NOVEL NONPARAMETRIC CONTROL CHARTS FOR MONITORING
MULTIVARIATE PROCESSES

by

PANITARN CHONGFUANGPRINYA

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2009
To my father, my mother, and my brother.
ACKNOWLEDGEMENTS

This dissertation is dedicated to my parents and my brother, Fusak, Patpuree and Tanatip Chongfuangprinya. I would like to express my gratitude to my father and my mother who always give me love, concern, support, strength, encouragement and everything. My sincere gratitude goes to my grandmothers, and my late grandfathers for their love and support. I am thankful to my brother for everything.

I am pleased to thank committee members, Dr. Sheik Imhran, Dr. Suk Young Kang, Dr. Sung Seek Moon, and Dr. Victoria C. P. Chen for their time and interest on my dissertation. I am thankful to the faculty and staff of the University of Texas at Arlington and Chulalongkorn University for various forms of support during my study. I am especially grateful to Dr. Pramote Dechaumphai for his advice and knowledge during undergraduate study and inspiration to finish Ph.D. degree. Additionally, I am thankful for the friendship of all COSMOSians, Panaya Rattakorn, Poovich Phaladiganon, and Weerawat Jitpitaklert. I especially appreciate my friend, Dr. Thuntee Sukchotrat who recommended and introduced me to my supervising professor. I am also thankful to him for his advice and commenting of my research. Special thank to Mr. Tommy Le and Mr. Kenzo Tran who make my Ph.D. study possible.

Finally, this dissertation would not have been possible without the expert guidance of my supervising professor. My deepest gratitude is to Dr. Seoung Bum Kim. I have been fortunate to have an advisor who gave me the time, knowledge and practical
advice. I am also thankful to him for commenting and helping me sort out the technical idea and details of my research. His patience and support helped me finish this dissertation.

    July 23, 2009
ABSTRACT

NOVEL NONPARAMETRIC CONTROL CHARTS FOR MONITORING MULTIVARIATE PROCESSES

PANITARN CHONGFUANGPRINYA, Ph.D.

The University of Texas at Arlington, 2009

Supervising Professor: Seoung Bum Kim

The objective of this dissertation is to develop novel nonparametric control charts to effectively monitor and diagnose multivariate processes with a minimal set of modeling assumptions. Statistical process control (SPC) is a set of procedures that uses statistical techniques to measure, analyze, and reduce process variation. A control chart, which is a special type of graph showing the results of periodic inspections over time, is the primary and most successful SPC tool in real-world applications. This dissertation proposes two nonparametric multivariate control charts based on (1) a support vector machines (SVM) algorithm and (2) a linkage ranking algorithm.

The first part of the dissertation proposes a new nonparametric multivariate control chart, called the SVM-PoC (Probability of Class) chart, which integrates a support vector machines algorithm, a bootstrap method, and a traditional control chart
technique. SVM-PoC charts use the PoC values from an SVM algorithm as the monitoring statistic. The control limits of SVM-PoC charts are obtained in a nonparametric way that employs the percentile of the PoC values estimated by a bootstrap method. The performance of the proposed control charts are compared with multivariate Hotelling’s $T^2$ control charts, which are widely used, through a simulation study under various scenarios. The results show that the proposed SVM-based control charts outperform Hotelling’s $T^2$ control charts in both normal and nonnormal situations.

The second part of the dissertation proposes a new nonparametric multivariate control chart based on a linkage ranking algorithm, a $k$LINK chart. This method constructs multivariate control charts based on the ranking of the new measurement relative to the training data. Simulations are performed to demonstrate the effectiveness of $k$LINK charts over Hotelling’s $T^2$ and ranking depth charts in nonnormal situations. Further, an exponential weighed moving average (EWMA) version of $k$LINK charts for increased sensitivity to small shifts was developed. The result showed that the EWMA-$k$LINK charts perform better than the original $k$LINK and multivariate exponentially weighted moving average (MEWMA) charts in detecting small shifts in the nonnormal cases.
# TABLE OF CONTENTS

| ACKNOWLEDGEMENTS | vi 
|------------------|------------------|
| ABSTRACT | vi 
| LIST OF ILLUSTRATIONS | xi 
| LIST OF TABLES | xiv 

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Statistical Process Control &amp; Control Charts</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Multivariate Control Charts</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Nonparametric Multivariate Control Chart</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Data Mining</td>
<td>5</td>
</tr>
<tr>
<td>1.5 Motivation and Contribution</td>
<td>5</td>
</tr>
<tr>
<td>1.6 Outline of the Thesis</td>
<td>6</td>
</tr>
<tr>
<td>2. INTEGRATION OF SUPPORT VECTOR MACHINES AND CONTROL CHARTS FOR MULTIVARIATE PROCESS MONITORING</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Integration of Support Vector Machines and Control Charts</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1 SVM Probabilities Estimation</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Bootstrap-Based Control Limits</td>
<td>14</td>
</tr>
<tr>
<td>2.2.3 Overall Comparison with Existing Methods</td>
<td>16</td>
</tr>
</tbody>
</table>
4. SUMMARY .............................................................................................................66

REFERENCES .........................................................................................................68

BIOGRAPHICAL INFORMATION ...........................................................................72
# LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 A control chart</td>
<td>2</td>
</tr>
<tr>
<td>2.1 Control boundaries from the SVM algorithm from (a) bivariate normal data and (b) bivariate gamma data</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Procedure to establish the control limits of SVM-PoC charts. The control limits are derived from the bootstrap-based $100(1-\alpha)^{th}$ percentile of the PoC values</td>
<td>15</td>
</tr>
<tr>
<td>2.3 The mean vector $\mu_{in}$ and the covariance matrix $S_{in}$ of multivariate normal distribution</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Three-dimensional PCA score plot of multivariate normal distribution. (a) N1, (b) N2, (c) N3, (d) N4, and (e) N5</td>
<td>21</td>
</tr>
<tr>
<td>2.5 Three-dimensional PCA score plot of multivariate gamma distribution. (a) G1, (b) G2, (c) G3, (d) G4, and (e) G5</td>
<td>22</td>
</tr>
<tr>
<td>2.6 Determination of the appropriate $k$ in $kNN$-PoC. (a) N2 and (b) G2</td>
<td>23</td>
</tr>
<tr>
<td>2.7 Multivariate control charts with corresponding control limits for the N3 scenario. (a) SVM-PoC chart, (b) LDA-PoC chart, (c) $kNN$-PoC chart, and (d) $T^2$ chart</td>
<td>24</td>
</tr>
<tr>
<td>2.8 Multivariate control charts with corresponding control limits for the G1 scenario. (a) SVM-PoC chart, (b) LDA-PoC chart, (c) $kNN$-PoC chart, and (d) $T^2$ chart</td>
<td>25</td>
</tr>
<tr>
<td>2.9 Average Type I and Type II error rates of the SVM-PoC charts. (a) under the N3 scenario, and (b) under the G1 scenario</td>
<td>26</td>
</tr>
<tr>
<td>2.10 Histogram of the SVM-PoC at the 90th percentile control limit from 1,000 replications. (a) N1, (b) N2, (c) N3, (d) N4, (e) N5, (f) G1, (g) G2, (h) G3, (i) G4, and (j) G5</td>
<td>27</td>
</tr>
</tbody>
</table>
2.11 Actual Type I and Type II error rates of all charts with standard error marks at $\alpha = 0.1, 0.5,$ and $0.9$. (a) N1, (b) N2, (c) N3, (d) N4, and (e) N5 .................................................................30

2.12 Actual Type I and Type II error rates of all charts with standard deviation marks at $\alpha = 0.1, 0.5,$ and $0.9$. (a) G1, (b) G2, (c) G3, (d) G4, and (e) G5 ..........31

2.13 Comparing average ARL0 and ARL1. (a) N1, (b) N2, (c) N4, (d) G1, (e) G2, and (f) G3 .................................................................34

2.14 Histogram of EWMA-SVM-PoC at the 90th percentile control limit from 1,000 replications. (a) N1, (b) N2, (c) N3, (d) G1, (e) G2, and (f) G3 ..........................35

3.1 A $k$LINK algorithm with the parameter $k = 1$, using the sample mean, applied to the bivariate gamma dataset. Numbers indicate the rankings.............42

3.2 Control boundaries of $k$LINK algorithm with $k = 1$ for bivariate gamma distribution with 200 in-control training data. (a) $\alpha = 0.1$ and (b) $\alpha = 0.2$ ..................................................................48

3.3 Control boundaries of $T^2$ algorithm for bivariate gamma distribution with 200 in-control training data. (a) $\alpha = 0.1$ and (b) $\alpha = 0.2$ ..................................................................48

3.4 Control boundaries of the ranking depth algorithm for bivariate gamma distribution with 200 in-control training data. (a) $\alpha = 0.1$ and (b) $\alpha = 0.2$ ..................................................................49

3.5 Simulation data from the bivariate normal distribution. (a) Training set, (b) N1, (c) N2, and (d) N3 ........................................................................52

3.6 Simulation data from the bivariate gamma distribution. (a) Training set, (b) G1, (c) G2, and (d) G3 ........................................................................53

3.7 Determination of the appropriate $k$ in the G2 scenario.............................................54

3.8 A $k$LINK chart for the G2 scenario ........................................................................56

3.9 Average Type I and Type II error rates of the $k$LINK charts. (a) N2 scenario, and (b) G2 scenario ........................................................................57

3.10 Actual Type I and Type II error rates of the $k$LINK, $T^2$, and ranking depth control charts. (a) N1, (b) N2, and (c) N3 ........................................59
3.11 Actual Type I and Type II error rates of the $k$LINK, $T^2$, and ranking depth control charts. (a) G1, (b) G2, and (c) G3 ..............................................60

3.12 Comparison of the average values of $ARL_0$ and $ARL_1$ among the EWMA-kLINK, $k$LINK, and EWMA-kLINK charts. (a) G1, and (b) G2.............................................................................................................62

3.13 Average values of Type I and Type II error rates (from 100 replications) of the $k$LINK charts with 100, 200, 300, 400, 500, and 1000 training data for (a) bivariate normal (N2) and (b) bivariate gamma (G2) cases.............................................................................................................63

3.14 Actual Type I and Type II error rates of the $k$LINK ($k=5$), $T^2$, and ranking depth (with Tukey depth) charts from the 10-dimensional gamma distribution. .................................................................64
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Comparison of key parameters of the SVM-PoC and other charts</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Comparison of the SVM-PoC and EWMA-SVM-PoC charts under $\text{ARL}_0$ and $\text{ARL}_1$</td>
<td>33</td>
</tr>
<tr>
<td>3.1 Comparison of Type I and Type II error rates for different $k$ and $\alpha$ in the G2 scenarios (Type II error rates are shown in parentheses)</td>
<td>55</td>
</tr>
</tbody>
</table>
1.1 Statistical Process Control & Control Charts

Statistical process control (SPC) is a set of procedures using statistical techniques to keep process variation as small as possible to ensure production quality. A control chart is one of the most successful SPC tools in real world applications (Amsden et al., 1998). Shewhart (1931) first developed a univariate control chart to use for a single attribute. A Shewhart’s \( \bar{x} \) chart uses the sample average of an attribute as a monitoring statistic. A typical control chart comprises of two major components, a monitoring statistic and a control limit. The purpose of the control chart is to estimate when there is evidence that supports the claim that the distribution of the quality characteristics has changed (e.g., shifted mean).

Figure 1.1 is an example of a control chart with an upper control limit set at \( \alpha = 0.2 \). The simulated data in this figure contain 20 observations in which first 10 observations are in control and the last 10 observations are out of control. It shows that a control chart can separate in-control observations from out-of-control observations quite well with two Type I errors and two Type II errors. Type I error occurs when the process is actually in control but the control chart signal an alarm. For example, in Figure 1.1., observations numbers seven and nine are actually in control but the control chart mistakenly generate signals. On the other hand, Type II error occurs when the
process is actually out of control but the control chart cannot detect the signal. In Figure 1.1, observations numbers 11 and 19 are the examples of Type II errors.

![Control Chart](image)

**Figure 1.1 A control chart.**

1.2 Multivariate Control Charts

In modern industries, the processes usually involve a number of process variables that are correlated with each other. Although individual univariate charts can be applied to separately monitor individual process parameter, reliance on such charts may lead to conclusions about multivariate problems that are both inefficiently derived and unsatisfactory in outcomes. In contrast, multivariate control charts take into account the relationship among the process variables and controls the overall probability of false alarm (Type I error rate) to monitor the processes (Woodall and Montgomery, 1999; Johnson and Wichern, 2007). The most widely used multivariate control chart is Hotelling’s $T^2$ control chart. Let $X=[x_1, x_2, \ldots, x_p]^T$ be a $p$-dimensional vector that
represents an observations from a process. The values of $T^2$ statistics can be calculated from the following equation (Hotelling, 1947):

$$T^2 = (\mathbf{X} - \bar{\mathbf{X}})^T \Sigma^{-1} (\mathbf{X} - \bar{\mathbf{X}}),$$

where $\bar{\mathbf{X}}$ and $\Sigma$ are, respectively, a sample mean vector and sample covariance matrix estimated from the in-control dataset. $T^2$ statistics follow the $F$-distribution with $p$ and $n-p$ degrees of freedom based on a multivariate normality assumption. The control limit of the $T^2$ control chart can be computed by

$$CL = \frac{p(n+1)(n-1)}{n^2 - np} F_{(\alpha, p, n-p)},$$

where $n$ and $p$ are, respectively, the number of observations and process variables. $\alpha$ is the Type I error rate that can be usually specified by the user. This control limit is used to monitor the future observations. The observation is considered to be out of control if the corresponding $T^2$ statistic exceeds the control limit.

$T^2$ control charts are known to be insensitive to small process shifts because the charts only use the information of the current observations and does not take into account the information of the past observations. Multivariate cumulative sum (MCUSUM) and multivariate exponentially weighted moving average (MEWMA) control charts use the information of both current and previous observations to increase the sensitivity of detecting small process shifts. A comprehensive review on multivariate control charts can be found in Bersimis et al., (2006).
1.3 Nonparametric Multivariate Control Chart

Most of multivariate control charts require the assumption, stating that the monitoring statistic follows a certain parametric distribution. However, most data from modern industries do not follow the parametric distribution. In order to address this limitation, nonparametric control charts that control the probabilities of false alarms no matter what the underlying distribution of the quality characteristics is have been developed. Liu (1995) proposed three nonparametric control charts based on ranking depth of the multivariate testing data relative to the multivariate training data and plot these ranks into the univariate control chart. These include $r$, $Q$, and $S$ charts. These charts are developed from the univariate $X$, $\bar{X}$, and CUSUM charts, respectively. Data depth is a measure of how central a data point is compared to a data cloud. Mahalanobis depth and simplicial depth (Liu, 1990) were used to construct these charts but more robust simplicial depth is more desirable for nonnormal data. Qiu (2008) proposed a nonparametric control chart based on log-linear modeling. The in-control distribution of the training data $X$ is estimated from the log-linear model of the transformed training data $Y$. Then, mean shift in the distribution of $Y$ is detected with a CUSUM procedure, appropriately designed. A comprehensive review on univariate nonparametric control charts can be found in Chakraborti et al. (2001).
1.4 Data Mining

Data mining is a process of extracting useful information from a large amount of data through the use of any relevant data analysis technique developed to help people make better decision (Tan et al., 2006). Data mining usually can be divided into two categories; unsupervised learning and supervised learning. Unsupervised methods rely solely on the input variables and do not take into account output information. The goal of unsupervised learning is to facilitate the extraction of implicit patterns and elicit the natural groupings within the data set without using any information from the output variable. On the other hand, supervised learning methods use information from both the input and output variables to generate the models that classify or predict the output values of future observations (Duda et. al., 2001; Hastie et al., 2001).

Data mining can be applied in various areas such as science, engineering, biomedicine, healthcare, and business. For examples, handwriting or finger print recognition, e-mail spam filtering, bioinformatics, and climate prediction are well known data mining applications.

1.5 Motivation and Contribution

There are limited methods of multivariate quality control chart that can effectively handle nonnormal situations. Although the existing nonparametric multivariate control charts perform well within the situation for which they were designed, there is no consensus exists about which of them best satisfies all conditions. As the limitations of control chart techniques become increasingly obvious in the face of ever more complex manufacturing processes, data mining algorithms, because of
their proven capabilities to effectively analyze and manage large amounts of data, have the potential to resolve the challenging problems in control charts. Despite the great potential of data mining algorithms for addressing the challenging problems in control charts, few efforts have been made to integrate data mining algorithms with control charts. In particular, both control charts and supervised learning algorithms in data mining have the same objective. In this dissertation we propose two novel nonparametric multivariate control charts that integrate data mining algorithms and control chart techniques. The experimental results from the rigorous simulation studies show that the proposed control charts can efficiently and effectively monitor multivariate processes in nonnormal situations.

1.6 Outline of the Thesis

The remainder of this dissertation is organized as follows. Chapter 2 proposes the novel multivariate control chart that integrates the support vector machine algorithm, bootstrap, and control chart techniques. Chapter 3 proposes a novel nonparametric control charts that uses a $k$-linkage ranking algorithm. Both chapters contain comprehensive literature reviews, clear motivations, methodologies, simulation results, and concluding remarks. Finally, Chapter 4 presents a conclusion of this dissertation.
CHAPTER 2
INTEGRATION OF SUPPORT VECTOR MACHINES AND CONTROL CHARTS FOR MULTIVARIATE PROCESS MONITORING

2.1 Introduction

Statistical process control (SPC) is a set of procedures using statistical techniques to measure, analyze, and reduce process variation. In other words, SPC is used to minimize process variation to ensure production quality. A control chart, which is a special type of graph showing the results of periodic minor inspections over time, is the primary and most successful SPC tool in real-world applications (Amsden et al., 1998). A control chart consists of two major components, a monitoring statistic and a control limit. When a control chart detects a special cause of a variation, an effort is made to find and remove the cause of the offending variation from the process so as to maintain product quality (Woodall and Montgomery, 1999).

Shewhart (1924) first developed a univariate control chart to use for a single attribute. A Shewhart’s $\bar{x}$ chart uses the sample average of an attribute as a monitoring statistic. However, in practice, an alarm might occur because of two or more attributes, or even because of a relationship between these attributes. In either case, detection requires monitoring two or more attributes simultaneously. Although univariate control charts can be applied to each individual characteristic, this technique may lead to unsatisfactory results when multivariate problems are involved (Bersimis et al., 2006; Mason et al., 1997). Multivariate control charts, which simultaneously incorporate
information about multiple process attributes, become more important and useful tools for modern applications. More precisely, multivariate control charts take into account the correlation of attributes and control the overall probability of a false alarm, which is called a Type I error rate (Woodall and Montgomery, 1999; Johnson and Wichern, 2007).

A Hotelling’s $T^2$ control chart ($T^2$ chart) is the most widely used multivariate chart (Hotelling, 1947). $T^2$ charts use the following $T^2$ statistic, which monitors a number of attributes simultaneously:

$$T^2 = (\mathbf{x} - \overline{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}}),$$

where $\overline{\mathbf{x}}$ and $\mathbf{S}$ are, respectively, a mean vector and a covariance matrix estimated from the historical dataset when the process is in control. $T^2$ charts signal when a statistically significant process change occurs. Under the fundamental assumption that the observations follow a multivariate normal distribution, the $T^2$ statistic follows an $F$-distribution with the corresponding parameters (Mason et al., 1997).

In the construction of $T^2$ charts, only in-control data are used to create the control limits. $T^2$ statistics for new observations are then calculated based on $\overline{\mathbf{x}}$ and $\mathbf{S}$ from in-control data. If the $T^2$ value corresponding to the new observation is lower than the control limit, it is considered in control; otherwise it is out of control. $T^2$ charts assume that the in-control group is the only population and this group is used to establish the control limits. This assumption has limited development of more efficient multivariate control charts that can take advantage of the available out-of-control data. Despite the clear fact that use of out-of-control information can significantly improve
efficiency, few efforts have been made to develop multivariate control charts that use this additional data.

Recently, several attempts have been made to convert multivariate SPC into supervised classification problems to serve the same purpose of process control (Hwang et al., 2007; Hu et al., 2007; Sukchotrat, 2008). Despite some successes with various categories of process data, all of these approaches were flawed in one way or another. Hwang et al., (2007) proposed the generation of artificial out-of-control data from the uniform distribution and the combining of these artificial data with in-control data to create a binary classification problem. This method has shown to be useful for a mix of continuous and categorical process variables. However, out-of-control data generated from uniform distributions are scattered randomly across not only the out-of-control region, but also the in-control region. Consequently, the binary classification model constructed with this inaccurate out-of-control information may be inappropriate for use. Hu et al., (2007) proposed generation of the out-of-control data based on the specific shift direction so as to make the control charts more sensitive to a specific type of fault. The main contribution of both of the aforementioned studies has been to attempt to convert multivariate control charts into supervised classification problems as a means to increase the performance of multivariate process control. However, neither study presented a clear way to establish a control limit, one of the important components of control charts.
Recently, Sukchotrat (2008) proposed probability of class (PoC) charts that take advantage of available out-of-control data. In contrast with the previous approaches that proposed generation of artificial out-of-control data, the label of the out-of-control data of this approach can be obtained either directly from the data or from a Phase I analysis (Sukchotrat, 2008). The PoC chart uses a monitoring statistic called a PoC that can be defined as a predicted probability that an observation belongs to a certain class; in-control class or out-of-control class. Further, the control limits of the PoC chart can be established and adjusted by a user-specified misclassification cost. Sukchotrat (2008) compared PoC charts that were based on a linear discriminant analysis (LDA) algorithm and a $k$-nearest neighbors ($k$NN) algorithm with multivariate $T^2$ charts. The PoC charts were found to outperform the multivariate $T^2$ charts. In a comparison of LDA-PoC charts with $k$NN-PoC charts, the latter outperformed the former, especially with nonnormal data.

In the present study we propose to use a support vector machines (SVM) algorithm with a bootstrap method to construct a PoC chart. In the proposed PoC chart, the monitoring statistic is the PoC from the SVM algorithm, and the control limits can be established from the percentiles of the PoC statistics, which is estimated by the bootstrap method (Efron and Tibshirani, 1993). SVM is one of the most widely used supervised learning algorithms and can efficiently handle high-dimensional data from nonnormal distributions (Hastie et al., 2001; Shawe-Taylor and Cristianini, 2000). The traditional SVM models did not provide output as probability information but as a class label. Platt (2000) used a supplementary sigmoid function to obtain the PoC in SVM.
Improved algorithms to obtain the PoC in SVM have been proposed (Lin et al., 2003; Wu et al., 2004).

Our simulation studies conducted in multivariate nonnormal scenarios showed the superiority of the proposed SVM-PoC chart over existing multivariate charts such as the $T^2$, LDA-PoC, and $k$NN-PoC.

2.2 Integration of Support Vector Machines and Control Charts

Like any other control charts, the two major components of the proposed SVM-PoC chart are the monitoring statistic and the control limit. Detailed descriptions of how these two components are presented in the following subsections:

2.2.1 SVM Probabilities Estimation

SVM uses geometric properties to obtain the hyperplane by solving a convex optimization problem that simultaneously minimizes the generalization error and maximizes the geometric margin (Burges, 1998). Given a training dataset $(x_i, y_i), i = 1, 2, \ldots, n$ where $y_i \in \{-1,1\}$ and a testing dataset $z$, the decision values can be obtained from the following equation:

$$f(z) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, z) + b,$$  \hspace{1cm} (2)

where $\alpha_i$ is a Lagrange multiplier that can be solved by quadratic programming under the constraint of cost, and $b$ is a parameter of the model that can be obtained from the following equation:
\[
\alpha_i \{ y_i \left( \sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b \right) - 1 \} = 0. \tag{3}
\]

In (3), \( K \) is a radial basis kernel function defined as
\[
K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2), \tag{4}
\]
where \( \gamma \) is a kernel parameter (Shawe-Taylor and Cristianini, 2000).

In order to obtain probabilities from the SVM model, we need to train not only the SVM model, but also an extra sigmoid function to transform \( f(\mathbf{z}) \) into probability forms. Platt (2000) used a parametric model of a sigmoid function with two parameters, \( A \) and \( B \), to find probabilities. By using \( I \) observations of in-control and \( O \) observations of out-of-control training data, \( A \) and \( B \) can be solved by the following optimizing problem:

\[
\text{Minimize} \left\{ -\sum_{i=1}^{n} \left( q_i \log(p_i) + (1-q_i) \log(1-p_i) \right) \right\}, \tag{5}
\]

where \( q_i \) is defined as
\[
q_i = \begin{cases} 
\frac{I+1}{I+2} & \text{if } y_i = +1 \\
\frac{1}{O+2} & \text{if } y_i = -1
\end{cases}, \quad i = 1, 2, \ldots, n, \tag{6}
\]

and \( p_i \) is defined as:
\[
p_i = \frac{1}{1+\exp(Af(\mathbf{x}_i) + B)}, \quad i = 1, 2, \ldots, n. \tag{7}
\]
A detailed solution strategy for (5) can be found in Platt (2000). Having found the parameters $A$ and $B$ that minimize (5), the SVM-PoC value of a testing observation $z_i$ can be obtained from the following equation:

$$\text{SVM-PoC}_i = \frac{1}{1 + \exp(Af(z_i) + B)}.$$ \hspace{1cm} (8)

Figure 2.1 shows examples of control boundaries from SVM for normal and gamma distributed data with a threshold of 0.5. That is, if SVM-PoC (8) is greater than or equal to 0.5, the corresponding observation was declared out of control; otherwise the observation would be treated as in control. In both the normal and gamma cases, we generated 900 in-control observations and 100 out-of-control observations. It can be seen that SVM can successfully discriminate between in-control and out-of-control observations in both normal and nonnormal (i.e., gamma) cases.

![Figure 2.1](image-url)  
Figure 2.1 Control boundaries from the SVM algorithm from (a) bivariate normal data and (b) bivariate gamma data.
2.2.2 Bootstrap-Based Control Limits

When the probability distribution of the monitoring statistic is known, the control limits of the control chart are established based on that probability distribution with a user-specified value (e.g., Type I error rates). In a PoC chart, however, the distribution of PoC is unknown. To address this issue, we propose a nonparametric method based on the bootstrap method to determine the control limits for a SVM-PoC chart. The bootstrap is a widely used resampling method for deriving statistical estimates when the population distribution is unknown (Efron and Tibshirani, 1993). Figure 2.2 shows a step-by-step description of the bootstrap-based procedure used to calculate the control limits of a SVM-PoC chart. The main idea is to sample the PoC values from the training data for $B$ times with replacements. The control limits are then determined based on the $100\cdot(1-\alpha)^{th}$ percentile of each of the bootstrap samples where $\alpha$ is the user-specified value with its range between 0 and 1. The following is more explicit description of the bootstrap-based procedure to establish the control limits of SVM-PoC charts.

1. Compute the PoC values from Phase I observations of size $m$ using (8).

2. Let $\text{PoC}^*_1, \text{PoC}^*_2, \ldots, \text{PoC}^*_m$ be a set of $m$ PoC values from the $i^{th}$ bootstrap sample ($i=1,\ldots, B$).

3. In each bootstrap sample, given a user-specified value $\alpha$, find the $100\cdot(1-\alpha)^{th}$ percentile value. With $B$ independent bootstrap samples, $B$ $100\cdot(1-\alpha)^{th}$ percentile values are obtained.
4. Calculate the control limit by taking an average of $B \times 100 \cdot (1-\alpha)^{th}$ percentile values (i.e., $\text{PoC}^*_{(100 \cdot (1-\alpha))}$).

5. In Phase II, for each testing observation, if the PoC value of a new observation exceeds $\text{PoC}^*_{(100 \cdot (1-\alpha))}$, declare the observation out of control.

---

**Figure 2.2** Procedure to establish the control limits of SVM-PoC charts. The control limits are derived from the bootstrap-based $100 \cdot (1-\alpha)^{th}$ percentile of the PoC values.
2.2.3 Overall Comparison with Existing Methods

Under the multivariate normal assumption, the monitoring statistic (1) in the $T^2$ chart follows the $F$-distribution. Further, the control limit of the $T^2$ chart can be computed by the $100\alpha\%$ tail-area of the $F$-distribution, where $\alpha$ is the Type I error rate that can be specified by the user.

In LDA-PoC charts, the monitoring statistic, defined as the probability that an observation $x$ belongs to class $i$, can be obtained from the following equation (Sukchotrat, 2008):

$$LDA-PoC_i = e^{-\frac{1}{2}(x-\bar{x}_i)^T S^{-1}(x-\bar{x}_i) + \pi_i}, \quad (9)$$

where $\bar{x}_i$, $S_p$, and $\pi_i$ are, respectively, the estimators of population mean, covariance matrix, and the numbers of observations in class $i$.

In the $k$NN-PoC chart, the probability that an observation $x$ belongs to class $i$ can be obtained from the following equation (Sukchotrat, 2008):

$$kNN-PoC_i = \frac{1}{k} \sum_{n=1}^{k} I(\omega^{(n)} = i), \quad (10)$$

where $\omega^{(n)}$, $n=1,2,\ldots, k$ are the classes of the $k$ nearest observations from the testing observation to the training dataset, and $I$ is the indicator function that returns the value one if the argument is true; otherwise the value is zero.

The control limits of both the LDA-PoC and $k$NN-PoC charts are established and adjusted by the misclassification cost, which is the ratio of Type I and Type II error rates. Our proposed SVM-PoC chart uses a control limit that can be obtained by the
100·(1-\(\alpha\))^{th} bootstrap percentile of the PoC values. Table 2.1 summarizes the monitoring statistics and the control limits of \(T^2\) charts, LDA-PoC charts, \(k\)NN-PoC charts, and SVM-PoC charts.

Table 2.1 Comparison of key parameters of the SVM-PoC and other charts

<table>
<thead>
<tr>
<th>Control Charts</th>
<th>Monitoring Statistics</th>
<th>Control Limit Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T^2) Charts</td>
<td>(T^2)</td>
<td>Type I error rate</td>
</tr>
<tr>
<td>LDA-PoC Charts</td>
<td>LDA-PoC</td>
<td>Misclassification cost</td>
</tr>
<tr>
<td>(k)NN-PoC Charts</td>
<td>(k)NN-PoC</td>
<td>Misclassification cost</td>
</tr>
<tr>
<td>SVM-PoC Charts</td>
<td>SVM-PoC</td>
<td>Bootstrap-based percentile of PoC</td>
</tr>
</tbody>
</table>

2.3 Simulation Study

A simulation study was conducted to evaluate the performance of the proposed SVM-PoC control chart and to compare it with other PoC charts (e.g., LDA-PoC and \(k\)NN-PoC) and \(T^2\) charts. We used an R package (www.r-project.org) to perform the simulation. To be precise, we used the function svm( ) from package e1071 (Dimitriadou et al., 2005), which is based on libsvm (Chang and Lin, 2001) with a radial basis kernel function. We used the radial basis kernel function because it has been generally used when no prior information about the data is available. Other parameters in SVM include the \(C\)-cost of constraints violation and \(\gamma\)-gamma, which were arbitrarily chosen as one and 0.1. It should be noted that the central purpose of the current study is not to find the best-tuned parameters of the SVM model but to propose the SVM-PoC chart as an improved version of the PoC chart.
2.3.1 Simulation Scenarios

For simulation, we generated data from the 10-dimensional multivariate normal distribution with the mean vector $\mu_{in}$ and the covariance matrix $S_{in}$ (Figure 2.3).

\[
\mu_{in} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\]

\[
S_{in} = \begin{bmatrix}
4.726 & -0.120 & 0.234 & -0.060 & -0.750 & 0.627 & 0.167 & -0.760 & 0.748 & 0.126 \\
-0.120 & 5.679 & 0.006 & -0.020 & -0.070 & 0.031 & -0.030 & 0.130 & 0.044 & -0.020 \\
0.234 & 0.006 & 3.428 & 0.026 & -0.050 & 0.138 & 0.107 & 0.001 & 0.151 & 0.070 \\
-0.060 & -0.020 & 0.026 & 5.053 & -0.090 & 0.077 & 0.001 & -0.140 & 0.096 & 0.004 \\
-0.570 & -0.070 & -0.050 & -0.090 & 7.883 & -0.050 & -0.190 & -0.350 & -0.030 & -0.110 \\
0.627 & 0.031 & 0.138 & 0.077 & -0.050 & 2.927 & 0.267 & 0.098 & 0.312 & 0.172 \\
0.167 & -0.030 & 0.107 & 0.001 & -0.190 & 0.267 & 4.492 & -0.210 & 0.310 & 0.084 \\
-0.760 & -0.130 & 0.001 & -0.140 & -0.0350 & 0.098 & -0.210 & 5.391 & 0.161 & -0.120 \\
0.748 & 0.044 & 0.151 & 0.096 & -0.030 & 0.312 & 0.310 & 0.161 & 2.424 & 0.198 \\
0.126 & -0.020 & 0.070 & 0.004 & -0.110 & 0.172 & 0.084 & -0.120 & 0.198 & 4.354
\end{bmatrix}
\]

Figure 2.3 The mean vector $\mu_{in}$ and the covariance matrix $S_{in}$ of multivariate normal distribution.

Further, to examine performance in cases of nonnormal data, we generated data from a 10-dimensional multivariate gamma distribution in which both the shape and the scale parameters were specified as one.

To train the model, we generated 1,000 training observations (Phase I data) containing 900 in-control and 100 out-of-control observations. We used the SVM-PoC values from 900 in-control observations to establish the control limits. Next, we generated 1,000 testing observations (Phase II data) in which the first 900 observations are in control and the last 100 are out of control to compare the performance of each method. To better reflect a real situation in which the number of out-of-control
observations is much smaller than the in-control observations, we assigned only 10% of
the data to be out of control.

To generate the out-of-control data, five types of mean shifts (N1, N2, N3, N4, and N5) were considered in the multivariate normal case and five types of magnitude of the shifts (G1, G2, G3, G4, and G5) were considered in the multivariate gamma case. Note that a change in variance was not considered in the present study. A summary of the simulation scenarios for the multivariate normal distribution and for the gamma distribution is as follows:

- N1 (very small mean shift): $\mu_{out} = \mu_{in} + 0.50\sigma_{in}$, $S_{in} = S_{out}$,
- N2 (small mean shift): $\mu_{out} = \mu_{in} + 0.75\sigma_{in}$, $S_{in} = S_{out}$,
- N3 (medium mean shift): $\mu_{out} = \mu_{in} + 1.00\sigma_{in}$, $S_{in} = S_{out}$,
- N4 (large mean shift): $\mu_{out} = \mu_{in} + 2.00\sigma_{in}$, $S_{in} = S_{out}$,
- N5 (very large mean shift): $\mu_{out} = \mu_{in} + 3.00\sigma_{in}$, $S_{in} = S_{out}$,
- G1 (very small magnitude of the shift): $\delta = 0.25$,
- G2 (small magnitude of the shift): $\delta = 0.50$,
- G3 (medium magnitude of the shift): $\delta = 0.75$,
- G4 (large magnitude of the shift): $\delta = 1.00$,
- G5 (very large magnitude of the shift): $\delta = 2.00$.

We used principal component analysis (PCA) to visualize the multivariate data simulated in the different scenarios. PCA allows the use of fewer dimensions to represent the structure of the original data without the loss of too much information (Jolliffe, 2002). Figures 2.4 and 2.5 display the three-dimensional score plots, showing
that the separation between the in-control and out-of-control observations becomes clearer as the degree of shift increases.

In $k$NN-PoC charts, we used Euclidean distance as a distance metric and determined parameter $k$ that leads to the minimum misclassification error rate. Examples to select the appropriate $k$ are shown for the N2 and G2 scenarios in Figure 2.6. It can be seen that maximum accuracy was achieved at $k=16$. Thus, 16 was chosen as $k$ in all scenarios with the $k$NN-PoC charts.

2.3.2 Construction of SVM-PoC Charts

Control limits in the proposed SVM-PoC chart are established by using percentiles of the PoC values based on the bootstrap method. To ensure comparability, the control limits for the LDA-PoC and $k$NN-PoC in this simulation were also established with the bootstrap percentile of PoC instead of the misclassification cost that was used in Sukchotrat (2008). The number of bootstrap replications $B$ is 1,000.

The SVM-PoC chart and the other charts for the N3 and G1 scenarios are shown in Figures 2.7 and 2.8. The 90th and 80th percentile control limits in the PoC charts were set with $\alpha$ as 0.10 and 0.20. In the $T^2$ chart, $\alpha$ as 0.10 and 0.20 provided the 90% and 80% control limits. These figures show that in the N3 and G1 scenarios, the proposed SVM-PoC chart separated in-control and out-of-control data quite well, but the LDA-PoC and $T^2$ charts can handle only the N3 scenario. In other words, the SVM-PoC chart can distinguish in-control data and out-of-control data better than the LDA-PoC and $T^2$ charts under multivariate gamma scenarios.
Figure 2.4 Three-dimensional PCA score plot of multivariate normal distribution. (a) N1, (b) N2, (c) N3, (d) N4, and (e) N5.
Figure 2.5 Three-dimensional PCA score plot of multivariate gamma distribution. (a) G1, (b) G2, (c) G3, (d) G4, and (e) G5.
Figure 2.6 Determination of the appropriate $k$ in $k$NN-PoC. (a) N2 and (b) G2.

2.3.3 Effect of the Control Limits

The control limits of the SVM-PoC chart were established by a user-specified $100\cdot(1-\alpha)^{th}$ percentile value ($\alpha$) estimated by the bootstrap method. Figure 2.9 illustrates how under multivariate normal and multivariate gamma scenarios, actual Type I and Type II error rates in SVM-PoC charts were changed by the user-specified percentile value in the $x$-axis. We ran 1,000 replications for each scenario to obtain the average Type I and Type II error rates. As expected, increases in $\alpha$ result in higher Type I error rates and lower Type II error rates. The proposed SVM-PoC chart yielded low Type I and Type II error rates when $\alpha$ was at 0.1. All scenarios provided the same trend with the higher shift yielding lower Type II error rates compared with the same Type I error rates. Control limits should be established according to the misclassification cost of the process. If the cost of a Type I error is higher than a Type II error, a lower $\alpha$ that provides a high control limit is preferred and vice versa.
Figure 2.7 Multivariate control charts with corresponding control limits for the N3 scenario. (a) SVM-PoC chart, (b) LDA-PoC chart, (c) $k$NN-PoC chart, and (d) $T^2$ chart.
Figure 2.8 Multivariate control charts with corresponding control limits for the G1 scenario. (a) SVM-PoC chart, (b) LDA-PoC chart, (c) $k$NN-PoC chart, and (d) $T^2$ chart.
Figure 2.9 Average Type I and Type II error rates of the SVM-PoC charts. (a) under the N3 scenario, and (b) under the G1 scenario.

Figure 2.10 shows the histogram of 1,000 bootstrap-derived 90th percentile control limits to examine the variability of the bootstrap estimates. It can be discovered from Figure 2.10 that in the multivariate normal and multivariate gamma scenarios, the variation diminishes as the shift gets larger. Because the maximum standard error for a control limit at any percentile and under any scenario is quite small (less than 0.05), the bootstrap percentile of SVM-PoC is reliable for use as a control limit.
Figure 2.10 Histogram of the SVM-PoC at the 90\textsuperscript{th} percentile control limit from 1,000 replications. (a) N1, (b) N2, (c) N3, (d) N4, (e) N5, (f) G1, (g) G2, (h) G3, (i) G4, and (j) G5.
2.3.4 Performance comparison

We ran 1,000 replications to determine the average of the Type I and Type II error rates with the SVM-PoC, LDA-PoC, kNN-PoC, and $T^2$ charts. Lower Type II error rates, given the similar level of Type I error rates, are considered to represent superior performance and a better method. Average Type I and Type II error rates for all methods under all scenarios are displayed in Figures 2.11 and 2.12.

Under very small to medium shifts in multivariate normal scenarios, and given the similar Type I error rates, the LDA-PoC chart delivered the best results and the $T^2$ chart provided the worst in a comparison of all methods (Figures 2.11(a), 2.11(b), and 2.11(c)). As shown in Figures 2.11(d) and 2.11(e), all methods were comparable in handling the larger shifts that occurred in the N4 and N5 scenarios. However, in coping with the very small shift that occurred in the multivariate gamma scenarios, shown in Figure 2.12(a), the proposed SVM-PoC chart performed best, and again the $T^2$ chart was the worst. As the magnitude of the shift grew, the performances of the LDA-PoC and kNN-PoC charts became more comparable to that of the SVM-PoC chart, but the $T^2$ chart still provided the worst performance. Obviously, the $T^2$ is the least efficient of the charts unless a shift is quite large. LDA-PoC charts surpassed SVM-PoC charts in the multivariate normal scenarios because LDA models are constructed based on the normality assumption. Note that the variation of Type I and Type II error rates from 1,000 replications is quite small, as we can see from the standard error marks at $\alpha = 0.1$, 0.5, and 0.9 in Figures 2.11 and 2.12. In other words, all four methods provided sufficiently stable results. The LDA-PoC charts seem to be the most robust method
because this approach provided such small standard errors that the marks are barely visible. When the shifts grow larger, the standard errors become smaller under all scenarios.

2.4 Exponentially Weighted PoC

Because a traditional $T^2$ control chart monitors and evaluates a current process based on the most recent observation, it may be insensitive to small process shifts. Exponentially weighted moving average (EWMA) charts, which accumulate information from the previous observations, were devised for increased sensitivity to small shifts (Lowry et al., 1992; Lucas and Saccucci, 1990; Stoumbos and Sullivan, 2002). Here we propose an EWMA version of the SVM-PoC chart (i.e., EWMA-SVM-POC). The monitoring statistic $Z_i$ can be computed from the following equation:

$$Z_i = \lambda \text{PoC}_i + (1 - \lambda)Z_{i-1},$$

(11)

where $\lambda$ is the smoothing parameter with a range between 0 and 1, and $Z_i$ is the EWMA-SVM-PoC for observation $i$. The starting value $Z_0$ can be obtained from the average value of PoCs from in-control training data. The proposed EWMA-SVM-PoC chart triggers an alarm when $Z_i$ exceeds the control limit. Smaller $\lambda$ makes the $Z_{i-1}$ relatively more important than PoC$_i$, therefore it is slower to discard an older observation, which has the effect of making it useful for detecting small shifts in a process. A larger $\lambda$ increases its usefulness for detecting larger shifts.
Figure 2.11 Actual Type I and Type II error rates of all charts with standard error marks at $\alpha = 0.1$, 0.5, and 0.9. (a) N1, (b) N2, (c) N3, (d) N4, and (e) N5.
Figure 2.12 Actual Type I and Type II error rates of all charts with standard deviation marks at $\alpha = 0.1$, 0.5, and 0.9. (a) G1, (b) G2, (c) G3, (d) G4, and (e) G5.
As with SVM-PoC charts, the control limits of EWMA-SVM-PoC charts are established by the $100\cdot(1-\alpha)^{th}$ percentile values of $Z$ as estimated by the bootstrap method where $\alpha$ is the Type I error rate. Here, we used 1,000 bootstrap samples to estimate the percentile values.

Because SVM-PoC charts performed quite well with large and very large shifts, our focus with an EWMA-SVM-PoC chart lies in evaluating its performance in detecting small shifts. Three of the smallest shift scenarios from each type of data (N1, N2, N3, G1, G2, and G3) were chosen to compare the performances of the SVM-PoC and EWMA-SVM-PoC charts. We used average run length (ARL) as a measure of performance. Two different types of ARL can be defined based on the process condition. In-control ARL ($ARL_0$) is defined as the expected number of observations needed for the chart to detect a shift in the in-control state, and out-of-control ARL ($ARL_1$) is the expected number of observations needed for the chart to detect a shift in the out-of-control state.

In our simulation, $ARL_0$ and $ARL_1$ were measured based on 1,000 replications with $\lambda$ at 0.4. Here we arbitrarily chose $\lambda=0.4$ because the main purpose is not to find the optimal parameter of the control chart. It can be seen from Table 2.2 and Figure 2.13 that in all scenarios, EWMA-SVM-PoC provides a lower $ARL_1$ than SVM-PoC at any given $ARL_0$. In general, we prefer a procedure that yields a lower $ARL_1$, given similar value of $ARL_0$. Therefore, it can be concluded that EWMA-SVM-PoC performed better than SVM-PoC with small shifts under both the multivariate normal and multivariate gamma scenarios. Note that in G3, the shift is large enough for SVM-
PoC chart to immediately detect the emergence of the out-of-control process, so in this case, SVM-PoC performs equally as well as EWMA-SVM-PoC.

Figure 2.14 displays the histogram of 1,000 bootstrap-derived 90\textsuperscript{th} percentile control limit, showing that the amount of variation diminishes as the shift grows larger in both the multivariate normal and multivariate gamma scenarios. Because the maximum standard error for any percentile control limit and for any scenario is less than 0.02, the bootstrap percentile of EWMA-SVM-PoC is reliable enough to be used as a control limit.

Table 2.2 Comparison of the SVM-PoC and EWMA-SVM-PoC charts under ARL\textsubscript{0} and ARL\textsubscript{1}

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>SVM-PoC</th>
<th>EWMA-SVM-PoC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARL\textsubscript{0}</td>
<td>ARL\textsubscript{1}</td>
</tr>
<tr>
<td>N1</td>
<td>17.456</td>
<td>2.332</td>
</tr>
<tr>
<td>N2</td>
<td>14.044</td>
<td>1.311</td>
</tr>
<tr>
<td>N3</td>
<td>23.42</td>
<td>1.116</td>
</tr>
<tr>
<td>G1</td>
<td>34.211</td>
<td>1.303</td>
</tr>
<tr>
<td>G2</td>
<td>27.269</td>
<td>1.053</td>
</tr>
<tr>
<td>G3</td>
<td>9.116</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 2.13 Comparing average $\text{ARL}_0$ and $\text{ARL}_1$. (a) N1, (b) N2, (c) N4, (d) G1, (e) G2, and (f) G3.
Figure 2.14 Histogram of EWMA-SVM-PoC at the 90\textsuperscript{th} percentile control limit from 1,000 replications. (a) N1, (b) N2, (c) N3, (d) G1, (e) G2, and (f) G3.

2.5 Conclusions

This study proposes a new multivariate SPC technique, the SVM-PoC chart, which integrates an SVM algorithm, a bootstrap method, and a traditional control chart technique. SVM-PoC charts use the PoC values from a SVM algorithm as the monitoring statistic. The control limits of SVM-PoC charts are obtained in a nonparametric way that employs the percentile of the PoC values estimated by a bootstrap method. In order to confirm the reliability of bootstrap estimates, the variability of the bootstrap estimates was studied using the simulated data. Further, in the simulation study, we compared the proposed SVM-PoC charts with other PoC...
control charts and $T^2$ control charts. This comparison revealed that the SVM-PoC charts outperformed the other charts under various scenarios involving nonnormal situations. In normally distributed data, LDA-PoC charts performed best. To increase the capability to detect small process shifts, we developed an EWMA version of the SVM-PoC charts. The simulation study of instances of small shifts showed that EWMA-SVM-PoC performed better than the SVM-PoC charts.

There are some interesting directions for future research. We believe that the proposed SVM-PoC charts can efficiently handle data that can be autocorrelated as well as a mix of continuous and categorical data. This conjecture is based on SVM’s proven capabilities to effectively analyze such data. One of our future studies will be to test this hypothesis.
CHAPTER 3
LINKAGE RANKING ALGORITHMS FOR NONPARAMETRIC
MULTIVARIATE QUALITY CONTROL

3.1 Introduction

Multivariate quality control monitors the quality of a production process that depends on several correlated quality characteristics. A sample quality characteristic is used to calculate measurements, such as sample means, which are then combined into a single statistic that may be plotted on a control chart. The main purpose of the control chart is to monitor the performance of a process over time to maintain the process in-control.

Most of the research performed in this field assumes the measurements that describe quality characteristics follow a multivariate normal distribution (Alt, 1985; Woodall and Ncube, 1985; Jackson, 1985; Mason et al., 1997; Lowry and Montgomery, 1995; Alloway and Raghavachari, 1991; Alt and Smith, 1988). A Hotelling’s $T^2$ control chart ($T^2$ chart) is the most widely used multivariate control chart (Hotelling, 1947), under the assumption that the observations follow a multivariate normal distribution. However, this assumption of normality is not always applicable. In particular, the distribution of the process variable may be highly skewed. In practice, use of transformed variables can be suggested. However, this task is difficult for a multivariate normal distribution, and statistical methods to effectively transform multivariate nonnormal data into multivariate normal data are very limited (Qiu, 2008)
When this assumed normality is not present, the calculated probabilities of Type I and Type II error rates derived from the control mechanisms are unreliable. Only a limited number of methods of multivariate quality control are available for use in nonnormal situations, which gives rise to the motivation to develop nonparametric methods for multivariate process control.

Nonparametric techniques control the probabilities of false alarms no matter what the underlying distribution of the quality characteristics is. In the absence of a distributional assumption, a nonparametric procedure requires training the data of the in-control measurements to represent the underlying distribution. Research in nonparametric multivariate quality control has been conducted by Kapatou and Reynolds (1994), Liu (1995), Hayter and Bush (1995), Bush (1996), Cheng et al. (2000), Stoumbos and Jones (2000), Qiu and Hawkins (2003), Stoumbos and Reynolds (2001), Beltran (2006), Qiu (2008). An overview of some of these methods can be found in Chakraborti et al. (2001).

A nonparametric $r$ chart (ranking depth chart) introduced by Liu (1995) is based on ranking the depth of the multivariate testing data relative to the multivariate training data and then plotting these ranks into the univariate control chart. Data depth is a measure of how central a data point is compared with a data cloud without any distribution assumption. Several data depths such as Mahalanobis depth and simplicial depth were used to construct $r$ charts (Liu, 1990). Later on, Beltran (2006) extended Liu’s ranking depth chart by introducing principle component analysis (PCA). His PCA-simplicial depth $r$ control charts were constructed by using simplicial depth ranks.
of the first and last principle components to improve sensitivity without any distribution assumptions. His method focused on the principle components that accounted for most of the variability of the data; thus, it focused on variability and correlation shifts.

Recently, Qiu (2008) proposed a nonparametric multivariate control based on log-linear modeling. He proposed to estimate the in-control distribution of the original training data \( X(i) \) by transforming \( X(i) \) into binary forms \( Y(i) \) to use a log-linear modeling approach and estimate the joint distribution of \( Y(i) \). Even though information is lost in the transformation, Qiu’s method has the capability to detect a shift in a location parameter vector (e.g., the median vector). However, when a large number of variables are involved, constructing log-linear models is a challenge. In such cases, the performance of log-linear models has been inconclusive.

Although all of these nonparametric control charts perform reasonably well in the situations for which they were designed, no consensus exists about which of them best satisfies all conditions. In the absence of consensus, we propose a new ranking algorithm, \( k \)-linkage ranking (\( k \)LINK), and construction of a \( k \)LINK chart. This is a logical, efficient, and robust control chart that can effectively monitor multivariate processes in nonnormal situations. The nonparametric method discussed in the present paper constructs control charts with quite flexible control boundaries based on the ranking of a new quality characteristic measurement related to the training data.

The outline of this chapter is follows. Section 2 describes the overview for the nonparametric procedure. Section 3 presents a proposed \( k \)LINK control chart. Sections 4 provides simulation results that show the effectiveness of the proposed nonparametric
procedures compared with existing multivariate charts, such as the $T^2$ and ranking depth charts in nonnormal situations. Section 5 presents an exponential weighed moving average version of a kLINK. Effect of sample size of training data and effect of high-dimensional data are in Section 6. Our conclusions are presented in Section 7.

3.2 Overview of the Procedure

The purpose of this study is to develop multivariate quality control methods that do not rely on an assumption of normally distributed quality characteristic measurements. Denote the vector of $p$ quality characteristic measurements by $\mathbf{x} = (x_1, x_2,\ldots, x_p)$, where $\mathbf{x}$ has covariance matrix $\Sigma$. Assume there exists an training data of $m$ independent, identically distributed observations, $\mathbf{x}^1,\ldots,\mathbf{x}^m$, where each is a $p$-dimensional vector of measurements, $\mathbf{x}^i = (x_{1i},\ldots,x_{pi})$ for $i = 1,\ldots, m$. Note that each measurement may be based on a single observation. The purpose of the quality control procedure is to test the null hypothesis that a new observation is from the same (unknown) distribution as the training data. If the null hypothesis is not true, a change has occurred within the process that has affected one or more of the quality characteristics, and the process should be declared out of control. When the process is operating at a constant mean and variability is caused only by unavoidable sources, the process is said to be in control. An out-of-control process is operating under assignable causes of variability. These assignable causes should be detected and eliminated.
In classical quality control, the training data (or Phase I data) are used to calculate control limits. Nonparametric procedures rely on training data to define the standard against which new observations are compared. When unknown distributions are involved, several procedures have been developed to determine if the training data can be considered in control. These procedures look for outliers and changes over time in the median of the training data (Bush, 1996). Given appropriately calibrated training data, our proposed nonparametric procedure calculates score statistics that rank the quality characteristic measurements of a new testing observation, \( x^0 \), relative to the observations in the training data. Specifically, \( x^0 \) is added to the training data to form the combined data of \( m + 1 \) observations; all \( m + 1 \) observations are then ranked, and the value plotted on the control chart is based on the resulting ranking of \( x^0 \).

For a graphical illustration of the proposed chart, consider the dataset, illustrated in Figure 3.1 (discussed in more detail in Section 4). The proposed \( k \)LINK method (that will be described later in detail) resulted in the rankings \( \{1, 2, \ldots, 100\} \) that label the observations. Low rankings indicate observations that are central to the combined data; the highest rankings indicate those that are at the fringes. In a two-dimensional case, a new observation can be plotted with the training data, and a visual judgment can be made as to whether or not the new observation comes from the same distribution. In higher dimensions, however, visualization becomes difficult. However, rankings are simple to interpret regardless of the number of quality characteristics.
Figure 3.1 A kLINK algorithm with the parameter $k=1$, using the sample mean, applied to the bivariate gamma dataset. Numbers indicate the rankings.

Scores and ranks are calculated for all observations in the combined data. A training data observation $x^i$ has score $S_i$ and the corresponding ranking $R_i$ ($i = 1, \ldots, m$). An observation $x_0$ has score $S_0$ and the corresponding ranking $R_0$. The rankings are obtained by simply ordering the scores (average ranks can be used if there are ties among the scores). A low rank $R_0$ indicates that the new observation is within the borders of the $p$-dimensional space, which is defined by the training data. If a new observation and all $m$ observations in the training data are from the same distribution (so that the process is in control), then $R_0$ is equally likely to take any value from one to
Large values of \( R_0 \) indicate that the new observation is outside the training data. The plausibility that the process is in-control can be calculated as

\[
\varphi = \frac{m + 2 - R_0}{m + 1}.
\]  

(1)

The \( \varphi \) represents the proportion of the \( m + 1 \) observations in the combined pool that have scores \( S_i \) no smaller than \( S_0 \). It is important to note that the smallest possible \( \varphi \) obtained from the nonparametric procedures is limited by the size of the training data.

To equalize the effects of quality characteristics with different variances and the relative contribution of correlated quality characteristics, the distance between two measurement observations in the multivariate space is commonly determined by the Mahalanobis distance between observations \( \mathbf{x}_i \) and \( \mathbf{x}_j \) (Bernstein, 1988; Johnson and Wichern, 1998):

\[
D^2 = (\mathbf{x}_i - \mathbf{x}_j)' \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j).
\]

When \( \Sigma \) is unknown, as we assume in this paper, the sample covariance matrix \( \mathbf{V} \) is used to calculate the distance.

\[
d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{x}_j).
\]

(2)

In the following section, we describe in detail a linkage ranking algorithm to create a nonparametric multivariate quality control chart.
3.3 Linkage Ranking Algorithm

A linkage ranking algorithm is based on cluster analysis and chaining. In cluster analysis, analysts are interested in separating the data into distinct groups. Chaining begins with one starting data point, and then observations are added to the chain one by one to form one large cluster instead of several smaller clusters. Observations are continually closer or more similar to the cluster than to the other observations; thus, no separation occurs and dissimilarities are not discovered. Although certain clustering techniques are designed to avoid this problem, the concept of chaining is the key to linkage ranking algorithms. Again, the fundamental question in quality control is whether a training dataset and a new observation have the same distribution, or in other words, whether they belong to the same cluster. If a new observation belongs to the cluster defined by the training data, then it will quickly be linked to the rest of the chain.

The chain begins at the center of the distribution, where it is represented by the central statistic, and then branches to all \( m + 1 \) observations in the combined data. The central statistic \( x^M = (x^M_1, ..., x^M_p) \) is intended to represent the middle of the distribution — for example, the sample mean or the sample median of the training data. At each step, the closest observation to the cluster is added. An observation’s distance to the cluster may be defined in several ways, which will be discussed later. The score statistic \( S_i \) is the order in which measurement realization \( x^i \) is linked to the chain.

\[
S_i = j \quad \text{when} \quad x^i \quad \text{is the} \quad j^{th} \quad \text{observation linked to the chain,}
\]

and hence, \( R_i = S_i \).
The chain will spread outward from the center to the fringes. The linkage ranking algorithm tends to allocate lower scores to observations in dense areas; this is because once one observation in a dense area is added to the chain, the others will soon follow. Having no close neighbors, observations in sparse areas will not be linked to the chain as quickly.

The central statistic is considered the initial member of the chain, although one may argue that once the first observation from the combined data is linked to the chain, then the central statistic should be dropped because it is not an actual observation. However, if we drop the central statistic at this point, then the observation that was second closest to the central statistic may not be the next observation to join the chain unless it also happens to be the closest observation to the first one linked. Thus, an observation’s distance to the central statistic should influence a decision about which observation will join the chain at the next iteration. As the chain grows, the weight of this decision will decrease. By the time the last few observations are joined to the chain, the decision to keep or discard the central statistic will no longer be influential. Because the control procedure is focused on the observations made at the fringes, keeping the central statistic may not be critical to the success of the procedure; however, its retention is recommended.

The question that remains is how to define the distance from an observation to the chain. The distance to the chain is a function of the distances to all the measurement observations in the chain. Suppose the chain has $g$ observations; then $m + 1 - g$ measurement observations remain to be linked to the chain. For an observation $x^j$ not in
the chain, calculate its Mahalanobis distance to every $x^i$ in the chain. Let $D_{i}^{(h)}$ be the $h^{th}$ smallest of these distances, $h = 1, \ldots, g$. Then calculate the distance from observation $x^i$ to the chain as

$$T_i = \sum_{h=1}^{k} D_{i}^{(h)},$$

where $k$ should be determined by the user. The linking rule is to add the observation $x^i$ with the smallest $T_i$. Thus, an observation’s distance to the chain is the minimum distance from it to any $k$ observations already in the chain.

A $k$-linkage ranking ($k$LINK) algorithm uses linking via the sum of $k$ distances in Equation (3) to calculate all the ranks $R_i$, and, subsequently, the rank $R_0$ for a new testing observation. Equation (1) is used to calculate the appropriate $\phi$.

The method based on linking can be thought of as a nearest-neighbor method. The parameter $k$ determines how many nearest neighbors are considered, and only observations already in the chain are a nonmember’s potential neighbors. When $k = 1$, the only consideration given to an observation joining the chain is its distance to any one member of the chain. When $k = g$, the observation selected to join depends on its distance to every point in the chain. Ideally, the center of the chain should remain close to the center of the training data. When $k$ is large, the center of the chain changes more slowly than with smaller values of $k$, which means that it takes more iterations for the mean of the chain to become significantly different from the mean of the combined data. However, it also takes more iteration to get back to the center of the training data once
the two means are no longer close. A summary of the $k$LINK algorithm is described as follows:

**Algorithm $k$LINK**

Specify $k$
Calculate central statistic $x^M$ of the training data.
Initialize the set of observations $x^i$ in the chain: CHAIN = \{M\}.
Initialize the set of observations $x^j$ not in the chain: NOTinCHAIN = \{0, 1,\ldots, m\}.
Initialize counter: RANK = 0.

**repeat**

for all observations $x^i$ such that $i \in$ NOTinCHAIN do

for all observations $x^j$ such that $j \in$ CHAIN do

Calculate $d_{ij}$ as in equation (2).
end for

Calculate $T_i$ as in equation (3).

if $T_i$ is the smallest so far then

Save index $i$ as $i^*$.
end if

end for

Add $x^{i^*}$ with smallest $T_{i^*}$ (saved) to CHAIN:
CHAIN = CHAIN ∪ \{i*\};
NOTinCHAIN = NOTinCHAIN \ ∪ \{i*\};
RANK = RANK +1;

$R_i = RANK$.

**until** all observations are linked in CHAIN.

Figures 3.2, 3.3, and 3.4 show examples of control boundaries from the $k$LINK (with $k = 1$), $T^2$, and a ranking depth (with simplicial depth) charts for the bivariate gamma distribution data with 200 simulated in-control training data. If the observation is inside the colored area, it is in control; otherwise the observation would be treated as out of control. We can see that the lower $\alpha$ results in a larger in-control boundary. It can be also observed that the $k$LINK chart produced more flexible control boundaries than
the $T^2$ and ranking depth charts. This implies that the $k$LINK chart can effectively control Type I and Type II error rates in nonnormal situations.

Figure 3.2 Control boundaries of $k$LINK algorithm with $k = 1$ for bivariate gamma distribution with 200 in-control training data. (a) $\alpha = 0.1$ and (b) $\alpha = 0.2$

Figure 3.3 Control boundaries of $T^2$ algorithm for bivariate gamma distribution with 200 in-control training data. (a) $\alpha = 0.1$ and (b) $\alpha = 0.2$. 
3.4 Simulation Study

A simulation study was conducted to evaluate the performance of the proposed kLINK control chart and to compare it with other control charts (e.g., $T^2$ and ranking depth). We used an R package (www.r-project.org) to perform the simulation. In particular, we used the function depth( ) from the R package “depth” (Masse and Plante, 2009) to implement ranking depth charts.

3.4.1 Simulation Scenarios

Two bivariate probability distributions were generated for the simulation study. We generated the data from the bivariate normal distribution with the mean vector $\mu_{in}$ and the covariance matrix $S_{in}$ as follows:

$$\mu_{in} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad S_{in} = \begin{bmatrix} 3.284 & 1.109 \\ 1.109 & 3.284 \end{bmatrix}.$$
Further, we generated the data from the bivariate gamma distribution in which both the shape and the scale parameters were specified as one. This particular set of shapes was devised to test the robustness of our proposed kLINK chart to nonnormality. Figure 3.1 illustrates the results of applying the kLINK algorithm with \( k = 1 \) to the multivariate gamma dataset. The number to the right of each observation is the ranking for that observation. Note that observations in the center of the plot score lowest, and observations closer to the fringes score highest. In addition, observations in dense areas tend to have similar scores, which is a desirable property of this nonparametric method.

To compare the performance of each method, we generated 500 in-control training observations and 200 testing observations in which the first 180 observations are in control and the last 20 observations are out of control.

To generate the out-of-control data, three types of shifts (N1, N2, and N3) were considered in the multivariate normal case, and three types of shifts (G1, G2, and G3) were considered in the multivariate gamma case. For univariate cases, the process shifts are generally expressed in terms of standard deviation. However, this may not be applicable in multivariate cases because shifts involve more than one process variable. In multivariate cases, shifts can be usually expressed in terms of the following noncentrality parameter \( \lambda \), a function of the magnitude of the shift \( \delta \) and the estimated covariance matrix \( S_{in} \):

\[
\lambda = \sqrt{\delta S_{in}^{-1} \delta}
\]  

(4)
In the present study we assume the covariance matrix has not changed and remains constant. The summary of the simulation scenario for multivariate normal distribution and gamma distribution is described as follows:

- N1 (small shift): $\lambda=1$,
- N2 (medium shift): $\lambda=2$,
- N3 (large shift): $\lambda=3$,
- G1 (small shift): $\lambda=1$,
- G2 (medium shift): $\lambda=2$,
- G3 (large shift): $\lambda=3$.

Figures 3.5 and 3.6, (which visualize the simulated datasets in different scenarios) show that the separation between in-control and out-of-control observations becomes clearer as the degree of shift increases.

As mentioned earlier, to construct the $k$LINK chart, we need first to determine parameter $k$. In general, one can try various $k$ and select the best $k$ that produces the smallest error rate. Tables 3.1 shows Type I and Type II error rates of the $k$LINK charts with different $\alpha$ from the G2 scenario. To find the optimal $k$, we generated 100 preliminary testing observations in which the first 50 observations are in control and the last 50 observations are out of control. The results shows that similar Type I and Type II error rates were obtained for different values of $k$, implying that $k$ does not play a significant role in constructing $k$LINK charts. In this paper, we use $k = 5$ for further analyses.
Figure 3.5 Simulation data from the bivariate normal distribution. (a) Training set, (b) N1, (c) N2, and (d) N3.
Figure 3.6 Simulation data from the bivariate gamma distribution.
(a) Training set, (b) G1, (c) G2, and (d) G3.
3.4.2 Construction of $k$LINK Charts

The $k$LINK charts were constructed from the simulated data. Figure 3.8 shows the $k$LINK chart in the G2 scenarios. The monitoring statistics are the $1-\varphi$ or the plausibility that the testing observations are out of control. The size of the combined data is 501, and the control limit (the horizontal solid line) is the desired $\alpha$, set here at 0.1 and 0.2. The values are reported as out of control if the corresponding monitoring statistics ($1-\varphi$) exceed the control limit (shown in Figure 3.8 in the horizontal solid line).
Table 3.1 Comparison of Type I and Type II error rates for different $k$ and $\alpha$ in the G2 scenarios (Type II error rates are shown in parentheses)

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>(1.00)</td>
<td>(0.98)</td>
<td>(0.94)</td>
<td>(0.98)</td>
<td>(0.96)</td>
<td>(0.96)</td>
<td>(0.96)</td>
<td>(0.96)</td>
<td>(0.96)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>(0.94)</td>
<td>0.00</td>
</tr>
<tr>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.88)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.84)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.72)</td>
<td>(0.78)</td>
<td>(0.82)</td>
<td>(0.82)</td>
<td>(0.80)</td>
<td>(0.80)</td>
<td>(0.80)</td>
<td>(0.80)</td>
<td>(0.80)</td>
<td>(0.80)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.70)</td>
<td>(0.72)</td>
<td>(0.78)</td>
<td>(0.76)</td>
<td>(0.76)</td>
<td>(0.76)</td>
<td>(0.78)</td>
<td>(0.78)</td>
<td>(0.78)</td>
<td>(0.80)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
<td>(0.58)</td>
<td>(0.60)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.56)</td>
<td>(0.52)</td>
<td>(0.50)</td>
<td>(0.50)</td>
<td>(0.54)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.60)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.56)</td>
<td>(0.50)</td>
<td>(0.50)</td>
<td>(0.50)</td>
<td>(0.50)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.56)</td>
<td>(0.56)</td>
<td>(0.52)</td>
<td>(0.48)</td>
<td>(0.48)</td>
<td>(0.48)</td>
<td>(0.48)</td>
<td>(0.48)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>0.09</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.50)</td>
<td>(0.52)</td>
<td>(0.52)</td>
<td>(0.46)</td>
<td>(0.44)</td>
<td>(0.46)</td>
<td>(0.46)</td>
<td>(0.46)</td>
<td>(0.44)</td>
<td>(0.44)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.10</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.44)</td>
<td>(0.42)</td>
<td>(0.42)</td>
<td>(0.40)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.40)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.11</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.42)</td>
<td>(0.36)</td>
<td>(0.34)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>0.12</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>0.13</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.28)</td>
<td>(0.30)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.32)</td>
<td>(0.30)</td>
<td>(0.30)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>0.14</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.30)</td>
<td>(0.28)</td>
<td>(0.26)</td>
<td>(0.26)</td>
<td>(0.26)</td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>0.15</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.22)</td>
<td>(0.22)</td>
<td>(0.18)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.18)</td>
<td>(0.18)</td>
<td>(0.18)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>0.16</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.22)</td>
<td>(0.16)</td>
<td>(0.12)</td>
<td>(0.14)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>0.17</td>
<td>0.08</td>
<td>0.08</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.22)</td>
<td>(0.16)</td>
<td>(0.10)</td>
<td>(0.08)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>0.18</td>
<td>0.08</td>
<td>0.10</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.22)</td>
<td>(0.16)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.19</td>
<td>0.12</td>
<td>0.12</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.06</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.18)</td>
<td>(0.16)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>0.20</td>
<td>0.14</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.10</td>
<td>0.12</td>
<td>0.12</td>
<td>0.10</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.18)</td>
<td>(0.16)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>
3.4.3 Effect of the Control Limits

The control limits used in the $k$LINK chart were established by a user-specified $\alpha$. Figure 3.9 illustrates how under multivariate normal and multivariate gamma scenarios, actual Type I and Type II error rates in the $k$LINK charts were changed by the user-specified $\alpha$ shown in the $x$-axis. We ran 500 replications for each scenario to obtain average Type I and Type II error rates. As expected, increases in $\alpha$ result in higher Type I error rates and lower Type II error rates. The proposed $k$LINK chart yielded low Type I and Type II error rates when $\alpha$ was between 0.05-0.2. All scenarios provided the same trend, with the larger shift yielding lower Type II error rates compared with the same Type I error rates. Because the maximum standard error for
Type I and Type II errors in any of the simulated scenarios is quite small (less than 0.006), 500 replications should be sufficient to draw a reliable conclusion.

![Graphs of Type I and Type II error rates](image)

Figure 3.9 Average Type I and Type II error rates of the $k$LINK charts. (a) N2 scenario, and (b) G2 scenario.

3.4.4 Performance Comparison

The proposed $k$LINK charts were compared with Hotelling’s $T^2$ charts, and ranking depth charts. Figures 3.10 and 3.11, respectively, show the comparative results in the normal and gamma scenarios. We ran 500 replications for each chart to determine the average Type I and Type II error rates. Given the same Type I error rates, lower Type II error rates were considered a better method.

In cases of normal distribution, all three control chart techniques produced comparable results. However, when gamma distribution was involved, the proposed $k$LINK charts outperformed the other two methods. Interestingly, in cases of gamma distribution, the ranking depth chart based on a nonparametric approach performed worse than the $T^2$ control chart. This unexpected result can be explained by the earlier
figures (Figures 3.2, 3.3, and 3.4) that show that the control boundary of the ranking depth chart is less flexible than the $k$-LINK algorithm. Furthermore, the control boundary of the ranking depth control chart is not even as effective as the $T^2$ control chart in the gamma distribution, an example of the skewed distribution. The maximum standard error of Type I and Type II error rates from 500 replications is 0.006, confirming our conclusions.

3.5 Exponentially Weighted $k$-LINK

Because a traditional $T^2$ control chart monitors and evaluates a current process based on the most recent observation, it may be insensitive to small process shifts. Multivariate exponentially weighted moving average (MEWMA) charts, which accumulate information from previous observations, were devised to provide robustness to nonnormality and increased sensitivity to small shifts (Lowry et al., 1992; Lucas and Saccucci, 1990; Stoumbos and Sullivan, 2002). Here we propose an exponentially weighted moving average (EWMA) version of the $k$-LINK chart (i.e., EWMA-$k$LINK). The monitoring statistic $Z_i$ can be computed from the following equation:

$$Z_i = \lambda(1-\phi) + (1-\lambda)Z_{i-1},$$  \hspace{1cm} (5)

where $\lambda$ is the smoothing parameter with a range between 0 and 1, and $Z_i$ is the EWMA-$k$LINK for observation $i$. The starting value $Z_0$ can be obtained from the average $1-\phi$ from in-control training data. The control limits of EWMA-$k$LINK charts are the desired $\alpha$. The proposed EWMA-$k$LINK chart signals an alarm when $Z_i$ exceeds the control limit.
Figure 3.10 Actual Type I and Type II error rates of the $k$LINK, $T^2$, and ranking depth control charts. (a) N1, (b) N2, and (c) N3.
Figure 3.11 Actual Type I and Type II error rates of the \( k \)LINK, \( T^2 \), and ranking depth control charts. (a) G1, (b) G2, and (c) G3.
Because $k$LINK charts performed quite well with medium and large shifts, our focus here on an EWMA-$k$LINK chart lies in evaluating its performance in detecting small shifts in a nonnormal scenario. Small- and medium-shift scenarios from bivariate gamma distribution data (G1 and G2) were used to compare the performances of the MEWMA, the $k$LINK and EWMA-$k$LINK charts. We used average run length (ARL) as a measure of performance. Two different types of ARL can be defined based on the condition of the process. In-control ARL ($ARL_0$) is defined as the expected number of observations needed for the chart to detect a shift in the in-control state; out-of-control ARL ($ARL_1$) is the number of observations expected to be necessary for the chart to detect a shift in the out-of-control state.

In our simulation, $ARL_0$ and $ARL_1$ were computed based on 500 replications with $\lambda$ at 0.25. Here the parameter $\lambda$ was chosen arbitrarily because the main purpose was not to find the optimal parameter of the EWMA control chart. Primarily, we prefer a procedure that provides a lower $ARL_1$, given a similar value of $ARL_0$. Both the $k$LINK and EWMA-$k$LINK charts were constructed with $k=5$. Figure 3.12 shows that in all scenarios and at any given $ARL_0$, the EWMA-$k$LINK charts produced a lower $ARL_1$ than the $k$LINK and MEWMA charts.
3.6. Discussions

3.6.1 Effect of the Sample Sizes of Training Data

The sample size of training data might affect the performance of $k$LINK charts. We studied the performance of six different sample sizes of training data (i.e., 100, 200, 300, 400, 500, and 1000) under medium shift in bivariate normal (N2) and bivariate gamma (G2) scenarios. The resulting average values of Type I and Type II errors from 100 replications of different sample sizes of training data were shown in Figure 3.13. The maximum standard error in this experiment is 0.004. It can be observed that $k$LINK charts produced the similar performance among different numbers of the training data, indicating that the performance of $k$LINK chart does not significantly affected by the number of the training data.
3.6.2 Effect of High-Dimensional Data

It is interesting to investigate the performance of $k$LINK charts in high-dimensional data sets. Here we generated 200 in-control training observations and 200 testing observations from the 10-dimensional gamma distribution. In the testing data, the first 180 observations are in control and the last 20 observations are out of control. Figure 3.14 shows the performance of the $k$LINK (with $k = 5$), $T^2$, and ranking depth (with Tukey depth) charts in terms of Type I and Type II error rates from 500 replications. The maximum standard error in this simulation is 0.0052, small enough to draw a reliable conclusion. For the ranking depth chart, we used Tukey depth because simplicial depth in the R package is limited for only two-dimensional data. In addition, Masse and Plante (2009) indicated that Tukey depth provides only approximation depth.
values in high-dimensional data sets. The result indicated that our $k$LINK charts outperformed both $T^2$ and ranking depth charts. Note that the ranking depth chart could not produce Type I error rates less than 0.6638. This is due to the limitation of ranking depth charts in high-dimensional data sets.

![Figure 3.14: Actual Type I and Type II error rates of the $k$LINK ($k=5$), $T^2$, and ranking depth (with Tukey depth) charts from the 10-dimensional gamma distribution.](image)

3.7. Conclusions

We have presented a new multivariate control chart technique ($k$LINK chart) when the measurement distribution is nonnormal. The proposed control chart is based on a $k$-linkage ranking algorithm that calculates the ranking of the new measurement relative to the in-control training data. Our simulation studies examined the performance of the proposed $k$LINK control chart and compared it ranking depth charts and Hotelling’s $T^2$ charts under various scenarios. The results showed that the $k$LINK
chart outperformed the ranking depth chart and the Hotelling’s $T^2$ chart in cases of nonnormal situations, and all three methods performed comparably in situations of normal distribution. To increase its capability to detect small process shifts, we developed an EWMA version of the $k$-LINK chart. The simulation study showed that the EWMA-$k$-LINK chart performed better than either the $k$-LINK or MEWMA charts in detecting small shifts in nonnormal cases.
CHAPTER 4

SUMMARY

This dissertation proposes nonparametric approaches for monitoring multivariate processes. In Chapter 2, a novel multivariate control chart that integrates an SVM algorithm, a bootstrap method, and a control chart technique is proposed. The proposed chart uses as the monitoring statistic the PoC values from an SVM algorithm. The control limits of SVM-PoC charts are obtained based on the percentile of the PoC values, estimated by a bootstrap method. In addition, in Chapter 3, we proposed a nonparametric multivariate quality control charts based on the new ranking algorithm, the $k$LINK chart. The proposed chart uses the ranking of the new measurement relative to the training data as the monitoring statistic. Both proposed charts serve as alternative options for nonparametric multivariate SPC that required less assumptions of the distribution of the data.

In order to confirm the robustness to nonnormality of the proposed control charts, the simulation studies showed that the proposed SVM-PoC and $k$LINK chart outperformed Hotelling’s $T^2$ and and other nonparametric control charts in the nonnormal cases. Further, the exponential weighted moving average version of SVM-PoC and $k$LINK charts were developed and the simulation studies showed that they performed better than the original SVM-PoC and $k$LINK charts, respectively, in detecting small shifts in nonnormality scenarios.
This research can be extended in many directions. We believe that the proposed SVM-PoC charts can efficiently handle autocorrelated, continuous, and categorical data, because SVM capabilities to effectively analyze such data. Also, optimum training data size and proportions of the out-of-control data can be another future research direction.
REFERENCES


69


BIOGRAPHICAL INFORMATION

Panitarn Chongfuangprinya began work for the degree of Doctor of Philosophy in Industrial Engineering at the University of Texas at Arlington in 2005. During his study, he worked as a graduate teaching assistant and graduate research assistant. His research interest is in the field of data mining and nonparametric statistical process control under nonnormality scenarios. He is also a member of Center of Stochastic Modeling, Optimization & Statistics (COSMOS) at UTA, Institute of Industrial Engineers (IIE) and Institute for Operation Research and Management Science (INFORMS). He holds Master of Science in Industrial Engineering from the University of Texas at Arlington and Bachelor of Engineering in Mechanical Engineering from Chulalongkorn University.