UTILIZING MLR TECHNIQUE IN MSPM SYSTEM

WAN HAIRANI BINTI WAN NA'AMAN

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

JANUARY 2014

ABSTRACT

The main purpose of this research is to propose a new MSPM technique, where the original variables are modelled into linear composites in order to reduce the number of variables in monitoring, where eventually it may also monitoring the performances. Hence, the objectives are to develop the conventional MSPM method for the original set of variables (System A), to develop the conventional MSPM method, which applies Multiple Linear Regression (MLR) technique (System B) and last but not least is to analyse the monitoring performances between System A and System B. By doing this research, it was proved that, it is able to justify that the developed and upgraded MSPM method is comparatively better than conventional PCA-based MSPM method in monitoring the multivariate of non-linear process. The research also show the development of advanced multivariate way of process monitoring in terms of variables points besides of samples scores. 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) where MLR technique is applies in phase I. All the mathematical model were developed into coding of matlab platform of version 7. The results were presented in the form of graphs or Shewart Chart of 95% and 99% confident limit with the Hotelling T-Squared Distribution and Squared Prediction Errors was considered has the statistical tools for observing. Both method which are PCA-based MSPM System and MLR-based MSPM System were compared it performance and had been analysed. From the results, MLR-based MSPM Sytem gives better performance interm of faster detection of fault compared with PCA-based MSPM System. As the conclusion, all the three objectives were achieved successfully.

Key Words: MSPM, PCA-based MSPM System, MLR-based MSPM Sytem

ABSTRAK

Maklamat utama kajian ini dijalankan adalah untuk mencadangkan satu teknik baru dalam sistem MSPM, dimana model komposit linear di bina untuk memantau pembolehubah asal yang diuji dalam kajian ini supaya tidak berubah malah mengurangkan bilangan pembolehubah yang dipantau, walaupon berkurang tetapi ia tidak mencacatkan proses pemantauan. Oleh itu, objektif untuk kajian ini adalah untuk membangunkan kaedah MSPM konvensional untuk set asal pembolehubah (sistem A), untuk membangunkan kaedah konvensional MSPM dimana teknik MLR di aplikasikan dalam sistem (Sistem B) dan akhir sekali adalah untuk menganalisis dan memantau perbandingan prestasi pemantauan antara sistem A dan sistem B. Dengan menjalankan kajian ini, ia telah membuktikan bahawa dengan membangunkan dan yang telah di ubah suai kaedah MSPM adalah lebih baik jika dibandingkan dengan kaedah konvensional PCA-berasakan MSPM dalam memantau multivariate bukan proses linear. Kajian ini juga telah menunjukkan perkembangan yang termaju dalam proses pemantauan cara multivariate dari segi pembolehubah selain daripada skor sampel. Prosedur-prosedur yang lengkap untuk mengesan dan mengenalpasti kerosakan terdiri daripada dua fasa utama iaitu sebagai model luar talian dan pemantauan (Fasa I) dan pemantauan dalam talian (Fasa II) di mana teknik MLR di aplikasikan dalam fasa I. Semua model matematik telah diubah menjadi pengekodan platform MATLAB versi 7. Hasil kajian telah dipersembahkan dalam bentuk graf atau Carta Shewart dengan 95% dan 99% had yakin dengan Pengagihan Hotelling T- Squared dan Squared Kesilapan Ramalan digunakan sebagai alat statistik untuk pemerhatian. Kedua-dua kaedah Sistem MSPM berasaskan PCA dan Sistem MSPM berasaskan MLR dibandingkan prestasinya dan dianalisis. Daripada keputusan kajian yang dicapai, Sistem MSPM berasaskan MLR memberikan prestasi yang lebih baik dari segi pengesanan kesilapan yang lebih cepat berbanding sistem MSPM berasaskan PCA. Sebagai kesimpulan , ketiga-tiga objektif telah dicapai dengan jayanya.

Kata Kunci: MSPM, Sistem MSPM berasaskan PCA, Sistem MSPM berasaskan MLR

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURE

LIST OF SYMBOLS

Y_i	Value of the Response Variable in the i^{th} Observation
β_0	Intercept Parameter,
β_1	Slope Parameter,
x_{1i}	Value Of The Independent Variable In The i^{th} Observation,
ε_i	Random Error Term Of The i^{th} Observation With Mean
k	Slope Parameter Associated With The k^{th} Variable,
R^2	Coefficient of Determination
T^2	Hotelling's T-Squared Distribution
C	Covariance
N	Samples
m	Variables
$\check{x}_{ji}$	Standardized Data for Variable 'i' at Sample 'j'
Λ	Eigenvalues Matrix for PCA.
σi	Standard Deviation for Variable 'i'.
α	Level of Confidence Limits.
P	Score Matrix

LIST OF ABBREVIATIONS

MSPM	Multivariate Statistical Process Monitoring
MLR	Multiple Linear Regression
PCA	Principle Component Analysis
SPC	Statistical Process Control
SQC	Statistical Quality Control
MSPC	Multivariate Statistical Process Control
CSTRwR	Continuous-stirred Tank Reactor with Recycle
PCs	Principle Components
MDS	Multidimensional Scaling
QR	Quantile Regression
NOC	Normal Operation Condition
SSE	Sum of Square Error
SPE	Squared Prediction Errors
NPC	Number of Principle Component
TEP	Tennessee Eastman Process
XMEAS	Vector of Process Variable Measurement
XMV	Vector of Manipulated Variables

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Nowadays, laboratory instruments produce great quantities of data. This creates a data overload and usually a big amount of these data are wasted. The problem is to compress and/or to extract relevant information. Generally, there is a great deal of correlated or redundant information in procedure measures. This information must be compressed in a manner that retains the essential information and is more easily displayed than each of the variables individually. Also, essential information often lies not in any individual variable but in how the variables change with respect to one another for example how they co-vary. In chemical and process industries, the most important thing is to produce the maximum amount of consistently high quality products as per requested and specified by the customers.

This is regarded as highly challenging due to the nature of the processes that always change over time and are also affected by various factors such as variations of raw materials as well as operating conditions, the presence of disturbances and also modification in the process technologies. Besides that, the influence of the surrounding factors including changing in market demands, the environmental impacts, restricting of the workforces and the unpredictable revolution in the management policies, may also to certain extent affect the product quality as well as the productivity of the production system. In any of the situations, one of the main critical problems is to promptly detect the occurring of faulty or abnormal operating conditions.

1.2 BACKGROUND OF STUDY

In process monitoring, the main objective is always to detect changes or departure behaviour from the normal process characteristics. Due to the nature of the processes that always change over time and are affected by several sources, process monitoring become more challenging in any chemical based industries.

Application of statistical methods in monitoring and control of industrial processes are included in a field generally known as statistical process control (SPC) or statistical quality control (SQC) (Damarla, 2011). The most widely used and popular SPC techniques involve univariate methods, that is, observing and analyzing a single variable at a time. Statistical Process Control (SPC) is an effective method of monitoring a process through the use of control charts. By collecting data from samples at various points within the process, variations in the process that may affect the quality of the end product or service can be detected and corrected, thus reducing waste as well as the likelihood of passing down to the customer. Thus, early detection and prevention of the problems are both crucial in this respect.

However, industrial quality problems are multivariate in nature, since they involve measurements on a number of variables simultaneously, rather than depending on one single variable. As a result, Multivariable Statistical Process Control (MSPC) system (Kano et al., 2001) is introduced, where a set of variables which are the manipulated variables and controlled variables are identified and the jointly monitored. In conclusion, early detection and diagnosis of process faults while the plant is still operating in a controllable region can help avoid abnormal event progression and reduce productivity loss.

1.3 PROBLEM STATEMENT

Monitoring and controlling a chemical process is a challenging task because it involves a huge number of variables. Usually, process monitoring is executed based on the principal-component analysis (PCA) technique, nevertheless it has its own limitation. As the number of variables grows, the fault detection performance tends to

be slow in progression, as well as, introduce greater complexity in the later stages especially in fault identification and diagnosing. Thus it is desirable to reduce those variables, while embedding them into a single measurement model, where the original variations can still be preserved.

1.4 RESEARCH OBJECTIVES

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

i. To develop the conventional MSPM method for the original set of variables (System A).
ii. To develop the conventional MSPM method, this applies Multiple Linear Regression (MLR) technique (System B).
iii. To analyse the monitoring performances between System A and System B.

1.5 RESEARCH QUESTIONS

i. Can the MLR technique sufficiently be used to model the original variables?
ii. How are the generic monitoring performances of the proposed method as compared to the traditional scheme?
iii. What is the optimized condition which must be complied in order to improve the new technique?

1.6 SCOPES OF STUDY

The research is based on multivariate statistical process monitoring (MSPM) where in this research; multiple linear regression method is used. 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:

i. Mainly focus to select only certain variable

ii. A Tennessee Eastman process is used for demonstration
iii. Shewhart control chart is chosen to show the progression of the monitoring statistics.
iv. All algorithms are developed and run based on Matlab version 7 platforms.

1.7 SIGNIFICANCE OF STUDY

This study produces a new idea on how to reduce the complexity of monitoring analysis by using MLR technique in modelling all the variables involved. The method is expected to improve the monitoring progressions especially in terms of fault detection sensitiveness.

1.8 REPORT ORGANIZATION

The proposed report is divided into five chapters which are the introduction, literature review, methodology, result and discussion and also conclusion. The first chapter renders an overview of statistical process control (SPC), multivariate process and their use in process monitoring. This chapter also presents the objectives of the present work, scope, the expected outcome and significance of the proposed study. The second chapter emphasizes on multivariate statistical process monitoring, fundamentals and theory of conventional MSPM system, classical SPC and univariate statistical tools, correlation analysis, multivariate statistical tools, principle component analysis (PCA), statistical associate with PCA models and MLR as an alternative solution for MSPM system. In chapter three, both conventional method and multiple linear regression method was been presented. Chapter four was discussing on the result and discussion of the research. Finally, conclusion and recommendation have been discussed in chapter five.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Since the last decade, design and development of data based model control has taken its momentum. This very trend owes an explanation. Identification and control of chemical process is a challenging task because of their multivariate, highly correlated and non-linear nature. Very often there are a large number of process variables are to be measured thus giving rise to a high dimensional data base characterizing the process of interest. To extract meaningful information from such a data base; meticulous pre-processing of data is mandatory. Otherwise those high dimensional dataset maybe seen through a smaller window by projecting the data along some selected fewer dimensions of maximum variability. This chapter will emphasize on the multivariate statistical process monitoring, fundamentals and theory of conventional MSPM system, classical SPC and univariate statistical tools, correlation analysis, multivariate statistical tools, principle component analysis (PCA), statistical associate with PCA models and MLR as an alternative solution for MSPM system.

2.2 MULTIVARIATE STATISTICAL PROCESS MONITORING

MSPM is considered as the best option in monitoring complex as well as considerably large scale industrial systems. There are also other methodologies such as model-based and knowledge-based techniques, but impractical particularly when concerning the huge scale and complexity issues (Chiang et al., 2001). In general, such difficulties are the rigidity, validity as well as difficulty in the development of first principle models, credibility of the process knowledge used as well as spurious decision

outcomes, and not to mention complications as well as inflexibility in updating recent information for the improvement of monitoring operation (Venkatasubramanian et al., 2003a; 2003b; 2003c). Nevertheless, these non-statistical process monitoring techniques may undoubtedly become more productive when concerning the diagnostic phase in contradiction to MSPM which is heavily dependent on the credibility of the process history data alone. Venkatasubramanian et al., (2003a; 2003b; 2003c) have also suggested that this can be modified further, perhaps by using a hybrid system that integrates various sets of techniques which works complementary with each other.

2.3 FUNDAMENTALS AND THEORY OF CONVENTIONAL MSPM SYSTEM

There are also other terminologies such as 'Multivariate Statistical Process Control' (MSPC) (Martin, et al., 1996; Kano et al., 2000; 2001; 2002; Bersimis et. al., 2007) or 'Multivariate Methods for Process Monitoring' (Kourti and MacGregor, 1995) or 'Statistical Process Control'(SPC) in multivariate process (MacGregor and Kourti, 1995) or 'Statistical Process Monitoring' (SPM) (Raich and Cinar, 1996) that have been used to represent the MSPM methodology. In other words, all of these systems denote the same monitoring mechanism that systematically utilizes statistical analysis in capturing the essential process information based on a correlation model from a set of variables of the collected historical normal operational process data (Yoon and MacGregor, 2000). Nevertheless, the depth of the monitoring scopes defined by those works differs from one to another.

Research by Smith (2002) illustrated that Process Control Analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since pattern in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. According to Bakshi (1998), PCA is a MSPC technique used for the purpose of data compression without losing any valuable information. Principal components (PCs) are transformed set of coordinates orthogonal to each other. The first PC is the direction of largest variation in the data set. The

projection of original data on the PCs produces the score data or transformed data as a linear combination of those fewer mutually orthogonal dimensions. PCA technique was applied on the auto-scaled data matrix to determine the principal eigenvectors, associated eigen values and scores or the transformed data along the principal components. The drawbacks are that the new latent variables often have no physical meaning and the user has a little control over the possible loss of information. Generally, PCA is a mathematical transform used to find correlations and explain variance in a data set.

2.4 CLASSICAL SPC AND UNIVARIATE STATISTICAL TOOLS

Statistical process control (SPC) involves monitoring the performance of a process over time to verify that it is remaining in a state of statistical control Marlin (2000) and Levinson (1990). Such a state of control is said to exist if certain process or product variables (usually small in number) remain close to their desired values and the only source of variation is common-cause variation, that is, variation which affects the process all the time and is essentially unavoidable within the current process. Abnormal process conditions are considered as events having special or assignable causes. Their occurances are identified and the sources (i.e. root cause) of such disturbances are eliminated.

The ultimate goal of both APC and SPC is to improve products and the processes used to make them. However, in contrast to APC, SPC achieves this goal by making the process less susceptible to future upsets. However, SPC alone cannot adequately control most process operations. The benefits are also long-term because although SPC uses experience and empirical models derived from real-time measurements, “control” is through infrequent manipulated variable movements and is usually not carried out in real-time. The techniques used in SPC can be loosely classified into (1) Analysis and (2) Regression. Analysis involves drawing useful conclusions from a single block of data. Regression involves building quantitative relationships between more than one block of data.

The foundations of SPC have been laid by Shewart (1931). The Shewart (see Figure 2.0), the cumulative sum (CUMSUM) by Hunter J (1986) and the exponentially weighted moving average (EWMA) (Woodward et.al., 1964) charts are widely used SPC tools in industry. Processes improvements can also be attained by using univariate and multivariate statistical tools to carry out data analysis or for building regression models.

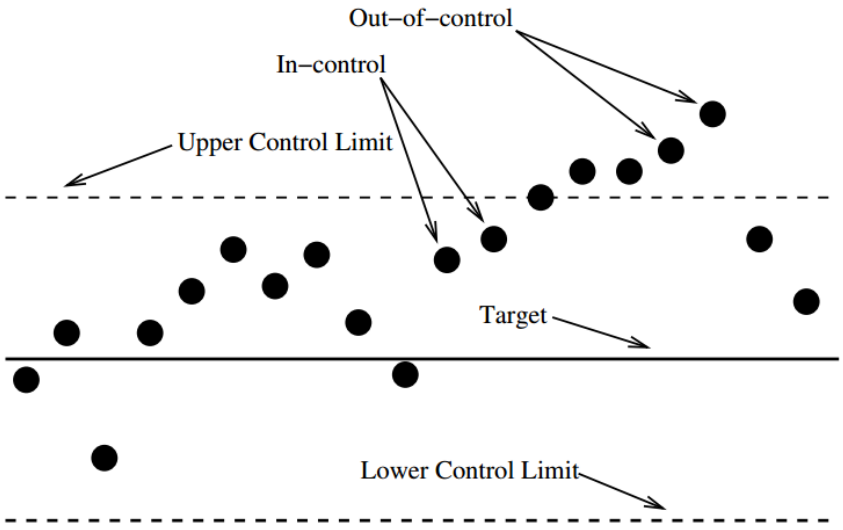

Figure 2.0: Shewart Chart of Control limit

Consider a data-set X of process measurements. X is of dimension $n_s \times n_y$ with x(t)(= $x_1(t)$, $x_2(t)$, ...$x_n(t)$) as the vector of measurement variables (or simply measurements) collected at sample time t. Typically the measurements contain noise and many of them are linearly dependant, i.e. there is redundancy in the data. As a result, n s and n y are very large, i.e. this data set is of a very high dimensionality. In order to reduce the dimensionality and identify the underlying correlation, several statistical techniques maybe employed.

2.5 CORRELATION ANALYSIS

In order to determine the redundancy in two sets of time-series data, it is necessary to determine whether two variables are correlated (i.e. parallel or collinear). There are several measures of redundancy. Consider two process measurements x and y, with means $\bar{x}$ and $\bar{y}$ respectivey, each a vector of length n_s. Covariance is defined as:

$$cov(x,y) = \frac{\sum_{i=1}^{n_s}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Correlation is usually expressed using the correlation coefficient ($r_{x,y}$) and is defined as:

$$r_{x,y} = \frac{\sum_{i=1}^{n_s}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sqrt{\sum_{i=1}^{n_s}(x_i - \bar{x})^2}\right)\left(\sqrt{\sum_{i=1}^{n_s}(y_i - \bar{y})^2}\right)}$$

when two different measurements are being compared, this quantity is known as the cross-correlation coefficient while when two time periods of the same measurement are being compared, the term used is auto-correlation coefficient. $r_{x,y}$ values range from -1 to +1. The extreme values can be interpreted as:

$$r_{x,y} = \begin{cases} 1 & \mathbf{x} \text{ and } \mathbf{y} \text{ are totally positively correlated} \\ 0 & \mathbf{x} \text{ and } \mathbf{y} \text{ are completely un-correlated i.e. independant} \\ -1 & \mathbf{x} \text{ and } \mathbf{y} \text{ are totally negatively correlated} \end{cases}$$

Intermediate values of $r_{x,y}$ reflects lesser degrees of correlation. It should be noted that correlation does not imply a causal relationship.

2.5.1 Multiple Linear Regressions (MLR)

Multiple Linear Regression (MLR) also known as Ordinary Least Squares (OLS) is an Inverse Least Squares (ILS) method. It calculates the pseudo inverse of X as:

$$\mathbf{X}^{+} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

2.5.2 Limitations of Univariate Techniques

Univariate techniques are not capable of dealing with collinear variables. Moreover, univariate techniques only perform some sort of noise averaging but do not remove noise in the truest sense.

2.6 MULTIVARIATE STATISTICAL TOOLS

Multivariate statistical techniques are becoming increasing popular in many diverse fields where they are variously referred to as Econometrics (in Economics), Biometrics (in Biology) or Chemometrics (in chemistry, particularly analytical chemistry). These techniques are used to perform a myriad of different tasks such as exploratory data analysis, pattern recognition, sample classification, discriminant analysis, data mining, bioinformatics, fault detection, etc. Although there may be slight differences in the nomenclature used, the underlying fundamental principles are the same. Owing to this similarity, research in this area is often cross-disciplinary. Extending classical SPC, which is traditionally univariate in nature, to multivariate cases, numerous researchers Kresta et al (1991), and Wise and Gallagher (1996) have recently discussed applications in chemical process analysis and control. Several multivariate statistical techniques such as Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares (PLS) and others may be employed.

2.6.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA), also known as the Karhunen-Loeve (KT) transform, was originally developed by Pearson (1901). It involves a matrix decomposition that gives rise to a new set of variables, known as the Principal Components (PCs), by transforming a given input matrix into two matrices. Another way of looking at this mathematical procedure is that PCA reveals the hidden “real” variables and so the PCs are often referred to as Latent Variables (LVs). Symbolically,

$$\left.\begin{aligned}
\mathbf{X} &\equiv \theta_1\mathbf{p}_1^T + \theta_2\mathbf{p}_2^T + \ldots + \theta_r\mathbf{p}_r^T + \ldots + \theta_l\mathbf{p}_l^T \\
&= \textstyle\sum_{i=1}^{r} \theta_i\mathbf{p}_i^T + \sum_{i=r+1}^{l} \theta_i\mathbf{p}_i^T \\
&= \textstyle\sum_{i=1}^{r} (-\theta_i)(-\mathbf{p}_i^T) + \sum_{i=r+1}^{l} (-\theta_i)(-\mathbf{p}_i^T) \\
&= \mathbf{\Theta}\mathbf{P}^T + \mathbf{E} \\
&= \hat{\mathbf{X}} + \mathbf{E}
\end{aligned}\right\}$$

i. The pi vectors are known as the PC factors or loadings. They are orthonormal (i.e.$p_i^T p_j = 0\ for\ i \neq j, p_j^T p_j = 1\ for\ i = j$) and provide the direction of the PCs. For each PC, there are as many loadings as there are variables in the input matrix (i.e. the P matrix is $n_{y \times r}$). The p_i vectors are the eigenvectors of the covariance matrix, i.e. for each p_i,

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_i$$

where the λ_is are the eigenvalues.

ii. The θ_i vectors are known as the PC scores. They are orthogonal (i.e.$\theta_i^T\theta_i = 0$ for $i \neq j$) and reflect the magnitude of the PCs. For each PC, there are as many scores as there are samples in the input matrix (i.e. the Θ matrix is $n_{s \times r}$). The score vector θ_i is the linear combination of the original X variables defined by p_i. In other words, the θ_i are the projections of X onto the p_i.

$$\theta_\mathbf{i} = \mathbf{X}\mathbf{p}_\mathbf{i}$$

iii. Each θi pTi pair is referred to as the ith PC. Their outer product forms a matrix of rank 1. They are arranged in order of decreasing eigenvalues (i.e. λ1 ≥ λ2 ≥· · · ≥ λr ≥ · · · ≥ λl). The variance captured by each PC is proportional to their eigen values. From a statistical standpoint, variance is equated with information. Hence, the first PC accounts for as much of the variability in the data as possible, and hence carries maximum information. Each succeeding PC accounts for as much of the remaining variability as possible.

iv. l is the maximum number of PCs. It is the smaller of the number of variables and the number of samples (i.e. l = min (n_s, n_y)). The PCA model, denoted by$\hat{X}$, is formed by retaining only a few PCs (i.e. r < l). The matrix formed by the min or components (r + 1 to l) is not included in the model truncated in this way and is referred to as the Residual Matrix (denoted by E). This matrix contains the un important variance or noise and any non-linearities, if present.

v. PCA is very closely related to Singular Value Decomposition (SVD). When all the components are retained (i.e. r = l) then the residual matrix vanishes. For this case, $\Theta = \mu\Sigma$ and P = V.

vi. Wise et al (1990) have argued that a theoretical connection between PCA and state space models exists. It has been demonstrated that, for processes where there are more measurements than significant states, variations in the process states appear primarily as variations in the PCA scores, while noise mainly affects the residuals. Hence, when limits on the PCA residuals are being derived, only the noise properties of the system have to be taken into consideration while the dynamics of the process do not have to be considered explicitly.

From a process analysis and control perspective, PCA offers several advantages. In all cases PCA derives its utility by determining the right value of r and discarding the trailing r + 1 to l PCs. PCA can be used to build models for prediction/estimation. For example, new values of the scores can be estimated from new measurements as follows. First a PCA model is built using the calibration data, i.e. X = ΘPT. Measurements from the new data are centered using the same mean and variance as the calibration data. For this scaled data, $X^{new} = \Theta^{new}P^T$ and so $X^{new}P = \Theta^{new}P^TP$. The loading vectors are orthonormal and so $P^T = P{-}1$. Hence, a new score vector can be estimated as $\theta^{new}_i = X^{new}p_i$. When PCA is used to solve linear regression problems the approach is known as Principal Component Regression (PCR). The pseudo inverse of X is:

$$\mathbf{X}^{+} = \mathbf{P}(\mathbf{\Theta}^{T}\mathbf{\Theta})^{-1}\mathbf{\Theta}$$

2.6.2 Statistics Associated With PCA Models

Several statistics associated with PCA models can also be used as measures to detect abnormal behavior in processes. Two of the most commonly used ones are the squared prediction error (SPE) and the Hotelling T^2. SPE is often referred to as the lack of fit statistic or the Q-residual. For the k^{th} sample from a set of measurements, these statistics are defined as:

$$Q_k = x_k(I - \mathbf{P}\mathbf{P}^T)x_k^T$$

$$T_k^2 = t_k \lambda^{-1} t_k^T$$

SPE indicates how well each sample conforms to the PCA model. As a result it also detects any new variations occurring in the process. The T^2 statistic is a measure of the variation in each sample within the model. In other words, it captures larger than normal variations.

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 (Zhao and Xu, 2004). Both the strength and weakness of PCA is that it is a non-parametric analysis. PCA is also commonly viewed as a Gaussian model; that is, the data is assumed to come from a Gaussian distribution.

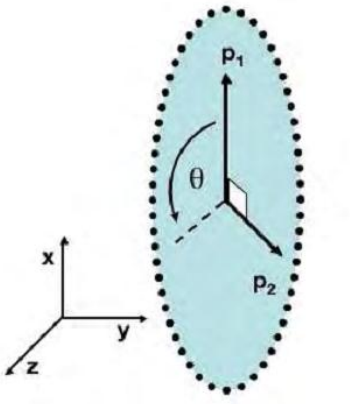

Figure 2.1: Data point which track a person on a ferris wheel

For example, from Shlens (2005) study, we can consider the recorded positions of a person on a ferris wheel over time in Figure 2.1. The probability distributions along the axes are approximately Gaussian and thus PCA finds (p1, p2), however according to Shlens, this answer might not be optimal. The most concise form of dimensional reduction is to recognize that the phase or angle along the ferris wheel contains all dynamic information. Thus, the appropriate parametric algorithms are to first convert the data to the appropriately centered polar coordinates and then compute PCA.

This prior non-linear transformation is sometimes termed a kernel transformation and the entire parametric algorithm is termed kernel PCA. Other common kernel transformations include Fourier and Gaussian transformations. This procedure is parametric because the user must incorporate prior knowledge of the structure in the selection of the kernel but it is also more optimal in the sense that the structure is more concisely described. One might envision situations where the principal components need not be orthogonal. Furthermore, the distributions along each dimension (xi) need not be Gaussian. If we are using a probabilistic interpretation of PCA, we might want to assume that the data is Gaussian because uncorrelated Gaussian random variables are also independent. Because PCA decorrelates the data, the resulting encodings in the basis of the principal components are independent. The random processes generating the encodings might then be thought of as the underlying independent causes of the data.

According to Nikolov (2010), in ICA, it is assumed that there are d independent, non-Gaussian random variables which is traditionally called sources, and that they are transformed by a mixing matrix W to give a measurement $x = (x^1, \ldots, x^d)$. This gives the relationship between the measurements x and the sources y:

$$x = Wy.$$

The goal of ICA then is to recover W and y given x, only by looking at the statistical structure of x. There is a lot to be said about this problem, and there are many techniques for solving it including maximization of nongaussianity, maximum likelihood, minimization of mutual information between components of the encoding,

maximization of mutual information between the data and the encodings, nonlinear decorrelation, and diagonalizing higher order cumulant tensors. Compared to independence, uncorrelatedness is a relatively weak statement to make about a set of random variables. However, uncorrelatedness can mean something stronger if we first pass the transformed data y through a nonlinearity and then decorrelate it. Whereas decorrelating the components of y involves only second-order statistics such as making the covariances zero, this nonlinearity brings higher order statistics into play when modeling the data.

2.7 MLR AS AN ALTERNATIVE SOLUTION FOR MSPM SYSTEM

Multiple linear regressions (MLR) are a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes also called the predictand, and the independent variables the predictors. MLR is based on least squares: the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized. MLR is probably the most widely used method in dendroclimatology for developing models to reconstruct climate variables from tree-ring series. Typically, a climatic variable is defined as the predictand and tree-ring variables from one or more sites are defined as predictors. The model is fit to a period – the calibration period – for which climatic and tree-ring data overlap.

In the process of fitting, or estimating, the model, statistics are computed that summarize the accuracy of the regression model for the calibration period. The performance of the model on data not used to fit the model is usually checked in some way by a process called validation. Finally, tree-ring data from before the calibration period are substituted into the prediction equation to get a reconstruction of the predictand. The reconstruction is a “prediction” in the sense that the regression model is applied to generate estimates of the predictand variable outside the period used to fit the data.

The uncertainty in the reconstruction is summarized by confidence intervals, which can be computed by various alternative ways. Regression has long been used in