DEVELOPMENT OF MULTIVARIATE STATISTICAL PROCESS MONITORING USING COMBINATION MLR-PCA METHOD

NORLIZA BINTI ISHAK

Thesis submitted in fulfillment of the requirement
for the award of the degree of Master of Engineering in Chemical

Faculty of Chemical and Natural Resources Engineering
UNIVERSITI MALAYSIA PAHANG

OCTOBER 2015

ABSTRACT

The most popular types of process monitoring systems is Multivariate Statistical Process Monitoring (MSPM) which has the most practical method in handling the complicated large scale processes. This is due to the ability of the system in maximizing the usage of abundant historial process data, in such a way that the original data dimensions are compressed and data variations preserved to certain extent in a set of transformed variables by way of linear combinations. Thus, the composite model is generally flexible regardless of the amount of variables that utilized. In this regard, conventional Principal Component Analysis (PCA) has been widely applied to conduct such compression function particularly for MSPM. However, the conventional PCA is a linear technique which results sometimes inappropriately employed especially in modeling processes that exhibit severe non-linear correlations. Therefore, a new solution is demanded, whereby the number of original variables can be reduced to certain extent (in terms of scales), while it still can maintain the variation as maximally as possible corresponding to the original, which are then transferable into monitoring statistics. One of the potential techniques available in addressing the issue is known as Multiple Linear Regression (MLR). The main objective of the technique is to predict a set of output values (criterion) based from a specified of linear function, which consists of a set of predictor. Therefore, the main multivariate data will be divided into two groups, which are the criterion and predictor categories. The study adopts, Tennessee Eastman Process (TEP) and Multiple Output and Multiple Input (MIMO) Pilot Plant System for demonstration. The general finding is that MLR-PCA normally employs less number of PCs compared to PCA, and thus, this will perhaps reduce complication during diagnosis. By adopting such approach, the monitoring task can be made simpler and perhaps more effective, in the sense that only those selected criterion variables (predicted values) will be taken for monitoring, while preserving the rest of the predictor value trends in the form of linear functions in association with the criterion variables. This study also shows that MLR-PCA works relatively better in terms of fault detection and identification against the conventional system.

ABSTRAK

Sistem pemantauan proses yang paling popular ialahPembolehubah Pemantauan Proses Statistik (MSPM) yang dianggap sebagai kaedah yang paling praktikal dalam menanggani sistem kompleks serta berskala besar. Hal ini adalah kerana keupayaan sistem dalam memaksimumkan penggunaan data proses dalam apa-apa cara bahawa dimensi data asal dimampatkan dan variasi data dikekalkan pada tahap tertentu dalam satu set pembolehubah yang diubah melalui kombinasi linear. Oleh itu, model komposit biasanya fleksibel tanpa mengira jumlah pembolehubah yang digunakan. Dalam hal ini, Analisis Komponen Utama konvensional (PCA) telah digunakan secara meluas untuk menjalankan fungsi mampatan itu terutamanya untuk MSPM. Walau bagaimanapun, PCA konvensional adalah teknik linear yang mana kaedah ini kadang-kadang tidak sesuai digunakan terutama dalam proses model kerana mempamerkan korelasi bukan linear yang tidak normal. Oleh itu, satu penyelesaian baru diperlukan, di mana bilangan pembolehubah asal boleh dikurangkan kepada tahap tertentu (dari segi skala), sementara ia masih boleh mengekalkan perubahan maksima sebagaimana ia mungkin sama dengan data yang asal, yang kemudiannya dipindahkan ke dalam statistik pemantauan. Salah satu teknik yang berpotensi untuk menangani isu ini dikenali sebagai Regresi Linear Berganda (MLR). Objektif utama teknik ini untuk meramalkan satu set nilai hasil (kriteria) berdasarkan dari yang dinyatakan fungsi linear, yang terdiri daripada satu set peramal. Oleh itu, data multivariat utama akan dibahagikan kepada dua kumpulan, iaitu kriteria dan kategori peramal. Kajian ini mengadaptasi Proses Tennessee Eastman (TEP) dan Multiple Output and Multiple Input (MIMO) Loji Pandu sebagai demostrasi. Kajian mendapati MLR-PCA biasanya menggunakan PC nombor yang minimum berbanding PCA dan sekaligus, hal ini berkemungkinan dapat mengurangkan kerumitan semasa diagnosis.Dengan mengguna pakai pendekatan ini, tugas pemantauan boleh dibuat lebih mudah dan mungkin lebih berkesan, dalam erti kata lain bahawa hanya pembolehubah kriteria tertentu sahaja (nilai meramalkan) akan diambil untuk dianalisis, di samping memelihara seluruh trend nilai peramal dalam bentuk fungsi linear bersama dengan pembolehubah kriteria. Kajian ini juga menunjukkan MLR-PCA berfungsi lebih baik dari segi pengesanan ralat dan mengenal pasti ralat berbanding kaedah konvensional.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CMDS	Classical Multidimensional Scaling
CSTRwR	Continuous Stirred Reactor with Recycle
CVA	Canonical Variate Analysis
DKPCA	Dynamic Kernel Principal Component Analysis
DPCA	Dynamic Principal Component Analysis
FA	Factor Analysis
FAC	False Alarm Cases
FAR	False Alarm Rate
FDT	Fault Detection Time
FI	Fault Identification
FIC	Flow rate Indicator Controller
ICA	Independent Component Analysis
ICs	Independent Components
IDV	Fault Cases
HPLC	High Performance Liquid Chromatography
KPCA	Kernel Principal Component Analysis
LPLS	Log-contrast Partial Least Square
MBPCA	Multi Block Principal Component Analysis
MDS	Multidimensional Scaling
MEWMA	Multivariate Exponentially - Weighted Moving Average
MIMO	Multiple Input Multiple Output
MLR	Multiple linear Regression
MPLS	Multi-way Partial Least Square
MSE	Mean Squared Error
MSKPCA	Multi Scale kernel Principal Component Analysis
MSPC	Multivariate Statistical Process Control
MSPCA	Multi Scale Principal Component Analysis
MSPM	Multivariate Statistical Process Monitoring

NCD	Number of Fault Cases Detected
ND	No detection
NFD	Number of Fastest Detection
NIPAL	Nonlinear Iterative Partial Least Square
NLPCA	Non-linear Principal Component Analysis
NOC	Normal Operating Condition
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCs	Principal Components
PLS	Partial Least Square
RPLS	Recursive Partial Least Square
SPC	Statistical Process Control
SPE	Squared Prediction Error
SPLS	Sparse Partial Least Square
SVD	Singular Value Decomposition
T^2	Hotteling's
TEP	Tennessee Eastman Process
TIC	Temperature Indicator Controller
XMEAS	Process Measurements
XMV	Process Manipulated Variables

LIST OF SYMBOLS

$\mathbf{A}$	Dissimilarity matrix for MDS.
$\boldsymbol{A}$	Unknown mixing matrix in ICA.
a	Number of PCs that retained in the PCA model.
$\mathbf{B}$	Regression coefficients of PCR.
$\mathbf{B_{MDS}}$	Double-centered matrix.
$\mathbf{b}$	Regression coefficient for MLR.
$\mathbf{C}$	Variance - covariance matrix.
$c_{m,m}$	Variance - covariance matrix at row m and column m.
$\boldsymbol{E}$	Residual matrix in ICA.
$\widetilde{e_i}$	The ith row vector in residual matrix.
$F_{A,n-A,\alpha}$	F distribution index with A and n-A degrees of freedom at α confident limit.
$\mathbf{I_m}$	Identity matrix.
l	Number of selected principal components.
lim_α	Control limits for T^2 and SPE statistics.
m	Number of samples for MDS.
m	Total number of variables.
$\bar{m}$	Means of the T^2 and SPE statistics.
$(MLR\ Statistics)j$	MLR statistics (SPE for the criterion variables) at a particular sampling time 'j'.
n	Number of samples.
$\mathbf{P}$	Loadings factor.
$\mathbf{P}$	Score matrix for PCA.
$\boldsymbol{p}$	Scores matrix for PLS.
$P_{i,j}$	The ith element for principal component j.
$\mathbf{P_a}$	Score matrix that containing the first a score vectors.
$(PredictorStatistics)j$	Predictor statistics at a particular sampling time 'j'.
$\boldsymbol{S}$	Independent component matrix in ICA.
SPE_α	α confidence limits for SPE parameter.

SPE_i	The contribution of the *i*th variable to SPE.
$(SPE_i)_j$	Contribution of the *i*th variable to MLR (SPE) statistics at particular sampling time 'j'.
$\mathbf{T}$	Orthogonal scores of the predictor variables.
$\mathbf{T_{nxl}}$	Matrix of the multivariate scores.
$\mathbf{t}$	Latent component matrix in PLS.
$\boldsymbol{T}_{(a)}$	Decomposing of data matrix into orthogonal scores.
T_α	*A* confidence limits for T^2 parameter.
$\boldsymbol{T^2}$	T squared value
T_i^2	T squared value at sample *i*.
$\mathbf{V}$	Eigenvector matrix for PCA.
$\mathbf{V}$	Loading matrix (a set of selected eigenvectors).
$\mathbf{V_1}$	Corresponding sets of eigenvectors in MDS.
$\mathbf{V_a}$	Loading matrix containing the first *a* loading vectors.
v	Variances of the T^2 and SPE statistics.
$\boldsymbol{w}$	Weight vectors.
wi	Aggregating weight.
$\mathbf{X}$	Predictor variables matrix.
$\boldsymbol{X}$	Data matrix for ICA.
$\mathbf{X_{nxm}}$	Data matrix where *n* and *m* are the number of samples and variables.
$\mathbf{X_{PCR}}$	Data scores obtained by the first principal components (PCs) in PCR.
$\mathbf{X_{PLS}}$	Predictor variables for PLS.
$\tilde{\mathbf{X}}$	Standardized multivariate data.
$\widehat{\boldsymbol{X}}$	Data matrix produces by the decompression of the data scores.
$\widetilde{\boldsymbol{X}}_{(a)}$	Variance of first principal components.
$x_{i,j}$	The *'i'*th individual origin data of variable *'j'*.
$\tilde{x}$	Standardized data.
$\bar{x}_j$	Mean of variable *'j'*.
$(X_i)_j$	Contribution of the *i*th variable to PCA statistics at particular sampling time 'j'.
$\mathbf{Y}$	Criterion variables.

$\mathbf{Y_{MLR}}$	New predicted criterion matrix.
$\mathbf{Z}$	Cartesian coordinate matrix.
z_α	Standard normal deviate corresponding to the upper (1-α) percentile.
α	Level of confidence limits.
σ_j	Standard variation of variable *'j'*.
λ_j	Eigenvalue corresponds to principal component *j*.
$\mathbf{\Lambda}$	Eigenvalues matrix for PCA.
$\mathbf{\Lambda}$	Loading matrix for FA.
$\Lambda_{\mathbf{1}}^{\frac{\mathbf{1}}{\mathbf{2}}}$	Diagonal matrix with all positive elements.
$\mathbf{\Psi}$	Diagonal matrix for FA.
$\mathbf{1}_{\boldsymbol{m}}$	Vector with element that equal to 1.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Monitoring is a continuous real-time task of determining the possible conditions of a physical system, recognizing and indicating inconsistencies of the behavior (Isermann, 2011). The application of statistical method in monitoring is widely uses in chemical-based industries that involves with series of unit operations in order to convert the input materials into desired products following the qualitative and quantitative specifications of the customers. This can be considered as highly challenging as the process subjects to be affected by various unstable conditions over the time of operation. Simply avoiding or slow response in such situations may result with the decadence of product quality and even leads to catastrophic events as well as risking the profitability of the company. Thus, it has always been imperative to have a systematic mechanism which can routinely manage all of these abnormal situations automatically. These problems can be addressed quite effectively by using the process monitoring system. The method normally functioned to conduct fault detection, fault identification and fault diagnosis tasks.

1.2 RESEARCH BACKGROUND

In general, there are two popular types of process monitoring systems available for industrial application, which are univariate and multivariate monitoring systems. Since most modern industrial processes are involving multivariate in nature especially in measurements on a number of characteristics, instead one single characteristic, univariate method provides little information regarding the mutual interactions and also do not function well for multivariable processes with highly correlated variables (Nomikos and MacGregor, 1995and Qin, 2012). The conventional univariate system, such as Statistical Process Control (SPC), has been mainly criticized for its limitation (particularly in the context of chemical-based operation), whereby it is only being operative under univariate analysis setting as well as a large number of control charts is always needed to be monitored concurrently (Bersismis et al., 2007). Besides, it always ignores the implications of harmonization between the output and input variables. Thus, the multivariate system such as Multivariate Statistical Process Monitoring (MSPM) can be regarded as the most practical method for handling complicated and large scale systems (Chiang et al., 2001).

As the process is in multivariate nature, the system will typically develop a model that correlates all of the variables simultaneously by using a set of normal operating condition (NOC) data that obtained from the historical process archive. In the other words, the system can utilize maximally all the process data stored for better use (Zhao et al., 2004). Besides, the system also has been popularly perceived as an advanced technique from the traditional SPC methodology. Unlike SPC, MSPM can extract useful information in terms of inter-related variable variations and represent it by using simplified parameters (normally in terms of multivariate scores and monitoring statistics). By applying the method, the monitoring operation can be executed much simpler (in the sense that only a small number of control charts are required) as well as critically consider the effect of all changes that contributed from various variables concurrently.

At the fundamental level application, there are two types of monitoring charts typically employed – Hotteling's *T2* and Squared Prediction Errors (SPE). The first represents conceptually the magnitude of deviation of the current sample from the center, whereas, the second analyzes the consistency of the current sample correlation according to the NOC model development. Both have been used complementary, whereby, control limits for both statistics are also computed accordingly. The main task would be to observe the progressions of both statistics on a control chart (usually Shewhart-type control chart) that constructed respectively. When the process is normal, all the statistics will remain below the control limit lines. However, whenever a fault event takes place in the process, the corresponding statistics will move away from the normal region until a point where it goes beyond the control limits specified (sooner or later). If the abnormal trend persists over a period of time, the system will then initiate the alarm which signifies that one or a combination of faulty event(s) has (have) actually occurred in the process (fault detection). Contribution plot is then applied, purposely to identify the main possible variables that either contributing or being affected from that particular abnormal event that detected. Lastly, further investigation is needed to critically sort out and finally diagnose the true source of the problem that related to that particular abnormal existence.

1.3 PROBLEM STATEMENT

MSPM normally utilizes linear-based principal component analysis (PCA) as the main technique of multivariate data compression. However, PCA sometimes is improperly used especially in modeling highly nonlinear processes as a high number of principal components (PCs) are always involved may lead to inefficient and unreliable monitoring performance reflected in false alarms and missed faults (Dong and McAvoy, 1996 and Žvokelj et al., 2011). If large variable are involved, then the PCs may also be selected considerably. As a result, Zhang et al., (1997) introduced non-linear PCA which based on the combination of neural network and principal curve but the computation is very demanding and it always requires a massive amount of data for creating the optimized NOC model (Yunus, 2012).

In other applications, Yunus and Zhang (2010a, b and c) as well as Yunus (2012) have developed three main frameworks of the MSPM by using the classical multidimensional scaling (CMDS) approach. Even though some improvements can be observed in terms of fault detection efficiency, those approaches employed different score projection (by way of variable scores) as opposed to the conventional PCA (by means of sample scores). It is argued that the CMDS approach cannot be effectively applied when it involves with a very large number of the variables as the variable scores cannot be re-produced as precisely as possible according to the pre-specified NOC configurations. This shows that, both techniques (PCA and MDS) suffered from technical difficulties particularly when handling large amount of variables in monitoring. Therefore, a new solution is demanded, whereby the number of original variables can be reduced to a certain extent (in terms of scales), while it can considerably maintain the original variation as largely as possible.

1.4 RESEARCH OBJECTIVES

One of the potential techniques available to address the issue is known as multiple linear regressions (MLR). The main objective of the technique is to predict a set of output values (criterion) based from a specified set of linear function, which consists of the predictor variables and to reduce the number of the dimensionality. Therefore, the main multivariate data will be divided into two groups, which are the criterion and predictor categories. By adopting such approach, the monitoring task can be made simpler and perhaps more effective, in the sense that only those criterion variables (predicted values) will be taken for monitoring, while preserving the rest of the predictor value trends in the form of linear functions. In light of this, the primary objectives of the study are:

i. To apply a basic MSPM system using conventional PCA, whereby the monitoring outcome of this system will be used as the benchmark performance in order to assess the credibility of the proposed system.
ii. To develop and investigate the performance of the MSPM using MLR method against the standard performance of MSPM-PCA. In particular, the monitoring data are divided into two main categories, whereby only the quality variables are used for monitoring.
iii. To analyze the performance of MLR-PCA by applying real process instrument data.

1.5 RESEARCH SCOPES

In order to accomplish the objectives, the scope of this study is focusing on the several criteria as follows:

i. This study considers the challenging case studies by using the Tennessee Eastman Process (TEP) and also the data from the Multiple Input and Multiple Output (MIMO) Plant for monitoring.
ii. The monitoring platform will be developed by using MATLAB 7.8.0:347 (R2009a) software.
iii. In applying the MLR method, the main multivariate data will be divided into two groups, which are the criterion and predictor categories.
iv. A number of comparative analyses between the proposed and conventional method are essential is evaluating the credibility of the new system's performance, particularly on the ground of number of cases detected, number of faster detection, fault detection and false alarm rate.
v. Fault identification is also conducted to complement the fault detection results.

1.6 RATIONALE AND SIGNIFICANT

It is expected that the monitoring performance can be performed efficiently as well as effectively using a small number of dimensions in comparison to the conventional monitoring outcomes. This is simply because the modified data contains less number of variables (in term of the criterion value of the MLR model) in contra to the original magnitude. As a result, fewer PCs or dimensions are expected to be used in the PCA models respectively. At the same time, it is also assumed that the original variations of the predictor variables can be significantly preserved and re-produced by the criterion variables during monitoring operation.

As a whole, the main contributions of this study are:

i. MLR technique can actually represent the behavior of the original data, in such a way it has the potential to be used in process monitoring.
ii. Newly system of MSPM framework which integrates MLR and PCA techniques under a single application will create the process system less complicated.
iii. The approach technique perhaps will improve the sensitivity of the fault detection performance comparative to the conventional approach particularly using less number of compressed PCs.
iv. Process monitoring based MLR framework potentially reduce the complexity during the fault identification as well as diagnosis operations.

1.7 REPORT ORGANIZATION

This thesis has been divided into seven main chapters. Chapter 1 represents the introduction of the study, including research background, the objectives, goals and contributions, the remaining chapters in thesis are organized as follows;

Chapter 2 begins with some theory regarding the fundamental of MSPM, which emphasizing on the theoretical aspect of PCA. Then, other different multivariate methods are discussed, which consists of PLS, MDS, MLR, and others. The discussion is then followed by explaining the corresponding issues and extensions, which have been conducted in process monitoring.

While, Chapter 3 explains the methodology of developing the monitoring systems based on conventional and enhanced techniques. All the related procedures will be explained in details. In this chapter also includes the process description and the types of fault cases on the case studies that chosen to be investigated - Tennessee Eastman Process (TEP) and MIMO Plant.

Chapter 4 commences with the analysis of the data collection, analysis and the results of the modeling framework for both methods for TEP and MIMO Plant.

Chapter 5 concludes the findings in this work and suggests possible directions for future research. The main consideration of this chapter is to provide the corresponding to the research objectives in Chapter 1. Lastly, Chapter 6 lists all the references used in this research.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Maintaining product quality at the highest level is of wide spread concern in the process industries nowadays. This basically means that highly proportion of specifications product is always desirable, and eventually, it has sparked motivation for the reduction of abnormal variability in the normal operation routine to the lowest level as possible. This situation also has led to an increase in the use of process monitoring technique, which traditionally involving Statistical Process Control (SPC) (Papazoglou, 1998; Montgomery, 2009 and Kumar, 2013). However, the traditional SPC technique cannot be applied effectively in a multivariate nature of complex processes, which typically consist of huge matrix of several characteristics rather than a vector measurement. A number of limitations on the usage of SPC have been discussed comprehensively by several authors including (MacGregor and Kourti, 1995; Montgomery, 1996; Papazoglou, 1998; Jackson, 2005and Behbahani et al., 2012).

In addressing this particular issues, Multivariate Statistical Process Monitoring (MSPM), has been found the most suitable approach as well as progressively develop and widely applied for process monitoring presently. This could be perhaps due to its ability in providing systematic diagnostic tool for the extensive monitoring of a large operation data by means of off-line as well as on-line operation, which typically includes fault detection, identification, diagnosis and also process recovery. Besides, as the system depends heavily on the statistical approach, therefore, it is also applicable to wide variety of processes (batch and continuous) regardless of size and environment.

Thus, this chapter outlines the generic overview of the MSPM system, which consisting of four main sections. The first is on the introduction and subsequently followed by two main sections on developing the MSPM framework. In particular, the first describes a comprehensive overview on the fundamentals as well as the normal procedures of implementing MSPM system. Meanwhile, the second explains on the extensions and also the relative importance of the system with respect to the other industrial monitoring techniques. At the end of the chapter, a critical review is provided especially to highlight the significance of the study in contributing to the current MSPM scopes.

2.2 FUNDAMENTAL OF MSPM

According to Bersismis et al., (2005) Multivariate Statistical Process Monitoring (MSPM), which also known as Multivariate Statistical Process Control (MSPC), refers to a systematic statistical modeling procedure that consistently analyzes the performance of the process (continuous and batch) under investigation, particularly in deciding whether the current operating condition is actually normal or abnormal. This has been performed typically by means of observing the progression of monitoring statistics, namely T^2 (magnitude of deviation) and SPE (consistency) on the control charts (Wise and Gallagher, 1996; Martin et al., 1996; Cinar et al., 2007and MacGregor and Cinar, 2012). Chiang et al., (2001) stated that there are four basic steps typically involved in any of process monitoring procedures as depicted in **Figure 2.1**.

The first element is fault detection where the aim is to initiate the alarm whenever fault(s) has (have) been detected by the monitoring system (especially in the case of consistent violation on the control limits is evidenced). Next, fault identification is a set of specific procedures that highlights all the main variables which have shown great influence on the fault that detected in the first stage. In the third step, another vast set of procedures will be conducted specifically in diagnosing on those identified variables that captured previously, particularly in defining the true nature of the cause that contribute to the fault (fault diagnosis). The last element in the process monitoring system is process recovery whereby this process takes the remedial actions in eliminating the causes that lead to the occurrence of faults.

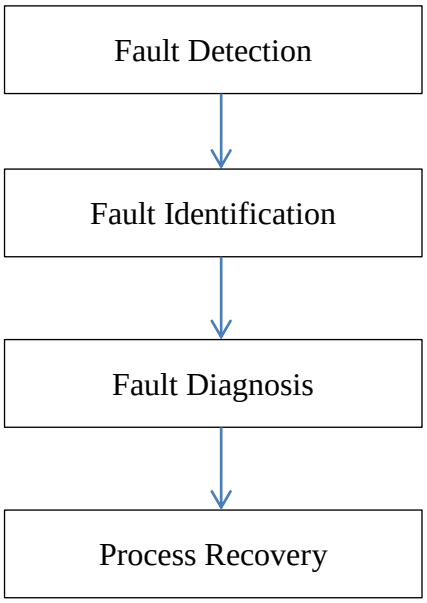

Figure 2.1: Procedures of Process Monitoring Systems (Chiang et al., 2001)

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Monitoring is a continuous real-time task of determining the possible conditions of a physical system, recognizing and indicating inconsistencies of the behavior (Isermann, 2011). The application of statistical method in monitoring is widely uses in chemical-based industries that involves with series of unit operations in order to convert the input materials into desired products following the qualitative and quantitative specifications of the customers. This can be considered as highly challenging as the process subjects to be affected by various unstable conditions over the time of operation. Simply avoiding or slow response in such situations may result with the decadence of product quality and even leads to catastrophic events as well as risking the profitability of the company. Thus, it has always been imperative to have a systematic mechanism which can routinely manage all of these abnormal situations automatically. These problems can be addressed quite effectively by using the process monitoring system. The method normally functioned to conduct fault detection, fault identification and fault diagnosis tasks.

1.2 RESEARCH BACKGROUND

In general, there are two popular types of process monitoring systems available for industrial application, which are univariate and multivariate monitoring systems. Since most modern industrial processes are involving multivariate in nature especially in measurements on a number of characteristics, instead one single characteristic, univariate method provides little information regarding the mutual interactions and also do not function well for multivariable processes with highly correlated variables (Nomikos and MacGregor, 1995and Qin, 2012). The conventional univariate system, such as Statistical Process Control (SPC), has been mainly criticized for its limitation (particularly in the context of chemical-based operation), whereby it is only being operative under univariate analysis setting as well as a large number of control charts is always needed to be monitored concurrently (Bersismis et al., 2007). Besides, it always ignores the implications of harmonization between the output and input variables. Thus, the multivariate system such as Multivariate Statistical Process Monitoring (MSPM) can be regarded as the most practical method for handling complicated and large scale systems (Chiang et al., 2001).

As the process is in multivariate nature, the system will typically develop a model that correlates all of the variables simultaneously by using a set of normal operating condition (NOC) data that obtained from the historical process archive. In the other words, the system can utilize maximally all the process data stored for better use (Zhao et al., 2004). Besides, the system also has been popularly perceived as an advanced technique from the traditional SPC methodology. Unlike SPC, MSPM can extract useful information in terms of inter-related variable variations and represent it by using simplified parameters (normally in terms of multivariate scores and monitoring statistics). By applying the method, the monitoring operation can be executed much simpler (in the sense that only a small number of control charts are required) as well as critically consider the effect of all changes that contributed from various variables concurrently.

At the fundamental level application, there are two types of monitoring charts typically employed – Hotteling's *T2* and Squared Prediction Errors (SPE). The first represents conceptually the magnitude of deviation of the current sample from the center, whereas, the second analyzes the consistency of the current sample correlation according to the NOC model development. Both have been used complementary, whereby, control limits for both statistics are also computed accordingly. The main task would be to observe the progressions of both statistics on a control chart (usually Shewhart-type control chart) that constructed respectively. When the process is normal, all the statistics will remain below the control limit lines. However, whenever a fault event takes place in the process, the corresponding statistics will move away from the normal region until a point where it goes beyond the control limits specified (sooner or later). If the abnormal trend persists over a period of time, the system will then initiate the alarm which signifies that one or a combination of faulty event(s) has (have) actually occurred in the process (fault detection). Contribution plot is then applied, purposely to identify the main possible variables that either contributing or being affected from that particular abnormal event that detected. Lastly, further investigation is needed to critically sort out and finally diagnose the true source of the problem that related to that particular abnormal existence.

1.3 PROBLEM STATEMENT

MSPM normally utilizes linear-based principal component analysis (PCA) as the main technique of multivariate data compression. However, PCA sometimes is improperly used especially in modeling highly nonlinear processes as a high number of principal components (PCs) are always involved may lead to inefficient and unreliable monitoring performance reflected in false alarms and missed faults (Dong and McAvoy, 1996 and Žvokelj et al., 2011). If large variable are involved, then the PCs may also be selected considerably. As a result, Zhang et al., (1997) introduced non-linear PCA which based on the combination of neural network and principal curve but the computation is very demanding and it always requires a massive amount of data for creating the optimized NOC model (Yunus, 2012).

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter explains the methodology of the conventional as well as the proposed method. The former utilizes Multivariate Statistical Process Monitoring (MSPM) System, which basically developed based on the procedures of Principal Component Analysis (PCA) technique. Meanwhile, the structure of the proposed system applies the integration of Multiple Linear Regression (MLR) into the original MSPM framework and that to be used for data prediction. This newly introduced monitoring system is still depending on the PCA to mainly compress the multivariate data.

In general, this chapter is divided into four main sections starting briefly with the introduction. Next, the PCA-MSPM procedures (Framework I) is explained and subsequently followed by the new procedures of MLR-PCA-MSPM (Framework II). All the model development for PCA and MLR-PCA are from the MATLAB toolbox. Then, the description of the Tennessee Eastman Process (TEP) case study which has been utilized to demonstrate the capability of the monitoring system that implemented in this study which the data was generated from a simulation work due to Chiang et al., (2001).

This chapter also presents description of the real case study, Multiple Input Multiple Output (MIMO) training system pilot plant which has been setup in Universiti Malaysia Pahang, Malaysia. As to ensure the true strength of the proposed monitoring system, MIMO plant was chosen, which conceptually represents the real plant operation data that originally produced through this project. The process contains several of fault cases, which are found suitable to be applied in this research.

3.2 FRAMEWORK I: PCA-MSPM SYSTEM PROCEDURES

The complete original procedures of the conventional MSPM framework can be obtained from Macgregor and Kourti (1995) as well as Raich and Cinar (1996), whereby it can be separated into two main phases as shown in **Figure 3.1** (Yunus, 2012). From **Figure 3.1**, the first phase is related to the model development of NOC data whereas the second facilities for monitoring of the new process data.

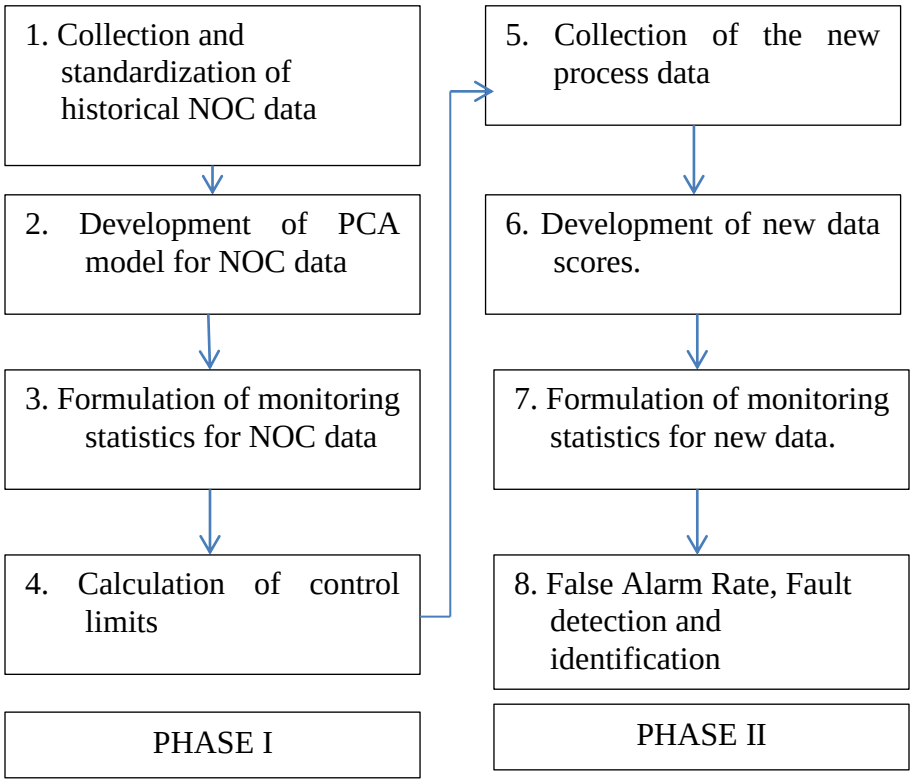

Figure 3.1: Generic MSPM Framework based on PCA.

3.2.1 PHASE I PROCEDURES

The first step of the first phase basically involves with collection of NOC data $\mathbf{X}_{mxn}$ (*m*: variables, *n*: samples), which normally conducted off-line based on the historical process data archive. The data are then standardized to zero mean and unit variance by using equation (3.1) until equation (3.3). All the procedures are due to MacGregor and Kourti, (1995); Nomikos and MacGregor, (1995) and Jackson, (2005). PCA typically utilizes variance-covariance or a matrix correlation measure of the normal operating condition (NOC) data matrix as the basis in developing the compressed multivariate configuration. According to Chiang et al., (2001), data standardization relates to capturing the data variation that extracted from the mean of data variables and scales it to unit variance. The data variables can be standardized by applying equation (3.1):