DATA MINING-DRIVEN APPROACHES FOR PROCESS MONITORING AND DIAGNOSIS

by

THUNTEE SUKCHOTRAT

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 2008

To my father, Pol.Lt.Gen. Dr. Wichianchot, my mother, Dr. Maneerat,
and my wife, Saranya Sukchotrat

ACKNOWLEDGEMENTS

I would like to thank the following persons who have supported my Ph.D. life and made this dissertation possible. I would like to express my sincerest gratitude to my supervising professor, Dr. Seoung Bum Kim, for invaluable advice and knowledge throughout my doctoral studies, and for sharing many experiences that are beneficial for both my Ph.D. and future life. There exist no words to describe how grateful I am to his kindness. I deeply appreciate Dr. Victoria C. P. Chen for critical reading and comments that have undoubtedly brought completion to many of my Ph.D. works. I am highly grateful to my supervising committee members, Dr. H. W. Corley, Dr. Jay Rosenberger, and Dr. Chien-Pai Han, for their interest and helpful advice on this dissertation. I would also like to thank Dr. Kwok-Leung Tsui (from Georgia Institute of Technology), Dr. Fugee Tsung (from Hong Kong University of Science and Technology), and Dr. George C. Runger (from Arizona State University), for providing constructive reviews and comments that have fulfilled the completion of my research.

I am grateful to Dr. H. W. Corley and Dr. Donald H. Liles for providing me a chance to be a part of the honorable society at The University of Texas at Arlington (UTA). I especially appreciate Dr. H. W. Corley who introduced and recommended me to my supervising professor. My appreciation is extended to Dr. K. R. Rao for making my graduate study at UTA possible. Also, I highly appreciate all the faculty members who taught me at Chulalongkorn University, especially Dr. Boonwa Thampitakkul who has also provided me admirable kindness and guidance on both of my study and life.

I would like to thank all COSMOSians, especially Drs. Aihong Wen, Chivalai Temiyasathit, Dachuan Thomas Shih, Durai Sundaramoorthi, Huiyuan Fan, Panaya Rattakorn, Prashant Tarun, Poovich Phaladiganon, Siriwat Visoldilokpun, and Weer-

awat Jitpitaklert, for helpful discussions and friendships. My special thanks go to my four best men, Santitarn Sathirathai, Amorpol Huvanandana, Roongrung Owachariyapitak, and Pituck Kijpalakorn, for always being with me and taking care of me during hard times. Many thanks also go to all of my seniors and friends in Texas, especially Nat Soontornvat, Yodchanan Wongsawat, Panitarn Chongfuangprinya, Krissda Prakalphakul, and Prattana Punnakitikashem, for always giving me useful advice and taking care of me. Finally, my deep gratitude goes to my parents and families for their inspiration, encouragement, and endless support. I would like to thank my brothers and sisters, Warong Sukchotrat, Sudaratana Wongweragiat, and Veerajet and Sutthaporn Chirachariyavej, for everything. I am indebted to my grandparents, aunts and uncles, especially Samroeng Vanasin, Prayoon Sukchotrat, Dr. Boon and Charuvarn Vanasin, Dr. Chinda Rojanasathit, Likit Khemkhaeng, and Chamsri Sukchotrat, for all of their love and support.

November 14, 2008

ABSTRACT

DATA MINING-DRIVEN APPROACHES FOR PROCESS MONITORING AND DIAGNOSIS

THUNTEE SUKCHOTRAT, Ph.D.

The University of Texas at Arlington, 2008

Supervising Professor: Seoung Bum Kim

The objective of this dissertation is to develop a new set of efficient process monitoring and diagnostic tools through their integration with data mining algorithms. Statistical process control (SPC) is one of the most widely used techniques for quality control. Although traditional SPC tools are effective in simple manufacturing processes that generate a small volume of independent data, these tools falter when confronted by the large streams of complex and correlated data found in modern manufacturing systems. As the limitations of SPC methodology become increasingly obvious in the face of ever more complex manufacturing processes, data mining, because of its proven capabilities to analyze and manage large amounts of data, has the potential to resolve the problems that are stretching SPC to its limits. This dissertation consists of three components.

First, we propose a new class of control charts that take advantage of available out-of-control information to improve the detection efficiency. The proposed charts integrate a traditional multivariate control chart technique with a supervised classification algorithm. We call the proposed chart the "Probability of Class (PoC) chart" because the values of the PoC, obtained from classification algorithms, are used as monitoring statistics. The control limits of PoC charts are established and

adjusted by the misclassification cost. Second, we propose a collection of new control charts, based on one-class classification algorithms to improve both phase I and phase II analyses in SPC. The proposed one-class classification-based control charts plots a monitoring statistic that represents the degree of being an outlier obtained through the one-class classification algorithm. The control limits of the proposed charts are established based on the empirical level of significance on the quantile estimated by the bootstrap method. Third, we propose a nonparametric false isolation approach in multivariate SPC through monitoring statistics obtained from the one-class classification-based control charts.

The monitoring statistics obtained from one-class classification are decomposed into individual components that reflect the contribution of individual variables to the fault signal. The threshold derived from the bootstrap-quantile estimated method can help indicate the significance of these variables. The novelty of this dissertation is the integration of perspectives from data mining, quality engineering, and statistics that recognizes their shared goals while highlighting their key differences, so as to enable new methodologies for overcoming longstanding research problems and challenges appearing in modern manufacturing/service systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Statistical Process Control

Statistical process control (SPC) is one of the key to improvement of the competitiveness of various industries and organizations [1]. The objective of SPC is to control quality in a process by detecting and reducing the process variabilities, such as delays, deviations, improper procedures, and incorrect estimates. These process variabilities can cause deterioration of products and services. SPC methodology is based on control charts [2], one of the primary techniques in SPC.

1.1.1 Control Charts

Control chart techniques were first developed by Shewhart [3] based on (but not identical to) the concept of statistical hypothesis testing [4] [5]. Control charts are the graphical tools for continuously monitoring a process in order to maintain the process in-control [6]. Control charts consist of two basic components. The first component is a monitoring statistic that has been plotted over a sample number or time. Monitoring statistics can be any measurable function of a sample that represents process quality in the form of variables or attributes such as the sample average and range. The second component of control charts is the control limit. The control limit is usually estimated from the underlying distribution of monitoring statistics that have been derived from in-control historical observations. The control limit is then used as the threshold for a new observation. Processes in which corresponding statistics to the process observations exceed the control limit are declared as out of control. In addition to the control limits, runs rules based on patterns of chart statistics may help identify the out-of-control process [7].

Shewhart-type control charts such as $\overline{X}$ and R charts calculate the monitoring statistics from current observations [4]. Non-Shewhart-type control charts such as cumulative sum (CUSUM) chart [8] and exponentially weighted moving average (EWMA) charts [9] accumulate information on current and past observations to improve the detection of small process shifts [10]. However, most control charts, including the aforementioned control charts, are univariate control charts for monitoring a single process variable [11].

1.1.2 Multivariate Control Charts

In practice, a process usually involves a number of process variables that are correlated with each other. Monitoring a multivariate process with several univariate control charts is inefficient. In fact, univariate control charts applied individually to each correlated variable may lead to incorrect interpretation of multivariate problems. Hotelling [12] first extended the use of univariate control charts to the solution of multivariate SPC problems. Hotelling's T^2 charts (T^2 charts) give a single graphical chart that simultaneously monitors all process variables in a multivariate process. T^2 charts plot the statistical distance (the Hotelling's T^2 statistic) of each observation computed from:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}), \tag{1.1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are a mean vector and a covariance matrix estimated from in-control historical observations. Given a user-specified level of significance (α), the control limit (CL) of the T^2 chart (for individual observations in which the subgroup size is one) is given by:

$$\text{CL} = \frac{p(n+1)(n-1)}{n(n-p)} \cdot F_{\alpha,p,n-p}, \tag{1.2}$$

where n is the sample size of the p-variate historical observations. $F_{\alpha,p,n-p}$ is the quantile function of an F distribution with p and $(n-p)$ degrees of freedom. If the T^2 statistic of the new observation exceeds this control limit, the observation (and the process) will be declared out of control.

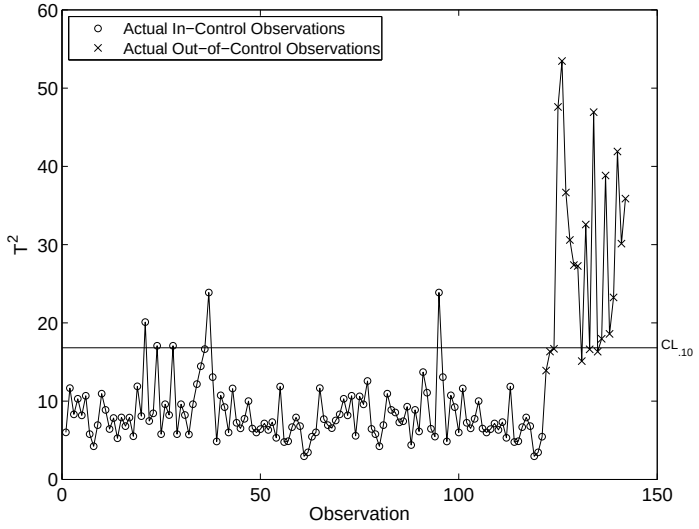


Figure 1.1. The T^2 chart constructed with a manufacturing data.

We illustrate a multivariate control chart by constructing a T^2 chart with data from a manufacturing process. The manufacturing process is characterized by nine process variables. The dataset contains 140 observations in which the last 20 observations are actually out of control. T^2 statistics that summarized all nine variables of each observation are plotted with the control limit in Figure 1.1. Five observations are incorrectly identified as out of control (i.e., a type I error). However, 10 observations are incorrectly identified as in control (i.e., a type II error). Moreover, it can be seen from Equation 1.2 that the T^2 statistic does not require any information given by previous observations. Thus, a T^2 chart could be regarded as a multivariate Shewhart chart [2]. Multivariate CUSUM and multivariate EWMA charts [13] are examples of multivariate non-Shewhart-type control charts developed from existing univariate control charts.

1.1.3 Phase I Statistical Process Control

In general, SPC can be divided into two phases. Phase I identifies the in-control process, and phase II continuously monitors the process over time. The objective of phase I analysis is to isolate the in-control data from unknown historical data from

a new process (or an existing process after adjustment) to which the control chart has never been applied [6]. Phase I analysis plots the unknown data that usually comprise the observations from a normal condition (i.e., in-control data) and abnormal conditions (i.e., out-of-control data and outliers) on a control chart. The trial control limits are constructed and the observations that exceed the control limits are investigated. If the observations are out of control, these out-of-control observations are eliminated. This process can be performed repeatedly until a clean in-control dataset is obtained.

The performance of a phase I application is usually measured in terms of its type I error rate (α) and type II error rate (β). The type I error rate can be estimated by the number of actual in-control observations that are incorrectly detected as out of control divided by the total number of actual in-control observations. The type II error rate can be estimated by the number of actual out-of-control observations that are incorrectly identified as in control divided by the total number of actual out-of-control observations. Given similar levels of type I error rates, those control charts that yield lower type II error rates are preferred. Montgomery [6] suggested that Shewhart-type control charts are effective in phase I analysis because of their effectiveness in detecting large shifts and outliers. Moreover, Shewhart-type control charts are easy to construct and interpret. The T^2 chart is an example of the multivariate Shewhart-type control chart widely used in phase I analysis. Phase I application of a T^2 chart recursively establishes the control limits and proceeds to remove the observations that exceed the control limits until no out-of-control observations are detected.

1.1.4 Phase II Statistical Process Control

The objective of phase II analysis is to correctly and quickly detect out-of-control observations and keep the process in control. Phase II monitors the ongoing process by constructing a control chart using the clean in-control dataset obtained in phase I [5]. The new observations are plotted on the control chart and those obser-

vations that exceed the control limits are identified as out of control. The occurrence of an out-of-control observation sets off an alarm that leads to an investigation of the process to find an assignable cause for the alarm signal.

In addition to type I and type II error rates, average run length (ARL) is an alternative performance measure for phase II applications [2]. ARL is the average number of observations required to trigger an alarm. In-control ARL (sometimes called ARL_0) and out-of-control ARL (sometimes called ARL_1) are calculated, respectively, under sequences of in-control and out-of-control processes. Control charts that yield higher ARL_0 with lower ARL_1 are preferred.

Most control charts were developed for phase II application, and relatively few attempts have been made to improve phase I applications [4]. Researchers should specify whether their studies focus on phase I or phase II analysis because these phases differ in their objectives. Moreover, it should be noted that the control limits of some control charts can differ, depending on whether they are used in phase I or phase II. This is because the distribution assumptions on the data may not be the same in both phases.

1.1.5 Fault Isolation in Multivariate Statistical Process Control

Once an out-of-control signal is triggered, it is important to interpret the signal and identify fault variables. This is because fault variables may contain useful information to support corrective actions. However, the interpretation of a signal is usually a problem with any multivariate control chart [14]. For example, a limitation of T^2 charts is that a T^2 statistic does not provide information that tells which variables contributed to the out-of-control signal.

Mason et al. [2] presented a series of orthogonal decompositions of Hotelling's T^2 statistic for fault isolation based on observations, called MTY's T^2 decomposition method. The decomposition leads to direct interpretation of out-of-control alarms. This facilitates the identification of fault variables once an out-of-control alarm is

reported by Hotelling's T^2 chart. More precisely, the MTY's decomposition method decomposes the T^2 statistics into individual components. The overall T^2 equals the summation of the conditional components. For example, $T^2 = T^2_{1|2} + T^2_{2|1}$ (for $p = 2$), $T^2 = T^2_{1|2,3} + T^2_{2|1,3} + T^2_{3|1,2}$ (for $p = 3$), and $T^2 = T^2_{1|2,...m} + T^2_{2|1,3,...m} + ... + T^2_{m|1,...m-1}$ (for $p = m$), where p is the number of process variables. If the overall statistic yields an out-of-control alarm, then it is possible to determine which components account for the large T^2 value. Each of these components can be compared with the following F distribution to determine if it is statistically significant:

$$\frac{n+1}{n} \cdot F_{\alpha,1,n-1}, \tag{1.3}$$

where n is the number of observations. We then can isolate the corresponding variables that contribute significantly to the out-of-control alarm. In short, the interpretation of the T^2 decomposition can be made as follow: If a component in the decomposition is significant in a statistical sense, it indicates that the corresponding variable is outside its normal range, and thus, we report this as a fault variable.

1.2 Data Mining

Modern researchers in various fields are confronted by an unprecedented wealth and complexity of data. However, the results available to these researchers through traditional data analysis techniques provide only limited solutions to complex situations. The approach to the huge demand for analysis and interpretation of these complex data is managed under the name of "data mining," or "knowledge discovery." Data mining is defined as the process of extracting useful information from large datasets through the use of any relevant data analysis techniques developed to help people make better decisions [15]. These data mining techniques themselves are defined and categorized according to their underlying statistical theories and computing algorithms [16]. This section discusses these various data mining methods and their applications.

1.2.1 Types of Data Mining

In general, data mining can be separated into three methodological categories: unsupervised learning, supervised learning, and semisupervised learning. Unsupervised methods rely solely on input variables (predictors) and do not take into account output (response) information. In unsupervised learning, the goal is to facilitate the extraction of implicit patterns and elicit the natural groupings within the dataset 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. The semisupervised method mixes the unsupervised and supervised methods to generate an appropriate classification/prediction model.

1.2.2 Unsupervised Learning Methods

Unsupervised learning methods attempt to extract important patterns from a dataset without using any information from the output variable. Clustering analysis, which is one of the unsupervised learning techniques, systematically partitions the dataset by minimizing within-group variation and maximizing between-group variation [16]. These variations can be measured based on a variety of distance metrics between observations in the dataset. Clustering analysis includes hierarchical and nonhierarchical algorithms.

Hierarchical clustering algorithms provide a dendrogram that represents the hierarchical structure of clusters [17]. At the highest level of this hierarchy is a single cluster that contains all of the observations. At the lowest level are clusters containing a single observation. Examples of hierarchical clustering algorithms are single linkage, complete linkage, average linkage, and Ward's method.

Nonhierarchical clustering algorithms achieve the purpose of clustering analysis without building a hierarchical structure. The k-means clustering algorithm is one of the most popular nonhierarchical clustering algorithms [18]. A brief summary of the

k-means clustering algorithm is as follows: Given k seed points, each observation is assigned to one of the k seed points close to the observation, which creates k clusters. Then, seed points are replaced with the mean of the currently assigned clusters. This procedure is repeated with updated seed points until the assignments do not change. The results of the k-means clustering algorithm depend on distance metrics, the number of clusters (k), and the location of seed points. Other nonhierarchical clustering algorithms include k-medoids and self-organizing maps.

Principal component analysis (PCA) is another unsupervised technique. It is widely used, primarily for dimensional reduction and visualization [17]. PCA is concerned with the covariance matrix of original variables, and the eigenvalues and eigenvectors are obtained from the covariance matrix. The product of the eigenvector corresponding to the largest eigenvalue and the original data matrix leads to the first principal component (PC), which expresses the maximum variance of the dataset. The second PC is then obtained using the eigenvector corresponding to the second largest eigenvalue. This process is repeated N times to obtain N PCs where N is the number of variables in the dataset. The PCs are not correlated with each other, and generally, the first few PCs are sufficient to account for most of the variations. Thus, the PCA plot of observations using these first few PC axes facilitates visualization of high-dimensional datasets.

1.2.3 Supervised Learning Methods

Supervised learning methods use both input and output variables to provide the model or rule that characterizes the relationships between the input and output variables [18]. Based on the characteristics of the output variable, supervised learning methods can be categorized as either regression or classification. In regression problems, the output variable is continuous so that the main goal is to predict the outcome values of an unknown future observation. In classification problems, the

output variable is categorical, and the goal is to assign existing labels to an unknown future observation.

Linear regression models have been widely used in regression problems because of their simplicity. Linear regression is a parametric approach that provides a linear equation to examine relationships of the mean response to a single variable or to multiple input variables [19]. Linear regression models are simple to derive, and the final model is easy to interpret. However, the parametric assumption of an error term in linear regression analysis often restricts its applicability to complicated multivariate data. Further, linear regression methods cannot be employed when the number of variables exceeds the number of observations. Multivariate adaptive regression spline (MARS) is a nonparametric regression method that compensates for the limitations of ordinary regression models. MARS is one of the few tractable methods for high-dimensional problems with interactions and estimates of an unknown relationship between a continuous output variable and a number of input variables. MARS is a data-driven statistical linear model in which a forward stepwise algorithm is first used to select the model term and is then followed by a backward procedure to prune the model. The approximation bends at "knot" locations to model curvature, and one of the objectives of the forward stepwise algorithm is to select the appropriate knots. Smoothing at the knots is an option that may be used if derivatives are desired.

Classification methods provide models to classify unknown observations according to the existing labels of the output variable. Traditional classification techniques include linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), which are based on Bayesian theory. Both the LDA and QDA assume that the dataset follows normal distribution. LDA generates a linear decision boundary by assuming that populations of different classes have the same covariance. QDA, on the other hand, does not have any restrictions on the equality of covariance between two populations and provides a quadratic equation that may be efficient for linearly nonseparable datasets.

Many supervised learning techniques, such as decision trees, support vector machines, k-nearest neighbors, and artificial neural networks, can handle both regression and classification problems. Decision tree models are highly popular in various areas because of their flexibility and interpretability. These models are flexible in that the models can efficiently handle both continuous and categorical variables in the model construction. The output of decision tree models is a hierarchical structure that consists of a series of if-then rules to predict the outcome of the response variable, thus facilitating the interpretation of the final model. From an algorithmic point of view, a decision tree model has a forward stepwise procedure that adds model terms, a backward procedure for pruning, and conducts variable selection by only including useful variables in the model.

Support vector machine (SVM) is another supervised learning model popularly used for both regression and classification problems. SVMs use geometric properties to obtain a separating hyperplane. This is done by solving a convex optimization problem that simultaneously minimizes the generalization error and maximizes the geometric margin between the classes [20]. Nonlinear SVM models can be constructed from kernel functions that include linear, polynomial, and radial basis functions.

Another useful supervised learning technique is k-nearest neighbors (kNNs), a type of lazy-learning (instance-based learning) technique [21]. kNNs do not require a trained model. Given a query point, the k closest points are determined. A variety of distance measures can be applied to calculate how close each point is to the query point. Then the k-nearest points are examined to find which of the most categories belong to the k-nearest points. Lastly, this category is assigned to the query point being examined. This procedure is repeated for all the points that require classification.

Finally, artificial neural networks (ANNs), inspired by the way biological nervous systems learn [22], are widely used for prediction modeling in many applications. ANN models are typically represented by a network diagram containing several layers

(e.g., input, hidden, and output layers) that consist of nodes. These nodes are interconnected with weighted connection lines. These weights are adjusted as training data are presented to the ANN during the training process. The neural network training process is an iterative adjustment of the internal weights to bring the network's output closer to the desired values through minimizing the mean squared error.

1.2.4 Semisupervised Learning Methods

Semisupervised learning approaches have received increasing attention in recent years. Chapelle et al. [23] described that "Semisupervised learning is halfway between supervised and unsupervised learning." Semisupervised learning methods create a classification model by using partial information from the labeled data. One-class classification is an example of a semisupervised learning method that can distinguish between the class of interest (target) and all other classes (outlier) [24]. In the construction of the classifiers, one-class classification techniques require only information from the target class. The applications of one-class classification include novelty detection, outlier detection, and imbalanced classification.

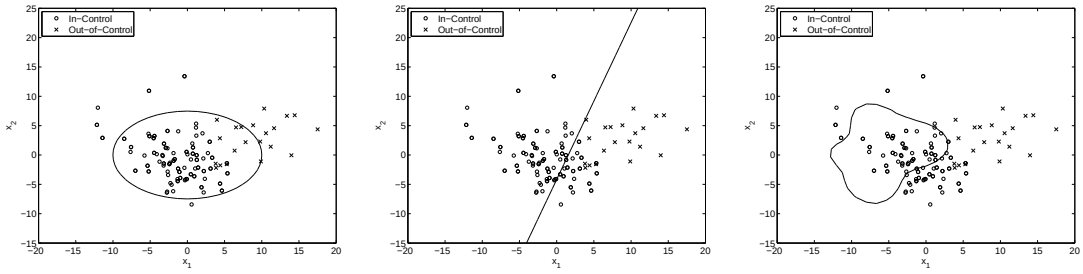
Support vector data description (SVDD) is a one-class classification technique that combines a traditional support SVM algorithm with a density approach [25]. SVDD produces a classifier to separate the target from the outliers. The decision boundary of SVDD is constructed from an optimization problem that minimizes the volume of the hypersphere from the boundary and maximizes the target data being captured by the boundary. The main difference between supervised and semisupervised classification methods is that the former generates a classifier to classify an unknown observation into the predefined classes and the latter creates a closed boundary around the target data in order to separate them from all other types of data.

1.2.5 Applications of Data Mining

Interest in data mining has increased greatly because of the availability of new analytical techniques with the potential to retrieve useful information or knowledge from vast amounts of complex data that were heretofore unmanageable. Data mining has a range of applications [26][27][28], including manufacturing, marketing, telecommunications, health care, biomedicine, e-commerce, and sports. In manufacturing, data mining methods have been applied to predict the number of product defects in a process and identify their causes. In marketing, market basket analysis provides a way to understand the behavior of profitable customers by analyzing their purchasing patterns. Further, unsupervised clustering analyses can be used to segment customers by market potential. In the telecommunication industries, data mining methods help sales and marketing people establish loyalty programs, develop fraud detection modules, and segment markets to reduce revenue loss. Data mining has received tremendous attention in the field of bioinformatics, which deals with large amounts of high-dimensional biological data. Data mining methods combined with microarray technology allow monitoring of thousands of genes simultaneously, leading to a greater understanding of molecular patterns. Clustering algorithms use microarray gene expression data to group the genes based on their level of expression, and classification algorithms use the labels of experimental conditions (e.g., disease status) to build models to classify different experimental conditions.

1.3 Motivation and Contribution

Control chart techniques in SPC and supervised/semisupervised techniques in data mining share a similar characteristic in terms of detecting the data of interest from a certain pool of data. To illustrate this similarity, Figure 1.2 displays the control boundary corresponding to the control limit of the T^2 chart and the decision boundaries of LDA and SVDD. These boundaries were constructed with a manufacturing dataset characterized by nine process variables. The dataset contains 120



(a) (b) (c)

Figure 1.2. The control/decision boundaries of the (a) T^2 chart, (b) LDA, and (c) SVDD constructed with a manufacturing data (reduced dimension).

in-control observations and 20 out-of-control observations. In order to facilitate visualization, PCA was applied to reduce the dimensions of the data before construction of the boundaries. Figure 1.2 (a) and (c) show that the observations located outside the control/decision boundary would be declared as out of control (or belonging to the out-of-control class). Figure 1.2 (b) indicates that the observations located on the right side of the decision boundary (i.e., linear classifier) would belong to the out-of-control class. It is clear that the T^2 control boundary plays the same role as the decision boundaries in the supervised and semisupervised learning methods, that is to detect or isolate the data class of interest.

A major advantage of control charts in SPC is to provide a practical approach to facilitate process monitoring and diagnosis. Data mining techniques have shown their effectiveness in solving different problems such as nonparametric problems, variable selection problems, and autoregressive problems. By combining data mining approaches with control charts, we gain a new class of process monitoring and diagnostic tools. This dissertation extends the application scope of both SPC and data mining techniques.

1.4 Outline of this Dissertation

The main chapters of this dissertation study distinctive approaches to the integration of data mining and SPC. To aid understanding, each of the main chapters

contains literature reviews, methodologies, numerical studies and results, concluding remarks, and references. Chapter 2 introduces a classification-based control chart to take advantage of available out-of-control observations to improve detection efficiency. Chapter 3 proposes an approach to combine one-class classification methods and control charts. A fault isolation approach for interpretation of out-of-control signals from one-class classification-based control charts is then proposed in Chapter 4.

The organization of the chapters is as follows:

Chapter 2: Classification-Based Control Charts (PoC charts) for Monitoring Multivariate Processes– Section 2.1 contains the introduction and review of the relevant literature. Section 2.2 describes the algorithm of the proposed PoC chart. Section 2.3 demonstrates the feasibility and effectiveness of the PoC chart through simulation results. Section 2.4 illustrates the implementation of the PoC chart using a case study with a real dataset. Section 2.5 discusses the preliminary results of implementing the proposed method on nonnormal data and its extension to exponentially weighted PoC charts. Section 2.6 presents concluding remarks for the chapter.

Chapter 3: Integration of One-Class Classification Methods and Control Charts– Section 3.1 introduces the chapter with a review of the literature on the topic. Section 3.2 presents SVDD-based control charts. Section 3.3 proposes a new K^2 chart that is based on the k-nearest neighbors data description algorithm. Section 3.4 provides a simulation study. Section 3.5 shows the capability of the proposed control charts in phase I application. Section 3.6 concludes the chapter with suggestions pertinent to the research findings and to more research.

Chapter 4: A Nonparametric Fault Isolation Approach Through One-Class Classification-Based Statistics in Multivariate SPC– Section 4.1 introduces fault isolation in multivariate SPC. Section 4.2 reviews existing fault isolation methods. Section 4.3 describes K^2 charts to facilitate discussion of the decomposition of K^2 statistics for fault isolation that are proposed in Section 4.4. Section 4.5 conducts experimental studies on the proposed fault isolation method. Section 4.6 demonstrates the use of

the proposed method through a real dataset. Section 4.7 presents conclusions reached through work covered in the chapter.

Last but not least, Chapter 5 summarizes this dissertation and lays out a path for future research.

CHAPTER 2

CLASSIFICATION-BASED CONTROL CHARTS FOR MONITORING MULTIVARIATE PROCESSES

2.1 Introduction

Statistical process control (SPC) is one of the widely used techniques for quality control. The main objective of SPC is to quickly detect the occurrence of special cause variations, so that corrective action may be taken before quality deteriorates and defective units are produced [5]. SPC utilizes control charts that monitor the performance of a process over time to maintain the process in-control. Univariate control charts are devised to monitor the quality of one process variable. Nowadays, a system usually involves a number of process variables that can be highly correlated with each other. Although individual univariate control charts can be applied to each individual variable, this may lead to inefficient and unsatisfactory conclusions for multivariate problems [2].

A number of studies have been devoted to developing multivariate control charts that are effective in terms of quickly detecting both small and large shifts of a mean vector while maintaining low false alarm rates [29][30]. The most widely used multivariate control chart is Hotelling's T^2 control chart [12]. Hotelling's T^2 control charts (T^2 charts) are efficient for monitoring a multivariate process because all process variables can be simultaneously monitored in a single chart that plots the T^2 statistics, computed from the following equation:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}), \tag{2.1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are a sample mean vector and a sample covariance matrix, respectively, estimated from in-control observations. For monitoring new observations, the corresponding T^2 statistic of each new observation is calculated based on $\bar{\mathbf{x}}$ and $\mathbf{S}$

in Equation 2.1. If the T^2 statistic of the observation is statistically large, then this observation is considered to be out of control. In other words, the status of new observations are determined by measuring how far these observations are from the scaled-mean estimated from in-control observations. Traditional control charts, including T^2 charts assume that the in-control group is the only population and can be used for measuring the degree of abnormality of new observations. This assumption has restricted the applicability of many multivariate statistical methods that can take advantage of available out-of-control data. In reality, there always are out-of-control data. Despite the clear fact that efficiency can be improved if the additional information is used, little effort has been made to use this information to develop control charts [31][32].

Recently, Hwang et al. [33] firstly attempted to convert the classical control chart problem into a supervised classification problem. They proposed to generate the out-of-control data from independent uniform distributions where the parameters are estimated by the maximum and minimum values of each variable plus/minus one standard deviation. The combination of the in-control and out-of-control data allows the application of two-class supervised classification methods. This approach has been shown to be useful when the relationships between the process variables are highly nonlinear and the process variables are a mix of continuous and categorical. However, because the out-of-control data generated from uniform distributions are scattered randomly across both the out-of-control and in-control the regions, the accuracy of the out-of-control label is somewhat questionable. Hu et al. [34] extended the idea of Hwang et al. [33] and obtained the control boundary using process knowledge on the specific faults. This study suggested simulating the out-of-control data based on the specific shift direction.

In the present study we extend previous works [33][34] and propose a collection of classification-based control charts that takes advantage of available out-of-control data. The proposed chart is constructed by plotting the values of probability of

class (PoC) as monitoring statistics and thus called PoC charts. PoC is based on the probability that an observation belongs to the class. Moreover, by adjusting the control limit, one can readily control type I and type II error rates. It should be mentioned that Hu et al. [34] also briefly discussed the possibility of using the estimated probability of out of control (one of the PoCs) from the classification method for monitoring a process. However, they did not clearly indicate that PoC can be used as a monitoring statistic for a multivariate control chart. We believe that this is an important SPC component of multivariate control charts that enable use of a univariate statistic to simultaneously monitor multiple process variables in a single graphical chart. Further, their study did not fully explore a capability to establish the control limit that indicates out-of-control observations.

Unlike the previous work that utilized artificially-generated out-of-control data [33][34], the proposed PoC chart assumes that out-of-control data are available (although there are few) and takes advantage of them. We believe this is a reasonable assumption because we do not live in a zero-defect world, and a real process always experiences change that can lead to out-of-control observations. However, if out-of-control information is unavailable, Phase I analysis can be performed on the historical dataset prior to the PoC chart. Observations that exceed the control limit of the T^2 chart established by the historical data are successively removed and can be labeled as out-of-control. It is true that these removed observations may not fully represent the out-of-control patterns. However, these are obviously different from in-control observations and thus can be labeled differently from in-control observations. It should be noted that the main intention of this paper is not to compare the performance between the PoC charts and traditional multivariate control charts (e.g., T^2 charts) because it is somewhat obvious that the performance can be improved using additional information (out-of-control information). The main purpose is to develop a new collection of efficient monitoring methodologies through their integration with data mining algorithms and explain the relationship between the decision boundary (con-

trol boundary) from classification methods and the control limits in the multivariate control charts.

2.2 PoC Charts

Various types of statistics can be used to construct control charts, where a statistic is defined as any measurable function of the sample. In the univariate process, the values of the sample average and the values of the sample range (the difference between the largest and smallest observations) are used as the monitoring statistics to construct the $\overline{X}$ and R charts, respectively. In the multivariate process, the T^2 statistic defined in Equation 2.1 is used to construct the T^2 chart. Aforementioned control charts belong to Shewhart control chart. For the example of non-Shewhart control chart, the exponentially weighted moving average chart uses an exponentially weighted moving average as the monitoring statistic [10].

The proposed PoC chart would be constructed by plotting a monitoring statistic called "Probability of Class." PoC can be defined as a predicted probability that an observation belongs to a certain class estimated from the supervised classification model. Let "PoC-In" be the probability that an observation belongs to the in-control class and "PoC-Out" be the probability that an observations belongs to the out-of-control class. Either PoC-In or PoC-Out can be used as the monitoring statistic for the PoC chart. Note that any classification methods can be used in the construction of PoC charts as long as they provide the predicted PoC. These methods include linear discriminant analysis, k-nearest neighbors, logistic regression, and classification trees, etc. Note that the PoC chart is a distribution-free procedure by when nonparametric classification models are employed. The next two subsections illustrate PoC charts based on two classification methods that are widely used in practice.

2.2.1 PoC charts Based on Linear Discriminant Analysis

We first illustrate the PoC chart based on the linear discriminant analysis (LDA) model. LDA can be performed by using the discriminant function [19]. Let $f(\mathbf{x}|\omega_i)$ be the conditional probability density function (pdf) for a random variable $\mathbf{x}^T = [x_1, x_2, ..., x_p]$ given class i, and let $P(\omega_i)$ be the prior probability of class i. According to Bayes' theorem, the pdf of the posterior distribution can be approximated by

$$f(\omega_i|\mathbf{x}) \propto f(\mathbf{x}|\omega_i)P(\omega_i). \tag{2.2}$$

LDA assumes $f(\mathbf{x})$ follows the p-dimensional multivariate normal distribution, which is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)}, \tag{2.3}$$

where μ is the mean vector and $\mathbf{\Sigma}$ is the $p \times p$ covariance matrix. By taking a log of Equation 2.2 and applying Equation 2.3, the following log-likelihood function is obtained:

$$\ln f(\omega_i|\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T\mathbf{S}_i^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{S}_i| + \ln\pi_i, \tag{2.4}$$

where $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$, and π_i are the estimators of population mean $\mu_\mathbf{i}$, covariance matrix $\mathbf{\Sigma_i}$, and the numbers of observations in class i. From Equation 2.4 by omitting the terms that are independent of i, the linear discriminant function becomes

$$\ln f(\omega_i|\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}_i)^T\mathbf{S}_p^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) + \ln\pi_i, \tag{2.5}$$

where $\mathbf{S}_p$ is the pooled covariance on the assumption that all populations have the same covariance matrix. The conditional probability that an observation $\mathbf{x}$ belongs to class i (η_{ω_i}) can be obtained by simply taking the exponential of Equation 2.6.

$$\eta_{\omega_i} = \pi_i e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{S}_p^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)}. \tag{2.6}$$

The LDA-PoC chart plots η_{ω_i}, a PoC statistic, of each observation over sample number or time.

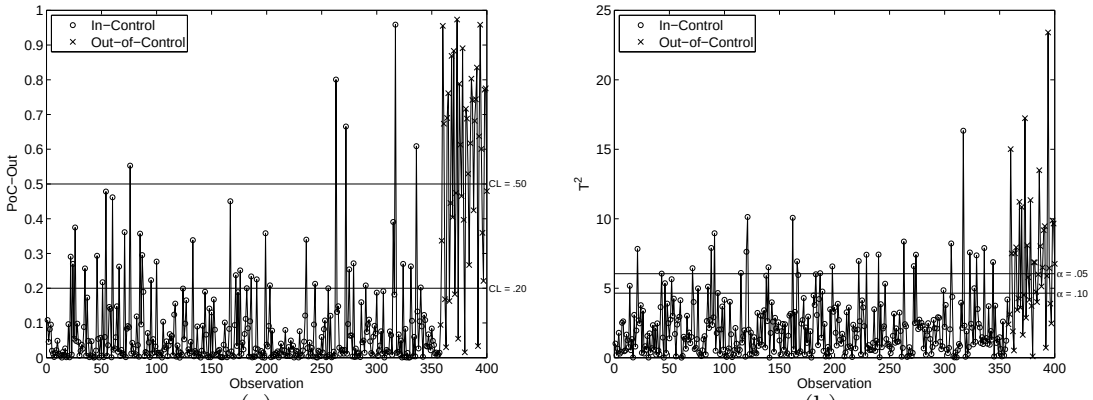


Figure 2.1. The (a) LDA-PoC chart and (b) T^2 chart from a bivariate normal distribution.

Simulated data were generated to illustrate the LDA-PoC chart. First, we generated 540 in-control and 60 out-of-control observations (with mean shifted by one standard deviation for all variables) from a bivariate normal distribution as training (Phase I) data. These data were used to establish the control limits for future monitoring. Note that only in-control observations were used to construct the control limit for the T^2 chart, while both in-control and out-of-control observations were used for the LDA-PoC chart. Next, we generated 400 additional observations (first 360 are in control and last 40 are out of control) as testing (Phase II) data. PoC-Out and T^2 statistics were calculated and plotted respectively on (a) the LDA-PoC chart and (b) the T^2-chart in Figure 2.1. It can be seen that a number of false alarms are observed in the plots because we used the small-mean-shift dataset. In order to clearly understand how the control limits can be established and adjusted, we show the two-dimensional plots of the original values, displaying the control boundaries from the LDA-PoC and T^2 charts (Figure 2.2). In T^2 charts, the elliptical contours of a bivariate normal distribution were obtained by the following equation:

$$(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) = \frac{p(n+1)(n-1)}{n(n-p)}F_{\alpha,p,n-p}, \tag{2.7}$$

where p and n are the numbers of variables and observations, respectively, and $F_{\alpha,p,n-p}$ is an F distribution with p and $(n-p)$ degrees of freedom. The size of the contour

depends on α on the right hand side of Equation 2.7. Hence, by changing α, one can adjust the control boundary (or control limit). Figure 2.2 displays two elliptical contours in which the corresponding α values are 0.1 and 0.05.

In LDA-PoC charts, control boundaries are established by the following linear equation:

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_p^{-1}x - (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \ln\left(\frac{\pi_2}{\pi_1}\right) + \ln\left(\frac{\gamma_{(1|2)}}{\gamma_{(2|1)}}\right) = 0, \qquad (2.8)$$

where $\bar{\mathbf{x}}_1$ and π_1 are the sample mean and the probability that an observation comes from class 1. Likewise, $\bar{\mathbf{x}}_2$ and π_2 are obtained from class 2. $\gamma_{(1|2)}$ represents the misclassification cost when an observation from class 2 (out-of-control class) is incorrectly classified as class 1 (in-control class). $\gamma_{(2|1)}$ represents the misclassification cost when an observation from class 1 (in-control class) is incorrectly classified as class 2 (out-of-control class). In LDA-PoC charts, the control boundary can be adjusted by specifying the misclassification cost. Thus, misclassification cost plays the same role as α in T^2 control charts. Figure 2.2 shows that how the control boundaries in the PoC chart are adjusted by misclassification cost, $\gamma_{(1|2)}$. By adjusting $\gamma_{(1|2)}$ from 0.5 to 0.2, more out-of-control points are detected. A higher level of $\gamma_{(1|2)}$ increases false positives and yields a larger type I error rate, but decreases the type II error rate.

2.2.2 PoC Charts Based on a k-Nearest Neighbors Algorithm

The control boundary of the LDA-PoC chart is established based on a normality assumption. However, many modern process data violate this normality assumption. Another limitation posed by linear control boundary in LDA-PoC chart is lack of detecting a fault generated from multiple directions. To addressed these issues, we illustrate the construction of the PoC chart based on a k-nearest neighbors (kNN) algorithm, a popularly used nonparametric approach. Figure 2.3 shows an example of boundaries from LDA and kNN from the data having a fault with multiple directions. It can be clearly seen that kNN can successfully discriminate between in-control and out-of-control observations by providing nonlinear control boundaries,

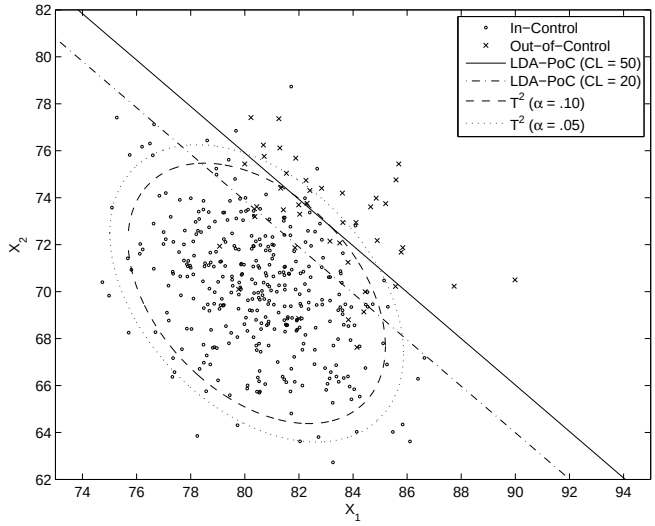


Figure 2.2. Control boundaries from the LDA-PoC chart and the T^2 chart from a bivariate normal distribution.

while LDA cannot. A kNN algorithm predicts the class of an object by analyzing its k nearest neighbors within the training data [17]. Figure 2.4 shows a simple example for calculating PoC-Out from the kNN algorithm. Let "o" and"x" be the actual in-control and out-of-control observations from a training set. Let "$*$" be a new testing observation that needs to be monitored. PoC-Out of this testing observation with $k = 3$ is calculated by the proportion of the out-of-control observations among the number of nearest neighborhood observations. Thus, in this example, PoC-Out is $\frac{1}{3}$.

In general, let $\omega^{(n)}, n = 1, 2, ..., k$ be the classes of the k observations from the training set that are nearest to the new testing observation. The probability that this observation belongs to class i, η_{ω_i}, can be calculated by

$$\eta_{\omega_i} = \sum_{n=1}^{k} \frac{\boldsymbol{I}(\omega^{(n)} = i)}{k}, \tag{2.9}$$

where $\boldsymbol{I}$ is the indicator function that returns the value 1 if the argument is true; otherwise 0. The values of η_{ω_i}, which correspond to PoC, are plotted to construct the control chart. In the kNN algorithm, the size of nearest neighbor, k, affects the performance of the kNN-PoC chart. One typical way is to determine the best k that leads to the minimum misclassification rate. In this example, we used k=30. Figure

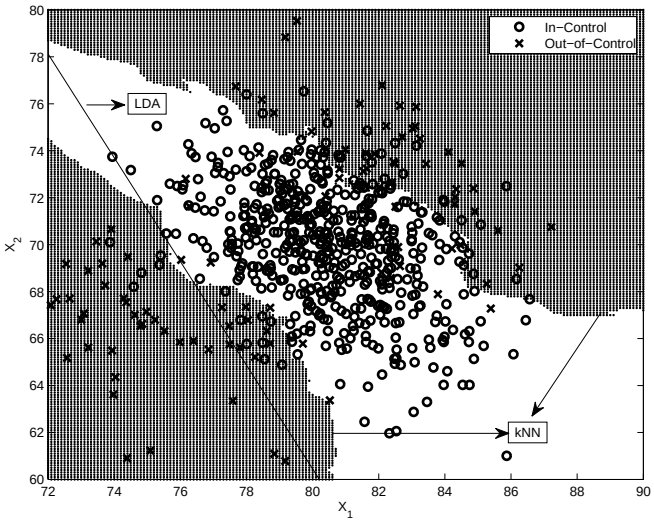


Figure 2.3. An example of control boundaries from LDA and kNN algorithms from a bivariate normal data having a fault with multiple directions.

2.5 illustrates the kNN-PoC (k=30) chart based on the same simulated data used in the previous section. The monitoring statistics PoC-Out are computed and plotted.

Like LDA-PoC charts, control boundaries of kNN-PoC charts can be adjusted by imposing a misclassification cost on each class. This can be explained by:

$$\gamma_{(1|2)} \sum_{n=1}^{k} \frac{\boldsymbol{I}(\omega^{(n)} = 1)}{k}, \tag{2.10}$$

where $\gamma_{(1|2)}$ is the misclassification cost when an observation from class 2 is incorrectly classified as class 1. Figure 2.6 shows that how the nonlinear control boundary of kNN is changed by the user-specified misclassification rate. The observations that are located in the shaded area are detected to be out-of-control. The kNN-PoC chart can detect more out-of-control observations as the control limit is changed from 0.5 to 0.2.

The potential issue of kNN-PoC is the scaling problem if we use non-scaled distances (e.g., Euclidean, city block) in a kNN algorithm because these distances depend on scale. In this case, the original variables should be scaled first. Note that there is implicit standardization in T^2 and LDA.

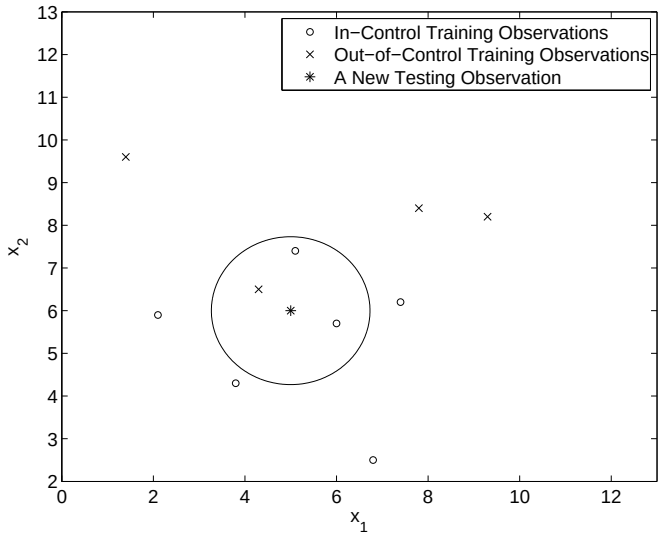


Figure 2.4. Illustration of calculating PoC-Out in the k-nearest algorithm.

2.3 Simulation Study

2.3.1 Simulation Scenarios

A simulation study was conducted to evaluate of the performance of PoC charts. For each simulation run, we generated an observation with nine variables based on a multivariate normal distribution with the mean vector μ_0 and the covariance matrix Σ_0, shown in Figure 2.7. The mean vector and covariance matrix were arbitrarily generated using MATLAB Statistics Toolbox (www.mathworks.com).

On the assumption that the number of out-of-control observations is much smaller than in-control observations, we assigned 10 percent of the simulated data to be out-of-control. To be specific, we generated 180 in-control and 20 out-of-control data points as the training (Phase I) set. We also generated additional 900 in-control and 100 out-of-control data points as the testing (Phase II) set. We measure the performance of control charts in terms of type I and type II errors. To generate the out-of-control data, four types of mean shifts (i.e., very small, small, medium, and large mean shifts) were considered. Note that we do not consider the variance change. Let μ_1 and Σ_1 be the mean and covariance matrix of the out-of-control data. The summary of simulation scenarios is described as follows:

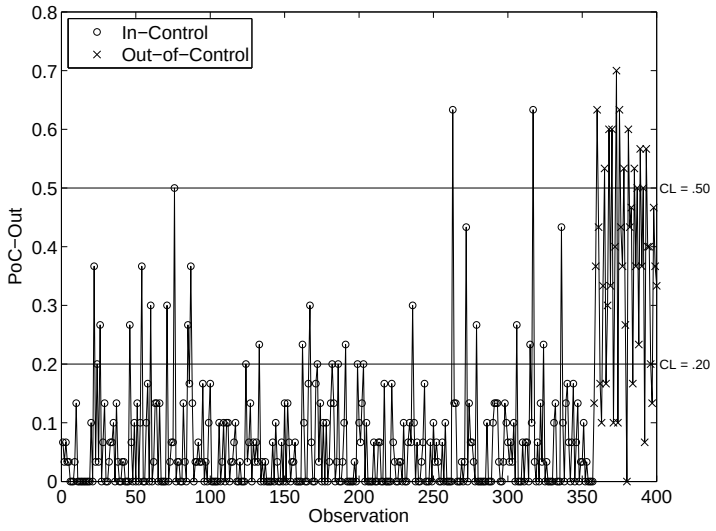


Figure 2.5. The kNN-PoC chart (k=30) of a sample from a bivariate normal distribution.

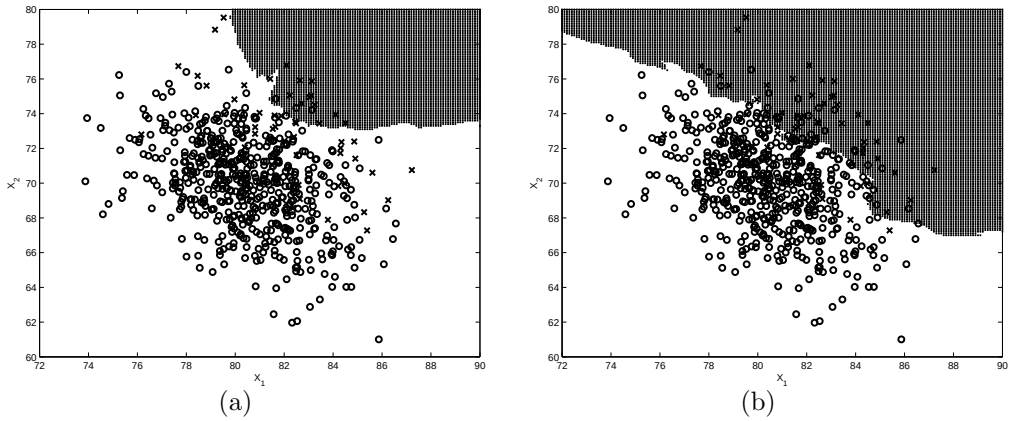


Figure 2.6. Control boundaries (a) CL = .50 and (b) CL = .20 of kNN-PoC charts (k=30) from a bivariate normal distribution.

$$\mu_0 = \begin{bmatrix} 68.128 & 37.948 & 83.180 & 50.281 & 70.947 & 42.889 & 30.462 & 18.965 & 19.343 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 8.6591 & 0.6301 & -0.1224 & -0.0188 & 0.6115 & -0.5917 & 1.5151 & 0.6003 & 1.1259 \\ 0.6301 & 6.3300 & 0.1095 & -0.3532 & -0.0043 & 0.9545 & -1.3764 & 0.5465 & -1.0935 \\ -0.1224 & 0.1095 & 6.0932 & 0.4329 & 0.5467 & -0.5517 & -0.5579 & -1.0738 & 1.0042 \\ -0.0188 & -0.3532 & 0.4329 & 8.2311 & -0.8174 & 0.5933 & -0.0533 & 1.4631 & 0.3283 \\ 0.61152 & -0.0043 & 0.5467 & -0.8174 & 6.6258 & -0.8168 & -0.3281 & 0.2996 & 0.1790 \\ -0.5917 & 0.9545 & -0.5517 & 0.5933 & -0.8168 & 5.0132 & 0.8866 & 0.2010 & -0.0503 \\ 1.5151 & -1.3764 & -0.5579 & -0.0533 & -0.3281 & 0.8866 & 7.3607 & 0.04585 & -0.8459 \\ 0.6003 & 0.5465 & -1.0738 & 1.4631 & 0.2996 & 0.2010 & 0.04585 & 8.1906 & 0.4657 \\ 1.1259 & -1.0935 & 1.0042 & 0.3283 & 0.1790 & -0.0503 & -0.8459 & 0.4657 & 4.0982 \end{bmatrix}$$

Figure 2.7. The mean vector μ_0 and the covariance matrix Σ_0 used in the simulation study.

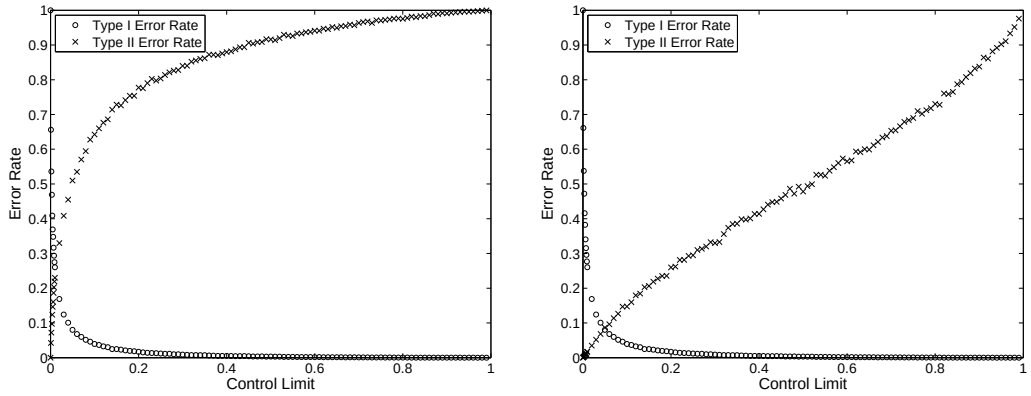


Figure 2.8. Type I and II error rates based on the control limits of LDA-PoC charts for (a) S1 and (b) S2 scenarios.

- S1 (very small mean shift): $\mu_1 = \mu_0 + 0.5\sigma_0$, $\Sigma_1 = \Sigma_0$,
- S2 (small mean shift): $\mu_1 = \mu_0 + 1\sigma_0$, $\Sigma_1 = \Sigma_0$,
- S3: (medium mean shift): $\mu_1 = \mu_0 + 2\sigma_0$, $\Sigma_1 = \Sigma_0$,
- S4: (large mean shift): $\mu_1 = \mu_0 + 3\sigma_0$, $\Sigma_1 = \Sigma_0$.

We used these simulated datasets to illustrate PoC charts in the following three subsections.

2.3.2 Effect of Control Limit in PoC Charts

As can be seen from earlier sections, the control limit of PoC charts can be adjusted by the user-specified misclassification cost. Figure 2.8 illustrates how actual type I and type II error rates are controlled by the control limit, indicated in the x-axis. The average values of type I and type II error rates from 100 simulation runs are presented against the different control limits. We displayed the results for only S1 and S2 scenarios because the error rates for S3 and S4 scenarios are too small to visualize. It can be seen that decreasing the control limit yielded larger type I error rates but produced smaller type II error rates.

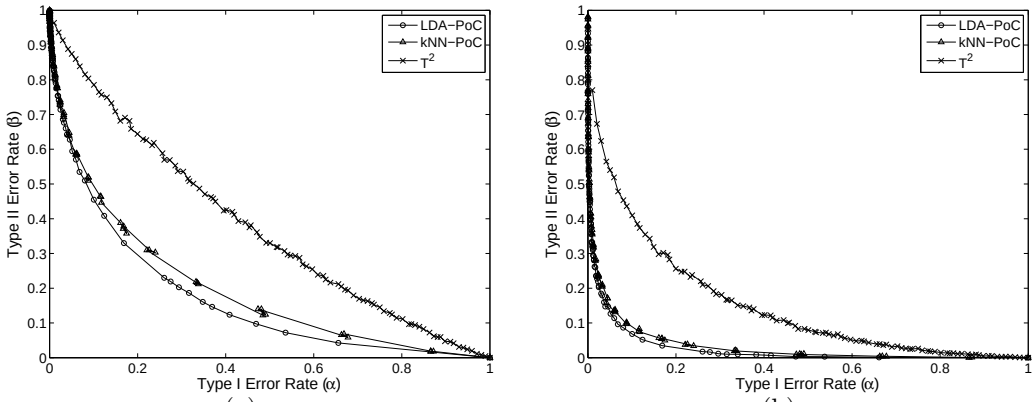


Figure 2.9. Behavior of type I and II error rates of the LDA-PoC, kNN-PoC, and T^2 charts for (a) S1 and (b) S2 scenarios.

2.3.3 Comparison Between PoC Charts and T^2 Charts

We compared the average values of type I and type II error rates from 100 simulations among LDA-PoC, kNN-PoC, and T^2 charts. In general, the performance of control charts can be measured by type II error rates given similar type I error rates. That is, the control chart method that yields a lower type II error rate can be considered as a better method. Figure 2.9 displays the behavior of type I and II error rates of the LDA-PoC, kNN-PoC, and T^2 charts. The result shows that the LDA-PoC and kNN-PoC charts produced smaller type II error rates than the T^2 chart, given similar type I error rates in the S1 and S2 scenarios. The same patterns were obtained from the S3 and S4 scenarios but the results are not shown here because the error rates are too small to clearly visualize. It should be noted that our simulation data were generated from the multivariate normal distribution and thus, the LDA-PoC chart produced the better results than the kNN-PoC, although not to a significant degree.

2.3.4 Effect of the Training Set Size

The size of training set and the proportion of out-of-control observations in the training set may affect the performance of the PoC chart. The simulation was

conducted using the LDA-PoC chart under the small-mean-shift scenario (S2). We considered five different sizes of the training set (i.e., 200, 400, 600, 800, and 1000), each with five different proportions of out-of-control observations (i.e., 0.01, 0.05, 0.10, 0.20, and 0.30). Table 2.1 shows the average errors from 100 simulation runs for different sizes of the training set and proportions of the out-of-control observations. The control limits were selected at the level that approximately gave type I error rate at 0.10. In general, larger training sets with higher proportions of out-of-control observations yielded smaller type II error rates. However, there are some interesting patterns between the size of training set and the proportion of out-of-control observations. Figure 2.10 displays a three-dimensional contour plot that facilitates the interpretation of this pattern. The x- and y-axes indicate the size of the training set and the proportion of out-of-control observations, respectively. The z-axis (the values on the contour plot) indicates the type II error rates. If the proportion of out-of-control observations is higher than 0.1, a similar type II error rate can be obtained regardless of the size of the training set. In other words, if the proportion of out-of-control observations is large enough, which is at 0.1 in this simulation, the performance of the PoC chart is not significantly affected by the size of training set. Similarly, with a training set of 1000, the PoC chart produced the same errors regardless of the proportion of out-of-control observations. In this simulation, with only 10 out-of-control observations (1%) out of 1000, the PoC chart works fine. Overall, this plot can be used as a way to determine the size of the training set and the proportion of out-of-control observations for achieving the targeted performance (e.g., type II error rate).

2.4 Case Study

In this section we illustrate PoC charts with a real dataset from a Wisconsin breast cancer study [35]. This dataset contains 569 observations, which is not time ordered. Each observation is characterized by 30 continuous variables and a two-class

Table 2.1. Average type I error rates (α) and type II error rates (β) of the LDA-PoC chart (the standard errors are shown inside the parentheses) with different sizes of training set and different proportions of out-of-control in the training set (S2 scenario)

Proportion of Out-of-Control		Size of Training Set				
		200	400	600	800	1000
0.01	α	.0718	.0815	.1013	.1015	.1015
		(.0045)	(.0039)	(.0042)	(.0038)	(.0035)
	β	.2120	.1340	.0910	.0840	.0530
		(.0151)	(.0112)	(.0095)	(.0094)	(.0076)
0.05	α	.1107	.1101	.1111	.1127	.1122
		(.0038)	(.0026)	(.0024)	(.0020)	(.0018)
	β	.0792	.0598	.0568	.0534	.0498
		(.0046)	(.0041)	(.0033)	(.0029)	(.0031)
0.10	α	.1136	.1115	.1142	.1153	.1145
		(.0025)	(.0021)	(.0017)	(.0017)	(.0017)
	β	.0569	.0534	.0481	.0474	.0469
		(.0028)	(.0027)	(.0017)	(.0024)	(.0023)
0.20	α	.1185	.1155	.1139	.1165	.1149
		(.0023)	(.0017)	(.0016)	(.0014)	(.0013)
	β	.0529	.0487	.0449	.0444	.0426
		(.0021)	(.0017)	(.0015)	(.0016)	(.0015)
0.30	α	.1143	.1157	.1166	.1127	.1146
		(.0021)	(.0016)	(.0016)	(.0014)	(.0013)
	β	.0505	.0482	.0442	.0467	.0457
		(.0015)	(.0013)	(.0012)	(.0012)	(.0013)

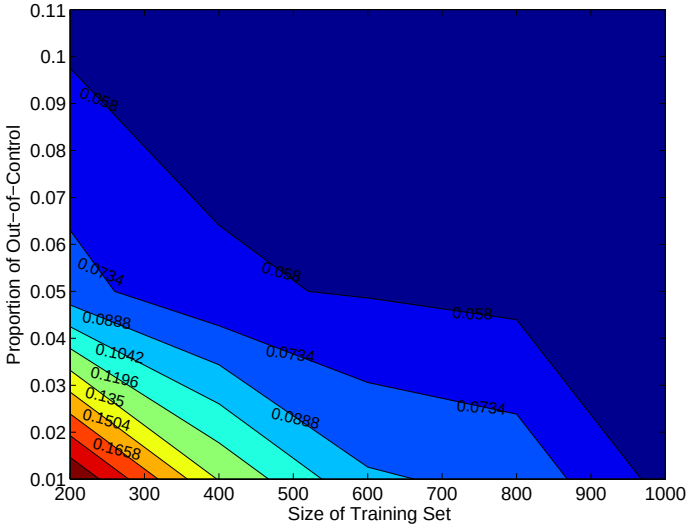


Figure 2.10. The contour plot of the type II error rates with different sizes of training set and proportions of out-of-control.

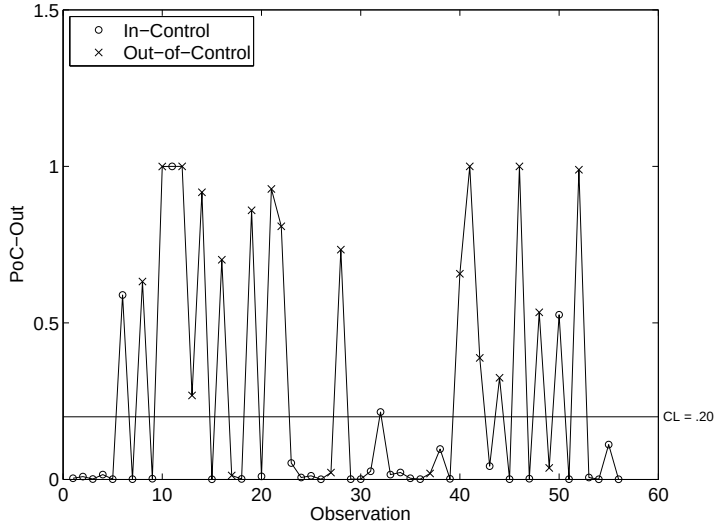


Figure 2.11. A supervised control chart of Wisconsin breast cancer data.

response variable (375 benign and 212 malignant). Without loss of generality, we considered the benign observations as in-control and the malignant observations as out-of-control.

The performances of the LDA-PoC and T^2 charts are reported in Table 2.2. To avoid the bias problem posed by the different choices of the training and testing sets, we used ten-fold cross validation to compute the error rates. The same proportion of out-of-control observations was used in each round of ten-fold cross validation. Figure 2.11 shows the LDA-PoC chart constructed using one of the testing sets from the ten-fold cross validation process. The observations that exceed the control limit (0.1) were considered to be out of control. We compared the LDA-PoC chart with the T^2 chart in terms of type II error rates given a similar type I error rate. The method that gives the lower type II error rate would be considered as the better method. We considered two cases, the first being a direct application of class labels and the second being the application of the recursive T^2 method to generate the labels. The recursive T^2 method eliminates the abnormal observations from the unlabeled data. In the construction of the LDA-PoC chart, the remaining and eliminated observations provide the labels for in-control and out-of-control, respectively. Table 2.2 shows that

Table 2.2. Average type I error rates (α) and type II error rates(β) of the LDA-PoC and T^2 charts on Wisconsin breast cancer data from ten-fold cross validation (the standard errors are shown inside the parentheses)

Case	Method	α	β
Use class information	LDA-PoC	.0287	.0426
		(.0061)	(.0132)
	T^2	.0745	.1232
		(.0122)	(.0161)
Not use class information	LDA-PoC	.2039	.1660
		(.0228)	(.0259)
	T^2	.2067	.1991
		(.0219)	(.0234)

the LDA-PoC chart produced smaller type II error rates than the T^2 chart in both cases, demonstrating the usefulness of PoC charts.

2.5 Discussion

2.5.1 Effect of Nonnormality on PoC Charts

PoC charts are robust to nonnormality when used with nonparametric classification methods, such as kNN. For illustrating the nonnormal case, we generated a binary banana-shaped dataset using MATLAB codes available from PRTools (www.prtools.org). This dataset does not obviously satisfy any existing parametric distributions (Figure 2.12). LDA-PoC and kNN-PoC charts were constructed with 180 in-control and 20 out-of-control observations, and a T^2 chart was constructed with the same 180 in-control data that were used in PoC charts. Further, we generated an additional 900 in-control and 100 out-of-control observations as the testing set. Figure 2.12 shows the control boundaries (estimated from the training set) of the three control charts embedded in a two-dimensional plot of the testing set. The kNN-PoC chart creates a nonlinear control boundary that well separates in-control and out-of-control observations generated from the nonnormal distribution. We displayed the resulting α and β from the LDA-PoC, kNN-PoC, and T^2 charts. The result indicated that the kNN-PoC charts outperformed the LDA-PoC and T^2 charts that

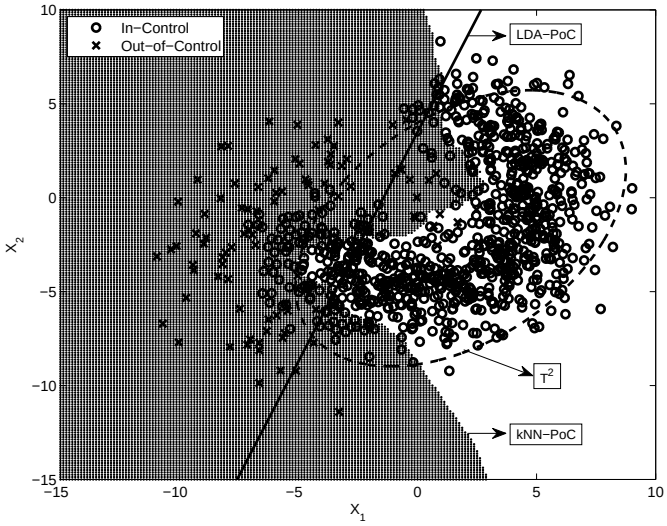


Figure 2.12. Control boundaries of LDA-PoC, kNN-PoC, and T^2 charts from a non-normal distribution dataset.

require the normality assumption (Figure 2.13). Our future work in this direction will involve a more extensive simulation study that explores various scenarios.

2.5.2 Exponentially Weighted PoC Charts

Shewhart's types of control charts are known to be insensitive to small process shifts because Shewhart control charts only use the information of the current observation and ignore the information of the past observations. Exponentially weighted moving average (EWMA) control charts use the information of both current and previous observations and thus increase the sensitivity of detecting small process shifts [10]. In multivariate processes, T^2 charts can be easily extended to multivariate EWMA control charts [36]. Hu and Runger [37] extended supervised control charts with artificial contrasts by incorporating time-weighted information. Similarly, one can readily incorporate the exponentially weighted factor (λ in Equation 2.11) into the PoC chart to improve the performance in detecting small process shifts.

$$z_t = \lambda \text{PoC-Out}_t + (1 - \lambda) z_{t-1}, \tag{2.11}$$

where $0 < \lambda \leq 1$. Note that a smaller value of λ can detect smaller shifts.

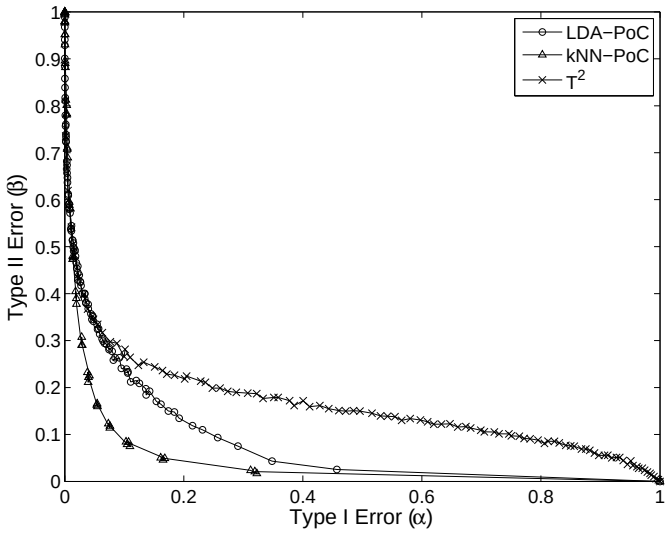


Figure 2.13. α and β of LDA-PoC, kNN-PoC, and T^2 charts from a nonnormal distribution set.

Table 2.3. Comparison of the LDA-PoC and EW-LDA-PoC charts in terms of ARL_0 and ARL_1

	LDA-PoC		EW-LDA-PoC	
Scenarios	ARL_0	ARL_1	ARL_0	ARL_1
S1	179.01	9.88	179.16	5.30
S2	273.36	2.06	270.15	1.34

Preliminary analysis was performed to demonstrate the feasibility of the exponentially weighted LDA-PoC (EW-LDA-PoC) charts. We compared the EW-LDA-PoC chart with the LDA-PoC chart under only the very-small-mean-shift (S1) and small-mean-shift scenarios (S2). We did not report the performance results under S3 (medium mean shift) and S4 (large mean shift) because we are particularly interested here in the performance under (very) small process shifts. Average run length (ARL), the average number of observations needed to signal a change was used to measure the performance. ARL_0 and ARL_1 were calculated under in-control and out-of-control processes, respectively. In general, we prefer the procedure that yields lower ARL_1 given the similar values of ARL_0. Here we arbitrarily choose $\lambda = 0.4$ for EW-LDA-PoC chart. Table 2.3 shows that the EW-LDA-PoC charts yields lower

ARL_1 than the LDA-PoC chart in both scenarios (S1 and S2). This demonstrates the effectiveness of using the exponentially weighted factor in PoC charts to improve the detection of (very) small process shifts. The future study will investigate the effective design of EW-PoC charts to give general recommendation to determine the control limits and the parameter of the chart (e.g., λ).

2.6 Conclusions

We have proposed PoC charts that combine classification algorithms with traditional control chart techniques. PoC charts take advantage of available out-of-control data. The label of out of control can be obtained either directly from the dataset or from a Phase I analysis. Then PoC statistics of individual observations are computed from classification algorithms and plotted. Further, the control limits of PoC charts can be adjusted by the user-specified misclassification cost. Experimental results with the simulated and real data showed the feasibility and the effectiveness of the proposed PoC chart. Although the main aim of this paper is not to compare PoC charts with traditional multivariate control charts, we found that PoC charts outperforms Hotelling's T^2 charts, especially in the small- and medium-mean-shift cases. The performance of PoC charts may depend on the size of training set and the amount of out-of-control data. A simulation study was conducted to examine the performance (type II error rates) of PoC charts under various scenarios of the size of training set and the proportion of out-of-control observations. The result indicated that a small proportion of out-of-control observations is sufficient to achieve the desired performance.

In summary, the advantage of using PoC charts is sixfold.

- PoC charts take advantage of available out-of-control data (or from Phase I study) to improve the detection efficiency in the complex process.

- The PoC, a monitoring statistic of PoC charts, is a univariate statistic that summarizes all process variables. Hence, PoC charts monitor multiple variables simultaneously in a single control chart.
- Because the POC is the probability values, the range of PoC charts is between 0 and 1. This provides better visualization especially in the large-mean-shift case.
- PoC charts combined with nonparametric classification methods do not require the normality assumption.
- PoC charts can handle multivariate process data containing both discrete and continuous values.
- PoC charts can readily interpret out-of-control signals using many of the procedures that have already been developed for variable selection.

CHAPTER 3

ONE-CLASS CLASSIFICATION-BASED CONTROL CHARTS FOR MULTIVARIATE PROCESS MONITORING

3.1 Introduction

Statistical process control (SPC) is one of the widely used techniques for quality control. The basic objective of SPC is to quickly detect the occurrence of special cause variation, so that the process can be investigated and corrective action may be taken before quality deteriorates and defective units are produced [4]. One of the important tools in SPC is a control chart that monitors the performance of a process over time to maintain the process in-control. In general, control chart problems in SPC can be divided into two phases [38][5]. Phase I analysis tries to isolate the in-control (normal) data from an unknown historical dataset and establish the control limits. Phase II analysis monitors the process using control charts derived from the "cleaned" in-control dataset from Phase I analysis. With a simple plot of a set of monitoring statistics that are derived from the original samples, the control chart can effectively determine whether or not a process is in a state of control. Examples of monitoring statistics include the sample average and the sample range. In addition to the monitoring statistics, another important component of control charts is control limits, which often are calculated based on the probabilistic distribution of the monitoring statistic.

Hotelling extended the univariate control chart to handle multivariate problems [39]. Hotelling's T^2 chart (T^2 chart) is a multivariate control chart that can monitor a multivariate process efficiently. T^2 charts use the T^2 statistic computed from the following equation:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}), \tag{3.1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are a sample mean vector and a sample covariance matrix determined from the in-control (Phase I) data. The T^2 statistic measures the distance between an observation and the scaled-mean estimated from the in-control data. Given that $\mathbf{x}$ follows a multivariate normal distribution, the T^2 statistic follows an F distribution [2][6]. In T^2 charts, the quantile 1-α of the F distribution is used as the control limit, where α is the user-specified level of significance. It is known that T^2 charts can effectively control type I and type II error rates when the underlying distribution of the process data is the multivariate normal distribution [6]. However, the distributional assumption of T^2 charts restrict their applicability to the nonnormal data, which can be found in many modern industries. A number of nonparametric control charts have been developed to address the limitation of the distributional assumptions [40][41][42][43]. However, no consensus exists about which of them best satisfies all conditions encountered in modern process systems. A detailed review of nonparametric control charts is beyond the scope of this paper.

As the limitations of current SPC techniques become increasingly obvious in the face of ever more complex processes, data mining algorithms, because of their proven capabilities to effectively analyze and manage large amount of data, have the potential to resolve the challenging problems in SPC. Despite the enormous popularity of data mining studies that have been conducted on a variety of applications, data mining techniques have not been thoroughly studied for application to control chart problems. In particular, one-class classification methods share a common goal with control charts because both methods assume that the in-control group (target group) is the only population and can be used for measuring the degree of abnormality of new observations.

Several studies have been undertaken recently with the goal of implementing one-class classification algorithms as an alternative to traditional control charts. Sun and Tsung [44] proposed kernel distance-based charts (K charts) based on a support vector data description (SVDD) algorithm. SVDD is a modified version of the origi-

nal support vector machines (SVMs) for solving one-class classification problems. K charts use a monitoring statistic derived from the distance between the new observation and the decision boundary generated by the SVDD algorithm. The control limits of K charts are established and adjusted from a parameter in the SVDD algorithm. Sun and Tsung's study revealed that K charts perform better than T^2 charts when the data deviates from normality. Kumar et al. [45] used another one-class SVM technique to construct robust K charts through normalized monitoring statistics. They showed that, in addition to the flexibility of nonnormal data, robust K charts can efficiently handle autocorrelated process data. Further, one-class SVM-based control charts have been applied to detect anomalies in computer-networking applications [46]. It is clearly laudable that the aforementioned studies proposed to use the monitoring statistic from the one-class SVM method. Thus, the construction of the charts does not require any distribution assumptions. However, they did not suggest an efficient way to establish the control limits, one of the major components in control charts.

This paper makes contributions in two aspects. First, we propose an efficient way to establish the control limits necessary to improve the existing one-class SVM-based control charts. Second, we propose new one-class classification-based control charts based on a k-nearest algorithm. Simulation studies were conducted to demonstrate the effectiveness of the proposed approaches in both the Phase I and Phase II analyses.

3.2 Support Vector Data Description (SVDD)-Based Control Charts

3.2.1 The SVDD Algorithm

An SVM is one of the supervised learning algorithms popularly used for both regression and classification problems. SVMs use geometric properties and obtain a separating hyperplane by solving a convex optimization problem that simultaneously minimizes the generalization error and maximizes the geometric margin between the

classes [20]. Nonlinear SVM models can be constructed from kernel functions that include linear, polynomial, and radial basis functions. SVDD is a mixture of SVM and the data description method for solving one-class classification problems [25]. SVDD provides a hypersphere boundary around the data. A brief summary of the SVDD algorithm is as follows: Let $\mathbf{a}$ be the center of the hypersphere. Let R^2 be the radius of the hypersphere (i.e., the distance from $\mathbf{a}$ to the boundary). Let $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$, for $i = 1, 2, \ldots, N$ be a sequence of p-variate training (target) observations. SVDD boundaries are constructed to minimize the volume of the hypersphere while maximizing the training observations captured by the hypersphere [25]. That is, the problem is to:

$$\text{Minimize } R^2 + C \sum_{i=1}^{N} \xi_i, \tag{3.2}$$

with the constraint:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \tag{3.3}$$

where $\xi_i > 0$ is the slack variable that allows $\mathbf{x}$ to be outside the hypersphere. R^2 is the distance from $\mathbf{a}$ to the boundary of the hypersphere. C controls the trade-off between the volume of the hypersphere and the misclassification errors. Tax and Duin [25] defined a user-specified parameter f that represents the fraction of the training data outside the decision boundary.

$$f = \frac{1}{NC}, \tag{3.4}$$

where N is the number of target observations. For instance, 80% of the training data points are supposed to be included in the SVDD boundary constructed with $f = 0.20$. When f is increased from 0.20 to 0.30, the volume of the hypersphere becomes smaller but the misclassification error in the target class becomes larger.

The problem in (3.2) can be solved by the following Lagrangian:

$$L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \{R^2 + \xi_i - (\|\mathbf{x}_i - \mathbf{a}\|^2)\} - \sum_{i=1}^{N} \gamma_i \xi_i, \tag{3.5}$$

where $\alpha_i \geq 0$ and $\gamma_i \geq 0$ are the Lagrange multipliers. Setting partial derivatives of L with respect to R, $\mathbf{a}$, and ξ_i and set to zero provides the following constraints:

$$\sum_{i=1}^{N} \alpha_i = 1, \tag{3.6}$$

$$\mathbf{a} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \tag{3.7}$$

$$\alpha_i = C - \gamma_i. \tag{3.8}$$

When substituting these constraints to (3.5), the optimization problem becomes:

$$L = \sum_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{ij} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j). \tag{3.9}$$

The solution, the set of α_i, $i = 1, 2, \ldots, N$, can be obtained by maximizing (3.9) subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{N} \alpha_i = 1$.

Like conventional SVM, the SVDD algorithm can generate more flexible decision boundaries by replacing inner product with kernel functions. For example, the following Gaussian kernel function can be replaced with the inner product in (3.9):

$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{S^2}), \tag{3.10}$$

where $S > 0$ is the width of the Gaussian kernel that controls the complexity of the SVDD boundary. Given a testing data point $\mathbf{z}$, D^2 that measures the distance between $\mathbf{z}$ and the center, $\mathbf{a}$ can be calculated by the following equation:

$$D^2 = K(\mathbf{z} \cdot \mathbf{z}) - 2 \sum_i \alpha_i K(\mathbf{z} \cdot \mathbf{x}_i) + \sum_{ij} \alpha_i \alpha_j K(\mathbf{x}_i \cdot \mathbf{x}_j). \tag{3.11}$$

For classification, a new observation $\mathbf{z}$ is classified as the target when D^2 is less than or equal to R^2.

To illustrate the control boundaries of SVDD, we generated a banana-shaped dataset using a MATLAB code available from PRTools [47]. The control boundaries with different values of parameters (f and S) in the SVDD algorithm were constructed from 180 in-control training observations (i.e., Phase I data). Figure 3.1 shows different SVDD boundaries embedded in two-dimensional plots of Phase I data. It can

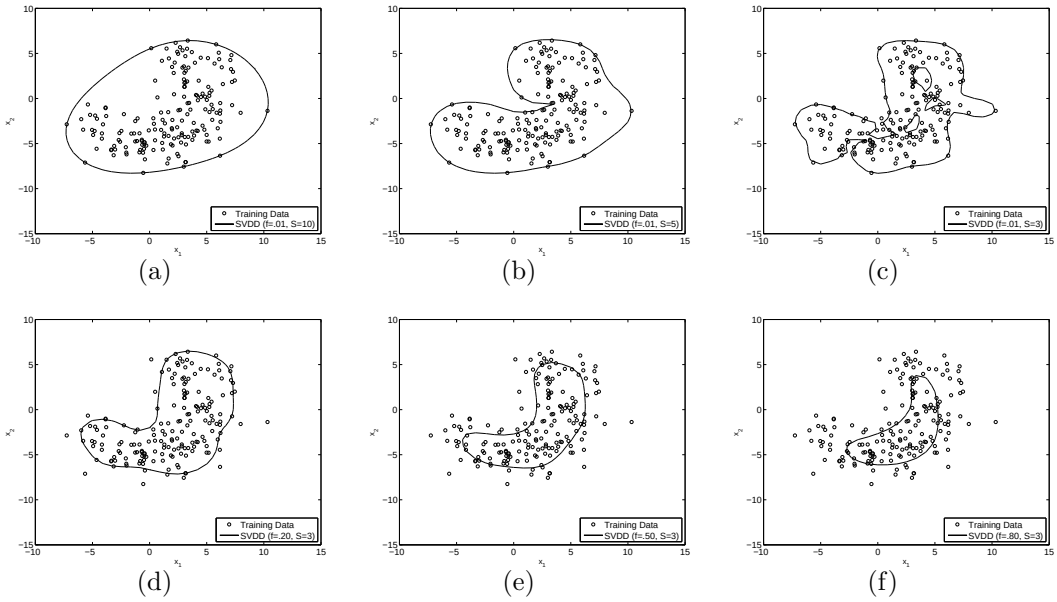


Figure 3.1. Control boundaries of SVDD obtained from different values of parameters; (a) f=0.01 and S=10, (b) f=0.01 and S=5, (c) f=0.01 and S=3, (d) f=0.20 and S=3, (e) f=0.50 and S=3, and (f) f=0.80 and S=3.

be seen from Figures 3.1 (a), (b), and (c) that given the same f value (f=0.01), the shape of the control boundary becomes smoother with larger S. One can choose an appropriate S that balances a tradeoff between oversmoothness and undersmoothness of the control boundary. In the present study, we tried some potential values of S and find the one that yields the smallest type I and type II error rates. Given the same S value (S=3), Figures 3.1 (c), (d), (e), and (f) show that the control boundary becomes tighter to the volume centroid with the larger f.

3.2.2 Existing Control Chart Methods based on the SVDD Algorithm

Several studies have implemented one-class classification methods in SPC problems. Sun and Tsung [44] proposed K charts to handle nonnormality problems by using the kernel distances obtained from the SVDD algorithm. They proposed to establish and adjust the control limits of the K chart by using f (or C), one of the parameters of the SVDD algorithm. Kumar et al. [45] proposed robust K charts, which are similar to K charts but use normalized kernel distances. One-class SVM-

based control charts were applied for anomaly detection in computer networks [46]. Although the aforementioned control charts use slightly different monitoring statistics, they are all based on the one-class SVM method.

One-class SVM (OCSVM)-based control charts can be constructed by plotting monitoring statistics (D^2) that measure the distance between new observations and the center of the hypersphere. The control limits (R^2) of OC-SVM charts are determined by f (or C). In other words, error rates in OC-SVM charts are adjusted by f. Large f values tend to yield larger a type I error rate because the algorithm utilizes less training data inside the boundary.

Figure 3.2 displays a T^2 chart and two OC-SVM charts corresponding to the control boundaries in Figures 3.1 (b) and (c). In these figures, the monitoring statistics of 400 Phase II data were plotted (the first 360 are in control and the last 40 are out of control). Note that the control limits of these charts were established by 180 Phase I data. In OC-SVM charts, it is interesting to observe that the user-specified f value affects not only the determination of the control limits, but also the calculation of the monitoring statistic. Note that two totally different control charts were obtained by the changing the value of f from 0.01 to 0.20 (Figures 3.2 (c) and (d)). This clearly demonstrates that f is inappropriate for establishing the control limits in OC-SVM charts. This limitation can be explained by Figure 3.1, showing that completely different control boundaries were obtained by changing the value of f from 0.01 to 0.20. As a consequence, an observation detected as out of control (or in control) may no longer be detected as out of control (or in control) as a reaction to the use of different values of f. In contrast, T^2 charts use the controlling value α that is independent of the monitoring statistic, T^2. Thus, the ellipse boundary of T^2 always captures more out-of-controls and yields a higher type I error rate with a larger α (Figures 3.1 (b) and (c)). Further, the same values of monitoring statistics are plotted in the T^2 chart regardless of α (Figure 3.2 (a)).

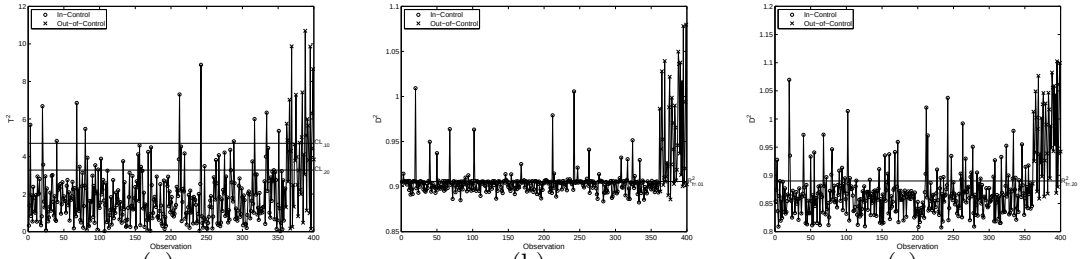


Figure 3.2. The (a) T^2 chart, (b) OC-SVM chart with f=0.01 and S=3, and (c) OC-SVM chart with f=0.20 and S=3.

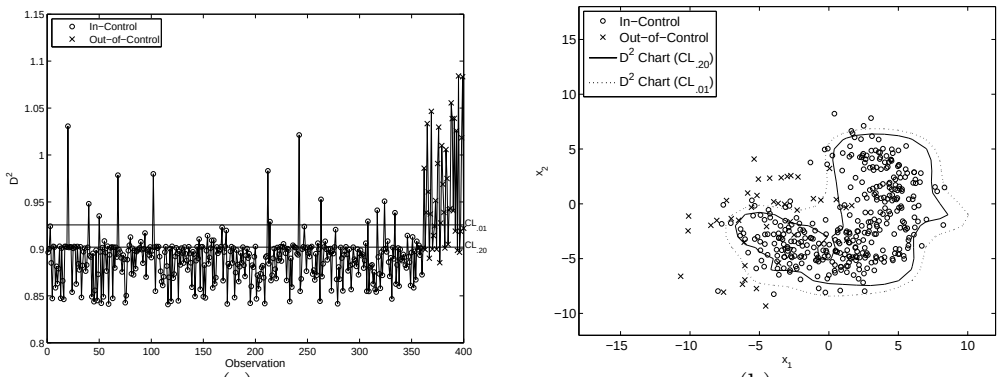
3.2.3 New Design Strategy of OC-SVM Charts

To address the limitations of the current OC-SVM control charts, we propose a new design strategy to establish the control limits in OC-SVM charts. We call the proposed chart D^2 charts. The control limits of D^2 charts are established and adjusted based on a quantile estimated by the bootstrap method, a widely used resampling method [48]. More precisely, let $D^2_{j1}, D^2_{j2}, \ldots, D^2_{jN}$ be a sequence of N monitoring statistics from j^{th} bootstrap sample (for $j = 1, ..., M$). Given a controlling value α ($0 < \alpha \leq 1$) and the ordered D^2 values, $D^2_{j(1)} < D^2_{j(2)} < \ldots < D^2_{j(N)}$, $\sum_j D^2_{j(i)}/M$ is used as the control limit where i is a roundup number of $N \cdot \alpha$. In other words, the control limits are established by using the D^2 values at 100×(1-α)th bootstrap percentile.

In summary, D^2 charts can be constructed as follows:

1. Specify the parameters f and S of the SVDD model from the training set.
2. Compute the D^2 statistic of each Phase I observation using (3.11).
3. Establish the control limits based on 100×(1-α)th percentile of the D^2 statistics estimated by the bootstrap method.
4. Monitor Phase II observations: Declare the observations out of control if the corresponding D^2 values exceed the control limit.

Figure 3.3 displays the D^2 chart and the corresponding control boundary. In the D^2 control chart, 180 in-control observations were used to estimate the control limits



(a) (b)

Figure 3.3. (a) The D^2 chart and (b) the corresponding control boundaries.

(bootstrap-based 99^{th} and 80^{th} percentiles of the D^2 statistics) and 400 D^2 statistics from Phase II observations were plotted. Figure 3.3 (b) shows the corresponding control boundary generated from the D^2 chart in Figure 3.3 (a). It can be seen that by increasing the α value from 0.01 to 0.20, more out-of-control observations were detected.

3.3 k-Nearest Neighbors Data Description (kNNDD)-Based Control Charts

The SVDD algorithm involves an optimization problem that requires a high computational load during the training process. The SVDD algorithm requires around 4.06 hours in one of our machines [1] to train the model using 4,000 bivariate observations. Because of the high computational cost, D^2 charts may not be efficient for a process that needs frequent retraining. In order to address this computational burden, we propose a new one-class classification-based control chart called a K^2 chart. The algorithm used in a K^2 chart requires about 5.42 seconds (on the same machine as the SVDD algorithm) to complete 4,000 bivariate training observations. K^2 charts are based on a k-nearest neighbors data description (kNNDD) method that solves one-class classification problems by estimating the local density of the data using a

[1] Intel(R) Core$^{\text{TM}}$ 2 Quad at 2.66 GHz with 4 GB of RAM

nearest neighbors algorithm [49][24]. A brief description of the kNNDD algorithm is presented in the following section.

3.3.1 kNNDD Algorithm

Let $\mathrm{NN}_i(\mathbf{z})$ be the i^{th} nearest neighbor training observation of a data point $\mathbf{z}$ that needs to be classified (or monitored). Let V be the volume of the hypersphere containing i nearest neighbor training observations. Let N be the size of the training set. The local density of $\mathbf{z}$ can be determined by:

$$d(\mathbf{z}) = \frac{i/N}{V\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}. \tag{3.12}$$

Similarly, the local density of $\mathrm{NN}_i(\mathbf{z})$ can be determined by:

$$d(\mathrm{NN}_i(\mathbf{z})) = \frac{i/N}{V\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}, \tag{3.13}$$

where $\mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))$ is the i^{th} nearest neighbor of $\mathrm{NN}_i(\mathbf{z})$ in the same training set. The kNNDD algorithm classifies $\mathbf{z}$ as the target class when the ratio of its local density of $\mathbf{z}$ (3.12) to the local density of $\mathrm{NN}_i(\mathbf{z})$ (3.13) is greater than or equal to 1, which can be explained as follows:

$$\frac{d(\mathbf{z})}{d(\mathrm{NN}_i(\mathbf{z}))} = \frac{\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{3.14}$$

To make the algorithm more robust, the average of k distances is considered (for $i = 1, ..., k$). Thus, (3.14) becomes:

$$\frac{\sum_{i=1}^{k} \|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{3.15}$$

In the kNNDD algorithm, the size of nearest neighbor, k, affects its performance. Figure 3.4 displays the control boundaries obtained by kNNDD with two different values of k. Decision boundary with $k = 30$ is fairly smooth compared to the control boundary obtained by using $k = 2$. One can search possible values of k and find an appropriate one that compromises a tradeoff between oversmoothness and undersmoothness of the control boundary. A previous study indicated that the proper range of k in the kNNDD algorithm is between 10 to 50 [49].

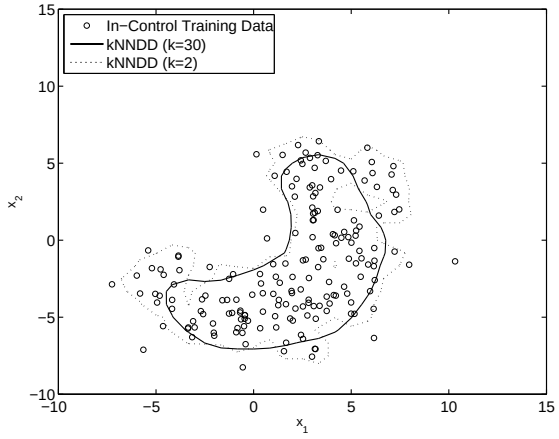


Figure 3.4. Control boundaries of kNNDD (with different k) constructed from the banana-shaped dataset.

3.3.2 K^2 Charts

To construct K^2 charts, the average distance between $\mathbf{z}$ and k nearest observations is calculated as follows:

$$K^2 = \frac{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}{k}. \tag{3.16}$$

K^2 values are then used as monitoring statistics. The control limits of the K^2 chart are obtained by the same method that we proposed in D^2 charts. Let $K^2_{j1}, K^2_{j2}, \ldots, K^2_{jN}$, for $j = 1, ..., M$ be a sequence of N monitoring statistics from jth bootstrap sample. Given a controlling value α $(0 \leq \alpha \leq 1)$ and the ordered K^2 values, $K^2_{j(1)} < K^2_{j(2)} < \ldots < K^2_{j(N)}$, $\sum_j K^2_{j(i)}/M$, is used as the control limit where i is a roundup number of $N \cdot \alpha$. The construction of K^2 charts is summarized as follows:

1. Specifiy the parameters k of the kNNDD algorithm from the training set.
2. Compute the K^2 statistic of each Phase I from (3.16).
3. Establish the control limits based on $100\times$(1-α)th percentile of the K^2 statistics estimated by the bootstrap method.
4. Monitor Phase II observations: Declare the observations out of control if the corresponding D^2 values exceed the control limits.

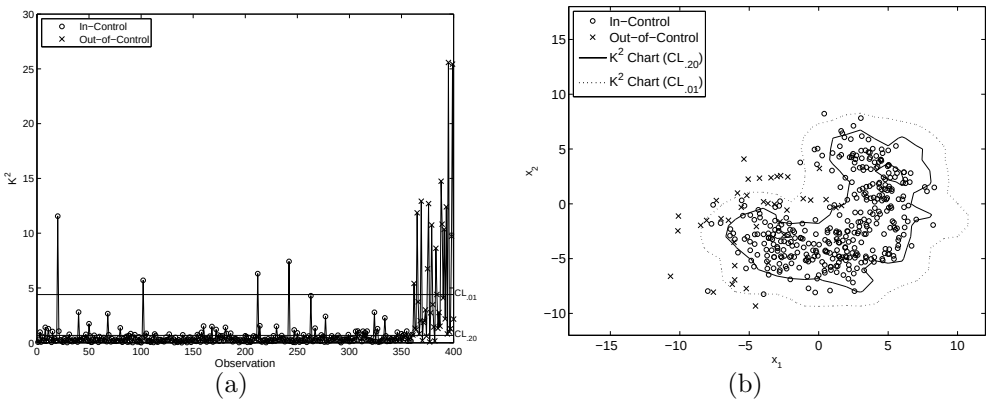


Figure 3.5. (a) The K^2 chart and (b) the corresponding control boundaries.

Figure 3.5 displays the K^2 chat (k=30) and the corresponding control boundary from the banana-shaped dataset. Two different control limits were calculated by estimated quantiles (0.99 and 0.80) from 5,000 bootstrap samples of 180 K^2 statistics. For monitoring Phase II observations, the K^2 value of each Phase II observation was plotted. Figure 3.5 (b) displays the control boundaries corresponding to the control limits embedded in a two-dimensional plot of the Phase II observations, showing that control charts become more sensitive as α increases.

3.4 Simulation Study

3.4.1 Simulation Setup

A simulation study was conducted to compare the performance among D^2, K^2, T^2, and OC-SVM charts. We generated the data based on the bivariate normal, bivariate t, and bivariate gamma and a banana-shaped dataset. For D^2 and OC-SVM charts, we used the width of Gaussian kernel, S=1 for the normal, t, and gamma cases and S=3 for the banana-shaped data. For K^2 charts, we used k=30. One thousand Phase II observations (900 in-control and 100 out-of-control) were monitored based on the control limits that were established by 200 Phase I observations. Let μ_0 and $\mathbf{\Sigma}_0$ be the mean vector and the covariance matrix of the in-control data. Let $\mu_1 = \mu_0 + \delta$ be

the mean vector of the out-of-control data. The magnitude of the shift δ is represented by the following noncentrality parameter λ:

$$\lambda = \sqrt{\delta^T \boldsymbol{\Sigma}_0^{-1} \delta}. \tag{3.17}$$

To generate the out-of-control data for the bivariate normal, bivariate t, and bivariate gamma distributions, two types of mean shifts (i.e., the medium mean shift $\lambda = 2$ and the large mean shift $\lambda = 3$) were considered. At a certain value of λ, all variables are shifted equally. Note that we do not consider the change in variance. We generated two different angles of banana shapes that represent the in-control and out-of-control data (please see [47] for more details on generating the banana-shaped dataset). The summary of simulation scenarios is described as follows:

- N_2, $\lambda = 2$: The medium-mean-shift case of the bivariate normal distribution with
  $\mu_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ and $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & .35 \\ .35 & 1 \end{bmatrix}$.
- N_2, $\lambda = 3$: The large-mean-shift case of the bivariate normal distribution with
  $\mu_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ and $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & .35 \\ .35 & 1 \end{bmatrix}$.
- $t_2(3)$, $\lambda = 2$: The medium-mean-shift case of the bivariate t distribution with three degrees of freedom.
- $t_2(3)$, $\lambda = 3$: The large-mean-shift case of the bivariate t distribution with three degrees of freedom.
- $Gam_2(1, 1)$, $\lambda = 2$: The medium-mean-shift case of the bivariate gamma distribution with the shape and scale parameters, where both of them are one.
- $Gam_2(1, 1)$, $\lambda = 3$: The large-mean-shift case of the bivariate gamma distribution with with the shape and scale parameters, where both of them are one.
- Banana-Shaped: A banana-shaped dataset with two different angles.

3.4.2 Control Limits

In contrast to existing OCSVM charts that use the parameter f of the SVDD algorithm to adjust the control limits, the control limits of D^2 and K^2 charts are adjusted by the empirical quantile, which is estimated by the bootstrap method. Figures 3.6 and 3.7 show how actual type I and type II error rates in the D^2, K^2, T^2, and OCSVM charts are controlled by the controlling factors (α or f), indicated in the x-axes. We used the average values of actual type I and type II error rates from 100 simulation runs. The standard errors of 100 simulations are relatively small (between .02 and .06), demonstrating that 100 simulations are enough to draw the meaningful conclusion. We presented the results for only N_2 and $Gam_2(1,1)$ scenarios, respectively, as examples of normal and nonnormal cases. In general, as the controlling factor increases, all control charts produced larger type I error rates but produced smaller type II error rates. The particularly strong positive correlation between the actual type I error rate and the controlling factor is desired. The proposed D^2 and K^2 charts satisfy this condition in both normal and nonnormal cases, but T^2 charts satisfy this condition in only normal cases. In both normal and nonnormal cases, OC-SVM charts failed to provide strong linear correlation between the actual type I error rate and the controlling factor. Moreover, type I and type II error rates may not be properly controlled by f as the size of target observations goes up in OC-SVM charts. Figure 3.8 shows OC-SVM charts, constructed from the N_2 with $\lambda = 2$ scenario using 300 and 400 target observations. It can be observed that type I and type II error rates seem to be constant over the different values of f. As we defined earlier in (3.4), f represents the fraction of the target data outside the decision boundary and has an inverse relationship with the total number of target observations. Thus, with the large number of target observations, the fraction of the target data (f) plays a little role in changing the control boundary, leading to relatively constant type I and type II error rates. These demonstrate that f is an inappropriate choice as the controlling factor in OC-SVM charts.

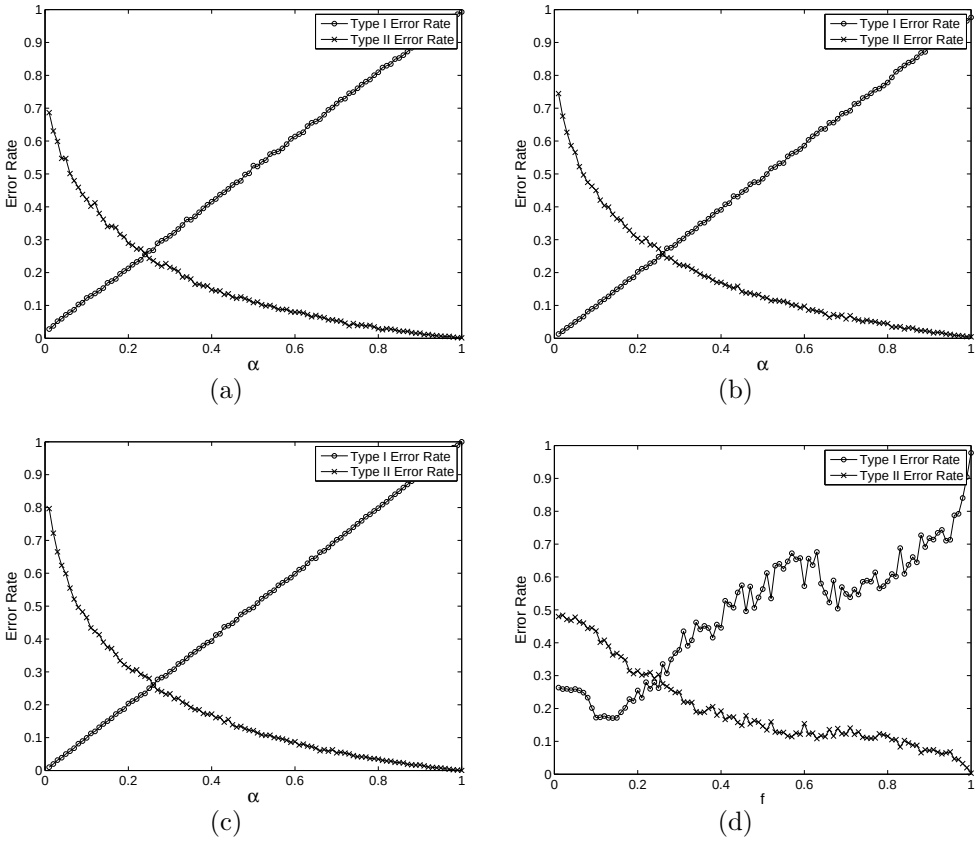


Figure 3.6. Average type I and type II error rates from (a) D^2, (b) K^2, (c) T^2, and (d) OC-SVM charts (N_2 with $\lambda = 2$ scenario).

3.4.3 Performance Comparisons

The average values of type I and type II error rates from 100 simulation runs among D^2, K^2, T^2, and OC-SVM charts were compared. The control chart that yields a lower type II error rate is considered a better method if the type I error rate is similar. Figures 3.9 displays the average rates of type I and type II error under all the simulation scenarios studied.

The result shows that the D^2 and K^2 charts produced smaller type II error rates than the T^2 chart, given similar type I error rates in the gamma and banana-shaped data scenarios. In the normal and t cases, all methods provide comparable performances. The range of standard errors of 100 simulation is between .02 and .08 for the normal, t, and Gamma cases, while much larger standard errors were obtained from OC-SVM (between .10 and .26). It should be noted that OC-SVM

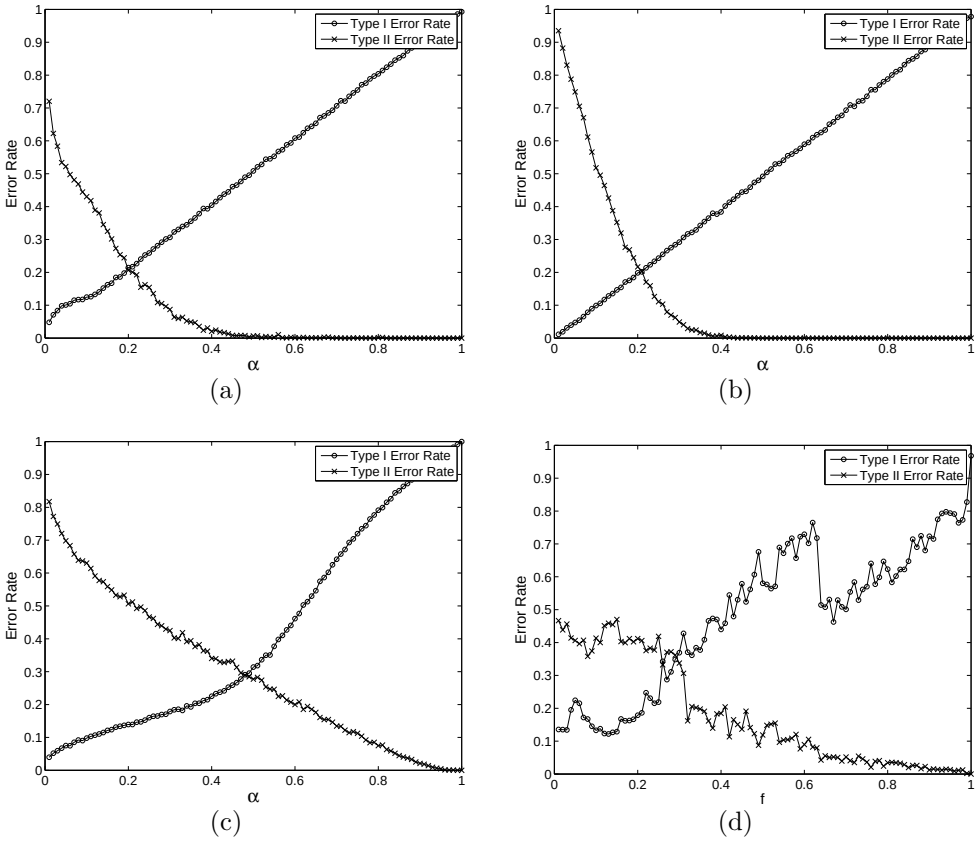


Figure 3.7. Average type I and type II error rates from (a) D^2, (b) K^2, (c) T^2, and (d) OC-SVM charts ($Gam_2(1,1)$ with $\lambda = 2$ scenario).

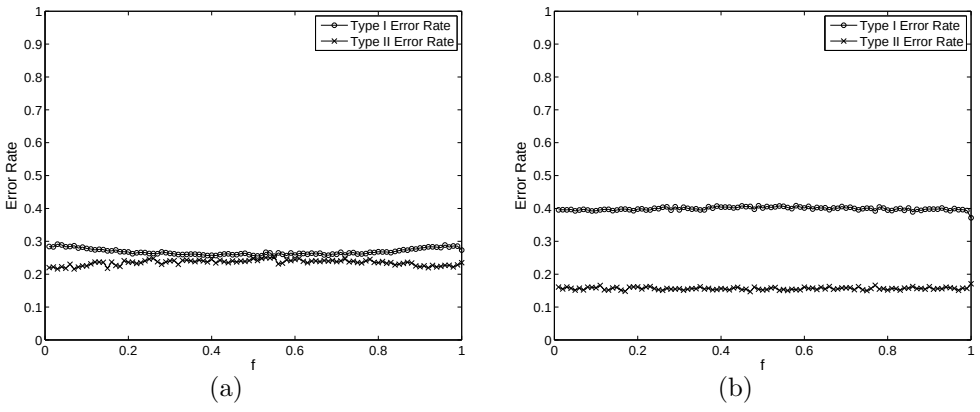


Figure 3.8. Average type I and type II error rates from OC-SVM charts when the number of Phase I observations is large upto (a) 300 observations and (b) 400 observations (N_2 with $\lambda = 2$ scenario).

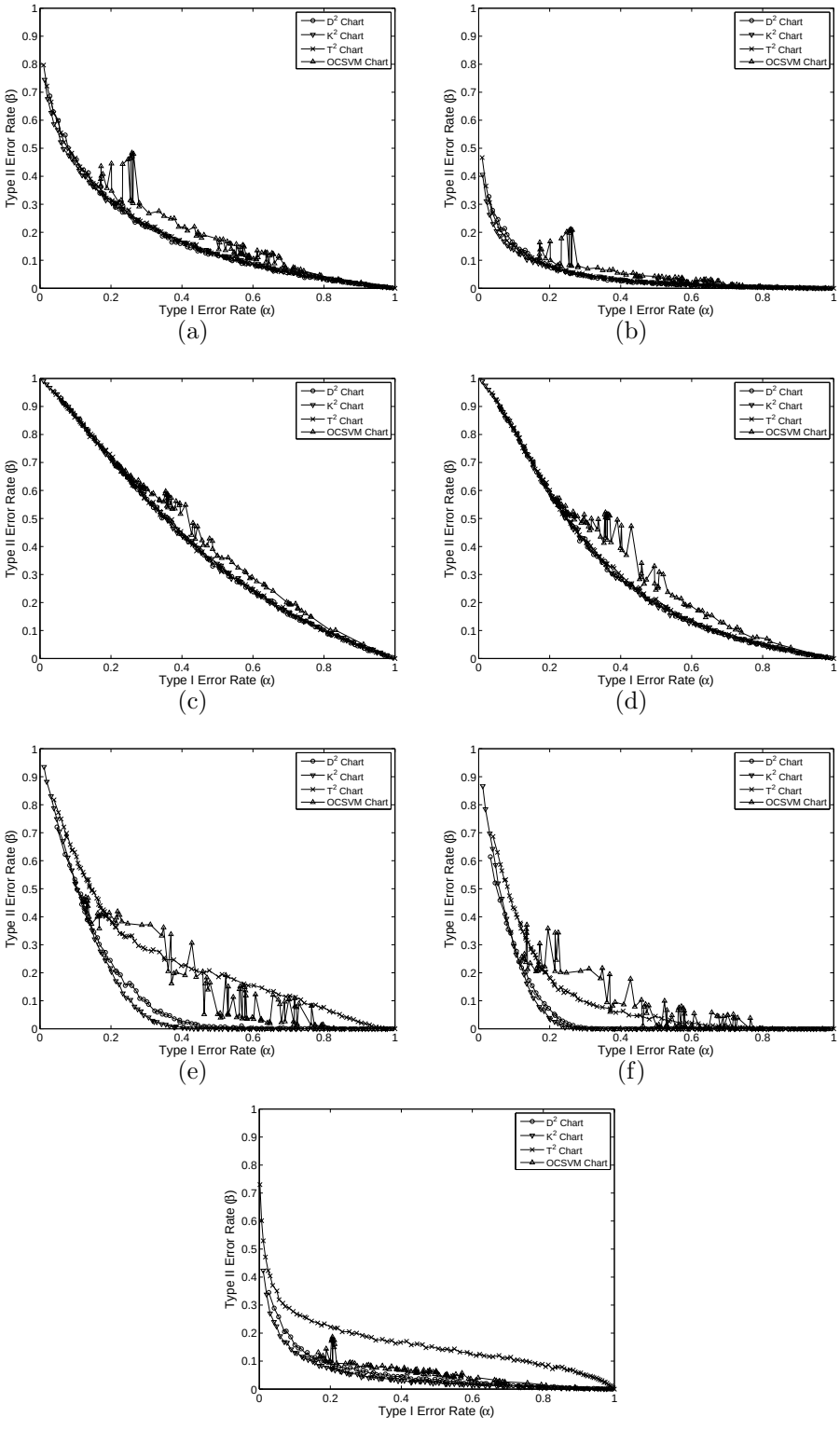


Figure 3.9. Type I and type II error rates of the D^2, K^2, T^2, and OC-SVM charts under the simulation scenarios studied; (a) N_2 with $\lambda = 2$, (b) N_2 with $\lambda = 3$, (c) $t_2(3)$ with $\lambda = 2$, (d) $t_2(3)$ with $\lambda = 3$, (e) $Gam_2(1,1)$ with $\lambda = 2$, (f) $Gam_2(1,1)$ with $\lambda = 3$, and (g) Banana-shaped scenarios.

charts produce very irregular type II error rates over type I error rates. Consequently, it is difficult to compare the performance of OC-SVM with other charts.

3.5 Phase I Application of D^2 and K^2 Charts

Phase I analysis separates the in-control data from the historical dataset, which is a mixture of the in-control and out-of-control data, in order to establish the reliable control limits for monitoring future observations. A simulation study was conducted to show the applicability of D^2 and K^2 charts for Phase I problems. We compared the performance of the D^2 and K^2 charts with the existing Phase I method that recursively removes the observations that exceed the control limits until no out-of-control observations are detected. In multivariate processes, this recursive procedure is performed by the Hotelling's T^2 control chart in the Phase I application [6].

We generated 200 historical observations from the bivariate normal, bivariate t, and bivariate gamma distributions and a banana-shaped dataset. We assigned 20 observations (out of 200) to be out of control where two different noncentrality parameters, λ=2 and $\lambda = 3$, were used for the bivariate normal, bivariate t, and bivariate gamma distributions. For the banana-shaped dataset, two different angles of banana shapes were used to represent the in-control and out-of-control data. The D^2 and K^2 charts were constructed with all 200 observations. The control charts removed the historical observations in which the statistics exceed the control limits. Analogous to D^2 and K^2 charts for Phase II analysis, $100 \times (1-\alpha)^{th}$ bootstrap percentiles of the D^2 and K^2 statistics of the historical data were used as control limits in Phase I analysis. The remaining observations were defined as in control. The observations that were actually in control but incorrectly removed were type I errors. The remaining observations that were actually out of control were type II errors.

We compared the performances of the D^2 and K^2 charts with the recursive T^2 in terms of type I and type II error rates (average values from 100 simulation

Table 3.1. Average values of type I error rate (α) and type II error rate (β) of the D^2 chart, the K^2 chart, and the recursive T^2 in Phase I application (average values of standard errors are shown inside the parentheses)

	D^2		K^2		T^2	
Scenarios	α	β	α	β	α	β
N_2, $\lambda = 2$	.1869	.3630	.1829	.3485	.1892	.3865
	(.0128)	(.0994)	(.0124)	(.1065)	(.0566)	(.1354)
N_2, $\lambda = 3$	.2124	.0800	.2183	.0825	.2137	.0915
	(.0103)	(.0674)	(.0097)	(.0561)	(.0664)	(.0810)
$t_2(3)$, $\lambda = 2$	.2219	.6830	.2144	.6790	.2248	.7165
	(.0119)	(.0932)	(.0122)	(.0970)	(.0426)	(.1071)
$t_2(3)$, $\lambda = 3$	.2038	.6290	.1995	.6375	.2069	.6230
	(.0132)	(.0949)	(.0126)	(.1013)	(.0444)	(.1436)
$Gam_2(1)$, $\lambda = 2$	.1996	.2410	.2101	.2825	.2089	.3250
	(.0171)	(.1307)	(.0140)	(.1196)	(.0596)	(.1969)
$Gam_2(1)$, $\lambda = 3$	.2204	.0380	.2208	.0715	.2335	.1200
	(.0100)	(.0556)	(.0116)	(.0905)	(.0781)	(.1482)
Banana-Shaped	.1751	.1680	.1771	.0865	.1791	.2380
	(.0118)	(.0886)	(.0127)	(.0721)	(.1026)	(.1211)

runs). Table 3.1 shows that the performances of the D^2 and K^2 charts are slightly better than recursive T^2 under the normal and t scenarios but they are comparable. Because the Hotelling's T^2 chart can effectively handle multivariate normal data, the recursive T^2 is also an appropriate method in normal distribution cases of Phase I analysis. However, in the gamma and banana-shaped data scenarios, the D^2 and K^2 charts produced smaller type II error rates than the recursive T^2 method. This clearly demonstrates that D^2 and K^2 charts are effective approaches to use for Phase I analysis in both normal and nonnormal cases.

3.6 Conclusions

We have proposed new multivariate control charts based on one-class classification algorithms. The proposed D^2 and K^2 charts obtain their monitoring statistics from the SVDD and kNNDD algorithms. The control limits are derived from the bootstrap-estimated quantile of monitoring statistics. The proposed control charts, because of their data-driven nature, can effectively describe reality, reflect the unique characteristics of the data being monitored, and require a minimal set of assump-

tions to construct a control chart. The comparative study from the simulated data shows that performances of the D^2 and K^2 charts were comparable to T^2 charts in the normal distribution case. However, D^2 and K^2 charts outperformed T^2 charts in nonnormal distribution cases. Moreover, we demonstrated the applicability and effectiveness of the D^2 and K^2 chart techniques for Phase I problems. There are several interesting directions for future research. One such direction is to extend our study to other one-class classification methodologies. Further research also can develop more efficient ways to establish control limits. A more comprehensive simulation study should be conducted to evaluate the efficacy and consequences of various scenarios, including the impact of variance changes.

CHAPTER 4

A NONPARAMETRIC FAULT ISOLATION APPROACH THROUGH ONE-CLASS CLASSIFICATION-BASED STATISTICS IN MULTIVARIATE SPC

4.1 Introduction

Statistical process control (SPC) is a widely used technique for quality control. The main objective of SPC is to detect the assignable causes so that the process can be corrected before quality deteriorates. SPC uses control charts that monitor the performance of a process over time to maintain the process in a state of statistical control. Univariate control charts are devised to monitor the quality of one process variable. In practice nowadays, however, a system usually involves a number of process variables that are correlated with each other. Although individual univariate control charts can be applied to individual variables, this may lead to inefficient and unsatisfactory conclusions for multivariate problems. Multivariate control charts take into account the relationships among variables to improve the detection of assignable causes. One of the most popular multivariate control charts is Hotelling's T^2 control chart (T^2 chart), a multivariate version of Schewart's univariate control charts. Let $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ be a p-dimensional vector represented an observation from a monitoring process. T^2 values are calculated from the following equation [12]:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}), \tag{4.1}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are a sample mean vector and a sample covariance matrix estimated from the historical dataset. The T^2 statistic follows an F distribution [12], given that $\mathbf{x}$ follows a multivariate normal distribution.

T^2 charts provide the control limits to monitor the process and detect any abnormal events allowing process improvement. However, because of the complexity of multivariate control charts, it is difficult to identify the causes of an out-of-control

alarm [14]. In other words, the multivariate control charts provide limited information about the contribution of each process variable for the out-of-control alarm. A number of fault isolation methods have been developed to address this issue. This method include T^2 statistic decomposition, U^2 statistic, adaptive regression adjusted scheme (ARA), and principal component analysis. A T^2 decomposition approach, called MTY's T^2 decomposition [2], decomposes the T^2 statistic into individual components that reflect the contribution of individual process variables responsible for the out-of-control signal. The contributed process variables can be determined based on the threshold, calculated by the quantile 1-α of an F distribution, where α is the user-specified level of significance.

Runger [50] proposed control charts using U^2 statistics to improve detection efficiency, assuming a known subset of variables contributed to the out-of-control signal. The U^2 chart approach reduces the dimensionality of the problem by determining the potential set of variables contributed at the design of the control chart. As a result, the assignable causes are initially addressed from the designing phase. Another approach to isolating the fault is to compute the relative indicator that reflects the contribution of the individual variables for the fault alarm [51]. This relative indicator approach can be considered as a special case of fault isolation using U^2 statistics when the subset of the contributed variables contains a single variable. Recently, Liu et al. [52] proposed an ARA scheme that requires an assumption on the number of contributed variables instead of the exact set of variables. The set of potential out-of-control variables is then determined through a generalized likelihood ratio test. Jackson [53] proposed to use transformed variables from principal component analysis (PCA) for constructing multivariate control charts based on T^2 and Q statistics. Various methods have been proposed to use PCA-based approaches for fault identification when an out-of-control signal is detected (e.g., [29][54]).

The aforementioned fault isolation approaches are based on the T^2 statistic for the interpretation of the out-of-control signal. As a result, all of the approaches require multivariate normality assumptions to draw reliable conclusions.

In the present study, we propose a nonparametric approach for fault isolation in multivariate SPC. The proposed approach is based on k-nearest-neighbors data description (kNNDD), one of the one-class classification algorithms [24]. Our proposed approach requires fewer assumptions than the existing fault isolation methods and thus efficiently accommodate the nonnormal datasets.

This paper is organized as follows. Section 4.2 reviews several existing methods for fault isolation in SPC. Section 4.3 describes the control chart based on the one-class classification algorithm. Section 4.4 presents the proposed K^2 decomposition algorithm for fault isolation in multivariate SPC. Section 4.5 provides a simulation study to explore the property of the proposed method and to compare it with the T^2 decomposition approach. In Section 4.6, we utilize a real dataset to demonstrate the effectiveness of the proposed algorithm. Finally, some conclusion remarks are given in Section 4.7.

4.2 Existing Fault Isolation Methods in Multivariate SPC

4.2.1 MTY's T^2 Decomposition

Mason et al. [55] provided the interpretation of an out-of-control signal given by T^2 charts. The approach (MTY's T^2 decomposition) partitions the overall T^2 statistic (Equation 4.1) into independent components as follows:

$$T^2 = T^2_{(p-1)} + T^2_{p|1,\ldots,p-1}. \tag{4.2}$$

Let the sample covariance matrix be expressed as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{x_{(p-1)}x_{(p-1)}} & \mathbf{S}_{x_{(p-1)}x_p} \\ \mathbf{S}^T_{x_{(p-1)}x_p} & s^2_p \end{bmatrix},$$

where $\mathbf{S}_{x_{(p-1)}x_{(p-1)}}$ is the $(p-1)\times(p-1)$ sample covariance matrix for the first $p-1$ variables, s_p^2 is the variance on the p^{th} variable, and $\mathbf{S}_{x_{(p-1)}x_p}$ is a $p-1$ dimensional vector containing the covariance between the p^{th} variable and the remaining $p-1$ variables. Let $\bar{\mathbf{x}}_{(p-1)}$ and $\bar{x}_p$, respectively, be the sample mean vector of the first $p-1$ variables and the sample mean of the p^{th} variable. The unconditional term $T^2_{(p-1)}$ in Equation 4.2 is computed as follows:

$$T^2_{(p-1)} = (\mathbf{x}_{(p-1)} - \bar{\mathbf{x}}_{(p-1)})^T \mathbf{S}^{-1}_{x_{(p-1)}x_{(p-1)}} (\mathbf{x}_{(p-1)} - \bar{\mathbf{x}}_{(p-1)}),$$

where $\mathbf{x}_{p-1} = [x_1, x_2, ..., x_{p-1}]^T$. The conditional term $T^2_{p|1,...,p-1}$ (Equation 4.2) is given by

$$T^2_{p|1,...,p-1} = \frac{x_p - \bar{x}_{p|1,...,p-1}}{s^2_{p|1,...,p-1}},$$

where

$$\begin{aligned} \bar{x}_{p|1,...,p-1} &= \bar{x}_p + b_p^T(\mathbf{x}_{(p-1)} - \bar{\mathbf{x}}_{(p-1)}), \\ s^2_{p|1,...,p-1} &= s_p^2 - \mathbf{S}^T_{x_{(p-1)}x_p} b_p^T, \\ b_p &= \mathbf{S}^{-1}_{x_{(p-1)}x_{(p-1)}} \mathbf{S}_{x_{(p-1)}x_p}. \end{aligned}$$

Next, the unconditional term in Equation 4.2 can be further partitioned into

$$T^2_{(p-1)} = T^2_{(p-2)} + T^2_{p-1|1,...,p-2}.$$

By iteratively partitioning in this manner, the MTY's decomposition of the T^2 statistic can be expressed as

$$T^2 = T_1^2 + T^2_{2|1} + T^2_{3|1,2} + ... + T^2_{p|1,2,...,p-1}, \tag{4.3}$$

where T_1^2 can be computed by

$$T_1^2 = \frac{(x_1 - \bar{x}_1)^2}{s_1^2}.$$

By using the permutation property, a number of decompositions of the T^2 are considered [55]. Each of the decomposed statistics follows the following F distribution with the corresponding degrees of freedom:

$$\frac{n+1}{n} \cdot F_{(1,n-1)}, \tag{4.4}$$

where n is the size of the dataset. Therefore, the decomposed T^2 statistic is determined whether it is significant, given a user-specified value α ($0 \leq \alpha \leq 1$). The major drawback of this approach is its excessive computation, although some computational loads can be reduced through a computing scheme proposed in [56].

4.2.2 Runger's U^2 Statistic and The Relative Indicator Approaches

If the subset of variables contributed to the signal is known, Runger's U^2 control charts (U^2 charts) [50] become an effective method to use. Let $\mathbf{y}$ be the vector of the subset of $\mathbf{x}$ containing variables that are not contributed to the signal. Let $\mu_{\mathbf{y}}$ and $\Sigma_{\mathbf{y}}$ be the population mean and the covariance matrix of $\mathbf{y}$. The U^2 statistic can be derived into the following from:

$$U^2 = (\mathbf{x} - \mu)^T \mathbf{\Sigma^{-1}} (\mathbf{x} - \mu) - (\mathbf{y} - \mu_{\mathbf{y}})^T \mathbf{\Sigma_{\mathbf{y}}^{-1}} (\mathbf{y} - \mu_{\mathbf{y}}). \tag{4.5}$$

The interpretation of a signal from a U^2 chart is automatically addressed as the subset of variables of $\mathbf{x}$ that is not defined in $\mathbf{y}$. However, this approach has a limitation in its usage when the subset of variables is difficult to define.

Another related fault isolation approach is to interpret a fault signal given by T^2 charts through relative indicators [51]. This approach isolates the signal by computing the indicator of the relative contribution of the j^{th} variable ($j = 1, ..., p$) to the overall T^2 statistic, $d_j T^2$.

$$d_j T^2 = T^2 - T^2_{1,2,...,j-1}, \tag{4.6}$$

where $T^2_{1,2,...,j-1}$ is the value of the T^2 statistic calculated from all variables except the j^{th} variable. An approximate suggested threshold for the $d_j T^2$ is $\chi^2_{\alpha,1}$, given an α value [51].

The relative indicator approach can be considered as a special case of using the U^2 statistic in the fault isolation. If $\mathbf{y} = [x_1, x_2, ..., x_{p-1}]$ for $\mathbf{x} = [x_1, x_2, ..., x_p]$ in the design of U^2 charts, both of the approaches will provide the same results. Furthermore, the relative indicator approach can be represented by MTY's T^2 de-

composition when considering only p conditional terms given the remaining $p-1$ variables as follows:

$$T^2_{1|2,...,p}, T^2_{2|1,3...,p}, ..., T^2_{p|1,...,p-1} = d_1T^2, d_2T^2, ..., d_pT^2. \tag{4.7}$$

4.2.3 The Adaptive Regression Adjusted Chart

Liu et al. [52] proposed the adaptive regression adjusted (ARA) chart. The ARA chart requires an assumption of the minimum number of in-control variables (that is, total number of variables - the maximum number of out-of-control variables) instead of specifying the subset of variables as required by the U^2 chart. Liu et al. [52] suggested applying an engineering judgment to determine the minimum number of in-control variables. The ARA approach determines the most likely subset of the in-control variables, $\mathbf{y}$, using a generalized likelihood ratio procedure. Let $\mathbf{z}$ be the vector of the subset of $\mathbf{x}$ containing the out-of-control variables, that is

$$\begin{aligned} [x_1, ..., x_q, x_{q+1}, ..., x_p] &= [y_1, ..., y_q, z_1, ..., z_{p-q}] \\ \mathbf{x}^T &= (\mathbf{y}^T, \mathbf{z}^T), \end{aligned}$$

where q is the minimum number of in-control variables. The sample mean ($\bar{\mathbf{x}}$) and covariance matrix ($\mathbf{S}$) of $\mathbf{x}$ are presented as follows:

$$\bar{\mathbf{x}} = [\bar{\mathbf{y}}^T, \bar{\mathbf{z}}^T]$$

$$\mathbf{S} = \begin{bmatrix} \mathbf{S_y} & \mathbf{S_{yz}} \\ \mathbf{S_{yz}} & \mathbf{S_z} \end{bmatrix}.$$

By finding the subset $\mathbf{y}$ that gives the smallest $(\mathbf{y}-\bar{\mathbf{y}})^T\mathbf{S_y^{-1}}(\mathbf{y}-\bar{\mathbf{y}})$, the most likely subset of the in-control variables for the observation can be obtained. The monitoring statistic $T^2_{\mathbf{z}|\mathbf{y}}$ of the ARA chart is computed as follows:

$$T^2_{\mathbf{z}|\mathbf{y}} = (\mathbf{z} - (\bar{\mathbf{z}} + \mathbf{S_{yz}}\mathbf{S_{yz}^{-1}}(\mathbf{y}-\bar{\mathbf{y}})))^T\mathbf{S^{-1}_{z|y}}(\mathbf{z} - (\bar{\mathbf{z}} + \mathbf{S_{yz}}\mathbf{S_{yz}^{-1}}(\mathbf{y}-\bar{\mathbf{y}}))), \tag{4.8}$$

where $\mathbf{S}_{\mathbf{z}|\mathbf{y}} = \mathbf{S}_{\mathbf{z}} - \mathbf{S}_{\mathbf{yz}}^{T}\mathbf{S}_{\mathbf{y}}^{-1}\mathbf{S}_{\mathbf{yz}}$. An out-of-control signal will be generated if $T^2_{\mathbf{z}|\mathbf{y}}$ exceeds the control limit determined by its probability distribution [52]. To isolate the signal, the individual decomposition of T^2 for $\mathbf{z}_j$ $(j = 1, ..., p-q)$ is determined:

$$T^2_{\mathbf{z}_j|\mathbf{y}} = (\mathbf{z}_j - (\bar{\mathbf{z}}_j + \mathbf{S}_{\mathbf{yz}_j}\mathbf{S}_{\mathbf{yz}_j}^{-1}(\mathbf{y} - \bar{\mathbf{y}})))^{T}\mathbf{S}_{\mathbf{z}_j|\mathbf{y}}^{-1}(\mathbf{z}_j - (\bar{\mathbf{z}}_j + \mathbf{S}_{\mathbf{yz}_j}\mathbf{S}_{\mathbf{yz}_j}^{-1}(\mathbf{y} - \bar{\mathbf{y}}))). \tag{4.9}$$

The out-of-control variables j can be determined by the large corresponding statistics $T^2_{\mathbf{z}_j|\mathbf{y}}$. A drawback of this approach is that there is no practical guide line to determine the minimum number of in-control variables.

4.2.4 Principal Component Analysis-Based Fault Isolation Method

Jackson and Mudholkar [57] proposed multivariate control charts based on PCA, which is a feature extraction method commonly used for dimensional reduction and visualization. A brief summary of the algorithm of PCA-based control charts is as follows: Let $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ be a p-dimensional vector represented an original observation from a monitoring process. Let $\mathbf{w}_j$ $(j = 1, ..., p)$ be the eigenvector (of the covariance matrix of $\mathbf{x}$) corresponding to the eigenvalue λ_j, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. The principal component $\mathbf{y}$ can be obtained from the product of the loading matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_p]^T$ and the original observation $\mathbf{x}$ as follows:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}. \tag{4.10}$$

To reduce the dimensionality, the first few principal components can be considered to represent the data without losing a significant amount of information. Let $\mathbf{y} = [y_1, y_2, ..., y_q]^T$ (for $q < p$) be the q-dimensional principal component of an observation. On the assumption that $\mathbf{x}$ follows a multivariate normal distribution, T^2 monitoring statistics based on PCA, PCA-T^2, can be obtained from the following relationship:

$$\text{PCA-}T^2 = \sum_{m=1}^{q} \frac{y_m^2}{s_m^2}, \tag{4.11}$$

where s_m^2 is the sample variance of y_m $(m = 1, ..., q)$. Next, Q statistics, which are measures of variances that are not captured by the q principal components, are

obtained from the sums of squares of the residuals corresponding to PCA as follows [57]:

$$\begin{aligned} Q &= [\mathbf{x} - \mathbf{V}^T\mathbf{y}]^T[\mathbf{x} - \mathbf{V}^T\mathbf{y}], \\ &= [\mathbf{x} - \hat{\mathbf{x}}]^T[\mathbf{x} - \hat{\mathbf{x}}], \end{aligned} \tag{4.12}$$

where $\mathbf{V}$ is the $q \times p$ matrix of the first q rows of the matrix $\mathbf{W}$.

If either the PCA-T^2 statistic (Equation 4.11) or the Q statistic (Equation 4.12) of an observation exceed their corresponding control limits derived from their reference distributions [57], the process is declared out of control. The variable contributed to the out-of-control signal from the PCA-T^2 statistic can be determined by the total contribution of the j^{th} variable, TC_j [54]:

$$\text{TC}_j = \sum_{m=1}^{q} \max(0, \frac{y_m^2}{s_m^2}\mathbf{W}_{mj}(x_j - \mu_j)), \tag{4.13}$$

where $\mathbf{W}_{mj}$ is the element at the row m and the column j of the matrix $\mathbf{W}$ and μ_j is the mean of the j^{th} variable (for $j = 1, ..., p$ and $m = 1, ..., q$). The contribution of the j^{th} variable to the signal from the Q statistic is measured through the squared prediction error (SPE) as computed by [54]

$$\text{SPE}_j = (x_j - \hat{x}_j)^2. \tag{4.14}$$

The variables that produce large TC or SPE values can be considered as the significant contributor [54].

The PCA-based process monitoring and diagnostic methods attempt to improve the detection efficiency of multivariate control charts by reducing the dimensionality. However, the PCA-based method may be ineffective when a fault occurred outside the subspace captured by the selected principal components [52].

4.3 K^2 Charts

K^2 charts use a monitoring statistic that represents the degree of being an outlier as obtained through a k-nearest-neighbors data description (kNNDD) algorithm.

The kNNDD algorithm estimates the local density of data using a nearest-neighbors algorithm to solve one-class classification problems [24][49]. Without loss of generality, the terminologies commonly preferred in SPC are used in this context, such as using the term "historical dataset" instead of "training dataset." Let $\mathrm{NN}_i(\mathbf{z})$ be the i^{th} nearest neighbor (in-control) historical observation of a new observation $\mathbf{z}$ that needs to be monitored. Let V be the volume of the hypersphere containing i nearest-neighbor historical observations. Let N be the size of the historical dataset. The local density of $\mathbf{z}$ can be computed by

$$d(\mathbf{z}) = \frac{i/N}{V\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}. \tag{4.15}$$

The local density of $\mathrm{NN}_i(\mathbf{z})$ can be computed by

$$d(\mathrm{NN}_i(\mathbf{z})) = \frac{i/N}{V\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}, \tag{4.16}$$

where $\mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))$ is the i^{th} nearest neighbor of $\mathrm{NN}_i(\mathbf{z})$ in the same historical dataset. Based on the kNNDD algorithm, $\mathbf{z}$ is more likely to be in control when its local density of $\mathbf{z}$ is higher than the local density of $\mathrm{NN}_i(\mathbf{z})$. In other words, $\mathbf{z}$ is determined as in control when the ratio of (4.15) to (4.16) is greater than or equal to 1 as shown:

$$\frac{d(\mathbf{z})}{d(\mathrm{NN}_i(\mathbf{z}))} = \frac{\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{4.17}$$

A more robust algorithm is acquired by considering the average of k distances (for $i = 1, ..., k$), which makes Equation 4.17 become

$$\frac{\sum_{i=1}^{k} \|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{4.18}$$

Prior to the construction of K^2 charts, the parameter k of the kNNDD algorithm should be determined to compromise a tradeoff between sensitivity and robustness of the algorithm. The kNNDD algorithm with a larger k would be less sensitive than the algorithm with a smaller k. A previous study suggested the proper range of k be between 10 to 50 [49].

The K^2 monitoring statistic used by the K^2 chart is then defined as the average distance between $\mathbf{z}$ and k nearest observations as follows:

$$K^2 = \frac{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}{k}. \tag{4.19}$$

The control limits of the K^2 charts are constructed based on a quantile estimation through the bootstrap method, a widely used resampling method [48]. More precisely, let $K^2_{r1}, K^2_{r2}, \ldots, K^2_{rN}$, for $r = 1, ..., M$ be a sequence of N monitoring statistics from the r^{th} bootstrap sample. Given a controlling value α ($0 \leq \alpha \leq 1$) and the ordered K^2 values, $K^2_{r(1)} < K^2_{r(2)} < \ldots < K^2_{r(N)}$, $\sum_r K^2_{r(R)}/M$, are used as the control limit where R is a roundup number of $N \cdot \alpha$. An observation is declared out of control if the corresponding K^2 values exceed the control limits.

4.4 Decomposition of K^2 Statistics for Fault Isolation

4.4.1 K^2 Decomposition Method

To interpret the fault from K^2 charts, we decompose the K^2 statistic into individual components that indicate the contribution of individual variables. Suppose that we consider a p-variate process, the overall K^2 statistic is calculated by considering all p variables in the process. The relative contribution of the j^{th} variable to the overall K^2 statistic, $K^2_{j|1,\ldots,j-1,j+1,\ldots,p}$ or d_jK^2, can be computed as follows:

$$d_jK^2 = K^2_{j|1,\ldots,j-1,j+1,\ldots,p} = K^2 - K^2_{1,\ldots,j-1,j+1,\ldots,p}, \tag{4.20}$$

where $K^2_{1,\ldots,j-1,j+1,\ldots,p}$ is the K^2 statistic calculated by considering all other variables except the j^{th} variable. The j^{th} variable corresponding to the large d_jK^2 value can be considered as the major contributor to the fault signal.

A numerical example taken from [58] was used to demonstrate the calculation of d_jK^2. Note that we chose $k = 3$ for the kNNDD algorithm. Fifty observations represent the measurements of the switch drums ($p = 5$). The first 35 observations (Obs1 to Obs35) serve as the phase I dataset to establish the control limits. Given that the fault alarm occurs at Obs48, the d_1K^2 of Obs48 can be calculated as follows:

1. By considering all five variables, the k-nearest neighbors of Obs48 ($\mathbf{z}$ = [13.065, 11.625, 14.923, 12.589, 12.446]^T) are Obs28 ($\text{NN}_1(\mathbf{z})$ = [16.615, 11.221, 14.151, 12.629, 10.601]^T), Obs8 ($\text{NN}_2(\mathbf{z})$ = [17.144, 12.254, 14.931, 13.715, 11.135]^T), and Obs1 ($\text{NN}_3(\mathbf{z})$ = [17.265, 11.788, 15.101, 13.903, 10.465]^T). Thus, the (Euclidean) distances between Obs48 to these nearest neighbors are 16.767, 20.021, and 23.349.
2. The overall K^2 value (or $K^2_{1,2,3,4,5}$) of Obs48 is then (16.767+20.021+23.349)/3 = 20.046.
3. By considering all variables except the 1^{st} variables, the k-nearest neighbors of Obs48 ($\mathbf{z}_{2,3,4,5}$ = [11.625, 14.923, 12.589, 12.446]^T) are Obs21, Obs8, and Obs28, which give $\text{NN}_1(\mathbf{z}_{2,3,4,5})$ = [11.575, 15.192, 11.809, 11.418]^T, $\text{NN}_2(\mathbf{z}_{2,3,4,5})$ = [12.254, 14.931, 13.715, 11.135]^T, and $\text{NN}_3(\mathbf{z}_{2,3,4,5})$ = [11.788, 15.101, 13.903, 10.465]^T. It can be seen that the nearest neighbors of an observation considering different sets of variables can be different. At this point, the distances between Obs48 to these nearest neighbors are 1.740, 3.382, and 4.165.
4. The $K^2_{2,3,4,5}$ value of Obs48 is then (1.740+3.382+4.165)/3 = 3.096.
5. Finally, the d_1K^2 (or $K^2_{1|2,3,4,5}$) is 20.046-3.096 = 16.950.

By computing in this manner, d_2K^2, d_3K^2, d_4K^2, and d_5K^2 are .195, .209, .999, and 7.234. By descending sorting the d_jK^2 values, the variables that mostly contributed to the fault alarm are the 1^{st}, 5^{th}, 4^{th}, 3^{rd}, and 2^{nd} variables, respectively.

Only a subset of variables can be considered in order to calculate the relative contribution of the j^{th} variable to subsets of variables. Examples of possible computations of K^2 decompositions considering subsets of variables using the switch drums data are as follows:

$$\begin{aligned} K^2_{1|2,3,5} &= K^2_{1,2,3,5} - K^2_{2,3,5} = 19.047 - 2.310 = 16.737 \\ K^2_{3|1,2,4} &= K^2_{1,2,3,4} - K^2_{1,2,4} = 12.812 - 9.621 = 3.191 \\ K^2_{5|1,4} &= K^2_{1,4,5} - K^2_{1,4} = 19.617 - 8.834 = 10.783. \end{aligned}$$

Although any interesting decompositions on subsets of variables can be computed, we recommend consideration of the contribution of the j^{th} variable to all other variables, d_jK^2 for j = 1,...,p. Most of the time, consideration only of d_jK^2 (j = 1,...,p) provides enough information on the contributed variable without excessive computation. If only the terms d_jK^2 (j = 1,...,p) are considered in our proposed approach, our procedure would be similar to Runger's relative indicator approach [51], differing only in that we apply the procedure on the K^2 statistic instead of the T^2 statistic.

4.4.2 Isolating Significant Variables from the K^2 Statistic

Once the decompositions of K^2 are obtained, a threshold can be incorporated to indicate the significant contributed variables. The threshold value of the K^2 decompositions can be calculated by a quantile estimate using the bootstrap method. Suppose that we have n training observations, we can obtain up to $N = n \cdot p$ of K^2 decompositions from the training set (when considering d_jK^2 with $j = 1, ..., p$). Let $K^2_{r1}, K^2_{r2}, ..., K^2_{rN}$, be a sequence of N K^2 decompositions from the r^{th} bootstrap sample (for $r = 1, ..., M$). By specifying a controlling value α ($0 < \alpha \leq 1$) and the ordered K^2 decomposition values, $K^2_{r(1)} < K^2_{r(2)} < ... < K^2_{r(N)}$, $\Sigma_r K^2_{r(R)}/M$ is used as the threshold where R is a roundup number of $N \cdot \alpha$. In other words, the threshold are determined by using the K^2 decomposition values at $100 \times (1 - \alpha)^{th}$ bootstrap percentile. The j^{th} variables corresponding to the d_jK^2 values that exceed the threshold are determined as the significant variables.

4.5 Simulation Study

4.5.1 Simulation Scenarios

A simulation study was conducted to evaluate the performance of the K^2 decomposition method and compare it with the T^2 decomposition method. The three-dimensional ($p = 3$) datasets generated based on the multivariate normal (N_3), mul-

tivariate log-normal ($LogN_3$), and multivariate gamma (Gam_3) were used. In the simulation, $k = 30$ was used for establishing K^2 control charts. For each simulation run, we generated 200 in-control observations and 1,000 out-of-control observations. Let μ_{I} and $\mathbf{\Sigma}_{\mathrm{I}}$ be the mean vector and the covariance matrix of the in-control data. Let $\mu_{\mathrm{O}} = \mu_{\mathrm{I}} + \delta$ be the mean vector of the out-of-control data. The magnitude of the shift is represented by the noncentrality parameter λ as follows:

$$\lambda = \sqrt{\delta^T \mathbf{\Sigma}_{\mathrm{I}}^{-1} \delta}, \tag{4.21}$$

where $\delta^T = [\delta_1, \delta_2, ..., \delta_p]$ and δ_j $(j = 1, ..., p)$ is the shift size in the j^{th} variable. At a certain value of λ (e.g., the small mean shift $\lambda = 0.5$ and the large mean shift $\lambda = 3$), the shift size in each variable δ_j $(j = 1, ..., p)$ after mean centering is provided for each case. For N_3 scenario, we generated observations based on a multivariate normal distribution with the zero mean vector and the following covariance matrix:

$$\Sigma_{\mathrm{I}} = \begin{bmatrix} 1.00 & 0.70 & 0.60 \\ 0.70 & 1.00 & 0.10 \\ 0.60 & 0.10 & 1.00 \end{bmatrix}.$$

The multivariate log-normal data ($LogN_3$ scenario) were generated using the same correlation matrix structure as used in N_3 scenario. Note that we do not consider the change in variance in either of the scenarios.

For Gam_3 scenario, we specified the shape and scale parameters of ones. Here, we considered the generalized multivariate gamma distribution as described in [59], which provides the same correlation value among each pair of variables. Regardless of the specific shifting variables, the similar degree of λ with the same number of shifting variables would provide the same results. The mean shift in negative directions is unavailable in this multivariate gamma distribution. Thus, the number of fault cases being considered in Gam_3 scenario is fewer than in the other scenarios.

4.5.2 Comparison Between the T^2 and K^2 decomposition methods

In this study, similarly to [52], correct fault isolation is defined as when the largest contribution variable (or set of variables) given by the corresponding decomposed statistic is the true fault variable. For instance, if we study the case that the second variable is shifted (e.g., $[\delta_1, \delta_2, \delta_3] = [0, .29, 0]$ for Case 2 of the N_3 scenario), the method that gives the largest corresponding statistic to the second variable (i.e., d_2K^2 or d_2T^2) will be defined as providing the correct fault isolation. We used the fault isolation error rates (one minus the correct fault isolation rates) as our performance measure. The average values of fault isolation error rates from 10,000 simulation runs of each scenario are reported in Table 4.1, Table 4.2, and Table 4.3. The standard errors from 10,000 simulations runs are less than .001.

Table 4.1 compares the fault isolation error rates between the K^2 decomposition and T^2 decomposition methods in a N_3 scenario and the results are graphically summarized in Figure 4.1. In the single-variable-shift cases, the T^2 decomposition method performed better than the K^2 decomposition method when the first variable is shifted (Case 1, 10, 19, and 28). Both methods produced the comparable results when the second variable is shifted (Case 2, 11, 20, and 29). When the third variable is shifted, the K^2 decomposition method outperformed the T^2 decomposition method. In cases in which two variables are shifted, the K^2 decomposition method performed better than the T^2 decomposition method when the first and second variables are shifted together in positive direction (Cases 4, 13, 22, and 31). On the other hand, the T^2 method performed better than the K^2 method when the first and second variables are shifted together in negative direction (Cases 7, 16, 25, and 34). When the third variable is shifted with either the first or the third variables, the K^2 method outperformed the T^2 method (Cases 5, 6, 8, 9, 14, 15, 17, 18, 23, 24, 26, 27, 32, 33, 35, and 36). As expected, both methods can correctly isolate true faulty variables as the shift size increases. The overall average error rates (taken from all 36 cases) of both methods are .5229 (the K^2 method) and .5744 (the T^2 method). It can be concluded

that the performances of the K^2 and T^2 decomposition methods are comparable in a multivariate normal case.

$LogN_3$ scenario illustrates the performances of the K^2 and T^2 methods in the case of a minor deviation from the normal distribution. Table 4.2 shows performance between the two decomposition methods in a $LogN_3$ scenario. To facilitate presentation and interpretation, the results are graphically displayed in Figure 4.2. It can be seen that the K^2 decomposition method outperformed the T^2 decomposition method when a single variable is shifted (Cases 1, 2, 3, 10, 11, 12, 19, 20, 21, 28, 29, and 30). In all of the cases in which two variables are shifted, the K^2 method performs better than the T^2 method, except when the first and the second variables are shifted together in negative directions (Cases 7, 16, 25, and 34). Both methods performed well when the shift sizes are larger. Overall, the K^2 method yielded smaller average error rate (taken from all 36 cases) than the T^2 decomposition method.

Table 4.3 compares the performances of the K^2 and T^2 methods in a Gam_3 scenario and the graphical summary of the results is displayed in Figure 4.3. In this scenario, the K^2 method outperformed the T^2 method in all cases. The overall average error rate of the K^2 method (.2945) is significantly smaller than the overall average error rate of the T^2 method (.5131). It should be noted that unlike the other scenarios, the correlation structure was not specified as the distribution parameter in the Gam_3 scenario. Thus, the results of the Gam_3 scenario may not be directly compared with the other scenarios in terms of overall average error rates.

4.5.3 Thresholds for the K^2 and T^2 Decomposition Methods

The effectiveness of the thresholds for K^2 and T^2 decomposition methods is illustrated through the cases selected from the previous section, which are Case 10 and Case 13 of the N_3 scenario and Case 3 and Case 4 of the Gam_3 scenario. For the K^2 decomposition method, each threshold was calculated by estimated quantiles from 1,000 bootstrap samples. The thresholds of the T^2 decomposition method were

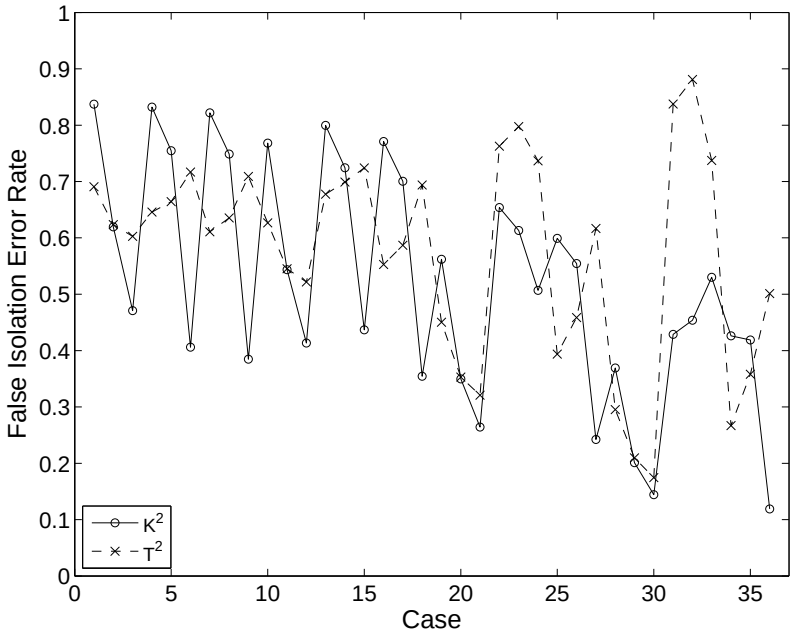


Figure 4.1. The performance comparison of fault isolation between the K^2 and T^2 decomposition methods in the N_3 scenario (average error rate from 10,000 simulation runs).

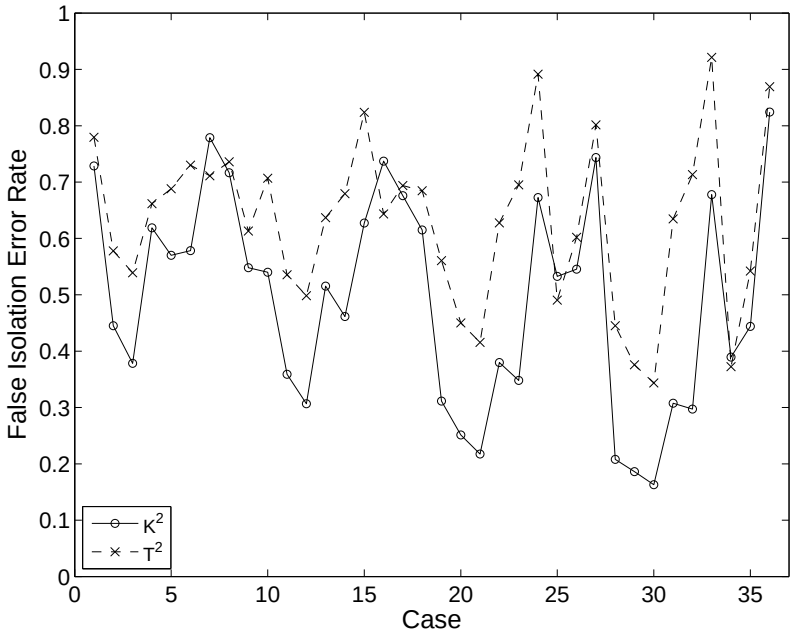


Figure 4.2. The performance comparison of fault isolation between the K^2 and T^2 decomposition methods in the $LogN_3$ scenario (average error rate from 10,000 simulation runs).

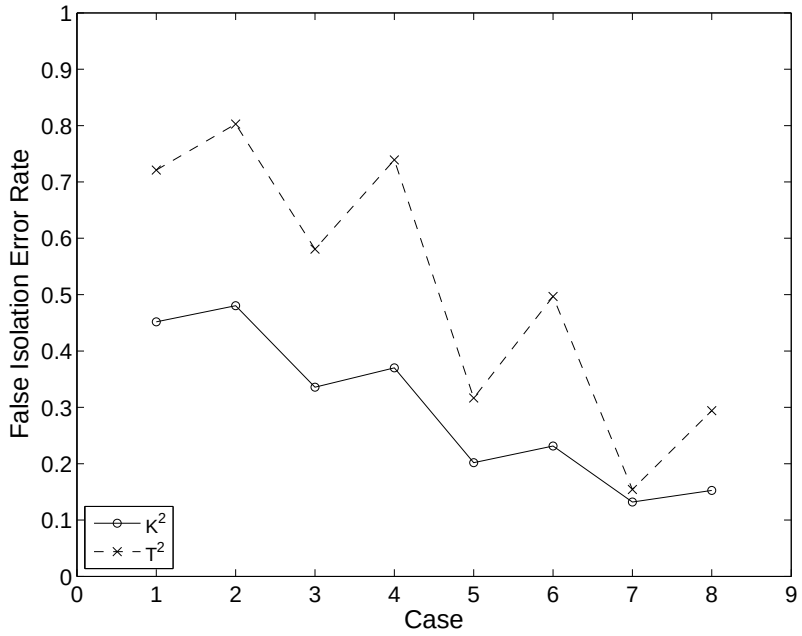


Figure 4.3. The performance comparison of fault isolation between the K^2 and T^2 decomposition methods in the Gam_3 scenario (average error rate from 10,000 simulation runs).

calculated from Equation 4.4. For both the K^2 and T^2 methods, the (i.e., $d_j K^2$ or $d_j T^2$, $j = 1, ..., p$) corresponding to the variable that exceed its threshold is indicated as significant.

A type I error rate is defined as the ratio of the number of in-control variables, which are incorrectly identified as contributed variables, to the total number of in-control variables. A type II error rate is defined as the ratio of the number of variables contributed to the out-of-control signal, which are not identified as contributed variables, to the total number of the variables contributed to the out-of-control signal.

Figure 4.4 shows how the trade-offs between actual type I and type II error rates in the K^2 and T^2 methods are controlled by their thresholds. The average values of actual type I and type II error rates from 10,000 simulation runs (for each value of α of which the standard errors are less than .001) are plotted. The method that yields a lower type II error rate should be considered the better method, given a similar type I error rate. Figures 4.4 (a) and (b) show that the T^2 method performs slightly

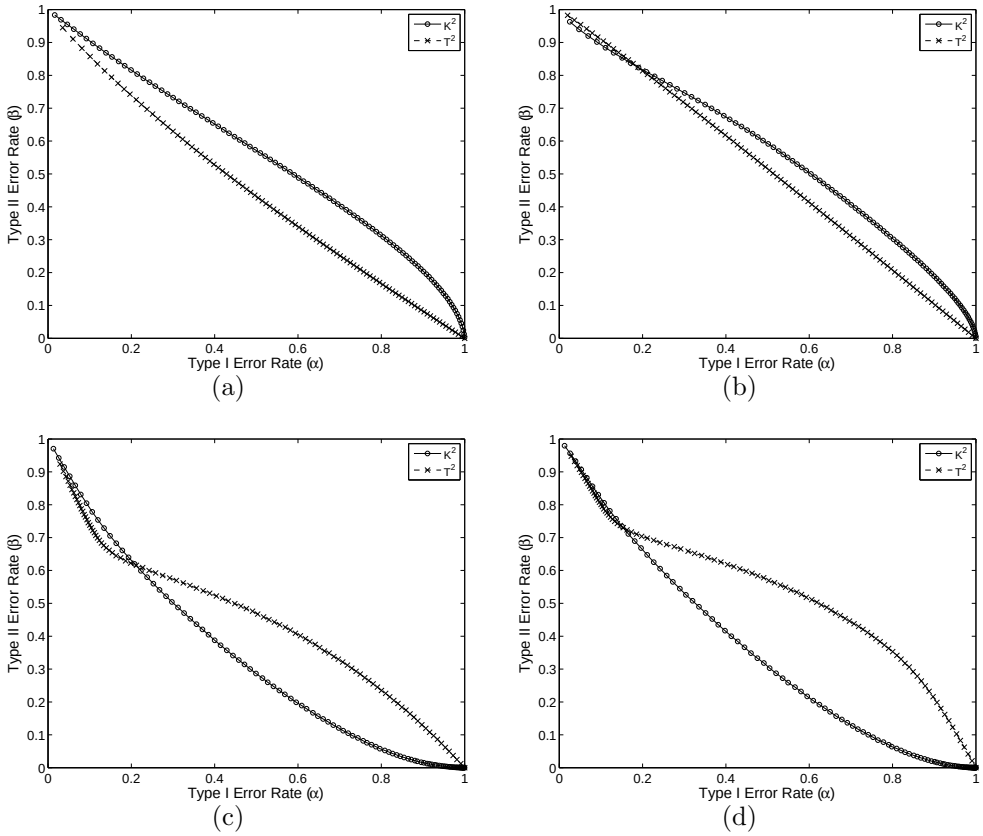


Figure 4.4. Behavior of type I and type II error rates of the K^2 and T^2 decomposition methods controlled by their thresholds for (a) N_3: Case 10, (b) N_3: Case 13, (c) Gam_3: Case 3, and (d) Gam_3: Case 4.

better than the K^2 method in the multivariate normal cases with single-variable and two-variable shifts. On the other hand, Figures 4.4 (c) and (d) show that the K^2 method mostly yielded smaller type II error rates than the T^2 method, implying that our proposed method can be effectively used in the nonnormal cases.

4.6 Case Study

An application of the proposed K^2 decomposition method is illustrated with a real dataset from a Wisconsin breast cancer study [35]. The dataset contains a total of 569 observations in which each of them is characterized by 30 continuous input variables and a two-class response variable (375 benign and 212 malignant

observations). Without loss of generality, the benign and malignant observations were considered as in control and out of control, respectively.

The proposed K^2 method was utilized to quantify the contribution of each individual variable on the out-of-control observations. The fault isolation results of five out-of-control observations (arbitrarily chosen) are reported in Table 4.4. An example of signal interpretation on the 24^{th} observation can be done by computing the d_jK^2 $(j = 1, ..., 30)$ values. The five highest d_jK^2 values determined that Variables 24, 4, 23, 14, and 3 (sorted in descending order by the higher corresponding statistics) are most likely to contribute to the out-of-control signals. In addition, our threshold indicated that only the first four variables (i.e., Variables 24, 4, 23, and 14) are significant at $\alpha = .01$.

Table 4.5 provides the fault isolation results using the T^2 decomposition method of the same five observations in order to compare to the results in Table 4.5. Although some variables (e.g., Variables 24, 4, and 14) can be detected by both of the methods for the 24^{th}, 135^{th}, 322^{nd} observations, both of the methods provide somewhat different results. The T^2 decomposition method also declared too many significant variables comparing to the K^2 decomposition method. For example, even such small α as .01, we can see that the T^2 decomposition declared 21 out of 30 variables to be significant in the 24^{th} observation. Moreover, from a Royston's multivariate normality test [60], we can conclude that the in-control data does not follow a multivariate normal distribution (p-value = 0). As a result, our proposed method should provide more reliable results than the T^2 decomposition method.

4.7 Conclusions

When an out-of-control signal is detected, SPC relies on operator intervention and cause-effect diagrams to find the root causes of process change. This is typically an experience-based and time-consuming effort without information-intensive representation of the manufacturing process and automatic diagnosis of process faults.

Although existing multivariate control charts provide control limits to monitor the process and detect any extraordinary events, it is a challenge to identify the causes of an out-of-control alarm in a multivariate setting. Most of the existing fault isolation approaches in multivariate SPC are based on the T^2 statistic, which limits them to handle only multivariate normal data.

This paper proposed a new nonparametric approach for fault isolation in multivariate SPC. The proposed approach decomposes the monitoring statistic, K^2 based on kNNDD one of the one-class classification algorithms. The K^2 decompositions are used to rank the importance of variables when a fault alarm is issued. The threshold established based on the bootstrap-quantile estimated method can be incorporated to determine the set of significant variables. The proposed approach requires a minimal set of assumptions that facilitates the fault isolation procedure in practice. The effectiveness of the proposed approach is demonstrated through our experimental studies with both simulated and real data. The performance of fault isolation using the K^2 decomposition method was comparable to the T^2 decomposition method with normal distribution data. When the data deviates from normality, the K^2 method outperforms the T^2 method.

As a result of exploring a decomposition procedure for the new nonparametric monitoring statistic, some interesting research directions arise. The approach that used to determine the subset of variables in the T^2 decomposition could be apply to determine the subset of variables in the K^2 decomposition. Advanced searching algorithms can be incorporated to the proposed approach to facilitate the indication of contributed variables when considering a large number of K^2 decomposition terms. Further research may apply the proposed procedure to another one-class classification-based monitoring statistic. Finally, the proposed approach can not only contribute to fault isolation in multivariate SPC, but also has a potential for the variable selection approach in one-class classification problems.

Table 4.1. Experimental design and simulation results (average error rate from 10,000 simulation runs) for N_3 scenario

					Error Rate	
Case	λ	δ_1	δ_2	δ_3	K^2	T^2
1	.5	.23	0	0	.8372	. 6909
2	.5	0	.29	0	. 6196	. 6236
3	.5	0	0	.33	.4710	.6026
4	.5	.40	.40	0	. 8321	. 6458
5	.5	.35	0	.35	.7546	.6643
6	.5	0	.17	.17	.4060	.7168
7	.5	.14	-.14	0	.8220	. 6353
8	.5	.14	0	-.14	.7488	.6353
9	.5	0	.33	-.33	.3847	.7091
10	1	.47	0	0	.7681	.6266
11	1	0	.59	0	.5435	.5451
12	1	0	0	.66	.4134	.5218
13	1	.80	.80	0	.7997	.6775
14	1	.71	0	.71	.7245	.6992
15	1	0	.35	.35	.4367	.7241
16	1	.27	-.27	0	.7710	.5527
17	1	.29	0	-.29	.7004	.5870
18	1	0	.66	-.66	.3544	.6939
19	2	.95	0	0	.5621	.4504
20	2	0	1.18	0	.3497	.3531
21	2	0	0	1.32	.2642	.3209
22	2	1.60	1.60	0	.6538	.7628
23	2	1.43	0	1.43	.6132	.7975
24	2	0	.70	.70	.5068	.7366
25	2	.55	-.55	0	.5993	.3936
26	2	.59	0	-.59	.5545	.4585
27	2	0	1.32	-1.32	.2422	.6166
28	3	1.42	0	0	.3690	.2954
29	3	0	1.77	0	.2012	.2095
30	3	0	0	1.99	.1443	.1747
31	3	2.40	2.40	0	.4287	.8376
32	3	2.14	0	2.14	.4538	.8811
33	3	0	1.06	1.06	.5301	.7374
34	3	.83	-.83	0	.4262	.2669
35	3	.88	0	-.88	.4188	.3580
36	3	0	1.99	-1.99	.1189	.5013
Overall Average					.5229	.5744

Table 4.2. Experimental design and simulation results (average error rate from 10,000 simulation runs) for $LogN_3$ scenario

					Error Rate	
Case	λ	δ_1	δ_2	δ_3	K^2	T^2
1	.5	.27	0	0	.7286	.7795
2	.5	0	.33	0	.4452	.5780
3	.5	0	0	.36	.3784	.5390
4	.5	.42	.42	0	.6188	.6617
5	.5	.38	0	.38	.5701	.6882
6	.5	0	.21	.21	.5785	.7302
7	.5	.18	-.18	0	.7789	.7112
8	.5	.19	0	-.19	.7165	.7360
9	.5	0	.38	-.38	.5481	.6132
10	1	.48	0	0	.5399	.7069
11	1	0	.57	0	.3588	.5356
12	1	0	0	.62	.3066	.4984
13	1	.71	.71	0	.5154	.6368
14	1	.66	0	.66	.4615	.6794
15	1	0	.38	.38	.6272	.8237
16	1	.35	-.35	0	.7371	.6433
17	1	.36	0	-.36	.6759	.6938
18	1	0	.65	-.65	.6149	.6847
19	2	.81	0	0	.3113	.5606
20	2	0	.93	0	.2512	.4502
21	2	0	0	1.00	.2174	.4155
22	2	1.13	1.13	0	.3798	.6275
23	2	1.05	0	1.05	.3479	.6950
24	2	0	.65	.65	.6726	.8915
25	2	.65	-.65	0	.5326	.4906
26	2	.67	0	-.67	.5455	.6020
27	2	0	1.03	-1.03	.7435	.8016
28	3	1.05	0	0	.2079	.4452
29	3	0	1.20	0	.1863	.3758
30	3	0	0	1.28	.1630	.3435
31	3	1.42	1.42	0	.3076	.6347
32	3	1.33	0	1.33	.2972	.7133
33	3	0	.87	.87	.6778	.9214
34	3	.90	-.90	0	.3894	.3725
35	3	.92	0	-.92	.4440	.5420
36	3	0	1.28	-1.28	.8243	.8692
Overall Average					.4916	.6303

Table 4.3. Experimental design and simulation results (average error rate from 10,000 simulation runs) for Gam_3 scenario

					Error Rate	
Case	λ	δ_1	δ_2	δ_3	K^2	T^2
1	.5	.73	0	0	.4518	.7210
2	.5	.64	.64	0	.4804	.8028
3	1	.99	0	0	.3358	.5807
4	1	.84	.84	0	.3701	.7392
5	2	1.45	0	0	.2017	.3165
6	2	1.19	1.19	0	.2313	.4968
7	3	1.84	0	0	.1319	.1542
8	3	1.50	1.50	0	.1526	.2940
Overall Average					.2945	.5131

Table 4.4. The fault isolation results of the K^2 decomposition approach of five out-of-control observations (arbitrarily chosen) from the Wisconsin breast cancer data

Observation Number	Five Most Likely Contributed Variables (*Significant Variable at $\alpha = .01$)	Total Significant Variables at $\alpha = .01$
24	24*, 4*, 23*, 14*, 3	4
135	24*, 4*, 14*, 23, 3	3
230	4*, 24*, 23, 22, 14	2
322	24*, 4*, 14*, 23, 3	3
563	24*, 4*, 22, 23, 2	2

Table 4.5. The fault isolation results of the T^2 decomposition approach of the chosen five out-of-control observations from the Wisconsin breast cancer data

Observation Number	Five Most Likely Contributed Variables (*Significant Variable at $\alpha = .01$)	Total Significant Variables at $\alpha = .01$
24	24*, 4*, 21*, 14*, 1*	21
135	24*, 21*, 4*, 28*, 8*	12
230	7*, 17*, 13*, 21, 23	3
322	14*, 11*, 24*, 21, 29	3
563	13*, 28*, 18*, 27*, 17*	13

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

The scope of application of data mining and SPC has been extended by the work undertaken in this dissertation. We introduced new data mining-driven approaches for process monitoring and diagnosis. In Chapter 2, we proposed classification-based control charts, PoC charts, that allow control chart techniques to utilize out-of-control observations through supervised learning algorithms. Although out-of-control information is available or obtainable in many processes, traditional control charts have never taken advantage of such information. By using this additional information, detection efficiency can be improved. In Chapter 3, one-class classification-based control charts are proposed to improve both phase I and phase II applications in SPC. The proposed approach facilitates the construction of control charts based on fewer assumptions than those used by traditional control charts. One-class classification-based control charts show their effectiveness over the traditional control charts when the data deviates from normal. Finally, in Chapter 4, a nonparametric fault isolation technique to interpret an out-of-control signal from one-class classification control charts is proposed. Our experimental studies show that the proposed fault isolation through one-class classification-based statistics overcomes the limitations of widely used fault isolation methods.

By constructing proposed control charts based on varieties of data mining algorithms, the advantages of the effectiveness and flexibility of data mining would be brought to solve more advanced SPC problems such as:

- Monitoring multivariate autocorrelated processes.
- Handling data with missing values.

- Accommodating a mixture of data formats; continuous, discrete, and categorical data.
- Dealing with all of the aforementioned issues simultaneously.

Once signals from such control charts are produced, our proposed decomposition procedure may be used to isolate the faults, which could be generated based on various monitoring statistics. In addition, the proposed decomposition procedure may be applied to variable selection problems in data mining.

Our proposed data mining-driven approaches for process monitoring and diagnosis may not only benefit SPC, but also could be extended to several applications that required monitoring and diagnostic tools. Furthermore, because SPC techniques were first developed based on the concept of statistical hypothesis testing, examples of applications that may be readily extended from our SPC research are "one-class classification-based hypothesis testing techniques." These are examples of our challenging future directions.

REFERENCES

[1] J. S. Oakland, *Statistical Process Control*, 5th ed. London, England: Butterworth-Heinemann, 2003.

[2] R. L. Mason and J. C. Young, *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia, PA: American Statistical Association and Society for Industrial and Applied Mathematics, 2002.

[3] W. A. Shewhart, *Economic Control of Quality of Manufactured Product*. Princeton, NJ: Van Nostrand Press, 1931.

[4] Z. G. Stoumbos, M. R. Reynolds, T. P. Ryan, and W. H. Woodall, "The state of statistical process control as we proceed into the 21^{st} century," *Journal of the American Statistical Association*, vol. 95, pp. 992–998, 2000.

[5] W. H. Woodall and D. C. Montgomery, "Research issues and ideas in statistical process control," *Journal of Quality Technology*, vol. 31, no. 4, pp. 376–386, 1999.

[6] D. C. Montgomery, *Introduction to Statistical Quality Control*, 5th ed. New York, NY: Wiley, 2005.

[7] C. W. Champ and W. H. Woodall, "Exact results for Shewhart control charts with supplementary runs rules," *Technometrics*, vol. 29, no. 4, pp. 393–399, 1987.

[8] E. S. Page, "Cumulative sum charts," *Technometrics*, vol. 3, no. 1, pp. 1–9, 1961.

[9] S. W. Robert, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.

[10] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–12, 1990.

[11] K. W. Linna, W. H. Woodall, and K. L. Busby, "The performance of multivariate control charts in the presence of measurement error," *Journal of Quality Technology*, vol. 33, no. 3, pp. 349–355, 2001.

[12] H. Hotelling, *Multivariate Quality Control*, ser. Techniques of Statistical Analysis, C. Eisenhart, M. W. Hastay, and W. A. Wallis, Eds. New York, NY: McGraw-Hill, 1947.

[13] C. A. Lowry and D. C. Montgomery, "A review of multivariate control charts," *IIE Transactions*, vol. 27, no. 6, pp. 800–810, 1995.

[14] S. J. Wierda, "Multivariate statistical process control – recent results and directions for future research," *Statistica Neerlandica*, vol. 48, no. 2, pp. 147–168, 1994.

[15] M. J. A. Berry and G. S. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York, NY: Wiley, 2000.

[16] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: a review," *SIGMOD Record*, vol. 34, no. 2, pp. 18–26, 2005.

[17] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* New York, NY: Springer, 2001.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: Wiley, 2001.

[19] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. Upper Saddle River, N.J.: Prentice Hall, 2002.

[20] V. N. Vapnik, *Statistical Learning Theory.* New York, NY: Wiley, 1998.

[21] T. M. Mitchell, *Machine Learning.* New York, NY: McGraw-Hill, 1997.

[22] E. Turban and J. E. Aronson, *Decision support systems and intelligent systems*, 6th ed. Upper Saddle River, NJ: Prentice Hall, 2001.

[23] O. Chapelle, A. Zien, and B. Scholkopf, Eds., *Semi-Supervised Learning.* Cambridge, MA: MIT Press, 2006.

[24] D. M. J. Tax, "One-class classification: Concept-learning in the absence of counter-examples," Ph.D. dissertation, Delf University of Technology, 2001.

[25] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[26] H.-P. Kriegel, K. Borgwardt, P. Kroger, A. Pryakhin, M. Schubert, and A. Zimek, "Future trends in data mining," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 87–97, 2007.

[27] G. P. Shapiro, "Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from university to business and analytics," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 99–105, 2007.

[28] D. Zhang and L. Zhou, "Discovering golden nuggets: data mining in financial application," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 34, no. 4, pp. 513–522, 2004.

[29] S. Bersimis, S. Psarakis, and J. Panaretos, "Multivariate statistical process control charts: an overview," *Quality & Reliability Engineering International*, vol. 23, no. 5, pp. 517–543, 2006.

[30] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control Engineering Practice*, vol. 3, no. 3, pp. 403–414, 1995.

[31] W. Jiang and K.-L. Tsui, "A theoretical framework and efficiency study of multivariate statistical process control charts," *IIE Transactions*, vol. 40, no. 7, pp. 650–663, 2008.

[32] M. C. Testik and G. C. Runger, "Multivariate one-sided control charts," *IIE Transactions*, vol. 38, no. 8, pp. 635–645, 2006.

[33] W. Hwang, G. Runger, and E. Tuv, "Multivariate statistical process control with artificial contrasts," *IIE Transactions*, vol. 39, no. 6, pp. 659–669, 2007.

[34] J. Hu and G. Runger, "Tuned artificial contrasts to detect signal," *International Journal of Production Research*, vol. 45, no. 23, pp. 5527–5534, 2007.

[35] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[36] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[37] J. Hu and G. Runger, "Time-based detection of changes to multivariate patterns," in *Proceedings of the INFORMS Artificial Intelligence and Data Mining Workshop*, 2006.

[38] W. H. Woodall, "Controversies and contradictions in statistical process control," *Journal of Quality Technology*, vol. 32, no. 4, pp. 341–350, 2000.

[39] H. Hotelling, *Techniques of statistical analysis*. New York, NY: McGraw-Hill, 1947, ch. Multivariate quality control, pp. 111–184.

[40] S. Bakir, "Distribution-free quality control charts based on signed-rank-like statistics," *Communications in Statistics: Theory and Methods*, vol. 35, pp. 743–757, 2006.

[41] S. Chakraborti, P. Van Der Laan, and S. T. Bakir, "Nonparametric control chart: an overview and some results," *Journal of Quality Technology*, vol. 33, no. 3, pp. 304–315, 2001.

[42] R. Y. Liu, K. Singh, and J. H. Teng, "DDMA-charts: nonparametric multivariate moving average control charts based on data depth," *Allgemeines Statistisches Archiv*, vol. 88, no. 2, pp. 235–258, 2004.

[43] P. Qiu and D. Hawkins, "A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions," *The Statistician*, vol. 52, no. 2, pp. 151–164, 2003.

[44] R. Sun and F. Tsung, "A kernel-distance-based multivariate control chart using support vector methods," *International Journal of Production Research*, vol. 41, no. 13, pp. 2975–2989, 2003.

[45] S. Kumar, A. K. Choudhary, M. Kumar, R. Shankar, and M. K. Tiwari, "Kernel distance-based robust support vector methods and its application in developing a robust K-chart," *International Journal of Production Research*, vol. 44, no. 1, pp. 77–96, 2006.

[46] Z. Zhang, X. Zhu, and J. Jin, "SVC-based multivariate control charts for automatic anomaly detection in computer networks," in *Proceedings of the Third International Conference on Autonomic and Autonomous Systems*, 2007.

[47] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D. M. J. Tax, "PRTools4: The MATLAB toolbox for pattern recognition," Delft University of Technology, Netherlands, 2007. [Online]. Available: http://www.prtools.org

[48] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, 1993.

[49] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the ACM SIGMOD 2000 international conference on management of data*, 2000.

[50] G. C. Runger, "Projections and the U^2 multivariate control chart," *Journal of Quality Technology*, vol. 28, no. 3, pp. 313–319, 1996.

[51] G. C. Runger, F. B. Alt, and D. C. Montgomery, "Contributors to a multivariate statistical process control chart signal," *Communications in Statistics: Theory and Methods*, vol. 25, no. 10, pp. 2203–2213, 1996.

[52] H. Liu, W. Jiang, A. Tangirala, and S. Shah, "An adaptive regression adjusted monitoring and fault isolation scheme," *Journal of Chemometrics*, vol. 20, pp. 280–293, 2006.

[53] J. E. Jackson, *A User Guide to Principal Components*. New York, NY: Wiley, 1991.

[54] T. Kourti and J. F. MacGregor, "Multivariate SPC methods for process and product monitoring," *Journal of Quality Technology*, vol. 28, pp. 409–428, 1996.

[55] R. L. Mason, N. D. Tracy, and J. C. Young, "Decomposition of T^2 for multivariate control chart interpretation," *Journal of Quality Technology*, vol. 27, no. 2, pp. 99–108, 1995.

[56] ——, "A practical approach for interpreting multivariate T^2 control chart signal," *Journal of Quality Technology*, vol. 29, no. 4, pp. 396–406, 1997.

[57] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[58] B. Flury and H. Riedwyl, *Multivariate Statistics: A Practical Approach.* London: Chapman & Hall, 1988.

[59] Z. G. Stoumbos and J. H. Sullivan, "Robustness to non-normality of the multivariate EWMA control chart," *Journal of Quality Technology*, vol. 41, no. 13, pp. 260–276, 2002.

[60] J. P. Royston, "Some techniques for assessing multivariate normality based on the shapiro-wilk w," *Applied Statistics*, vol. 32, no. 2, pp. 121–133, 1983.

BIOGRAPHICAL STATEMENT

Thuntee Sukchotrat received a B.E. degree in Industrial Engineering from Chulalongkorn University in 2002. In 2003, he worked as an engineer in the technical standard department at Advanced Info Service PCL, Thailand. In 2005, he completed an M.S. degree in Industrial and Manufacturing Systems Engineering from the University of Texas at Arlington (UTA), where he has continued pursuing a Ph.D. degree and working as a graduate research assistant. His current research interest is applications of data mining and multivariate statistical process control. He is a member of the Center of Stochastic Modeling, Optimization & Statistics (COSMOS) at UTA, Institute for Operation Research and Management Science (INFORMS), and Institute of Industrial Engineers (IIE). His web address is http://www.thuntee.com.