UNIVERSITY OF STRATHCLYDE
DEPARMENT OF PURE AND APPLIED CHEMISTRY

MONITORING OF COMPLEX NONSTATIONARY INDUSTRIAL PROCESSES

Atakan Sahin

A thesis submitted to the Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, in part fulfilment of the regulations for the degree of Doctor of Philosophy

FEB 2020

COPYRIGHT

Signed:

Date:

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Dr. Alison Nordon for her unlimited patience and helps through my PhD studies. Her support and encouragement are one of the critical motivations to complete this thesis. I am also grateful to Prof. Dr. Julian Morris and Dr. Jiazhu Pan for their guidance in my PhD studies.

I gratefully acknowledge the financial support from European Union's Horizon 2020 research and innovation programme named 'ModLife' (Advancing Modelling for Process-Product Innovation, Optimization, Monitoring and Control in Life Science Industries) under the Marie Sklodowska-Curie grant agreement number 675251.

In addition to academic support, I also sincerely thank my girlfriend, Emma-Louise for her years of love and support, and the jokes.

February 2020 Atakan Sahin

ABSTRACT

Monitoring of complex manufacturing processes using multivariate statistical process control (MSPC) is becoming more important. However, classical MSPC is restricted to stationary data while most industrial processes are nonstationary. One way of addressing nonstationary data is to calculate the difference between consecutive time series data samples. However, this can cause loss of dynamic information, resulting in inadequate process monitoring or a reduced fault detection capability. Cointegration analysis has recently been adopted for process monitoring of nonstationary processes. However, the first applications considered only nonstationary variables whereas complex industrial processes contain both stationary and nonstationary variables. Furthermore, there is inefficiency in the modelling when dealing with higher level nonstationary time series. This particular issue can be solved by using common-trend residuals-based monitoring. However, the use of different models requires a number of control charts to be monitored by a data analyst. To solve these issues, a multi-level multi-factor model is proposed for the monitoring of complex continuous and batch industrial processes. The method uses a combination of principal component analysis (PCA), and cointegration and common-trend models at the 1st level, and then a PCA model at the 2nd level to monitor the combined stationary outputs from the 1st level. The method is tested with ramp and step type fault functions on continuous and batch process simulations, and compared with conventional PCA and cointegration based approaches. The findings show that the multi-level multi-factor model can provide better fault detection rates compared to conventional PCA and cointegration based approaches. In addition, a parameter tuning scheme based on the big-bang big-crunch global optimisation algorithm is used to select the optimum parameters for the multi-level multi-factor model when applied to continuous and batch processes. This not only improves the model's performance but also assists with its practical application in an industrial environment.

TABLE OF CONTENTS

Page

ABBREVIATIONS

2D	**:** Two Dimensional
3D	**:** Three Dimensional
ACF	**:** Auto-Correlation Function
ADF	**:** Augmented Dickey-Fuller
ANN	**:** Artificial Neural Network
AR	**:** Auto Regressive
ARMA	**:** Auto Regressive – Moving-Average
ARIMA	**:** Auto Regressive Integrated Moving-Average
BB-BC	**:** Big-Bang Big-Crunch
CUSUM	**:** Cumulative Sum
CPV	**:** Cumulative Percentage of Variance
CSTH	**:** Continuous Stirred Tank Heater
CSTR	**:** Continuous Stirred Tank Reactor
CVA	**:** Canonical Variate Analysis
DF	**:** Dickey-Fuller
DPCA	**:** Dynamic Principal Component Analysis
EG	**:** Engle-Granger
EWMA	**:** Exponentially Weighted Moving Average
FDI	**:** Fault Detection and Isolation
ICA	**:** Independent Component Analysis
IID	**:** Independent Identical Distributed
IoT	**:** Internet of Things
KPSS	**:** Kwiatkowski, Phillips, Schmidt, and Shin
LCL	**:** Lower Control Limit
MA	**:** Moving Average
MSPC	**:** Multivariate Statistical Process Control
NIPALS	**:** Nonlinear Iterative Partial Least Squares
OLS	**:** Ordinary Least Squares
PC	**:** Principal Component
PCA	**:** Principal Component Analysis
PLS	**:** Projection to Latent Structures / Partial Least Squares
PP	**:** Philips-Perron
RPLS	**:** Recursive Partial Least Squares
SPC	**:** Statistical Process Control
SPE	**:** Squared Prediction Error
SQC	**:** Statistical Quality Control
SVD	**:** Singular Value Decomposition
SVM	**:** Support Vector Machine
TEP	**:** Tennessee Eastman Process
UCL	**:** Upper Control Limit
VAR	**:** Vector Auto Regressive
VARMA	**:** Vector Auto Regressive Moving Average
VECM	**:** Vector Error Correction Model

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. INTRODUCTION

1.1 Motivation

Information technologies have had many improvements in data storage and its usage in the last decades and it will have more, according to recent technological breakthroughs. Data storage and data recording will be much easier than now after the ongoing internet of things (IoT) revolution. Industry 4.0 is a well-known name for the IoT revolution. As the data acquisition process has become easier, this has resulted in a data explosion for several sectors such as informatics and bioinformatics (Ashton, 2009). According to Gartner's 2018 report on strategic technology trends, one of the most promising tools is intelligent analytics (Panetta, 2017), which includes multivariate statistical analysis.

In today's manufacturing practices, innovation is the primary strategy for improvement; sometimes, the survival of the companies. The manufacturing paradigm in the pharmaceutical industry has evolved from quality-by-testing to quality-by-design (Lopes and Sarraguça, 2018). This has resulted in process analytical technologies becoming more important at all stages in manufacturing. The Made Smarter Review, which was commissioned by the UK government, looks at digitalisation of UK industry by 2030 and what is required across different sectors (Department for Business, Energy & Industrial Strategy, 2017). One of the critical points identified for improvements in the performance of food manufacturing was a need for data-driven real-time decision support systems.

Multivariate statistical process control (MSPC) techniques provide tools for the comprehensive on-line monitoring of manufacturing processes and the on-line detection of process malfunctions, and are capable of being applied to both continuous and batch processes. MSPC techniques play an essential role in maintaining the quality of manufacturing by providing data-driven real-time decision support systems. However, the capability of classical MSPC, based on projection-based methods such as principal component analysis (PCA) and partial least squares (sometimes termed projection to latent structures) (PLS), is restricted to stationary systems/variables. Most industrial processes are nonstationary in nature, which may be caused by seasonal changes, processes that involve filling and emptying, throughput changes, the presence of unmeasured disturbances, and operator interventions, etc. However,

dealing with nonstationary variables has only recently started to receive increasing attention. One way of addressing nonstationary data is to calculate the difference between consecutive time series data samples or use of difference based autoregression models such as autoregressive integrated moving average (ARIMA). ARIMA comes with an enormous computational burden for multivariate processes because of the number of variables. It is also known that variable differencing can lead to the loss of dynamic information.

A promising tool called cointegration has been adopted into process monitoring to cope with nonstationarity. Cointegration is arguably the most effective way of handling the nonstationary characteristics of data and was proposed to formulate the problem of the existence of linear equilibria. It has been used extensively in the area of econometrics, and more recently, in several disciplines of science and engineering to reflect any long-run information, which can be easily removed via de-trending and differencing. Even though cointegration analysis is a powerful tool, it can give rise to a cointegration matrix of low rank when there is a large number of nonstationary variables if there is high level nonstationary present. This issue can be solved through the use of a common-trend model, which can model the nonstationary factors remaining in the low rank cointegration matrix. However, it gives rise to another problem which is the increased number of the control charts associated with all of the different models. It is a disadvantage compared to conventional MSPC approaches that only require a single control chart based on T^2 and Squared Prediction Error (SPE) metrics. On the other hand, the cointegration residuals-based monitoring method is applicable only to the nonstationary variables; these are only a part of the data from complex industrial processes, which also comprise stationary variables. Therefore, a new process monitoring method for fault detection purposes is required that can be used with batch or continuous processes, processes that exhibit both stationary and nonstationary (including high-level) characteristics, and the output of the model can be displayed in a single control chart comprised of the T^2 and SPE metrics.

1.2 Objectives

The research presented in this thesis is part of the ModLife (Advancing Modelling for Process-Product Innovation, Optimization, Monitoring and Control in Life Science Industries) project, a H2020 innovative training network (ITN) funded under the Marie

Sklodowska-Curie grant agreement number 675251. The overall goal of ModLife is to develop advanced model-based optimisation, monitoring and control as enabling technologies for bioprocess-product development and innovation tailored for the needs of the life science industries. The ModLife ITN aims to develop the next generation of high-performance computing tools and in-situ measurements for increasing the efficiency, innovation and competitiveness of Europe's life sciences and processing industries.

The overall aim of the research described in this thesis is to develop a methodology for monitoring of complex industrial processes that comprise stationary and nonstationary variables. The methodology must also be applicable to both continuous and batch processes. More specifically, the objectives of the research can be summarised as follows:

1. To develop a new process monitoring approach for fault detection that can be used with complex continuous industrial processes.
2. To adapt and extend the method devised in objective 1 for use with complex batch industrial processes.
3. To develop a parameter tuning scheme based on a global optimisation algorithm to determine the optimum design parameters for multi-level multi-factor models when used for monitoring of batch and continuous processes.
4. To apply the methods developed in objectives 1 to 3 to example continuous and batch processes, and to compare the performance of the new method against current state-of-the-art methods reported in the literature.

1.3 Contributions to Knowledge

A new process monitoring approach, termed multi-level multi-factor, has been devised for monitoring of continuous processes. This method was then extended to enable monitoring of batch processes through incorporation of a step to divide the batch into multiple phases, with each phase taken as a continuous process. The multi-level multi-factor model consists of 2 levels and 4 sub-models across the 2 levels. The sub-models include PCA models for each level, and cointegration and common-trend residuals-based process monitoring models at the 1st level. In the 1st level, the stationary variables are modelled by PCA while the nonstationary variables are modelled by

cointegration and common-trend models to determine the stationary factors from both modelling techniques for the 2nd level PCA model. The model combines the advantages of three existing approaches, namely PCA, cointegration and common-trend models, and the performance advantage over current state-of-the-art methods has been demonstrated using simulations of continuous and batch processes. The multi-level multi-factor method presented is the first method based on cointegration analysis, which considers all process variables (stationary and nonstationary) in the continuous process monitoring.

The multi-level multi-factor model has a number of design parameters such as the number of principal components (PCs) for the 1st and 2nd PCA models, phase length for multi-phase batch process monitoring, and the combination of nonstationary variables for the cointegration models when the number of the nonstationary variables exceeds 12; the Johansen test that is used in cointegration analysis of multivariate systems only supports 12 or less variables. The simple reason for this limitation is that no mathematician has computed the critical values for more than 12 variables. To assist the data analyst with the selection of all of the design parameters and to provide optimum performance of the models, a parameter tuning scheme based on big bang-big crunch global optimisation algorithm has been developed and applied to example continuous and batch processes. Use of the optimisation algorithm will assist with the practical implementation of the multi-level multi-factor method.

1.4 The Layout of the Thesis

The thesis comprises 7 chapters. This chapter (Chapter 1) gives the motivation, and the aims and objectives of the research described in the following chapters. The contributions to knowledge made by this thesis are also presented.

Chapter 2 provides an overview of process performance monitoring and projection based MSPC techniques for process performance monitoring. It gives an introduction to statistical process control (SPC), MSPC and well-known conventional MSPC methods (PCA and PLS), and the metrics used in control charts for process monitoring. Furthermore, the data characteristics of complex industrial processes are summarised with a critical review of the literature highlighting the limitations of currently proposed methods.

Chapter 3 discusses the topics of nonstationarity, unit root tests to find nonstationarity, and cointegration, which is arguably the most effective tool for the modelling of nonstationary variables. Process monitoring techniques based on cointegration and common-trend residuals are introduced and reviewed as a means of monitoring of nonstationary variables.

Chapter 4 presents monitoring techniques for continuous processes: conventional PCA, dynamic PCA (DPCA), cointegration residuals-based model, common-trend residuals-based model and the new multi-level multi-factor model. A comparison of the performance of the models for fault detection has been conducted using data from a continuous stirred tank heater simulator

Chapter 5 presents monitoring techniques for batch processes: multi-PCA, a multi-level model and an extension of the new multi-level multi-factor model to accommodate multi-phase modelling. A comparison of the performance of the models for fault detection has been performed using data from an industrial penicillin simulator.

Chapter 6 presents a parameter tuning scheme, based on the BB-BC global optimisation algorithm, to enable the design of optimum multi-level multi-factor models. It helps to search for several design parameters through the multi-level multi-factor model, which is quite troublesome for a data analyst to carry out manually.

Finally, Chapter 7 concludes with a summary of the research and its industrial impact, along with suggestions for future work.

2. MULTIVARIATE STATISTICAL PROCESS CONTROL (MSPC)

2.1 Overview

In this section, an overview of multivariate statistical process control (MSPC) techniques is presented. The projection-based techniques such as principal component analysis (PCA) and projection to latent variables or sometimes termed partial least squares (PLS) are discussed. Even though such conventional MSPC methods are powerful, they are not appropriate for some complex industrial processes.

Statistical process control (SPC) addresses a range of techniques which are used to investigate whether a particular manufacturing or production process is operating in a state of '*statistical*' control where products must meet manufacturing standards with little, if not zero, variability. A production system that shows only common cause variation is said to be statistically in-control. Shewhart, one of the originators of the early control charts defined a state of control as "a phenomenon will be said to be controlled when, through the use of experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction within limits means that we can state, at least approximately, that the probability that the observed phenomenon will fall within the given limits" (Shewhart, 1930). Sometimes abnormal variabilities within the manufacturing process may be present as a result of errors such as machine error, operator error differing process feed materials, etc.

In the early days, SPC tended to be associated more with the monitoring of individual quality characteristics of the product, i.e. statistical quality control (SQC). In control charts, limits were defined for the measured process or the quality parameter of the end product under the natural process variability. When all the measurements fall within the pre-specified control limits, the process is said to be '*statistically-in-control*' (Montgomery, 2001). Any abnormalities caused by unexpected process variation gives rise to the chart crossing the determined control limits. Investigations can proceed to determine and correct the reason for that occurrence. A simple statistically-in-control chart is shown in Figure 2.1.

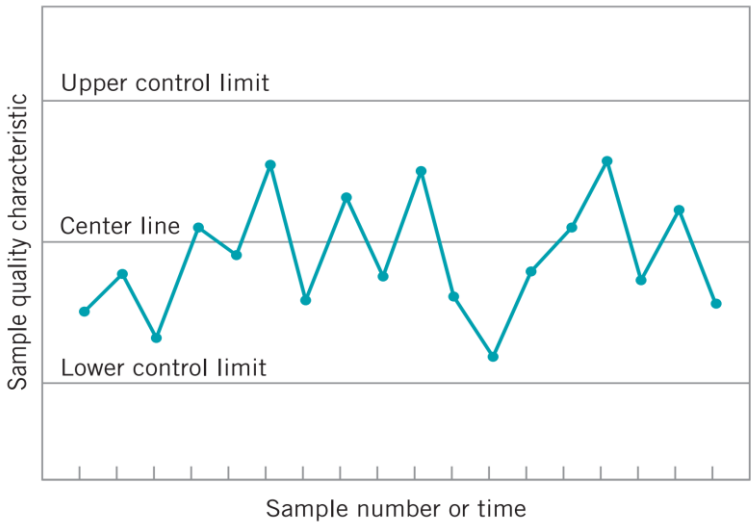

Figure 2.1: Example of a univariate control chart. Figure based on data from (Montgomery, 2001).

Control charts are closely related to statistical hypothesis testing. Control charts test the new observation to determine whether it falls within the control limits or not. If it falls within the control limits, the process is said to be '*statistically-in-control*', and the null hypothesis cannot be rejected. In contrast, if the new value crosses the control limits, it implies that the process is out-of-control and the null-hypothesis should be rejected. The most commonly applied univariate tools reflect the Shewhart principles of mean and range, cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control charts.

2.1.1 Univariate Statistical Process Monitoring

The first and probably the most known control chart was introduced by Walter Shewhart (Shewhart, 1930). Since then, they have found many applications in the process industries. The Shewhart control charts are a family of monitoring tools, which can be used to check the statistics of the mean or variability of key variables inside the processes. The upper and lower control limits for monitoring the mean value (target value) of a variable are given by:

$$CL_{Shewhart} = \hat{\mu} \pm A\hat{\sigma} \tag{2.1}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated values of the mean and standard deviation of the monitored variables used in the modelling. A is the level constant that dictates the capability of the chart. The Shewhart control chart is typically used to detect significant

shifts ($> 3\hat{\sigma}$) in processes. The statistics in the Shewhart chart can be plotted either with a single measurement or the average of the previous n measurements. Montgomery suggested the use of four or five measurements within the range of $2\hat{\sigma}$ (Montgomery, 2001). Having control limits farther from the centre line decreases the probability of a type-I error, where type-I refers to states where a point is falling beyond the control limits. Conversely, wider control limits also increase the probability of a type-II error where type-II refers to states where a point is falling between the control limits when the process is really out of control. On the other hand, moving the control limits closer to the centre line gives rise to the opposite effects: an increase in the risk of a type-I error, and a decreases in type-II error (Montgomery, 2001).

Shewhart control charts have been in use for more than 75 years. Additional charts such as the cumulative sum (CUSUM) chart introduced by Page (Page, 1954) and the exponentially moving average (EWMA) chart presented by Roberts (Roberts, 1959) are typical 'memorising' control charts which produce better results for small shift detections that can frequently occur.

The CUSUM chart has been studied by many authors, including Woodall and Adams (Woodall and Adams, 1993) and Ewan (Ewan, 1963). The CUSUM chart combines the cumulative sum of the deviation between the previous samples and the target value and current information from the measurement. Consequently, it is more effective than Shewhart charts in detecting small changes from the mean value. The CUSUM chart is founded on the assumption that the measured process variables are stationary.

Montgomery provided a formal definition for stationarity (Montgomery, 2001). A time series can be considered stationary if (i) the expected value of the time series is not dependent on the time, and (ii) the autocovariance function defined by $Cov(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})$ is only a function of k, not a time where $\boldsymbol{x}_t$ is the sample observation and k is any lag. The impact of nonstationary data, which is against any predetermined target, unlike stationary data, is discussed in the following sections.

Time series dependency refers to the dependence of the observation $\boldsymbol{x}_{t+1}$ on the previous observation $\boldsymbol{x}_t$. Autocorrelation is used to name such a reliance where the distribution of the measured process variables is normal with a known mean and variance $N(\mu, \sigma^2)$ (Montgomery, 2001).

2.1.2 Extensions of Univariate Control Charts for Multivariate Processes

Univariate control charts provide a tool to monitor process performance. However, the use of these charts is limited since they only consider one variable at a time. Complex processes, on the other hand, exhibit interactions between the variables which can give rise to misleading information in univariate control charts.

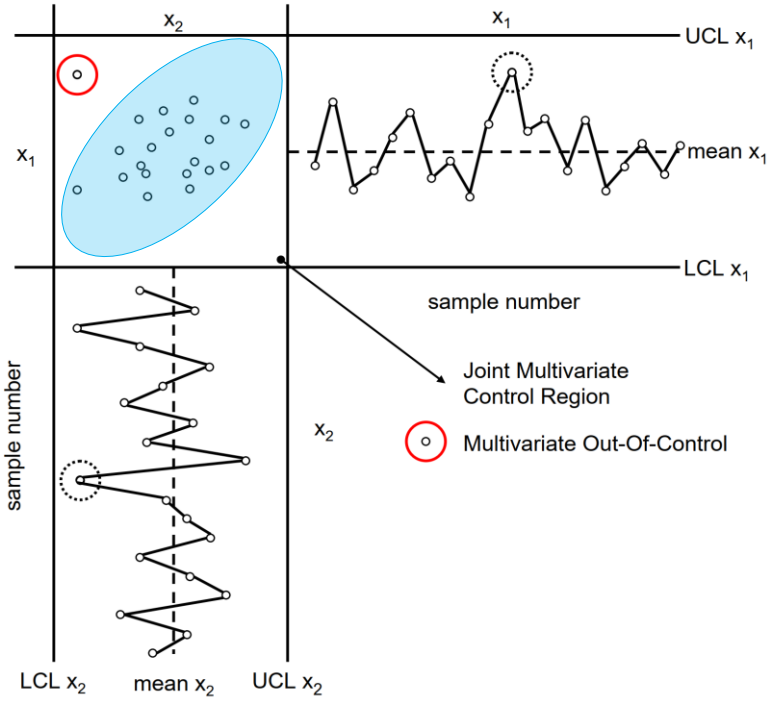

Figure 2.2: Illustration for the multivariate ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) monitoring with defined control eclipse, and upper control limit (UCL) and lower control limit (LCL).

Figure 2.2 illustrates two process variables both with univariate mean control charts, upper control limits (UCLs) and lower control limits (LCLs). Both charts show that the monitored variables are in control. However, one point shows an abnormality within the joint multivariate control region.

Hotelling established multivariate process control techniques in 1947 (Hotelling, 1947). It represents the multivariate counterpart of the Shewhart control chart based on Hotelling's T^2 statistics as a pioneering study of MSPC.

2.2 Multivariate Statistical Process Control Using Projection Based Techniques

The standard multivariate procedures to reduce the dimensionality of the process variables are projection techniques like PCA and PLS models. Both are model-based approaches using a historical data set that is assumed to be in control (Bersimis,

Psarakis and Panaretos, 2007). The first studies and applications of multivariate methods were made by John MacGregor's group (Kresta, Macgregor and Marlin, 1991) and Barry Wise's group (Wise and Ricker, 1991). Following the model determination, future samples are checked in the diagnosis part of modelling to evaluate whether the sample fits the model or not. They can handle process variables and quality variables. The PCA approach provides the basis of MSPC based fault detection and diagnosis when only the process variables ($\boldsymbol{X}$) are available. A PLS model is developed using the process ($\boldsymbol{X}$) and quality variables ($\boldsymbol{Y}$). Figure 2.3 summarizes the usage of both representations (Qin, 2012). The main focus of this study is PCA based monitoring techniques for fault detection and diagnosis. In the following sections, conventional PCA and PLS models are discussed for process monitoring purposes.

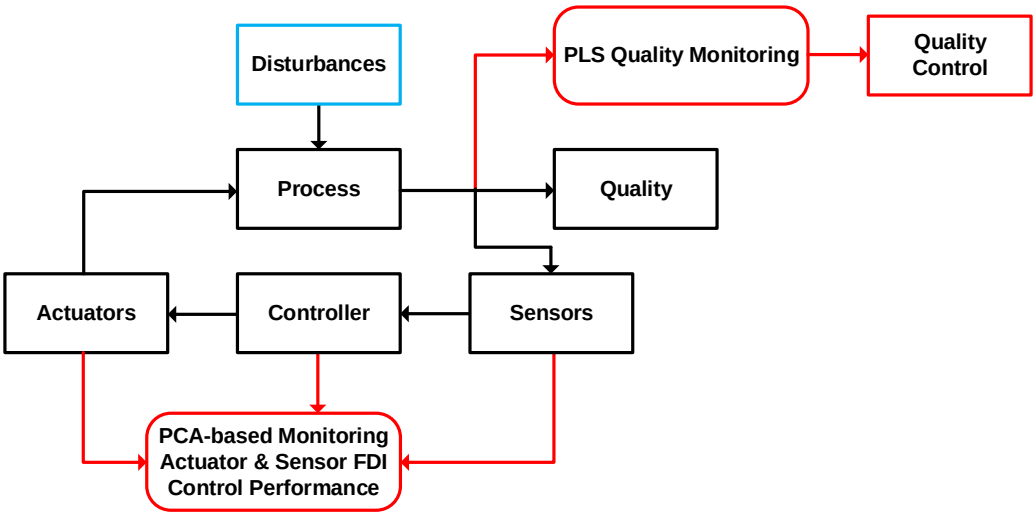

Figure 2.3: Process and quality monitoring problem in fault detection and diagnosis

2.3 Principal Component Analysis

The use of PCA is arguably the most popular MSPC methodology after the adoption by Hotelling of the T^2 statistic (Abdi and Williams, 2010). Originally, PCA was developed by Pearson to find the closest line and planes to the variables. PCA analyses the variance-covariance structure of a data matrix $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ to extract important information that is a linear combination of $\boldsymbol{X}$ and to express it as a set of new orthogonal variables called *principal components* where N is the number of the variables and M is the number of samples (Goodall and Jolliffe, 1988).

Figure 2.4 illustrates the orthogonal PCs that create a plane which is also perpendicular to the third PC if it exists. Here, the first PC represents the most significant amount of variation of the data. The methodology is directly applicable to continuous processes and can be extended to batch processes.

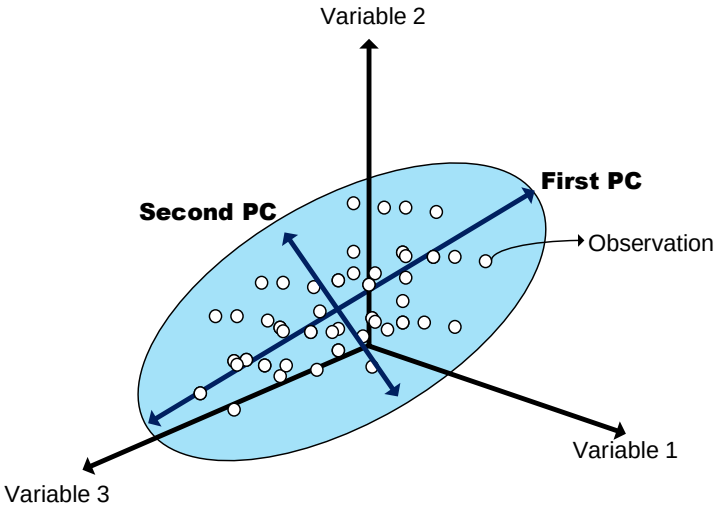

Figure 2.4: Graphical representation of principal component analysis.

In general, $\boldsymbol{X}$ will be pre-processed before the analysis. Mostly, this pre-processing is done by standardizing the data to transform the data onto unit scale (mean and variance of each row are equal to 0 and 1). PCA performs the eigen decomposition on the covariance matrix ($cov(\boldsymbol{X}) \in \mathbb{R}^{N\times N}$) where each element represents the covariance between two variables. The covariance matrix can be represented as below:

$$cov(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1N} \\ \vdots & \ddots & \vdots \\ \sigma_{N1} & \cdots & \sigma_{NN} \end{bmatrix} \tag{2.2}$$

where the sample covariance between two variables, j and k, is calculated as follows:

$$\sigma_{jk} = \frac{1}{M-1}\sum_{i=1}^{M}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{2.3}$$

where $\bar{x}$ is the mean of the variable. PCA decomposes $\boldsymbol{X}$ as a sum of the outer product of the vectors $\boldsymbol{t}_r$ and $\boldsymbol{p}_r$ where $\boldsymbol{t}_r$ and $\boldsymbol{p}_r$ are the scores and loading vectors, respectively, for the r^{th} component.

$$\boldsymbol{X} = \sum_{r=1}^{R} \boldsymbol{p}_r \boldsymbol{t}_r^T + \boldsymbol{E} = \boldsymbol{P}\boldsymbol{T}^T + \boldsymbol{E} \tag{2.4}$$

$\boldsymbol{E} \in \mathbb{R}^{N \times M}$ represents the model residuals, or model errors, in addition to the statistical model. R is the maximum number of principal components under the condition of $R = \min(M, N)$. Here, the score matrix $\boldsymbol{T} \in \mathbb{R}^{M \times R} = (\boldsymbol{t}_1, \dots, \boldsymbol{t}_R)$ contains information on how the samples are related to each other, while the loading matrix $\boldsymbol{P} \in \mathbb{R}^{N \times R} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_R)$ defines the interrelation of the variables. Since the columns of $\boldsymbol{T}$ are orthogonal ($\boldsymbol{t}_r^T \boldsymbol{t}_{r-1} = 0$) and the loading columns are orthonormal ($\boldsymbol{p}_r^T \boldsymbol{p}_{r-1} = 0$ and $\boldsymbol{p}_r^T \boldsymbol{p}_r = 1$), the covariance matrix can be written as (Qin, 2003):

$$cov(\boldsymbol{X}) = \boldsymbol{P}^T \boldsymbol{\Lambda} \boldsymbol{P} \tag{2.5}$$

where

$$\boldsymbol{\Lambda} = \frac{1}{M-1} \boldsymbol{T}^T \boldsymbol{T} = diag\{\lambda_1, \dots, \lambda_N\} \tag{2.6}$$

where λ_i are the eigenvalues of the covariance matrix in descending order of the represented variation of the data. When M is very large, then the eigenvalues can be found from the sample variance of the ith score vector as follows:

$$\lambda_1 = \frac{1}{M-1} \boldsymbol{t}_i^T \boldsymbol{t}_i \tag{2.7}$$

Decomposition of $\boldsymbol{X}$ can be performed by either nonlinear iterative partial least squares (NIPALS) (Wold, Esbensen and Geladi, 1987) or the singular value decomposition (SVD) algorithm. These approaches are discussed in more detail in Appendix-A. Selection of the number of principal components to be used is an essential step that impacts the PCA model.

The cumulative percentage of variance (CPV) is a measure of how much variation is captured by the first r PCs:

$$CPV_r = \left({\textstyle\sum_{i=1}^{r} \lambda_i} \Big/ {\textstyle\sum_{i=1}^{N} \lambda_i} \right) 100\% \tag{2.8}$$

The number of PCs selected can be based on criteria chosen by the data analyst.

2.4 Projection to Latent Structures

Projection to latent structures or partial least squares is a regression technique based on two data blocks, the input ($\boldsymbol{X} \in \mathbb{R}^{N\times M}$) and output ($\boldsymbol{Y} \in \mathbb{R}^{N_o\times M}$) matrices where N_o is the number of output variables. In general, the data sets mentioned above include state variables for $\boldsymbol{X}$ and quality measurements for $\boldsymbol{Y}$.

PLS projects $\boldsymbol{X}$ and $\boldsymbol{Y}$ to a low dimensional space defined by L latent variables:

$$\begin{cases} \boldsymbol{X} = \sum_{i=1}^{L} \boldsymbol{p}_i \boldsymbol{t}_i^T + \boldsymbol{E} = \boldsymbol{P}\boldsymbol{T}^T + \boldsymbol{E} \\ \boldsymbol{Y} = \sum_{i=1}^{L} \boldsymbol{q}_i \boldsymbol{u}_i^T + \boldsymbol{F} = \boldsymbol{Q}\boldsymbol{U}^T + \boldsymbol{F} \end{cases} \tag{2.9}$$

where $\boldsymbol{T} \in \mathbb{R}^{M\times L} = (\boldsymbol{t}_1, \dots, \boldsymbol{t}_L)$ and $\boldsymbol{U} \in \mathbb{R}^{M\times L} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_L)$ are the latent score matrices with $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$ vectors and $\boldsymbol{P} \in \mathbb{R}^{N\times L} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_L)$ and $\boldsymbol{Q} \in \mathbb{R}^{N_0\times L} = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_L)$ are the loading matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively (Geladi and Kowalski, 1986). The number of latent factors is often determined by cross-validation.

The latent vectors $(\boldsymbol{t}_i)$ is computed sequentially by the NIPALS algorithm from the data to maximise the covariance between the deflated input such as $\boldsymbol{X}_i = \boldsymbol{X}_{i-1} - \boldsymbol{p}_{i-1}\boldsymbol{t}_{i-1}^T$ with the initial assumption $\boldsymbol{X}_1 = \boldsymbol{X}$. A new set of latent variables is constructed to represent the linear combination of $\boldsymbol{X}$ while unlike PCA, PLS tries to reduce dimensionality into a few pairs of latent variables. Details of the NIPALS algorithm (Wold, 1975) can be found in Appendix-B.

2.5 Fault Detection and Fault Detection Metrics

In typical MSPC applications, fault detection forms the first step. It helps to judge if an abnormality happened in a process or not. If the monitoring statistics exceed the determined limit, a fault alarm can be raised. Fault detection is then followed by fault identification and diagnosis. The methodology of process performance monitoring is summarised in Figure 2.5 (Ge, Song and Gao, 2013).

Typically, the Hotelling's T^2 statistic and the squared prediction error (SPE) (or Q statistic) are used to assess the variability in process performance monitoring. Residual space and principal component space can be tracked by SPE and Hotelling's T^2,

respectively. In addition to these two well-known metrics, some combined metrics have also been proposed (Raich and Çinar, 1996; Yue and Qin, 2001). The global Mahalanobis distance test can also be used in both metrics. However, since process data are highly cross or autocorrelated, which makes the variances of the residual components close to zero, use of the Mahalanobis distance in residual space is not the first choice (Qin, 2012).

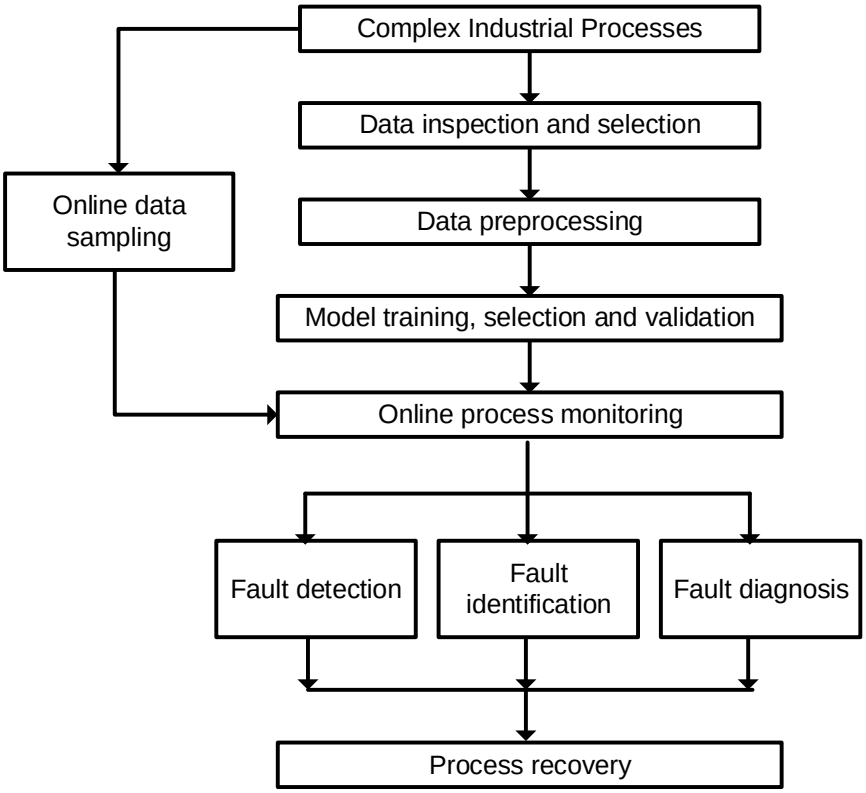

Figure 2.5: Data-based process performance monitoring methodology (adapted from (Ge, Song and Gao, 2013)).

2.5.1 Hotelling's T^2 Statistics

Hotelling's T^2 statistic measures variations in principal component spaces. If the number of data samples, M, is vast, the variance-covariance matrix ($\mathbf{\Sigma}$) of in-control data are accurately known; T^2 index can be defined for the i^{th} sample as follows:

$$(T^2)_i = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \mathbf{\Sigma}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{2.10}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{N\times 1}$ is a column vector of $\boldsymbol{X}$ comprising N variables for the i^{th} sample and $\bar{\boldsymbol{x}} \in \mathbb{R}^{1\times 1}$ is the mean vector of the variables. Owing to accurate estimation of mean and covariance, the T^2 index can be well approximated by a χ^2 distribution with R (number of PCs) degrees of freedom (Qin, 2012):

$$(T^2)_\alpha \leq \chi^2_{R,\alpha} \tag{2.11}$$

where α is the significance level of the distribution; however, most of the time, $\boldsymbol{\Sigma}$ cannot be known accurately. This gives rise to use of the estimated variance-covariance ($\boldsymbol{S}$) or sample covariance matrix from training or historical samples:

$$(T^2)_i = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{2.12}$$

On the other hand, because of collinear or highly correlated process variables owing to the effects of feedback control, inversion of the variance-covariance matrix may result in instability. That does not apply to PC scores since they are orthogonal and uncorrelated (Kourti and MacGregor, 1996). Therefore, T^2 can be found by using PCA scores as follows:

$$\boldsymbol{T}^2 = \boldsymbol{P}^T \boldsymbol{X} \boldsymbol{\Lambda}^{-1} \boldsymbol{X}^T \boldsymbol{P} = \boldsymbol{T}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{T} \tag{2.13}$$

where $\boldsymbol{\Lambda} = \boldsymbol{T}^T \boldsymbol{T}/(M-1)$ is the sample covariance matrix under the condition that the process data are normal and has a multivariate normal distribution. T^2 is related to a F distribution which is the ratio of two independent χ^2 scores:

$$\frac{R(M-1)}{(M-R)} F_{R,(M-R);\alpha} \tag{2.14}$$

where $F_{R,(M-R);\alpha}$ is the F distribution with R and $(M-R)$ degrees of freedom.

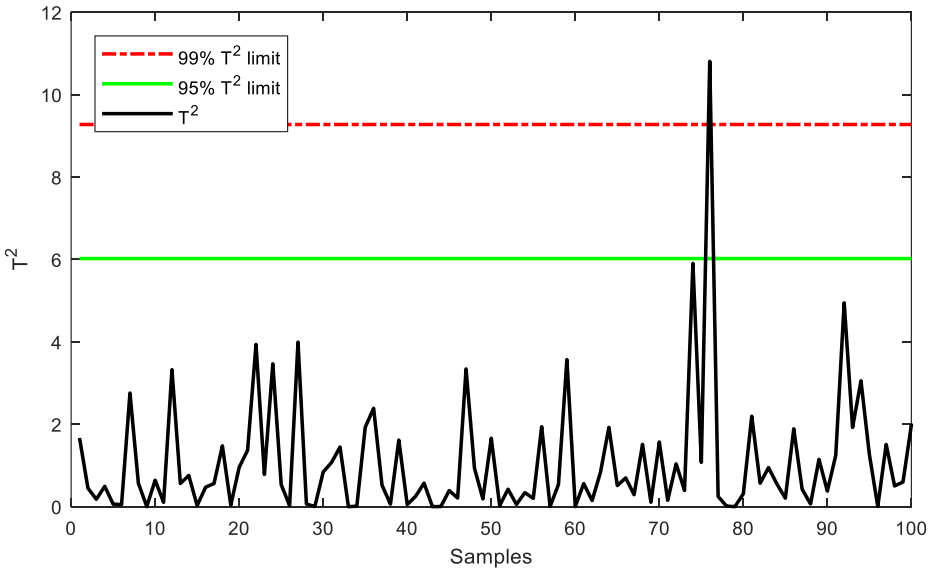

Figure 2.6: Example of a Hotelling's T^2 monitoring chart.

Figure 2.6 shows an example monitoring chart based on Hotelling's T^2 with control limits. Here, one sample lies outside the action limit; therefore, this sample needs to be further interrogated using the contribution plots.

2.5.2 Squared Prediction Error

Hotelling's T^2 is typically used to monitor variation within the principal component space. A second monitoring metric that tracks residual space is the SPE or Q-statistic. It is the squared difference between the observed and the predicted values from the normal representation:

$$SPE_i = (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^T(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i) = (\boldsymbol{x}_i - \boldsymbol{P}\boldsymbol{t}_i^T)^T(\boldsymbol{x}_i - \boldsymbol{P}\boldsymbol{t}_i^T) \tag{2.15}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{N\times 1}$ and $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{N\times 1}$ are the original and estimated variable vectors, respectively, for the i^{th}sample, and $\boldsymbol{t}_i \in \mathbb{R}^{1\times N}$ is the score vector for the i^{th} sample. It can also be defined as the squared perpendicular distance of a multivariate observation from the reduced principal component space in a matrix form as follow(Qin, 2003):

$$\boldsymbol{SPE} \cong \|(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^T)\boldsymbol{X}\| \tag{2.16}$$

A process is considered abnormal if

$$SPE_i \geq Q_\alpha \tag{2.17}$$

where Q_α defines the upper control limit for the SPE with a significance level α (Jackson and Mudholkar, 1979):

$$Q_\alpha = \theta_1\left(\frac{z_{(1-\alpha)}\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2} + 1\right)^{\frac{1}{h_0}} \tag{2.18}$$

where

$$\theta_i = \sum_{k=R+1}^{N} \lambda_k^i, \qquad i = 1,2,3$$
$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{2.19}$$

where λ_k^i is the eigenvalue, i and k refer to indexes of the power and largest eigen order, respectively, and $z_{(1-\alpha)}$ is the standard normal deviate or z-score for the $(1-\alpha)$ percentile. This limit is derived under the condition that a sample vector $\boldsymbol{x}$ from $\boldsymbol{X}$ follows a multivariate normal distribution. An alternative limit for the SPE has been defined by Nomikos and MacGregor (Nomikos and MacGregor, 1995b) by using (Box, 1954) by using χ^2 distribution:

$$Q_\alpha^2 = g\chi_{h,\alpha}^2 \tag{2.20}$$

where

$$g = \frac{\theta_2}{\theta_1}, \qquad h = \theta_1^2/\theta_2 \tag{2.21}$$

2.5.3 The Asymmetric Role of SPE and T^2 in the Performance Monitoring

Both statistics (or metrics), as mentioned earlier, can be used for process performance monitoring; however, it is worth pointing out that they measure different properties of the process and their roles are not designed to be symmetric to each other. In the case of process performance monitoring, SPE is preferable to T^2 which is the opposite to that for quality control. Some instances require SPE and T^2 in a combined form as a single metric (Qin, 2003).

The SPE statistic measures variability, which breaks the typical process correlation indicated by an abnormal situation. The T^2 statistics measures the distance to the origin in PC subspace. In many complex industrial processes, the PC subspace contains normal variation with the significant variance described by the PC representation and the residual subspace containing mainly noise. Due to noise characteristics, the T^2 normal region is defined as larger than the SPE normal region. Therefore, it takes a much larger fault magnitude to exceed the T^2 control limit. On the other hand, the normal region defined by the SPE control limit includes residual components which are mainly noise. Furthermore, small to moderate faults can easily exceed the SPE control limits.

An example taken from Qin et al. (Qin, 2003) can help to illustrate the difference between the two metrics. In Figure 2.7, measurement of the inlet and outlet flow rates of a unit are represented with two abnormalities given by the filled with black and blue

circles. The steady-state data are described by the T^2 region and the PCA model with one PC is depicted by the 45° line in Figure 2.7(b). An abnormal sample (denoted by the filled blue circle) deviates from the normal line and has a large SPE, and so is detected by an SPE chart. However, the T^2 index is still in-control for this sample. On the other hand, an abnormality (denoted by the filled black circle as a throughput sample) can cause an increase in T^2 alone which states that the change is consistent with the model but it may be just a shift of operating region which is not an error. This does not cause a rise in SPE, and therefore, use of the SPE rather than the T^2 metric is preferred for fault detection.

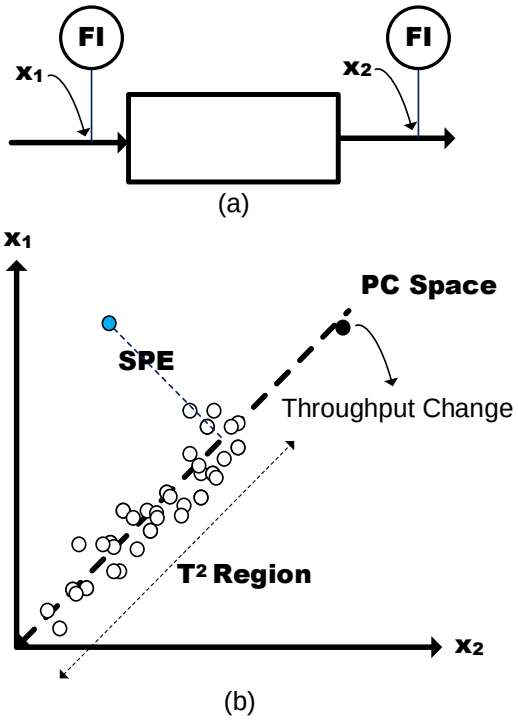

Figure 2.7: (a) Inlet and outlet flow measurement illustration, (b) measurement representation with principal component space with two abnormalities (filled blue and black circles).

Another difference between PC space and residual space is the nonstationarity of the estimated and original variables. Data from complex industrial processes are rarely normally distributed and stationary. As nonstationary variables tend to show significant variability and, the principal component subspace usually captures this large variability in the model, monitoring charts based on the T^2 metric can exhibit a significant number of false alarms for nonstationary data. More comprehensive control limits are also required for nonstationary data, which can cause an increase in the undetected fault rates. The PC subspace usually captures the nonstationary parts of the

signals due to the high variability, therefore, the use of the T^2 metric can incur false alarms due to the nonstationarities.

2.6 Complex Industrial Processes and Data Characteristics

Industrial processes can be divided into two main groups: continuous and batch processes. A continuous process is a flow production method performed around the optimum state most of the time. On the other hand, batch processes have finite operation duration, strictly following process specifications. From an MSPC view, batch processes have an extra-dimension for different batches to keep batch-to-batch variations, therefore their data ($\underline{\boldsymbol{X}} \in \mathbb{R}^{I \times N \times M}$) is three dimensional ($3D$) as illustrated in Figure 2.8.

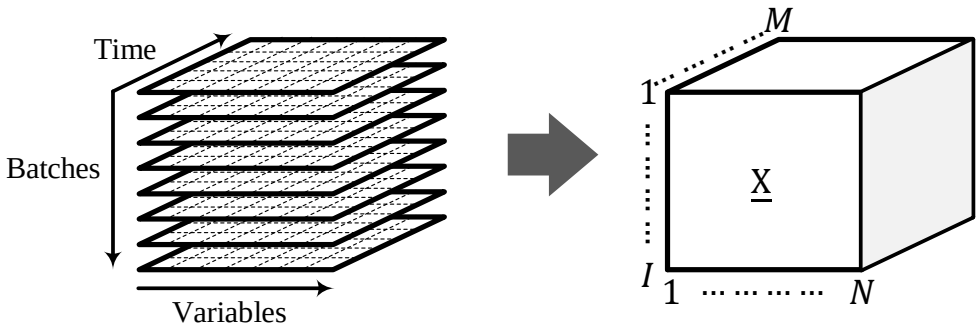

Figure 2.8: 3D data representation of a batch process.

Batch process monitoring using MSPC approaches were pioneered by Nomikos and MacGregor (1994; 1995b). Nomikos et al. (1995a) proposed multiway PCA and then multiway PLS for performance monitoring of batch processes. Multiway methods provide data unfolding of $3D$ data into two dimensional ($2D$) data where projection based techniques can work. An alternative approach for unfolding the data matrix was proposed by Wold et al. (1998).

In Figure 2.9, the various approaches for unfolding the batch data are illustrated. Nomikos and MacGregor (1994) proposed an unfolding method illustrated in Figure 2.9(a). Each column vector contains the measurements for variables from at each time point over all batches, and each row comprises the measurements for all variables from one batch. This allows the comparison of the performance of each batch at a specific time point against a group of '*normal*' batches. However, this unfolding can be

problematic with score calculation when online monitoring is needed where the entire trajectory is not completed before the end of the batch.

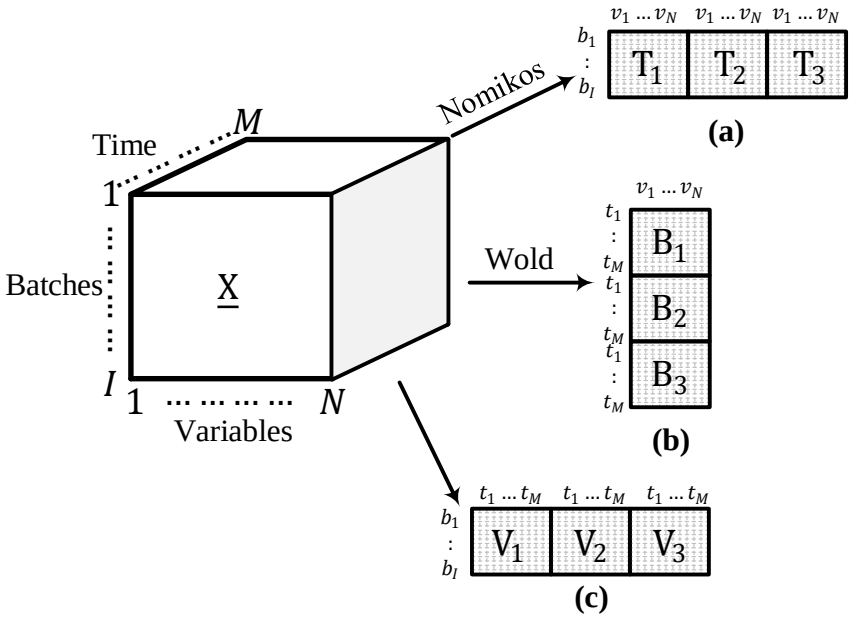

Figure 2.9: Illustration of different unfolding approaches. (a) Time-wise unfolding (Nomikos' approach), (b) batch-wise unfolding (Wold's approach), and (c) variable-wise unfolding

An alternative approach, given in Figure 2.9(b) was proposed by Wold et al. (1998). Here, each row contains measurements of the variables at a particular time point in a batch and each column comprises measurements of each variable across all batches and time points. This allows monitoring of the score trajectories of a variable from all batches and it does not need any filling approach for missing data in online monitoring unlike Nomikos's approach. However, mean centring and scaling of the unfolded data matrix do not remove the mean trajectories since it only captures the covariance among the variables which is not major interest for performance monitoring. The last approach is illustrated in Figure 2.9(c) where each row contains measurements of the same variables through the batches and each row comprises the measures for variables from one batch. Similarly, mean centring and scaling does not remove the mean trajectories since it captures the covariance only for time t. In online monitoring, it causes missing data in each data block before the process end. Therefore, it is unlikely to be chosen by the data analyst. Kourti (2003) has discussed a detailed statistical analysis of multiway monitoring and grade transitions for batch.

The given techniques consider the entire batch as a single object. This can lead to modelling problems when the characteristics of the variables change over time. Since

each phase has its own underlying features, variables can exhibit significantly different behaviours over the phases. Multi-phase behaviour has been studied widely in the last decade after Undey and Cinar (2002) presented a multi-stage multi-phase statistical monitoring method for batch processes. Multi-PCA modelling has been proposed for multi-phase batch monitoring and allows different PCA different models to be used for each phase (Lu, Gao and Wang, 2004). Stubbs et al. (2013) proposed an interval model in combination with a multiway model for a fed-batch penicillin simulator. This method required the data analyst to select which variables are required within each phase. A dynamic monitoring algorithm based on time lag shift has been proposed by Chen et al. (2002). The use of lagged variables extended subspace identification and has also been applied to continuous processes (Yao and Gao, 2008, 2009).

2.6.1 Data Characteristics in the Complex Industrial Processes

2.6.1.1 Non-Gaussian Distributions

Traditional projection-based MSPC techniques make the inherent assumption that a process data are normal distributed, i.e. have a Gaussian distribution. Data from most industrial processes do not follow a normal distribution exactly. The control limits may be inaccurate and thus unable to represent the typical operation region because of inadequate distributions. Figure 2.10 illustrates different types of distributions that can be exhibited by complex processes. Several enhancements have been proposed to conventional MSPC techniques to overcome inaccurate limit problems arising from non-Gaussian distributions.

Independent component analysis (ICA), proposed to look at components from both statistically independent, non-Gaussian and Gaussian variables, was introduced by Li and Wang (Li and Wang, 2002) for monitoring, then combined with PCA by Kano (Kano *et al.*, 2004). PCA can impose the first and second moments of information, namely the mean and variance of the data, which is the basis of the Gaussian distribution. However, non-Gaussian data characteristics may show skewness and kurtosis, which are the third and fourth moments. ICA may reveal more meaningful information from data by handling the higher order moments. However, ICA can cause a selection problem for the independent components where it may result in different independent components in replicate analysis of the same data. Several extensions including hybrid modelling have been proposed by Li and Wang to control these

different characteristics where each component is independent (Li and Wang, 2002). Even though the ICA model can extract high-order moments and provide independent latent variables, randomised components from the random initialisation may result in unstable monitoring performance. There are also difficulties in selecting the number of components, which directly affects the control limits for ICA-based monitoring.

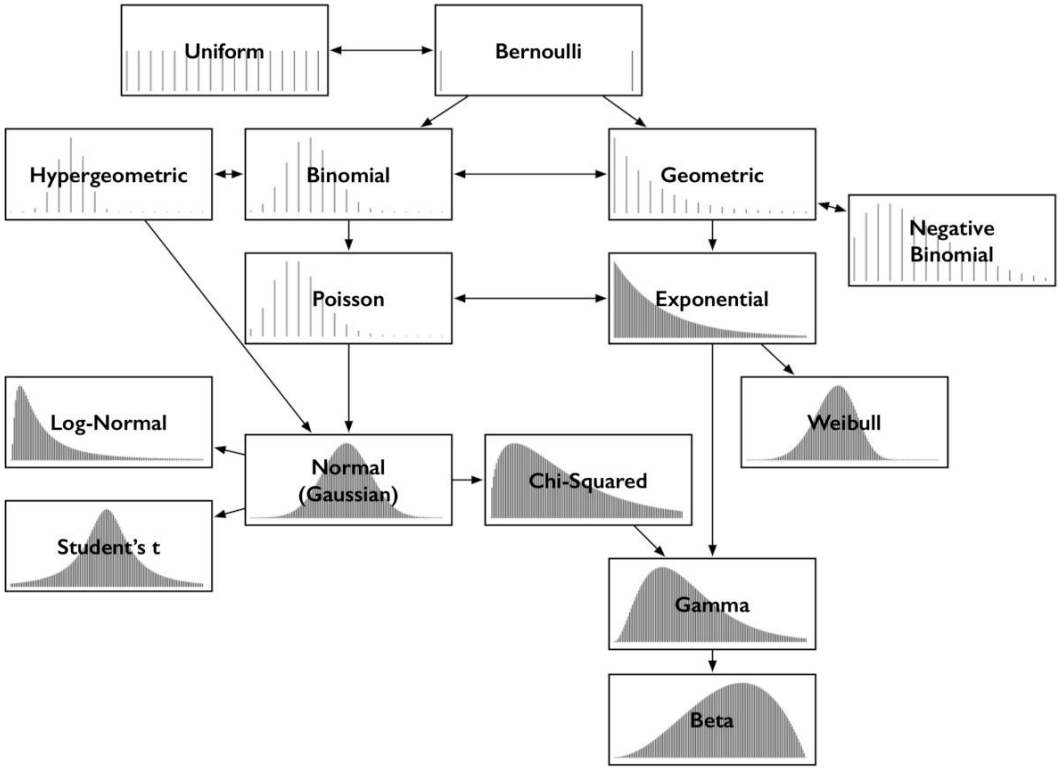

Figure 2.10: Illustration of discrete and continuous distributions.

A Gaussian mixture model was proposed by Chen and Lui (1999), which can be described by several local linear models through employing the expectation-maximization algorithm. Later, this method was improved by a maximum-likelihood PCA modelling framework (Sang *et al.*, 2005). However, difficulties with model training and the definition of the number of local models have limited its use. The given methods for analysis of data with non-Gaussian characteristics have some advantages for particular data sets, though they have limited general applicability. However, they cannot solve issues with correlation and dependencies in the data, which is critical for process monitoring.

2.6.1.2 Nonlinearity

A large number of industrial processes have linearly correlated variables as a result of stable production and operate within a small region of steady conditions; however,

modern industrial processes can run under various operating conditions. Furthermore, relationships among the variables can be more complicated, and so linear modelling may not function well even if linearisation around steady-state conditions are applied. The models built on this type of data become more specific and complicated, working only on data that are similar to that used to build the model. Therefore, having a nonlinear model for one specific process may only enable singular usage for that process.

Several local linear models, built on an approximation of the nonlinear process can be a first choice with the application of PCA (Kerschen and Golinval, 2002). However, difficulties can arise with the determination of the number of models. Adaptive and recursive models that can also handle the slow-varying processes have been studied widely. A recursive exponentially weighted PLS model was proposed (Dayal and MacGregor, 1997) that was followed by the development of an adaptive monitoring scheme, which incorporated recursive PCA to update the mean of the training data, the number of the PCs and the control limits (Li *et al.*, 2000). Moving window approaches that use a N step ahead horizon strategy have been applied, which overlaps to some extent with the work of Li et al. (Wang, Kruger and Irwin, 2005).

A data-based approach based on the application of neural networks for nonlinear PCA was proposed and used the principal curve method; this shapes the line of PCs with regards to the samples (Dong and Mcavoy, 1996). A five-layer neural network has been designed and the principal curve used to calculate the scores of the nonlinear PCA. An extension for a correlated data set, which combined neural network-based nonlinear PCA and time lag shifts, was proposed to monitor complex processes such as the Tennessee Eastman process (TEP) (Chen and Liao, 2002). To enhance the process monitoring performance, a hierarchical neural network based on a fuzzy clustering method has been designed and applied to the TEP (Eslamloueyan, 2011). Even though the nonlinear PCs can be directly obtained, the training of the nonlinear PCA model requires prior knowledge such as the number of PCs. This makes training of the models difficult particularly where building of neural network models is already more time consuming than conventional PCA models.

A kernel is a way of computing the dot product of two vectors in some (possibly high dimension) feature space. It provides a mapping between the given features to another feature space. The similarity between the foundation of projection techniques such as

PCA and PLS, and kernels gives rise to potential hybrid models. Kernel PCA has been proposed to avoid nonlinear mapping involved in computational intelligence models such as neural networks (Odiowei and Cao, 2010). As for conventional PCA, two monitoring metrics have been constructed to enable separate monitoring of the system and the noisy part of the process. The training of the kernel PCA model is easier than a nonlinear PCA model since the nonlinear optimisation problem is eliminated. However, a kernel selection with parameter tuning requires the data analyst's attention.

Linear approximation approaches by several local linear models is another type of nonlinear method. A linear subspace model has been proposed, which divides the nonlinear process into several linear subspaces. The modelling is based on PCA decomposition. Multi-phase models based on multiway PCA and conventional PCA for the monitoring of batch processes can also deal with nonlinearities in the data as they are based on local modelling (Chen and Liu, 2002; Lu, Gao and Wang, 2004; Yao and Gao, 2008, 2009; Stubbs, Zhang and Morris, 2013). Compared to nonlinear modelling via neural networks and kernel-based models, the determination of the linear approximation method seems much more straightforward because model interpretation is easier for linear approximation methods.

On the other hand, it is difficult to determine the number of local models as well as the time intervals over which individual local models apply for a process. This requires more attention from the data analyst to work on which is time-consuming. Therefore, parameter tuning methods are necessary for the determination of the range and number of local models.

2.6.1.3 Correlated and Dependent Variables

When the value of the observations such as $\boldsymbol{x}_{t+k}$ depends on the value of observations $\boldsymbol{x}_t$, it is said to be dependent where $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+k}$ represent a sample and time shifted sample vector, respectively, of the same variables and k is a lag. The term autocorrelation can express this dependency. Cross-correlation describes the dependence between different variables. One of the diagnostic methods to investigate autocorrelation in a data series is to extract graphical information from the scatter plot of all the data pairs $(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})$. For each delay, the autocorrelation coefficient ρ can be calculated as below:

$$\rho_k = \frac{cov(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})}{var(\boldsymbol{x}_t)} \tag{2.22}$$

A collection of ρ_k represents the autocorrelation function (ACF). The example given in Figure 2.11 shows the ACF function of a moving average (MA) model of a time series. The stationary characteristics of the data and its autocorrelation properties are often closely related. If a time series displays a sharp increase in the mean with time, autocorrelation analysis will indicate a high positive autocorrelation structure. Nonstationary data series are called auto-dependent rather than autocorrelated (Box and Narasimhan, 2010).

Complex industrial processes show autocorrelations because of feedback control systems, random noise and process disturbances. However, most of the conventional MSPC techniques are based on the assumption that the process samples are independent of each other, i.e., stationary. If the dynamic information of the process is not incorporated into the model, the fault detection model may be misleading. Several enhancements have been proposed to improve the monitoring performances of dynamic processes that show autocorrelation and time-dependent data characteristics.

A dynamic PCA (DPCA) model has been used for disturbance rejection and isolation for complex industrial processes (Ku, Storer and Georgakis, 1995). A DPCA model is extracted from data sets that have several lagged time data samples of each variable. Having lagged variation of the process inside the DPCA model extracts the autocorrelation of the process variables. Subsequent studies have utilised hybrid models comprising DPCA with ICA and kernel PCA to tackle nonlinearities (Jia *et al.*, 2010; Stefatos and Hamza, 2010). Rato and Reis (2013b) have discussed an extension for the determination of the number of lags and Vanhatalo et al. (2017) have proposed another one based on autocorrelation relationships. DPCA is designed to account for the dynamic structures. However, the resulting scores from DPCA may still be autocorrelated, and possibly cross-correlated (De Ketelaere, Hubert and Schmitt, 2015). An autoregressive moving average (ARMA) filter was applied to remove the autocorrelations and improve the model performance (Rato and Reis, 2013a). DPCA approaches are easy to implement as many existing models can be used. However, they may not efficiently model the correlated and dependent behaviours of the data (Kruger, Zhou and Irwin, 2004).

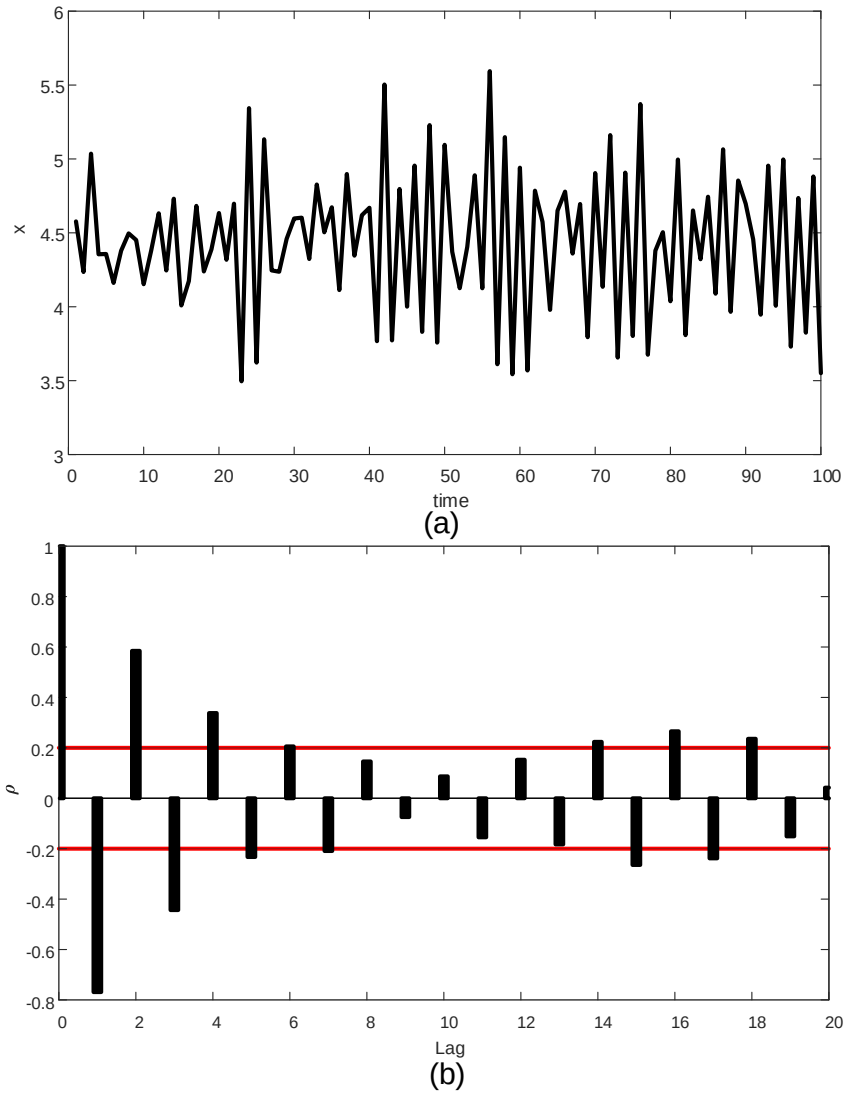

Figure 2.11: Example of a negatively correlated process $\boldsymbol{x}_t = 8 - 0.8\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$, with (a) a time series plot, (b) an autocorrelation function for each lag.

State-space model-based methods such as canonical variate analysis (CVA) take serial correlations into account in the dimension reduction step, which also aims to maximise of the correlation statistics (Negiz and Çinar, 1997; Russell, Chiang and Braatz, 2000a). Subspace identification method based on a state-space model has been compared with DPCA and autoregressive model-based approaches (Xie and Kruger, 2006). The subspace identification model has also been improved by using Kalman filters to capture further autocorrelation in the data. The state-space model-based methods can model both autocorrelation and cross-correlation of the data but it is difficult to determine the number of states.

Adaptive/recursive methods and multi-phase models can also be applied to correlated data sets for monitoring purposes, as already discussed in Section 2.6.1.2. However, model updating of adaptive models is carried out blind, which may include data

exhibiting a fault. Furthermore, multiphase models require knowledge of the number of phases and the time intervals of the individual phases, which requires further advances in the modelling methods.

2.6.1.4 Nonstationarity

A time series is stationary if it displays the same statistical behaviour in time and can be characterised with a constant probability distribution whose joint probability distributions do not change with time. The stochastic properties of stationary processes are unaffected by time. Figure 2.12 represents examples of stationary and nonstationary processes.

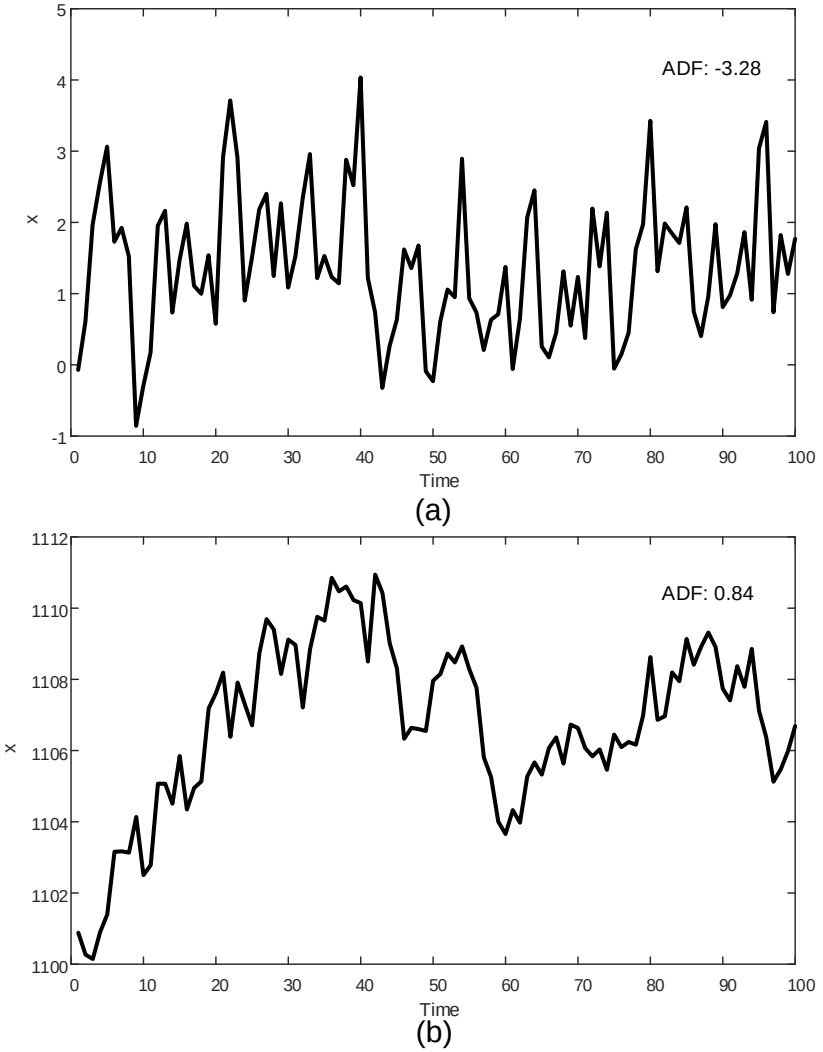

Figure 2.12: Example of (a) a stationary process $\boldsymbol{x}_t = 1 + 0.3\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$, and (b) a nonstationary process $\boldsymbol{x}_t = 11 + \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$ (adapted from (Montgomery, Jennings and Kulahci, 2015)).

Generally speaking, when a complex industrial process operates under normal operating conditions, the variables follow a multivariate, stationary, stochastic process

approximately. However, through regulatory and feedback control, the recorded variables exhibit serial correlations, and time-varying and nonstationary characteristics. Typical causes of such nonstationarities may originate from changes in the operating conditions, grade changes, and variations in the feed rates. Furthermore, some variables of batch and fed-batch processes in the pharmaceutical industry typically exhibit nonstationary characteristics along with stationarities. For these processes, the variables are not strictly stationary and the statistical characteristics usually show slow time-varying behaviors.

Even though time-varying and autocorrelated variables have been studied extensively by the MSPC techniques, nonstationary variables have received little attention until the last decade. One of the first reports of nonstationarity in process monitoring was in a publication on recursive PLS (RPLS) algorithm (Wang, Kruger and Lennox, 2003). Here, the mean and variance of the processes were updated by RPLS. A well-known ARIMA model can also model nonstationary variables; however, the training of multivariate models is difficult because of the number of parameters involved. It can also lead to loss of dynamic information from variable differencing.

A promising tool to cope with the modelling of nonstationary variables is cointegration analysis, which has been adopted from econometrics. Cointegration analysis was used for process monitoring by Chen et al. (2009) following the first mention of its use for the monitoring of nonstationary processes by Xu et al. (2007); the study involved condition monitoring of a fluid catalytic cracking unit. This was followed by the publication of a cointegration-based method for process monitoring. Li et al. (2014) proposed metrics for monitoring techniques based on cointegration residuals using the TEP as an example; however, the method only used the nonstationary variables from all of the variables available to monitor the process. Fault diagnosis with cointegration analysis has more recently been employed on the TEP (Sun, Zhang, Zhao and Gao, 2017). Use of a combination of common-trend and cointegration analysis with different performance metrics was proposed by Lin et al. (2017), which was subsequently enhanced with the Chigira procedure (Lin et al., 2019). However, this still required more than one control chart, based on statistical metrics, to be followed. The first application of cointegration analysis and conventional PCA to the study of batch processes used data from a penicillin simulation (Zhang, Zhao and Gao, 2019). Dominant trend-based logistic regression has also been compared to cointegration

analysis-based approaches for the study of nonstationary processes (Shang et al., 2017).

Although the methods mentioned above have proved to be effective, the application of the techniques apart from that in Zhang's study has so far been limited to nonstationary variables (Zhang, Zhao and Gao, 2019). The models were built using only nonstationary variables. Therefore, any faults on stationary variables were not considered, which is against the fundamentals of MSPC. A common-trend model helps to monitor high-level nonstationary variables, which cannot be modelled by cointegration analysis. However, the use of a common-trend model as described by Lin required the determination and monitoring of several control charts, each comprised of T^2 and SPE metrics. This creates contradictions when the control charts do not detect the same fault dynamics. Furthermore, the use of only cointegration analysis of nonstationary variables may not be sufficient to model high-level nonstationary variables such as batch processes. Therefore, a study is needed to extend the use of cointegration analysis within a monitoring scheme that can be applied to both stationary and nonstationary (including high-level) variables.

2.7 Conclusions

In this chapter, an overview of process performance monitoring has been given and a wide range of projection-based MSPC techniques for process monitoring has been presented. MSPC is an extensive monitoring tool that considers all variables together, unlike univariate SPC. Projection-based techniques such as PCA and PLS form the basis of MSPC owing to their dimension reduction and variance representation capabilities. Process monitoring with PCA-based approaches is constructed around two metrics namely, T^2 and the SPE. However, the use of the SPE is preferred due to its use for residual space monitoring and its monitoring performance with nonstationary variables.

Generally, complex industrial processes have various data characteristics, such as non-Gaussian distributions, nonlinearities, autocorrelations and nonstationarities. To date, several studies have been carried out to address those specific problems in both continuous and batch processes. However, good process monitoring must deal not only with one data characteristic but all data behaviours simultaneously. Furthermore, the conveniences provided by MSPC such as the coverage of all variables throughout the

process and the requirement for only one control chart should not be given up. It is worth noting that one control chart can be based on several metrics such as T^2 and SPE from the same source of signal. This being the case, a method that combines stationary and nonstationary variable modelling and retains the best features of MSPC is proposed as a means of monitoring complex industrial process, and will be described in the following chapters.

3. MODELLING OF NONSTATIONARY VARIABLES

3.1 Overview

Nonstationarity occurs in complex industrial processes due to, for example, seasonal changes, the presence of disturbances, operator inventions, etc. Fault detection by monitoring of nonstationary process data is becoming an increasing challenge. It is not possible to consider a complex process as stationary over all variables. However, dealing with nonstationary variables has only recently started to receive some attention.

Wang et al. (2003) discussed the monitoring of processes that exhibited nonstationary and/or time varying behaviour. They showed that the application of RPLS algorithms together with a recursive structure for the mean and variances of the processes, together with confidence limits, had problems with detecting incipient faults (Wang, Kruger and Lennox, 2003). It is well-known that an ARIMA model can represent nonstationary variables (Box, Luceño and Paniagua-Quiñones, 2011); however, that comes with a vast computational burden for multivariate processes. It is also known that variable differencing, another approach to cope with nonstationarity, can lead to the loss of dynamic information.

Cointegration analysis is a promising tool to model nonstationary variables by establishing long-run equilibria between nonstationary variables. The use of cointegration in process monitoring was proposed by Chen et al. (2009). Cointegration is arguably the most effective ways of handling the nonstationary characteristics of data, and was proposed to formulate the existence of linear equilibria (Engle and Granger, 1987). Applied economists used cointegration analysis to cope with the difficulties that arise when the data contains a unit roots, which indicates nonstationarity. It has been extensively used in the area of econometrics, and more recently, in several disciplines of science and engineering to ensure that it reflects any long-run information which can be easily removed via de-trending and differencing. It is also worthy of note that, Robert F. Engle and Clive Granger's contribution of cointegration to econometric modelling and analysis was awarded a Nobel Prize for Economics in October 2003.

Cointegration has been studied intensely in areas where there are complex systems such as econometrics (Engle and Granger, 1987; Stock and Watson, 1988; Harris and Sollis, 2003; Kirchgässner and Wolters, 2007), structural health monitoring (Cross, Worden and Chen, 2011; Antoniadou, Cross and Worden, 2013), computational systems for tool wear monitoring (Wang *et al.*, 2014) and process system engineering for process monitoring (Chen, Kruger and Leung, 2009; Sun, Zhang, Zhao and Gao, 2017; Zhang, Zhao and Gao, 2019). Chen et al. (2009) demonstrated the use of cointegration analysis for the monitoring of an industrial distillation unit, which exhibited nonstationary behaviours. Recently, cointegration has also been used in combination with conventional PCA for monitoring of multi-phase batch processes (Zhang, Zhao and Gao, 2019).

Even though cointegration is a powerful tool for the building of long-run relationships between time series, cointegration analysis can result in a low rank cointegration matrix, which does not allow the use of all latent stationary and nonstationary factors, even when the number of nonstationary variables is high. This issue can be solved through the use of a common-trend model. The common-trend model representation was proposed by Stock and Watson (1988) with the connection between cointegration and a common-trend model being derived by linear stochastic trends (Johansen, 1988; Stock and Watson, 1988). There exists a duality between common-trend and cointegration analysis. Cointegration analysis restricts the number of independent trends, namely the rank of the cointegration matrix, and the observed variables from all the independent trends. The rank of cointegration matrix also determines the number of the cointegration residuals vectors which is the basis of the cointegration residuals-based process monitoring methods. In order to extend the effectiveness of cointegration analysis and prevent loss of the ability to detect faulty process dynamics, a common-trend model can be used to form the independent trends. This was demonstrated for a continuous industrial melter process through the use of a common-trend model implemented with a forecast recovery filter to obtain stationary factors (Lin, Kruger and Chen, 2017). This work was further extended with another study addressing the use of the Chigira procedure for the monitoring of nonstationary and dynamic trends for practical process fault diagnosis (Lin et al., 2019).

In this chapter, the definition of nonstationarity is given and the different unit root tests that are used to determine if a variable is nonstationary or not are described. The

concept of cointegration and modelling techniques such as the Engle-Granger approach for a single equation method, and the Johansen test for multivariate systems are introduced with examples.

3.2 Stationary and Nonstationary Time Series

In the broadest sense, stationarity is a statistical property of data or a defined process that does not change over time. It is a common pattern in different science fields for approximation of complex phenomena. Therefore, it has become a common assumption for many practices and tools in time series analysis.

Two different kinds of stationarity can be distinguished. If the common distribution function of a stochastic process does not change in time, the process can be categorised as strictly stationary. The strictly stationary concept is difficult to apply in practice; therefore, only weak stationarity and stationarity in the second moments are considered. Weak stationarity only requires the shift-invariance in time of the first (mean) and second (auto-covariance), moments for all time points. It implies that the process has the same mean at all-time points and the covariance between the variables at any two time points $(\boldsymbol{x}_t, \boldsymbol{x}_{t+k})$ depend only on k, not time.

The types of stationarity of a stochastics process $(\boldsymbol{x}_t)$ can be defined for the corresponding moments starting from the first moment, the mean, as given below (Kirchgässner and Wolters, 2007):

- *Mean stationary*: If the expected value is constant for all t $(\mathrm{E}[\boldsymbol{x}_t] = \mu)$ then the process is mean stationary.

- *Variance stationary*: If the expected value is constant and finite for all t $(\mathrm{var}[\boldsymbol{x}_t] = \mathrm{cov}[\boldsymbol{x}_t, \boldsymbol{x}_t] = \mathrm{E}[(\boldsymbol{x}_t - \mu)^2] = \sigma^2 \text{ and } \sigma^2 < \infty)$ then the process is variance stationary.

- *Covariance stationary*: If the expected value is only a function, not dependent in time t on the time difference between the corresponding two samples $\left(\mathrm{cov}[\boldsymbol{x}_t, \boldsymbol{x}_k] = \mathrm{E}[(\boldsymbol{x}_t - \mu_t)(\boldsymbol{x}_k - \mu_k)] = f(|k - t|)\right)$, then the process is covariance stationary.

- *Weak stationary*: If a stochastic process is mean and covariance stationary, then it is weak stationary. In other words, the given stationary definitions describe the weak stationarity.

Figure 3.1 illustrates types of weak stationarities.

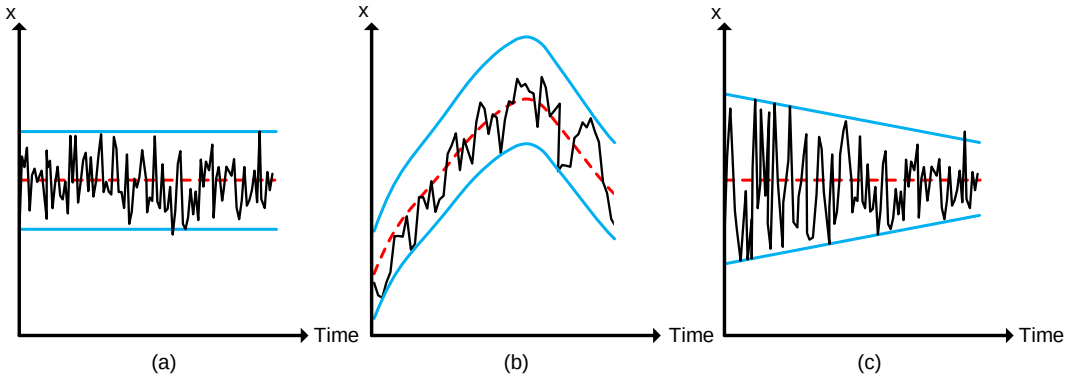

Figure 3.1: Constancy in the mean and variance to illustrate weak stationarities. (a) stationary mean and stationary variance, (b) nonstationary mean and stationary variance, and (c) stationary mean and nonstationary variance.

In time series analysis, the question of whether the model should be estimated by a single equation approach such as ordinary least squares (OLS) or as a system estimator where it is necessary to consider the underlying properties of the time series. A model containing nonstationary variables may lead to a problem of spurious regression (Harris and Sollis, 2003). Furthermore, nonstationary variables may lead to an increase in false alarm rates because of slow time-varying characteristics in process monitoring. Complex industrial processes rarely behave in a stationary manner (Ketelaere et al., 2011). In econometrics, nonstationarity plays a vital role in affecting the performance of a financial time series model because of the time-varying characteristics. Consequently, helpful tools to handle nonstationary data have been heavily influenced by econometricians.

Suppose a variable x_t is generated by a first-order autoregressive (AR) process:

$$\boldsymbol{x}_t = \theta \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{3.1}$$

where, $\boldsymbol{x}_t$ depends on the value at the previous time, $\boldsymbol{x}_{t-1}$, and a disturbance $(\boldsymbol{\epsilon}_t)$ is a normal distribution with zero mean and σ^2 variance. The variable $\boldsymbol{x}_t$ is stationary if $\theta < 1$ and nonstationary if $\theta = 1$. Moreover, if $\theta > 1$, then it is nonstationary and

tends to $\pm\infty$. *A unit root process* defines a process whose first difference is stationary. Equation (3.1) can be described as a unit root process when $\theta = 1$. Consequently, searching for the existence of the condition $\theta = 1$ is called the *unit root test*. The difference between some of these processes is illustrated in Figure 3.2.

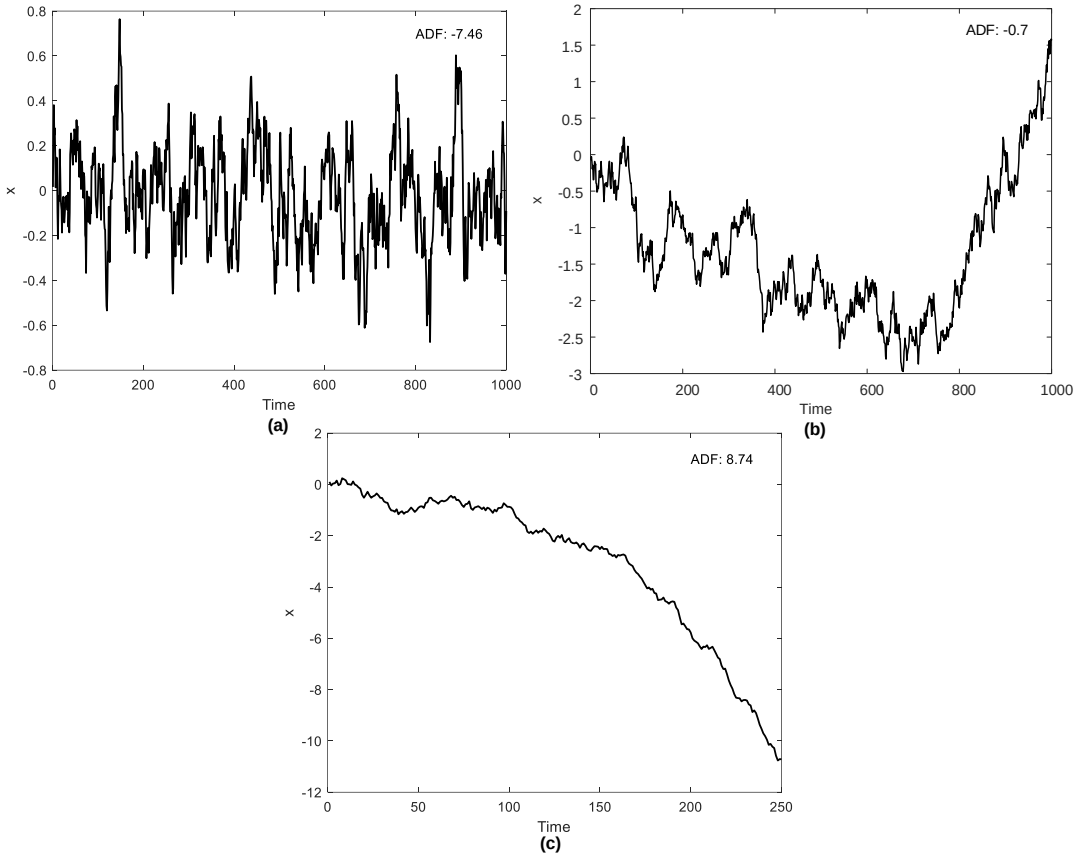

Figure 3.2 : Illustration of AR(1) processes (a) $\theta = 0.9$, (b) $\theta = 1$, (c) $\theta = 1.01$ and $\epsilon \sim N(0, 0.1^2)$.

A stationary series tends to return to its mean value and varies around the mean, while a nonstationary series might have different mean values for different samples through time. Assuming $\theta = 1$ and rearranging Equation (3.1) with a starting initial value $\boldsymbol{x}_{t-k}$:

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-k} + \sum_{j=0}^{k-1} \boldsymbol{\epsilon}_{t-j} \tag{3.2}$$

$\boldsymbol{x}_t$ is the current value and depends on the initial value and all disturbances produced by $\boldsymbol{\epsilon}$ lie between $t - k + 1$ and t. The variance of $\boldsymbol{x}_t$ is time dependent and $t\sigma^2$

increases to infinity. Thus, it does not converge to a mean value and the expected time to return the original mean value is infinite.

3.3 Testing for Unit Roots

If a variable contains a unit root, then it is nonstationary and unless it combines with other nonstationary variables via methods like cointegration to form a stationary cointegration relationship, using it in methods or models that consist of stationary variables can be detrimental to the fault detection capability or could falsely imply the existence of a meaningful relationship.

There are several ways of testing for the existence of a unit root. The Dickey-Fuller (DF) test and the extended, the version augmented DF (ADF) test will be the main focus here for testing the null hypothesis that the time series contains a unit root. Even though there are several tests that have been proposed, only four will be discussed in this section.

3.3.1 The Dickey-Fuller Test

The Dickey-Fuller (DF) test was the first unit root test proposed to test the null hypothesis that there is the presence of a unit root in an AR model of a given time series and thus the process is not stationary (Dickey and Fuller, 1979). The simplest form of the DF test can be defined by using AR(1):

$$\boldsymbol{x}_t = \theta \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{3.3}$$

or

$$(1 - L)\boldsymbol{x}_t = \Delta \boldsymbol{x}_t = (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{3.4}$$

where Δ is a difference operator such that $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$, and L is a lag operator where $L\boldsymbol{x}_t = \boldsymbol{x}_{t-1}$. Either variant of the test is applicable where the null hypothesis is

$$H_0: \quad \theta \geq 1 \tag{3.5}$$

versus the alternative hypothesis:

$$H_1: \quad \theta < 1 \tag{3.6}$$

A standard approach to test a hypothesis is the t-test. However, because of the nonstationarity of the data, the computed statistics do not follow a standard t-distribution. Thus, a DF distribution can be calculated using the samples which are generated by Monte Carlo techniques. This procedure is achieved by fixing $\theta = 1$ and then adding randomly generated ϵ_t to the normal distribution. The rejection percentages of the model are also calculated based on this approach. These are the critical values to reject a null hypothesis of a unit root at various significance levels such as 5% and 1% based on the DF distribution of

$$(\hat{\theta} - 1)/\mathrm{SE}(\hat{\theta}) \tag{3.7}$$

where $\hat{\theta}$ is an estimator of θ and $\mathrm{SE}(\hat{\theta})$ is the standard deviation of the estimated parameter. Some critical values are tabulated in Table 3.1 for different models (Harris and Sollis, 2003). The DF test has two other versions which are different in terms of the model of the unit root test. These are based on Equation (3.4) with additional drift (a_0):

$$\Delta \boldsymbol{x}_t = a_0 + (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{3.8}$$

and with drift and a deterministic time trend:

$$\Delta \boldsymbol{x}_t = a_0 + a_1 t + (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{3.9}$$

The size and type of the test can significantly affect the results. Thus, a-priori knowledge or structured strategies can be used for the type of test to allow the best fitting test. Philips and Perron (Phillips and Perron, 1988) suggested a sequential testing procedure for the model types given in Table 3.1, and starts with the use of Equation (3.9) and continues with eliminating unnecessary nuisance parameters. The testing stops when the test can reject the null hypothesis of a unit root.

Table 3.1: Critical values of Dickey-Fuller tests for $m = 100$ (Harris and Sollis, 2003, p. 47).

Model	Hypothesis	Critical Values for 95%	99%
$\Delta \boldsymbol{x}_t = a_0 + a_1 t + (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$	$\theta = 1$	-3.51	-4.04
	$\theta = 1, a_1 = 0$	6.49	8.73
	$\theta = 1,$ $a_0 = a_1 = 0$	4.88	6.5
$\Delta \boldsymbol{x}_t = a_0 + (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$	$\theta = 1$	-2.89	-3.51
	$\theta = 1, a_0 = 0$	4.71	6.70
$\Delta \boldsymbol{x}_t = (\theta - 1)\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$	$\theta = 1$	-1.95	-2.60

3.3.2 The Augmented Dickey-Fuller Test

In the DF test, the parameter estimation might be biased because the test cannot guarantee that $\boldsymbol{\epsilon}_t$ is white noise. Dickey and Fuller extended the DF test to the so-called ADF test (Dickey and Fuller, 1981). The ADF test also covers AR(p) processes whereas the DF test is only for AR(1) processes. Assume that $\boldsymbol{x}_t$ follows a p^{th} order AR process:

$$\boldsymbol{x}_t = \phi_1 \boldsymbol{x}_{t-1} + \phi_2 \boldsymbol{x}_{t-2} + \cdots + \phi_p \boldsymbol{x}_{t-p} + \boldsymbol{\epsilon}_t \tag{3.10}$$

or

$$\Delta \boldsymbol{x}_t = \theta^* \boldsymbol{x}_{t-1} + \theta_1 \Delta \boldsymbol{x}_{t-1} + \theta_2 \Delta \boldsymbol{x}_{t-2} + \cdots + \theta_{p-1} \Delta \boldsymbol{x}_{t-p+1} + \boldsymbol{\epsilon}_t \tag{3.11}$$

where $\theta^* = (\phi_1 + \cdots + \phi_p) - 1$, $\theta_i = (\phi_1 + \cdots + \phi_i) - 1$ for $i = 1, \dots, p-1$ and $\boldsymbol{\epsilon}_t \sim N(0, \sigma^2)$. The null hypothesis which indicates that x_t contains a unit root is:

$$H_0: \quad \theta^* \geq 0 \tag{3.12}$$

in contrast with the alternative test:

$$H_1: \ \theta^* < 0 \tag{3.13}$$

The ADF *t*-statistic calculation is the same as given below:

$$\theta^*/\mathrm{SE}(\theta^*) \tag{3.14}$$

The critical values shown in Table 3.1 are still valid for the ADF test. However, it is only strictly accurate for large samples. Banerjee (Banerjee *et al.*, 1993, p. 106) showed that the significance levels for small samples are not the same as those that are under the strong assumptions of the simple DF model. Based on Equation (3.11) with additional drift (a_0), the deterministic time trend is:

$$\Delta \boldsymbol{x}_t = \theta^* \boldsymbol{x}_{t-1} + \sum_{i=1}^{p-1} \theta_i \Delta \mathbf{x}_{t-i} + a_0 + a_1 t + \boldsymbol{\epsilon}_t \tag{3.15}$$

Here, the choice of p is important and should be sufficient to capture the correlation structure in the data. The selection of the lag variable p has been discussed in several studies (Harris, 1992; Banerjee *et al.*, 1993; Greene, 2017) as it affects rejection of the null hypothesis when it is true. According to the study by Said and Dickey (1984), the lag variable p should be $m^{1/3}$ where m is the sample size.

3.3.3 The Philips-Perron Test

An alternative approach to the ADF test was suggested by Phillips and Perron (Phillips and Perron, 1988) termed the PP test. It differs from the ADF test by adding robustness to deal with correlations. The Phillip's Z-test $\left(Z(\tau_\mu)\right)$ is the t-statistic associated with testing the null hypothesis $\theta = 1$ in Equation (3.3) and is calculated as:

$$Z(\tau_\mu) = (S_\epsilon/S_{Ml})\tau_\mu \ - \frac{1}{2}(S_{Ml}^2 - S_\epsilon^2)\left\{ S_{Ml}\left[M^2 \sum_{t=2}^{M} (\boldsymbol{x}_{t-1} - \boldsymbol{x}_{-1})^2 \right]^{1/2} \right\}^{-1} \tag{3.16}$$

where x_{-1} is an additional term to ensure non-negativity for the estimation of S_{Ml} (non-negativity is not guaranteed for finite numbers of samples), and

$$S_{\epsilon}^{2} = M\sum_{t=1}^{M}(\boldsymbol{\epsilon}_t)^2 \ ,$$
$$S_{Ml}^{2} = M^{-1}\sum_{t=1}^{M}(\boldsymbol{\epsilon}_t)^2 + 2M^{-1}\sum_{t=1}^{l}\sum_{t=j+1}^{M}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t-j} \tag{3.17}$$

where l is the lag truncation parameter to ensure that the autocorrelation of the residuals is fully captured.

The critical values for this statistic are the same as for τ_{μ} in Table 3.1 and $Z(\tau_{\mu})$ reduces the ADF test statistic (τ_{μ}) when autocorrelation is not present. This also makes $S_{\epsilon}/S_{Tl} = 1$.

In contrast to the ADF and PP tests, the Kwiatkowski, Philips, Schmidt and Shin (KPSS) test (Kwiatkowski *et al.*, 1992) assumes that the null hypothesis is stationary around a mean or a linear trend while the alternative hypothesis is the presence of the unit root. This is a significant difference since it is possible for a time series to be nonstationary, and still have no unit root yet be trend-stationary.

3.4 Cointegration

For years, econometricians have not considered the effects of nonstationarity as they did not consider that economic time series data might be *integrated*. Granger (Granger and Newbold, 1974) showed that the use of traditional statistical procedures proposed for the investigation of stationary stochastic time series might be an issue in such situations.

If a time series is differenced d times to become a stationary time series, then it contains d unit roots, and it is integrated of order d. It is denoted as $I(d)$. Assume that there are two time series $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and both are $I(d)$. Generally, any linear combination of $I(d)$ time series will also be $I(d)$; however, if there is $\boldsymbol{\beta}$ that defines regression between $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ where a disturbance term $\boldsymbol{\epsilon}_t$ is defined as:

$$\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - \boldsymbol{\beta}\boldsymbol{x}_t \tag{3.18}$$

then it has a lower order of integration such as $I(d-b)$ where $b > 0$. This phenomenon is termed *cointegration* of order (d, b) between $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ by Engle and

Granger (1987). Consequently, if $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ were $I(1)$ and $\boldsymbol{\epsilon}_t \sim I(0)$, then this is called *cointegration* of order $CI(1,1)$. The introduction of cointegration has had a massive impact on econometrics and financial time series analysis and Engle and Granger were awarded the Nobel Prize for Economics in 2003 (NobelPrize.org, n.d.).

The long-run properties of a time series are said to be cointegrated as they are linked to form an equilibrium relationship spanning the long-run. Thus, if $\boldsymbol{x}_t \sim I(1)$ and $\boldsymbol{y}_t \sim I(0)$, then these cannot be cointegrated as $I(0)$ suggests a constant mean while $I(1)$ tends to drift over time. As a result, an error term defined between $\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - \boldsymbol{\beta x}_t$ would not be a constant over any period of time. On the other hand, having a mixture of different order time series is possible in the case when three or more time series exist in the model. In this case, the cointegration order between them must be the lower order series for both bilateral and multivariate models (Engle and Granger, 1987).

3.4.1 Cointegration in Single Equations

The Engle-Granger (EG) approach or EG two-step method refers to a calculation of the bivariate cointegration model following (Engle and Granger, 1987):

1. Assuming both time series have unit roots, find the linear approximated relationship between $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ via ordinary least squares (OLS). Find the errors $(\boldsymbol{\epsilon}_t)$.
2. Test $\boldsymbol{\epsilon}_t$ for unit root existence via unit root test methods.

In discussing cointegration, it has been shown that if two time series are $I(d)$ then the estimation of the long-run relationship between them can be represented with a static model:

$$\boldsymbol{y}_t = \boldsymbol{\beta x}_t + \boldsymbol{\epsilon}_t \tag{3.19}$$

Estimation of $\boldsymbol{\beta}$ in Equation (3.19) can be found by using OLS. It achieves a consistent estimate of the long-run steady-state relationship between variables. Furthermore, the OLS estimator is *consistent* in the presence of a deterministic trend. Consider Equation (3.19); the OLS estimator will be

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{t=1}^{M} \boldsymbol{x}_t \boldsymbol{y}_t}{\sum_{t=1}^{M} (\boldsymbol{x}_t)^2} = \boldsymbol{\beta} + \frac{\sum_{t=1}^{M} \boldsymbol{x}_t \boldsymbol{\epsilon}_t}{\sum_{t=1}^{M} (\boldsymbol{x}_t)^2} \tag{3.20}$$

To observe the convergence, assume $\boldsymbol{y}_t = \boldsymbol{\beta} t + \boldsymbol{\epsilon}_t$ as the estimator for the constant term (t) converges at the usual rate, and consider

$$var(\widehat{\boldsymbol{\beta}}) = \sigma_\epsilon^2 \left(\frac{\sum_{t=1}^{M} t^2}{(\sum_{t=1}^{M} t^2)^2} \right) = \sigma_\epsilon^2 \left(\frac{1}{\sum_{t=1}^{M} t^2} \right) = \frac{\sigma_\epsilon^2}{M(M+1)(2M+1)/6} \quad (3.21)$$

This will converge to zero much faster than the usual OLS estimator with stationary $I(0)$ variables.

The second step of the EG approach for testing the null hypothesis of whether or not $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are cointegrated is to check $\boldsymbol{\epsilon}_t \sim I(1)$ against the alternative $\boldsymbol{\epsilon}_t \sim I(0)$. Even though there are several methods to test this, Engle and Granger advocated the ADF test by using:

$$\Delta \hat{\boldsymbol{\epsilon}}_t = \theta^* \hat{\boldsymbol{\epsilon}}_{t-1} + \sum_{i=1}^{p-1} \theta_i \Delta \hat{\boldsymbol{\epsilon}}_{t-1} + a_0 + a_1 t + \boldsymbol{\omega}_t \quad (3.22)$$

where $\hat{\boldsymbol{\epsilon}}_t$ is obtained by estimation of Equation (3.19) and $\boldsymbol{\omega}_t \sim N(0, \sigma^2)$. While testing with the ADF test, the inclusion of linear and/or constant terms in the test regression is a question that needs to be answered. Hansen's results (Hansen, 1992) obtained using Monte Carlo experiments show that having a time trend in the test model results in a loss of test power, which leads to under-rejections whether $\hat{\boldsymbol{\epsilon}}_t$ contains a deterministic trend or not. The test power indicates the probability that the test rejects the null hypothesis when the alternative hypothesis is true. Therefore, this form of testing should be based on $a_1 = 0$.

Furthermore, the standard DF tables tabulated in Table 3.1 tend to over-reject the null hypothesis for two reasons. Firstly, OLS estimates $\hat{\boldsymbol{\epsilon}}_t$ to have the smallest variance that makes the $\hat{\boldsymbol{\epsilon}}_t$ look as stationary as possible. Secondly, the test distributions are affected by the number of regressors (n) which can also change by additional characteristics in the model type such as the trend and constant. MacKinnon (MacKinnon, 1991) proposed the critical values for the response surfaces given in Table 3.2 for particular tests for a set of parameters to link all of the problems mentioned via the following relation:

$$C(p) = \phi_\infty + \phi_1 M^{-1} + \phi_2 M^2 \quad (3.23)$$

where $C(p)$ is the $p\%$ critical value. For example, estimation of a 5% critical value for 105 observations and $n = 3$ can be calculated as $-3.74 - (8.35/105) - (13.41/105^2) \approx -3.82$ where the calculation is given by Equation (3.22) with inclusion of a constant (a_0) but not a trend (a_1).

Table 3.2: Response surfaces for critical values of the cointegration test (MacKinnon, 1991).

n	Model	Significance level / %	ϕ_∞	ϕ_1	ϕ_2
1	No constant, no trend	1	-2.56	-1.96	-10.04
		5	-1.93	-0.39	0.0
		10	-1.61	-0.18	0.0
1	Constant, no trend	1	-3.43	-5.99	-29.25
		5	-2.86	-2.73	-8.36
		10	-2.56	-1.43	-4.48
1	Constant, trend	1	-3.96	-8.35	-47.44
		5	-3.41	-4.03	-17.83
		10	-3.12	-2.41	-7.58
3	No constant, no trend	1	-4.29	-13.79	-46.37
		5	-3.74	-8.35	-13.41
		10	-3.45	-6.24	-2.79

For example, three times series taken from a dataset comprising interest rates in Canada (uk.mathworks.com, n.d.) are illustrated in Figure 3.3. Their t-statistics are calculated as $[-0.6463, -0.068, 0.246]$ by the ADF test and the critical value is -1.947. Therefore, the null hypotheses for all three time series were accepted and they are all nonstationary. An OLS estimator resulted in $\boldsymbol{\epsilon} = [+1.0 \ -2.22 + 1.07]\boldsymbol{X} + 1.23$ for this particular estimation with $n = 3$ where $\boldsymbol{X}$ is the time series matrix of interest rates. Figure 3.3(b) shows the estimated cointegration relationship where the t-statistic of the ADF test is -3.93, and thus the null hypothesis is rejected; therefore, the cointegrated data are stationary.

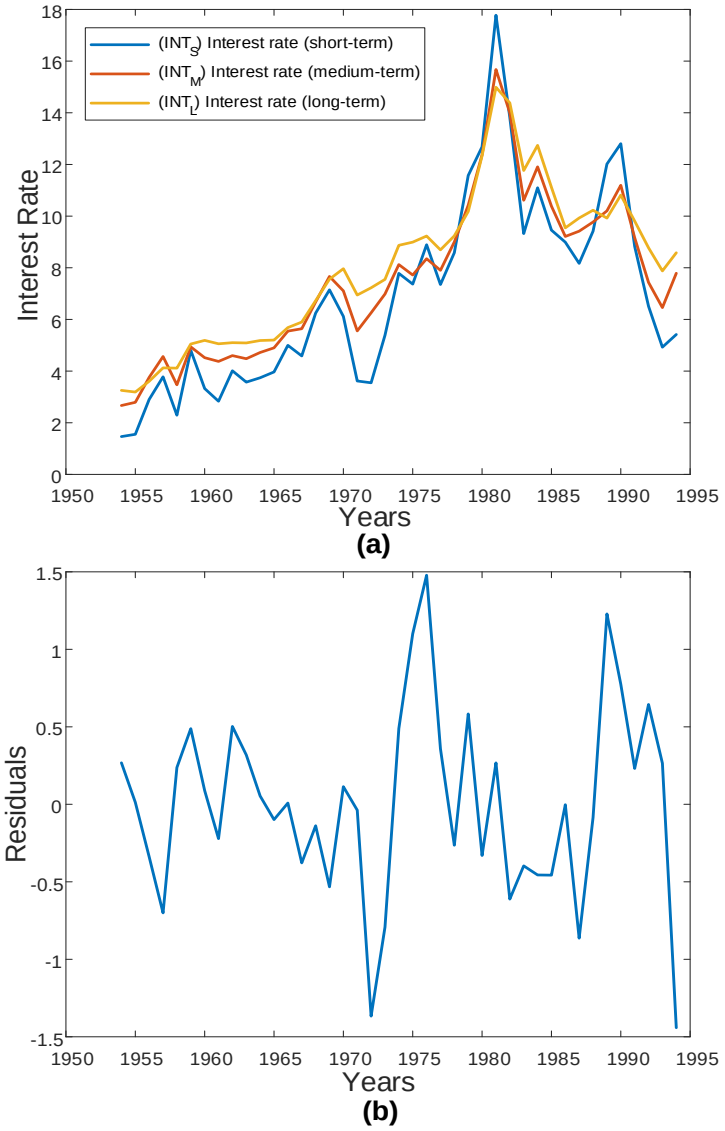

Figure 3.3: Illustration of (a)interest rates in Canada, and (b) the estimated cointegration relationship using the Engle-Granger model.

3.4.2 Cointegration in Multivariate Systems

Cointegration testing with a single equation can be quite problematic if there are $n_{ns} > 2$ nonstationary variables to be modelled where n_{ns} is the number of nonstationary variables. This arises because of the possibility that there is more than one cointegration relationship present and a single relationship can be misleading. On the other hand, the Johansen test allows use of more than one cointegration relationship by using a multivariate vector autoregression (VAR) model (Johansen, 1988). The maximum number of cointegration relationships that can be modelled is given by $n_{ns} - 1$ and can be up to 11, as the calculated critical values for Johansen test models

is limited to 12 nonstationary variables in the current literature (Harris and Sollis, 2003).

Defining a vector $\boldsymbol{z}_t$ of N potentially endogenous variables as an unrestricted VAR model involving up to k lags:

$$\boldsymbol{z}_t = \boldsymbol{A}_1\boldsymbol{z}_{t-1} + \cdots + \boldsymbol{A}_k\boldsymbol{z}_{t-k} + \boldsymbol{u}_t \tag{3.24}$$

where $\boldsymbol{z}_t \in \mathbb{R}^{n_{ns}\times 1}$ and $\boldsymbol{A}_i \in \mathbb{R}^{n_{ns}\times n_{ns}}$ is a matrix of parameters where $\boldsymbol{u}_t \sim N(0, \sigma^2)$. It can be reformulated into a vector error correction model (VECM) from:

$$\Delta\boldsymbol{z}_t = \boldsymbol{\Gamma}_1\Delta\boldsymbol{z}_{t-1} + \cdots + \boldsymbol{\Gamma}_{k-1}\Delta\boldsymbol{z}_{t-k+1} + \boldsymbol{\Pi}\boldsymbol{z}_{t-k} + \boldsymbol{u}_t \tag{3.25}$$

where

$$\begin{gathered}\boldsymbol{\Gamma}_i = -(\boldsymbol{I} - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_i), \quad i = 1, \dots, k-1 \\ \boldsymbol{\Pi} = -(\boldsymbol{I} - \boldsymbol{A}_1 - \cdots - \boldsymbol{A}_k)\end{gathered} \tag{3.26}$$

Here, the VECM contains both short and long-run information of $\boldsymbol{z}_t$ via the estimates of $\widehat{\boldsymbol{\Gamma}}_i$ and $\widehat{\boldsymbol{\Pi}}$, respectively. As $\boldsymbol{\Pi}$ stands for the long-run information, it can be separated into two terms as follows:

$$\boldsymbol{\Pi} = \boldsymbol{\alpha}\mathbf{B}^T \tag{3.27}$$

where $\boldsymbol{\alpha}$ represents the speed of adjustment to a non-equilibrium and $\mathbf{B}$ is a matrix of the long-run coefficients. Thus there are, $n_{ns} - 1$ cointegration relationships in the multivariate model which can be expressed as follows:

$$\boldsymbol{\xi}_t = \mathbf{B}^T\boldsymbol{z}_{t-k} \tag{3.28}$$

Assuming that $\boldsymbol{z}_t$ is a vector of nonstationary $I(1)$ variables, then the terms $\Delta\boldsymbol{z}_{t-i}$ (i.e., Equation (3.25)) are $I(0)$. It follows that $\boldsymbol{\Pi}z_{t-k}$ must also be stationary and $I(0)$ where $\boldsymbol{u}_t$ is already defined to be white noise and $\boldsymbol{u}_t \sim I(0)$. Three cases can achieve this condition; first, all variables of $\boldsymbol{z}_t$ are already stationary which it is not relevant in the present context; second is the instance where $\boldsymbol{\Pi} \in \mathbb{R}^{n_{ns}\times n_{ns}}$ is a zero matrix because there are no linear combinations of $\boldsymbol{z}_t$ that are $I(0)$. The third and final case is the existence of r cointegration vector in $\mathbf{B}$ where it provides $\mathbf{B}^T\boldsymbol{z}_{t-k} \sim I(0)$. Here, $\mathbf{B}$ forms r linearly independent combinations of the variables in $\boldsymbol{z}_t$ where each is stationary

under the condition $r \leq n_{ns} - 1$. The remaining $n_{ns} - r$ columns of $\mathbf{B}$ forms common-trends where they are $I(1)$. To ensure $\mathbf{\Pi} \boldsymbol{z}_{t-k}$ is $I(0)$, only the cointegration vectors in $\mathbf{B}$ must reflect Equation (3.25). The values of the $I(0)$ vectors are arranged by $\boldsymbol{\alpha}$ with the smallest elements in the last $n_{ns} - r$ columns. Furthermore, Johansen showed that it is possible to combine some of the $I(1)$ vectors with combinations of $I(2)$ variables to form stationary vectors known as *polynomial cointegration*. However, this may not be valid for all cointegration relationships. There is another way to use common-trend models in process monitoring, which will be discussed in the following sections.

Consequently, testing for cointegration amounts to checking the rank of $\mathbf{\Pi}$ or finding the number of r linearly independent columns for $\mathbf{\Pi}$. This estimation is also known as *reduced rank regression* because $r \leq (n_s - 1)$, and it is usually not possible to apply ordinary regression techniques on Equation (3.25) and was proposed by Johansen (Johansen, 1988).

As highlighted by Harris and Sollis, it is common to assume that the data are nonstationary and cointegration relationships need to be found to avoid the problem of spurious regressions (Harris and Sollis, 2003). In comparison to the field of econometrics, it is typical in process monitoring applications to assume that the data are stationary and so PCA-based techniques and a control chart based on T^2 can be used. However, this is not valid in applications wherein the PCA-based techniques perform poorly in the modelling of nonstationary variables and T^2 performs arguably poorer in comparison to the SPE for the monitoring of complex industrial processes.

It is also possible that a cointegration relationship is present when there is a mix of $I(0), I(1)$ and $I(2)$ variables in the model. Furthermore, stationary $I(0)$ variables might well play a pivotal role in building a sensible long-run equilibrium between nonstationary variables (Johansen, 1995). However, this information is weakened by the current process monitoring techniques based on cointegration residuals.

Equation (3.25) represents the VECM which is a combination of short and long-run characteristics that is a useful tool in econometrics. By regressing $\boldsymbol{z}_{t-k}$ and $\Delta \boldsymbol{z}_t$ separately and respectively on the right-hand side of Equation (3.29), the effect of short-run dynamics can be removed:

$$\Delta \boldsymbol{z}_t = \boldsymbol{P}_1 \Delta \boldsymbol{z}_{t-1} + \cdots + \boldsymbol{P}_{k-1} \Delta \boldsymbol{z}_{t-k+1} + \boldsymbol{R}_{0t} \tag{3.29}$$

where $\boldsymbol{R}_{0t}$ is difference errors, and

$$\boldsymbol{z}_{t-k} = \boldsymbol{T}_1 \Delta \boldsymbol{z}_{t-1} + \cdots + \boldsymbol{T}_{k-1} \Delta \boldsymbol{z}_{t-1} + \boldsymbol{R}_{kt} \tag{3.30}$$

where $\boldsymbol{R}_{kt}$ is the levels errors. Then, these two can be used to form residual matrices:

$$\boldsymbol{S}_{ij} = \frac{1}{M} \boldsymbol{R}_{it} \boldsymbol{R}_{jt}^T, \qquad i, j = 0, k \tag{3.31}$$

Johansen (1988) showed that the eigenvectors corresponding to the first r largest eigenvalues from the maximum likelihood estimation of $\boldsymbol{\beta}$ are:

$$|\boldsymbol{\lambda} \boldsymbol{S}_{kk} - \boldsymbol{S}_{k0} \boldsymbol{S}_{00}^{-1} \boldsymbol{S}_{0k}| = 0 \tag{3.32}$$

where $\boldsymbol{\lambda}$ is an eigen matrix from the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n_{ns}}$ wherein the corresponding eigenvectors are $\widehat{\boldsymbol{V}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_n)$ $\left(\widehat{\boldsymbol{V}}^T \boldsymbol{S}_{kk} \widehat{\boldsymbol{V}} = \boldsymbol{I}\right)$. The first r elements of $\widehat{\boldsymbol{V}}$ determines the linear combinations of the stationary relationship as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_r)$. These represent the cointegration vectors because the eigenvalues are the largest squared canonical correlation between $\boldsymbol{R}_{0t}$ and $\boldsymbol{R}_{kt}$.To determine the rank r the estimates of all distinct $\widehat{\boldsymbol{\beta}}_i^T \boldsymbol{z}_t (i = 1, \ldots, n_{ns})$ are calculated. Here, the combination of $I(1)$ from $\boldsymbol{z}_t$ and $I(0)$ from $\Delta \boldsymbol{z}_t$ elements of Equation (3.25) results in high correlations. Such a combination can only be created by the difference of $I(1)$ and $I(0)$. Therefore, the cointegration vectors must in themselves be $I(0)$ in order to achieve higher correlations. As a result, the magnitude of the eigenvalues $\left(\hat{\lambda}_i\right)$ represents how strongly correlated the cointegration relation is with the stationary part of the model. The remaining $n_{ns} - r$ combinations are theoretically uncorrelated as they still show $I(1)$ characteristics as a common-trend where $\hat{\lambda}_i = 0$ for $i = r + 1, \ldots, n_{ns}$. Johansen (Johansen, 1992) also showed that the relationship between the eigenvalues and $\boldsymbol{\alpha}$ is given by:

$$\hat{\lambda}_i = \widehat{\boldsymbol{\alpha}}_i^T \boldsymbol{S}_{00}^{-1} \widehat{\boldsymbol{\alpha}}_{\boldsymbol{i}} \tag{3.33}$$

where

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{S}_{0k}\widehat{\boldsymbol{\beta}} \tag{3.34}$$

Following the estimation of $\widehat{\boldsymbol{\beta}}$, the procedure can continue with the rank testing. The null hypothesis defined for $n_{ns} - r$ unit-roots remaining after r cointegration vectors can be defined as follows:

$$H_0: \quad \lambda_i = 0, \qquad i = r+1, \dots, n_{ns} \tag{3.35}$$

The trace statistics defined to test the null hypothesis, which is the comparison of the log of the maximised likelihood function of the restricted and unrestricted model where the restriction is the number of cointegration vectors denoted as r can be defined as:

$$\lambda_{trace} = -2\log(Q) = -M\sum_{i=r+1}^{n_{ns}} \log(1-\hat{\lambda}_i) \tag{3.36}$$

where $r = 0,1, \dots, n_{ns} - 2, n_{ns} - 1$ and Q is the ratio of the restricted maximised likelihood to the unrestricted maximised likelihood. Another test called the maximal eigenvalue or λ_{max} statistic also tests the significance of the largest λ_r:

$$\lambda_{max} = -M\log(1-\hat{\lambda}_{r+1}) \tag{3.37}$$

where $r = 0,1, \dots, n_{ns} - 2, n_{ns} - 1$. Both of them test the null hypothesis, which is the existence of r cointegration vector against the presence of $r+1$ cointegration vectors.

The dataset comprising interest rates in Canada, which is illustrated in Figure 3.3(a) was also evaluated with Johansen test. The Johansen test resulted in $r = 2$ which is the maximum rank that it can take ($r_{max} = n - 1 = 3 - 1 = 2$). The residuals were calculated as $\boldsymbol{\xi}_t = \mathbf{B}^T\boldsymbol{z}_{t-1} = \begin{bmatrix} 0.077 & 1.753 \\ -2.14 & -3.72 \\ 2.056 & 1.716 \end{bmatrix}^T \boldsymbol{z}_{t-1}$ where $\boldsymbol{z}$ is the time series matrix of interest rates. Figure 3.4 shows the estimated cointegration residuals where the t-statistics of ADF test are -2.40 and -1.31 and the critical value is -1.239. Therefore, the null hypothesis is rejected and so, the data are stationary.

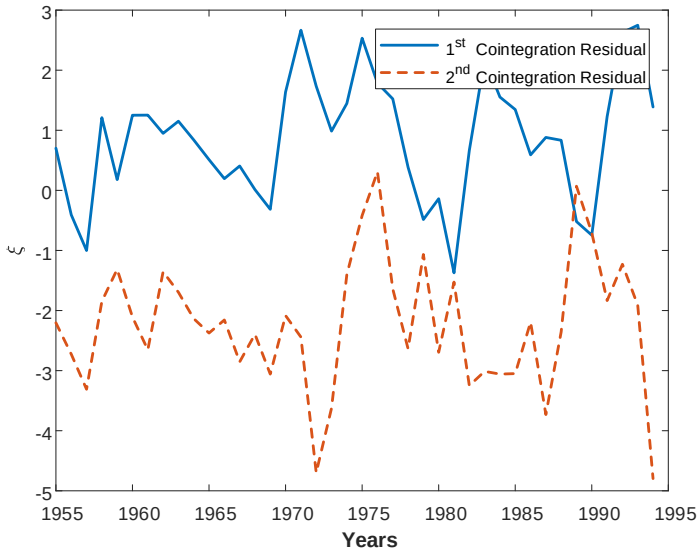

Figure 3.4: Illustration of estimated cointegration residuals using the Johansen model.

3.4.3 Common-trend Representation

Common-trends are present when the cointegration vectors do not establish stationary cointegration space between $\boldsymbol{z}_{t-k}$ and **B.** Another representation for the cointegrated variables is known as the *common-trend representation* or *dynamic factor model* proposed by Stock and Watson (Stock and Watson, 1988):

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{\beta}_{\perp}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{x}_t &= \boldsymbol{x}_{t-1} + \boldsymbol{u}_t \end{aligned} \tag{3.38}$$

where $\boldsymbol{\epsilon}_t$ is a stationary time series, $\boldsymbol{\beta}_{\perp} \in \mathbb{R}^{n_{ns} \times (n_{ns}-r)}$ is the loading matrix, $\boldsymbol{x}_t \in \mathbb{R}^{(n_{ns}-r)\times 1}$ is a random walk process and u_t is an independent identical distributed (IID) process. Even though, Stock and Watson's model has a unique estimation procedure for the determination of $\boldsymbol{\beta}_{\perp}$, a common-trend representation will be used for further procedures. These are derived by establishing a connection between cointegration analysis and the common-trend model. More precisely, it was shown in Section 3.4.2 that $\boldsymbol{z}_t$ could be modelled using Equation (3.24), where the cointegration matrix is given by calculating Equation (3.27). Here, **B** consist of $\boldsymbol{\beta}$ and the unused cointegration vectors, which have no cointegration relationship, retain the common-trends. Furthermore, $\boldsymbol{\beta}_{\perp} \in \mathbb{R}^{n_{ns} \times (n_{ns}-r)}$ is the orthogonal complement of $\boldsymbol{\beta} \in \mathbb{R}^{n_{ns} \times r}$ that helps to model common-trends (Lin, Kruger and Chen, 2017).

Escribano and Pena (Escribano and Peña, 1994) demonstrated that the relationship between $\boldsymbol{\beta}_{\perp}$ and $\boldsymbol{\beta}$ can be used to represent $\boldsymbol{z}_t$ in the form of a Kasa decomposition:

$$\boldsymbol{z}_t = \boldsymbol{\beta}_\perp[\boldsymbol{\beta}_\perp^T\boldsymbol{\beta}_\perp]^{-1}\boldsymbol{\beta}_\perp^T\boldsymbol{z}_t + \boldsymbol{\beta}[\boldsymbol{\beta}^T\boldsymbol{\beta}]^{-1}\boldsymbol{\beta}^T\boldsymbol{z}_t \quad (3.39)$$

where $\boldsymbol{\beta}_\perp^T\boldsymbol{z}_t$ and $\boldsymbol{\beta}^T\boldsymbol{z}_t$ are the identified nonstationary and stationary factors, respectively. The further usage of the identified nonstationary factors for process monitoring is detailed in Chapter 4 to 6.

3.5 Conclusions

This chapter has discussed a wide range of topics about nonstationarity and the modelling techniques that are applicable to nonstationary data: cointegration starting from econometrics to process systems engineering. Even though some studies have proposed methods to deal with nonstationarity in the area of process monitoring, there had not been a breakthrough in the monitoring of nonstationary variables from complex industrial processes until the use of cointegration analysis. Its capability in nonstationary variable modelling has been proved not only in its main area of application, econometrics, but also in areas such as process system engineering and construction.

The use of cointegration starts with the identification of the nonstationary variables. Several testing techniques have been introduced. The most well-known and popular tool is the ADF test because of its ease-of-use. The cointegration model can then be estimated by the Engle-Granger approach or the Johansen Test. In this modelling study, the Johansen test is favoured as it can be used to analyse multivariate variables.

In contrast, some nonstationary characteristics might not be involved in the model by the Johansen test, which is mostly $I(1)$ and $I(2)$ series, and so can be modelled by the common-trend representation. Through the use of common-trend residuals-based process monitoring, the dynamic information remaining in the unused cointegration vectors can be used for monitoring.

4. MONITORING CONTINUOUS PROCESSES USING COINTEGRATION BASED APPROACHES

4.1 Overview

Continuous processes were the initial focus of MSPC techniques where the process variability is within defined limits to provide continuous excellence in manufacturing. Most MSPC techniques are implemented on continuous processes first then batch or fed-batch processes.

Chapter 2 introduced the main concepts of MSPC and PCA based techniques within the broader literature. Some of the detailed descriptions of MSPC techniques in Chapter 2 are particularly relevant to this chapter. Classical MSPC aims to detect deviations of typical process behaviour during two distinct phases of the process measurements called offline training and online diagnosis. Offline training is model training which is the practice of retrospectively evaluating whether a previously completed process was statistically in-control. Likewise, online diagnosis is the practice of determining whether new observations from the process are in-control as they are obtained in real-time. During both phases, time dependence in the form of autocorrelation and/or nonstationarity can be present. Autocorrelation arises when the in control measurements within one-time series are not serially independent, while nonstationarity arises when the parameters governing a process, such as the mean or covariance, change over time. In this chapter, time dependency and nonstationarity will be the main focus.

Control charts based on PCA have been successfully applied in high dimensional data sets when the data are not time-dependent. This is also called static PCA because the model contains no dynamic components when the training data set consists of samples for time t. Therefore, no attempt is made to model a relationship between variables at different time points (autocorrelations), and the PCA model cannot be adjusted for changes in the underlying parameters (nonstationarity). However, complex industrial processes show time dependency and nonstationary characteristics due to the use of feedback control systems, process disturbances, changes in the operating conditions, variations in the feed rates, etc.

DPCA is proposed as a solution to the problem of autocorrelated data by using different time points in the training data set and aims to extract independent noise from the uncorrelated variable space; however, it is still not a solution for nonstationarity. Cointegration and common-trend residuals-based process monitoring approaches offer solutions to the nonstationary data modelling problem.

Starting from the first implementation of a cointegration residuals-based monitoring method, one of the main problems is that it is only applicable to nonstationary variables while complex industrial processes contain both stationary and nonstationary variables. The second problem is inefficiency in the modelling when higher level nonstationary time series are present. This can be solved by using common-trend residuals-based monitoring; however, it gives rise to another problem which is the increased number of control charts associated with all of different models. MSPC became a popular technique after SPC as it required fewer charts. Current methods reported in the literature require the usage of several control charts at the same time, which is a disadvantage compared to conventional MSPC approaches that only require a single control chart consisting of T^2 and SPE performance metrics. To solve all these problems, a new multi-level multi-factor model based on cointegration analysis and a PCA model is proposed as a means of providing a single control chart composed of T^2 and SPE performance metrics for the monitoring of complex industrial processes exhibiting both stationary and nonstationary characteristics.

In this chapter, a continuous stirred-tank heater (CSTH) is used to compare PCA, DPCA, cointegration residuals-based monitoring, common-trend residuals-based monitoring, and finally the new multi-level multi-factor method for monitoring of a continuous process. The performance of the methods presented is evaluated using 2 different examples to assess the capability of the monitoring methods to detect different types of faults.

4.2 Introducing the Continuous Stirred Tank Heater Simulator

The continuous stirred tank heater system (CSTH) is a subsystem of the most advanced complex systems (Thornhill, 2008). The CSTH has been tested by several process performance monitoring techniques (Ding, 2008, 2014; Yu and Qin, 2008; Wang, Yin and Kaynak, 2014; Kundu, Kundu and Damarla, 2017). In the simulation of a CSTH, hot and cold water is mixed from the input pipes and then it is heated by steam through

a heating coil. Even though the simulation does not involve a chemical reaction, it comprises sensors, electrical units, valves and a heat exchanger. As the system can be described as second order plus a dead time, the time dependent and nonstationary characteristics can be seen through the time. A realistic model of a CSTH was proposed by Thornhill and validated on a pilot plant at the University of Alberta (Thornhill, Patwardhan and Shah, 2008).

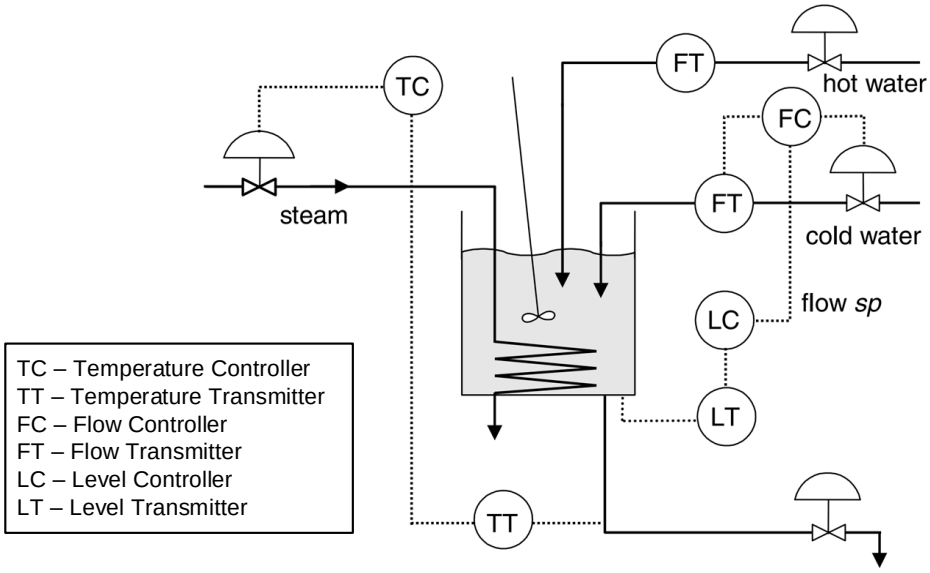

Figure 4.1: A schematic of the continuous stirred tank heater.

The configuration of the system is illustrated in Figure 4.1. Because the CSTH is well mixed, the temperature of a liquid in the tank is the same as the temperature of a liquid in the outlet pipe. The tank has a circular cross section with a volume of 8 L and height of 50 cm. The inputs or utilities namely hot and cold water come from a shared service; therefore, the pressure of the water can fluctuate between $60 - 80$ psi. Control valves in the plant have pneumatic actuators which use a compressed air supply with a pressure of $3 - 15$ psi. Furthermore, the flow instruments are orifice plates with a differential pressure transmitter that works between a $4 - 20$ mA output.

The dynamic volumetric and heat balances are:

$$\frac{dV(x)}{dt} = f_{cw} + f_{hw} - f_{out}(x) \tag{4.1}$$

where x is the level, V is the volume of water, f represents the flow rate of the liquid where cw, hw and out denote the cold water inlet, the hot water inlet and the outlet flow rates, respectively. Typically, there is expected to be a linear relationship between

the level and volume; however, the volume occupied by the heating coils in the lower half of the tank shows nonlinear characteristics. The top of the heating coils is 16.9 cm from the base of the tank and above this height, the relationship between volume and level becomes linear.

$$\frac{dH}{dt} = W_{st} + h_{cw}\rho_{cw}f_{cw} + h_{hw}\rho_{hw}f_{hw} - h_{out}\rho_{out}f_{out}(x) \tag{4.2}$$

where H is the total enthalpy in the tank, and h and ρ represent the specific enthalpy and density of the liquids, respectively. In the well-mixed case $h_{out} = H/V\rho_{out}$. The set temperatures for hot and cold water are 50 ℃ and 24 ℃, respectively. Here, the manual outlet valve was fixed at 50% under normal operating conditions and the flow rate (m^3s^{-1}), which was calculated empirically from experiments, is given by:

$$f_{out}(x) = 10^{-4}\left(\left(0.1013\sqrt{(55 + x)}\right) + 0.02037\right) \tag{4.3}$$

where the outlet valve was 55 cm below the bottom of the tank and the head of the water. 0.1013 was calculated from the slope of the best fit straight line to the calibration graph for the outlet flow rate, and 0.02037 was determined from the vertical axis intercept. Furthermore, W_{st} is the heat inlet flow rate from steam and it depends on the steam valve setting. Because of the heat exchange and heat transfer coefficient could not be measured, the relationship is determined empirically. The heat balance for steady-state operation with only cold water as the inlet with $f_{cw} = f_{out}$ is given by:

$$W_{st} = h_{out}\rho_{out}f_{out}(x) - h_{cw}\rho_{cw}f_{cw} \tag{4.4}$$

The calculation for W_{st} is tabulated for values between $0 - 15.04$ kJs^{-1} (Thornhill, Patwardhan and Shah, 2008). The cold water valve dynamics were found as a first order lag with time delay of 1 s. and time constant of the valve is 3.8 s. Therefore, the valve transfer function is:

$$MV(s) = \frac{e^{-s}}{3.8s + 1}OP(s) \tag{4.5}$$

where $MV(s)$ is the valve position and s represents a domain transformed by a Laplace transformation from a continuous time signal which can be represented as $s = j\omega$ in

the Fourier transform where ω is frequency and $j = \sqrt{-1}$. In closed loop, $OP(s)$ is the controller output while in open loop it is a valve demand signal applied to the valve. The inputs for electronic signals, which correspond to the steam and cold water valves, are in the range $4 - 20$ mA. The measurements form the temperature, level and cold flow sensors are on the same scale. The variables utilised for process monitoring of the CSTH are listed in Table 4.1.

Table 4.1: CSTH variables for process monitoring.

Number	Definition	Number	Definition
1	Cold water valve (mA)	8	Hot water valve (mA)
2	Steam valve (mA)	9	Temperature (mA)
3	Cold water temperature (mA)	10	Hot water temperature (mA)
4	Cold water flow (mA)	11	Hot water flow (mA)
5	Cold water flow ($m^3\ s^{-1}$)	12	Hot water flow ($m^3\ s^{-1}$)
6	Temperature (℃)	13	Overflow
7	Level (mA)	14	Level (cm)

4.3 Monitoring Techniques for Continuous Processes

4.3.1 Principal Component Analysis

Figure 4.2 illustrates the relationship between data pre-treatment and a PCA model that uses all variables without considering the stationary characteristics of the data.

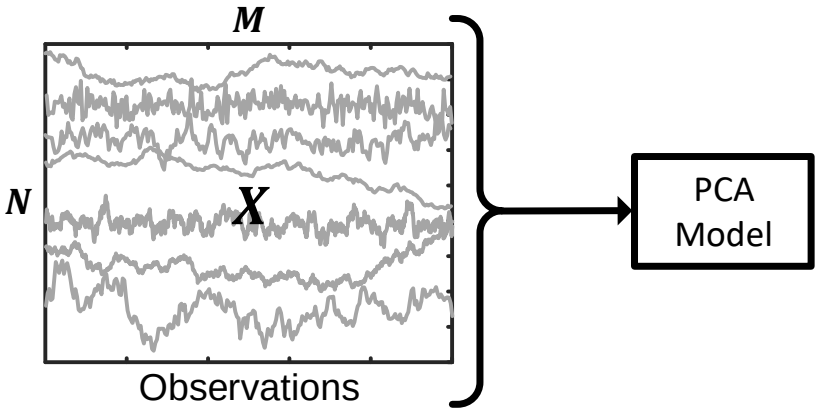

Figure 4.2: Illustration of the treatment of data , $\boldsymbol{X}$, which comprises both stationary and nonstationary variables by PCA.

A PCA model has only one design parameter which is the number of principal components (PCs). A reliable model must represent sufficient explained variance to capture the fault signature while avoiding false alarms. It dictates the performance of the model built on the training data set in terms of type-I errors, which is the false alarm rate, and type-II errors, which is the missed fault rate. A description of the use of a PCA model for process monitoring is given in Section 2.3. Moreover, $T^2$and SPE metrics are defined in Section 2.5 for further information.

4.3.2 Dynamic Principal Component Analysis

Correlated and dependent variables are one of the distinct characteristics of complex industrial process data. One way to address autocorrelation is to perform first-order differencing. However, this can cause issues in terms of the dynamic information of the data as parts of the data might be whitened after differencing. In the case of step faults, differencing will reveal the significant difference between the previous and current operating point. However, differenced time series reveal that change on only one sample as the step faults reach its final point in a consecutive time. Furthermore, a single sample that shows the step fault, can be considered as an outlier in the control chart as the process goes back to the normal operation region due to the feedback controllers.

DPCA is an extension of PCA to handle autocorrelated data sets (Ku, Storer and Georgakis, 1995). It was a continuation of an earlier study where ARIMA models were used to fit the data to reduce cross-correlation between the variables (Ku, Storer and Georgakis, 1994). However, use of ARIMA models is complex as the number of parameters for each sub-model needs to be estimated for a high number of variables in a multivariate process. On the other hand, DPCA combines a PCA model's ability to cope with high dimensionality and ARIMA's lag operators to deal with autocorrelations. Assume that the data matrix for DPCA is:

$$\boldsymbol{X}_L^T = [\boldsymbol{X}_t^T\ \boldsymbol{X}_{t-1}^T\ \boldsymbol{X}_{t-2}^T \dots \boldsymbol{X}_{t-l}^T\ ] \tag{4.6}$$

where l is the lag parameter and

$$\boldsymbol{X}_t^T = \begin{bmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & & \vdots \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ x_{M,1} & \cdots & x_{M,N} \end{bmatrix}, \quad \boldsymbol{X}_{t-1}^T = \begin{bmatrix} 0 & 0 & 0 \\ x_{1,1} & & x_{1,N} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ x_{M-1,1} & \cdots & x_{M-1,N} \end{bmatrix} \tag{4.7}$$

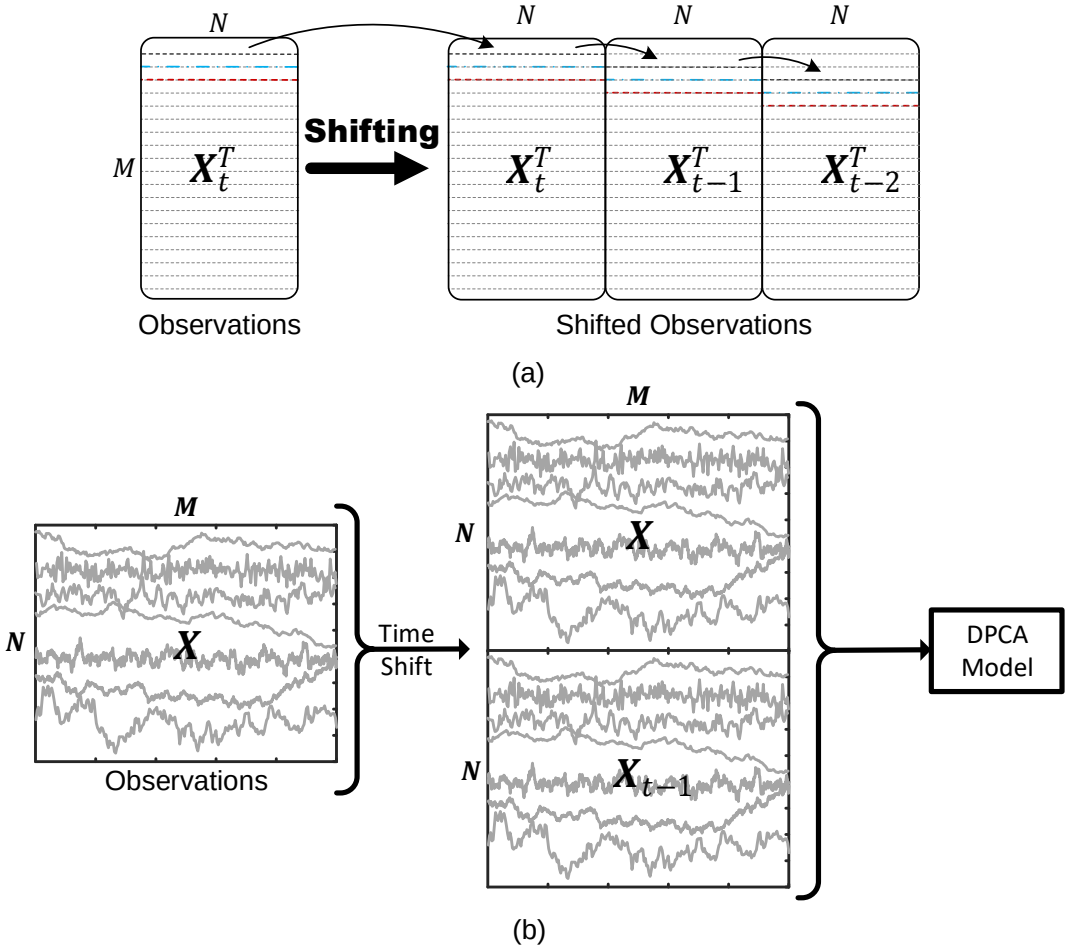

Figure 4.3: (a) A schematic representation of DPCA use with variables lagged twice, $[\boldsymbol{X}_t\ \boldsymbol{X}_{t-1}\ \boldsymbol{X}_{t-2}\ ]$, and (b) DPCA use with variables lagged once, $\boldsymbol{X}_{t-1}$, with data, $\boldsymbol{X}$, that comprises both stationary and nonstationary variables.

The use of the lag operator for each variable is illustrated in Figure 4.3. One of the critical steps to provide a DPCA model is to select the number of the time lag shifts as it can profoundly affect the explained variance level per principal component. A simple way to determine the optimum number of the time lag shifts for a DPCA model is to examine the ACFs of each score. Depending on the existence of autocorrelation, an additional time lag shift can be added until enough lags have been added to sufficiently reduce the auto and cross-correlations between PCA scores. This does not aim to reduce all correlations starting from the first PC. It rather tries to create

correlated and correlation free score spaces. However, this is extremely cumbersome because of the number of variables.

Consequently, Ku (Ku, Storer and Georgakis, 1995) provided an algorithm given in Table 4.2 to determine how many time lag shifts can define the new linear relationship (r_{new}). It begins with no lags and sequentially increases the number of the lag until the maximum number of the time lag shifts is reached. The number of linear relations (r_l) will be the total number of variables minus the number of PCs. When a new time lag shift does not reveal an important linear relationship between the PCs, the algorithm stops and the previous lag number is selected. The suggested time lag shift is typically one or two. For example, the algorithm suggests use of $l = 1$ for a continuous stirred tank reactor simulation which has 7 variables (Malik, 1998).

Table 4.2: Linear relationship determination algorithm for dynamic principal component analysis (Ku, Storer and Georgakis, 1995).

1: Set the lag parameter $l = 0$.
2: Form the new data matrix $\boldsymbol{X}_L^T = [\boldsymbol{X}_t^T\ \boldsymbol{X}_{t-1}^T\ \boldsymbol{X}_{t-2}^T \dots \boldsymbol{X}_{t-l}^T\ ]$.
3: Perform PCA and calculate *all* the principal component scores.
4: Set the comparison index $j = N(l+1)$ and relation rank $r_l(l) = 0$ where N is the number of variables.
5: If jth eigenvalue from PCA nearly equal to zero, then continue to Step 7.
6: Set $j = j + 1$ and $r_l(l) = r_l(l) + 1$, then go back to Step 5.
7: Calculate number of a new relationship between PCs:

$$r_{new}(l) = r_l(l) - \sum_{i=0}^{l-1}(l - i + 1)r_{new}(i) \tag{4.8}$$

8: If $r_{new}(l) \leq 0$, then go to Step 10, otherwise, go to next step.
9: Set $l = l + 1$, then go to Step 2.
10: Stop

DPCA provides dynamic modelling but it does not aim to give uncorrelated score variables. The score variables will still be autocorrelated and probably cross-correlated even though there is no cross-correlation present between variables. Furthermore, another significant but mostly ignored feature explained by Ku is "For a dynamic system, after properly selecting the number of constraining relations, Q-statistics can be calculated from the *independent* noise space. If data are auto-correlated, T^2 will still be calculated from the correlated variables. Since T^2 represents the movement of

the data in the multidimensional space, it contains important information about the process although the variables from which it is calculated are not independent". The take-home message from here is to use the independent noise space to determine a SPE metric. Therefore, after choosing an appropriate number of principal components, the training data can be transformed into loading and score matrices:

$$\boldsymbol{X}_L = \sum_{r=1}^{R} \boldsymbol{p}_r \boldsymbol{t}_r^T + \boldsymbol{E} = \boldsymbol{P}\boldsymbol{T}^T + \boldsymbol{E} \tag{4.9}$$

It follows that the determination of T^2 which is given by:

$$(\boldsymbol{T^2})_{DPCA} = \boldsymbol{X}^T \boldsymbol{P} \boldsymbol{\Lambda}^{-1} \boldsymbol{X}^T \boldsymbol{P}^T \boldsymbol{X} \tag{4.10}$$

In the Q-statistic monitoring, the use of independent noise space is equivalent to uncorrelated score variables. Therefore, some modification is needed to the SPE equation as follows:

$$\boldsymbol{SPE}_{PCA} \cong \|(\boldsymbol{I} - \boldsymbol{P}_{ind}\boldsymbol{P}_{ind}^T)\boldsymbol{X}\| \tag{4.11}$$

where $\boldsymbol{P}_{ind}$ consist of loading vectors which correspond to the independent space. Similarly, the upper limit calculation in Equation (2.17) and (2.18) can be performed using a modified calculation of θ:

$$\theta_i = \sum_{k=1}^{R_{ind}} \lambda_k^i, \qquad i = 1,2,3 \tag{4.12}$$

where R_{ind} is the number of independent score vectors.

4.3.3 Cointegration Residuals-based Process Monitoring

In complex industrial processes, cointegration models can describe the long-run relationships between variables of the process operating under normal conditions. The cointegration model can provide the stationary residuals space from the nonstationary variables as long as the working conditions for the trained model are not changed. However, it breaks the long-run equilibrium when a fault occurs and the residual sequence might become nonstationary.

Process monitoring of nonstationary variables using cointegration residuals, illustrated in Figure 4.4., was proposed by Chen et al. (2009), and the control limits were proposed by Li et al. (2014). Before starting to train the models, the number of nonstationary variables (n_{ns}) needs to be known for the nonstationary variables, $\boldsymbol{X}_{NS} \in \mathbb{R}^{n_{ns} \times M}$. This can be achieved using the unit root tests introduced in Section 3.3. The cointegration model can be defined as a long-run equilibrium of the process when the process operates under normal conditions. The Johansen test is the most common multivariate cointegration tool introduced in Section 3.4.2 to find the cointegration matrix $\boldsymbol{\beta}$.

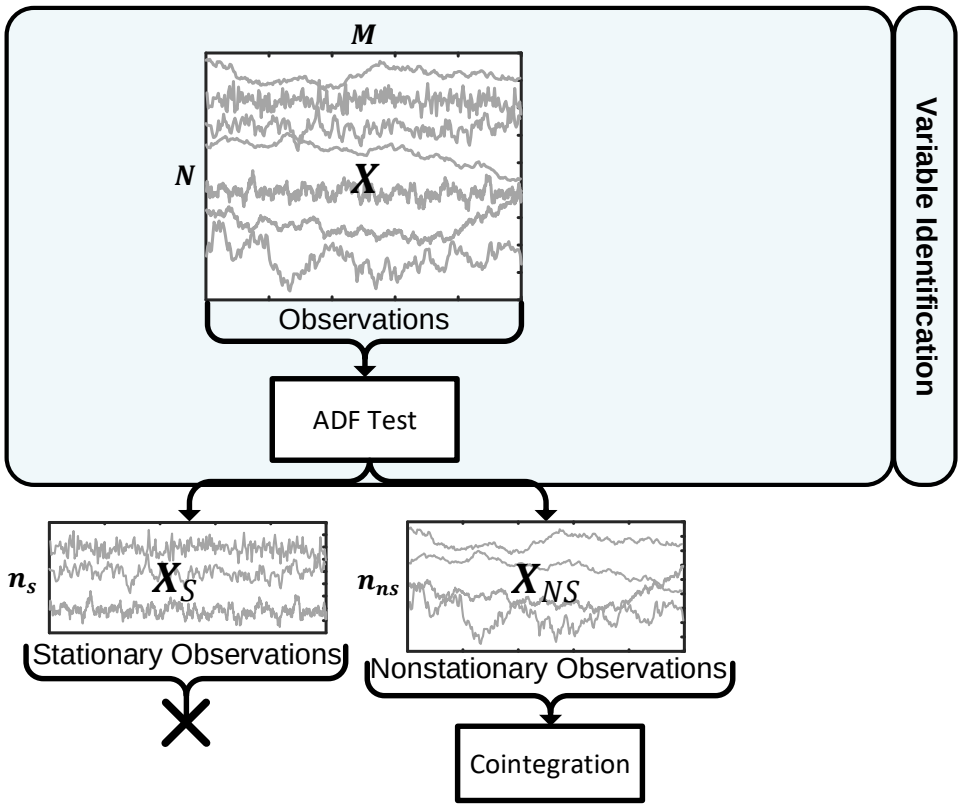

Figure 4.4: Illustration of process monitoring method based on cointegration residuals.

It has been shown that $I(1)$ variables $(\boldsymbol{x}_t)$ can combine with the cointegration matrix to estimate stationary residuals as given below:

$$\boldsymbol{\xi}_t = \boldsymbol{\beta}^T \boldsymbol{X}_{NS} \tag{4.13}$$

where $\boldsymbol{\xi}_t$ is called the residual equilibrium sequence and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \dots, \boldsymbol{\beta}^{(R)}\right)$ is the cointegration matrix under the condition $\boldsymbol{\xi}_t = \boldsymbol{\beta}^T \boldsymbol{X}_{NS} \sim I(d-b)$ where $d \geq b > 0$. R describes the rank cointegration relationship.

The T^2 metric for cointegration residuals-based monitoring proposed by Li et al. (2014) is based on a residual equilibrium series as follows:

$$(\boldsymbol{T^2})_{CA} = \boldsymbol{\xi}_t^T \boldsymbol{\Lambda}_{CA}^{-1} \boldsymbol{\xi}_t \tag{4.14}$$

where $\boldsymbol{\Lambda}_{\mathrm{CA}} = \sum_{k=1}^{M} \boldsymbol{\xi}_k^T \boldsymbol{\xi}_k \, /M$ is the sample covariance matrix of the residuals and M is the sample size. T^2 shows the changes in the variation of the common stationary variables on the nonstationary variables $(\boldsymbol{X}_{NS})$. Similar to the well-known T^2, the F-distribution is also used for providing the control limits:

$$\frac{R(M-1)(M+1)}{M(M-R)} F_{\alpha}(R, M-R) \tag{4.15}$$

where α is the significance level of the F-distribution $F_{\alpha}(R, M-R)$ with degrees of freedom, R and $M-R$. The Johansen test guarantees that the residuals from the first R cointegration vectors are stationary. The remaining $n_{NS} - R$ cointegration vectors still represent a significant portion of the process characteristics where it consists of the variables of $I(1)$ and $I(2)$. In some cases, the Johansen test may result in a smaller number of cointegration vectors that might cause loss of effectiveness of monitoring of nonstationary variables. In such circumstances, a common-trend model is a helpful tool to evaluate the unused cointegration relationship estimated by the Johansen test.

4.3.4 Cointegration and Common-trend Residuals-based Process Monitoring

The use of common-trend residuals for process monitoring was introduced by Lin (Lin, Kruger and Chen, 2017). The key feature of their model is that it included the common-trend residuals in addition to cointegration residuals. However, it is noted that the combined monitoring schemes require separate control limits and charts to monitor the common-trend and cointegration residuals, as illustrated in Figure 4.5.

A common-trend model requires cointegration analysis to generate the cointegration matrix. It requires a set of vectors $\boldsymbol{\beta}_{\perp} \in \mathbb{R}^{n_{ns} \times (n_{ns}-R)}$ such that it is perpendicular to the cointegration matrix $\boldsymbol{\beta} \in \mathbb{R}^{n_{ns} \times R}$ and $[\boldsymbol{\beta}, \boldsymbol{\beta}_{\perp}]$ has full rank. Therefore, $\boldsymbol{\beta}_{\perp}$ is the orthogonal complement of $\boldsymbol{\beta}$ and is termed the common-trend loading matrix. Common-trend residuals can be represented as:

$$\boldsymbol{e}_t = \boldsymbol{\beta}_{\perp} \boldsymbol{X}_{NS} \tag{4.16}$$

where $\boldsymbol{e}_t$ represents the nonstationary residuals from cointegration analysis. The residuals represent the stochastic trends caused by nonstationarity but were not involved in the residuals for cointegration analysis. As $\boldsymbol{\beta}_\perp$ is the complement of $\boldsymbol{\beta}$, the rank is dependent on $\boldsymbol{\beta}$, which is proven in Section 3.4.3.

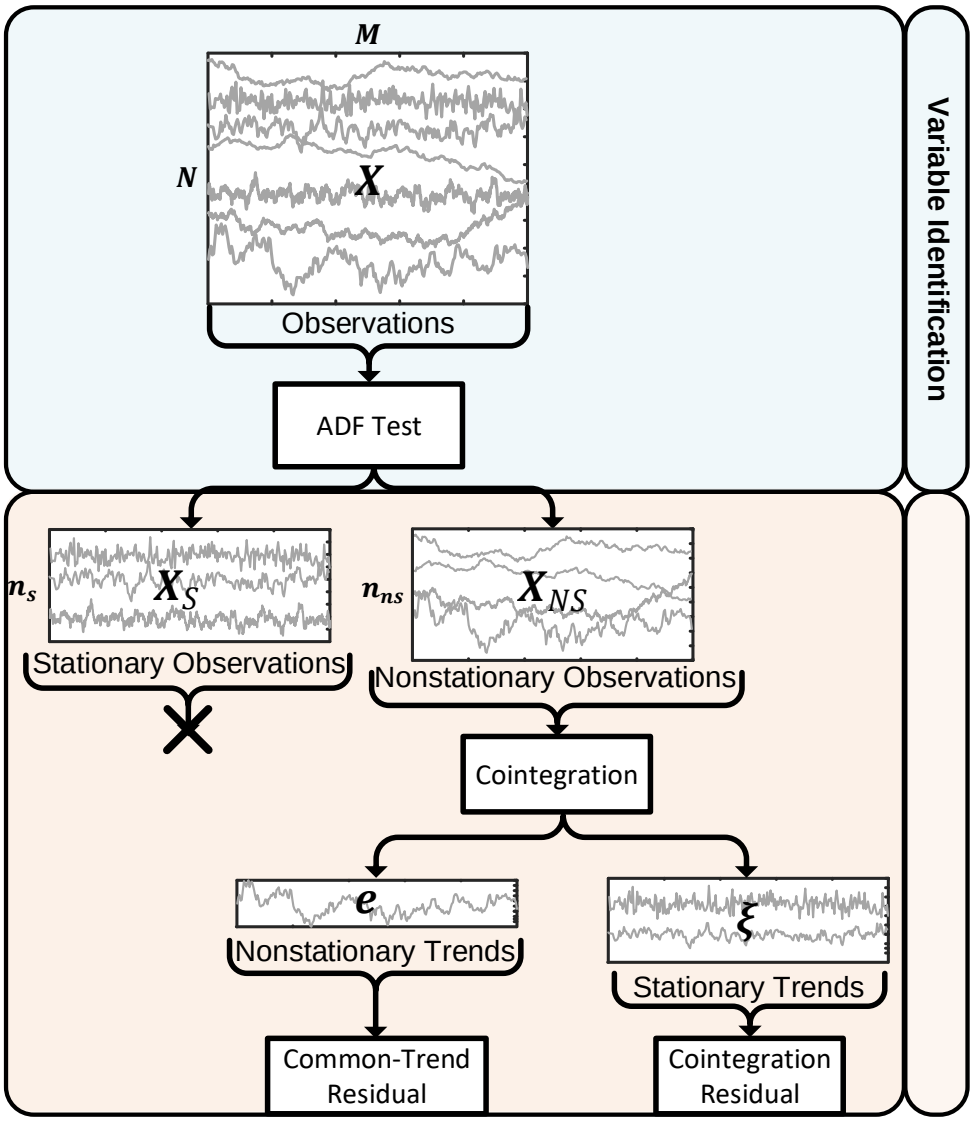

Figure 4.5: Illustration of process monitoring method based on common-trend and cointegration residuals.

Nonstationary factors that were obtained from the common-trend model are not involved in the cointegration analysis residuals due to nonstationarity but they are still integrated of order $d = 1$ (Lin, Kruger and Chen, 2017). The difference operator converts $d = 1$ variables into stationary by using $\Delta \boldsymbol{e}_t = \boldsymbol{e}_t - \boldsymbol{e}_{t-1}$ where $\Delta \boldsymbol{e}_t$ is stationary. A sample point of $\boldsymbol{e}_t$ can be represented via a $VAR(p)$ process:

$$\boldsymbol{e}_t = \boldsymbol{A}_1\boldsymbol{e}_{t-1} + \cdots + \boldsymbol{A}_p\boldsymbol{e}_{t-p} + \boldsymbol{\varsigma}_t \tag{4.17}$$

where $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_p$ are the autoregression matrices and $\boldsymbol{\varsigma}_t$ denotes a random error vector which has a multivariate normal distribution with $N(0, \boldsymbol{\Lambda}_{CT})$. The VECM can be obtained after the subtraction $\boldsymbol{e}_{t-1}$ from both sides of Equation (4.17) and rearrangement of the terms as follows:

$$\Delta\boldsymbol{e}_t = \boldsymbol{\Pi}\boldsymbol{e}_{t-1} + \boldsymbol{\Theta}_1\Delta\boldsymbol{e}_{t-1} + \cdots + \boldsymbol{\Theta}_{p-1}\Delta\boldsymbol{e}_{t-p+1} + \boldsymbol{\varsigma}_t \tag{4.18}$$

where

$$\begin{aligned} \boldsymbol{\Pi} &= -\left(\boldsymbol{I}_{n_{ns}} - \boldsymbol{A}_1 - \cdots \boldsymbol{A}_p\right) \\ \boldsymbol{\Theta}_i &= -\left(\boldsymbol{A}_{i+1} + \cdots + \boldsymbol{A}_p\right), \qquad i = 1, \ldots, p-1 \end{aligned} \tag{4.19}$$

Here, $\boldsymbol{\varsigma}$ is stationary and does not contain any serial correlation as it is whitened by the autoregression. Hence, it is useful to use as a source for the monitoring metrics. The T^2 metric can also be defined for $\boldsymbol{\varsigma}$ as below:

$$\boldsymbol{T}_{CT}^2 = \boldsymbol{\varsigma}_t^T \boldsymbol{\Lambda}_{CT}^{-1} \boldsymbol{\varsigma}_t \tag{4.20}$$

where $\boldsymbol{\Lambda}_{\mathrm{CT}} = \sum_{k=1}^{M} \boldsymbol{\varsigma}_k^T \boldsymbol{\varsigma}_k \,/M$ is the sample covariance matrix of the residuals where M is the sample size. $\boldsymbol{T}_{CT}^2$ shows the variation of the change of the nonstationary variables in the model. Furthermore, the control limits can be determined as follows:

$$\frac{(n_{ns} - R)(M-1)(M+1)}{M(M - n_{ns} + R)} F_\alpha(n_{ns} - R, M - n_{ns} + R) \tag{4.21}$$

where α is the significance level of the F-distribution $F_\alpha(n_{ns} - R, M - n_{ns} + R)$ with degrees of freedom $n_{ns} - R$ and $M - n_{ns} + R$.

A nonstationary time series can be converted into a stationary time series by first-order differencing but that might also cause some loss in information related to the process dynamics as discussed in earlier sections. Lin et al. (2017) proposed the use of a forecast recovery filter based on vector autoregressive moving average (VARMA) model. Filtering has two issues: first, it is specifically designed for step faults. Secondly, it is often not practical to realise step changes in the process systems. As a result, having a differenced relatively small number of variables has the same impact as much as having highly specialised filtering.

On the other hand, a relatively small number of variables that have common-trends might be close to 1 or sometimes 0; thus, having only a common-trend model for monitoring will not be the answer for effective monitoring of nonstationary processes. Accordingly, this gives rise to the search for a multi-level multi-factor model that can cover all type of characteristics in the processes, which requires only one control chart to be monitored.

4.3.5 Multi-level Multi-factor Process Monitoring Model

A multi-level multi-factor model, which covers all types of data characteristics, and requires only one monitoring chart is needed. A single chart based on T^2 and SPE metrics is easier to monitor than multiple charts; the latter may clash in terms of identifying a sample as an abnormality or an outlier. Having a multi-level multi-factor model that uses specific models for different type of data characteristics will also provide enhanced performance in terms of earlier detection of faults and increased fault detection rates compared to conventional and cointegration residual-based approaches, which focus only on the nonstationary variables.

The multi-level multi-factor process monitoring model consists of two modelling levels and a data pre-treatment stage to identify the stationary of the data. The two modelling levels include three techniques namely conventional PCA, cointegration residuals-based process monitoring and common-trend residuals-based process monitoring. The sequence of the procedure is illustrated in Figure 4.6. Level assignment helps to build a monitoring chart that allows the monitoring of all of the process using a single chart. More precisely; firstly, conventional PCA is a well-known and useful tool for process monitoring, which can deal with stationary and IID variables. However, PCA cannot be used with time-varying and nonstationary data, which has been discussed in several studies (Ku, Storer and Georgakis, 1995; Chen, Kruger and Leung, 2009; Li, Qin and Yuan, 2014; Shang *et al.*, 2017). Secondly, cointegration residuals-based process monitoring is a useful tool when nonstationary variables are considered. However, there can be issues with cointegration analysis such as a low or zero cointegration rank, when the nonstationary variables exhibit high-level nonstationarity, which limits the effectiveness of cointegration analysis.

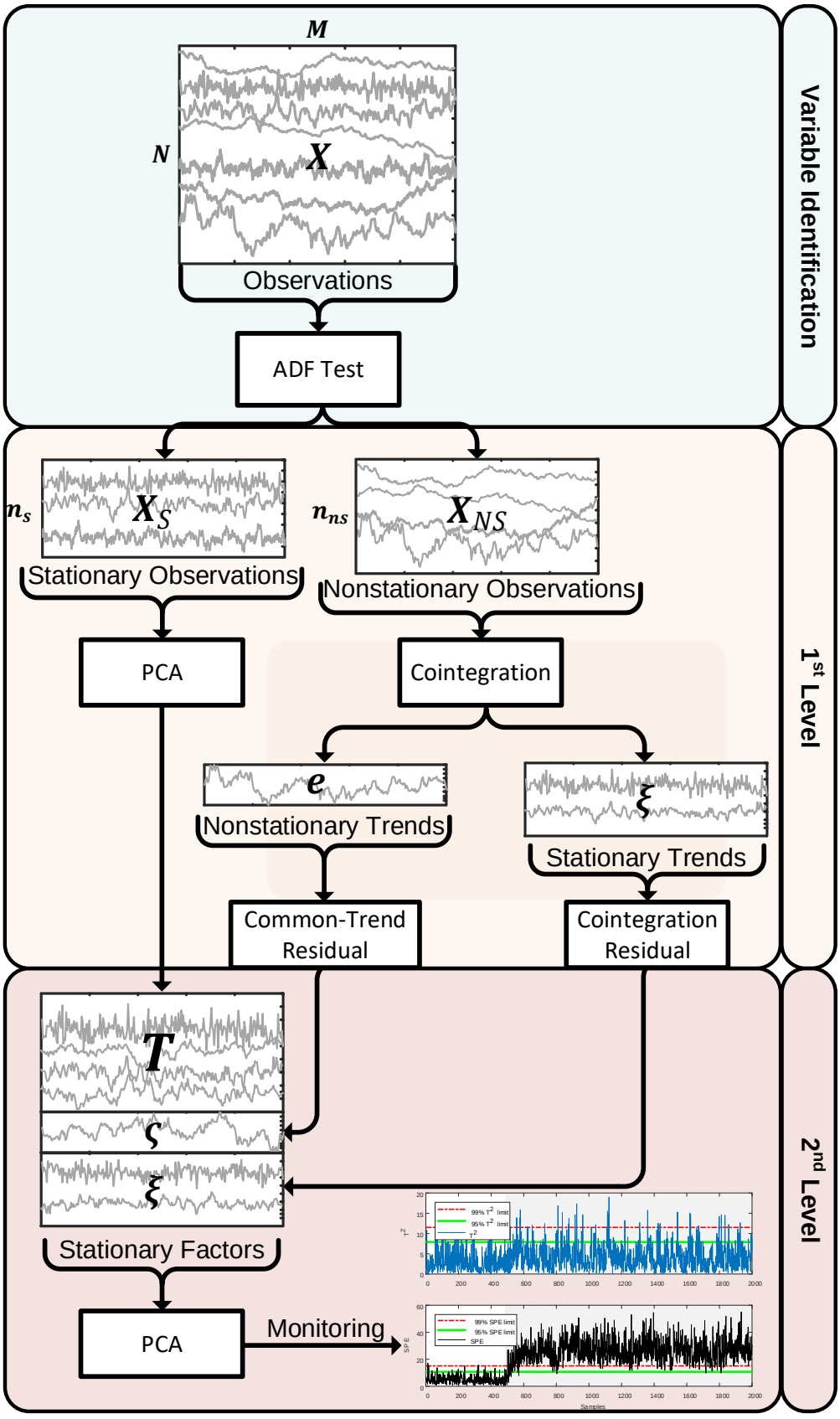

Figure 4.6: Illustration of process monitoring method based on a multi-level multi-factor model.

Moreover, a cointegration relationship might not exist for high cointegration degrees such as $I(2)$ in combination with $I(1)$ (Lin, Kruger and Chen, 2017). It increases the possibility of having common-trends in the cointegration analysis, which cannot be used in cointegration residuals. Lastly, common-trends residuals are helpful whereby the cointegration degree is high and cointegration analysis is enough to tackle all nonstationary variables. However, no studies have demonstrated that common-trends always exist or that common-trend residuals-based approaches will be superior to cointegration residuals-based approaches for process monitoring.

Prior to building the multi-level multi-factor model, the ADF test is applied to the variables ($\boldsymbol{X}$) collected from a complex industrial process. The ADF test searches for the unit root of the following regression model:

$$\Delta \boldsymbol{x}_t = \theta^* \boldsymbol{x}_{t-1} + \sum_{i=1}^{p-1} \theta_i \Delta \boldsymbol{x}_{t-i} + \boldsymbol{\epsilon}_t \tag{4.22}$$

where p is the order of the AR model, $\theta^* = \left(\phi_1 + \cdots + \phi_p\right) - 1$, $\theta_i = (\phi_1 + \cdots + \phi_i) - 1$ for $i = 1, \ldots, p-1$, $\boldsymbol{\epsilon}_t \sim N(0, \sigma^2)$, for the AR parameter ϕ and $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$. The null hypothesis of the existence of the unit root where $\theta^* = 1$ compares the test against the alternative hypothesis with the presence of the stationarity under the condition $\theta^* < 1$. This will result in identification of the variable as nonstationary or stationary, and so is called variable identification. Further information about unit root tests is given in Section 3.3.

The first level of the technique starts with cointegration analysis on the nonstationary variables $\left(\boldsymbol{X}_{NS} = \left[\boldsymbol{x}_{NS}^{(1)}, \ldots, \boldsymbol{x}_{NS}^{(n_{ns})}\right]^T \in \mathbb{R}^{n_{ns} \times M}\right)$. Cointegration residuals from the long-run equilibrium can be generalised as follows:

$$\boldsymbol{\xi} = \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{X}_{NS} \tag{4.23}$$

where $\boldsymbol{\beta}$ is the cointegration matrix and if it has rank R such that $\boldsymbol{\beta} \in \mathbb{R}^{n_{ns} \times R} = [\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(R)}]$, it represents R linearly independent stationary vectors from the cointegration analysis; therefore $\boldsymbol{\xi} \in \mathbb{R}^{R \times M}$ describes stationary factors that can be modelled using PCA in the 2nd level. Estimation of $\boldsymbol{\beta}$ can be done by following Equation (3.27) in Section 3.4.2.

Another stationary factor produced in the 1st level by cointegration analysis comes from the common-trend residuals which can be represented as

$$\boldsymbol{e} = \boldsymbol{\beta}_{\perp}\boldsymbol{X}_{NS} \tag{4.24}$$

where $\boldsymbol{e}$ represents the nonstationary residuals. The residuals represent the stochastic trends caused by nonstationarity but were not involved in the residuals for cointegration analysis. As $\boldsymbol{\beta}_{\perp}$ is the complement of $\boldsymbol{\beta}$, the rank is dependent on $\boldsymbol{\beta}$. The stationary factors from the common-trend model ($\boldsymbol{\varsigma}$) can be determined after auto regression whitens them as represented in the previous section.

The last stationary factor produced in the 1st level arises from conventional PCA. The stationary factors or t-scores ($\boldsymbol{T}_S$) can be derived by the PCA model as follows

$$\boldsymbol{X}_S = \sum_{r=1}^{R_S} \boldsymbol{p}_r \boldsymbol{t}_r^T + \boldsymbol{E}_S = \boldsymbol{P}_S \boldsymbol{T}_S^T + \boldsymbol{E}_S \tag{4.25}$$

where $\boldsymbol{X}_S \in \mathbb{R}^{n_s \times M}$, n_s is the number of stationary variables determined by the ADF test, and $\boldsymbol{P}_S$ is the loading matrix with $\boldsymbol{E}_S$ being the estimation error matrix. R_S is the maximum number of principal components under the condition of $R_S = \min(n_s, M)$.

The second and final level of the multi-level multi-factor model combines the factors from the 3 sub-models from 1st level, which are all stationary and hence can be analysed by PCA. The stationary factors can be described as follows:

$$\widetilde{\boldsymbol{X}} = [\boldsymbol{\xi}, \boldsymbol{\varsigma}, \boldsymbol{T}_s] \tag{4.26}$$

Similar to the PCA model described previously, it uses the same SVD or NIPALS decomposition to give:

$$\widetilde{\boldsymbol{X}} = \sum_{r=1}^{R_{Sec}} \boldsymbol{p}_r \boldsymbol{t}_r^T + \widetilde{\boldsymbol{E}} = \widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{T}}^T + \widetilde{\boldsymbol{E}} \tag{4.27}$$

Here, the number of factors involved in $\widetilde{\boldsymbol{X}}$ varies depending on the percentage of the explained variance chosen for the PCA model and the rank of the cointegration analysis in the 1st level.

The T^2 performance monitoring metric, which covers all process variables can be stated as follows:

$$\boldsymbol{T}^2 = \widetilde{\boldsymbol{P}}^T \widetilde{\boldsymbol{X}} \boldsymbol{\Lambda}^{-1} \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{P}} \tag{4.28}$$

where $\boldsymbol{\Lambda} = \boldsymbol{T}^T \boldsymbol{T}/(M-1)$ is the sample covariance matrix under the condition that the process is normal and the data has a multivariate normal distribution. Hence, T^2 becomes related to a F distribution:

$$\frac{R_{Sec}(M-1)}{(M-R_{Sec})} F_{R,(M-R_{Sec});\alpha} \tag{4.29}$$

where $F_{R,(M-R);\alpha}$ is the F distribution with R and $(M-R)$ degrees of freedom and α is the significance level of the F distribution. The T^2 metric measures the distance of the sample to the origin in PC subspace which can be explained by the model parameters. On the other hand, the SPE metric measures variability which breaks the typical process correlation indicated by an abnormal situation based on a squared difference between the observed and the predicted values from the normal representation and can be defined as follows:

$$Q_i = SPE_i = (x_i - \hat{x}_i)^2 = \left(x_i - \widetilde{\boldsymbol{P}} \tilde{t}_i^T\right)^2 \tag{4.30}$$

where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ are the original and estimated variable vectors for the i^{th} sample, and $\tilde{\boldsymbol{t}}_i$ is the score vector for the i^{th} sample of $\widetilde{\boldsymbol{T}}$. The upper control limits for the SPE under a significance level α can be calculated (Jackson and Mudholkar, 1979) via:

$$Q_i = \theta_1 \left(\frac{z_{(1-\alpha)}\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \tag{4.31}$$

where $\theta_i = \sum_{k=R+1}^{N} \lambda_k^i, i = 1,2,3$ and $h_0 = 1 - (2\theta_1\theta_3)/(3\theta_2^2)$. Here, λ_k^i is the eigenvalue from residuals, i and k refer indexes of the power and largest eigen order, $z_{(1-\alpha)}$ is the standard normal deviate or z-score regarding $(1-\alpha)$ percentile. This limit is derived under the condition that a sample vector $\boldsymbol{x}$ from $\widetilde{\boldsymbol{X}}$ follows a multivariate normal distribution.

The first part of the multi-level multi-factor model requires offline training of the model with the data that was gathered under normal steady-state conditions. The critical steps of the proposed multi-level multi-factor model are listed in Table 4.3.

Table 4.3: Offline training of the multi-level multi-factor model.

1: Identify the stationarity and nonstationarity of each variables using the ADF test.
2: Set up nonstationary ($\boldsymbol{X}_{NS}$) and stationary ($\boldsymbol{X}_{S}$) variables matrices.
3: Use Equation (3.25) to obtain $\boldsymbol{\beta}$.
4: Calculate $\boldsymbol{\xi}$ via Equation (4.13)
5: Find $\boldsymbol{\beta}_{\perp}$ and calculate $\boldsymbol{e}$ via Equation (4.16) and then whiten $\boldsymbol{e}$ via Equation (4.18) to find $\boldsymbol{\varsigma}$.
6: Use Appendix-A on ($\boldsymbol{X}_{S}$) to find $\boldsymbol{T}_{s}$ via Equation (4.25).
7: Set up stationary factors ($\widetilde{\boldsymbol{X}}$) via Equation (4.26)
8: Use Appendix-A on ($\widetilde{\boldsymbol{X}}$) to find $\widetilde{\boldsymbol{T}}$ via Equation (4.27).
9: Calculate the upper control limits.

Following the determination of the model and the upper control limits, online process monitoring can then be conducted as listed in Table 4.4.

Table 4.4: Online process monitoring using the multi-level multi-factor model.

1: For each sampling point (t).
2: Calculate $\boldsymbol{\xi}_{t}$ via Equation (4.13)
3: Calculate $\boldsymbol{\varsigma}_{t}$ via Equation (4.16) then Equation (4.18)
4: Calculate $\boldsymbol{T}_{s_t}$ via Equation (4.25).
5: Set up stationary factors ($\widetilde{\boldsymbol{X}}_{t}$) via Equation (4.26)
6: Calculate $\widetilde{\boldsymbol{T}}_{t}$ via Equation (4.27).
7: Calculate T^2 via Equation (4.28) and SPE via Equation (4.30)
8: If the sampling point exceeds the defined upper limits, then raise an alarm
9: End if it is the last point

4.4 Application to Continuous Stirred Tank Heater

The purpose of the use of the CSTH is to evaluate the effectiveness of the models on different types of errors. Here, closed-loop control is implemented in the standard form of a proportional plus integral (PI) controller for both level and temperature control. Even though the initial variables are described in the paper by Thornhill et al. (Thornhill, Patwardhan and Shah, 2008) via the Simulink files provided, any changes

made to the Simulink file can affect the steady-state of the process. This can be solved by using signal limits for integral windups and then assigning the final states that describe the new steady state as the initial tuning of the process. To do this, the final states should be saved using the configuration parameters under the simulation toolbar.

Two different types of error function are defined and applied for testing. The first is a step function used on the inlet hot water temperature, and the second is a ramp function applied to the cold water valve. In this thesis, all simulation and process monitoring studies have been performed using MATLAB® R2017b (Mathworks, Natwick, USA) on an Intel® Core™ i7-6700HQ CPU @2.60 GHz in conjunction with the MATLAB® statistics and machine learning toolbox and the econometrics toolbox.

4.4.1 Model Training

A training data set was collected for steady-state operation of the process under normal operating conditions with a 1 second sampling time for 2000 samples. Even though 2000 samples is quite a short period of time for operation of the heat exchange system, it is sufficient to show the differences in the fault detection capabilities of the different methods. In the literature, there is no exact definition for the percentage of the explained variance that must be used in a PCA model. For example, Russell et al. used between $11-20$ PCs (out of a maximum of 42 PCs) in a PCA model to describe $40-72\%$ of the variance in the data from a simulation of the TEP (Russell, Chiang and Braatz, 2000b). Alternatively, Rato et al. used 17 principal components to describe $68-70\%$ of the variance in the process (Rato and Reis, 2013c).

Further information about the selection of PCs for all given models are given in Appendix-C. A PCA model was trained with 2 PCs, which explained 54.5% of the total variance; the first five PCs explained $32.2, 22.3, 19.7, 15.1$ and 5.1% of the variance in the data. The T^2 and SPE charts for training performance are illustrated in Figure 4.7 and tabulated in Table 4.5.

The linear relationship determination algorithm for DPCA suggested that $l = 2$ lag shifts should be used, with regards to the algorithm in Table 4.2. A linear relationship determination found 9 and 1 new linear relationships between the PCs for each lagged additional training data set, which states the autocorrelations in each added data matrix as a time lag shift. The DPCA model was trained with 3 PCs, which explained 42.9% of the total variance; the first 5 PCs out of a total of 42 explained $18.9, 12.8, 11.3, 11.1$

and 10.4% of the variance in the data. In the training of the DPCA model, the performance of the model with 3 PCs explained 42.9% and 4 PCs explained 54% were arguably closed to each other as explained in Appendix-C. The model with 3 PCs is selected to have a model that is close to 45%, which is the percentage that have been used in most of the models in this study. The T^2 and SPE charts for training performance are given in Figure 4.8 and tabulated in Table 4.5. It can be concluded from Table 4.5, that the training performance is acceptable in terms of the type-1 error rate (below 1%) for both PCA-based approaches in Figure 4.7 and Figure 4.8.

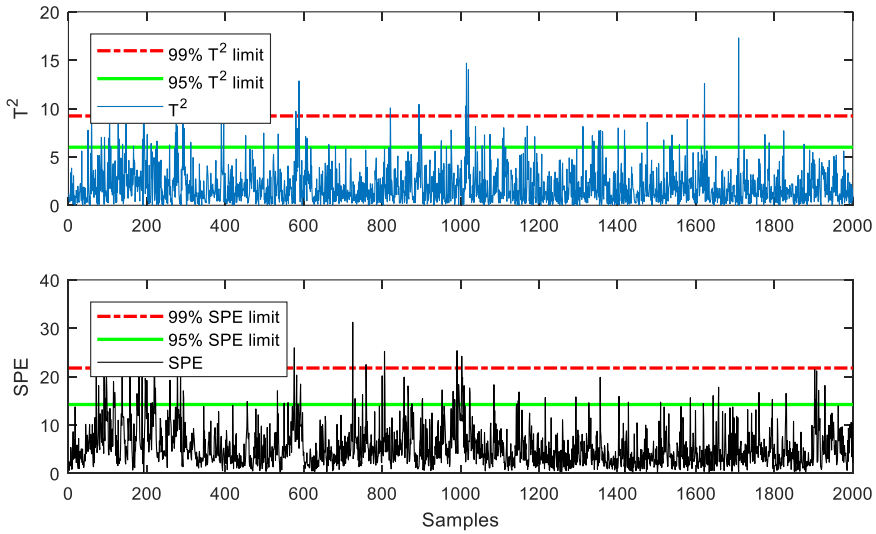

Figure 4.7: T^2 and SPE metrics for a PCA model built using training data from the CSTH.

Table 4.5: A comparison of the offline training performance of different models for the CSTH process.

Metric Name (significance level)	**Type I error (%)**				
	PCA	DPCA	Cointegration Residuals	Common-trend Residuals	Multi-level Multi-factor
SPE(1%)	**0.6**	0.8	–	–	**0.6**
T^2(1%)	0.7	0.5	0.8	**0.4**	0.8

Process monitoring methods based on cointegration residuals requires variable identification prior to performing the training procedure. Table 4.6 lists the identity of the variables based on a 5% significance level and an AR order 12. 9 of the 14 variables were assigned as nonstationary and hence, require cointegration residuals-based process monitoring methods.

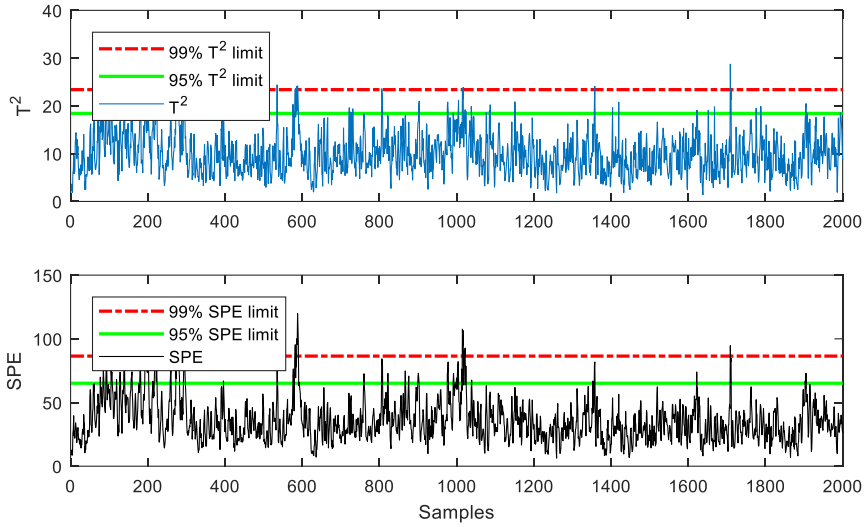

Figure 4.8: T^2 and SPE metrics for a DPCA model built using training data from the CSTH.

Table 4.6: Result of the ADF test applied to CSTH training data.

Variable No	AR Order	t Statistic	Critical Value	Test Result
1	12	−0.1188	−1.9416	Nonstationary
2	12	−0.4772	−1.9416	Nonstationary
3	12	−3.1265	−1.9416	Stationary
4	12	−4.8339	−1.9416	Stationary
5	12	−0.0250	−1.9416	Nonstationary
6	12	−0.0738	−1.9416	Nonstationary
7	12	−0.0785	−1.9416	Nonstationary
8	12	−0.0825	−1.9416	Nonstationary
9	12	−0.02955	−1.9416	Nonstationary
10	12	−4.1265	−1.9416	Stationary
11	12	−3.3779	−1.9416	Stationary
12	12	−0.9295	−1.9416	Nonstationary
13	12	−5.3781	−1.9416	Stationary
14	12	−0.1089	−1.9416	Nonstationary

The rank (R) for the cointegration matrix was determined to be 5 where the maximum rank is 8. Therefore, $\boldsymbol{\beta}$ was defined as $\mathbb{R}^{9\times5}$ where $\boldsymbol{\beta}_{\perp}$ can be defined as $\mathbb{R}^{9\times4}$ perpendicular to $\boldsymbol{\beta}$. The T^2 chart for training performance is illustrated in Figure 4.9 and tabulated in Table 4.5. Following the determination of $\boldsymbol{\beta}_{\perp}$, the T^2 chart for the training performance of the common-trend residuals can be found in Figure 4.10 and Table 4.5. Compared to PCA-based models, cointegration and common-trend residuals-based models performed similar or better in terms of type-I error rates.

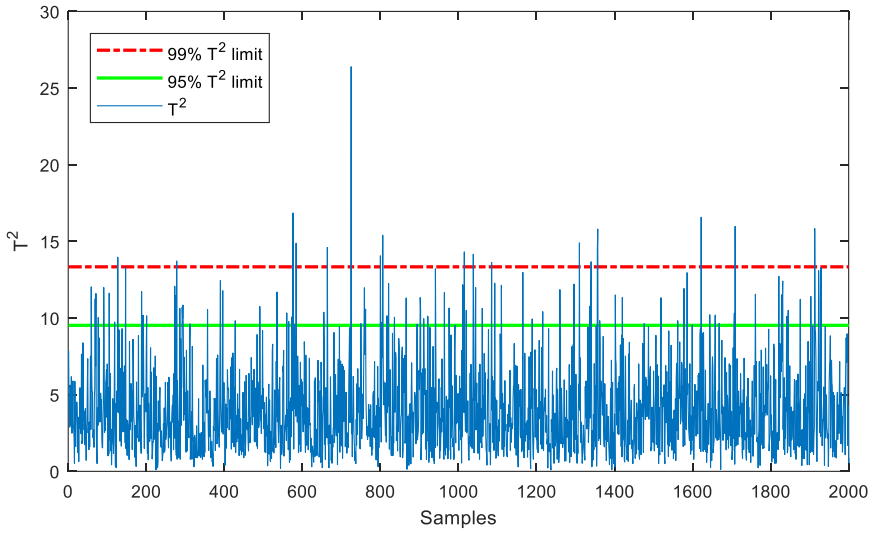

Figure 4.9: T^2 metric for a cointegration residuals model built using the training data from CSTH.

A PCA model built that using stationary variables ($\boldsymbol{X}_S$) was trained with 1 PC that explained 72.2% of the total variance; the first 5 PCs explained 72.2, 25.0, 2.7, 0.001 and 0.001% of the variance in the data. The stationary factors $(\widetilde{\boldsymbol{X}})$ comprise 1 score vector ($\boldsymbol{T}$) where $\boldsymbol{\xi}$ and $\boldsymbol{\varsigma}$ consist of 5 and 4 vectors, respectively.

The 2nd level PCA model was trained with 2 PCs (out of a total of 10), which explained 47.9% of the total variance; the first 5 PCs explained 25.9, 22.0, 16.8, 13.4 and 11.1% of the data. The T^2 and SPE charts for training performance are illustrated in Figure 4.11 and tabulated in Table 4.5. The performance of the multi-level multi-factor model was comparable to that of the cointegration residuals-based model in terms of the type-I error rate; the multi-level multi-factor model utilizes the SPE statistic for monitoring, which is the favoured metric in process monitoring for fault detection.

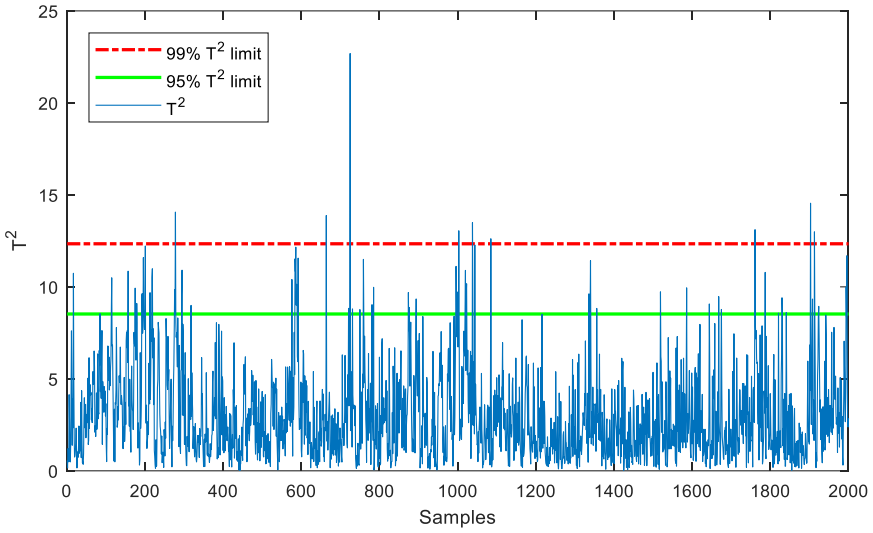

Figure 4.10: T^2 metric for a common-trend residuals model built using the training data from CSTH.

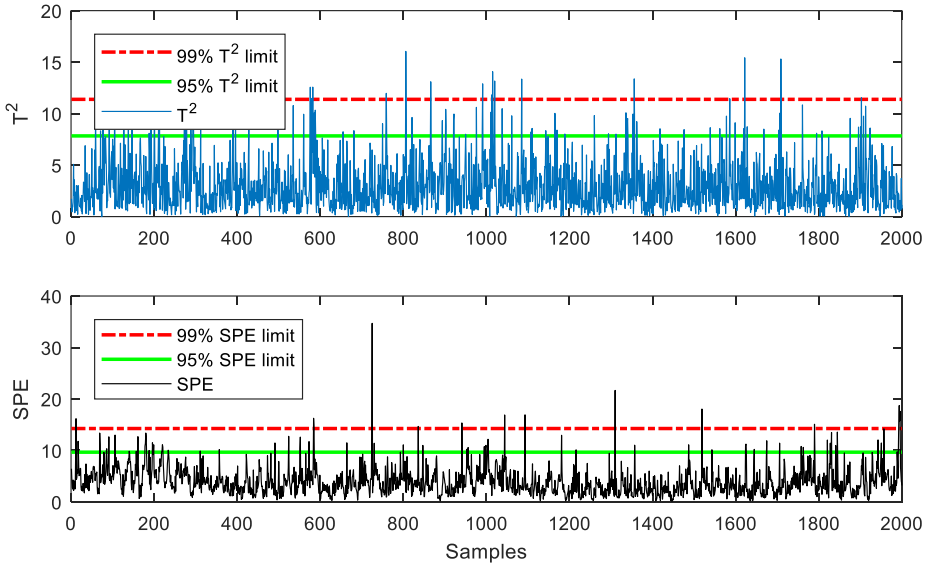

Figure 4.11: T^2 and SPE metrics for the multi-level multi-factor model built using the training data from CSTH.

4.4.2 Model Testing

For the first case, an electrical sensor current fault was applied in the form of a step function at sample number 500. The error increases the hot water temperature of the inlet by $+1$ ℃, where the range of the input varies between $48 - 52$ ℃; thus reflecting a 2% change. Faulty zones are highlighted on the figures with grey shading. It should be noted that the process operates under feedback control which tries to regulate the process deviations from the normal operating conditions. That being the case, the fault is created to test the ability of the monitoring system to detect instant and continuous changes.

For the second case, a ramp function with a slope $10^{-8}\ m^3 s^{-1}$ was defined as a valve malfunction on the flow of the cold water, which is stationary according to Table 4.6. The names of the variables are shown in Table 4.1. A ramp function displays a slow-varying trend, which shows nonstationary characteristics. By using the given models with this example, their performance with a nonstationary fault that is applied on the stationary variable is tested.

The results obtained using a PCA model for a step function type fault are shown in Figure 4.12 and tabulated in Table 4.7. The reasons for not using the T^2 metric for process monitoring, which are discussed in Section 2.5.3, can be seen in this example. A 2% change of 1 ℃ is not sufficient to impact on the model variance as can be seen in Figure 4.12. PCA can detect the abnormality via the SPE at sample number 509 but this comes with some fluctuation around the control limits. Even though PCA can identify the step function type fault, the magnitude of the SPE metric was not enough to exceed the upper control limits for all of the samples which causes type-2 errors.

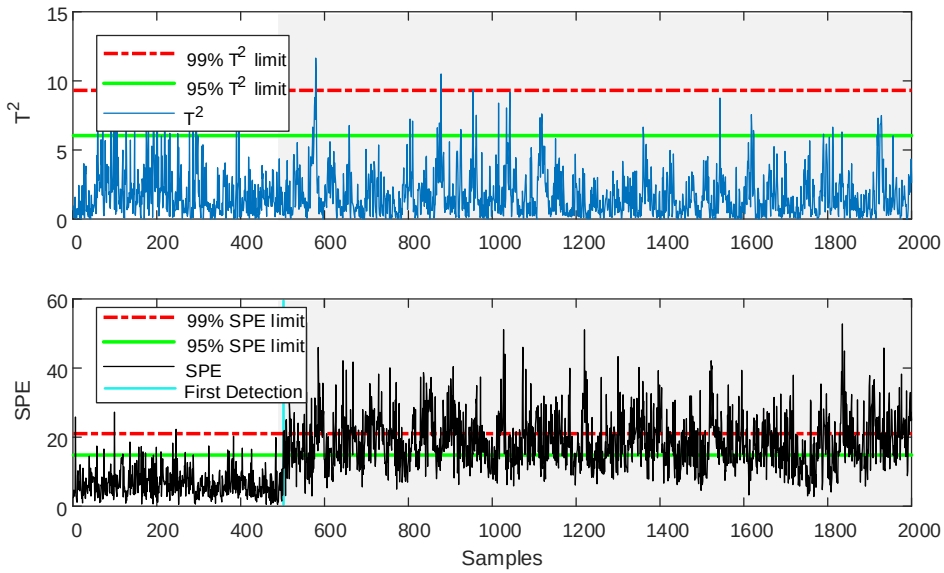

Figure 4.12: T^2 and SPE metrics obtained using PCA with test data from the CSTH exhibiting a step function type fault. The fault is first detected at sample number 509 using the SPE metric (indicated by turquoise vertical line).

The T^2 and SPE metrics for the DPCA model for a step function type fault are illustrated in Figure 4.13 and tabulated in Table 4.7. DPCA is able to detect the fault signature with both T^2 and SPE with some type-2 errors due to the additional time lag shift variables. One way to solve the missed fault signature and reduce the type-2 error rate in the T^2 chart is to increase the percentage variance explained but this comes with

an increase in the type-I errors in the no-fault zones in the control charts for both PCA and DPCA models. Thus, the problems affecting PCA also affect the DPCA.

Table 4.7: Comparison of the online diagnosis performance in terms of error rate (%) of different models for the monitoring of the CSTH process exhibiting step and ramp function type faults.

Case	Metric Name (%)	Error Type	Models				
			PCA	DPCA	Cointegration Residuals	Common -trend Residuals	Multi-level Multi-factor
Step	SPE(1%)	Type-I	0.6	**0**	–	–	**0**
	T^2(1%)		0.1	**0**	0.8	**0**	**0**
	SPE(1%)	Type-II	66.20	81.67	–	–	**5**
	T^2(1%)		99.80	65	50.93	78.80	97
Ramp	SPE(1%)	Type-I	1.2	13.40	–	–	**0.4**
	T^2(1%)		3.80	**2**	1.60	1	2.60
	SPE(1%)	Type-II	13.27	10.40	–	–	**0.8**
	T^2(1%)		29.87	18.80	**0.8**	15.13	3

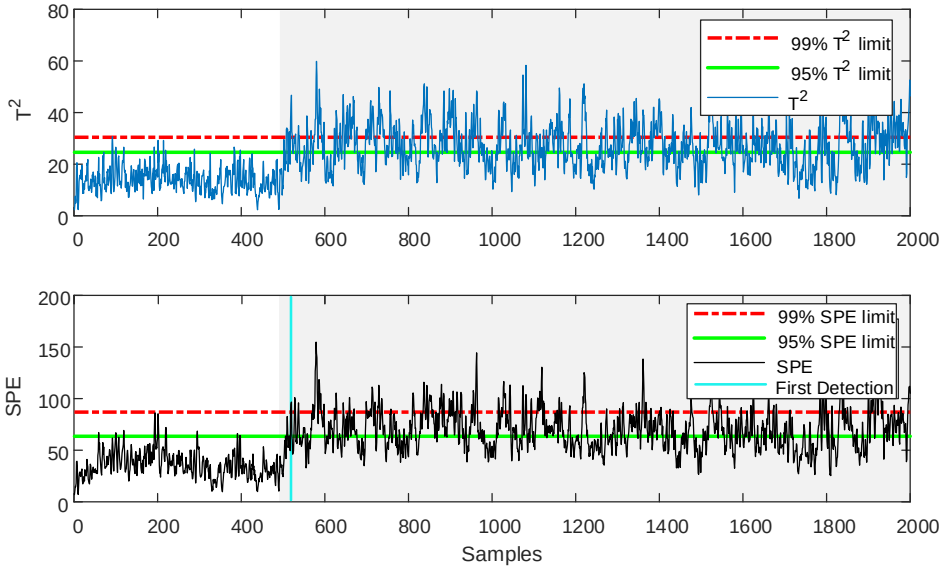

Figure 4.13: T^2 and SPE metrics obtained using DPCA with test data from the CSTH exhibiting a step function type fault. The fault is first detected at sample number 513 using the SPE metric (indicated by turquoise vertical line).

The results obtained for the detection of a step function type fault using cointegration residuals and common-trend residuals are illustrated in Figure 4.14 and Figure 4.15,

respectively, and are tabulated in Table 4.7. The fault is detected at sample number 535 and 550 by the cointegration and common-trend residuals, respectively. The CTSH simulation contains some high-level nonstationary variables that are challenging to model using a cointegration relationship, which can cause reduced rank, but a common-trend model can be used to extract the stationary factors from the unused cointegration vectors. Nevertheless, neither method is sufficient for detection of the fault signature with an acceptable type-II error rate. This might be because the most affected variables from the step function type fault are stationary variables, namely the hot and cold water flows.

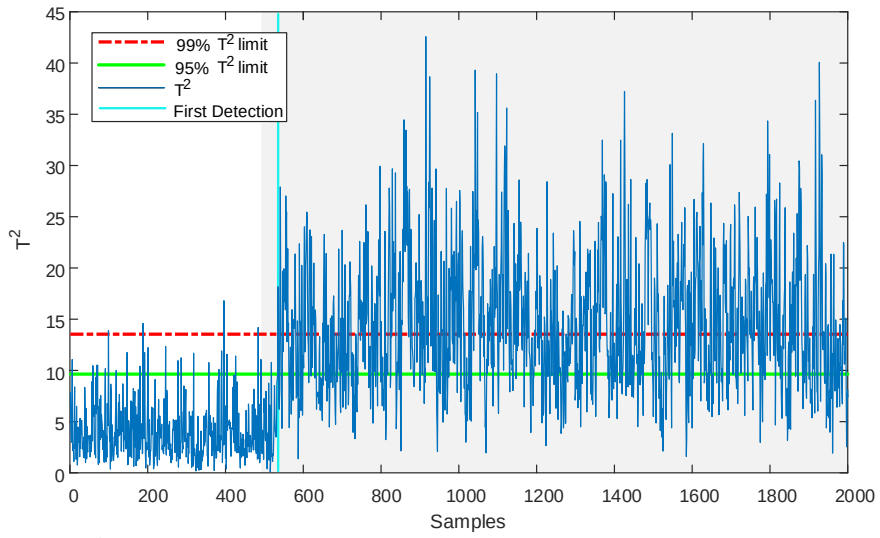

Figure 4.14: T^2 metric obtained using cointegration residuals with test data from the CSTH exhibiting a step function type fault. The fault is first detected at sample number 535 (indicated by turquoise vertical line).

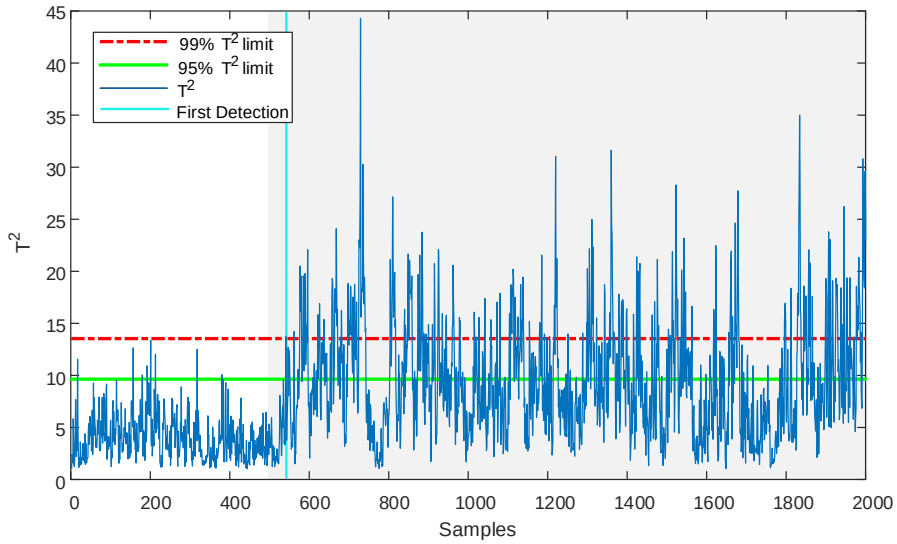

Figure 4.15: T^2 metric obtained using common-trend residuals with test data from the CSTH exhibiting a step function type fault. The fault is first detected at sample number 550 (indicated by turquoise vertical line).

The results obtained for the detection of the step function type fault using the multi-level multi-factor model are shown in Figure 4.16 and tabulated in Table 4.7. The effectiveness of the multi-level multi-factor model for fault detection via the SPE metric demonstrates the superiority of this method with step type function errors. The performance of the multi-level multi-factor model is improved compared to separate cointegration and common-trend residuals-based methods owing to the combination of these methods with PCA, which is applied to the stationary variables where some of the fault resides. As a result, the fault is detected by the multi-level multi-factor model at sample number 503, via the SPE control chart. As expected, the multi-level multi-factor model cannot detect any particular fault via the T^2 chart as the error is not enough to vary above the defined T^2 confidence bounds.

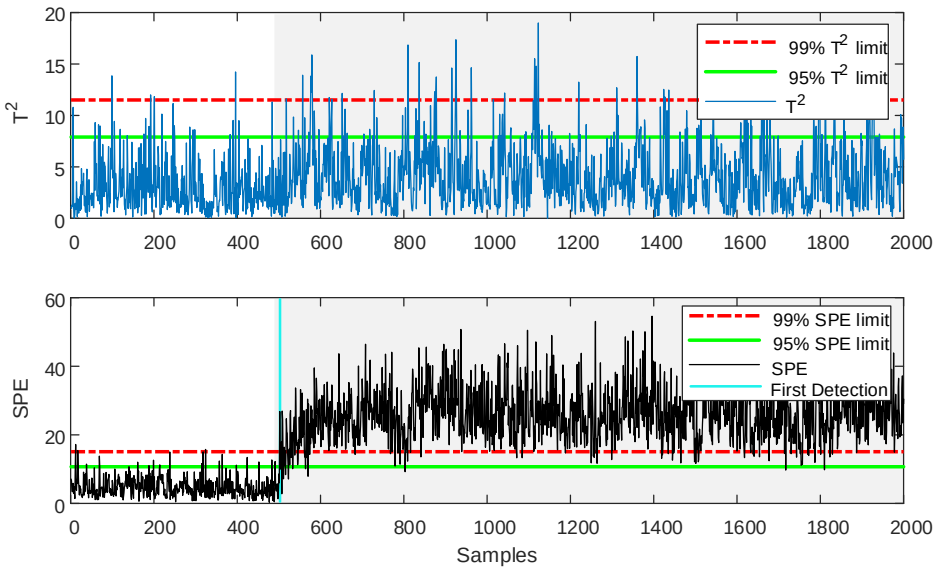

Figure 4.16: T^2 and SPE metrics obtained using the multi-level multi-factor method with test data from the CSTH exhibiting a step function type fault. The fault is first detected at sample number 503 using the SPE metric (indicated by turquoise vertical line).

The results obtained using PCA for the detection of a ramp function type fault are shown in Figure 4.17 and tabulated in Table 4.7. Even though the T^2 metric is able to detect the fault, the time to do so is expected to be later than that for the SPE as the fault is a ramp type function. This is due to the inefficiency of the T^2 metric for monitoring of nonstationary variables as previously discussed. PCA can detect an abnormality via the SPE chart at sample number 590 but this comes with some type-II errors until around sample number 700.

The T^2 and SPE metrics obtained using the DPCA model for detection of a ramp function type fault are illustrated in Figure 4.18. DPCA can detect the fault signature with both T^2 and SPE earlier than PCA owing to its capabilities with additional time lag shifts. The performance of DPCA is tabulated in Table 4.7, where it shows better performance than PCA with regards to type-II errors. However, it causes some normal operation samples to go beyond the control limits, which caused an increase in the type-I error in the SPE control chart.

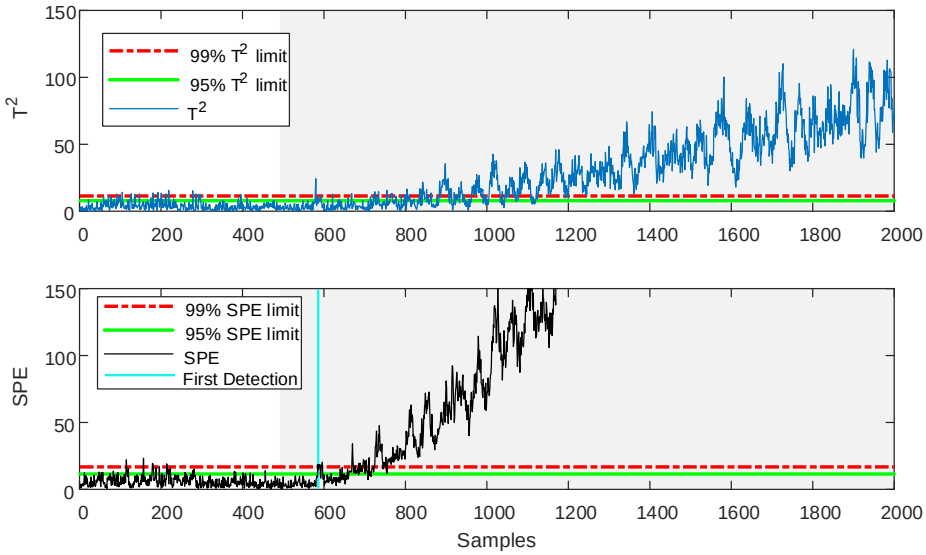

Figure 4.17: T^2 and SPE metrics obtained using PCA with test data from the CSTH exhibiting a ramp function type fault. The fault is first detected at sample number 590 using the SPE metric (indicated by turquoise vertical line).

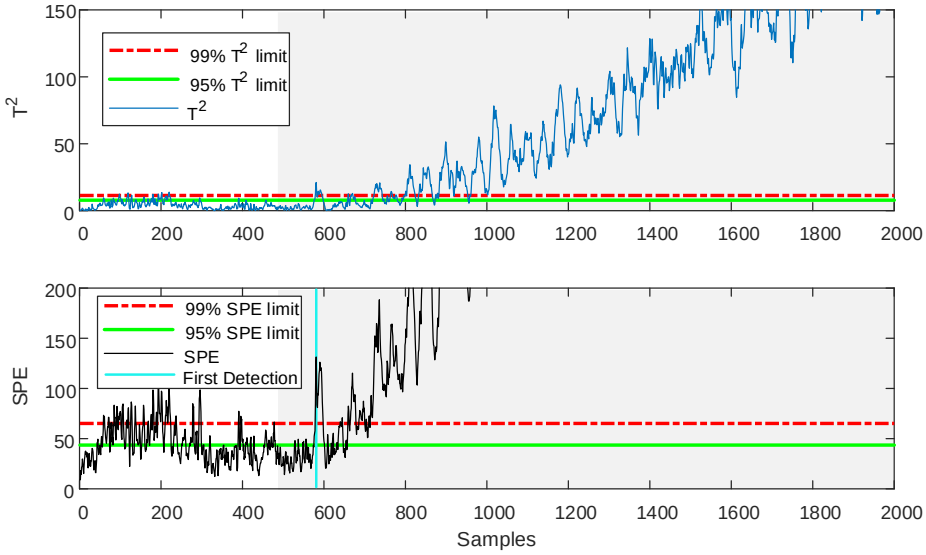

Figure 4.18: T^2 and SPE metrics obtained using DPCA with test data from the CSTH exhibiting a ramp function type fault. The fault is first detected at sample number 587 using the SPE metric (indicated by turquoise vertical line).

The results obtained for the detection of the ramp function type fault using the cointegration residuals and common-trend residuals are illustrated in Figure 4.19 and Figure 4.20, respectively, and are tabulated in Table 4.7. The fault is first detected at sample number 513 and 580 for cointegration and common-trend residuals, respectively. Cointegration residuals are able to detect the ramp type of error in a short period of time whereas common-trend residuals were not. This indicates that most of the nonstationarity arising from the ramp function is captured via cointegration analysis. It is noted that the common-trend model can still detect the fault.

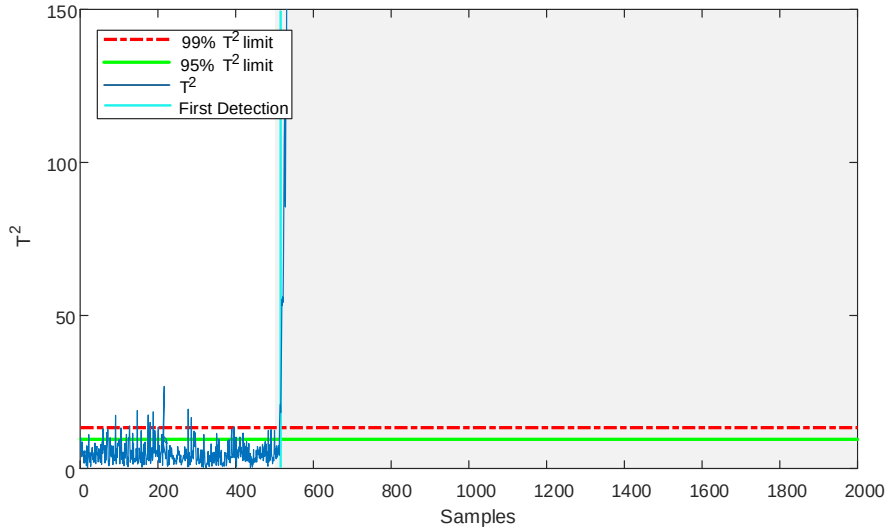

Figure 4.19: T^2 metric obtained using cointegration residuals with test data from the CSTH exhibiting a ramp function type fault. The fault is first detected at sample number 513 (indicated by turquoise vertical line).

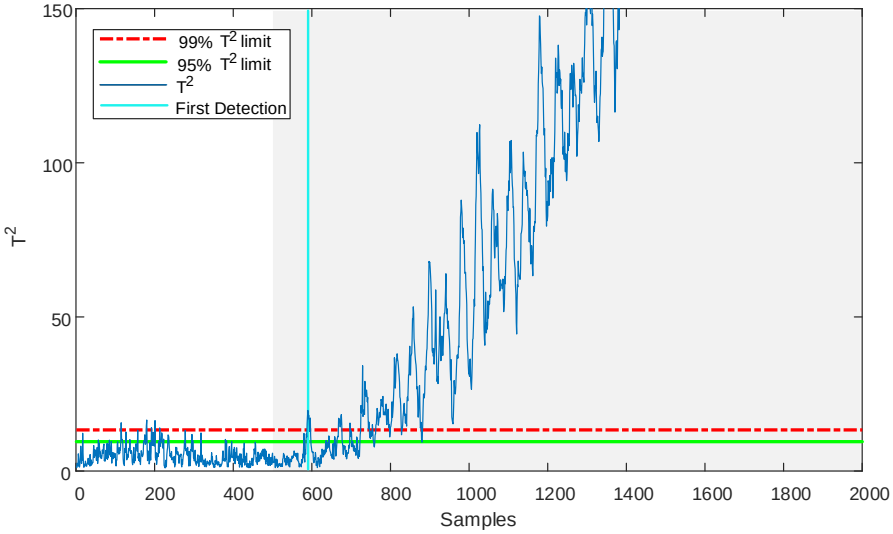

Figure 4.20: T^2 metric obtained using common-trend residuals with test data from the CSTH exhibiting a ramp function type fault. The fault is first detected at sample number 580 (indicated by turquoise vertical line).

Finally, Figure 4.21 shows the T^2 and SPE metrics obtained using the multi-level multi-factor model of a ramp function type fault. Here, a clear advantage of cointegration residuals is also transferred to the multi-level multi-factor model. The detection performance is tabulated in Table 4.7 where the results obtained for the SPE are the same as those obtained for the T^2 metric with cointegration residuals.

As it can be seen from Table 4.7, the effectiveness of the multi-level multi-factor model is proven on both step and ramp type function faults as it models both stationary and nonstationary characteristics in the data. It is worth noting that cointegration residuals and common-trend residuals-based approaches cannot be compared with other methods via the SPE metric because only a modified T^2 metric is available for these models. However, the effectiveness of the multi-level multi-factor model compared to other methods can be demonstrated if the best detection rates by either the T^2 and SPE metrics are used as a comparison.

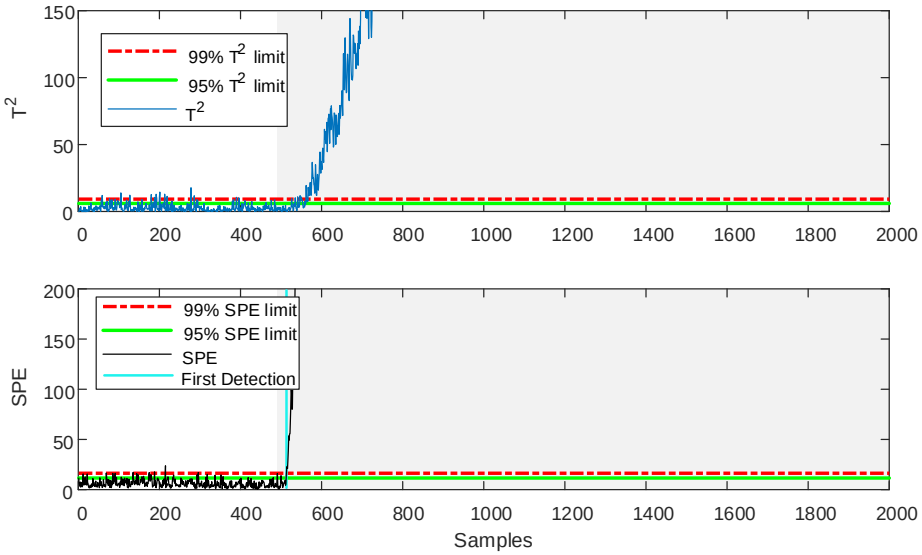

Figure 4.21: T^2 and SPE metrics obtained using the multi-level multi-factor model with test data from the CSTH exhibiting a ramp function type fault. The fault is first detected at sample number 543 using the SPE metric (indicated by turquoise vertical line).

4.5 Conclusions

This chapter has explored a new multi-level multi-factor process monitoring method for use with complex, continuous industrial processes. The method requires a data pre-treatment step, which is then followed by two-levels of modelling. The procedure requires the identification of the stationary and nonstationary variables in the data set. Following this, the data are then divided into two groups: stationary data are handled

by the PCA model and nonstationary data are handled by cointegration analysis for the determination of cointegration and common-trend residuals, which are stationary. Such models transform the data into stationary factors, which can then be modelled by PCA in the 2nd and final level of the multi-level multi-factor model.

The multi-level multi-factor model solves 3 problems: firstly, a need for a model that can be applied to data exhibiting a wide range of characteristics, not only nonstationary variables like cointegration residuals-based monitoring but also stationary variables. Secondly, the problem in modelling nonstationary variables with a high degree of nonstationarity is addressed by using both cointegration and common-trend residuals at the same time. Thirdly and lastly, a confusion that comes with multiple monitoring charts is removed through the use of a single control chart based on T^2 and SPE performance metrics. This is the first report of a method based on cointegration analysis that can be applied to both nonstationary and stationary data, and its use for the monitoring of continuous processes has been successfully demonstrated using a CSTH simulation. The effectiveness of the multi-level multi-factor model was evaluated with two different types of faults. The proposed multi-level multi-factor model showed its superiority against PCA, DPCA, cointegration and common-trend residuals-based monitoring methods.

5. MONITORING BATCH PROCESSES USING COINTEGRATION-BASED APPROACHES

5.1 Overview

Unlike continuous processes, batch and semi-batch processes are characterised by a prescribed processing of materials for a finite duration of time. Even though feedback control is applicable for some variables, it often cannot be applied to correct disturbances in a timely manner during a batch. Therefore, techniques that can provide insights into variables and their relationships through their statistical properties may improve the product quality and minimize batch to batch differences. MSPC techniques such as PCA and PLS have been applied successfully to batch processes through multiway PCA (Nomikos and MacGregor, 1994; Nomikos and MacGregor, 1995b). Multiway PCA considers the entire batch as a single object due to the conversion of the 3D data cube to a 2D representation by unfolding methods. This can cause modelling problems for multi-phase batch processes where the characteristics of the process variables change from phase to phase. Since each phase has its own underlying characteristics, the variables can exhibit significantly different behaviours over the duration of the batch. Multi-model PCA approaches have been proposed to cope with these differences. For example, such approaches were used to monitor two phases of an exothermic batch chemical reactor (Kosanovich *et al.*, 1994; Dong and McAvoy, 1995). To extend the performance of multi-phase modelling, several methods have been developed and used to handle different problems in batch process monitoring such as phase transitions and variable selection for each phase (Kourti, 2003; Zhao et al., 2007; Ng and Srinivasan, 2009).

Although several techniques have been proposed, most of the existing methods assume that all statistical characteristics of the data, gathered from the normal operating conditions are sufficiently covered by the models. However, in real-life situations, the nonstationarity issue widely exists in complex industrial processes (Chen, Kruger and Leung, 2009; Sun, Zhang, Zhao and Sun, 2017). Cointegration approaches have proved their effectiveness in continuous process in several studies (Chen, Kruger and Leung, 2009; Lin, Kruger and Chen, 2017). One of the issues of cointegration approaches is that they are applicable only to nonstationary variables whereas complex industrial processes contain both stationary and nonstationary variables. In a recent

study of batch process monitoring, a combination of cointegration residuals and a PCA was suggested (Zhang, Zhao and Gao, 2019). However, the presence of high-level nonstationary variables within the cointegration model can affect the rank of the cointegration matrix due to impractical modelling of the cointegration relationship established. Therefore, cointegration analysis can give rise to a cointegration matrix with low or zero rank owing to high-level nonstationarity. This can, however, be solved by using common-trend residuals-based monitoring as was demonstrated in Chapter 4.

In this chapter, an industrial penicillin simulation is introduced and, used to compare different performance monitoring approaches: multi-PCA, a multi-level model and the multi-level multi-factor model. The different monitoring approaches are applied to 2 case studies to evaluate their performance for a complex industrial batch process exhibiting different kinds of faults.

5.2 Introducing the Industrial Penicillin Simulator

Filamentous micro-organisms are the primary source of commercial quantities of secondary metabolites. In penicillin production, which is the target product in this case, the endpoint is defined by the maximum yield. It is a billion-dollar industry which was pioneered through the development of deep-tank fermentation in the 1940s (Goldrick *et al.*, 2015). There have been research activities conducted with laboratory-scale equipment to develop mathematical models. The PenSim simulator is a well-known benchmark fed-batch process for the replication of penicillin production (Birol, Ündey and Çinar, 2002). However, such a simple model structure is not sufficient to capture complex process dynamics even though several process monitoring studies have been carried out using this simulation. Recently, a more realistic simulation of an industrial-scale penicillin simulator (Goldrick, 2015), illustrated in Figure 5.1, was validated using real data from a $100{,}000\ L$ fed-batch process (Goldrick *et al.*, 2015). It provides a realistic simulation of faults with the faults introduced directly affecting the process states in comparison with the PenSim simulator (Birol, Ündey and Çinar, 2002).

The simulation considers the growth, morphology, metabolic production and degeneration of the biomass during a submerged P. chrysogenum fermentation. The simulation divides the internal structure of the biomass, or hyphae, into four separate regions: actively growing region (R_0), non-growing region (R_1), degenerated region

(R_3), which is formed through vacuolation, and autolysed region (R_4). The vacuoles are defined as the vacuole region (R_2) (Goldrick *et al.*, 2015).

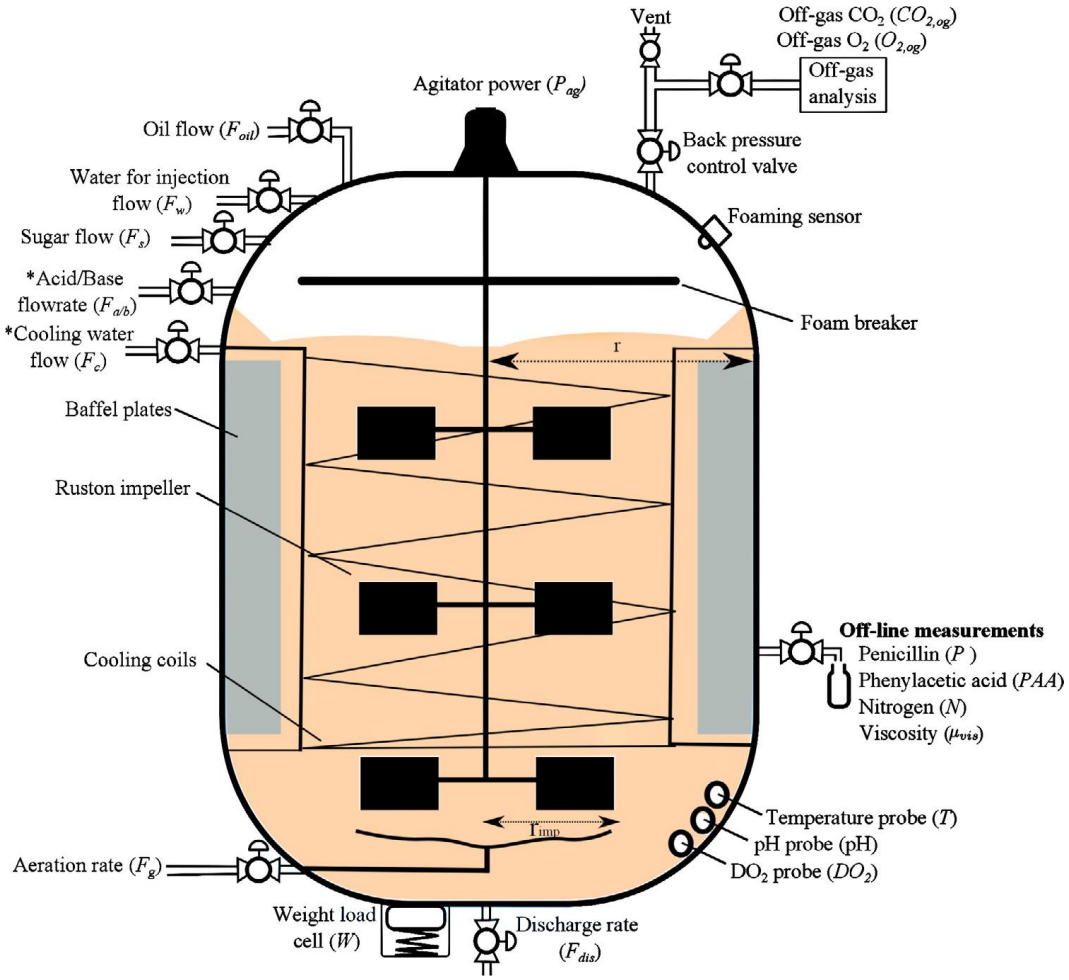

Figure 5.1: Schematic of a bioreactor with process inputs and outputs (Goldrick *et al.*, 2015).

The total biomass (X_{Bio}) is given by:

$$X_{Bio} = R_0 + R_1 + R_3 + R_4 \tag{5.1}$$

with the growing regions (R_0) given by:

$$\frac{dR_0}{dt} = r_b - r_{diff} - \frac{F_{in}R_0}{V} \tag{5.2}$$

the non-growing regions (R_1) given by:

$$\frac{dR_1}{dt} = r_e - r_b + r_{diff} - r_{deg} - \frac{F_{in}R_1}{V} \tag{5.3}$$

the degenerated regions (R_3) given by:

$$\frac{dR_3}{dt} = r_{deg} - r_a - \frac{F_{in}R_3}{V} \tag{5.4}$$

the autolysed regions (R_4) given by:

$$\frac{dR_4}{dt} = r_a - \frac{F_{in}R_4}{V} \tag{5.5}$$

and the product formation or penicillin yield (P) given by:

$$\frac{dP}{dt} = r_P - r_h - \frac{F_{in}P}{V} \tag{5.6}$$

where $r_{b,diff,e,deg,a,P,h,m}$ is the rate of branching, differentiation, extension, degeneration, autolysis, product formation, product hydrolysis and maintenance, respectively. F_{in} represents all the process inputs. The fermenter volume (V) is calculated by:

$$\frac{dV}{dt} = F_S + F_{oil} + F_{PAA} + F_a + F_b + F_w - F_{evp} - F_{dis} \tag{5.7}$$

where F_{PAA} is the flow rate of the phenyl acetic acid and F_A and F_b are the flow rates of the acid and base, respectively. F_w is the flow rate of the water for injection which is typically used to reduce the broth viscosity. F_{evp} is the evaporation rate of the fermenter and F_{dis} is the volume discharged from the vessel during production to maintain the volume within its maximum working capacity. F_s and F_{oil} represents the sugar and soybean oil feed rate. The substrate consumption(sub) is given by:

$$\frac{dsub}{dt} = -Y_{s/X}r_e + Y_{s/X}r_b - m_s r_m - Y_{s/P}r_p + \frac{F_s c_s}{V} + \frac{F_{oil}c_{oil}}{V} - \frac{F_{in}sub}{V} \tag{5.8}$$

The batch time is represented by t. $Y_{s/X}$ and $Y_{s/P}$ represents the substrate yield coefficients of biomass and penicillin, respectively, and m_s is the substrate maintenance term. Dissolved oxygen (DO_2) is a key macro-nutrient used by the micro-organism for growth, maintenance and metabolic production:

$$\frac{dDO_2}{dt} = -\mu_x X Y_{O_2/X} - \mu_P P Y_{O_2/P} - m_{O_2}X + k_L a(DO_2^* - DO_2) - \frac{DO_2 dV}{V d_t} \tag{5.9}$$

The first three terms represent the oxygen uptake rate (OUR) which accounts for oxygen being consumed during biomass growth ($\mu_X X$), maintenance $\left(m_{O_2} X\right)$ and penicillin production ($\mu_P P$). $Y_{O_2/X}$and $Y_{O_2/P}$ are the oxygen yield coefficients for biomass and penicillin, respectively. μ_X represents the rate of change of the growing, non-growing, degenerated and autolysed regions. Oxygen transfer rate is the product of the volumetric mass transfer coefficient ($k_L a$) and the difference between the dissolved oxygen concentration (DO_2) and the oxygen saturation concentration (DO_2^*). Dissolved carbon dioxide ($CO_{2,L}$) is an important and often overlooked variable in fermentation modelling with accumulation of $CO_{2,L}$ in the broth reported to be detrimental to cell growth and productivity:

$$\frac{dCO_{2,L}}{dt} = \delta_{c/o} k_L a(CO_{2,L}^* - CO_{2,L}) - \frac{CO_{2,L} dV}{V d_t} \tag{5.10}$$

where $\delta_{c/o}$ is the ratio of the carbon dioxide to oxygen mass transfer coefficients. The second most abundant nutrient in fermentation media is generally the nitrogen source as it is critical for growth and required for metabolic production of product. The primary source of nitrogen for each batch is generally contained in the starting media and consumed throughout the batch. This consumption is modelled using a material balance, which considered the nitrogen consumed during biomass growth and maintenance and that utilised for penicillin production, defined as:

$$\frac{dN}{dt} = \frac{F_{oil} c_{N_{oil}} + F_{PAA} c_{N_{PAA}} + N_{shots} c_{N_{shots}}}{V} - \mu_X X Y_{N/X} - \mu_P X Y_{N/P} - m_N X - \frac{N}{V}\frac{dV}{dt} \tag{5.11}$$

The simulation considers the nitrogen composition of the input feeds, F_{oil} and F_{PAA} in addition to the ammonia sulphate salts (N_{shots}), which are added to rapidly increase the nitrogen concentration. The nitrogen concentration in these inputs (i) is represented by c_{N_i}. The yield coefficients of nitrogen to biomass and penicillin were represented as $Y_{N/X}$and $Y_{N/P}$, respectively.

Some fermentations require the addition of a precursor to ensure metabolic production of a desired product. This is particularly important for the industrial case study presented here, where phenylacetic acid (PPA) is added to supply the desired side

chain for penicillin synthesis. A simplified PAA uptake rate based on the biomass growth, penicillin production and penicillin maintenance is modelled here as:

$$\frac{dPAA}{dt} = -\frac{F_{PAA}c_{PAA}}{V} - Y_{PAA/P}\mu_P P - Y_{PAA/X}\mu_X X - m_{PAA}P - \frac{PAA}{V}\frac{dV}{dt} \quad (5.12)$$

where F_{PAA} is the flow rate of the PAA and c_{PAA} is the feed solution concentration. $Y_{PAA/P}$ and $Y_{PAA/X}$are the yield coefficients of PAA for penicillin and biomass, respectively, and m_{PAA} is a maintenance term related to the concentration of penicillin. Off-gas analysis involves monitoring the exhaust gas leaving the head space of the fermenter. The off-gas concentration of oxygen $\left(O_{2_{out}}\right)$ was measured in terms of a percentage (%):

$$\frac{dO_{2_{out}}}{dt} = \frac{Q_{g_{in}}O_{2,in} - Q_{g_{out}}O_{2,out} - k_L a(DO_2^* - DO_2)V_L}{(29V_g/22.4)} \quad (5.13)$$

where $Q_{g_{in}}$ and $Q_{g_{out}}$ are taken as the mass flow rate of air in and out, respectively. $O_{2,in}$ and $O_{2,out}$ are the oxygen concentrations in the gas inlet and outlet, respectively. V_g is the volume of the gas in the vessel taken as εV_L which is converted into a mass here to work out $O_{2,out}$ in terms of a %. Similarly, the off-gas calculation of CO_2 was calculated as:

$$\frac{dCO_{2_{out}}}{dt} = \frac{Q_{g_{in}}CO_{2,in} - Q_{g_{out}}CO_{2,out} + CER_X}{(29V_g/22.4)} \quad (5.14)$$

where CER_X is the carbon evolution rate taken here as directly related to the biomass: $CER_X = VX/Y_{CO_2-X}$ with Y_{CO_2-X} as the yield coefficient of CO_2. Measuring the CO_2 and O_2 concentrations in the off-gas analysis allows for the calculation of the oxygen uptake rate (OUR) and the carbon evolution rate (CER) which are valuable tools in the real-time monitoring of fermentations:

$$\begin{aligned} OUR &= \frac{32}{22.4}F_{g_{in}}\left(O_{2_{in}} - O_{2_{out}}\frac{N_{2_{in}}}{1 - O_{2_{out}} - CO_{2_{out}}}\right) \\ CER &= \frac{44}{22.4}F_{g_{in}}\left(CO_{2_{out}}\frac{N_{2_{in}}}{1 - O_{2_{out}} - CO_{2_{out}}} - CO_{2_{in}}\right) \end{aligned} \quad (5.15)$$

The simulation based on the equations above, was validated using the batch records from an industrial case study using the parameter values and presented in the case study (Goldrick *et al.*, 2015). Here, the sugar flow (F_s), soybean oil flow (F_{oil}), water for injection (F_w), aertion rate (F_g), vessel back pressure (P_1) and discharge rate (F_{dis}) are manipulated variables which were interpolated using a sampling time of $0.2\ h$ and used as the simulation inputs. A 0.2 hours sampling time was used in the data generation. The simulation is validated by 10 batches. The variables marked with the asterisk in Figure 5.1 represent the variables not recorded within the 10 batch records.

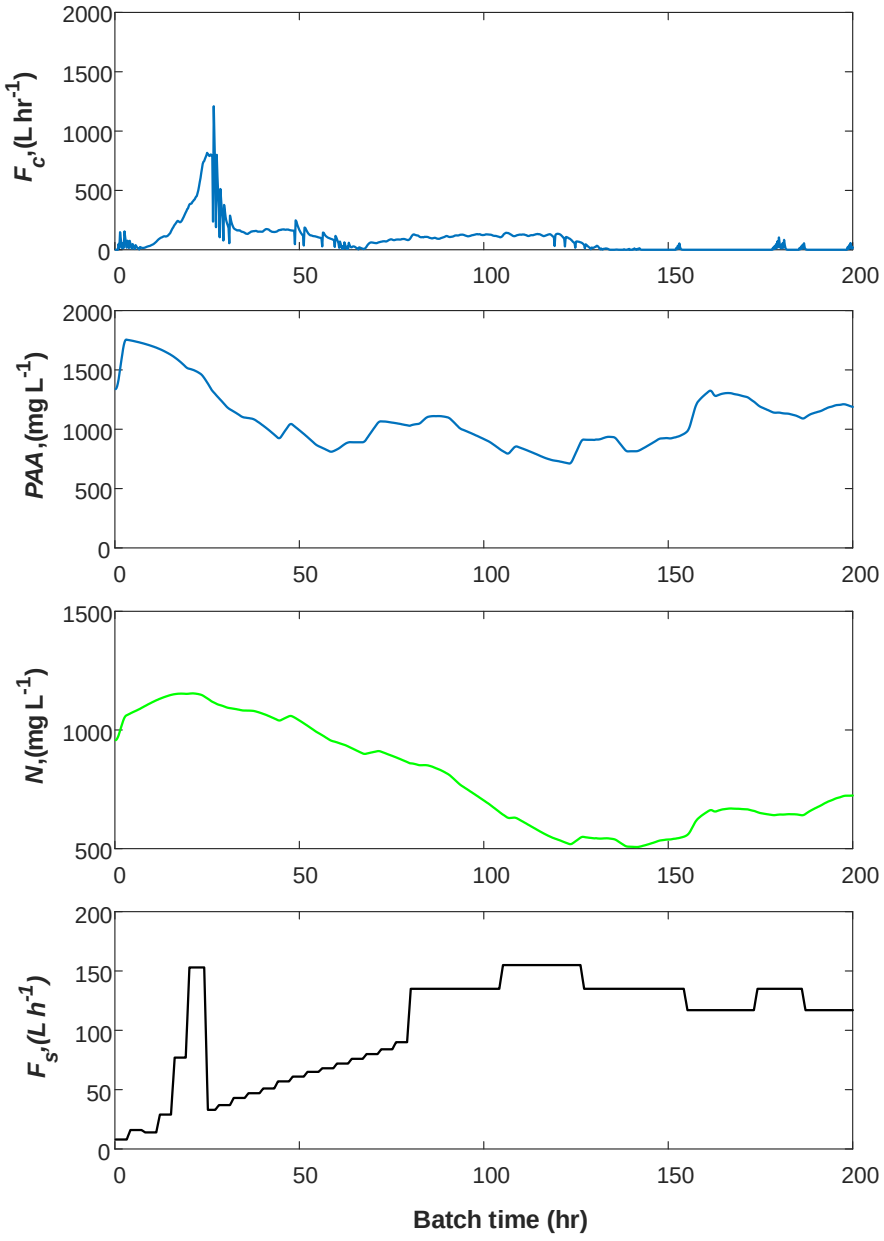

Figure 5.2: Illustration of coolant flow (F_c), phenyl acetic acid concentration (PAA), nitrogen concentration (NH_3) and substrate flow rate (F_s) for the simulation of penicillin production under normal operating conditions (Batch 1 in the study by Goldrick *et al.*, 2015).

The penicillin simulator gives an opportunity to follow different variables that have different characteristics such as binary, continuous, non-continuous, stationary, nonstationary, etc. Some variables from the simulation are illustrated in Figure 5.2 to show the differences between the different time ranges. Furthermore, some variables are selected for process monitoring depending on their effect on the process. The variables used in this study are tabulated in Table 5.1. Some variables such as OUR and CER are calculated from the measured variables. These calculations are stated above. Moreover, the penicillin fermentation process is a multiphase process, which has been divided into 5 phases in several studies for process monitoring purposes (Zhang, Zhao and Gao, 2019). The five phases adopted in this study were selected as $(0-30\,, 30-68.2,\ 68.2-102.2,\ 102.2-133.4,\ 133.4-\ 200\ hrs+)$.

Table 5.1: Variables monitored in the industrial penicillin simulator.

No	Variable	Acronym	No	Variable	Acronym
1	Coolant Flow	F_c	**11**	Phenyl acetic acid Concentration	PAA
2	Dissolved Oxygen Level	DO_2	**12**	Nitrogen Concentration	Nit
3	Biomass Concentration	X_{Bio}	**13**	Oxygen Uptake Rate	OUR
4	Penicillin Yield	P	**14**	Oxygen Percentage	O_2
5	Vessel Volume	V	**15**	Carbon Evolution Rate	CER
6	Vessel Weight	W	**16**	Biomass Specific Growth Rate	μ_X
7	pH	pH	**17**	Penicillin Specific Growth Rate	μ_P
8	Vessel Temperature	T	**18**	Dissolved CO_2	$CO_{2,L}$
9	Generated Heat	Q	**19**	Viscosity	μ_{vis}
10	CO_2 Percent in Outgas	$CO_{2_{out}}$			

5.3 Monitoring Techniques for Batch Processes

5.3.1 Multi-Principal Component Analysis

Multi principal component analysis (multi-PCA) is a multi-phase approach that establishes different PCA models for different phases/stages using conventional PCA

modelling without considering the nonstationary of the data. The approach was first described in a study by DuPont and John MacGregor (Kosanovich *et al.*, 1994; Lu, Gao and Wang, 2004; Zhang, Zhao and Gao, 2019). The PCA model of the c^{th} phase can be described by:

$$\boldsymbol{X}^c = \sum_{r=1}^{R_M^c} \boldsymbol{p}_r \boldsymbol{t}_r^T + \boldsymbol{E}^c = \boldsymbol{P}^c (\boldsymbol{T}^c)^T + \boldsymbol{E}^c \tag{5.16}$$

where $\boldsymbol{X}^c \in \mathbb{R}^{N \times m_c}$, N and m_c are the number of variables and samples in the c^{th} phase, respectively, $\boldsymbol{P}^c$ is the loading matrix and $\boldsymbol{E}^c$is the estimation error matrix. R_S^c is the maximum number of principal components for the c^{th} phase under the condition of $R_M^c = \min(N, m_c)$. The T^2 metric for the c^{th} phase can be defined by:

$$(\boldsymbol{T}^c)^2 = (\boldsymbol{T}^c)^T (\boldsymbol{\Lambda}_M^c)^{-1} \boldsymbol{T}^c \tag{5.17}$$

and calculation of the control limit for the T^2 metric at a confidence limit, α, is given by:

$$\frac{R_M(m_c - 1)}{m_c - R_M} F_\alpha(R_M, m_c - R_M) \tag{5.18}$$

where R_M is the number of PCs, $\boldsymbol{\Lambda}_{\mathrm{M}}^c = (\boldsymbol{T}^c)^T \boldsymbol{T}^c/(m_c - 1)$ is the sample covariance matrix of scores where m_c is the sample size of the phase. α is the significance level of the F-distribution with degrees of freedom R_M and $m_c - R_M$. The SPE metric for the c^{th} phase can be defined as:

$$SPE^c = (\boldsymbol{E}^c)^{\boldsymbol{T}} \boldsymbol{E}^c \tag{5.19}$$

The limits are dependent on the eigenvalues as given below:

$$SPE_{UCL}^c = \left(\frac{z_{(1-\alpha)}\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \tag{5.20}$$

where $\theta_i = \sum_{k=R_M+1}^{m_c} \lambda_k^i$, $i = 1,2,3$ and $h_0 = 1 - (2\theta_1\theta_3)/(3\theta_2^2)$. Here, λ_k^i is the eigenvalue from residuals, i and k refer to indexes of the power and the largest eigen

order, respectively, and $z_{(1-\alpha)}$ is the standard normal deviate or z-score regarding $(1-\alpha)$ percentile.

5.3.2 Multi-level Process Monitoring Method

Use of a multi-level model has been proposed for the monitoring of multi-phase batch processes, which combines the cointegration residuals and t-scores from multiple PCA models (Zhang, Zhao and Gao, 2019). It requires one control chart with the control metrics and limits determined by a PCA model trained on the stationary factors gathered from the PCA model and the cointegration residuals.

Batch processes have a three dimensional data structure that combines the number of batches (I), variables (N) and samples (M) $\underline{\boldsymbol{X}} \in \mathbb{R}^{I\times N\times M}$. Figure 5.3 shows a schematic diagram of the data unfolding. A key property of batch processes is that they can have different phases of operation, e.g. exhibiting different characteristics. A *concurrent identification procedure* has been proposed to determine the unit roots for the variable identification (Zhang, Zhao and Gao, 2019).

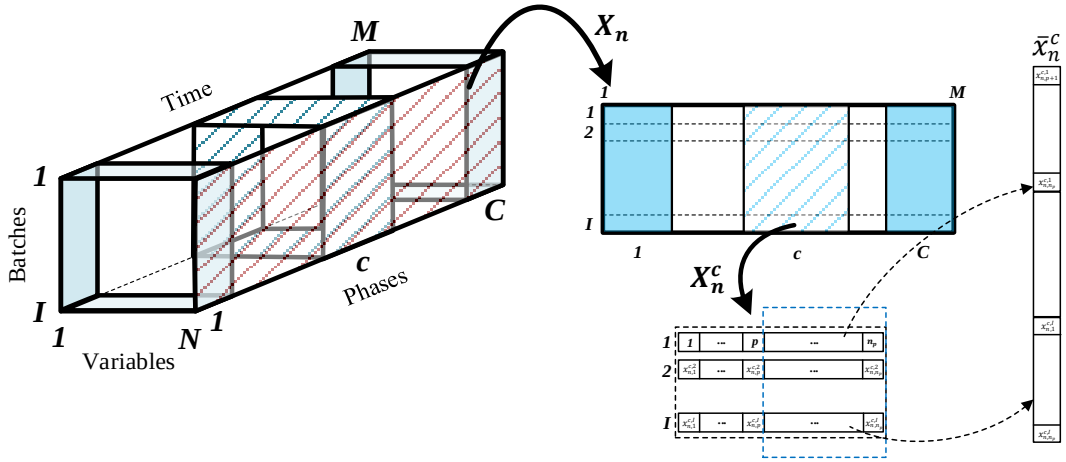

Figure 5.3: Illustration of data unfolding for the identification process.

The concurrent identification procedure employed is based on the ADF test on unfolded variables of $\underline{\boldsymbol{X}}$, or the unit root test. The ADF test is applied to the concatenated variables for the same phase across each batch. Thus, the ADF test is applied $C \times N$ times to identify the stationary and nonstationary variables in each process phase as shown in Figure 5.3, where C is the number of phases. The following regression model can be derived from Equation (3.15) in Section 3.3.2:

$$\begin{cases} \boldsymbol{x}_{n,t}^{c,1} = \mu_n + \theta_n^* \boldsymbol{x}_{n,t-1}^{c,1} + \sum_{k=1}^{p} \theta_k \Delta \boldsymbol{x}_{n,t-k+1}^{c,1} + \boldsymbol{\epsilon}_{n,t}^{c,1} \\ \vdots \\ \boldsymbol{x}_{n,t}^{c,I} = \mu_n + \theta_n^* \boldsymbol{x}_{n,t-1}^{c,I} + \sum_{k=1}^{p} \theta_k \Delta \boldsymbol{x}_{n,t-k+1}^{c,I} + \boldsymbol{\epsilon}_{n,t}^{c,I} \end{cases} \tag{5.21}$$

where μ is a constant, $\boldsymbol{\epsilon}$ is the IID noise representation, and θ_n^* and θ_k are the regression variables for the n^{th} sample $(n = 1, \dots, N)$. $\Delta \boldsymbol{x}_{n,t}^{c,i} = \boldsymbol{x}_{n,t}^{c,i} - \boldsymbol{x}_{n,t-1}^{c,i}$ is the difference operator where p is the lag term, i is the batch number $(i = 1, \dots, I)$ and c is the phase number $(c = 1, \dots, C)$.

An OLS model can be used for the estimation of the regression parameters θ^* and θ:

$$\left[\hat{\mu}_n, \hat{\theta}_n, \hat{\alpha}_1, \dots, \hat{\alpha}_p\right]^T = (\boldsymbol{\Gamma}^T \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \overline{\boldsymbol{x}}_n^c \tag{5.22}$$

where $\overline{\boldsymbol{x}}_n^c = \left[x_{n,1}^{c,1}, \dots, x_{n,1}^{c,I}\right]^T$ is the unfolded vector of the n^{th} variable in the c^{th} phase. $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_I]^T$ is the regression matrix:

$$\boldsymbol{\Gamma}_i = \begin{bmatrix} 1 & x_{n,p}^i & \Delta x_{n,p}^i & \cdots & \Delta x_{n,1}^i \\ 1 & x_{n,p+1}^i & \Delta x_{n,p+1}^i & \cdots & \Delta x_{n,2}^i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,m_c-1}^i & \Delta x_{n,m_c-1}^i & \cdots & \Delta x_{n,m_c-p}^i \end{bmatrix} \tag{5.23}$$

Here, $\boldsymbol{\Gamma_i}$ represents the selected variables of the i^{th} batch, and therefore, Γ is the generalized version of the regression variables where $\Delta \boldsymbol{x}_{n,p}^i = \boldsymbol{x}_{n,p}^i - \boldsymbol{x}_{n,p-1}^i$, p is the order of the regression and m_c is the length of the phase. In practice the ADF test for variable identification in batch processes is different from the original ADF test applied to continuous time series because of the repetition of variables across each of the process phases.

Following the identification of stationary and nonstationary variables, the training data set $(\boldsymbol{X})$ is divided into two groups: $\boldsymbol{X}_{NS}$ and $\boldsymbol{X}_S$ as illustrated in Figure 5.4. The multi-level method uses two models: a PCA model for the stationary variables $(\boldsymbol{X}_S)$ and a cointegration analysis model for the nonstationary variables. The stationary factors $\left(\widetilde{\boldsymbol{X}}_{ML}\right)$ gathered from the t-scores of the PCA model and the cointegration residuals from cointegration analysis are used as a training data set for the 2nd level PCA model.

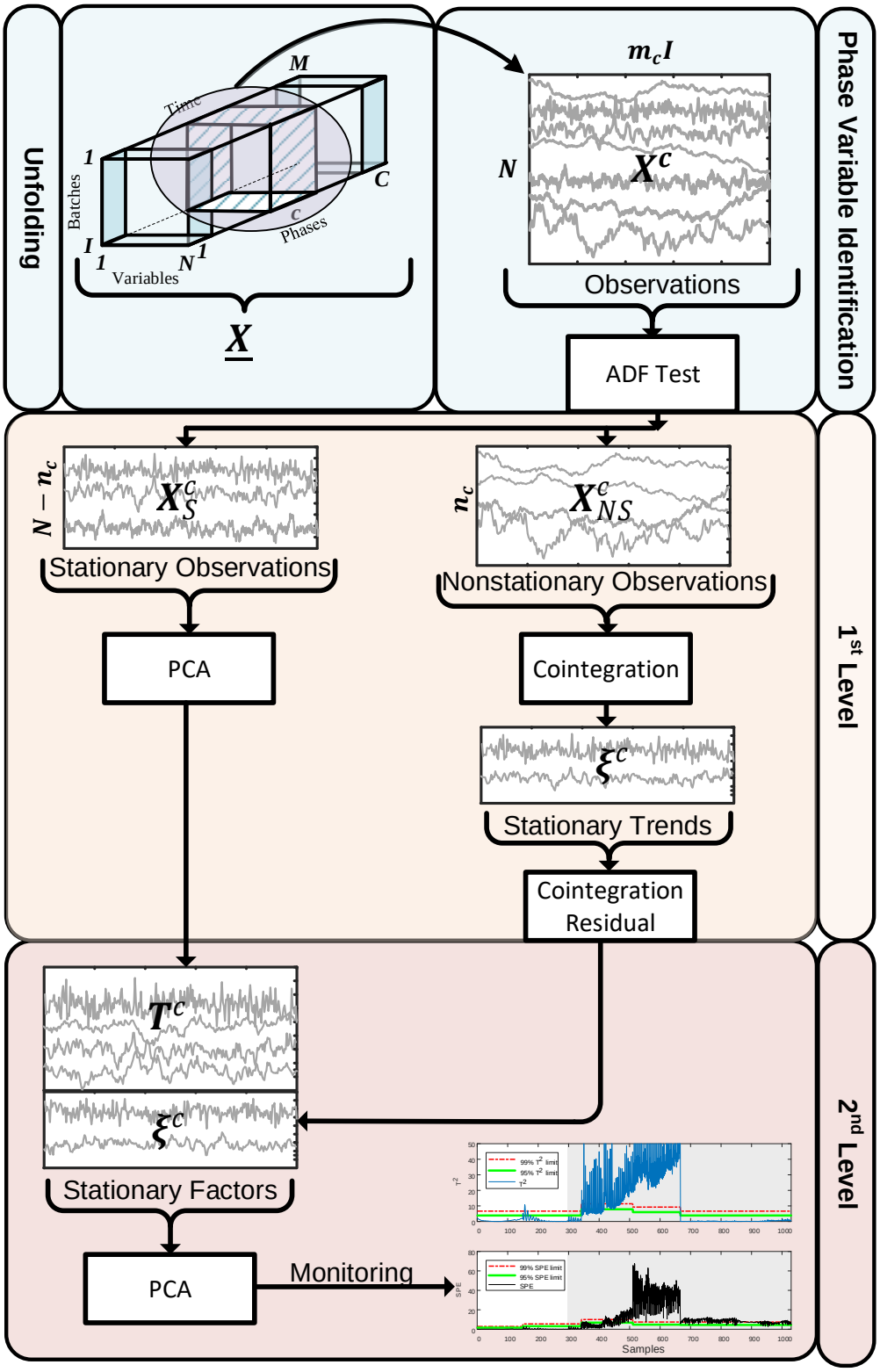

Figure 5.4: Illustration of batch process monitoring method based on a multi-level model.

The multi-level method uses several models for each phase, which trained separately in terms of the models and upper limits for control charts. Thus, 3 models must be

trained: 2 models for the 1st level consisting of PCA and cointegration analysis, and a PCA model at the 2nd level for the stationary factors. The stationary cointegration residuals can be represented for the c^{th} phase as below:

$$\boldsymbol{\xi}^c = (\boldsymbol{\beta}^c)^T \boldsymbol{X}_{NS}^c \tag{5.24}$$

where $\boldsymbol{X}_{NS}^c \in \mathbb{R}^{n_c \times m_c}$ represents the nonstationary variables where n_c and R_c are the number of the nonstationary variables and the cointegration rank for the c^{th} phase. $\boldsymbol{\beta}^c \in \mathbb{R}^{n_c \times R_c}$ is the cointegration matrix determined by the Johansen test according to Section 3.4.2.

The T^2 metric for process monitoring based on cointegration residuals is given by:

$$(\boldsymbol{T}_{CA}^c)^2 = (\boldsymbol{\xi}^c)^T (\boldsymbol{\Lambda}_{CA}^c)^{-1} \boldsymbol{\xi}^c \tag{5.25}$$

where $\boldsymbol{\Lambda}_{\mathrm{CA}}^{\mathrm{c}} = \sum_{k=1}^{m_c} (\boldsymbol{\xi}_k^c)^T \boldsymbol{\xi}_k^c / m_c$ is the sample covariance matrix in the residuals. $\boldsymbol{T}_{CA}^c$ provides information about the changes of the variations between the nonstationary variables ($\boldsymbol{X}_{NS}^c$). Similar to the previously defined upper control limits, the control limits for cointegration analysis in the c^{th} phase can be calculated as

$$\frac{R_c(m_c-1)(m_c+1)}{m_c(m_c-R_c)} F_\alpha(R_c, m_c - R_c) \tag{5.26}$$

where α is the significance level of the F-distribution $F_\alpha(R_c, m_c - R_c)$ with degrees of freedom R_c and $m_c - R_c$.

A conventional PCA model in the 1st level determines the t-scores ($\boldsymbol{T}_S$) for the stationary factors matrix $(\widetilde{\boldsymbol{X}}_{ML})$. Decomposition of the stationary variables ($\boldsymbol{X}_S$) can be described for the c^{th} phase:

$$\boldsymbol{X}_S^c = \sum_{r=1}^{R_S^c} \boldsymbol{p}_r \boldsymbol{t}_r^T + \boldsymbol{E}_S^c = \boldsymbol{P}_S^c (\boldsymbol{T}_S^c)^T + \boldsymbol{E}_S^c \tag{5.27}$$

where $\boldsymbol{X}_S^c \in \mathbb{R}^{(N-n_c)\times m_c}$, n_c is the number of the nonstationary variables determined by the concurrent ADF test, $\boldsymbol{P}_S^c$ is the loading matrix and $\boldsymbol{E}_S^c$ is the estimation error matrix for the c^{th} phase. R_S^c is the maximum number of PCs for the c^{th} phase where

$R_S^c = \min(N - n_c, m_c)$. T^2 can be defined for the stationary variables in the 1st level model for the c^{th} phase:

$$(\boldsymbol{T}_s^c)^2 = (\boldsymbol{T}_S^c)^T (\boldsymbol{\Lambda}_{PCA}^c)^{-1} \boldsymbol{T}_S^c \tag{5.28}$$

and the upper control limits:

$$\frac{R_{PCA}(m_c - 1)}{m_c - R_{PCA}} F_\alpha(R_{PCA}, m_c - R_{PCA}) \tag{5.29}$$

where R_{PCA} is the number of the PCs for the 1st level PCA model, $\boldsymbol{\Lambda}_{\mathrm{PCA}}^c = (\boldsymbol{T}_s^c)^T \boldsymbol{T}_s^c/(m_c - 1)$ is the sample covariance matrix of scores. α is the significance level of the F-distribution with degrees of freedom R_{PCA} and $m_c - R_{PCA}$. Furthermore, the SPE metric can be determined as:

$$SPE^c = (\boldsymbol{E}_S^c)^{\boldsymbol{T}} \boldsymbol{E}_S^c \tag{5.30}$$

and the upper control limits:

$$\left(\frac{z_{(1-\alpha)}\sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \tag{5.31}$$

where $\theta_i = \sum_{k=R_{PCA}+1}^{m_c} \lambda_k^i$, $i = 1,2,3$ and $h_0 = 1 - (2\theta_1\theta_3)/(3\theta_2^2)$. Here, λ_k^i is the eigenvalue from residuals, i and k refer indexes of the power and largest eigen order, $z_{(1-\alpha)}$ is the standard normal deviate or z-score regarding $(1 - \alpha)$ percentile. Sections 2.3 and 2.5 provide further information on PCA modelling.

The stationary factor for the multi-level model for the c^{th} phase ($\widetilde{\boldsymbol{X}}_{ML}^c$) can be represented as:

$$\widetilde{\boldsymbol{X}}_{ML}^c = [\boldsymbol{\xi}^c, \boldsymbol{T}_S^c] \tag{5.32}$$

and the 2nd level PCA model on the stationary factors can be represented as

$$\widetilde{\boldsymbol{X}}_{ML}^c = \sum_{r=1}^{R_{ML}^c} \boldsymbol{p}_{ML} \boldsymbol{t}_{ML}^T + \widetilde{\boldsymbol{E}}_{ML}^c = \widetilde{\boldsymbol{P}}_{ML}^c \left(\widetilde{\boldsymbol{T}}_{ML}^c\right)^T + \widetilde{\boldsymbol{E}}_{ML}^c \tag{5.33}$$

where R_{ML}^{c} is the number of PCs for the c^{th} phase and $\widetilde{\boldsymbol{P}}_{ML}^{c}$, $\widetilde{\boldsymbol{T}}_{ML}^{c}$ and $\widetilde{\boldsymbol{E}}_{ML}^{c}$ are the loading, score and error matrix of the 2nd level PCA model, respectively. The $\widetilde{\boldsymbol{T}}_{ML}^{c}$. and SPE_{ML} metrics can be derived from the 1st level PCA model from the given matrices and the number of PCs (R_{ML}^{c}).

5.3.3 Multi-level Multi-factor Process Monitoring Method for Batch Processes

The multi-level model described in Section 5.3.2 was the first method to propose the use of PCA and cointegration analysis together for batch processes. The approach provides a final PCA model that combines cointegration residuals and PCA scores. However, cointegration analysis is ill-suited to the modelling of data exhibiting higher level nonstationary characteristics. In some cases, cointegration analysis can end up with only one cointegration relationship even though the cointegration matrix has a higher rank. Some cointegration relationships built by cointegration analysis can result in an ill-suited situation for the monitoring of nonstationary variables where the t-score vectors can dominate in terms of the number of variables in the 2nd level PCA model. To solve this issue, the use of common-trend models has been adopted along with cointegration analysis and PCA.

The concurrent identification method detailed in the previous section is used for variable identification for the multi-level multi-factor model. The model is illustrated in Figure 5.5. In comparison to the multi-level model, the 1st level modelling techniques include common-trend residuals-based process monitoring method, which is applied to the nonstationary variables ($\boldsymbol{X}_{NS}$). Therefore, the 2nd level PCA model uses different stationary factors to those in the multi-level model.

The multi-level multi-factor model uses the same models in the 1st level as for the multi-level method with the addition of a common-trend residuals-based process monitoring method. A common-trend model is applied to the unused vectors in the cointegration matrix to improve the performance of the monitoring of the nonstationary variables. Cointegration analysis determines the rank of cointegration matrix according to the cointegration relationship of the nonstationary variables, which also equals the number of stationary factors gathered from the cointegration residuals. The common-trend model allows the multi-level multi-factor model to analyse the nonstationary space, that is not used by cointegration model. After whitening of the common-trend residuals, the variables transform into stationary factors.

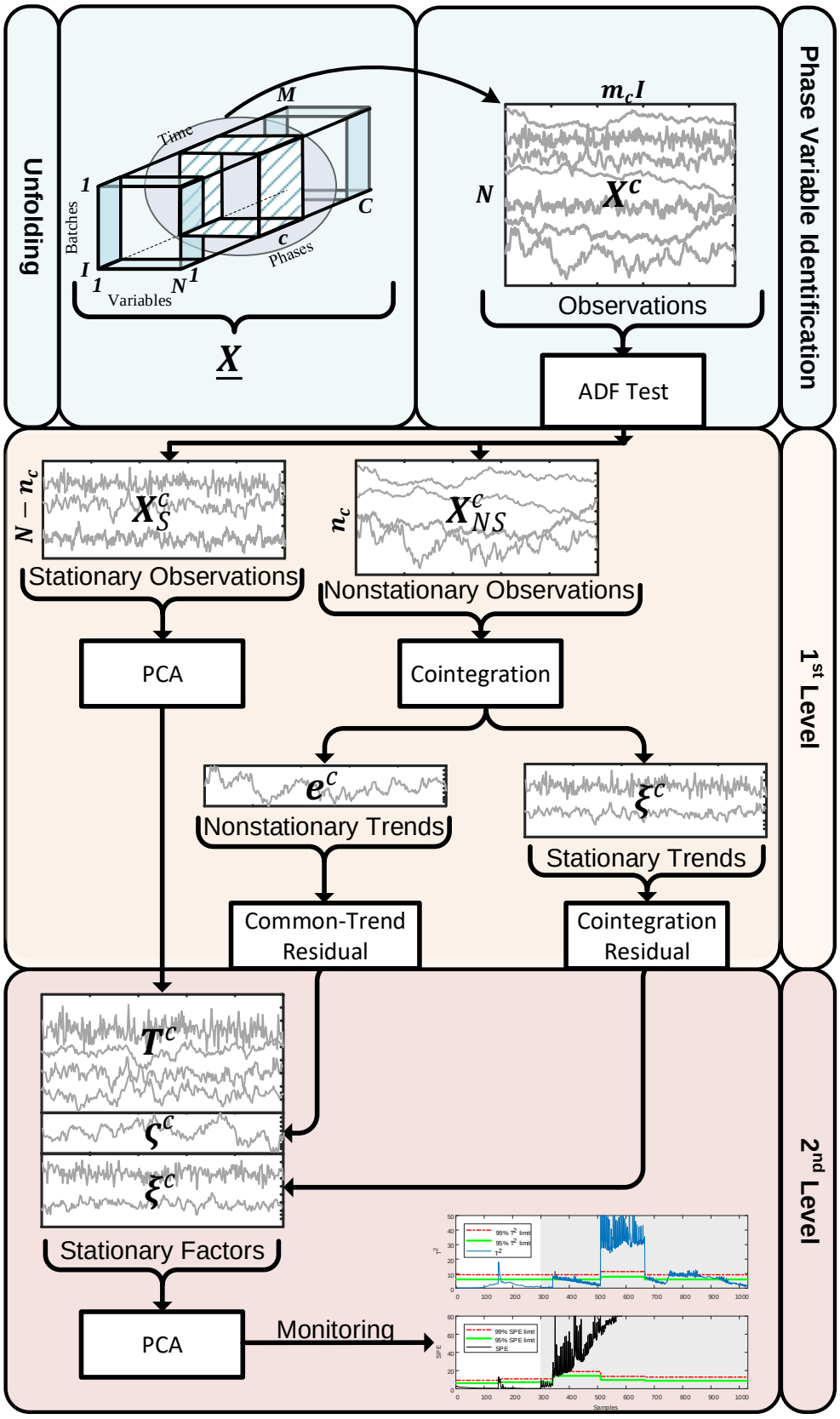

Figure 5.5: Illustration of batch process monitoring method based on a multi-level multi-factor model.

In addition to the stationary factors $\boldsymbol{\xi}^c$ and $\boldsymbol{T}_s^c$ described by the multi-level model for the c^{th} phase, the stationary factors from the common-trend residuals ($\boldsymbol{\varsigma}^c$) require the same cointegration matrix built by the cointegration model for the nonstationary variables ($\boldsymbol{X}_{NS}$) in Section 5.3.2. The cointegration matrix has a maximum rank of $N-1$. Low rank matrices imply the existence of common-trends. Assuming that the rank for the c^{th} phase is R_c, the first R_c columns of the cointegration matrix represent the relationship between nonstationary variables to obtain stationary factors. The remaining $n_{ns}-R_c$ cointegration vectors still represent a large portion of the process characteristics. For some cases in batch process monitoring, cointegration analysis may result in a smaller cointegration rank, which may cause loss of valuable information.

A common-trend model requires a set of vectors $\boldsymbol{\beta}_{\perp}^c \in \mathbb{R}^{n_c \times (n_c - R_c)}$ such that the matrix is perpendicular to the ranked cointegration matrix $\boldsymbol{\beta}^c \in \mathbb{R}^{n_c \times R_c}$ and $[\boldsymbol{\beta}^c, \boldsymbol{\beta}_{\perp}^c]$ has full rank. Common-trend residuals for the c^{th} phase can be represented as:

$$\boldsymbol{e}^c = (\boldsymbol{\beta}_{\perp}^c)^T \boldsymbol{X}_{NS}^c \tag{5.34}$$

where $\boldsymbol{e}^c$ represents the nonstationary residuals and a sample point of $\boldsymbol{e}^c$ can be represented via a VAR process, discussed in Section 4.3.4, to extract stationary factors ($\boldsymbol{\varsigma}^c$). The T^2 metric can be defined as:

$$(\boldsymbol{T}_{CT}^c)^2 = (\boldsymbol{\varsigma}^c)^T (\boldsymbol{\Lambda}_{CT}^c)^{-1} \boldsymbol{\varsigma}^c \tag{5.35}$$

where $\boldsymbol{\Lambda}_{\mathrm{CT}}^{\mathrm{c}} = \sum_{k=1}^{m_c} (\boldsymbol{\varsigma}_k^c)^T \boldsymbol{\varsigma}_k^c / m_c$ is the sample covariance matrix and the UCLs are given by:

$$\frac{(n_{ns} - R_c)(m_c - 1)(m_c + 1)}{m_c(m_c - n_c + R_c)} F_{\alpha}(n_c - R_c, m_c - n_c + R_c) \tag{5.36}$$

where α is the significance level of the F-distribution $F_{\alpha}(n_c - R_c, m_c - n_c + R_c)$ with degrees of freedom $n_c - R_c$ and $m_c - n_c + R_c$. The stationary factors arising from the 1st level of the multi-level multi-factor model are given by the stationary factors from the 3 sub-models:

$$\widetilde{\boldsymbol{X}}^c = \left[\widetilde{\boldsymbol{X}}_{ML}^c, \boldsymbol{\varsigma}^c\right] = [\boldsymbol{\xi}^c, \boldsymbol{T}_s^c, \boldsymbol{\varsigma}^c] \tag{5.37}$$

Note that cointegration residuals ($\boldsymbol{\xi}^c$) and t-score ($\boldsymbol{T}_S^c$) from 1st level are the same stationary factors gathered from the 1st level of the multi-level modelling $\left(\widetilde{\boldsymbol{X}}_{ML}^c\right)$. The 2nd PCA model applied to the stationary factors $\left(\widetilde{\boldsymbol{X}}^c\right)$ can be represented as:

$$\widetilde{\boldsymbol{X}}^c = \sum_{r=1}^{R_{Sec}^c} \boldsymbol{p}_r \boldsymbol{t}_r^T + \widetilde{\boldsymbol{E}}^c = \widetilde{\boldsymbol{P}}^c \left(\widetilde{\boldsymbol{T}}^c\right)^T + \widetilde{\boldsymbol{E}}^c \tag{5.38}$$

where R_{Sec}^c is the number of PCs for the c^{th} phase and $\widetilde{\boldsymbol{P}}^c$, $\widetilde{\boldsymbol{T}}^c$ and $\widetilde{\boldsymbol{E}}^c$ are the loading, score and error matrix of the 2nd level PCA model, respectively. The T^2 and SPE metrics can be derived from the 1st level PCA model by using the given matrices and the number of PCs (R_{Sec}^c). Sections 2.3 and 2.5 provide the PCA modelling for further information.

Unlike the continuous process version of the multi-level multi-factor model, the model for batch processes includes unique variable identification and additional modelling as each of the phases requires different models. In Table 5.2, the offline training procedure of the multi-level multi-factor model for batch processes is summarized. This requires batch data sets collected under normal operating conditions.

Table 5.2: Offline training of the multi-level multi-factor model for batch processes.

1: Set up phase lengths (m_C). Set $c = 1$ and $n = 1$
2: Compute Equation (5.23) for regression matrix Γ.
3: Solve OLS in Equation (5.22) and determine θ for nonstationarity.
4: If $n < N$, then set $n = n + 1$ and go to Step 2
5: Solve Equation (3.25) to obtain $\boldsymbol{\beta}^c$.
6: Calculate $\boldsymbol{\xi}^c$ via Equation (5.25)
7: Find $\boldsymbol{\beta}_\perp^c$ and calculate $\boldsymbol{e}^c$ via Equation (5.34) and then whiten $\boldsymbol{e}^c$ via Equation (4.18) to find $\boldsymbol{\varsigma}^c$.
8: Use Appendix-A on ($\boldsymbol{X}_S^c$) to find $\boldsymbol{T}^c$ via Equation (5.27).
9: Set up stationary factors ($\widetilde{\boldsymbol{X}}^c$) via Equation (5.37)
10: Use Appendix-A on ($\widetilde{\boldsymbol{X}}^c$) to find $\widetilde{\boldsymbol{T}}^c$ via Equation (5.38).
11: Calculate the upper control limits.
12: If $c < C$ then set $c = c + 1$ and $n = 1$ go to Step 2.

Following the determination of the upper control limits and the model, process monitoring can continue with the online diagnosis listed in Table 5.3.

Table 5.3: Online diagnosis of the multi-level multi-factor model for batch processes.

1 : For each sampling point (t) set $c = 1$
2 : Calculate $\boldsymbol{\xi}_t^c$ via Equation (5.25)
3 : Calculate $\boldsymbol{\varsigma}_t^c$ via Equation (5.34) then Equation (4.18)
4 : Calculate $\boldsymbol{T}_t^c$ via Equation (5.27).
5 : Set up stationary factors $(\widetilde{\boldsymbol{X}}_t)$ via Equation (5.37)
6 : Calculate $\widetilde{\boldsymbol{T}}_t^c$ via Equation (5.38).
7 : Calculate T^2 via Equation (5.28) and SPE via Equation (5.30)
8 : If any of them exceed the upper limits that defined, then raise an alarm
9 : If $t < m_c$ go to Step 2, else $c = c + 1$ and go to Step 2 until the last sampling point.
10: End if it is the last point

5.4 Application to Industrial Penicillin Simulator

The industrial-scale penicillin simulator was used to compare three methods that have been introduced: multi-PCA, a multi-level model and the multi-level multi-factor model. Nine standard batches were selected from the real fermentation data sets and cut to 200 hours for the training of the methods.

Two examples of process faults were created to evaluate the performance of the different methods. The first comprised a fault on the temperature sensor, which affects directly the variables that are measured (see Table 5.1). The fault is a ramp function starting at the time of the initiation of the fault. The second was a fault on the substrate feed rate, which was not monitored directly but has an indirect effect on the variables that were monitored (see Table 5.1).

The multi-PCA method is a multiphase monitoring approach, which establishes a different PCA model for each phase but does not consider the nonstationarity of the data. Each PCA model can explain a different percentage of variance so as to vary the sensitivity of the model through time for the phases. However, for simplicity, the explained variance level of the multi-PCA models was set to be the same as that for the PCA model employed in the first level of the multi-level multi-factor method. The explained variance was selected to be 45% to provide a better performance on the type-1 error rate for the training data sets compared to other fixed variance percentages (see Appendix-D for further information about PC selection). It is noted that each of the operation phases has its own variance distributions given by the PCs. Therefore,

even the same level of explained variance can result in a different number of PCs and upper control limits.

In contrast, the multi-level and multi-level multi-factor models share some of sub-models from the 1st level of modelling. In addition to these same models, the multi-level multi-factor model uses common-trend models. In comparison to the multi-level model, the multi-level multi-factor model uses common-trend models where cointegration analysis has difficulties in building long-run equilibria to make nonstationary $\left(I(1)\right)$ variables stationary $\left(I(0)\right)$, which can be monitored using standard PCA.

5.4.1 Model Training

Training of the sub-models for each phase was performed using the data from the normal batches. Multi-PCA requires one parameter to be set, which is the number of principal components required to explain the selected variance in the data, 45% in this case.

It is noted that since the multiphase batch process is divided into different operating phases, each phase is modelled by a different model. This gives rise to differences in the control limits for each phase. Thus, it is normal to see various control limits throughout the different phases. For both PCA and cointegration analysis, T^2 is based on the F-distribution-based ranking for the control limits while the SPE uses eigenvalues for control limits with PCA. The procedure for training of the models is shown in Table 5.2. Details on how model variance was selected is given in Appendix-D.

First training was performed using multi-PCA and the results obtained for the monitoring of a normal batch are shown in Figure 5.6. The multi-PCA model combines 5 different PCA models, which are tabulated in Table 5.4. Figure 5.6 shows the residuals collected from one of the nine training data sets using a cross validation approach. Most of the T^2 metrics lie below the control limits and a few samples went above the control limits in the SPE chart. This issue can be resolved by having a different variance level for each phase. However, the main purpose of this study is to compare the same parameters for the same models in the different methods, thus the variance level was kept the same for each PCA model in the study.

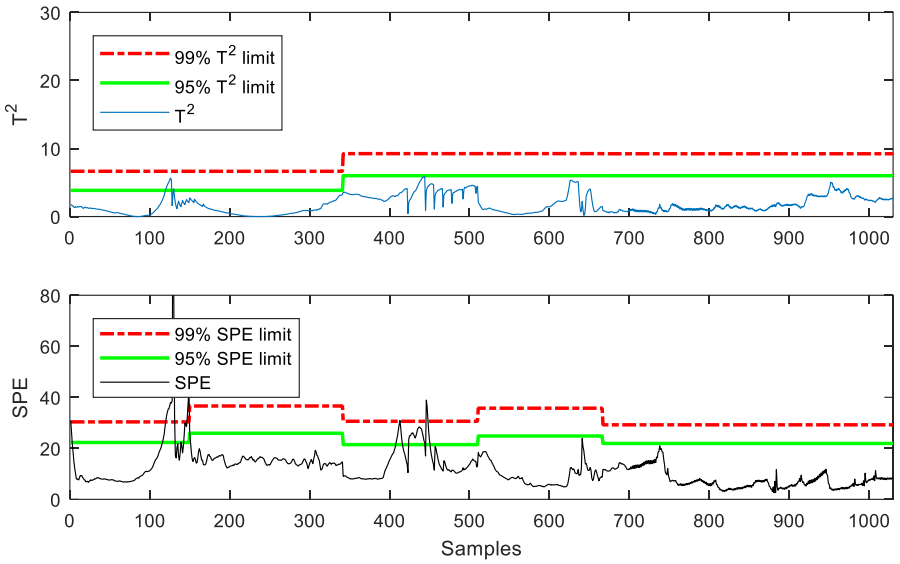

Figure 5.6: T^2 and SPE metrics obtained for a multi-PCA model built using a training batch exhibiting normal operation.

Table 5.4: Number of principal components selected for each PCA model.

Model Name	**Phase Number**	**Number of PCs**	**Variance in the Data** (%)
Multi-PCA	1	2	**44.9**, **12.4**, 9.2, 8.5, 6.2
	2	2	**39.5**, **17.1**, 10.6, 10.6, 6.4
	3	2	**31.0**, **19.7**, 14.8, 8.9, 7.8
	4	3	**23.2**, **20.3**, **17.5**, 10.7, 8.9
	5	3	**28.2**, **15.2**, **10.5**, 10.4, 6.7
1st level PCA	1	1	**60.5**, 13.5, 11.9, 6.6, 3.1
	2	1	**55.0**, 14.8, 11.4, 6.8, 5.4
	3	2	**26.8**, **20.1**, 19.1, 14.1, 8.7
	4	2	**33.0**, **24.3**, 18.2, 11.1, 9.4
	5	2	**26.6**, **18.4**, 12.4, 9.3, 8.0
2nd level PCA (Multi-level)	1	1	**78.3**, 21.7
	2	1	**58.3**, 41.6
	3	3	**21**, **5**, **18.5**, **16.7**, 16.6, 14.8
	4	2	**33.0**, **27.4**, 22.6, 17.0
	5	1	**53.8**, 33.3, 12.9
2nd level PCA (Multi-level Multi-factor)	1	2	**43.8**, **24.0**, 15.2, 9.4, 7.3
	2	2	**35.4**, **25.5**, 17.8, 12.4, 8.7
	3	2	**35.1**, **14.6**, 10.5, 8.8, 7.8
	4	3	**24.9**, **19.5**, **15.8**, 11.2, 8.9
	5	2	**34.6**, **26.5**, 16.6, 9.0, 7.6

It is common to see different characteristics in the variables through the different time phases in batch processes and so it is necessary to check the stationary characteristics of each variable during all time phases. These tests can show different results for the

same variable across different time phases, which has led to the development of the multi-level multi-factor process monitoring scheme where different models are used for each variable set and time phase. The variables that exhibit nonstationarity during phases 1 to 5 of the batch process are tabulated in Table 5.5.

Having different control limits for the SPE and T^2 metrics is possible because of the different number of PCs selected for each phases. For example, the T^2 metric in Figure 5.6 shows two different levels for the control limits. This difference arises from the selected number of PCs for the multi-PCA model as stated in Table 5.4. However, this is not the case for the SPE metric as it uses eigenvalues of the different samples through the phases.

Table 5.5: Nonstationary variables in each phase of the batch process given by the industrial penicillin simulator.

No	**Phase Range (PR)**		**Nonstationary Variables**
	In hours	In samples	
1	$0-30$	$0-150$	$[x_1, x_6, x_7, x_8, x_{12}, x_{13}, x_{18}]$
2	$30-68.2$	$150-341$	$[x_1, x_8, x_{12}, x_{13}, x_{17}, x_{18}]$
3	$68.2-102.2$	$341-511$	$[x_1, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}$
4	$102.2-133.4$	$511-667$	$[x_1, x_2, x_3, x_5, x_9, x_{10}, x_{12}, x_{13}, x_{17}, x_{18}]$
5	$133.4-200+$	$667-1000+$	$[x_4, x_6, x_{12}, x_{13}, x_{17}, x_{18}]$

The first phase of the process is the growth period of the penicillin fermentation. The biomass and weight show clear trends in that period. The first 30 hours of the process has the lowest penicillin concentration because of the initiation of the production. In that phase $X_{NS}^1 = [x_1, x_6, x_7, x_8, x_{12}, x_{13}, x_{18}]$ has been identified as being nonstationary. The second phase is subject to growth of the penicillin culture and hence, increasing concentration. Changes in PAA are also possible because of the optimum control of the flow of phenyl acetic acid, which is a major challenge in penicillin fermentations. It is known that high levels of PAA in the culture are toxic to the biomass by inhibiting growth and penicillin production (Goldrick *et al.*, 2015). F_c keeps its nonstationary characteristics with $T(x_8)$, $Nit(x_{12})$, $OUR(x_{13})$ and $CO_{2,L}(x_{18})$. However, $V(x_5)$ and $W(x_6)$ show stationary characteristics which means that these variables can be monitored by classical PCA. On the other hand, $\mu_P(x_{17})$

changes from being stationary to nonstationary for the 2nd phase: $X_{NS}^{2} = [x_1, x_8, x_{12}, x_{13}, x_{17}, x_{18}]$. The 3rd phase is the saddle point for the different variables, with an increase in the number of nonstationary variables: $X_{NS}^{3} = [x_1, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}]$. After the 3rd phase, the process tends to exhibit steady state regarding the natural limits of the penicillin growth Here; $W, Q(x_9), PAA(x_{11})$, $O_2(x_{14})$ and viscosity became stationary and the following variables were identified as nonstationary: $X_{NS}^{4} = [x_1, x_2, x_3, x_5, x_9, x_{10}, x_{12}, x_{13}, x_{17}, x_{18}]$. For the final (5th) phase, there is a reduction in the number of nonstationary variables, which is expected as more variables exhibit steady-state characteristics. On the other hand, penicillin yield and vessel weight show nonstationary characteristics due to the end of the growing phase where penicillin concentration can be reduced in that phase and weight can be fluctuated from changes (Goldrick *et al.*, 2015). The nonstationary variables in the final phase are given by: $X_{NS}^{5} = [x_4, x_6, x_{12}, x_{13}, x_{17}, x_{18}]$.

The T^2 and SPE metrics obtained using the multi-level model for the training data set are shown in Figure 5.7. As for the multi-level multi-factor model, the multi-level model consists of two levels of modelling, which includes PCA and cointegration analysis. The 1st level PCA models for each phase are tabulated in Table 5.4. Differences in the control limits for each phase result from the rank of the PCA models. The rank of the cointegration matrix for the 5 different phases was determined as: $[1, 1, 4, 2, \text{and } 1]$ using the Johansen test. The ranks show the existence of common-trends within the cointegration analysis, which allows the use of a common-trend model and the number of nonstationary variables for each phase can be assigned as $[7, 6, 12, 10, \text{and } 6]$. The 2nd level PCA models for each phase are also tabulated in Table 5.4. As can be seen from the number of PCs used in the 2nd level PCA model in the multi-level model, a small number of factors are integrated into the 2nd PCA model.

The results obtained for the training data using the multi-level multi-factor model are shown in Figure 5.8. Here, the different sub-models from the 1st and 2nd levels of the multi-level scheme are presented to show the differences between them. Figure 5.8(a) shows the T^2 metric for the cointegration residuals-based model for the nonstationary variables. Figure 5.8(b) shows the T^2 metric for the common-trend residuals-based model for the nonstationary variables. The control limits are different for these two approaches as they complement each other. The last sub-model from the 1st level of the multi-level scheme is shown in Figure 5.8(c), and shows the T^2 and SPE metrics

after application of PCA to only the stationary variables. The 1st level PCA model combines 5 different PCA models which are tabulated in Table 5.4. In the final (2nd level) stage, the residuals from the three sub-models from the 1st level are combined and analysed using PCA. The T^2 and SPE metrics for the 2nd level PCA model are shown in Figure 5.8(d). Modelling by PCA is possible as the residuals from the 1st level are stationary. Again, it combines 5 different PCA models, which are tabulated in Table 5.4. In comparison to the multi-level model, more stationary factors that describe the process dynamics are taken into the 2nd level model, which can be then used to monitor these dynamics.

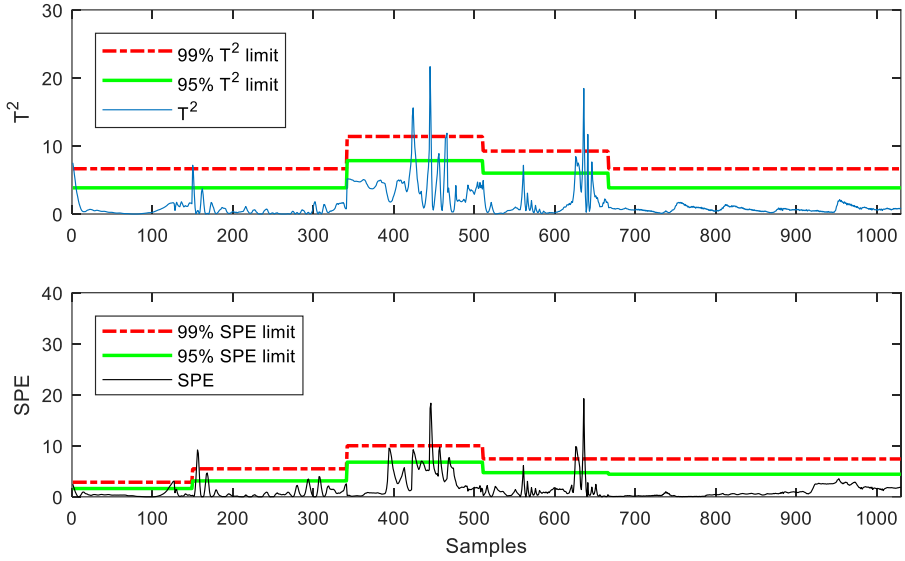

Figure 5.7: T^2 and SPE metrics obtained for a multi-level model built using a training batch exhibiting normal operation.

A performance comparison of the three methods is summarized in Table 5.6. 99% SPE limits were selected as the basis for the comparison as the SPE metric is used to detect deviations that are not explained by the model. The statistical comparison is carried out using the false alarm rate (type-I error) as it is training data. It can be seen that the performance of the three methods is comparable.

Table 5.6: Performance comparison of the multi-level multi-factor, multi-level, and multi-PCA methods based on the type-1 error rate for training data exhibiting normal operation.

Case Study	**Type I error (%)**		
	Multi-level Multi-factor	Multi-level	Multi-PCA
Training	1.8	2.1	2.3

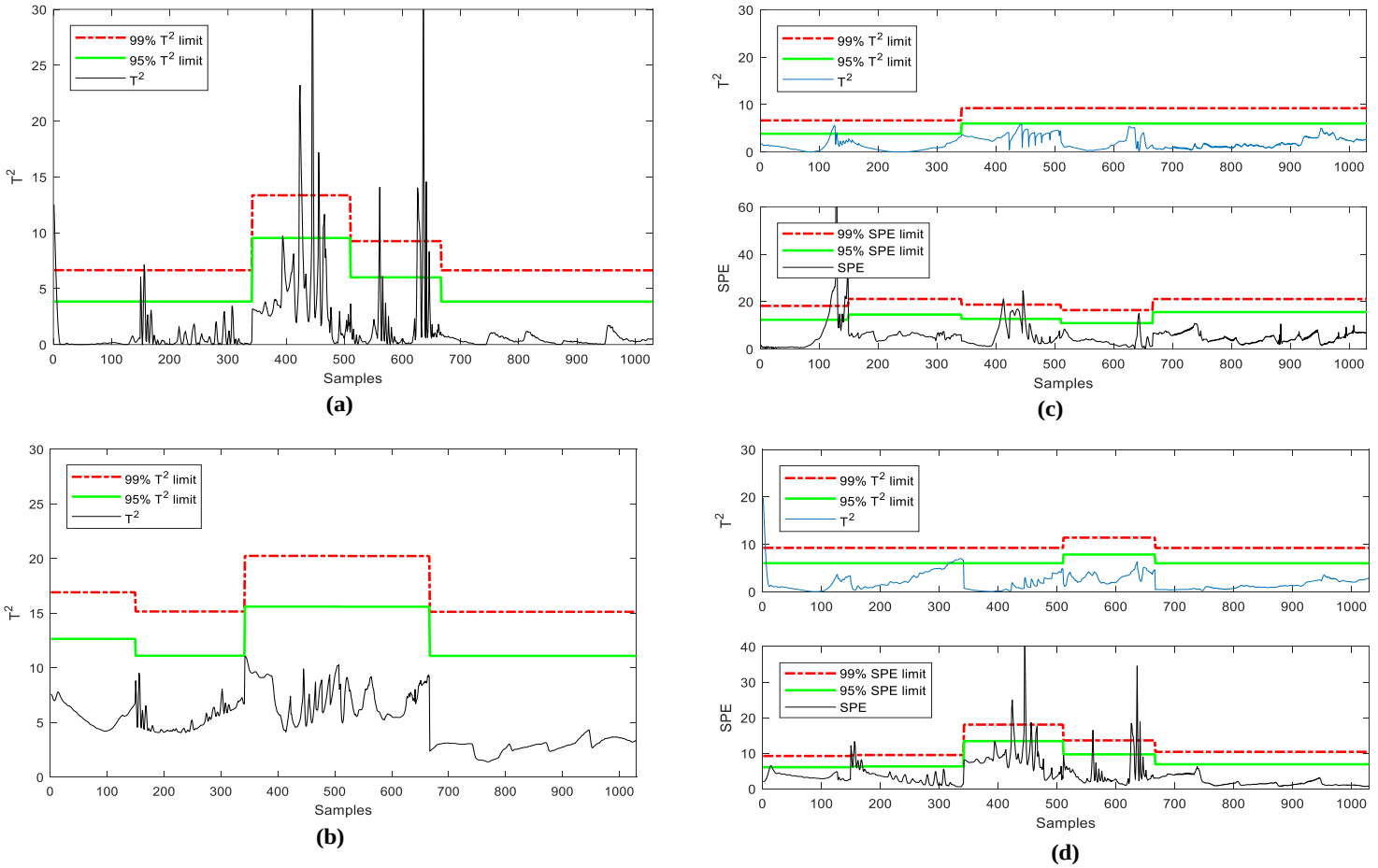

Figure 5.8: Metrics obtained using the multi-level multi-factor model for a training batch exhibiting normal operation. (a) T^2 metric for cointegration analysis at the 1st level, (b) T^2 metric for common-trend model at the 1st level, (c) T^2 and SPE metrics for PCA at the 1st level, and (d) T^2 and SPE metrics for PCA at the 2nd level.

5.4.2 Model Testing

The 3 models were applied to test data sets generated using the industrial penicillin simulator. In the first case, a sensor error on the temperature was defined to be a ramp function starting from sample number 200. This allowed a maximum error of 2% on the measured temperature value. Faulty zones are highlighted in the figures with grey shading as shown, for example, in Figure 5.9 and Figure 5.10. This fault tests the monitoring systems for an error that could result in a plant shutdown due to the increasing magnitude of the error. It is worth noting that the temperature sensor fault shows a clear trend characteristic, which is nonstationary.

The second case was selected as an example of a fault that had an indirect effect on the variables that were monitored. The substrate feed rate was not monitored and therefore, the models need to detect the error from the measured variables indirectly (see Table 5.5). The substrate (sugar) feed rate is changed to 20 $L\,h^{-1}$ at sample number 300. Note that the fault examples used here have already been defined in the simulation (Goldrick *et al.*, 2015).

The results obtained using multi-PCA for detection of a temperature sensor error are shown in Figure 5.9. The T^2 statistic, as expected, was not able to detect the fault during the first three phases of the batch (samples $0-511$) where the vessel temperature exhibits nonstationary characteristics. The ramp function is considered as a common-trend fault and so no samples exceeded the T^2 control limits that define the variance for the normal operating conditions. It is also noted that owing to the PCA model's high explained variance, the T^2 metric is unable to detect nonstationary characteristics. After the process entered the 4th phase (samples $511-667$), the SPE metric goes above the control limits for the first time at sample number 535. The slow time-varying fault characteristics from the ramp function made the fault difficult to detect by multi-PCA owing to the dynamic relationship between the temperature and the other variables. However, after entering the 5th phase (samples $667-1000+$), both metrics go above the control limits.

It is expected that the temperature sensor error would affect some variables that are monitored such as the vessel temperature (x_8). However, the sensor error may be evident in more variables than just the vessel temperature, as the process comprises a number of cross correlated variables. Considering that the vessel temperature is

identified as nonstationary for the first three phases (samples $0-511$) and shows stationary characteristics for the last two phases (samples $511-1000+$), it can be concluded that when the characteristics of the variables change in the different phases, this can affect the performance of conventional models.

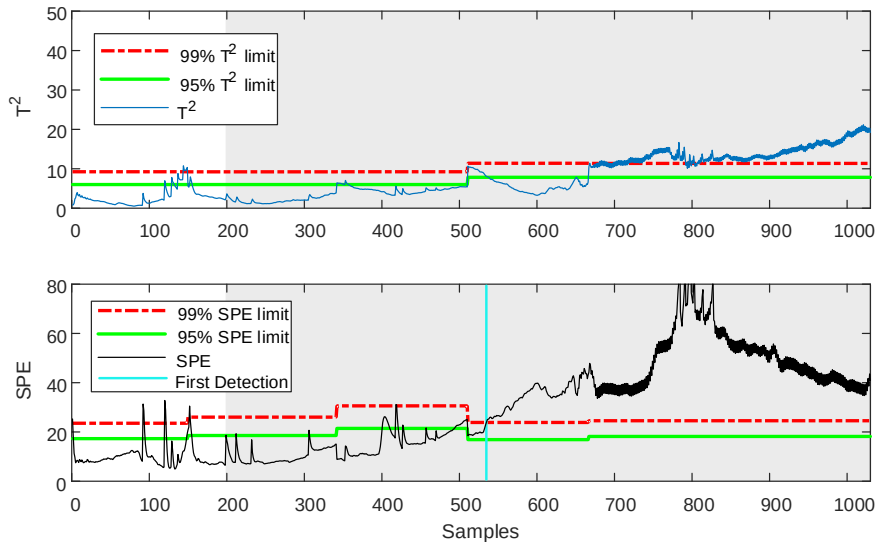

Figure 5.9: T^2 and SPE metrics obtained using multi-PCA for a batch exhibiting a temperature sensor error. The fault is first detected at sample number 535 using the SPE metric (indicated by turquoise vertical line).

Figure 5.10 shows the T^2 and SPE results obtained using a multi-level model for data exhibiting a temperature sensor fault. The multi-level model uses cointegration residuals and PCA scores to build a model. Thus, it uses the stationary factors that are later used to derive the T^2 and SPE metrics observed in Figure 5.11(a) and Figure 5.11(c). Here, different characteristics are modelled by the different models. As can be seen from Figure 5.11(a) and Figure 5.11(c), the PCA and cointegration models are effective on different phases. Detection of the fault using the SPE metric occurs in the 3rd phase (samples $341-511$) at sample number 400 as, shown in Figure 5.10. This occurs earlier than when a fault is detected by multi-PCA. Even though some samples in the 3rd phase (samples $341-511$) are below the control limit, the multi-level model has a better fault detection rate than multi-PCA. As can be seen from Figure 5.11(a) and Figure 5.11(c), the PCA model cannot detect the fault until the 5th phase (samples $667-1000+$). Thus, the effective stationary factors in the multi-level model for fault detection were the cointegration residuals.

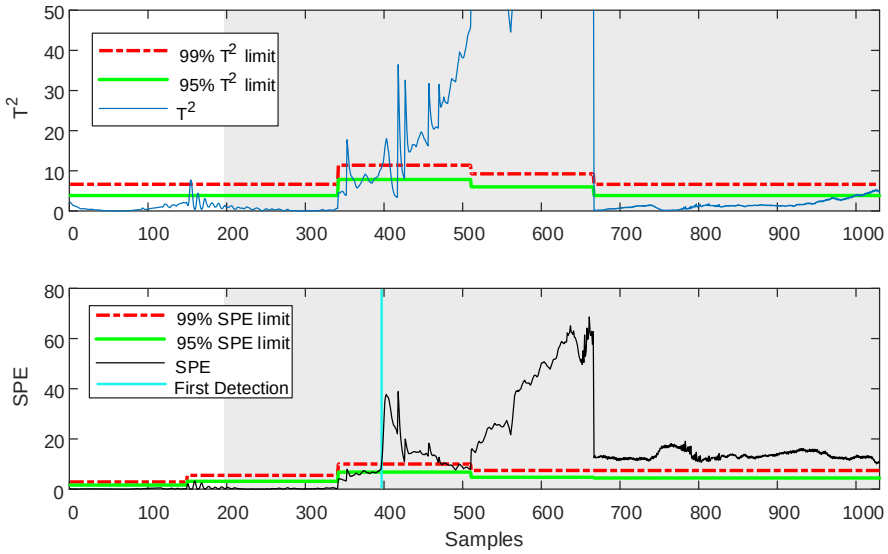

Figure 5.10: T^2 and SPE metrics obtained using a multi-level method for a batch exhibiting a temperature sensor error. The fault is first detected at sample number 400 using the SPE metric (indicated by turquoise vertical line).

Figure 5.11(a) shows the T^2 metric for the cointegration residuals-based process monitoring results for the nonstationary variables. Figure 5.11(b) shows the T^2 metric for the common-trend residuals for the nonstationary variables. Figure 5.11(c) represents the T^2 and SPE metrics for the application of PCA to the stationary variables. As for the multi-level model, the PCA model is effective when the variable that is the most affected by the fault, variable x_8, shows stationary characteristics. On the other hand, it is better to use cointegration and common-trend residuals-based process monitoring models during the first 3 phases (samples $0 - 511$) where variable x_8 exhibits nonstationary characteristics, illustrated in Figure 5.11(a) and (b). The results obtained for detection of the temperature sensor fault using the multi-level multi-factor model are shown in Figure 5.11(d). The results obtained for the three sub-models at the 1st level and the final model at the 2nd level are also shown.

The temperature sensor fault was first detected by the multi-level multi-factor model at the beginning of the 3rd phase (sample number 341), as can be seen in Figure 5.11 (d) which shows the residuals from the 2nd level PCA model that were built from the residuals of the 3 sub-models from the 1st level (see Figure 5.11 (a) to Figure 5.11(c)). The multi-level multi-factor model was able to detect the temperature sensor error earlier than the multi-level method owing to the common-trend residuals involved in the 2nd PCA model.

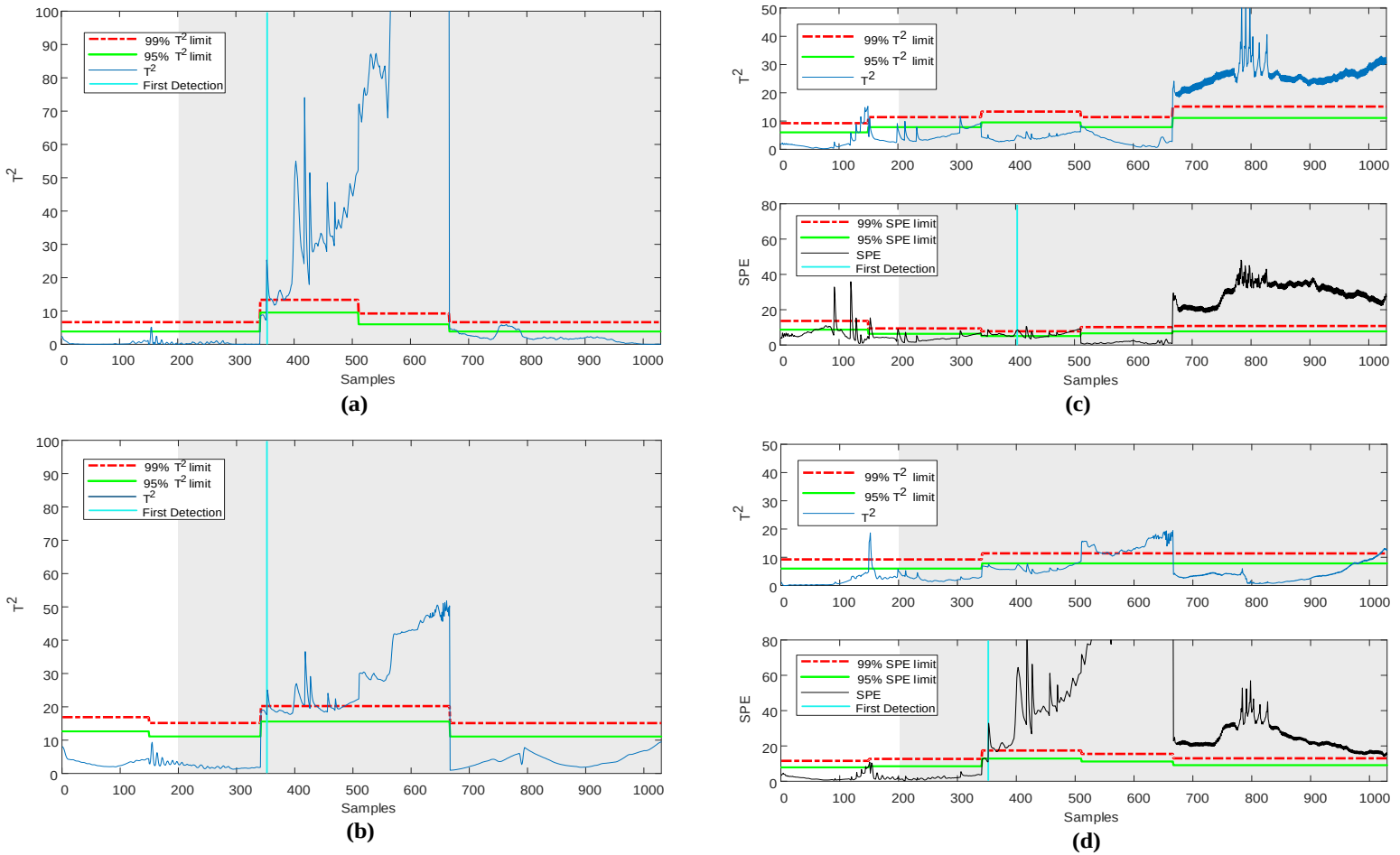

Figure 5.11: Metrics obtained using the multi-level multi-factor model for a batch exhibiting a temperature sensor error. (a) T^2 metric for cointegration analysis at the 1st level, (b) T^2 metric for common-trend model at the 1st level, (c) T^2 and SPE metrics for PCA at the 1st level, and (d) T^2 and SPE metrics for PCA at the 2nd level. The turquoise vertical line indicates when the fault was first detected.

It is worth observing that there are some peak points, which appear at around sample number 800. The ramp characteristic changes into a standard input (2%) without any increase at sample number 800. On the other hand, the poor performance of T^2 with nonstationary data is well-known; thus, the use of the SPE for the 2nd level model is recommended. A performance comparison of the given models is given in Table 5.7.

In the second example, a fault was introduced to the substrate feed rate, which is not part of the list of measured variables in Table 5.1. Hence, this fault has to be detected indirectly via those variables that are measured. The substrate (sugar) feed rate (F_s) was changed to $20\ L\ h^{-1}$ at sample 300, which is just before the change from the 3rd to 4th phase in the process.

The results obtained using multi-PCA for detection of a substrate feed error are shown in Figure 5.12. Oscillations in the residuals are common for cointegration based approaches because they model x_t in the long-run with stationary Δx_t variables (see Equation (3.25)) (Chen, Kruger and Leung, 2009; Li, Qin and Yuan, 2014). The oscillations observed in Figure 5.12 arise from variable x_{10}, which denotes the percentage of CO_2 in the outgas that is partly related to the substrate feed rate as discussed in Sections 2.8 and 3.5 of the original publication of Goldrick et al.(2015). Furthermore, the higher oscillations in the cointegration residuals are also reported in several studies such as in the monitoring of the distillation unit (Chen, Kruger and Leung, 2009). Although the multi-PCA model detected some abnormalities at the end of the 2nd phase (sample number 341), they were not well detected and did not continue into the 3rd phase (samples $341 - 511$). This can be related to the characteristics of x_{10} which shows nonstationarity in the 3rd and 4th phases (samples $341 - 667$). Consequently, the first continuous detection of the fault using a multi-PCA model occurred at sample number 480 with some type-II errors.

The T^2 and SPE metrics obtained using a multi-level model for data exhibiting a substrate feed rate error are illustrated in Figure 5.13. Here, the first detection of a fault in the SPE chart occurs at sample number 420; however, it continues with some type-II errors until sample number 450 for the 3rd and 4th phases ($341 - 667$). Even though, x_{10} is stationary for the 5th phase, the PCA model contributes to keeping the SPE metric above the control limit with some type-II errors. It should be noted that the substrate feed rate fault does not have a direct effect on x_{10}. This can be followed

through the monitoring of variables exhibiting higher degree nonstationarity using common-trend models.

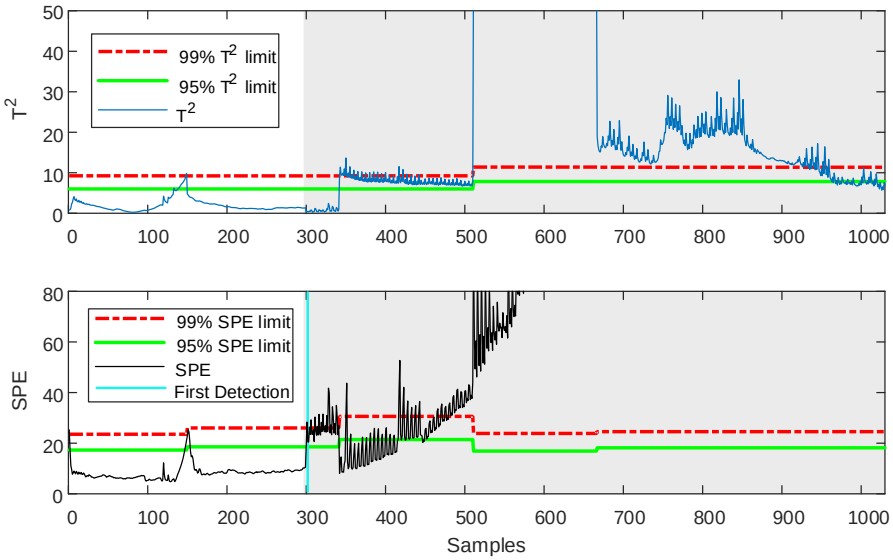

Figure 5.12: T^2 and SPE metrics obtained using multi-PCA for a batch exhibiting a substrate feed rate error. The fault is first detected at sample number 302 using the SPE metric (indicated by turquoise vertical line).

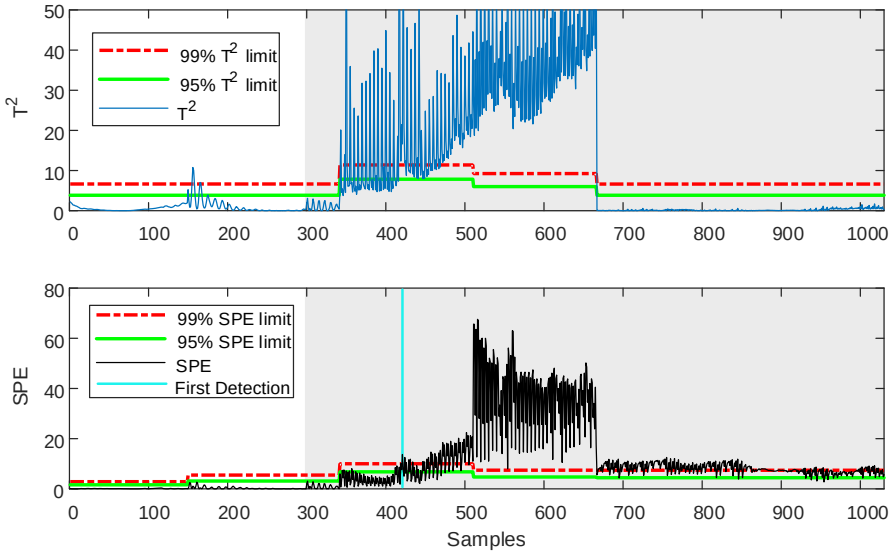

Figure 5.13: T^2 and SPE metrics obtained using the multi-level method for a batch exhibiting a substrate feed rate error. The fault is first detected at sample number 420 using the SPE metric (indicated by turquoise vertical line).

The T^2 and SPE charts obtained using the multi-level multi-factor model for substrate feed rate fault are shown in Figure 5.14(d). The results obtained for the three sub-models at the 1st level and the final model at the 2nd level are given. Figure 5.14(a) shows the T^2 metric for the cointegration residuals-based monitoring results for the nonstationary variables. Figure 5.14(b) shows the T^2 metric for the common-trend residuals for the nonstationary variables. Figure 5.14(c) shows the T^2 and SPE metrics

from application of PCA to the stationary variables where the fault detection is evident via the SPE metric during the 3rd and 5th phases (samples $300-341$ and $667-1000+$). Even though the multi-level multi-factor model uses the same factors as those shown in Figure 5.14(a) and Figure 5.14(c), it is evident that the earlier fault detection observed in Figure 5.14(d) by the multi-level multi-factor model compared to the multi-level model arises from use of the common-trend residuals illustrated in Figure 5.14(c). Note that the rank of the cointegration matrices were found to be $1, 1, 4, 2,$ and 1 for each phase (out of $7, 6, 12, 10$ and 6 nonstationary variables) which shows the existence of higher degree nonstationarity within the nonstationary variables. The rank of the cointegration matrix can be zero or a small number compared to the large number of nonstationary variables that are to be modelled. This is because of the impracticality of cointegration analysis modelling to build linear cointegration relationships when high-level nonstationarity is present. The number of factors used in the common-trend residuals-based monitoring was $6, 5, 8, 8$ and 5, which complement the cointegration matrix assigned by the Johansen test.

In addition to early detection of the error in the substrate feed rate at around sample number 350, the multi-level multi-factor method can extract information from the higher level nonstationary variables as observed in the final phase (samples $667-1000+$) in Figure 5.14(c) and Figure 5.14(d). This compensates for some deficiencies of cointegration modelling when the level of nonstationarity is high. Therefore, the use of the common-trend model is needed when higher nonstationarity levels are considered such as in batch processes.

Table 5.7 summarizes the fault detection results for the test data sets obtained using the three different methods: the multi-level multi-factor, multi-level and multi-PCA models. 99% SPE limits were selected for the comparison as the SPE statistic tends to detect deviations that are not explained by the model. The statistical comparison shows that the multi-level multi-factor method has better performance in terms of both false alarm rate (type-I error) and missing alarm rate (type-II error) in comparison with conventional multi-PCA and the multi-level model.

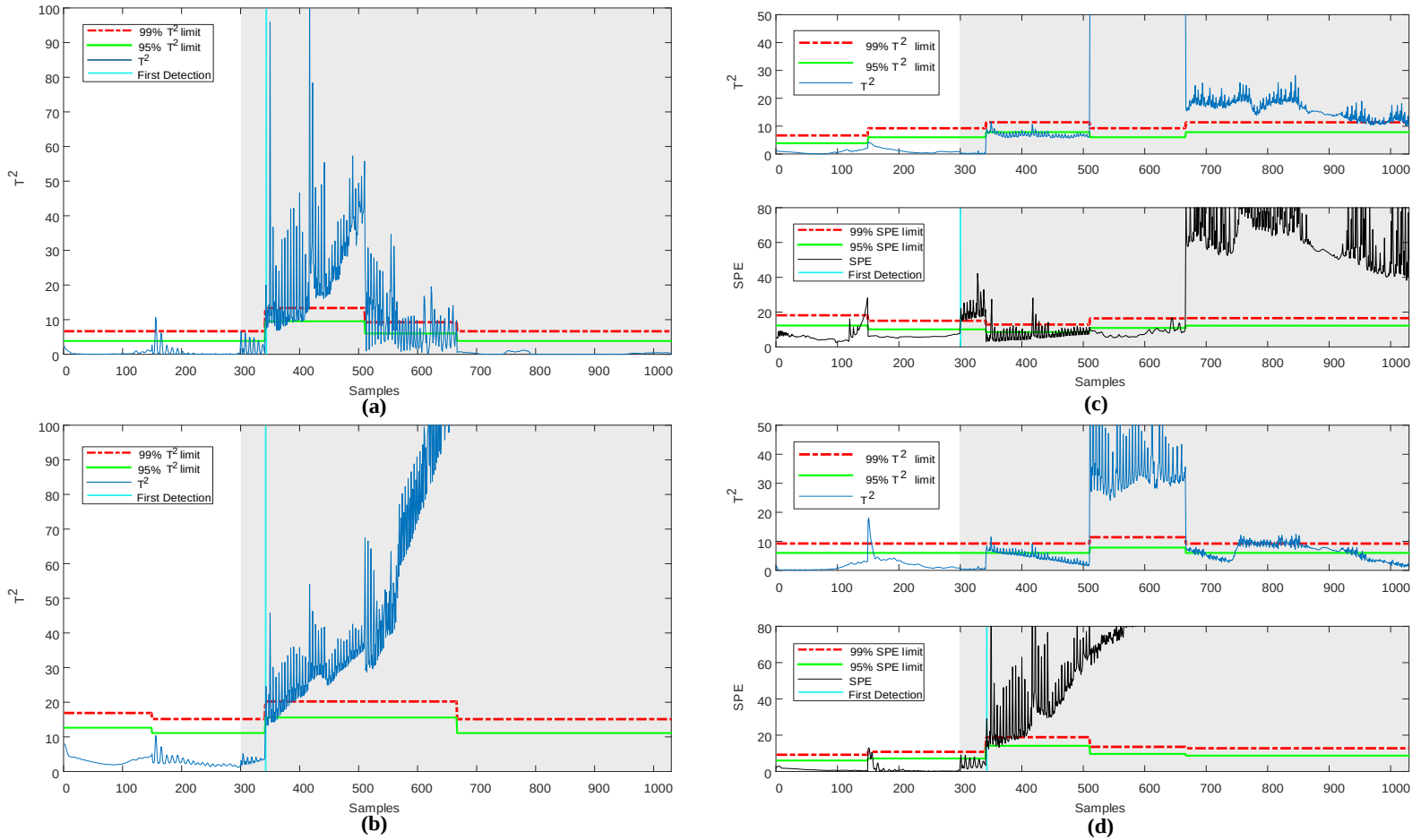

Figure 5.14: Metrics obtained using the multi-level multi-factor model for a batch exhibiting a substrate feed rate error. (a) T^2 metric for cointegration analysis at the 1st level, (b) T^2 metric for common-trend model at the 1st level, (c) T^2 and SPE metrics for PCA at the 1st level, and (d) T^2 and SPE metrics for PCA at the 2nd level. The turquoise vertical line indicates when the fault was first detected.

Table 5.7: Type I and type II errors for the multi-level multi-factor, multi-level, and multi-PCA models for detection of a temperature sensor error and a substrate feed rate error in the test data sets.

Case Study	Type I error (%)		
	Multi-level Multi-factor	Multi-level	multi-PCA
Temperature	–	–	3.2
Substrate Feed	1.3	1.1	–
Case Study	**Type II error (%)**		
	Multi-level Multi-factor	Multi-level	multi-PCA
Temperature	**20.9**	27.3	39.7
Substrate Feed	**10.3**	37.5	19.6

5.5 Conclusions

A multi-level multi-factor process monitoring scheme has been developed that can deal with both stationary and nonstationary variables for multi-phase batch processes. The multi-level multi-factor scheme consists of two levels. In the 1st level, the statistical information underpinning the nonstationary variables is extracted by cointegration analysis and a common-trend model while PCA extracts information from the stationary variables. The final part (2nd level) of the monitoring scheme is based on conventional PCA as the factors from the 1st level models are stationary and hence can be modelled by PCA.

The advantage of the multi-level multi-factor method is that it incorporates a common-trend model as can be seen directly in cases where there is a low rank cointegration matrix which can be caused by high level nonstationary characteristics such as $I(2)$. A low rank cointegration matrix means that fewer variables are monitored by methods such as the multi-level model (Zhang, Zhao and Gao, 2019). Thus it is conjectured that the multi-level-multifactor model has better performance than previously reported models for the monitoring of nonstationary variables such as those found in complex multi-phase batch processes. The multi-level multi-factor method has been used to detect two types of faults in the simulation of the industrial scale penicillin fermentation process, demonstrating its potential for general applicability to the monitoring of batch and semi-batch processes.

This chapter validates the multi-level multi-factor model using data sets gathered from a realistic fermentation process simulation. However, the model needs to be evaluated using data sets gathered from actual industrial processes before it can be considered for real-time process monitoring.

6. PARAMETER TUNING FOR MULTI-LEVEL MULTI-FACTOR MODELLING

6.1 Overview

The multi-level multi-factor model proved its effectiveness on several test cases for both continuous (Chapter 4) and batch processes (Chapter 5). The model has two PCA models; one for each level. This requires optimisation of one design parameter per model, the number of PCs, which can be set on the basis of explained variance.

In the Chapters 4 and 5, the models were designed according to their performance with training data, i.e., the type-I error. However, the proposed model's level structure means that the two models should not be considered independently. The number of the PCs in the 1st level PCA model directly determines the number of t-score vectors, which is one of the sources of the stationary factors for the 2nd level modelling. Designing models separately where there is only one metric to be considered is also time consuming for a data analyst. All of the possible combinations must be tried to find the optimum parameters for the multi-level multi-factor model.

The multi-level multi-factor method models the different phases of batch processes via a series of local models to cope with nonlinearities through the time domain. It also affects the data identification procedure where the characteristics of the variables can change within the determined phases. Furthermore, the model that is responsible for these variables can also be changed with regards to the time period covered by the phases. However, designing the phase length for the multi-level multi-factor model is another exhaustive task for the data analyst considering the possibilities for each time range within the process.

The multi-level multi-factor model truly supports multivariate cases unlike the other cointegration residuals-based approaches (Chen, Kruger and Leung, 2009; Li, Qin and Yuan, 2014; Lin, Kruger and Chen, 2017; Sun, Zhang, Zhao and Gao, 2017). Multiple cointegration models can be used as the source of the stationary factors for the second level modelling when there are more than 12 nonstationary variables. It is known that the cointegration rank depends on the combination of the nonstationary variables involved in the cointegration model where the rank may increase or decrease according to this combination (Harris and Sollis, 2003). Therefore, the performance of the multi-

level multi-factor model may increase through selection of a better combination of the nonstationary variables for each cointegration model. However, this search for the optimum combination is another demanding task for the analyst.

Global optimization algorithms aim to find the best global solution for a model, in the possible presence of multiple local optima. Starting from the genetic algorithm (GA), several heuristic global optimization algorithms have been proposed for different search or optimization problems (Erol and Eksin, 2006). Problems such as that described, which requires optimisation of 3 properties (the number of PCs in the PCA models at both levels, the phase length for batch processes and selection of the nonstationary variables for each cointegration model when there are more than 12 nonstationary variables), can be solved by using a global optimization algorithm. The big-bang big-crunch (BB-BC) algorithm is one of the most promising optimization algorithms to tackle problems such as this.

In the MSPC literature, global optimization algorithms have been used in several cases. For example, Gao and Hou (2016) used a GA along with support vector machines (SVMs) for fault diagnosis, based on clustering approaches, with the TEP; faults were detected using PCA models. GA-based kernel PCA (KPCA) was studied by Jiang and Yan (2018) to search for the optimum variables for inclusion in the KPCA model, which is arguably unnecessary for PCA when used as a MSPC technique. Particle swarm optimization (PSO) has been used to train artificial neural networks (ANNs) and SVMs for the modelling of vibration data (Samanta and Nataraj, 2009). Similarly, Jia et al. (2012) used a GA to search for the optimum KPCA parameters for the monitoring of a penicillin simulation. Xie and Kruger (2006) used PSO to tackle the convergence problem of independent component analysis (ICA). These examples show the use of global optimization algorithms for the exhaustive search of the parameter space.

In this chapter, a parameter tuning method based on the BB-BC global optimization algorithm, is proposed for optimisation of the parameters for the multi-level multi-factor model. It manages several parameter search problems, as mentioned above. The application of the method to the examples presented in Chapters 4 and 5 and the TEP will be evaluated to show its capability with both continuous and batch processes.

6.2 Big Bang-Big Crunch Optimization Algorithm

The big bang-big crunch (BB-BC) optimization algorithm is a global optimization algorithm inspired by the formation of the universe, namely Big Bang and Big Crunch theories (Erol and Eksin, 2006). Big Bang is the first step where the solution candidates are randomly distributed over the search space. It forms the population matrix by choosing the solution candidates. It is then followed by the Big Crunch phase where a contraction procedure calculates a centre of mass for the defined population as described in Table 6.1.

Table 6.1: Big Bang-Big Crunch optimization algorithm.

Step 1 (*Big Bang Phase*): M candidates are initially generated at random in the limited search space.
Step 2: The cost function values of all population are computed.
Step 3 (*Big Crunch Phase*): The centre of mass is calculated.
Step 4: New population is calculated around the new point assigned in Step 3.
Step 5: Go to Step 2 until the stopping criteria is met.

The first population of Big-Bang is randomly generated over the entire search space like any other evolutionary search algorithm. The populations for each iteration are randomly distributed around the centre of mass or the best fit individual candidate. In the Big Crunch phase, a contraction procedure is applied. Here, each individual population member is associated with a cost function value, which determines the best fit value for the next starting point of the Big Bang phase. Another strategy for the determination of the future population is the use of the centre of the mass instead of the best fit:

$$\boldsymbol{x}_c = \frac{\sum_{i=1}^{M}(1/f^i)\boldsymbol{x}_i}{\sum_{i=1}^{M}(1/f^i)} \tag{6.1}$$

where $\boldsymbol{x}_c$ and $\boldsymbol{x}_i$ are the vectors of the centre of the mass and the candidate, respectively, f^i is the cost function value of $\boldsymbol{x}_i$ and M is the population size. The new generation of the population for the next Big Bang iteration is normally distributed around $\boldsymbol{x}_c$:

$$\boldsymbol{x}_{new} = \boldsymbol{x}_c + \frac{r\alpha(\boldsymbol{x}_{max} - \boldsymbol{x}_{min})}{k} \tag{6.2}$$

where r is a normal random number, and α is a limiting parameter for the size of the search space. x_{max} and x_{min} are predefined upper and lower limits for the search space, respectively, and k is the number of iterations. Here, the connection between iteration number and the limits for the prescribed search boundaries, provides a shrink searching space close to the stopping criteria for the fine search.

As is the case for any evolutionary optimization algorithm, the BB-BC algorithm continues until a specified stopping criteria is met. Commonly used stopping criteria are: (i) maximum number of iterations, (ii) maximum run time, and (iii) minimum convergence goal for the fitness value.

In the original paper describing the BB-BC optimization algorithm, simulation results on benchmark test functions were reported and compared to those obtained with a GA; this demonstrated the quick convergence capability of the BB-BC algorithm. Thus, the BB-BC optimization algorithm was chosen for this study.

6.3 Parameter Tuning for Multi-level Multi-factor Model

6.3.1 Parameters

Multi-level multi-factor modelling of continuous and batch was presented in Chapter 4 and Chapter 5, respectively. For the sake of generalisation, the following sections describes optimisation of the multi-level multi-factor model for batch processes, as it combines several continuous models (one for each phase of the batch process). When considering its application to continuous processes, $c = 1$.

The T^2 and SPE metrics are calculated for the multi-level multi-factor model. The $(T^c)^2$ statistic is established to monitor the dominant subspace, which is the PCs space for the c^{th} phase according to Equation (5.28), and the SPE statistic is established to monitor the residual subspace according to Equation (5.30) using R_{Sec}^{C} as the number of PCs for the 2nd level PCA. The metrics provide local modelling of the defined phases in the 2nd level model. The metrics available at the 1st level of modelling vary because of the data characteristics and sub-models built to deal with these characteristics. Therefore, the optimization algorithm utilises the metrics from the 2nd level of the multi-level multifactor model as the response factor for the evaluation of the fault detection capability.

Multi-level multi-factor modelling has four design spaces that can be changed depending on the detection requirements. They are indicated in Figure 6.1 by the uppercase letters, A to D, and are highlighted with grey shading. These are: (A) the number of the PCs in the 1st level PCA model by choosing the minimum CPV (p_1), which models stationary variables, (B) the number of the PCs in the 2nd level PCA model by choosing the minimum CPV (p_2), which models the stationary factors gathered from the 1st level sub-models, (C) the number and length of the phases, which determines the variables characteristics, and (D) selection of the nonstationary variables for each cointegration model ($\boldsymbol{Ord}$) when there are more than 12 nonstationary variables. These searches can be combined to improve performance of the optimum model apart from (C), which is only applicable to batch processes.

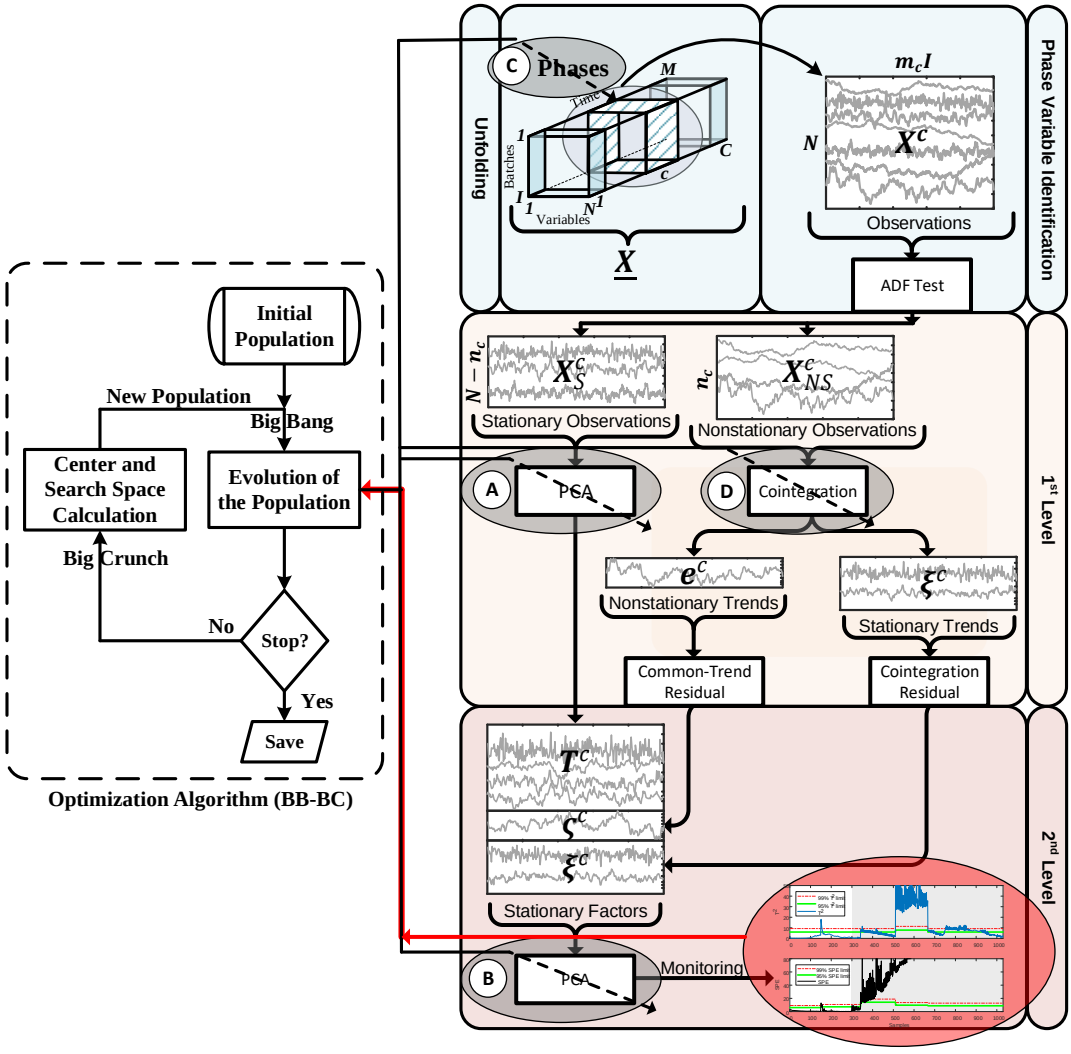

Figure 6.1: Illustration of the parameter tuning scheme for the four design spaces (denoted A to D) of a multi-level multi-factor method for the monitoring of batch processes, where (A) is the number of PCs for the 1st level PCA, (B) is the number of PCs for the 2nd level PCA, (C) is the phase lengths for the batch processes, and (D) is the selection of the nonstationary variables for each cointegration model.

Given features affect the fault detection capability of the multi-level multi-factor model according to the following decision logic, which is used to determine the status of the process:

$$\begin{cases} (T^c)^2 < (T^c)^2_{UCL} \; and \; SPE^C < SPE^C_{UCL} \Rightarrow Fault-free \\ (T^c)^2 \geq (T^c)^2_{UCL} \; or \; SPE^C > SPE^C_{UCL} \Rightarrow Fault \; alarm \end{cases} \tag{6.3}$$

where $(T^c)^2_{UCL}$ is the upper control limit for the T^2 statistic for the c^{th} phase, SPE^C_{UCL} is the upper control limit for the SPE statistic for the c^{th} phase. Fault detection is conducted based on the SPE statistic in Chapter 5 including phase information. Considering the same approach here, the fault detection depends on

$$\boldsymbol{SPE^C} = \|\boldsymbol{I} - \widetilde{\boldsymbol{P}}^c\left(\widetilde{\boldsymbol{P}}^c\right)^{\boldsymbol{T}}\widetilde{\boldsymbol{X}}^c\| \tag{6.4}$$

where $\widetilde{\boldsymbol{P}}^c \in \mathbb{R}^{R^c_{Sec}\times(n_c+R^c_S)}$ is the loading matrix of the 2nd level PCA model, R^c_{Sec} is the number of the PCs in the 2nd level PCA model, n_c is the number of the nonstationary variables, and R^c_S is the number of the PCs in the 1st level PCA model. Monitoring based on the SPE^c statistic depends on the control limits (SPE^c_{UCL}) determined in Equation (5.20) where

$$\theta_i = \sum_{k=R^c_{Sec}+1}^{N} \lambda_k^i, \qquad i = 1,2,3$$
$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{6.5}$$

It is worth noting that the number of PCs determined on the basis of the CPV, given in Equation (2.7), affects both SPE^c and SPE^c_{UCL}. However, it is subject to $1 \leq R \leq N-1$ where R is the number of PCs and N is the number of variables. Given limits helps to have a loading vector $\left(\widetilde{\boldsymbol{P}}^c \in \mathbb{R}^{N\times R}\right)$ where $R \geq 1$. Furthermore, having $N-1$ eigenvalues provides real SPE control limits by leaving at least one eigenvalues outside of the model. Following determination of the SPE metric, it can be seen that monitoring based on type-I and type-II error rates is highly dependent on the number of PCs in both PCA models, involved in the multi-level multi-factor modelling, and the phase lengths. Therefore, a systematic search of these parameters will further improve the performance of the multi-level multi-factor model for process monitoring. To provide this search, the BB-BC global optimization algorithm is applied to the

multi-level multi-factor method for the monitoring of batch processes as illustrated in Figure 6.1.

In the global optimization-based parameter tuning procedure, the manipulated variables consist of the number of PCs, the phases lengths and combination of the nonstationary variables for each cointegration model if the number of nonstationary variables exceed 12. The number of PCs can be defined through the minimum CPV and can take values between $[0 \text{ and } 100]$ but the value should be between the value of CPV for the first and last eigenvalues of the related PCA model. The maximum number of phases must be limited before performing the search by assigning (C_{cons}). It is possible to find a lower number of phases than the maximum number. However, having no limit for the number of phases may negatively affect the optimization performance. As also mentioned in Chapter 5, different batch data sets are cut to have same number of the samples. Therefore, the search procedure constrains the candidates to have the same number of samples in total. The sum of the phase length candidates $(\boldsymbol{PR})$ must be M where the phase length candidates may vary between $[0 \text{ and } M]$.

Another search procedure may be required when the number of nonstationary variables exceeds 12, which is the maximum number of variables supported by Johansen's test for cointegration analysis. In this case, more than one cointegration analysis model must be established on the nonstationary variables. Here, two factors can affect the performance of the multi-level multi-factor model: the number of nonstationary variables for each cointegration analysis model and the variables chosen for the models. It is known that the combination of the nonstationary variables may affect their nonstationarity characteristics (Harris and Sollis, 2003). Therefore, searching for how many cointegration analysis models are needed with which nonstationary variables constitutes another parameter tuning problem.

6.3.2 Cost Function

A cost function defines an objective to be reached by the optimization algorithm. This can be reachable or impossible to reach by the algorithm. The algorithm stops searching if the defined cost value is reached. In this study, the main goal is to improve the false alarm rate through training and test data sets, which are given to the global optimization algorithm to build an optimized multi-level multi-factor model. The data sets must cover all the possible characteristics of the process and errors. A cost

function can be defined with only one set of type-I and type-II errors. However, a truly global optimisation must have several data sets, which cover all working conditions and the defined error characteristics within the process. Cost function can be defined for the training data set, which includes data sets from different scenarios; faulty and fault-free. Assume that K different data sets represent two zones (fault-free and faulty) through the time-scale where $0 < i < t_f$ covers the fault-free zone and $t_f \leq i < t_{end}$ covers the faulty zone for time index i. Consequently, a cost function can be defined as follows:

$$J(p_1, p_2, \boldsymbol{PR}, \boldsymbol{Ord}) = \sum_{k=1}^{K} \sum_{i=1}^{M} \frac{j_{k,i}}{M} \tag{6.6}$$

where M is the number of the samples for the data set from scenario k and $j_{k,i}$ is:

$$\begin{aligned} \underset{0<i<t_f}{j_{k,i}} &= \begin{cases} 1, & if\ SPE^c > SPE^c_{UCL} \\ 0, & if\ SPE^c \leq SPE^c_{UCL} \end{cases} \\ \underset{t_f<i<t_{end}}{j_{k,i}} &= \begin{cases} 1, & if\ SPE^c < SPE^c_{UCL} \\ 0, & if\ SPE^c \geq SPE^c_{UCL} \end{cases} \end{aligned} \tag{6.7}$$

where SPE^C is dependent on $\widetilde{\boldsymbol{P}}^c \in \mathbb{R}^{R^c_{Sec} \times (n_c + R^c_S)}$ and $\widetilde{\boldsymbol{X}}^c = [\boldsymbol{\xi^c}, \boldsymbol{T^c_S}, \boldsymbol{\varsigma^c}]$.

$$\begin{aligned} R^c_S &= \min_r \left(\left(\sum_{i=1}^{r} \lambda_i \Big/ \sum_{i=1}^{N-n_c} \lambda_i \right) 100 \geq p_1 \right) \\ R^c_{sec} &= \min_r \left(\left(\sum_{i=1}^{r} \lambda_i \Big/ \sum_{i=1}^{n_c+R^c_S} \lambda_i \right) 100 \geq p_2 \right) \end{aligned} \tag{6.8}$$

where p_1 and p_2 are the selected minimum cumulative percentage of variances for the 1st and 2nd PCA models in the multi-level multi-factor method. Both variables can take values between $[0, 100]$ by applying the condition above. $1 < c < C$ represents the related phase division, which can also be tuned by design space C to find phase lengths ($\boldsymbol{PR} = (m_1, \dots, m_c, \dots, m_C)$). Therefore, a phase length (m_c) can take a value between $[0, M]$. Furthermore, if $n_c > 12$, the partition of $\boldsymbol{X}^c_{NS} \in \mathbb{R}^{n_c \times m_c}$ into the matrices by $\boldsymbol{Ord}$ up to 12 variables for each Johansen model affects the determination of $\boldsymbol{\xi^c}$ and $\boldsymbol{\varsigma^c}$. Consequently, the optimisation problem can be defined as:

$$\underset{p_1,p_2,\boldsymbol{PR},\boldsymbol{Ord}}{\operatorname{argmin}} \; J(p_1,p_2,\boldsymbol{PR},\boldsymbol{Ord})$$

$$s.t. \quad \left(\lambda_1 \Big/ \sum_{i=1}^{N-n_c} \lambda_i\right) 100 < p_1 < \left(\sum_{i=1}^{N-n_c-1} \lambda_i \Big/ \sum_{i=1}^{N-n_c} \lambda_i\right) 100$$

$$\left(\lambda_1 \Big/ \sum_{i=1}^{n_c+R_S^c} \lambda_i\right) 100 < p_2 < \left(\sum_{i=1}^{n_c+R_S^c-1} \lambda_i \Big/ \sum_{i=1}^{n_c+R_S^c} \lambda_i\right) 100 \tag{6.9}$$

$$C = C_{cons} \parallel C < \frac{M}{2}; \; \sum_{j=1}^{C} m_j = M$$

$$\boldsymbol{Ord} = (\boldsymbol{Ord}_1, \dots, \boldsymbol{Ord}_l), l \leq \frac{n_c}{2}; if \; o \in \boldsymbol{Ord_1}, then \; o \notin \boldsymbol{Ord} - \boldsymbol{Ord_1}$$

where C_{cons} is a constant selected for the maximum number of phases which can be assigned before parameter tuning. l represents the number of the Johansen model that can be assigned before parameter tuning. The tuning of $\boldsymbol{Ord}$ can also be seen as a distribution of the nonstationary variables into the Johansen models. Therefore, an element of $\boldsymbol{Ord}$ can be an integer value between $[0, n_{ns})$. The conditions given above cover all search spaces (A to D) for the tuning of the parameters $(p_1, p_2, \boldsymbol{PR}, \boldsymbol{Ord})$. Therefore, the parameters must be considered separately if only specific search spaces are of interest. For example, only p_1 and p_2 would be considered for consideration of search spaces A and B.

6.4 Application to a Continuous Stirred Tank Heater

6.4.1 Model Optimisation

A simulation of the continuous stirred tank heater was described and studied in Chapter 4. The parameters that were searched by the BB-BC optimization algorithm for the CSTH data sets are defined in Section 6.3.1. Here, two design spaces, A and B were searched using the parameter tuning method represented in Figure 6.1. The characteristics of the variables in the training data set cannot be changed throughout the process as the training data set represents the normal operating conditions of the process. Therefore, there is no need to check for changes in the nonstationarity of the variables. The multi-level multi-factor model for the CSTH simulator comprised two PCA models; one for each level and 45% of the variance was retained in both models in Chapter 4. The percentage of variance to be retained in the two PCA models was optimized using data representing normal operating conditions and that exhibiting a

temperature sensor fault simulated using a step type function, which changed the inlet hot water temperature by $+1$ ℃ . Therefore, the number of scenarios that were used in parameter tuning was $K = 2$. Parameter tuning via the BB-BC optimization algorithm was achieved using 10 iterations for every 20 populations of candidate variables. The number of iterations varies according to a linear relationship with the number of manipulated parameters. This is also related to the complexity of the search space. Tuning of the two parameters (p_1, p_2) required 10 iterations and took 83.2 seconds. The change in the cost function with the number of iterations for the optimum model parameters can be seen in Figure 6.2.

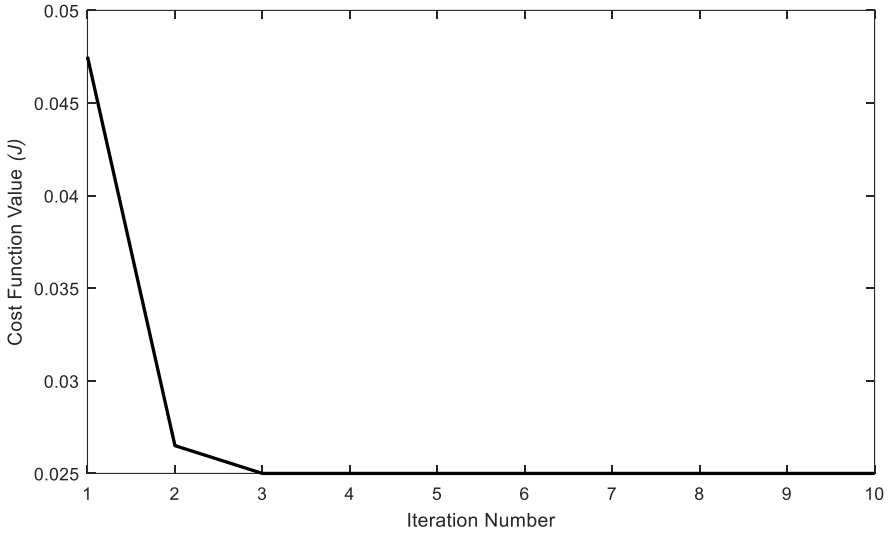

Figure 6.2: Cost function for the multi-level multi-factor model of data from the CSTH.

In the multi-level multifactor model developed in Chapter 4, 45% of the variance was retained in both PCA models. In comparison, the BB-BC algorithm suggested retention of 85.5% and 58.8% of the variance in the data for the 1st level and 2nd level PCA models, respectively, in the optimum multi-level multi-factor model. The results obtained for the training data using the suggested parameters are shown in Figure 6.3. The explained variance of the training data set is given in Section 4.4.1 where 4 PCs were required to describe 85.5% of the variance in the data at the 1st level and 4 PCs are required to describe 58.8% of the variance in the data at the 2nd level of the multi-level multi-factor model.

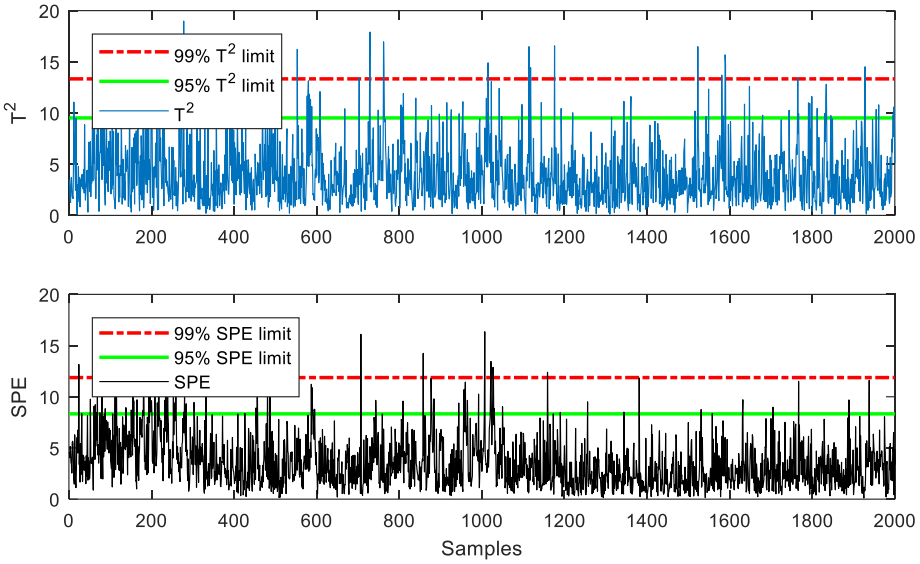

Figure 6.3: T^2 and SPE metrics for the optimum multi-level multi-factor model built using training data from the CSTH exhibiting normal operating conditions.

A temperature sensor error that exhibits a step function type fault was used to tune the parameters. The T^2 and SPE metrics are shown in Figure 6.4. The optimum model provides better detection of both types of faults. For the step function type fault, the type-II error rate is reduced to 1.4% from 5% in the SPE control chart. The results are compared in Table 6.2 where the model that was developed in Chapter 4 is termed a fixed-variance model.

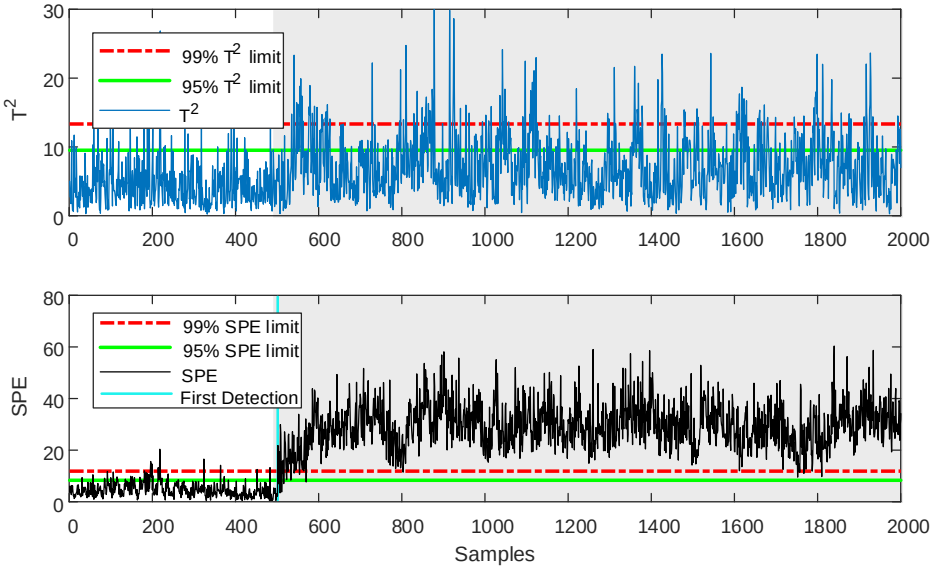

Figure 6.4: T^2 and SPE metrics for the optimum multi-level multi-factor model with parameter tuning data from the CSTH that exhibits a step function type fault. The fault was first detected at sample number 502 using the SPE metric (indicated by turquoise vertical line).

Table 6.2: Online diagnosis performance of the fixed variance and optimum multi-level multi-factor models for the CSTH process.

Case	Metric Name (Confidence Limit%)	Error Type	Models	
			Fixed variance	Optimum
Step	SPE(1%)	Type-I	**0**	**0**
	T^2(1%)		**0**	**0**
	SPE(1%)	Type-II	5	**1.4**
	T^2(1%)		97	**89.2**
Ramp	SPE(1%)	Type-I	**0.4**	0.6
	T^2(1%)		2.60	**1**
	SPE(1%)	Type-II	0.8	**0.7**
	T^2(1%)		3	**1.1**

6.4.2 Model Performance

Evaluation of the capability of the optimum model for fault detection was evaluated using data exhibiting a ramp function type fault, which is not used for the parameter tuning. The fault is a valve malfunction on the flow of cold water. The T^2 and SPE metrics are shown in Figure 6.5 for the ramp function type fault. The results are compared in Table 6.2 where the model that was developed in Chapter 4 is termed a fixed-variance model. For the ramp function type fault, the early detection performance was improved and the type-II error rate is reduced to 0.7% from 0.8% in the SPE chart. Here, it can be seen that use of a higher level of explained variance in the 1st level PCA model improves the performance of the multi-level multi-factor model. Rather than testing the metric performance of each possible portion of variance explained by the models, the parameter tuning scheme can seek the connection between different design parameters, which explained the percentage of variance for both PCA models. The suggested number of PCs shows that the 1st level PCA model should not be considered as a conventional PCA model for process monitoring. The use of the 1st level PCA model should aim to extract process dynamics into t-scores,

which is one of the sources of the stationary factors for the 2nd level PCA model. Furthermore, the 2nd level PCA model can represent the process dynamics, conventional PCA modelling where the explained variance level is selected to be around 50%.

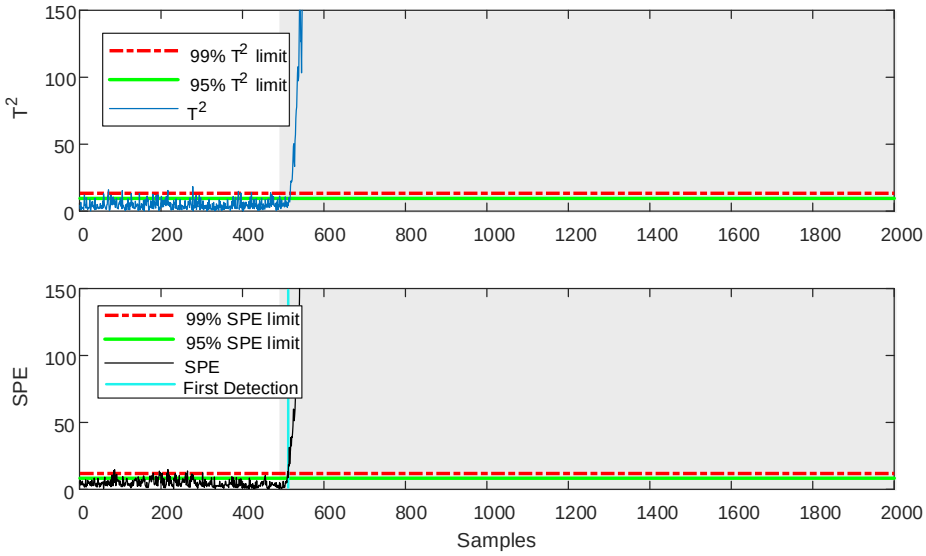

Figure 6.5: T^2 and SPE metrics for the optimum multi-level multi-factor model with the test data from the CSTH that exhibits a ramp function type fault. The fault was first detected at sample number 516 using the SPE metrics (indicated by turquoise vertical line).

6.5 Application to the Industrial Penicillin Simulator

6.5.1 Model Optimisation

The industrial penicillin simulator was described in Chapter 5. The multi-level multi-factor model designed for the industrial penicillin simulator includes 5 phases and the number of nonstationary variables for each phases does not exceed 12. Thus, 3 design spaces, A, B and C, were searched using the parameter tuning scheme represented in Figure 6.1. As the process is of the batch type and the expected number of nonstationary variables for each phase is lower than 12, only phase length and the number of the PCs for the 1st and 2nd level PCA models in the multi-level multi-factor model are subject to the optimization procedure.

The data in Chapter 5 is used here for the optimisation study. An additional test data set was simulated to provide 3 fault examples for evaluation of model performance in Section 6.5.2. The data for parameter tuning comprised 9 standard batches, which were cut to a length of 200 hours, for training and two test data sets. Therefore, the number

of scenarios was assigned as $K = 11$. The first fault comprised a monitoring error on the temperature sensor, which affects directly the variables that are given in Table 5.1. The fault shows ramp function characteristics starting from the first initiation of the fault at sample 200. The second fault was an error on the substrate feed rate starting from sample number 300, which was not monitored directly but has an indirect effect on the variables that were monitored (see Table 5.1). Another fault on the substrate feed rate was also defined from sample number 500, and used in the parameter tuning procedure. Section 5.4 describes the phases and their lengths, and the fault characteristics. The five phases adopted in this study were selected as $(0-30, 30-68.2,\ 68.2-102.2, 102.2-133.4, 133.4-200\ hrs+)$. The number of PCs was chosen on the basis of explained variance which was set to 45% for all of the PCA models in Chapter 5.

Global optimization via the BB-BC algorithm was achieved using 500 iterations for every 10 populations of candidate variables. Tuning of the three parameters $(p_1, p_2, \boldsymbol{PR})$ required 500 iterations and took ~ 190 minutes; C_{cons} was set to 5 for the sake of simplicity. The change in the cost function with the number of iterations for the optimum model parameters can be seen in Figure 6.6.

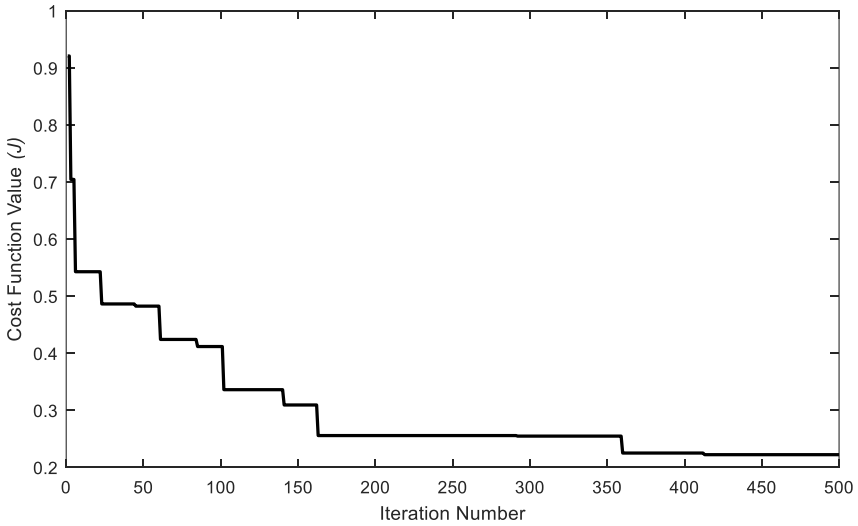

Figure 6.6: Cost function for the best result, which gives the optimum parameters for the multi-level multi-factor model of data from the industrial penicillin simulation.

The phase lengths suggested by the BB-BC algorithm are given in Table 6.3, and compared to those utilised in Table 5.5 in Chapter 5, which were selected on the basis of the different stages in a fermentation process (i.e. expert user knowledge). There are some changes in the number of nonstationary variables when the phases are selected

through parameter tuning due to changes in the phase lengths. The 1st phase is shortened to 3.6 hours compared to 30 hours. The 2nd phase is between $3.6 - 35.6$ hours, and shows similar variable characteristics in terms of nonstationarity to those in the 1st phase. The 3rd phase identified by the parameter tuning procedure has close time boundaries to those of the 2nd phase identified by the expert user. However, there is also some overlap with the 3rd phase in terms of the variables exhibiting nonstationary characteristics. Similar to the 3rd phase of the optimum model, the 4th phase of the optimum model has close time boundaries with the 3rd phase identified by the expert user. The final phase identified by parameter tuning is a combination of the 4th and 5th phases, identified by the expert user.

Table 6.3: Nonstationary variables in each phase for the fixed variance (with expert user knowledge) and optimum model for the industrial penicillin simulator.

	Fixed Variance Model (with expert user knowledge)		**Optimum Model**	
No	**Phase Range (PR)**			
	In hours	In samples	In hours	In samples
1	$0 - 30$	$0 - 150$	$0 - 3.6$	$0 - 18$
2	$30 - 68.2$	$150 - 341$	$3.6 - 35.6$	$19 - 178$
3	$68.2 - 102.2$	$341 - 511$	$35.6 - 74.4$	$178 - 372$
4	$102.2 - 133.4$	$511 - 667$	$74.4 - 108.4$	$372 - 542$
5	$133.4 - 200 +$	$667 - 1000 +$	$108.4 - 200 +$	$542 - 1000 +$
No	**Nonstationary Variables**			
1	$[x_1, x_6, x_7, x_8, x_{12}, x_{13}, x_{18}]$		$[x_5, x_6, x_7, x_8, x_9, x_{10}, x_{17}, x_{18}]$	
2	$[x_1, x_8, x_{12}, x_{13}, x_{17}, x_{18}]$		$[x_1, x_6, x_8, x_{12}, x_{13}, x_{17}, x_{18}]$	
3	$[x_1, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}]$		$[x_1, x_6, x_7, x_9, x_{10}, x_{11}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}]$	
4	$[x_1, x_2, x_3, x_5, x_9, x_{10}, x_{12}, x_{13}, x_{17}, x_{18}]$		$[x_1, x_3, x_5, x_8, x_9, x_{12}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}]$	
5	$[x_4, x_6, x_{12}, x_{13}, x_{17}, x_{18}]$		$[x_1, x_4, x_6, x_{12}, x_{13}, x_{17}]$	

In comparison with the phase assignment by an expert, the optimum model shortened the length of the 1st phase (from $0 - 150$ to $0 - 18$) and extended the length of the 5th phase (from $667 - 1000 +$ to $542 - 1000 +$). This being the case, the middle phases were of comparable length for the optimum model. More detailed searches can also be

performed using a greater number of phases; the number of phases was initially assigned as 5 for this search as the fixed variance model that utilised expert user knowledge comprised 5 phases.

The rank of the cointegration matrix for the 5 different phases was determined as: $[1, 1, 5, 2, \text{and } 3]$ using the Johansen test. This allows the use of the common trend model as the number of nonstationary variables for each phase is: $[7, 6, 6, 9, \text{and } 3]$. Here the rank of the cointegration models, which was $[1, 1, 4, 2, \text{and } 1]$, is larger compared to that in the fixed-variance multi-level multi-factor model. The number of nonstationary factors for the common-trend model is $[6, 5, 8, 8, \text{and } 5]$. The optimum model shows that a longer 5^{th} phase helps to establish 2 more cointegration relationships between nonstationary variables. Another change occurs in the 3^{rd} phase (samples $178 - 372$, which is similar to 2^{nd} phase used in Chapter 5 (samples $150 - 341$)). If the limits for both phases are assumed to be the same, the rank of that phase increases to 5 from 1. Even though some variables were identified as nonstationary, this does not mean that they can establish a cointegration relationship easily. Higher level nonstationarities make construction of a model harder for $I(1)$ nonstationary series. The parameter tuning scheme helps to search optimum phase lengths, which helps the cointegration models to increase their performance in regards to building linear relationships between nonstationary variables.

In addition to the phase lengths, the number of PCs were also searched through the observation of the percentage of the explained variance. Unlike the fixed variance models where 45% of the variance in the data was retained for both models, retention of 84.55% and 19.05% of the variance in the data was suggested for the 1^{st} level and 2^{nd} level PCA models. The results obtained for the training data using the suggested parameters are shown in Figure 6.7. Figure 6.7(a) and Figure 6.7(b) show the T^2 metrics from cointegration analysis and the common-trend residuals for the nonstationary variables. The last sub-model from the 1^{st} level of the multi-level scheme is shown in Figure 6.7 (c), which shows the T^2 and SPE metrics of the PCA model for the stationary variables only. The optimally designed PCA models in terms of PCs are tabulated in Table 6.4. In comparison with the fixed variance model training performance, the type-I error rate is reduced to 1.12% from 1.80%.

Similar to the CSTH example, the parameter tuning scheme suggested retention of a high portion of the variance for the 1^{st} level PCA model and a low portion of the

variance for the 2nd level PCA model. In comparison to the CSTH example, a 3rd design space (C) was searched using the parameter tuning scheme in addition to searching for the optimum number of PCs (design space (A) and (B)). The type-I error rate for the optimum model with the training data set was 0.8 whereas that for the fixed variance model presented in Chapter 5 was 0.7 (see Table 5.6). As can be seen from the error rates, changes in the percent variance for the 1st and 2nd level models did not affect the performance of the multi-level multi-factor model. Having a high portion of variance in the 1st level PCA model provided all of the extracted dynamics into the 2nd level PCA model via the stationary factors. It should be noted that, due to the full rank feature of the combination of the cointegration model and common-trend model, modelling of nonstationary variables within the multi-level multi-factor model does not have any dimensionality reducing features. This being the case, the t-scores must evolve into stationary factors as much as possible because the univariate dynamics from each factor becomes a part of the multivariate dynamics. Therefore, loss of dynamic information in the stationary variables modelled by 1st PCA is possible, if a small number of PCs is selected. The given optimisation results also support this approach.

Table 6.4: Number of principal components and corresponding variances for the optimum multi-level multi-factor model for the industrial pennicillin simulator.

Model Name	Phase Number	Number of PCs	Variance	Variance in the Data (%)
1st level PCA	1	4	88.4	**53.1, 16.7, 10.7, 8.0**, 4.4
	2	3	84.6	**54.3, 20.4, 9.9**, 6.4, 3.5
	3	5	90.7	**30.5, 23.9, 18.0, 10.8, 7.4**
	4	4	92.6	**47.9, 20.1, 14.1, 10.5, 4.2**
	5	7	87.4	**23.2, 17.4, 14.9, 11.9, 8.8, 6.6, 5.**
2nd level PCA	1	1	36.4	**36.4**, 19.4, 13.0, 10.3, 9.2
	2	1	31.6	**31.6**, 23.2, 19.4, 11.4, 6.7
	3	1	30.6	**30.6**, 12.6, 11.3, 9.7, 8.8
	4	1	23.9	**23.9**, 20.8, 12.9, 11.3, 8.2
	5	1	21.0	**21.0**, 14.5, 13.9, 11.7, 8.9

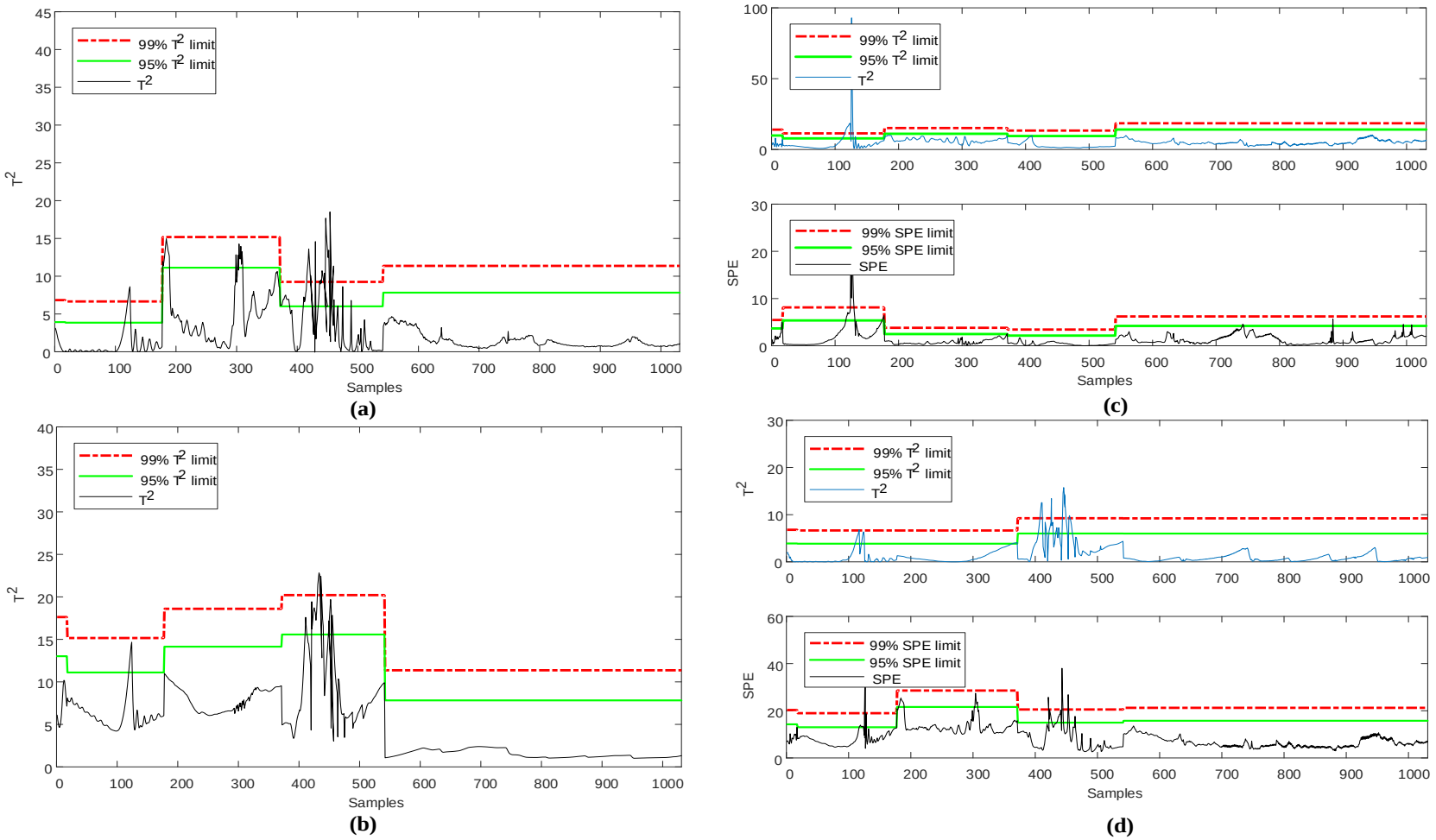

Figure 6.7: Metrics obtained using the optimum multi-level multi-factor model for a training batch exhibiting normal operation. (a) T^2 metric for cointegration analysis at the 1st level, (b) T^2 metric for common-trend model at the 1st level, (c) T^2 and SPE metrics for PCA at the 1st level, and (d) T^2 and SPE metrics for PCA at the 2nd level.

The type-I errors observed in Figure 5.8(d) around sample number 650 are modelled properly in the optimum model (Figure 6.7(d)) due to the merging of the phases. However, the type-I error at sample number 450 still exists. High magnitude type-I errors are reduced by only searching phases that affect the performance of cointegration models. Two faulty process examples, which were analysed in Chapter 5, were used to evaluate the training of the parameter tuning scheme.

The T^2 and SPE charts obtained for data exhibiting a temperature sensor error using the 2nd level PCA model of the optimum multi-level multi-factor model are shown in Figure 6.8 where the fixed variance model results can be found in Figure 5.11(d). Here, the fault shows ramp function type characteristics starting from sample number 200 until the end. The optimum model provides early detection of the fault, while the fixed variance model could not detect it. Detection of the fault occurred approximately 100 samples earlier. However, in the last phase at around sample number 700, some type-II errors occurred for the optimum model. Even though the optimum model detected the fault earlier, due to the performance of the last phase, the type-II error rates for the optimum model are comparable to those for the fixed variance model as tabulated in Table 6.5.

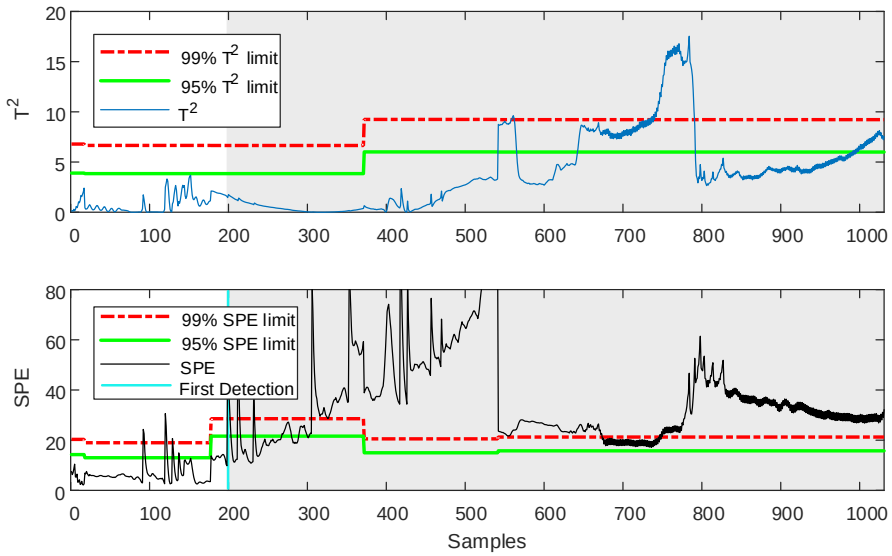

Figure 6.8: T^2 and SPE metrics obtained using the optimum multi-level multi-factor model for a batch exhibiting a temperature sensor error. The fault was first detected at sample number 200 using the SPE metrics (indicated by turquoise vertical line).

The T^2 and SPE charts obtained for data exhibiting a substrate feed error rate case using the 2nd PCA model of the optimum multi-level multi-factor model is shown in Figure 6.9 where the fixed variance model results can be found in Figure 5.14(d). Here,

a substrate feed error represents a step function type fault starting from sample number 300. The optimum model detected the fault at sample number 300, which was earlier than the fixed variance model where the fault was detected at sample number 350. The optimum model helps to reduce the type-II error rate by nearly 30% for this example as tabulated in Table 6.5.

Table 6.5: Type I and type II errors for the fixed variance and optimum multi-level multi-factor models for detection of different types of faults.

Case Study	Type I error (%)		Type II error (%)	
	Fixed variance	Optimum	Fixed variance	Optimum
Temperature	**0**	2.3	**20.91**	21.10
Substrate Feed	**0.01**	**0.01**	10.3	**7.24**
Aeration Rate	1.5	**0**	6.5	**0**
Vessel Pressure	6	**5**	51.01	**0**
Base Flow Rate	**1.57**	1.85	100	**64**

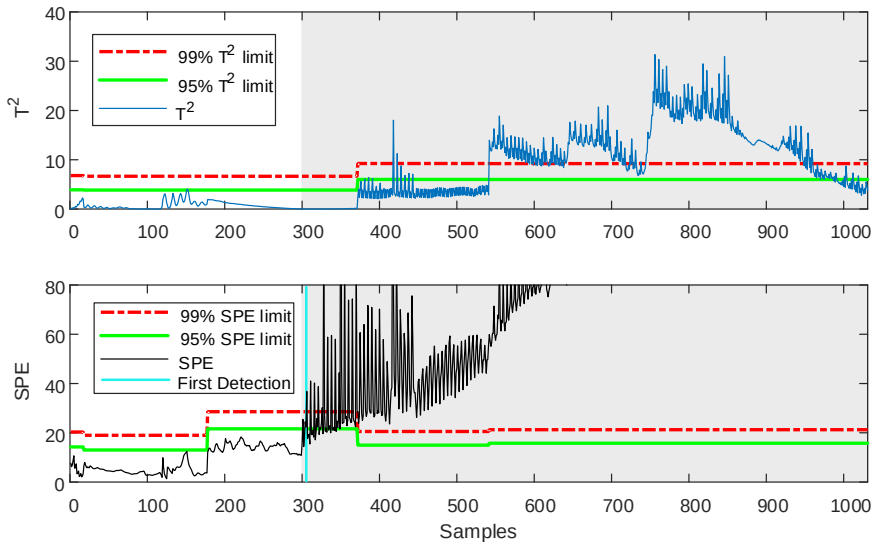

Figure 6.9: T^2 and SPE metrics obtained using the optimum multi-level multi-factor model for a batch exhibiting a substrate feed rate error. The fault was first detected at sample number 300 using the SPE metrics (indicated by turquoise vertical line).

6.5.2 Model Performance

The data exhibiting errors in the temperature sensor and the substrate feed rate were utilised within the parameter tuning step of the optimisation algorithm. Thus, to evaluate the process monitoring capability of the optimised model using data that had not been used within the optimisation procedure, 3 additional batches exhibiting faults were generated using the industrial penicillin simulator. The faults used in this section

were also defined in the simulation package (Goldrick et al., 2015). Therefore, they are used according to their fault type such as ramp or step function type faults, and magnitudes. The performance of the fixed variance and optimum models was then compared.

In the first example, an aeration rate fault, which shows step function type characteristics, was introduced at sample 200 where the rate was set to $22\ Lh^{-1}$. The aeration rate ($F_{g_{in}}$) is not directly monitored by the multi-level multi-factor model as it is not in the monitored variables in Table 5.1. The effects of the aeration rate can be followed via Equation (5.15) where it is related to the oxygen uptake rate ($OUR - x_{13}$) and carbon evolution rate ($CER - x_{15}$). According to the nonstationary variables for each phase in Table 6.3, the only difference in these two variables occurred in the 4th phase where x_{15} showed nonstationary characteristics. A comparison of the fixed variance and optimum models is illustrated in Figure 6.10 and tabulated in Table 6.5. Here, the optimum model provides better detection without any type-I or type-II errors. The fixed variance model detected variation changes at the beginning of the fault but due to differences in the number of PCs and the cointegration models (2nd phase for fixed variance and the cointegration rank equals 1 and 3rd level for the optimum model and the cointegration rank equals 5), the T^2 statistic could not exceed the UCLs.

In the second example, a vessel back pressure fault, which shows step function type fault characteristics, was introduced at sample 600 where the air head pressure was set to $2\ bar$. The pressure is a manipulated variable used in the sequential batch control strategy of the industrial penicillin simulator to control O_2 concentration, which can be followed from Equations (5.14) and (5.15). Similar to previous example it has arguably some effects on OUR and CER. This example was selected to test different types of faults that have not been used in the training of the parameter tuning scheme. The comparison can be observed in Figure 6.11. Compared to Figure 6.11(a), the optimum model detected the fault without any type-II errors whereas the fixed variance model detected the fault only in the 4th phase (samples $511 - 667$) and some other samples around sample number 850. Similar to the previous example, the fixed variance model has a similar performance in the 4th phase (higher magnitude in comparison with the others in the previous example) but the same performance was not continued in the last phase. In contrast, the optimum model assigned only one phase between sample numbers $542 - 1000 +$. This also helped the optimum model

to retain its performance starting from the first sign of the fault to the end. The type-I errors are a problem for both the fixed variance and optimum models but the optimum model shows better performance in terms of the type-I error rates as well.

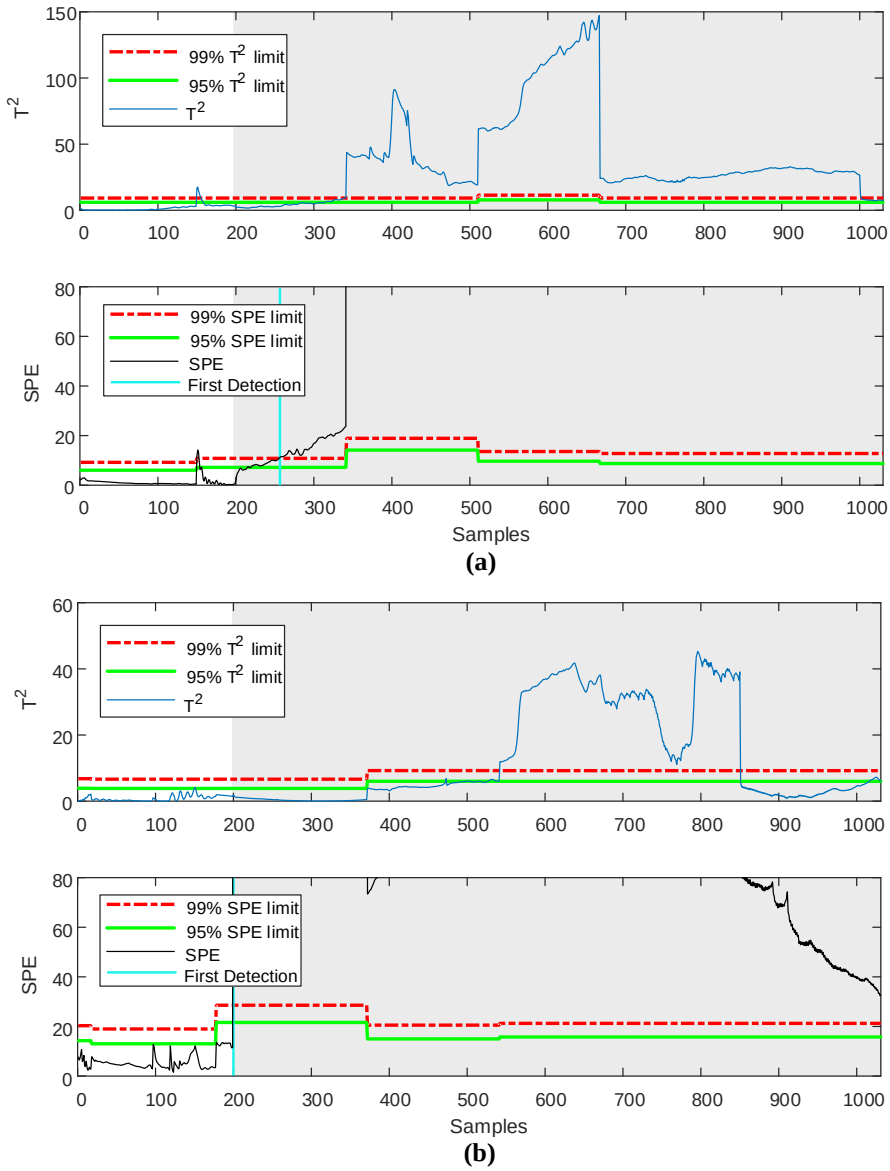

Figure 6.10: T^2 and SPE metrics obtained using (a) the fixed variance, and (b) the optimum multi-level multi-factor model for a batch exhibiting an aeration rate error. The turquoise vertical lines indicate when the fault was first detected.

In the final example, the base flow rate fault, which shows step function type characteristics, was introduced at sample 700 where the base flow rate (F_b) was set to 5 Lh^{-1}. The base solution is a manipulated variable to control pH level and it is activated when the pH level (x_7) has decreased under 0.05 from the pH set-point. x_7 is a stationary variable, according to both phase distributions in the phase that the fault occurred. As with the previous examples, this fault type that is effective in the last

phase was not involved in the parameter tuning example. Sample number 700 occurs in the last phase of both of the models. Thus, the last phase of the modelling performance is compared in this example. As can be seen from Figure 6.12(a), it is not possible to detect the fault at any point using the fixed variance model while the optimum model detects the fault at around sample number 900 but with some type-I errors that are not detected in the fixed variance model.

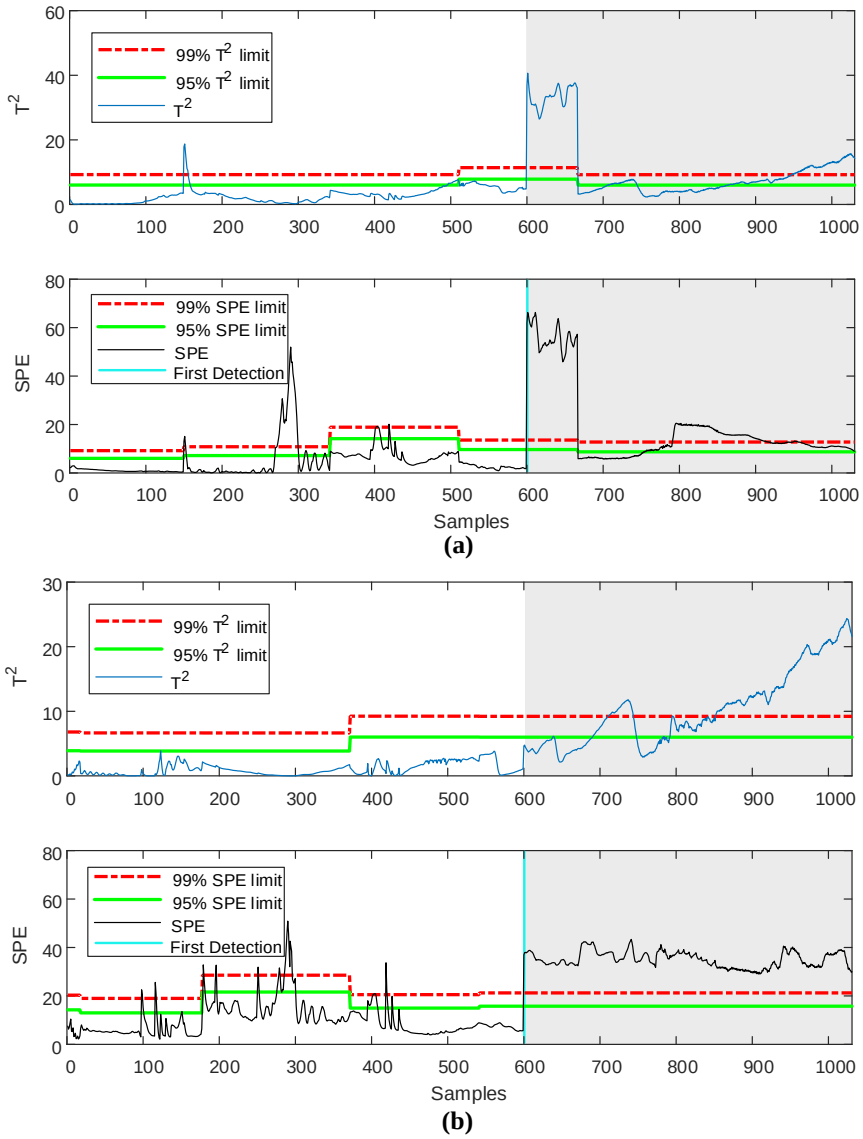

Figure 6.11: T^2 and SPE metrics obtained using (a) the fixed variance, and (b) the optimum multi-level multi-factor model for a batch exhibiting a vessel back pressure error. The turquoise vertical line indicates when the fault was first detected.

The optimum model shows better performance for detection of all faults terms of the type-II error rate comparison shown in Table 6.5. Even though the temperature fault was detected earlier than with the fixed variance model (the optimum model detected the fault at sample number 308 compared to sample number 359 for the fixed variance

model), the optimised model has type-II errors in the final phase (between samples 680 to 750). This is the only case where the performance of the optimum and fixed variance models was comparable. By using parameter tuning, the effective PCs and factors were discovered. From a MSPC point of view, use of a model that explains between $40-50\%$ of the total variance, is capable of detecting most of the faults. However, due to a combination of the cointegration residuals and the multi-level characteristics, the 1st level model performs better for fault detection when it explains a higher percentage of the variance. In comparison, a lower percentage variance is preferred for the 2nd level model to achieve good fault detection rates.

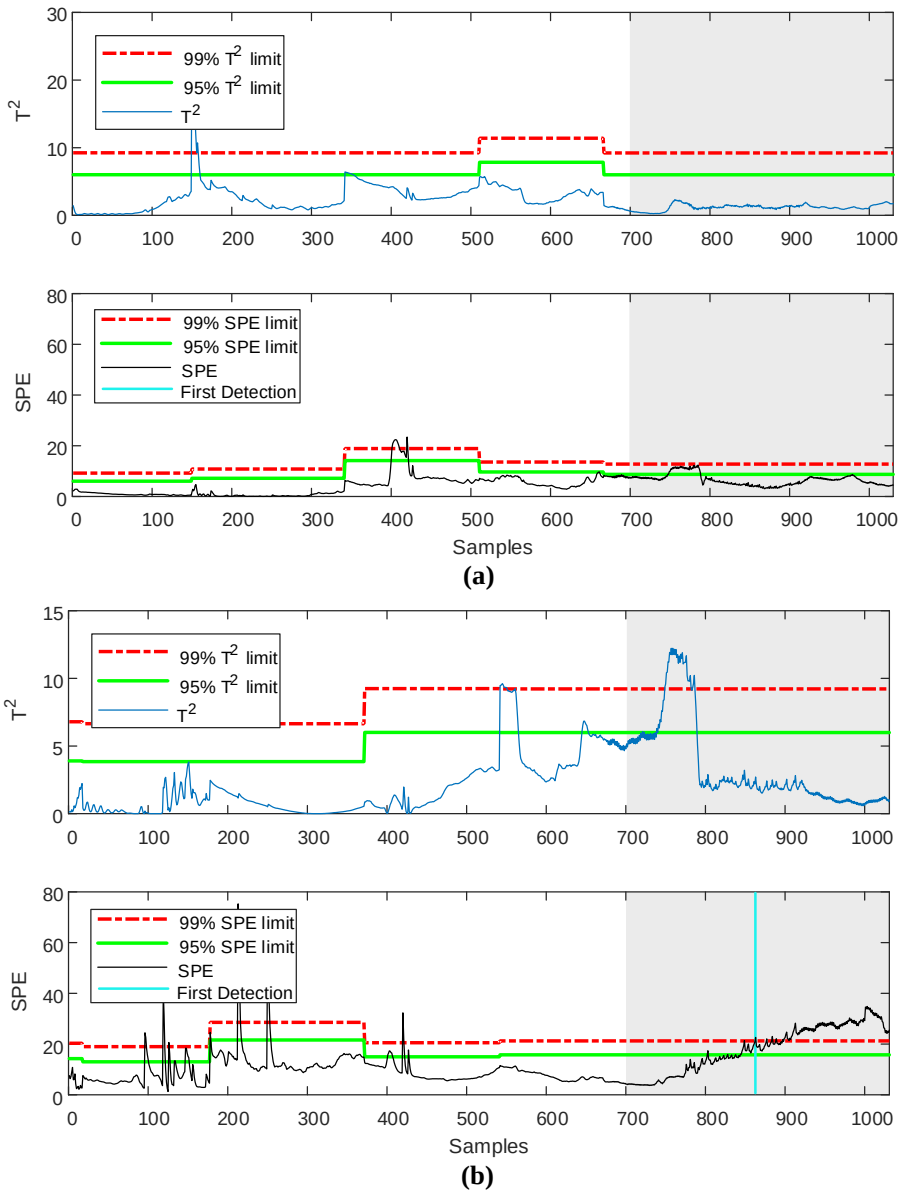

Figure 6.12: T^2 and SPE metrics obtained using (a) the fixed variance, and (b) the optimum multi-level multi-factor model for a batch exhibiting a base flow rate error. The turquoise vertical line indicates when the fault was first detected.

6.6 Application to the Tennessee Eastman Process

The CSTH example only required the search for the optimum number of PCs required for the two PCA models, while the industrial penicillin example (batch process) required the search for both the optimum number of PCs required for the two PCA models and the length of the phases of the batch. For processes that comprise more than 12 nonstationary variables, an additional search parameter, the combination of nonstationary variables for each cointegration model, is needed. This is because use of the Johansen test is limited to use with 12 or less nonstationary variables. As discussed by Harris and Sollis in their consideration of the order of integration of the variables, the cointegration rank depends on the combination of nonstationary variables involved in the cointegration model where the rank may increase or decrease according to the combination (Harris and Sollis, 2003). Searching different combinations of the nonstationary variables may change the cointegration rank and the performance of the cointegration residuals in the multi-level multi-factor model. Therefore, exploration of this property is investigated using a continuous process simulator called the Tennessee Eastman Process (TEP), which consists of 53 variables.

6.6.1 Introducing the Tennessee Eastman Process Simulator

The Tennessee Eastman process (TEP) simulator emulates a continuous chemical process. Downs and Vogel proposed it as a test problem for a list of potential applications such as plant control, optimization and monitoring (Downs and Vogel, 1993). It is a popular test problem within the chemometrics community and has been used widely for the development of MSPC techniques (Ku, Storer and Georgakis, 1995; Rato and Reis, 2013c; Sun, Zhang, Zhao and Gao, 2017). A revised version of the TEP was proposed by Bathelt et al. to cover the randomness problem in the previous version (Bathelt, Ricker and Jelali, 2015).

The process contains five major units as illustrated in Figure 6.13: a reactor, a stripper, a compressor, a vapour-liquid separator, and a condenser. It produces two liquid products (G and H) and a liquid by-product (F) from four gas reactants (A, C, D, and E) from the irreversible exothermic reactions (Downs and Vogel, 1993). Reactants, A, D and E, flow into the reactor then the reactor feeds the condenser. The vapour-liquid separator separates the substances then the stripper separates the remaining A, D and

E from the liquid and another reactant, C, is added to the product. This is followed by the exit of the final product given below (Capaci *et al.*, 2019):

$$\begin{cases} A(g) + C(g) + D(g) \rightarrow G(liq) \\ A(g) + C(g) + E(g) \rightarrow H(liq) \\ A(g) + E(g) \rightarrow F(liq) \\ 3D(g) \rightarrow 2F(liq) \end{cases} \tag{6.10}$$

The TEP simulator contains 41 measured and 12 manipulated variables, which are listed in Table 6.6. The sample time of the simulator is 1.8 seconds, however, the sample time for recording variables was chosen as 3 minutes.

Table 6.6: Measured and manipulated variables of the Tennessee Eastman Process.

No	Variable Description	No	Variable Description
1	A feed (Stream 1)	28	Component F (Stream 6)
2	D feed (Stream 2)	29	Component A (Stream 9)
3	E feed (Stream 3)	30	Component B (Stream 9)
4	A and C feed (Stream 4)	31	Component C (Stream 9)
5	Recycle flow (Stream 8)	32	Component D (Stream 9)
6	Reactor feed rate (Stream 6)	33	Component E (Stream 9)
7	Reactor pressure	34	Component F (Stream 9)
8	Reactor level	35	Component G (Stream 9)
9	Reactor temperature	36	Component H (Stream 9)
10	Purge rate (Stream 9)	37	Component D (Stream 11)
11	Product separator temperature	38	Component E (Stream 11)
12	Product separator level	39	Component F (Stream 11)
13	Product separator pressure	40	Component G (Stream 11)
14	Product separator underflow (Stream 10)	41	Component H (Stream 11)
15	Stripper level	42	D feed flow (Stream 2)
16	Stripper pressure	43	E feed flow (Stream 3)
17	Stripper separator underflow (Stream 11)	44	A feed flow (Stream 1)
18	Stripper temperature	45	A and C feed flow (Stream 4)
19	Stripper steam flow	46	Compressor recycle valve
20	Compressor work	47	Purge valve (Stream 9)
21	Reactor cooling water outlet temperature	48	Separator liquid flow (Stream 10)
22	Stripper cooling water outlet temperature	49	Stripper liquid product flow(Stream 11)
23	Component A (Stream 6)	50	Stripper steam Valve
24	Component B (Stream 6)	51	Reactor cooling water flow
25	Component C (Stream 6)	52	Condenser cooling water flow
26	Component D (Stream 6)	53	Agitator
27	Component E (Stream 6)		

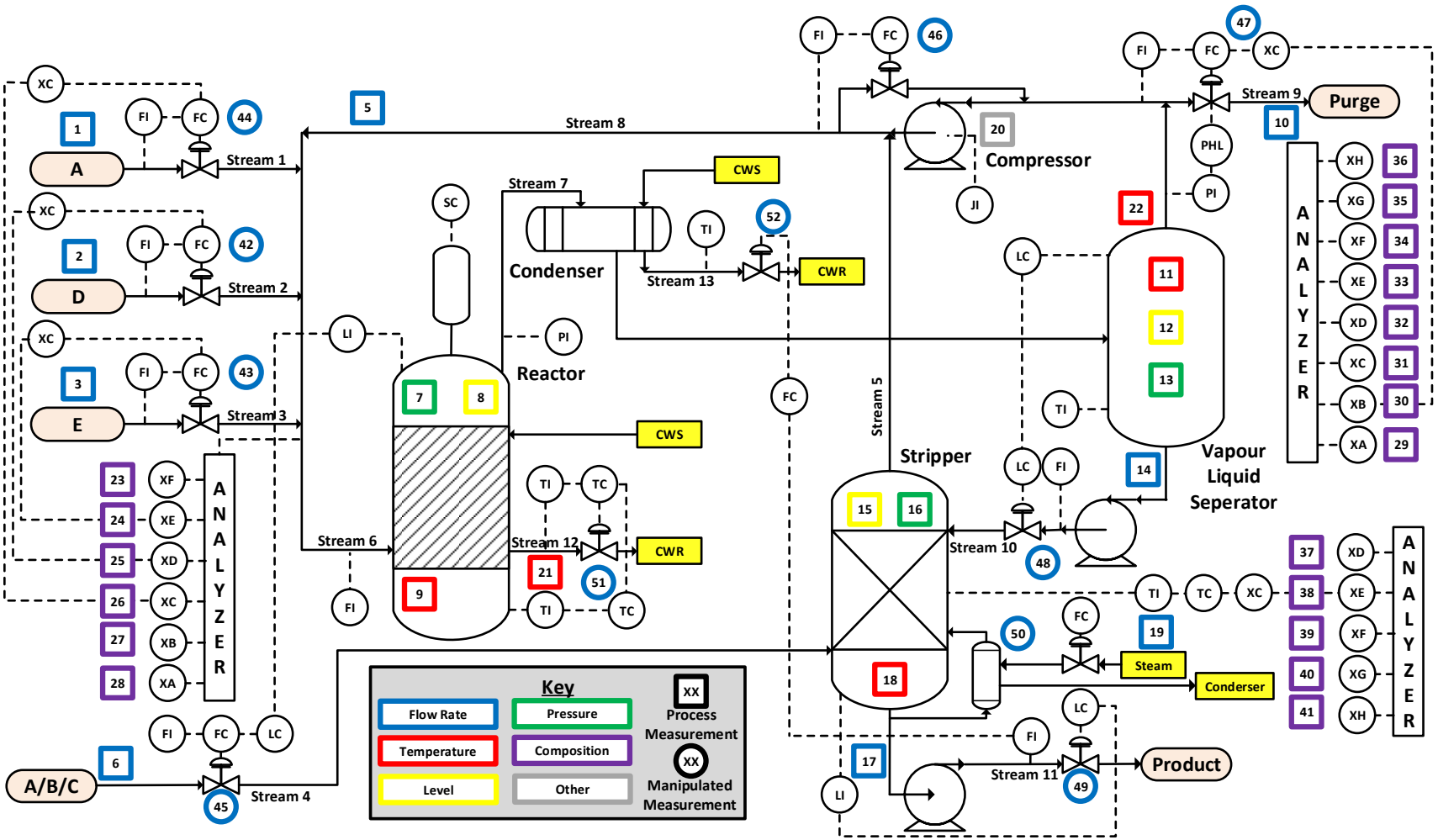

Figure 6.13: Illustration of the Tennessee Eastman Process simulator.

6.6.2 Multi-level Multi-factor Model on Tennessee Eastman Process

A multi-level multi-factor model was established on the training data set gathered from normal operation for 1200 samples. Data identification was performed using the ADF test based on an $AR(11)$ model with a significance level of 0.05. 2 manipulated variables, x_{50} and x_{53}, were not included as they were constant for this study. 18 variables were identified as stationary, with the remaining 33 variables identified as nonstationary. The stationary variables are: $x_4, x_5, x_6, x_9, x_{12}, x_{14}, x_{19}, x_{20}, x_{26}, x_{27}, x_{31}, x_{33}, x_{35}, x_{36}, x_{37}, x_{39}, x_{46}$ and x_{47}.

A PCA model for the 1st level was trained with 8 PCs, which explained 50.5% of the total variance where PCs 1 to 10 explained $7.47, 6.54, 6.50, 6.35, 6.07, 5.93, 5.86, 5.78, 5.62$, and 5.38% of the variance in the data. Therefore, the stationary factors $(\tilde{\boldsymbol{X}})$ were described using 8 score vectors ($\boldsymbol{T}$). The detailed PCA model selections for multi-level multi-factor model is presented in Appendix-E, which shows the number of PCs are chosen that retains closest to 45% of the variance in the data for the 1st and 2nd level PCA models.

There are 33 nonstationary variables to be modelled by cointegration analysis via the Johansen test, where one model can support only 12 variables. Therefore, at least 3 cointegration models are needed to produce stationary cointegration residuals from all of the nonstationary variables. The rank (R) for the cointegration matrix was [8 9 5] (from [12 12 9] nonstationary variables for each model). Therefore, the cointegration residuals matrix $(\boldsymbol{\xi})$ consists of 22 vectors. Hence, the rank of the perpendicular matrices for the common-trend model was [4 3 4], and the common-trend residuals matrix $(\boldsymbol{\varsigma})$ consists of 11 vectors. The 2nd level PCA model was trained with 9 PCs out of 41, which explained 47.9% of the total variance where PCs 1 to 10 explained 8.83, 8.37, 6.15, 5.34, 4.53, 4.12, 3.82, 3.39, 3.27 and 2.99% of the variance in the data. The T^2 and SPE charts for the fixed variance multi-level multi-factor model built using training data is illustrated in Figure 6.14; both metrics exhibited type-I error rates of 1%.

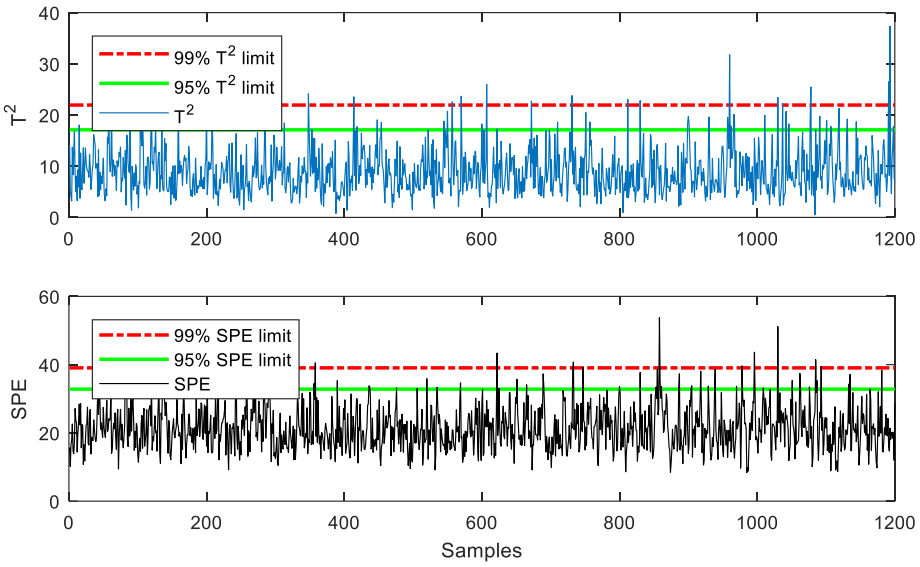

Figure 6.14: T^2 and SPE metrics for the fixed variance multi-level multi-factor model built using training data from the TEP.

6.6.3 Model Optimisation

As described in the previous applications of the parameter tuning system, the BB-BC optimization algorithm can be used to search for the optimum number of PCs for both PCA models. In addition to optimization of these two parameters, it is known that the combination of the nonstationary variables inside the cointegration model may affect the cointegration rank through the cointegration relationship between them. Therefore, the BB-BC optimization algorithm was also used to search for the optimum combination of nonstationary variables to use in each cointegration model. This search can be improved by increasing the number of cointegration models. However, this gives rise to an increase in the computational time of the algorithm. To simplify the search for the optimum combination of nonstationary variables to employ in the cointegration models, the number of cointegration models was set to 3 and the number of variables in each model was set to 12, 12 and 9 nonstationary variables for each cointegration model. Therefore, BB-BC searches 3 nonstationary variable set combinations for cointegration analysis ((D) in Figure 6.1). The design space for the optimum number of PCs ((A) and (B) in Figure 6.1) was searched using the level of variance explained in the data, and can take any value between [0 100]. When considering the optimum combination of nonstationary variables, any nonstationary variables can take part in only one cointegration model. Thus, the BB-BC algorithm searches for the optimum combination for each cointegration model. In contrast, the

fixed variance model uses the sequential order of the nonstationary variables for the cointegration models (first 12 nonstationary variables from Table 6.6 are modelled by the 1st cointegration model, the 2nd 12 nonstationary variables for the 2nd model and the remaining for the 3rd model).

The TEP simulator enables generation of several faults. The parameter tuning scheme can only be tested if some fault cases are used for the training of the parameter tuning procedure and some of the faults used for model performance testing are not used in the development of the parameter tuning scheme. The TEP simulator runs faults with only one binary ($0\ or\ 1$) control parameter. Even though some set point changes were suggested in the original publication (Downs and Vogel, 1993), benchmark faults are not described by their magnitudes. The faults given in this section are used without any change in the TEP simulator. The following scenarios were selected for training the parameter tuning: (i) A/C feed ratio error, B composition is constant in stream 4 (step type), (ii) B composition error, A/C ratio is constant in stream 4 (step type), (iii) D feed temperature error in stream 2 (Step type), (iv) C feed flow error in stream 4 (random variation), (v) E feed flow error in stream 3 (random variation), (vi) idv-16 unknown error. Note that idv represents the fault number of the defined error in the TEP. Therefore, the number of the selected cases was assigned as $K = 7$ which also includes data depicting normal operating conditions. Some of the selected faults, such as (i) and (ii), can be detected with nearly a 100% success rate using the fixed variance model. These faults were combined with other faults, such as (iii), which has a lower detection rate using the fixed variance model.

500 iterations for every 20 populations of candidate variables were used for this study. Tuning for three parameters ($p_1, p_2, \boldsymbol{Ord}$) required 500 iterations took 115.23 minutes; l was set to 3. The change in the cost function with the number of iterations for the optimum model parameters can be seen in Figure 6.15.

In comparison with the fixed variance multi-level multi-factor model that retains closest to 45% of the variance In the data for the 1st and 2nd level PCA models, the BB-BC algorithm suggested retention of 93.2% and 32.0% of the variance in the data for the 1st level and 2nd level PCA models, respectively, in the optimum multi-level multi-factor model. The results obtained for the training data using the suggested parameters are shown in Figure 6.16; both metrics exhibited type-I error rates of 0.7%. The 1st level PCA model explained 93.2% of the total variance, requiring 16 PCs, and

the 2nd level PCA model explained 32% of the total variance, requireing 8 PCs, in the multi-level multi-factor model.

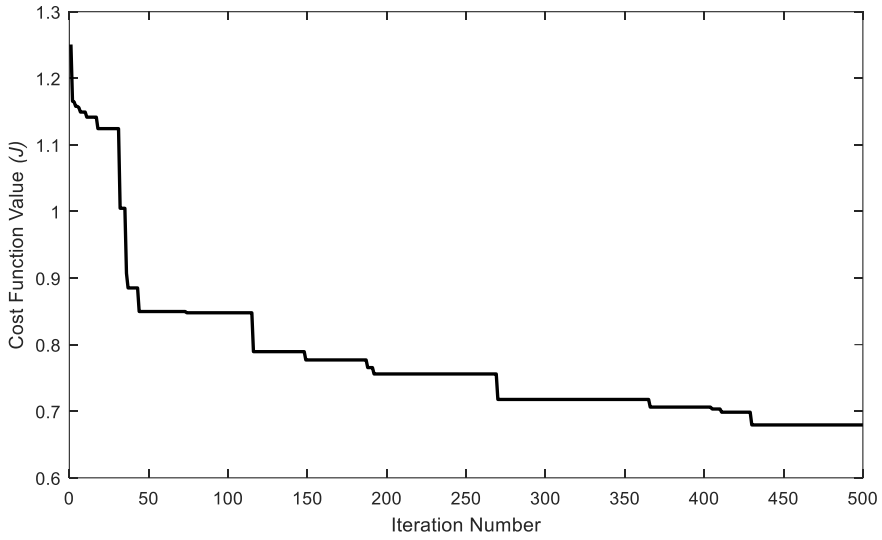

Figure 6.15: Cost function change for the best run that gives the optimum parameters for the multi-level multi-factor model of data from the TEP.

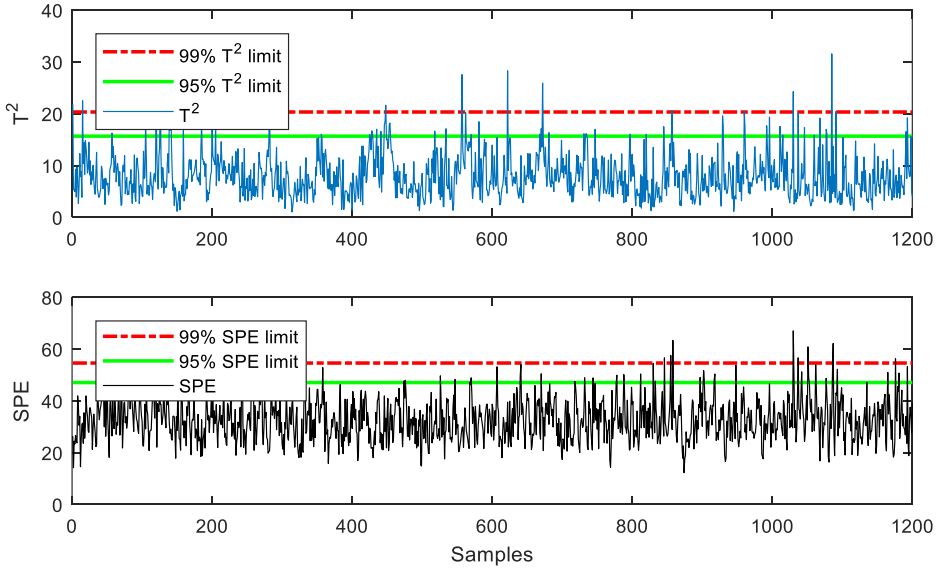

Figure 6.16: T^2 and SPE metrics for the optimum multi-level multi-factor model built using training data from the TEP.

The optimum combination of the nonstationary variables for the cointegration models obtained using the BB-BC optimisation algorithm is listed in Table 6.7. Furthermore, the cointegration ranks for the sub-models were [11 10 7]. Hence, the rank of the perpendicular matrices for the common-trend model were [1 2 2].

Table 6.7: Combinations of nonstationary variables used in the cointegration models within fixed variance and optimum multi-level multi-factor models

No	Fixed Variance Model	Optimum Model
1	$[x_1, x_2, x_3, x_7, x_8, x_{10}, x_{11}, x_{13}, x_{15}$ $x_{16}, x_{17}, x_{18}]$	$[x_1, x_2, x_{16}, x_{17}, x_{18}, x_{22}, x_{24}, x_{34},$ $x_{38}, x_{41}, x_{45}, x_{49}]$
2	$[x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{28}, x_{29}, x_{30}$ $x_{32}, x_{34}, x_{38}, x_{40}]$	$[x_7, x_{10}, x_{11}, x_{15}, x_{21}, x_{28}, x_{30}, x_{40},$ $x_{43}, x_{44}, x_{51}, x_{52}]$
3	$[x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{48}, x_{49}, x_{51}, x_{52}]$	$[x_3, x_8, x_{13}, x_{23}, x_{25}, x_{29}, x_{32}, x_{42}, x_{48}]$

The performance of the fixed variance and optimum models for detection of five different types of faults are compared in Table 6.8. The first 3 cases have been used in the parameter tuning scheme and the first two faults have already been detected with nearly a 100% success rate. 3 fault cases are detailed to show the performance differences given in Table 6.8. Additional two errors are selected from the unused fault cases in the parameter tuning.

Table 6.8: Type I and type II errors for the fixed variance and optimum multi-level multi-factor models for detection of different types of faults.

Case	Models							
	Fixed variance				Optimum			
	Type-I		Type-II		Type-I		Type-II	
	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2
A/C feed ratio and B composition constant	0.8	1.2	**0.3**	0.06	**0.7**	1.4	**0.3**	0.25
B composition and A/C ratio constants	**0.9**	1.1	1.8	2.9	**0.9**	1.3	**1.7**	1.1
D feed temperature	**2.8**	1	81.1	99	3.1	0.9	**54.7**	98
A,B and C feed composition	1.8	1.1	35.1	31.3	**1**	1.7	**18.3**	22.5
A and C feed pressure	1.2	1.1	10.3	44.4	**0.9**	1	**8.9**	15.5

The D feed temperature error was used in the training of the parameter tuning scheme. Even though it is defined on D feed (x_2), it is related to the secondary feature of x_2, which is not directly monitored. The reactor temperature (x_9) is affected indirectly, which is stationary according to Table 6.7. Therefore, selection of the number of PCs

(search space (A) and (B)) is more important than the nonstationary variable combination (search space (D)) for this particular error. The T^2 and SPE metrics for the fixed variance and optimum models are illustrated in Figure 6.17 and Figure 6.18, respectively. The optimum model provided a ~30% improvement in terms of SPE metric (type-II error rate reduced to 54.7% from 81.1%) as tabulated in Table 6.8. The increase in the fault detection rate is attributed to differences in the number of PCs rather than the cointegration rank, which is changed after the new combination of the nonstationary variables.

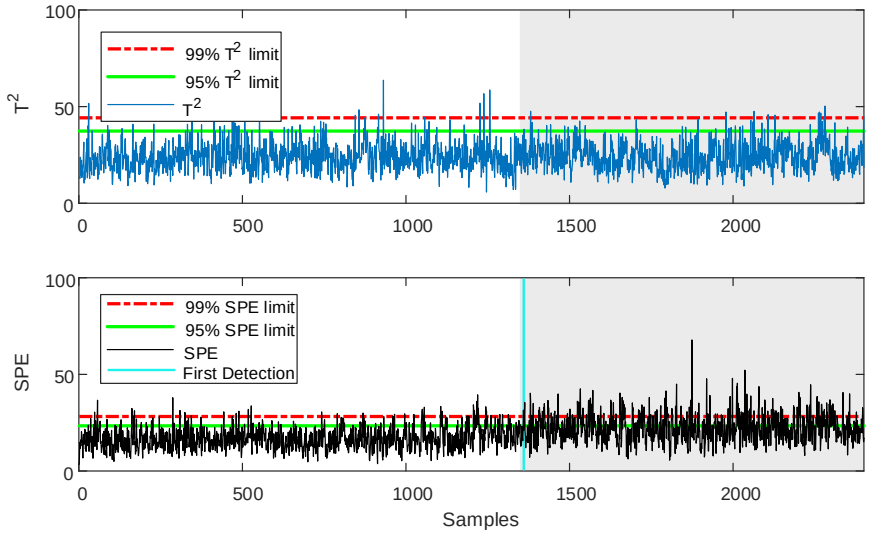

Figure 6.17: T^2 and SPE metrics obtained using the fixed variance multi-level multi-factor model with test data from the TEP exhibiting a fault in the D feed temperature. The fault was first detected at sample number 1410 using the SPE metrics (indicated by turquoise vertical line).

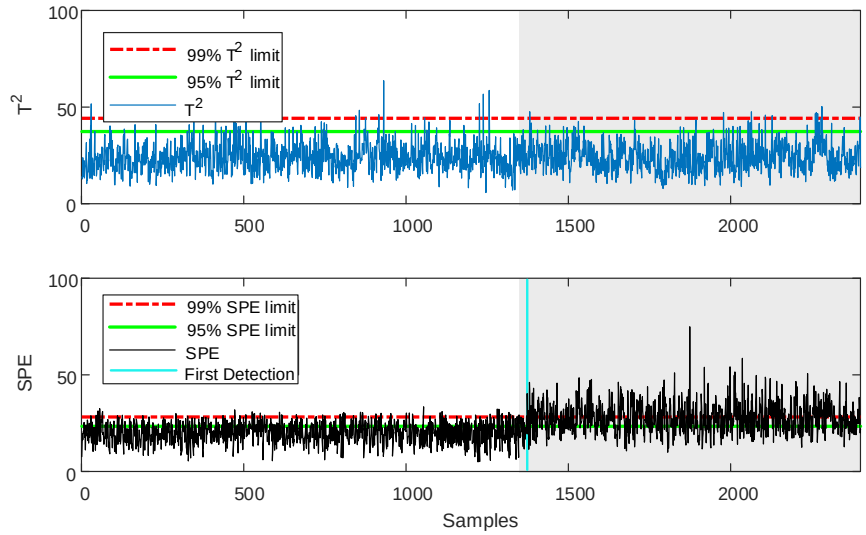

Figure 6.18: T^2 and SPE metrics obtained using the optimum multi-level multi-factor model with test data from the TEP exhibiting a fault on the D feed temperature. The fault was first detected at sample number 1425 using the SPE metrics (indicated by turquoise vertical line).

6.6.4 Model Performance

In this section, the model is evaluated using two examples that were not used in the parameter tuning. The first fault case is an error in the A, B and C feed composition on stream 4. They are the reactants of the production of G, H, and F. Therefore, several variables can affect the fault signatures such as $x_{23}, x_{24}. x_{25}$ in the reactor, $x_{29}, x_{30}, x_{31}, x_{34}, x_{35}, x_{36}$ in the purge gas analysis and more. Therefore, the combination of the nonstationary variables in each cointegration model is effective for this case. The T^2 and SPE metrics are given in Figure 6.19 and Figure 6.20 for the fixed variance and optimum models, respectively.

The optimum model was able to detect the fault using the SPE metric, giving a higher fault detection rate than the fixed variance model. The performance metrics for both models are listed in Table 6.8. Here, the optimum model improved the performance by ~40% in terms of SPE metric (type-II error rate reduced to 18.3% from 35.1%).

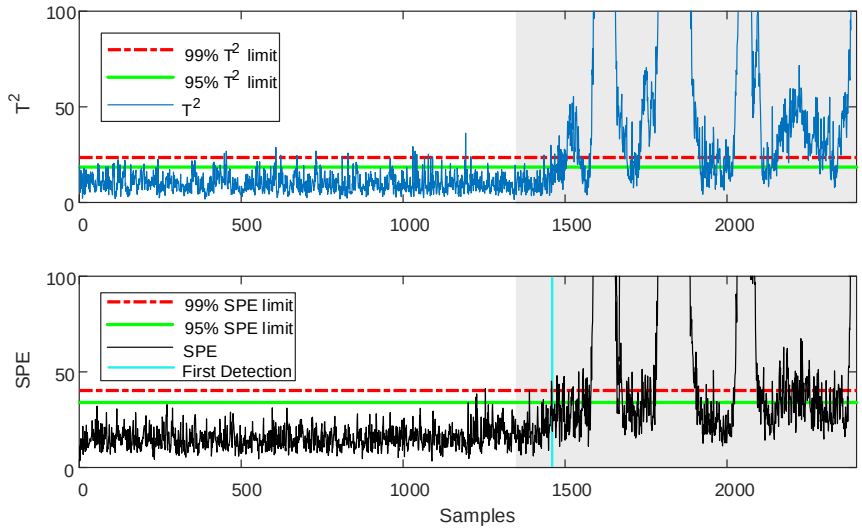

Figure 6.19: T^2 and SPE metrics obtained using the fixed variance multi-level multi-factor model with test data from the TEP exhibiting a fault on the A, B and C feed composition. The fault was first detected at sample number 1470 using the SPE metrics (indicated by turquoise vertical line).

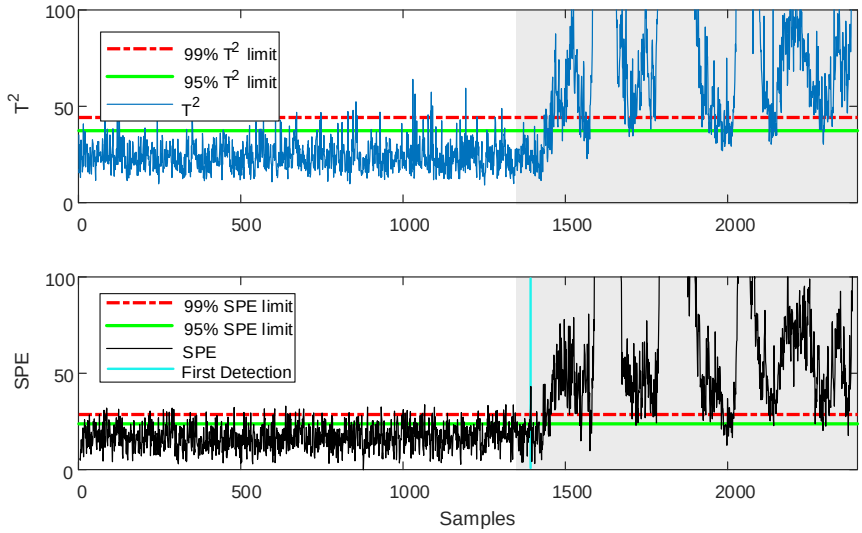

Figure 6.20: T^2 and SPE metrics obtained using the optimum multi-level multi-factor model with test data from the TEP exhibiting a fault on the A, B and C feed composition. The fault was first detected at sample number 1435 using the SPE metrics (indicated by turquoise vertical line).

The second fault case is an error in the A and C feed pressure in stream 4. These reactants are connected to the stripper and are manipulated via the variables of the reactant. x_4 and x_{45} are the two variables that can be affected by this error as they are stationary and nonstationary, respectively. Figure 6.21 and Figure 6.22 represent the T^2 and SPE metrics for the fixed variance and optimum models, respectively. The results are also tabulated in Table 6.8. The fault detection rates is increased in the fault zone by ~10% (type-II error rate for SPE metric reduced to 8.9% from 10.3%) and the faulty alarm rate in the normal operating region is decreased by ~25% (type-I error rate for SPE metric reduced to 0.93% from 1.21%). Changes in the numbers of PCs in the PCA models also changed the variance representation of the optimum model as can be seen from the T^2 metric performance in Figure 6.22. Hence, the parameter tuning system has improved the capability and effectiveness of the multi-level multi-factor model for fault detection in complex nonstationary industrial processes.

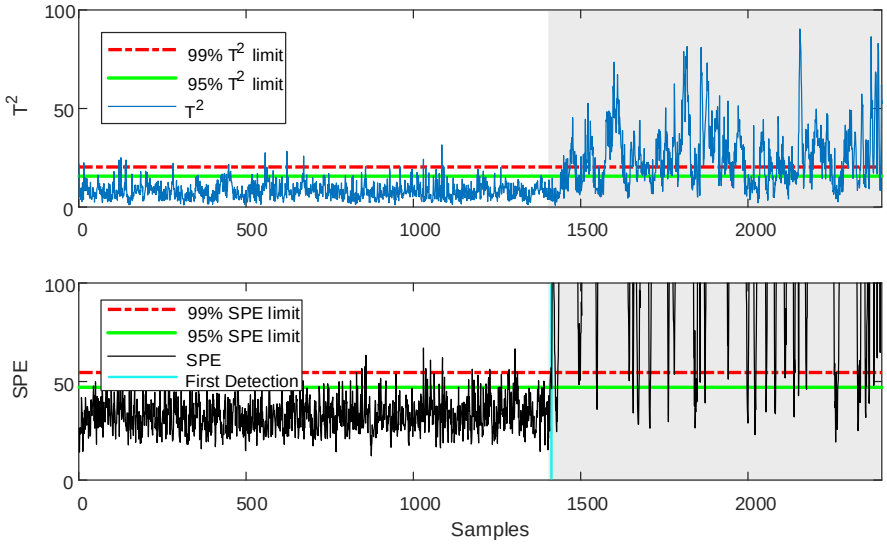

Figure 6.21: T^2 and SPE metrics obtained using the fixed variance multi-level multi-factor model with test data from the TEP exhibiting a fault on the A and C feed pressure. The fault was first detected at sample number 1405 using the SPE metrics (indicated by turquoise vertical line).

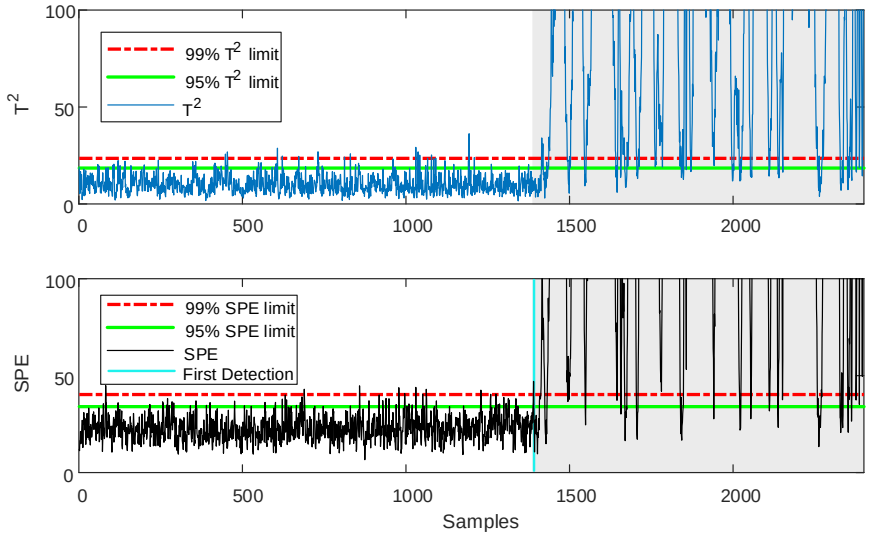

Figure 6.22: T^2 and SPE metrics obtained using the optimum multi-level multi-factor model with test data from the TEP exhibiting a fault on the A and C feed pressure. The fault was first detected at sample number 1403 using the SPE metrics (indicated by turquoise vertical line).

6.7 Conclusion

In this chapter, a parameter tuning system based on the BB-BC optimization algorithm was proposed for optimisation of multi-level multi-factor models. It solves several problems associated with the multi-level multi-factor model in terms of parameter optimisation for the different sub-models. These parameter optimisations can be categorised into four design spaces for the multi-level multi-factor model. These can be listed as: (A) the number of the PCs for the 1st level PCA model, (B) the number of

the PCs for the 2nd level PCA model, (C), phase length in batch process monitoring using multi-level multi-factor model, and (D) combination of the nonstationary variables for the cointegration models when the number of nonstationary variables exceeds 12 because the Johansen test is limited to use with 12 or fewer nonstationary variables.

The effects of the parameter optimisation on the performance of the multi-level multi-factor model was demonstrated on different cases starting from the simplest one. The CSTH simulation, used in Chapter 4, was optimised in terms of the number of PCs for both the 1st and 2nd level PCA models of the multi-level multi-factor model (design space (A) and (B)). Up to a 65% improvement in the type-II error rate for the SPE metric was observed in the fault detection rate base on the SPE statistic, The industrial penicillin simulator, used in Chapter 5, was optimised in terms of the number of PCs for both the 1st and 2nd level PCA models and the phase lengths of the corresponding batch processes. Up to a 35% improvement in the type-II error rate for the SPE metric was observed in the fault detection rate based on the SPE. The final case is tested on the TEP simulation, which has 53 variables. It was optimised in terms of the number of PCs for both the 1st and 2nd level PCA models and the combination of the nonstationary variables for the cointegration analysis models. Up to 40% improvement in the type-II error rate for the SPE metric was observed using the parameter tuning scheme for this example.

The computational time for the parameter tuning procedures is dependent upon the search space complexity. The initial search spaces (A) and (B) are the simplest in comparison to spaces (C) and (D) with respect to computational time. On the other hand, parameter tuning for the phase length through the search spaces A, B and C takes more time to complete the same number of iterations in comparison with the other search space combinations because of the involvement of the ADF test for each candidate.

A parameter tuning scheme that provides an insight into effective modelling of the multi-level multi factor model is presented. The scheme can also be implemented for online monitoring where the parameters can be updated after each time period of a batch.

7. CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

Multivariate statistical process control (MSPC) techniques are of key importance to different industries. They provide comprehensive on-line monitoring of manufacturing processes and the on-line detection of process malfunctions, and are capable of being applied to both continuous and batch processes. Today's competitive manufacturing practices require companies to be more efficient, more sustainable and consistent about quality assurance. MSPC techniques are contributing to the digital transformation of industries across a range of sectors in response to Industry 4.0. MSPC plays an essential role in intelligent analytics by providing quality maintenance, fault detection and diagnostic systems using data-driven real-time decision support systems.

The MSPC idea was born through the use of projection based latent variables modelling such as principal component analysis (PCA) and partial least squares (PLS) for process monitoring. In comparison to SPC, MSPC techniques aim to interpret high dimensional data by extracting data dynamics into reduced dimensionalities. However, the applicability of classical MSPC (the projection-based methods such as PCA and PLS) is restricted to stationary systems/variables. In contrast, complex industrial processes are highly nonstationary in nature. This being the case, dealing with nonstationary variables is starting to gain increased attention, and the existence of nonstationary variables accompanied by stationary variables has only recently started to be addressed. A ground zero approach to address nonstationarity in the data is to calculate the difference between consecutive time series data samples or use of the difference-based ARIMA model. It is a known fact from the econometrics field that variable differencing can lead to loss of dynamic information in the long-run, which contains valuable information.

One of the valuable findings from the research on nonstationary time series in the econometrics field is cointegration analysis. Cointegration analysis is arguably the most effective way of handling nonstationarities, which formulates the problem of the linear equilibrium along with nonstationary variables by using long-run equilibria. The effectiveness of cointegration analysis has been proved not only in its initial application area of econometrics but also in several disciplines of engineering such as process systems and construction. One of the modelling techniques used to extract

cointegration relationships is the Johansen test, which supports up to 12 nonstationary variables. The number of cointegration relationships can decrease down to zero because of the high-level of nonstationarity such as $I(2)$ variables. Even though cointegration analysis is a powerful tool to extract cointegration relationships, the rank of the cointegration matrix can be low despite there being a high number of nonstationary variables. This is because of the presence of higher level nonstationary time series. A common-trend model can solve the unrepresented cointegration relationship. However, this means another control chart must be used in addition to the cointegration residuals-based approach. This is a disadvantage compared to conventional MSPC approaches that only require a single control chart based on T^2 and SPE performance metrics.

Nonstationary process monitoring has been studied by dividing the data into two groups, stationary and nonstationary. Cointegration residuals-based approaches along with common trend residuals-based approaches can only model the nonstationary variables of processes. This gives rise to another issue for monitoring of complex industrial process, where nonstationary variables are present along with stationary variables. The current literature on the use of cointegration analysis models for process monitoring has just considered the nonstationary variables and the faults defined on these variables. This is limited by the definition stationary test as they give a binary result. It is another disadvantage compared to conventional MSPC approaches such as PCA that consider all of the process variables (assuming that they are stationary without testing them).

A multi-level multi-factor process monitoring model is proposed to solve the given problems in the application to continuous processes within Chapter 4. It consists of 2 level modelling along with data pre-treatment at the beginning to divide the process variables into two groups on the basis of their characteristics. The 1st level of modelling consists of three sub-models to cover the monitoring of all process variables, which solves one of the issues with cointegration analysis for process monitoring. Furthermore, the method uses a common-trend model in addition to PCA and a cointegration model to propose an alternative for the case of high-level nonstationary characteristics. On the other, it uses a 2nd PCA model in the 2nd level to model stationary factors gathered from the 1st level sub-models: PCA, cointegration and common-trend models. By using the T^2 and SPE metrics, it requires only one control

chart to be monitored by the operator, which is one of the advantages of MSPC techniques. The design of the models is completed based on the type-1 error rate performances in the training data set, and the same percent variance was used for the PCA models to provide similar features for comparison with other models. Other models that were compared include conventional PCA, dynamic PCA (DPCA), and cointegration and common trend residuals based approaches. DPCA is an extension of conventional PCA using a time lag shift to cope with time dependency and autocorrelations.

The continuous stirred tank heater (CSTH) simulator was used to compare the multi-level multi-factor model and the given models in Chapter 4. The CSTH is a second order plus dead time system, which shows time dependency and nonstationarity on the variables. A comparison of models was made on two different type of faults where the multi-level multi-factor model proved its superiority against its counterparts. In comparison with cointegration and common-trend models, the multi-level multi-factor model was able to detect faults for both stationary and nonstationary variables whereas the individual cointegration and common trend models were only able to detect a ramp function type fault, which is nonstationary. The multi-level multi-factor model also proved its superiority against conventional PCA and DPCA, which was proposed to cope with time dependent data. Therefore, the multi-level multi-factor model is the first reported method in the MSPC literature for the monitoring of continuous processes that combines both stationary and nonstationary variable modelling with only one control chart.

Multi-phase approaches are a solution to cope with time-varying and nonlinearities, especially in batch process modelling. The multi-level multi-factor model is enhanced with the power of nonlinear modelling using multi-phase approaches in Chapter 5. One of the counterparts of the multi-level multi-factor model for batch processes is the multi-level model. It uses only PCA and cointegration for the modelling of stationary and nonstationary variables, respectively. Unlike continuous processes, batch or fed-batch processes can have high-level nonstationary variables. The high-level nonstationarity among the nonstationary variables can cause a low rank cointegration matrix, which indicates the impracticality of modelling these types of variables by cointegration analysis. A common-trend model within the multi-level multi-factor method provides an opportunity to use unused cointegration vectors. This also

provides extra stationary factors for the 2nd level PCA model, which can be selected through the choice of the number of principal components (PCs).

An industrial penicillin simulator was used to compare the performance of the multi-level multi-factor, multi-level, and multi-PCA models for the monitoring of batch processes. The industrial penicillin simulation depicts a complex batch processes, which contains both stationary and nonstationary characteristics within its variables. The phase lengths were assigned using expert knowledge to build multi-phase models. The ability to detect two different types of error using the different models was evaluated. Two comparisons were made between the methods. The first one was between conventional PCA-based multi-PCA and the multi-level multi-factor model. Multi-PCA could detect the step function type fault better than the ramp function type fault. However, the multi-level multi-factor model possessed earlier detection capability for both fault type due to the ability to include all types of variables in the model. Secondly, the comparison between a multi-level and the multi-level multi-factor model indicates the effect of the use of common trend models. Again, the multi-level multi-factor model proved the need for common-trend models in the modelling of nonstationary variables where high-level nonstationary is present such as for complex batch processes. Therefore, Chapter 5 clearly demonstrated the effectiveness of incorporating a common-trend model within the multi-level multi-factor model.

Interpretability of models has been a secondary theme throughout the thesis up until the last chapter as the models have been designed separately by the author. There are several search spaces within the multi-level multi-factor model, which can be searched to select the appropriate parameters. Those can be grouped into 4: (A) the number of PCs in the 1st level PCA model, which models stationary variables, (B) the number of the PCs in the 2nd level PCA model, which models the stationary factors gathered from the 1st level sub-models, (C) the number and length of the phases, which determines the variables characteristics, and (D) selection of the nonstationary variables for each cointegration model when there are more than 12 nonstationary variables. A parameter tuning scheme based on the big bang-big crunch (BB-BC) global optimization algorithm is proposed to improve the multi-level multi-factor model by changing the given parameters above considering not only one parameter, but all parameters to have better fault detection capability.

The models designed for the different examples in Chapter 4 and 5 were compared with the optimum model found by the BB-BC algorithm. The results showed that a high percent variance 1st level PCA model and a low percent variance (or similar to that used in Chapters 4 and 5 (45%)) 2nd level PCA model provides improved fault detection. This also showed the similar principle that inclusion of the common-trend residuals improved the performance of the multi-level multi-factor model in comparison with the multi-level model, which used only PCA and cointegration analysis. Extra t-scores variables provided by the number of the PCs, gave an opportunity for the 2nd level PCA model to better represent the process. Two other cases were also investigated. One of them was the batch process that was used in Chapter 5. This time phase lengths (search space (C)) were searched using a parameter tuning scheme to improve both modelling of the stationary and nonstationary variables as the characteristics of the variables can change within the phases. Optimising the phase lengths improved the performance of the multi-level multi-factor model. A final case was the Tennessee Eastman process (TEP), which has 53 recorded variables. It is a perfect testbed to try multiple cointegration models for continuous process monitoring and tuning of the optimum combination of nonstationary variables for these models. The results showed that the combination of nonstationary variables in each model can affect the rank of the cointegration matrix and the performance of the cointegration residuals-based monitoring.

The parameter tuning scheme is an important requirement for the practical deployment of the method, which helps non-experts in the field to build and deploy the multi-level multi-factor model. Unlike a training environment, changes in the environment, especially seasonal, may occur which requires tuning of the calibration model. The parameter tuning scheme makes the method self-sufficient and self-adaptive for any process, which is a unique feature of the multi-level multi-factor model. Simple parameter tuning, including the search spaces (A) and (B), can be done in no more than 10 minutes. This does not require any change in the process before-hand apart from the validation of the model. Deployment of the other search spaces such as (C) and (D) may increase the optimisation time up to an hour. In this study, the tuning for the search spaces (A) and (B) took 83.2 seconds while that for (A), (B) and (C) took 190 minutes and (A), (B) and (D) took 115 minutes. Note that, the optimisation time and the number of training batches / data sets have a linear relationship. Therefore, the

number of scenerios (K) and number of the samples (M) must be considered as a computational load. The optimisation time can be increased due to the size of the data set. Therefore, selection of the training data set must be done carefully. This being the case, the implementation of the parameter tuning scheme within the multi-level multi-factor model can change the impracticality of the current model calibration process.

7.2 Future Work

The work presented in this thesis includes model presentation and model verification on different complex industrial process data sets obtained from the simulation of the processes described. However, before the multi-level multi-factor model can be used in practice, it has to be validated thoroughly via data sets obtained from actual industrial processes. It has to be ensured that the developed algorithms generate relevant results under all circumstances both on simulated and real data sets. In this work, three different process simulation have been used. From a practical point of view, some of the faults characteristics defined within the processes may or may not be realistic in the real process. For example, in practice, step functions do not exist in physical environment. A transition form of the variable between initial and final value of the step changes always exists within the nature, especially for high magnitude signals. This being case a model validation and new model studies on the real data collected from complex industrial process is required to improve and test the proposed work. Therefore, complex processes that show high level nonstationary such as fermentation processes would be a good start for model validation using real process data.

In addition, it is a prerequisite for a validation procedure that the multi-level multi-factor model can be tested in real-time on complex industrial process. This includes the application of the parameter tuning scheme along with the validated model. Only then can the real effect of the use of the multi-level multi-factor be estimated for process monitoring and model tuning. Practical validation or an in depth theoretical validation may lead to improvements. The more practical real-time data that are available, the better the behaviour of the models developed can be investigated.

Another theme to investigate in more detail is the development of new heuristic algorithms for the parameter tuning scheme. Note that the current version requires some inputs from the operator to the algorithm such as the maximum number of phases

for batch process modelling using the multi-level multi-factor model. Some new criteria can be defined that changes with the number of phases within the search space such as one based on the statistics score of the ADF tests. Furthermore, parameter self-tuning can be implemented easily by incorporating tuning loops fed by new data sets. This can be done automatically followed by an abnormality detection in the process or within a time period determined by the process expert.

Similar to the current form of the multi-phase structure, the phases can be searched by a clustering algorithm. A study was reported on batch process monitoring based on fuzzy segmentation using Gath-Geva clustering (Tanatavikorn and Yamashita, 2017) proposed to phase length for several PCA models. Similar searching procedure based on fuzz c-means clustering or Gath-Geva clustering can be implemented into the search space (C) in addition to other optimisation parameters in the parameter tuning scheme.

For the cointegration and common-trend residuals-based process monitoring models, the significance of the residual vectors is an unknown factor for process monitoring. As an example from PCA, eigenvalues can derive the importance of the corresponding score vector. However, cointegration residuals do not have such a relationship with any defined parameters. A parameter that can provide a selection criterion between the cointegration residuals would be helpful to design a detailed model. This can also change the performance of the multi-level multi-factor model and the effects of the selection can also be searched through the parameter tuning scheme based on the BB-BC optimisation algorithm.

A multi-level multi-factor model built only from process variables might not be able to monitor advanced features, which cannot be measured by physical measurements, especially in biopharmaceutical manufacturing. In such cases, use of spectroscopic data of the process may help advance process monitoring. Nevertheless, it requires fusion techniques to combine and evaluate the data sets from physical measurements and spectroscopic device such as Raman. A recent study published on the industrial penicillin simulator (used in Chapters 5 and 6) reports the addition of Raman spectroscopic data to the simulator (Goldrick *et al.*, 2019). This gives an opportunity to develop a fusion technique between the multi-level multi-factor model of the process data and Raman spectroscopic data, which would be valuable to research.

Another direction in cointegration research for process monitoring might be the use of the Chigira procedure. The Johansen test is the conventional cointegration testing method, which can support up to 12 variables. More than 12 variables requires additional modelling as presented in the application of the TEP using the multi-level multi-factor model. A test procedure for estimating the cointegration rank on the basis of PCA has been proposed to support cointegration modelling of more than 12 variables (Chigira, 2008). Therefore, it seems worthy to test the Chigira procedure for use in process monitoring cases that include more than 12 nonstationary variables such as the TEP.

In conclusion, proposing the multi-level multi-factor model was the first stage of the model development. Some further steps must follow (as described above) to see the impacts of modelling both nonstationary and stationary variables in industry.

APPENDIX-A

Nonlinear Iterative Partial Least Squares (NIPALS)

The NIPALS algorithm decomposes $\boldsymbol{X}$ sequentially through each principal components. The algorithm for the sequential order can be followed from Table A.1.

Table A.1: NIPALS for PCA.

1: Scale and subtract the averages from $\boldsymbol{X}$. Set $r = 1$.

2: Select a $\boldsymbol{t}_r$ for the score vector with the largest variance.

3: Calculate a loading vector as $\boldsymbol{p}_r = {\boldsymbol{X}\boldsymbol{t}_r}/{\boldsymbol{t}_r^T\boldsymbol{t}_r}$.

4: Normalize the loading vector to the unit length $\boldsymbol{p}_r = {\boldsymbol{p}_r}/{\|\boldsymbol{p}_r\|}$

5: Calculate a new score vector $\boldsymbol{t}_r = {\boldsymbol{X}^T\boldsymbol{p}_r}/{\boldsymbol{p}_r^T\boldsymbol{p}_r}$

6: Check the convergence., for instance using the sum of squared differences between all elements in two consecutive score vectors. If converge, continue to 7, otherwise return to 3. If convergence is not met by a specified number of iterations, break anyway.

7: Form the residuals $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{p}_r\boldsymbol{t}_r^T$. Use $\boldsymbol{E}$ as $\boldsymbol{X}$, set $r = r + 1$ until the maximum number of principal components met, then return to 3.

Singular Value Decomposition

The SVD is a generalization of the eigen decomposition. It decomposes a rectangular matrix into three simple matrices: one diagonal and two orthonormal.

$$\boldsymbol{X}^T = \boldsymbol{P}^T\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^T \tag{A.1}$$

where

- $\boldsymbol{U}$ is the normalized eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$(i.e $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$). The columns of $\boldsymbol{U}$ are called the left singular vectors of $\boldsymbol{X}$.
- $\boldsymbol{V}$ is the normalized eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$(i.e $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$). The columns of $\boldsymbol{V}$ are called the right singular vectors of $\boldsymbol{X}$. $\boldsymbol{V}$ also represents the loading matrix ($\boldsymbol{P}$) for PCA.
- $\boldsymbol{\Delta}$ is the diagonal matrix of the singular values, $\boldsymbol{\Delta} = \boldsymbol{\Lambda}^{1/2}$where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues of the matrix $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$.

APPENDIX-B

Similar to the algorithm described for PCA, the NIPALS algorithm decomposes $\boldsymbol{X}$ and $\boldsymbol{Y}$ as sequentially through each principal components. The algorithm for the sequential order can be followed from Table B.1.

Table B.1: NIPALS for PLS.

1: Scale and subtract the averages from $\boldsymbol{X}$and $\boldsymbol{Y}$. Set $r = 1$.

2: Select a u_r for the output score vector with the largest variance for column $\boldsymbol{Y}$.

3: Calculate an input loading vector as $\boldsymbol{w}_r = {\boldsymbol{X}\boldsymbol{u}_r}/{\boldsymbol{u}_r^T\boldsymbol{u}_r}$.

4: Normalize $\boldsymbol{w}_r$ to the unit length $\boldsymbol{w}_r = {\boldsymbol{w}_r}/{\|\boldsymbol{w}_r\|}$

5: Calculate a new input score vector $\boldsymbol{t}_r = {\boldsymbol{X}^T\boldsymbol{w}_r}/{\boldsymbol{w}_r^T\boldsymbol{w}_r}$

6: Calculate an output loading vector as $\boldsymbol{q}_r = {\boldsymbol{Y}t_r}/{\boldsymbol{t}_r^T\boldsymbol{t}_r}$.

7: Normalize $\boldsymbol{q}_r$ to the unit length $\boldsymbol{q}_r = {\boldsymbol{q}_r}/{\|\boldsymbol{q}_r\|}$

8: Calculate a new output score vector $u_r = {\boldsymbol{Y}^T\boldsymbol{q}_r}/{\boldsymbol{q}_r^T\boldsymbol{q}_r}$

9: Check the convergence., for instance using the sum of squared differences between all elements in two consecutive score vectors of $\boldsymbol{u}$. If converge, continue to 10, otherwise return to 3. If convergence is not met by a specified number of iterations, break anyway.

10: Calculate the loading vector of $\boldsymbol{X}$ as $\boldsymbol{p}_r = {\boldsymbol{X}t_r}/{\boldsymbol{t}_r^T\boldsymbol{t}_r}$.

11: Calculate the score coefficient as $\boldsymbol{d}_r = {\boldsymbol{u}_r^T\boldsymbol{t}_r}/{\boldsymbol{t}_r^T\boldsymbol{t}_r}$.

12: Form the residuals $\boldsymbol{X}_{New} = \boldsymbol{X} - \boldsymbol{p}_r\boldsymbol{t}_r^T$.

13: Form the residuals $\boldsymbol{Y}_{New} = \boldsymbol{Y} - \boldsymbol{d}_r\boldsymbol{q}_r\boldsymbol{t}_r^T$.

14: Replace $\boldsymbol{X}_{New}$ as $\boldsymbol{X}$ and the same for $\boldsymbol{Y}$, set $r = r + 1$ until the maximum principal components met, then return to 3.

APPENDIX-C

The selection criterion for the number of principal components (PCs) for the models, which includes the PCA model is detailed in this appendix. There are several ways to determine the number of PCs, such as defining the maximum limit for the cumulative explained variance, the training data error rate comparison, looking for a 'knee' in the scree plot, limits for the minimum eigenvalue, and cross-validation. Here, a conventional PCA model trained with 2 PCs (selected on the basis of the 'knee' in the scree plot and the type-I error) explained 54.48% of the data. Figure C.1 represent a scree plot of the eigenvalues extracted from CSTH data by PCA. Table C.1 shows the type-I error rate for the corresponding cumulative variance percentage, where it can be seen that the selection of 45% gives marginally the best results. Therefore, to select 45% of the variance in the data, 2 PCs are required which actually explains 54.48% of the variance in the data.

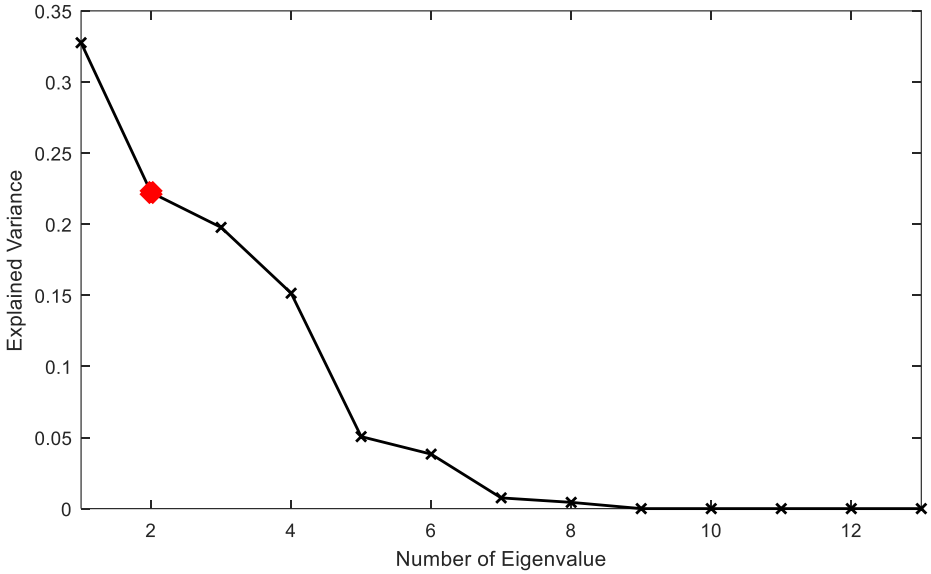

Figure C.1: Scree plot for the eigenvalues from the PCA model of CSTH data collected under normal operating conditions. The red dot represents the number of PCs selected.

Table C.1: Offline training performance of PCA models with corresponding cumulative explained variance for the CSTH process.

Metric Name (level)	Variance / Type I error (%)				
	45%	55%	65%	75%	85%
SPE(1%)	**0.6**	**0.7**	0.7	0.95	0.95
T^2(1%)	**0.7**	**0.65**	**0.65**	0.85	0.85
Number of PCs	2	3	3	4	4

Ku argues in favour of using parallel analysis for choosing the number of the PCs, but it is also recommended that the number of PCs for SPE metric can selected in regards to the score vectors, which are independent or nearly independent in time (Ku, Storer and Georgakis, 1995). Figure C.2 represents the scree plot of the eigenvalues of the DPCA model where the knee is at the 2nd and 7th PCs. On the other hand

Table C.**2** shows the training performance with respect to the type-I error. Consequently, the DPCA model is trained according to the error rate performance with selection of 35% of the cumulative explained variance. The model with 3 PCs that explained 42.95% of the variance in the data was selected to have a model that is close to 45%, which is the percentage that have been used in most of the models in this study.

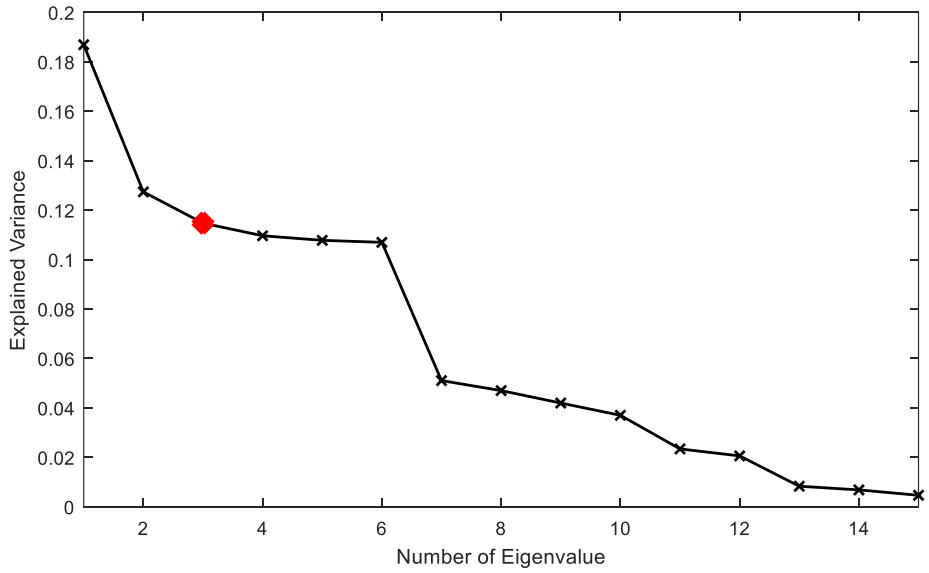

Figure C.2: Scree plot for the eigenvalues from the DPCA model for CSTH data collected under normal operating conditions. The red dot represents the number of PCs selected.

Table C.2: Offline training performance of DPCA models with corresponding cumulative explained variance for the CSTH process.

Metric Name (level)	Variance / Type I error (%)				
	35%	45%	55%	65%	75%
SPE(1%)	**0.6**	**0.7**	0.7	0.9	0.95
T^2(1%)	**0.7**	**0.65**	**0.65**	0.85	0.85
Number of PCs	3	4	6	7	9

The training of the multi-level multi-factor model is two part. Firstly, the number of PCs for the 1st level PCA model needs to be determined. There are 5 stationary variables in the 1st level PCA model and Figure C.3 represents the scree plot for the eigenvalues, where it can be seen that only one eigenvalue explained more than 70% of the variance in the data. From the classical MSPC point of view, selection of the first PC is sufficient to represent the data. Table C.3 also shows that the type-I error is largely unaffected by the number of PCs selected. Therefore, the 1st level PCA model was trained with 1 PC, which explained 72.24% of the variance in the data. Even though most of the models trained with 45% of the variance in regards to the number PCs, 1st level PCA model for CSTH data is used 1 PC that explained more than half of the variance.

Secondly, the number of PCs for the 2nd level PCA model needs to be determined. In this study, this was determined independently from the 1st level PCA model. However, it is known that selection of the first model affects the performance of the 2nd level PCA model. The 1st level PCA model gave rise to stationary factors, which consisted of 10 variables. Figure C.4 represents the scree plot for the eigenvalues when a 'knee' is not evident. Table C.4 tabulates the training performance in terms of the type-I error for the CSTH data with the corresponding cumulative explained variance. As a result 2 PCs were selected for the 2nd level PCA model, which explained 47.4% of the variance in the data.

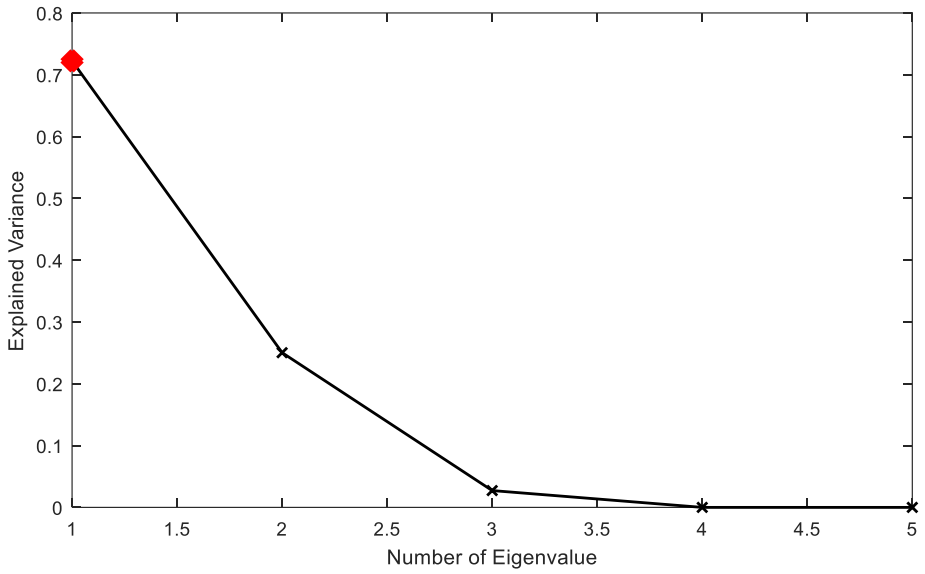

Figure C.3: Scree plot for the eigenvalues from the 1st level PCA model of the multi-level multi-factor model for CSTH data representing normal operating conditions. The red dot represents the number of PCs selected.

Table C.3: Offline training performance of the 1st level PCA models for the multi-level multi-factor model with corresponding cumulative explained variance for the CSTH process.

Metric Name (level)	**Variance / Type I error (%)**		
	70%	90%	98%
SPE(1%)	**0.8**	0.8	0.9
T^2(1%)	**0.65**	0.65	0.65
Number of PCs	1	2	3

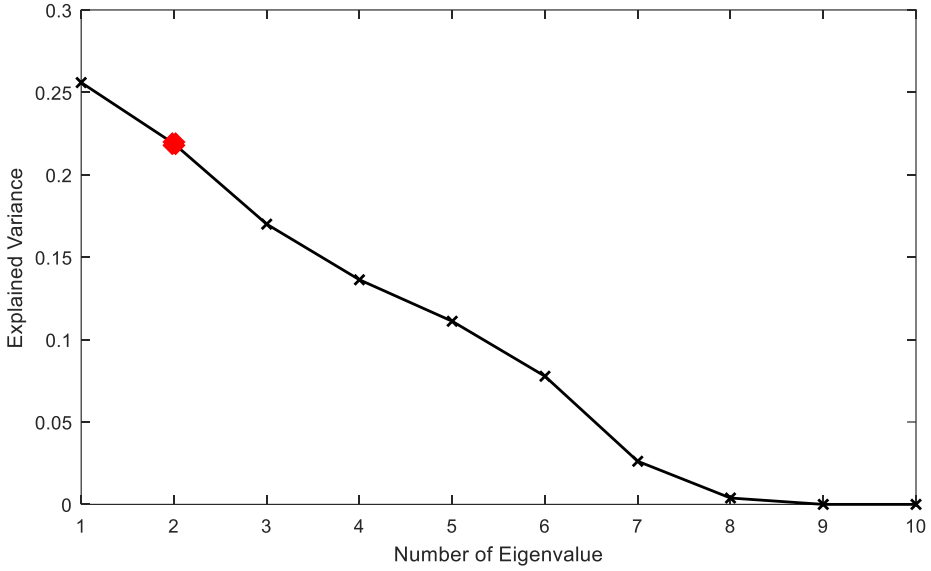

Figure C.4: Scree plot for the eigenvalues from the 2nd level PCA model of the multi-level multi-factor model for CSTH data representing normal operating conditions. The red dot represents the number of PCs selected.

Table C.4: Offline training performance of the 2nd level PCA models for the multi-level multi-factor model with corresponding cumulative explained variance for the CSTH process.

Metric Name (level)	**Variance / Type I error (%)**				
	45%	55%	65%	75%	85%
SPE(1%)	**0.6**	0.8	0.9	0.9	1.0
T^2(1%)	**0.8**	0.95	0.90	0.90	1.05
Number of PCs	2	3	4	4	5

APPENDIX-D

Chapter 5 compares 3 different multi-phase based models when applied to data from the industrial penicillin simulator. These models can be listed as: multi-PCA, multi-level, and multi-level multi-factor. In comparison to continuous processes, each model is designed and built for a specific phase. However, for the sake of simplicity a constant percent variance value was selected for each phase.

The scree plot for the multi-PCA model for each phases is illustrated in Figure D.1. The figures given below is an illustration of the information given in Table 5.4. However, due to the multi-phase complexity, selection of the number of PCs according to the eigenvalue distribution for each phase is a troublesome procedure. Therefore, the training performances in terms of type-1 errors for different percent variances were evaluated for data from the industrial penicillin simulation. Table D.1 shows the type-1 error for both T^2 and the SPE. Therefore, a fixed variance level of 45% was selected for each phase based on this training performance.

The multi-level and multi-level multi-factor models use the same stationary variable model, which is the 1st level PCA model. Similar to one constant variance selection for all phases in multi-PCA, Figure D.2 illustrates the scree plot of the eigenvalues of the stationary variables modelled by the 1st level PCA model. However, the number of PCs is selected with a fixed variance of 45% according to the training performance of the 1st level PCA model tabulated in Table D.2. Furthermore, the scree plot of the 2nd level PCA model is illustrated in Figure D.3. The number of the PCs was selected according to the performance tabulated in Table D.3. Note that, due to the small number of PCs, more than 75% explained the variance contained in all of the eigenvalues. Therefore, the SPE limits cannot be calculated for this example.

Similarly, a selection of one constant variance selection for all phases is also applied to the multi-level multi-factor model training. The scree plot for 2nd level PCA model in the multi-level multi-factor method is illustrated in Figure D.4 and the number of the PCs was selected as 45% according to the performance tabulated in Table D.4

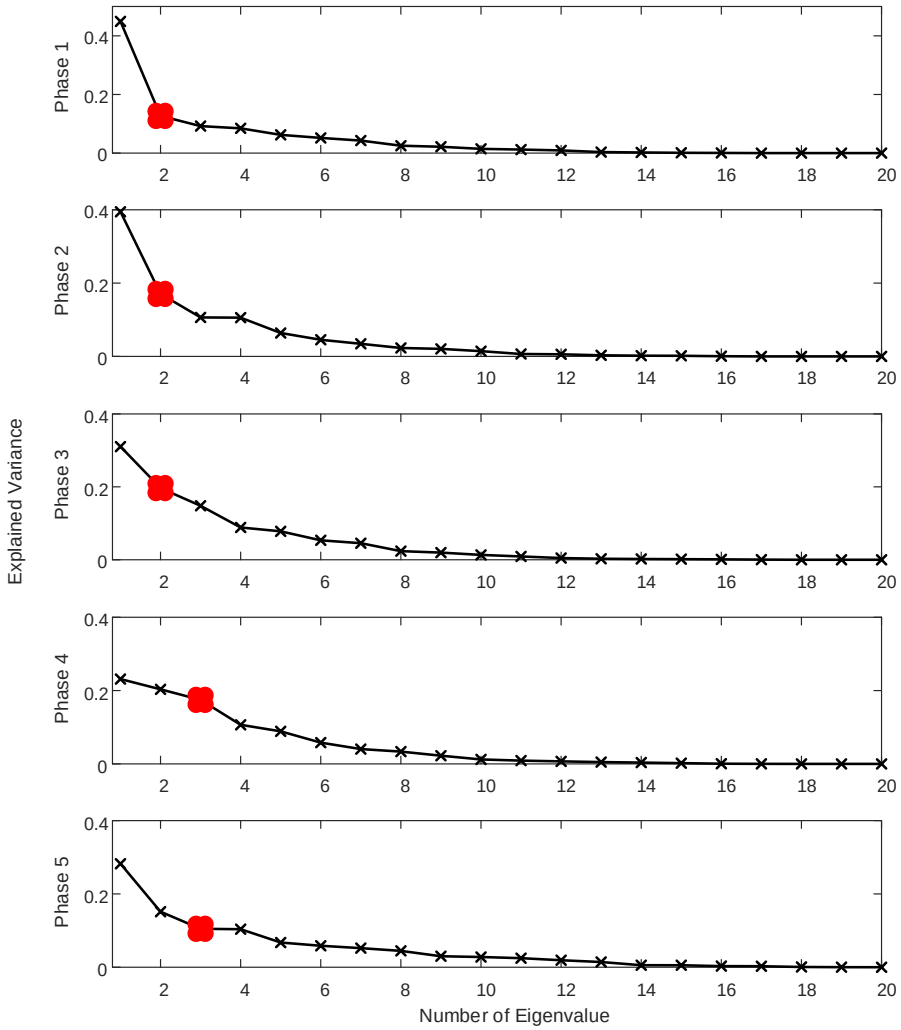

Figure D.1: Scree plot for the eigenvalues from the multi-PCA model for industrial penicillin simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table D.1: Offline training performance of the multi-PCA models with corresponding cumulative explained variance for the industrial penicillin simulator.

Metric Name (level)	**Variance / Type I error (%)**				
	45%	55%	65%	75%	85%
SPE(1%)	**2. 3**	2.91	3.59	3.59	3.39
T^2(1%)	**0**	0	0	0	0
Number of PCs	[**2 2 2 3 3**]	[2 2 3 3 4]	[3 3 3 4 5]	[4 4 5 5 6]	[6 6 6 6 8]

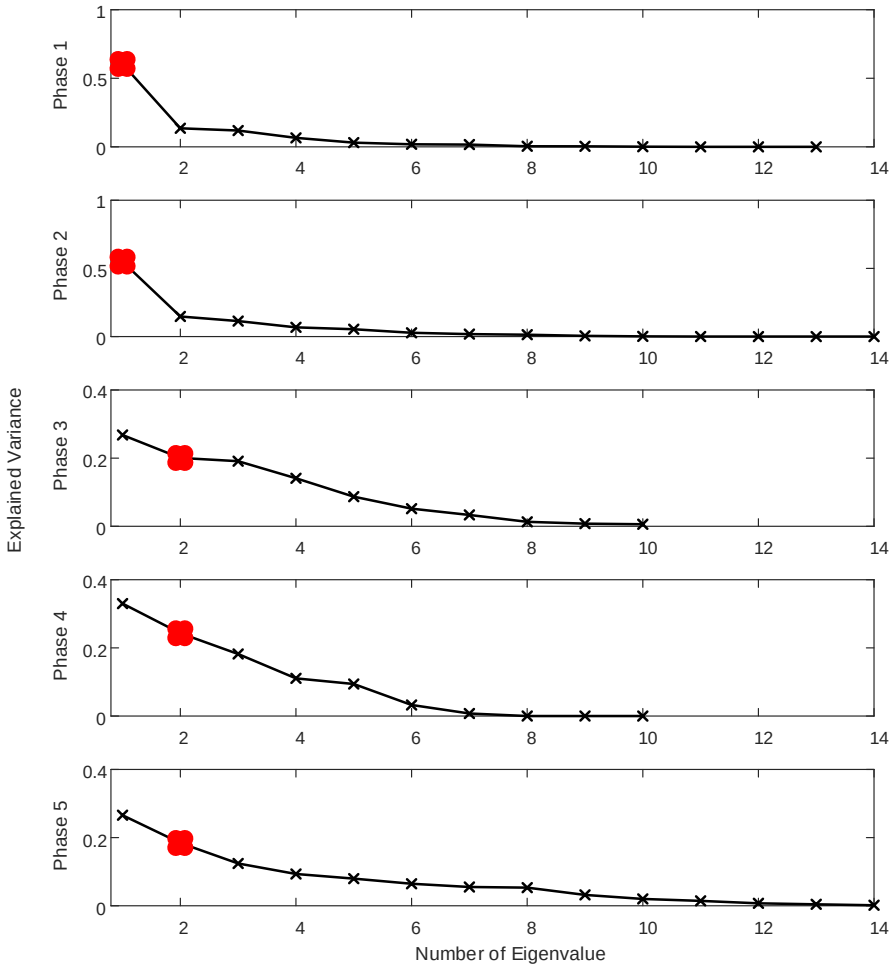

Figure D.2: Scree plot for the eigenvalues from the 1st level PCA models for industrial penicillin simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table D.2: Offline training performance of the 1st level PCA models with corresponding cumulative explained variance for the industrial penicillin simulator.

Metric Name (level)	**Variance / Type I error (%)**				
	45%	55%	65%	75%	85%
SPE(1%)	**2.42**	4.75	3.88	3.68	4.38
T^2(1%)	**0**	0	0.29	3.66	4.14
Number of PCs	[**1 1 2 2 2**]	[1 2 3 2 3]	[2 2 3 3 4]	[3 3 4 3 6]	[3 4 5 4 7]

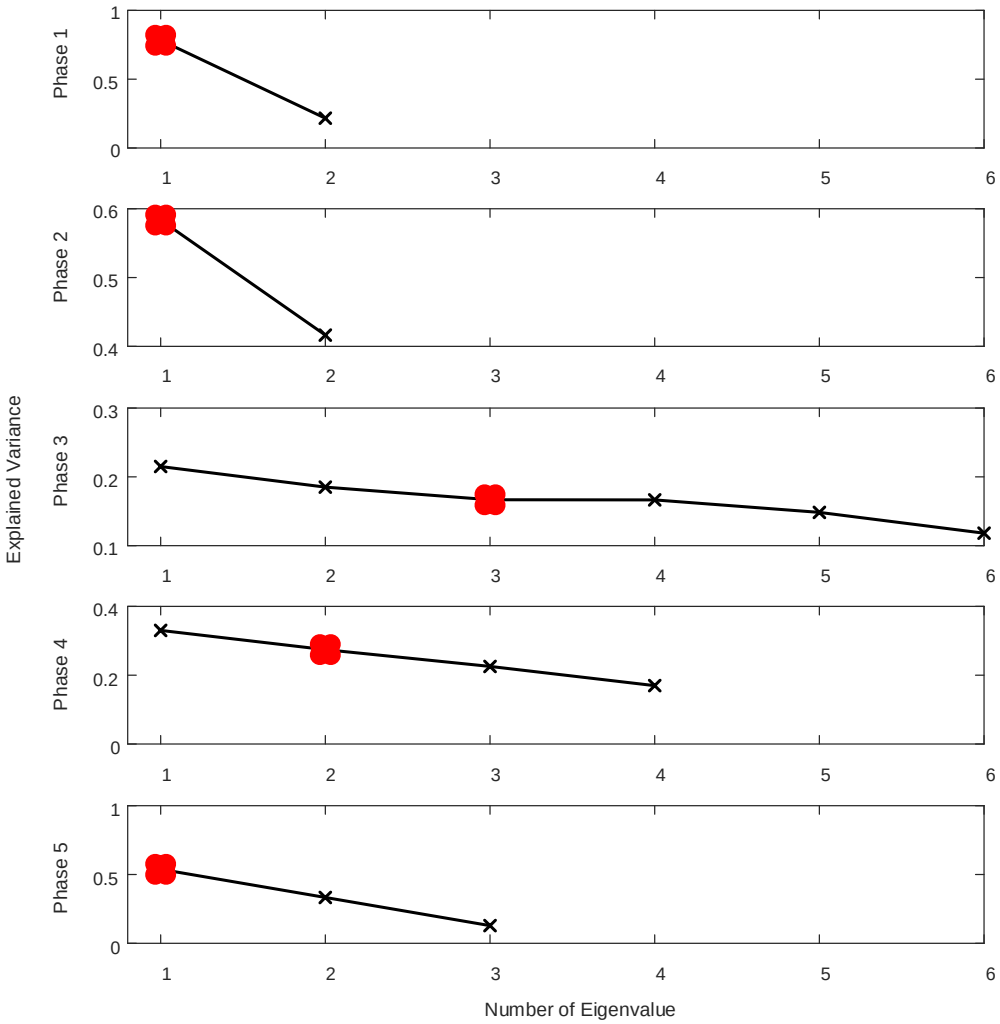

Figure D.3: Scree plot for the eigenvalues from the 2nd level PCA model for a multi-level model for industrial penicillin simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table D.3: Offline training performance of the 2nd level PCA model for multi-level models with corresponding cumulative explained variance for the industrial penicillin simulator.

Metric Name (level)	**Variance / Type I error (%)**				
	45%	55%	65%	75%	85%
SPE(1%)	**2.10**	2.22	~	~	~
T^2(1%)	**0.58**	0.77	~	~	~
Number of PCs	[**1 1 3 2 1**]	[1 1 3 2 2]	[1 2 4 3 2]	[1 2 5 3 2]	[2 2 5 4 2]

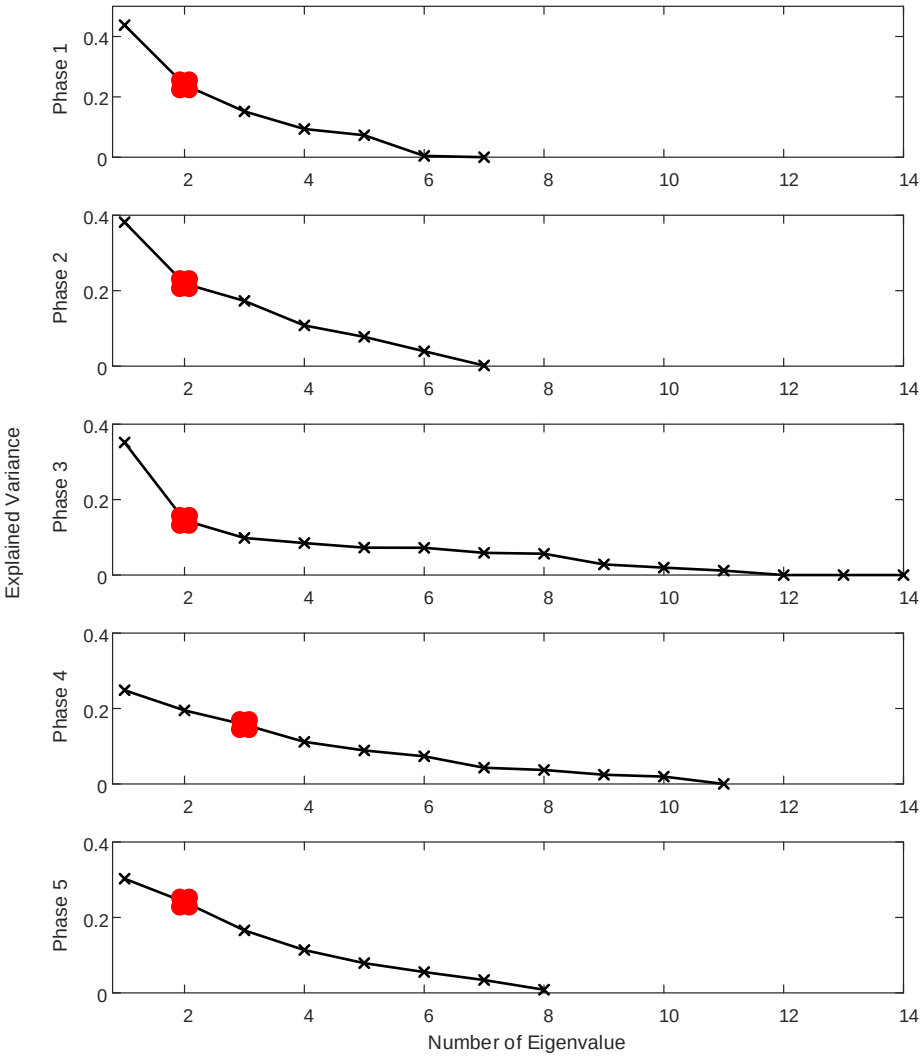

Figure D.4: Scree plot for the eigenvalues from the 2nd level PCA model for multi-level multi-factor model for industrial penicillin simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table D.4: Offline training performance of the 2nd level PCA model for multi-level multi-factor models with corresponding cumulative explained variance for the industrial penicillin simulator.

Metric Name (level)	**Variance / Type I error (%)**				
	45%	55%	65%	75%	85%
SPE(1%)	**1.80**	2.33	2.52	1.83	1.93
T^2(1%)	**0.38**	0.38	0.38	1.94	1.58
Number of PCs	[**2 2 2 3 2**]	[2 2 3 3 2]	[2 3 4 4 3]	[3 3 5 5 3]	[4 4 7 6 4]

APPENDIX-E

The multi-level multi-factor model built for the TEP simulator is detailed in this appendix. The scree plot for the 1st level PCA model is illustrated in Figure E.1. A fixed explained variance of 45% was chosen according to the training performance tabulated in Table E.1. Here, the performance for 55% is very close to 45%. However, a lower percentage was chosen to avoid noise in the test cases.

Furthermore, the scree plot for the 2nd level PCA model is illustrated in Figure E.2. A fixed variance of 45% was chosen according to the training performance tabulated in in Table E.2. Here a percent variance of 75% has a lower type-I error rate than a percent variance of 45%. However, 22 PCs for monitoring can be troublesome according to the MSPC point of view with regards to avoidance of noise.

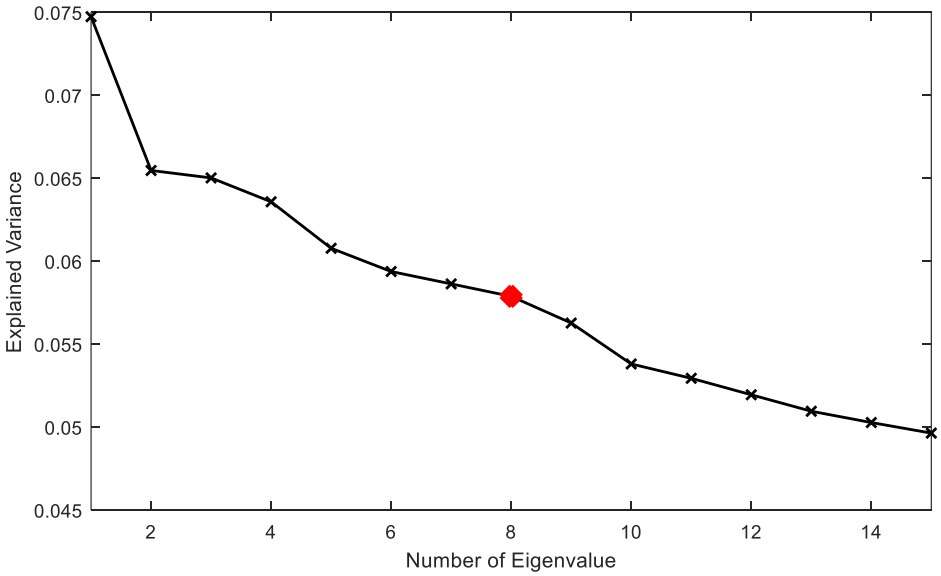

Figure E.1: Scree plot for the eigenvalues from the 1st level PCA model for the TEP simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table E.1: Offline training performance of the 1st level PCA model for multi-level multi-factor models with corresponding cumulative explained variance for the TEP process.

Metric Name (level)	Variance / Type I error (%)				
	35%	45%	55%	65%	75%
SPE(1%)	2.25	**2.08**	2.16	2.25	2
T^2(1%)	1.41	**0.83**	**0.83**	1.16	1
Number of PCs	6	**8**	9	11	13

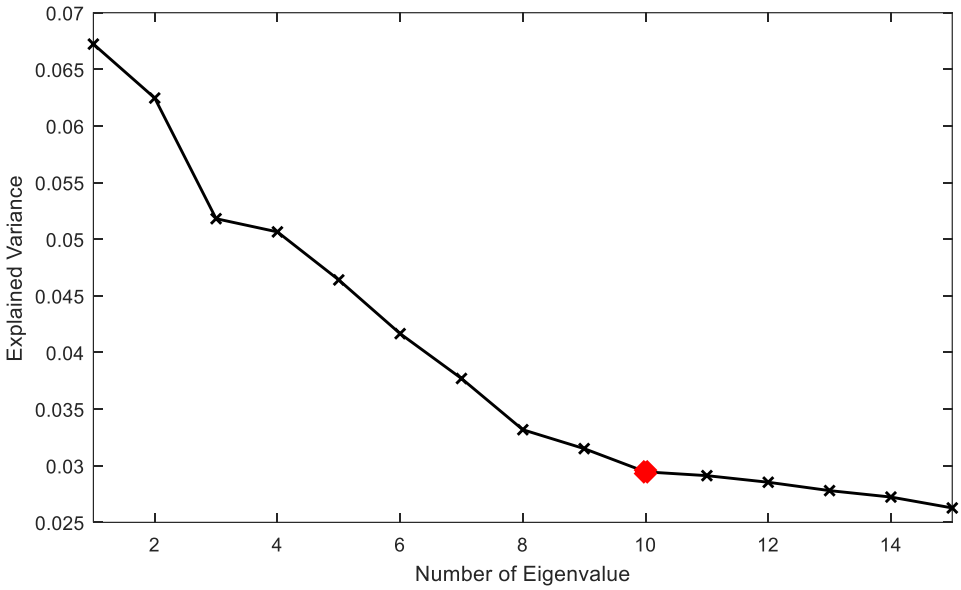

Figure E.2: Scree plot for the eigenvalues from the 2nd level PCA model for TEP simulator data representing normal operating conditions. The red dots represent the number of PCs selected.

Table E.2: Offline training performance of the 2nd level PCA model for multi-level multi-factor models with corresponding cumulative explained variance for the TEP process.

Metric Name (level)	Variance / Type I error (%)				
	35%	45%	55%	65%	75%
SPE(1%)	1.33	1.25	1.41	1.08	**0.9**
T^2(1%)	**0.66**	**0.66**	**0.66**	0.91	0.71
Number of PCs	**10**	14	18	22	27

BIBLIOGRAPHY

Abdi, H. and Williams, L. J. (2010) 'Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*. doi: 10.1002/wics.101.

Antoniadou, I., Cross, E. J. and Worden, K. (2013) 'Cointegration and the Empirical Mode Decomposition for the Analysis of Diagnostic Data', *Key Engineering Materials*, 569–570, pp. 884–891. doi: 10.4028/www.scientific.net/KEM.569-570.884.

Ashton, K., 2009. That 'internet of things' thing. *RFID journal*, *22*(7), pp.97-114.

Banerjee, A., Dolado, J.J., Galbraith, J.W. and Hendry, D., (1993) 'Co-integration, error correction, and the econometric analysis of non-stationary data', *OUP Catalogue*. p. 344. doi: 10.1093/0198288107.001.0001.

Bathelt, A., Ricker, N. L. and Jelali, M. (2015) 'Revision of the Tennessee eastman process model', in *IFAC-PapersOnLine*. doi: 10.1016/j.ifacol.2015.08.199.

Bersimis, S., Psarakis, S. and Panaretos, J. (2007) 'Multivariate statistical process control charts: an overview', *Quality and Reliability Engineering International*, 23(5), pp. 517–543. doi: 10.1002/qre.829.

Birol, G., Ündey, C. and Çinar, A. (2002) 'A modular simulation package for fed-batch fermentation: penicillin production', *Computers & Chemical Engineering*, 26(11), pp. 1553–1565. doi: 10.1016/S0098-1354(02)00127-8.

Box, G. E. P. (1954) 'Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification', *The Annals of Mathematical Statistics*. doi: 10.1214/aoms/1177728786.

Box, G. E. P., Luceño, A. and Paniagua-Quiñones, M. D. C. (2011) *Statistical Control by Monitoring and Adjustment: Second Edition*, *Statistical Control by Monitoring and Adjustment: Second Edition*. doi: 10.1002/9781118164532.

Box, G. and Narasimhan, S. (2010) 'Rethinking Statistics for Quality Control', *Quality Engineering*, 22(2), pp. 60–72. doi: 10.1080/08982110903510297.

Capaci, F., Vanhatalo, E., Kulahci, M., Bergquist, B. (2019) 'The revised Tennessee Eastman process simulator as testbed for SPC and DoE methods', *Quality Engineering*. 31(2), pp. 212-229. doi: 10.1080/08982112.2018.1461905.

Chen, J. and Liao, C. M. (2002) 'Dynamic process fault monitoring based on neural network and PCA', *Journal of Process Control*. doi: 10.1016/S0959-1524(01)00027-0.

Chen, J. and Liu, J. (1999) 'Mixture Principal Component Analysis Models for Process Monitoring', *Industrial & Engineering Chemistry Research*, 38(4), pp. 1478–1488. doi: 10.1021/ie980577d.

Chen, J. and Liu, K. (2002) 'On-line batch process monitoring using dynamic PCA and dynamic PLS models', *Chemical Engineering Science*, 57, pp. 63–75. doi: 10.1016/S0009-2509(01)00366-9.

Chen, Q., Kruger, U. and Leung, A. Y. T. (2009) 'Cointegration testing method for monitoring nonstationary processes', *Industrial and Engineering Chemistry Research*,

48(7), pp. 3533–3543. doi: 10.1021/ie801611s.

Chigira, H. (2008) ‘A test of cointegration rank based on principal component analysis’, *Applied Economics Letters*. doi: 10.1080/13504850600722096.

Cross, E. J., Worden, K. and Chen, Q. (2011) ‘Cointegration: a novel approach for the removal of environmental trends in structural health monitoring data’, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2133), pp. 2712–2732. doi: 10.1098/rspa.2011.0023.

Dayal, B. S. and MacGregor, J. F. (1997) ‘Recursive exponentially weighted PLS and its applications to adaptive control and prediction’, *Journal of Process Control*. doi: 10.1016/S0959-1524(97)80001-7.

Department for Business, Energy & Industrial Strategy (2017). *Made Smarter Review*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/made-smarter-review.

Dickey, D. A. and Fuller, W. A. (1979) ‘Distribution of the Estimators for Autoregressive Time Series With a Unit Root’, *Journal of the American Statistical Association*, 74(366), p. 427. doi: 10.2307/2286348.

Dickey, D. A. and Fuller, W. A. (1981) ‘Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root’, *Econometrica*. doi: 10.2307/1912517.

Ding, S. X. (2008) *Model-based fault diagnosis techniques: Design schemes, algorithms, and tools*. doi: 10.1007/978-3-540-76304-8.

Ding, S. X. (2014) ‘Data-driven design of monitoring and diagnosis systems for dynamic processes: A review of subspace technique based schemes and some recent results’, *Journal of Process Control*. doi: 10.1016/j.jprocont.2013.08.011.

Dong, D. and Mcavoy, T. J. (1996) ‘Nonlinear principal component analysis - Based on principal curves and neural networks’, *Computers and Chemical Engineering*. doi: 10.1016/0098-1354(95)00003-K.

Dong, D. and McAvoy, T. J. (1995) ‘Multi-stage batch process monitoring’, in *Proceedings of the American Control Conference*. doi: 10.1109/acc.1995.531208.

Downs, J. J. and Vogel, E. F. (1993) ‘A plant-wide industrial process control problem’, *Computers and Chemical Engineering*. doi: 10.1016/0098-1354(93)80018-I.

Engle, R. F. and Granger, C. W. J. (1987) ‘Co-Integration and Error Correction: Representation, Estimation, and Testing’, *Econometrica*, 55(2), p. 251. doi: 10.2307/1913236.

Erol, O. K. and Eksin, I. (2006) ‘A new optimization method: Big Bang-Big Crunch’, *Advances in Engineering Software*. doi: 10.1016/j.advengsoft.2005.04.005.

Escribano, A. and Peña, D. (1994) ‘Cointegration and common factors’, *Journal of Time Series Analysis*. doi: 10.1111/j.1467-9892.1994.tb00213.x.

Eslamloueyan, R. (2011) ‘Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee-Eastman process’, *Applied Soft Computing Journal*. doi: 10.1016/j.asoc.2010.04.012.

Ewan, W. D. (1963) ‘When and How to Use Cu-Sum Charts’, *Technometrics*, 5(1), pp. 1–22. doi: 10.1080/00401706.1963.10490055.

Gao, X. and Hou, J. (2016) ‘An improved SVM integrated GS-PCA fault diagnosis

approach of Tennessee Eastman process', *Neurocomputing*. doi: 10.1016/j.neucom.2015.10.018.

Ge, Z., Song, Z. and Gao, F. (2013) 'Review of Recent Research on Data-Based Process Monitoring', *Industrial & Engineering Chemistry Research*, 52(10), pp. 3543–3562. doi: 10.1021/ie302069q.

Geladi, P. and Kowalski, B. R. (1986) 'Partial least-squares regression: a tutorial', *Analytica Chimica Acta*, 185(C), pp. 1–17. doi: 10.1016/0003-2670(86)80028-9.

Goldrick, S. (2016). *IndPenSim*. [online] www.industrialpenicillinsimulation.com. Available at: http://www.industrialpenicillinsimulation.com/ [Accessed 22 Feb. 2020].

Goldrick, S., Stefan, A., Lovett, D., Montague, G., Lennox, B. (2015) 'The development of an industrial-scale fed-batch fermentation simulation', *Journal of Biotechnology*, 193, pp. 70–82. doi: 10.1016/j.jbiotec.2014.10.029.

Goldrick, S. Duran-Villalobos, C.A., Jankauskas, K., Lovett, D., Farid, S.S. and Lennox, B. (2019) 'Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process', *Computers and Chemical Engineering*. doi: 10.1016/j.compchemeng.2019.05.037.

Goodall, C. and Jolliffe, I. T. (1988) 'Principal Component Analysis', *Technometrics*, 30(3), p. 351. doi: 10.2307/1270093.

Granger, C. W. J. and Newbold, P. (1974) 'Spurious regressions in econometrics', *Journal of Econometrics*. doi: 10.1016/0304-4076(74)90034-7.

Greene, W. H. . (2017) *Econometric analysis, 8th ed., Prentice Hall*.

H. Hotelling (1947) 'Multivariable Quality Control—Illustrated by the Air Testing of Sample Bombsight', in *Techniques of Statistical Analysis*.

Hansen, B. E. (1992) 'Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends', *Journal of Econometrics*. doi: 10.1016/0304-4076(92)90081-2.

Harris, R. I. D. (1992) 'Testing for unit roots using the augmented Dickey-Fuller test. Some issues relating to the size, power and the lag structure of the test', *Economics Letters*. doi: 10.1016/0165-1765(92)90022-Q.

Harris, R. and Sollis, R. (2003) *Applied Time Series Modelling and Forecasting*. Chichester, UK: Wiley. Available at: https://www.wiley.com/en-gb/Applied+Time+Series+Modelling+and+Forecasting-p-9780470844434.

Jackson, J. E. and Mudholkar, G. S. (1979) 'Control procedures for residuals associated with principal component analysis', *Technometrics*. doi: 10.1080/00401706.1979.10489779.

Jia, M., Chu, F., Wang, F. and Wang, W. (2010) 'On-line batch process monitoring using batch dynamic kernel principal component analysis', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2010.02.004.

Jia, M., Xu, H., Liu, X. and Wang, N. (2012) 'The optimization of the kind and parameters of kernel function in KPCA for process monitoring', *Computers and Chemical Engineering*. doi: 10.1016/j.compchemeng.2012.06.023.

Jiang, Q. and Yan, X. (2018) 'Parallel PCA–KPCA for nonlinear process monitoring',

Control Engineering Practice. doi: 10.1016/j.conengprac.2018.07.012.

Joe Qin, S. (2003) 'Statistical process monitoring: basics and beyond', *Journal of Chemometrics*. doi: 10.1002/cem.800.

Johansen, S. (1988) 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control*, 12(2–3), pp. 231–254. doi: 10.1016/0165-1889(88)90041-3.

Johansen, S. (1992) 'Determination of cointegration rank in the presence of a linear trend', *Oxford Bulletin of Economics and Statistics*, 54(3), pp. 383–397. doi: 10.1111/j.1468-0084.1992.tb00008.x.

Johansen, S. (1995) 'A stastistical analysis of cointegration for i(2) variables', *Econometric Theory*. doi: 10.1017/S0266466600009026.

Kano, M., Tanaka, S., Hasebe, S., Hashimoto, I. and Ohno, H. (2004) 'Combined Multivariate Statistical Process Control', *IFAC Proceedings Volumes*. doi: 10.1016/s1474-6670(17)38745-1.

Kershen, G. and Golinval, J.-C. (2002) 'Non-linear generalization of principal component analysis: from a global to a local approach', *Journal of Sound and Vibration*, 254(5), pp. 867–876. doi: 10.1006/jsvi.2001.4129.

Ketelaere, B. De, Mertens, K., Mathijs, F., Diaz, D.S. and Baerdemaeker, J.D. (2011) 'Nonstationarity in statistical process control - Issues, cases, ideas', *Applied Stochastic Models in Business and Industry*, 27(4), pp. 367–376. doi: 10.1002/asmb.911.

Ketelaere, B., Hubert, M. and Schmitt, E. (2015) 'Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data', *Journal of Quality Technology*. doi: 10.1080/00224065.2015.11918137.

Kirchgässner, G. and Wolters, J. (2007) *Introduction to modern time series analysis*, *Introduction to Modern Time Series Analysis*. doi: 10.1007/978-3-540-73291-4.

Kosanovich, K. A., Piovoso, M.J., Dahl, K.S., MacGregor, J.F. and Nomikos, P. (1994) 'Multi-way PCA applied to an industrial batch process', in *Proceedings of 1994 American Control Conference - ACC '94*. IEEE, pp. 1294–1298. doi: 10.1109/ACC.1994.752268.

Kourti, T. (2003) 'Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions', in *Journal of Chemometrics*. doi: 10.1002/cem.778.

Kourti, T. and MacGregor, J. F. (1996) 'Multivariate SPC Methods for Process and Product Monitoring', *Journal of Quality Technology*, 28(4), pp. 409–428. doi: 10.1080/00224065.1996.11979699.

Kresta, J. V., Macgregor, J. F. and Marlin, T. E. (1991) 'Multivariate statistical monitoring of process operating performance', *The Canadian Journal of Chemical Engineering*, 69(1), pp. 35–47. doi: 10.1002/cjce.5450690105.

Kruger, U., Zhou, Y. and Irwin, G. W. (2004) 'Improved principal component monitoring of large-scale processes', *Journal of Process Control*, 14(8), pp. 879–888. doi: 10.1016/j.jprocont.2004.02.002.

Ku, W., Storer, R. H. and Georgakis, C. (1994) 'Uses of state estimation for statistical process control', *Computers and Chemical Engineering*. doi: 10.1016/0098-1354(94)80093-6.

Ku, W., Storer, R. H. and Georgakis, C. (1995) 'Disturbance detection and isolation by dynamic principal component analysis', *Chemometrics and Intelligent Laboratory Systems*, 30(1), pp. 179–196. doi: 10.1016/0169-7439(95)00076-3.

Kundu, M., Kundu, P. K. and Damarla, S. K. (2017) *Chemometric monitoring: Product quality assessment, process fault detection, and applications, Chemometric Monitoring: Product Quality Assessment, Process Fault Detection, and Applications*. doi: 10.1201/b21069.

Kwiatkowski, D., Phillips, P.C., Schmidt, P. and Shin, Y. (1992) 'Testing the null hypothesis of stationarity against the alternative of a unit root', *Journal of Econometrics*. doi: 10.1016/0304-4076(92)90104-y.

Li, G., Qin, S. J. and Yuan, T. (2014) *Nonstationarity and cointegration tests for fault detection of dynamic processes, IFAC Proceedings Volumes (IFAC-PapersOnline)*. IFAC. doi: 10.3182/20140824-6-ZA-1003.00754.

Li, R. F. and Wang, X. Z. (2002) 'Dimension reduction of process dynamic trends using independent component analysis', *Computers and Chemical Engineering*. doi: 10.1016/S0098-1354(01)00773-6.

Li, W., Yue, H.H., Valle-Cervantes, S. and Qin, S.J. (2000) 'Recursive PCA for adaptive process monitoring', *Journal of Process Control*, 10(5), pp. 471–486. doi: 10.1016/S0959-1524(00)00022-6.

Lin, Y., Kruger, U., Gu, F., Ball, A. and Chen, Q. (2019) 'Monitoring nonstationary and dynamic trends for practical process fault diagnosis', *Control Engineering Practice*. doi: 10.1016/j.conengprac.2018.11.020.

Lin, Y., Kruger, U. and Chen, Q. (2017) 'Monitoring Nonstationary Dynamic Systems Using Cointegration and Common-Trends Analysis', *Industrial & Engineering Chemistry Research*, 56(31), pp. 8895–8905. doi: 10.1021/acs.iecr.7b00011.

Lopes, J. A. and Sarraguça, M. C. (2018) 'Data Processing in Multivariate Analysis of Pharmaceutical Processes', in *Multivariate Analysis in the Pharmaceutical Industry*. Elsevier, pp. 35–51. doi: 10.1016/B978-0-12-811065-2.00002-3.

Lu, N., Gao, F. and Wang, F. (2004) 'Sub-PCA modeling and on-line monitoring strategy for batch processes', *AIChE Journal*, 50(1), pp. 255–259. doi: 10.1002/aic.10024.

MacKinnon, J. G. (1991) 'Critical Values for Cointegration Tests', in Engle, R. F. and Granger, W. J. (eds) *Long-Run Economic Relationships: Readings in Cointegration*. New York: Oxford University Press, pp. 267–277.

Malik, S. A. (1998) 'Book Review: Nonlinear Process Control, M. A. Henson and D. E. Seborg (eds), Prentice Hall PTR, Upper Saddle River, NJ 07458, xii+432 pp. ISBN 0-13-625179-x.', *International Journal of Robust and Nonlinear Control*, 8(7), pp. 643–645. doi: 10.1002/(SICI)1099-1239(199806)8:7<643::AID-RNC340>3.0.CO;2-K.

Montgomery, D. (2001) 'Introduction To Statical Quality Control', *Angewandte Chemie International Edition*, 40(6), p. 9823. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Montgomery, D. C., Jennings, C. L. and Kulahci, M. (2015) *Introduction to Time series Analysis and Forcasting, Statewide Agricultural Land Use Baseline 2015*. doi: 10.1017/CBO9781107415324.004.

Negiz, A. and Çinar, A. (1997) 'Pls, balanced, and canonical variate realization techniques for identifying VARMA models in state space', *Chemometrics and Intelligent Laboratory Systems*, pp. 209–221. doi: 10.1016/S0169-7439(97)00035-X.

Ng, Y. S. and Srinivasan, R. (2009) 'An adjoined multi-model approach for monitoring batch and transient operations', *Computers & Chemical Engineering*, 33(4), pp. 887–902. doi: 10.1016/j.compchemeng.2008.11.014.

NobelPrize.org. (n.d.). *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2003*. [online] Available at: https://www.nobelprize.org/prizes/economic-sciences/2003/ [Accessed 22 Feb. 2020].

Nomikos, P. and MacGregor, J. F. (1994) 'Monitoring batch processes using multiway principal component analysis', *AIChE Journal*, 40(8), pp. 1361–1375. doi: 10.1002/aic.690400809.

Nomikos, P. and MacGregor, J. F. (1995a) 'Multi-way partial least squares in monitoring batch processes', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/0169-7439(95)00043-7.

Nomikos, P.and MacGregor, J. F. (1995b) 'Multivariate SPC charts for monitoring batch processes', *Technometrics*, 37(1), pp. 41–59. doi: doi:10.1016/0967-0661(95)00014-L.

Odiowei, P. E. P. and Cao, Y. (2010) 'Nonlinear Dynamic Process Monitoring Using Canonical Variate Analysis and Kernel Density Estimations', *Industrial Informatics, IEEE Transactions on*, 6(1), pp. 36–45. doi: 10.1109/TII.2009.2032654.

Page, E. S. (1954) 'Continuous Inspection Schemes', *Biometrika*, 41(1/2), p. 100. doi: 10.2307/2333009.

Panetta, K. (2017). *Smarter With Gartner*. [online] Gartner.com. Available at: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/.

Phillips, P. C. B. and Perron, P. (1988) 'Testing for a unit root in time series regression', *Biometrika*. doi: 10.1093/biomet/75.2.335.

Qin, S. J. (2012) 'Survey on data-driven industrial process monitoring and diagnosis', *Annual Reviews in Control*. Elsevier Ltd, 36(2), pp. 220–234. doi: 10.1016/j.arcontrol.2012.09.004.

Raich, A. and Çinar, A. (1996) 'Statistical Process Monitoring and Disturbance Diagnosis in Multivariable Continuous Processes', *AIChE Journal*. doi: 10.1002/aic.690420412.

Rato, T. J. and Reis, M. S. (2013a) 'Advantage of using decorrelated residuals in dynamic principal component analysis for monitoring large-scale systems', *Industrial and Engineering Chemistry Research*. doi: 10.1021/ie3035306.

Rato, T. J. and Reis, M. S. (2013b) 'Defining the structure of DPCA models and its impact on process monitoring and prediction activities', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2013.03.009.

Rato, T. J. and Reis, M. S. (2013c) 'Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR)', *Chemometrics and Intelligent Laboratory Systems*. Elsevier B.V., 125, pp. 101–108. doi: 10.1016/j.chemolab.2013.04.002.

Roberts, S. W. (1959) ‘Control Chart Tests Based on Geometric Moving Averages’, *Technometrics*, 1(3), pp. 239–250. doi: 10.1080/00401706.1959.10489860.

Russell, E. L., Chiang, L. H. and Braatz, R. D. (2000a) ‘Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis’, *Chemometrics and Intelligent Laboratory Systems*, 51, pp. 81–93.

Russell, E. L., Chiang, L. H. and Braatz, R. D. (2000b) ‘Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis’, pp. 81–93.

Said, S. E. and Dickey, D. A. (1984) ‘Testing for unit roots in autoregressive-moving average models of unknown order’, *Biometrika*. doi: 10.1093/biomet/71.3.599.

Samanta, B. and Nataraj, C. (2009) ‘Use of particle swarm optimization for machinery fault detection’, *Engineering Applications of Artificial Intelligence*. doi: 10.1016/j.engappai.2008.07.006.

Sang, W. C., Martin, E.B., Morris, A.J. and Lee, I.B. (2005) ‘Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture’, *Industrial and Engineering Chemistry Research*. doi: 10.1021/ie049081o.

Shang, J., Chen, M., Ji, H., Zhou, D., Zhang, H. and Li, M. (2017) ‘Dominant trend based logistic regression for fault diagnosis in nonstationary processes’, *Control Engineering Practice*. Elsevier Ltd, 66(April), pp. 156–168. doi: 10.1016/j.conengprac.2017.06.011.

Shewhart, W. A. (1930) ‘Economic Quality Control of Manufactured Product 1’, *Bell System Technical Journal*, 9(2), pp. 364–389. doi: 10.1002/j.1538-7305.1930.tb00373.x.

Stefatos, G. and Hamza, A. Ben (2010) ‘Dynamic independent component analysis approach for fault detection and diagnosis’, *Expert Systems with Applications*. doi: 10.1016/j.eswa.2010.06.101.

Stock, J. H. and Watson, M. W. (1988) ‘Testing for Common Trends’, *Journal of the American Statistical Association*, 83(404), pp. 1097–1107. doi: 10.1080/01621459.1988.10478707.

Stubbs, S., Zhang, J. and Morris, J. (2013) ‘Multiway Interval Partial Least Squares for Batch Process Performance Monitoring’, *Industrial & Engineering Chemistry Research*, 52(35), pp. 12399–12407. doi: 10.1021/ie303562t.

Sun, H., Zhang, S., Zhao, C. and Gao, F. (2017) ‘A Sparse Reconstruction Strategy for Online Fault Diagnosis in Nonstationary Processes with No a Priori Fault Information’, *Industrial & Engineering Chemistry Research*, p. acs.iecr.7b00156. doi: 10.1021/acs.iecr.7b00156.

Sun, H., Zhang, S., Zhao, C. and Sun, Y. (2017) ‘Fault isolation method for nonstationary industrial processes’, *Proceedings of Control And Decision Conference (CCDC)*, pp. 6637–6642.

Tanatavikorn, H. and Yamashita, Y. (2017) ‘Batch Process Monitoring Based on Fuzzy Segmentation of Multivariate Time-Series’, *Journal of Chemical Engineering of Japan*, 50(1), pp. 53–63. doi: 10.1252/jcej.16we193.

Thornhill, N. (2008). *The Stirred Tank Heater Simulation*. [online] personal-pages.ps.ic.ac.uk. Available at: http://personal-

pages.ps.ic.ac.uk/~nina/CSTHSimulation/index.htm [Accessed 22 Feb. 2020].

Thornhill, N. F., Patwardhan, S. C. and Shah, S. L. (2008) 'A continuous stirred tank heater simulation model with applications', *Journal of Process Control*. doi: 10.1016/j.jprocont.2007.07.006.

Undey, C. and Cinar, A. (2002) 'Statistical Monitoring of Multistage, Multiphase Batch Processes', *IEEE Control Systems*. doi: 10.1109/MCS.2002.1035216.

uk.mathworks.com. (n.d.). *Data Sets and Examples - MATLAB & Simulink - MathWorks United Kingdom*. [online] Available at: https://uk.mathworks.com/help/econ/data-sets-and-examples.html [Accessed 22 Feb. 2020].

Vanhatalo, E., Kulahci, M. and Bergquist, B. (2017) 'On the structure of dynamic principal component analysis used in statistical process monitoring', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2017.05.016.

Wang, G., Liu, C., Cui, Y. and Feng, X. (2014) 'Tool wear monitoring based on cointegration modelling of multisensory information', *International Journal of Computer Integrated Manufacturing*, 27(5), pp. 479–487. doi: 10.1080/0951192X.2013.814162.

Wang, G., Yin, S. and Kaynak, O. (2014) 'An LWPR-based data-driven fault detection approach for nonlinear process monitoring', *IEEE Transactions on Industrial Informatics*. doi: 10.1109/TII.2014.2341934.

Wang, X., Kruger, U. and Irwin, G. W. (2005) 'Process monitoring approach using fast moving window PCA', *Industrial and Engineering Chemistry Research*. doi: 10.1021/ie048873f.

Wang, X., Kruger, U. and Lennox, B. (2003) 'Recursive partial least squares algorithms for monitoring complex industrial processes', *Control Engineering Practice*. doi: 10.1016/S0967-0661(02)00096-5.

Wise, B. M. and Ricker, N. L. (1991) 'Recent advances in nultivariate statistical process control improving robustness and sensitivity', *IFAC Symposium on Advanced Control of Chemical Processes, Toulouse, France, October*.

Wold, H. (1975) 'Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach', *Journal of Applied Probability*. doi: 10.1017/s0021900200047604.

Wold, S., Antti, H., Lindgren, F. and Öhman, J. (1998) 'Orthogonal signal correction of near-infrared spectra', in *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/S0169-7439(98)00109-9.

Wold, S., Esbensen, K. and Geladi, P. (1987) 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/0169-7439(87)80084-9.

Woodall, W. H. and Adams, B. M. (1993) 'The statistical design of cusum charts', *Quality Engineering*, 5(4), pp. 559–570. doi: 10.1080/08982119308918998.

Xie, L. and Kruger, U. (2006) 'Statistical Processes Monitoring Based on Improved ICA and SVDD', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 1247–1256. doi: 10.1007/11816157_156.

Xu, Z. and Chen, Q. (2007) ‘Nonstationary system monitoring using cointegration testing method’, *Key Engineering Materials*, 347, pp. 245-250. doi: 10.4028/www.scientific.net/KEM.347.245.

Yao, Y. and Gao, F. (2008) ‘Subspace identification for two-dimensional dynamic batch process statistical monitoring’, *Chemical Engineering Science*, 63(13), pp. 3411–3418. doi: 10.1016/j.ces.2008.04.007.

Yao, Y. and Gao, F. (2009) ‘Multivariate statistical monitoring of multiphase two-dimensional dynamic batch processes’, *Journal of Process Control*. doi: 10.1016/j.jprocont.2009.07.003.

Yu, J. and Qin, S. J. (2008) ‘Multimode process monitoring with bayesian inference-based finite Gaussian mixture models’, *AIChE Journal*. doi: 10.1002/aic.11515.

Yue, H. H. and Qin, S. J. (2001) ‘Reconstruction-based fault identification using a combined index’, *Industrial and Engineering Chemistry Research*. doi: 10.1021/ie000141+.

Zhang, S., Zhao, C. and Gao, F. (2019) ‘Incipient Fault Detection for Multiphase Batch Processes With Limited Batches’, *IEEE Transactions on Control Systems Technology*, 27(1), pp. 103–117. doi: 10.1109/TCST.2017.2755580.

Zhao, C., Wang, F., Lu, N. and Jia, M. (2007) ‘Stage-based soft-transition multiple PCA modeling and on-line monitoring strategy for batch processes’, *Journal of Process Control*. doi: 10.1016/j.jprocont.2007.02.005.