UTILISING MULTIPLE LINEAR REGRESSION (MLR) TECHNIQUE IN MULTIVARIATE STATISTICAL PROCESS MONITORING (MSPM) SYSTEM

MOHD FARID BIN MOHD NA'AIM

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of
Bachelor of Chemical Engineering

Faculty of Chemical & Natural Resources Engineering
UNIVERSITI MALAYSIA PAHANG

JULY 2014

ABSTRACT

Nowadays, modern process plants are following the trend of highly integrated and complex processes and highly instrumented chemical processes. In the most chemical processes, the need of monitoring various process variables is driven by the large amount of data produced by the instruments. Eventually, data will overload and wasted. Therefore, something needs to be done to monitor process data in lesser dimension as well as retain the most variations present. Principal Component Analysis (PCA) based MSPM technique is introduced to help us monitor and control processes. However, we have to retain too much principal component (PC) scores which complicate the fault detection operation. Thus, new MSPM technique, Multiple Linear Regression (MLR) is introduced to be utilized together with PCA to monitor predictor variables from criterion variables by equation that relates both variables. Hence, we only need to monitor criterion variables. The hypothesis for this study is if MLR is implemented, the lesser the number of variables need to be monitored. This proposed method is applied to the on-line monitoring of a simulated continuous stirred tank reactor with recycle (CSTRwR) from case study of Zhang, Martin, & Morris(1995). MATLAB software is utilized in this study. The general framework of fault detection comprises Phase I and Phase II. Phase I starts from normalisation of NOC data, PC scores formulated, monitoring statistics SPE and T^2 are calculated and lastly 95% and 99% control limits developed. Phase II starts from standardisation of fault data with respect to NOC data, PC scores developed, monitoring statistics SPE and T^2 are developed and lastly fault detection using control limits developed in Phase I. System A is the CSTRwR monitored by PCA-based MSPM system for the original set of variables. System B is the CSTRwR monitored by new MLR-PCA based MSPM system for the criterion variables which are the main product variables. Dynamic model was developed in Phase I. Then, fault data was introduced in the Phase II for fault detection. For System A, using abrupt fault data No.1 (F01a), faults were detected successfully by monitoring five variables out of 13 variables meanwhile System B only monitor two variables out of three variables with almost identical outcomes. Hence, new MSPM technique, MLR was successfully proven to be an efficient monitoring tool with quick detection and isolability while retaining as much as possible variations in lesser dimension.

ABSTRAK

Dasawarsa ini, loji proses moden mempunyai proses bersepadu dan kompleks. Kebanyakan proses kimia, memerlukan pemantauan banyak pembolehubah proses hasil dorongan jumlah data yang besar oleh instrumen. Akhirnya, data ini melebihi muatan dan disia-siakan begitu sahaja. Sesuatu perlu dilakukan untuk memantau data proses dalam dimensi lebih kecil dan juga mengekalkan variasi data. Sistem Pemantauan Pelbagai Pembolehubah secara Statistik (MSPM) berasaskan Analisis Komponen Utama (PCA) diperkenalkan untuk membantu kami memantau dan mengawal proses. Namun, terlalu banyak komponen utama (PC) Skor merumitkan operasi pengesanan data rosak. Oleh itu, teknik MSPM baru, Regresi Linear Berganda (MLR) dimanfaatkan dbersama-sama dengan PCA untuk memantau pembolehubah peramal melalui pembolehubah kriteria oleh persamaan yang berkaitan kedua-dua pembolehubah. Jadi, kita hanya perlu memantau pembolehubah kriteria. Hipotesis kajian ini adalah jika MLR dilaksanakan, maka kurang bilangan pembolehubah perlu dipantau. Kaedah ini dicadangkan untuk pemantauan atas talian berterusan reaktor tangki simulasi dikacau dengan kitar semula (CSTRwR) daripada kajian kes Zhang, Martin, & Morris (1995). Perisian MATLAB digunakan dalam kajian ini. Rangka kerja umum pengesanan data rosak terdiri daripada Fasa I dan Fasa II. Fasa I bermula dari normalisasi data NOC, skor PC dirumuskan, pemantauan statistik SPE dan T^2 dikira dan akhir sekali 95% dan 99% had kawalan. Fasa II bermula dari penyeragaman data rosak dengan data NOC, skor PC dibangunkan, pemantauan statistik SPE dan T^2 dibangunkan dan pengesanan data rosak menggunakan had kawalan dibangunkan dalam Fasa I. Sistem A ialah CSTRwR dipantau oleh sistem MSPM berasaskan PCA daripada set asal pembolehubah. Sistem B adalah CSTRwR dipantau oleh sistem MSPM baru MLR-PCA berasaskan daripada pemboleh ubah kriteria. Model dinamik telah dibangunkan dalam Fasa I. Kemudian, data rosak telah diperkenalkan pada Fasa II untuk mengesan data rosak. Sistem A, menggunakan data mendadak No.1 data (F01a), data rosak berjaya dikesan dengan memantau lima pembolehubah daripada 13 pembolehubah dan Sistem B dengan dua pembolehubah daripada tiga pembolehubah dengan hasil yang hampir sama. Oleh itu, sistem MSPM baru MLR-PCA telah berjaya terbukti menjadi alat pengawasan yang cekap.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CSTRwR	Continuous Stirred Tank Reactor with Recycle
F01a	Abrupt Fault Data No.1
F01i	Incipient Fault Data No.1
F02a	Abrupt Fault Data No.2
F02i	Incipient Fault Data No.2
F03a	Abrupt Fault Data No.3
F03i	Incipient Fault Data No.3
MLR	Multiple Linear Regression
MSPM system	Multivariate Statistical Process Monitoring system
NOC	Normal Operating Condition
PCA	Principal Component Analysis
PC scores	Principal Component scores
SPC	Statistical Process Control
SPE	Squared Prediction Error

1 INTRODUCTION

1.1 Motivation and statement of problem

Recently, in order to meet the minimum performance requirement for process plants, we need to be a step forward than others competitors in industrial world since it is getting harder and harder to maintain the performance of process plants. More stringent environmental and safety regulations, stronger competition, product reliability, consistency level and rapidly changing economic climate are among the main thing in upgrading product specification quality. Adding to the problem is that modern process plant following the trend of highly integrated and complex processes and highly instrumented chemical processes. Thus, the main challenge is to maintain the production quality and meet the market demand while in the same time, we have an abundance process variables needed to be monitored so that they do not deviate too much from the set points. In the most chemical processes, the need of monitoring various process variables is driven by the production of large amount of data by the instruments. Eventually, data will overload and most of the cases, we are unable to retrieve any useful information from it and this is termed as wastage of data. Since it is stated by Raich & Cinar, (1996) that most chemical process operations are multivariable continuous processes with collinearities among the process variables, therefore there is some approach that need to be taken in order to monitor process data in lesser dimension and in the same time retain the variations present as much as possible.

Process monitoring techniques can be used to solve the problem arise systematically. Nowadays, process monitoring system are categorized into several type such as First – Principle Process Monitoring, Knowledge – Based Process Monitoring, Pattern Recognition Process Monitoring and in this paper, the techniques that will be used is Multivariate Statistical Process Monitoring (MSPM). The traditional Statistical Process Control (SPC) concept is used widely in industries in order to determine abnormality of chemical process but it can only monitor a few variables while there are thousands of variables which are also interdependence among each other. Hence, chemometric approach is implemented in MSPM system in this paper. Wise & Gallagher, (1995) mentioned that chemometrics is the science of relating measurements made on a chemical system to the state of the system via application of mathematical or statistical methods. Wise and Gallagher, (1995) also pointed out that Principal Component Analysis (PCA) is a favourite tool of chemometricians for data compression and information extraction. The main concept of PCA is to reduce the dimensionality of data as well as retain as much as possible the variations present as much as possible.

In the previous parts, we have mentioned about the monitoring and controlling process in the industry which needed us to monitor abundance of process data at the same time and retrieved valuable information to be synthesized. PCA is introduced as the tool to help us monitor and control processes. However, PCA main limitation is we have to retain too much principal component (PC) scores and this could complicate the fault diagnosis operation. Hence, there is a need to utilize other methods together with PCA so that we can overcome the issue rise. Thus, in this study, Multiple Linear Regression (MLR) is introduced and it is defined as a mathematical tool that quantifies the relationship between a dependent variable and one or more independent variables as reported by Guillen-Casla et. al (2011). MLR divide the original data into two variables, mainly criterion variables and predictor variables. MLR will enable us to monitor predictor variables from criterion variables by equation that relates both variables.

1.2 Objectives

The main objectives of this research is to propose a new MSPM technique, where in order to reduce the number of variables in monitoring, the original variables are modelled into linear composites, where eventually also enable us to monitor performances. Hence, the objectives are

a) To develop the conventional MSPM method using Principal Component Analysis (PCA) for the original set of variables (System A).
b) To develop the conventional MSPM method using Principal Component Analysis (PCA), which utilise Multiple Linear Regression (MLR) technique. (System B).
c) To analyse the monitoring performances between System A and System B.

1.3 Research Questions

The research questions are formulated to guide our discussions and arguments in this study. These research questions are closely related to the research objectives above.

i) Can the original variables be modelled sufficiently by the utilization of MLR technique?
ii) Is there any significant differences between the monitoring performances of the proposed method compared to the conventional MSPM system?
iii) What are the requirements for the optimized operating condition to be complied in order to increase the efficiency of newly proposed technique?

1.4 Significance of Study

This study focus on the reduction of dimensionality of process data in process monitoring analysis by the utilization of MLR technique before implementing the conventional MSPM system which commonly used PCA as tools for data extraction and compression. The findings of this study are vital to assist and ease the burden of monitoring which usually involve huge number of variables with high dimensionality and complexity. Process engineer may among those that can benefit from this study. The findings from this study may bring out faster fault detection sensitiveness in chemical processes, thus enable the corrective actions to be taken faster and reduce the damages caused by the disturbances changes or set point changes in the MSPM system.

1.5 Scopes of Study

This paper is based on multivariate statistical process monitoring (MSPM) where in this paper, Multiple Linear Regression (MLR) method is utilised with Principal Component Analysis (PCA). The method will relate the variables of the process with the process itself and also it will relate certain variables on the controller in the system. The scopes of the study are:

- Mainly focus on the criterion variables for monitoring
- A continuous-stirred tank reactor with recycle (CSTRwR) system, Zhang, Martin, & Morris, (1995) is used for demonstration, whereby the faults are consisting of abrupt and incipient.
- Shewhart control chart is chosen to show the progression of the monitoring statistics which consists of T^2 chart and Squared Prediction Error (SPE) chart.
- All algorithms are developed and run based on Matlab version 7 platforms.
- The duration period for this study is two semester of academic session from 9 September 2013 until 30 May 2014.
- The process data obtained is readily available data from the case study of CSTRwR system, Zhang, Martin, & Morris, (1995).

1.6 Organization of this thesis

This thesis report is divided into five chapters which are introduction, literature review, methodology, results & discussion and conclusion & recommendations. The first chapter is the introductory part of the study. The background of the MSPM system is discussed and the current issues of the process monitoring is highlighted together with the motivation that drive this study is mentioned. The problem statement gives the potential solution of the current issues by utilizing the MLR technique in the MSPM system. Next, the research objectives and research questions of the study are clearly stated together with the significance of study and scopes of the study.

The second chapter discuss on the fundamental and theory of the PCA, historical development of PCA and limitations and extensions of PCA. After that, MLR is introduced and explained and the potential solutions of limitations of PCA which can be overcome by the utilization of MLR technique before the implementation of the PCA. The third chapter will briefly explain on the detailed description on the PCA-based conventional MSPM system (System A) and the PCA-based conventional MSPM utilized with MLR technique (System B).

The fourth chapter will present to us the results and discussion the simulation works through utilisation of MATLAB coding in order to achieve the objectives listed in this research. After that, detailed discussion of the overall monitoring results using both systems will be provided. The last chapter, chapter five concludes all the works done in the research and provides us a short summary so that we could review and improve the mistakes done in this study.

2 LITERATURE REVIEW

2.1 Introduction

The fast developing in terms of design of data-based model control has started since the early 90's. The natures of chemical processes recently which are highly integrated, complex, multivariate and non-linear make the fault detection, identification, diagnosis and monitoring processes become extremely difficult. Thus, data recording become more frequent and failing to retrieve useful information can lead to 'data wastage'. Hence, the most common monitoring technique used which is Multivariate Statistical Process Monitoring (MSPM) is taken into account in this paper. Principal Component Analysis (PCA) is introduced as a key step to carrying out MSPM (Chen., Bandoni., & Romagnoli, 1996). In this chapter, emphasis will be on the fundamentals and theory behind PCA, limits and extension of PCA and Multiple Linear Regression (MLR) and the complementary relationship between PCA and MLR to be explored.

2.2 Fundamentals and Theory of PCA

According to Jolliffe, (2002), PCA can be defined as the dimensional reduction of high dimension data and in the same time, the variations in the data are retained. It was introduced by Pearson, (1901) and further developed by Harold Hotelling in 1930s. "The objectives of PCA are data summarising, classification of variables, outlier detection , early warning of potential malfunctions and 'fingerprinting' for fault detection and it is used to summarise data with minimal loss of data" (Martin, Morris, & J.Zhang, 1996). PCA is applied in almost every discipline, chemistry, biology, engineering, meteorology and others in terms of process optimization, quality control and data visualization.

Theoretical Steps of PCA modelling

Let X = x_1,x_2,. ..,x_n, be an m-dimensional data set describing either the process variables or the quality information as stated by Zhang, Martin, & Morris, (1996). Then, the normal operating condition data (NOC data) is standardised as shown in the Equation (1) below.

$$x_{ij} = \frac{x_{ij} - x_{j,mean}}{\sigma(x_i)}, j = 1, \dots, m, i = 1, \dots, n \qquad (1)$$

We can see that it is vital to scale the NOC data first before performing the PCA such that,

> "Principal component analysis depends critically upon the scales used to measure the variables. If we consider a set of multivariate data where the variables, x_l, x_2, x_3, ..., x_m, are of completely different types, for example pressures, temperatures, flow rates, etc., then the structure of the principal components derived from this data set will depend essentially upon the arbitrary set of units of measurement. If there are large differences between the variances of x_l, x_2, x_3, ..., x_m, those variables whose variances are large will tend to dominate the first few principal components. It is found that in practice these variables may not be of prime importance in detecting process malfunctions. This lack of scale invariance implies that care needs to be taken when scaling the data. Different scaling routines can produce different results. Three possible ways to scale the data are: select 'natural units' by ensuring all the variables measured are of the same type; variables can be mean-centred; or the variables can be scaled to zero mean and unit variance". (Martin, Morris, & J.Zhang, 1996)

The standardized data describing normal conditions are stored in a (n x m) matrix X=[z_{ij}], with the m sensor values for each observation arranged on each of n rows as mentioned by Lewin, (1995).Next, The (m x m) correlation matrix, A, which is symmetrical and positive definite, is formed:

$$A = X^T X \tag{2}$$

The eigenvalues, λ_j and eigenvectors, p_j of A are then computed in decreasing order of magnitude ($\lambda_1 > \lambda_2 > \ldots > \lambda_m$). The original data can then be expressed in terms of the eigenvectors, which define the principal component directions:

$$X = t_1 p_1{}^T + t_2 p_2{}^T + \ldots + t_k p_k{}^T + \ldots + t_m p_m{}^T \tag{3}$$

Where $t_j = Xp_j$ is the (n x 1) score vector, a projection of the data onto the j-th principal component vectors. An approximate model, X compromising of the first k terms of (3) will capture the most of the observed variance in X if the data is correlated. Number of principal components should be determined by the eigenvalues. After the numbers of principal components have been determined, thus the data matrix X can be represented by:

$$X = TP = \left[\hat{T}\ \tilde{T}\right]\left[\hat{P}\ \tilde{T}\right]^T \tag{4}$$

where matrix T contained retained principal components and ignored principal components meanwhile matrix P contained retained eigenvectors and ignored eigenvectorsas pointed out by Harrou, Nounou, Nounou, & Madakyaru, (2012). Expanding the equation, we get

$$\boldsymbol{X} = \hat{\boldsymbol{T}}\hat{\boldsymbol{P}}^T + \tilde{T}\tilde{P}^T = \boldsymbol{X}\hat{P}\hat{\boldsymbol{P}}^T + \boldsymbol{X}\left(\boldsymbol{I}_m - \hat{\boldsymbol{P}}\,\hat{\boldsymbol{P}}^T\right) \tag{5}$$

where matrix $\hat{\boldsymbol{X}}$ represent the modelled variation of X based on the k first components and matrix E represent the variations corresponding to process noise.

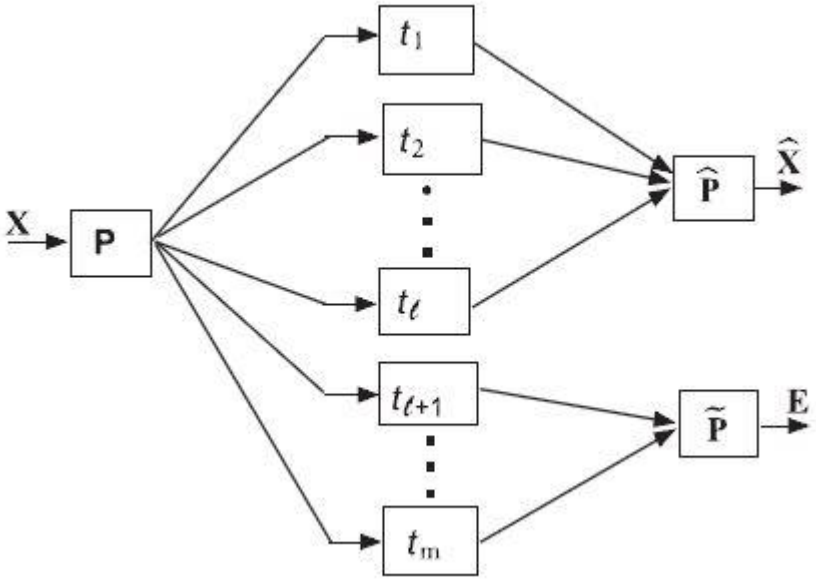

Figure 2. 1 Representation of data as a sum of approximate and error parts using PCA

The estimation of the matrices $\widehat{\boldsymbol{X}}$ and **E** is shown in Figure 2.1. Measured vector **x** can be expressed as the sum of two orthogonal parts, approximated vector $\widehat{\boldsymbol{X}}$ and residual vector $\widetilde{\boldsymbol{X}}$ using the PCA.

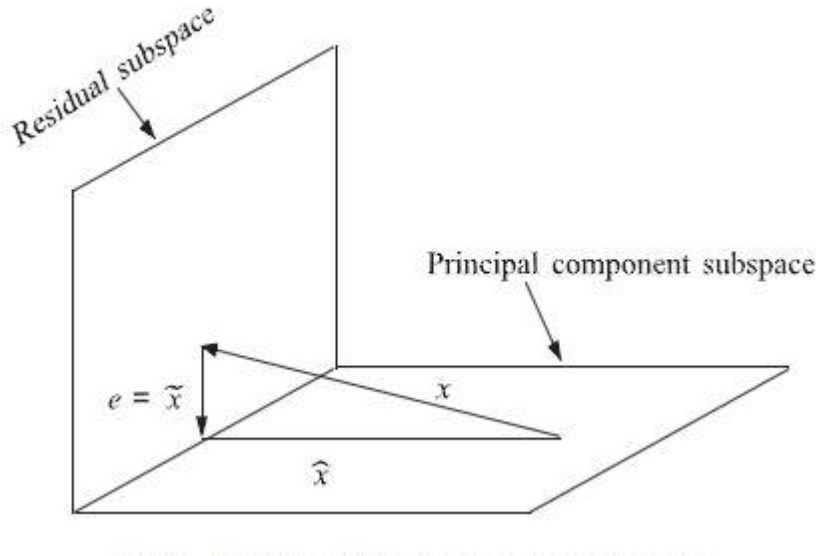

Figure 2. 2 Decomposition of measurement vector

Figure 2.2 shows the decomposition of measurement vector. In a fault – free situation, the residual vector, $\widetilde{\boldsymbol{X}}$ is commonly negligible but it can significantly increase if a fault is present. In process monitoring, fault detection process using PCA will initially use fault – free data to construct PCA model. Then, the PCA model will use detection indices Hotelling's T^2 and Squared Prediction Error (SPE) to detect faults. Next, after the principal components are developed, the monitoring statistics are developed which are Hotelling's T^2 and Squared Prediction Error (SPE).

$$T^2 = x^T P\widehat{\boldsymbol{A}}^{-1}\hat{P}^T x \qquad (6)$$

Hotelling, (1933) defined Hotelling's T^2 as statistic which measure the variations in the principal components at different sampling time. $\widehat{\boldsymbol{A}}$ is diagonal matrix containing the eigenvalues correlated to the retained k principal components. A fault declared if the value of T^2 in the new testing data exceeds the values of T^2 developed in PCA model.

On the other hand, Q statistic or SPE statistic defined as the measurement of the projection on the data on the residual subspace, which provides an overall measure of how a data sample fits the PCA model, (Wise & Gallagher, 1995).

$$Q_i = \boldsymbol{e}_i \boldsymbol{e}_i^T = \boldsymbol{x}_i\left(\boldsymbol{I} - \boldsymbol{P}_k \boldsymbol{P}_k^T\right)\boldsymbol{x}_i^T \qquad (7)$$

where $\mathbf{e_i}$ is the i^{th} row of **E**, P_k is the matrix of the first k loadings vectors retained in the PCA model (where each vector is a column of P_k) and **I** is the identity matrix of appropriate size (n x n).

2.3 Historical development of PCA

The history of statistic techniques often gets stuck with the origin where they were started. However, in terms of PCA, it is widely recognised that PCA was described by Pearson, (1901) and further developed by Hotelling, (1933). Hotelling's paper consists of two parts. The first part is parallel to the approach introduced by Pearson, (1991) which more focused on the best fit lines and planes on a set of points in p-dimensional space together with the optimization of geometric problems that also considered to lead to PCs. Meanwhile, the other part of Hotelling, (1933) adopted the approach which works with the standard algebraic derivation.

Pearson, (1901) grabbed the attention by saying that the application of his approach can easily solve numerical problems and added that even for four or more variables, the calculations might get complicated, still they are solveable. These comments were made 50 years before the availability of computer spreaded across the world. Meanwhile, among Hotelling's best contribution was to suggest the terms 'components' in order to differentiate the other uses of the word 'factor' in mathematic.

The components that are derived in Hotelling's approach are named 'principal components. That is the starting point of the evolutions and further development on PCA theoretically and this represents general growth of statistical techniques. In this case, we need to take into the consideration that since computing power is needed by PCA, thus the spread of PCA happened coincidently with expansion of introduction of electronic computers. Pearson, (1991) can be optimistic with his comments on the easiness of solving numerical problems using his method, but to be frank, it is not feasible to work out PCA using hands unless the variables are less than four. The best out of the PCA can only be exploited for larger number of variables, so only after the invention of computer; PCA can be used to its full potential.

2.4 Limits and extension of PCA

Although PCA is good for linear or almost linear problems, it fails to deal well with the significant intrinsic nonlinearity associated with real-world processes. Hence, nonlinear extensions of PCA have been investigated by different researchers, such as Dong & McAvoy, (1996) and (Shi, 2011).

Dong & McAvoy, (1996) expressed that, in non-linear process, PCA might discard minor components which may contain useful information. This is due to variables that have minimal impacts in the first two principal components but might dominate lower order component as interpreted by Martin, Morris, & J.Zhang, (1996). We can see here that we need to come out with some approach to improve the efficiency of PCA in MSPM system.

2.5 Fundamentals and theory of Multiple Linear Regressions (MLR)

"Multiple Linear Regression (MLR) is a regression model that contains more one regressor variable" (Runger & Montgomery, 2011). It is utilised in the situation of more than one predictor variable, the situation that happens commonly in modern chemical processes. Frequently, MLR models used as approximating functions that give insight into relationship between criterion and predictor variables as stated by Runger & Montgomery, (2011).

Theoretical steps of Multiple Linear Regression (MLR)

This is the general form of multiple linear regression:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad (8)$$

where for a set of *i* observations,

Y_i = the criterion variable

β_0 = a coefficient

β_1; β_2;.. β_p = the coefficients of X_{i1}, $X_{i2..}$ X_{ip} independent variables (predictor variables)

ε_i = residual error (difference between observations and predicted values).

The hypotheses required to apply multiple linear regression as mentioned by Agirre-Basurko, Ibarra-Berastegi, & Madariaga, (2006) are:

i) the predictor variables must be independent

ii) the residual errors ε_i must be independent and they must be normally distributed, with 0 mean and σ^2 constant variance.

The observations $\{X_{i1}, X_{i2},.,X_{ip}, Y_i\}_{i = 1,2,.,n}$ are helpful in the estimation of the parameters β and they form the calibration set. The least square method is the usual technique used to estimate the parameters.

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{N}(X_{i1} - \overline{X_{i1}})(\widehat{Y_i} - \bar{Y})\sum_{i=1}^{N}(X_{i2} - \overline{X_{i2}})^2 - \sum_{i=1}^{N}(X_{i2} - \overline{X_{i2}})(\widehat{Y_i} - \bar{Y})\sum_{i=1}^{N}(X_{i1} - \overline{X_{i1}})(X_{i2} - \overline{X_{i2}})}{\sum_{i=1}^{N}(X_{i1} - \overline{X_{i1}})^2\sum_{i=1}^{N}(X_{i2} - \overline{X_{i2}})^2 - \left(\sum_{i=1}^{N}(X_{i1} - \overline{X_{i1}})(X_{i2} - \overline{X_{i2}})\right)^2} \quad (9)$$

In the Equation (9), the formula is formulated for a least squares estimated coefficient in an equation with two independent variables based on Baker paper (2013). Since more than two independent variables as demonstrated in Zhang case study which has ten predictor variables, the formula can get even more complicated. Hence, it is more efficient if matrix algebra used. However, that still not make them easier to be calculated in the spreadsheet. This is where MATLAB as a specific computer program utilised for the calculation of multiple linear regression coefficients from the criterion and predictor variables. Hence, the equation for the predicted value is:

$$\widehat{Y_i} = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_p X_{ip} \quad (10)$$

where,

$\widehat{Y_i}$ = the criterion variable

b_i = the estimationsof the β_i parameters

2.6 Historical Development of MLR

The history of multiple of linear regression (MLR) technique cannot be exactly dated back to its origin but the term regression come from Francis Galton in late 1880s which originally used the term in biology. It is his student, Pearson (1901) was given the credit later on to move the term to a more general statistical context. His contiued research on 1924 eventually introduced residuals method of least squares regression instead of trend ratio from least square trend regression used before that to remove time trend. Several years later, Metzler (1940) combined both the residual and trend square least regression.

After the removal of time trend, the regression of the demand curve was done. It was done by a pair of Henry Schultz, Pearson's former student and Henry Moore, PhD supervisor of Schultz. They used multiple - correlation and 'line of best fit' methods. Multiple – correlation method, which is referred to as 'multivariate linear regression' in today's textbook terminology was used by Moore in a search for 'dynamic law of demand in its complex form'. However, Schultz found out measurement errors prevalent in economic data (since he was one of the distingusihed advocates for general equilibrium economics) and decided to choose 'line of best fit' method over the least squares method.

Stern objection given from Gilboy (1931) which claimed Schultz's method failed to specify the supply demand curve and eventually the price – quantity points on the scatter diagram could not be a demand curve. Plus, she accused that Schultz's empirical demand curve was static on the ground that a successful detrending that could remove all the dynamic elements. Meanwhile, interpreting the residuals of regressions becomes a serious problem in 1920s and 1930s and this problem was viewed from both residuals on measurement errors and disturbances in variables perspective. Later on, it was concluded that there would be much easier if economic theory provided some prior knowledge about the true relation. Consequently, Kyun (2006) concluded that least square method was proposed which was developed as a mean of approximate representation and thus belongs to mathematics in general and not exclusively to statistics.

2.7 Limits and extension of MLR

The goal of the regression analysis is to determine the values of the parameters of the regression equation and then to quantify the goodness of the fit in respect of the dependent variable Y. The main advantages of MLR over other multivariate projection methods are computational simple and the capability of deriving coefficients which directly relate to the original data like remarked by Qin, Liu, Liu, & Tong, (2009). Nevertheless, MLR also limit to certain situations only. Sometimes, MLR are over fitting the data, dimensionality of data, poor prediction and inability to work on ill conditional data as observed by Qin, Liu, Liu, & Tong, (2009). Hence, something needs to be done to improve the disadvantages of MLR.

2.8 Potentials of utilising MLR in MSPM system

In order to make fault detection process become more easier and faster, MLR can be fitted before PCA to build empirical models from non-linear experimental data which can serve as approximating functions to reduce number of the criterion variables that exist in variation-described PCA models. Hence, it can remove indirect effect of variables which dominate the minor components in PCA but do not have impact in the first two principal components. Above inference can be obtained from the integration of different researchers' thoughts such as Placca et al. (2010), Camdevyren et al. (2005) and Martin, Morris, & J.Zhang, (1996). This approach can be the potential solution of the inadequacy of application of PCA in non-linear chemical processes. Previous works done by the researchers mentioned earlier show that MLR are more utilised in the other method of monitoring technique.

However, in this study, the focus is given on the function of MLR to divide the original variables, X_0 into criterion variables, Y and predictor variables, X*. The criterion variables are the output or quality variables meanwhile predictor variables are the input or disturbances variables. Both criterion and predictor variables are related by equations which are alternative tools to reduce the number of the monitoring variables that exist before we implement PCA to detect 'out of control' status. Thus, monitoring processes become easier and detection time of the fault may be shortened.