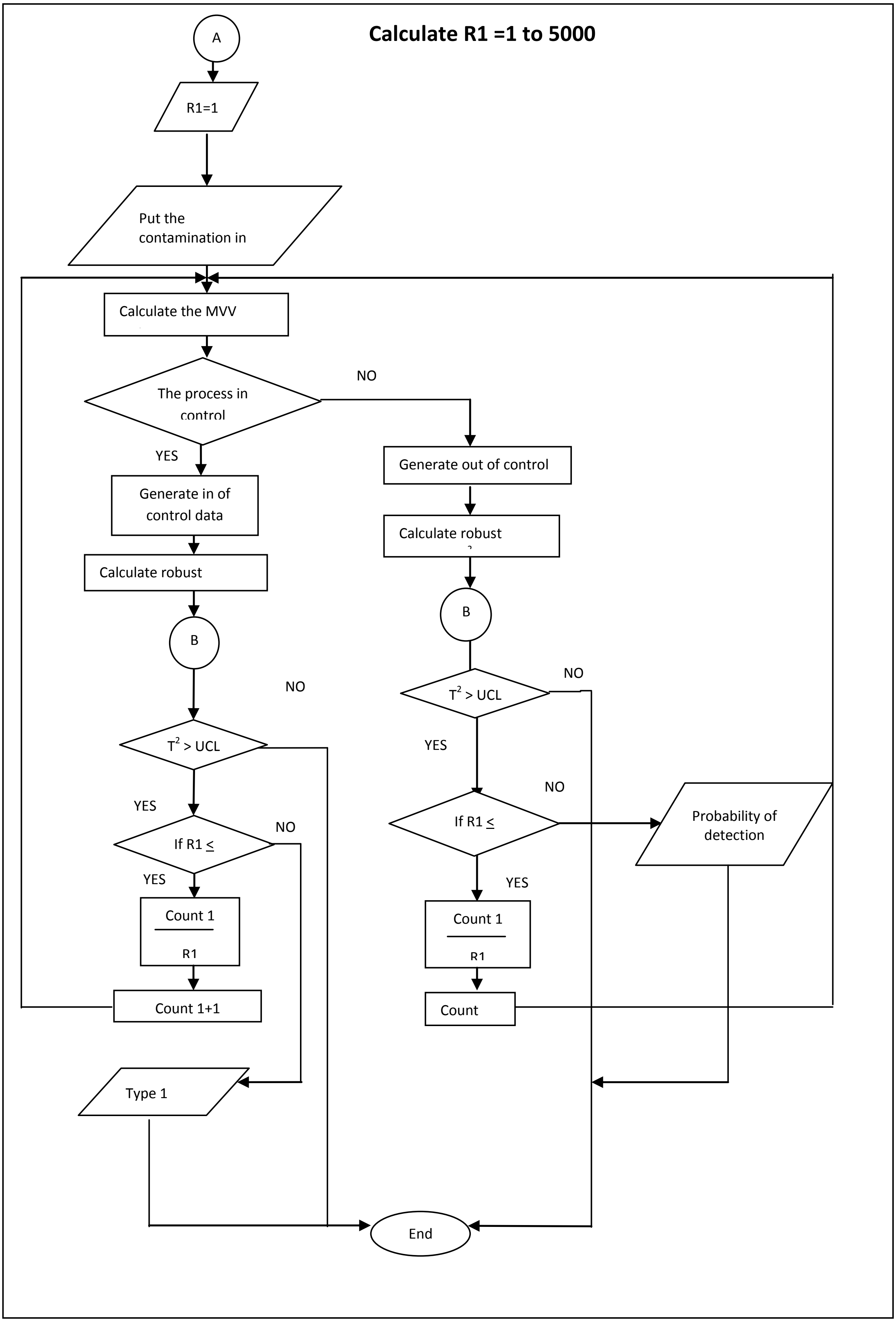
Ye, N., Emran, S. M., Chen, Q. & Vilbert, S. (2002). Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE transactions on computers, 51(7), 810-820.

Zuo, Y. (2006). The Frontiers in Statistics. In Peter Bickel on his 65th Birthday, *Robust location and scatter estimators in multivariate analysis*. Imperial College Press.

Appendix A

FLOW CHART FOR PROCESS OF CALCULATING T^2_{MVV}




Calculate R1 =1 to 5000
A
R1=1
Put the contamination in
Calculate the MVV
NO
The process in control
YES
Generate in of control data
Generate out of control
Calculate robust
Calculate robust
B
B
NO
NO
T^2 > UCL
T^2 > UCL
YES
YES
NO
NO
If R1 ≤
If R1 ≤
YES
YES
Probability of detection
Count 1
R1
Count 1
R1
Count 1+1
Count
Type 1
End

Appendix B

PROGRAM FOR MVV ESTIMATOR

```
function [T,S]= real_MVV(x)
epsilon=10^-5; delta=10e-15;
[n,p]=size(x);
h=floor((n+p+1)/2);% break down point 50%
%h=0.75*n;
rep=500;
SC2=zeros(p,p,500);
TC2=zeros(500,p);
TraceSo=zeros(500,1);
%This condition for h<n, p>=2 and n<=600
%Process of choosing the initial observation (starting subset,(p+1)subset)
%of Ho. This process repeate 500 times.
for k=1:rep
    Iadd=1;
    DetSo=0;
    ho=p+Iadd;
        while DetSo<delta
         Ho=x(rn(1:ho),:);
         To = mean(Ho);
        So=cov(Ho)*(ho-1)/ho;
        DetSo=det(So); Iadd=Iadd+1;
    end
    clear DetSo Iadd h1;
    d=zeros(n,1);
    for m=1:2
        for i=1:n;
            d(i)=(x(i,:)-To)/So*(x(i,:)-To)';
        end
    H1 = x(pi(1:h),:);
    T1=mean(H1);
    a1=(h-1)/h;
    S1=cov(H1)*a1;
    To=T1; So=S1;
    end
TraceSo(k,1)=trace(S1^2);
SC2(:,:,k)=S1;
TC2(k,:)=T1;
end
[TraceSoSort,pi500]=sort(TraceSo);
SCon=zeros(10,1);
TCon=zeros(10,1);
```

Appendix C

PROGRAM FOR HOTELLING T^2

```
% This program calculates the type I error for the Hotelling T square control charts
clear all;
R=5000; R1=1000;   N=200; P=2;

pi=0.2; %percent of outliers
%m=[3 3];
m=[5 5];

Scov1=zeros(P,P,R1);
xbar1=zeros(R1,P);
S=zeros(P,P,R);
T=zeros(R,P);
%meanminusmean=zeros(R1,P);
ROUND=floor(pi*N);
T21=zeros(R,1);
Looping to get the UCL value
for r=1:R
     seed = 3985+r;
     rand('seed',seed)
     randn('seed',seed);
     Z=randn(N+1,P);  %generate random data set
    [T,S]=real_MVV(Z(1:N,:));  %Recall the subroutine result
     meanminusmean(r,:)= Z(N+1,:)-T;
    T21(r,1)=meanminusmean(r,:)/S * (meanminusmean(r,:))'; %T2
 end
for r1=1:R1
    seed = 95395+r1;
    rand('seed',seed);
    randn('seed',seed);
    Z=randn(N,P);

    %contaminate
    Data11=[Z(1:ROUND,:)+repmat(m,ROUND,1);Z(ROUND+1:N,:)];%Case A
   [xbar1(r1,:),Scov1(:,:,r1)]=real_MVV(Data11);

end
```

```
%Phase II
T221=zeros(R1,1);
T222=zeros(R1,1);
for r1=1:R1
    seed = 15391+r1;
    rand('seed',seed);
    randn('seed',seed);
    Z1=randn(1,P);    Z2=Z1+m;
    Data1MinusMean=Z1-xbar1(r1,:);
    Data2MinusMean=Z2-xbar1(r1,:);
    T221(r1)= Data1MinusMean/Scov1(:,:,r1)*(Data1MinusMean)';
T222(r1)= Data2MinusMean/Scov1(:,:,r1)*(Data2MinusMean)';
end

Count1=0;
Count2=0;
for i=1:R1
   if( T221(i)>UCL)
      Count1=Count1+1;
   end
    if( T222(i)>UCL)
       Count2=Count2+1;
    end
end
 typeerror=Count1/R1;
 ProbDetect=Count2/R1;
```