APPROVAL DOCUMENT

UNIVERSITI MALAYSIA PAHANG

CENTER FOR GRADUATE STUDIES

We certify that the thesis entitled Development of PCA-based Fault Detection System Based on Various of NOC Models for Continuous-based Process is written by Mohamad Yusup bin Abd Wahab. We have examined the final copy of this thesis and in our opinion it is fully adequate in terms of scope and quality for the award of the degree of Bachelor of Chemical Engineering and Natural Resources. We herewith recommend that it be accepted in fulfilment of the requirement for the degree of Bachelor specializing in Chemical Engineering and Natural Resources.

Name of External Examiner   Signature

Institution:

Name of Internal Examiner   Signature

Institution:

DEVELOPMENT OF PCA-BASED FAULT DETECTION SYSTEM BASED ON VARIOUS OF NOC MODELS FOR CONTINUOUS-BASED PROCESS

MOHAMAD YUSUP BIN ABD WAHAB

A thesis submitted in the fulfillment of the requirements for the award of the degree of Bachelor of Chemical Engineers

Faculty of Chemical and Natural Resources Engineering
University Malaysia Pahang

JULY, 2012

SUPERVISOR'S DECLARATION

"I hereby declare that I have read this thesis and in my opinion this thesis has fulfilled the qualities and requirements for the award of Bachelor of Chemical Engineering"

Signature : ………………………………….....

Name of Supervisor : Dr. Mohd Yusri bin Mohd Yunus

Position :

Date : 1st July 2012

STUDENT'S DECLARATION

I declare that this thesis entitled "Development of PCA-based Fault Detection System Based on Various of NOC Models for Continuous-based Process" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree"

Signature : ……………………………………………......

Name : MOHAMAD YUSUP BIN ABD WAHAB

ID Number : KA09026

Date : 1st July 2012

To my father and beloved mother

ACKNOWLEDGEMENT

In the name of Allah S.W.T the Most Beneficent and the Most Merciful. The deepest sense of gratitude to the Almighty for the strength and ability to complete this project. Infinite thanks I brace upon Him.

I would like to take this opportunity to express my sincere appreciation to my supervisor Dr Mohd Yusri bin Mohd Yunus, for encouragement, guidance, morale support, and critics in bringing this project fruition. Without him outstanding support and interest, this thesis would not been at the best it would right now.

I would also like to express my deepest appreciation to my mother whom always support me and motivate me to complete this final year project.

Last but not least, I am also indebt to Faculty of Chemical and Natural Resources Engineering for the usage of workstation lab for my research. My sincere appreciation also extends to all my colleagues, housemates, and friends whom had provided assistance at various occasions.

Finally to individuals who has involved neither directly nor indirectly in succession of this thesis. Indeed I could never adequately express my indebtedness to all of them. Thank you.

ABSTRACT

Multivariate Statistical Process Control (MSPC) technique has been widely used for fault detection and diagnosis. Currently, contribution plots are used as basic tools for fault diagnosis in MSPC approaches. This plot does not exactly diagnose the fault, it just provides greater insight into possible causes and thereby narrow down the search. Hence, the cause of the faults cannot be found in a straightforward manner. Therefore, this study is conducted to introduce a new approach for detecting and diagnosing fault via correlation technique. The correlation coefficient is determined using multivariate analysis techniques, namely Principal Component Analysis (PCA). In order to overcome these problems, the objective of this research is to develop new approaches, which can improve the performance of the present conventional MSPC methods. The new approaches have been developed, the Outline Analysis Approach for examining the distribution of Principal Component Analysis (PCA) scores, the Correlation Coefficient Approach for detecting changes in the correlation structure within the variables. This research proposed PCA Outline Analysis Control Chart for fault detection. The result from the conventional method and ne approach were compared based on their accuracy and sensitivity. Based on the results of the study, the new approaches generally performed better compared to the conventional approaches, particularly the PCA Outline Analysis Control Chart.

ABSTRAK

Kawalan Proses Statistik multivariat (MSPC) teknik telah digunakan secara meluas untuk mengesan kerosakan dan diagnosis. Pada masa ini, plot sumbangan digunakan sebagai alat asas untuk diagnosis kerosakan dalam pendekatan MSPC. Plot ini tidak tepat mendiagnosis kerosakan, ia hanya memberikan gambaran yang lebih besar ke dalam sebab-sebab yang mungkin dan dengan itu mengecilkan carian. Oleh itu, punca kesalahan yang tidak boleh ditemui dalam cara yang jelas dan mudah. Oleh itu, kajian ini dijalankan untuk memperkenalkan satu pendekatan baru untuk mengesan dan mendiagnosis kerosakan melalui teknik korelasi. Pekali korelasi ditentukan menggunakan teknik analisis multivariat, iaitu Analisis Komponen Utama (PCA). Bagi mengatasi masalah ini, objektif kajian ini adalah untuk membangunkan pendekatan baru, yang boleh meningkatkan prestasi kaedah konvensional yang hadir MSPC. Pendekatan baru telah dibangunkan, Analisis Pendekatan Ringkasan untuk memeriksa taburan Analisis komponen prinsipal (PCA) skor, Pendekatan Korelasi Pekali untuk mengesan perubahan dalam struktur korelasi dalam pembolehubah. Kajian ini mencadangkan PCA Panjang Analisis Carta Kawalan untuk mengesan kerosakan. Hasil dari kaedah konvensional dan pendekatan ne berbanding berdasarkan ketepatan dan sensitiviti mereka. Berdasarkan hasil kajian itu, pendekatan baru yang dilakukan secara umumnya lebih baik berbanding dengan pendekatan konvensional, terutamanya Carta Kawalan Analisis Rangka PCA.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

CHAPTER 1

INTRODUCTION

1.1 Background of Study

1.1.1 Statistical process control (SPC)

Statistical process control (SPC) involves using statistical techniques to measure and analyze the variation in processes. Most often used for manufacturing processes, the intent of SPC is to monitor product quality and maintain processes to fixed targets. Statistical quality control refers to using statistical techniques for measuring and improving the quality of processes and includes SPC in addition to other techniques, such as sampling plans, experimental design, variation reduction, process capability analysis, and process improvement plans. Therefore, the true meaning of SPC could also be represented conceptually by a diagram as shown in **Figure 1.1**, whereby those basic domains in **Table 1.1** have been integrated to serve as an integrated function, known as process monitoring.

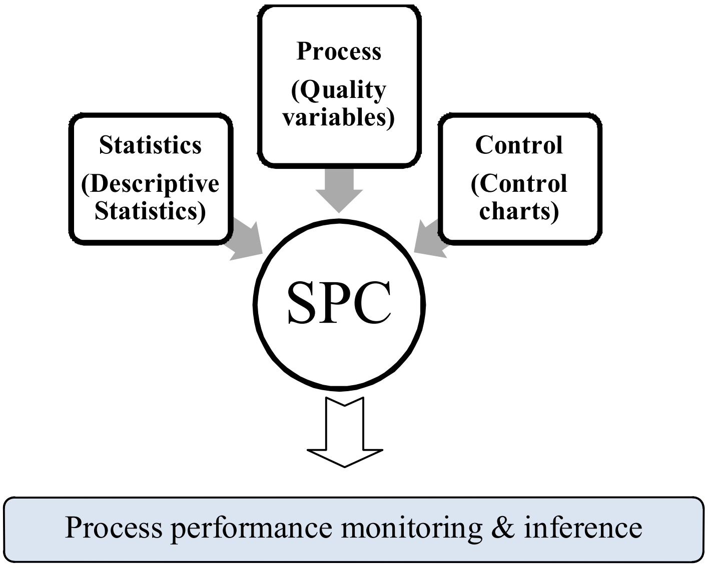

Figure 1.1: Concept of SPC

Table 1.1 Definitions of the Domains used in SPC

Domains	**Definition**
Statistical	Drawing conclusions using scientific or mathematical approach of analyzing data.
Process	The whole combination of people, equipment, materials, methods and environment working together to produce output; any work area that has identifiable, measurable output.
Control	Making something behave in a predictable consistent manner.

As the main focus is to monitor the process performance over time, functionally, it could also be used to detect those possible large process variations, or namely faults, in the quality products or outputs which can then leads to the prevention of fault sources from the inputs. In short, SPC is a useful technique especially to detect, diagnose and eventually performing process stabilization for any given process, in terms of providing an early warning for plant operators as well as gaining the deeper understanding on the process behaviour that tend to be achieved in the process. Due to this, from the commercial and quality assurance point of view, therefore, the aims of using SPC are mainly to (Wetherill and Brown, 1991):

i. improve quality (as a whole) by means of reduced manufacturing costs and increased customer satisfaction, and also,
ii. to increase productivity (process and products) by maximizing output value relative to inputs.

A primary tool used for SPC is the control chart, a graphical representation of certain descriptive statistics for specific quantitative measurements of the manufacturing process. These descriptive statistics are displayed in the control chart in comparison to their "in-control" sampling distributions. The comparison detects any unusual variation in the manufacturing process, which could indicate a problem with the process. Several different descriptive statistics can be used in control charts and there are several different types of control charts that can test for different causes, such as how quickly major vs. minor shifts in process means are detected. Control charts are also used with product measurements to analyze process capability and for continuous process improvement efforts. Over time, other process-monitoring tools have been developed, including:

- Cumulative Sum (CUSUM) charts: the ordinate of each plotted point represents the algebraic sum of the previous ordinate and the most recent deviations from the target.
- Exponentially Weighted Moving Average (EWMA) charts: each chart point represents the weighted average of current and all previous subgroup values, giving more weight to recent process history and decreasing weights for older data.

1.1.2 Multivariate Statistical Process Control (MSPC)

Multivariate Statistical Process Control (MSPC) (Kresta *et al.*, 1991; MacGregor, 1994; Kourti and MacGregor, 1995; MacGregor and Kourti, 1995; Kasonovich *et al.*, 1996; Wise and Gallagher, 1996; Wold *et al.*, 1999 and Lennox, 2001) is known generally as an upgraded technique, from which, it was emerged as a result of reformation in conventional Statistical Process Control (SPC) method. The significant enhancement made to MSPC is that it takes into consideration of multivariate behaviour among all variable relations simultaneously as opposed to the conventional schemes in SPC. In order to appreciate the distinctions between MSPC and SPC fundamentally, the first section will elaborate more on this issue according to their individual basic domains respectively. Next, the essentials of MSPC implementations are then to be discussed. In addition, an overview of the advanced projects conducted for MSPC are also presented subsequently.

Therefore, in order to apply MSPC successfully, those domains of SPC should be considered as well as understood especially from the angle of multivariate standpoint. Basically, SPC concerns with one basic question and that is 'given a sequence of data sets, did all come from the same population?' (Wheeler and Chambers, 1990).Therefore, as to answer this question thoroughly with respect to MSPC method, there are also other several related questions need to be answered fundamentally according to those three major fields of SPC, which are:

i. Descriptive statistics:
   a. Where the values are centred based on multivariate manner?
   b. How to measure the spread out of those multivariate values?
ii. Quality Variables
   a. How can the outcomes of the certain procedures be described when the system internally affect to each others?
   b. What are the techniques used to represent the values described by the multivariate descriptive statistics?
iii. Control charts
   a. How can the properties of the sample be used to infer the stability of the given population?
   b. What are the shapes of distribution of the values described by the multivariate descriptive statistics if the system has been randomized?
   c. What are the criteria to distinguish between two different populations under a single sample observation?

Hence, the following discussions are focusing on to answer those questions according to the MSPC groundwork. Before continue to those sections, as the mathematical basis of any multivariate analysis is on data matrices (Green and Carroll,1976; Chatfield and Collins,1980; Marvin,1982; Cox and Cox,1994; Borg and Groenen,1997) the followings show some of the references in presenting those mathematical symbols, that they might be relevance in this report:

i. For any given matrices, technically, they are denoted by boldface capital symbols.
ii. Boldface lowercase letters represent form for vectors.
iii. Every element of either matrices or vectors and other mathematical terms including scalar values will be written in the form of *italic lowercase*.

Thus, it obviously shows that SPC is not really a suitable monitoring method for any kind of multivariate process. As a result, the objectives of having MSPC is therefore should be understood based on the following arguments:

i. MSPC is the only method, of which, the data is treated simultaneously into a single monitoring by way of reducing the dimensionality of the data observed without losing any of important information.
ii. The technique is capable of capturing on the directionality of process variation information especially on how all the variables are behaving relative to one another. (MacGregor and Kourti, 1995).
iii. By using the multivariate method, the presence of noise levels could be reduced through averaging (MacGregor and Kourti, 1995).
iv. This approach can reduce the burden of constructing a large amount of single-variable control charts and enable detecting events that are impossible or difficult to detect from single-variable control charts (Phatak, 1999).

On top of those objectives, the original goals of SPC are also been considered as well as carried together, such a way that the ultimate aim is to increase productivity of high quality products for multivariate process, which is illustrated in **Figure 1.2**.

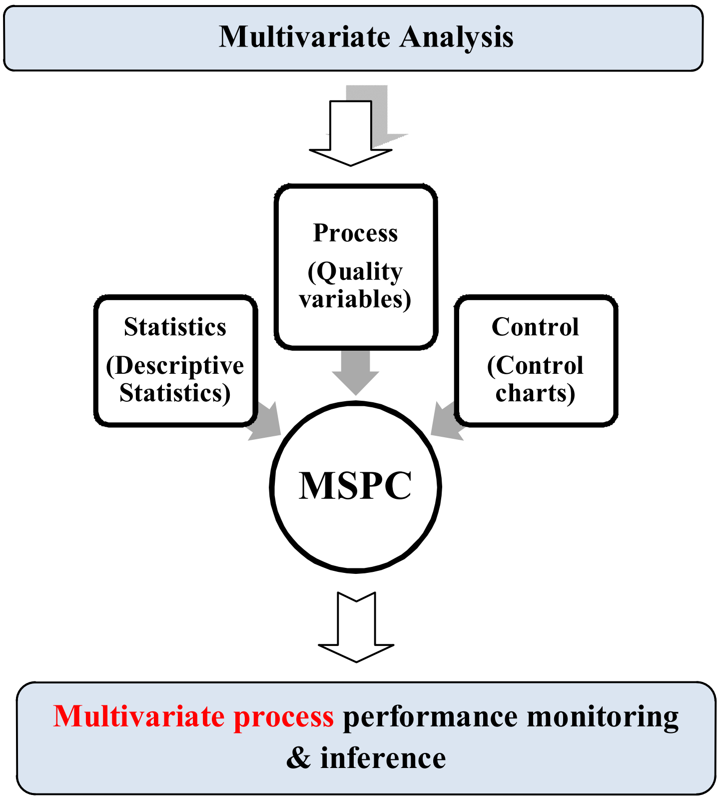

Figure 1.2: Concept of MSPC

1.2 Problem Statement

In order to ensure the successfulness of any operation, it is important to detect process upsets, equipment malfunctions or other special events as early as possible and then to diagnose and remove the factors that cause those events. However, Zhao et al., (2004) mentioned that a process which is having multiple operating modes tends trigger continuous warning signal even when the process itself is operating under another steady-state. SPC is rather complicated, in the sense that there are many individual control charts need to be monitored because the collection number of variables observed is also exist in a large quantity. MSPC is the only method, of which, the data is treated simultaneously into a single monitoring by way of reducing the dimensionality of the data observed without losing any of important information.

1.3 Objective of Study

The main purpose of this research is to study the impact of applying various modes of normal operating condition (NOC) in terms of the number of samples and variable variations on the process monitoring performance for continuous-based process. Therefore, the main objectives of this research are:

i. To develop the conventional MSPM method based on a single NOC
ii. To implement the conventional MSPM method based on different modes of NOC.
iii. To analyze the monitoring performance between system (i) and (ii).

1.4 Research Question

i. What is the main impact of reducing the number of samples as well as variations on the monitoring performance?
ii. What are the criteria should be used in selecting the NOC model?

1.5 Scope of study

Scope of propose study are on the development of PCA-based fault detection system based on various modes of NOC models for continuous-based process. There are three main scope will be investigated using MATLAB.

i. The conventional MSPM method will be develop based on single NOC mode. The linear PCA algorithm is used for reducing the multivariate data dimensions.

ii. The MSPM will be run traditionally by implementing different mode, which in this research is on two modes. According to Zhao et al. (2004),in spite of the success of applying PCA based MSPM tools to process data for detecting abnormal situations, when these tools are applied to a process with multiple operating modes, many missing and false alarms appear even when the process itself under other steady-state nominal operating conditions.

iii. As all data have been obtained, it will be analyze further with two multivariate control charts namely Hotelling's T2 and Squared Prediction Errors (SPE) statistic for the fault detection operation.

1.6 Rational and Significance

In this research, effort mainly concentrates on breaking through the current limitation and the further application of MSPC on a multivariate continuous chemical process. The main contributions of this research are:

1. Application of MSPC tools on the fault detection and diagnosis.
2. An Eigenvalue-eigenvector PCA approach had been used for developing Principals Components model.
   - Modified Process Fault Detection and Diagnosis, mechanisms are also developed based on the Outline Analysis.

1.7 Contributions

i. A new set of criteria is proposed for selecting the optimized NOC data for monitoring.

ii. As a result of (i), the monitoring performance can be enhanced in terms of missing and false alarm.

1.8 Organization of This Report

The new monitoring algorithm has been proposed in this study by developing PCA-based fault detection system based on various modes of NOC models for continuous-based process. Hence, this report is divided into five main chapters. The first chapter discusses the background of the works which includes the problem statement, objectives, scopes and contributions**.** **Chapter II** which is literature review describes the fundamental of MSPC and justification of applying PCA in MSPM frameworks. **Chapter III** explains the research methodology of this study. **Chapter IV** presents some of the preliminary results. Conclusions and further research works are given in **Chapter V.**

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

According to Venkatasubramaniam, Rengaswamy, Kavuri and Yin (2003) MSPM tools are data driven technique that generally reduce the dimension of process data and extract key features and trends that are of interest to plant personnel. MSPM tools used to reduces dimensions of process data, like PCA and subsequent refinements, which have show great success. In chapter 2, we will discuss on the fundamental or theory of process monitoring on MSPM using PCA tools, process monitoring issues and extension and justification of applying PCA in MSPM frameworks. Lastly, a summary is given at the end of this chapter.

2.2 Fundamental of PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is one of the most common multivariate analyses applied in the MSPC area (Jackson, 1991, MacGregor and Kourti, 1995; Zhang *et al.*, 1997, Gnanadesikan, 1997). In particular, PCA can be described by means of either using mathematical representation or graphical representation. Firstly, from the mathematical point of view, PCA is a multivariate projection technique, which can transform a set of original variables $[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m]$ to a set of new variables $[\mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_m]$. Generally, these newly formed variables are called Principal Components, PC for short (Jackson, 1991), whereby they build the individual linear combinations of the original variables which is simplified as follows:

$$\mathbf{P}_{n\times m} = \mathbf{X}_{n\times m}\mathbf{V}_{m\times m} \tag{2.1}$$

Initially, the original matrix, **X** has *m* variables with each variable has *n* number of measurements. The data are arranged in the form of *n* x *m*, where the measurements of a variable are organized in the form of a column vector, which is shown as follows:

$$\mathbf{X}_{n\times m} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \tag{2.2}$$

Next, **V** is known as the eigenvector matrix, whereby it gives the weighting function in forming the linear combinations of the original variables. The eigenvector matrix, **V**, contains eigenvectors or also known as loading vectors $\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_m$. Each eigenvector $\mathbf{v}_m$ is a column vector which contains the arrangement of elements $\mathbf{v}^T_m$ = $[v_{1,m}, v_{2,m} \ldots v_{m,m}]^T$, as denoted in the following matrix:

$$\mathbf{V}_{m\times m:} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_m] = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,m} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,m} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ v_{m,1} & v_{m,2} & \cdots & v_{m,m} \end{bmatrix} \tag{2.3}$$

Lastly, **P** is the principal components scores matrix, in which it contains *n* scores for each of the principal components, as given subsequently as:

$$\mathbf{P} = [\mathbf{P}_1 \quad \cdots \quad \mathbf{P}_m] \tag{2.4}$$

$$= \begin{bmatrix} x_{1,1}v_{1,1} + x_{1,2}v_{2,1} + \cdots + x_{1,m}v_{m,1} & \cdots & x_{1,1}v_{1,m} + x_{1,2}v_{2,m} + \cdots + x_{1,m}v_{m,m} \\ x_{2,1}v_{1,1} + x_{2,2}v_{2,1} + \cdots + x_{2,m}v_{m,1} & \cdots & x_{2,1}v_{1,m} + x_{2,2}v_{2,m} + \cdots + x_{2,m}v_{m,m} \\ \vdots & \vdots & \vdots \\ x_{n,1}v_{1,1} + x_{n,2}v_{2,1} + \cdots + x_{n,m}v_{m,1} & \cdots & x_{n,1}v_{1,m} + x_{n,2}v_{2,m} + \cdots + x_{n,m}v_{m,m} \end{bmatrix} \tag{2.5}$$

Thus, as far as multivariate calculations are concerned, **P**, will play the role of quality variables rather than individual variables of **X**, in the MSPC method. The original data, **X**, are also could be predictable backed from the calculated PC element, by which, the procedures are:

$$\mathbf{X}_{n\times m}\mathbf{V}_{m\times m} = \mathbf{P}_{n\times m} \tag{2.6}$$

$$\mathbf{X}_{n\times m}\mathbf{V}_{m\times m}\mathbf{V}^{T}_{mxm} = \mathbf{P}_{n\times m}\mathbf{V}^{T}_{mxm} \tag{2.7}$$

$$\mathbf{X}_{n\times m}\mathbf{I}_{m\times m} = \mathbf{P}_{n\times m}\mathbf{V}^{T}_{mxm} \tag{2.8}$$

$$\mathbf{X}_{n\times m} = [\mathbf{p_1} \quad \mathbf{p_2} \quad \cdots \quad \mathbf{p}_m]\begin{vmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \vdots \\ \mathbf{v}_m \end{vmatrix} \tag{2.9}$$

If all the PCs are used to represent the original variables, the original raw data matrix is reproduced back as shown in equation (2.24). However, in this situation, the purpose for using Principal Components Analysis as data dimension reduction technique will be lost. In order to maintain the uniqueness of this technique, only several PCs will be used to represent most of the original data variation. Therefore, if '*a*' of Principal Components are decided to be retained with $a < m$, then equation (2.24) can be adjusted and written as follows:

$$\mathbf{X}_{n\times m} = \mathbf{P}_{a}\mathbf{V}^{T}_{a} + \mathbf{P}_{m-a}\mathbf{V}^{T}_{m-a} \tag{2.10}$$

$$\mathbf{X}_{n\times m} = \mathbf{P}_{a}\mathbf{V}^{T}_{a} + \mathbf{E} \tag{2.11}$$

The retained principal components [p_1, p_2, ..., p_a], which form the P V term, are associated with systematic variation in data while the residual principal components [p_{a+1}, p_{a+2}, ..., p_m], which form the residual matrix E are considered of containing measurement errors (Seborg *et al.*, 1996). Therefore, PCA is a multivariate analysis technique that could use less number of newly formed variables to represent the original data variations without losing significant information, in which, information here is referred to data variation.

For the graphical representation of PCA, the linear combinations of the original variables in forming the new variables are actually representing selection of a new coordinate system with [P_1, P_2, ..., P_m] as the new axes obtained by rotating the original system with x_1, x_2, ..., x_n as the coordinate axes. The new axes represent the direction with maximum variability and provide a simpler and more parsimonious description of the variance-covariance matrix or correlation matrix (Johnson and Wichern, 1992). Figure 2.1 and 2.2 are prepared to give graphical representations of PCA

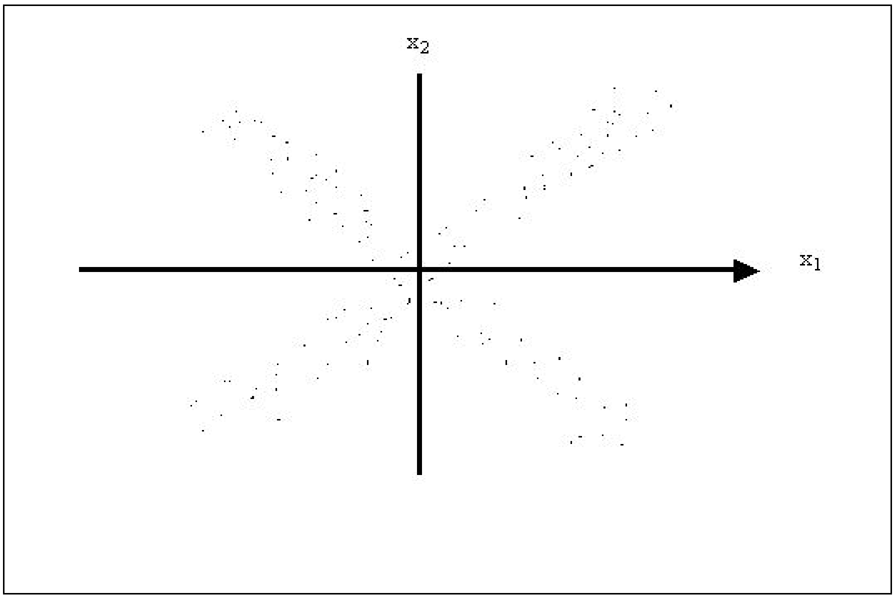

Figure 2.1: System for Two Variables Distribution

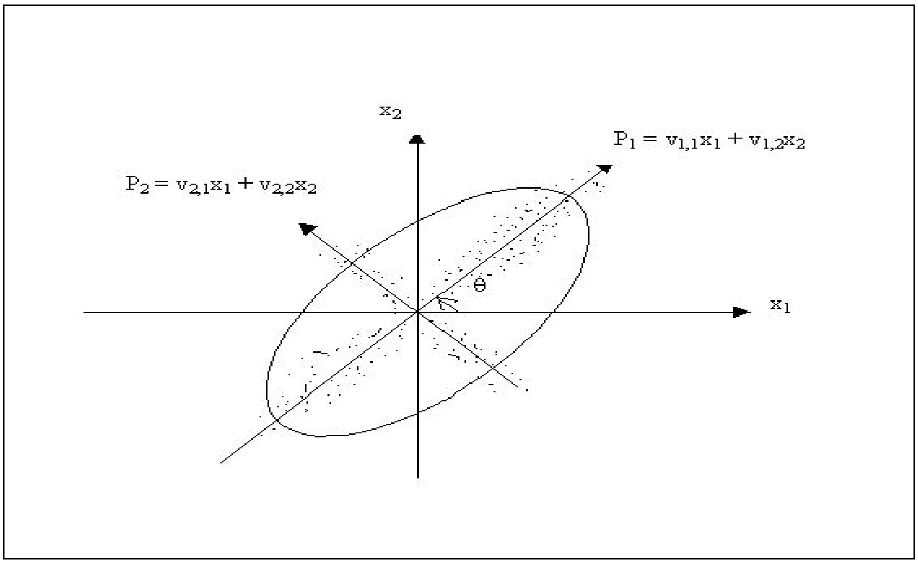

Figure 2.2: Graphical Representation of PCA

2.3 Fundamentals / Theory of Process Monitoring on MSPM Using PCA Tools

Reformation and upgrading of conventional Statistical Process Control (SPC) method has produce MSPC. MSPC tools such as principal component analysis (PCA) were used to reduce the explaining dimension of the process data. Maestri et al. say this method has show great success and particularly suited to data set comprising correlated and collinear variables. Ge and Song (2008) define process data as different group based, for instance, on variation in the operating capacity, seasonal variations or changes in the feedstock characteristics and also on modifications in the operation strategies. From a geometric point of view, whenever such as a change occurs, the process data tend to group into a new cluster in a different location in the high dimensional space containing the process normal operating region. However when the data is considered belong to a unique normal operating region, the volume of this region becomes incorrectly large. Zhao et al, (2006) say this region will lead to an increasing number of missing and false alarm. According to Zhao et al, (2004) when PCA based MSPC tools applied to a process with multiple operating modes, many missing and false alarm can appear even when the process itself is operating under other steady-state nominal operating conditions. Particularly this technique is for reducing the number of dimensions used from the original data as well as projected them into a number of uncorrelated variables, by means of forming the appropriate linear combinations of the original variables. Hence, MSPC is the only method where the data is treated simultaneously by way of reducing the dimensionality of the data observed without losing any of important information. In addition, this method can reduce the burden of constructing a large amount of single-variable control charts and enable detecting events that are impossible or difficult to detect from the single-variable control charts (Phatak, 1999).

According to Venkatasubramaniam et al, (2003) multivariate statistical techniques are powerful tool that capable to compressing data and reducing its dimensionality. Hence the essential information is retained and easy to analyze than the original huge data set. Moreover, it is able to handle noise and correlation to extract true information effectively. Initially, PCA method is proposed by Pearson (1901) later, it been develop by Hotelling (1947). This is a standard multivariate technique which has been including in many textbooks (Jackson, 1991; Anderson, 1984) and research paper (Wold, Esbensen and Geladi, 1987; Wold, 1978). Venkatasubramaniam et al, (2003) say PCA is based on orthogonal decomposition of the covariance matrix of the process variables along directions that explain the maximum variation of the data. Yu and Zhang say this method involved a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables, which are called principal component.

2.4 Extension Of Principal Component Analysis

There are many extension of Principle Component Analysis (PCA) which is some of these is Kernel of PCA, Multiway-PCA, Multi-Scale PCA, Moving PCA, Multi-Block PCA and many more.

2..4.1 Multiway-PCA

Conventional PCA is best for analyzing a two-dimensional matrix of data collected from a steady state process, containing linear relationships between the variables. Since these conditions are often not satisfied in practice, several extensions of PCA have been developed. Nomikos and MacGregor (1994) proposed Multi-Way PCA, which allows the analysis of a multi-dimensional matrix. Multi way method organized the data into time ordered block which each represent a single sample or process run. Three dimensional array data (I, batch samples x J, process variables x K, time) is decomposed to two dimensional array (I x JK) data for easier analysis (Wise and Gallagher, 1996). Projection of these three dimensions data into two dimensions makes this method suitable and widely applied for batch processes. Nomikos and MacGregor (1994) used simulated data obtained from a semibatch reactor to monitor the process.

Multi-Way PCA was applied to industrial batch polymerization reactor using Hotelling's T2 chart for fault detection and contribution plots for fault diagnosis

(MacGregor and Kourti, 1995; Nomikos, 1996; Kourti et al., 1996). Multi-Way PCA was also applied using process data collected from an industrial fed-batch fermentation process (Lennox et al., 2001). Lopes and Monezes (1998) applied this method for an industrial antibiotic production processes to detect faulty batches. Martin et al. (1999) used batch polymerization reactor to illustrate the implementation of Multi-Way PCA. Instead of using T^2 statistic, M^2statistics was used to determine the confidence bound for data not normally distributed. Martin and Morris (1996) proposed M^2 statistics that an empirical density based approach which the bounds calculated is based on density estimation.

Under the Multi-Way PCA monitoring framework, Q statistic is used as the fault detection tool to detect abnormal batch variation whereas contribution plots are used as the fault diagnosis tool to isolate the faulty process variables that are responsible for the out-of-control situation. The disadvantage of contribution plot is that, they cannot isolate the causes automatically without the presence of confidence limit. The plant personnel need to decide whether the out-of-control situation is due to single or multiple faulty process variables.

2.4.2 Multi-Block PCA

Extensions of basic PCA to handle very large processes via Multi-Block PCA were made by MacGregor et al. (1994). This method permits easier modeling and interpretation of a large matrix by decomposing it into smaller matrices or blocks. The Multi-Block PCA enables plant wide monitoring. The monitoring framework for fault detection and diagnosis can be viewed in Figure 2.1.

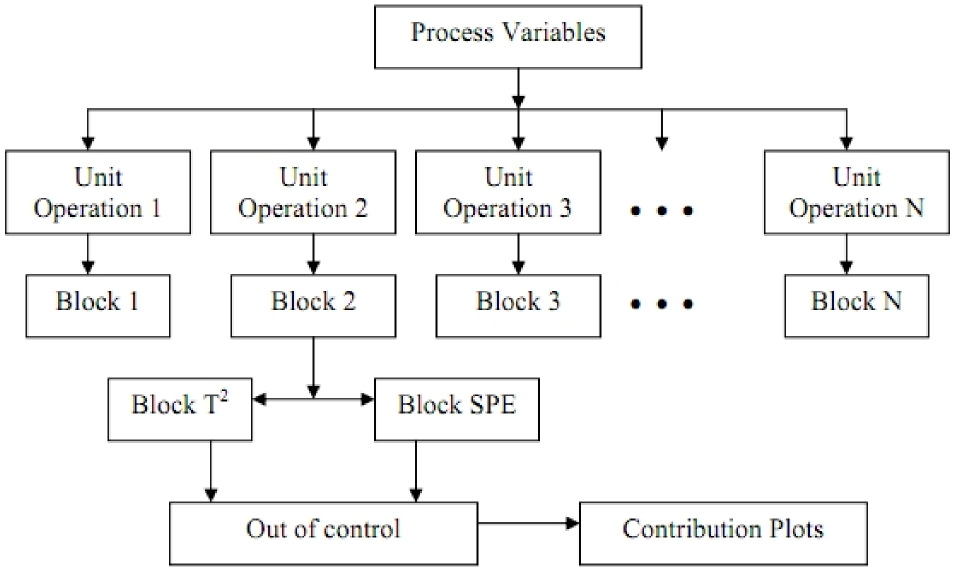

Figure 2.3: Multi-block PCA monitoring framework

The process variables are divided into several blocks with respective to specific unit operation. Block Hotellings' T^2 control chart and Q statistic control chart are used to detect out-of-control situation while contribution plots are used to isolate faulty process variables that cause the out-of-control situation. This approach relies on contribution plots for fault diagnosis; the shortcomings of this method still exist and could not offer complete fault isolation.

2.4.3 Moving PCA

The concept of Moving Principal Components Analysis is based on the idea that a change of correlation in between process variables can be detected by monitoring the directions of principal components (Kano et al., 2000a; Kano et al., 2001b). In order to evaluate the change of direction of each principal component, an index based of the inner product between two principal components is defined. The index proposed in MPCA contained information on the current PCs directions with reference PCs directions to detect any non-conformance situation to the reference models. Previous known fault data sets are used to construct PCA models corresponded to various kinds of fault situation to diagnose the fault cause. The drawbacks of this method are as follow:

i. There are a lot of PCA models, which representing various past fault situations are needed.
ii. Sufficient data have to be collected in order to construct PC models for each fault situation.

2.4.4 Dissimilarity, DISSIM

Kano et al. (2000d) proposed monitoring method based on process data distribution known as DISSIM. DISSIM method is based on the idea that a change of operating condition can be detected by monitoring a distribution of time-series data, which reflects the corresponding operating condition. The degree of dissimilarity between data sets is determined in DISSIM method (Kano et al., 2000a; Kano et al., 2000b; Kano et al., 2001a). Dissimilarity index was defined to evaluate the difference between two data quantitatively. Dissimilarity index control chart is used for fault detection. This index contains the information of the current data distribution with reference data distribution.

For fault diagnosis, each historical known fault data set is used to construct the known fault PCA models, which represents specific known fault data distribution. Besides, a similarity index is introduced to compare the current fault data distribution to each previous known fault PCA models. The proposed method is limited to PCs, which have similar variances because the index cannot function well if the PCs are changed. Other drawbacks of this method are sufficient data is required to construct every known fault PCA models and non-ability to isolate new fault. For fault diagnosis purpose, a contribution of each process variable to the dissimilarity index is introduced for identifying the variables that contribute significantly to an out-of-control value of the index (Kano et al., 2000c). However, there is difficult to identify exact fault causes for the process, which has many feedback control loops and the process variables are complicatedly related to each other.

2.4.5 Multi-Scale PCA

Bakshi (1998) developed Multi-Scale Principal Components Analysis, MSPCA by combining PCA and wavelet analysis. PCA has the ability to extract the relationship between the process variables and de-correlate cross correlation while wavelet analysis has the ability to extract events at different scales, compress deterministic features in a small number of relatively large coefficients, and approximately decorrelate a variety of stochastic processes (Bakshi, 1999). MS-PCA methodology determines separate PCA models at each scale to identify the scales where significant events occur. MS-PCA method has been applied for fault detection in industrial Fluidized Catalytic Cracker Unit, FCCU. Results showed that MS-PCA detects the faulty condition faster than conventional PCA using Hotelling's T^2statistic and Q statistic but the weakness of this method is that he didn't propose fault diagnosis method.

Kano et al. (2000a) applied MS-PCA to monitor problems of a simple two dimension matrix array data obtained from Tennessee Eastman Challenge process. Other researchers, Misra et al. (2002) proposed the combination of PCA and wavelet analysis. In essence, the MS-PCA approach is the same as proposed by Bakshi (1998). However, some differences have been introduced in their study such as multi-scale fault identification technique to identify the type of fault and sensor validation approach to serve as an early warning in case a fault of large magnitude is present. An industrial gas phase tubular reactor system used in this work for process fault diagnosis and sensor fault detection. The outcomes showed that the proposed method was able to detect and identify faults and abnormal events earlier than the conventional PCA approach. The disadvantage of this method is that, it requires basic understanding of the physical and chemical principles governing the process operation to help in clustering the highly correlated variables together before constructing the PCA model. Multi-scale fault identification does not provide the limits for contribution plot.

2.4.6 Kernel of PCA

Some extension of PCA is nonlinear principle components (NLPCA) or also Kernel PCA (KPCA). According to Vidal, Ma, and Sastry, (2005) KPCA is method of identifying a nonlinear manifold from sample points. NLPCA is a standard solution based on embedding the first data into a higher space, then applying PCA. As a result it will give large dimension space, so the eigen value is being decomposition or also known as kernel matrix.

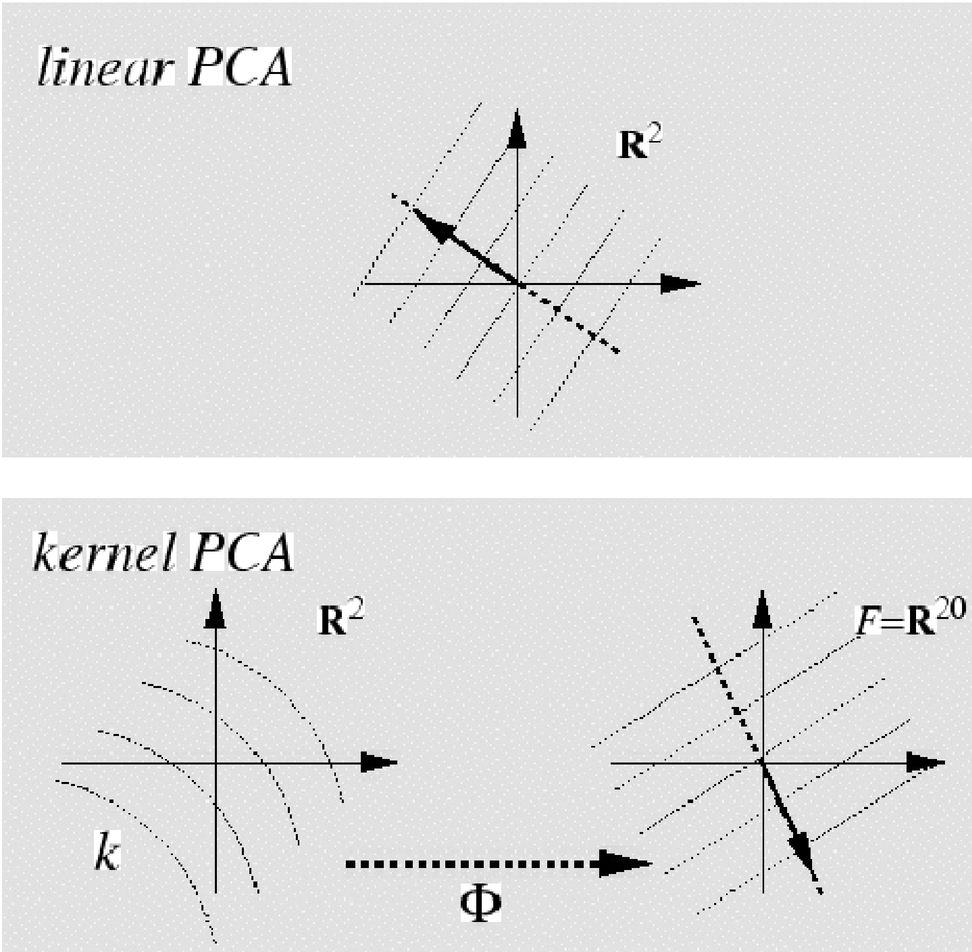

Figure 2.4: Linear PCA and Kernel PCA

From Figure 2.4 above, it show the basic idea of kernel PCA. By using a nonlinear function *k* instead of the standard d dot product, we implicitly perform PCA in a possibly high dimensional space F which is nonlinearly related to input space. The dotted lines are contour lines of constant feature value. Suppose that the number of observations *m* exceeds the input dimensionality *n*. In linear PCA, most samples are nonzero eigen values (Welling, nd). While for Kernel PCA variable will be nonzero eigen values. Thus, this is not necessarily a dimensionality reduction (Scholkopf, Smola and Muller, 2001). Furthermore, it may not be possible to find an exact preimage in input space of a reconstructed pattern based on a few of the eigenvectors. One of the disadvantages of KPCA is that, in practice, it is difficult to determine which kernel function to use because the choice of the kernel naturally depends on the nonlinear structure of the manifold to be identified (Vidal, Ma, and Sastry, 2005). In fact, learning kernels is an active topic of research in machine learning.

2.5 Extension of Multivariate Statistical Process Monitoring

Besides PCA, there also have several more extension of MSPM such as Projection to Latent Structures (PLS), Independent Component Analysis (ICA), Subspace Identification and many more.

2.5.1 Projection to Latent Structures (PLS)

Projection to latent structures or partial least squares (PLS) is a multivariable statistical regression method based on projecting or viewing the information in a high dimensional data space down onto a low dimensional one defined by some latent variables (Zhao et al., 2006). Abdi (2010) say PLS is a recent technique that combines features and generalizes PCA and multiple linear regressions. Zhao et al. (2000) support Abdi statement as PLS is one of the most powerful linear regression techniques to deal with noisy and highly correlated data. . Its goal is to predict a set of dependent variables from a set of independent variables or predictors. This prediction is achieved by extracting from the predictors a set of orthogonal factors called latent variables which have the best predictive power (Abdi, 2010).

PLS already has been successfully applied in diverse fields including process monitoring and quality control and identification of process dynamics & control with a limited number of observations available (Lee et al, 2006). When dealing with nonlinear systems, this approach assumes that the underlying nonlinear relationship between predictor data and response data can be approximated by quadratic PLS (QPLS) or neural network based PLS (NNPLS) while retaining the outer mapping framework of linear PLS algorithm and matrices were auto-scaled before they were processed by PLS algorithm (Wold, 2005). PLS model consists of outer relations which data are expressed in terms of their respective scores and inner relations that link the data to the data in the latent subspace. PLS finds the latent variables from the measured data by capturing the largest variance in the data and achieves the maximum correlation between the predictor variables and response variables.

A tutorial description along with some examples on the PLS model was provided by Geladi and Kowalaski (1986). PLS reduces the dimensionality of the measured data, finds the latent variables from the measured data by capturing the largest variance in the data and achieves the maximum correlation between the predictor X variables and response Y variables. In PLS based process dynamics, the inner relationship between variance and scores. The process dynamics in latent subspace could not be well identified by linear or quadratic relationships. For multivariable processes, the Partial least squares (PLS) controllers offer the opportunity to be designed as a series of SISO controllers (Qin and McAvoy (1992, 1993). Because of the diagonal structure of the dynamic part of the PLS model, input-output pairings are automatic. Series of SISO controllers designed on the basis of the dynamic models identified into latent subspaces and embedded in the PLS framework are used to control the process. Till date there is no reference on NNPLS controllers in the open literature though PLS & NNPLS based process identification, PLS controllers are well documented. The quality of the prediction obtained from a PLS regression model is evaluated with cross-validation techniques such as the bootstrap and jackknife. There are two main variants of PLS regression which is the most common one separates the roles of dependent and independent variables and the second one is used mostly to analyze brain imaging data that gives the same roles to dependent and independent variables.

2.5.2 Independent Component Analysis (ICA)

Hyvarinen (n.d) identified independent component analysis (ICA) is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. It is a generative models for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown (Lee et al., 2004). The latent variables are assumed non Gaussian and mutually independent and they are called the independent components of the observed data (Lee et al., 2003). ICA seeks to extract these independent components as well as the mixing matrix of coefficients.

Although ICA can be looked upon a useful extension of PCA, its objective differs from that of PCA. Bakshi (1998) say PCA is a dimensionality reduction technique that reduces the data dimension by projecting the correlated variables onto a smaller set of new variables that are uncorrelated and retain most of the original variance. However, its objective is only to correlate variables, not to make them independent. PCA can only impose independence up to second order statistics information which is mean and variance (Kano et al., 2004). While constraining the direction vectors to be orthogonal, whereas ICA has no orthogonality constraint and involves higher-order statistics, for an example it not only correlates the data for second order statistics but also reduces higher order statistical dependencies (Yoo et al., 2004). Hence, ICs reveal more useful information from observed data than principal components (PCs).

The data analyzed by ICA could originate from many different kinds of application fields, including digital images, document databases, economic indicators and psychometric measurements (Kano et al., 2004). In many cases, the measurements are given as a set of parallel signals or time series; the term blind source separation is used to characterize this problem. Typical examples are mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, or parallel time series obtained from some industrial process (Lee at al., 2004). A number of applications of ICA have been reported in speech processing, biomedical signal processing, machine vibrationanalysis, nuclear magnetic resonance spectroscopy, infrared optical source separation, radio-communications, and so on (Girolami, 1999). Kano et al. (2004) say the key idea or motivation of using ICA is that the monitoring performance can be improved by focusing on essential variables that drive a process.

2.5.3 Subspace Identification

Subspace methods or also known as 4SID or Subspace State Space System Identifcation are used to find linear state space model from experimental data as a relatively new alternative to widely used regression methods such as ARX, ARMAX and many more (Trnka and Havlena, 2005). Subspace identification algorithms are widely known and appreciated for their quick and reliable estimation of linear models based on available input/output measurements (Goethals et al. 2004). Much of the reliability of subspace identification algorithms is attributed to the fact that a model is obtained, solely by using numerically reliable matrix and vector manipulations such as projections and singular value decompositions. This is different from the classical predictor error methods, which usually involve with the minimization of a non-convex cost-function. There will be no guarantee that the obtained local minimum yields a good model. Treasure et al. (2003) say subspace identification can determine a set of state variables for describing process dynamics, produce a reduced set of variables to monitor process performance and offer contribution charts to diagnose anomalous behavior. This is demonstrated by an application study to a realistic simulation of a chemical process (Treasure et al., 2003). This point of view on subspace methods will show suitable fields for their application, where we can take advantage of their good properties like numerical robustness implemented by QR and SVD factorization, implicit rank reduction, non-iterative algorithm and few user parameters (Trenka et al., n.d.).

Goethals et al. (2004) say although subspace identification algorithms are fast and robust, a major drawback is that their use is mostly restricted to the class of linear systems. Some attempts to extend the use of subspace identification algorithms to nonlinear systems have been made in the past for general nonlinear systems, or more restricted model structures such as bilinear models, Wiener models, and Hammerstein models. Overschee and Moor (1996) say that Subspace identification algorithms are based on concepts from system theory, (numerical) linear algebra and statistics. The subspace identification approach does not suffer from any of these inconveniences. The only parameter to be user-specified is the order of the model, which can be determined by inspection of certain singular values.

Ljung et al. (1993) say when implemented correctly, subspace identification algorithms are fast, despite the fact that they are using QR and singular value decompositions. As a matter of fact, they are faster than the "classical" identification methods, such as Prediction Error Methods, because they are not. Hence there are also no convergence problems. Moreover, numerical robustness is guaranteed precisely because of these well-understood algorithms from numerical linear algebra. As a consequence, the user will never be confronted with hard-to-deal-with-problems such as lack of convergence, slow convergence or numerical instability (Chiuso et al., 2004).

2.6 Summary

In conclusion, this chapter review and discuss on fundamental of MSPM using PCA tools, and some of extension of PCA and MSPM. Moreover Chapter three will illustrates the basic of methodology and techniques applied to achieve the objectives.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Generally MSPC procedure could be categorized into two phases. For phases 1, it involves the off-line monitoring operation, while for phases 2 it involves on-line monitoring operation. However, for this study before undergo this two phase, the normal operating condition sample data is structured based on the assumption that has been made. The Figure 3.1 below show the procedures involves in MSPC to detect, identified, diagnosis and recover the fault detection.

The complete procedures of fault detection and identification comprise of two main phases namely as off-line modelling and monitoring (phase I) and on-line monitoring (phase II).

The main steps of MSPM system:

i. Fault detection: to designate the departure of observed samples from an acceptable range using a set of parameters.
ii. Fault identification: identifying the observed process variables that are most relevant to the fault which is typically identified by using the contribution plot technique.
iii. Fault diagnosis: specifically determines the type of fault which has been significantly (and should be also validated) contribute to the signal.
iv. Process recovery: remove the cause(s) that contribute to the detected fault.

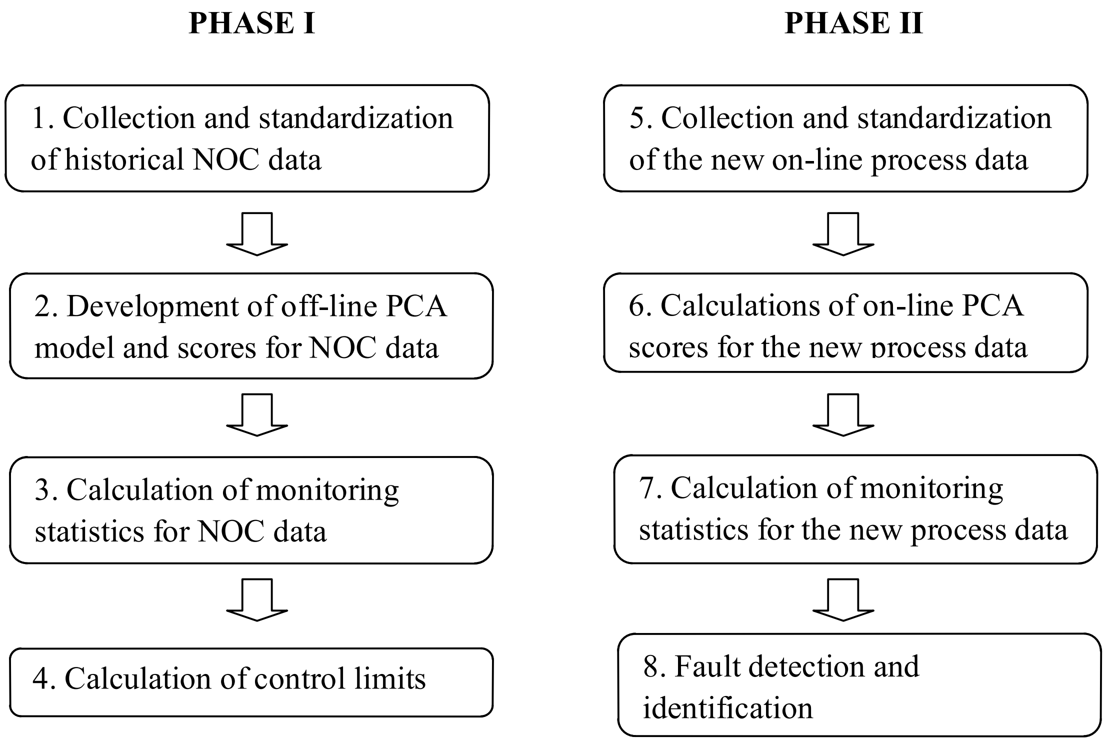

Figure 3.1: Two main phases namely as off-line modelling and monitoring (phase I) and on-line monitoring (phase II)

3.2 Phase I: Off-line Modelling and Monitoring

- Firstly, a set of normal operation condition (NOC) data, $\mathbf{X}_{n\times m}$ (n: samples, m: variables), are identified off-line based on the historical process data archive.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \tag{3.1}$$

- NOC simply implies that the process is operated at the desired setting condition and produces satisfactory products that meet the qualitative as well as quantitative specified standard (Martin et al., 1996).

- Then, the data are then standardized to zero mean and unit variance with respective to each of the variables because PCA results depend on data scales.

$$\breve{x}_{j,i} = \frac{\left(x_{j,i} - \bar{x}_i\right)}{\sigma_i} \tag{3.2}$$

- In the second step, the development of PCA model for the NOC data requires the establishment of a set of variance-covariance matrix, $\mathbf{C}_{mxm}$.

$$C = \frac{1}{n-1}\breve{X}'\breve{X} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,m} \end{bmatrix} \tag{3.3}$$

- **C** is then transformed into a set of basic structures of eigen-based formula.

$$C = V \Lambda V^{T} \tag{3.4}$$

- Finally, the PCA model of can be simply developed by:

$$\mathbf{P} = \breve{\mathbf{X}}\ \mathbf{V} \tag{3.5}$$

$$\mathbf{P} = \left[\mathbf{p_1} \quad \cdots \quad \mathbf{p_m}\right]$$
$$= \begin{bmatrix} \breve{x}_{1,1}v_{1,1} + \cdots + \breve{x}_{1,m}v_{m,1} & \cdots & \breve{x}_{1,1}v_{1,m} + \cdots + \breve{x}_{1,m}v_{m,m} \\ \vdots & \cdots & \vdots \\ \breve{x}_{n,1}v_{1,1} + \cdots + \breve{x}_{n,m}v_{m,1} & \cdots & \breve{x}_{n,1}v_{1,m} + \cdots + \breve{x}_{n,m}v_{m,m} \end{bmatrix} \tag{3.6}$$

- The following equation presents a measure of data variations captured by the first a principal components (Jolliffe, 2002).

$$k = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_a}{\lambda_1 + \lambda_2 + \ldots + \lambda_a + \ldots + \lambda_m} \tag{3.7}$$

- The third step basically involves calculation of the Hotelling's T^2 and SPE monitoring statistics.

$$T_i^2 = \sum_{j=1}^{a} \frac{p_{i,j}^2}{\lambda_j} \tag{3.8}$$

$$\begin{aligned} \tilde{\mathbf{E}} &= \breve{\mathbf{X}} - \hat{\mathbf{X}} \\ &= \breve{\mathbf{X}} - \mathbf{P}_a \mathbf{V}_{\mathbf{a}}^{\mathrm{T}} \\ &= \breve{\mathbf{X}} - \breve{\mathbf{X}} \mathbf{V}_a \mathbf{V}_a^{\mathbf{T}} \\ &= \breve{\mathbf{X}} \left(\mathbf{I} - \mathbf{V}_a \mathbf{V}_a^{\mathbf{T}} \right) \end{aligned} \tag{3.9}$$

$$SPE_i = \tilde{\mathbf{e}}_{\mathbf{i}} \tilde{\mathbf{e}}_{\mathbf{i}}^T \tag{3.10}$$

- The final task in phase I (4th step) deals with developing the control limits for both of the statistics.

$$T_\alpha = \frac{A(n-1)}{(n-A)} F_{A,n-A,\alpha} \tag{3.11}$$

$$SPE_\alpha = \theta_1 \left(\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_1 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \tag{3.12}$$

3.3 Phase II: On-line Monitoring

For the second phases, all the first phases of fault detection procedure will be developed for on-line application. However for the last step which is step 8 there are two main operations which have to be conducted separately which is fault detection and fault identification. Fault detection is a result of an occurrence of a special event that is not in conformance to the common cause nature. This fault detection will be declared if monitoring statistics exceeding its respective control limit for a pre-defined successive number of samples. While, for the fault identification is based on contribution plot. Typically, the circumstance of the on-line MSPC application always involves monitoring the real-time states of process condition. In other words, the dynamic behavior of the operation conditions should be analyzed in a real time manner to reflect the process status immediately. Even though, the procedures are still considering, more or less, the major steps taken during the off-line application development previously. There are two main operations which have to be conducted separately - fault detection and fault identification.

Fault detection:

- A fault situation is regarded as a result of an occurrence of a special event that is not in conformance to the common cause nature .
- Technically, a fault situation will be declared if either of the monitoring statistics exceeding its respective control limit for a pre-defined successive number of samples.

3.4 Summary

In conclusion, there are four main steps of MSPC system which is fault detection, fault identification, fault diagnosis and process recovery. This method will designated the departure of observed samples from an acceptable range using a set of parameters. Moreover, it will identify the observed process variables that most relevant to the fault by using the contribution plot technique. This will specifically determine the type of fault which has been significantly contributed to the signal.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

Initially some work has been done, that specially emphasize on the integration of process monitoring algorithms based on the conventional PCA. Hence, several works are represented in this chapter. Firstly, a description on the case study is briefly explained for the particular analysis. Next, is discussion on the result of NOC data and the tested fault data. The tested data are discussed for abrupt and incipient. Finally, is the summary of the chapter.

4.2 Case Study of CSTR System

This case study is looking on a continuous stirred tank reactor (CSTR) system, which the schematic diagram of a simulation is shown in Figure 4.1 below (Zhang, 2006).

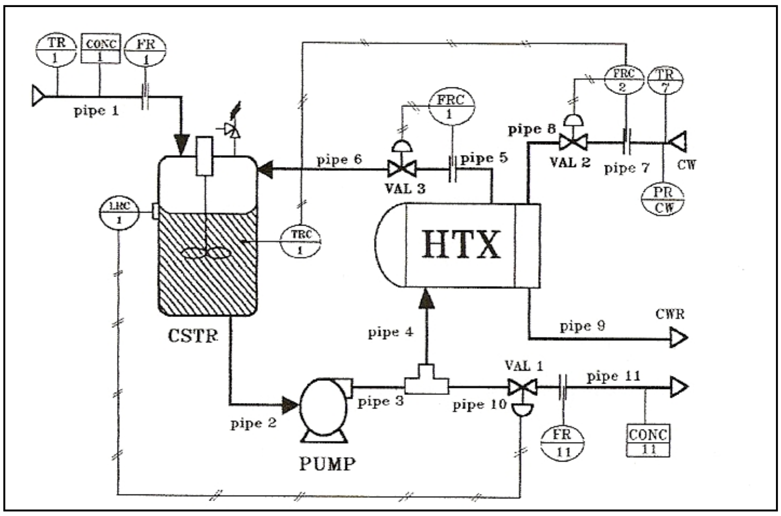

Figure 4.1: CSTR system

In this system, it applied an irreversible heterogeneous catalytic exothermic reaction between reactant A and product B in the reactor vessel. This system purpose to maintain the product concentration by installing three separated control loops in the process, which it consist of tank temperature, tank level and recycling flow variables. The fed of cold water flow rate to the heat exchanger is manipulating via a cascade control, in order to control temperature in the reactor. Moreover, the flow rate of the product is controlled to maintain reactor level. Finally, the recycle flow rate is indirectly manipulating the controller to maintain product composition in the reactor. A recent study shows that a set of multiple neural networks algorithms has been developed to enhance the reliability of fault diagnosis operation for this system (Zhang, 2006).In this process, there are ten on-line measured process variables and three controller outputs. As a result, thirteen on-line information sources are considered as listed in Table 4.1.

Table 4.1: List of variables in the CSTR system for monitoring

Process			Instruments		
No.	Variables	Variable Names	No.	Variables	Variable Names
1	V1	Tank temperature	11	V11	Controller 1
2	V2	Tank level	12	V12	Controller 3
3	V3	Feed temperature	13	V13	Controller 2
4	V4	Inlet flow rate			
5	V5	Recycle flow rate			
6	V6	Outlet flow rate			
7	V7	Cooling water flow rate			
8	V8	Product concentration			

9	V9	Feed concentration
10	V10	Heat exchanger entrance pressure

On top of those objectives, the original goals of SPC are also been considered as well as carried together, such a way that the productivity of multivariate process monitoring is improved.

Fault detection and diagnosis become more important in chemical industry because it presented a better way of fault handling: Monitoring and management. For this purpose, a good Fault Detection and Diagnosis mechanism must fulfill the characteristics such as early prediction, sensitive and accurate. As a major Fault Detection and Diagnosis tools, MSPC play a significant rule in multivariate data processing. MSPC had been used to generate the relationship among the interacted variable and reduce the complexity of the data matrix. This part is more focus on the application and developing of PCA. Several PCA extensions are developed and applied in chemical engineering area. The result reported form other researchers shown that, the performance of the MSPC is improving.

4.3 Normal Operating Condition Data Collection

This simulation has identified and collects a set of normal operating condition. This set of NOC data containing 100 measurements of 13 variables. By using PCA algorithm, the standardized NOC data is analysed. This analysis is performing to identify the required number of PCs as to reduce the dimension of the multivariate data.

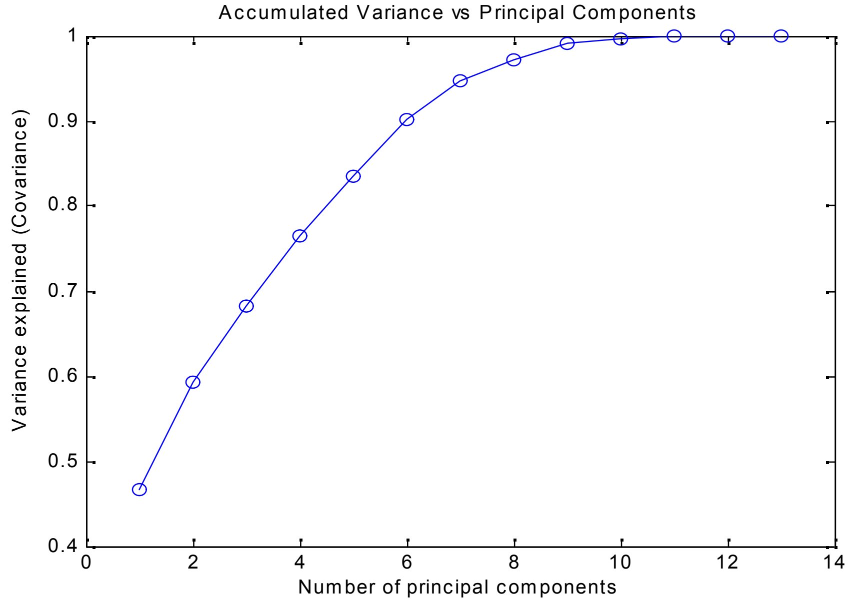

Figure 4.2: Accumulated data variance explained by different PCs

From figure 4.2 above, show the accumulated data variance explained by different PCs. Its shows that at least six PCs are needed to represent over 90% of the total NOC data variance and four PCs are required to explain over 80% of the total variance. Therefore, for this particular case study, six PCs are retained in the PCA model for the calculation of NOC scores. The Hoteling's T2 statistic and SPE statistic was then to be calculated and plotted together with the 95% and 99% confidence limits.

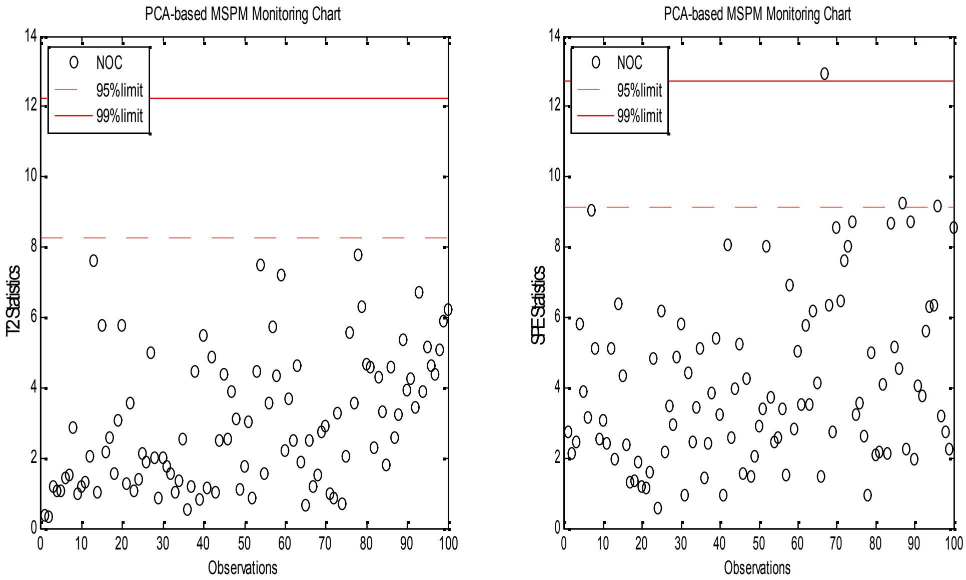

Figure 4.3: (a) T2 statistic for NOC data and (b) SPE statistic for NOC data

Figure 4.3 illustrate graph of NOC data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the NOC data is below the confident limits however foe SPE there are one sample outside the limit boundaries. The data is still normal as the fault is considered when 3 samples in series are out from the boundaries. Thus, T2 statistic and SPE statistics illustrate normal operating condition.

This NOC data has been test by reduce the sample using 50 measurements with 13 variables. By using PCA algorithm, the standardized NOC data test is analysed. This analysis objective is to identify the required number of PCs as to reduce the dimension of the multivariate data. Same with NOC data, the Hoteling's T2 statistic and SPE statistic was calculated and plotted together with the 95% and 99% confidence limits.

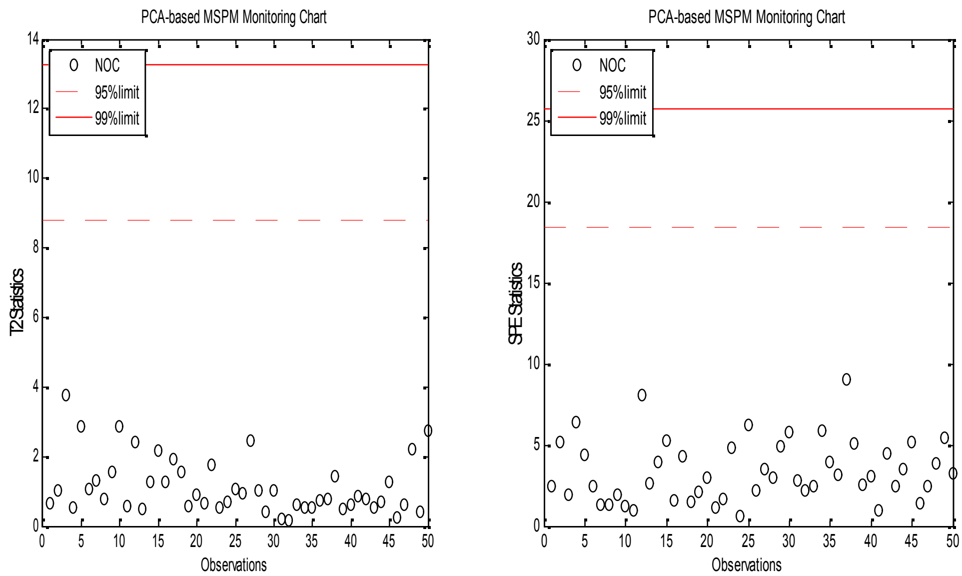

Figure 4.4: (a) T2 statistic for NOC data and (b) SPE statistic for NOC data

Figure 4.4 show statistic for NOC data test which is T2 statistics and SPE statistics. This figure show that for both statistics the process are normal as all the statistics NOC data test is below the confident limit.

4.4 Fault data collection

The system also subjects to be affected from several malfunction conditions as summarized in Table 4.2.

Table 4.2: List of abnormal operations in CSTR

Fault Cases	Fault Causes
1	Pipe 2 or 3 is blocked or pump fails
2	Pipe 10 or 11 is blocked or control valve 1 fails low

For each fault presented in Table 4.2, both abrupt and incipient faults are considered. An abrupt fault indicates a sudden change or step change in a process variable or parameter and typically it maintains over the operation time until the cause is completely removed. Detecting this kind of malfunctions should be easy for any multivariate monitoring system as the deviations are usually very obvious. On the other hand, an incipient fault depicts a kind of fault that gradually deviates from the normal setting. Thus, the monitoring system typically takes a while to detect these particular abnormal behaviours. In particular, all the faults were introduced at sample 2 and the sampling time was fixed at 4 seconds.

By concentrated to fault 1 and 2 for both abrupt and incipient, analyzing on the contributed variable to the fault and the impact will be discussed. Firstly, abrupt for fault 1 show the contributed variable are variable 10 which is heat exchanger entrance pressure. From figure 4.5(a) i when the pump fail or the pipe 2 or 3 is blocked, the flow rate enter the heat exchanger entrance pressure are low. From the fault data that have been run also show that variable 10 are first detected and have a high contributed to the fault. While, the impact are variable 5 which is recycle flow rate. As the flow rate of water are low due to pipe 2 or 3 or pump fail hence the recycling flow rate are also low. This impact and contributed variable to fault are show on figure 4.5 (a) ii.

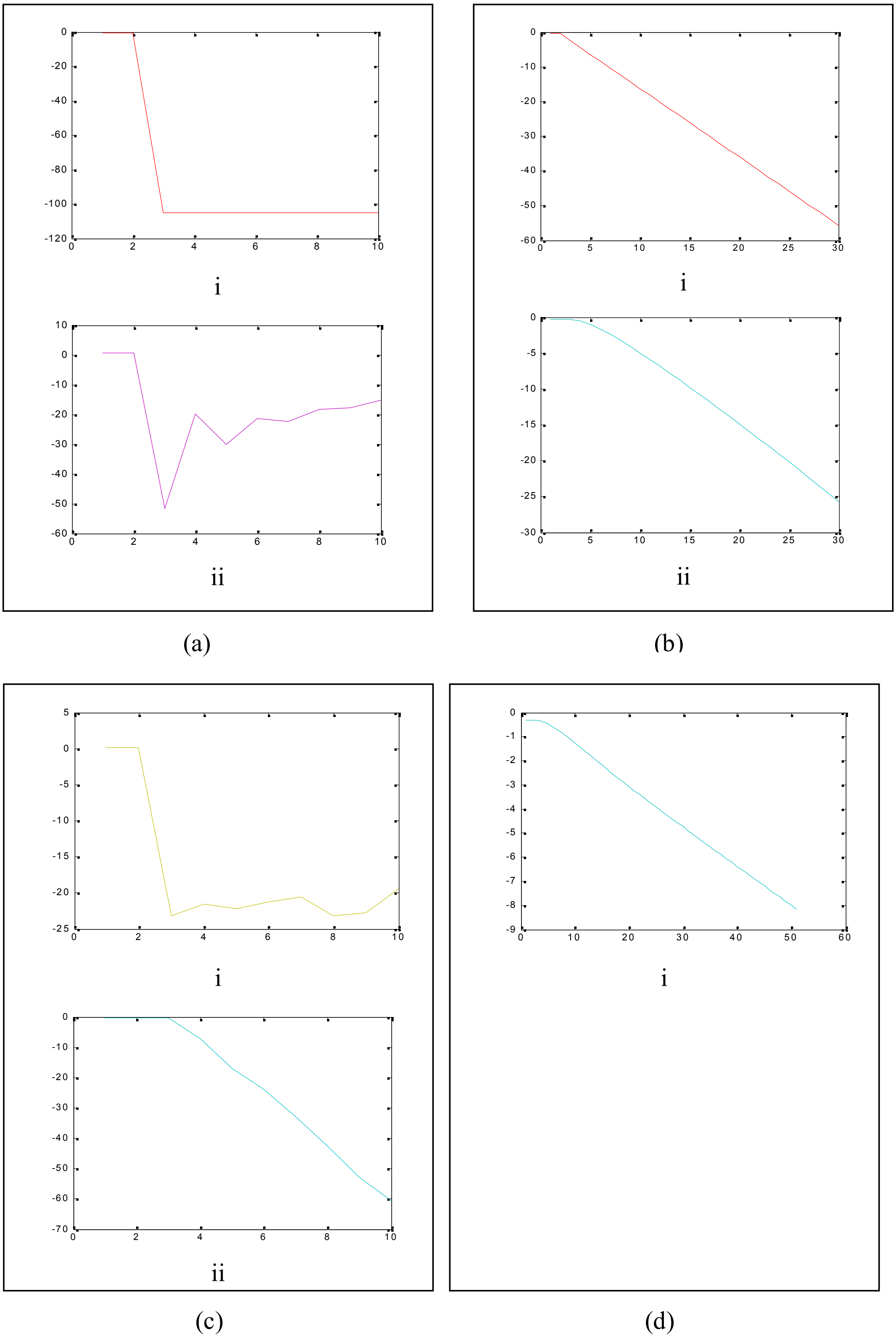

Figure 4.5: (a) Fault 3 abrupt (b) Fault 3 incipient (c) Fault 4 abrupt (d) Fault 4 incipient

Next, for the incipient Fault 1 also heat exchanger entrance pressure which is variable 10 are show high contributed to fault as shown in figure 4.5 (b) i. But however, the impact is at variable 11 which is valve 1 as shown on figure 4.5 (b) ii. As the water flow rate decrease because of pump fail or pipe 2 or 3 blockage hence, the valve will open little and outlet flow rate are decreasing.

While for the fault 4 abrupt, the figure 4.5 (c) i show the variable 6 which is outlet flow rate become low due pipe 10 or 11 is blocked or control valve 1 fails low. This has result the outlet flow rate to decrease as the pipe is failed or blockage. Meanwhile the impact is at variable 11 which is valve 1 as show on figure 4.5 (c) ii. As the pipe 10or 11 fail the valve 1 will open little and resulting the flow rate decrease. Moreover, for incipient fault 4, variable 11 which are valve 1 shown the most contributed variable show at figure 4.5 (d). Same with abrupt, because the pipe 10 or 11 is blocked or control valve 1 fails low, it result the outlet flow rate to decrease.

Result for T^2 and SPE for the the fault no.1

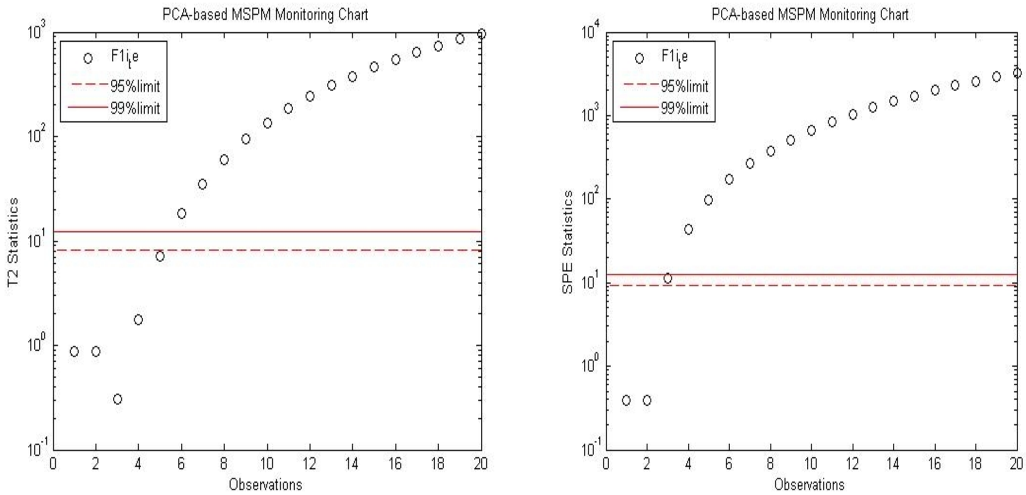

Figure 4.6: Results of PCA-based MSPM system for incipient fault

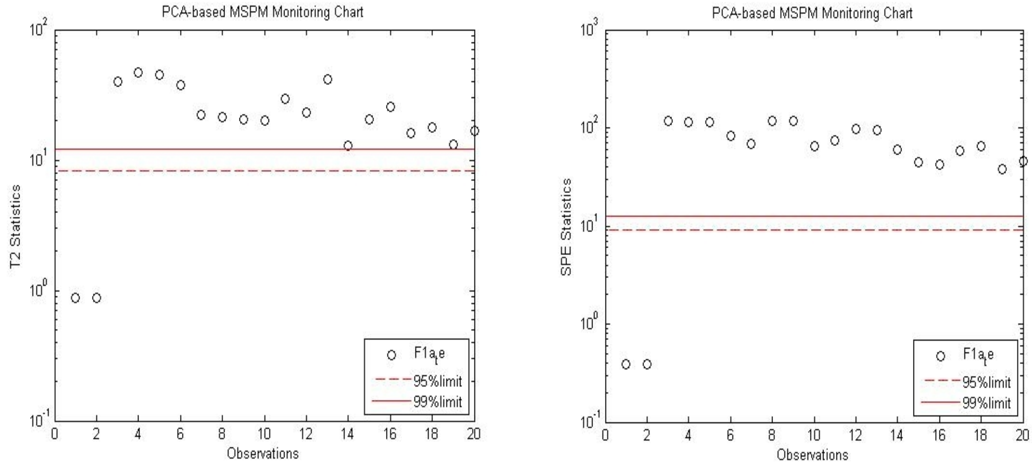

Figure 4.7: Results of PCA-based MSPM system for Abrupt fault.

Figure 4.6 and illustrate graph of PCA-based MSPM system for incipient data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for incipient fault data is only five variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only five early data is normal and next data is out of control or fault. However for SPE there are only three sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

For Figure 4.7 illustrate graph of PCA-based MSPM system for abrupt data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for abrupt fault data is only two variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only two early data is normal and next data is out of control or fault. However for SPE is also the same result as T^2 there are only two sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

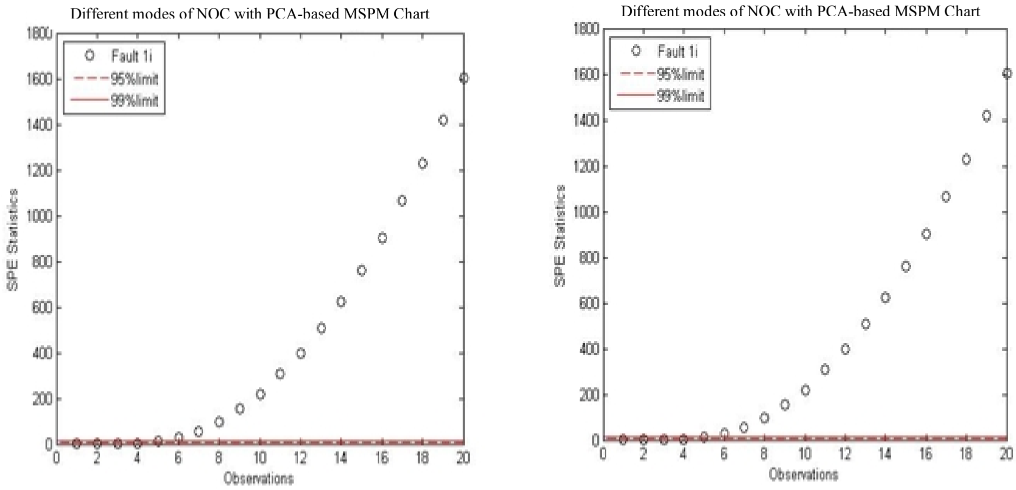

Figure 4.8: Results of different modes of NOC with PCA-based MSPM system for incipient fault

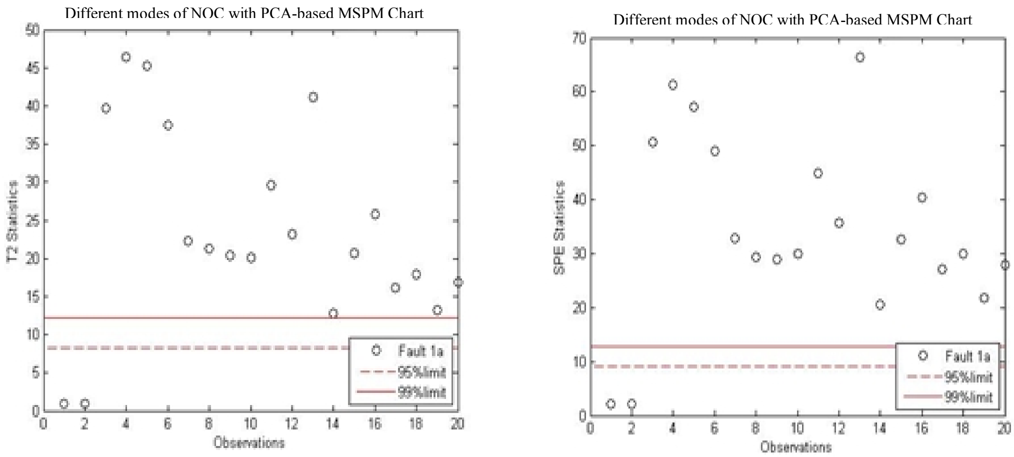

Figure 4.9: Results of different modes of NOC with PCA-based MSPM system for Abrupt fault

This NOC data has been test by reduce the sample using 50 measurements with 13 variables. By using PCA algorithm, the standardized NOC data test is analysed. This analysis objective is to identify the required number of PCs as to reduce the dimension of the multivariate data. Same with NOC data, the Hotelling's T2 statistic and SPE statistic was calculated and plotted together with the 95% and 99% confidence limits.

Figure 4.8 and illustrate graph of different modes of NOC with PCA-based MSPM system for incipient fault for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for incipient fault data is have five variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only five early data is normal and next data is fault. The variable is slow to achieve a fault decision. A. However for SPE there have four sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system. Is the same because the variable is slow to achieve the fault. Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

For Figure 4.9 illustrate graph of PCA-based MSPM system for abrupt data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for abrupt fault data is only two variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only two early data is normal and next data is out of control or fault. However for SPE is also the same result as T^2 there are only two sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

Result for T^2 and SPE for the the fault no.2

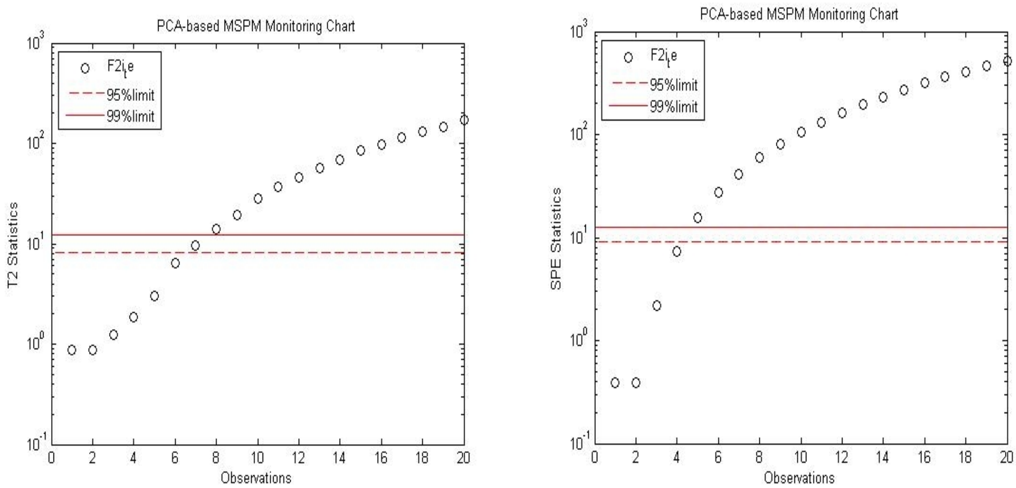

Figure 4.10: Results of PCA-based MSPM system for incipient fault

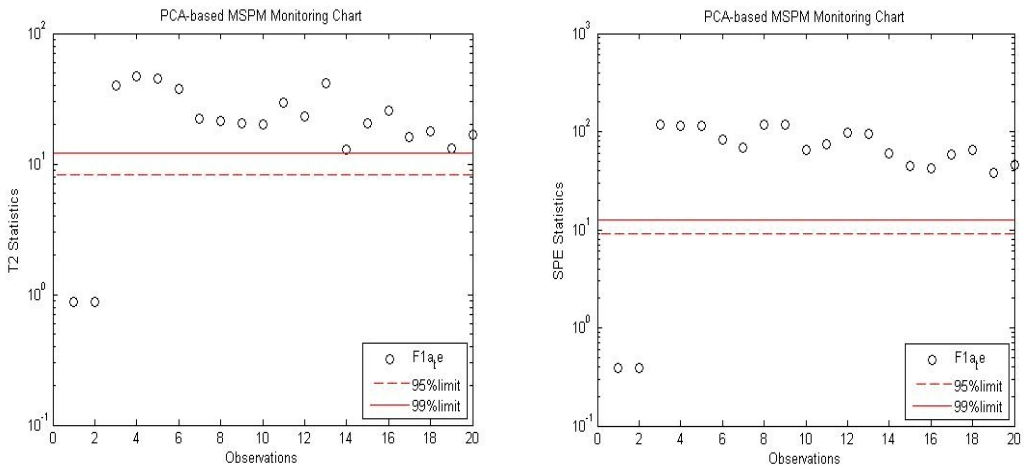

Figure 4.11: Results of PCA-based MSPM system for Abrupt fault.

Figure 4.10 and illustrate graph of PCA-based MSPM system for incipient data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for incipient fault data is only seven variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only seven early data is normal and next data is out of control or fault. The variable is slow to achieve the fault. However for SPE there are only four sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

For Figure 4.11 illustrate graph of PCA-based MSPM system for abrupt data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for abrupt fault data is only two variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only two early data is normal and next data is out of control or fault. However for SPE is also the same result as T^2 there are only two sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

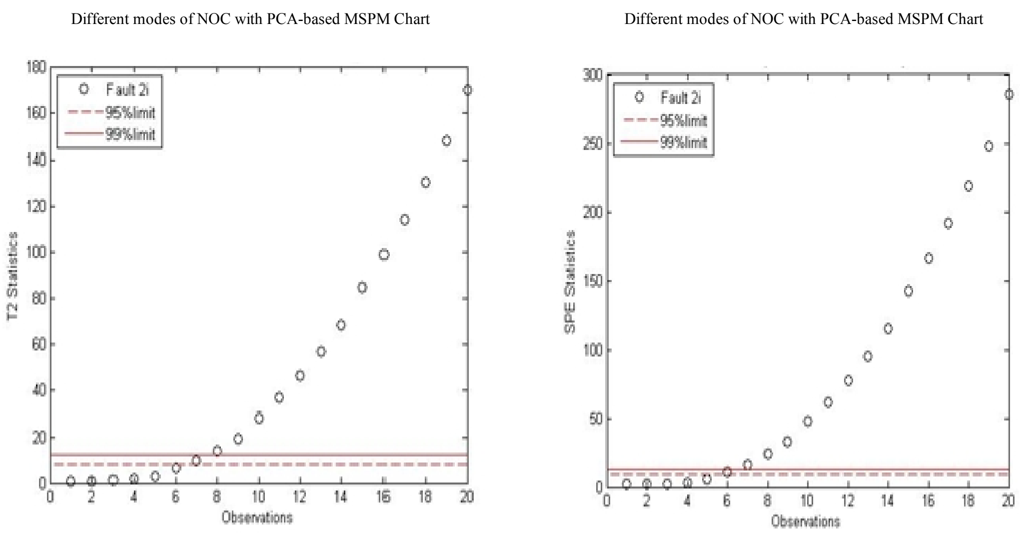

Figure 4.12: Results of different modes of NOC with PCA-based MSPM system for incipient fault

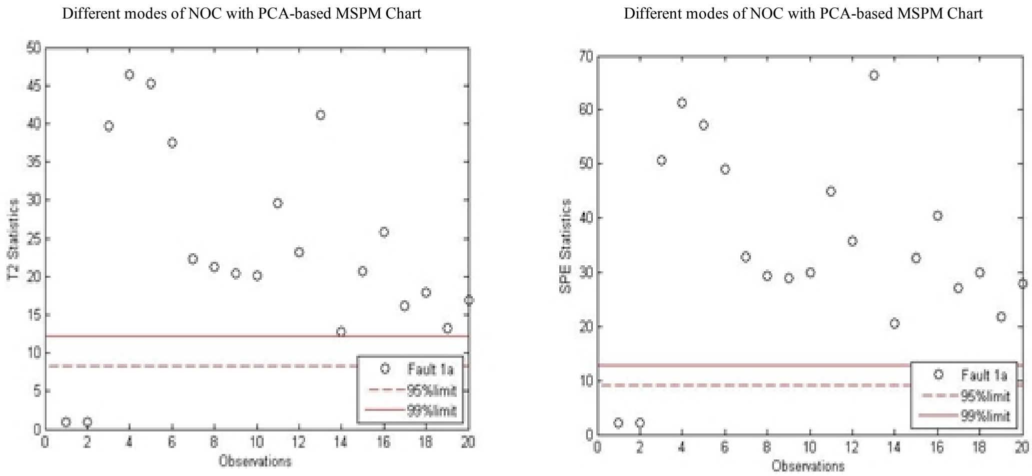

Figure 4.13: Results of different modes of NOC with PCA-based MSPM system for Abrupt fault

This NOC data has been test by reduce the sample using 50 measurements with 13 variables. By using PCA algorithm, the standardized NOC data test is analysed. This analysis objective is to identify the required number of PCs as to reduce the dimension of the multivariate data. Same with NOC data, the Hotelling's T2 statistic and SPE statistic was calculated and plotted together with the 95% and 99% confidence limits.

Figure 4.12 and illustrate graph of different modes of NOC with PCA-based MSPM system for incipient fault for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for incipient fault data is have five variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only five early data is normal and next data is fault. The variable is slow to achieve a fault decision. However for SPE there have four sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system. Is the same because the variable is slow to achieve the fault. Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

For Figure 4.13 illustrate graph of PCA-based MSPM system for abrupt data for T2 statistic and SPE statistic. It can be seen that the T2 statistic for the PCA-based MSPM system for abrupt fault data is only two variable is below the limits. The limits have sets on 95% and maximum limit 99%. So the results we can see only two early data is normal and next data is out of control or fault. However for SPE is also the same result as T^2 there are only two sample is inside on the limits level. The next data is also out of control for the incipient fault of PCA-based MSPM system . Thus, T2 statistic and SPE statistics illustrate abnormal operating condition.

4.5 Summary

A simulation of CSTR process is applying the conventional PCA to monitor the process. The main conventional PCA results have been discussed initially, which includes both of the NOC and fault data. For the fault data, both incipient and abrupt are discussing by looking at the contribution causes and impact, and also the T2 statistic and SPE statistics.

CHAPTER 5

CONCLUSION AND RECOMENDATION

5.1 Conclusions

In this research, MSPM using PCA tools is introduced. Some of the extension of MSPM and PCA is be review and the basic methodology to approach the proposed has been illustrated. The core technique to formulate the multivariate dimensional data reduction has been developing in order to approach the objectives using conventional PCA technique. The main goal in carrying out this study is to implement the conventional MSPM method based on different modes of NOC and analyze it with the conventional PCA technique on single NOC data. Based on the review on literature review there are many more method and technique to formulated multivariate data reduction. Every method has its own advantages and disadvantages. This research has proposed to run the traditional PCA by analyzing it with single NOC data and different modes of NOC data. Based on preliminary result get it has shown the technique has affected the fault detection solutions. However, it does not mean that it is an excellent method for non-linear process monitoring. Therefore, more analyses are required. A new fault detection method can be developed which is based on dissimilarity matrix application by integrating it with conventional PCA in MSPM system. By the way, PCA based MSPM slightly performed better in fault detection specifically for incipient faults compared to different modes of NOC with PCA based MSPM.

5.1 Recomendations

The results may valid only for CSTR system. It is recommended for future research to use data from other chemical processing systems such as PBR, PFR or other known chemical reactors. By the way more faults should be tested to come up with much more concrete conclusion.

REFERENCES

Chen, J., Liao, C.M., Lin, F.R.J. and Lu, M.J. (2001). “PCA Based Control Charts with Memory Effect for Process Monitoring.” *Industrial and Engineering Chemistry Research.* Vol. 40. 1516-1527.

Dunia, R., Qin, S.J., Edgar, T.F. and McAvoy, T.J. (1996). “Use of Principal Component Analysis for sensor fault identification.” *Computers and Chemical Engineering*. Vol. 20, supplement. S713 – S718.

Felder, R.M. and Rousseau, R.W. (1986). “Elementary Principles of Chemical Processes.” John Wiley and Sons.

Isermann, R. (1997). “Supervision, Fault Detection and Fault Diagnosis Methods – An Intorduction.” *Control Engineering Practice*. Vol. 5(5). 639 – 652.

Kourti, T. dan MacGregor, J.F. (1996). “Recent Developments in Multivariate SPC Methods for Monitoring and Diagnosing Process and Product Performance.” *Journal of Quality Technology*, Vol.28 (4). 409-428.

Kresta, J., MacGregor, J.F. and Martin, T.E. (1991). “Multivariate Statistical MonitoringOf Process Operation Performance.” *Canadian Journal of Chemical Engineering.* Vol.69. 34-47

Manish, M., Yue, H.H., Qin, S.J. and Ling, C. (2002). “Multivariate Process Monitoring and Fault Diagnosis by Multi-scale PCA.” *Computers and Chemical Engineering*. Vol.26. 1281-1293.

Quesenberry, C.P. (1997). “SPC methods for quality improvement.” John Wiley and Sons, Inc. 142-144, 274.

Stephanopoulos, G. (1984). "Chemical Process Control: An Introduction to Theory and Practice." 1st Edition, Englewood Cliffs, New Jersey, Massachussets Institute of Technology

J.F MacGregor & T. Kourti. (1995). Statistical Process Control of Multivariate Process. Control Eng. Practice, Vol.3, pp 403-414.

E.B Martin, A.J Morris, & J. Zhang. (1996). Process performance monitoring using multivariate statistical process control. IEE proc.-Control Theory Appl. Vol. 143, No.2, pp 132-144.

Leonard A. Doty. (1996). Statistical Process Control. Industrial Press Inc. Retrieved on December 12, 2011 from Google Books database.

Robert Lee Mason & John C. Young. (2001). Multivariate Statistical Process Control with Industrial Application. ASA SIAM. Retrieved on December 12, 2011 from Google book database..

I.T Jollieffe. (2002). Principle Component Analysis: Second Edition. Springer.

Bernhard Scholkopf, Alaxander Smola, & Klaus-Robert Muller. (n.d). Nonlinear Component Analysis Eigen Value Problem, Max-Planck-Institute für biologische Kybernetik.

Hyun-Woo Cho. (2006). Identification of contributing variables using kernel-based discriminant modeling and reconstruction. Expert System with Application,Vol 33, Issue 2, Pages 274-285.

Pirter M. Kroonenberg. (1983). Three-Mode Principle Component Analysis. DSWO Press.

Layth C. Alwan & Harry V. Roberts. (1988). Time-Series Modelling for Statistical Process Control. Journal of Business & Economic Statical Process Control, Vol. 6, No.1.

APPENDIX

Calculation the NOC data using the MATLAB software. The calculation data was install in the MATLAB to compare the result the conventional MSPM method based on a single NOC with the conventional MSPM method based on different modes of NOC.

```
%TE NOC
%PCA
[Sd_N,Me,St]=autosc(Data_NOC);
Sd_N_te=scal(Data_NOC_te,Me,St);
Sd_F3a_te=scal(Data_F03a_te,Me,St);
Sd_F4a_te=scal(Data_F04a_te,Me,St);
Sd_F3i_te=scal(Data_F03i_te,Me,St);
Sd_F4i_te=scal(Data_F04i_te,Me,St);
nv=13;
  ns=100;
  C=Sd_N'*Sd_N;
  [U_cov,S_cov,V_cov]=svd(C);
  [u,s,v]=svd(Sd_N);
  for i=1:nv
   Pc_cov(i)=sum(diag(S_cov(1:i,1:i)))/sum(diag(S_cov(1:nv,1:nv)));
  end
  figure
  plot(Pc_cov,'o')
  hold
  plot(Pc_cov)
  xlabel('Number of principal components')
```

```
  ylabel('Variance explained (Covariance)')
  title('Accumulated Variance vs Principal Components')
  npc=3;
  Sco_NOC=Sd_N*V_cov(:,1:npc);
  Sco_NOC_te=Sd_N_te*V_cov(:,1:npc);

  T2_95_lim=npc*(ns-1)*finv(0.95,npc,(ns-npc))/(ns-npc); % 95% control
limit for T2
  T2_99_lim=npc*(ns-1)*finv(0.99,npc,(ns-npc))/(ns-npc); % 99% control
limit for T2

  T2_npc=[];
  for i=1:npc
      Lambd(i)=((S_cov(i,i)/(ns-1)));
      T2=(Sco_NOC(:,i).^2)/((S_cov(i,i)/(ns-1)));
      T2_npc=[T2_npc;T2];
  end
  D=[];
  D=[D;reshape(T2_npc,ns,npc)];
  T2_NOC=sum(D');
  Lambd1=S_cov(1,1)/(ns-1);
  Lambd2=S_cov(2,2)/(ns-1);
  Lambd3=S_cov(3,3)/(ns-1);

T2_NOC_test=Sco_NOC(:,1).^2/Lambd1+Sco_NOC(:,2).^2/Lambd2+Sco_NOC(:,3)
.^2/Lambd3;
  % Graph for Tesis
  figure
  plot(T2_NOC,'ko')
```

```
hold
plot([0;ns],T2_95_lim*[1;1],'r--')
plot([0;ns],T2_99_lim*[1;1],'r-')
xlabel('Observations')
ylabel('T2 Statistics')
legend({'NOC' '95%limit' '99%limit'}, 'Location','NorthWest');
title('PCA-based MSPM Monitoring Chart')


Th_1=sum(diag(S_cov(npc+1:nv,npc+1:nv)/(ns-1)));
Th_2=sum(diag(S_cov(npc+1:nv,npc+1:nv)/(ns-1)).^2);
Th_3=sum(diag(S_cov(npc+1:nv,npc+1:nv)/(ns-1)).^3);
H_0=1-2*Th_1*Th_3/(3*Th_2^2);
Q_95=Th_1*(1.645*sqrt(2*Th_2*H_0^2)/Th_1+1+Th_2*H_0*(H_0-
1)/Th_1^2)^(1/H_0);
Q_99=Th_1*(2.326*sqrt(2*Th_2*H_0^2)/Th_1+1+Th_2*H_0*(H_0-
1)/Th_1^2)^(1/H_0);
%Th_1=sum(diag(S_cov(npc+1:nv,npc+1:nv)));
%Th_2=sum(diag(S_cov(npc+1:nv,npc+1:nv)).^2);
%Th_3=sum(diag(S_cov(npc+1:nv,npc+1:nv)).^3);
%H_0=1-2*Th_1*Th_3/(3*Th_2^2);
%Q_95=Th_1*(1.645*sqrt(2*Th_2*H_0^2)/Th_1+1+Th_2*H_0*(H_0-
1)/Th_1^2)^(1/H_0);
%Q_99=Th_1*(2.326*sqrt(2*Th_2*H_0^2)/Th_1+1+Th_2*H_0*(H_0-
1)/Th_1^2)^(1/H_0);
th1=sum(diag(s(npc+1:13,npc+1:13).^2/(ns-1)));
th2=sum(diag(s(npc+1:13,npc+1:13).^2/(ns-1)).^2);
th3=sum(diag(s(npc+1:13,npc+1:13).^2/(ns-1)).^3);
h0=1-2*th1*th3/(3*th2^2);
```

```
  q_cl_95=th1*(1.645*sqrt(2*th2*h0^2)/th1+1+th2*h0*(h0-
1)/th1^2)^(1/h0);
  q_cl_99=th1*(2.326*sqrt(2*th2*h0^2)/th1+1+th2*h0*(h0-
1)/th1^2)^(1/h0);

  NOC_Norp=Sco_NOC*V_cov(:,1:npc)';
  Err_NOC=(Sd_N-NOC_Norp).^2;
  Spe_NOC=sum((Err_NOC)');
  figure
  plot(Spe_NOC,'ko')
  hold
  plot([0 ns], Q_95*[1 1],'r--')
  plot([0 ns], Q_99*[1 1],'r-')
  xlabel('Observations')
  ylabel('SPE Statistics')
  legend({'NOC' '95%limit' '99%limit'}, 'Location','NorthWest');
  title('PCA-based MSPM Monitoring Chart')

  sdnp=Sco_NOC*v(:,1:npc)'; % predicting original variables values
from pc's
  err_sdn=sdn-sdnp; % score erros (original - prediction)
  spe_n=sum((err_sdn.^2)')'; % sum of squared prediction error by
samples


%CSTR F1
%PCA
  no=20;
  Sco_F1a_te=Sd_F1a_te*V_cov(:,1:npc);
```

```
T2_F1a_te=Sco_F1a_te(:,1).^2/Lambd1+Sco_F1a_te(:,2).^2/Lambd2+Sco_F1a_
te(:,3).^2/Lambd3;
  figure
  semilogy(T2_F1a_te,'ko')
  hold
  semilogy([0;no],T2_95_lim*[1;1],'r--')
  semilogy([0;no],T2_99_lim*[1;1],'r-')
  xlabel('Observations')
  ylabel('T2 Statistics')
  legend({'F1a_te' '95%limit' '99%limit'}, 'Location','NorthWest');
  title('PCA-based MSPM Monitoring Chart')

  F1a_Norp=Sco_F1a_te*V_cov(:,1:npc)'; % predicting original variables
from PCs
  Err_F1a_te=(Sd_F1a_te-F1a_Norp).^2; % score errors (original -
prediction)
  Spe_F1a_te=sum((Err_F1a_te)'); % sum of SPE by samples
  figure
  semilogy(Spe_F1a_te,'ko')
  hold
  semilogy([0;no], Q_95*[1 1],'r--')
  semilogy([0;no], Q_99*[1 1],'r-')
  xlabel('Observations')
  ylabel('SPE Statistics')
  legend({'F1a_te' '95%limit' '99%limit'}, 'Location','NorthWest');
  title('PCA-based MSPM Monitoring Chart')
```