DISTRIBUTION-FREE FAULT IDENTIFICATION AND ANOMALY DETECTION IN HIGH-DIMENSIONAL DATA

by

MEHMET TURKOZ

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Industrial and Systems Engineering

Written under the direction of

Myong K. Jeong and Elsayed A. Elsayed

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

MAY, 2018

ABSTRACT OF THE DISSERTATION

Distribution-Free Fault Identification and Anomaly Detection in High-Dimensional Data

By MEHMET TURKOZ

Dissertation Directors:

Myong K. Jeong and Elsayed A. Elsayed

Quality engineering is an essential activity in production processes and its objective is to ensure the quality of the products throughout the production stages. Many processes have several attributes that need to be continuously monitored to detect any variable changes in the production process. We refer to the monitoring process with several quality characteristics as multivariate statistical process control (MSPC). Most of the quality control procedures assume that the characteristics of the process follow normal distributions; however, this is a limiting assumption since the underlying distribution of the processes may not be normal.

In this dissertation, we present procedures to identify the faulty variables and detect anomalies in MSPC with high dimensional data when the underlying distribution of the process is unknown. We first propose a distribution-free adaptive step-down (DFASD) procedure, which is motivated by a well-known data description method called support vector data description (SVDD). This data description procedure includes the support

vectors which identify the hypersphere boundary for the available data by using the kernel concept. In a high-dimensional process, identifying the variable or a subset of variables, which cause an out-of-control (OC) signal, is a challenging issue in quality engineering. DFASD procedure utilizes conditional statistics for the identification of faulty variables. The proposed DFASD procedure selects a variable having no significant evidence of a change at each step based on the variables that are selected in the previous steps. The proposed DFASD stops when there are no longer variables to classify to the unchanged set. Therefore, it concludes the variables which are not in the unchanged set as changed variables.

We then present a new distribution-free fault identification procedure based on Bayesian inference which is called Bayesian SVDD (BSVDD). While the traditional SVDD assumes that the process parameters are constants to be determined, the center of hypersphere may be considered as a random vector with inherent randomness based on a given training dataset. We introduce a Bayesian approach for SVDD by assuming that a transformed data into the higher dimensional space follow normal distribution. A distance from a point to the center of the hypersphere is inversely proportional to the likelihood in the proposed model. This is because SVDD is a special case of the proposed BSVDD model, which improves SVDD by utilizing the precise prior knowledge. Therefore, by combining proposed BSVDD with an adaptive step-down procedure, we drive a new BSVDD based fault identification procedure for the MSPC. This is the first research to identify the faulty variables by using the distribution-free approach based on Bayesian inference.

We also present an anomaly detection procedure which is easily applicable in detecting anomalies in multimode processes. Traditional quality control procedures assume that normal observations are obtained from a single distribution. However, due to the complexities of modern industrial processes, the observations might have multiple operating modes. In other words, normal observations may be obtained from more than one distribution. In such cases, conventional quality control procedures might trigger false alarms while the process is indeed in another operating mode. We propose a generalized support vector-based anomaly detection procedure called *n*-class SVDD which can be used to determine the anomalies in multimode processes. The proposed procedure constructs *n* hyperspheres by considering the relationship among modes. In addition, we introduce a generalized Bayesian framework by not only considering the prior information from each mode but also the relationships among the modes.

Finally, we present a new Bayesian procedure for anomaly detection in multi-class data. The existing procedures for anomaly detection mostly take only the normal information into account. However, the anomaly information is often available from the engineering knowledge and the historical data of the process. The performance of the anomaly detection procedures can be improved when available anomaly data are utilized to obtain data description. We propose a multi-class Bayesian SVDD model that takes anomaly data into consideration when the anomaly data are available and an appropriate prior distribution of the anomaly data is obtained.

ACKNOWLEDGEMENTS

I would like to acknowledge and present my sincere thanks to my advisors Professor Myong K. Jeong and Professor Elsayed A. Elsayed for their excellent guidance and encouragement during my dissertation and research. Their patient guidance has provided the most motivation during my PhD study. I am very honored to be given the opportunity to have worked with them and been accepted as a member of their academic family. I am further fortunate to have the most valuable committee members, Professor Hoang Pham and Professor Ming-ge Xie who have helped me beyond their expected role. I also would like to thank them for their insightful comments and encouragement.

My sincere thanks also go to my mother, Hatice Turkoz and my father, Hayrullah Turkoz who raised me with such love and with a belief that almost anything is possible through hard work. I would like to also thank my brother, Rafet Turkoz, for his endless support, help and believing in me. I would like to extend my gratitude to my relatives, especially Banu Turkoz, Arzu Turkoz, Lutfi Turkoz, Murat Okyay, Ozgur Can Okyay, Adnan and Rafia Okyay, Umut Gemici, Yasar and Gungor Gemici.

I acknowledge all my friends from Rutgers University for being great colleagues and friends. I would like to especially thank Boyoung Park who always wishes me the best and supports me. In addition, I would like to thank my faithful friend, Sangahn Kim, for his brotherhood and friendship. I would also like to thank my close friends Mejdal Alqahtani, Jaeseungn Baek, Semih Cetindag and Akin Ozdemir.

Finally, I would like to acknowledge the Turkish Ministry of National Education for funding my research. I appreciate the opportunity to be a laureate of this prestigious scholarship. I am also thankful to the Scientific and Technological Research Council of Turkey (TUBITAK) for supporting my studies in Turkey.

DEDICATION

To my family – The Turkoz Family,

to the memory of my wonderful grandmother, Sefika Turkoz.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Motivation of the Work

The higher expectations of customers and the globalization of the world economy have further emphasized the role of quality engineering and process control. Therefore, industrial organizations have paid significant attention to the control and monitoring of the quality characteristics of products. Controlling and monitoring of the quality characteristics have been mostly achieved by using the Statistical process control (SPC) techniques.

In the quality engineering research studies, SPC charts have been widely used to monitor product quality and detect changes in product and process parameters that occur due to the variability of the production processes. The variability can be reduced by using SPC charts which improve the quality of products (Montgomery, 2007). The SPC charts mainly focus on the detection of assignable causes of variation but do not identify or adjust the cause of variability which is a limited approach for process improvement.

Shewhart (1931) sequential testing chart is the base for charts and procedures that have been proposed in the literature such as exponentially weighted moving average (EWMA) and cumulative sum (CUSUM). In addition, other researchers have introduced extensions of these charts (Crowder, 1989, Ewan, 1963, Gan, 1991, Hawkins, 1991, Hawkins, 1993,

Hawkins and Olwell, 1997, Lucas and Saccucci, 1990, Roberts, 1959, Woodall and Adams, 1993, Kim *et al.*, 2013, Park *et al.*, 2012, Jeong *et al.*, 2006, Abdella *et al.*, 2017).

These charts are mainly used for the univariate case; however, most of the products have more than one quality characteristics and most of the production processes are correlated to each other. To monitor these kinds of processes, multivariate statistical process control (MSPC) charts have been introduced such as Hotelling's chi-square, multivariate exponentially weighted moving average (MEWMA) and multivariate CUSUM charts (Crosier, 1988, Lowry *et al.*, 1992, Ngai and Zhang, 2001, Pignatiello and Runger, 1990, Kim *et al.*, 2014). These charts monitor the quality characteristics simultaneously by considering the correlations among the product's characteristics and the process parameters.

Despite the fact that these procedures are only effective in monitoring and controlling of the processes under the normality assumption, in a realistic case which is the violation of the normality assumption, there has been an increasing concern about their effectiveness in monitoring such processes. To overcome the drawback of non-normality, several nonparametric or distribution-free control charts have been designed to monitor the univariate processes (Bakir, 2004, Bakir, 2006, Bakir and Reynolds, 1979, Janacek and Meikle, 1997, Li *et al.*, 2010, Liu *et al.*, 2014, Park *et al.*, 1987, Wang *et al.*, 2016). Several multivariate nonparametric or distribution free control charts are also introduced (Chen *et al.*, 2016, Liu, 1995, Sun and Tsung, 2003, Zhou *et al.*, 2015, Zi *et al.*, 2013, Zou and Tsung, 2011, Zou *et al.*, 2012, Cho *et al.*, 2006).

Once MSPC charts signal for the detection of process changes, identifying faulty variables becomes a more challenging issue. Even though MSPC charts identify the abnormal cases, they have limited ability to identify the variables that cause the out-of-control (OC) signal because the monitoring statistics are calculated by considering all variables. There have been efforts to identify a root cause of the process abnormality. The first attempt to identify the faulty variables is developed by Doganaksoy *et al.* (1991). However, the main drawback of their procedure is due to the lack of the correlation effect. To overcome this drawback, several procedures have been proposed (Das and Prakash, 2008, Hawkins, 1991, Hawkins, 1993, Kim *et al.*, 2016b, Mason *et al.*, 1995, Mason *et al.*, 1997, Runger, 1996, Runger *et al.*, 1996, Sullivan *et al.*, 2007). These procedures assume that the process follows a multivariate normal distribution.

The design of the above fault identification procedures is based on the assumption that the quality control process follows multivariate normal distribution. However, when the underlying process distribution is unknown or limited, these procedures may be inefficient and unable to identify the faulty variables. In this dissertation, we intend to investigate fault identification methods that do not assume any specific probability distribution for the process or product observation.

Most of the research on MSPC procedures is based on the assumption that a process has a single operating region and follows a unimodal distribution. However, if a process has multiple operating regions, the performance of the conventional MSPC procedures may be inefficient. To overcome the inefficiency of the traditional MSPC procedures, several

studies have been introduced based on different approaches such as Gaussian mixture models, partial least squares, principle component analysis (PCA) and independent component analysis (ICA) (Choi *et al.*, 2004, Xie and Shi, 2012, Xu *et al.*, 2014, Xu and Deng, 2016, Yu and Qin, 2008, Zhao *et al.*, 2004, Zhao *et al.*, 2006). In addition to the statistical approaches, data mining based approaches have also been introduced (Kang and Kim, 2013, Kang *et al.*, 2016). Both of these data mining approaches are based on the *k*-nearest neighbor data description. The *k*-nearest neighbor data description approaches are suitable for spherical data distribution (Yu *et al.*, 2002), yet having spherical data distribution is not common in real-life applications. In this dissertation, we intend to detect the anomalies in multimode processes regardless of the number of classes or types of the distributions.

1.2 Problem Description and Assumptions

There are two types of SPC techniques, traditional SPC and multimode SPC. In the traditional SPC, identifying the faulty variables is considered one of the important areas of research in the MSPC. It is important to identify the faulty variables in processes under the assumption that the distribution of the characteristics of the processes is unknown, which makes the faulty identification much more challenging. Although the non-normality assumption is realistic, there is limited research that investigates such processes to identify the faulty variables. In this dissertation, we assume that the underlying distribution of the process is unknown or distribution-free.

In a process control, a variable is called a faulty variable when its mean is changed. Therefore, in this dissertation, we identify the variables of the out-of-control (OC) observation whose mean has changed. We will also use different terminologies which represent the variables, such as ‘quality characteristics’, ‘quality features’, ‘process variables’. Moreover, it is assumed that only a single observation is sampled and its faulty variables are identified at each sampling epoch.

In contrast to the traditional SPC, multimode SPC procedures perform well if a process has multiple operating regions. Most of the multimode SPC procedures are highly dependent on the type of data, mostly the spherical distribution, which makes them inefficient if the data deviates from the spherical distribution. In this dissertation, we assume that the underlying distributions of multimode processes can be obtained from any distributions.

In addition, it is assumed that a set of in-control data is available which are used to set up the parameters (Phase I) of the proposed method, such as critical values. We identify the faulty variables of the OC observation (Phase II) after setting up the parameters of the procedures. Therefore, we intend to concentrate on both Phases I and II.

1.3 Approaches to Identify Faulty Variables in MSPC and Detect Anomalies in Multimode Processes

In research on fault identification of MSPC processes and anomaly detection in multimode process, a process is assumed to follow some parametric distributions, specially normal or spherical distribution. The statistical properties of the most of fault identification procedures and multimode processes are valid if the normality assumption is satisfied. However in some cases, information about the underlying distribution of a given process is limited or unknown; such cases pose a challenge for quality engineering because the properties (critical values and threshold values) of the parametric fault identification and multimode processes procedures may be affected. With the unknown distribution assumption, we will introduce different fault identification procedures as well as an anomaly detection procedure which can be used to detect anomalies in multimode processes.

1.3.1. Distribution-Free Based Fault Identification

Despite existing parametric methods' success under the normality assumption, there has been an increasing concern about situations when the normality assumption does not hold. A limited number of nonparametric or distribution-free procedures have been introduced by relaxing the normality assumption. The existing nonparametric methods for detection of process changes are based on the *k*-nearest neighbor data description and hybrid novelty score (Kim *et al.*, 2011, Tuerhong and Kim, 2011), which are built under the assumption that there is only one variable change. These existing distribution-free

fault identification procedures will be reviewed prior to introducing proposed fault identification procedures in Chapter 2.

We first review the data description method, support vector data description (SVDD), which is the fundamental of our proposed methods. SVDD is used to represent the data when the underlying distribution of the data is unknown. In spite of the wide applications of the SVDD, there has been limited research on the fault identification procedure. We adopt and propose a distribution-free based fault identification procedure and compare its performance with other existing parametric and distribution-free fault identification procedures.

1.3.2. Bayesian Framework for Fault Variable Identification

Traditional fault identification procedures such as T^2 decomposition, regression-adjusted variables and distribution-free procedures have been applied to identify the faulty variables in industrial processes characterized by several measurable parameters. However, the true parameters of these methods can be random variables. It would be reasonable to assign a prior distribution to these parameters. The Bayesian approach focuses on determining the optimal policy to identify the parameters based on the posterior probability obtained from available in-control data. Therefore, in this dissertation, we propose a Bayesian distribution-free fault identification procedure. This the first research for distribution free fault identification method using Bayesian inference.

1.3.3. Generalized Support Vector Data Description with Bayesian Framework

In real-life problems, identifying anomalies is important as they may have significant information about a procedure. Several anomaly detection procedures are introduced to identify anomalies. Among them, SVDD procedure has gained more attention and inspired a lot of researchers. The existing SVDD procedures assume that the normal data (operating mode) consist of one or two classes. However, a process can operate on more than two modes. In this dissertation, we propose *n*-class SVDD which is independent of the number of classes. The proposed procedure constructs *n* hyperspheres by considering relationships among classes. In addition, conventional SVDD procedures are built by ignoring prior information. Thus, we introduce a Bayesian framework by not only considering the prior information of each class but also the relationships among the classes.

1.4 Dissertation Outline

This dissertation is organized as follows. Chapter 2 provides a review of the relevant literature about parametric and non-parametric fault identification procedures as well as anomaly detection procedures. In Chapter 3, we introduce a distribution-free fault identification procedure by combining the data description procedure SVDD and the adaptive step-down (ASD) procedure. Chapter 4 proposes a Bayesian support vector data description (BSVDD). In addition, by combining BSVDD with an adaptive step-down procedure, we propose an BSVDD based distribution-free faulty variable identification procedure and we show the superiority of the fault identification procedure through

different simulation studies. In Chapter 5, an SVDD based anomaly detection procedure called *n*-class SVDD is introduced to identify the anomalies, which can be adapted to multimode processes. In addition, we introduce generalized Bayesian SVDD procedure which is based on a proposed *n*-class SVDD. In Chapter 6, we introduce a multi-class Bayesian SVDD model that takes anomaly data into consideration when the anomaly data are available. Finally, in Chapter 7, we discuss the conclusions and the future research topics related to the thrust of this dissertation.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces a comprehensive review of work related to the research being investigated in this dissertation. We present relevant research for identification of faulty variables in high-dimensional multivariate processes for both parametric and distribution-free diagnosis procedures and anomaly detection procedures which can be used to identify anomalies in multimode processes. We review different methodologies for these procedures and discuss their advantages and limitations. We begin with the description of the parametric methods for faulty variable identification.

2.1 Parametric Methods for Identification of Faulty Variables

In quality engineering, multivariate statistical process control (MSPC) charts have been widely used to monitor product quality in multi-dimensional process and to detect process changes. One of the advantages of the MSPC charts, such as Hotelling's chi-square chart, multivariate exponentially weighted moving average (MEWMA) (Lowry *et al.*, 1992), multivariate CUSUM (Crosier, 1988) is to monitor multiple variables simultaneously. However, they have limited abilities in identifying the sources of variability in the production processes since these charts monitor the quality characteristics simultaneously by considering the correlations among the product's characteristics and the process parameters. In addition, they are also limited in identifying

the variables that cause out-of-control (OC) signal since the monitoring statistics are calculated by considering all process variables.

Suppose that a process has a total of *p* variables, once the MSPC chart signals an alarm with the detection of process abnormality, then identification of the faulty variables among the *p* variables is of interest since it often provides important information in process diagnosis and in taking corrective actions to adjust the process. Therefore, fault diagnosis in statistical process control remains an important and challenging issue for the quality engineers, specially in high dimensional data.

Several investigators attempted to address the identification of the faulty variables. For example, Doganaksoy *et al.* (1991) propose an individual test statistic for each variable to identify the faulty ones. This method, however, is known to ignore the effect of the correlation between variables and is difficult to implement in high-dimensional processes.

To consider the correlation effect, Hawkins (1991, 1993) proposes a novel method based on regression-adjusted variables. In this method, it is assumed that the process follows a multivariate normal distribution, $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, and OC occurs due to the shift in a single variable *i* by changing the distribution of the process to $N(\boldsymbol{\mu}_0 + \delta \mathbf{e}_i, \boldsymbol{\Sigma}_0)$ where $\mathbf{e}_i$ is the column vector with entries 1 in row *i* and 0 elsewhere and δ is related to the size of shift. This procedure obtains the residual of variable x_i to identify whether x_i is a faulty

variable. The procedure is also built on the assumption that only one variable changes. Therefore, it may not perform well if more than one variable change, or when a single variable strongly correlated with other variables is shifted (Das and Prakash, 2008, Hawkins, 1993).

A similar research to Hawkins (1993) is introduced by Runger *et al.* (1996), which investigates the contribution of a subset of variables causing a change in the mean shift. Based on the knowledge of a subset with unchanged variables, relative contribution of the complement of this subset can be calculated. The proposed procedure calculates the contribution of each variable to the mean shift by determining the difference between the overall T^2 statistic and $T^2_{1,\ldots,j-1,j+1,\ldots,p}$ statistic. The contribution of variable j can be expressed as follows:

$$T^2_{j|1,2,\ldots,j-1,j+1,\ldots,p} = T^2 - T^2_{1,\ldots,j-1,j+1,\ldots,p}\,,\ \forall j = 1,\ldots,p$$

where $T^2_{1,\ldots,j-1,j+1,\ldots,p}$ is the T^2 statistic using all variables except the j^{th} variable. It is shown that $T^2_{j|1,2,\ldots,j-1,j+1,\ldots,p}$ follows $\chi^2_{1,\alpha}$ with the degree of freedom one, where α is the significance level. However, the identification of the subset of variables which do not cause a mean shift is not straightforward. Therefore, this procedure is meaningful only when we have knowledge of a set of unchanged variables surely, otherwise it may lead to poor use of the procedure in fault identification.

Mason *et al.* (1995, 1997) propose a T^2 decomposition method based on the regression adjustment method (or conditional T^2), which is known as the Mason–Tracy–Young (MTY). The MTY method decomposes the T^2 statistic into the combinatorial number of conditional statistics to identify a changed variable. The T^2 decomposition-based approaches theoretically work well. However, their calculation complexity makes the procedures impractical for a large number of variables since these procedures consider $p!$ decompositions. To reduce the computational complexity, Sullivan *et al.* (2007) propose an algorithm which considers ${}_pC_k$ (*p* choose *k*) number of subsets, where *k* is the size of a set, and calculates the T^2 statistics for all of these subsets for *p*-variate observation. However, for large p, the computational complexity still remains an obstacle even though this approach reduces the computational work compared to the MTY method.

When MSPC charts detect abnormality, in a high dimensional process, it is assumed that only a few variables, one or small set of variables, are responsible for a process shift which is known as the sparsity property (Zou and Qiu, 2009). Based on the variable selection methods and the sparsity assumption, Wang and Jiang (2009) and Zou and Qiu (2009) introduce process monitoring and diagnosis schemes. A procedure based on Least Absolute Shrinkage and Selection Operator (LASSO) for fault identification is proposed by Zou *et al.* (2011) by using an adaptive LASSO-type penalty function. Since this procedure is based on the maximum likelihood estimation (MLE) approach, its performance is dependent on the number of available OC observations obtained from the

observations only after the estimated change point. If the size of a shift is not sufficiently small, the estimated change point is close to the OC signaled point, thus a small number of samples for the MLE-based fault identification is obtained. Therefore, the LASSO-based procedure may not perform well because of the small number of OC observations (Kim *et al.*, 2016b).

Kim *et al.* (2016b) propose an adaptive step-down procedure under the sparsity assumption for identifying variables whose means are shifted. This procedure selects a variable having no significant evidence of a change at each step based on the variables that are selected in the previous steps. The algorithm stops when there are no variables to classify as the unchanged set. Therefore, it identifies the variables which are not in the unchanged set as changed variables. Instead of using $p!$ number of decompositions, it reduces the computational times using the identified unchanged variables.

2.2 Distribution-Free Methods for Identification of Faulty Variables

The parametric methods for fault identification procedures are based on the fundamental assumption that the process data follow multivariate normal distribution. However, it is well-known that, in many real life applications, the underlying process distribution is unknown or non-normal. Mostly, the process may have a highly skewed distribution. It may be suggested that the non-normal data can be transformed to multivariate normal data. However, this transformation may not be efficient (Qiu, 2008). In this case,

statistical properties of commonly used fault identification procedures, designed to perform well under the normality assumption could be highly affected.

Although the limitations occurring due to the non-normality is crucial, the research work that addresses this problem is sparse. To overcome the concern of the normality assumption, Kim *et al.* (2011), by taking advantages of data description technique, propose a nonparametric fault identification method based on the *k*-nearest neighbor data description, which is called the K^2 decomposition method. In this method, an out of control observation is observed using the K^2–chart (Sukchotrat *et al.*, 2009). The control chart statistic, K^2, is defined as the average distance to the *k*-nearest neighbors. If the K^2 statistic is greater than a predetermined threshold value, then a new observation is considered as an out of control observation. For the identification process, the contribution of each variable is calculated considering the difference between the overall K^2 and the $K^2_{1,\ldots,j-1,j+1,\ldots,p}$ statistics. The contribution of variable j is obtained as follows:

$$K^2_{j|1,2,\ldots,j-1,j+1,\ldots,p} = K^2 - K^2_{1,\ldots,j-1,j+1,\ldots,p)} (\forall j = 1,\ldots, p)$$

where $K^2_{1,\ldots,j-1,j+1,\ldots,p}$ is the K^2 statistic in the reduced space considering all variables except the j^{th} variable. The variables are identified as faulty variables if the

corresponding conditional K^2 statistic is greater than a predetermined threshold value. The threshold value is obtained using the bootstrap method (Efron and Tibshirani, 1994).

In addition, Tuerhong and Kim (2011) propose another distribution-free fault identification method based on a hybrid novelty score (HNS) to calculate a test statistic. Based on the decomposed HNS values, the variables are identified as faulty variables if the corresponding conditional HNS statistic is greater than the threshold value obtained using the bootstrap method.

The contribution of each variable is determined by using the monitoring statistics of those approaches depend on the regression adjustment-based method, which illustrate good performance only in the case with a single mean shift, and their performance deteriorates as the number of faulty variables increase. The main disadvantage of these two methods is that they include the *k*-nearest data description procedure whose performance decreases in a high dimensional data (Tax and Duin, 2004). Consequently, if the process has high dimensional data, the performance of detection of faulty variables might deteriorate. Therefore, there is a need for a suitable fault identification scheme particularly when more than one variable changes in a high dimensional process.

2.3 Bayesian Approach for Identification of Faulty Variables

Existing distribution-free fault identification and parametric methods control the identification procedure through the fixed parameters for a given dataset, which may lead

to unstable parameter estimations. In this case, it would be reasonable to assume that a solution path of the parameter estimation can vary in a probabilistic way. For example, when the dataset includes several outliers around a specific direction, existing procedures may be biased by those outliers resulting in a poor performance. Li *et al.* (2008) introduce a causation-based decomposition fault identification method which integrates the traditional MTY decomposition with a Bayesian causal network by defining the causal relationship between variables. Recently, Tan and Shi (2012) introduce a Bayesian based approach to determine the mean shift and the direction of the shift to identify the root causes. Thus, a smaller number of decompositions may be obtained. Even though these approaches may be appropriate for some processes, they may not be appropriate in certain processes which have unknown causal or Bayesian hierarchical property. In addition, these approaches are built on the multivariate normality assumption. In the case of nonnormality, the identification performance of these procedures may decrease. Even though nonnormality is an important issue specially for real life problems, to date, this problem has received little attention in the research of distribution-free fault identification. Therefore, it will be interesting and beneficial for quality engineers to investigate Bayesian based distribution free fault identification procedures in high-dimensional process.

2.4 Anomaly Detection Procedures

In various applications of real-life problems, identifying the patterns which do not conform to normally behaved patterns is important since these patterns indicate critical

and significant information that can be used towards taking action to recover processes or the applications. This kind of pattern is mostly called an anomaly or an outlier. In the literature, several procedures based on classification based, nearest neighbor based, clustering based, nearest neighbor based, and statistical procedures have been developed to detect anomalies (Chandola *et al.*, 2009).

Among the classification based procedures, one-class classification procedures have a prominent place in the literature. These procedures describe data by obtaining a decision rule based on all the training observations. One of the well-known data description procedure called support vector data description (SVDD) proposed by Tax and Duin (1999) describes data with a hypersphere with minimal volume by transforming original observations into a new space using kernel functions. If the original data is complex, use of kernel trick improves the power of SVDD. SVDD has long been a question of great interest in a wide range of applications mainly focusing on the detection of the anomalies as well as other real-life problems such as face recognition, image processing, pattern detection and quality control (Bovolo *et al.*, 2010, Lee *et al.*, 2006, Ning and Tsung, 2013, Kang *et al.*, 2012, Jeong *et al.*, 2012, Lee *et al.*, 2014, Kim *et al.*, 2016a, Shin *et al.*, 2012).

Most of the SVDD procedures assume that all training observations are obtained from a single known distribution. However, in-real life problems, target data may belong to more than one class. To overcome this drawback, Huang *et al.* (2011) introduce a procedure called two-class SVDD (TC-SVDD) in which normal data consists of two-

classes. However, in many real-world applications, normal data may be obtained from more than two classes. Thus, existing procedures may not recognize the differences between the classes, which results in poor anomaly detection performance. Therefore, in this dissertation, we propose a generalized SVDD procedure based on a Bayesian framework. The proposed procedure can be easily applicable to identify the anomalies in multimode processes.

CHAPTER 3

DISTRIBUTION-FREE ADAPTIVE STEP-DOWN APPROACH FOR FAULT IDENTIFICATION

3.1 Introduction

In quality control area, multivariate statistical process control (MSPC) chart is the primarily used technique to monitor product quality in high-dimensional processes and detect process mean shift or variance change. Even though MSPC charts take advantage of monitoring several process variables simultaneously by considering the correlation among multiple variables, it has a limited ability to identify the variables causing the out-of-control (OC) signal (Doganaksoy *et al.*, 1991). To overcome this limitation, several researchers have developed approaches for identifying faulty variables when the process shift occurs.

Doganaksoy *et al.* (1991) propose a test statistic for each variable to identify the faulty variables. However, this approach does not consider the effect of the correlation between variables. In addition, Hawkins (1991, 1993) proposes an approach based on regression-adjusted variables considering the correlations among variables. This approach is effective under the condition that only one variable is changed. However, when more than one variable are shifted, the identification performance can be significantly decreased (Hawkins, 1991).

Runger (1996) develops a U^2 control chart based on the subset of the variables not causing the mean shift. Based on the subset of unchanged variables, relative contribution of complement of this subset is determined to identify the faulty variables. However, this approach is not suitable when either the subset of unchanged variables contains a large number of variables or the subset is unknown because the identification of the subset of variables which do not cause a mean shift is not straightforward. In addition, Runger *et al.* (1996) introduce another approach which investigates the contribution of variables causing the mean shift. This approach is a polynomial time algorithm for the calculation of conditional T^2 statistic of each variable. Runger's approach is a special case of Hawkins' regression-adjustment approach when only one variable is considered.

By drawing on the concept of conditional T^2, Mason *et al.* (1995, 1997) propose a T^2 decomposition approach which is known as the Mason–Tracy–Young (MTY). This approach decomposes the T^2 statistic into conditional T^2 statistic to identify a changed variable. For large number of variables, however, the computational complexity makes the MTY approach impractical since it needs to examine $p!$ decompositions. To reduce the computational complexity, Sullivan *et al.* (2007) propose an approach which evaluates every possible subset of variables. For *p*-variate observation, this algorithm considers $\binom{p}{k}$ number of subsets where *k* is the size of a set and calculates the T^2 statistic of these subsets. The computational complexity, for large p, still remains an

obstacle even though this approach reduces the computational effort compared to the MTY.

When a process alarms an OC signal, in high-dimensional process, only a few variables are responsible for a process shift, which is known as sparsity property (Zou and Qiu, 2009). Based on the sparsity assumption, Kim *et al.* (2016b) propose an adaptive step-down (ASD) approach for identification of faulty variables by using the knowledge from the identified unchanged variables. Instead of using $p!$ decompositions, they reduce the computational times using the identified unchanged variables.

Despite existing approaches' success under normality assumption, there has been an increasing concern about the normality assumption, it is impractical. Several distribution-free approaches have been proposed (Sukchotrat *et al.*, 2009, Sun and Tsung, 2003) to monitor processes which follow non-normal distribution. However, these approaches may be inefficient and unable to identify the faulty variables in a process. Kim *et al.* (2011) propose a nonparametric fault identification approach based on the k-nearest neighbor data description. Their approach is similar to the Hawkins' approach in determining the contribution of each variable. In addition, Tuerhong and Kim (2011) propose another distribution-free fault identification approach based on a hybrid novelty score (HNS). In this approach, they use a hybrid novelty score to calculate a test statistic. However, because existing nonparametric approaches depend on Hawkins' regression-adjusted approach which performs well when only one variable shifts, the performance of

fault identification could significantly decrease when the number of faulty variables is more than one. Moreover, both approaches use the *k*-nearest neighbor data description for the fault identification approach. Yu *et al.* (2002) indicate that the *k*-nearest neighbor data description approach is suitable for spherical data distribution. On the other hand, having spherical data distribution is not common in real-life applications. Therefore, the power of the approaches based on the *k*-nearest neighbor data description may decrease when the data deviate from the spherical distribution.

Therefore, in this chapter, we propose a distribution-free fault identification approach based on support vector data description (SVDD)-based test statistic. Sun and Tsung (2003) show the superiority of SVDD-based control chart when identifying more than two variables, SVDD uses kernel approaches that provide the advantage of dealing with high-dimension data. In addition, Tax and Duin (2004) show the superiority of the SVDD compared to the *k*-nearest data description in a high-dimensional data.

The proposed distribution-free fault identification approach combines SVDD with an ASD approach to identify the changed variables. The proposed approach has the following distinctive advantages compared with existing ones. First, the proposed approach initially selects the unchanged variables in each step then eventually identifies the changed variables using the set of the selected unchanged variables. This approach reduces the computational times when a few variables are changed in a high dimensional process. Second, the proposed approach is not sensitive to the correlation between

variables, leading to stable performance regardless of the number of changed variables. We discuss the advantages of the proposed approach in further details in Section 3.4.

This chapter is organized as follows. In Section 3.2, we review both parametric and distribution-free existing approaches. In Section 3.3, we propose a distribution-free adaptive step-down fault identification (DFASD) approach based on SVDD. In Section 3.4, the performance of the proposed approach is demonstrated with extensive simulation studies followed by the conclusion in Section 3.5.

3.2 Existing Parametric and Distribution-Free Approaches

In this section, we briefly review some of the existing parametric and distribution-free approaches. The parametric approaches for the fault identification approach are based on the normality assumption while distribution-free approaches have been developed without assumptions about the distributions of the parameters.

Runger *et al.* (1996) propose an approach which depends on the contribution of each variable to the mean shift by determining the difference between overall T^2 statistic and $T^2_{1,\ldots,j-1,j+1,\ldots,p}$ statistic. The contribution of variable j can be expressed as follows:

$$T^2_{j|1,2,\ldots,j-1,j+1,\ldots,p} = T^2 - T^2_{1,\ldots,j-1,j+1,\ldots,p} \ (\forall j = 1,\ldots,p) \tag{3.1}$$

where $T^2_{1,\ldots,j-1,j+1,\ldots,p}$ is calculated by T^2 statistic using all variables except the j^{th} variable. It is shown that $T^2_{j|1,2,\ldots,j-1,j+1,\ldots,p}$ follows a χ^2_1 with degree of freedom one. Because this approach is closely related to the regression-adjustment approach, performance of this approach could deteriorate when the number of changed variables is greater than one.

To reduce computational effort of identifying a few changed variables in high dimensional process, Kim *et al.* (2016b) propose an ASD approach, which is based on the normality of observations, for fault variable identification approach which utilizes unchanged variables in each step by using conditional $T^2_{j|\hat{\Gamma}}$ statistic. In the i^{th} step, unchanged variable is selected as follows:

$$\gamma_i = \underset{j \notin \hat{\Gamma}}{\operatorname{argmin}} T^2_{j|\hat{\Gamma}} \tag{3.2}$$

where $T^2_{j|\hat{\Gamma}} = T^2_{\hat{\Gamma}\cup\{j\}} - T^2_{\hat{\Gamma}}$ and $\hat{\Gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_{i-1}\}$ represents the set of unchanged variables identified in the previous steps. When the algorithm stops, it concludes that the variables which are not in $\hat{\Gamma}$ are considered faulty variables.

On the other hand, if the data do not follow multivariate normal distribution, parametric approaches may perform poorly. To overcome this drawback, Kim *et al.* (2011) propose a distribution-free fault identification K^2 decomposition approach based on the *k*-nearest neighbor data description. In K^2 decomposition approach, an out of control observation

is observed using K^2 chart (Sukchotrat *et al.*, 2009). The control chart statistic is defined as the average distance to the *k*-nearest neighbors.

$$K^2 = \frac{\sum_{i=1}^{k} \left\| \mathbf{z} - NN_i(\mathbf{z}) \right\|}{k} \tag{3.3}$$

where $NN_i(\mathbf{z})$ is the i^{th} nearest neighbor of the new observation $\mathbf{z}$. If the K^2 statistic is greater than the predetermined threshold value, then this new observation is considered as an out of control observation. For the identification process, the contribution of each variable is calculated considering the difference between overall K^2 and $K^2_{1,\ldots,j-1,j+1,\ldots,p}$ statistic. The contribution of variable *j* is obtained as follows:

$$K^2_{j|1,2,\ldots,j-1,j+1,\ldots,p} = K^2 - K^2_{1,\ldots,j-1,j+1,\ldots,p} (\forall j = 1,\ldots,p) \tag{3.4}$$

where $K^2_{1,\ldots,j-1,j+1,\ldots,p}$ is the K^2 statistic in the reduced space considering all variables except the j^{th} variable. The variables are identified as faulty variables if the corresponding conditional K^2 statistic is greater than the threshold value. The threshold value is obtained using the bootstrap approach.

In addition, Tuerhong and Kim (2011) propose a distribution-free approach based on a HNS decomposition approach. Instead of using K^2 decomposition statistic, the HNS decomposition approach calculates a test statistic as follows:

$$HNS(\mathbf{z}) = D^{k}_{average-distance}(\mathbf{z}) \times \left(\frac{2}{1+\exp\left(-D^{k}_{convex-hull}(\mathbf{z})\right)} \right) \tag{3.5}$$

where $D^{k}_{average-distance}(\mathbf{z}) = \frac{1}{k}\sum_{j=1}^{k} \left\| \mathbf{z} - \mathbf{x_j} \right\|$ is the average distance to the *k* nearest neighbors and $D^{k}_{convex-hull}(\mathbf{z}) = \left\| \mathbf{z} - \sum_{j=1}^{k} W_j \mathbf{x_j} \right\|$ is the distance to the convex hull obtained by *k* nearest neighbors. In this equation, convex hull is obtained from the following quadratic optimization problem:

$$\min_{W} \left(D^{k}_{convex-hull}(\mathbf{z}) \right)^2 = \left| \mathbf{z} - \sum_{j=1}^{k} W_j \mathbf{x_j} \right|^2 \tag{3.6}$$

$$s.t. \ \sum_{j=1}^{k} W_j = 1, W_j \geq 0, \forall j$$

The effect of the variable *j* on the OC signal is evaluated as follows:

$$HNS_{j|1,2,..,j-1,j+1,\ldots,p}(\mathbf{z}) = HNS(\mathbf{z}) - HNS_{1,\ldots,j-1,j+1,\ldots,p}(\mathbf{z}), \forall j = 1,\ldots,p) \tag{3.7}$$

Based on the decomposed HNS values, the variables are identified as faulty variables if the corresponding conditional HNS statistic is greater than the threshold value which is obtained using the bootstrap approach.

The main limitation of existing distribution-free approaches is their sensitivity to the number of changed variables. If the number of shifted variables is greater than one, those approaches do not perform well because the control statistic based on regression adjustment might be distorted when unchanged variables are not well identified.

3.3 The Distribution Free Fault Variable Identification Using Adaptive Step-Down Approach (DFASD)

To overcome some of the drawbacks of the existing approaches such as the presence of correlation effects between variables, effects of the number of changed variables, and extensive computational efforts in high dimensional data, we propose an DFASD for fault identification. The proposed approach combines SVDD-based test statistic with an ASD approach which selects an unchanged variable at each step based on the variables that are selected in previous steps.

3.3.1 Support Vector Data Description–Based Test Statistic

Given in-control data $\left\{\mathbf{x_i}|\mathbf{x_i} \in R^p\right\}, i=1,\ldots,N,$ the main goal of SVDD is to find a hypersphere which covers the data with minimal volume, with center $\mathbf{a}$ and radius R (Tax and Duin, 1999). Misclassification in the in-control data is allowed by introducing variable ε_i to penalize outliers for the largest distance between $\mathbf{x}_i$ and $\mathbf{a}$. The primal formulation is constructed as follows:

$$\begin{aligned} &\min R^2 + C\sum_{i=1}^{N}\varepsilon_i \\ &s.t.\ \ \left\|\mathbf{x_i} - \mathbf{a}\right\|^2 \le R^2 + \varepsilon_i,\ \varepsilon_i \ge 0\ \forall i \in \{1,\ldots,N\} \end{aligned} \tag{3.8}$$

where C is the regularization parameter which adjusts the volume of the sphere by considering the number of in-control observations that fall outside the boundary. The dual formulation is obtained by using the Lagrangian function:

$$L\left(R,\mathbf{a},\varepsilon_i\right) = R^2 + C\sum_{i=1}^{N}\varepsilon_i - \sum_{i=1}^{N}\tau_i\left(R^2 + \varepsilon_i - \left(\mathbf{x_i} - \mathbf{a}\right)^T\left(\mathbf{x_i} - \mathbf{a}\right)\right) - \sum_{i=1}^{N}\gamma_i\varepsilon_i \tag{3.9}$$

where $\tau_i \ge 0$ and $\gamma_i \ge 0$ are the Lagrange multipliers. By taking the partial derivatives of Lagrangian function:

$$\frac{\partial L}{\partial R} = 0 \rightarrow \sum_{i=1}^{N}\tau_i = 1 \tag{3.10}$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0 \rightarrow \mathbf{a} = \frac{\sum_{i=1}^{N}\tau_i\mathbf{x}_i}{\sum_{i=1}^{N}\tau_i} = \sum_{i=1}^{N}\tau_i\mathbf{x}_i \tag{3.11}$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \rightarrow C - \tau_i - \gamma_i = 0 \tag{3.12}$$

From the Eq. (3.12), $\tau_i = C - \gamma_i$ and since $\tau_i \geq 0, \gamma_i \geq 0$, Langrange multipliers $\gamma_i$can be removed when we demand that $0 \leq \tau_i \leq C$. In this case, dual formulation is constructed by substituting the Eqs. (3.10), (3.11) and (3.12) into the Eq. (3.9).

$$\begin{aligned} &\max \sum_{i=1}^{N} \tau_i \left(\mathbf{x_i} \cdot \mathbf{x_i}\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} \tau_i \tau_j \left(\mathbf{x_i} \cdot \mathbf{x_j}\right) \\ &s.t.\ 0 \leq \tau_i \leq C, i = 1, 2, \ldots, N, \sum_{i=1}^{N} \tau_i = 1 \end{aligned} \tag{3.13}$$

In-control points corresponding to the positive τ_i are called support vector and they are placed on the boundary or outside of the boundary. After SVDD boundary is obtained from the in-control data, a new observation **z** is considered out of control if the distance from **z** to the center **a** is greater than the radius R. This is based on the new M statistics which is defined by Eq. (3.14).

$$M = \|\mathbf{z} - \mathbf{a}\|^2 = \left(\mathbf{z} - \sum_{i=1}^{N} \tau_i \mathbf{x_i}\right)^T \left(\mathbf{z} - \sum_{i=1}^{N} \tau_i \mathbf{x_i}\right) \tag{3.14}$$

$$= \mathbf{z^T z} - 2\sum_{i=1}^{N} \tau_i \mathbf{x_i^T z} + \sum_{i,j=1}^{N} \tau_i \tau_j \mathbf{x_i x_j}$$

If the inner product is kernelized by a kernel function, more suitable boundary to cover the data can be obtained, thus the new formulation is given by Eq. (3.15) as follows:

$$\begin{aligned} &\max \sum_{i=1}^{N} \tau_i K\left(\mathbf{x_i}, \mathbf{x_i}\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} \tau_i \tau_j K\left(\mathbf{x_i}, \mathbf{x_j}\right) \\ &s.t.\ 0 \le \tau_i \le C, i = 1, 2, \ldots, N,\ \sum_{i=1}^{N} \tau_i = 1 \end{aligned} \tag{3.15}$$

Where $K(\cdot)$ is a kernel function. Using Eq. (3.14), we obtain a new *M* statistic as shown in Eq. (3.16):

$$M = K\left(\mathbf{z}, \mathbf{z}\right) - 2\sum_{i=1}^{N} \tau_i K\left(\mathbf{x_i}, \mathbf{z}\right) + \sum_{i,j=1}^{N} \tau_i \tau_j K\left(\mathbf{x_{i,}} \mathbf{x_j}\right) \tag{3.16}$$

3.3.2 Distribution Free Adaptive Step-Down Approach

In this section, we propose an DFASD which utilizes the set of unchanged variables using conditional *M* statistic to identify the faulty variables. Starting with an empty set, we identify an unchanged variable in each step having no significant evidence of being changed. In the i^{th} step, the unchanged variable is selected as follows:

$$\beta_i = \underset{j \notin \hat{\Gamma}}{\operatorname{argmin}}\, M_{j|\hat{\Gamma}} \tag{3.17}$$

where $M_{j|\hat{\Gamma}} = M_{j \cup \hat{\Gamma}} - M_{\hat{\Gamma}}$ and $\hat{\Gamma} = \{\beta_1, \beta_2, \ldots, \beta_{i-1}\}$ is the set of variables assigned as unchanged set by the previous steps. When $i = 2$, the set $\hat{\Gamma}$ contains only one variable. Thus $M_{\hat{\Gamma}} = M_j$, where M_j is the unconditional statistic of an individual variable which

is obtained in the first step. To calculate $M_{j\cup\hat{\Gamma}}$ and $M_{\hat{\Gamma}}$, we use the same support vectors obtained from the full dimension. For example, if $p=3$, $M_{1|23}$ can be expressed as

$$M_{1|23} = M_{123} - M_{23} \tag{3.18}$$

Vectors $\mathbf{x_i} = (x_{i1}, x_{i2}, x_{i3}) \in S$, where S is a set of support vectors and corresponding dual variables $\{\tau_i > 0\}$ are obtained by solving the optimization problem in Eq. (3.15) for the full dimension. In this case, for a new observation $\mathbf{z} = (z_1, z_2, z_3)$

$$M_{123} = K(\mathbf{z},\mathbf{z}) - 2\sum_{i\in\boldsymbol{S}} \tau_i K(\mathbf{z},\mathbf{x_i}) + \sum_{i,j\in\boldsymbol{S}} \tau_i\tau_j K(\mathbf{x_i},\mathbf{x_j}) \tag{3.19}$$

$$M_{23} = K(\mathbf{z}',\mathbf{z}') - 2\sum_{i\in\boldsymbol{S}} \tau_i K(\mathbf{z}',\mathbf{x}'_{\mathbf{i}}) + \sum_{i,j\in\boldsymbol{S}} \tau_i\tau_j K(\mathbf{x}'_{\mathbf{i}},\mathbf{x}'_{\mathbf{j}}) \tag{3.20}$$

where $\mathbf{x}'_{\mathbf{i}} = (x_{i2}, x_{i3})$ and $\mathbf{z}' = (\mathrm{z}_2, \mathrm{z}_3)$.

Figure 3.1 shows the flow chart of the proposed approach.

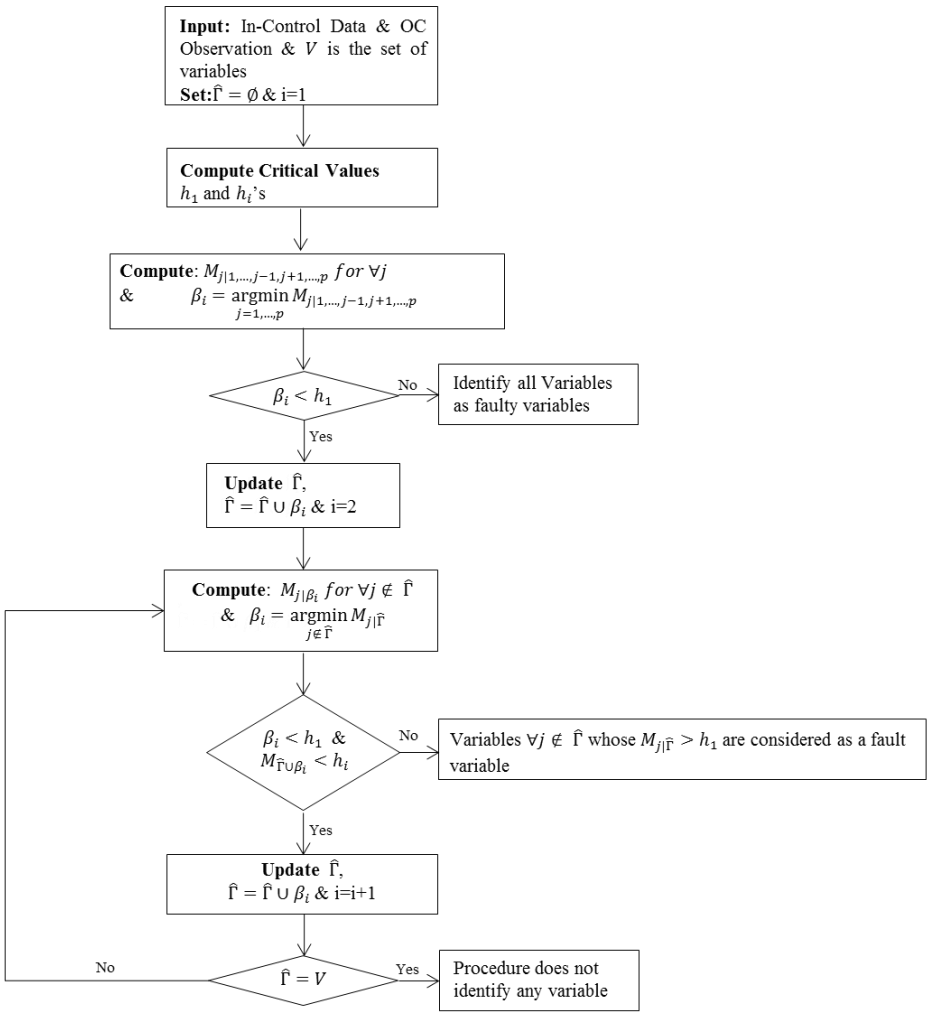


Figure 3.1 Flowchart of the proposed DFASD approach

3.3.2.1 Initial Variable Selection

Since the proposed approach depends on identification of unchanged variables, in each step an unchanged variable is identified according to the Eq. (3.17). Since $\hat{\Gamma}$ is empty in the first step, the first unchanged variable is identified as follows:

$$\beta_1 = \underset{j=1,\ldots,p}{\operatorname{argmin}} M_{j|1,\ldots,j-1,j+1,\ldots,p} \tag{3.21}$$

where $M_{j|1,\ldots,j-1,j+1,\ldots,p}$ is the conditional M statistic of the j^{th} variable. In this case, $M_{j|1,\ldots,j-1,j+1,\ldots,p} = M - M_{1,\ldots,j-1,j+1,\ldots,p}$ where $M_{1,\ldots,j-1,j+1,\ldots,p}$ is the M statistic of the reduced space spanned by variables $\{1,\ldots,j-1,j+1,\ldots,p\}$.

In the proposed DFASD approach, there are two rules to identify faulty variables. The first rule depends on conditional $M_{\beta_i|\hat{\Gamma}}$. If the minimum conditional $M_{\beta_i|\hat{\Gamma}}$ is greater than pre-determined threshold value h_1, then all $j \notin \hat{\Gamma}$ are considered as changed variables. The second rule is: if $M_{\hat{\Gamma}\cup\beta_i}$ is greater than the threshold value h_i, then the variables $j \notin \hat{\Gamma}$ whose $M_{j|\hat{\Gamma}} > h_1$ are considered as a faulty variable. This rule prevents adding changed variable to the set of unchanged variables $\hat{\Gamma}$. When $M_{\hat{\Gamma}\cup\beta_i} > h_i$ or $M_{\beta_i|\hat{\Gamma}} > h_1$, we stop and conclude that the changed variables are all identified. However, if $M_{\hat{\Gamma}\cup\beta_i} < h_i$ and $M_{\beta_i|\hat{\Gamma}} < h_1$, β_i is identified as unchanged variable. In the first step, if $\min_{j=1,\ldots,p} M_{j|1,\ldots,j-1,j+1,\ldots,p}$ is greater than the threshold value h_1, then algorithm stops and all variables are identified as faulty variables.

3.3.2.2 Determination of Threshold Values

To determine the faulty variables, appropriate threshold values should be established. If $M_{\beta_i|\hat{\Gamma}}$ is greater than the threshold value h_1, then all $j \notin \hat{\Gamma}$ are identified as faulty variables. Since the M statistic is obtained in a distribution-free manner, it is difficult to conduct parametric inference on the M decomposition values. To determine a threshold value h_1, in this chapter, bootstrap resampling approach is used. For the determination of h_1, we use the following steps:

Step 1: Assume that there are *p*-variate N in-control observations.

Step 2: Calculate all the possible conditional $M_{\beta_i|\hat{\Gamma}}$ statistics. There are $n_1 = N \times \sum_{i=2}^{p} \binom{p}{i}$ conditional statistic (For simplicity $M_{\beta_i|\hat{\Gamma}} = M_C$).

Step 3: Using bootstrap, obtain $M_{c,r,1}, M_{c,r,1}, \ldots, M_{c,r,n_1}$ conditional statistics from the r^{th} bootstrap sample $(r = 1, \ldots, T)$.

Step 4: Sort the conditional statistics of each bootstrap sample $M_{c,r,(1)} < M_{c,r,(2)} < \ldots < M_{c,r,(n_1)}$

Step 5: In each bootstrap sample, find the s^{th} statistic where $s = n_1 \times (1-\alpha_1)$ and α_1 is the significance level for faulty variables (s is rounded to the closet integer).

Step 6: Calculate the threshold value by taking the average of s^{th} statistics:

$$h_1 = \frac{1}{T} \sum_{r=1}^{T} M_{c,r,(s)}$$

In addition, to prevent adding a changed variable to the set of unchanged variables $\hat{\Gamma}$, we need to satisfy $M_{\hat{\Gamma} \cup \{\beta_i\}} < h_i$. The threshold value h_i can be obtained by using a bootstrap resampling approach which is similar to the calculation of h_1. In Step 1, instead of calculating all the possible conditional $M_{\beta_i|\hat{\Gamma}}$ statistic, all possible $M_{\hat{\Gamma} \cup \beta_i}$ statistic are calculated where $\left|\hat{\Gamma} \cup \{\beta_i\}\right| = i$ (for simplicity $M_{\hat{\Gamma} \cup \beta_i} = M_U$). There is $n_2 = N \times (p-1)$ statistic. In Step 5, find the K^{th} statistic where $s = n_2 \times (1-\alpha_2)$ and α_2 is a significance level for the group of unchanged variables. Therefore, h_i is obtained as follows:

$$h_i = \frac{1}{T} \sum_{r=1}^{T} M_{U,r,(s)}$$

Under the normal operating conditions, the contribution of each variable in the first step of ASD algorithm is approximately equal since the algorithm calculates the contribution

of one variable conditioning on all other variables. In the following steps, algorithm checks the only one variable's effect conditioning on the previously identified unchanged variables. For example, $h_1^{(i)}$ shows the critical value obtained in step $i(i>1)$. In step i, critical values are obtained by conditioning on previously identified $i-1$ variables. Therefore, according to our set up, these critical values should be similar to the each other because the critical values are obtained based on the only one variable's effect by removing the effect of previously identified $i-1$ unchanged variables.

To compare the $h_1^{(i)}$ values, we conduct a simulation study and its results are shown in Table 3.1. We generate datasets that follow three distributions multivariate gamma multivariate lognormal and multivariate normal. By using the proposed bootstrap approach, we obtain the original h_1 values. In addition, we use the similar bootstrap approach to obtain $h_1^{(i)}$ values for each step. The h_1 and $h_1^{(i)}$ results obtained from the bootstrap approach are shown in Table 3.1.

Table 3.1 Comparison of critical values for different datasets ($p = 3$)

Step 1 ($M_{j\|1,2,\ldots,j-1,j,\ldots,p}$)	Step 2 ($M_{i\|j}$)	Step 3 ($M_{k\|i,j}$)
$h_1^{(1)}$	$h_1^{(2)}$	$h_1^{(3)}$
Multivariate Gamma (Original $h_1 = 0.0323$)		
0.0322	0.0323	0.0322
Multivariate Normal (Original $h_1 = 0.2868$)		
0.2869	0.2870	0.2869
Multivariate Lognormal (Original $h_1 = 0.8836$)		
0.8800	0.8845	0.8800

Table 3.1 shows the three different critical values obtained in each step. It concludes that even considering each step individually, it does not change the overall performance since the $h_1^{(i)}$ values are almost the same as the critical value h_1. Therefore, decomposing critical value, h_1 as $h_1^{(1)}, h_1^{(2)}$ and $h_1^{(3)}$ may not affect the performance significantly considering the complexity of computations to obtain individual $h_1^{(i)}$.

3.4 Simulation Study

3.4.1 Simulation Setup

In this section, we evaluate the performance of the proposed approach for faulty variables compared with that of existing approaches such as ASD, K^2 decomposition approach and HNS decomposition by using both normality and non-normality dataset. For non-normal distributions, we take several common distributions into our experiments such as

multivariate gamma and multivariate lognormal distribution. Moreover, for more irregular distributed data shape, multivariate banana-shaped data is considered in the experiment. For lognormal distribution dataset, we assume mean vector $\boldsymbol{\mu}_0 = 0$ and covariance $\boldsymbol{\Sigma}_0 = \left[\sigma_{ij}\right]_{1\le i,j\le p}$, where $\sigma_{ii} = 0.1$, $\sigma_{ij} = 0.1\rho$, and $|\rho| < 1$ in order to generate the data from a narrow range. This would be assumed for the other distributions as well. To generate a gamma distribution dataset, the same approach used in Stoumbos and Sullivan (2002) is employed by assuming the shape and scale parameters to have values of one. The banana-shaped dataset shown in Figure 3.2 is generated by the approach described in Duin *et al.* (2000). For instance, a six-dimensional dataset is obtained by integrating three two-dimensional banana-shaped datasets. In each simulation run, 500 in-control and 1000 out of control observations are generated. The critical values of each approach are determined by using in control datasets.

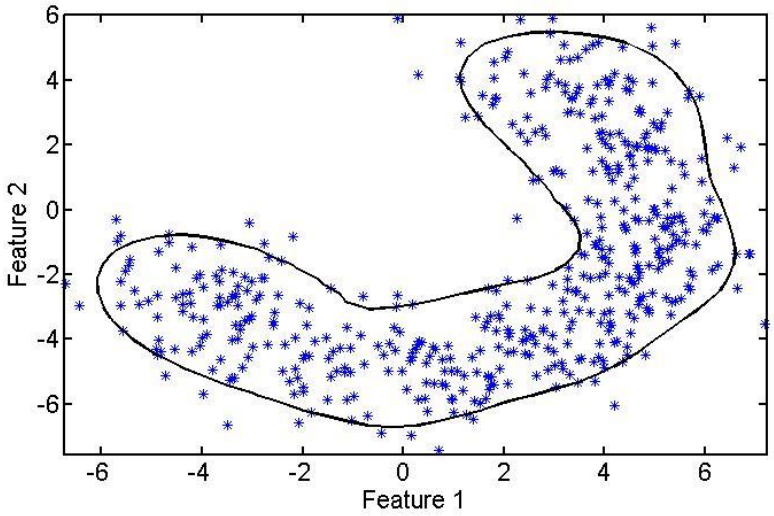


Figure 3.2 Two-dimensional banana-shaped data

A mean vector for out-of-control data is defined as $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta}$ where $\boldsymbol{\delta} = \left[\delta_1 \delta_2 \ldots \delta_p \right]$ and δ_i represents the shift size of the i^{th} variable. For example, $\{d, 0, 0, \ldots\}$ indicates that mean shift occurs at the first variable and the size of the shift is given by the standard deviation of corresponding variable. However, for the six dimensional banana-shaped data, we determine the shift size as an additive mean because the theoretical standard deviation is not straightforward. Moreover, for the optimization problem in Eq. (3.15), the parameter C is determined using the following equation (Tax and Duin, 2004): $C \cong \frac{1}{N \times (\, fraction of\ outliers)}$. This equation means that some of the in-control data are allowed to be outside of the SVDD boundary by controlling the C value. Tax and Duin

(2004) show that Eq. (3.15) is optimized by choosing the optimum *C* value. In our experiments, we set the fraction of outliers to α_1 value.

3.4.2 Measures of Performance

For performance comparisons, in this chapter, two performance measures are used. The first measure checks whether the identification result matches mean shift vector. This is evaluated using the correctness rates (CR) which is defined as

$$\mathrm{CR} = \frac{\sum_{i=1}^{n} I_{\Gamma=\hat{\Gamma}}}{n} \tag{3.22}$$

where n is the number of identifications and I is the indicator function. However, CR may not be suitable as the number of variables increases. In this sense, we adopt the second measurement called the expected error rates (EER) in mean shift (Zou *et al.*, 2011):

$$\mathrm{EER} = E\left(\frac{\text{Number of errors}}{\text{Number of variables}}\right) \tag{3.23}$$

Through EER, all variables in the observation vector are checked one by one whether they are correctly identified or not. Therefore, by using these two performance measures together, we can evaluate the effectiveness of the proposed approach in identifying the shifted variables.

3.4.3 Choice of Kernel Function for DFASD

The performance of SVDD is strictly based on the choice of the kernel function and parameters of the kernel function which directly controls the nonlinear mapping of the features. Therefore, choices of the kernel function and kernel parameters play an important role in the performance of the proposed DFASD approach. Ning and Tsung (2013) mention that it would be easy to choose the appropriate kernel and its parameters for two dimensional data. However, in high-dimension, it would be difficult to obtain the optimal kernel function with appropriate parameters. Therefore, it is not easy to suggest the specific kernel function for the SVDD based approaches specially for high-dimensional cases. In addition, different kernel functions have their own advantages and disadvantages according to the data properties such as the dimension, correlation and dispersion. Therefore, it is more difficult to suggest one optimal kernel function. One of the popular kernel functions is called Gaussian kernel function which is defined as:

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2w^2}\right) \tag{3.24}$$

where w is a parameter of the Gaussian kernel. For the SVDD-based control chart such as K-chart, Ning and Tsung (2013) propose an approach to choose the parameters of the Gaussian kernel based on the cross validation. Another traditional kernel function is a polynomial kernel function defined as follows:

$$K(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^T\mathbf{y} + 1\right)^d \tag{3.25}$$

where d represents the order of the polynomial kernel.

With these advantages and disadvantages of different kernel functions, the practitioners can choose the proper function based on their engineering knowledge. Once the appropriate kernel function is selected, then the complexity parameter that controls the feature of kernel is determined. Many criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC) and general cross validation (GCV) can be applied to obtain the optimal parameters. In this chapter, we compare the CR performances of DFASD approach under polynomial and Gaussian kernel setting. For simplicity, we use the polynomial kernel function as

$$K(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \tag{3.26}$$

where $\phi(\mathbf{x}) = \left(x_1, x_2, \ldots, x_p, \frac{x_1^2}{2!}, \frac{x_2^2}{2!}, \ldots, \frac{x_p^2}{2!}, \ldots, \frac{x_1^d}{d!}, \frac{x_2^d}{d!}, \ldots, \frac{x_p^d}{d!} \right)$. Tables 3.2 and 3.3 demonstrate the performances for different kernel functions.

Table 3.2 The effect of kernel functions according to different parameters under multivariate gamma distribution ($p = 5$)

Kernel Type		Polynomial kernel with parameter d				Gaussian kernel with parameter w				
Direction	Size	2	3	4	5	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}
{d 0 0}	1.0 σ	0.1360	0.1090	0.0961	0.1327	0.0760	0.1340	0.1280	0.1450	0.1110
	2.0 σ	0.2070	0.1910	0.1947	0.1918	0.0630	0.2010	0.2030	0.1790	0.1610
	3.0 σ	0.4350	0.4246	0.4047	0.4328	0.1230	0.5000	0.6170	0.4310	0.4220
{d 0 d}	1.0 σ	0.0870	0.0858	0.0844	0.0823	0.0724	0.0990	0.0968	0.0877	0.0913
	2.0 σ	0.2520	0.2431	0.2442	0.2471	0.1514	0.2268	0.2535	0.2548	0.2410
	3.0 σ	0.4950	0.4882	0.4459	0.4703	0.1696	0.3014	0.4821	0.4859	0.4792

Table 3.3 The effect of kernel functions according to different parameters under banana-shaped data ($p = 6$)

Kernel Type		Polynomial kernel with parameter d				Gaussian kernel with parameter w				
Direction	Size	2	3	4	5	5^{4}	5^{5}	5^{6}	5^{7}	5^{8}
{d 0 0}	1.0	0.1628	0.1582	0.1558	0.1506	0.1370	0.1310	0.1330	0.1350	0.1370
	2.0	0.3079	0.3011	0.3057	0.2971	0.2740	0.2860	0.2690	0.2430	0.2350
	3.0	0.4152	0.3904	0.4037	0.4150	0.3420	0.3580	0.3440	0.3770	0.3730
{d 0 d}	1.0	0.0485	0.0472	0.0431	0.0467	0.0390	0.0460	0.0570	0.0360	0.0390
	2.0	0.1340	0.1488	0.1348	0.1451	0.1010	0.1150	0.1130	0.1120	0.1160
	3.0	0.2459	0.2182	0.2113	0.2194	0.1930	0.1830	0.2210	0.1820	0.1720

In Tables 3.2 and 3.3, the results suggest the use of a polynomial kernel with the second order for gamma distribution and Gaussian kernel with the appropriate parameter from the table for banana-shaped data. Although the tables provide a guide for choosing the kernel type and their parameters, discussing many different kernel functions in this

chapter is beyond the scope of this research. Therefore, in this chapter, we use the kernel defined in Eq. (3.26) due to the similar results obtained from the simulations.

3.4.4 Simulation Results

Tables 3.4 and 3.5 compare the performace of the proposed DFASD approach, K^2 and HNS decomposition and ASD versions of K^2 and HNS decomposition approaches (K^2-ASD and HNS-ASD) under multivariate gamma and multivariate normal distributions, respectively. Combining ASD with existing K^2 or HNS decomposition approach affects the performance of these two approaches as shown in Tables 3.4 and 3.5.

Table 3.4 Performance comparison of DFASD, K^2 decomposition, HNS decomposition, K^2-ASD and HNS-ASD with multivariate gamma distribution with $p = 3$

Shift		DFASD		K^2		HNS		K^2 ASD		HNS ASD	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER	CR	EER
{d 0 0 }	1.0 σ	0.2325	0.4101	0.4350	0.2679	0.4680	0.2454	0.4940	0.2435	0.4324	0.2586
	2.0 σ	0.3237	0.3380	0.6644	0.1365	0.7635	0.1012	0.8927	0.0524	0.8519	0.0637
	3.0 σ	0.5640	0.1893	0.5705	0.1530	0.7813	0.0824	0.9228	0.0349	0.8856	0.0459
{d 0 d }	1.0 σ	0.2372	0.3158	0.0650	0.4930	0.2296	0.3601	0.1057	0.4241	0.1760	0.3762
	2.0 σ	0.4577	0.2081	0.0707	0.4687	0.2650	0.3370	0.1809	0.3563	0.3437	0.2555
	3.0 σ	0.6946	0.1193	0.0579	0.4342	0.2590	0.2594	0.1612	0.3579	0.3543	0.2435

Table 3.5 Performance comparison of DFASD, K^2 decomposition, HNS decomposition, K^2-ASD and HNS-ASD with multivariate normal distribution with $p = 3$

Shift		DFASD		K^2		HNS		K^2 ASD		HNS ASD	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER	CR	EER
{d 0 0 }	1.0 σ	0.3783	0.3136	0.3003	0.2626	0.3602	0.2566	0.6710	0.1783	0.6172	0.2037
	2.0 σ	0.6374	0.1720	0.3393	0.2264	0.4215	0.2061	0.8727	0.0668	0.8279	0.0864
	3.0 σ	0.8014	0.0899	0.2818	0.2519	0.3754	0.2342	0.8717	0.0677	0.8435	0.0801
{d 0 d }	1.0 σ	0.3082	0.3454	0.0061	0.6162	0.1233	0.5103	0.0472	0.5887	0.2236	0.4700
	2.0 σ	0.5617	0.2254	0.0009	0.6226	0.1470	0.4642	0.0160	0.6866	0.1915	0.4899
	3.0 σ	0.8066	0.0815	0.0004	0.5535	0.1941	0.3362	0.0035	0.5564	0.2346	0.3417

Tables 3.4 and 3.5 indicate that the proposed DFASD approach is comparable with existing K^2 and HNS decomposition approaches for large shift when more than one variable changes. On the other hand, when more than one variable changes, the proposed approach outperforms the existing approaches. In addition, results from Tables 3.4 and 3.5 show that combining ASD with K^2 or HNS slightly improves the performance of original approaches. Although K^2-ASD and HNS-ASD improves the performance of original K^2 or HNS approaches, DFASD still outperforms K^2-ASD and HNS-ASD. Therefore, in this chapter, for the simulation studies, DFASD is compared only with K^2 and HNS decomposition approaches.

Table 3.6 compares the performace of the proposed DFASD approach and the ASD approach under the six dimensional multivariate normal distribution. The performance results are an average of 1,000 simulation runs of each scenario. Table 3.6 shows that the

ASD approach outperforms the DFASD approach because ASD approach is designed for normally distributed data. However, as the number of the shifted variable increases, the performance of ASD significantly decreases because ASD is based on the sparsity assumption while our proposed DFASD has a stable performance regardless of the number of shifted variables.

Table 3.6 Performance comparison of DFASD and ASD under multivariate normal data $(p = 6)$

Shift		DFASD		ASD	
Direction	Size	CR	EER	CR	EER
{d 0 0 0 0 0}	1.0σ	0.2624	0.2889	0.5256	0.1017
	1.5σ	0.4208	0.2057	0.7781	0.0478
	2.0σ	0.5509	0.1534	0.9100	0.0194
	2.5σ	0.6524	0.1200	0.9584	0.0083
	3.0σ	0.7148	0.0998	0.9706	0.0053
{d 0 d 0 0 0}	1.0σ	0.1468	0.3138	0.1230	0.2591
	1.5σ	0.2752	0.2427	0.3523	0.1751
	2.0σ	0.4283	0.1863	0.6598	0.0875
	2.5σ	0.5924	0.1339	0.8927	0.0281
	3.0σ	0.7056	0.0983	0.9728	0.0072
{d 0 d 0 d 0}	1.0σ	0.1129	0.3709	0.0292	0.4647
	1.5σ	0.2233	0.3013	0.1368	0.4051
	2.0σ	0.3800	0.2255	0.3599	0.3169
	2.5σ	0.5672	0.1490	0.5562	0.2612
	3.0σ	0.7082	0.0965	0.5729	0.2677

Table 3.7 presents the performance of the banana-shaped data with six dimensions. Since the banana-shaped data are deviated from the normal distribution, it is difficult to obtain

an ellipsoid boundary. From Table 3.7, it is apparent that the performance of the proposed DFASD is better than that of existing approaches in all cases. All of the distribution-free approaches outperform ASD in this scenario as expected. In addition, the performance of K^2 decomposition approach deteriorates and performance of HNS approach does not perform as well as the proposed approach when the number of shifted variables increases because the existing approaches are not optimal for identifying multiple changed variables.

Table 3.7 Performance comparison of DFASD, ASD, K^2 decomposition and HNS decomposition with six dimensional banana-shaped data

Shift		DFASD		ASD		K^2		HNS	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER
{d 0 0 0 0 0 }	1.0	0.1628	0.2401	0.0065	0.1837	0.0519	0.2463	0.0142	0.5194
	1.5	0.2363	0.2124	0.0086	0.1806	0.0738	0.2306	0.0217	0.4844
	2.0	0.3079	0.1857	0.0186	0.1768	0.1119	0.2123	0.0350	0.4623
	2.5	0.3656	0.1651	0.0397	0.1713	0.1648	0.1921	0.0537	0.4312
	3.0	0.4152	0.1477	0.0754	0.1640	0.2258	0.1732	0.0600	0.4035
{d 0 d 0 0 0 }	1.0	0.0485	0.3277	0.0000	0.3421	0.0040	0.3700	0.0178	0.4903
	1.5	0.0868	0.2890	0.0000	0.3387	0.0031	0.3536	0.0299	0.4532
	2.0	0.1340	0.2546	0.0001	0.3346	0.0041	0.3357	0.0400	0.4202
	2.5	0.2395	0.2277	0.0007	0.3276	0.0080	0.3163	0.0541	0.3901
	3.0	0.2459	0.2061	0.0028	0.3176	0.0153	0.2966	0.0690	0.3628
{d 0 d 0 d 0 }	1.0	0.0260	0.4231	0.0000	0.5021	0.0004	0.5027	0.0249	0.4755
	1.5	0.0377	0.3763	0.0000	0.4996	0.0002	0.4924	0.0312	0.4462
	2.0	0.0502	0.3385	0.0000	0.4953	0.0002	0.4783	0.0406	0.4160
	2.5	0.0868	0.2993	0.0000	0.4876	0.0003	0.4615	0.0517	0.3885
	3.0	0.1203	0.2723	0.0001	0.4744	0.0007	0.4401	0.0655	0.3646

The performance of distribution-free approaches from multivariate lognormal and gamma distribution scenarios for the three dimensional datasets are illustrated in Tables 3.8 and Table 3.9. It is shown that the DFASD approach is comparable with other distribution-free approaches when one variable is shifted with a large shift. In addition, the DFASD approach outperforms others significantly when two variables are shifted. Since the test statistic in the K^2 decomposition and HNS decomposition approaches are developed based on Hawkins' regression-adjusted variables approach, they only perform well when only one variable changes. These results provide further support for the hypothesis that the proposed DFASD is not affected by the number of shifted variables.

Table 3.8 Performance comparison of DFASD, K^2 decomposition and HNS decomposition under multivariate lognormal when $p = 3$

Shift		DFASD		K^2		HNS	
Direction	Size	CR	EER	CR	EER	CR	EER
{d 0 0}	1.0σ	0.1930	0.4506	0.3502	0.2600	0.3374	0.2746
	1.5σ	0.2575	0.3738	0.3737	0.2249	0.3727	0.2346
	2.0σ	0.3820	0.2890	0.3086	0.2400	0.3148	0.2518
	2.5σ	0.5651	0.2021	0.2308	0.2691	0.2353	0.2951
	3.0σ	0.7465	0.1233	0.1795	0.3009	0.1911	0.3300
{d 0 d}	1.0σ	0.1858	0.4935	0.0156	0.6368	0.0882	0.5738
	1.5σ	0.2530	0.4607	0.0085	0.6365	0.0876	0.5688
	2.0σ	0.3588	0.3791	0.0071	0.5949	0.0823	0.4949
	2.5σ	0.5252	0.2609	0.0062	0.5329	0.0802	0.4202
	3.0σ	0.7366	0.1335	0.0053	0.4793	0.0779	0.3495

Table 3.9 Performance comparison of DFASD, K^2 decomposition and HNS decomposition under multivariate gamma when $p = 3$

Shift		DFASD		K^2		HNS	
Direction	Size	CR	EER	CR	EER	CR	EER
{d 0 0}	1.0σ	0.2325	0.4101	0.4350	0.2679	0.4680	0.2454
	1.5σ	0.2727	0.3788	0.5743	0.1866	0.6120	0.1674
	2.0σ	0.3237	0.3380	0.6644	0.1365	0.7635	0.1012
	2.5σ	0.4046	0.2785	0.6414	0.1343	0.7763	0.0981
	3.0σ	0.5640	0.1893	0.5705	0.1530	0.7813	0.0824
{d 0 d}	1.0σ	0.2372	0.3158	0.0650	0.4930	0.2296	0.3601
	1.5σ	0.3422	0.2591	0.0662	0.4820	0.2362	0.3551
	2.0σ	0.4577	0.2081	0.0707	0.4687	0.2650	0.3370
	2.5σ	0.5695	0.1693	0.0651	0.4590	0.2814	0.3160
	3.0σ	0.6946	0.1193	0.0579	0.4342	0.2590	0.2594

To explore the advantage of the proposed DFASD approach, five dimensional dataset, which follow a multivariate normal, multivariate lognormal and multivariate gamma distribution, respectively, are employed when one variable is shifted and when three variables are shifted. In Table 3.10, we show that when only one variable shifts, the K^2 decomposition approach is better than that of DFASD approach specially for smaller size shifts. However, when three variables are shifted, the proposed DFASD approach outperforms the K^2 decomposition approach in all cases. The CR of K^2 decomposition approach, for example, approaches zero when there are three faulty variables. These

results further support that the proposed DFASD approach shows superiority when compared with the other distribution-free approaches when several variables are faulty.

Table 3.10 Performance comparison of DFASD and K^2 decomposition when $p = 5$

		Multivariate Normal				Multivariate Lognormal				Multivariate Gamma			
Shift		DFASD		K^2		DFASD		K^2		DFASD		K^2	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER	CR	EER	CR	EER
{d 0 0 0 0}	1.0σ	0.288	0.293	0.392	0.157	0.132	0.392	0.529	0.136	0.136	0.380	0.498	0.177
	1.5σ	0.443	0.214	0.507	0.110	0.201	0.308	0.657	0.086	0.162	0.341	0.717	0.099
	2.0σ	0.569	0.160	0.563	0.092	0.328	0.233	0.664	0.077	0.207	0.297	0.804	0.062
	2.5σ	0.668	0.125	0.581	0.087	0.518	0.167	0.642	0.078	0.280	0.249	0.835	0.047
	3.0σ	0.736	0.100	0.575	0.089	0.688	0.115	0.622	0.081	0.435	0.175	0.840	0.041
{d 0 d 0 d}	1.0σ	0.153	0.390	0.000	0.592	0.068	0.523	0.001	0.594	0.087	0.345	0.006	0.497
	1.5σ	0.267	0.319	0.000	0.609	0.119	0.477	0.000	0.628	0.157	0.291	0.003	0.497
	2.0σ	0.417	0.240	0.000	0.619	0.221	0.383	0.000	0.631	0.252	0.247	0.001	0.506
	2.5σ	0.587	0.158	0.000	0.618	0.416	0.254	0.000	0.613	0.355	0.215	0.001	0.520
	3.0σ	0.736	0.095	0.000	0.602	0.665	0.129	0.000	0.584	0.495	0.160	0.001	0.530

In addition, we explore the effect of α_1 and α_2 by using the multivariate normal distribution. Table 3.11 provides the CR performance of the DFASD approach by combining α_1 with α_2. This table is quite revealing in several ways of choosing the α values. It is shown that identifying the large shift is accomplished successfully with different α_1 values when only one variable is changed. On the other hand, when more than one variable is changed, smaller α_1 can not identify the changed variables. Since the critical value is large for small α_1,i.e., conditional M statistic is smaller than the critical values, so that error rate of misdetection increases. Moreover, if we adjust α_1 value too

high, the critical value decreases. Thus, the error rate of false identification increases. To choose α_2, we need to consider the stopping criteria, $M_{\hat{\Gamma}\cup\beta_i} > h_i$ or $M_{\beta_i|\hat{\Gamma}} > h_1$. If α_2 value is too large, h_i critical value becomes too small, so that algorithm may stop at the beginning of the approach. On the other hand, if $\alpha_2 \geq \alpha_1$, the critical value h_i becomes small, so that α_2 can terminate the algorithm without the following steps. Based on this experiments, $\alpha_1 = \alpha_2$ yields the best performance in identifying faulty variables. In all of our experiments, we use $\alpha_1 = \alpha_2 = 0.1$. We also check the effect of α_1 and α_2 for different datasets and obtain the similar patterns as in multivariate normal case.

Table 3.11 Effect of α_1 and α_2 to the performance of the DFASD approach

		$\alpha_1 = 0.005$			$\alpha_1 = 0.05$			$\alpha_1 = 0.1$		
α_2		0.005	0.05	0.1	0.005	0.05	0.1	0.005	0.05	0.1
Direction	Size					$\rho = 0.75$				
{d 0 0}	1.0σ	0.2958	0.3162	0.2798	0.3699	0.3725	0.3650	0.3745	0.3835	0.3783
{d 0 0}	2.0σ	0.5085	0.5439	0.5511	0.6312	0.6336	0.6253	0.6379	0.6398	0.6374
{d 0 0}	3.0σ	0.6597	0.6413	0.6573	0.7919	0.7855	0.7906	0.7976	0.8014	0.8014
{d 0 d}	1.0σ	0.2527	0.2561	0.2671	0.3161	0.2992	0.3077	0.3150	0.3105	0.3082
{d 0 d}	2.0σ	0.4246	0.4365	0.4423	0.5659	0.5537	0.5588	0.5780	0.5446	0.5617
{d 0 d}	3.0σ	0.6901	0.6776	0.6445	0.8038	0.7980	0.7941	0.8035	0.8094	0.8066
						$\rho = 0.5$				
{d 0 0}	1.0σ	0.3578	0.3551	0.3636	0.4169	0.4179	0.4278	0.4197	0.4251	0.4315
{d 0 0}	2.0σ	0.5744	0.5696	0.5759	0.6589	0.6578	0.6618	0.6694	0.6746	0.6693
{d 0 0}	3.0σ	0.6669	0.6757	0.6971	0.7724	0.7740	0.7796	0.7818	0.7816	0.7854
{d 0 d}	1.0σ	0.3142	0.3090	0.3074	0.3649	0.3574	0.3639	0.3700	0.3799	0.3726
{d 0 d}	2.0σ	0.4985	0.5124	0.5089	0.6277	0.6213	0.6208	0.6319	0.6343	0.6300
{d 0 d}	3.0σ	0.6760	0.6893	0.6909	0.8032	0.7958	0.8113	0.8140	0.8081	0.8121
						$\rho = 0.25$				
{d 0 0}	$1.0\,\sigma$	0.3673	0.3715	0.3938	0.4041	0.4056	0.3977	0.4115	0.4021	0.4090
{d 0 0}	2.0σ	0.6028	0.6668	0.5577	0.6353	0.6268	0.6246	0.6374	0.6356	0.6367
{d 0 0}	3.0σ	0.6798	0.6836	0.6697	0.7456	0.7537	0.7465	0.7473	0.7512	0.8846
{d 0 d}	1.0σ	0.2980	0.3070	0.2968	0.3477	0.3542	0.3557	0.3535	0.3586	0.3580
{d 0 d}	2.0σ	0.5322	0.5211	0.5244	0.6074	0.6141	0.6018	0.6215	0.6195	0.6298
{d 0 d}	$3.0\,\sigma$	0.6825	0.7149	0.6898	0.7932	0.7958	0.8006	0.8064	0.8047	0.8055

Tables 3.12 and 3.13 show the correlation effect of the DFASD approach, K^2 and HNS under multivariate normal and lognormal distribution, respectively. The results from these extensive simulations demonstrate that the proposed approach is robust to correlation. These results show that, in most cases, correlation does not affect to the

performance of the proposed approach. However, K^2 and HNS depend highly on the correlation because the correlation among 'nearest k' data points strongly affects to the computation of the statistic.

Table 3.12 Correlation effect of DFASD, K^2 and HNS decomposition approaches under multivariate normal distribution when $p = 3$

Multivariate Normal							
Shift		DFASD		K^2		HNS	
Direction	Size	CR	EER	CR	EER	CR	EER
$\rho = 0.75$							
{d 0 0 }	1.0 σ	0.3783	0.3136	0.3003	0.2626	0.3602	0.2566
	2.0 σ	0.6374	0.1720	0.3393	0.2264	0.4215	0.2061
	3.0 σ	0.8014	0.0899	0.2818	0.2519	0.3754	0.2342
{d 0 d }	1.0 σ	0.3082	0.3454	0.0061	0.6162	0.1233	0.5103
	2.0 σ	0.5617	0.2254	0.0009	0.6226	0.1470	0.4642
	3.0 σ	0.8066	0.0815	0.0004	0.5535	0.1941	0.3362
$\rho = 0.50$							
{d 0 0 }	1.0 σ	0.3756	0.3361	0.3489	0.2818	0.3174	0.3108
	2.0 σ	0.7081	0.1373	0.5130	0.1732	0.4670	0.1990
	3.0 σ	0.7845	0.0910	0.5758	0.1448	0.5633	0.1575
{d 0 d }	1.0 σ	0.3617	0.3204	0.0326	0.5048	0.1995	0.4083
	2.0 σ	0.5918	0.1805	0.0537	0.4833	0.2625	0.3587
	3.0 σ	0.8157	0.0741	0.0791	0.4519	0.4165	0.2420
$\rho = 0.25$							
{d 0 0 }	1.0 σ	0.4024	0.2999	0.4383	0.2764	0.3580	0.3068
	2.0 σ	0.6308	0.1525	0.6684	0.1276	0.5609	0.1703
	3.0 σ	0.7558	0.0917	0.7435	0.0891	0.6465	0.1274
{d 0 d }	1.0 σ	0.3760	0.2869	0.1008	0.4060	0.2678	0.3436
	2.0 σ	0.5520	0.1938	0.1557	0.3703	0.3631	0.2863
	3.0 σ	0.7982	0.0751	0.3032	0.2796	0.6114	0.1537

Table 3.13 Correlation effect of DFASD, K^2 and HNS decomposition approaches under multivariate lognormal distribution when $p = 3$

Multivariate Lognormal							
Shift		DFASD		K^2		HNS	
Direction	Size	CR	EER	CR	EER	CR	EER
$\rho = 0.75$							
{d 0 0 }	1.0 σ	0.1930	0.4506	0.3502	0.2600	0.3374	0.2746
	2.0 σ	0.3820	0.2890	0.3086	0.2400	0.3148	0.2518
	3.0 σ	0.7465	0.1233	0.1795	0.3009	0.1911	0.3300
{d 0 d }	1.0 σ	0.1858	0.4935	0.0156	0.6368	0.0882	0.5738
	2.0 σ	0.3588	0.3791	0.0071	0.5949	0.0823	0.4949
	3.0 σ	0.7366	0.1335	0.0053	0.4793	0.0779	0.3495
$\rho = 0.50$							
{d 0 0 }	1.0 σ	0.3185	0.3925	0.3633	0.2907	0.2803	0.3251
	2.0 σ	0.4659	0.2937	0.4736	0.1984	0.3435	0.2562
	3.0 σ	0.6808	0.1463	0.4586	0.1876	0.3572	0.2419
{d 0 d }	1.0 σ	0.2655	0.3559	0.0666	0.4939	0.1906	0.4218
	2.0 σ	0.5102	0.2410	0.0986	0.4621	0.2659	0.3438
	3.0 σ	0.7672	0.1026	0.1253	0.3883	0.3035	0.2573
$\rho = 0.25$							
{d 0 0 }	1.0 σ	0.3217	0.3773	0.3972	0.2975	0.2713	0.3492
	2.0 σ	0.4881	0.2475	0.5775	0.1729	0.3930	0.2463
	3.0 σ	0.6181	0.1542	0.6768	0.1155	0.4471	0.2080
{d 0 d }	1.0 σ	0.2503	0.3534	0.1528	0.4061	0.2215	0.3643
	2.0 σ	0.5073	0.2071	0.3417	0.2819	0.4790	0.2179
	3.0 σ	0.7622	0.0921	0.5105	0.1931	0.5795	0.1494

To present the advantage of the proposed DFASD approach, we investigate its performance using a high-dimensional dataset. In high-dimensional cases, the CR would be possibly very small because it would be counted as ‘incorrect’ even if only one miss or false identification occurs. For this reason, we introduce two more measurements for

high-dimensional cases. These two measurements enable us to evaluate the performance in terms of both missed identification and false identification. They are ER1 and ER2, which are defined as;

$$\text{ER1} = \frac{\textit{the number of miss identified variables}}{\textit{the number of fault variables}}$$

and

$$\text{ER2} = \frac{\textit{the number of false identified variables}}{\textit{the number of unchanged variables}}$$

In low dimensional cases with reasonable CR, there is no need to consider EER, ER1 and ER2. However, as dimension increases, the CR might be no longer useful to see the performance difference. Thus, we use ER1 and ER2 when CR is not comparable due to its small value in high-dimensional cases.

Tables 3.14 and 3.15 show the performances of the DFASD and the K^2 decomposition approach under multivariate lognormal and gamma distributions, respectively. Similar to the previous results, when only one variable shifts, the performance of the K^2 decomposition approach outperforms the proposed DFASD approach in most of the cases. However, the performance reverses when more than one variable changes. In a more than one variable shift case, the K^2 decomposition approach mostly identifies unchanged

variables as changed variable, thus the ER2 of the K^2 decomposition approach is mostly greater than the ER2 of the DFASD approach. Therefore, the CR performance of the K^2 decomposition approach is significantly small. On the other hand, both approaches mostly identify the changed variables correctly, therefore their ER1 values are comparable.

Table 3.14 Performance of the DFASD and the K^2 decomposition approaches under multivariate lognormal distribution when $p = 10$

Shift				DFASD				K^2			
x_1	x_3	x_5	x_7	CR	EER	ER1	ER2	CR	EER	ER1	ER2
1.00	0.00	0.00	0.00	0.5690	0.1254	0.2290	0.1139	0.6780	0.0409	0.0010	0.0453
0.00	0.88	0.00	0.00	0.4340	0.1569	0.4100	0.1288	0.6470	0.0470	0.0050	0.0517
0.00	0.00	0.95	0.00	0.5530	0.1319	0.2330	0.1207	0.6510	0.0459	0.0010	0.0509
0.00	0.00	0.00	1.46	0.6350	0.1016	0.0000	0.1129	0.7830	0.0272	0.0000	0.0302
1.12	1.04	0.00	0.00	0.6440	0.1017	0.1195	0.0973	0.3190	0.0951	0.0020	0.1184
1.22	0.00	1.20	0.00	0.6800	0.0870	0.0205	0.1036	0.4600	0.0685	0.0050	0.0844
1.50	0.00	0.00	1.50	0.6810	0.0835	0.0000	0.1044	0.2520	0.1014	0.0000	0.1268
0.00	1.43	1.38	0.00	0.6460	0.0918	0.0000	0.1148	0.4240	0.0782	0.0020	0.0973
0.00	1.50	0.00	1.50	0.6830	0.0829	0.0000	0.1036	0.3780	0.0822	0.0015	0.1024
0.00	0.00	2.88	2.93	0.6460	0.0918	0.0000	0.1148	0.0260	0.2049	0.0000	0.2561

Table 3.15 Performance of the DFASD and the K^2 decomposition approaches under multivariate gamma distribution when $p = 10$

Shift				DFASD				K^2			
x_1	x_3	x_5	x_7	CR	EER	ER1	ER2	CR	EER	ER1	ER2
1.00	0.00	0.00	0.00	0.4410	0.1004	0.0000	0.1116	0.8630	0.0159	0.0000	0.0177
0.00	0.88	0.00	0.00	0.4830	0.0873	0.0000	0.0970	0.3590	0.0722	0.0000	0.0802
0.00	0.00	0.95	0.00	0.4260	0.1057	0.0000	0.1174	0.8980	0.0125	0.0000	0.0139
0.00	0.00	0.00	1.46	0.4430	0.0986	0.0000	0.1096	0.8340	0.0190	0.0000	0.0211
1.12	1.04	0.00	0.00	0.5140	0.0784	0.0000	0.0980	0.0070	0.1907	0.0000	0.2384
1.22	0.00	1.20	0.00	0.4480	0.0959	0.0000	0.1199	0.0710	0.1690	0.0000	0.2112
1.50	0.00	0.00	1.50	0.4480	0.0959	0.0959	0.1199	0.0040	0.2441	0.0000	0.3051
0.00	1.43	1.38	0.00	0.4330	0.1012	0.0000	0.1265	0.0210	0.1846	0.0000	0.2308
0.00	1.50	0.00	1.50	0.4350	0.1006	0.0000	0.1258	0.2650	0.1050	0.0000	0.1313
0.00	0.00	2.88	2.93	0.4330	0.0997	0.0000	0.1246	0.0000	0.4981	0.0000	0.6226

Tables 3.16, 3.17 and 3.18 illustrate the performances of the proposed DFASD approach and the K^2 decomposition approach under multivariate gamma distribution with $p = 30,50$ and 100, respectively. The CR values of the both approaches are mostly zero as the dimension increases, which is the expected result of the high dimensional observations. In Tables 3.16, 3.17 and 3.18, the changed variables are mostly identified correctly leading to the zero ER1 values. Moreover, the large ER2 values indicate that the K^2 decomposition approach deteriorates in the identification of the unchanged variables. Although the CR values are not comparable in high-dimensional cases, ER1 and ER2 demonstrate the superiority of the proposed identification approach.

Table 3.16 Performance of the DFASD and the K^2 decomposition approaches under multivariate gamma distribution when $p = 30$

Shift					DFASD			K^2		
x_1	x_2	x_3	x_4	x_5	CR	ER1	ER2	CR	ER1	ER2
0.45	0.65	0.00	0.00	0.00	0.0280	0.0000	0.1375	0.0140	0.0005	0.1462
0.55	0.00	0.46	0.00	0.00	0.0210	0.0000	0.1390	0.0050	0.0015	0.1516
0.88	0.00	0.00	1.06	0.00	0.0290	0.0000	0.1306	0.0050	0.0000	0.1586
1.21	1.04	0.00	0.00	0.00	0.0230	0.0000	0.1360	0.0006	0.0000	0.1565
1.15	0.00	0.00	1.46	0.00	0.0200	0.0000	0.1389	0.0000	0.0000	0.2132
1.98	0.00	0.00	0.00	2.24	0.0250	0.0000	0.1407	0.0000	0.0000	0.2268
1.15	0.80	1.46	0.00	0.00	0.0440	0.0000	0.1241	0.0000	0.0000	0.2215
1.25	0.00	1.36	0.00	1.28	0.0240	0.0000	0.1438	0.0000	0.0000	0.2468

Table 3.17 Performance of the DFASD and the K^2 decomposition approaches under multivariate gamma distribution when $p = 50$

Shift					DFASD			K^2		
x_1	x_2	x_3	x_4	x_5	CR	ER1	ER2	CR	ER1	ER2
0.25	0.00	0.68	0.00	0.00	0.0000	0.0865	0.1458	0.0000	0.2330	0.1638
0.75	0.75	0.00	0.00	0.00	0.0010	0.0000	0.1468	0.0010	0.0000	0.1499
1.25	0.00	0.00	1.15	0.00	0.0030	0.0000	0.1459	0.0000	0.0000	0.1620
1.95	0.00	0.00	2.28	0.00	0.0000	0.0000	0.1471	0.0000	0.0000	0.2287
0.45	0.55	0.00	0.00	0.65	0.0020	0.0000	0.1480	0.0000	0.0087	0.1771
0.75	0.00	0.65	0.95	0.00	0.0020	0.0000	0.1486	0.0000	0.0007	0.1717
1.20	0.95	0.48	0.00	0.00	0.0030	0.0000	0.1516	0.0000	0.0000	0.2205
1.18	0.00	1.25	0.00	1.33	0.0240	0.0000	0.1438	0.0000	0.0000	0.2489

Table 3.18 Performance of the DFASD and the K^2 decomposition approaches under multivariate gamma distribution when $p = 100$

Shift					DFASD			K^2		
x_1	x_2	x_3	x_4	x_5	CR	ER1	ER2	CR	ER1	ER2
0.28	0.60	0.00	0.00	0.00	0.0000	0.2530	0.1272	0.0000	0.3050	0.1334
0.50	0.74	0.00	0.00	0.00	0.0000	0.0000	0.1377	0.0000	0.0000	0.1505
0.74	0.95	0.00	0.00	0.00	0.0000	0.0000	0.1376	0.0000	0.0000	0.1341
0.89	0.00	0.85	0.00	0.00	0.0000	0.0000	0.1381	0.0000	0.0000	0.1443
1.25	0.00	1.35	0.00	0.00	0.0000	0.0000	0.1364	0.0000	0.0000	0.1627
0.00	1.96	0.00	2.45	0.00	0.0000	0.0000	0.1370	0.0000	0.0000	0.1695
1.12	2.05	1.66	0.00	0.00	0.0000	0.0000	0.1340	0.0000	0.0000	0.1812
1.18	0.00	1.35	0.00	1.26	0.0000	0.0000	0.1382	0.0000	0.0000	0.1732

3.5 Conclusions

Fault diagnosis in statistical process control is a challenging issue for the processes in a high-dimensional space. Although many fault isolation approaches have been introduced, mostly based on T^2, they assume the underlying distribution of the process follow a multivariate normal distribution. Even though there have been several distribution free approaches, they are not effective when the numbers of changed variables are large. In this chapter, we propose a distribution-free fault variable identification approach by combining an SVDD-based test statistic with an ASD approach. The proposed DFASD approach identifies unchanged variables step by step under no significant evidence of a change and eventually obtains the changed variables by considering at most $\frac{p\times(p+1)}{2}$ decompositions. The proposed approach is superior to the existing approaches when the one variable is shifted with a large shift. In addition, the simulation experiments using different distribution datasets demonstrate that the proposed DFASD approach

outperforms existing distribution free approaches such as the K^2 decomposition and the HNS decomposition approach when the number of shifted variable is more than one. We also introduce two performance measures for high-dimensional data: ER1 and ER2 which demonstrate the superiority of the proposed DFASD approach. Moreover, the proposed approach is not sensitive to the correlation between variables, leading to a stable performance regardless of the number of changed variables.

As a future research, it is interesting to analyze the effects of different kernel types with different parameters. Another future research is the extension of the proposed approach to distribution-free variable selection-based control charts.

CHAPTER 4

BAYESIAN FRAMEWORK FOR FAULT VARIABLE IDENTIFICATION

4.1 Introduction

Multivariate statistical process control (MSPC) charts are widely used to monitor product quality and detect process changes in multi-dimensional processes. MSPC charts such as Hotelling's chi-square chart, multivariate exponentially weighted moving average (MEWMA) (Lowry *et al.*, 1992) and multivariate CUSUM (Crosier, 1988) are used to monitor multiple variables simultaneously by taking into account the correlation among variables and are designed only to detect process shifts but not to provide information about the cause of the shift (Jiang and Tsui, 2008, Wang and Jiang, 2009). Once MSPC charts detect a process abnormality, identifying faulty variables is of significant interest for engineers as it often provides important diagnostic information and enables taking corrective actions. However, identification of faulty variables is as a challenging issue (Kim *et al.*, 2016b, Li *et al.*, 2008, Zou *et al.*, 2011).

Approaches for identifying faulty variables have been investigated by many researchers. For example, Doganaksoy *et al.* (1991) propose an individual test statistic for each variable to identify the faulty variables ignoring the effect of the correlation among variables. Hawkins (1991, 1993) proposes another approach that considers the correlations among variables based on regression-adjusted variables. This approach is only effective when one variable is changed. However, when several variables are shifted

simultaneously, or when even a single variable that is strongly correlated with other variables is shifted, the identification performance may significantly decrease (Das and Prakash, 2008, Kim *et al.*, 2016b).

Runger (1996) expands the regression adjustment approach by investigating the contribution of a subset of variables to the mean shift. Based on the prior knowledge of a subset with unchanged variables, relative contributions of the complement of this subset can be calculated. However, in practice, the prior knowledge of a set of unchanged variables may not be available. In addition, Runger *et al.* (1996) propose a polynomial time algorithm for the calculation of conditional T^2 statistic of each variable to investigate the contribution of the variable causing the mean shift. Based on the regression adjustment procedure (or conditional T^2), Mason *et al.* (1995, 1997) propose a T^2 decomposition procedure known as the Mason–Tracy–Young (MTY). It decomposes the T^2 statistic into the combinatorial number of conditional statistic to identify a changed variable. Although the T^2 decomposition-based approaches theoretically work well, but they are impractical for a large number of variables because such approaches consider $p!$ decompositions (where p is the number of process parameters). Sullivan *et al.* (2007) propose an algorithm which considers ${}_pC_k$ number of subsets, where k is the size of the subset of the variables, and calculates the T^2 statistic for all of subsets for p-variate observation. However, the computational complexity, for large p, remains an obstacle for practical implementations even though this approach reduces the computational burden compared to MTY.

When a process with multiple variables is operating under abnormal conditions, only one or a small subset of variables would be possibly responsible for the process shift; this is known as the sparsity property (Jiang *et al.*, 2012, Wang and Jiang, 2009, Zou and Qiu, 2009). Based on the sparsity assumption, Kim *et al.* (2016b) propose an adaptive step-down (ASD) procedure for diagnosis of faulty variables by identifying unchanged variables one by one in each iteration. It has an advantage in terms of computational intensity over MTY.

The fault identification procedures mentioned above are based on the assumption that the underlying distribution of the process follows the multivariate normal distribution. In many real life applications, however, the underlying process distribution is usually unknown. There is limited research that addresses the identification procedure for faulty variables when the underlying probability distribution of a process is unknown or follows a general distribution. Kim *et al.* (2011) propose a distribution-free fault identification procedure based on the k-nearest neighbor data description. Tuerhong and Kim (2011) propose another distribution-free fault identification procedure based on a hybrid novelty score (HNS) obtained by using k-nearest neighbors to calculate a test statistic.

However, existing distribution-free fault identification procedures have the following limitations. First, the k-nearest neighbor data description procedure is suitable only for spherical data distribution, which is not common in real-life applications (Yu *et al.*, 2002). Therefore, when the data deviate from the spherical distribution, the performance of the faulty variable identification procedures based on the k-nearest neighbor data

description may decrease. Second, in calculating the contribution of each variable to the process change, the test statistic of those approaches depend on the regression adjustment-based procedure, which may perform significantly poorly when multiple shifts cause abnormality of the process, although it has a good identification performance when only one variable shifts. Third, existing approaches control the identification procedure with a given dataset in a deterministic way, which may provide limited information of the result.

Therefore, in this chapter, we propose a novel distribution-free fault identification procedure based on a Bayesian framework when a few variables are responsible for the process shift. A new test statistic based on Bayesian support vector data description (BSVDD), which is a kernel based one class classification procedure, is proposed and the changed variables are identified based on an efficient algorithm with significant computational advantage. The proposed procedure would not be limited to the spherical distribution by mapping the data into high-dimensional kernel space with an appropriate kernel function. Moreover, the proposed BSVDD considers a probabilistic behavior of the parameters by taking ‘prior knowledge’ into account in the Bayesian framework. In this chapter, we improve the capability of the procedure based on the interpretation of the parameters with the prior distributions. Accordingly, we propose a local density degree function to select the parameters, which is more interpretable and offers better performance in the identification of faulty variables than the existing procedures.

This chapter is organized as follows. After the review of SVDD in Section 4.2, we propose the distribution-free BSVDD procedure based on a Bayesian statistic in Section 4.3. In Section 4.4, simulation studies and results are demonstrated. In Section 4.5, we apply the proposed procedure in a real-life case study of bolts' dimensions monitoring, followed by the conclusion in Section 4.6.

4.2 Support Vector Data Description (SVDD)

There are several procedures to describe the data. SVDD is an effective procedure for describing irregularly patterned data (Ning and Tsung, 2013). For a given data set $\mathbf{D} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{R}^p, i = 1, ..., m\}$, with respect to the *p*-dimensional vector $\mathbf{x}$, SVDD finds a hypersphere which covers the data with minimal volume, with a center $\mathbf{a}$ and a radius R (Tax and Duin, 1999). To allow the misclassification in $\mathbf{D}$, ε_i is introduced to penalize outliers for large distances between $\mathbf{x}_i$ and $\mathbf{a}$. The primal formulation of this problem is constructed as follows:

$$\begin{aligned} &\min \quad R^2 + C\sum_{i=1}^{m} \varepsilon_i \\ &s.t. \quad \left\|\mathbf{x}_i - \mathbf{a}\right\|^2 \le R^2 + \varepsilon_i, \quad \varepsilon_i \ge 0 \ \ \forall i \in \{1, ..., m\} \end{aligned}$$

where C is the regularization parameter which controls the volume of the hypersphere by adjusting the number of observations that are located outside the boundary. The Dual formulation is obtained by using the Lagrangian function:

$$L\left(R,\mathbf{a},\varepsilon_i\right)=R^2+C\sum_{i=1}^{m}\varepsilon_i-\sum_{i=1}^{m}\tau_i\left(R^2+\varepsilon_i-\left(\mathbf{x}_i-\mathbf{a}\right)^T\left(\mathbf{x}_i-\mathbf{a}\right)\right)-\sum_{i=1}^{m}\gamma_i\varepsilon_i \tag{4.1}$$

where $\tau_i \geq 0$ and $\gamma_i \geq 0$ are Lagrangian dual variables. Taking partial derivatives of Lagrangian function, we obtain

$$\frac{\partial L}{\partial R}=0\rightarrow\sum_{i=1}^{m}\tau_i=1,\ \frac{\partial L}{\partial \mathbf{a}}=0\rightarrow\mathbf{a}=\frac{\sum_{i=1}^{m}\tau_i\mathbf{x}_i}{\sum_{i=1}^{m}\tau_i}=\sum_{i=1}^{m}\tau_i\mathbf{x}_i,\ \frac{\partial L}{\partial \varepsilon_i}=0\rightarrow C-\tau_i-\gamma_i=0 \tag{4.2}$$

In this case, the dual formulation in Eq. (4.3) is constructed by substituting Eq. (4.2) into Eq. (4.1).

$$\begin{aligned}\max\quad & \sum_{i=1}^{m}\tau_i\left(\mathbf{x}_i\cdot\mathbf{x}_i\right)-\sum_{i=1}^{m}\sum_{j=1}^{m}\tau_i\tau_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)\\ s.t.\quad & \sum_{i=1}^{m}\tau_i=1,\ 0\leq\tau_i\leq C\quad\forall i\in\{1,\ldots,m\}\end{aligned} \tag{4.3}$$

Data points corresponding to the positive τ_i are called support vectors and they are placed on or outside the boundary. Based on the SVDD boundary with a given C from the in-control data, a new observation $\mathbf{z}$ can be classified as "in" or "out" of the data boundary by checking the distance from $\mathbf{z}$ to the center of the hypersphere. The inner product in Eq. (4.3) can be replaced by a kernel function $K(\mathbf{x}_i,\mathbf{x}_i)$, defining

$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, and thus a more suitable nonlinear boundary to cover the data can be obtained (Tax and Duin, 1999).

4.3 Proposed Methodology

4.3.1 Bayesian SVDD with Non-Normal Prior for Fault Identification

Although the traditional SVDD assumes that the parameters are fixed constants to be determined, the resultant center **a** may be a random vector with inherent randomness based upon a given training dataset. With an underlying solution from SVDD in Eq. (4.2), the mean **a** is determined as a weighted average of the training data. Ghasemi *et al.* (2016) introduce a Bayesian approach in SVDD by assuming that a transformation through $\phi(\cdot)$ maps the data into higher dimensional space in which the transformed data follow a Gaussian distribution with mean $\mathbf{a} = \sum_i \tau_i \phi(\mathbf{x}_i)$ and an identity covariance. Since the distance of a point to the center of a hypersphere is inversely proportional to the likelihood in the weighted Gaussian model, SVDD is a special case of the weighted Gaussian model, which improves SVDD by utilizing precise prior knowledge. Thus, the unknown parameter τ_i can be estimated through a Bayesian approach with a proper prior distribution for $\boldsymbol{\tau}$ (Ghasemi *et al.*, 2016).

However, their approach assumes that prior distribution $p(\tau_i)$ is normally distributed, which is inappropriate because τ_i is defined in $0 \le \tau_i \le C$ to keep convexity in the SVDD, and the constraint $\sum_{i=1}^{m} \tau_i = 1$ must be satisfied. Therefore, in this chapter, we propose an

improved Bayesian SVDD which is more realistic with a proper prior distribution of the dual variable τ_i expressed as a truncated exponential as follows:

$$p(\tau_i \mid C) = \frac{\theta_i e^{-\theta_i \tau_i}}{1 - e^{-C\theta_i}}, \quad 0 \le \tau_i \le C.$$

Similar to Ghasemi *et al.* (2016), it is assumed that training data mapped into a higher dimensional kernel space follow a Gaussian distribution, i.e., $\phi(\mathbf{x}_j) \sim N\left(\sum_i \tau_i \phi(\mathbf{x}_i), \sigma^2 \mathbf{I}\right)$. Then the likelihood probability given parameter $\boldsymbol{\tau}$ becomes

$$p(\mathbf{D} \mid \boldsymbol{\tau}) = \prod_{i=1}^{m} \frac{1}{\left(2\pi\right)^{\hat{p}/2} \sigma^{\hat{p}}} e^{-\frac{1}{2\sigma^2}\left\| \phi(\mathbf{x}_i) - \sum_{j=1}^{m} \tau_j \phi(\mathbf{x}_j) \right\|_2^2}.$$

Maximizing a posterior (MAP) is derived by the typical Bayesian rule as

$$p(\boldsymbol{\tau} \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau})}{p(\mathbf{D})},$$

where $\mathbf{D}$ is a set of training data. Since $p(\mathbf{D})$ is a normalizing constant independent of $\boldsymbol{\tau}$, it can be ignored so that

$$p(\boldsymbol{\tau} \mid \mathbf{D}) \propto p(\mathbf{D} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau}). \tag{4.4}$$

The solution of MAP is given by

$$\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\tau}} p(\boldsymbol{\tau} \mid \mathbf{D}). \tag{4.5}$$

By taking a logarithm and using the relationship in Eq. (4.4), then Eq. (4.5) is equivalent to

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} 2^{-1} \times \left(-2\sigma^{-2}\boldsymbol{\tau}^T \mathbf{B}\mathbf{1} + m\sigma^{-2}\boldsymbol{\tau}^T \mathbf{K}\boldsymbol{\tau} + 2\sum_{i=1}^{m} \theta_i \tau_i \right)$$

where $\mathbf{B}$ is a diagonal matrix and $\mathbf{B}_{i,i} = \sum_j \mathbf{K}_{i,j}$, $\mathbf{1}$ is an $m \times 1$ vector with all ones, and $\mathbf{K}$ is the kernel matrix in which the $(i, j)^{th}$ element of the matrix is defined as $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

The scale parameter σ^2 is proportionally associated with the radius R in the feature space, and the radius is inversely proportional to C. Since C is inversely proportional to the number of training data m (Tax and Duin, 2004), the relationship $p(\mathbf{D} \mid \boldsymbol{\tau}, \sigma^2) \propto p(\mathbf{D} \mid \boldsymbol{\tau}, m)$ holds for the given data set. Then, Eq. (4.4) can be rewritten by replacing the probability $p(\mathbf{D} \mid \boldsymbol{\tau}, m)$, and the solution $\boldsymbol{\tau}$ is obtained by the following optimization problem:

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} 2^{-1} \times \left(-2m^{-1}\boldsymbol{\tau}^T \mathbf{B1} + \boldsymbol{\tau}^T \mathbf{K}\boldsymbol{\tau} + 2\sum_{i=1}^{m} \theta_i \tau_i \right). \tag{4.6}$$

Note that the optimization problem must satisfy the same constraints as those in SVDD with an association between SVDD and Bayesian SVDD.

$$s.t. \quad \sum_{i=1}^{m} \tau_i = 1, \; 0 \le \tau_i \le C \quad \forall i \in \{1, ..., m\}$$

From the structure of the objective function in Eq. (4.6), we conjecture that the determination of parameters would play a critical role in optimization (see Appendix A. for detailed derivations).

4.3.1.1 Determination of the Prior Parameters

In order to identify the proper values of θ_i, its interpretation must be carried out in advance. The optimization problem in Eq. (4.6) is reduced to the original SVDD when the parameter θ_i is defined as follows:

$$\theta_i = \frac{1}{m}\sum_{j=1}^{m} \mathbf{K}_{ij} - \frac{1}{2}\mathbf{K}_{ii} \tag{4.7}$$

Note that if the data point is inside the boundary, the corresponding τ_i is zero, otherwise, corresponding τ_i ($\leq C$) is nonzero. Since θ_i is inversely proportional to the mean of the truncated exponential distribution, it should have a small value for data points far from the center $\mathbf{a}$ in the feature space and vice versa.

In Eq. (4.7), the first term represents the average distance from i^{th} data point to all data points including its own distance to the center. Thus, it can be seen as a relative distance to the other data points. Moreover, the second term $\mathbf{K}_{ii}$ in Eq. (4.7) represents an absolute distance to the center. Therefore, Eq. (4.7) can be interpreted in a way that a small value would be assigned to the data points located around the edge of the data set because their relative and absolute distances are both large resulting in a small θ_i. On the other hand, the points in a dense area which are close to the center will have a large relative distance but a small absolute value resulting in a large θ_i.

In addition, we propose a new procedure to determine a prior parameter θ_i based on the interpretation above. Now we define a local density degree of an observation in a given data set. Suppose that

$$\nu_i = \exp\left\{-\kappa \times \frac{\left\|\mathbf{x}_i - \mathbf{x}_i^{\,k}\right\|}{\min_{j \in D}\left\|\mathbf{x}_j - \mathbf{x}_j^{\,k}\right\|}\right\}, \quad i \in \{1,\ldots,m\} \tag{4.8}$$

where $\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|$ denotes the distance between i^{th} and its k^{th} neighbor point, and κ is a parameter that controls the weight of density. It is straightforward that ν_i is inversely proportional to the density, i.e., a point which is located in a dense area and has a small ratio of $\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\| / \min_{j \in D} \left\|\mathbf{x}_j - \mathbf{x}_j^k\right\|$ leading to the large value of ν_i, and vice versa. Thus, it holds the association $\theta_i \propto \nu_i$. Throughout the chapter, we assume $\theta_i = \nu_i$. In addition, we determine κ as discussed in Section 4.

4.3.2 A Framework for Identifying Faulty Variables

4.3.2.1 Construction of the Conditional Statistic and Diagnosis Procedure

In many parametric and nonparametric fault identification procedures, the marginal effect of each variable or the set of the variables is calculated using decomposition as

$$M_{\Psi_1|\Psi_0} = M_{\Psi_0 \cup \Psi_1} - M_{\Psi_0}, \tag{4.9}$$

where Ψ_0 and Ψ_1 are the complementary sets in $\Psi \equiv \{\Psi_0 \cup \Psi_1\} \subset S$, and S is a set consisting of all variables. In the general framework of decomposition, measurements of distance can be well incorporated into Eq. (4.9). For example, when $M = T^2$, the statistic becomes the decomposition of T^2 . The k -nearest neighbor and other distance measurements such as HNS are also replaceable with M (Kim *et al.*, 2011, Tuerhong and Kim, 2011).

In a higher dimensional space, the data point $\mathbf{x}_i$ is transformed to $\phi(\mathbf{x}_i)$. Since the transformed data can be viewed as surrounding the center $\mathbf{a} = \sum_i \tau_i \phi(\mathbf{x}_i)$ in the feature space, the distance to the center in the feature space can be measured as $\|\phi(\mathbf{z}) - \mathbf{a}\|_2^2$. In this chapter, we determine our measurement of M as

$$\begin{aligned} M &= \|\phi(\mathbf{z}) - \mathbf{a}\|_2^2 \\ &= K(\mathbf{z},\mathbf{z}) - 2\sum_{i\in S} \tau_i K(\mathbf{x}_i,\mathbf{z}) + \sum_{i,j\in S} \tau_i \tau_j K(\mathbf{x}_i,\mathbf{x}_j) \end{aligned}$$

Then, for an arbitrary set A, the statistic M_A is calculated as

$$M_A = K(\mathbf{z}_A,\mathbf{z}_A) - 2\sum_{i\in S} \tau_i K(\mathbf{x}_{i,A},\mathbf{z}_A) + \sum_{i,j\in S} \tau_i \tau_j K(\mathbf{x}_{i,A},\mathbf{x}_{j,A})$$

where $\mathbf{z}_A$ is an $|A|\times 1$ dimensional vector, and $|A|$ is the cardinality of the set $A(\subset S)$. Thus, the conditional statistic $M_{\Psi_1|\Psi_0}$ in Eq. (4.9) can be viewed as a contribution of a set of variables in Ψ_1 over the variables in Ψ. For example, $M_{j|\Gamma}$, where Γ contains all variables without j^{th} variable, can be interpreted as a contribution of j^{th} variable over all other variables.

Once the effect of the variables is quantified, we identify unsuspicious variables (variables that are unlikely cause process changes) rather than suspicious ones since it would be reasonably easier to identify unchanged variables under the sparsity assumption

than to identify the changed variables directly. Moreover, several traditional identification procedures suffer from the computational issue with the combinatorial number of steps in the algorithms. To overcome this issue, we adopt an ASD procedure recently introduced by Kim *et al.* (2016b). Even though ASD is developed for parametric conditions, we modify the conditional statistic as in Eq. (4.9) and investigate the ASD procedure under nonparametric conditions. The algorithm conducts sequential tests with predetermined threshold values from the in-control dataset as follows:

Algorithm: BSVDD procedure for fault identification

***Initialize*:** $\beta_1 = \arg\min_{j \in S} M_{j|1,\ldots,j-1,j+1,\ldots,p}$; $i = 1$;

If $\beta_1 > h_1$

Faulty Variables $= \{1, \ldots, p\}$

Else

Do

$\hat{\Gamma} := \hat{\Gamma} \cup \{\beta_i\}$;

$i := i + 1$;

$\beta_i = \arg\min_{j \notin \hat{\Gamma}} M_{j|\hat{\Gamma}}$

While $M_{\beta_i|\hat{\Gamma}} < h_1$ and $M_{\hat{\Gamma} \cup \{\beta_i\}} < h_i$

Faulty Variables = Find $\{ k \in S - \hat{\Gamma} : M_{k|\hat{\Gamma}} > h_1 \}$

End

The algorithm seeks the least contributor at inception and the most contributors at the end, one by one in each iteration. Thus, the variables not classified into the set $\hat{\Gamma}$ after the end of the algorithm are considered faulty variables. This approach significantly reduces the computational complexity specially in high dimensional processes. Indeed, the

computational complexity of the proposed procedure is at most $O\left(p^2\right)$ which is quite efficient when compared to the existing decomposition procedures.

4.3.2.2 Determination of the Thresholds

Throughout the steps of diagnosis, the values of the thresholds, h_1 and h_i play an important role since they control the behavior of the solution path. Since the test statistic is obtained in a nonparametric manner according to the kernel type and the original data distribution, determining the exact distributions of the test statistic for the determination of threshold values is much more challenging. In this chapter, we determine the values of the thresholds by applying a bootstrap resampling procedure, which has been widely used to provide an accurate inference of unknown distribution (Efron and Tibshirani, 1994). Assume that there are *m* numbers of *p*-variate in-control observations. The following procedure is proposed for the determination of general threshold h_1 that to classifies a variable into the set of unchanged variables $\hat{\Gamma}$.

- Step 1: Calculate all the possible conditional $M_{\beta_i|\hat{\Gamma}}$ statistics. There are $n_1 = m \times \sum_{i=2}^{p} \binom{p}{i}$ conditional statistic. Denote $M_{\beta_i|\hat{\Gamma}} = M_C$ for simplicity.
- Step 2: $M_{c,r,1}, M_{c,r,2}, \ldots, M_{c,r,n_1}$ are n_1-conditional statistic from the r^{th} bootstrap sample $\left(r = 1, \ldots, T\right)$.

- Step 3: Sort the conditional statistic of each bootstrap sample in ascending order, i.e.,

$$M_{c,r,(1)} < M_{c,r,(2)} < \ldots < M_{c,r,(n_1)} \ .$$

- Step 4: Find the s^{th} percentile value where $s = n_1 \times (1-\alpha_1)$ and α_1 is the significance level for faulty variables.
- Step 5: The threshold value is calculated by taking the average of s^{th} statistic:

$$h_1 = \frac{1}{T}\sum_{r=1}^{T} M_{c,r,(s)}$$

In addition, to prevent adding a changed variable into the set of unchanged variables $\hat{\Gamma}$, the condition $M_{\hat{\Gamma}\cup\{\beta_i\}} < h_i$, where i is the cardinality of the set $\hat{\Gamma}\cup\{\beta_i\}$, must be satisfied. The threshold value h_i can be obtained by using a bootstrap resampling procedure similar to the calculation of h_1. Instead of calculating all the possible conditional $M_{\beta_i|\hat{\Gamma}}$ statistic in Step 1, all possible $M_{\hat{\Gamma}\cup\beta_i}$ statistic are calculated with each significance level α_i. Since the manner of all calculations after the second iteration is identical, we consider that $\alpha_i = \alpha_2$ for $i \geq 2$.

4.4 Performance Assessment

In this section, we demonstrate various experiments to assess the performance of the proposed BSVDD-based fault identification procedure and compare it with the existing fault identification procedures such as T^2 decomposition with ASD procedure (Kim *et*

al., 2016b), K^2 and the HNS decomposition (Kim *et al.*, 2011, Tuerhong and Kim, 2011). These decomposition techniques can be interpreted in the same framework given in Eq. (4.9) for calculating the marginal effect of the subset of variables. For the T^2 decomposition with ASD procedure, we follow the guidelines defined in Kim *et al.* (2016b) to determine the threshold parameters for K^2 as an average distance to the *k*-neighbors, and the number of neighbors is used as suggested in Kim *et al.* (2011). Throughout this chapter, we represent T^2 decomposition with ASD, K^2 and HNS decomposition as simply T^2, K^2 and HNS, respectively.

To determine the value of parameter C, we adopt the suggestion by Tax and Duin (2004) using the following $C \cong (m \cdot f)^{-1}$, where f is a fraction of outliers. Because the threshold value h_1 can be determined by this fraction, f is equal to α_1. Threshold values are obtained from 200 in-control data and 10,000 out-of-control observations are taken and averaged for the performance calculation.

For performance comparisons, we use correctness ratio which is defined as follows:

$$\mathrm{CR} = \frac{\sum_{i=1}^{n} I(\Gamma \equiv \hat{\Gamma})}{n}$$

where *n* is the number of identifications and $I(\cdot)$ is the indicator function which equals 1 when correctly identified and equals zero otherwise. However, CR may not be

appropriate for every case because the indicator function only considers the case when $\Gamma \equiv \hat{\Gamma}$. Thus, it would be zero even if 99% of the shifted variables are correctly identified in one out-of-control observation. To supplement this strictness, we introduce another measurement as the expected error rates (EER) in the mean shift (Zou *et al.*, 2011) as

$$\text{EER} = E\left(\frac{\text{Number of errors}}{\text{Number of variables}} \right).$$

In EER, we check all the variables one by one in the observation vector to determine whether they are correctly identified. Therefore, by using both performance measures together, we can properly assess each procedure in identifying the shifted variables.

4.4.1 Performance Comparison under Conventional Non-Normal Dataset

In literature on distribution-free fault identification, several well-known distributions are tested to evaluate proposed methodologies in non-normal data environment such as multivariate gamma, multivariate lognormal, and some irregular patterned data. Gamma distributed data are obtained by utilizing the procedure explained in Stoumbos and Sullivan (2002) by choosing the shape and scale parameters as one. On the other hand, lognormal observation $\mathbf{x}$ is generated according to the normal variable $\mathbf{y}$ and $\mathbf{x} = \exp(\mathbf{y})$. The mean shift is obtained as $\boldsymbol{\mu}_1 = \boldsymbol{\mu} + \boldsymbol{\delta}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)$, where δ_i represents the shift size of variable *i*. In the comparison tables, for instance, $\{s, 0, 0, \ldots\}$ indicates that mean shift occurs at the first variable and the size of the additive shift is δ_1.

Tables 4.1 and 4.2 show the performance of BSVDD against other fault identification approaches under the multivariate gamma and lognormal distributions with five dimensional datasets, respectively.

Table 4.1 Performance comparisons of BSVDD, SVDD, T^2 and K^2 procedures under multivariate gamma distribution

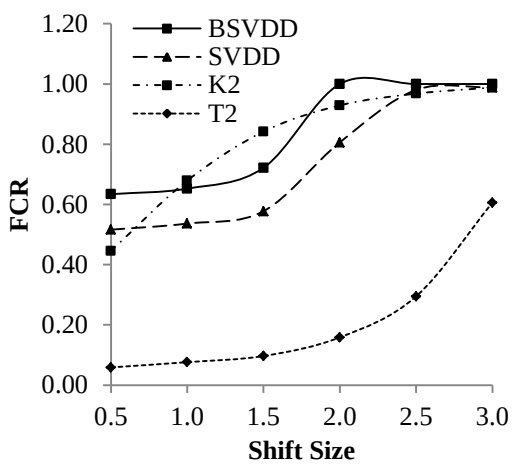
Shift		T^2		K^2		SVDD		BSVDD	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER
{s 0 0 0 0}	0.50	0.0501	0.2031	0.3561	0.2244	0.1225	0.3760	0.1694	0.3568
	1.00	0.0675	0.1931	0.5822	0.1307	0.1790	0.3046	0.2688	0.2548
	1.50	0.0890	0.1855	0.7461	0.0756	0.2532	0.2389	0.4131	0.1804
	2.00	0.1486	0.1722	0.8169	0.0514	0.4598	0.1788	0.7091	0.1082
	2.50	0.2829	0.1444	0.8415	0.0411	0.6555	0.1294	0.7195	0.0996
	3.00	0.5926	0.0822	0.8486	0.0363	0.6750	0.1219	0.7192	0.1025
{s 0 s 0 0}	0.50	0.0056	0.3835	0.0326	0.3543	0.0563	0.4170	0.0684	0.3848
	1.00	0.0116	0.3735	0.0598	0.3014	0.1146	0.3770	0.1530	0.2938
	1.50	0.0184	0.3692	0.1240	0.2512	0.2000	0.3101	0.2950	0.2159
	2.00	0.0357	0.3588	0.3176	0.1814	0.4196	0.1929	0.6994	0.0901
	2.50	0.0903	0.3285	0.5064	0.1271	0.6709	0.0958	0.7667	0.0720
	3.00	0.3065	0.2302	0.4799	0.1232	0.7157	0.0830	0.7639	0.0723
{s 0 s 0 s}	0.50	0.0006	0.5755	0.0004	0.5093	0.0592	0.4489	0.0680	0.3768
	1.00	0.0013	0.5694	0.0004	0.4923	0.1210	0.3914	0.1513	0.2862
	1.50	0.0008	0.5733	0.0001	0.5050	0.1895	0.3602	0.2573	0.2328
	2.00	0.0034	0.5667	0.0005	0.5275	0.3318	0.2549	0.6750	0.0779
	2.50	0.0103	0.5415	0.0005	0.5152	0.474	0.1663	0.8121	0.0474
	3.00	0.0214	0.5100	0.0019	0.4961	0.5008	0.1602	0.8084	0.0482

Table 4.2 Performance comparisons of BSVDD, SVDD, T^2 and K^2 procedures under multivariate lognormal distribution

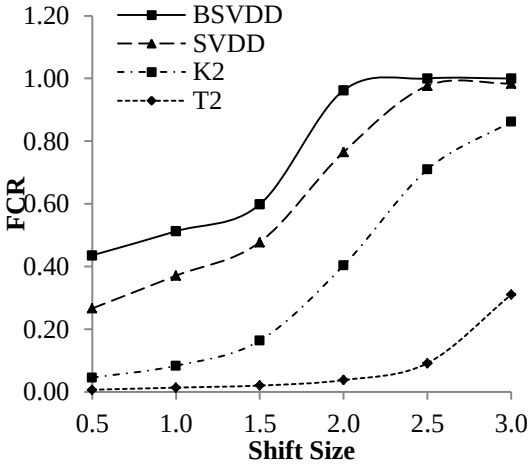
Shift		T^2		K^2		SVDD		BSVDD	
Direction	Size	CR	EER	CR	EER	CR	EER	CR	EER
{s 0 0 0 0}	0.50	0.0129	0.2120	0.1852	0.3358	0.0608	0.3736	0.0846	0.4199
	1.00	0.0119	0.2050	0.5050	0.1872	0.0792	0.3043	0.2477	0.2816
	1.50	0.0110	0.2004	0.7197	0.0952	0.0982	0.2736	0.5849	0.1408
	2.00	0.0156	0.1987	0.7696	0.0696	0.1917	0.2529	0.6399	0.1202
	2.50	0.0322	0.1948	0.7523	0.0660	0.4127	0.2001	0.6415	0.1172
	3.00	0.0902	0.1835	0.7085	0.0707	0.5895	0.1588	0.6375	0.1136
{s 0 s 0 0}	0.50	0.0003	0.4027	0.0256	0.4105	0.0265	0.4550	0.0750	0.3646
	1.00	0.0002	0.4008	0.0046	0.3714	0.0418	0.5161	0.2433	0.2578
	1.50	0.0006	0.3992	0.0091	0.3331	0.1081	0.3678	0.6744	0.0966
	2.00	0.0013	0.3975	0.0483	0.2789	0.3106	0.2464	0.7223	0.0749
	2.50	0.0029	0.3951	0.2439	0.2223	0.5057	0.1667	0.7432	0.0696
	3.00	0.0102	0.3911	0.3246	0.1926	0.6001	0.1347	0.7636	0.0644
{s 0 s 0 s}	0.50	0.0000	0.5913	0.0073	0.5243	0.0473	0.5022	0.1450	0.3461
	1.00	0.0001	0.5943	0.0032	0.5481	0.0802	0.5238	0.3953	0.2264
	1.50	0.0000	0.5959	0.0009	0.5891	0.1274	0.5453	0.7626	0.0832
	2.00	0.0004	0.5940	0.0009	0.6162	0.2485	0.4140	0.8173	0.0505
	2.50	0.0001	0.5922	0.0002	0.6285	0.3855	0.3097	0.8188	0.0502
	3.00	0.0006	0.5913	0.0003	0.6053	0.4724	0.2731	0.8057	0.0520

The proposed BSVDD outperforms other procedures because BSVDD utilizes the density-based prior knowledge. In addition, the performance of T^2 is poor because the two data sets are far from being normal. Interestingly, K^2 underperforms others when more than one variable is shifted because it is based on the regression adjustment procedure, which shows poorer performances as the number of faulty variables increases. Since the indicator function in CR always assumes the value 1 when all changed and unchanged variables are correctly identified, it cannot provide sufficient information to

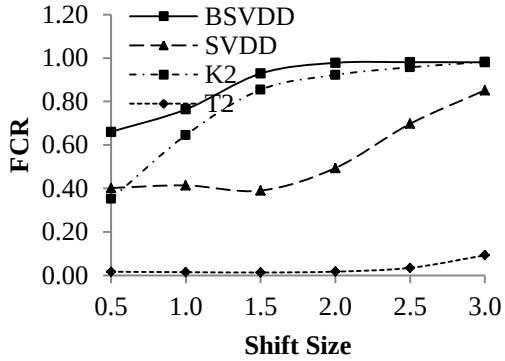
determine the exact number of faulty variables correctly, specially as the number of variables increases. Therefore, we introduce another performance measure, named the fault correctness ratio (FCR), which focuses only on the changed variables. FCR performance measure determines whether the changed variables are correctly identified. As shown, CR considers both changed and unchanged variables. However, if one is interested only in the possibly changed variables, CR may not be sufficient to result in a strong conclusion about the changed variables. In addition, EER considers both changed and unchanged variables; it may not provide broad information to decide whether changed variables are identified correctly, either. However, FCR is more appropriate to make conclusions about only the changed variables. Figure 4.1 illustrates the FCR performance.



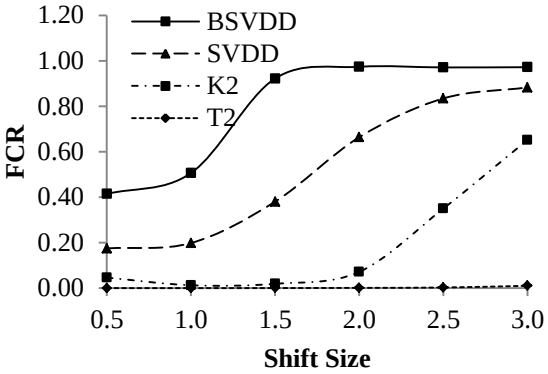
(a) Gamma distribution: one variable shift



(b) Gamma distribution: two variable shift



(c) Lognormal distribution: one variable shift



(d) Lognormal distribution: two variable shift

Figure 4.1 FCR performances for multivariate gamma and lognormal distribution

Figure 4.1 shows that the performance measures of BSVDD and K^2 are comparable in many scenarios because K^2 identifies several changes correctly but not exactly, while FCR of T^2 still remains almost zero. It also demonstrates that the distribution-free approaches perform better when normality assumption is ignored. Moreover, Figure 4.1 demonstrates the FCR performance measures of the BSVDD and the existing procedures

when one and two variables are changed. When one variable changes, Figure 4.1 (a) demonstrates that, for the gamma distribution, BSVDD identifies the changed variable more accurately than the K^2 decomposition procedure for small shift and that BSVDD performance is still comparable with K^2 decomposition procedure for moderate shift. However, the BSVDD and SVDD outperform other approaches in case of large shifts. In addition, when two variables change, Figure 4.1 (b) indicates that the FCR performance measures of the BSVDD and SVDD are much higher than the existing procedures, but BSVDD is significantly better than SVDD due to the use of prior information. Figures 4.1 (c) and (d) also illustrate the FCR performance for the multivariate lognormal distribution when one and two variables change, and show a similar performance to multivariate gamma distribution.

In fact, the performance of the proposed procedure can be controlled by the weight parameter $\kappa(\geq 0)$. When κ approaches infinity, ν_i approaches zero, which leads to ignoring the term $\sum_{i=1}^{m} \theta_i \tau_i$ in Eq. (4.6). Moreover, as κ becomes smaller, ν_i approaches 1, i.e., it assigns equal weights to τ_i. By the nature of optimization problem of Eq. (4.6), as the value of ν_i approaches 1, it tends to have more sparse support vectors and vice versa. Since the magnitude of ν_i is inversely proportional to the density, the sensitivity of the procedure can be controlled by the parameter κ. In this chapter, we assign the parameter weight from zero to one to determine the effect of κ.

Tables 4.3 and 4.4 show the CR performance of BSVDD for different $\kappa(\leq 1)$ parameters under the multivariate gamma and lognormal distributions, respectively. CR values are obtained by evaluating different κ values using cross-validation for given data sets. From these tables, it is clear that, for a given data set, the CR performance of the BSVDD is better for κ values between 0.2 and 0.4. Thus, we choose an appropriate κ value, which is 0.3 for multivariate gamma and multivariate lognormal distributed data.

Table 4.3 CR performance of BSVDD with different κ parameters under multivariate gamma distribution with $p = 5$

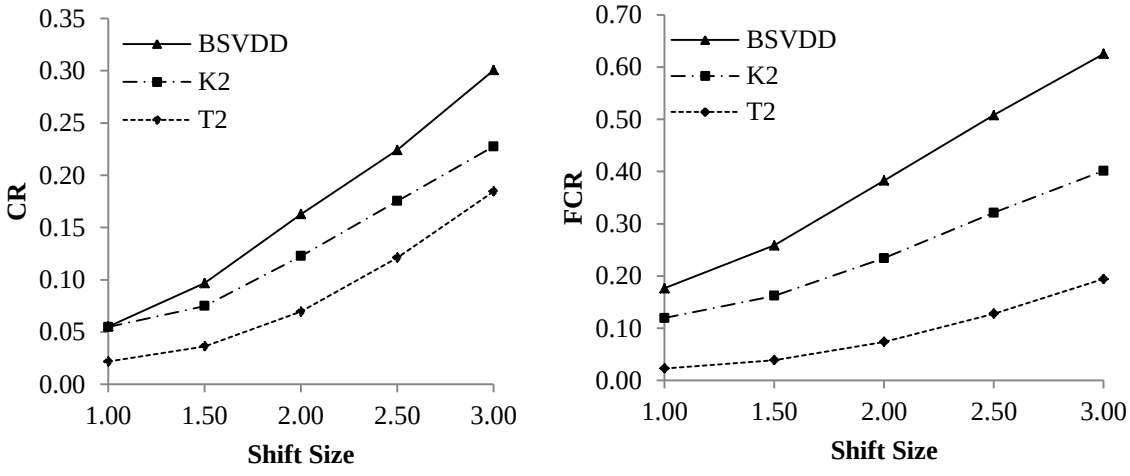
Shift		CR				
Direction	Size	κ =0.2	κ =0.4	κ =0.6	κ =0.8	κ =1
{s 0 0 0 0}	1.00	0.2586	0.2604	0.2545	0.2580	0.2508
	2.00	0.7134	0.7142	0.7133	0.7115	0.7121
	3.00	0.7289	0.7280	0.7238	0.7157	0.7227
{s 0 s 0 0}	1.00	0.1531	0.1571	0.1503	0.1522	0.1513
	2.00	0.7256	0.7184	0.7142	0.7027	0.7075
	3.00	0.7654	0.7644	0.7641	0.7607	0.7678
{s 0 s 0 s}	1.00	0.1559	0.1550	0.1520	0.1445	0.1532
	2.00	0.6913	0.6816	0.6810	0.6769	0.6929
	3.00	0.8156	0.8157	0.8120	0.8078	0.8131

Table 4.4 CR performance of BSVDD with different κ parameters under multivariate lognormal distribution with $p = 5$

Shift		CR				
Direction	Size	κ =0.2	κ =0.4	κ =0.6	κ =0.8	κ =1
{s 0 0 0 0}	1.00	0.2538	0.2538	0.2529	0.2528	0.2529
	2.00	0.6439	0.6446	0.6424	0.6431	0.6435
	3.00	0.6380	0.6379	0.6355	0.6362	0.6370
{s 0 s 0 0}	1.00	0.2443	0.2434	0.2435	0.2441	0.2453
	2.00	0.7182	0.7188	0.7090	0.7105	0.7108
	3.00	0.7623	0.7683	0.7565	0.7663	0.7645
{s 0 s 0 s}	1.00	0.3759	0.3744	0.3729	0.3723	0.3734
	2.00	0.8193	0.8178	0.8173	0.8101	0.8109
	3.00	0.8061	0.8067	0.7936	0.8058	0.8037

In addition to the above non-normal data sets, we also consider the irregularly shaped data in our simulation study. Duin *et al.* (2000) introduce a procedure for obtaining an irregular two dimensional data called banana-shaped dataset. This procedure is used to generate six-dimensional datasets which are obtained by integrating three two-dimensional banana-shaped datasets. Since this data set has an irregular shape, it is difficult to represent it with an ellipsoid boundary. The CR and FCR performance of the banana-shaped data are presented in Figure 4.2 for cases when one variable changes. Figure 4.2 (a) shows that CR performance of BSVDD and K^2 are not comparable, specially for small shifts. However, in Figure 4.2 (b), FCR performance clearly shows that the proposed BSVDD outperforms the existing procedures in all cases. In addition, the performance of all of the distribution-free procedures is significantly higher than the T^2 procedure, which is an expected result due to the underlying assumptions of T^2

procedure. Therefore, BSVDD still outperforms the existing procedures and the performance of both T^2 and K^2 decreases when the number of shifted variables increases.



(a) (b)

Figure 4.2 CR (a) and FCR (b) performances of banana shaped data when one variable is changed

4.4.2 Performance Assessment with Generalized Non-Normal Data

Since the proposed BSVDD procedure does not assume any specific distribution, it is of interest to study the effects of deviation from normality on the proposed procedure. The multivariate skew normal (MSN) distribution takes advantage of general set-up for comparison under non-normal dataset by controlling skewness. In the following

subsection, we briefly review the MSN distribution and show the experimental results based on different skewness parameters.

4.4.2.1 Multivariate Skew Normal Distribution

A *p*-dimensional vector $\mathbf{x}$ follows the MSN distribution with the density function defined as $2\Phi_p(\mathbf{y}-\boldsymbol{\varepsilon};\boldsymbol{\Sigma})\boldsymbol{\Phi}(\mathbf{d}^T\mathbf{w}^{-1}(\mathbf{y}-\boldsymbol{\varepsilon}))$, $(\mathbf{x}\in\mathbf{R}^p)$ by defining $\mathbf{y}=\boldsymbol{\varepsilon}+\mathbf{w}\mathbf{x}$, where $\Phi_p(\mathbf{y}-\boldsymbol{\varepsilon};\boldsymbol{\Sigma})$ is the *p*-dimensional normal density with location and scale parameters, $\boldsymbol{\mu}=(\mu_1,...,\mu_p)^T$, $\mathbf{w}=(w_1,...,w_p)$ respectively, and correlation matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Phi}(\cdot)$ is the $N(0,1)$ distribution function. The parameter $\mathbf{d}$ is defined as *p*-dimensional skewness parameter. If $\mathbf{d}$ is a vector of zeros, $\boldsymbol{\Phi}(\mathbf{d}^T\mathbf{x})$ equals to $1/2$. Therefore, the density function defined above is reduced to *p*-dimensional normal distribution $N_p(0,\boldsymbol{\Sigma})$ (Azzalini and Capitanio, 1999, Azzalini and Dalla Valle, 1996). Throughout this chapter, we use the notation $\mathbf{y}\sim SN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{d})$ when $\mathbf{y}$ follows an MSN distribution. Gupta *et al.* (2004) obtain mean and covariance of a vector and the mean of $\mathbf{y}\sim SN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{d})$ which is based on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{d}$. To obtain the shifted observations for the simulation study, we only focus on $\boldsymbol{\mu}$. If $\boldsymbol{\mu}$ is the in-control mean, then the shifted observation is obtained from $\mathbf{y}\sim SN_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma},\mathbf{d})$ where $\boldsymbol{\mu}_1=\boldsymbol{\mu}+\boldsymbol{\delta}$ and $\boldsymbol{\delta}=(\delta_1,...,\delta_p)$.

In addition, skewness parameter **d** shows the deviance of the data from normality. When **d** deviates from zero, the data also deviate from normality. Figure 4.3 shows the effect of skewness parameter on a bivariate normal distribution.

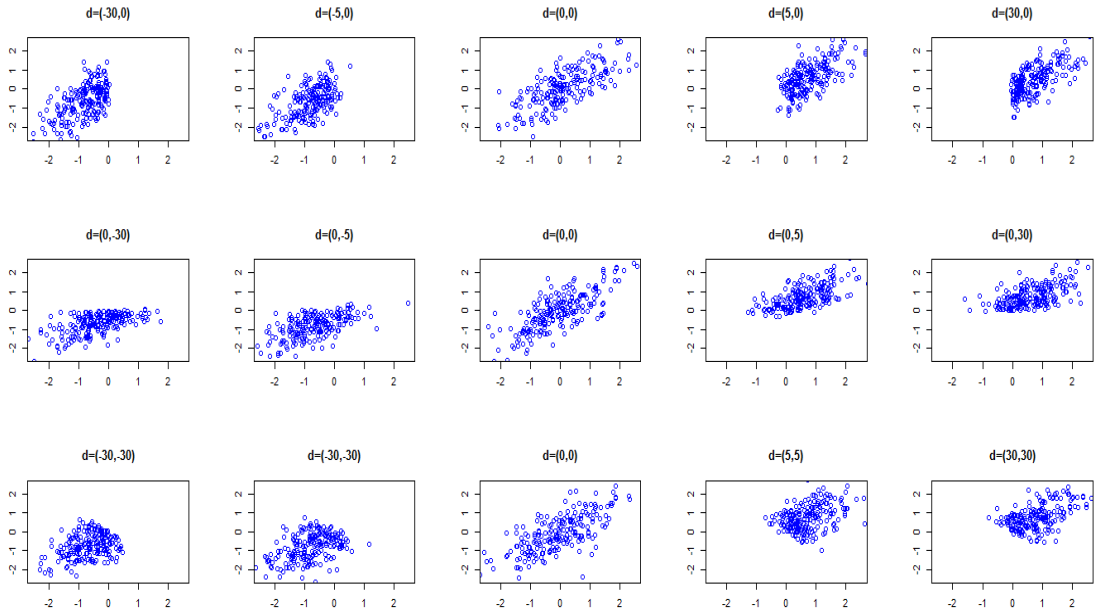


Figure 4.3 Effect of skewness parameter on a bivariate normal distribution

The first row in Figure 4.3 shows the change in the first parameter of **d** in a bivariate data. Increasing the first parameter makes the distribution more skewed towards the positive **x** axis. However, decreasing the first parameter increases the skewness towards the negative **x** axis. A similar pattern occurs in the second row which shows the effect of second variable. The third row shows the changes when both skewness parameters are

changed. It is also clear that changing the skewness parameter not only changes the mean but also changes the variance.

4.4.2.2 Performance of the BSVDD: Multivariate Skew Normal Data

In this section, we compare the proposed BSVDD fault identification with the existing procedures by using MSN distribution for different values of the skewness parameter **d**. As discussed, if **d** is a vector of zeros, the corresponding data follow multivariate normal distribution. First, we compare the proposed procedure with the existing procedures under the normality assumption, i.e., $\mathbf{d} = \mathbf{0}$.

In the simulation study, we generate the in-control data from MSN distribution assuming the same $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for all of the multivariate skew normal cases. Therefore, the only parameter used to obtain the multivariate skew normal data is **d**. We assume $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \left[\sigma_{ij}\right]_{1 \le i, j \le p}$ where $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.75$. Therefore, the mean shift is obtained as $\boldsymbol{\mu}_1 = \boldsymbol{\mu} + \boldsymbol{\delta}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)$ where δ_i represents the shift size of variable i.

Table 4.5 compares the CR performace of the proposed procedure and the existing procedures including the parametric procedure T^2 under the five dimensional multivariate normal distribution generated by MSN distribution by choosing **d** as a vector of zero. Table 4.5 shows that the T^2 procedure outperforms the existing procedures when the sparsity assumption is satisfied since it is designed for the normally

distributed data. Kim *et al.* (2016b) indicate that it may perform poor when the sparsity assumption is violated. In addition, since K^2 decomposition and HNS decomposition procedures are based on the regression-adjusted variables (Hawkins, 1991), the performances of K^2 decomposition and HNS decomposition procedures are comparable to the BSVDD and the SVDD procedures when only one variable changes, specially for small shifts. On the contrary, the BSVDD and the SVDD performance measures are comparable with T^2, specially, for small shifts when two variables change. However, CR of T^2 increases when the shift size is large. The most significant benefit of the BSVDD and the SVDD procedures occurs when the number of shifted variables increases. For example, when three variables are shifted, the performance of the BSVDD significantly outperforms the existing procedures including the SVDD.

Table 4.5 CR performance comparison of BSVDD and existing procedures under multivariate normal distribution

Direction	Size	T^2	K^2	HNS	SVDD	BSVDD
{s 0 0 0 0}	0.5	0.3999	0.2024	0.0250	0.1284	0.1969
	1.0	0.7100	0.3653	0.0601	0.2889	0.3725
	1.5	0.8674	0.4627	0.0850	0.4278	0.5152
	2.0	0.9488	0.5224	0.1009	0.5743	0.6238
{s s 0 0 0}	0.5	0.0497	0.0239	0.0345	0.0698	0.0894
	1.0	0.1741	0.0433	0.0536	0.1712	0.2123
	1.5	0.3525	0.0691	0.0706	0.3163	0.3266
	2.0	0.6079	0.0913	0.0911	0.4522	0.4626
{s s s 0 0}	0.5	0.0034	0.0053	0.0207	0.0394	0.0728
	1.0	0.0099	0.0012	0.0699	0.1029	0.1783
	1.5	0.0315	0.0000	0.0945	0.1616	0.2833
	2.0	0.0609	0.0000	0.1208	0.2161	0.4135

Table 4.6 demonstrates the CR performance of the proposed and the existing procedures under MSN distribution with nonzero skew parameters. Two different data sets are considered by choosing **d**=(-2, 1, -2, 1,-2) and **d**= (2, -1, 2, -1, 2). By choosing this setting in which skewness parameters can be considered as symmetric, it is possible to observe how the proposed procedure responses to positive and negative skewness. From Table 4.6, we observe that the BSVDD shows "good" performance in most of the cases. However, for the large shift, T^2 performs "well" since the data can be represented by an ellipsoid when OC signal is far from the boundary. On the other hand, when more than one variable shifts, the BSVDD outperforms the existing procedures due to the violation of the sparsity assumption.

Table 4.6 CR performance comparisons of BSVDD and existing procedures under MSN data with different skewness parameters

Direction	Size	T^2	K^2	HNS	SVDD	BSVDD
{s 0 0 0 0}	0.5	0.1099	0.2192	0.0600	0.0764	0.1437
	1.0	0.3522	0.4229	0.1366	0.2482	0.3564
	1.5	0.6522	0.5482	0.1931	0.4625	0.5469
	2.0	0.8594	0.5893	0.2341	0.6503	0.6599
{s s 0 0 0}	0.5	0.0045	0.0111	0.0489	0.0544	0.0905
	1.0	0.0399	0.0159	0.1026	0.1926	0.2486
	1.5	0.1835	0.0439	0.1507	0.4062	0.4370
	2.0	0.5045	0.0871	0.1964	0.6059	0.6203
{s s s 0 0}	0.5	0.0003	0.0022	0.0377	0.0432	0.0448
	1.0	0.0053	0.0011	0.0770	0.1630	0.1665
	1.5	0.0516	0.0004	0.1284	0.3409	0.3724
	2.0	0.1878	0.0002	0.1948	0.4995	0.6078

(a) Skewness parameter $\mathbf{d} = (-2,\ 1, -2,\ 1, -2)$

Direction	Size	T^2	K^2	HNS	SVDD	BSVDD
{s 0 0 0 0}	0.5	0.2075	0.3797	0.1150	0.0840	0.2314
	1.0	0.4844	0.5914	0.2207	0.2387	0.4309
	1.5	0.7674	0.6744	0.2674	0.4523	0.5760
	2.0	0.9121	0.6860	0.2854	0.6558	0.6659
{s s 0 0 0}	0.5	0.0063	0.0110	0.0478	0.0550	0.0831
	1.0	0.0513	0.0076	0.0721	0.1907	0.2264
	1.5	0.2238	0.0280	0.1055	0.4031	0.4098
	2.0	0.5543	0.0649	0.1403	0.6018	0.6097
{s s s 0 0}	0.5	0.0007	0.0014	0.0288	0.0375	0.0273
	1.0	0.0044	0.0002	0.0521	0.1172	0.1128
	1.5	0.0335	0.0001	0.0987	0.2379	0.3080
	2.0	0.1075	0.0000	0.1795	0.3134	0.5772

(b) Skewness parameter $\mathbf{d} = (2, -1,\ 2, -1,\ 2)$

Table 4.7 demonstrates the CR performance of the proposed and the existing procedures with the same set up as in Table 4.5 with negative shift. Results for the negative shift are similar to the positive shift case. Briefly, when only one variable changes, K^2 decomposition outperforms other procedures in small shifts and the performance of ASD increases with large shifts. However, when more than one variable changes, the performance of the K^2 decomposition and T^2 decreases significantly and the BSVDD outperforms both procedures.

Table 4.7 Performance comparisons of BSVDD and existing procedures under MSN data with negative shifts

Shifted Variables {1}				Shifted Variables {1,2}			
Shift size	T^2	K^2	BSVDD	Shift size	T^2	K^2	BSVDD
	$\mathbf{d}=(-2,\ 1,-2,\ 1,-2)$				$\mathbf{d}=(-2,\ 1,-2,\ 1,-2)$		
-0.5	0.1232	0.2548	0.1294	-0.5	0.0068	0.0073	0.0416
-1.0	0.3421	0.4901	0.2972	-1.0	0.0575	0.0088	0.1607
-1.5	0.6321	0.6111	0.4811	-1.5	0.2427	0.0333	0.3801
	$\mathbf{d}=(0,0,0,0,0)$				$\mathbf{d}=(0,0,0,0,0)$		
-0.5	0.3819	0.2207	0.1872	-0.5	0.0678	0.0174	0.1151
-1.0	0.6754	0.3693	0.3622	-1.0	0.2136	0.0263	0.2360
-1.5	0.8605	0.4718	0.5023	-1.5	0.4182	0.0500	0.3595
	$\mathbf{d}=(2,-1,\ 2,-1,\ 2)$				$\mathbf{d}=(2,-1,\ 2,-1,\ 2)$		
-0.5	0.1290	0.2935	0.1772	-0.5	0.0080	0.0131	0.0848
-1.0	0.3704	0.5304	0.3875	-1.0	0.0672	0.0154	0.2525
-1.5	0.6528	0.6796	0.5473	-1.5	0.2607	0.0492	0.4374

4.5 Case Study: Monitoring the Change of Bolt Dimensions

In this section, we demonstrate the performance of the proposed procedure by applying it to a real dataset from an automated monitoring system for bolts' dimensions reported by Kim *et al.* (2017) using image-processing techniques. Bolts are placed on a conveyor belt system and infrared sensors are used to detect the presence of a bolt when it reaches the inspection station and cameras are triggered to take images of the bolts. Image processing is performed and several dimensions of the bolt are compared with corresponding threshold values. When any of the dimensions exceed the specified threshold values, the bolt is automatically removed from the conveyor using a diversion mechanism. We consider the bolts to have no faults if the dimensions are within the acceptable interval of the threshold values. Figure 4.4 shows the bolt inspection system.

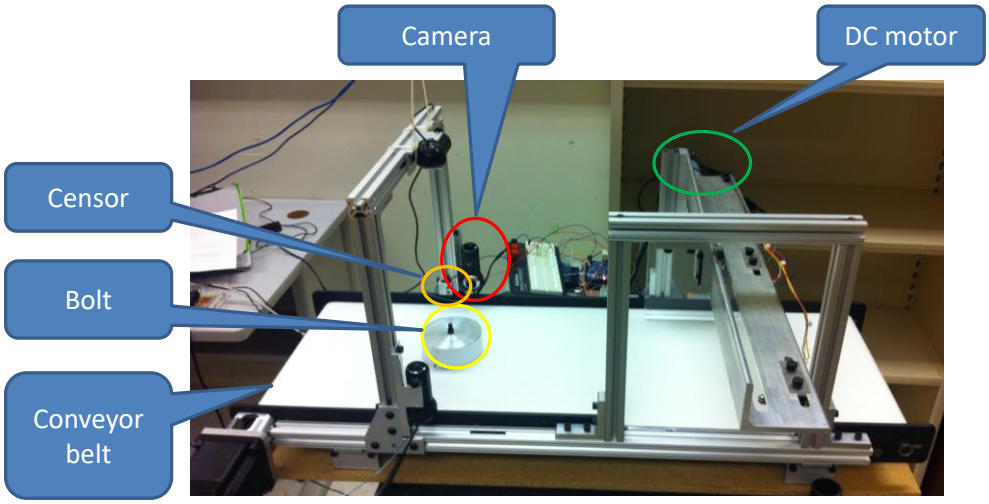


Figure 4.4 Conveyor belt system for measuring bolt dimensions automatically

In this experiment, we observe four characteristics: the head width (x_1), the head height (x_2), the bolt width (x_3), and the length of the bolt (x_4) as shown in Figure 4.5.

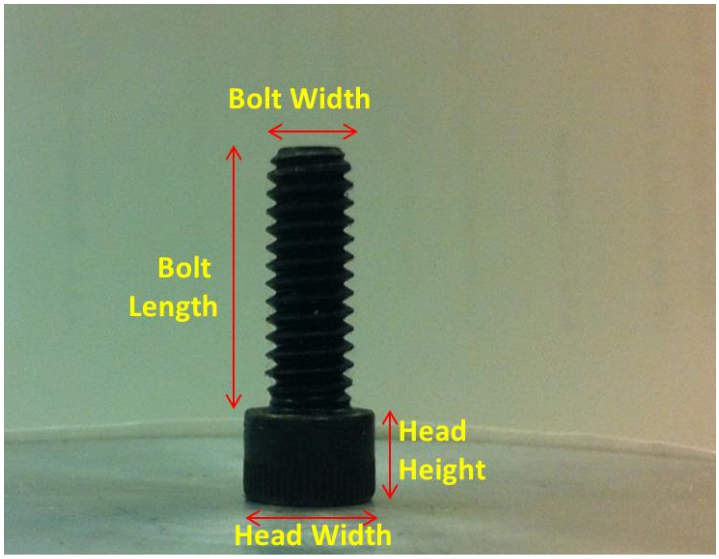


Figure 4.5 Measurements from bolt image

Sixty in-control observations are collected. In this experiment, it is desired to maintain bolt characteristics as close as possible to the target values $\boldsymbol{\mu}_0 = (0.3673,\ 0.2449,\ 0.2502,\ 0.7346)$ and $\sigma = (0.0018\ 0.0063\ 0.0032\ 0.0075)$. The correlation matrix for the in-control data is obtained as:

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & -0.1853 & 0.3231 & 0.2026 \\ -0.1853 & 1 & 0.1025 & -0.9511 \\ 0.3231 & 0.1025 & 1 & -0.1516 \\ 0.2026 & -0.9511 & -0.1516 & 1 \end{pmatrix}$$

In this study, fifteen out of control observations are obtained with a positive shift change for the mean of $\mathbf{x}_2$ and $\mathbf{x}_4$ to be 0.2575 and 0.7466, respectively (In-control and out-of-control observations are shown in Appendix B.). These observations are utilized in the proposed BSVDD and K^2-decomposition procedures. Among fifteen out-of-control

observations, K^2-decomposition procedure identifies all the variables correctly in 8 out of 15 (53.33%) observations. The proposed BSVDD-based approach identifies all the variables correctly in 13 out of 15 (86.67%) observations. These results show that the proposed BSVDD procedure significantly improves the fault identification ability when compared with the K^2-decomposition.

4.6 Conclusions

Fault diagnosis in statistical process control is a challenging issue. Although many fault detection procedures have been introduced, mostly based on T^2, they assume that the underlying distribution of the process follows a multivariate normal distribution. In this chapter, we propose a BSVDD based distribution-free faulty variable identification procedure which decreases the computational complexity of faulty variable identification in high-dimensional processes. We propose a local density degree of an observation by assigning different weights depending on the relative distance between the observations and their k^{th} neighbors in order to determine the prior parameters of the BSVDD.

Experiments with diverse data sets show an important feature of the BSVDD procedure, which is the robustness to the non-normal data, specially for irregularly patterned data, in terms of fault detections. This feature is important in practice when the type of the distribution is unknown. In addition, the proposed approach shows better performance when the number of shifted variables is greater than one for non-normal distributions.

This work can be extended for both monitoring and diagnosis in multistage processes as a future research. Since fault identification can be used for variable selection, the extension of the proposed procedure to variable selection problems would be an interesting direction for future research.

CHAPTER 5

GENERALIZED SUPPORT VECTOR DATA DESCRIPTION WITH BAYESIAN FRAMEWORK

5.1 Introduction

Identifying patterns that do not conform to normally behaved patterns is important since, in variety of applications, these patterns indicate critical and significant information that can be used to take actions to improve the applications. This kind of pattern is called anomaly or outlier. Several procedures based on classification-based, nearest neighbor-based, clustering-based, and statistical procedures have been developed to detect anomalies (Chandola *et al.*, 2009).

Among the classification-based procedures, improved versions of support vector machine (SVM) proposed by (Vapnik, 1995) are used to detect the anomalies. Schölkopf *et al.* (2001) introduce a concept based on the SVM by transforming the features via a kernel function. This procedure assumes that the origin is the only member of the second class and separates the other objects from the origin by drawing a hyperplane. This method is improved by considering all points “close enough” to the origin as second class along with the origin (Li *et al.*, 2003).

A particular case of the classification-based procedures called one-class classification procedures have a prominent place in the literature. They assume that all training observations are obtained from a certain distribution and describe the data with a boundary obtained around the normal patterns. One-class classification procedures identify the pattern as anomaly if it is placed outside of the learned boundary.

One of the promising data description procedure called support vector data description (SVDD) proposed by Tax and Duin (1999) finds a hypersphere which describes the data with minimal volume by transforming original observations into a new space using kernel functions. Usage of kernel transformation improves the power of SVDD specially if the original data are complex. SVDD has long been of a great interest in a wide range of applications mainly focusing on the detection of the anomalies as well as other real-life problems such as face recognition, image processing, pattern detection and quality control (Bovolo *et al.*, 2010, Lee *et al.*, 2006, Ning and Tsung, 2013).

Since SVDD has emerged as a powerful approach to identify the anomalies and several researches based on the origin of SVDD have been proposed to improve the traditional SVDD. By drawing on the concept of local density, Lee *et al.* (2005) are able to improve the SVDD by introducing weight for each observation using the local densities. In addition, Lee *et al.* (2007) introduce similar procedure by obtaining the local densities using the nearest neighbor and Parzen window approaches (Parzen, 1962). Recently, Ghasemi *et al.* (2016) introduce an SVDD with Bayesian approach (BSVDD) by assuming that a transformed data in the higher dimensional space follow a Gaussian

distribution. They assume the dual variables of the SVDD follow Gaussian distribution and show the superiority of BSVDD. The procedures mentioned above only use the normal observation. However, Tax and Duin (2004) introduce a procedure that also utilizes the negative samples (objects which should be rejected). This procedure, called negative SVDD (NSVDD), introduces an additional constraint for the negative samples by forcing them to be outside of the boundary.

In addition, extensive studies of SVDD have been introduced for classification. Lee and Lee (2007) introduce the classification procedure where the decision boundaries are based on the posterior probability distribution obtained from the SVDDs for each class. Mu and Nandi (2009) introduce a multistage multiclass classification procedure for obtaining the decision boundaries based on a combination of linear discriminant analysis and nearest-neighbor obtained from the NSVDDs for each class. Both of these procedures construct the boundary for each class by ignoring the interaction among classes. Kang and Cho (2012) propose a binary classification algorithm, named support vector class description (SVCD). Unlike the above classification procedures, SVCD penalizes the other class observations if they are not classified into their corresponding classes. However, these classification approaches do not identify the anomalies. In addition, most of the SVDD procedures for anomaly detection mentioned above assume that the target data has only one class of normal data. However, in many real-life problems, normal data may consist of more than one distribution or class. In this case, applying the traditional SVDD procedures may ignore the differences between the classes. To overcome this drawback, Huang *et al.* (2011) introduce an anomaly detection

procedure in which the normal data consists of two-classes called two-class SVDD (TC-SVDD).

However, existing one-class and two-class SVDD procedures have the following limitations. First, existing SVDD procedures are based on the assumption that normal data consist of one or two-classes. However, in many real-world applications, normal data may be obtained from more than two classes. Thus, existing procedures may not recognize the differences between the classes, and may result in a poor anomaly detection performance. Second, the existing deterministic SVDDs do not reflect the prior or domain-specific knowledge when they are applied to the real-world problems that are not considered to be in the same domain.

In this chapter, we propose a generalized SVDD procedure which simultaneously finds the hyperspheres which describe each class accurately by including as many as possible of its class observations. Regardless of the number of classes, the proposed procedure identifies the anomalies, based on the relative distance to the center of each hypersphere of each class. Moreover, we introduce a Bayesian framework for generalized SVDD procedure. The procedure considers a probabilistic behavior of the parameters by taking ‘prior knowledge’ from each class. Determination of the prior and posterior probabilities is a key factor for each class. Finally, we obtain a closed form expression for the proposed procedure by specifying some of the prior distributions.

This chapter is organized as follows. After the review of the existing SVDD procedures in Section 5.2, we propose the generalized SVDD procedure in Section 5.3. In Section 5.4, we propose the generalized Bayesian SVDD procedure. In Section 5.5, simulation studies and results are demonstrated. In Section 5.6, we apply the proposed procedure in a case study of a Continuous Stirred Tank Heater (CSTH), followed by the conclusions in Section 5.7.

5.2 Benchmark Procedures

By relaxing the normality assumption, we pursue an appropriate model to describe the data more accurately. Among a number of data description techniques, SVDD is an effective method for describing irregularly patterned data. For a given data set $\mathbf{D} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{R}^p, i = 1, ..., N\}$, SVDD finds a hypersphere which covers the data with minimal volume, with a center $\mathbf{a}$ and a radius R (Tax and Duin, 1999). To allow the misclassification in $\mathbf{D}$, we introduce ε_i to penalize outliers that have large distances between $\mathbf{x}_i$ and $\mathbf{a}$. The primal formulation of the problem is constructed as follows:

$$\begin{aligned} &\min \quad R^2 + C\sum_{i=1}^{N} \varepsilon_i \\ &s.t. \quad \left\|\mathbf{x}_i - \mathbf{a}\right\|^2 \le R^2 + \varepsilon_i, \quad \varepsilon_i \ge 0 \ \ \forall i \in \{1, ..., N\} \end{aligned}$$

where C is the regularization parameter which adjusts the volume of the hypersphere by considering the number of observations that are located outside the boundary. Dual formulation is obtained by using the Lagrangian function:

$$L\left(R,\mathbf{a},\varepsilon_i\right)=R^2+C\sum_{i=1}^{N}\varepsilon_i-\sum_{i=1}^{N}\tau_i\left(R^2+\varepsilon_i-\left(\mathbf{x}_i-\mathbf{a}\right)^T\left(\mathbf{x}_i-\mathbf{a}\right)\right)-\sum_{i=1}^{N}\gamma_i\varepsilon_i \tag{5.1}$$

where $\tau_i \geq 0$ and $\gamma_i \geq 0$ are Lagrangian dual variables. By taking partial derivatives of the Lagrangian function, we obtain

$$\frac{\partial L}{\partial R}=0\rightarrow\sum_{i=1}^{N}\tau_i=1,\ \frac{\partial L}{\partial \mathbf{a}}=0\rightarrow\mathbf{a}=\frac{\sum_{i=1}^{N}\tau_i\mathbf{x}_i}{\sum_{i=1}^{N}\tau_i}=\sum_{i=1}^{N}\tau_i\mathbf{x}_i,\ \frac{\partial L}{\partial \varepsilon_i}=0\rightarrow C-\tau_i-\gamma_i=0 \tag{5.2}$$

In this case, dual formulation in Eq. (5.3) is constructed by substituting Eq. (5.2) into Eq. (5.1).

$$\begin{aligned}\max\quad & \sum_{i=1}^{N}\tau_i\left(\mathbf{x}_i\cdot\mathbf{x}_i\right)-\sum_{i=1}^{N}\sum_{j=1}^{N}\tau_i\tau_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)\\ s.t.\quad & \sum_{i=1}^{N}\tau_i=1,\ 0\leq\tau_i\leq C\quad\forall i\in\{1,...,N\}\end{aligned} \tag{5.3}$$

Data points corresponding to the positive τ_i are called support vectors and they are placed on the boundary or outside the boundary. Based on the SVDD boundary with given C from the in-control data, a new observation $\mathbf{z}$ can be classified as in or out of the data boundary by checking the distance from $\mathbf{z}$ to the center of hypersphere. The inner products can be replaced by a kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$, defining $K(\mathbf{x}_i,\mathbf{x}_j)=\left\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\right\rangle$, then a more suitable boundary to cover the data can be obtained (Tax and Duin, 1999).

5.3 Generalized *n*-Class SVDD

In real life problems, in-control observations can be formed in more than one or two classes. In this case, existing SVDD procedures do not differentiate between different classes and do not obtain a description for each class. Therefore, in this section, we introduce multiclass SVDD called *n*-class SVDD (*n*-SVDD).

The *n*-SVDD is based on the assumption that the target data set contains *n* classes of objects. For given *n* classes, $\mathbf{D}_1 = \left\{\mathbf{x}_1^{(1)},\ldots,\mathbf{x}_{N_1}^{(1)}\right\}, \mathbf{D}_2 = \left\{\mathbf{x}_1^{(2)},\ldots,\mathbf{x}_{N_2}^{(2)}\right\},\ldots,\mathbf{D}_n = \left\{\mathbf{x}_1^{(n)},\ldots,\mathbf{x}_{N_n}^{(n)}\right\}$, the goal of the *n*-SVDD is to find *n* hyperspheres which cover each class with minimal volume, with centers $\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n$ and radiuses $R_1,R_2,\ldots,R_n$. The primal formulation of the *n*-SVDD is constructed as follows:

$$\begin{aligned} &\min \textstyle\sum_{k=1}^{n} R_k^2 \\ &s.t.\ \left\|\mathbf{x}_i^{(k)} - \mathbf{a}_k\right\|^2 \le R_k^2 \ \ k = 1,\ldots,n\ ,\ i = 1,\ldots N_k \\ &\quad \left\|\mathbf{x}_i^{(m)} - \mathbf{a}_k\right\|^2 \ge R_k^2 \ \ m \ne k,\ \ m,k = 1,\ldots,n \quad\ \ i = 1,\ldots,N_m \end{aligned} \tag{5.4}$$

By the optimization problem in Eq. (5.4), every observation in class *n* falls inside the hypersphere *n* (constraint 1) and falls outside the other hyperspheres (constraint 2). However, to allow for some observations to be outside its class's boundary or to be inside the other classes' boundary, we introduce penalty functions $\varepsilon_i^{(k)}$ and $\varsigma_i^{(m,k)}$. Then, the primal formulation of an optimization problem is given as follows:

$$
\begin{aligned}
&\min \sum_{k=1}^{n} R_k^2 + \sum_{k=1}^{n}\left(C_k \sum_{i=1}^{N_k} \varepsilon_i^{(k)}\right) + \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\left(B_{(m,k)} \sum_{i=1}^{N_m} \varsigma_i^{(m,k)}\right) \\
&s.t.\ \left\|\mathbf{x}_i^{(k)} - \mathbf{a}_k\right\|^2 \le R_k^2 + \varepsilon_i^{(k)}\ \ k = 1,\ldots,n\ ,\ i = 1,\ldots N_k \\
&\qquad \left\|\mathbf{x}_i^{(m)} - \mathbf{a}_k\right\|^2 \ge R_k^2 - \varsigma_i^{(m,k)}\ \ m \neq k,\ \ m,k = 1,\ldots,n\ ,\ \ i = 1,\ldots,N_m \\
&\varepsilon_i^{(k)}, \zeta_i^{(m,k)} \ge 0\ \ \forall i,k,m
\end{aligned}
\tag{5.5}
$$

where C_k and $B_{(m,k)}$ are the regularization parameters which control the volume of the hyperspheres.. The two constraints ensure that if an observation $\mathbf{x}_i^{(k)}$ falls outside of the hypersphere *k*, the objective function increases by $C_k\varepsilon_i^{(k)}$ and if an observation $\mathbf{x}_i^{(m)}$ falls inside the hypersphere *k* the objective function increases by $B_{(m,k)}\varsigma_i^{(m,k)}$. Table 5.1 shows the notations used to obtain *n*-SVDD.

Table 5.1 Notations of *n*-SVDD

R_k	: radius of class *k* , (*k*=1,..,*n*)
$\mathbf{a}_k$	: center of class *k* , (*k*=1,..,*n*)
$\varepsilon_i^{(k)}$	: penalty given to training samples of class *k* which lie outside the hypersphere *k*.
$\zeta_i^{(m,k)}$	: penalty given to training samples of class *m* which lie inside the hypersphere *k*.

Dual formulation of *n*-SVDD is obtained by introducing the following Lagrangian function as shown in Eq. (5.6).

$$
\begin{aligned}
L\left(R_n, \mathbf{a}_n, \varepsilon_i^{(k)}, \zeta_i^{(m,k)}\right) = & \sum_{k=1}^{n} R_k^2 + \sum_{k=1}^{n}\left(C_k \sum_{i=1}^{N_k} \varepsilon_i^{(k)}\right) + \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\left(B_{(m,k)} \sum_{i=1}^{N_m} \zeta_i^{(m,k)}\right) \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_k} a_i^{(k)}\left[R_k^2 + \varepsilon_i^{(k)} - \left(\mathbf{x}_i^{(k)} - \mathbf{a}_k\right)^2\right] \\
& - \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sum_{i=1}^{N_m} \beta_i^{(m,k)}\left[\left(\mathbf{x}_i^{(m)} - \mathbf{a}_k\right)^2 - R_k^2 + \zeta_i^{(m,k)}\right] \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_k} \gamma_i^{(k)} \varepsilon_i^{(k)} - \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sum_{i=1}^{N_m} \gamma_i^{(m,k)} \zeta_i^{(m,k)}
\end{aligned}
\tag{5.6}
$$

where $\alpha_i^{(k)} \geq 0, \beta_i^{(m,k)} \geq 0, \gamma_i^{(k)} \geq 0$ $\alpha_i^{(k)} \geq 0, \beta_i^{(m,k)} \geq 0, \gamma_i^{(k)} \geq 0$ and $\gamma_i^{(m,k)} \geq 0$. By taking partial derivatives of the Lagrangian function $L\left(R_n, \mathbf{a}_n, \varepsilon_i^{(k)}, \zeta_i^{(m,k)}\right)$, the following equations are obtained.

$$
\begin{aligned}
& \frac{\partial L}{\partial R_k} = 0 \Rightarrow \sum_{i=1}^{N_k} \alpha_i^{(k)} - \sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m} \beta_i^{(m,k)} = 1 \\
& \frac{\partial L}{\partial \mathbf{a}_k} = 0 \Rightarrow \mathbf{a}_k = \sum_{i=1}^{N_k} \alpha_i^{(k)} \mathbf{x}_i^{(k)} - \sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m} \beta_i^{(m,k)} \mathbf{x}_i^{(m)} \\
& \frac{\partial L}{\partial \varepsilon_i^{(k)}} = 0 \Rightarrow C_k - \gamma_i^{(k)} - \alpha_i^{(k)} = 0 \Rightarrow 0 \leq \alpha_i^{(k)} \leq C_k \\
& \frac{\partial L}{\partial \zeta_i^{(m,k)}} = 0 \Rightarrow B_{(m,k)} - \gamma_i^{(m,k)} - \beta_i^{(m,k)} = 0 \Rightarrow 0 \leq \beta_i^{(m,k)} \leq B_{(m,k)} \\
& \forall m, k
\end{aligned}
\tag{5.7}
$$

Therefore, we obtain the dual formulation in Eq. (5.8) by substituting Eq. (5.7) into Eq. (5.6).

$$\begin{aligned}\max & \sum_{m=1}^{n}\sum_{i=1}^{N_m}\alpha_i^{(m)}\left(\mathbf{x}_i^{(m)}\cdot\mathbf{x}_i^{(m)}\right)-\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}\left(\mathbf{x}_i^{(m)}\cdot\mathbf{x}_i^{(m)}\right)\\ & -\sum_{k=1}^{n}\left(\sum_{i=1}^{N_k}\alpha_i^{(k)}\mathbf{x}_i^{(k)}-\sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}\mathbf{x}_i^{(m)}\right)^2\\ s.t.\ & \sum_{i=1}^{N_k}\alpha_i^{(k)}-\sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}=1,\quad \forall k=1,\ldots,n\\ & 0\le\alpha_i^{(k)}\le C_k,\quad \forall i,k\\ & 0\le\beta_i^{(m,k)}\le B_{(m,k)},\ \forall m,k,\ m\neq k=1,\ldots,n\end{aligned} \tag{5.8}$$

The vectors $\mathbf{x}_i^{(k)}$ with $\alpha_i^{(k)}>0$ and $\mathbf{x}_i^{(m)}$ with $\beta_i^{(m,k)}>0$ $(m\neq k)$ are called support vectors for class *k*. For an in-control observation $\mathbf{x}_i^{(k)}$, if $\alpha_i^{(k)}=0$ and $0<\beta_i^{(k,m)}<B_{(k,m)}$, then $\mathbf{x}_i^{(k)}$ is inside hypersphere *k* and lies on the hypersphere $m\,(m\neq k)$. Briefly, a support vector $\mathbf{x}_i^{(k)}$ with non-zero $\alpha_i^{(k)}$ and non-zero $\beta_i^{(k,m)}$ is located on or outside of hypersphere *k*.

We obtain the R_k by calculating the distance from the center $\mathbf{a}_k$ of the hypersphere to any of the support vectors of class *k* except the support vectors with $\alpha_i^{(k)}=C_k$ and $\beta_i^{(m,k)}=B_{(m,k)}$. Let $\mathbf{x}_s$ be a support vector of class *k* with the conditions $0<\alpha_s^{(k)}<C_k$ and $0<\beta_i^{(m,k)}<B_{(m,k)}$. Then,

$$\begin{aligned}R_k^2 &=\left\|\mathbf{x}_s-\mathbf{a}_k\right\|^2\quad,\forall k=1,\ldots,n\\ &=\left(\mathbf{x}_s\cdot\mathbf{x}_s\right)-2\sum_{i=1}^{N_k}\alpha_i^{(k)}\left(\mathbf{x}_s\cdot\mathbf{x}_i^{(k)}\right)+2\sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}\left(\mathbf{x}_s\cdot\mathbf{x}_i^{(m)}\right)+\sum_{i=1}^{N_k}\sum_{i=1}^{N_k}\alpha_i^{(k)}\alpha_j^{(k)}\left(\mathbf{x}_i^{(k)}\cdot\mathbf{x}_j^{(k)}\right)\\ &+\sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\sum_{k\neq t=1}^{n}\sum_{j=1}^{N_t}\beta_i^{(m,k)}\beta_j^{(t,k)}\left(\mathbf{x}_i^{(m)}\cdot\mathbf{x}_j^{(t)}\right)-2\sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\sum_{j=1}^{N_k}\beta_i^{(m,k)}\alpha_j^{(k)}\left(\mathbf{x}_i^{(m)}\cdot\mathbf{x}_j^{(k)}\right)\end{aligned}$$

To decide whether a new observation $\mathbf{z}$ is anomaly or not, we introduce a decision rule based on the knowledge obtained from the *n* hyperspheres, such as R_k, $\mathbf{a}_k$. Introduced decision rule obtains the distances from a new observation $\mathbf{z}$ to the center of the each hyperspheres, $\mathbf{a}_k$. Therefore, a new observation $\mathbf{z}$ is called anomaly if Eq. (5.9) is satisfied

$$\|\mathbf{z}-\mathbf{a}_k\|^2 > R_k^2 \quad \forall k = 1,\ldots,n. \tag{5.9}$$

This equation shows that a new observation $\mathbf{z}$ an anomaly if it is not placed in any of the hyperspheres.

By replacing the dot products with kernel function, we can obtain the flexible boundaries. The inner products, $\mathbf{x}_i \cdot \mathbf{x}_j$ can be replaced by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, defining $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. In this chapter, we use Gaussian kernel function which is defined as:

$$K(\mathbf{x},\mathbf{y}) = exp\left(\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2w^2}\right)$$

where w is a parameter of the Gaussian kernel.

5.4 Bayesian *n*-Class SVDD

In this section, we introduce the Bayesian framework for generalized *n*-SVDD. In traditional SVDD procedures, parameters are deterministic. However, the centers of each hypersphere $(\mathbf{a}_k, k=1,\ldots,n)$ can be a random variable which allows us to obtain a probabilistic interpretation. Therefore, in this section, we introduce the Bayesian *n*-class SVDD (B-*n*SVDD).

B-*n*SVDD approach assumes that in-control observation $\mathbf{x}_i^{(k)}$ $\left(k=1,\ldots,n\right)$ is transformed into higher dimensional space through $\phi(\cdot)$ and the transformed data $\phi\left(\mathbf{x}_i^{(k)}\right)$ follows a Gaussian distribution.

$$\phi(\mathbf{x}_i^{(k)}) \sim N\left(\sum_{i=1}^{N_k}\alpha_i^{(k)}\mathbf{x}_i^{(k)} - \sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}\mathbf{x}_i^{(m)}, \sigma_{kk}^2\mathbf{I}\right).$$

Distance of a point $\mathbf{x}_i^{(k)}$ to the center of a hypersphere *k* is inversely proportional to the likelihood in the weighted Gaussian model. Thus, *n*-SVDD is a special case of the weighted Gaussian model, which improves *n*-SVDD by utilizing precise prior knowledge.

The unknown weight parameters $\boldsymbol{\alpha}$ is estimated through a Bayesian approach with a proper prior distribution for $\boldsymbol{\alpha}$ which is obtained from $\boldsymbol{\alpha}_i^{(k)}$ and $\boldsymbol{\beta}_i^{(m,k)}$ (dual variables of the *n*-SVDD).

$$\boldsymbol{\alpha}^{(k)} = \left(\alpha_1^{(k)} \ldots \alpha_{Nk}^{(k)}\right)^T, \boldsymbol{\beta}^{(m,k)} = \left(\beta_1^{(m,k)} \ldots \beta_{N_m}^{(m,k)}\right)^T$$

B-*n*SVDD approach assumes that the prior distributions of the dual variables $\alpha_i^{(k)}$ and $\beta_i^{(m,k)}$ are Gaussian distributions which is the conjugate prior of the likelihood below

$$p\left(\boldsymbol{\alpha}^{(k)}\right) = \frac{1}{\left(2\pi\right)^{N_k/2}\sigma^{N_k}} e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{\alpha}^{(k)}-\mathbf{m}^{(k)}\right\|_2^2}, \; p\left(\boldsymbol{\beta}^{(m,k)}\right) = \frac{1}{\left(2\pi\right)^{N_m/2}\sigma^{N_m}} e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{\beta}^{(m,k)}-\mathbf{m}^{(m,k)}\right\|_2^2}$$

Since it is assumed that training data in each class $\mathbf{x}_i^{(k)}$ mapped into a higher dimensional kernel space follow a Gaussian distribution, the likelihood probability given parameter $\boldsymbol{\alpha}$ becomes

$$p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right) = \prod_{j=1}^{n}\left[\prod_{k=1}^{N_j}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\sigma_{jj}{}^{\hat{p}}} e^{-\frac{1}{2\sigma_{jj}^2}\left\|\phi\left(\mathbf{x}_k^{(j)}\right)-\left(\sum_{i=1}^{N_j}\alpha_i^{(j)}\mathbf{x}_i^{(j)}-\sum_{j\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,j)}\mathbf{x}_i^{(m)}\right)\right\|_2^2}\right]$$

Maximizing a posterior (MAP) is derived by the typical Bayesian rule as

$$p(\boldsymbol{\alpha}\,|\,\mathbf{D}) = \frac{p(\mathbf{D}\,|\,\boldsymbol{\alpha})\,p(\boldsymbol{\alpha})}{p(\mathbf{D})},$$

where $\mathbf{D} = \mathbf{D}_1 \cup \ldots \cup \mathbf{D}_n$ is a set of training data. Since $p(\mathbf{D})$ is a normalizing constant independent of $\boldsymbol{\alpha}$, it can be ignored so that

$$p(\boldsymbol{\alpha} \mid \mathbf{D}) \propto p(\mathbf{D} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\alpha}). \tag{5.10}$$

The solution of MAP is given by

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} p(\boldsymbol{\alpha} \mid \mathbf{D}). \tag{5.11}$$

We obtain the MAP solution by taking the logarithm and using the relationships in Eq. (5.10) and Eq. (5.11).

If the number of classes is two ($n=2$), the MAP solution is obtained as in Eq. (5.12) (in this equation $\boldsymbol{\beta}^{(m,k)}$ is replaced with $\boldsymbol{\beta}^{(m)}$).

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \left[2^{-1} \times \begin{pmatrix} 2\sum_{m=1}^{2} \sigma_{mm}^{-2} \left(\alpha^{(m)}\right)^T \mathbf{B}^{(m)} \mathbf{1}^{(m)} - 2\sum_{m=1}^{2}\sum_{m\neq k=1}^{2} \sigma_{kk}^{-2} \left(\beta^{(m)}\right)^T \mathbf{B}^{(m,k)} \mathbf{1}^{(m,k)} \\ -\sum_{k=1}^{2} N_k \sigma_{kk}^{-2} \left(\alpha^{(k)}\right)^T \mathbf{K}^{(k)} \alpha^{(k)} - \sum_{m=1}^{2}\sum_{m\neq k=1}^{2} N_m \sigma_{mm}^{-2} \left(\beta^{(k)}\right)^T \mathbf{K}^{(k)} \beta^{(k)} \\ +2\sum_{m=1}^{2}\sum_{m\neq k=1}^{2} N_m \sigma_{mm}^{-2} \left(\alpha^{(m)}\right)^T \mathbf{K}^{(m,k)} \beta^{(k)} \\ -\sigma^{-2}\left(\sum_{i=1}^{2} \boldsymbol{\alpha}^{(i)^T} \boldsymbol{\alpha}^{(i)} - 2\boldsymbol{\alpha}^{(i)^T} \mathbf{m}^{(i)}\right) - \sigma^{-2}\left(\sum_{i=1}^{2}\sum_{i\neq j=1}^{2} \boldsymbol{\beta}^{(i)^T} \boldsymbol{\beta}^{(i)} - 2\boldsymbol{\beta}^{(i)^T} \mathbf{m}^{(i,j)}\right) \end{pmatrix} \right] \tag{5.12}$$

where $\mathbf{B}_{ii}^{(m)} = \sum_j \mathbf{K}_{ij}^{(m)}$, $\mathbf{1}^{(m)}$ and $\mathbf{1}^{(m,k)}$ are $1 \times N_m$ vector with all ones and $\mathbf{B}_{ii}^{(m,k)} = \sum_j \mathbf{K}_{ij}^{(m,k)}$ (see Appendix C. for detailed derivation of Eq. (5.12)). Kernel matrix $\mathbf{K}$ is identified for two classes as follows:

$\mathbf{K} = \begin{bmatrix} \mathbf{K}^{(1)} & \mathbf{K}^{(1,2)} \\ \mathbf{K}^{(2,1)} & \mathbf{K}^{(2)} \end{bmatrix}$ where $\mathbf{K}^{(k)}$ is a $N_k \times N_k$ matrix and for each element of $\mathbf{K}^{(k)}$ matrix is obtained by the kernel distances of each element in class *k*.

$$\mathbf{K}_{ij}^{(k)} = K\left(\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(1)}\right) = \left\langle \phi\left(\mathbf{x}_i^{(k)}\right), \phi\left(\mathbf{y}_j^{(k)}\right) \right\rangle \quad \forall \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(k)} \in \mathbf{D}_k$$

$$\mathbf{K}_{ij}^{(m,k)} = K\left(\mathbf{x}_i^{(m)}, \mathbf{y}_j^{(k)}\right) = \left\langle \phi\left(\mathbf{x}_i^{(m)}\right), \phi\left(\mathbf{y}_j^{(k)}\right) \right\rangle, \forall \mathbf{x}_i^{(m)} \in \mathbf{D}_m, \forall \mathbf{y}_j^{(k)} \in \mathbf{D}_k \quad \forall m \neq k = 1,2$$

Note that the optimization problem needs to satisfy the same constraints of the original optimization problem in Eq. (5.8). Thus, B-2SVDD satisfies the following constraints:

$$\sum_{i=1}^{N_1} \alpha_i^{(1)} - \sum_{i=1}^{N_2} \beta_i^{(2)} = 1 \quad , \quad \sum_{i=1}^{N_2} \alpha_i^{(2)} - \sum_{i=1}^{N_1} \beta_i^{(1)} = 1$$
$$0 \leq \alpha_i^{(1)} \leq C_1 \quad , \quad 0 \leq \alpha_i^{(2)} \leq C_2$$
$$0 \leq \beta_i^{(1)} \leq B_{(1,2)} \quad , \quad 0 \leq \beta_i^{(2)} \leq B_{(2,1)}$$

The optimization problem of B-2SVDD is represented in a matrix form as follows:

$$\min 2^{-1}\left(\boldsymbol{\alpha}^{(1)}\ \boldsymbol{\alpha}^{(2)}\ \boldsymbol{\beta}^{(1)}\ \boldsymbol{\beta}^{(2)}\right)\begin{pmatrix} N_1\sigma_{11}^{-2}\mathbf{K}^{(1)}+\sigma^{-2}\mathbf{1} & 0 & 0 & -N_1\sigma_{11}^{-2}\mathbf{K}^{(1,2)} \\ 0 & N_2\sigma_{22}^{-2}\mathbf{K}^{(2)}+\sigma^{-2}\mathbf{1} & -N_2\sigma_{22}^{-2}\mathbf{K}^{(2,1)} & 0 \\ 0 & -N_2\sigma_{22}^{-2}\mathbf{K}^{(1,2)} & N_2\sigma_{22}^{-2}\mathbf{K}^{(1)} & 0 \\ -N_1\sigma_{11}^{-2}K^{(2,1)} & 0 & 0 & N_1\sigma_{11}^{-2}K^{(2)} \end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}^{(1)}\\ \boldsymbol{\alpha}^{(2)}\\ \boldsymbol{\beta}^{(1)}\\ \boldsymbol{\beta}^{(2)}\end{pmatrix}$$

$$+2^{-1}\left(\boldsymbol{\alpha}^{(1)}\ \boldsymbol{\alpha}^{(2)}\ \boldsymbol{\beta}^{(1)}\ \boldsymbol{\beta}^{(2)}\right)\begin{pmatrix} -2\sigma_{11}^{-2}diag\left(\mathbf{B}^{(1)}\right)-2\sigma^{-2}\mathbf{m}^{(1)} \\ -2\sigma_{22}^{-2}diag\left(\mathbf{B}^{(2)}\right)-2\sigma^{-2}\mathbf{m}^{(2)} \\ 2\sigma_{22}^{-2}diag\left(\mathbf{B}^{(1,2)}\right)-2\sigma^{-2}\mathbf{m}^{(1,2)} \\ 2\sigma_{11}^{-2}diag\left(\mathbf{B}^{(2,1)}\right)-2\sigma^{-2}\mathbf{m}^{(2,1)} \end{pmatrix}$$

$$s.t.\ \begin{pmatrix}\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{-1}\\ \mathbf{0} & \mathbf{1} & \mathbf{-1} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}^{(1)}\\ \boldsymbol{\alpha}^{(2)}\\ \boldsymbol{\beta}^{(1)}\\ \boldsymbol{\beta}^{(2)}\end{pmatrix}=\begin{pmatrix}1\\ 1\end{pmatrix}$$

$$0\le\alpha_i^{(1)}\le C_1\ ,\ \ 0\le\alpha_i^{(2)}\le C_2$$

$$0\le\beta_i^{(1)}\le B_{(1,2)}\ ,\ 0\le\beta_i^{(2)}\le B_{(2,1)}$$

B-2SVDD is a quadratic programming formulation and it has an optimum solution.

5.4.1 Determination of the Prior Parameters

In the proposed optimization problem, the parameters of the distribution of $\boldsymbol{\alpha}$ are needed to be determined. To obtain some insight, we recall the original TC-SVDD concept. We provide the reasoning only for the elements of class 1 because the reasoning for class 2 is then straightforward to determine the parameters for the two classes $\boldsymbol{\alpha}=\left(\boldsymbol{\alpha}^{(1)},\boldsymbol{\beta}^{(1)},\boldsymbol{\alpha}^{(2)},\boldsymbol{\beta}^{(2)}\right)$.

It is known that if the vector $\mathbf{x}_i^{(1)}$ is inside the boundary, the corresponding $\alpha_i^{(1)}$ is zero. Since boundary covers the densest area (note that $\mathbf{x}_i^{(1)}$ which has $\alpha_i^{(1)} = 0$ is close to the most of the vectors). Therefore, the sum of the kernel distances between $\mathbf{x}_i^{(1)}$ and other vectors in class 1 is small. However, if $\mathbf{x}_i^{(1)}$ vector is outside of the boundary, then $\alpha_i^{(1)}$ equals C_1. Since outside of the boundary is a less dense area, the sum of the kernel distance between $\mathbf{x}_i^{(1)}$ and other vectors in class 1 is large. In addition, if the $\mathbf{x}_i^{(1)}$ vector is on the boundary, corresponding $\alpha_i^{(1)}$ satisfies $0 < \alpha_i^{(1)} < C_1$. This area is also less dense. Since the $\mathbf{x}_i^{(1)}$ vector is on the boundary, the sum of the kernel distance between $\mathbf{x}_i^{(1)}$ and other vectors is also large.

Therefore it is reasonable to adjust $\boldsymbol{\alpha}^{(1)}$ by assigning a smaller weight to vectors if it is inside the boundary and a larger weight for the others. Therefore, we can determine the $\mathbf{m}^{(1)}$ as follows:

$$m_i^{(1)} \propto -\sum_{j \in \mathbf{D}_1} \mathbf{K}_{i,j}^{(1)}$$

where $\mathbf{m}^{(1)} = (m_1^{(1)}, ..., m_{N_1}^{(1)})$. Accordingly, it is expected that support vectors to be set with larger values. Therefore, sparse and accurate solution boundary is obtained. To control the sparsity of the solution $m_i^{(1)}$ is chosen as $m_i^{(1)} = -\left(\sum_{i=1}^{N_1} \mathbf{K}_{i,j}^{(1)}\right)^{\nu}$ where ν $(0 < \nu \leq 1)$.

In addition, class 2 observations, $\mathbf{x}_i^{(2)}$, which satisfy $\beta_i^{(2)} > 0$ are also the support vector for class 1. These second kind of support vectors lie on the boundary of class one if $\mathbf{x}_i^{(2)}$ satisfies $0 < \beta_i^{(2)} < B_{(2,1)}$ or inside the boundary of class 1 if $\mathbf{x}_i^{(2)}$ satisfies $\beta_i^{(2)} = B_{(2,1)}$. Thus, we can recognize that the second kind of support vectors are close the class 1 observations rather than class 2 observations. Therefore, we can determine $\mathbf{m}^{(2,1)}$ as follows:

$$m_i^{(2,1)} \propto -\sum_{j \in \mathbf{D}_1} \mathbf{K}_{i,j}^{(2,1)}$$

Thus, $m_i^{(2,1)}$ is chosen as $m_i^{(2,1)} = -\left(\sum_{i=1}^{N_1} \mathbf{K}_{i,j}^{(2,1)}\right)^{v}$ where v $(0 < v \leq 1)$.

5.4.2 Effect of the Penalty Parameters

In this section, we analyze the effects of different penalization and Gaussian kernel parameters as shown in Figure (5.1). Two banana-shaped data are obtained with equal size of sixty three observations (blue and red points for class 1 and class2, respectively). B-2SVDD boundaries are obtained with a Gaussian kernel function with width parameter w. Similar to the traditional SVDD, increase of *w* increases the volume of the hypersphere which decreases the number of support vectors and draws a simple boundary. On the other hand, if the *w* becomes smaller, then the boundaries become more complex, thus it may cause over fitting since every normal observation becomes a support vector. In addition, we can see the effects of C_1 and C_2 from Figure 5.1. Large C_1 (C_2)

increases the volume of the hypersphere 1 (2). This fact can also be explained using Eq. (5.5). Large C_1 or C_2 tends to increase the objective function although it is a minimization problem. To remove the effect of C_1 and C_2, most of the observations are placed inside the boundary by assigning the penalty functions zero (ε_i's). Thus, the volume of the hypersphere increases. On the other hand, small C_1 and C_2 decrease the effect of the penalty. Thus, some of the observations are placed outside of the boundary resulting in a smaller hypersphere.

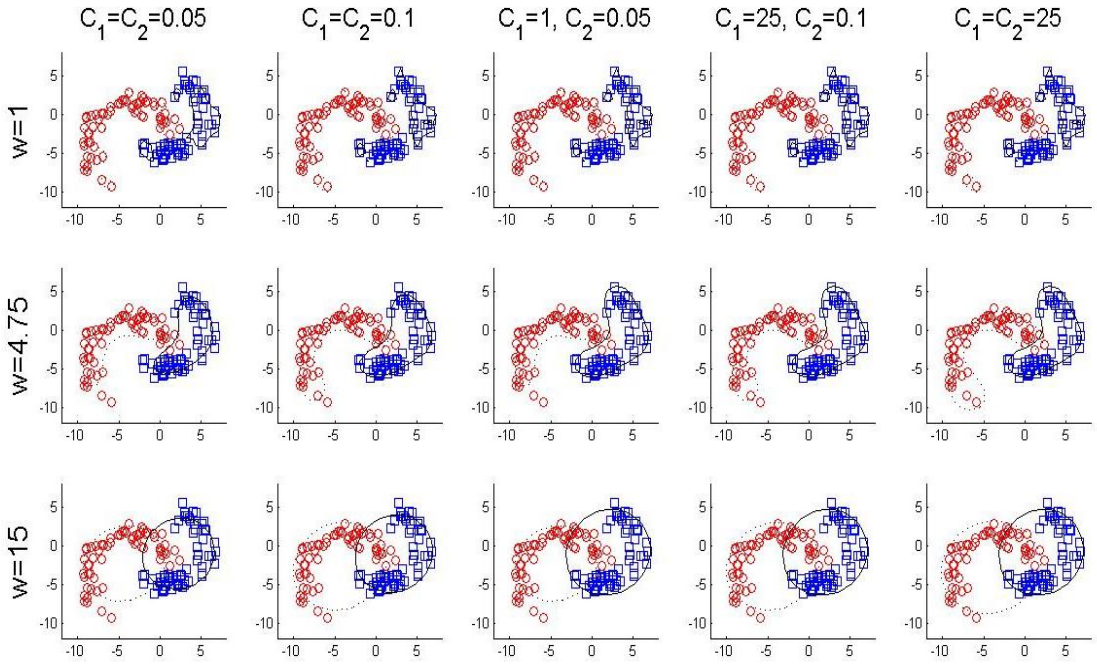


Figure 5.1 Effect of penalty parameters (C_1, C_2) on B-2SVDD

5.5 Simulation Study

In this section, we provide a comparison of the proposed B-2SVDD method with TC-SVDD, traditional SVDD and BSVDD using both an artificial data set and a real data set.

We consider two different performance measures: the total classification accuracy rate (TCAR) is used for the determination of the best parameters with validation dataset and the classification accuracy rate for the testing dataset which is defined as follows:

$$TCAR = \frac{\#of\ true\ classifications\ in\ Class\ 1\ and\ Class\ 2 + \#of\ true\ identifications\ of\ Outliers}{total\ number\ of\ observations}$$

However TCAR is not suitable for the performance comparison of the traditional SVDD and the Bayesian SVDD. Therefore, we introduce another performance measure called overall classification accuracy rate (OCAR) which considers the Class 1 and Class 2 observations as one class.

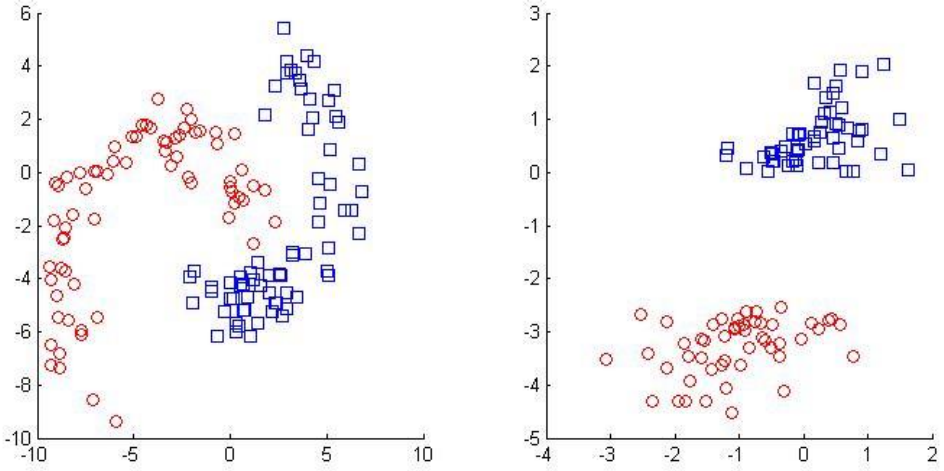
$$OCAR = \frac{\#of\ true\ classification\ for\ combined\ class + \#of\ true\ identification\ in\ Outlier}{total\ number\ of\ observations}$$

In this study, 10-fold cross validation is used for the performance measures for all data sets since available data are not separated as training and testing data. The parameters of the B-2SVDD and benchmarks procedures are optimized based on the intervals in Table 5.2.

Table 5.2 Parameters for each algorithm

Algorithm	Parameters	Candidates
B-2SVDD	Kernel Parameter	2^{-2}, 2^{-1}, 2^{0},1,1.2,1.25, 1.5, 2^{1}, 2^{2}, 5, 5.5, 6, 2^{3}, 10^{1}, 10^{2}
	C1,C2	10^{-3}, 10^{-2}, 0.025, 0.05, 0.075,10^{-1}, 2^{-2}, 2^{-1}, 2^{0} ,5^{2}
	B(1,2), B(2,1)	0.05, 0.05
TC-SVDD	Kernel Parameter	2^{-2}, 2^{-1}, 2^{0}, 1.5, 2^{1}, 2^{2}, 5, 5.5, 6, 2^{3}, 10^{1}, 10^{2}
	C1,C2	10^{-3}, 10^{-2}, 0.025, 0.05, 0.075,10^{-1}, 2^{-2}, 2^{-1}, 2^{0} ,5^{2}
	B(1,2), B(2,1)	0.05, 0.05
SVDD	Kernel Parameter	2^{-2}, 2^{-1}, 2^{0}, 1.5, 2^{1}, 2^{2}, 5, 5.5, 6, 2^{3}, 10^{1}, 10^{2}
	C	10^{-3}, 10^{-2}, 0.025, 0.05, 0.075,10^{-1}, 2^{-2}, 2^{-1}, 2^{0} ,5^{2}
BSVDD	Kernel Parameter	2^{-2}, 2^{-1}, 2^{0}, 1.5, 2^{1}, 2^{2}, 5, 5.5, 6, 2^{3}, 10^{1}, 10^{2}
	C	10^{-3}, 10^{-2}, 0.025, 0.05, 0.075,10^{-1}, 2^{-2}, 2^{-1}, 2^{0} ,5^{2}
	V	0.1 ,0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

In this part, two classes of banana-shaped data in two-dimensions, introduced by Duin *et al.* (2000), is sampled with the size of each class of objects being 70 (Figure 5.2 (a)). Two different multivariate normal data sets are generated with means [5 -5] and [-5 -4] with a size of 50 for each class as outlier data sets. In addition to the banana-shaped data, we use multivariate skew normal (MSN) distribution which takes advantage of general set-up for comparison under non-normal dataset by controlling skewness (Azzalini and Capitanio, 1999, Azzalini and Dalla Valle, 1996). Two different target data are obtained from multivariate skew normal data with size 100 for each of the classes (Figure 5.2 (b)) as well as the abnormal data which are generated from the three normal distributions and placed between the normal classes with the size of each abnormal class being 30.



(a) (b)

Figure 5.2 (a) Banana-shaped data (b) MSN data

Table 5.3 illustrates the performance of B-2SVDD against other benchmark procedures under MSN and banana-shaped data with two dimensional datasets. The results show that B-2SVDD outperforms the other support vector data description methods of traditional SVDD, BSVDD and TC-SVDD.

Table 5.3 Performance comparisons of SVDD, BSVDD, TC-SVDD and B-2SVDD under two-dimensional MSN and banana-shaped data

Data	MSN		Banana-Shaped	
Performance	OCAR	TCAR	OCAR	TCAR
SVDD	0.9870	NA	0.9614	NA
BSVDD	0.9870	NA	0.9623	NA
TC-SVDD	0.9870	0.9870	0.9649	0.9640
B-2SVDD	0.9890	0.9890	0.9675	0.9649

Table 5.4 demonstrates the performances of the proposed and the existing procedures under the MSN distribution with nonzero skew parameters. Two classes of MSN data with the sizes 100 in five-dimensions are obtained by choosing the skewness parameters as $\mathbf{d}$=(-2, 1, -2, 1,-2) and $\mathbf{d}$= (2, -1, 2, -1, 2) . By choosing this setting in which skewness parameters can be considered as symmetric, it is possible to observe how the proposed procedure responds to positive and negative skewness. Two hundred outliers are obtained by shifting the mean of the original mean. From Table 5.4, we show that the B-2SVDD outperforms the existing procedures.

Table 5.4 Performance comparisons of SVDD, BSVDD, TC-SVDD and B-2SVDD under five-dimensional MSN data

Data	MSN	
Performance	OCAR	TCAR
SVDD	0.9395	NA
BSVDD	0.9400	NA
TC-SVDD	0.9445	0.9409
B-2SVDD	0.9482	0.9414

5.6 Case Study: Continuous Stirred Tank Heater (CSTH)

The stirred tank heater plant is used in the Department of Chemical and Materials Engineering at the University of Alberta (Thornhill *et al.*, 2008). In this plant, hot and cold water are mixed and heated using steam through a heating coil. The mixed water is drained from the tank through a long pipe as shown in Figure 5.3. Hot and cold water are

well stirred and mixed in the tank by assuming the temperature in the tank to be the same as the outflow temperature and keeping the volume of the water level in the tank as the desired objective. The tank has a circular cross section with a volume of 8l cm^3 and height of 50 cm.

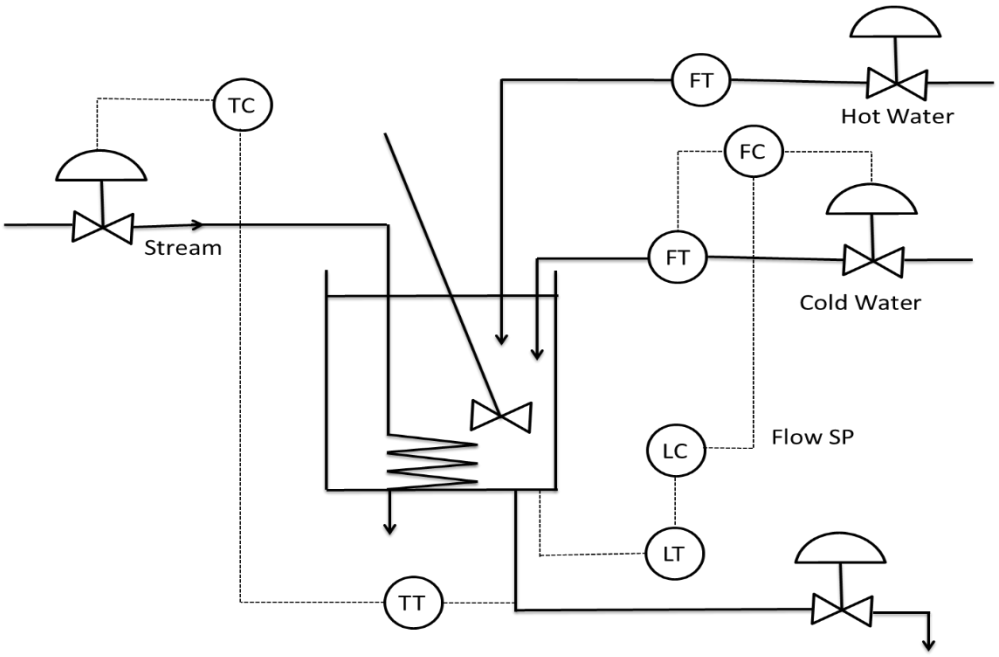


Figure 5.3 Continuous Stirred Tank Heater (Thornhill *et al.*, 2008)

The steam temperature (TC), the thermocouple temperature (TT), the cold water flow (FC), the heating flow (FT), the level controller (LC), and the heating level (LT) are the instruments used for operation of the plant. The cold and hot water (CW and HW) enters the plant with 60–80 psi, and the hot water boiler is heated by the steam supply. Control valves in the CSTH plant are controlled by the air pressure using 3–15 psi compressed air

supply. Flow instruments show signals with a nominal 4–20 mA output. The signals of level instruments are also shown using the same scale, mA.

The inputs of the process are the CW, HW and steam valve demands, and the outputs consist of the electronic measurements from the level, cold and hot water flow and temperature instruments. Disturbances of the experiment include a deterministic oscillatory disturbance to the cold water flow rate, a random disturbance to the water level, and the temperature measurement noise.

The aim of this study is to determine the dynamic responses of the outputs for specified inputs. For a given input values, it is desired to keep the volume and the outflow temperature in steady conditions. In other words, the system is considered in normal operating condition if the desired volume and outflow temperature are obtained by adjusted input values. It is not always possible to have the desired volume and outflow temperature even when the input values are adjusted. Therefore, if the desired values are obtained for the adjusted input values, then the process operates in normal conditions. This kind of operating conditions is called a normal operating mode. Two operating conditions of CSTH are shown in Table 5.5.

Table 5.5 Operating conditions of CSTH plant

Input Variables	Mode 1	Mode 2
Desired Volume	12	12
Desired Temperature	10.5	10.5
HW Valve	0	5.5

The CSTH plant achieves the steady state conditions where the desired variables are satisfied if the adjustments are chosen as one of the modes in Table 5.5.

In this study, we obtain the data from three controlled variables, water level, temperature, and CW flow. Two modes of CSTH are obtained under normal operating modes with the size of each mode of objects being 100. In addition, abnormal observations are obtained from each mode with sizes of 100 by introducing sudden step change of -0.75 into the level measurement at steps 101 to 200. The plots of the normal and abnormal data are given in Figure 5.4.

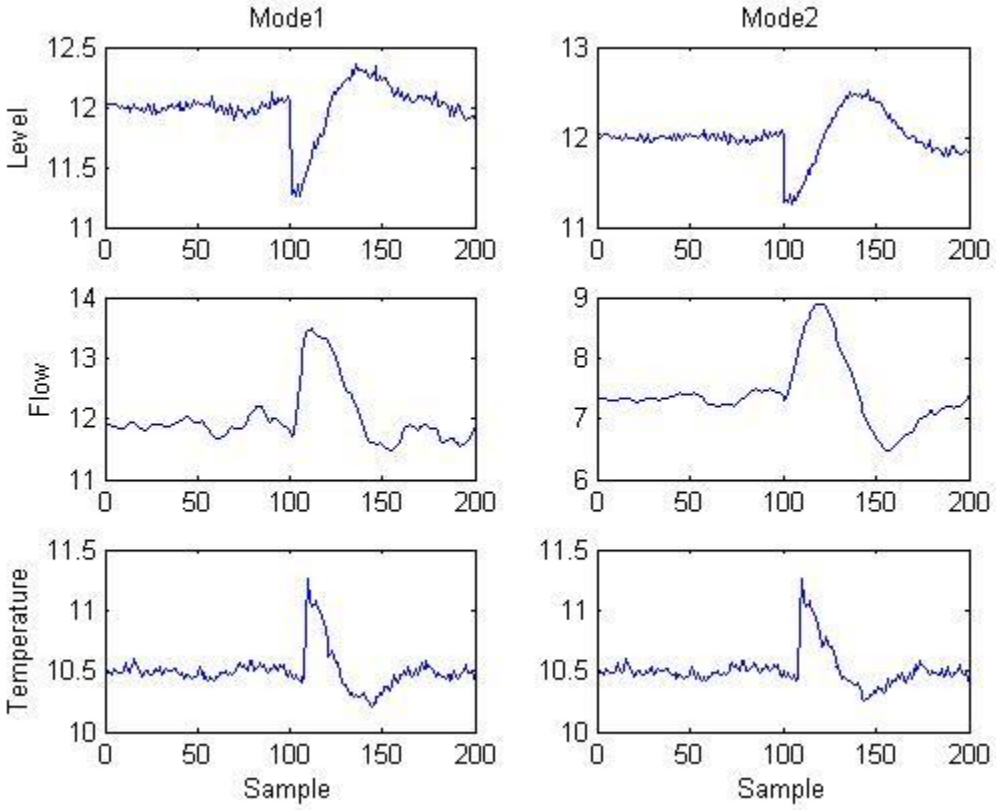


Figure 5.4 CSTH normal and abnormal data

By using the water level, CW flow and temperature, we obtain the three-dimensional vector for each observation, that is, total 400 vectors, 100 normal and 100 abnormal for each modes. The plot of normal observations is shown in Figure 5.5.

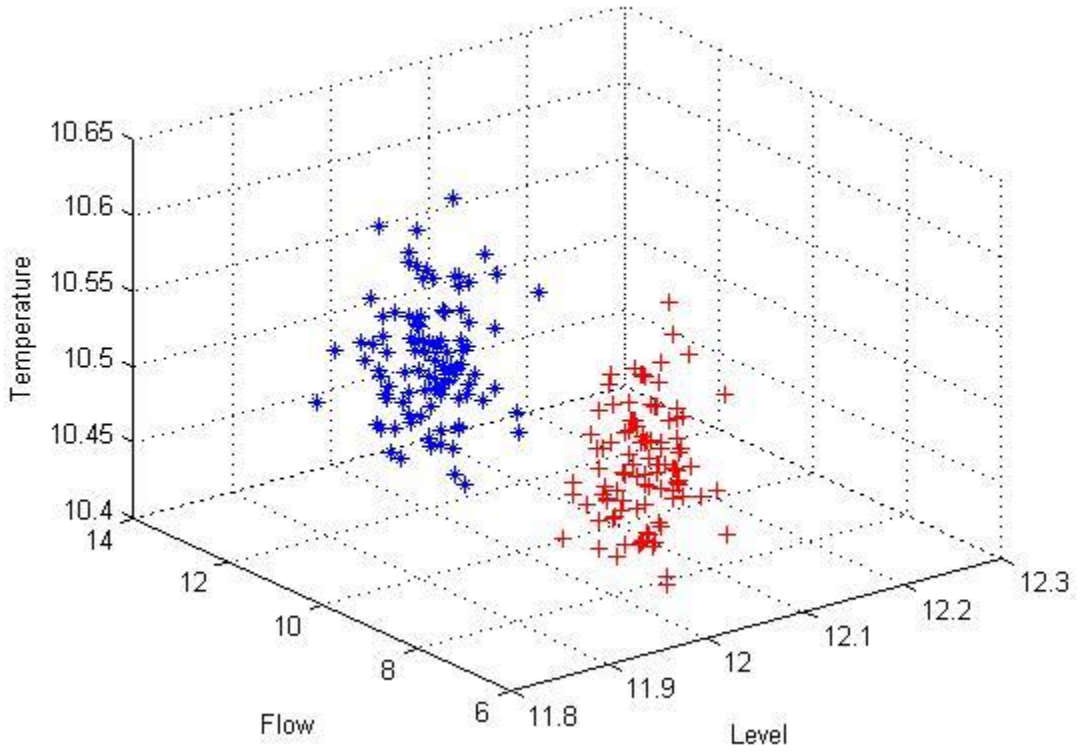


Figure 5.5 Mode 1 (*) and Mode 2 (+)

By using the three-dimensional observations, the boundaries are obtained for each SVDD procedure. Table 5.6 shows the results obtained from each of the SVDD approaches. As shown in Figure 5.5, these data are not complicated as the banana-shaped data set and the modes are well separated. The procedures identify the normal observations correctly. Therefore, two performance measures provide the same conclusion. The results show that B-2SVDD outperforms the existing procedures.

Table 5.6 Performance comparisons of TC-SVDD and B-2SVDD for CSTH data

Data	CSTH	
Performance	OCAR	TCAR
TC-SVDD	0.8982	0.8982
B-2SVDD	0.9309	0.9309

5.7 Conclusions

In real-life problems, identifying the anomalies is important due to the fact that they may have significant information. Several anomaly detection procedures are introduced to identify anomalies in data. Among them, the support vector data description (SVDD) procedure has gained more attention. Most of the SVDD procedures are built assuming that the normal data have only one or two-classes. However, this is not a general case if the data have more than two-classes.

In this chapter, we propose a generalized SVDD procedure called *n*-SVDD which is independent of the number of classes. The proposed procedure finds *n* hyperspheres. Each hypersphere keeps as many as observations of the same class inside the boundary and tries to keep other class's observations outside the hypersphere. In addition, we propose a Bayesian SVDD procedure by assuming that the normal data consist of two-classes. B-*n*SVDD is built assuming the transformed variables and the prior distributions follow normal distribution.

Experiments with diverse data sets show that the B-*n*SVDD is superior to the existing SVDD procedures. In future research, this work can be extended to determine the appropriate prior distributions which have a direct effect on the performance of the B-*n*SVDD procedure. This can also be extended to cases where the boundaries among classes are not “crisp” and the sizes of the classes are significantly different.

CHAPTER 6

MULTI-CLASS BAYESIAN SUPPORT VECTOR DATA DESCRIPTION WITH ANOMALIES

6.1 Introduction

In Chapter 5, the anomaly detection procedure based on SVDD in Bayesian framework is developed. It is designed to obtain the boundary that separates the anomaly data from the normal data. This chapter proposes a multi-class Bayesian SVDD model that takes anomaly data into consideration when the anomaly data are available and an appropriate prior distribution of the anomaly data is obtained.

In real-life applications, having information about the anomaly data is more desirable than knowing only normal patterns. For instance, anomalies in credit card transaction data may indicate a stolen credit card or identity theft (Aleskerov *et al.*, 1997). Anomalies in MRI images may reveal tumor formations (Spence *et al.*, 2001). An anomalous pattern in a computer network may imply that computer is hacked and sends out sensitive data to an unknown destination (Kumar, 2005). Anomaly information which includes critical and significant indicators can be used effectively to fix or improve applications (Chandola *et al.*, 2009). Several attempts have been made to develop anomaly detection procedures based on classification, nearest neighbor, clustering, and statistical procedures (Chandola *et al.*, 2009, Hodge and Austin, 2004).

Among the anomaly detection techniques, significant attention has been focused on the provision of support vector machine (SVM) proposed by Vapnik (1995). SVM transforms data into the kernel space and detects anomalies by separating other objects from the origin with a hyperplane (Amer *et al.*, 2013, Erfani *et al.*, 2016, Li *et al.*, 2003, Schölkopf *et al.*, 2001, Sotiris *et al.*, 2010). Another prominent technique is one-class classification procedures, a particular case of the classification-based procedures (Moya *et al.*, 1993). One of the one-class classification procedure is called support vector data description (SVDD), proposed by Tax and Duin (1999). SVDD identifies a given pattern as anomalous if it is placed outside of the decision boundary. SVDD procedures are widely used in the literature and several versions of SVDD have been developed (Bovolo *et al.*, 2010, Ghasemi *et al.*, 2016, Kang and Cho, 2012, Lee and Lee, 2007, Lee *et al.*, 2005, Lee *et al.*, 2007, Lee *et al.*, 2006, Ning and Tsung, 2013).

The SVDD procedures for anomaly detection are based on the assumption that the target data have only one class of normal data. However, many real-life data consist of more than one distribution. To consider the differences between the classes, Huang *et al.* (2011) introduce an anomaly detection procedure by assuming normal data consisting of two-classes called two-class SVDD (TC-SVDD).

The procedures mentioned above are only based on normal data. However, the performance of these procedures can be improved when available anomaly data are utilized to obtain data description. Tax and Duin (2004) introduce a procedure that considers anomaly observations. This procedure introduces extra constraints based on

anomaly observations to the traditional SVDD and it incorporates the anomaly observations into the training procedure to improve the boundary description. Thus, normal data are placed within a sphere and anomaly data are placed outside it. This procedure is based on the L1-norm of slack variables with two regularization parameters. Mu and Nandi (2009) improve this procedure in two ways by involving different forms of slack vectors. The first extension includes the L2-norm for slack variables with two regularization parameters and the second extension introduces only one regularization parameter. However, these procedures assume that the normal data are obtained from one distribution.

This chapter introduces a new Bayesian procedure of anomaly detection in multi-class data where the prior distribution of the anomaly is known. The proposed procedure simultaneously finds the hyperspheres for each class by including as many of its class observations as possible and keeps the other class observations and anomalies outside the boundary. To construct the Bayesian framework, a probabilistic behavior of the parameters is considered by taking 'prior knowledge' of each class as well as the anomaly distribution into account.

The proposed procedure differs from the existing procedures as follows: First, the existing procedures for anomaly detection mostly take only the normal information into account (Moya *et al.*, 1993). However, the anomaly information is often available from the engineering knowledge and the historical data of the process. Thus, the proposed procedure describes the multi-class normal data more accurately by considering the

anomalies. Second, the existing deterministic procedures do not reflect the prior or domain-specific knowledge when they are applied to the real-world problems that are not considered to be in the same domain. On the other hand, the prior information of the anomaly can be effectively applied through the Bayesian framework leading to more precise data description. It is expected that a more accurate boundary for the normal data can be obtained by considering the location of the anomaly data. For example, if we have knowledge of the anomaly locations, the proposed procedure attempts to obtain a boundary of normal data in a way to avoid the anomaly data. In addition, even when the anomaly data are placed in unexpected locations, the information of those anomalies are helpful to better describe the normal data because using the information of anomalies makes the boundary of the normal data much tighter.

This chapter is organized as follows. After the review of the existing procedures in Section 6.2, we propose a new Bayesian procedure of anomaly detection in multi-class data environment in Section 6.3. In Section 6.4, we conduct simulation studies for performance comparison with existing procedures. In Section 6.5, we illustrate the proposed procedure in a real-life case study of a continuous stirred tank heater (CSTH), followed by conclusions and future research directions in Section 6.6.

6.2 Preliminaries

Given data set $\mathbf{D} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{R}^p, i = 1, ..., N\}$ where $\mathbf{x}_i$ is observation i, the SVDD procedure finds a hypersphere by covering the data with minimal volume, with a center

$\mathbf{a}$ and a radius R (Tax and Duin, 1999). By allowing the misclassification in $\mathbf{D}$, some observations can be kept outside the hypersphere. The SVDD is formulated as follows:

$$\begin{aligned} &\min \quad R^2 + C\sum_{i=1}^{N}\varepsilon_i \\ &s.t. \quad \left\|\mathbf{x}_i - \mathbf{a}\right\|^2 \le R^2 + \varepsilon_i, \quad \varepsilon_i \ge 0 \ \ \forall i \in \{1,...,N\} \end{aligned} \tag{6.1}$$

where C is the regularization parameter, which controls the volume of the hypersphere by keeping some of the observations outside the boundary. ε_i is the slack variable and N is the number of observations in the data set. Dual formulation is obtained by using the Lagrangian function:

$$\begin{aligned} &\max \ \sum_{i=1}^{N}\tau_i\left(\mathbf{x}_i \cdot \mathbf{x}_i\right) - \sum_{i=1}^{N}\sum_{j=1}^{N}\tau_i\tau_j\left(\mathbf{x}_i \cdot \mathbf{x}_j\right) \\ &s.t. \quad \sum_{i=1}^{N}\tau_i = 1, \ 0 \le \tau_i \le C \quad \forall i \in \{1,...,N\} \end{aligned} \tag{6.2}$$

The observations with positive τ_i are called support vectors and are placed on the boundary or outside the boundary. SVDD's decision rule is based on checking the distance from an observation to the center. A new observation $\mathbf{z}$ can be classified as "in" or "out" of the data boundary based on the distance from $\mathbf{z}$ to the center of the hypersphere. The inner product in Eq. (6.2) can be replaced by a kernel function $K(\mathbf{x}_i, \mathbf{x}_i)$, defining $K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \right\rangle$, and thus a more suitable nonlinear boundary to cover the data can be obtained (Tax and Duin, 1999).

6.3 Multi-Class SVDD with Anomaly Observations

In some of the real life applications, available data consist of more than one class along with a few anomaly data. In such situations, utilizing the existing SVDD techniques may not provide a good boundary representation for the available data. Therefore, this section introduces a new SVDD procedure called *n*-class SVDD with anomalies (*n*SVDD-A).

The proposed *n*SVDD-A assumes that the target data set consists of *n* normal classes $\mathbf{D}_1 = \left\{\mathbf{x}_1^{(1)},\ldots,\mathbf{x}_{N_1}^{(1)}\right\}, \mathbf{D}_2 = \left\{\mathbf{x}_1^{(2)},\ldots,\mathbf{x}_{N_2}^{(2)}\right\},\ldots,\mathbf{D}_n = \left\{\mathbf{x}_1^{(n)},\ldots,\mathbf{x}_{N_n}^{(n)}\right\}$ and a few anomaly data $\mathbf{D}_A = \left\{\mathbf{x}_1^{(A)},\ldots,\mathbf{x}_{N_A}^{(A)}\right\}$. By utilizing the anomaly data, the goal of the *n*SVDD-A for given *n* classes is to find *n* hyperspheres which cover each class with minimal volume, with centers $\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n$ and radii $R_1, R_2,\ldots, R_n$. The primal formulation of the *n*SVDD-A is constructed as follows:

$$
\begin{aligned}
&\min \textstyle\sum_{k=1}^{n} R_k^2 \\
&s.t.\ \left\|\mathbf{x}_i^{(k)} - \mathbf{a}_k\right\|^2 \le R_k^2 \ \ k = 1,\ldots,n\ ,\ i = 1,\ldots N_k \\
&\qquad \left\|\mathbf{x}_i^{(m)} - \mathbf{a}_k\right\|^2 \ge R_k^2 \ \ m \ne k,\ \ m,k = 1,\ldots,n,\ i = 1,\ldots,N_m \\
&\qquad \left\|\mathbf{x}_i^{(A)} - \mathbf{a}_k\right\|^2 \ge R_k^2 \ \ k = 1,\ldots,n,\ i = 1,\ldots,N_A
\end{aligned}
\tag{6.3}
$$

where N_k is the size of the class k ($k = 1,\ldots,n$) and N_A is the size of the anomaly data.

With the optimization problem above, each observation in class *n* is placed inside the hypersphere *n* (constraint 1) and falls outside the other hyperspheres (constraint 2). In

addition, all of the anomalies are placed outside of each hypersphere (constraint 3). However, by introducing penalties $\varepsilon_i^{(k)}$ and $\varsigma_i^{(m,k)}$, some observations can be placed outside of its class's boundary or inside of the other classes' boundary. In addition, to allow some anomalies to be placed inside the one of the hyperspheres, we introduce the O penalty. Then, the primal formulation of the proposed optimization problem is obtained as follows:

$$
\begin{aligned}
&\min \sum_{k=1}^{n} R_k^2 + \sum_{k=1}^{n}\left(C_k \sum_{i=1}^{N_k} \varepsilon_i^{(k)} + O\sum_{i=1}^{N_A} \varsigma_i^{(A_k)}\right) + \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\left(B\sum_{i=1}^{N_m} \varsigma_i^{(m,k)}\right) \\
&s.t.\ \left\|\mathbf{x}_i^{(k)} - \mathbf{a}_k\right\|^2 \le R_k^2 + \varepsilon_i^{(k)} \quad k = 1,\ldots,n\ ,\ i = 1,\ldots N_k \\
&\quad \left\|\mathbf{x}_i^{(m)} - \mathbf{a}_k\right\|^2 \ge R_k^2 - \varsigma_i^{(m,k)} \ \ m \neq k,\ \ m,k = 1,\ldots,n\ ,\ \ i = 1,\ldots,N_m \\
&\quad \left\|\mathbf{x}_i^{(A)} - \mathbf{a}_k\right\|^2 \ge R_k^2 - \varsigma_i^{(A_k)} \ \ k = 1,\ldots,n\ ,\ \ i = 1,\ldots,N_A \\
&\quad \varepsilon_i^{(k)}, \zeta_i^{(m,k)}, \varsigma_i^{(A_k)} \ge 0\ \ \forall i,k,m
\end{aligned}
\tag{6.4}
$$

where C_k, B and O are called regularization parameters which control the volume of the hyperspheres. The first two constraints ensure that if an observation $\mathbf{x}_i^{(k)}$ falls outside the hypersphere *k*, the objective function increases by $C_k\varepsilon_i^{(k)}$ and if an observation $\mathbf{x}_i^{(m)}$ falls inside the hypersphere *k*, the objective function increases by $B\varsigma_i^{(m,k)}$. In addition, if the anomaly observation $\mathbf{x}_i^{(A)}$ is placed inside the hypersphere *k*, the objective function increases by $O\varsigma_i^{(A_k)}$. Table 6.1 shows the notations used to obtain *n*SVDD-A.

Table.6.1 Notations of *n*SVDD-A

R_k	: radius of class k , (k=1,..,n)
$\mathbf{a}_k$	: center of class k , (k=1,..,n)
$\varepsilon_i^{(k)}$	: penalty given to training sample i of class k which lies outside hypersphere k.
$\zeta_i^{(m,k)}$	: penalty given to training sample i of class m which lies inside hypersphere k.
$\zeta_i^{(A_k)}$	: penalty given to anomaly sample i which lies inside hypersphere k.

By introducing the Lagrangian function, dual formulation of *n*SVDD-A is obtained.

$$
\begin{aligned}
L\left(R_n, \mathbf{a}_n, \varepsilon_i^{(k)}, \zeta_i^{(m,k)}\right) = & \sum_{k=1}^{n} R_k^2 + \sum_{k=1}^{n}\left(C_k \sum_{i=1}^{N_k} \varepsilon_i^{(k)} + O\sum_{i=1}^{N_A} \zeta_i^{(A_k)}\right) \\
& + \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\left(B\sum_{i=1}^{N_m} \zeta_i^{(m,k)}\right) \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_k} a_i^{(k)}\left[R_k^2 + \varepsilon_i^{(k)} - \left(\mathbf{x}_i^{(k)} - \mathbf{a}_k\right)^2\right] \\
& - \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sum_{i=1}^{N_m} \beta_i^{(m,k)}\left[\left(\mathbf{x}_i^{(m)} - \mathbf{a}_k\right)^2 - R_k^2 + \zeta_i^{(m,k)}\right] \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_A} \tau_i^{(A_k)}\left[\left(\mathbf{x}_i^{(A)} - \mathbf{a}_k\right)^2 - R_k^2 + \zeta_i^{(A_k)}\right] \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_k} \gamma_i^{(k)}\varepsilon_i^{(k)} - \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sum_{i=1}^{N_m} \gamma_i^{(m,k)}\zeta_i^{(m,k)} \\
& - \sum_{k=1}^{n}\sum_{i=1}^{N_0} \gamma_i^{(A_k)}\zeta_i^{(A_k)}
\end{aligned}
\tag{6.5}
$$

where $\alpha_i^{(k)} \geq 0, \beta_i^{(m,k)} \geq 0, \gamma_i^{(k)} \geq 0$, $\gamma_i^{(m,k)} \geq 0$ and $\gamma_i^{(A_k)} \geq 0$. By taking partial derivatives of the Lagrangian function $L\left(R_n, \mathbf{a}_n, \varepsilon_i^{(k)}, \zeta_i^{(m,k)}, \zeta_i^{(A_k)}\right)$, the following equations are obtained.

$$
\begin{aligned}
&\frac{\partial L}{\partial R_k} = 0 \Rightarrow \sum_{i=1}^{N_k} \alpha_i^{(k)} - \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} - \sum_{i=1}^{N_A} \tau_i^{(A_k)} = 1 \\
&\frac{\partial L}{\partial \mathbf{a}_k} = 0 \Rightarrow \mathbf{a}_k = \sum_{i=1}^{N_k} \alpha_i^{(k)} \mathbf{x}_i^{(k)} - \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} \mathbf{x}_i^{(m)} - \sum_{i=1}^{N_A} \tau_i^{(A_k)} \mathbf{x}_i^{(A)} \\
&\frac{\partial L}{\partial \varepsilon_i^{(k)}} = 0 \Rightarrow C_k - \gamma_i^{(k)} - \alpha_i^{(k)} = 0 \Rightarrow 0 \le \alpha_i^{(k)} \le C_k \\
&\frac{\partial L}{\partial \zeta_i^{(m,k)}} = 0 \Rightarrow B - \gamma_i^{(m,k)} - \beta_i^{(m,k)} = 0 \Rightarrow 0 \le \beta_i^{(m,k)} \le B \\
&\frac{\partial L}{\partial \zeta_i^{(A_k)}} = 0 \Rightarrow O - \gamma_i^{(A_k)} - \tau_i^{(A_k)} = 0 \Rightarrow 0 \le \tau_i^{(A_k)} \le O \\
&\forall m, k
\end{aligned}
\tag{6.6}
$$

Therefore, by substituting Eq. (6.6) into Eq. (6.5), the dual formulation in Eq. (6.7) is obtained.

$$
\begin{aligned}
&\max \sum_{m=1}^{n} \sum_{i=1}^{N_m} \alpha_i^{(m)} \left(\mathbf{x}_i^{(m)} \cdot \mathbf{x}_i^{(m)} \right) - \sum_{m=1}^{n} \sum_{m \neq k=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} \left(\mathbf{x}_i^{(m)} \cdot \mathbf{x}_i^{(m)} \right) - \sum_{m=1}^{n} \sum_{i=1}^{N_m} \tau_i^{(A_m)} \left(\mathbf{x}_i^{(A)} \cdot \mathbf{x}_i^{(A)} \right) \\
&\quad - \sum_{k=1}^{n} \left(\sum_{i=1}^{N_k} \alpha_i^{(k)} \mathbf{x}_i^{(k)} - \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} \mathbf{x}_i^{(m)} - \sum_{i=1}^{N_A} \tau_i^{(A_k)} \mathbf{x}_i^{(A)} \right)^2 \\
&\textit{s.t.} \ \sum_{i=1}^{N_k} \alpha_i^{(k)} - \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} - \sum_{i=1}^{N_A} \tau_i^{(A_k)} = 1, \ \ \forall k = 1, \ldots, n \\
&\quad 0 \le \alpha_i^{(k)} \le C_k, \ \ \forall i = 1, \ldots, N_k, \forall k = 1, \ldots, n \\
&\quad 0 \le \beta_i^{(m,k)} \le B, \ \forall m, k, \ \forall i = 1, \ldots, N_m \\
&\quad 0 \le \tau_i^{(A_k)} \le O, \ \ \forall i = 1, \ldots, N_A, \forall k = 1, \ldots, n
\end{aligned}
\tag{6.7}
$$

The vectors $\mathbf{x}_i^{(k)}$ with $\alpha_i^{(k)} > 0$, $\mathbf{x}_i^{(m)}$ with $\beta_i^{(m,k)} > 0$ $(m \neq k)$ and $\mathbf{x}_i^{(A)}$ with $\tau_i^{(A_k)} > 0$ are called support vectors for class *k*. For an in-control observation $\mathbf{x}_i^{(k)}$, if $\alpha_i^{(k)} = C_k$ ($0 < \alpha_i^{(k)} < C_k$), then $\mathbf{x}_i^{(k)}$ is outside the hypersphere *k* (on the boundary of hypersphere *k*). If $\beta_i^{(m,k)} = B$ ($0 < \beta_i^{(m,k)} < B$), $\mathbf{x}_i^{(m)}$ is inside the hypersphere *k* (on the boundary of

hypersphere *k*). In addition, for an anomaly observation $\mathbf{x}_i^{(A)}$, if $\tau_i^{(A_k)} = O$ $(0 < \tau_i^{(A_k)} < O)$, $\mathbf{x}_i^{(A)}$ is placed inside hypersphere *k* (on the boundary of hypersphere *k*).

To find the radius of each hypersphere, R_k, we calculate the distance from the center $\mathbf{a}_k$ of the hypersphere to any of the support vectors of class *k* except the support vectors with $\alpha_i^{(k)} = C_k$, $\beta_i^{(m,k)} = B$ and $\tau_i^{(A_k)} = O$. Let $\mathbf{x}_s$ be a support vector of class *k* with one of the following conditions; $0 < \alpha_s^{(k)} < C_k$, $0 < \beta_i^{(m,k)} < B$ and $0 < \tau_i^{(A_k)} < O$. Then,

$$
\begin{aligned}
R_k^2 &= \left\| \mathbf{x}_s - \mathbf{a}_k \right\|^2 \quad , \forall k = 1, \ldots, n \\
&= \left(\mathbf{x}_s \cdot \mathbf{x}_s \right) - 2 \sum_{i=1}^{N_k} \alpha_i^{(k)} \left(\mathbf{x}_s \cdot \mathbf{x}_i^{(k)} \right) + 2 \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \beta_i^{(m,k)} \left(\mathbf{x}_s \cdot \mathbf{x}_i^{(m)} \right) + 2 \sum_{i=1}^{N_A} \tau_i^{(A_k)} \left(\mathbf{x}_s \cdot \mathbf{x}_i^{(A)} \right) \\
&+ \sum_{i=1}^{N_k} \sum_{i=1}^{N_k} \alpha_i^{(k)} \alpha_j^{(k)} \left(\mathbf{x}_i^{(k)} \cdot \mathbf{x}_j^{(k)} \right) + \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \sum_{k \neq t=1}^{n} \sum_{j=1}^{N_t} \beta_i^{(m,k)} \beta_j^{(t,k)} \left(\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(t)} \right) \\
&- 2 \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \sum_{j=1}^{N_k} \beta_i^{(m,k)} \alpha_j^{(k)} \left(\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(k)} \right) - 2 \sum_{i=1}^{N_k} \sum_{j=1}^{N_A} \alpha_i^{(k)} \tau_i^{(A_k)} \left(\mathbf{x}_s \cdot \mathbf{x}_j^{(A)} \right) \\
&+ 2 \sum_{k \neq m=1}^{n} \sum_{i=1}^{N_m} \sum_{j=1}^{N_A} \beta_i^{(m,k)} \tau_j^{(A_k)} \left(\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(A)} \right) + \sum_{i=1}^{N_A} \sum_{i=1}^{N_A} \tau_i^{(A_k)} \tau_j^{(A_k)} \left(\mathbf{x}_i^{(A)} \cdot \mathbf{x}_j^{(A)} \right)
\end{aligned}
$$

The new observation $\mathbf{z}$ is called an anomaly if the following Eq. (6.8) is satisfied, i.e., the distance from observation $\mathbf{z}$ to the center of each hypersphere, $\mathbf{a}_k$ is greater than the corresponding radius R_k

$$
\left\| \mathbf{z} - \mathbf{a}_k \right\|^2 > R_k^2 \quad \forall k = 1, \ldots, n. \tag{6.8}
$$

To obtain flexible boundaries, the inner products, $(\mathbf{x}_i \cdot \mathbf{x}_j)$ can be replaced by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, defining $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. In this chapter, we use Gaussian kernel function defined as:

$$K(\mathbf{x}, \mathbf{y}) = exp\left(\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2w^2} \right)$$

where w is a parameter of the Gaussian kernel.

The optimization problem of *n*SVDD-A can be represented in a matrix form as follows:

$$\begin{aligned}
&\min \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{A} \\
&s.t \ \ \mathbf{S}\boldsymbol{\theta} = \mathbf{1} \\
&\quad 0 \le \alpha_i^{(k)} \le C_k, \quad \forall i = 1, \ldots, N_k, \forall k = 1, \ldots, n \\
&\quad 0 \le \beta_i^{(m,k)} \le B, \ \forall m, k, \ \forall i = 1, \ldots, N_m \\
&\quad 0 \le \tau_i^{(A_k)} \le O, \quad \forall i = 1, \ldots, N_A, \forall k = 1, \ldots, n
\end{aligned}$$

with $\mathbf{H} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} & -\mathbf{Z}^T \\ \mathbf{Y} & \mathbf{X} & \mathbf{Z}^T \\ -\mathbf{Z} & \mathbf{Z} & \mathbf{T} \end{pmatrix}$, $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^{(1)} \ldots \boldsymbol{\alpha}^{(n)} \boldsymbol{\beta}^{(1,2)} \ldots \boldsymbol{\beta}^{(n,n-1)} \boldsymbol{\tau}^{(A_1)} \ldots \boldsymbol{\tau}^{(A_n)} \right)^T$, $\mathbf{S} = \left(\mathbf{I} \,|\, \hat{\mathbf{I}} \,|\, \mathbf{I}_A \right)$, and

$$\mathbf{A} = \left(-diag(\mathbf{K}^{(1)}) \ \cdots \ -diag(\mathbf{K}^{(n)}) \ \ diag(\mathbf{K}^{(1)}) \ \cdots \ diag(\mathbf{K}^{(n)}) \ \ diag(\mathbf{K}^{(A)}) \ \cdots \ diag(\mathbf{K}^{(A)}) \right)^T.$$

In this formulation, $\boldsymbol{\alpha}^{(k)} = \left(\alpha_1^{(k)} \ldots \alpha_{N_k}^{(k)} \right)$, $\boldsymbol{\beta}^{(m,k)} = \left(\beta_1^{(m,k)} \ldots \beta_{N_m}^{(m,k)} \right)$, $\boldsymbol{\tau}^{(A_k)} = \left(\tau_1^{(A_k)} \ldots \tau_{N_A}^{(A_k)} \right)$

$$\mathbf{I}=\begin{bmatrix}\mathbf{e}_1^T & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{e}_2^T & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e}_n^T\end{bmatrix},\ \hat{\mathbf{I}}=\begin{bmatrix}\mathbf{0} & -\mathbf{e}_2^T & \cdots & -\mathbf{e}_n^T\\ -\mathbf{e}_1^T & \mathbf{0} & \cdots & -\mathbf{e}_n^T\\ \vdots & \vdots & \ddots & \vdots\\ -\mathbf{e}_1^T & -\mathbf{e}_2^T & \cdots & \mathbf{0}\end{bmatrix},\ \mathbf{I}_A=\begin{bmatrix}-\mathbf{e}_A^T & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & -\mathbf{e}_A^T & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{e}_A^T\end{bmatrix},$$

$$\mathbf{X}=\begin{bmatrix}\mathbf{K}^{(1)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^{(2)} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}^{(n)}\end{bmatrix},\ \mathbf{Y}=\begin{bmatrix}\mathbf{0} & -\mathbf{K}^{(1,2)} & \cdots & -\mathbf{K}^{(1,n)}\\ -\mathbf{K}^{(2,1)} & \mathbf{0} & \cdots & -\mathbf{K}^{(2,n)}\\ \vdots & \vdots & \ddots & \vdots\\ -\mathbf{K}^{(n,1)} & -\mathbf{K}^{(n,2)} & \cdots & \mathbf{0}\end{bmatrix}$$

$$\mathbf{Z}=\begin{bmatrix}\mathbf{K}^{(A,1)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^{(A,2)} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}^{(A,n)}\end{bmatrix}\ \mathbf{T}=\begin{bmatrix}\mathbf{K}^{(A)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^{(A)} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}^{(A)}\end{bmatrix}$$

where $\mathbf{e}_k$ is a column vector with size $N_k \times 1$ where $k=1,2,\ldots,n,A$ and $\mathbf{0}$ is a matrix of zeros. $\mathbf{K}^{(k)}$ and $\mathbf{K}^{(k,m)}$ are matrices with sizes $N_k \times N_k$ and $N_k \times N_m$, respectively. Each element of $\mathbf{K}^{(k)}$ and $\mathbf{K}^{(k,m)}$ are obtained by the kernel distances of each element in class *k* and *m*.

$$\mathbf{K}_{ij}^{(k)}=K\left(\mathbf{x}_i^{(k)},\mathbf{y}_j^{(k)}\right)=\left\langle\phi\left(\mathbf{x}_i^{(k)}\right),\phi\left(\mathbf{y}_j^{(k)}\right)\right\rangle\quad\forall\mathbf{x}_i^{(k)},\mathbf{y}_j^{(k)}\in\mathbf{D}_k$$

$$\mathbf{K}_{ij}^{(m,k)}=K\left(\mathbf{x}_i^{(m)},\mathbf{y}_j^{(k)}\right)=\left\langle\phi\left(\mathbf{x}_i^{(m)}\right),\phi\left(\mathbf{y}_j^{(k)}\right)\right\rangle,\forall\,\mathbf{x}_i^{(m)}\in\mathbf{D}_m,\ \forall\mathbf{y}_j^{(k)}\in\mathbf{D}_k\ \ \forall m\neq k=1,2,\ldots,n,A$$

6.3.1 Bayesian Framework of Multi-Class SVDD with Anomaly Observations

This section introduces the Bayesian framework for generalized *n*SVDD-A. In traditional SVDD procedures, parameters are deterministic. However, the center of the hyperspheres, $\mathbf{a}_k$, can be a random variable which allows us to obtain a probabilistic interpretation of the anomalies. Therefore, in this section, we introduce a Bayesian *n*-class SVDD with anomalies (B-*n*SVDD-A).

In the B-*n*SVDD-A approach, we assume that in-control observation $\mathbf{x}_i^{(k)}$ $\left(k=1,\ldots,n\right)$ is transformed into the higher dimensional space through $\phi(\cdot)$ and the transformed data $\phi\left(\mathbf{x}_i^{(k)}\right)$ follow a Gaussian distribution.

$$\phi(\mathbf{x}_i^{(k)}) \sim N\left(\sum_{i=1}^{N_k}\alpha_i^{(k)}\mathbf{x}_i^{(k)} - \sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)}\mathbf{x}_i^{(m)} - \sum_{i=1}^{N_A}\tau_i^{(A_k)}\mathbf{x}_i^{(A)}, \sigma_{kk}^2\mathbf{I}\right).$$

In the weighted Gaussian model, distance of an observation $\mathbf{x}_i^{(k)}$ to the center of a hypersphere *k* is inversely proportional to the likelihood. Thus, *n*SVDD-A is a special case of the weighted Gaussian model, which improves *n*SVDD-A by utilizing prior knowledge. We estimate the unknown weight parameters $\boldsymbol{\alpha}$ through a Bayesian approach with a proper prior distribution for $\boldsymbol{\alpha}$, obtained from the dual variables of the *n*SVDD-A ($\alpha_i^{(k)}, \beta_i^{(m,k)}$ and $\tau_i^{(A_k)}$). B-*n*SVDD-A approach assumes that the prior distributions of the dual variables $\alpha_i^{(k)}, \beta_i^{(m,k)}$ and $\tau_i^{(A_k)}$ are Gaussian distributions which is the conjugate prior of the likelihood below

$$p\left(\boldsymbol{\alpha}^{(k)}\right)=\frac{1}{\left(2\pi\right)^{N_k/2}\,\sigma^{N_k}}e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{\alpha}^{(k)}-\mathbf{m}^{(k)}\right\|_2^2},\ p\left(\boldsymbol{\beta}^{(m,k)}\right)=\frac{1}{\left(2\pi\right)^{N_m/2}\,\sigma^{N_m}}e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{\beta}^{(m,k)}-\mathbf{m}^{(m,k)}\right\|_2^2},$$

$$p\left(\boldsymbol{\tau}^{(k)}\right)=\frac{1}{\left(2\pi\right)^{N_A/2}\,\sigma^{N_A}}e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{\tau}^{(A_k)}-\mathbf{m}^{(A_k)}\right\|_2^2}$$

Assuming that the training data information in each class $\mathbf{x}_i^{(k)}$ follow a Gaussian distribution in kernel space, we can obtain the likelihood probability function given parameter $\boldsymbol{\alpha}$ as follows:

$$p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)=\prod_{j=1}^{n}\left[\prod_{k=1}^{N_j}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\,\sigma_{jj}{}^{\hat{p}}}e^{-\frac{1}{2\sigma_{jj}^2}\left\|\phi\left(\mathbf{x}_k^{(j)}\right)-\left(\sum_{i=1}^{N_j}\alpha_i^{(j)}\mathbf{x}_i^{(j)}-\sum_{j\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,j)}\mathbf{x}_i^{(m)}-\sum_{i=1}^{N_A}\tau_i^{(Aj)}\mathbf{x}_i^{(A)}\right)\right\|_2^2}\right]$$

By using the Bayesian rule, maximizing a posterior (MAP) is derived as follows:

$$p(\boldsymbol{\alpha}\,|\,\mathbf{D})=\frac{p(\mathbf{D}\,|\,\boldsymbol{\alpha})\,p(\boldsymbol{\alpha})}{p(\mathbf{D})},$$

where $\mathbf{D}=\mathbf{D}_1\cup\ldots\cup\mathbf{D}_n$ is a set of training data. Since $p(\mathbf{D})$ is a normalizing constant independent of $\boldsymbol{\alpha}$, it can be ignored, thus

$$p(\boldsymbol{\alpha}\,|\,\mathbf{D})\propto p(\mathbf{D}\,|\,\boldsymbol{\alpha})\,p(\boldsymbol{\alpha}). \tag{6.9}$$

The MAP solution is obtained by

$$\hat{\boldsymbol{\alpha}}=\arg\max_{\boldsymbol{\alpha}}\,p(\boldsymbol{\alpha}\,|\,\mathbf{D}) \tag{6.10}$$

which can be obtained as in Eq. (6.11).

$$\widehat{\boldsymbol{\alpha}} = \arg\max 2^{-1} \begin{pmatrix} 2\sum_{m=1}^{n}\sigma_{mm}^{-2}\left(\alpha^{(m)}\right)^T\mathbf{B}^{(m)}\mathbf{1}^{(m)} - 2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sigma_{kk}^{-2}\left(\beta^{(m,k)}\right)^T\mathbf{B}^{(m,k)}\mathbf{1}^{(m,k)} \\ -\sum_{k=1}^{n}N_k\sigma_{kk}^{-2}\left(\alpha^{(k)}\right)^T\mathbf{K}^{(k)}\alpha^{(k)} - \sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\beta^{(k,m)}\right)^T\mathbf{K}^{(k)}\beta^{(k,m)} \\ +2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\alpha^{(m)}\right)^T\mathbf{K}^{(m,k)}\beta^{(k,m)} - 2\sum_{m=1}^{n}\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{B}^{(A,m)}\mathbf{1}^{(A,m)} \\ +2\sum_{m=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A,m)}\alpha^{(m)} \\ -2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A,k)}\beta^{(k,m)} - \sum_{m=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A)}\tau^{(A_m)} \\ -\sigma^{-2}\left(\sum_{i=1}^{n}\boldsymbol{\alpha}^{(i)^T}\boldsymbol{\alpha}^{(i)} - 2\boldsymbol{\alpha}^{(i)^T}\mathbf{m}^{(i)}\right) - \sigma^{-2}\left(\sum_{i=1}^{n}\sum_{i\neq j=1}^{n}\boldsymbol{\beta}^{(i,j)^T}\boldsymbol{\beta}^{(i,j)} - 2\boldsymbol{\beta}^{(i,j)^T}\mathbf{m}^{(i,j)}\right) \\ -\sigma^{-2}\left(\sum_{i=1}^{n}\boldsymbol{\tau}^{(A_i)^T}\boldsymbol{\tau}^{(A_i)} - 2\boldsymbol{\tau}^{(A_i)^T}\mathbf{m}^{(A_i)}\right) \end{pmatrix}_{\boldsymbol{\alpha}} \tag{6.11}$$

where $\mathbf{B}_{ii}^{(m)} = \sum_j \mathbf{K}_{ij}^{(m)}$, $\mathbf{1}^{(m)}$ and $\mathbf{1}^{(m,k)}$ are $1\times N_m$ vector with all ones $\mathbf{B}_{ii}^{(A,k)} = \sum_j \mathbf{K}_{ij}^{(A,k)}$ and $\mathbf{B}_{ii}^{(m,k)} = \sum_j \mathbf{K}_{ij}^{(m,k)}$ (see Appendix D for detailed derivation of Eq. (6.11)).

Since we use the dual formulation to obtain the B-*n*SVDD-A, it should satisfy the same constraints of the original optimization problem in Eq. (6.7). Thus, B-*n*SVDD-A has the following constraints:

$$\sum_{i=1}^{N_k}\alpha_i^{(k)} - \sum_{k\neq m=1}^{n}\sum_{i=1}^{N_m}\beta_i^{(m,k)} - \sum_{i=1}^{N_A}\tau_i^{(A_k)} = 1, \quad \forall k = 1,\ldots,n$$
$$0 \le \alpha_i^{(k)} \le C_k, \quad \forall i = 1,\ldots,N_k, \forall k = 1,\ldots,n$$
$$0 \le \beta_i^{(m,k)} \le B, \ \forall m,k, \ \forall i = 1,\ldots,N_m$$
$$0 \le \tau_i^{(A_k)} \le O, \quad \forall i = 1,\ldots,N_A, \forall k = 1,\ldots,n$$

The optimization problem of B-*n*SVDD-A can be represented in a matrix form as follows:

$$\min 2^{-1}\left(\boldsymbol{\theta}^T\mathbf{M}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{E}\right)$$
$$s.t \ \ \mathbf{S}\boldsymbol{\theta} = \mathbf{1}$$
$$0 \le \alpha_i^{(k)} \le C_k, \quad \forall i = 1,\ldots,N_k, \forall k = 1,\ldots,n$$
$$0 \le \beta_i^{(m,k)} \le B, \ \forall m,k, \ \forall i = 1,\ldots,N_m$$
$$0 \le \tau_i^{(A_k)} \le O, \quad \forall i = 1,\ldots,N_A, \forall k = 1,\ldots,n$$

with $\mathbf{M} = \begin{pmatrix} \mathbf{P} & \mathbf{R}^T & -\mathbf{V}^T \\ \mathbf{R} & \mathbf{Q} & \mathbf{U}^T \\ -\mathbf{V} & \mathbf{U} & \mathbf{\Pi} \end{pmatrix}$, $\mathbf{E} = 2\begin{pmatrix} -\sigma_{11}^{-2}diag(\mathbf{B}^{(1)}) - \sigma^{-2}\mathbf{m}^{(1)} \\ \vdots \\ -\sigma_{11}^{-2}diag(\mathbf{B}^{(n)}) - \sigma^{-2}\mathbf{m}^{(n)} \\ \sigma_{11}^{-2}diag(\mathbf{B}^{(1,2)}) - \sigma^{-2}\mathbf{m}^{(1,2)} \\ \vdots \\ \sigma_{11}^{-2}diag(\mathbf{B}^{(n,n-1)}) - \sigma^{-2}\mathbf{m}^{(n,n-1)} \\ \sigma_{11}^{-2}diag(\mathbf{B}^{(A,1)}) - \sigma^{-2}\mathbf{m}^{(A_1)} \\ \vdots \\ \sigma_{11}^{-2}diag(\mathbf{B}^{(A,n)}) - \sigma^{-2}\mathbf{m}^{(A_n)} \end{pmatrix}$

$$\mathbf{P} = \begin{bmatrix} N_1\sigma_{11}^{-2}\mathbf{K}^{(1)} + \sigma^{-2}\mathbf{1}_{N_1\times N_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & N_2\sigma_{22}^{-2}\mathbf{K}^{(2)} + \sigma^{-2}\mathbf{1}_{N_2\times N_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & N_n\sigma_{nn}^{-2}\mathbf{K}^{(n)} + \sigma^{-2}\mathbf{1}_{N_n\times N_n} \end{bmatrix}$$

$$\mathbf{Q}=\begin{bmatrix} \sum_{i=2}^{n} N_i\sigma_{ii}^{-2}\mathbf{K}^{(1)}+\sigma^{-2}\mathbf{1}_{N_1\times N_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sum_{2\neq i=1}^{n} N_i\sigma_{ii}^{-2}\mathbf{K}^{(2)}+\sigma^{-2}\mathbf{1}_{N_2\times N_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{i=1}^{n-1} N_i\sigma_{ii}^{-2}\mathbf{K}^{(n)}+\sigma^{-2}\mathbf{1}_{N_n\times N_n} \end{bmatrix}$$

$$\mathbf{R}=\begin{bmatrix} \mathbf{0} & -N_2\sigma_{22}^{-2}\mathbf{K}^{(1,2)} & \cdots & -N_n\sigma_{nn}^{-2}\mathbf{K}^{(1,n)} \\ -N_1\sigma_{11}^{-2}\mathbf{K}^{(2,1)} & \mathbf{0} & \cdots & -N_n\sigma_{nn}^{-2}\mathbf{K}^{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ -N_1\sigma_{11}^{-2}\mathbf{K}^{(n,1)} & -N_2\sigma_{22}^{-2}\mathbf{K}^{(n,2)} & \cdots & \mathbf{0} \end{bmatrix}$$

$$\mathbf{V}=\begin{bmatrix} N_1\sigma_{11}^{-2}\mathbf{K}^{(A,1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & N_2\sigma_{22}^{-2}\mathbf{K}^{(A,2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & N_n\sigma_{nn}^{-2}\mathbf{K}^{(A,n)} \end{bmatrix}$$

$$\mathbf{U}=\begin{bmatrix} \mathbf{0} & N_1\sigma_{11}^{-2}\mathbf{K}^{(A,2)} & \cdots & N_1\sigma_{11}^{-2}\mathbf{K}^{(A,n)} \\ N_2\sigma_{22}^{-2}\mathbf{K}^{(A,1)} & \mathbf{0} & \cdots & N_2\sigma_{22}^{-2}\mathbf{K}^{(A,n)} \\ \vdots & \vdots & \ddots & \vdots \\ N_n\sigma_{nn}^{-2}\mathbf{K}^{(A,1)} & N_n\sigma_{nn}^{-2}\mathbf{K}^{(A,2)} & \cdots & \mathbf{0} \end{bmatrix}$$

$$\mathbf{\Pi}=\begin{bmatrix} N_1\sigma_{11}^{-2}\mathbf{K}^{(A)}+\sigma^{-2}\mathbf{1}_{N_A\times N_A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & N_2\sigma_{22}^{-2}\mathbf{K}^{(A)}+\sigma^{-2}\mathbf{1}_{N_A\times N_A} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & N_n\sigma_{nn}^{-2}\mathbf{K}^{(A)}+\sigma^{-2}\mathbf{1}_{N_A\times N_A} \end{bmatrix}$$

where $\boldsymbol{\theta}$, $\mathbf{S}$ and each element of $\mathbf{K}^{(k)}$ and $\mathbf{K}^{(k,m)}$ $\left(\forall m\neq k=1,2,\ldots,n,A\right)$ are obtained as described in Section 6.3.

6.3.2 Parameter Settings of the Prior Distribution

To solve the above optimization problems, we need to know the parameters of the distribution of $\boldsymbol{\alpha}$ beforehand. We determine the parameters of $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\beta}^{(k,m)}$ ($k \neq m = 1,\ldots,n$) by adopting the procedure explained in Ghasemi *et al.* (2016). Thus, the parameters of $\alpha_i^{(k)}$ and $\beta_i^{(m,k)}$ are determined as follows:

$$m_i^{(k)} = -\left(\sum_{j=1}^{N_k} \mathbf{K}_{i,j}^{(k)}\right)^{v} \text{ where } k = 1,2,\ldots,n \text{ and } v\ (0 < v \leq 1)$$

and

$$m_i^{(k,m)} = -\left(\sum_{j=1}^{N_m} \mathbf{K}_{i,j}^{(k,m)}\right)^{v} \text{ where } k \neq m = 1,2,\ldots,n \text{ and } v\ (0 < v \leq 1).$$

In addition to the parameters of $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\beta}^{(k,m)}$, we also need to determine the parameters of $\boldsymbol{\tau}^{(A_k)}$ ($k = 1,\ldots,n$). Anomaly class observations, $\mathbf{x}_i^{(A)}$, which satisfy $\tau_i^{(A_1)} > 0$ are also the support vector for class 1. These support vectors lie on the boundary of class 1 if $0 < \tau_i^{(A_1)} < O$, or they are placed inside hypersphere 1 if $\tau_i^{(A_1)} = O$. Therefore, we can determine $m_i^{(A_1)}$ for $\tau_i^{(A_1)}$ as follows:

$$m_i^{(A_1)} \propto -\sum_{j \in \mathbf{D}_1} \mathbf{K}_{i,j}^{(A,1)}$$

Thus, $m_i^{(A_1)}$ is chosen as $m_i^{(A_1)} = -\left(\sum_{j=1}^{N_1} \mathbf{K}_{i,j}^{(A,1)}\right)^{v}$ where $v\ (0 < v \leq 1)$.

6.4 Performance Comparison

In this section, we demonstrate the performance of the proposed procedure by conducting simulations on some artificial data sets and a real data set. The performance of the proposed procedures is compared with existing SVDD procedure and its variants such as traditional SVDD, TC-SVDD and Bayesian TC-SVDD (BTC-SVDD).

The performances of different procedures are compared using the total anomaly accuracy ratio (TAAR) for the testing dataset under the same error ratio of normal classes. TAAR is defined as follows:

$$TAAR = \frac{\#\,of\ true\,identifications\,in\,anomalies}{total\ number\ of\ anomalies}$$

In this study, 10-fold cross validation is used to measure the performances of all data sets. The parameters of the proposed and benchmark procedures are optimized based on cross validation procedure introduced by Kang and Cho (2012) .

6.4.1 Simulated Examples

6.4.1.1 Banana-Shaped Data

Two classes of banana-shaped data in two-dimensions, introduced by Duin *et al.* (2000), are well sampled with the size of each class of objects being 70. Four different anomaly data sets with the size 50 are generated from multivariate normal distribution with means

[-1,-6.2] (a), [0,-6.5] (b), [4,-5] (c) and [4.5,-5] (d) and covariance 0.25 times identity matrix as shown in Figure 6.1 (a,b,c,d). By using the normal data set (blue dot) and anomaly data (yellow dot), we obtain the 2SVDD-A and B-2SVDD-A boundaries for each case. For the testing data, 100 observations are obtained for each case from the same distributions as anomaly data sets.

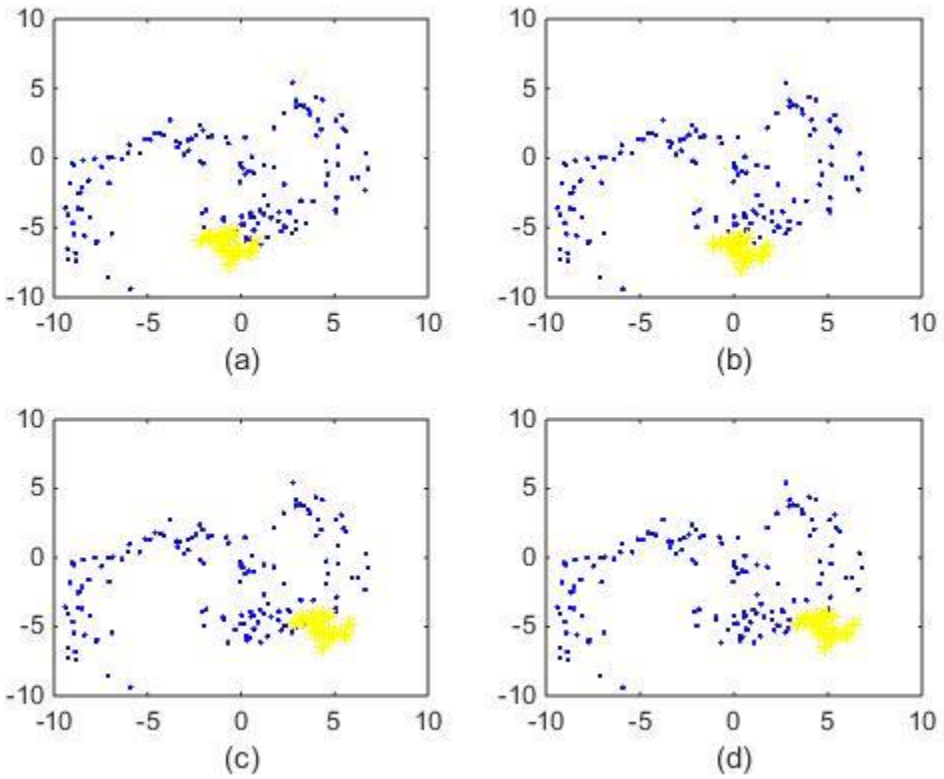


Figure 6.1 Banana-shaped data (blue dot) with anomaly data (yellow dot). (a), (b), (c) and (d) denote the location of the anomaly data obtained from the means [-1,-6.2], [0,-6.5], [4,-5] and [4.5,-5], respectively.

Figure 6.2 shows the boundaries of the proposed procedure for each case. Note that the proposed SVDD gives two descriptions for the training data set that contains two classes of the training data and the anomaly set simultaneously.

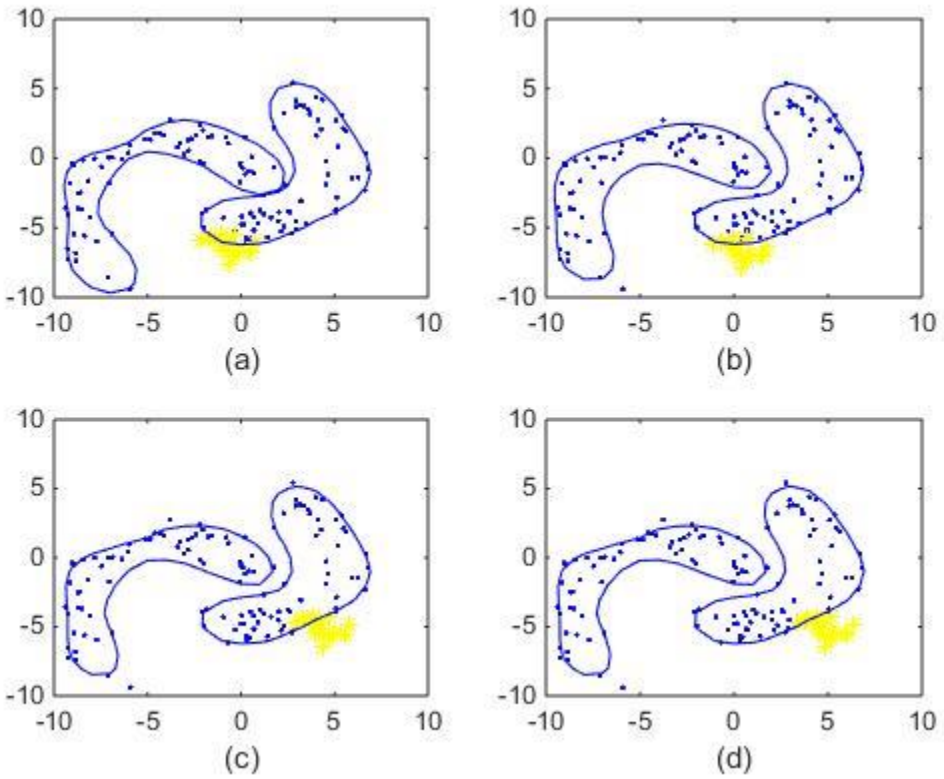


Figure 6.2 Illustration of the proposed procedure applied to the banana-shaped data set. (a), (b), (c) and (d) denote the boundaries of the cases where the anomalies are obtained with means [-1,-6.2], [0,-6.5], [4,-5] and [4.5,-5], respectively.

Table 6.2 illustrates the performance of the proposed procedures against other benchmark procedures under banana-shaped data for the cases a, b, c, d explained above. The results show that *2*SVDD-A outperforms traditional SVDD and TC-SVDD. For the Bayesian procedures, proposed B-*2*SVDD-A procedure outperforms other existing Bayesian SVDD procedures. In addition, Figure 6.1 shows that the anomaly data are closest to the target data in the case (a) and farthest from the target data in the case (d). Table 6.2 shows that each procedure's performance increases and becomes similar to the others when the

anomaly data are far from the normal data. However, the proposed B-*2*SVDD-A still outperforms other procedures even if anomaly data are close to the normal data.

Table 6.2 Performance comparisons (TAAR) under banana-shaped data when testing data are same as anomaly data

	Cases			
Procedure	a	b	c	d
SVDD	0.6120	0.6800	0.7780	0.9120
TC-SVDD	0.6400	0.7020	0.7880	0.9120
BTC-SVDD	0.7360	0.8060	0.8680	0.9400
2SVDD-A	0.6590	0.7229	0.8267	0.9362
B-2SVDD-A	0.7800	0.8524	0.9000	0.9543

In addition, Figure 6.3 illustrates the normal (banana-shaped), anomaly and testing data sets. In this study, anomaly points (yellow dot) and normal points (blue dot) are used to obtain the boundary for 2SVDD-A and B-*2*SVDD-A. Anomaly data are obtained from multivariate normal distribution with mean [4,-5] and covariance 0.25 times identity matrix (a). Testing points (red dot) are generated from the normal distribution for the cases b, c and d with means [-7.7, 0.5], [6.8,-2] and [0,-6.5] and same covariance matrix as anomaly data, respectively.

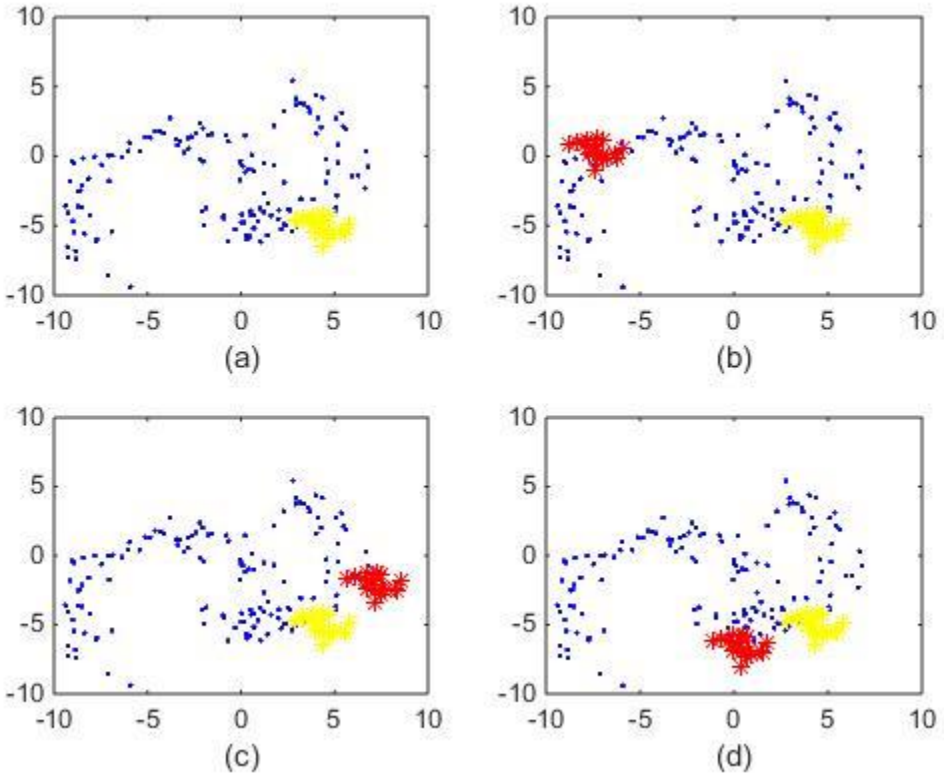


Figure 6.3 Banana-shaped data (blue dot) with anomaly (yellow dot) and testing data (red dot). (a) denotes the location of the anomaly data obtained from the mean [4,-5]. (b), (c) and (d) denote the location of the testing data obtained from the means [-7.7, 0.5], [6.8,-2] and [0,-6.5], respectively.

Table 6.3 shows that B-*2*SVDD-A outperforms other procedures regardless of the location of the anomalies for the cases b, c and d explained above. In addition, *2*SVDD-A's performance improvement is significantly more than traditional SVDD and TC-SVDD. If the testing data are far from the normal data, the performances of the procedures become similar to one another.

Table 6.3 Performance comparisons (TAAR) under banana-shaped data when testing data are different than anomaly data

	Cases		
Procedure	b	c	d
SVDD	0.4800	0.6420	0.6825
TC-SVDD	0.5060	0.6440	0.7040
BTC-SVDD	0.5760	0.7420	0.8060
2SVDD-A	0.5810	0.6724	0.7295
B-2SVDD-A	0.6162	0.7876	0.8543

6.4.1.2 Multivariate Skew Normal Distribution

In addition to the banana-shaped data, we also obtain data from multivariate skew normal (MSN) distribution. A *p*-dimensional vector $\mathbf{x}$ follows the MSN distribution with the density function defined as $2\Phi_p(\mathbf{y}-\boldsymbol{\varepsilon};\boldsymbol{\Sigma})\boldsymbol{\Phi}(\mathbf{d}^T\mathbf{w}^{-1}(\mathbf{y}-\boldsymbol{\varepsilon}))$, $(\mathbf{x}\in\mathbf{R}^p)$ by defining $\mathbf{y}=\boldsymbol{\varepsilon}+\mathbf{w}\mathbf{x}$, where $\Phi_p(\mathbf{y}-\boldsymbol{\varepsilon};\boldsymbol{\Sigma})$ is the *p*-dimensional normal density with location and scale parameters, $\boldsymbol{\mu}=(\mu_1,...,\mu_p)^T$, $\mathbf{w}=(w_1,...,w_p)$ respectively. Correlation matrix is denoted by $\boldsymbol{\Sigma}$, and $\boldsymbol{\Phi}(\cdot)$ is the $N(0,1)$ distribution function. The parameter $\mathbf{d}$ is defined as *p*-dimensional skewness parameter. If $\mathbf{d}$ is a vector of zeros, $\boldsymbol{\Phi}(\mathbf{d}^T\mathbf{x})$ equals to $1/2$. Therefore, the density function defined above is reduced to *p*-dimensional normal distribution $N_p(0,\boldsymbol{\Sigma})$ (Azzalini and Capitanio, 1999, Azzalini and Dalla Valle, 1996). Throughout this paper, we use the notation $\mathbf{y}\sim SN_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{d})$ when $\mathbf{y}$ follows an MSN distribution. Gupta *et al*. (2004) show how to obtain mean and covariance of a vector and

the mean of $\mathbf{y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})$ which is based on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{d}$. To obtain the shifted observations for the simulation study, we only focus on $\boldsymbol{\mu}$. If $\boldsymbol{\mu}$ is the in-control mean, then the shifted observation is obtained from $\mathbf{y} \sim SN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}, \mathbf{d})$ where $\boldsymbol{\mu}_1 = \boldsymbol{\mu} + \boldsymbol{\delta}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)$. In addition, skewness parameter $\mathbf{d}$ shows the deviance of the data from normality. When $\mathbf{d}$ deviates from zero, the data also deviate from normality.

Two different MSN data sets with size 100 for each of the classes are utilized as normal data (Figure 6.4 (a)). We also generate different anomaly data sets from normal distribution with means [0, -2.5], [0.485, 2] and [1, -0.1] and covariance 0.01 times identity matrix as shown in Figure 6.4 (b, c, d) and obtain testing points from the same distribution as anomaly data.

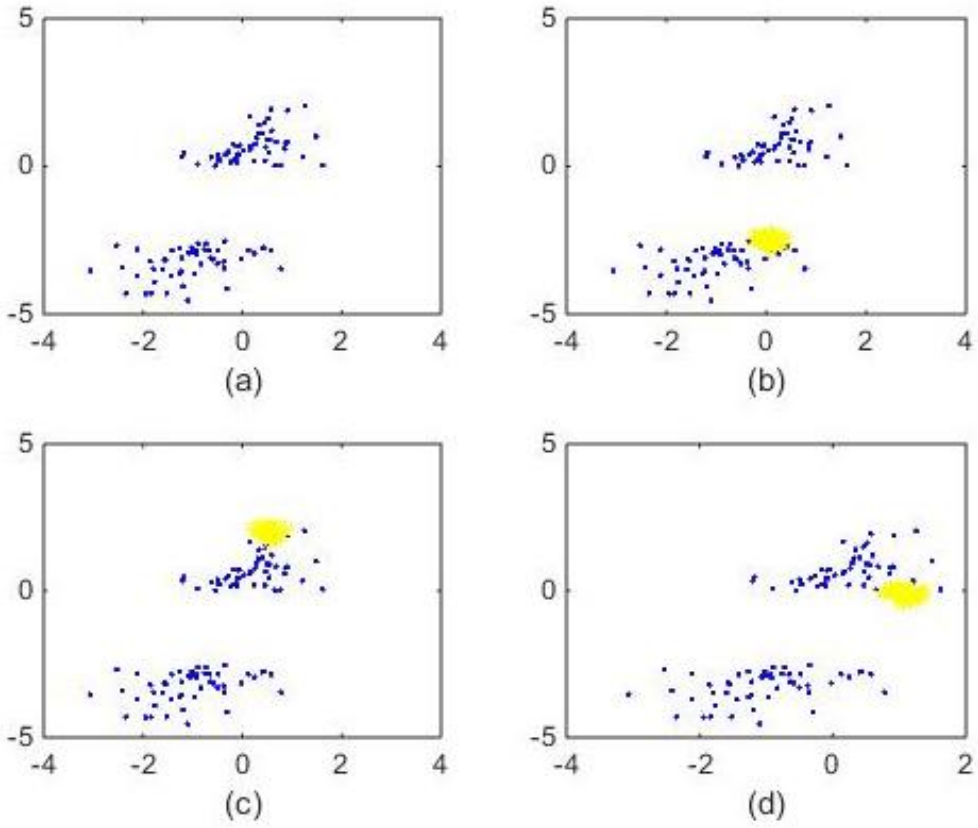


Figure 6.4 Multivariate skew normal data (blue dot) with anomaly data (yellow dot). (a) denotes the two classes multivariate skew normal data. (b), (c) and (d) denote the location of the anomaly data obtained from the means [0, -2.5], [0.485, 2] and [1, -0.1], respectively.

Performances of the different procedures are summarized in Table 6.4 for the cases b, c and d. Similar to the banana-shaped case, B-*2*SVDD-A outperforms all of the existing SVDD procedures and *2*SVDD-A outperforms other non Bayesian SVDD procedures. In addition, the performances of the SVDD procedures become similar to one another when the anomaly data are placed far from the normal data.

Table 6.4 Performance comparisons (TAAR) under two-dimensional MSN data when testing data are same as anomaly data

	Cases		
Procedure	b	c	d
SVDD	0.5611	0.6578	0.8367
TC-SVDD	0.6467	0.6667	0.8422
BTC-SVDD	0.7533	0.7700	0.8756
2SVDD-A	0.6978	0.7935	0.9141
B-2SVDD-A	0.8043	0.8707	0.9609

In addition to the two dimensional skew normal data, Table 6.5 summarizes the performances of the proposed and the existing procedures under two classes of MSN data sets with sizes 100 in five-dimension. Five-dimensional normal data are obtained by choosing the skewness parameters as **d**=(-2, 1, -2, 1,-2) and **d**= (2, -1, 2, -1, 2). By choosing this setting in which skewness parameters can be considered as symmetric, it is possible to observe how the proposed procedure responds to positive and negative skewness. Three different anomaly data sets are used to obtain the boundaries for *2*SVDD-A and B-*2*SVDD-A by shifting the mean of normal MSN distribution as 1*[1 1 1 0 0] (a), 1.25*[1 1 1 0 0] (b) and 1.5*[1 1 1 0 0] (c). Eighty testing points are obtained in the same way as obtaining anomaly data. The results show the same performance pattern as in the previous simulation studies. B-*2*SVDD-A outperforms all of the existing SVDD procedures and the performances of the SVDD procedures become similar to one

another for the large shift. In addition, we observe that increasing the dimension of the data does not affect the pattern of the results.

Table 6.5 Performance comparisons (TAAR) under five-dimensional MSN data

	Cases		
Procedure	a	b	c
SVDD	0.3700	0.5675	0.7500
TC-SVDD	0.3825	0.5763	0.7512
BTC-SVDD	0.4825	0.6875	0.8725
2SVDD-A	0.4707	0.6585	0.8122
B-2SVDD-A	0.5683	0.7793	0.8854

6.5 Case Studies: Continuous Stirred Tank Heater

In this section, we use the same case study introduced in Chapter 5.6, namely, continuous stirred tank heater (CSTH). We obtain the data from three controlled variables, level, temperature, and CW flow. The two modes of CSTH are obtained under normal operating modes with the size of each mode of objects being 100 as explained in Chapter 5.6. In addition, three different anomaly data sets are obtained from Mode 1 each with size 20 by introducing step change of -0.5 (a), -0.75 (b) and -1 (c) into the level measurement. In addition, eighty testing points are obtained from Mode 1 in the same way that the anomaly data are obtained.

By using the normal and anomaly data, the boundaries are obtained for each B-2SVDD-A and 2SVDD-A procedures. Table 6.7 shows the results obtained from the each SVDD approaches. These results demonstrate that the proposed B2SVDD procedure significantly improves the ability of anomaly detection compared to the other SVDD procedures. It is also clear that the performances of the SVDD procedures become similar to one another while the shift size increases.

Table 6.7 Performance comparisons (TAAR) for CSTH data

	Cases		
Procedure	a	b	c
TC-SVDD	0.6762	0.8250	0.8938
BTC-SVDD	0.6937	0.8463	0.9025
2SVDD-A	0.6939	0.8451	0.9049
B-2SVDD-A	0.8293	0.9073	0.9634

6.6 Conclusions

Identifying anomalies is crucial since they may include significant information for a given process or real-life problems. When normal data consist of more than one class and some anomaly data are available, existing SVDD procedures may not be efficient for detection of anomalies.

In this chapter, we propose a generalized SVDD procedure called *n*SVDD-A which utilizes the anomaly information and is independent from the number of classes. The

proposed procedure determines *n* number of hyperspheres. Each hypersphere keeps as many corresponding observations as possible inside its boundary and places observations of other classes and anomalies outside hypersphere. Furthermore, we introduce a Bayesian SVDD procedure by assuming that *n*-classes normal data and some anomaly data are available. B-*n*SVDD-A is built by assuming that transformed variables and prior distributions follow normal distribution in the feature space. Experimental results based on different data sets show that B-*n*SVDD-A is superior to the existing SVDD procedures in terms of detection of anomalies.

In future research, it would be reasonable to develop a method that obtains the probabilistic outputs of B-*n*SVDD-A, which may provide even more information when detecting the anomalies. In addition, a new framework can be developed by decreasing the number of the parameters, which allows us to decrease computational time of parameter selection.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This dissertation proposes and subsequently develops procedures for distribution-free fault identification and anomaly detection in high-dimensional data. This chapter presents the summary and conclusions of this dissertation and describes the possible future research related to this dissertation.

7.1 Summary and Conclusions

7.1.1 Distribution-Free Adaptive Step-Down Approach for Fault Identification

In chapter 3, we introduce a distribution-free adaptive step-down approach for faulty variable identification. The proposed procedure integrates adaptive step-down approach with an SVDD-based test statistic. The proposed procedure called DFASD selects a variable having no significant evidence of a change based on the p variables that are selected in previous steps and eventually obtains the changed variables. This strategy can reduce computational times when a few variables are changed in a high-dimensional process. In addition, the proposed procedure is robust to the correlations between variables, resulting in stable performance regardless of the number of changed variables. The experiment results with diverse dataset demonstrate superiority of the proposed distribution-free procedure.

7.1.2 Bayesian Framework for Fault Variable Identification

In chapter 4, we propose a Bayesian approach for fault identification that addresses the limitations posed by the normality assumption, a distribution-free procedure. The proposed approach is based on Bayesian support vector data description (BSVDD) and the faulty variables are identified based on an efficient algorithm with significant computational advantage. In addition we also propose a local density degree function to assign the parameters of the prior distribution. The proposed local density degree function is more interpretable and offers better performance in the identification of faulty variables than the existing procedures. Experiments with diverse data sets show that BSVDD procedure is robust to the non-normal data, specially for irregularly patterned data, in terms of fault detections. This feature is important in practice when the type of the distribution is unknown.

7.1.3 Generalized Support Vector Data Description with Bayesian Framework

Chapter 5 introduces an anomaly detection procedure which is independent of the number of classes, called *n*-SVDD. Regardless of the number of classes, anomalies are identified based on the relative distance to the center of each hypersphere of each class. In addition, we introduce a Bayesian framework for the proposed *n*-SVDD procedure by assuming the transformed variables and the prior distributions follow normal distribution. Experiments with diverse data sets show that the proposed Bayesian procedure is superior to the existing SVDD procedures.

7.1.4 Multi-Class Bayesian Support Vector Data Description with Anomalies

In chapter 6, we propose a multi-class Bayesian procedure called B-*n*SVDD-A for anomaly detection. The proposed procedure describes the multi-class normal data more accurately by considering the prior information of the anomaly data. Regardless of the location of the anomalies, the information of those anomalies is helpful to better describe the normal data since anomaly information makes the boundary of the normal data much tighter. To show the superiority of the proposed procedure, we conduct simulation studies with diverse data sets and a real-life case study of a continuous stirred tank heater (CSTH). These studies indicate that the proposed procedure is superior to the existing SVDD procedures in terms of detection of anomalies.

7.2. Future Research

In Chapters 3 and 4, this dissertation focuses on identification of faulty variables with high dimensional data when the underlying distribution of the process is obtained from one class. These works can be extended by studying the identification of the faulty variables where underlying distribution of the process is obtained from multi-class data. In addition, in future work, we may extend our procedures for multimode multistage processes.

In Chapters 5 and 6, we introduce anomaly detection procedures under the assumption that there are given n number of classes (normal data consist of n distributions) and anomaly data. In particular, the introduced procedures can be extended for the case where

the number of classes is known but the labels of the observations are unknown. Therefore, extended procedures can identify the classes and anomalies simultaneously under the assumption that the number of classes is given.

In addition, when there is no information available about the classes, it is challenging to obtain the critical values for multi-class anomaly detection procedures. Thus, it is important to know the number of classes of the normal data to detect anomalies in multi-class environment. Therefore, we can extend Chapters 5 and 6 to detect anomalies in multi-class environment where there is no class information.

Appendix A. Derivation of Eq. (4.6)

Starting with the Eq. (4.5),

$$\begin{aligned}\hat{\boldsymbol{\tau}} &= \arg\max_{\boldsymbol{\tau}} p(\boldsymbol{\tau} \mid \mathbf{D}) \\ &= \arg\max_{\boldsymbol{\tau}} \ln\left(p(\boldsymbol{\tau} \mid \mathbf{D})\right) \\ &= \arg\max_{\boldsymbol{\tau}} \left(\ln\left(p(\mathbf{D} \mid \boldsymbol{\tau})\right) + \ln\left(p(\boldsymbol{\tau})\right)\right)\end{aligned} \tag{A.1}$$

Two terms in Eq. (A.1) can be written as

$$\begin{aligned}\ln\left(p(\mathbf{D} \mid \boldsymbol{\tau})\right) &= \ln\left(\prod_{i=1}^{m} \frac{1}{(2\pi)^{\hat{p}/2}\sigma^{\hat{p}}} e^{-\frac{1}{2\sigma^2}\left\|\phi(\mathbf{x}_i)-\sum_{j=1}^{m}\tau_j\phi(\mathbf{x}_j)\right\|_2^2}\right) \\ &= \ln\left(\left(\frac{1}{(2\pi)^{\hat{p}/2}\sigma^{\hat{p}}}\right)^m\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}\mathbf{K}_{ii} - 2\sum_{i=1}^{m}\sum_{j=1}^{m}\tau_j\mathbf{K}_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\tau_j\tau_k\mathbf{K}_{jk}\right) \\ &= \ln\left(\left(\frac{1}{(2\pi)^{\hat{p}/2}\sigma^{\hat{p}}}\right)^m\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}\mathbf{K}_{ii} - 2\boldsymbol{\tau}^T\mathbf{B}\mathbf{1} + m\boldsymbol{\tau}^T\mathbf{K}\boldsymbol{\tau}\right)\end{aligned}$$

$$\ln\left(p(\boldsymbol{\tau})\right) = \ln\left(\prod_{i=1}^{m} p(\tau_i \mid C)\right) = \ln\left(\prod_{i=1}^{m}\frac{\theta_i e^{-\theta_i\tau_i}}{1-e^{-C\theta_i}}\right) = \ln\left(\prod_{i=1}^{m}\frac{\theta_i}{1-e^{-C\theta_i}}\right) - \sum_{i=1}^{m}\theta_i\tau_i$$

The results of $\ln(p(\mathbf{D} \mid \boldsymbol{\tau}))$ and $\ln\left(p(\boldsymbol{\tau})\right)$ are substituted in Eq. (A.1) as follows:

$$\hat{\boldsymbol{\tau}} = \arg\max_{\boldsymbol{\alpha}} \left(\ln\left(p\left(\mathbf{D} \mid \boldsymbol{\tau} \right) \right) + \ln\left(p\left(\boldsymbol{\tau} \right) \right) \right)$$
$$= \arg\min_{\boldsymbol{\alpha}} 2^{-1} \times \left(-2\sigma^{-2} \boldsymbol{\tau}^T \mathbf{B}\mathbf{1} + m\sigma^{-2} \boldsymbol{\tau}^T \mathbf{K}\boldsymbol{\tau} + 2\sum_{i=1}^{m} \theta_i \tau_i \right)$$

where $\mathbf{B}$ is a diagonal matrix and $\mathbf{B}_{i,i} = \sum_j \mathbf{K}_{i,j}$, $\mathbf{1}$ is a $m \times 1$ vector with all ones, and $\mathbf{K}$ is the kernel matrix in which $(i, j)^{th}$ component of the matrix is defined as $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Since C is inversely proportional to the number of training data m (Tax and Duin, 2004), the relationship $p(\mathbf{D} \mid \boldsymbol{\tau}, \sigma^2) \propto p(\mathbf{D} \mid \boldsymbol{\tau}, m)$ holds for given data set. Therefore, by using m instead of σ^2, the Eq. (4.6) is obtained.

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} 2^{-1} \times \left(-2m^{-1} \boldsymbol{\tau}^T \mathbf{B}\mathbf{1} + \boldsymbol{\tau}^T \mathbf{K}\boldsymbol{\tau} + 2\sum_{i=1}^{m} \theta_i \tau_i \right).$$

Appendix B. Raw Data of Bolt Measurements

Table B.1 In-control observations

Sample Number	x_1	x_2	x_3	x_4
1	0.36827	0.23753	0.25885	0.73546
2	0.37060	0.24296	0.25457	0.73476
3	0.36323	0.25096	0.25227	0.72170
4	0.36885	0.25071	0.24852	0.72502
5	0.36787	0.25409	0.24995	0.72280
6	0.36495	0.24696	0.24808	0.73149
7	0.36652	0.23617	0.25224	0.73826
8	0.36792	0.23990	0.24933	0.74028
9	0.37374	0.23415	0.25543	0.74557
10	0.37228	0.25622	0.24819	0.72031
11	0.36487	0.24266	0.24742	0.73358
12	0.37276	0.24599	0.25128	0.73093
13	0.36861	0.24286	0.24615	0.73417
14	0.36719	0.25047	0.25212	0.72510
15	0.36859	0.23933	0.25138	0.73799
16	0.36693	0.23646	0.24935	0.74291
17	0.36708	0.23624	0.24534	0.74545
18	0.36998	0.24618	0.25536	0.73093
19	0.36984	0.24216	0.25261	0.73433
20	0.36985	0.24203	0.25067	0.73668
21	0.36851	0.25290	0.25171	0.72312
22	0.36513	0.24811	0.24832	0.72555
23	0.36859	0.24529	0.24583	0.73434
24	0.37023	0.25283	0.25187	0.72384
25	0.36818	0.23935	0.24780	0.73868
26	0.36916	0.24800	0.24872	0.72997
27	0.36861	0.24922	0.24794	0.72806
28	0.36675	0.24375	0.24817	0.73099
29	0.36783	0.24589	0.24465	0.73120
30	0.36588	0.23860	0.25164	0.73771
31	0.36890	0.23676	0.25206	0.73949
32	0.36524	0.24689	0.24901	0.73292
33	0.36538	0.25062	0.24937	0.72299
34	0.36584	0.26185	0.24835	0.71166
35	0.36200	0.24421	0.24984	0.72965

36	0.36989	0.24438	0.25139	0.72990
37	0.36789	0.24401	0.24836	0.73260
38	0.36594	0.23381	0.25243	0.74262
39	0.36977	0.24058	0.25071	0.74030
40	0.36422	0.23579	0.24573	0.74125
41	0.36712	0.25022	0.24966	0.72334
42	0.36687	0.23968	0.24695	0.74147
43	0.36787	0.24515	0.24724	0.73474
44	0.36786	0.24116	0.25777	0.73407
45	0.36574	0.24779	0.25440	0.72393
46	0.36725	0.24122	0.25077	0.73458
47	0.36700	0.24813	0.24654	0.72822
48	0.36843	0.24874	0.24866	0.72846
49	0.36927	0.25422	0.25171	0.72083
50	0.36930	0.24240	0.25360	0.73523
51	0.36575	0.23267	0.24420	0.74676
52	0.36744	0.23961	0.24289	0.73595
53	0.36511	0.25470	0.24534	0.71892
54	0.36530	0.23956	0.24948	0.73555
55	0.36729	0.25086	0.25187	0.72355
56	0.37006	0.24388	0.25320	0.73211
57	0.36591	0.25469	0.24978	0.72051
58	0.36797	0.23233	0.25009	0.73893
59	0.36689	0.24394	0.24844	0.73185
60	0.36931	0.23612	0.25329	0.74408

Table B.2 Out of control observations

Sample Number	x_1	x_2	x_3	x_4
1	0.36511	0.25464	0.24799	0.74997
2	0.37090	0.25729	0.25017	0.74833
3	0.36739	0.25926	0.25054	0.74562
4	0.36542	0.25847	0.25446	0.74427
5	0.36801	0.25625	0.25036	0.74659
6	0.36783	0.25884	0.25450	0.74476
7	0.36460	0.25406	0.25036	0.75071
8	0.36947	0.25623	0.25773	0.74769
9	0.36714	0.26044	0.24735	0.74205
10	0.37005	0.26412	0.25263	0.73650
11	0.37118	0.25561	0.24887	0.74715
12	0.36578	0.25435	0.24800	0.75066
13	0.36740	0.25282	0.24338	0.75712
14	0.36786	0.25458	0.25043	0.74833
15	0.36699	0.25749	0.24721	0.74831

Appendix C. Derivation of Eq. (5.12)

Starting with the Eq. (5.11),

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \ln\left(p\left(\boldsymbol{\alpha}\,|\,\mathbf{D}\right)\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)p\left(\boldsymbol{\alpha}\right)\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \left[\ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)\right)+\ln\left(p\left(\boldsymbol{\alpha}\right)\right)\right]\end{aligned} \tag{C.1}$$

Two terms in Eq. (C.1) can be written as

$$\ln\left(p\left(\boldsymbol{\alpha}\right)\right) = \ln\left(\left(\frac{1}{\left(2\pi\right)^{(N_1+N_2)/2}\,\sigma^{N_1+N_2}}\right)\right) - \frac{1}{2\sigma^2}\left[\begin{array}{l}\left(\sum_{i=1}^{2}\boldsymbol{\alpha}^{(i)^T}\boldsymbol{\alpha}^{(i)} - 2\boldsymbol{\alpha}^{(i)^T}\mathbf{m}^{(i)} + \mathbf{m}^{(i)^T}\mathbf{m}^{(i)}\right) + \\ \left(\sum_{i=1}^{2}\sum_{i\neq j=1}^{2}\boldsymbol{\beta}^{(i)^T}\boldsymbol{\beta}^{(i)} - 2\boldsymbol{\beta}^{(i)^T}\mathbf{m}^{(i,j)} + \mathbf{m}^{(i,j)^T}\mathbf{m}^{(i,j)}\right)\end{array}\right]$$

$$\begin{aligned}\ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)\right) &= \ln\left(\begin{array}{l}\prod_{k=1}^{N_1}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\,\sigma_{11}^{\hat{p}}}\,e^{-\frac{1}{2\sigma_{11}^2}\left\|\phi\left(\mathbf{x}_k^{(1)}\right)-\left(\sum_{j=1}^{N_1}\alpha_j^{(1)}\phi\left(\mathbf{x}_j^{(1)}\right)-\sum_{j=1}^{N_2}\beta_j^{(2)}\phi\left(\mathbf{x}_j^{(2)}\right)\right)\right\|_2^2} \\ \times\prod_{k=1}^{N_2}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\,\sigma_{22}^{\hat{p}}}\,e^{-\frac{1}{2\sigma_{22}^2}\left\|\phi\left(\mathbf{x}_k^{(2)}\right)-\left(\sum_{j=1}^{N_2}\alpha_j^{(2)}\phi\left(\mathbf{x}_j^{(2)}\right)-\sum_{j=1}^{N_1}\beta_j^{(1)}\phi\left(\mathbf{x}_j^{(1)}\right)\right)\right\|_2^2}\end{array}\right) \\ &= \ln\left(\prod_{k=1}^{N_1}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\,\sigma_{11}^{\hat{p}}}\,e^{-\frac{1}{2\sigma_{11}^2}\left\|\phi\left(\mathbf{x}_k^{(1)}\right)-\left(\sum_{j=1}^{N_1}\alpha_j^{(1)}\phi\left(\mathbf{x}_j^{(1)}\right)-\sum_{j=1}^{N_2}\beta_j^{(2)}\phi\left(\mathbf{x}_j^{(2)}\right)\right)\right\|_2^2}\right) \\ &\quad + \ln\left(\prod_{k=1}^{N_2}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\,\sigma_{22}^{\hat{p}}}\,e^{-\frac{1}{2\sigma_{22}^2}\left\|\phi\left(\mathbf{x}_k^{(2)}\right)-\left(\sum_{j=1}^{N_2}\alpha_j^{(2)}\phi\left(\mathbf{x}_j^{(2)}\right)-\sum_{j=1}^{N_1}\beta_j^{(1)}\phi\left(\mathbf{x}_j^{(1)}\right)\right)\right\|_2^2}\right)\end{aligned}$$

The results of $\ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)\right)$ and $\ln\left(p\left(\boldsymbol{\alpha}\right)\right)$ are substituted in Eq. (C.1). Therefore, the Eq. (5.12) is obtained as follows:

$$\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \left[2^{-1} \times \left(\begin{array}{l} 2\sum_{m=1}^{2} \sigma_{mm}^{-2} \left(\alpha^{(m)}\right)^{T} \mathbf{B}^{(m)} \mathbf{1}^{(m)} - 2\sum_{m=1}^{2} \sum_{m\neq k=1}^{2} \sigma_{kk}^{-2} \left(\beta^{(m)}\right)^{T} \mathbf{B}^{(m,k)} \mathbf{1}^{(m,k)} \\ -\sum_{k=1}^{2} N_k \sigma_{kk}^{-2} \left(\alpha^{(k)}\right)^{T} \mathbf{K}^{(k)} \alpha^{(k)} - \sum_{m=1}^{2} \sum_{m\neq k=1}^{2} N_m \sigma_{mm}^{-2} \left(\beta^{(k)}\right)^{T} \mathbf{K}^{(k)} \beta^{(k)} \\ +2\sum_{m=1}^{2} \sum_{m\neq k=1}^{2} N_m \sigma_{mm}^{-2} \left(\alpha^{(m)}\right)^{T} \mathbf{K}^{(m,k)} \beta^{(k)} \\ -\sigma^{-2} \left(\sum_{i=1}^{2} \boldsymbol{\alpha}^{(i)^T} \boldsymbol{\alpha}^{(i)} - 2\boldsymbol{\alpha}^{(i)^T} \mathbf{m}^{(i)} \right) - \sigma^{-2} \left(\sum_{i=1}^{2} \sum_{i\neq j=1}^{2} \boldsymbol{\beta}^{(i)^T} \boldsymbol{\beta}^{(i)} - 2\boldsymbol{\beta}^{(i)^T} \mathbf{m}^{(i,j)} \right) \end{array} \right) \right]$$

Appendix D. Derivation of the Eq. (6.11)

Starting with the Eq. (6.10),

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \ln\left(p\left(\boldsymbol{\alpha}\,|\,\mathbf{D}\right)\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)p\left(\boldsymbol{\alpha}\right)\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \left[\ln\left(p\left(\mathbf{D}\,|\,\boldsymbol{\alpha}\right)\right)+\ln\left(p\left(\boldsymbol{\alpha}\right)\right)\right]\end{aligned} \tag{D.1}$$

Two terms in Eq. (D.1) can be written as

$$\begin{aligned}p\left(\boldsymbol{\alpha}\right) &= \left(\prod_{i=1}^{n} p\left(\boldsymbol{\alpha}^{(i)}\right)\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{n} p\left(\boldsymbol{\beta}^{(i,j)}\right)\right)\left(\prod_{i=1}^{n} p\left(\boldsymbol{\tau}^{(A_i)}\right)\right) \\ &= \left(\frac{1}{\left(2\pi\right)^{\frac{1}{2}\sum_{i=1}^{n} N_i}\ \sigma^{\sum_{i=1}^{n} N_i}}\right) e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\|\boldsymbol{\alpha}^{(i)}-\mathbf{m}^{(i)}\right\|_2^2} \times \\ &\quad \left(\frac{1}{\left(2\pi\right)^{\sum_{i=1}^{n} N_i}\ \sigma^{2\sum_{i=1}^{n} N_i}}\right) e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{\beta}^{(i,j)}-\mathbf{m}^{(i,j)}\right\|_2^2} \times \left(\frac{1}{\left(2\pi\right)^{\frac{nN_A}{2}}\ \sigma^{nN_A}}\right) e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\|\boldsymbol{\tau}^{(Ai)}-\mathbf{m}^{(A_i)}\right\|_2^2}\end{aligned}$$

$$\ln\left(p\left(\boldsymbol{\alpha}\right)\right) = \ln\left(\frac{1}{\left(2\pi\right)^{\frac{3}{2}\sum_{i=1}^{n} N_i+\frac{nN_A}{2}}\ \sigma^{3\sum_{i=1}^{n} N_i+nN_A}}\right) - \frac{1}{2\sigma^2}\left[\begin{array}{l}\left(\sum_{i=1}^{n}\boldsymbol{\alpha}^{(i)^T}\boldsymbol{\alpha}^{(i)} - 2\boldsymbol{\alpha}^{(i)^T}\mathbf{m}^{(i)} + \mathbf{m}^{(i)^T}\mathbf{m}^{(i)}\right)+ \\ \left(\sum_{i=1}^{n}\sum_{i\neq j=1}^{n}\boldsymbol{\beta}^{(i,j)^T}\boldsymbol{\beta}^{(i,j)} - 2\boldsymbol{\beta}^{(i,j)^T}\mathbf{m}^{(i,j)} + \mathbf{m}^{(i,j)^T}\mathbf{m}^{(i,j)}\right) \\ +\left(\sum_{i=1}^{n}\boldsymbol{\tau}^{(A_i)^T}\boldsymbol{\tau}^{(i)} - 2\boldsymbol{\tau}^{(A_i)^T}\mathbf{m}^{(A_i)} + \mathbf{m}^{(A_i)^T}\mathbf{m}^{(A_i)}\right)\end{array}\right]$$

$$\ln\left(p\left(\mathbf{D}\mid\boldsymbol{\alpha}\right)\right)=\ln\left(\prod_{i=1}^{n}\prod_{k=1}^{N_i}\frac{1}{\left(2\pi\right)^{\frac{\hat{p}}{2}}\sigma_{ii}^{\hat{p}}}e^{-\frac{1}{2\sigma_{ii}^{2}}\left\|\phi\left(\mathbf{x}_k^{(i)}\right)-\left(\sum_{m=1}^{N_i}\alpha_m^{(i)}\mathbf{x}_m^{(i)}-\sum_{l\neq i=1}^{n}\sum_{j=1}^{N_l}\beta_j^{(l,i)}\mathbf{x}_j^{(l)}-\sum_{j=1}^{N_A}\tau_j^{(A_i)}\mathbf{x}_j^{(A)}\right)\right\|_2^2}\right)$$

The results of $\ln\left(p\left(\mathbf{D}\mid\boldsymbol{\alpha}\right)\right)$ and $\ln\left(p\left(\boldsymbol{\alpha}\right)\right)$ are substituted in Eq. (D.1). Therefore, the Eq. (6.11) is obtained as follows:

$$\widehat{\boldsymbol{\alpha}}=\arg\max_{\boldsymbol{\alpha}}\left[2^{-1}\times\left(\begin{array}{l}
2\sum_{m=1}^{n}\sigma_{mm}^{-2}\left(\alpha^{(m)}\right)^T\mathbf{B}^{(m)}\mathbf{1}^{(m)}-2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}\sigma_{kk}^{-2}\left(\beta^{(m,k)}\right)^T\mathbf{B}^{(m,k)}\mathbf{1}^{(m,k)}\\
-\sum_{k=1}^{n}N_k\sigma_{kk}^{-2}\left(\alpha^{(k)}\right)^T\mathbf{K}^{(k)}\alpha^{(k)}-\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\beta^{(k,m)}\right)^T\mathbf{K}^{(k)}\beta^{(k,m)}\\
+2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\alpha^{(m)}\right)^T\mathbf{K}^{(m,k)}\beta^{(k,m)}-2\sum_{m=1}^{n}\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{B}^{(A,m)}\mathbf{1}^{(A,m)}\\
+2\sum_{m=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A,m)}\alpha^{(m)}\\
-2\sum_{m=1}^{n}\sum_{m\neq k=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A,k)}\beta^{(k,m)}-\sum_{m=1}^{n}N_m\sigma_{mm}^{-2}\left(\tau^{(A_m)}\right)^T\mathbf{K}^{(A)}\tau^{(A_m)}\\
-\sigma^{-2}\left(\sum_{i=1}^{n}\boldsymbol{\alpha}^{(i)^T}\boldsymbol{\alpha}^{(i)}-2\boldsymbol{\alpha}^{(i)^T}\mathbf{m}^{(i)}\right)-\sigma^{-2}\left(\sum_{i=1}^{n}\sum_{i\neq j=1}^{n}\boldsymbol{\beta}^{(i,j)^T}\boldsymbol{\beta}^{(i,j)}-2\boldsymbol{\beta}^{(i,j)^T}\mathbf{m}^{(i,j)}\right)\\
-\sigma^{-2}\left(\sum_{i=1}^{n}\boldsymbol{\tau}^{(A_i)^T}\boldsymbol{\tau}^{(A_i)}-2\boldsymbol{\tau}^{(A_i)^T}\mathbf{m}^{(A_i)}\right)
\end{array}\right)\right]$$

REFERENCES

ABDELLA, G. M., AL-KHALIFA, K. N., KIM, S., JEONG, M. K., ELSAYED, E. A. and HAMOUDA, A. M. 2017. Variable Selection-based Multivariate Cumulative Sum Control Chart. *Quality and Reliability Engineering International,* 33**,** 565-578.

ALESKEROV, E., FREISLEBEN, B. and RAO, B. Cardwatch: A neural network based database mining system for credit card fraud detection. Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997, 1997. IEEE, 220-226.

AMER, M., GOLDSTEIN, M. and ABDENNADHER, S. Enhancing one-class support vector machines for unsupervised anomaly detection. Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, 2013. ACM, 8-15.

AZZALINI, A. and CAPITANIO, A. 1999. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 61**,** 579-602.

AZZALINI, A. and DALLA VALLE, A. 1996. The multivariate skew-normal distribution. *Biometrika,* 83**,** 715-726.

BAKIR, S. T. 2004. A distribution-free Shewhart quality control chart based on signed-ranks. *Quality Engineering,* 16**,** 613-623.

BAKIR, S. T. 2006. Distribution-free quality control charts based on signed-rank-like statistics. *Communications in Statistics—Theory and Methods,* 35**,** 743-757.

BAKIR, S. T. and REYNOLDS, M. R. 1979. A nonparametric procedure for process control based on within-group ranking. *Technometrics,* 21**,** 175-183.

BOVOLO, F., CAMPS-VALLS, G. and BRUZZONE, L. 2010. A support vector domain method for change detection in multitemporal images. *Pattern Recognition Letters,* 31**,** 1148-1154.

CHANDOLA, V., BANERJEE, A. and KUMAR, V. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR),* 41**,** 15.

CHEN, N., ZI, X. and ZOU, C. 2016. A distribution-free multivariate control chart. *Technometrics,* 58**,** 448-459.

CHO, H.-W., JEONG, M. K. and KWON, Y. 2006. Support vector data description for calibration monitoring of remotely located microrobotic system. *Journal of Manufacturing Systems,* 25**,** 196-208.

CHOI, S. W., PARK, J. H. and LEE, I.-B. 2004. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Computers* and *chemical engineering,* 28**,** 1377-1387.

CROSIER, R. B. 1988. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics,* 30**,** 291-303.

CROWDER, S. V. 1989. Design of exponentially weighted moving average schemes. *Journal of Quality Technology,* 21**,** 155-162.

DAS, N. and PRAKASH, V. 2008. Interpreting the out-of-control signal in multivariate control chart—a comparative study. *The International Journal of Advanced Manufacturing Technology,* 37**,** 966-979.

DOGANAKSOY, N., FALTIN, F. W. and TUCKER, W. T. 1991. Identification of out of control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics-Theory and Methods,* 20**,** 2775-2790.

DUIN, R., JUSZCZAK, P., PACLIK, P., PEKALSKA, E., DE RIDDER, D., TAX, D. and VERZAKOV, S. 2000. A matlab toolbox for pattern recognition. *PRTools version,* 3**,** 109-111.

EFRON, B. and TIBSHIRANI, R. J. 1994. *An introduction to the bootstrap,* CRC press.

ERFANI, S. M., RAJASEGARAR, S., KARUNASEKERA, S. and LECKIE, C. 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition,* 58**,** 121-134.

EWAN, W. D. 1963. When and how to use cu-sum charts. *Technometrics,* 5**,** 1-22.

GAN, F. 1991. EWMA control chart under linear drift. *Journal of Statistical Computation and Simulation,* 38**,** 181-200.

GHASEMI, A., RABIEE, H. R., MANZURI, M. T. and ROHBAN, M. H. 2016. A bayesian approach to the data description problem. *arXiv preprint arXiv:1602.07507*.

GUPTA, A. K., GONZÁLEZ-FARíAS, G. and DOMíNGUEZ-MOLINA, J. A. 2004. A multivariate skew normal distribution. *Journal of multivariate analysis,* 89**,** 181-190.

HAWKINS, D. M. 1991. Multivariate quality control based on regression-adiusted variables. *Technometrics,* 33**,** 61-75.

HAWKINS, D. M. 1993. Regression adjustment for variables in multivariate quality control. *Journal of Quality Technology,* 25**,** 170-182.

HAWKINS, D. M. and OLWELL, D. H. 1997. Inverse Gaussian cumulative sum control charts for location and shape. *The Statistician***,** 323-335.

HODGE, V. and AUSTIN, J. 2004. A survey of outlier detection methodologies. *Artificial intelligence review,* 22**,** 85-126.

HUANG, G., CHEN, H., ZHOU, Z., YIN, F. and GUO, K. 2011. Two-class support vector data description. *Pattern Recognition,* 44**,** 320-329.

JANACEK, G. and MEIKLE, S. 1997. Control charts based on medians. *Journal of the Royal Statistical Society: Series D (The Statistician),* 46**,** 19-31.

JEONG, M. K., LU, J.-C. and WANG, N. 2006. Wavelet-based SPC procedure for complicated functional data. *International Journal of Production Research,* 44**,** 729-744.

JEONG, Y.-S., KANG, I.-H., JEONG, M.-K. and KONG, D. 2012. A new feature selection method for one-class classification problems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42**,** 1500-1509.

JIANG, W. and TSUI, K.-L. 2008. A theoretical framework and efficiency study of multivariate statistical process control charts. *IIE Transactions,* 40**,** 650-663.

JIANG, W., WANG, K. and TSUNG, F. 2012. A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. *Journal of Quality Technology,* 44**,** 209.

KANG, I., JEONG, M. K. and KONG, D. 2012. A differentiated one-class classification method with applications to intrusion detection. *Expert Systems with Applications,* 39**,** 3899-3905.

KANG, J. H. and KIM, S. B. 2013. A clustering algorithm-based control chart for inhomogeneously distributed TFT-LCD processes. *International Journal of Production Research,* 51**,** 5644-5657.

KANG, J. H., YU, J. and KIM, S. B. 2016. Adaptive nonparametric control chart for time-varying and multimodal processes. *Journal of Process Control,* 37**,** 34-45.

KANG, P. and CHO, S. 2012. Support vector class description (SVCD): Classification in kernel space. *Intelligent Data Analysis,* 16**,** 351-364.

KIM, B., JEONG, Y.-S., TONG, S. H., CHANG, I.-K. and JEONG, M.-K. 2016a. Step-down spatial randomness test for detecting abnormalities in DRAM wafers with multiple spatial maps. *IEEE Transactions on Semiconductor Manufacturing,* 29**,** 57-65.

KIM, J., AL-KHALIFA, K., JEONG, M., HAMOUDA, A. and ELSAYED, E. 2014. Multivariate statistical process control charts based on the approximate sequential χ 2 test. *International Journal of Production Research,* 52**,** 5514-5527.

KIM, J., AL-KHALIFA, K., PARK, M., JEONG, M., HAMOUDA, A. and ELSAYED, E. 2013. Adaptive cumulative sum charts with the adaptive runs rule. *International Journal of Production Research,* 51**,** 4556-4569.

KIM, J., JEONG, M. K., ELSAYED, E. A., AL-KHALIFA, K. and HAMOUDA, A. 2016b. An adaptive step-down procedure for fault variable identification. *International Journal of Production Research,* 54**,** 3187-3200.

KIM, S., JEONG, M. K. and ELSAYED, E. A. 2017. Generalized smoothing parameters of a multivariate EWMA control chart. *IISE Transactions,* 49**,** 58-69.

KIM, S. B., SUKCHOTRAT, T. and PARK, S.-K. 2011. A nonparametric fault isolation approach through one-class classification algorithms. *IIE Transactions,* 43**,** 505-517.

KUMAR, V. 2005. Parallel and distributed computing for cybersecurity. *IEEE Distributed Systems Online,* 6.

LEE, D. and LEE, J. 2007. Domain described support vector classifier for multi-classification problems. *Pattern Recognition,* 40**,** 41-51.

LEE, K., KIM, D.-W., LEE, D. and LEE, K. H. 2005. Improving support vector data description using local density degree. *Pattern Recognition,* 38**,** 1768-1771.

LEE, K., KIM, D.-W., LEE, K. H. and LEE, D. 2007. Density-induced support vector data description. *IEEE Transactions on Neural Networks,* 18**,** 284-289.

LEE, K., KIM, N. and JEONG, M. K. 2014. The sparse signomial classification and regression model. *Annals of Operations Research,* 216**,** 257-286.

LEE, S.-W., PARK, J. and LEE, S.-W. 2006. Low resolution face recognition based on support vector data description. *Pattern Recognition,* 39**,** 1809-1812.

LI, J., JIN, J. and SHI, J. 2008. Causation-Based T^ sup 2^ Decomposition for Multivariate Process Monitoring and Diagnosis. *Journal of Quality Technology,* 40**,** 46.

LI, K.-L., HUANG, H.-K., TIAN, S.-F. and XU, W. Improving one-class SVM for anomaly detection. Machine Learning and Cybernetics, 2003 International Conference on, 2003. IEEE, 3077-3081.

LI, S.-Y., TANG, L.-C. and NG, S.-H. 2010. Nonparametric CUSUM and EWMA control charts for detecting mean shifts. *Journal of Quality Technology,* 42**,** 209.

LIU, L., TSUNG, F. and ZHANG, J. 2014. Adaptive nonparametric CUSUM scheme for detecting unknown shifts in location. *International Journal of Production Research,* 52**,** 1592-1606.

LIU, R. Y. 1995. Control charts for multivariate processes. *Journal of the American Statistical Association,* 90**,** 1380-1387.

LOWRY, C. A., WOODALL, W. H., CHAMP, C. W. and RIGDON, S. E. 1992. A multivariate exponentially weighted moving average control chart. *Technometrics,* 34**,** 46-53.

LUCAS, J. M. and SACCUCCI, M. S. 1990. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics,* 32**,** 1-12.

MASON, R. L., TRACY, N. D. and YOUNG, J. C. 1995. Decomposition of T2 for multivariate control chart interpretation. *Journal of quality technology,* 27**,** 99-108.

MASON, R. L., TRACY, N. D. and YOUNG, J. C. 1997. A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology,* 29**,** 396.

MONTGOMERY, D. C. 2007. *Introduction to statistical quality control*, John Wiley and Sons.

MOYA, M. M., KOCH, M. W. and HOSTETLER, L. D. 1993. One-class classifier networks for target recognition applications. Sandia National Labs., Albuquerque, NM (United States).

MU, T. and NANDI, A. K. 2009. Multiclass classification based on extended support vector data description. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39**,** 1206-1216.

NGAI, H.-M. and ZHANG, J. 2001. Multivariate cumulative sum control charts based on projection pursuit. *Statistica Sinica***,** 747-766.

NING, X. and TSUNG, F. 2013. Improved design of kernel distance–based charts using support vector methods. *IIE transactions*, 45**,** 464-476.

PARK, C., PARK, C., REYNOLDS JR, M. R. and REYNOLDS JR, M. R. 1987. Nonparametric procedures for monitoring a location parameter based on linear placement statistics. *Sequential Analysis*, 6**,** 303-323.

PARK, M., KIM, J., JEONG, M., HAMOUDA, A., AL-KHALIFA, K. and ELSAYED, E. 2012. Economic cost models of integrated APC controlled SPC charts. *International Journal of Production Research*, 50**,** 3936-3955.

PARZEN, E. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33**,** 1065-1076.

PIGNATIELLO, J. J. and RUNGER, G. C. 1990. Comparisons of multivariate CUSUM charts. *Journal of quality technology*, 22**,** 173-186.

QIU, P. 2008. Distribution-free multivariate process control based on log-linear modeling. *IIE Transactions*, 40**,** 664-677.

ROBERTS, S. 1959. Control chart tests based on geometric moving averages. *Technometrics*, 1**,** 239-250.

RUNGER, G. C. 1996. Projections and the U 2 multivariate control chart. *Journal of Quality Technology*, 28**,** 313-319.

RUNGER, G. C., ALT, F. B. and MONTGOMERY, D. C. 1996. Contributors to a multivariate statistical process control chart signal. *Communications in Statistics--Theory and Methods*, 25**,** 2203-2213.

SCHÖLKOPF, B., PLATT, J. C., SHAWE-TAYLOR, J., SMOLA, A. J. and WILLIAMSON, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13**,** 1443-1471.

SHEWHART, W. A. 1931. *Economic control of quality of manufactured product*, ASQ Quality Press.

SHIN, K. S., JEONG, Y.-S. and JEONG, M. K. 2012. A two-leveled symbiotic evolutionary algorithm for clustering problems. *Applied Intelligence*, 36**,** 788-799.

SOTIRIS, V. A., PETER, W. T. and PECHT, M. G. 2010. Anomaly detection through a bayesian support vector machine. *IEEE Transactions on Reliability,* 59**,** 277-286.

SPENCE, C., PARRA, L. and SAJDA, P. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on, 2001. IEEE, 3-10.

STOUMBOS, Z. G. and SULLIVAN, J. H. 2002. Robustness to non-normality of the multivariate EWMA control chart. *Journal of Quality Technology,* 34**,** 260.

SUKCHOTRAT, T., KIM, S. B. and TSUNG, F. 2009. One-class classification-based control charts for multivariate process monitoring. *IIE transactions,* 42**,** 107-120.

SULLIVAN, J. H., STOUMBOS, Z. G., MASON, R. L. and YOUNG, J. C. 2007. Step-down analysis for changes in the covariance matrix and other parameters. *Journal of Quality Technology,* 39**,** 66.

SUN, R. and TSUNG, F. 2003. A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research,* 41**,** 2975-2989.

TAN, M. H. and SHI, J. 2012. A Bayesian approach for interpreting mean shifts in multivariate quality control. *Technometrics,* 54**,** 294-307.

TAX, D. M. and DUIN, R. P. 1999. Support vector domain description. *Pattern recognition letters,* 20**,** 1191-1199.

TAX, D. M. and DUIN, R. P. 2004. Support vector data description. *Machine learning,* 54**,** 45-66.

THORNHILL, N. F., PATWARDHAN, S. C. and SHAH, S. L. 2008. A continuous stirred tank heater simulation model with applications. *Journal of Process Control,* 18**,** 347-360.

TUERHONG, G. and KIM, S. B. A nonparametric fault isolation approach through hybrid novelty score. Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on, 2011. IEEE, 261-266.

VAPNIK, V. 1995. The nature of statistical learning theory Springer New York Google Scholar.

WANG, D., ZHANG, L. and XIONG, Q. 2016. A nonparametric CUSUM control chart based on the Mann-Whitney statistic. *Communications in Statistics-Theory and Methods*.

WANG, K. and JIANG, W. 2009. High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology,* 41**,** 247.

WOODALL, W. H. and ADAMS, B. M. 1993. The statistical design of CUSUM charts. *Quality Engineering,* 5**,** 559-570.

XIE, X. and SHI, H. 2012. Dynamic multimode process modeling and monitoring using adaptive Gaussian mixture models. *Industrial* and *Engineering Chemistry Research,* 51**,** 5497-5505.

XU, X., XIE, L. and WANG, S. 2014. Multimode process monitoring with PCA mixture model. *Computers* and *Electrical Engineering,* 40**,** 2101-2112.

XU, Y. and DENG, X. 2016. Fault detection of multimode non-Gaussian dynamic process using dynamic Bayesian independent component analysis. *Neurocomputing,* 200**,** 70-79.

YU, J. and QIN, S. J. 2008. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE Journal,* 54**,** 1811-1829.

YU, K., JI, L. and ZHANG, X. 2002. Kernel nearest-neighbor algorithm. *Neural Processing Letters,* 15**,** 147-156.

ZHAO, S. J., XU, Y. M. and ZHANG, J. 2004. A multiple PCA model based technique for the monitoring of processes with multiple operating modes. *Computer Aided Chemical Engineering,* 18**,** 865-870.

ZHAO, S. J., ZHANG, J. and XU, Y. M. 2006. Performance monitoring of processes with multiple operating modes through multiple PLS models. *Journal of process Control,* 16**,** 763-772.

ZHOU, M., LIU, L., GENG, W. and ZHOU, J. 2015. Multivariate Control Chart Based on Multivariate Smirnov Test. *Communications in Statistics-Simulation and Computation,* 44**,** 1600-1611.

ZI, X., ZOU, C., ZHOU, Q. and WANG, J. 2013. A directional multivariate sign EWMA control chart. *Quality Technology* and *Quantitative Management,* 10**,** 115-132.

ZOU, C., JIANG, W. and TSUNG, F. 2011. A LASSO-based diagnostic framework for multivariate statistical process control. *Technometrics,* 53**,** 297-309.

ZOU, C. and QIU, P. 2009. Multivariate statistical process control using LASSO. *Journal of the American Statistical Association,* 104**,** 1586-1596.

ZOU, C. and TSUNG, F. 2011. A multivariate sign EWMA control chart. *Technometrics,* 53**,** 84-97.

ZOU, C., WANG, Z. and TSUNG, F. 2012. A spatial rank-based multivariate EWMA control chart. *Naval Research Logistics (NRL),* 59**,** 91-110.