Observability and Economic aspects of Fault Detection and Diagnosis Using CUSUM based Multivariate Statistics

by

Mohamed Bin Shams

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Chemical Engineering

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis including any required final versions, as accepted by the examiners.

I understand that my thesis made electronically available to the public.

Abstract

This project focuses on the fault observability problem and its impact on plant performance and profitability.

The study has been conducted along two main directions. First, a technique has been developed to detect and diagnose faulty situations that could not be observed by previously reported methods. The technique is demonstrated through a subset of faults typically considered for the Tennessee Eastman Process (TEP); which have been found unobservable in all previous studies. The proposed strategy combines the cumulative sum (CUSUM) of the process measurements with Principal Component Analysis (PCA). The CUSUM is used to enhance faults under conditions of small fault/signal to noise ratio while the use of PCA facilitates the filtering of noise in the presence of highly correlated data. Multivariate indices, namely, T^2 and Q statistics based on the cumulative sums of all available measurements were used for observing these faults. The $ARL_{o.c}$ was proposed as a statistical metric to quantify fault observability. Following the faults detection, the problem of fault isolation is treated. It is shown that for the particular faults considered in the TEP problem, the contribution plots are not able to properly isolate the faults under consideration. This motivates the use of the CUSUM based PCA technique previously used for detection, for unambiguously diagnose the faults. The diagnosis scheme is performed by constructing a family of CUSUM based PCA models corresponding to each fault and then testing whether the statistical thresholds related to a particular faulty model is exceeded or not, hence, indicating occurrence or absence of the corresponding fault.

Although the CUSUM based techniques were found successful in detecting abnormal situations as well as isolating the faults, long time intervals were required for both detection and diagnosis. The potential economic impact of these resulting delays motivates the second main

objective of this project. More specifically, a methodology to quantify the potential economical loss due to unobserved faults when standard statistical monitoring charts are used is developed.

Since most of the chemical and petrochemical plants are operated under closed loop scheme, the interaction of the control is also explicitly considered. An optimization problem is formulated to search for the optimal tradeoff between fault observability and closed loop performance. This optimization problem is solved in the frequency domain by using approximate closed loop transfer function models and in the time domain using a simulation based approach.

The optimization in the time domain is applied to the TEP to solve for the optimal tuning parameters of the controllers that minimize an economic cost of the process.

Acknowledgment

First, I give my most enthusiastic and highest praise to Allah, the giver of my life and all of the good things I have experienced. He alone has given me the resources, the strength, the patience and the love of people that I have needed.

I owe my deepest gratitude to Professor Hector M. Budman. I appreciate his continual help, encouragement and insistence during the course of this work. I would also like to show my sincere thanks to Professor Thomas A. Duever for his continuous interest and raising morale. This thesis would not have been possible without having them together.

I would like to thank the members of my PhD examining committee: Professor Jose A. Romagnoli, Professor Kumaraswamy Ponnambalam, Professor Ali Elkamel, and Professor Luis Ricardez for their time, comments and valuable suggestions.

I would also like to thank my current and previous friends in the process systems engineering group and the other research groups at the University of Waterloo, especially Walid Al-Gherwi and Rosendo Diaz.

Finally, I wish to express my gratitude to the University of Bahrain for its support and sponsorship during my post-graduate studies wish greatly contributed to the successful completion of this study.

Dedication

T*o the memory of my honorable father,*

to my beloved mother,

to my warmhearted brothers and sisters,

to my lovely fiancée.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Process disturbances or faults if unobserved have a serious impact on process profitability, product quality, safety, productivity and pollution level. A fault may be defined as a deviation of at least one variable from an acceptable level (Isermann, 2006). As such, the fault is a state that may lead to malfunction or failure of the system. A fault diagnostic system is composed mainly of a detection algorithm to detect an abnormal operation due to the occurrence of one or more faults followed by a diagnosis scheme to isolate the particular fault. Efficient and advanced automated diagnostic systems that detect, identify, diagnose and consequently correct these abnormal process behaviors are of great importance to modern complex chemical industries. Whenever a process encounters a fault, the effect of this fault is propagated to all or some of the process variables. Most of the available fault detection algorithms involve comparing the observed behavior of the process to the corresponding output of a reference model where the latter may be mechanistic, empirical or a combination of both. If the fault is observable, the fault detection scheme will generate related symptom patterns which in turn are fed to the fault diagnosis scheme to precisely determine the root cause of the observed behavior.

Different problems associated with the fault detection and diagnosis problem have been reported in the literature. Despite their practical importance, *distinguishability* and *observability* have not been studied thoroughly in the literature.

Fault observability is related to the detection phase and it is interpreted as a delay in identifying the occurrence of a fault, i.e. the detection of abnormal operation. The distinguishability problem arises when it is desired to isolate the particular occurring fault but the system exhibits similar responses in variables used for detection while different faults occur.

Accordingly, an observable fault is defined as the one in which its effects can be observed from the available set of measurement variables. Although fault observability is very much to the fault detection phase, lack of observability can affect the fault diagnosis resolution, that is, the degree of uniqueness achieved in diagnosing a particular fault i.e. distinguishability. This project focuses on observability issues associated with the fault detection and diagnosis problem.

When known faulty data are passed through a calibrated fault detection model, whether this model is analytical, empirical or semi-empirical based, the monitored statistic should indicate the presence of a fault. If the detection algorithm does not signal the fault, then the specific fault is referred to as unobservable with that particular algorithm. Sometimes this means that the measured process variables do not contain enough information regarding the fault and more representative variable(s) should be used for detection. This is very much related to the sensor distribution or sensor selection design problem (Raghuraj *et al*., 1999). In other cases, a lack of observability may indicate that certain process variables need to be emphasized within the detection algorithm at the cost of other variables. This is especially important when those variables used for detection are tightly controlled to satisfy quality requirements and as a result of this, the information regarding the fault remains buried in the background noise. Then, these variables need to be emphasized so they can be observed.

For example, variables that are perfectly controlled during a closed loop operation do not exhibit large variability and if not properly emphasized or if excluded from the dataset used to build the

detection model, relevant information will be lost and the fault will not be detected. On the other hand, from a control performance perspective, small variability is generally required. Hence the requirement for good fault detection on the one hand and good control on the other tend to oppose each other. Therefore a tradeoff must be sought between closed loop performance and fault observability. This tradeoff has been addressed in the current project through the solution of an optimization problem.

On the other hand, as will be shown in the Tennessee Eastman Process (TEP), the signal to noise ratio is a key factor affecting the observability of the fault. Often, a fault occurs with very small magnitude and, most of the time, is completely buried in the measurement noise making its detection a very difficult task. On the other hand, even small faults with small magnitudes, if they persist for a long time, might have an important economical impact which motivates the need for their early detection. Hence, it can be seen that fault observability is affected by different factors. These factors lead to faults with different levels of observability, from completely unobservable faults which are mostly related to the sensor selection problem, to observable faults that can be detected by a variety of techniques. As a result, a closer investigation of those factors that affects detection with different levels of observability need to be considered. In the case of inevitable lack of observability, a methodology is required to mitigate the possible economical impact associated with problem.

On the basis of the above, in the current project the fault observability problem is considered from different perspectives. The following section summarizes the objective of the current project.

1.2 The objectives of the research

The following main objectives were sought in the current project:

(1) Development of new detection and diagnosis algorithm to deal with faults that have been found unobservable by other reported techniques. The algorithm is demonstrated on a subset of faults in the Tennessee Eastman Process (Downs and Vogel, 1993).

(2) Investigation of the trade-offs between fault observability and economics in the fault detection problem in chemical processes.

(3) Enhancement of fault observability in large scale industrial processes using proper tuning of feedback controllers while maintaining acceptable closed loop performance.

1.3 The contribution of the current project

The contributions of this project presented in each chapter can be summarized as follows. Chapter 2 contains an up to date literature review that cover the main aspects of fault detection, identification and diagnosis problem. The fault detection and diagnosis techniques can be broadly classified into three main categories: (1) analytical methods, e.g. observer based methods (2) empirical methods, e.g. univariate and multivariate statistical based methods (3) semi-empirical methods which combine different features from the two aforementioned methods. Most chemical and petrochemical processes are monitored on a frequent basis resulting in large amounts of measured data. This makes empirical methods, in particular, multivariate statistics more attractive for large scale industrial processes. Although the empirical based method is

adopted in this project, overviews of the methods based on either analytical or semi-empirical models are also briefly reviewed. In addition, previous attempts to tackle the fault detection-control interaction will be addressed in this chapter.

Chapter 3 demonstrates the fault observability problem through a realistic industrial simulator, namely, the Tennessee Eastman Process (TEP). A new statistical monitoring scheme, namely, CUSUM based PCA is proposed to tackle a subset of the TEP process that could not be observed using the fault detection methodologies previously reported in the literature. The proposed method was successful in detecting, identifying and diagnosing these faults.

Chapter 4 is motivated by the resulting significant detection's delays associated with the CUSUM based PCA algorithm. Whether a fault is not observed or at least required long time to be observed, in both cases there are economic losses associated with this lack of observability.

A methodology is proposed to quantify the potential economical losses due to unobserved faults, especially, when standard statistical monitoring techniques are used. The proposed methodology explicitly considers the inevitable interaction between feedback control and fault detection through the solution of an optimization problem.

Chapter 5 generalizes the proposed methodology given in chapter 4 by applying it to an industrial scale problem, namely, TEP simulator.

Chapter 6 concludes with detailed recommendations for future wok on the following topics: (1) the effect of nonlinearity on fault the observability ; (2) the optimal designed experiments for the identification and diagnosis of faults when empirical statistical models are used ; (3) the need for simultaneously optimizing plant design, feedback controllers and fault detection and diagnosis algorithms.

Most of the findings in this project have been presented at refereed publications and oral presentations at conferences and meetings as follows:

Refereed Publications

Bin Shams, M., Budman, H., Duever, T. (2010) "Finding a trade-off between observability and economics in the fault detection of chemical processes", ***Computer and Chemical Engineering***, (In Press)

Bin Shams, M., Budman, H., Duever, T. (2010) "Enhancing fault observability using feedback control", ***Industrial and Engineering Chemistry Research***, (In Press)

Bin Shams, M., Budman, H., Duever, T. (2010) "Detection, Identification and Diagnosis using CUSUM based PCA", ***Chemical Engineering Science***, (Submitted)

Bin Shams, M., Budman, H., Duever, T. (2010) "Fault detection using CUSUM based techniques with application to the Tennessee Eastman Process", ***Proceedings of the 9th International Symposium on Dynamic and Control of Process systems* (*DYCOPS*),** Leuven, Belgium

Oral presentations at conferences and meeting

Bin Shams, **M**, Duever, T., Budman, H. (2010) "Fault detection and Diagnosis: Observability and Isolability", ***60th Canadian Chemical Engineering Conference***, Saskatoon, Saskatchewan, Canada

Bin Shams, M., Budman, H., Duever, T. (2010) "Fault detection using CUSUM based techniques with application to the Tennessee Eastman Process", ***International Symposium on Dynamic and Control of Process systems* (*DYCOPS*),** Leuven, Belgium

Bin Shams, **M.**, Duever, T., Budman, H. (2010) "CUSUM based techniques for fault detection of chemical processes", ***Control and Statistics Conference***, Waterloo, Ontario, Canada

Bin Shams, **M.**, Duever, T.; Budman, H. (2009) "Fault detection in chemical process: observability and economic considerations", ***CORS/INFORMS Joint International meeting***, Toronto, Ontario, Canada

Bin Shams, **M**, Duever, T., Budman, H. (2009) "Observability and economics considerations in fault detection", ***Control and Statistics Conference***, Kingston, Ontario, Canada

Chapter 2

Literature Review

Automatic distributed control and model predictive control systems have brought great benefits to chemical and petrochemical industries. However, a very important control task in managing process plants still remains largely a non-automated activity performed by human operators. This is the task of responding to abnormal events in a process. This involves the timely detection of abnormal events, diagnosing its causal origins and then taking appropriate supervisory control decisions and/or actions to bring the process back to a normal safe operating state.

Closely associated with the area of fault detection and diagnosis is the research area of gross error detection, that is the identification of faulty or failed sensors in the process (Isermann, 2006). Related to gross error detection, data reconciliation or rectification is the task of providing estimates for the true values of sensor readings using material and energy balances (Romagnoli and Sanchez, 2000). However, since fault diagnostic systems can also be used to predict sensor failures, both gross error detection and data reconciliation can be considered as special cases of a more general fault diagnostic problem.

Fault detection and diagnosis is generally conducted on the basis of some mechanistic empirical or semi-empirical model of the process. Since most chemical processes exhibit nonlinearities, nonlinear models have been often used for detection and diagnosis purposes, especially when wide ranges of operating conditions are considered. However, in situations

where the process is being operated under closed loop in a close neighborhood of an operating point, linear models are almost always adequate (Kourti, 2002).

In general the diagnostic decision making process can be viewed as a series of transformations or mappings on process measurements. Fig.2.1 depicts the distinct transformation that the process data go through during the fault detection and diagnosis problem. The measurement space is a space of measurements x_1, x_2, … x_m; where $x_i \in R^n$ represents the input to the diagnostic systems. The feature space is the space of the points $y = (y_1, y_2, ..., y_i)$ where $y_i \in R^z$ is the i^{th} feature obtained as a function of the measurements. The dimension z may be less, equal or greater than n, depends on the features' model used. These measurements are analyzed and combined with the aid of *a priori* process knowledge to extract useful features about the process behavior. The mapping from the feature space to the decision space is usually developed to meet some objective function such as minimizing the misdetection or misclassification rates (Venkatasubramanian *et al*., 2003a). This transformation is achieved either using a discriminant function or in some cases by using simple threshold functions. The decision space is a space of points $d = [d_1, d_2, ..., d_k]$,where k is the number of decision variables obtained by suitable transformation of the feature space. The class space shown in Fig 2.1 is a set of integers $c = [c_1, c_2, ..., c_m]$,where m is the number of fault classes. The class space is thus the final interpretation of the diagnostic system delivered to the user. The transformation from decision space to class space is also performed using threshold functions, or symbolic reasoning (Chiang *et al*., 2001). Fault detection and diagnosis techniques can be broadly classified into three general categories: analytical methods, empirical and semi-empirical based methods (Chiang *et al*., 2001; Isermann, 2006; Venkatasubramanian, 2003). In the following sections, the main features of each method are highlighted.

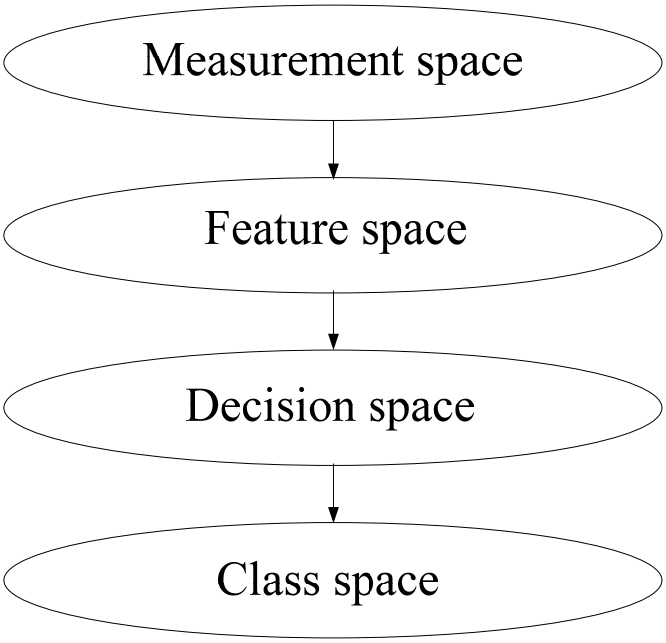


Fig.2.1 The general diagnostic decision making process

2.1 Analytical based methods

An obvious approach to detect process faults is to compare the process behavior with a process model describing the nominal normal process behavior, i.e. in the absence of faults. Then, the difference between the process measurements and the model predictions can be expressed by residuals that describe discrepancies between the process's actual behavior and the normal operation predicted by the model.

Relying on an explicit mechanistic model of a monitored plant, all analytical based fault detection and diagnosis methods require generating inconsistencies between the actual and model based predicted behavior.

In general, the problem of fault detection and diagnosis is one of identifying the state of a process based on its behavior as manifested through its sensors and actuators measurements.

When a fault or change occurs, a change in the relationship among these observed variables with respect to the model based predictions occurs, resulting in nonzero residuals. Although dynamic systems are continuous processes and all chemical engineering processes are nonlinear, almost all of the practically reported diagnostic tools use sampled data and assume linearity (Venkatasubramanian *et al*., 2003a; Yoon and MacGregor, 2001; Chiang *et al*., 2001). Thus, their application to nonlinear system requires a model linearization around a selected operating point.

In analytical based approaches, the effects of faults and other inputs such as disturbances and noise can be mathematically considered as either additive or multiplicative contributions. For instance, when these contributions are accounted for in additive fashion a discrete time state space model can be used as follows:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + B_f f(t) + B_d d(t) + B_n n(t) \\ y(t) &= Cx(t) + Du(t) + D_f f(t) + D_d d(t) + D_n n(t) \end{aligned} \tag{2.1}$$

where subscripts f, d, n are used for the matrices associated with fault, disturbances and noise effects, respectively. By introducing the shift operator defined by $q.x(t) = x(t+1)$ the above state space representation can be rewritten in terms of transfer functions with respect to the different inputs as follows:

$$y(t) = P(q).u(t) + P_f(g).f(t) + P_d(q).d + P_n(q)n(t) \tag{2.2}$$

where

$$\begin{aligned} P(q) &= C(qI - A)^{-1}B + D \\ P_f(q) &= C(qI - A)^{-1}B_f + D_f \\ P_d(q) &= C(qI - A)^{-1}B_d + D_d \\ P_n(q) &= C(qI - A)^{-1}B_n + D_n \end{aligned} \tag{2.3}$$

Equation (2.2) describes the effect of additive faults *f*, disturbances *d* and noise *n* on the plant's outputs. The transfer functions are assumed to be time invariant.

Alternatively, some faults are best modeled by assuming that their effect is entering the equations in a multiplicative form as follows:

$$\begin{aligned} x(t+1) &= (A+\Delta A)x(t) + (B+\Delta B)u(t) \\ y(t) &= (C+\Delta C)x(t) + (D+\Delta D)u(t) \end{aligned} \tag{2.4}$$

by introducing the shift operator in (2.4), this equation can be written in transfer function form as

$$y(t) = P^o(q)u(t) \tag{2.5}$$

$$P^o(q) = (C+\Delta C)(qI - A - \Delta A)^{-1}(B+\Delta B) + D + \Delta D \tag{2.6}$$

where $P^o(q)$ is the true transfer function for the physical system. Then, the discrepancy ΔP between the nominal model and the true system can be reformulated based on (2.6) as follows

$$\Delta P(q) = P^o(q) - P(q) = (C+\Delta C)(qI - A - \Delta A)^{-1}(B+\Delta B) + D + \Delta D - (C(qI-A)^{-1}B + D) \tag{2.7}$$

The discrepancy may be due to parametric faults or modeling errors. The latter may be due to inaccuracy in some of the physical parameters, unmodeled dynamics caused by simplifying the actual higher order system with a lower order model, approximating a nonlinear model with a linear model or may be due to simplifying assumptions in the derivation of a first principles model of the plant. So, in the absence of additive faults' contributions, disturbance and noise, as given by (2.2), the plant output would be:

$$y(t) = (P(q) + \Delta P(q)).u(t) = P(q)u(t) + \Delta P.u(t) \tag{2.8}$$

Equation (2.8) shows why the discrepancy in (2.7) is referred to as multiplicative with respect to $u(t)$ rather than additive. Analytical methods can be further classified into three main methods, namely, parameter estimation, state observer and consistency or parity relations. An overview of these three techniques is given in the following subsections.

2.1.1 Parameter estimation

Parameter estimation methods are appropriate if the process faults are associated with changes in model parameters (e.g. multiplicative faults) and mathematical models of the system are available. The model parameters are generally unmeasured, but can be estimated using standard parameter estimation techniques (Ljung, 1999; Isermann, 2006). Constructing the models from first principles models facilitates relating the model parameters directly to parameters that have physical meaning. Thresholds can be placed on the individual differences between the nominal model parameters values and the parameter estimates, or on some combination of these differences (Isermann, 2006; Chiang *et al*., 2001), where nominal values refer to those expected under normal or non faulty operating conditions. A comparison of the parameter estimates to their nominal values is carried out by computing the differences

$$\Delta p_j = p_j - \hat{p}_j \tag{2.9}$$

where p_j and $\hat{p}_j$ are the nominal value and the estimated value for the physical parameter *j*. Even if no faults are occurring in the plant, the Δp_j will not be equal to zero due to the noise. In other word Δp_j is a stochastic variable and a threshold must be used to indicate whether a fault has occurred or whether the changes in Δp_j are due to noise. A fault is detected when a single Δp_j is larger than some threshold and the parameter associated with the threshold violation is the one associated with the fault. This process monitoring procedure can be made more sensitive to slow drifts by applying exponential moving averages or cumulative sums on the parameter differences Δp_j (Chiang *et al*., 2001; Venkatasubramanian *et al*., 2003b).

2.1.2 State observer

Several approaches have been proposed for fault detection based on the classical Luenberger state observer, Kalman filter and the so called output observer (Ljung, 1999). The observer based methods reconstruct the output of the system from the measurement or subset of the measurement and either the state estimation error or the output error is used as a residual vector for plant status monitoring. The observer based methods are appropriate if the faults are associated with changes in actuators, sensors or unmeasurable state variables, that is, they are especially appropriate for detecting and isolating additive faults. To use these methods, a detailed mathematical model for the plant is required, preferably derived from first principles so that the states in the state space equations have a physical interpretation. For the states that are measured, a residual can be defined as the difference between the estimated state and the measured state. For states that cannot be directly measured, the residuals are defined based on the differences between the estimated plant output and the measured plant output. Based on thresholds on the residuals of the state variables or output variables, abrupt changes in operation related to process faults can be detected. Although the aforementioned techniques assume the linearity of the system under consideration, different nonlinear state estimators are available (Rawling and Bakshi, 2006). Among these techniques, Extended Kalman Filter (EKF) has received most of the attention due to its simplicity and effectiveness in tackling many nonlinear systems. The main drawbacks of EKF estimator are that they often result in a suboptimal solution due to the use of linearization at each time interval and inability to accurately incorporate physical state constrains. A class of estimators which do not require explicit linearization has been recently proposed e.g. particle filtering. Since they belong to the Markov Chain Monte Carlo (MCMC) based techniques, the obvious drawback in their use is their

associated computational cost. These nonlinear state estimators have been recently exploited for state feedback control applications (Prakash *et al*., 2010). However, their capabilities for fault detection and diagnosis have not been yet investigated.

2.1.3 Parity relations

In addition to the use of observers for the identification of faults based on residuals, another source to generate and test residuals is a parity relation. In this latter method a general equation for residual generation is formulated as a function of the observed input/output measurements as follows:

$$r(t) = V(q)u(t) + W(q)y(t) \tag{2.10}$$

where $r(t)$ is the residual vector and $V(q)$ and $W(q)$ are transfer function matrices and q is the shift operator. The advantage of parity relations over states observers based methods is their capability to *decouple* faults from other disturbances and noises.

Considering equation (2.2) the residual should be zero when the unknown inputs in (2.2) (the faults, $f(t)$, $d(t)$ and noise $n(t)$) are zero. Then, substituting (2.2) into (2.10) and setting the unknown inputs to zero results in the following equality:

$$V(q).u(t) + W(q).P(q).u(t) = 0 \tag{2.11}$$

To satisfy this equation for all inputs $u(t)$, it is necessary to satisfy

$$V(q) = -W(q).P(q) \tag{2.12}$$

Substituting (2.12) into (2.10) gives

$$r(t) = -W(q).[y(t) - P(q).u(t)] \tag{2.13}$$

where the transfer function P (q) can be obtained from the linearization of first principles models. Thus, specifying the transfer function W (q) is the key step in the determination of the parity relations. Substituting (2.2) into (2.13) gives the residual r (t) in terms of the unknown inputs:

$$r(t) = W(q).[P_f(q).f(t) + P_d(q).d(t) + P_n(q)n(t)] \quad (2.14)$$

Equation (2.14) is called the *parity relation* which determines the dependence of the residual on the faults, disturbances and noises.

Since analytical methods such as the one described above are generally based on detailed first principles, they can incorporate physical understanding into the process monitoring scheme to facilitate the fault diagnosis process. On the other hand, as the process becomes larger and more complicated, formulating an accurate model of the plant may become a very expensive task and the resulting models may be computationally demanding for real-time process monitoring. This motivates the use of empirical data-driven methods which are more suitable for large scale systems such as chemical plants as explained in the following section.

2.2 Empirical based methods

These methods rely on models built from data acquired from the system under normal or faulty condition depends on whether the model will be used for detection or diagnosis, respectively. Methods that use univariate and multivariate statistical control charts belong to this category. The general statistical control problem considers a repetitive statistical hypothesis testing problems where each problem is characterized by n process variables, x_1, x_2,...x_n. If n=1, the problem is univariate, otherwise it is multivariate. Because of the presence of noise and unmeasured disturbances other than the process disturbances related to the faults, these quality

variables are generally assumed as random variables. If the monitored variables are not cross correlated among themselves, then a process can be monitored using *n* different univariate charts. However, since most of the chemical and petrochemical processes are characterized by their interconnectivity nature, the monitored variables are generally highly correlated. This makes multivariate statistical charts a better approach (Kourti and MacGregor, 1995; Negiz and Cinar, 1997).

2.2.1 Conventional statistical control charts

For monitoring one quality variable x, where $x \sim N(\mu_0, \sigma_0^2)$, a Shewhart control chart for the mean can be built with the following control limits:

$$\begin{aligned} UCL &= \mu_0 + Z_{\alpha/2}(\sigma_0/\sqrt{n}), \\ CL &= \mu_0 \\ LCL &= \mu_0 - Z_{\alpha/2}(\sigma_0/\sqrt{n}) \end{aligned} \tag{2.15}$$

where μ_0, σ_0 and n are the in-control mean, in-control standard deviation and the number of samples, respectively. $Z_{\alpha/2}$ is the value of the standard normal distribution exceeded with probability $\alpha/2$. A natural multivariate extension to the univariate chart is the Hotelling's multivariate control chart (Hotelling, 1947). This procedure assumes *p* quality variables that are jointly distributed as *p independently normally distributed* random variables. The Hotelling's multivariate control chart is used to test that a statistically significant shift in the mean has occurred as per the following test

$$\chi_i^2 = (x_i - \mu_0)^T \sum{}^{-1} (x_i - \mu_0) > h_1 \tag{2.16}$$

where h_1, $\sum$ and μ_0 are the control limit, in-control data covariance matrix and the in-control mean vector, respectively. On the other hand, when the values of $\sum$ and μ_0 are not exactly

known, the in-control sample covariance matrix (S) and mean vector $\bar{x}$ are used instead for the calculation of the Hotelling statistics. The UCL can be calculated exactly as a function of the *F* distribution and given as:

$$\frac{p.(m+1).(m-1)}{m^2 - mp}.F_{\alpha,p,m-p} \tag{2.17}$$

where $F_{\alpha,p,m-p}$ denotes the *F*-percentile and *p* ,*(m-p)* and *m* are the numerator degrees of freedom, denominator degrees of freedom and the number of samples estimated during the in the normal in-control phase (calibration stage), respectively.

Just as it is important to monitor the process *mean* vector μ in the multivariate case it is also important to monitor the process *variability*. The process variability is described by the $p \times p$ covariance matrix. Two main procedures are available to calculate this matrix (Alt, 1984; Montgomery, 1997). The first approach is based on conducting repeated significance tests of the hypothesis that the process covariance matrix is equal to a particular matrix of constants denoted by $\sum$. If this procedure is used, the monitored statistic on the control chart for the i^{th} sample is

$$W_i = -p.n + p.n.\ln(n) - n.\ln(|A_i|/|\sum|) + tr(\sum^{-1}.A_i) \tag{2.18}$$

where, $A_i = (n-1).S_i$ and S_i is the sample covariance matrix for sample i . For this statistic, the upper control limit is given by $UCL = \chi^2_{\alpha,p(p+1)/2}$.

The second approach is based on the sample generalized variance $|S|$. This statistic is widely used to measure the multivariate dispersion. The control limits are defined by considering $|S|$ as a random variable, the mean $E(|S|)$, the variance $V(|S|)$ and by assuming that most of the probability distribution of $|S|$ is contained in the interval $E|S| \pm 3.\sqrt{V(|S|)}$ (Montgomery, 1997). The latter approach is widely used to measure multivariate variability.

Because Hotelling's charts are based on the most recent observations, they are insensitive to small and moderate shifts in the mean vector or variability (covariance) matrix. Several multivariate cumulative sums (CUSUM) based procedures and Exponentially Weighted Moving Average (EWMA) procedures have been reported (Alt, 1984; Croiser, 1988; MacGregor and Kourti, 1995, Lowry *et al*., 1992). Woodall and Ncube (1985) recommended monitoring each of the *p* quality characteristic individually with cumulative sum charts. Croiser (1988) proposed two different multivariate CUSUM charts. He used the in-control average run length ($ARL_{i.c}$) and the out-of-control average run length ($ARL_{o.c}$) as performance indices, where the subscripts (*i.c*) and (*o.c*) stand for in-control and out-of-control, respectively. $ARL_{i.c}$ is defined as the average number of points the process remains in-control when the system is operated normally. On the other hand, $ARL_{o.c}$ is defined as the average number of points elapsing between the occurrences of a fault until the detection of an out of control condition. Accordingly, one expects a good performance when a control chart is characterized by large $ARL_{i.c}$ and small $ARL_{o.c}$. Based on (Crosier, 1988) the multivariate CUSUM method can be based on the following statistics

$$
\begin{aligned}
&C_i = \sqrt{(S_{i-1} + x_i)^T \sum{}^{-1} (S_{i-1} + x_i)} \\
&S_i = 0 \qquad\qquad if\ C_i \le k_1 \\
&S_i = (S_{i-1} + x_i).(1 - k_1 / C_i) \qquad if\ C_i > k_1
\end{aligned}
\tag{2.19}
$$

where *i* is the sample number , S_0 =0 and k_1>0. This MCUSUM serve to indicate an out of control condition when $\gamma_i = \sqrt{S_i^T . \sum{}^{-1} . S_i} > h_2$

The multivariate EWMA given by (Lowry *et al*, 1992) is also a natural extension of the univariate EWMA defined as

$$Z_i = R.x_i + (I - R).Z_{i-1} \tag{2.20}$$

where $Z_0 = 0$ and R=diag(r_1,$r_{2,\ldots}r_p$), $0<r_j\leq 1$, j=1,2…p. The MEWMA chart gives an out of control signal when $T_i^2 = Z_i^T \sum_{Z_i}^{-1} Z_i > h_3$

where again $h_3>0$ is chosen to achieve a predefined $ARL_{i.c}$ and $\sum_{Z_i}$ is the covariance matrix of Z_i (Lowry *et al*., 1992). A key drawback of multivariate CUSUM methods used in the literature for quality control is that they have not explicitly addressed the correlations among the variables and therefore they have been sensitive to noise, collinearity and unmeasured disturbances. One of the contributions of the current thesis is to explicitly address this problem by combining CUSUM based charts with Principal Component Analysis based statistics in order to reduce the sensitivity of the resulting model to noise due to the correlation among the data. This is further explained in the following section.

2.2.2 Latent variables methods

The above multivariate statistics assumed that the monitored variables are neither autocorrelated nor crosscorrelated. For chemical and petrochemical processes that frequently involve recycle streams and heat integration schemes and are often operated under closed loop control, the assumption of independence among variables is rarely accurate. Methods based on multivariate statistical analysis, in particular, Principal Component Analysis (PCA) and Projection to Latent Structures (PLS), have been used to deal with the problem of cross correlation. There are different versions of PCA/PLS models reported in the literature (Venkatasubramanian *et al*., 2003c). An enhanced version of these techniques based on wavelet analysis has been also proposed to deal with both crosscorrelated and autocorrelated variables (Bakshi, 1998) as well as with robustness issues (Chen *et al*., 1996; Wang and Romagnoli, 2005). Combining wavelet analysis with PCA provides multi-resolution or multiscale

capabilities to the fault detection and isolation schemes. In particular, it enhances the performance of the detection and isolation when the frequency of the fault is confined to a specific frequency band i.e. frequency localization, otherwise it is comparable to single scale algorithms (Yoon and MacGregor, 2004). In addition, these multivariate statistical techniques have the capability of reducing the dimensionality of the problem at hand by summarizing the original variables using new variables i.e. latent variables (Jackson, 1991; MacGregor *et al*., 1994, Kourti, 2002; Venkatasubramanian, 2003c). These methods are especially attractive when the available process measurements are highly correlated but only a small number of faults have to be monitored. When the process data are highly correlated, these data can be projected onto a smaller variable space described by the principal components or latent variables, thus reducing the dimensionality of the problem. Ku *et al*., (1995) proposed a dynamic version of the PCA model using a lagged version of the data matrix used to identify the PCA model. To Account for the nonlinearity associated with most chemical processes, Kramer, (1991) proposed a neural network based PCA model. In addition to the input and output layers, he used an internal layer called *bottleneck* to provide a dimensionality reduction capability. Maulud *et al*., (2006) proposed a multiscale nonlinear PCA that combines the Kramer's method with wavelet analysis. Dong and McAvoy, 1996 combined the nonlinear generalization of the PCA known as the principal curves (Hastie *et al*., 1989) with neural networks. Alternatively, (Lee *et al*., 2004) proposed the Kernel based PCA. The latter can be considered more attractive than the neural network based PCA, since it requires straightforward solution of an eigenvalue problem, similar to the one solved for conventional static PCA. Since most of these techniques are based on the idea of static PCA models this technique will be explained in detail.

The conventional linear PCA model is identified from normal operation of the system. The basic idea of principal component analysis is to reduce the dimensionality of highly correlated data, while retaining as much as possible the variability present in the data. This is achieved by transforming the measured data to a new set of variables, i.e. the *principal components*, which are ordered such that the first principal component explains most of the variability present in the original variables while successive components represent progressively lower variability values. It is assumed that **x** is a vector of a large number of *n* random variables referred to as measurement signals. The vector **x** may contains inputs and/or output variables of a process. Then a reduced set of a considerably smaller number $a < m$ variables is searched which preserve most of the information given in the variances and covariance of these variables. This is accomplished by calculating a set of orthogonal vectors in the directions where most of the data variation occurs. To illustrate the problem, 8 measurements of two variables $x_1(k)$ and $x_2(k)$ are considered. The measurements are presented in two vectors

$$\mathbf{x}_1^T=[x_1(1)\ x_1(2)\ \ldots\ x_1(8)]$$

$$\mathbf{x}_2^T=[x_2(1)\ x_2(2)\ \ldots\ x_2(8)] \tag{2.21}$$

A data matrix

$$X=[\mathbf{x}_1\ \mathbf{x}_2] \tag{2.22}$$

is formed containing all measured data within the coordinate system $(\mathbf{x}_1|\mathbf{x}_2)$. As Fig. 2.2 shows, the measured data fluctuate in both directions of coordinate $\mathbf{x}_1$ and $\mathbf{x}_2$.

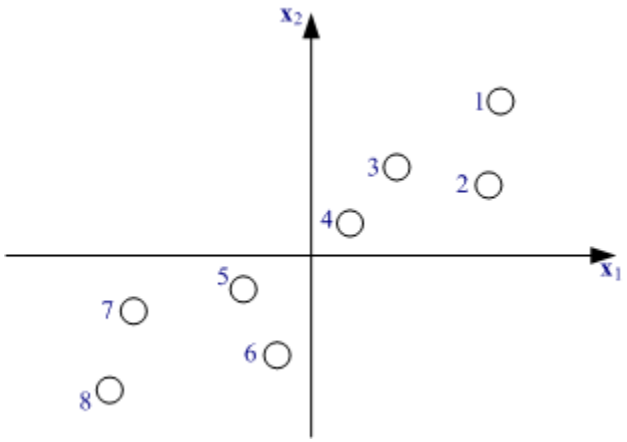


Fig.2.2 Plot of the 8 measurements of two variables

Then a transformation to a new two dimensional coordinate system ($\mathbf{t}_1$, $\mathbf{t}_2$) is sought, in which the variances of the data are the largest in the direction of $\mathbf{t}_1$ and second largest along the direction of $\mathbf{t}_2$, where $\mathbf{t}_1$ and $\mathbf{t}_2$ are referred to as the first and second principal components. A further condition is that ($\mathbf{t}_1$, $\mathbf{t}_2$) forms an orthogonal coordinate system as depicted in Figure 2.3. The new data can then be described by a matrix T= [$\mathbf{t}_1$ $\mathbf{t}_2$].

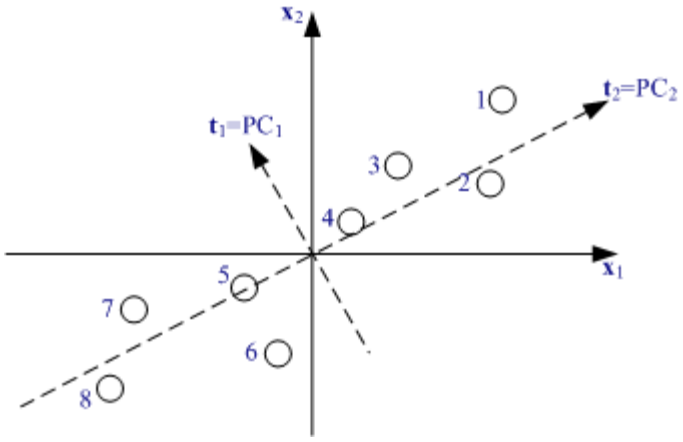


Fig.2.3 Plot of the 8 measurements of two variables along the two principal components ($\mathbf{t}_1$, $\mathbf{t}_2$)

For a general *n*-dimensional case, the PCA algorithm can be summarized as follows:

Transform a data matrix with *n* variables and *m* measurements, $X=[\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_m]$ into a new data matrix $T=[\mathbf{t}_1,\mathbf{t}_2\ldots\mathbf{t}_a]$, with also *m* measurements, but of smaller dimension $a < n$ .This can be obtained through the use of a transformation matrix P as follows:

$$T_{[\,m\,x\,a]}=X_{[\,m\,x\,n]}.P_{[\,n\,x\,a]} \tag{2.23}$$

$$P=[\mathbf{p}_1\ \mathbf{p}_2\ \ldots\mathbf{p}_a] \tag{2.24}$$

As this transformation is orthonormal, it satisfies

$$P^T P=I \tag{2.25}$$

Therefore according to (2.23), also

$$X=T\,P^T \tag{2.26}$$

The T matrix is referred to as the *score matrix* and P is the *loading matrix*. Equation (2.26) can be rewritten as follows

$$X=\mathbf{t}_1\,\mathbf{p}_1{}^T+\mathbf{t}_2\,\mathbf{p}_2{}^T+\ldots+\mathbf{t}_a\,\mathbf{p}_a{}^T=\sum \mathbf{t}_j\,\mathbf{p}_j{}^T \tag{2.27}$$

To find the elements $\mathbf{p}_j$, of the transformation matrix P which explains the variability in decreasing order of importance a stepwise optimization has to be solved. For each step *j* with

$$\mathbf{t}_j=X\,\mathbf{p}_j \tag{2.28}$$

a maximal variance of data $\mathbf{t}_j$ means

$$max\ \mathbf{t}_j{}^T\,\mathbf{t}_j=max\ (X\,\mathbf{p}_j)^T\,(X\,\mathbf{p}_j)=max\ \mathbf{p}_j{}^T\,X^T\,X\,\mathbf{p}_j \tag{2.29}$$

under the constraint

$$\mathbf{p}_j{}^T.\mathbf{p}_j=1 \tag{2.30}$$

A standard approach for this optimization problem is to use of the method of Lagrange multipliers. If the function $f(\mathbf{p}_j)$ has to be maximized under the condition $g=p_j{}^T p_j-1=0$, the corresponding loss function is defined as follows

$$V=f(\mathbf{p}_j)-\lambda_j.g(\mathbf{p}_j) \tag{2.31}$$

where λ_j is the Lagrange multiplier. This lead to

$$V= \mathbf{p}_j^T X^T X \mathbf{p}_j - \lambda_j(\mathbf{p}_j^T\mathbf{p}_j-1) \tag{2.32}$$

$$dV/d\mathbf{p}_j=2X^TX\mathbf{p}_j-2\lambda_j \mathbf{p}_j= 0 \tag{2.33}$$

$$[X^T X- \lambda_j I]\mathbf{p}_j=0 \tag{2.34}$$

$$A= X^T X \tag{2.35}$$

$$[A- \lambda_j I] \mathbf{p}_j=0 \tag{2.36}$$

Thus, the solution has the form of a classical eigenvalue problem. A is proportional to the covariance matrix for zero mean variables of the measured data, λ_j is an eigenvalue and $\mathbf{p}_j$ is an eigenvector of the matrix A. From (2.36) it follows that

$$\mathbf{p}_j^T A \mathbf{p}_j = \mathbf{p}_j^T \lambda_j \mathbf{p}_j \tag{2.37}$$

and inserting in (2.29) yields for the maximal variance

$$max\ \mathbf{t}_j^T \mathbf{t}_j = max\ \mathbf{p}_j^T \lambda_j \mathbf{p}_j \tag{2.38}$$

Therefore, the eigenvalue is proportional to the variance explained by the corresponding eigenvector.

Since most of the dimensionality reduction techniques are scale dependent, before applying PCA a pretreatment of the process data are required. For example, one might have to remove those variables which have no information relevant to monitoring the process. However, the most important pretreatment step is autoscaling. Process data need to be scaled to avoid particular variables to dominate the outcome if absolute values would be used and to ensure that each variable is given equal weight within the correlation between observations and inferred faults. Autoscaling is accomplished by subtracting from each variable their samples mean

assuming that for a linear process the fault related information is mostly contained in the deviations of the data from the mean. Then, the resulting mean centered variables are divided by their corresponding standard deviation. This step scales each variable to a unit variance and avoids the model predictions being dominated by the high variance variables if absolute values would be used.

Let the pretreated data matrix be $X \in R^{m \times n}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{n2} & \cdots & x_{mn} \end{bmatrix} \tag{2.39}$$

where m is the number of samples and n is the number of variables. Then X can be decomposed as

$$X = TP^T + E \tag{2.40}$$

where the columns of P ($n \times a$) are the principal component loadings and the columns of T ($m \times a$) are the scores corresponding to each observation. The residual is then expressed as

$$E = X(I - PP^T) \tag{2.41}$$

For fault detection, the T^2 statistics based on the first a principal components is defined as

$$T_a^2 = \sum_{i=1}^{a} \frac{t_i^{\,2}}{\lambda_i} \tag{2.42}$$

where λ_i is the i^{th} eigenvalue of the covariance matrix of X. The T^2 statistic given in (2.42) can be viewed as an ellipsoid in a dimensional space. Confidence limits for T^2 at confidence level $(1-\alpha)$ can be related to the F-distribution as follows

$$T_{m,a}^2 = \frac{(m-1)a}{m-a} F_{a,m-a} \tag{2.43}$$

where $F_{a,m-a}$ is the upper 100α % critical point of the F distribution with a and (m-a) degrees of freedom.

Although the first principal components often captures a large part of the variability in the data, monitoring process variables by T^2 based on the first PC's is generally not sufficient. The first PC's generally captures the steady state correlation and common cause variations where the latter refer to the predicted variability and/or the background noise (Kourti and MacGregor, 1995). If a totally new type of event occur which was not present in the reference data used to identify the reference model, then additional PC's will become significant and the new observation vector x_i will move off the calibrated plane. Such new event can be detected by computing the squared prediction error (SPE). Let $x_i(1\times n)$ denotes the i^{th} multivariate observation whose corresponding score is t_i=x_i.P. The prediction from the PCA model for x_i is given by $\hat{x}_i = t_i.P^T = x_i P.P^T$. Then, the p dimensional error vector is given by $e_i = x_i - \hat{x}_i$ and the corresponding SPE or Q is then defined as follows

$$Q = e_i.e_i^T \tag{2.45}$$

The SPE can be thought of as a measure of the plant model mismatch. The confidence limits for SPE are given by Jackson and Mudholkar, (1991). This test suggests the existence of abnormal condition when $SPE > Q_\alpha$, where Q_α is defined as follows

$$Q_\alpha = \Theta_1\left(1+\frac{c_\alpha.h_0.\sqrt{2.\Theta_2}}{\Theta_1}+\frac{\Theta_2.h_0(h_0-1)}{\Theta_1^2}\right)^{\frac{1}{h_0}} \tag{2.46}$$

$$\Theta_i = \sum_{j=p+1}^{n}\lambda_j^i; \quad for\, i = 1,2,3 \tag{2.47}$$

$$h_0 = 1-\frac{2.\Theta_1.\Theta_3}{3.\Theta_2^2} \tag{2.48}$$

c_α is the confidence limits for the $(1-\alpha)$ percentile in a normal distribution. The confidence limits calculated in equation (2.43) are based on the assumptions that the measurements are time independent and multivariate normally distributed. Thus, the confidence limits in (2.45)-(2.47) were derived assuming that the errors are identically and independently normally distributed with zero mean. In practice, the normality assumption maybe very inaccurate but this may not affect significantly the results due to central limit theorem considerations. The highly autocorrelated data would certainly affect the confidence limits for T^2 chart. Also, improper choice of the number of PC's to be used in the model may result in autocorrelated residuals, which will affect the confidence limits for the SPE chart.

2.2.3 Higher order statistics (HOS)

Similar to PCA, most of the statistical monitoring techniques are based on first and second order statistics such as the mean, variance, autocorrelation function and power spectrum. Their extensive use is due to the ease of their implementation as well as their interpretation which is very much related to the characteristics of the signal, like centered tendency, spread and energy. Second order measures are a subset of the Higher Order Statistics (HOS) of the signal. The *third* and *fourth* order measures contain additional information about the time series, which does not show up in the second order measures.

Choudhury *et al*.2006 used the HOS for fault detection, in particular, for valve stiction. He showed that these measures can be used to quantify the deviation from *gaussianity* that is associated with the resulting nonlinear behavior following the occurrence of valve stiction. Stiction usually manifests itself as a limit cycle in the controlled and manipulated variables. They showed, both through simulations and industrial data, that with aid of the two indices, the non

gaussianity index (NGI) and the non linearity index (NLI), valve stiction can be effectively detected (Choudhury *et al*., 2004; Choudhury *et al*., 2006).

The bispectrum is the frequency domain counterparts of the third order cumulant. The cumulant is defined as the natural logarithm (log_e) of the moment generating function of a certain random variable (Mendel, 1991).

Specifically, one can show mathematically that the cumulant is a nonlinear combination of the first, second and third order moments (Mendel, 1991). Intuitively, it measures the *skewness* of a random variable's probability density function. There are two methods to estimate the bispectrum, namely, the direct and indirect methods. Practically the indirect method is preferable since it is solely require the Fourier transform of the signal's third cumulant.

To improve the statistical properties of the bispectrum estimator, the normalized bispectrum (also known as bicoherence) is used. A potential disadvantage of the bispectrum for detection of nonlinear limit cycle oscillations is that a limit cycle may have symmetrical waveforms (e.g. square wave or triangle wave) which may cause lack of detection since the bispectrum of a symmetrical waveform is zero. The proficiency of bicoherence in detecting oscillatory faults considered in this project, e.g. valve stiction, has been experimented, however, with no satisfactory outcomes. Therefore, no results are shown for brevity.

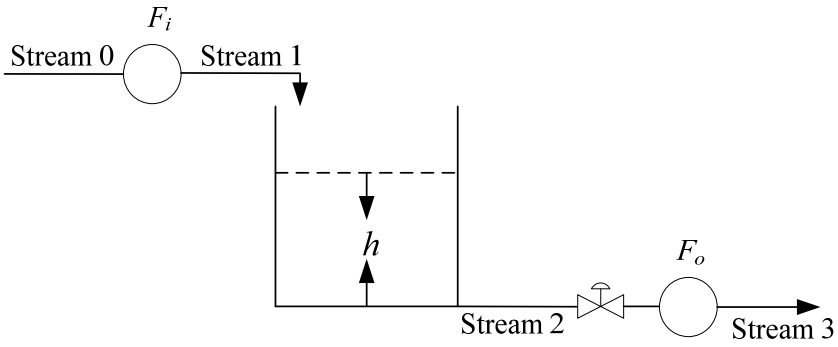
2.3 Semi-empirical based methods

In this class of models, measurements are combined with qualitative knowledge about the process to build a model for fault detection and diagnosis purposes. Causal analysis, expert systems and pattern recognition algorithms are some examples of semi-empirical based methods.

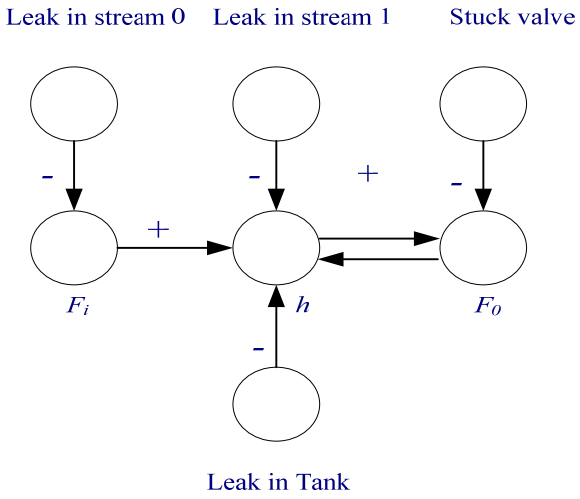
2.3.1 Causal analysis

Causal analysis is based on a causal relationship between fault symptoms and the process variables, and it is primarily used for fault diagnosis. The signed directed graph (SDG) is a qualitative model based approach for fault diagnosis that incorporates causal analysis. It is a map showing the relationship of the process variables and it is also reflects the behavior of the equipment involved as well as general system topology (Venkatasubramanian *et al*., 2003b). A SDG for a gravity flow tank system is shown in Figure (2.4). Nodes can correspond to process variables, specific sensors, system faults, components or subsystem failure. A threshold is defined for each node. The nodes takes up values of (0), (-) or (+) when its corresponding value is normal, above or below its threshold, respectively. Arc signs associated with each directed arc can take values (+) and (-) representing whether the cause and effect change in the same direction or the opposite direction respectively. For example, when a leak in (stream 1) in figure 2.4 occurs, it will decrease the liquid height and as a result a (-) sign is used on the arc. The goal of utilizing the SDG for diagnosing faults is to locate the possible *root nodes* representing the system faults based on the observed symptoms. To achieve this, the measured node deviations are propagated from the *effect nodes* to *cause nodes* via consistent arcs, until the *root nodes* are identified. An arc is defined to be consistent if the product of the sign of the *cause node*, the sign of the arc and the sign of the *effect node* is positive. For complex and large scale systems, a SDG for the process can be developed from the system's equations. Alternatively, the SDG can also be developed based on the knowledge of the process from an expert or historical data. The main drawbacks of using SDG include the lack of resolution, potentially long computing times and assumptions of a single fault at each time. Vedam and Venkatasubramanian (1999) used the SDG to automate the interpretation of the contribution plot (Miller *et al*., 1993) when PCA

model is used for detection. This is in fact necessary since the contribution plot is based on a correlation model and is consequently lacking causality that is necessary for isolating the fault. Therefore, SDG can supplement the contribution plots for better diagnosis capabilities.



(*a*)



(*b*)

Fig.2.4 (*a*) A gravity flow tank system. (*b*) Signed directed graph for the gravity tank system

2.3.2 Expert system

An experienced engineer or domain expert is generally capable of diagnosing faults in a much shorter time than an inexperienced operator because the experience they have accumulated

about the process i.e. *a prior* knowledge. To assist the personnel to diagnose faults, the expert's experience can be formulated as a set of *if-then* rules, which can be used to build an expert system. This is referred to as a *shallow knowledge expert* system in contrast to the *deep knowledge expert* system which is based on a formal mathematical model of the system or a complete behavioral description of its components during faulty and normal cases (Chiang *et al*., 2001). A combined version of both knowledge based expert system i.e. *shallow* knowledge and *deep* knowledge expert systems is usually used for added reliability (Chiang *et al*., 200l; Venkatasubramanian *et al*., 2003b).

2.3.3 Pattern recognition

Pattern recognition methods for fault detection and diagnosis define the relationship between the data patterns and fault classes without modeling the internal process states explicitly. These approaches include artificial neural networks (ANN), and self organizing maps (Baughman and Liu, 1995; Vaidyanathan and Venkatasubramanian, 1991). Empirical based methods are dimensionality reduction techniques based on rigorous multivariate statistics, whereas neural network and self organizing map are black box methods that learn the patterns based entirely from training session (Chiang *et al*., 2001; Bhushan and Romagnoli, 2008).

One common way to use neural network for fault diagnosis is to assign the input neurons to process variables and the output neurons to fault indicators or classes. During the calibration phase the network uses the error in the output values to update the weights (Venkatasubramanian *et al*., 2003c). This is known as *supervised* learning. Neural networks can represent complex nonlinear relationships and are good at classifying phenomena into preselected categories used in

the training process. However, their reasoning ability is limited and that motivates its combination with expert systems or fuzzy logic to enhance its performance.

Some of the main limitations of neural networks may be summarized as follows: (1) Long training times which might make the neural network impractical. (2) The requirement of large amount of training data to guarantee high fidelity during the testing phase. (3) Most training techniques are capable to tune the network but they do not guarantee that the network will work properly, for example, getting trapped in local minima during the training phase.

On the other hand, support vector machine arises recently as a strong competitive for the neural network based classification techniques (Chiang *et al*., 2004; Mahadevan and Shah, 2009). The idea of using SVM of separating two classes is to find support vectors i.e. subset of the training points, to define bounding planes in which the margin between both planes is maximized. In the case where the data are *not* linearly separable, the *kernel trick* is used (Bishop, 1995). That is, the original low dimensional data is mapped into higher dimensional feature space where the construction of a linear classifier becomes possible. This mapping is achieved using the so called kernel functions e.g. linear, polynomial, radial basis etc. The main advantage that favorite SVM over the ANN based classifier is the convexity of the optimization problem, since the problem of finding the support vectors can be posed as a quadratic program with linear constraints, hence local minima are avoided. The main drawback of SVM is that it cannot be used when the classes are larger than 10, since this increase the computation burden (Chiang *et al*., 2004).

2.4 The control-detection interaction problem

Most of the available fault detection and diagnosis systems, in particular empirical based techniques, are implemented in a supervisory level on top of the available control system. The survey papers (Gertler, 1988; Himmelblau, 1978; Willsky, 1976) provide early summary of work in this area of fault detection and diagnosis, while (Venkatasubramanian *et al*., 2003a; Venkatasubramanian *et al*., 2003b; Venkatasubramanian *et al*., 2003c; Qin, 2003) provide a more recent account. Despite the breadth of these survey papers, the issue of control and diagnostic interaction has not been addressed in any of them. In most of the literature, it is unclear, what is the effect the control system has on the diagnostic system. Among the few attempts, Jacobson and Nett (1991) proposed the four parameters controller setup as a generalization of the two degrees of freedom controllers. Tyler and Morari (1994) reformulated the four degrees of freedom controller into a general framework, for which tools from optimal and robust control were applied. They conclude that when uncertain plants are used in synthesizing a model based controller, a separated design does not lead to a reasonable diagnostic and the control and diagnostic must be synthesized simultaneously. Yoon and MacGregor (2001) showed how the effect of the feedback control may affect the steady state signatures used in diagnosis. Furthermore, they concluded that the presence of feedback or feed forward control complicates the identification and the diagnosis of faults, since they cause simple faults to be propagated through the system. Gertler and Cao (1994) study the affect of control on the detection problem from the modeling point of view. In particular, they studied the role of the control algorithms e.g. feedback and ratio controls, in providing the required excitation necessary to identify the PCA model used for the detection. They suggested changing the set point of the feedback control and the ratio coefficient for the ratio control for improved identification.

Recently, Benjamin *et al*, (2009) used a nonlinear state feedback algorithm to force distinct responses and symptoms to enhance faults distinguishability. Recently, Mahdi and Mehrdad, (2010) use the loop-shaping base techniques to study the tradeoffs between fault detection and control objectives.

More importantly, the interaction between standard feedback control algorithm e.g. PI controllers, and the standard statistical fault detection method has not been addressed, in particular, in the context of fault observability.

Accordingly, an additional novel contribution of this project is to explicitly address, the interaction between standard feedback control algorithms and the widely popular statistical monitoring charts in particular, in the context of fault observability.

2.5 Summary

On the basis of the above literature review, fault detection and diagnosis techniques were broadly categorized into three types: (1) analytical based techniques; (2) empirical based techniques; (3) and semi-empirical based techniques. The choice between these techniques highly depends on the problem at hand. Considering that dynamic models for large scale industrial processes are either not available or very difficult to obtain, data-driven empirical based techniques have been found more suitable for large scale industrial processes. Associated with the empirical based techniques is the issue of fault observability. In spite of its practical significance, the existing literature has not explicitly addressed this issue in the particular context of fault detection. The observability problem is often exacerbated by the fact that most of the proposed statistical monitoring approaches are used with variables that are involved in feedback control strategies; either as manipulated or controlled variables. Thus, from the detection and

diagnosis perspectives, the feedback controller has generally not been explicitly considered in the analysis of the diagnostic system's performance.

This project mostly concerns with the problem of fault observability. In particular, the main contributions of this project can be summarized into three main points presented correspondingly in three different chapters of the thesis.

(1) In chapter 3 the issue of fault observability has been demonstrated using a subset of the Tennessee Eastman Process (TEP) faults which have been found unobserved with all statistical monitoring approaches previously reported. From an industrial perspective, undetected faults have an impact on profit and its prompt detection is of paramount significance. Accordingly, a new statistical based technique that combine CUSUM and PCA based monitoring charts is proposed to detect and consequently diagnose these faults.

(2) Although faults can be detected by using the CUSUM based techniques proposed in Chapter 3, it was found that certain faults required a long period of time for detection corresponding to the time needed for the error to be integrated through the CUSUM operation to a level that surpasses a statistical threshold. To address this problem, in chapter 4 an approach is proposed for enhancing detection through proper tuning of feedback controllers that pair variables also used for detection. Thus, to enhance the detection characteristic of statistical monitoring charts operated under close loop processes, a new optimization based methodology that explicitly considers the interaction between feedback control and standard statistical monitoring charts is proposed. This optimization takes into account the potential economic costs related to variability, wear of actuators, design costs and costs related to the lack of observability of faults. The use of a frequency domain based approach allows different fault's frequencies and noise levels to be considered.

(3**)** Finally, chapter 5 addresses the applicability of the strategy proposed in chapter 4 to a realistic industrial scale simulator, namely, the Tennessee Eastman Process (TEP). Since the analysis in Chapter 4 was based on frequency response it was found suboptimal for highly nonlinear processes such as the TEP. To address this, a simulation based approach was proposed for solving the integrated optimization of control tuning and detection performance proposed in Chapter 4.

Chapter 3

Fault Detection, Identification and Diagnosis using CUSUM based PCA

(*Adopted form Bin Shams et al., 2010c, Chemical Engineering Science, Submitted***)**

3.1 Overview

In this chapter, a cumulative sum based statistical monitoring scheme is used to monitor a particular set of the Tennessee Eastman Process (TEP) faults that could not be properly detected or diagnosed with other fault detection and diagnosis methodologies previously reported. T^2 and Q statistics based on the cumulative sums of all available measurements were successful in observing these three faults. For the purpose of fault isolation, contribution plots were found to be inadequate when similar variable responses are associated with different faults. Fault historical data is then used in combination with the proposed CUSUM based PCA model to unambiguously characterize the different fault signatures. The proposed CUSUM based PCA was successful in detecting, identifying and diagnosing both individual as well as simultaneous occurrences of these faults.

3.2 Introduction

Fault observability and distinguishability are desirable properties for any detection and diagnosis system. Fault observability is relevant to the detection phase and it can be viewed as the ability of a detection system to detect, by using the available process measurements, abnormal process operation due to the occurrence of one or more faults. On the other hand,

distinguishability is related to the ability of a monitoring system to diagnose or isolate a particular fault by using the available measurements especially when the system exhibits similar responses in these measurements for different faults (Qin, 2003; Benjamin *et al*., 2008). Thus, observability refers to the ability of detecting abnormal operation whereas distinguishability refers to the ability to identify the particular fault or faults causing abnormal operation. Different methods have been proposed in the literature for fault detection and fault diagnosis (Venkatasubramanian *et al*., 2003a; Venkatasubramanian *et al*., 2003b; Venkatasubramanian *et al*., 2003c). These methods can be broadly categorized into three main classes: (1) Analytical methods which are solely based on first-principles models, e.g. observer based techniques; (2) Empirical methods e.g. univariate and multivariate statistical methods and (3) Semi-empirical methods, which combine empirical models with prior knowledge about the system under consideration for example through the use of expert systems or fuzzy rules (Chiang *et al*., 2001; Bhushan and Romagnoli, 2008). Each of these methods has its own advantages and disadvantages depending on the problem. A number of researchers suggest combining these methods to improve detection. For examples, (Chiang and Braatz, 2003; Lee *et al*., 2004) have observed that data driven analysis is enhanced by incorporating fundamental causal relationships among variables. Analytical methods require the use of first-principle models that may be often complex and difficult to obtain and calibrate, thus making them less attractive for large scale systems. Therefore, this work focuses on the use of empirical methods for detection and isolation. Since data in chemical processes generally exhibit high correlation in time and cross-correlation among variables, multivariate statistical methods such as latent variable methods have been proposed for fault detection and diagnosis since they can deal effectively with these problems e.g. Principal Components Analysis (MacGregor and Kourti, 1995).

Fault *isolation* or fault *diagnosis* is used interchangeably in the literature. The difference is that when the classification problem is resolved with the help of historical fault data, the corresponding procedure is referred to as a diagnosis; otherwise, it is referred to as isolation. The most widely used method for fault isolation is the contribution plot (Miller *et al*., 1993) which does not use historical fault data. It depicts the contribution of each process variable to the monitored statistics. That is, it identifies those variables that are most correlated with the fault in question. For that reason the term fault identification is sometimes used when fault isolation is accomplished using a contribution plot. Its effectiveness is limited to simple faults e.g. sensor and actuator faults (Yoon and MacGregor, 2001; Qin 2003). Researchers have used contribution plots combined with other methods to enhance the fault diagnosis procedure. For example, Dunia and Qin, 1998 proposed a fault identification index based on the fault reconstruction square prediction error (FRSPE). The smallest FRSPE is obtained for the reconstructed fault in question. Raich and Cinar, 1997, proposed distance and angle metrics to diagnose process disturbances. Yoon and MacGregor, 2001 proposed an angle based metric called the joint angle plot. Ku *et al*, (1995) use different dynamic principal component analysis (DPCA) based models to characterize each fault. All these methods need historical fault data to precisely diagnose faults e.g. fault vector directions. It will be shown in this work that the contribution plot is not sufficient to accurately isolate faults in specific situations where the measured variables behave similarly during the occurrence of different faults.

This chapter proposes the application of Cumulative-Sum (CUSUM) based models in combination with PCA for the detection and diagnosis of faults in the Tennessee Eastman problem (TEP) (Downs and Vogel, 1993). Location CUSUM (LCS) and Scale CUSUM (SCS) in combination with PCA based models were used to detect three particular faults that have been

found unobservable by other algorithms previously applied to the TEP (Cheng *et al*., 2010; Chiang *et al.*, 2001; Chiang and Braatz, 2001; Ding *et al.*, 2009; Ku *et al.*, 1995; Zhang, 2009).

In a previous study, Bin Shams *et al*., (2010d, Appendix B), have shown that an algorithm that combines CUSUM transformations with the Hotelling's T^2 statistics is able to detect abnormal operation following the occurrences of faults. However, the previous study had two key limitations. First, the previously proposed methodology required *a priori* selection of a subset of measurements that were highly correlated to each particular fault. These variables were selected by using process knowledge. Second, the previous study does not consider the diagnosis problem associated with detected faults.

The current study expands the previous work by proposing a detection algorithm that uses all the measurements available at the plant thus bypassing the need to select a priori a set of particular measurements for detection. Then, an algorithm that based on the combination of PCA and CUSUM for detection, identification and diagnosis is presented.

This chapter is organized as follows: definitions and an overview of the faults considered in the Tennessee Eastman Process (TEP) are given in section 3.3. Description of the proposed CUSUM based techniques for the detection and diagnosis are given in section 3.4. The results of applying the proposed CUSUM based strategy are discussed in section 3.5. Conclusions are given in section 3.6.

3.3 Definition and Methods

3.3.1 Out-of-control Average Run Length ($ARL_{o.c}$)

Most monitoring data driven techniques are based on the statistical hypothesis-testing principle. Two types of errors occur when performing hypothesis testing namely type I and type

II errors. A type I error occurs when a control chart indicates a fault in the absence of it, whereas a type II error occurs when a control chart fails to identify the occurrence of a fault (Montgomery, 1997). Observability of a fault is referred to as the ability to detect the fault from the chosen set of measurements. As proposed in a previous study, the out-of-control Average Run Length ($ARL_{o.c}$) is used as a statistical measure to gauge observability (Bin Shams *et al*, 2010a). The subscript (*o.c*) stands for *out of control*. The $ARL_{o.c}$ is defined as the average number of points that must be sampled or plotted before the chart signals the occurrence of a fault and it is a function of the probability of a type II error (β) occurring, that is

$$ARL_{o.c} = f(\beta) \tag{3.1}$$

For example, if in response to a certain fault, the estimated $ARL_{o.c}$= 1, the fault would be detected, on the average, after the first sample following the onset of the fault. On the other hand, an $ARL_{o.c}$= *infinity* or a very large number implies that the fault is unobservable or it takes a long time to observe it. The value of the $ARL_{o.c}$ depends on the type of chart that is used for monitoring. Several analytical expressions are available for simple statistical charts (Montgomery, 1997). For other type of charts, different approaches to estimate the $ARL_{o.c}$ based on the Markov chain approach have been proposed e.g. (Brook and Evans, 1972) but in practice, the $ARL_{o.c}$ is usually estimated from simulations conducted with random realizations of the disturbances (Woodall and Ncube, 1985). The latter approach is adopted in the current study.

3.3.2 The Cumulative sum (CUSUM) based control charts

A key disadvantage of Shewhart like control charts often used for detection is that they only use current time-interval information while not accounting for time history. Hence, those charts are relatively insensitive to small shifts in the process variables especially for small signal

to noise ratio. These shortcomings motivate the use of other alternatives such as the univariate or the multivariate version of the CUSUM based charts. Four types of statistics are combined in the current study for either fault detection or fault isolation. Specifically, location cumulative sum (LCS), scale cumulative sum (SCS) and the PCA based statistical measures, T^2 and Q. The current study proposes the use of a combined version of these statistics as described in the following section. The LCS and SCS algorithms are examples of univariate statistics while the T^2 and the Q are multivariate statistics. Both the LCS and SCS are performed using the following two statistics, corresponding to a two-sided hypothesis test (Hawkins and Olwell, 1998):

$$C_i^+ = \max[0, C_{i-1}^+ + x_i - (\mu_{i.c} + k)] \tag{3.2}$$

$$C_i^- = \max[0, C_{i-1}^- + (\mu_{i.c} - k) - x_i] \tag{3.3}$$

$$C_0^+ = C_0^- = 0$$

where k , $\mu_{i.c}$, C_i^+ and C_i^- are the slack variable, the *in control* mean , and the upper and the lower CUSUM statistics, respectively. The role of the slack variables is to introduce robustness with respect to noise. At every new sample, the statistics in equations (3.2) and (3.3), account for the accumulated sums of small deviations. These summations are corrected using the slack variable and compared to zero using the (max) operation. When either one of the two statistical measures in equations (3.2) and (3.3) exceeds a threshold H, the process is considered to be out of control. Following their respective definitions, the LCS is especially effective for detecting changes in the average whereas the SCS is suitable for detecting changes in variability. Guidelines for the selection of k and H have been reported (Hawkins and Olwell, 1998; Montgomery, 1997). Typically k is selected to be half of the expected shift in either μ or σ. H is determined so that a

prespecified $ARL_{o.c}$ is achieved. It should be noticed that when using equations (3.2) and (3.3), the LCS uses the original raw data x_i, whereas the SCS uses the following standardized quantity

$$x_i = \frac{\sqrt{|x_i|} - 0.822}{0.349} \tag{3.4}$$

A derivation of the quantities in equation (3.4) is given in Appendix A. Although LCS and SCS can be applied to individual measurements, there are many situations in which a single representative statistic for more than one variable is necessary. This is especially important when it is desired to present the operators with information in a compact form to simplify their monitoring activities during operation. In that case multivariate statistical measures can be used based on the univariate CUSUM's as explained in the following section.

3.3.3 Statistical monitoring with Principal Component Analysis (PCA)

For a process with n measurement variables, one alternative is to use n univariate control charts to monitor the process. As mentioned above, to simplify the presentation of information, a second alternative consists of using a principal component analysis (PCA) model to produce T^2 and Q charts for monitoring the n variables simultaneously. PCA involves the computation of loadings and scores using the covariance matrix of data $X \in R^{m \times n}$; where n is the number of variables and m is the total number of samples. If the original variables are correlated, it is possible to summarize most of the variability present in the n variables space in terms of a lower p dimensional subspace ($p<<n$) where p represents the number of the principal components that are retained to explain the majority of the variability in the data. If only two principal components are found to be significant, two dimensional score plots are used (i.e. T_1 versus T_2).

For more than two principal components, Hotelling T^2 and Q statistics are usually used to monitor the process. The T^2 statistic, based on the first p PCs, is defined as

$$T^2 = \sum_{i=1}^{p} \frac{t_i^2}{\lambda_i} \tag{3.5}$$

where λ_i is the i^{th} eigenvalue of the covariance matrix of the original data matrix. Confidence limits for T^2 at confidence level $(1-\alpha)$ are related to the F-distribution as follows:

$$T_{m,p}^2 = \frac{(m-1)p}{m-p} F_{p,m-p} \tag{3.6}$$

where $F_{p,m-p}$ is the upper $100\alpha\%$ critical point of the F-distribution with p and $(m-p)$ degrees of freedom. Monitoring the process variables by the T^2 values based on the first p principal components is often not sufficient since it only helps to detect whether or not the variation is within the plane defined by the first p principal components which generally capture steady state correlations corresponding to normal operation of the process. If a new event, not present in the calibration data used to identify the reference model occurs, then additional principal components may become significant and the new observation vector x_i will move off the calibrated plane. Such new events can be detected by computing the squared prediction error or Q statistic. Let $x_i \in R^n$ denote the i^{th} multivariate observation vector whose corresponding score is $t_i = x_i P$. The prediction from the PCA model for x_i is given by $\hat{x}_i = t_i P^T = x_i P P^T$. Then, the p dimensional error vector is given by $e_i = x_i - \hat{x}_i$ and the corresponding Q is defined as follows

$$Q = e_i . e_i^T \tag{3.7}$$

Accordingly, Q can be viewed as a measure of plant-model mismatch. The confidence limits for Q are given by Jackson (1991). This test suggests the existence of an abnormal condition when $Q > Q_\alpha$, where Q_α is defined as follows

$$Q_\alpha = \Theta_1 \left(1 + \frac{c_\alpha . h_0 . \sqrt{2.\Theta_2}}{\Theta_1} + \frac{\Theta_2 . h_0 (h_0 - 1)}{\Theta_1^2} \right)^{\frac{1}{h_0}} \tag{3.8}$$

$$\Theta_i = \sum_{j=p+1}^{n} \lambda_j^i; \quad for\, i = 1,2,3 \tag{3.9}$$

$$h_0 = 1 - \frac{2.\Theta_1.\Theta_3}{3.\Theta_2^2} \tag{3.10}$$

c_α are the confidence limits for the $(1-\alpha)$ percentile in a standard normal distribution. These confidence limits are calculated based on the assumptions that that the measurements are time independent and multivariate normally distributed.

Most of the statistical monitoring techniques available use static PCA as its basic building block (Bakshi, 1998; Wang and Romagnoli, 2005; Kramer, 1993).

3.3.4 Tennessee Eastman Process (TEP): previous detection and diagnosis approaches

The Tennessee Eastman process has been widely used as a benchmark problem to compare various monitoring solutions (Cheng *et al*., 2010; Chiang *et al*., 2001; Ding *et al*., 2009; Ku *et al*., 1995; Lee *et al*., 2004; Zhang and Zhang, 2010). The process is open loop unstable due to the process's exothermic reaction. Beside the reactor, the process has four main unit operations, as shown in Fig. 3.1: condenser, compressor, separator and stripper. The process produces two liquid products (G and H) and one by-product (F) from four gaseous reactants (A, C, D, E) and an inert (B). The original open loop FORTRAN code was provided by Downs and Vogel, 1993. The simulations of the plant were done with the second decentralized control structure proposed in Lyman and Georgakis (1995). The second control structure is characterized by selecting the variable that will manipulate the production rate and accordingly arranges the inventory control

loops. It consists of 19 SISO loops with PID based controllers. Proportional controllers are used for the separators and stripper levels. Series cascaded arrangements are used for the reactor level or composition control. The slave controllers use P action, while the master uses PI action. The rest of controllers are PI controllers. Tyreus and Luyben settings have been used to tune all the controllers.

Different monitoring techniques have been tested and reported for the TEP. These techniques have shown different capabilities in detecting the majority of the 20 faults generally assumed for the process. However, all of these previously reported techniques have consistently failed in detecting the three particular faults described in Table 3.1.

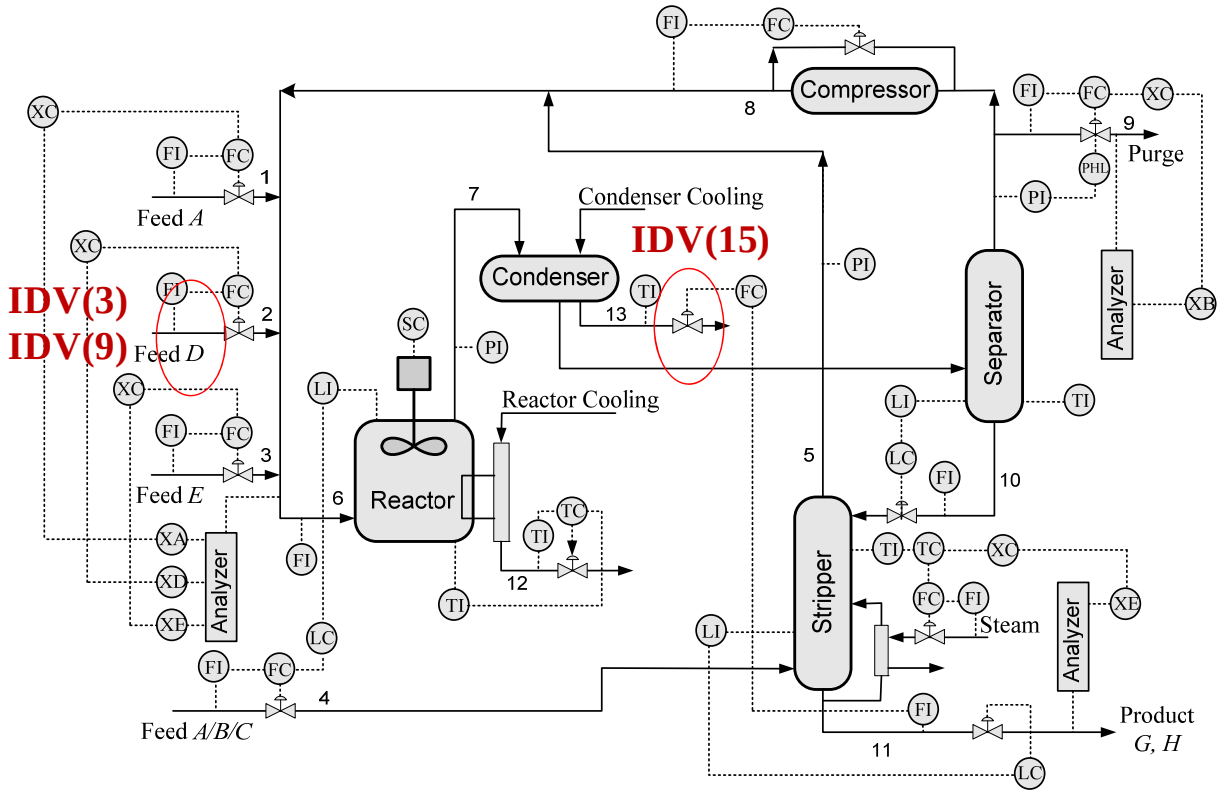


Fig.3.1 Tennessee Eastman Process (TEP) with the second control structure described in (Lyman and Georgakis, 1995). The circles indicates the location of the three faults described in Table.1

Table.3.1: The unobservable faults of the TEP process

Fault No.	Description	Characteristic
IDV(3)	D feed temperature	Step Change
IDV(9)	D feed temperature	Random variation
IDV(15)	Condenser cooling water valve	Valve stiction

Almost, all of the methods previously applied to the TEP were of a multivariate nature. Figure 3.2 shows the results of the application of static PCA for the TEP using the T^2 and Q statistics. The bounds of normal operation corresponding to a 95% and a 99% confidence levels corresponding to the no fault situation, are shown as dotted lines in Fig.3.2 and are calculated by equation (3.6) and (3.8).

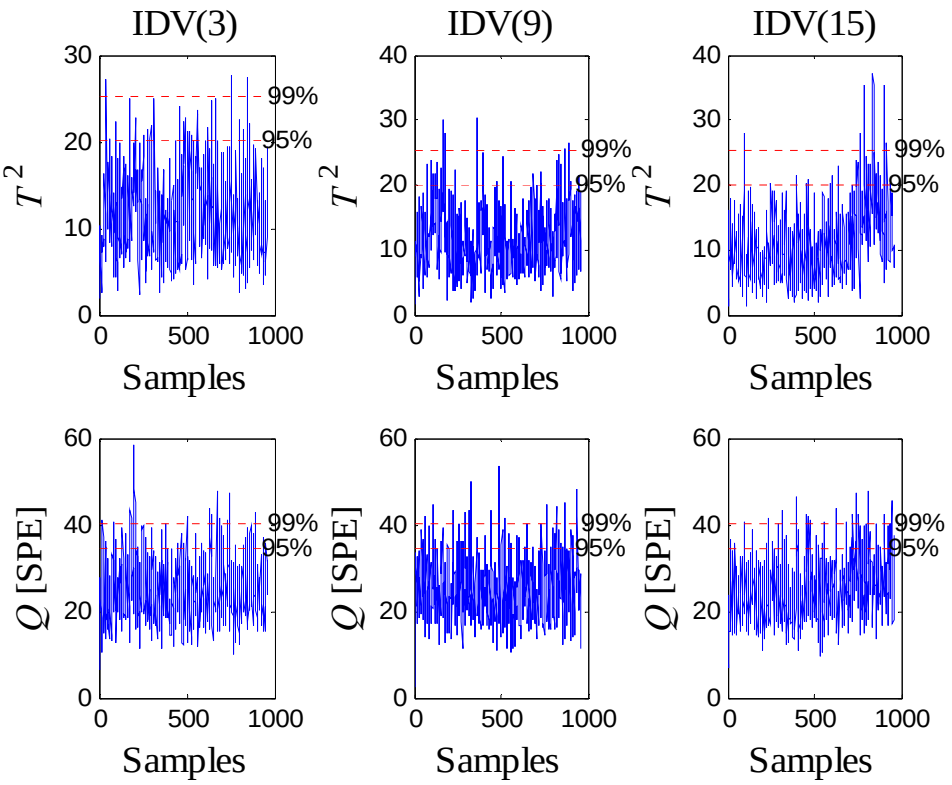


Fig.3.2 Monitoring the TEP faults using T^2 statistic based on the PCA. Top: IDV (3), left: IDV (9), right: IDV (15).

The meaning of these bounds is that if either the T^2 or Q is above these bounds after the occurrence of the fault, then the fault has been successfully detected. For the plots in Fig. 3.2 the corresponding faults were introduced at time=160 samples. However, as shown in Fig. 3.2, both T^2 and Q statistics fail to surpass the thresholds after the onset of either one of the 3 faults in Table 3.1, i.e. IDV(3), IDV(9) and IDV(15). The incidental threshold violations shown in Fig.3.2 are attributed to the Type I error associated with the underlying hypothesis testing. Hence, it can be concluded that these faults cannot be detected by static PCA. It should be noticed that when a PCA are used, p is replaced with a in equation (3.6), where a is the number of principal components retained in the PCA model.

Dynamic principal component analysis (DPCA) proposed by Ku *el al*; (1995) was used for the detection and diagnosis of the TEP faults. The DPCA has the advantage of taking into account information along several time intervals in contrast to the conventional static PCA which is based solely on data collected at the current time. Accordingly, DPCA is more suitable for dynamic systems. However, the latter was also unsuccessful in observing these three faults. The results of the application of DPCA are not shown here for brevity.

The inability of previous techniques to detect the three faults given in Table 3.1 motivates the use of cumulative sum measures. The Multivariate Cumulative Sum (MCUSUM) algorithm using all of the available TEP measurements was also initially tested (MacGregor and Kourti, 1995; Woodall and Ncube, 1985) but this technique was also unable to detect these three faults. The resulting lack of observability when using specific techniques is attributed to the statistically insignificant changes in the process mean or the process variance or combination of both exhibited by the system when these faults occur.

3.4 CUSUM based Multivariate Statistics

3.4.1 Fault Detection using CUSUM based PCA

In a previous study, three faults of the TEP process, namely IDV(3): A step change in the D feed stream's temperature, IDV(9): A random variation in the D feed stream's temperature and IDV(15): A stiction in the condenser cooling water valve; have been observed using univariate CUSUM based Hotelling T^2 (Bin Shams *et al*, 2010d; Appendix B). These faults were especially chosen since it was shown in other studies that they could not be detected by other fault detection techniques applied to the TEP problem. Figure.3.2 illustrates the difficulty associated with detecting and diagnosing these three faults using the space of three measurements that are greatly affected by the presence of these three faults, namely, XMV(10): reactor cooling water flow, XMV(11): condenser cooling water flow, and XMEAS(21): reactor cooling water outlet temperature.

As stated in the introduction, the previous study was based on a subset of measurements that were identified, using process knowledge, to have significant correlation with respect to each one of the faults. It can be seen from Fig.3.3 that there is a significant overlapping between the responses of these three variables both during the presence and absence of the fault, thus making it very difficult to both detect and isolate these three faults.

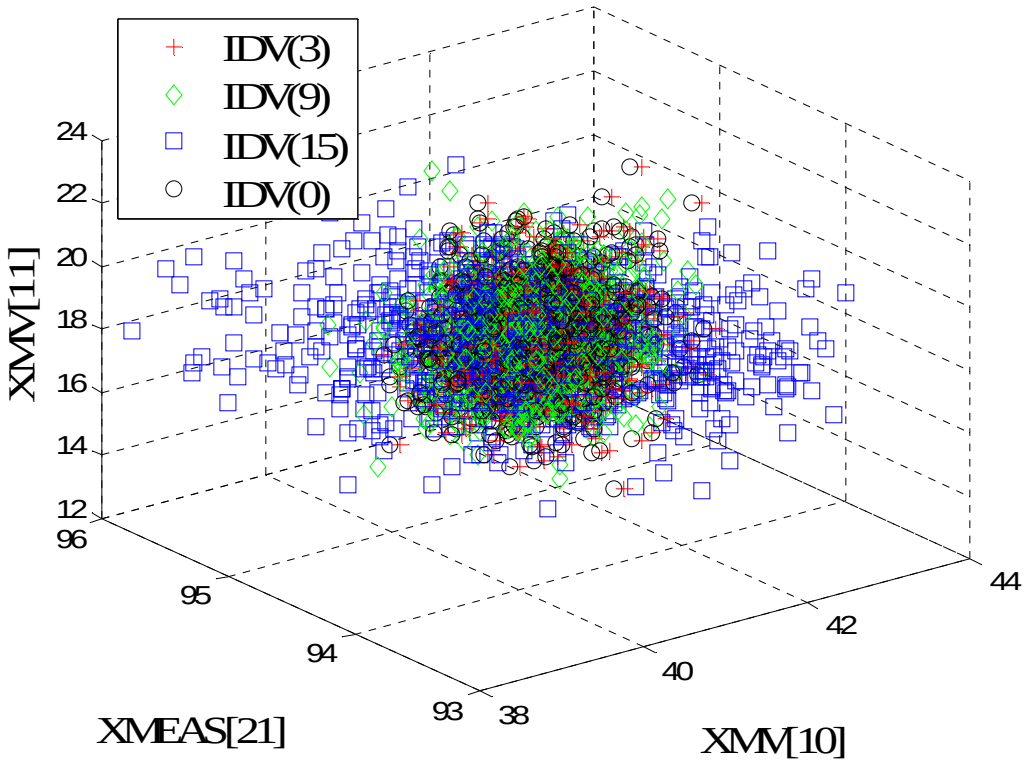


Fig.3.3 The three variables which are greatly affected by the presence of the three faults of the TEP process. The strong overlap between the three faulty states i.e. IDV (3), IDV (9), IDV (15) and the normal condition IDV (0) makes the detection and diagnosis of these three faults a challenging task.

The detection algorithm in the previous work (Bin Shams *et al*., 2010d, Appendix B), was based on particular set of variables that had to be chosen a priori. Since there is no systematic way to identify these variables and since there may be an information loss when using a subset of the available measurement, all the available measurements are used in the current work for detection and diagnosis. As schematically described in Fig.3.4, two matrices are initially formed.

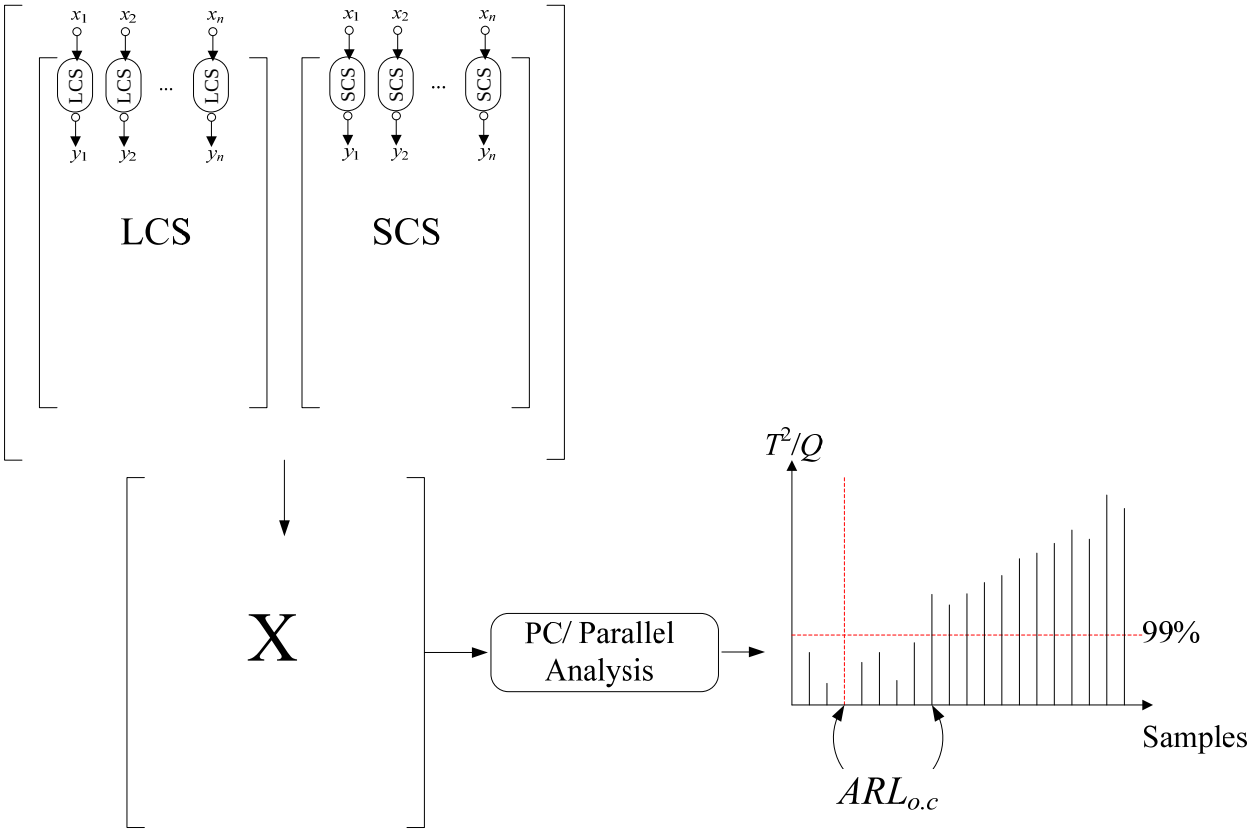


Fig.3.4 The proposed CUSUM based statistics. The LCS and the SCS are performed on each sample vector. PCA is performed on the augmented matrix. The score and the residual spaces are monitored using the T^2 and Q statistics.

One of them contains in its columns the LCS CUSUM's and the other one contains the SCS CUSUM's of all the measurements. For example for the TEP problem, 52 measured variables are available. Thus each one of the two aforementioned matrices contains 52 columns and a number of rows corresponding to the number of time intervals at which the measurements were collected. Then, a new matrix is formed by appending together the SCS and LCS matrices side by side resulting in a matrix with 104 columns. To account for collinearity across the variables, the principal components are calculated. Principal components that contribute significantly to the total variability of the data are determined using *parallel analysis* (Chiang *et al*., 2001). The T^2 is then used to monitor the space spanned by these components. In addition,

the Q statistics is used to monitor the space spanned by those components not included in the calibrated model. Thus, any variation within the space defined by the PCA model is monitored using the T^2. On the other hand, lack of correlations as represented by the PCA model is detected using the Q statistic. Either or both statistics are used to detect the presence of any abnormality whenever their instantaneous values exceed the corresponding critical values at a given significance level.

3.4.2 Fault Isolation/Diagnosis using CUSUM based PCA

The contribution plot has been the main tool used for fault isolation (Miller *et al*., 1993; Qin, 2003). Two contribution plots are usually used to identify those variables affected by the presence of faults, namely, the contribution plots related to either T^2 or Q statistics. For $X \in R^{m \times n}$, the total contribution of variable j to the Q statistic at each sampling instant i is given by:

$$Cont_{ij} = e_{ij}^2 \tag{3.11}$$

On the other hand, the contribution of variable j to the T^2 statistic for $a << n$ principal components at each sampling instance i is given by:

$$Cont_{ij} = \sum_{k=1}^{a} \left(\frac{p_{jk}^2}{\lambda_k} x_{ij}^2 + \frac{2p_{jk}}{\lambda_k} \left(\sum_{\substack{r=1 \\ r \neq j}}^{n} p_{rk} x_{ir} \right) x_{ij} + \frac{1}{\lambda_k} \left(\sum_{\substack{r=1 \\ r \neq j}}^{n} p_{rk} x_{ir} \right)^2 \right) \tag{3.12}$$

where p_{jk} and λ_k are the jk element of the loading matrix and the k-th eigenvalue, respectively. To obtain the total contribution of the variables j within a specific time period, the corresponding $Cont_{ij}$ is summed over the required time window. As shown in equation (3.12) the contribution of the j variable to the T^2 statistic consists of three terms. The first term includes solely variable j and the second term contains a cross-product between variable j and the rest of variables. The last term does not contain x_{ij}, and therefore does not affect the conclusion drawn about variable

contribution. In practice, it is omitted. Although very simple to build, a fundamental drawback associated with the contribution plot is the lack of precision in isolating the correct fault as illustrated later in the case study. The main reasons is that contribution plots are based on a non-causal statistical correlation model that does not take into consideration the cause and effect correlations between the process variables. An alternative to the contribution plots proposed in the current study is the use of a set of models that are based on the CUSUM transformation combined with PCA based statistics as follows. Assume a model PCA*j* is designed to detect a particular fault f_j. This model is trained using data generated when the fault f_j occurs. The faulty data can be obtained from a historical data base or using a designed experiments. This data characterize the steady state correlations structure between the process measurements when fault f_j occurs. Then, if the critical limits determined for the T^2 and Q statistics are exceeded this can be interpreted as a situation where fault f is *not* active. Figure 3.5 depicts the proposed CUSUM based PCA strategy. The misdetection rate (MR%) is defined as the percentage of samples below the control limit after the fault occurrence, i.e. $n_b/n_t \times 100\%$; where n_b and n_t are the number of points below the threshold and the total number of samples following the fault onset, respectively. Accordingly, when low misdetection rate is associated with T^2 or Q statistics, using the CUSUM-PCA model corresponding to fault f_j, this implies that the process is either operated under normal conditions or alternatively it experiences a fault different than that particular fault f_j. Therefore, higher misdetection rate will indicate that the acquired measurements are in accordance with the model of fault f_j, that is, fault f_j occurs. Assuming a total of n_f faults, then a total of n_f CUSUM based PCA models are required to isolate each one of these faults.

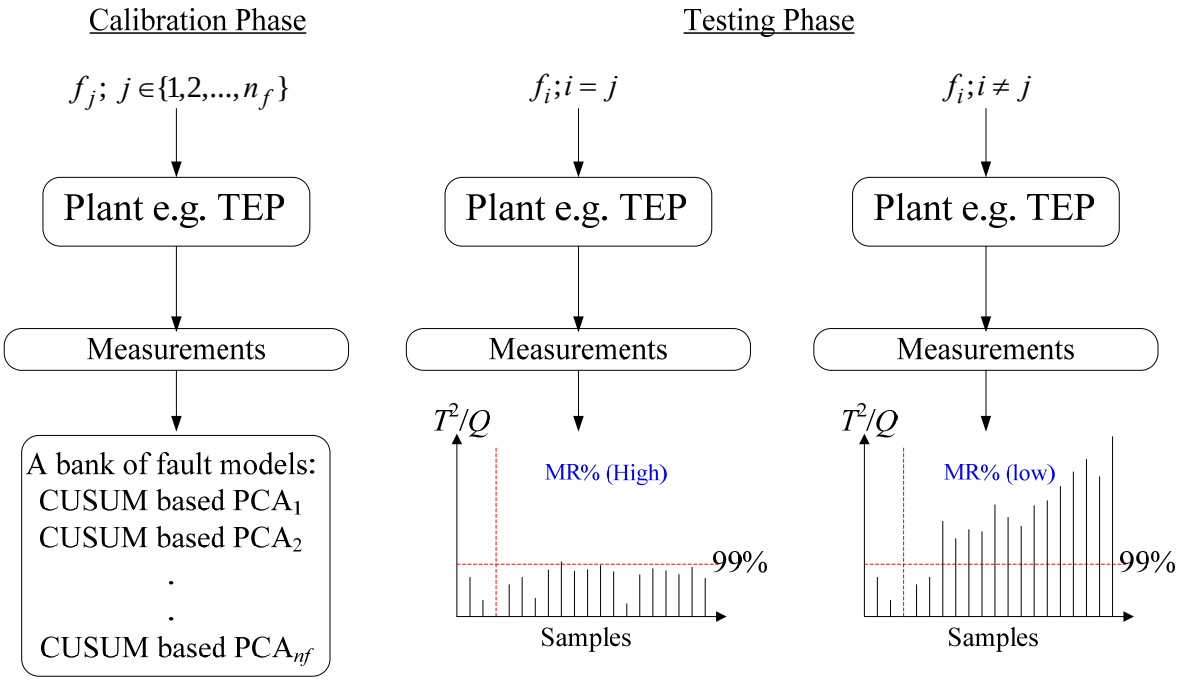


Fig.3.5 The proposed CUSUM based diagnosis strategy.

3.5 Results and Discussions

The sampling frequency for the CUSUM based chart is (1/180) Hz (3 min. time intervals). In all the following simulations, the faults are introduced after 160 samples, that is, after 8 hours of a normal operation. Different noise realizations were tested and used to calculate the average run lengths ($ARL_{o.c}$).

For comparison reason, figure 3.6 depicts the normal condition. Figure 3.7, Figure 3.8, Figure 3.9 and Figure 3.10 depict the successful detection of IDV (3), IDV (9) IDV (15) and the simultaneous occurrence of IDV(3) and IDV(15), respectively when the CUSUM of all the measurements are used as explained in Section 3.4.1.

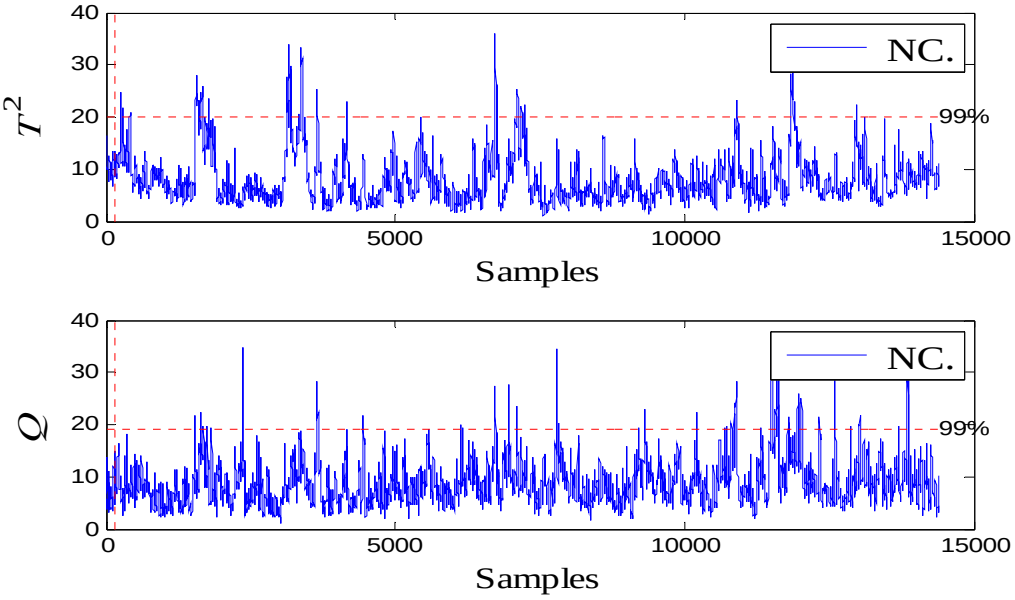


Fig.3.6 The T^2 and Q statistics based on the CUSUM based PCA for Normal Condition; horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

The vertical dashed line represents the onset of the fault whereas the horizontal dashed line represents the 99% confidence limit. Both T^2 as well as Q statistics are calculated, where the first one serves to identify a departure from the variables normal condition values, the second serves to indicate a departure from the steady state correlation. As can be seen from these figures, different faults affect the two monitored statistics differently. For example, it is clear in Fig.3.8 that the fault is better detected by the Q statistics since T^2 alone is not a sufficiently accurate indicator to detect this fault. This illustrates the need to use both T^2 and Q statistics to identify the presence of a fault with certainty. Table 3.2 shows the $ARL_{o.c}$ associated with the detection of each one of the 3 faults.

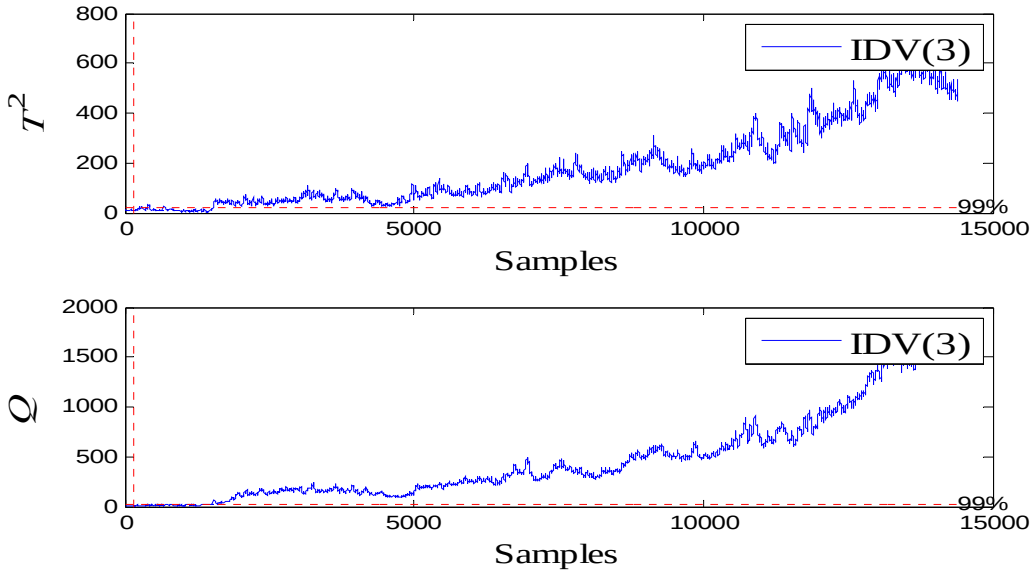


Fig.3.7 The T^2 and Q statistics based on the CUSUM based PCA for IDV (3); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

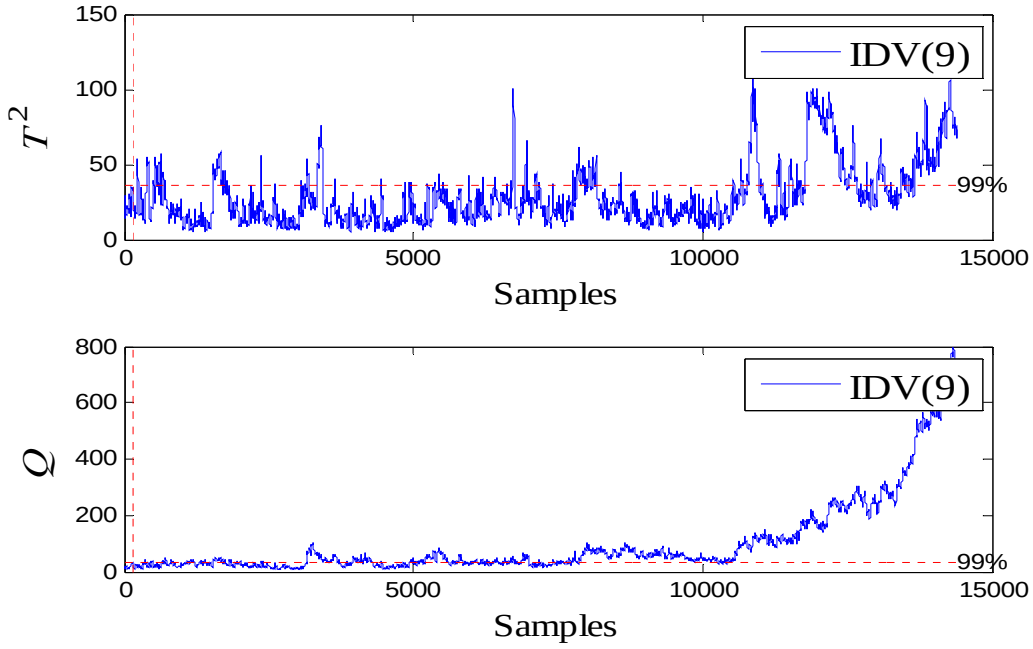


Fig.3.8 The T^2 and Q statistics based on the CUSUM based PCA for IDV (9); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

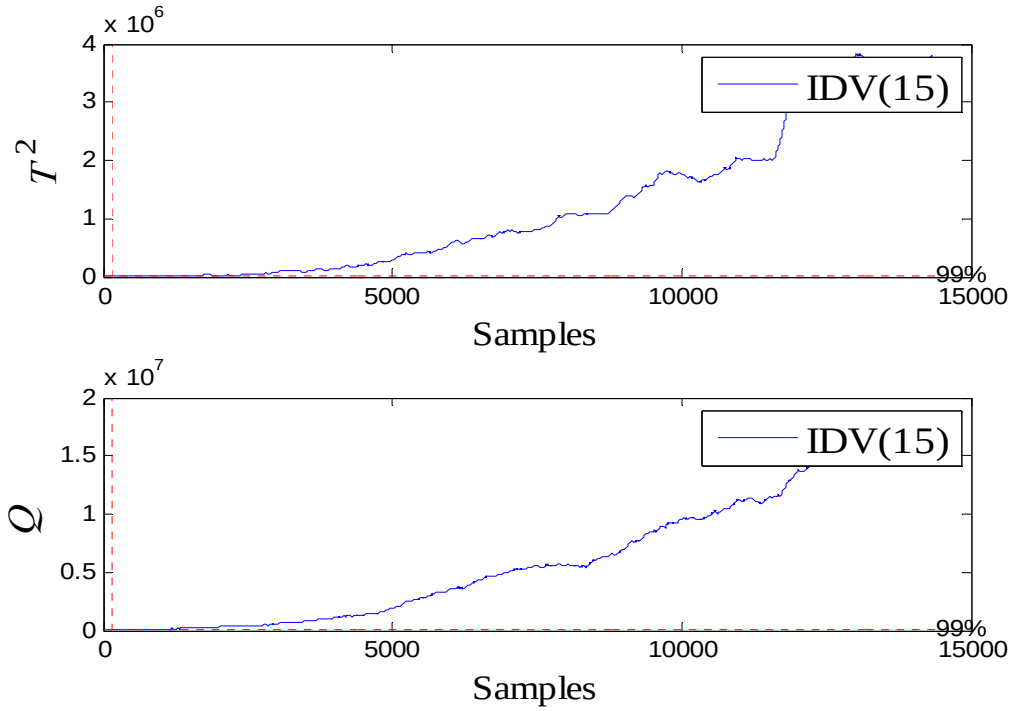


Fig.3.9 The T^2 and Q statistics based on the CUSUM based PCA for IDV (15); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

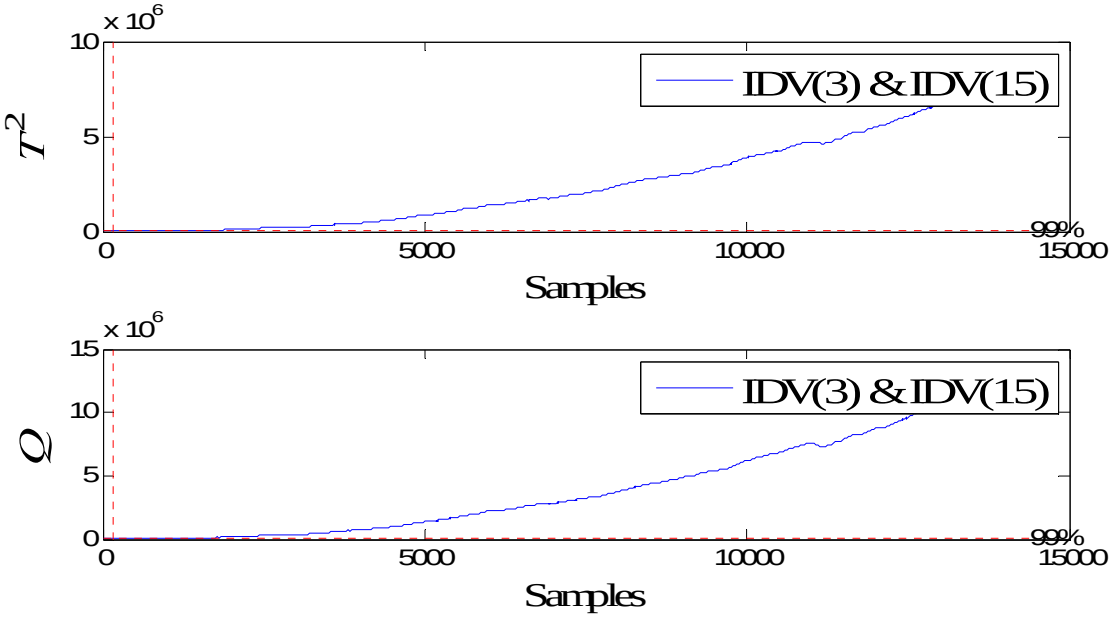


Fig.3.10 The T^2 and Q statistics based on the CUSUM based PCA for IDV (3) & IDV (15); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

Table 3.2: The estimated $ARL_{o.c}$ for the CUSUM based T^2 and Q statistics

Fault No.	Statistics	$^{*}ARL_{o.c}$ (hr)
IDV(3)	T^2 (Q)	467.6 (222.9)
IDV(9)	T^2 (Q)	143.8 (127.9)
IDV(15)	T^2 (Q)	0.6 (0.6)
IDV(3) & IDV(15)	T^2 (Q)	0.6 (0.6)

* All $ARL_{o.c}$ are calculated after the onset of the faults (i.e. after 8 hours)

As can be seen, there is a long delay associated with the detection of these faults, especially with IDV (3) and IDV (9). However, it can be argued that slow detection is preferable as compared to no detection at all.

Once any of the three faults is detected, it is desired to isolate the occurred fault, i.e. to identify those variable most correlated with occurred faults.

Figures 3.11, 3.12 and 3.13 depict the CUSUM based T^2 and Q contribution plots. Figure 3.11 shows the significant contribution of measurement 51: reactor cooling water flow, when IDV (3) occurs. Fig.3.12 shows that measurement 21: reactor cooling water outlet temperature, contributes significantly to the Q statistic when IDV (9) occurs. Fig.3.13 depicts the contribution of measurement 22 or the separator cooling water outlet temperature in the presence of IDV (15). Although the contribution plots emphasize those variables most related with the corresponding faults, there are situations where the contribution plots may be misleading. To demonstrate this situation, IDV (4): a step change in the reactor cooling water inlet temperature is used (Downs and Vogel, 1993). As can be seen from Fig.3.11 and Fig.3.14 the CUSUM based contribution plots are not actually helpful in isolating the root cause for their corresponding faults, i.e. IDV (3) and IDV (4).

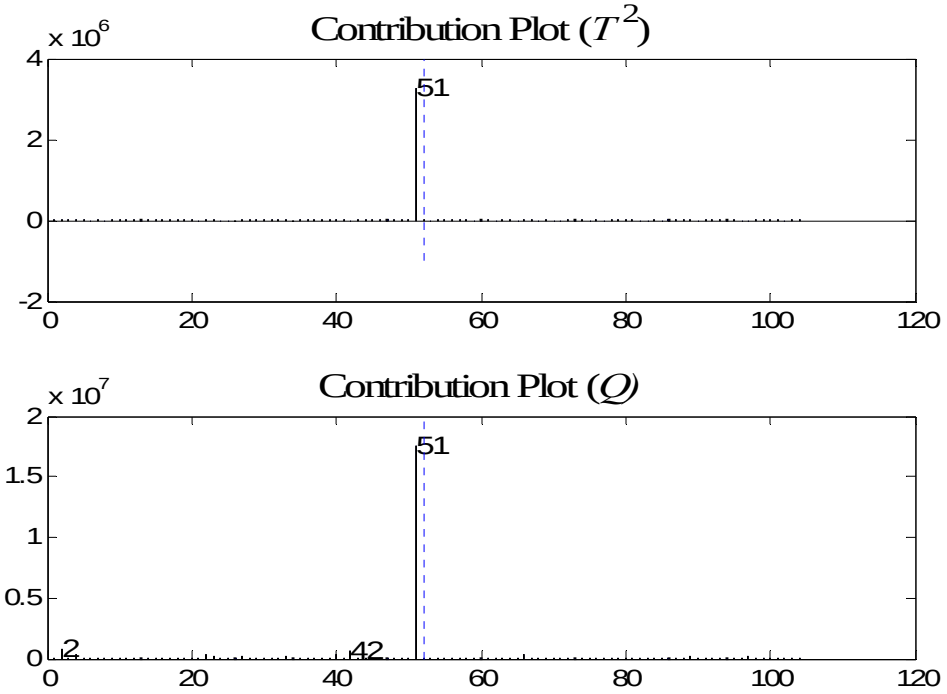


Fig.3.11 Contribution plot for IDV (3)

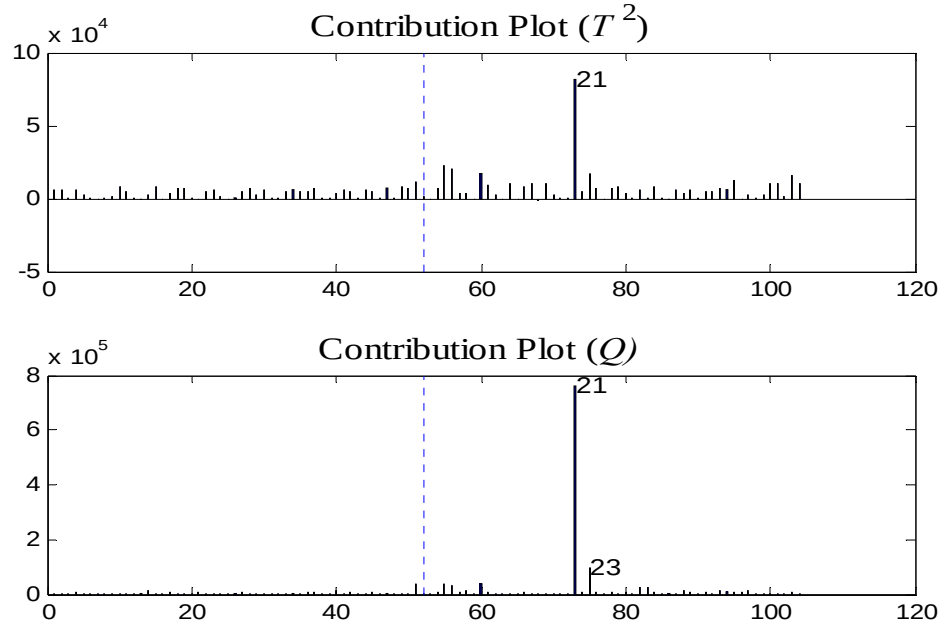


Fig.3.12 Contribution plot for IDV (9)

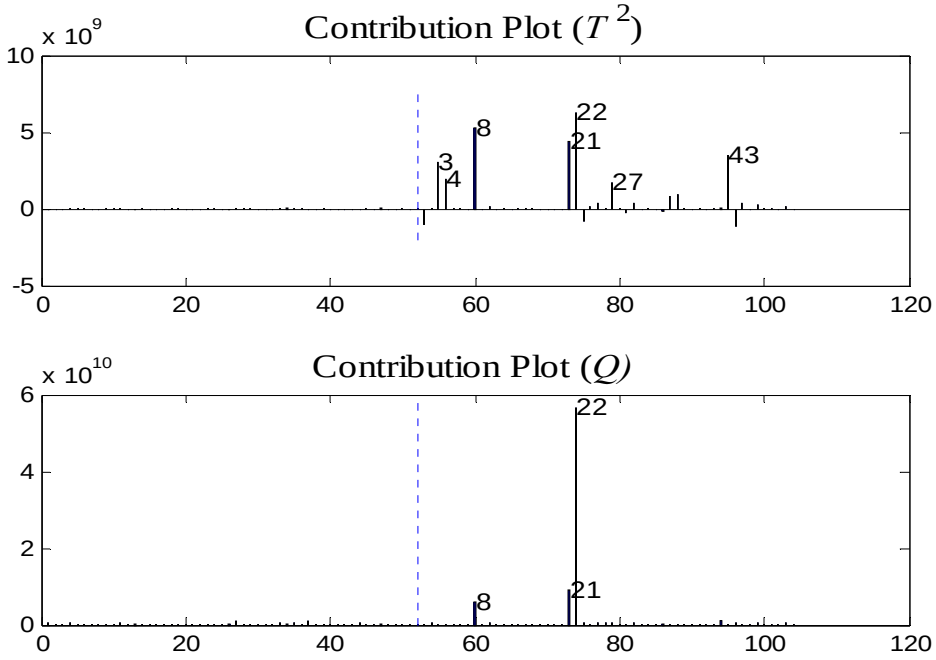


Fig.3.13 Contribution plot for IDV (15)

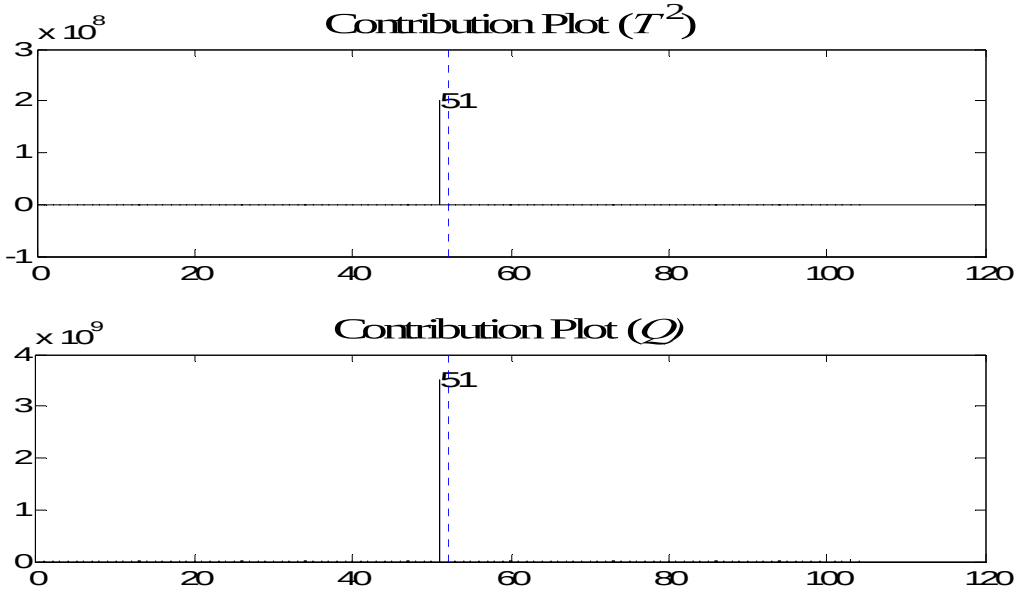


Fig.3.14 Contribution plot for IDV (4)

In fact a fault misclassification is inevitable; since both contribution plots, i.e. Fig.3.11 and Fig.3.14, choose the same variable i.e. measurements 51 (reactor cooling water flow) as a possible cause for completely two different faults. In addition, since contribution plots are based on non-causal correlation model, pinpointing the variables that are correlated with the occurred fault is all what contribution plots can provide. The determination of the type of the fault is still an ambiguous.

An enhanced fault diagnosis can be achieved by identifying PCA models based on the CUSUM based PCA as explained in section 3.4.2. Four PCA models are identified for the individual occurrence of faults IDV (3), IDV (9) and IDV (15) plus one corresponding to the scenario where IDV (3) and IDV (15) are occurring simultaneously.

Fig.3.15 depicts the diagnosis results of IDV (3). The model is calibrated using the faulty data when IDV (3) occur. This figure is composed of a total of 6 subplots. Each row of the subplots shows the T^2 and Q responses corresponding to the occurrence of each one of the 3 faults namely, IDV (3), IDV (9) and IDV (15), respectively. It can be seen that the first row show that the measurements are in accordance with the calibrated model, i.e. the responses are within the limits of the model indicating the occurrence of IDV (3). On the other hand in the second and the third row, the T^2 and Q critical limits of the models corresponding to IDV (9) and IDV (15), respectively are exceeded. From these plots, it can be concluded that IDV (3) is most likely the experienced fault, that is, a step change in the D feed temperature whereas the other two faults IDV(9) and IDV(15) are not active.

Comparing to the contribution plots in distinguishing between IDV (3) and IDV (4), a better diagnosis performance is obtained when the proposed PCA models are used as illustrated

in Fig.3.16. In the latter, the CUSUM based PCA model is calibrated using a data generated when IDV (3) is active.

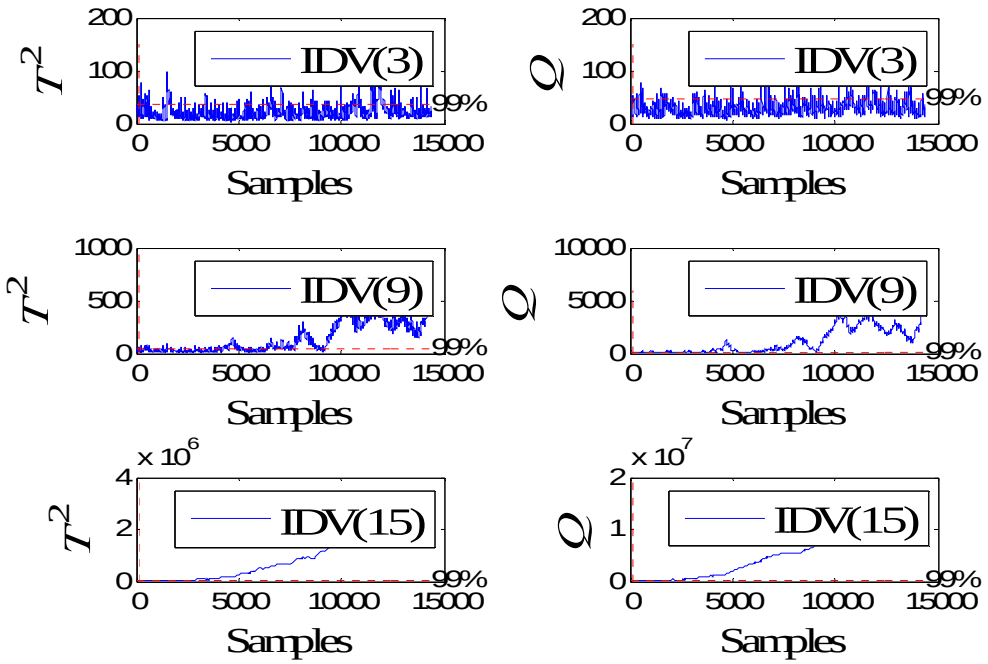


Fig.3.15 The diagnosis results for IDV (3), First row: IDV (3), Second row: IDV (9), Third row: IDV (15). The higher misdetection rate at the first row indicates the occurrence of IDV (3).

It can be seen from Figure 3.16 that misdetection rates of 94.66% and 96.67% are obtained for the T^2 and Q. That is, the corresponding thresholds are not exceeded for the CUSUM based PCA model trained with IDV (3) active, whereas these same thresholds are not exceeded 2.52% and 0.78% of the times for T^2 and Q, respectively when IDV (4) is active implying that most likely fault IDV(3) is the one active.

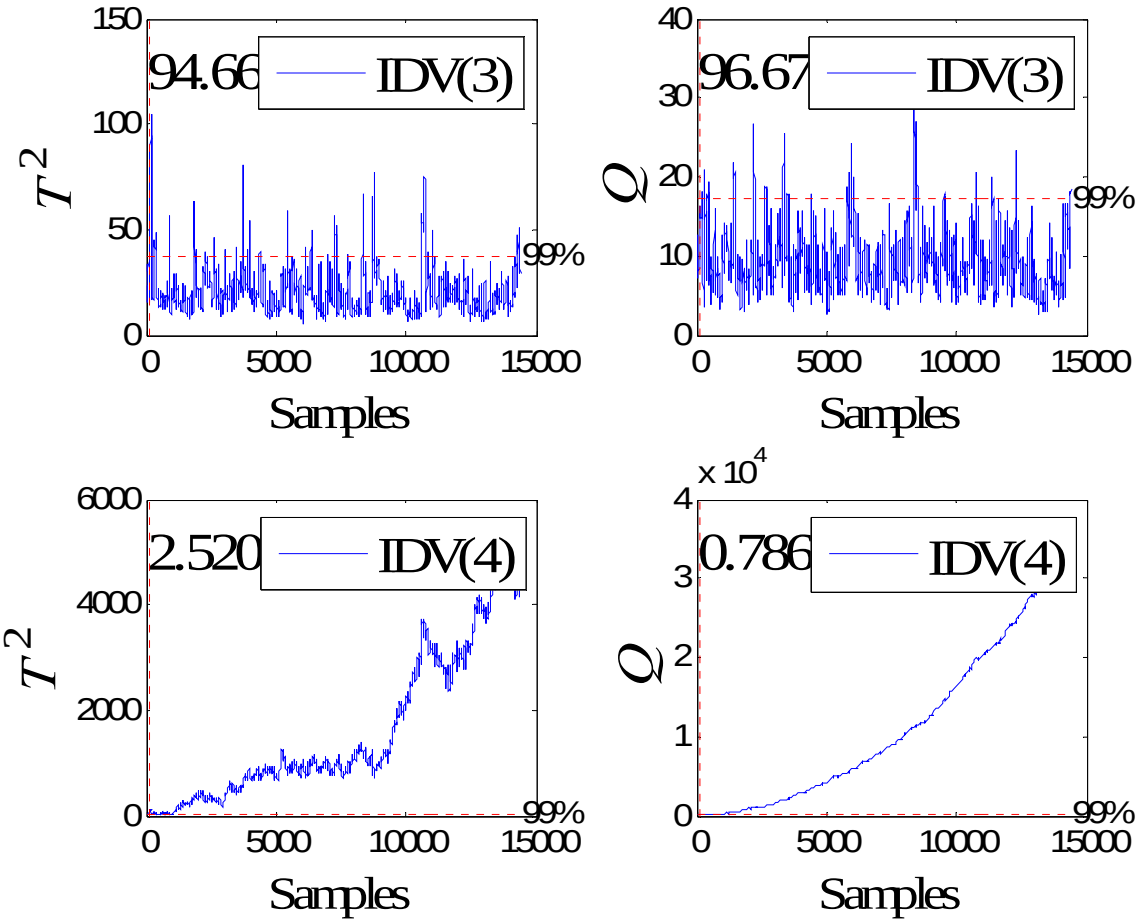


Fig.3.16 The fault historical data of IDV (3) clears the ambiguity associated with the contribution plots, Fig.3.11 and Fig.3.14. The first row: IDV (3). The second row: IDV (4).

Figure 3.17 shows data generated when IDV (15) occur. It can be seen in the third row of subplots that the critical limits of T^2 and Q are exceeded indicating that IDV (15) is the active fault i.e. a stiction in the condenser cooling water valve. Fig.3.17 illustrate the precise diagnosis of the simultaneous occurrence of IDV (3) and IDV (15) using the proposed approach. The necessity of considering both the T^2 and Q statistics in diagnosing these faults can be explained as follows. The first column of subplots in Fig.3.15 or Fig.3.16 would suggest that all of the faults are possible fault candidates. In that case the Q statistic is used to assess which fault is

active. The need for using the Q statistics is further reinforced by the fact that not all the variation caused by the fault can be observed within the score space captured by the CUSUM based PCA model.

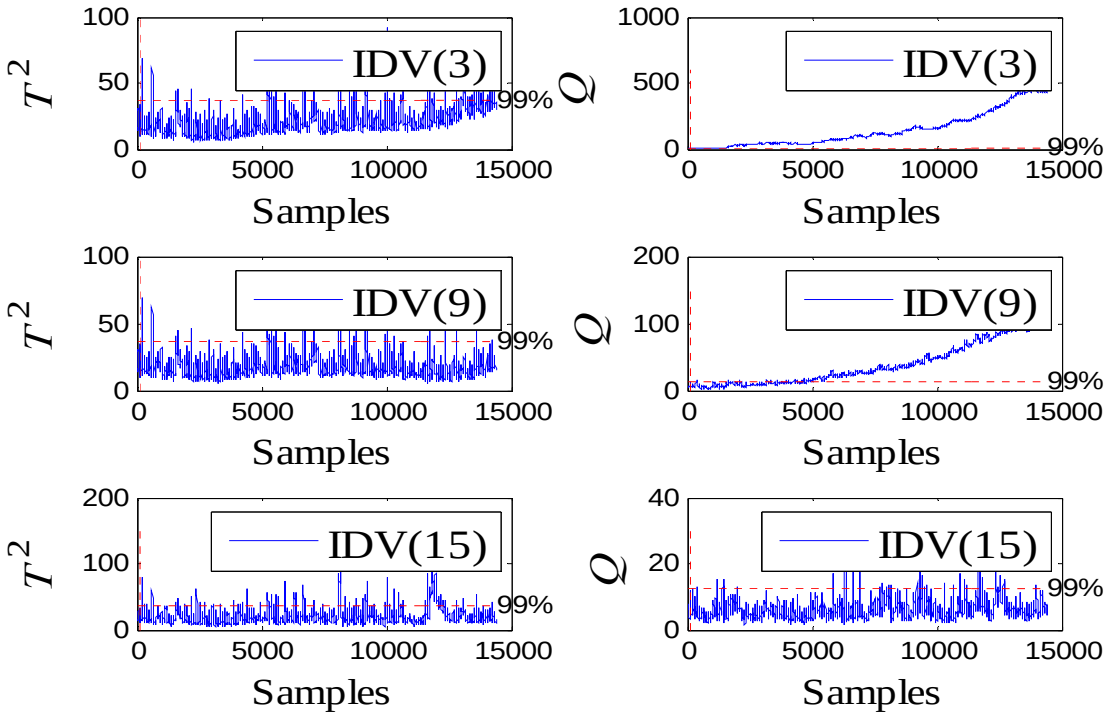


Fig.3.17 The diagnosis results for IDV (15), First row: IDV (3), Second row: IDV (9), Third row: IDV (15). The higher misdetection rate at the third row indicates the occurrence of IDV (15).

Table 3.3, Table 3.4 and Table 3.5 summarize the diagnosis results for faults IDV (3), IDV (15) and for the simultaneous occurrence of both faults, respectively. Finally, the results using the proposed CUSUM based PCA is compared with the fault diagnosis results obtained when static PCA is applied directly to the data without CUSUM transformations of the measurements.

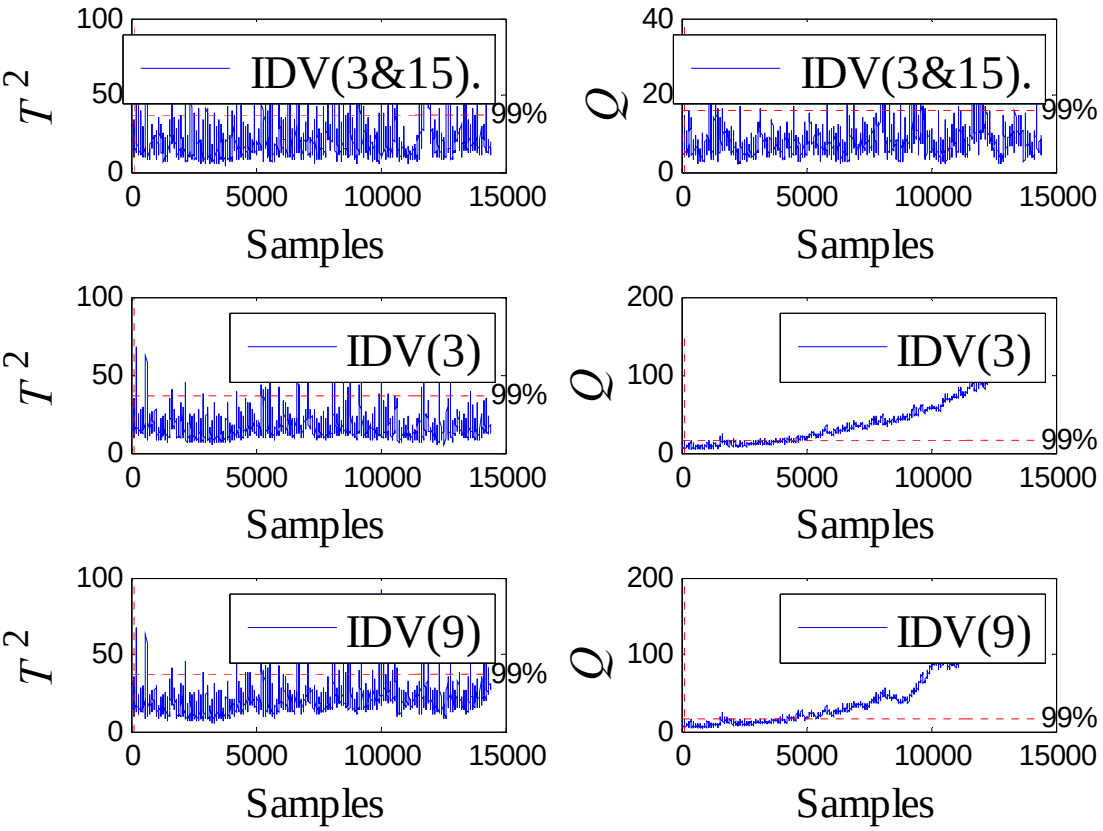


Fig.3.18 The diagnosis results for the simultaneous occurrence of IDV (3) & IDV (15); First row: IDV (3) & IDV (15), Second row: IDV (3), Third row: IDV (9). The higher misdetection rate at the third row indicates the simultaneous occurrence of IDV (3) & IDV (15).

Typically it would be desirable that the decision column would have only one entry with value equal to one in the row corresponding to the active fault indicating a clear diagnosis of this fault. An entry of one in the decision column was assigned when the misdetection rate of the corresponding fault, i.e. the response variables do not exceed the T^2 and Q thresholds is more than 90% of the total number of samples following the onset of fault. This value is found reasonable since a much lower value was typically obtained when discrepancies occur. Any additional entry with value equal to one in a fault other than IDV (3) thus implies

misclassification of the fault under consideration. It can see from Table 3.3, Table 3.4 and Table 3.5 that when static PCA models for each of the three faults is used to characterize each fault, the occurring fault is always misclassified as can be seen from the decision column i.e. all entries with ones. For example, in Table 3.3, when IDV (3) has occurred, the diagnosis scheme based solely on static PCA indicates the presence of all of the faults, which is, in fact, a wrong decision. The application of the proposed CUSUM based statistics (T^2 and Q) shows a much superior capability compared to PCA as indicated by the single entry in the decision column. Similar results are obtained for IDV (15) as well as for the simultaneous occurrence of both faults as can be seen in Table 3.5.

Table 3.3: Misdetection rates; Calibration model: IDV (3)

Fault Active	**Misdetection rate (PCA)**		**Decision**	**Misdetection rate (CUSUM-PCA)**		**Decision**
	T^2	Q		T^2	Q	
IDV(0)	99.11%	99.17%	1	19.49%	11.50%	0
IDV(3)	99.12%	99.20%	1	93.65%	96.33%	1
IDV(9)	98.89%	99.01%	1	24.92%	6.79%	0
IDV(15)	19.38%	55.01%	0	0.39%	0.08%	0
IDV(3 &15)	18.50%	55.38%	0	0.39%	0.08%	0

Table 3.4: Misdetection rates; Calibration model: IDV (15)

Fault Active	Misdetection rate (PCA)		Decision	Misdetection rate (CUSUM-PCA)		Decision
	T^2	Q		T^2	Q	
IDV(0)	99.94%	99.67%	1	86.00%	13.80%	0
IDV(3)	99.95%	99.68%	1	76.81%	12.82%	0
IDV(9)	99.90%	99.58%	1	85.95%	13.55%	0
IDV(15)	99.00%	99.08%	1	98.60%	99.65%	1
IDV(3 &15)	99.17%	99.07%	1	90.06%	20.86%	0

Table 3.5: Misdetection rates; Calibration model: IDV (3 & 15)

Fault Active	Misdetection rate (PCA)		Decision	Misdetection rate (CUSUM-PCA)		Decision
	T^2	Q		T^2	Q	
IDV(0)	99.90%	99.62%	1	89.81%	14.06%	0
IDV(3)	99.90%	99.60%	1	89.88%	13.51%	0
IDV(9)	99.89%	99.53%	1	89.74%	13.96%	0
IDV(15)	98.72%	98.96%	1	93.82%	32.18%	0
IDV(3 &15)	99.05%	99.10%	1	98.49%	98.25%	1

3.6 Conclusions

A new approach based on the combination of univariate CUSUMs and multivariate PCA statistics is proposed for the fault detection and diagnosis problem. The approach has been demonstrated using a subset of the Tennessee Eastman Process faults that have been typically found unobservable or undistinguishable when using other detection or diagnosis techniques. Following successful detection of the faults, the variables related to the faults are identified. In addition, due to the overlapping nature of the measured variables that are strongly correlated to these three faults, the contribution plots were found inadequate in some instances to precisely locate the root cause of the faults. Instead, the use of a family of PCA models trained with CUSUM transformations of all the available measurements collected during individual or simultaneous occurrence of the faults were found effective in correctly diagnosing these faults.

Chapter 4

Finding a Trade-Off between Observability and Economics in the Fault Detection of Chemical Processes

(*Adapted from Bin Shams et al., 2010a.Computer and Chemical Engineering, in press*)

4.1 Overview

This chapter presents a methodology to quantitatively gauge the potential economical loss due to unobserved faults when standard statistical monitoring charts are used. It is shown that in closed loop operation, a shorter time for detection may result from retuning the controller at the expense of higher product variability. Accordingly, an optimization approach is proposed for finding a trade-off between the economic losses resulting from lack of detection and losses resulting from higher product variability. In order to account for faults with different frequency contents, the method is applied in the frequency domain. The proposed optimization based methodology is later validated in the time domain.

4.2 Introduction

The need for efficient and profitable operation in chemical industries requires the use of efficient process monitoring strategies. Venkatasubramanian *et al*., (2003a) emphasized that the petrochemical industry looses over $20 billion per year due to inappropriate reaction to abnormal process behavior. Thus, faults have a serious impact on process economy, product quality, safety, productivity and pollution level. A fault may be defined as a deviation of at least one variable from an acceptable level (Isermann, 2006). The survey papers (e.g. Himmelblau, 1978;

Isermann, 1984; Gertler, 1988; Willsky, 1976) provide a summary of early work in this area and (Venkatasubramanian *et al*., 2003a, 2003b, and 2003c) provide a more recent account. Most of the available fault detection algorithms involve comparing the observed behavior of the process to the corresponding output of a reference model which may be mechanistic, empirical or semi empirical (Venkatasubramanian *et al*., 2003a). If the fault is observable, the fault detection scheme will generate fault symptom patterns which in turn are fed to the fault diagnosis scheme to determine the root cause of the observed abnormal behavior.

A fault diagnostic system is composed of a detection algorithm followed by a diagnosis scheme. An observable fault is defined as one that can be detected or observed from the chosen set of measurement variables in spite of the background noise. Lack of observability will result in a suboptimal operation due to the presence of an undetected fault.

When data is collected from a process while a fault is occurring, the application of a given statistical model to these data, either univariate or multivariate, is supposed to indicate the presence of the fault. If the statistical model fails to provide indication of the fault this may indicate that the specific fault cannot be observed with that particular model. The most common reasons for this lack of observability are as follows: (a) the measured process variables exhibit low signal to noise ratios and (b) the measured variables do not contain sufficient information regarding this fault and more representative variable(s) should be used for detection (Raghuraj *et al*., 1999; Kourti, 2002). The latter reason is especially important when those variables used for detection are tightly controlled to satisfy quality requirements resulting in lack of information with respect to the fault detection scheme. Then, in order to detect a fault, it may be required to increase the variability, for example, by detuning the controller, so the fault can be observed.

On the other hand, detuning the controller causes deterioration of closed loop performance and possible loss of profit due to higher product variability. Hence, there is a tradeoff between fast fault detection on the one hand and good control on the other. Most of the available fault detection systems, in particular data driven techniques, are implemented as a supplement to the available control system. Despite the significant amount of research in fault detection, the topic of the interaction between control and fault diagnosis has not been extensively studied in particular in the context of fault observability and fault distinguishability. Jacobson and Nett (1991) proposed a four parameter controller setup as a generalization of the two degrees of freedom controllers and Tyler and Morari (1994) reformulated the four degrees of freedom controller into a general framework for which tools from optimal and robust control were applied. The main conclusion of their studies was that when uncertain plants are used in synthesizing a model based controller, the control and diagnostic systems must be synthesized simultaneously. The main drawbacks of these approaches are: (a) they did not use standard fault diagnostic algorithms (e.g. Exponential weighted moving average (EWMA), Cumulative sum (CUSUM), Principal component analysis (PCA), Partial least square (PLS) etc.) and (b) they did not address the economic impact of unobservable faults. The focus of this work is to investigate the simultaneous design of controller and fault diagnosis scheme to enhance fault observability while mitigating through control the impact of unobserved faults.

This work addresses these topics as follows:

1- A tabular CUSUM and T^2-PCA based algorithms are used for detection for univariate and multivariate cases, respectively. Under low signal to noise ratio, it is shown that these algorithms require a certain period of time to detect certain classes of faults. Accordingly, the observability of the fault is related to its duration or alternatively to its frequency.

2- The tuning parameters of the closed loop controller are optimized to achieve an optimal tradeoff between economic losses that may result when high frequency faults (relative to a statistical monitoring chart) and closed-loop variability are experienced.

The chapter is organized as follows. In section 4.3, definitions and theoretical background are presented. The details of the algorithm and the models are given in section 4.4. To illustrate the methodology, a simulation example based on an endothermic continuous stirred tank reactor is presented in section 4.5. Analysis and discussion of the result are presented in section 4.6 followed by conclusions.

4.3 Preliminaries and theoretical background

4.3.1 The Tabular CUSUM

The cumulative sum chart is an efficient control chart that accumulates information collected over current and past samples. The Tabular CUSUM is performed using the following two statistics

$$C_j^+ = \max\left(0, C_{j-1}^+ + x_i - (\mu_{incontrol} + K)\right) \tag{4.1}$$

$$C_j^- = \max\left(0, C_{j-1}^- + (\mu_{incontrol} - K) - x_i)\right) \tag{4.2}$$

where C_0^- and C_0^+ are equal to zero. When either C_j^- or C_j^+ exceed a threshold H, the process is out of control. Based on an extensive study, Montgomery (1997) provides guidelines for the selection of K and H. In general, $K=k.\sigma_{in\ control}$ and $H=h.\sigma_{in\ control}$, where $\sigma_{in\ control}$ is the in-control standard deviation. A common choice for k and h are $\delta/2$ and 5, respectively. The variable δ represents the expected shift in standard deviation units induced in the monitored variable.

4.3.2 Principal Component analysis (PCA)

For a process with n measurement variables, one alternative is to use n univariate control charts to monitor the process. In order to simplify the presentation of information, a second alternative consists of using a principal component analysis (PCA) model to produce T^2 and Q charts for monitoring the n variables simultaneously. PCA involves the computation of loadings and scores using the covariance matrix of data $X \in R^{m \times n}$; where n is the number of variables and m is the total number of samples (MacGregor and Kourti, 1995). If the original variables are correlated, it is possible to summarize most of the variability present in the n variables space in terms of a lower p dimensional subspace ($p<<n$). Here, p represents the number of the principal components. If only two principal components are found, two dimensional score plots are used (i.e. T_1 versus T_2). For more than two principal components, Hotelling T^2 and Q statistics are usually used to monitor the process. The T^2 statistics based on the first p PCs is defined as

$$T^2 = \sum_{i=1}^{p} \frac{t_i^2}{\lambda_i} \tag{4.3}$$

where λ_i is the i-th eigenvalue of the covariance matrix of the original data matrix. Confidence limits for T^2 at confidence level $(1-\alpha)$ are related to the F-distribution as follows:

$$T_{m,p}^2 = \frac{(m-1)p}{m-p} F_{p,m-p} \tag{4.4}$$

; where $F_{p,m-p}$ is the upper $100\alpha\%$ critical point of the F-distribution with p and $(m-p)$ degrees of freedom. Monitoring the process variables by the T^2 values based on the first p principal components is not sufficient since this will only help detect whether or not the variation is within the plane defined by the first p principal components which generally captures steady state correlation. If a new event which was not present in the calibration data used to identify the

reference model occurs, then additional principal components may become significant and the new observation vector x_i will move off the calibrated plane. Such new events can be detected by computing the squared prediction error or Q statistic. Let $x_i \in R^n$ denote the i^{th} multivariate observation vector whose corresponding score is $t_i = x_i.P$. The prediction from the PCA model for x_i is given by $\hat{x}_i = t_i.P^T = x_i P.P^T$. Then, the p dimensional error vector is given by $e_i = x_i - \hat{x}_i$ and the corresponding Q is defined as follows

$$Q = e_i.e_i^T \tag{4.5}$$

Accordingly Q can be thought of as a measure of plant-model mismatch. The confidence limits for Q are given by Jackson (1991). This test suggests the existence of abnormal condition when $Q > Q_\alpha$, where Q_α is defined as follows

$$Q_\alpha = \Theta_1 \left(1 + \frac{c_\alpha . h_0 . \sqrt{2.\Theta_2}}{\Theta_1} + \frac{\Theta_2 . h_0 (h_0 - 1)}{\Theta_1^2} \right)^{\frac{1}{h_0}} \tag{4.6}$$

$$\Theta_i = \sum_{j=p+1}^{n} \lambda_j^{\,i}; \quad for\, i = 1,2,3 \tag{4.7}$$

$$h_0 = 1 - \frac{2.\Theta_1.\Theta_3}{3.\Theta_2^2} \tag{4.8}$$

c_α are the confidence limits for the $(1-\alpha)$ percentile in a standard normal distribution. These confidence limits are calculated based on the assumptions that that the measurements are time independent and multivariate normally distributed.

4.3.3 $ARL_{out\ of\ control}$ as an observability measure and its relation to the fault's frequency

The average run length ($ARL_{out\ of\ control}$) is used in the current work to quantify the observability of the fault. The $ARL_{out\ of\ control}$ is defined as the average number of sampled points

that elapsed between the onset of the fault and its detection. The $ARL_{out\ of\ control}$ is found to be a function of β the probability of type II error, that is, the probability of not detecting a fault although the fault has occurred. For a Shewhart chart, it is straightforward to show that

$$ARL_{out\ of\ control} = \frac{1}{1-\beta} \tag{4.9}$$

The denominator of equation (4.9) is called the power of the detection. For other types of statistical charts, different approaches have been reported in the literature to estimate the $ARL_{out\ of\ control}$. Among these, a Markov Chain based approach has been used (e.g. Brook and Evans, 1972; Lucas and Crosier, 1982). However, in practice, the $ARL_{out\ of\ control}$ is usually obtained using simulations for random realizations of the disturbances (Woodall and Ncube, 1985). The latter approach is adopted in the current study.

To illustrate the use of the $ARL_{out\ of\ control}$ as an observability index a simple example is proposed as follows. The example consists of a unity gain process given by $y = x+n$, where y is the measured value to be used for detection, x is the actual fault and n is white noise with μ=0 and $\sigma^2_{in\ control}$ =1. The magnitude of the signal y is equal to or less than the noise level. Two square-wave like faults with period T=24 samples and T=3 samples, respectively are considered as candidate faults x to be identified by the fault detection algorithm. These faults are shown in Fig. 4.1a and Fig. 4.2a, respectively.

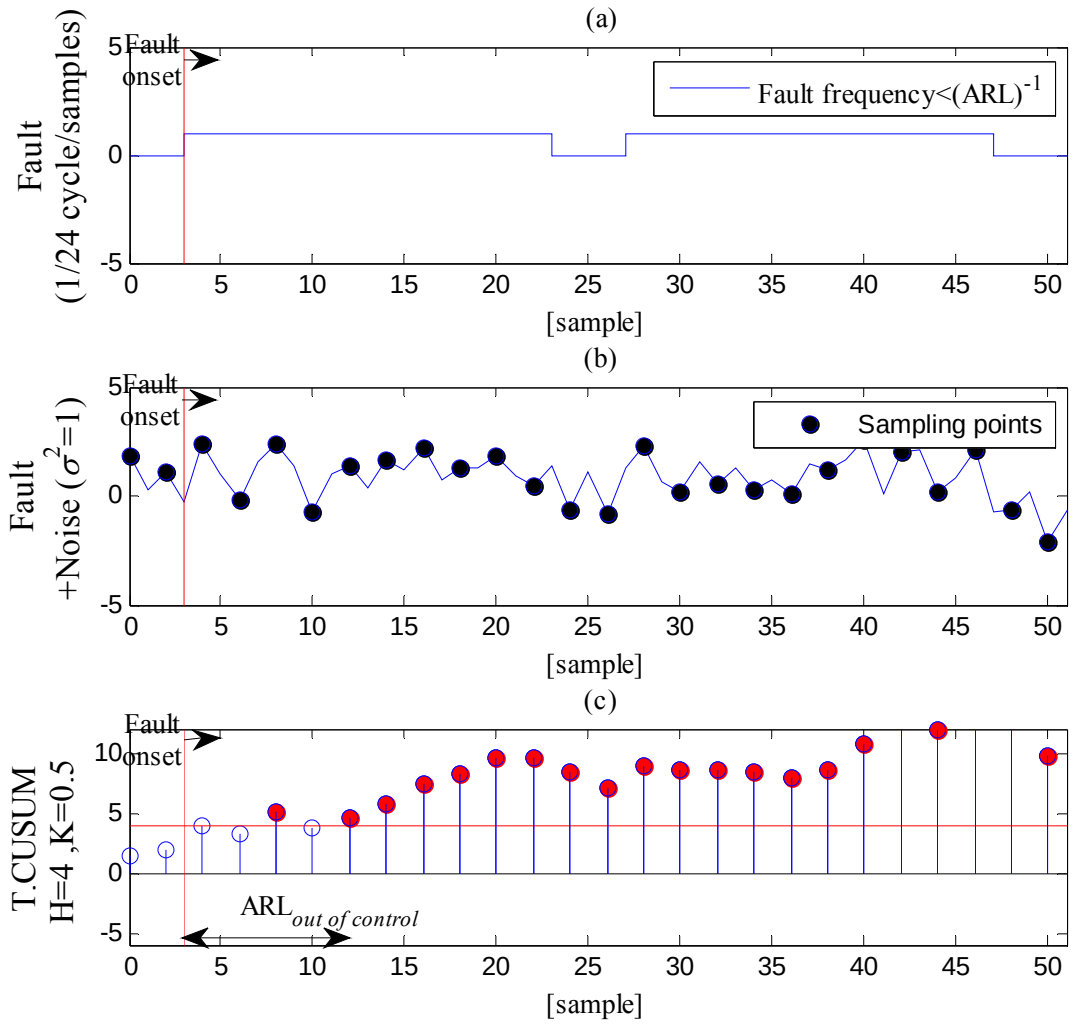


Fig. 4.1 The dependency of the tabular CUSUM $ARL_{out\ of\ control}$ on the frequency of the fault. (a) The faulty signal with a frequency content smaller than the tabular CUSUM $(ARL_{out\ of\ control})^{-1}$. (b) Noise added to the faulty signal where dark circles represent the sampling instants. (c) The monitoring of the faulty signal using tabular cusum. Dark circles represent the detection of the faulty samples. Fault onsets at sample=3.

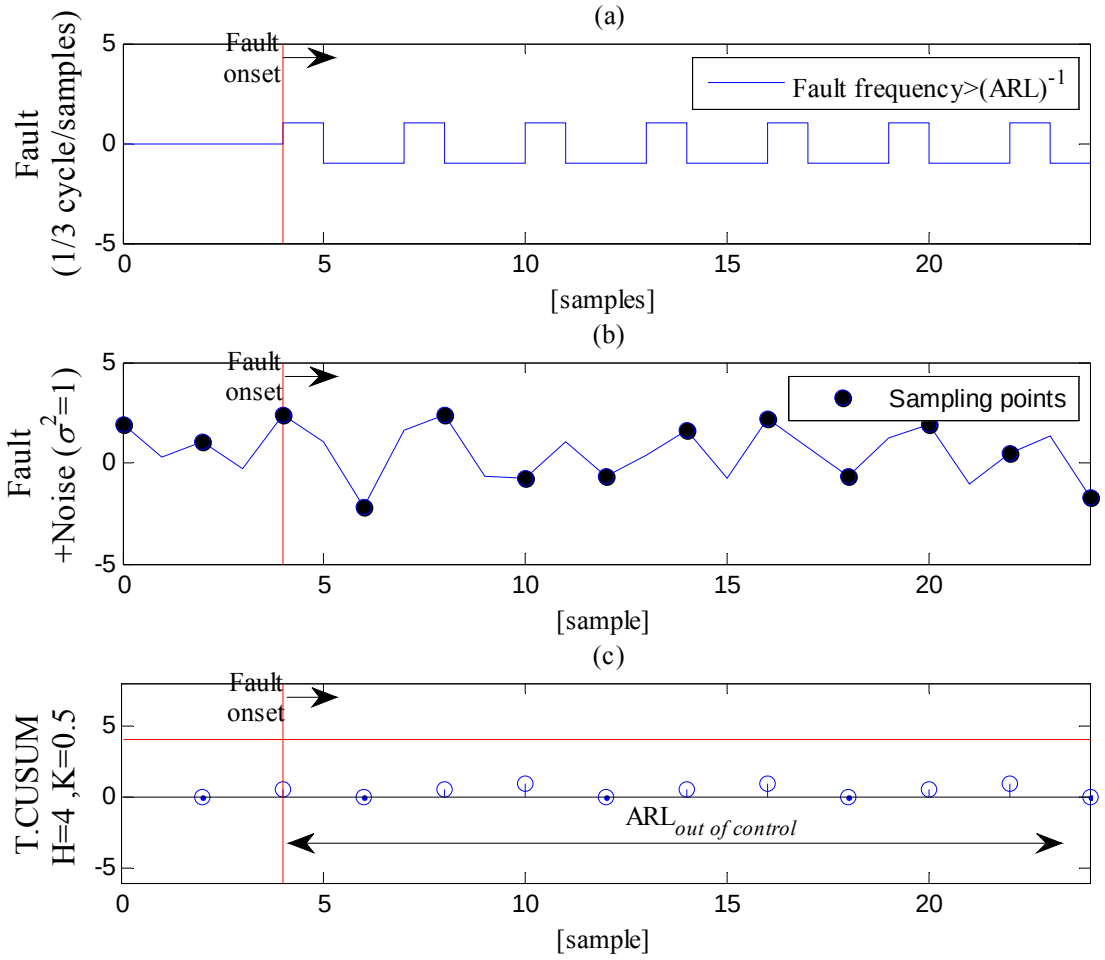


Fig. 4.2 The dependency of the tabular CUSUM $ARL_{out\ of\ control}$ on the frequency of the fault. (a) The fault signal with a frequency content greater than the tabular CUSUM $(ARL_{out\ of\ control})^{-1}$. (b) Noise added to the faulty signal where dark circles represent the sampling instant. (c) The monitoring of the faulty signal using tabular cusum. No violation of the threshold has found. Fault onsets at sample=4.

Following the addition of noise, the resulting measured values of y corresponding to these two faults are given in Fig. 4.1b and Fig. 4.2b, respectively, where the solid circles represent measurements obtained at discrete sampling intervals. Considering the magnitude of the square signals and the magnitude of the noise, the theoretical $ARL_{out\ of\ control}$ is determined by the expected shift and the chosen set of parameters as explained in section 4.3.1 (i.e. K and H). In this example, the theoretical $ARL_{out\ of\ control}$= 8.38 samples where k=0.5 and h=4 (Montgomery, 1997). That is, K=0.5 and H=4, since $\sigma^2_{in\ control}$ =1. Based on this value of $ARL_{out\ of\ control,}$ it is expected that a square signal with a period that is longer than the $ARL_{out\ of\ control}$ will exceed the threshold H shown by a horizontal solid line in Fig. 4.1c and therefore will be detected as shown in this figure. On the other hand, pulses of shorter duration than the calculated $ARL_{out\ of\ control}$ will not exceed the threshold H and therefore, they will not be detected as shown in Fig. 4.2c. Theoretically, an $ARL_{out\ of\ control}$ = infinity will be obtained if the fault is not observable regardless of its period, while an $ARL_{out\ of\ control}$=1 will result if the fault can be observed on the first sample following the onset of the fault. Accordingly, the observability of faults can be estimated from the relative values of the period of the square wave versus the estimated theoretical $ARL_{out\ of\ control}$ value (8.38 samples). Thus, the fault observability is related to the period of the square signal. Alternatively, it can also be related to the inverse of the period, i.e., the frequency. In that case, the frequency of the fault can be compared with the *inverse* of the estimated theoretical $ARL_{out\ of\ control}$ to assess whether the fault can be detected or not. Correspondingly, the inverse of $ARL_{out\ of\ control}$ is used in this work to assess observability as a function of the fault's frequency. This observability calculation is then used within an optimization problem formulated for finding a tradeoff between fault detection ability and closed loop control performance. The formulation of this problem is detailed in the following section.

4.4 Optimal tuning of the closed loop controllers supervised by a fault detection algorithm

This section proposes a method to tune a system composed of a statistical fault detection algorithm and closed loop controllers. Since the tuning of these controllers are expected to affect both the speed of detection of the fault and the variability in manipulated and controlled variables, optimal tuning parameters are sought from the solution of an optimization problem that considers costs related to the lack of detection and the variability. To achieve this objective it is proposed to solve the following optimization problem:

$$\min_{\lambda} J = a_1\gamma_1 + a_2\gamma_2 + a_3\gamma_3 \tag{4.10}$$

where (γ_1) is the variability associated with quality characteristics variables, (γ_2) is the variability associated with the operating costs of the process, e.g. costs of utilities and (γ_3) is the variability associated with the rate of the control changes which can be generally related to the cost of wear of actuators. The weight a_i determines the relative contribution of each variability to the total variability of the system and λ is a vector containing the decision variables. Two cases were considered for λ as follows: (a) λ includes the controller tuning parameters only and (b) λ includes both the controller tuning parameters and the Tabular CUSUM parameters, that is, k and h. The coefficients in the objective function a_i represent the expected repair/dissatisfaction cost at maximum tolerable variability $(\gamma_i^{\max})$. This definition of a_i is motivated by the proportionality constant in the well known Taguchi's loss function (Montgomery, 1997; Ross, 1988). These costs are problem specific. For instance, the cost of utilities (second term in the RHS of equation 4.10) is obtained from the expected cost of utilities per unit of variability for the utility source under consideration, e.g. steam etc. On the other hand the cost assigned to the rate of change of a valve (third term in the RHS of equation 4.10) could be obtained from the cost of the valve and from the expected valve life for a specific variability.

In addition, since it is easier to characterize the faults entering the system by a frequency dependent distribution, the variabilities (γ_i) are calculated in the frequency domain using Parseval's identity as follows:

$$\gamma_i = \left(\frac{1}{2\pi} \int_{-\omega_l}^{+\omega_u} \left| \overbrace{G_i(j\omega) l(j\omega)}^{y_i(j\omega)} \right|^2 d\omega \right) . \alpha(j\omega) = \int_0^t \left| y_i(t) \right|^2 dt$$

$$\alpha(\omega) = \begin{cases} 1 & \omega > (ARL_{out\ of\ control})^{-1} \\ 0 & otherwise \end{cases} \tag{4.11}$$

where $G_i(j\omega) \in \Omega$ is the closed loop transfer function that relates a specific fault to an output y_i, and belongs to the set of identifiable transfer functions Ω. To solve the problem in (4.10) it is necessary to find the closed loop transfer functions G_i between the fault and the variables related to product quality, variables related to the operating costs and variables related to actuators used for closed loop control. The input signal $l(j\omega)$ is a frequency dependent description of the faults that has to be identified off-line from historical process data. As explained in the previous section, the length of time necessary to detect a fault can be quantified by the $ARL_{out\ of\ control}$ and its inverse $(ARL_{out\ of\ control})^{-1}$ can be effectively used to distinguish between fault frequencies that can be observed, using a particular statistical monitoring chart, from fault frequencies that are unobservable. This can be further illustrated by using Fig.4.3. For example, if the theoretical $ARL_{out\ of\ control}$ for a particular control chart is 10 samples, its reciprocal (0.1cycle/sample) determines the boundary value between the observed and unobserved fault's frequencies as shown in Fig. 4.3. The role of the switching function $\alpha(\omega)$ in equation (4.11) is to assign costs only to the faults that can be observed as opposed to faults that cannot be observed based on their corresponding frequencies.

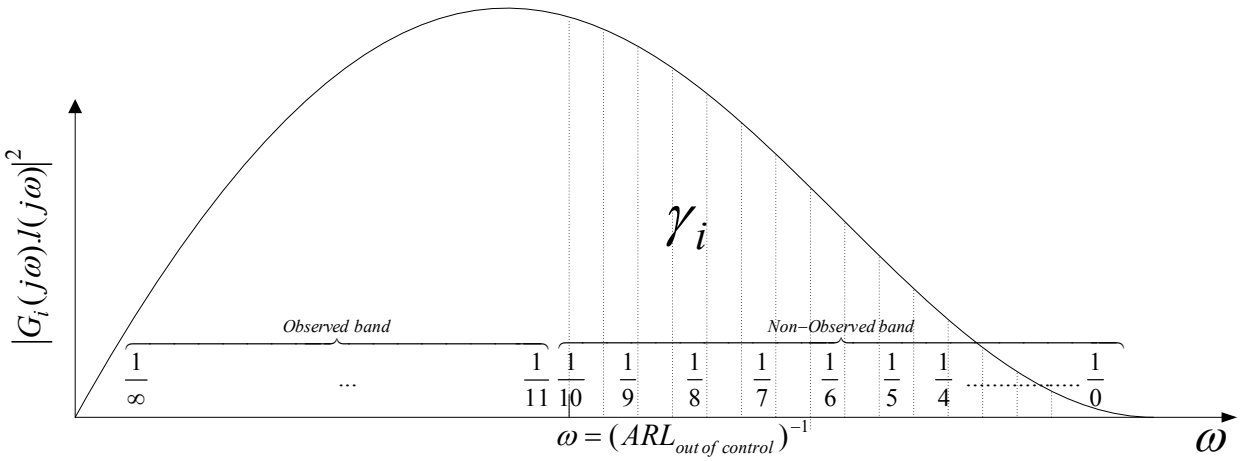


Fig.4.3 The included domain of the $\alpha(\omega)$ in the definition of (γ_i) in equation (4.11)

Regarding the choice of the magnitude of the fault Δ_{fault}, the smallest expected value for a particular system can be selected so as to obtain the longest $ARL_{out\ of\ control}$ which will result in the worst case cost of undetected faults.

Accordingly, the variabilities in equation (4.11) and the corresponding costs in equation (4.10) account only for the costs incurred due to lack of detection of the fault. An implicit simplifying assumption made in this work is that when a fault can be detected, the cost incurred during the time between the onset of the fault and its detection is negligible as compared to the cost of unobservable faults. This assumption is based on the premise that if the fault can be detected, the faulty situation can be addressed, whereas when faults cannot be detected, they have a long term economic impact that is not being mitigated.

The proposed optimization methodology may be modified to account for situations when two or more faults occur. In that case the worst case scenario can be sought by considering the superposition of costs as follows:

$$\min_{\lambda} J_{total} = J_{f1} + J_{f2} + ... + J_{fk} \tag{4.12}$$

To summarize, the solution to the optimization problem proposed in equation (4.10) for one fault or the minimization of the cost in (4.12) for two or more faults proceed as per the following steps:

1- Assume initial guesses for the parameters of the controller or/and monitoring charts.

2- Simulate the dynamic model without noise and identify stable closed loop transfer functions for a RBS (Random Binary Sequence) input signal. The frequency content of the RBS signal is representative of the plant's faults. Stability is tested through the identified closed loop transfer function poles.

3- For a given noise level, simulate the closed loop dynamic model for different noise realizations and calculate the $ARL_{out\ of\ control}$ for a specific fault or a set of faults.

4- Calculate variabilities in the frequency domain per equation (4.11).

5- Minimize the objective function as per equation (4.10) with respect to the decision variable vector (controller/monitoring parameters). A detailed flowchart summarizing this procedure is given in Fig. 4.4. Finally, an additional possible use of the optimization results is to assess whether a required level of profitability can be obtained by simply retuning the controllers. For example, if the minimal value of J^* in equation (4.10) is larger than a pre-specified acceptable cost, this will signify that retuning of the controllers is not sufficient to achieve the desired profitability and instead, other sensors have to be selected for fault detection or a plant design change may be required. On the other hand, if the value of J^* is smaller than the pre-specified cost, then the fault under investigation would be considered of minimal economic significance and its detection would be deemed less critical from an economical point of view.

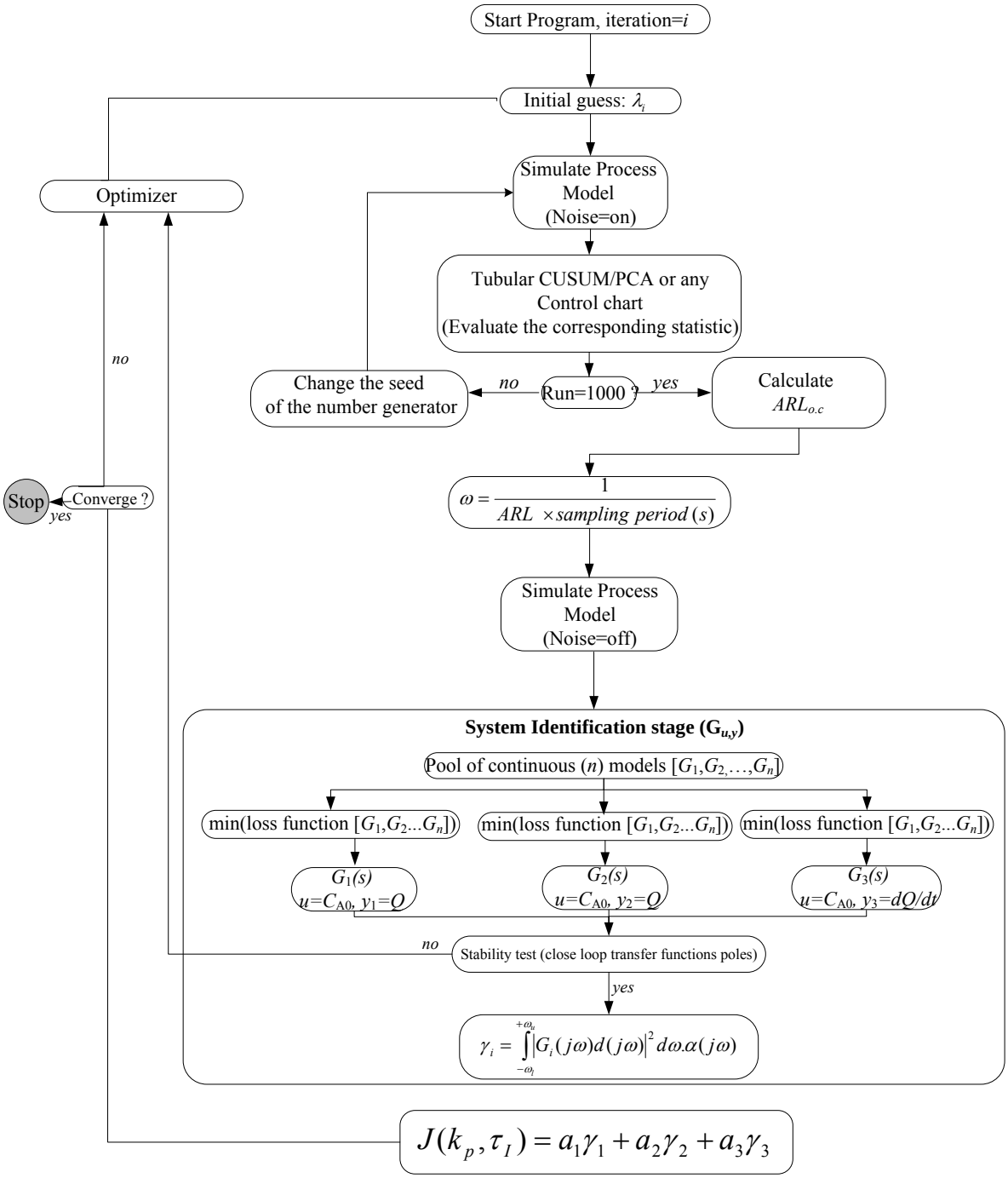


Fig. 4.4 The flow diagram of the proposed scheme to assess the economical significance of unobserved faults

4.5 Case study: (Continuous Stirred Tank Reactor, CSTR)

The proposed methodology and optimization problem given by equation (4.11) are illustrated for a non isothermal CSTR system (Riggs, 1999). Fig.4.5 depicts the CSTR with a concentration control loop as well as the variables for which faults are considered. Table.4.1 shows the process parameters and the operating conditions used in the simulation.

The control objective was to maintain the outlet concentration at certain required level by manipulating the heat duty *Q*. The model and PI control algorithm are described by the following equations:

$$V_r \frac{dC_A}{dt} = \frac{F}{\rho}(C_{A0} - C_A) - V_r k_0 C_A e^{-\frac{E}{RT}} \tag{4.13}$$

$$V_r \rho C_v \frac{dT}{dt} = FC_p(T_0 - T) - V_r \Delta H k_0 C_A e^{-\frac{E}{RT}} + Q \tag{4.14}$$

$$Q(t) = Q_s - k_p(C_{A.set} - C_A(t)) - k_p / \tau_I \int_{t_0}^{t} (C_{A.set} - C_A(t^*))dt^* \tag{4.15}$$

The noise in all measured variables was assumed to be uniformly distributed with variance (σ^2_{noise}) . Different levels of noise were tested in order to assess the performance of the proposed methodology for different signal to noise ratios. For this problem, the outputs y_i that were considered in the optimization problem given by equation (4.10) are the outlet concentration (*CA*) that is viewed as a quality related variable, the heat duty (*Q*) that represents the operating cost and the rate of change of the heat duty (*dQ/dt*) that is related to the actuator wear cost. To evaluate the variabilities (γ_i) in *CA*, *Q* and *dQ*/*dt* with respect to changes in the inlet concentration, random binary sequences in the inlet concentration, are used to identify the corresponding three closed loop transfer functions $G_i(j\omega)_{u,y_i}$.

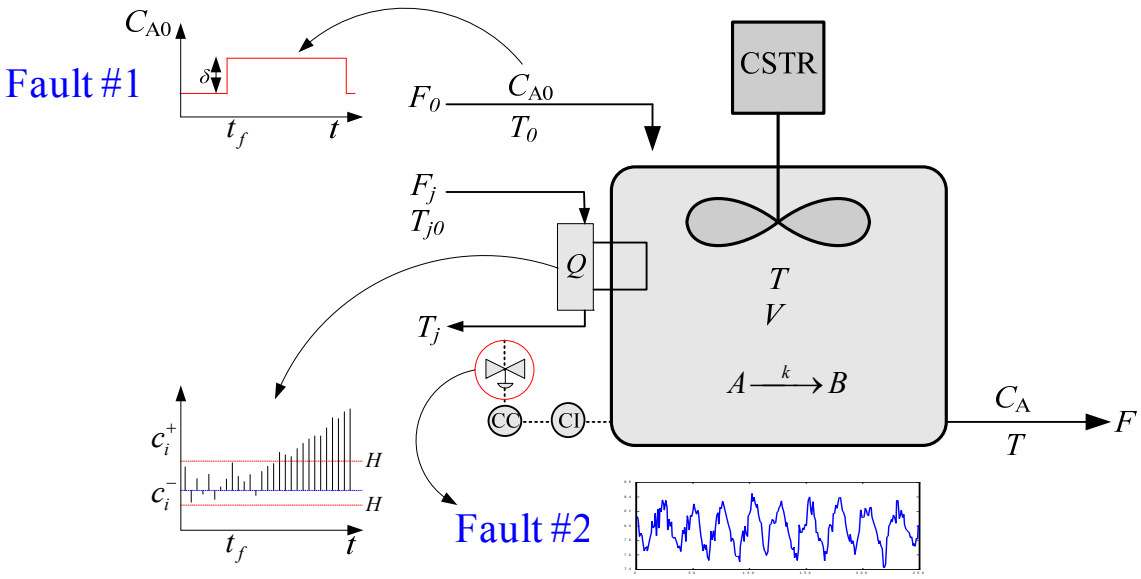


Fig.4.5 The Endothermic Continuous Stirred Tank Reactor (CSTR) with a concentration control loop and two types of typical industrial faults

Table 4.1 Variable descriptions and parameters settings used in the CSTR simulation

Symbol	Description
V_r	Reactor volume (100 L)
F	Mass feed rate (10 kg/s)
C_{A0}	Inlet concentration of component A (1.0 gmoles/L)
ρ	Density of the reactor feed and product (1.0 cal/g)
C_p	Heat capacity, reactor feed and product (1.0 cal/g/K)
C_v	Assumed equal to C_p
ΔH	Heat of reaction (160000 cal/gmoles)
T_0	Feed temperature (400 K)
E/R	Normalized activation energy (20000 K)
C_A	Reactant concentration (0.25 gmoles/l)
k_0	Rate constant (1.97E24 s^{-1})
T	Reactor temperature (350 K)
Qss	Heat addition rate (700000 cal/s)

The input concentration signal was assumed to be limited to a frequency bandwidth of 0-2 Hz. The identification of the transfer functions was performed using the MATLAB® system identification toolbox. A set of transfer function of different orders was considered. Several step responses were simulated with the nonlinear dynamic model to estimate the order of the system. In most of the cases, transfer functions with one zero and two poles were found adequate to model the input-output relation. For each input/ output pair, a model discrimination criterion is used to select the model with the smallest loss function value, i.e. normalized sum of square errors. Although the original problem is described by nonlinear equations, the linear approximation is justified by the fact that the models represent closed loop behavior; hence, the variables remain in a close neighborhood of the nominal operating condition.

Most industrial faults manifest themselves as either a change in the mean or by an increase in the variability of the process. To illustrate both situations, two types of faults are implemented in this case study and are shown in Fig.4.5. The first fault is a square wave variation in the inlet concentration with frequency equal 0.002 Hz and different possible magnitudes (Δ_{fault}). This low frequency signal would appear as a step change within the time period of simulation. The second fault is the occurrence of stiction in the concentration control loop valve. A valve stiction usually manifests itself as a periodic fluctuation in manipulated or controlled variables (Choudhury *et al*., 2005; Horch, 1999; Thornhill, 2005). To simulate the stiction behavior of a valve, the model given in (Choudhury *et al*., 2005) and schematically described in Fig. 4.6 is used. It is a two-parameter (*S* and *J*) data driven model. The parameter *S* is the dead band of the valve given as a percentage of the controller output span. *J* is referred to as the *slip jump* and it represents the sudden release of potential energy, stored in the actuator chamber due to high

static friction, in the form of kinetic energy as the valve starts to move. The latter is given as a percentage of the manipulated variable span.

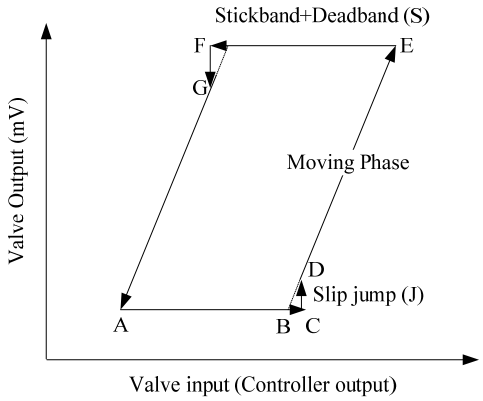


Fig.4.6 Typical input/output behavior of a sticky valve (Choudhury *et al*., 2005).

The parameter *J* determines the aforementioned fluctuating behavior. This model was found consistent with a typical "sticky" valve behavior and has been validated using real industrial data (Choudhury *et al*., 2005).

The identification of the closed loop systems was performed using the MATLAB® identification toolbox. The optimization problems were solved with the MATLAB® optimization routine fminsearch. Due to its non convexity, the problem has been solved for different randomly generated initial guesses, and the lowest cost function value among all the runs is chosen as the optimum. The CPU time required to solve a single optimization problem in this study was approximately 1.5 h on an Intel® Core ™ 2 Duo desktop processor (2.66 GHz and 2.0 GB RAM). The proposed approach is expected to offer advantages for the solution of larger problems since the variability has been calculated by a semi-analytical approach (Parseval's integral). As such this approach is expected to be computationally more efficient than calculating variability based

on dynamic simulations that need to be conducted for a very large number of possible disturbances.

4.6 Results and Discussion

4.6.1 Tabular CUSUM: Fault#1 (A square wave variation in the reactor inlet concentration)

Based on the dynamics of the process and the bandwidth of the fault signals the sampling interval for the measurements is chosen to be 3 seconds. Two different fault detection algorithms were considered: (a) Tabular CUSUM based solely on the manipulated variable i.e. the heat duty and (b) T^2 statistic based on the multivariate PCA model. The latter is based on three measured variables: the temperature (T), the heat duty (Q) and the outlet concentration (CA). A square wave variation in the inlet concentration of magnitude Δ_{fault} = 0.035gmol/l is simulated. The tabular CUSUM parameters are chosen as explained in section 4.3.1. Specifically, K=0.5δ, where δ is calculated by averaging, at each iteration, the first and last 200 points of the monitored variable (Q) and then taking their difference. On the other hand, H=5$\sigma_{in\ control}$, where $\sigma_{in\ control}$ is determined by the chosen level of noise.

First, using the tabular CUSUM algorithm, the optimization problem in Equation (4.10) is solved for different levels of measurement noise. The cost function given in equation (4.11) for this particular problem is composed of three terms, namely, a cost associated with the variabilities in outlet concentration, in the heat consumption and in the rate of change of the heat load which may be related to the wear of the valve. Tables 4.2 and 4.3 show the optimal costs obtained from

the solution of problem (4.10) when the fault magnitudes are Δ_{fault} = 0.035 and Δ_{fault} = 0.100 gmol/l, respectively.

Table 4.2 Summary of the results for fault#1 [square wave in the inlet concentration Δ_{fault}=0.035, f = 0.002Hz] using the Tabular CUSUM

Noise level [%]	k^*_p [*cal.l/s/gmol*]	$\tau^*_I (60)^{-1}$ [*min*]	$ARL^*_{out\ of\ control}$	Cost function (J^*)
0.1	5000	0.612	27.36	32488
0.2	75508	0.507	37.80	68590
0.3	51596	0.632	55.92	82976

Table 4.3 Summary of the results for fault#1 [square wave in the inlet concentration Δ_{fault} = 0.10, f = 0.002Hz] using the Tabular CUSUM

Noise level [%]	k^*_p [*cal.l/s/gmol*]	$\tau^*_I (60)^{-1}$ [*min*]	$ARL^*_{out\ of\ control}$	Cost function (J^*)
0.1	408890	2.716	22.56	20134
0.2	385010	2.287	23.82	22482
0.3	458690	2.266	24.72	26369

To investigate the effect of the signal to noise ratio effect on the observability of faults, different noise levels have been considered. The identified model structure has been fixed for all noise level to block any variability that could arise due to model structure variation. As shown in these two tables, the cost decreases in both Tables 4.2 and 4.3 as the signal to noise ratio increases since, as expected, it is easier to detect faults with large signal to noise ratios thus reducing the impact of lack of observability. The same effect can be also seen when comparing faults of different sizes for the same amount of noise. For example, when the noise level is 0.2%, the combined cost associated with loss due to variability in concentration, cost of steam and the cost of valve wear is reduced by 67.2% when the fault size is increased from Δ_{fault} =0.035 gmol/l to Δ_{fault} =0.100 gmol/l (see Tables 4.2 and 4.3). Also, as expected, for the same level of noise the

$ARL^*_{out\ of\ control}$ increases as the magnitude of the fault decreases: $ARL^*_{out\ of\ control}$ is 23.82 s for a fault magnitude of Δ_{fault} = 0.10 gmol/l and 37.80 s for a fault magnitude of Δ_{fault} =0.035 gmol/l.

To validate the results of the optimization problem, the nonlinear dynamic model was simulated in the time domain. The objective of this validation was to assess the appropriateness of Parseval's identity (equation 4.11) that is based on a linear approximation to estimate the nonlinear system variabilities. The second case (i.e. 0.2 %) from Table 4.2 is considered. A random binary sequence with amplitude of 0.035gmol/l and frequency bandwidth between $(37.8)^{-1}$Hz to infinity has been designed in the inlet concentration to the reactor. The three variables, namely, the outlet concentration (*CA*), heat duty (*Q*) and the rate of change of the heat duty (*dQ/dt*) are simulated in the time domain with the actual nonlinear CSTR model equations and using the optimal controller's tuning parameters presented in Table 4.2 for 0.2% noise level. For each variable, the right hand side of equation (4.11) is calculated by integrating the square of the variability over time and the total cost is then determined using equation (4.10). Although the costs calculated with Parseval's identity and with time based simulations are relatively close, the cost calculated from the time domain simulation was less than that calculated in the frequency domain with an error of 8.9 % (J^*= 62444 versus J^*= 68590) thus verifying that the frequency based calculation provides an upper bound.

In addition, to test the sensitivity of the optimal solutions, 10% and 20% perturbations are introduced in the optimal values of the tuning parameters (i.e. k^*_p and τ^*_I) and the cost in the time domain is recalculated . It was found that the value of the objective function in the neighborhood of the optimal solution is the lowest and the value of the cost function increases as the controller

parameters deviate from the optimal one. In particular, the values of the cost function are J= 66155 and J= 70349 for 10% and 20% perturbations in the controller parameters, respectively compared to J^*= 62444 for the optimal set of tuning parameters.

An additional study was conducted where the parameters of the tabular CUSUM, namely k and h were used as design variables together with the tuning parameters of the controller to minimize the cost function. From equation (4.1) and (4.2), it can be seen that the calculated statistics are affected by the value of $K=k\delta$. In this study, k instead of K has been included as decision variable, since δ (the deviation in the monitored variable) is recalculated at each iteration as explained in Section 4.6.1. Similarly, h instead of H is considered, since $\sigma_{in\ control}$ is fixed to achieve the desired signal to noise ratio (see Section 4.3.1). These two parameters have been bounded within ±10% of their recommended values. It is noticed from simulations that the use of values much larger than the recommended ones result in significant deterioration in the tabular CUSUM chart performance. For instance, for a 0.2% noise level and Δ_{fault}= 0.035gmol/l, the undetected cost J^*=66060 is obtained with the following optimal values for the design variables: k^*_p= 95477, τ^*_I = 0.699, h^*=4.90 and k^*= 0.490. Thus, at the given fault magnitude and noise level, the cost has been decreased by 3.68%, compared to the result obtained when only the controller tuning parameters are used as decision variables. It is also noticed that the optimal values of h^* and k^* are very close to the one recommended by Montgomery (1997). Based on the fact that the improvement in cost when the h^* and k^* are optimized together with the controller parameters is models and in order to decrease the complexity of the optimizations, only the controller tuning parameters were considered as decision variables in the rest of the case studies.

4.6.2 T^2-PCA: Fault#1 (A square wave variation in the reactor inlet concentration)

To demonstrate the applicability of the proposed method to handle a multivariate monitoring problem, the optimization problem was solved for a case where a multivariate static PCA is used for fault detection. Three measurements were chosen for formulating the PCA detection algorithm: reactor temperature (T), outlet concentration (CA) and the manipulated variable (Q). Using the parallel analysis algorithm (Chiang et al., 2001), two principal components were found adequate to capture more than 97% of the variability of the system. The thresholds for fault detection were calculated from equations (4.4) and (4.6). Either one of the two complementary statistics, i.e., T^2 and Q, may be used to calculate the $ARL_{out\ of\ control}$ when a PCA is used for fault detection. T^2 is the sum of square of the scores and represents the variability captured by the PCA model while Q is the sum of squares of the PCA model errors. The choice of one of them within the framework of the proposed methodology depends on the situation at hand. In the current study the T^2 is used, since it shows more sensitivity in detecting the two faults under consideration. Table 4.4 summarizes the optimal cost function values for PCA when the fault magnitude is Δ_{fault}= 0.035gmole/l.

Table 4.4 Summary of the results for fault#1[square wave in the inlet concentration Δ_{fault} =0.035, f = 0.002Hz] using the PCA

Noise level [%]	k^*_p [*cal.l/s/gmol*]	$\tau^*_I (60)^{-1}$ [*min*]	$ARL^*_{out\ of\ control}$	Cost function (J^*)
0.1	5082	1.042	37.38	1354.1
0.2	12475	1.650	58.44	6109.9
0.3	5309.6	0.919	75.54	8033.4

Once again, the cost associated with the undetected faults is decreased as the signal to noise ratio increases. It should be noticed that the $ARL^{*}_{out\ of\ control}$ is larger, for all noise level, when T^2 PCA based statistic is used as compared to the case where the Tabular CUSUM is used. This result is expected, since with this small magnitude, i.e. Δ_{fault}= 0.035gmol/l, the integrating nature of the tabular CUSUM will indicate the occurrence of the fault faster than the case where T^2 is used since the latter is solely based on the instantaneous measurement vector.

4.6.3 Tabular CUSUM / T^2-PCA: Fault #2 (stiction in the steam control valve)

This case study illustrates the application of the proposed methodology in the presence of stiction in the control valve. The data-driven model given in (Choudhury *et al*., 2005) has been used. As mentioned in section 4.5, the behavior of the model is characterized by the two parameters *J* and *S*, see Fig.4.6. The effect of both the *J* and *S* parameters on the observability of the stiction fault was studied. It was found that *J* only affects the amplitude of the limit cycle resulting with a sticky valve. As shown in Fig. 4.7, as the slip-jump (*J*) increases the amplitude of the limit cycle increases. For a fixed *slip*-jump value (*J*), the increase in the *S* value increases the $ARL_{out\ of\ control}$ since the role of the stick-band (*S*) is to introduce a delay in the output of the stiction model. In this case study, the values of *J* and *S* are chosen to be 0.034 and 0.2, respectively. A summary of the results for this case is shown in Table.4.5. In contrast with Fault #1, a large cost results with this fault when using the Tabular CUSUM. The reason is that the Tabular CUSUM is especially suitable, for the detection of small persistent biases in variables, whereas in the presence of stiction, the variables are fluctuating around an approximately zero

mean value, thereby making the fault hard to detect using the tabular CUSUM. On the other hand, T^2 -PCA charts are suitable for detecting variability changes, since the T^2 statistics is based on the system covariance matrix.

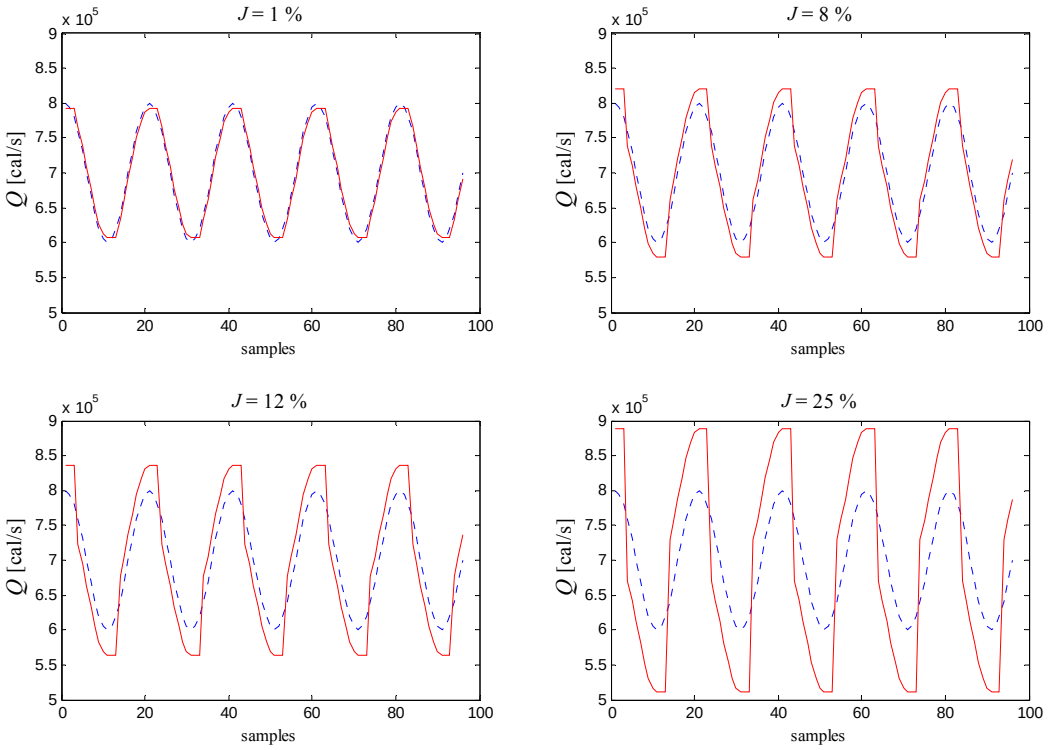


Fig.4.7 The effect of the slip jump (J) on the amplitude of the limit cycle. The dashed line depicts the controller output whereas the solid line represents the output from the stiction model. In all of the above subplots, S= 3%

Table 4.5 Summary of the results for fault#2 (stiction in the reactor heating valve); J=0.034 and S=0.2.

Chart type	Noise level [%]	k^*_p [*cal.l/s/gmol*]	$\tau^*_I (60)^{-1}$ [*min*]	$ARL^*_{out\ of\ control}$	Cost function (J^*)
Tabular CUSUM	0.1	13781	0.807	160.44	58991
T^2(PCA)	0.1	11785	4.805	4.140	0.2365

4.6.4 Tabular CUSUM: Simultaneous occurrence of Fault#1 and Fault#2

Finally, to handle the situation where more than one fault is experienced, the superposition of the costs corresponding to the two faults (equation 4.12) is minimized with respect to the controller parameters. For 0.1% noise level, a minimum cost of J^*= 307890 is obtained when k^*_p= 44341 and τ^*_I = 2.560. This cost is significantly higher than the sum of the two costs obtained when the faults were treated separately, see Table 4.1 and Table 4.5. This indicates a large sensitivity in the observability of the faults with respect to the controller's tuning parameters and it also verifies the fact that the Tabular CUSUM is not very suitable for detecting stiction as shown in subsection 4.6.3.

4.7. Conclusions

In this chapter, a methodology has been developed to quantify the cost associated with faults with different degrees of observability. The main objective was to consider the economical consequences associated with theses faults. The proposed methodology minimizes, over the frequency domain, the cost associated with the quality characteristic variable(s), operating cost and the cost associated with the control changes while adjusting the controller tuning parameters and the parameters of the fault detection algorithm. Observability of the fault has been gauged using the concept of $ARL_{out\ of\ control}$ and incorporated within the proposed framework. The method has been tested using an endothermic continuous stirred tank reactor (CSTR). Two faults have been considered, a low frequency square wave in the inlet concentration and valve stiction. Tabular CUSUM and multivariate PCA were used for detection. Each one of these methods was found to perform better for a particular type of fault. For example, the CUSUM was found

suitable for detecting faults that consist on changes in mean whereas multivariate PCA is more suitable for detecting changes in both mean and variance. The results have been validated in the time domain to test the suitability for using Parseval's theorem in quantifying the chosen variabilities.

Chapter 5

Enhancing Fault-Observability by the Use of Feedback Control

(*Adopted form Bin Shams et al., 2010b, Industrial and Engineering Chemistry Research, in press***)**

5.1 Overview

This chapter deals with detection of faults in the Tennessee Eastman problem; that have been found unobservable by previous studies. Hotelling's-T^2 charting based on the cumulative sums of the faults' relevant variables was successful in detecting these faults, however, with significant delays. In order to reduce these delays it is proposed to retune the feedback controllers involving controlled and manipulated variables used for detection. An optimization based methodology that searches for an optimal trade-off between fault detection and control performance is formulated. The resulting controller design is compared with a design previously reported in the literature showing that significant reductions in the detection delays and overall reductions in plant's costs can be achieved by proper tuning of controllers. Both individual and simultaneous occurrence of faults is considered.

5.2 Introduction

An important aspect for the economical and safe operation of a chemical process is the rapid detection and removal of malfunctions or faults. A fault may be defined as a deviation of at least one variable from an acceptable level (Isermann, 2006). Different methods have been proposed in the literature for fault detection and fault diagnosis (Isermann, 2006; Venkatasubramanian *et al*., 2003a; Venkatasubramanian *et al*., 2003b; Venkatasubramanian *et*

al., 2003c; Chiang *et al*., 2001). These methods can be broadly categorized into three main classes: (1) Analytical methods which are solely based on first-principles models e.g. observer based techniques; (2) empirical methods e.g. univariate and multivariate statistical methods and (3) Semi-empirical methods, which combine empirical models with prior knowledge about the system. Each of these methods has its own advantages and disadvantages depending on the problem. Analytical methods require the use of first-principle models, thus making them less attractive for large scale systems. Therefore, they are not considered in the current work and instead an empirical method is used. A number of researchers suggest combining these methods to improve detection. For example, Chiang *et al*., (2003) and Lee *et al*., (2004) have observed that empirical methods are enhanced if knowledge of the process is used to account for the fundamental causal relationships among variables. The method used in the current work makes use of process knowledge to enhance detection. Different problems associated to the fault detection and diagnosis problems have been reported in the literature (Isermann, 2006; Venkatasubramanian *et al*., 2003a; Venkatasubramanian *et al*., 2003b; Venkatasubramanian *et al*., 2003c; Chiang *et al*., 2001). Distinguishability and Observability are two central problems related to the fault detection and diagnosis problem (Qin, 2003). Distinguishability problem arises when a system exhibits similar responses in variables used for detection while different faults occur (Qin, 2003; Benjamin *et al*., 2008; Lou *et al*., 2003). On the other hand, fault observability is related to the detection phase and it is interpreted as a delay in identifying the occurrence of a fault. This work is focused on the observability problem and proposes a methodology to mitigate it through feedback control.

A recent study by Bin Shams *et al*., (2010d, Appendix B) has proposed the application of Cumulative-Sum (CUSUM) based models for the detection of faults in the Tennessee Eastman

problem (TEP), (Downs and Vogel, 1993). Bin Shams *et al*. (2010d, Appendix B) proposed the application of Location CUSUM (LCS) and Scale CUSUM (SCS) based models to detect three particular faults that have been found unobservable by other algorithms previously applied to the TEP (Chiang *et al*., 2001; Chiang *et al*., 2003; Lee *et al*., 2004; Cheng *et al*., 2010; Ding *et al*., 2009; Zhang *et al*., 2009; Ku *et al*., 1995). After demonstrating the detection capability of the univariate CUSUM based methods for each one of the three faults, a Hotelling's T^2 chart based on a cumulative sum of the observations was proposed for the individual or simultaneous detection of these three faults. The latter was found as essential when large numbers of correlated variables are considered.

To quantify fault observability when using the CUSUM based statistics, Bin Shams *et al*., (2010a, Chapter 4) proposed the out of control Average Run Length ($ARL_{o.c}$). Their proposed CUSUM based statistics was successful in detecting the three unobservable faults (i.e. compared to other methods reported in the literatures); however, a long time of detection, i.e. large $ARL_{o.c}$ values were obtained. To mitigate this problem and since the faults are observed from variables that are embedded within control loops, the current work studies the effect of controllers' tuning parameters on fault observability.

Despite the evident interaction between fault detection and control, the research on fault detection and control algorithms evolved separately, mainly because of the challenges associated with each of these problems (Tyler and Morari, 1994). The interaction between control and detection stems from the fact that most of the monitored variables are either process variables (PV) or manipulated variables (MV) operated within feedback control loops. Since the detection is highly dependent on the dynamic response of the monitored variables, there is a possibility to speed up the detection of the faults by re-tuning the controllers involving these variables. The

tradeoffs between fault detection and control performance generally arise from the fact that faster detection requires higher variability in the variables used for detection whereas higher variability generally translates into poorer control associated to lower product uniformity or higher wear of actuators. Bin Shams *et al*., (2010a, Chapter 4) have recently addressed the interaction between control and fault detection for a simple chemical process by using linear transfer function models identified in closed loop operation. However, the previous approach by Bin Shams *et al*., (2010a, Chapter 4) cannot be easily generalized to a more complex process such as the TEP since it requires the identification of a large number of models and since it is limited to linear systems only.

The current study proposes a numerical simulation based approach for finding an optimal trade-off between control and fault detection. The tuning parameters of the controllers involving the variables used for fault detection are used as optimization variables. The fault detection algorithm is based on CUSUM statistics. Since the simulations are conducted with the full nonlinear dynamic model of the process, assumptions of linearity are not needed as in the previous study of Bin Shams *et al*., (2010a, Chapter 4). Moreover, the tedious identification of transfer functions required by the previous study, are avoided in the current approach.

The chapter is organised as follows: a description of the CUSUM statistics and the metric used to gauge fault observability are given in section 5.3. Section 5.4 illustrates the observability problem through three particular faults in the Tennessee Eastman Process and demonstrates the ability of the CUSUM based statistics in tackling these faults. In section 5.5, the interaction between feedback control and fault detection is discussed. The methodology for simultaneously

optimizing fault detection and feedback control to alleviate the problem of observability is given in Section 5.6. Conclusions are given in Section 5.7.

5.3 Preliminaries

5.3.1 The Cumulative sum (CUSUM) based control charts

Shewhart charts are often used for assessment of faults (MacGregor and Kourti, 1995; Montgomery, 1997). A key disadvantage with these statistics is that they only use current time-interval information while not accounting for the entire time history. Hence, those charts are relatively insensitive to small shifts in the process variables for small signal to noise ratio. These shortcomings have motivated the use of other alternatives such as the univariate or the multivariate version of the CUSUM based charts (Bin Shams *et al*., 2010d, Appendix B). Three types of statistical charts are used in this paper. Specifically, location cumulative sum (LCS), scale cumulative sum (SCS) and the Hotelling's T^2. The LCS and SCS algorithms are examples of univariate statistics while the Hotelling's T^2 is a multivariate statistic. Both the LCS and SCS are performed using the following two statistics, corresponding to a two sided hypothesis test (Montgomery, 1997)

$$C_i^+ = \max[0, C_{i-1}^+ + x_i - (\mu_{i.c} + K)] \tag{5.1}$$

$$C_i^- = \max[0, C_{i-1}^- + (\mu_{i.c} - K) - x_i] \tag{5.2}$$

$$C_0^+ = C_0^- = 0$$

where K, $\mu_{i..c}$, C_i^+ and C_i^- are the slack variable, the *in control* mean , the upper and the lower CUSUM statistics, respectively. The subscript (*i.c*) stands for the in-control state. The role of the slack variables is to introduce robustness to noise. At every new sample, the statistics' in equations; 5.1 and 5.2, result in the accumulations of small deviations in the mean (LCS) or

small changes in the variability (SCS). These accumulations are corrected using the slack variable and compared to zero using the (max) operation. When either one of the two statistics in equations; 5.1 and 5.2, exceed a threshold H, the process is considered to be out of control. Following their respective definitions, the LCS is especially effective for detecting changes in the average whereas the SCS is suitable for detecting changes in variability. Guidelines for the selection of K and H have been reported (Montgomery, 1997). Typically K is selected to be half of the expected shift in either μ or σ in standard deviation units. H is determined so that a prespecified average run length $ARL_{o.c}$, to be defined in the following section, is achieved. The latter is also determined in standard deviation units. It should be noticed that when using equations; 5.1 and 5.2, the LCS uses the original raw data x_i, whereas the SCS uses the following standardized quantity:

$$x_i = \frac{\sqrt{|x_i|} - 0.822}{0.349} \tag{5.3}$$

A derivation of the quantities in equation (5.3) is given in Appendix A. Although LCS and SCS can be applied to individual measurements, there are many situations in which a pooled representative statistic for more than one variable is necessary. This is especially important when it is desired to present the operators with compact information to simplify the monitoring activities for the process. For that purpose, when the monitored variables are normally and statistically independent, the Hotelling's T^2 can be used. The Hotelling's T^2 statistics and the upper and lower control limits are given by:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \, \mathbf{S}^{-1} \, (\mathbf{x} - \bar{\mathbf{x}}) \tag{5.4}$$

$$UCL = \frac{p(m+1)(m-1)}{(m^2 - mp)} . F_{\alpha, p, m-p} \tag{5.5}$$
$$LCL = 0$$

where p is the number of monitored variables, m is the total number of samples, $\mathbf{x}$ is the sample vector, $\bar{\mathbf{x}}$ is the in-control mean, S is the estimated covariance matrix and F is the critical value of F distribution at α significance level . In the current work the cumulative sum statistics' are combined together into one statistic, namely, the Hotelling's T^2 as described in the following section.

5.3.2 Multivariate CUSUM-based T^2

To simplify the presentation of detection data to plant personnel, it is often advantageous to summarize the information of several charts into one single chart. To that purposes, Bin Shams *et al*., (2010d, Appendix B) proposed the use of a combined version of the CUSUM algorithms as illustrated in Fig.5.1.

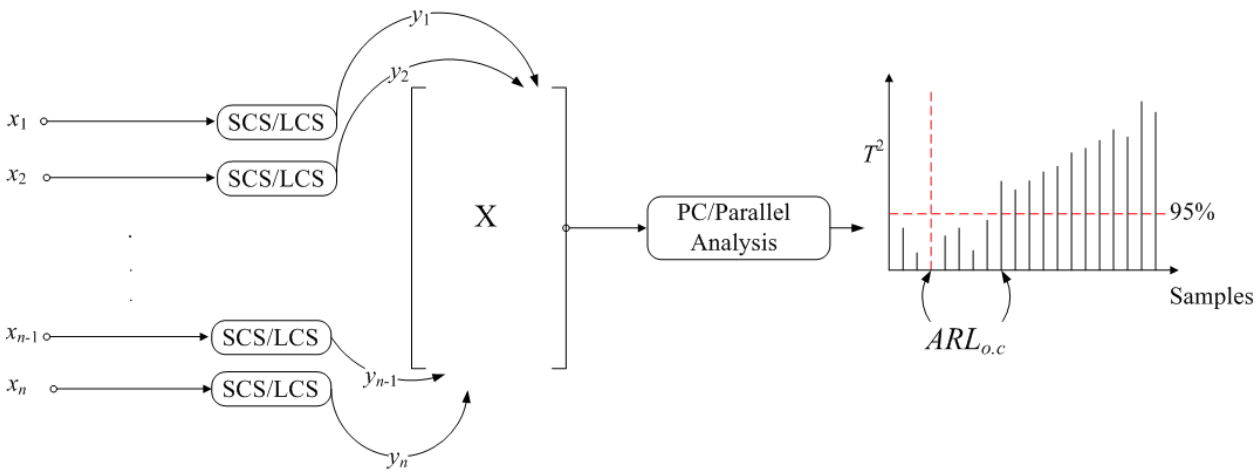


Fig.5.1 The Proposed CUSUM based T^2

The proposed CUSUM- based T^2 performs LCS and SCS on identified fault relevant variables x_i. Then, the transformed variables are stacked in $X \in R^{n \times m}$ matrix where n is the number of samples and m the number of the fault's relevant variables. Because of the interconnected nature of the chemical plants, process measurements are most likely cross

correlated. To account for any possible collinearity, the principal components of the X matrix are estimated using Principal Components Analysis (PCA) and only those representing the major variability are retained. The latter step can be determined with different tests (Chiang *et al*., 2001). In this work the Parallel Analysis method is used. The Parallel Analysis is an enhanced version of the Scree test. The reduction order is determined by comparing the Eigenvalues profile of the variables matrix $X \in R^{n \times m}$ to the Eigenvalues profile for a statistically independent matrix $Y \in R^{n \times m}$. The reduction order is the point at which the two profiles intersect. This is the point where significant process variation is separated from random noise and any possible linear cross correlation between the variables. If the candidate variables to be used for detection are found to be uncorrelated, they are directly used for monitoring through equation (5.4). However, when the variables are found to be correlated, the scores corresponding to the principal components of the X matrix are used instead of the original variables.

5.3.3 Out of Control Average Run Length ($ARL_{o.c}$) as an observability index

Observability of a fault is referred to as the ability to detect the fault from the chosen set of measurements. On the basis of a previous study by Bin Shams *et al*. (2010a, Chapter 4), the out-of-control Average Run Length ($ARL_{o.c}$) was proposed as a statistical measure to quantify the observability of faults when a statistical chart is used for monitoring. The subscript ($o.c$) stands for out-of-control. The $ARL_{o.c}$ is defined as the average number of points that must be sampled or plotted before the chart signals a violation of some pre-specified threshold (Montgomery, 1997). Most statistical monitoring techniques are based on the statistical hypothesis-testing principle that involve two types of errors, namely, type I and type II errors. A type I error occurs when a control chart indicates a fault in the absence of it, whereas a type II error occurs when a

control chart fails to identify the occurrence of a fault (Montgomery, 1997). Mathematically, $ARL_{o.c}$ is a function of the probability of type II error (β), that is

$$ARL_{o.c} = f(\beta) \tag{5.6}$$

For simple statistical charts such as the univariate Shewhart chart, $ARL_{o.c}$ has a closed form expression (Montgomery, 1997). However, for most of the statistical charts, numerical based approaches have been reported in the literature to estimate $ARL_{o.c}$ and they can be classified broadly into two categories: (1) Markov chain approach and (2) simulation of random realizations of the disturbances. The latter approach is adopted in the current study.

Due to their integrating nature, cumulative sum based techniques require some time before a fault can be detected, especially if the changes are very small. Accordingly, the $ARL_{o.c}$ is a suitable metric to quantify this expected delay in detection in terms of the number of sampling intervals that takes the response to surpass a pre-specified threshold. For example, if in response to a certain fault the estimated $ARL_{o.c}$= 1, the fault would be detected, on the average, after the first sample following the onset of the fault. On the other hand, when $ARL_{o.c}$= *infinity* or a very large number the fault is unobservable or it takes a long time to observe it.

Knowledge of the $ARL_{o.c}$ can be used to establish a frequency limit such as when the frequency of the fault is beyond this limit, it cannot be detected. To illustrate this property of the $ARL_{o.c}$, a simple example is given. Consider a unity gain process given by $y = x+n$, where y is the measured value to be used for detection, x is the actual fault and n is a white noise with μ=0 and $\sigma^2_{i.c}$=1. The magnitude of the signal y is equal or less than the noise level. Two square-wave like faults with period T=24 samples and T=3 samples, respectively are considered as candidate faults x to be monitored by the fault detection algorithm. These faults are shown in Fig. 5.2a and Fig. 5.3a, respectively.

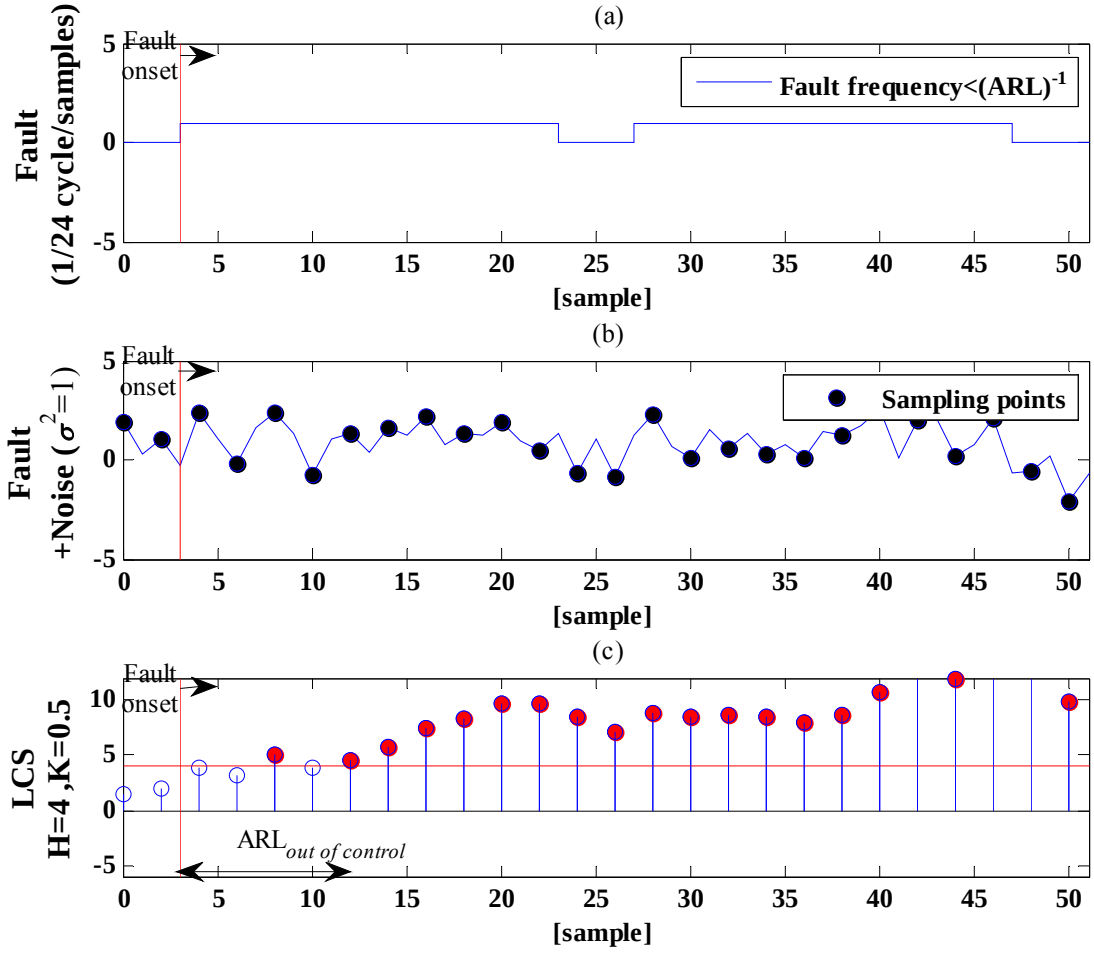


Fig.5.2 The dependency of the LCS $ARL_{o.c}$ on the fault frequency. (a) Fault with frequency smaller than $(ARL_{o.c})^{-1}$. (b) Noise added to the fault signal. (c) Monitoring using LCS, where dark circles depicts the faulty signals. $ARL_{o.c}$ ~ 8 samples (Montgomery, 1997)

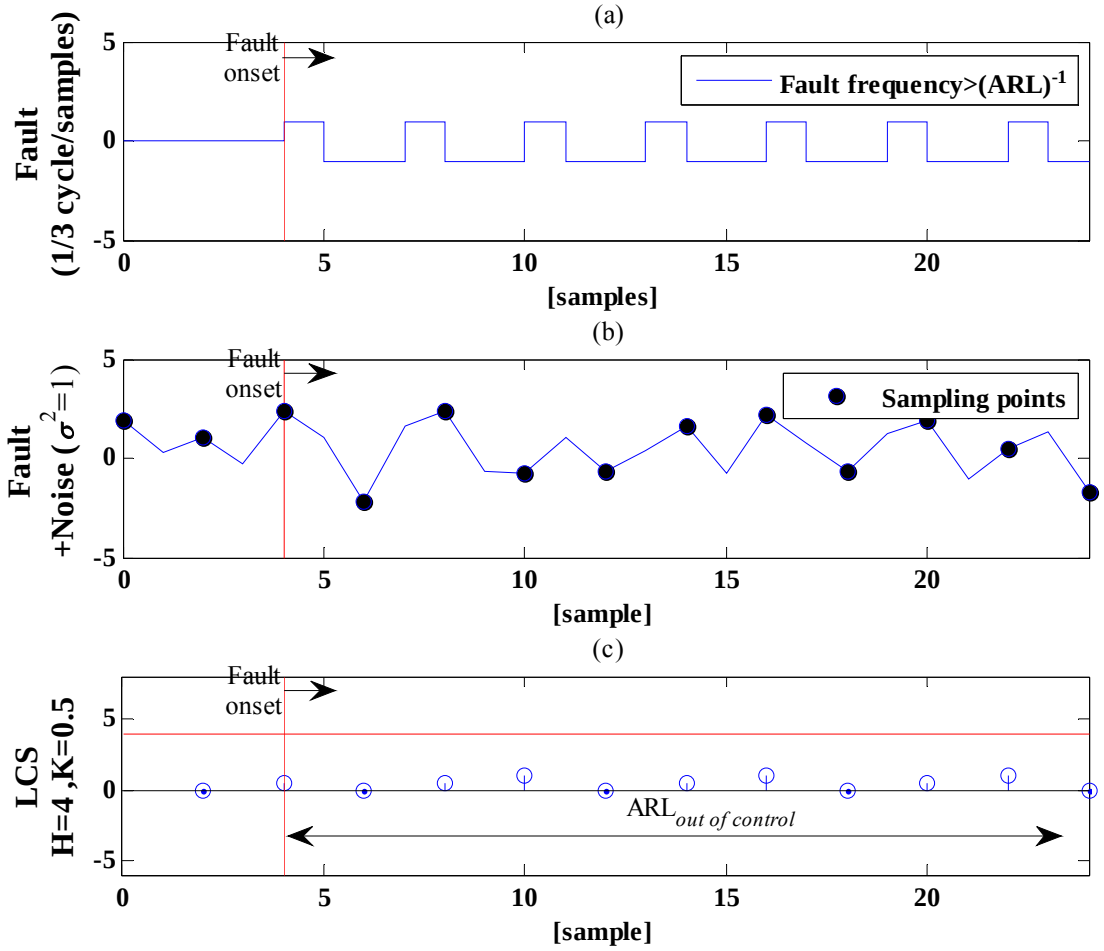


Fig.5.3 The dependency of the LCS $ARL_{o.c}$ on the fault frequency. (a) Fault with frequency larger than $(ARL_{o.c})^{-1}$. (b) Noise added to the fault signal. (c) Monitoring using LCS, where dark circles depicts the faulty signals. $ARL_{o.c}$ ~ 8 samples (Montgomery, 1997)

Following the addition of noise, the resulting measured values of y corresponding to these two faults are given in Fig. 5.2b and Fig. 5.3b, respectively, where the solid circles represent measurements obtained at discrete sampling intervals. For the implemented LCS, Montgomery (1997), gives the theoretical $ARL_{o.c}$= 8.38 when K=0.5, H=4, when $\sigma^2_{i.c}$ =1. Based on this value of $ARL_{o.c,}$ it is expected that square pulses that are of longer duration than the $ARL_{o.c}$=8.38 will exceed the threshold H shown by a horizontal solid line in Fig. 5.2c and therefore will be detected as shown in this figure. On the other hand, pulses of shorter duration than the calculated $ARL_{o.c}$ will not exceed the threshold H and therefore, they will not be detected as shown in Fig. 5.3.c. Accordingly, the observability of faults can be estimated from the relative values of the period of the square wave versus the estimated theoretical $ARL_{o.c}$ value (8.38 samples) or alternatively to the inverse of the period, i.e. the frequency. In that case, to assess observability the frequency of the fault can be compared with the *inverse* of the estimated theoretical $ARL_{o.c}$. Since it is common to characterize disturbances or faults by their frequency content, $(ARL_{o.c})^{-1}$ will be used within the optimization problem described later in the paper to quantify the faults' frequency bands for which these faults will remain unobservable by the detection algorithm.

5.4. Tennessee Eastman Process (TEP)

The problem of fault observability will be illustrated using an industrial process simulator referred to in the literature as the Tennessee Eastman Process (TEP). TEP has been proposed by Downs and Vogel (1993), and has been used as a benchmark problem in several studies to compare various control (Lyman and Georgakis, 1995, McAvoy and Ye, 1994; Ricker and Lee, 1995); and monitoring solutions (Chiang *et al*., 2001; Chiang *et al*., 2003; Lee *et al*., 2004; Cheng *et al*., 2010; Ding *et al*., 2009; Zhang *et al*., 2009; Ku *et al*., 1995). It consists of five

major unit operations, as shown in Fig. 5.4: reactor, condenser, compressor, separator and stripper. The process produces two liquid products (G and H) and one by-product (F) from four gaseous reactants (A, C, D, E) and an inert (B). Based on the required product mix and production rate, the plant can be operated according to six different modes of operation.

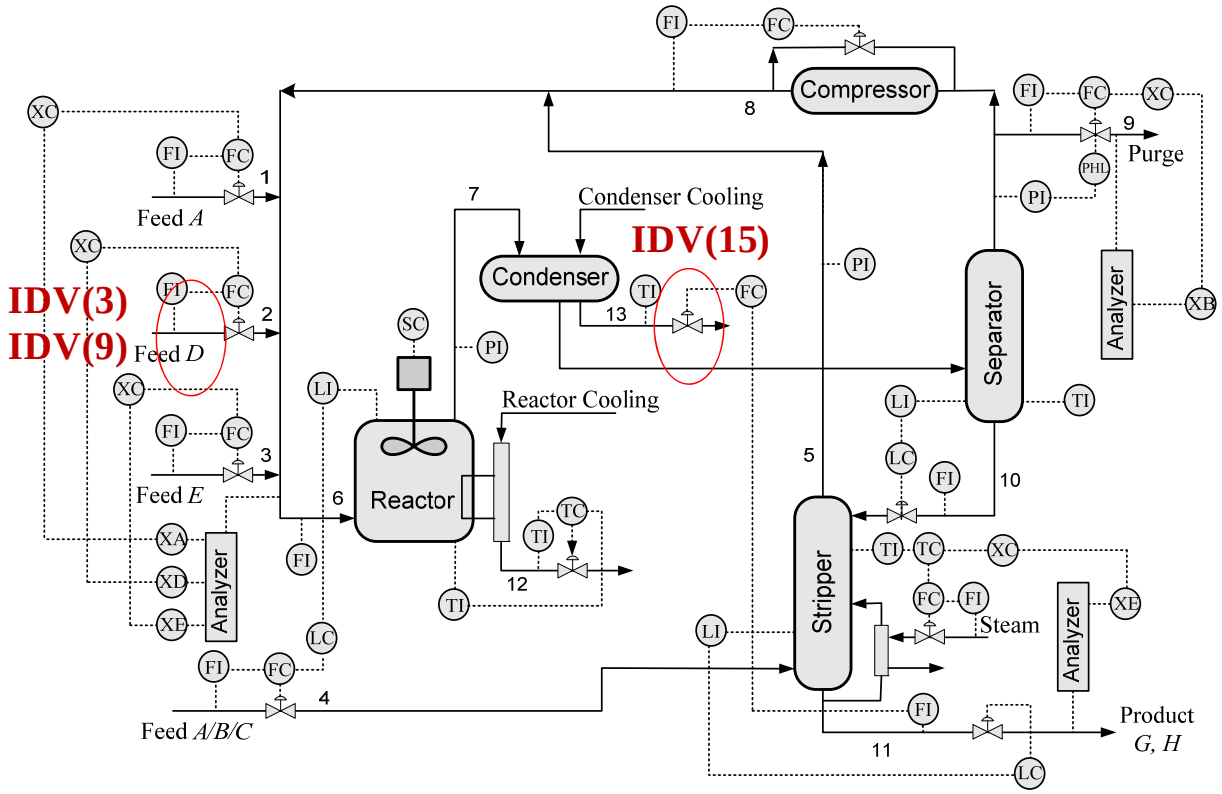


Fig.5.4 Tennessee Eastman process with the second control scheme described in (Lyman and Georgakis, 1995); the circles indicate the location of the three faults described in Table 5.1

The original open loop FORTRAN code was provided by Downs and Vogel (1993). The process is open loop unstable because of the exothermic reaction that takes place in the reactor; hence it cannot be operated in manual mode. Several decentralized control structures have been proposed for the TEP and the structure used by Lyman and Georgakis (1995) was used in this

work. Different monitoring techniques have been tested on a total of 15 particular faults defined for the TEP (Chiang *et al*., 2001; Chiang *et al*., 2003; Lee *et al*., 2004; Cheng *et al*., 2010; Ding *et al*., 2009; Zhang *et al*., 2009; Ku *et al*., 1995); and listed in Table 5.1.

Table.5.1 TEP faults (Unobservable faults are emphasized with bold font)

Faults	**Description**	**Nature**
IDV(0)	Normal operating condition	
IDV(1)	A/C feed ratio, B composition constant (stream 4)	Step
IDV(2)	B composition, A/C ratio constant (stream 4)	Step
IDV(3)	**D feed temperature**	**Step**
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss (stream 1)	Step
IDV(7)	C header pressure loss	Step
IDV(8)	A,B,C feed composition (stream 4)	Random variation
IDV(9)	**D feed temperature**	**Random variation**
IDV(10)	C feed temperature	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	**Condenser cooling water valve**	**Sticking**

These techniques have shown different capabilities in detecting the majority of these faults. However, all of these previously reported techniques have consistently failed in detecting three particular faults listed in Table 5.1 and referred heretofore as faults IDV(3), IDV(9) and IDV(15).

The lack of observability associated with these faults has been generally attributed to the statistically insignificant changes occurring in the monitored variables following the onset of the 3 faults. For fairness, it should be stressed that in most of the reported work, the detection was based on small number of current and past time intervals thus the entire time histories of the measurements were not considered for detection as done in the CUSUM calculations used in the current study. However, the fact remains that these faults have not been detected in previous studies whereas they may be significant negative economic or operational effects from their occurrence. Thus it is still very relevant to attempt to detect them. The next section describes how these faults can be detected with CUSUM based techniques.

5.4.1 The CUSUM based charting approach for faults IDV (3), IDV (9) and IDV (15)

Bin Shams *et al*., (2010d, Appendix B) summarized most of the monitoring solutions proposed in the literature for the Tennessee Eastman Process faults. The inability of previous techniques to detect the 3 faults emphasized in Table 1, namely, IDV (3), IDV (9) and IDV (15), motivates the use of the cumulative sum based measures. Initially, a Multivariate Cumulative Sum (MCUSUM) (MacGregor and Kourti, 1995) that used all of the available TEP measurements was used for detection but it was found to be unable to detect these three faults. It was hypothesized that this inability was due to the occurrence of large amount of noise

combined with the high correlation among process variables. Thus, the multivariate CUSUM based statistics presented in section 5.3.2 was used and it was successful in detecting the three faults (Bin Shams *et al*., 2010d, Appendix B). In principle all the monitored variables could be used for detection. However, it was found that due to small signal to noise ratio in most variables combined with the fact that the variables are highly correlated in a nonlinear fashion, using all the variables did not resulted in improved detection of the 3 faults.

Although for the TEP case it is relatively easy to identify the variables that are most relevant for each fault, the identification of these variables for other large scale systems solely based on process knowledge may not be always a trivial task. To do this selection it is proposed to use contribution plots (Miller *et al*., 1998) based on CUSUM based T^2 for all available measurements, to systematically identify those variables. For that purpose, an augmented matrix is used that contains both the LCS and SCS of all the variables, i.e. a total of 104 columns (2 × 52 measurements). After applying PCA to this augmented matrix, the contribution plots can be obtained. For example, the top and bottom plots in Fig.5.5 show that the dominant contributions correspond to the LCS of XMV [10] (reactor cooling water flow) and to the SCS of XMEAS [21] (reactor cooling water exit temperature) for faults IDV (3) and fault IDV (9) respectively. The top and bottom plots in Fig.5.5 show the contribution of XMV [10] and XMEAS [21] for IDV (3) and IDV (9), respectively. The horizontal axis contains numbers from 1 to 104 corresponding to the 52 LCS' and 52 SCS' values used for the calculations of the contributions. The large values in the plot correspond to the dominant contributions. It was found as expected that the most dominant variable for detection of fault 3 is an LCS indicator since the fault manifests itself as a change in bias whereas for fault 9 an SCS indicator is required since the fault involves a change in variability. Table 5.2 shows the fault-variables pairing for these three faults

as found from the contribution plots. Accordingly, the LCS algorithm was applied to XMV (10) for fault 3 whereas the SCS algorithm was applied to XMEAS (21) for fault 9 and XMV (11) for fault 15, and then the corresponding cumulative sums were used to drive the Hotelling T^2 statistics defined in equation (5.4) where **x** is a vector sample composed of the 3 corresponding cumulative sums. Although the T^2 statistic based on the augmented matrix could be used within the proposed optimization based methodology, for simplicity this has not been done since the matrix has to be inverted at each optimization step resulting in significantly higher computational effort as compared to the case where the most relevant variables for each particular fault are used. The use of three separate control charts was appropriate for monitoring the 3 faults; however, it is often convenient for practical purposes as mentioned above to monitor the process with a smaller number of charts. For example, Fig.5.6 and Fig.5.7 depict the ability of the proposed CUSUM based T^2 statistics in observing IDV (9) and simultaneous occurrence of IDV (3) and IDV (15), respectively. Table 5.3 gives the estimated $ARL_{o.c}$ values for the three faults using the univariate and multivariate CUSUM based statistics (Bin Shams *et al*., 2010d, Appendix B). The calculated $ARL_{o.c}$ is only an approximation since a more precise estimate of the $ARL_{o.c}$ requires averaging over a large number of noise realizations as done later in the case study. A problem associated with these results is the relatively long periods of time required to detect the occurrence of the faults. The immediate implication is that only faults that are of longer durations than the corresponding $ARL_{o.c}$ values, or alternatively of smaller frequency than $1/ ARL_{o.c}$, can be detected using the CUSUM based statistics as explained in the previous section. Thus, faults with duration shorter than the $ARL_{o.c}$ would go undetected and a cost may be associated with this lack of observability. To partially address this problem a method is

proposed in the next section to retune the controllers so as to reduce the detection times while maintaining suitable control of the process.

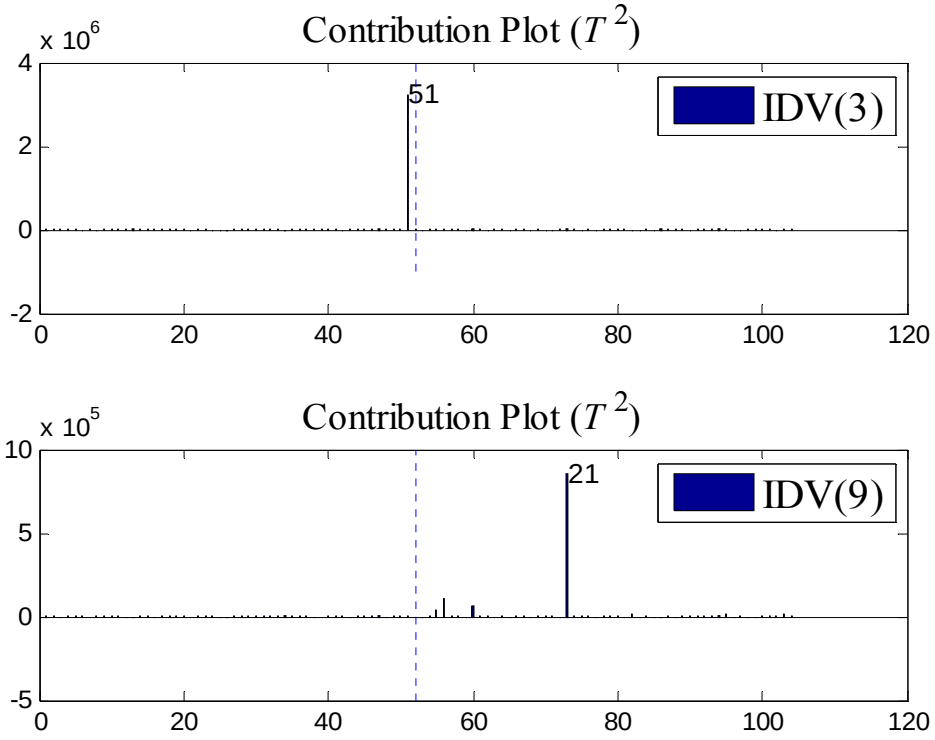


Fig. 5.5 The contributions of all of the variables to CUSUM based T^2. Top plot: IDV (3); bottom plot: IDV (9). The horizontal axis includes a total of 104 variables corresponding to the LCS's and the SCS's of all 52 measurements respectively. The vertical dotted line separates the LCS and SCS corresponding sets.

Table.5.2 The unobservable faults/process variables pairing

Faults	Measurements*	Description
IDV(3)	XMV(10)	Reactor cooling water flow
IDV(9)	XMEAS(21)	Reactor cooling outlet temp.
IDV(15)	XMV(11)	Condenser cooling water flow

*The variable measurements as appeared in (Down and Vogel, 1993)

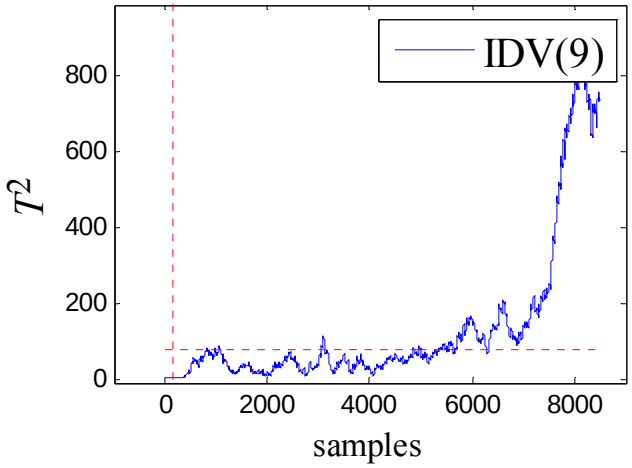


Fig. 5.6 The Hotelling'T^2 for IDV (9); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

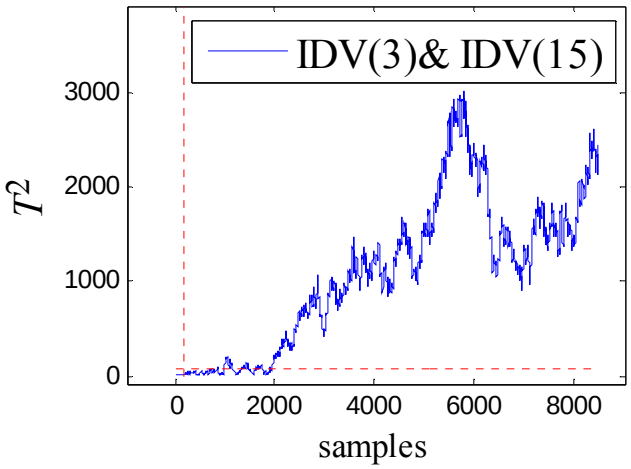


Fig. 5.7 The Hotelling'T^2 for the simultaneous occurrence of IDV (3) and IDV (15); horizontal and vertical lines represent the statistical limit and the fault onset, respectively

Table.5.3 The estimated $ARL_{o.c}$ for the LSC, SCS and T^2

Fault	Statistics	*$ARL_{o.c}$ (hr)
IDV(3)	LCS	127.05
IDV(9)	SCS	8.20
IDV(15)	SCS	41.00
IDV(3)	T^2	102.40
IDV(9)	T^2	276.05
IDV(15)	T^2	89.65
IDV(3) & IDV(15)	T^2	41.30

*All $ARL_{o.c}$ are calculated from after onset of the faults (i.e. after 8 hours)

5.5 An approach for finding an optimal trade-off between fault detection and control

To shorten the detection time given by the $ARL_{o.c}$, it is proposed to retune the corresponding controller. To illustrate the interaction between control and fault detection, IDV (15) and its corresponding controller are considered, i.e. the controller that manipulates the condenser's cooling water valve XMV (11). The $ARL_{o.c}$ corresponding to fault IDV (15) and the variability in XMV (11) are plotted as a function of the controller proportional gain (K) and shown in Fig.5.8.

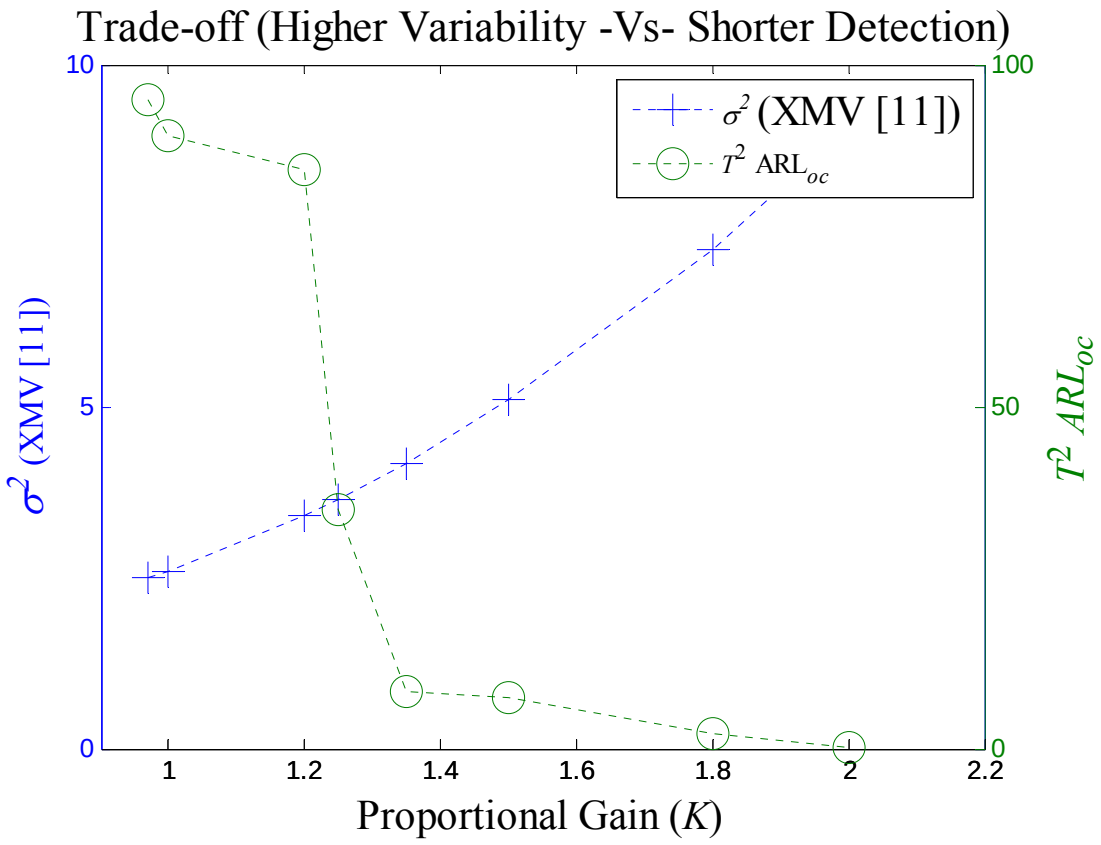


Fig. 5.8 The change in variability and the T^2-$ARL_{o.c}$ as a function of the condenser controller's gain, XMV [11].

As can be seen from Fig.5.8, there is a significant interaction between the control, as manifested by the variability in the manipulated variable, and the detection scheme, e.g. the multivariate CUSUM based T^2 statistics in this case. Tuning of the controller significantly reduces the $ARL_{o.c}$ that would be required to observe IDV (15), but at the expense of significant degradation in controller's performance as shown by the increased variability in the manipulated variable value. This variability may translate into significant wear of the corresponding valve. Thus, there is a motivation to seek for a trade-off between the costs associated with detection speed and closed loop performance. An objective function that represents such trade-off can be formulated as follows:

$$\min_{\lambda} J = a_1\gamma_1 + a_2\gamma_2 + a_3\gamma_3 \tag{5.7}$$

$$\gamma_i = \sigma^2_{y_i} = \sum_{l_b}^{u_b} (y - \bar{y})^2 / (u_b - l_b) \tag{5.8}$$

; where y_i represent the controlled variables, the manipulated variables and the rate change of the manipulated variables for i =1, 2 and 3, respectively, and the corresponding γ_i represent the variability associated with these variables as calculated by equation (5.8). A key assumption in this work is that, if a fault can be detected, the faulty situation can be immediately corrected thus there is no cost associated with observable faults. Thus, costs are assumed to be associated to faults that cannot be observed and therefore persist for a long time without being fixed. Accordingly, J is formulated to include only the costs associated with undetected faults, which as per the discussion in the previous section, are those faults with frequencies larger than $(ARL_{o.c})^{-1}$. The tuning parameters of the controller are given by λ and they are the decision variables for the optimization problem. The lower and upper bounds of summation l_b and u_b are the time samples at the onset of the fault and at the end of simulation, respectively.

The coefficients in the objective function a_i represent actual economic costs associated to variability in manipulated and controlled variables and to the rate of change in manipulated variable. These costs are problem specific and they should be chosen based on knowledge of the process. For instance, the cost of variability in manipulated variables (second term in the RHS of equation 5.7) may be related to cost of utilities e.g. changes in steam, cooling water flow etc. On the other hand the cost assigned to the rate of change of a valve (third term in the RHS of equation 5.7) could be obtained from the cost of the valve and from the expected valve life for a specific variability level.

Although the objective function above was formulated for a particular fault and a particular controller, it can be generalized to the case where multiple faults may occurred, either sequentially one after the other or simultaneously, and where multiple controllers are retuned to optimize the problem. Accordingly two objective functions will be considered. The first one assumes that the faults occur simultaneously and accordingly the cost function is given as follows:

$$\min_{\lambda_1,\lambda_2,...,\lambda_k} (J \mid f_1, f_2, ..., f_k) = a_1\gamma_1 + a_2\gamma_2 + a_3\gamma_3 \tag{5.9}$$

The second proposed objective function assumes that the faults occur sequentially one after another and accordingly it is defined as follows:

$$\min_{\lambda_1,\lambda_2,...,\lambda_k} J_{total} = J_{f_1} + J_{f_2} + ... + J_{f_k} \tag{5.10}$$

In the latter, faults occur sequentially and it is assumed that one fault occurs for some period of time after which it is removed before another fault occurs.

In summary, the optimization problem consists of minimizing the objective functions given by either equations 5.7 for one fault and 5.9 or 5.10 for multiple faults, with respect to the tuning parameters of the controllers. These minimizations are done based on numerical simulations of

the process under closed loop control for different realizations of the faults. To ensure that the cost function is evaluated for faults that cannot be observed, the frequency band of the faults introduced in the simulations in the form of an RBS (random binary signal) is given between a lower limit of $(ARL_{o.c})^{-1}$ to an upper limit of Nyquist frequency (ω_u), i.e. $0.5/T$ where T is the sampling interval used in the numerical simulations. To reduce the computation time within the proposed scheme, the numerical simulations of the TEP model are performed in FORTRAN. However, the optimization and the processing of the results are done in MATLAB®. To achieve this, a program was designed to allow the two platforms to communicate. One of the key sources of computational burden in the optimization problem is the calculation of the $ARL_{o.c}$ which requires conducting repeated simulations of the closed loop model of a particular fault for different noise realizations. However, since these calculations are independent from each other, it was possible to exploit the parallel computation capabilities of MATLAB® to speed up this task.

The optimization problem is solved as per the following steps:

1. The controller parameters (λ_k) and the random number generator's seed are initialized.
2. The IDV (j) is set and the Tennessee Eastman differential/algebraic equations are solved (FORTRAN), where $j \in \{3,9,15\}$, the three unobservable faults.
3. The simulated data is retrieved in MATLAB® where CUSUM based-T^2 is calculated and the out-of-control run length ($RL_{o.c}$) for a single realization is estimated.
4. The seed of the random number generator is changed and steps two and 3 are repeated until the maximum number of runs is reached.
5. $ARL_{o.c}$ is calculated.
6. A random binary sequence (RBS) is designed with frequency content $[(ARL_{o.c})^{-1}\ \omega_u]$.

7. The RBS_j signal replaces the IDV (j) and the Tennessee Eastman process is re-simulated in (FORTRAN).

8. Equation (5.8) is evaluated, for a given RBSj and λ_k and equation (5.9) is minimized until convergence. A detailed flow chart summarising the abovementioned steps is given in Fig. 5.9.

It should be noticed that the methodology proposed above is based on simulations of a first principle model of the system. When such a model is not available, the method could still be applied by using empirical models of the different units that may be identified from input output data plant data. Then, these empirical models could be interconnected with the corresponding controller models by using appropriate software, e.g. Simulink® (MathWorks) to provide a closed loop plant simulator for which the method above could be readily applied.

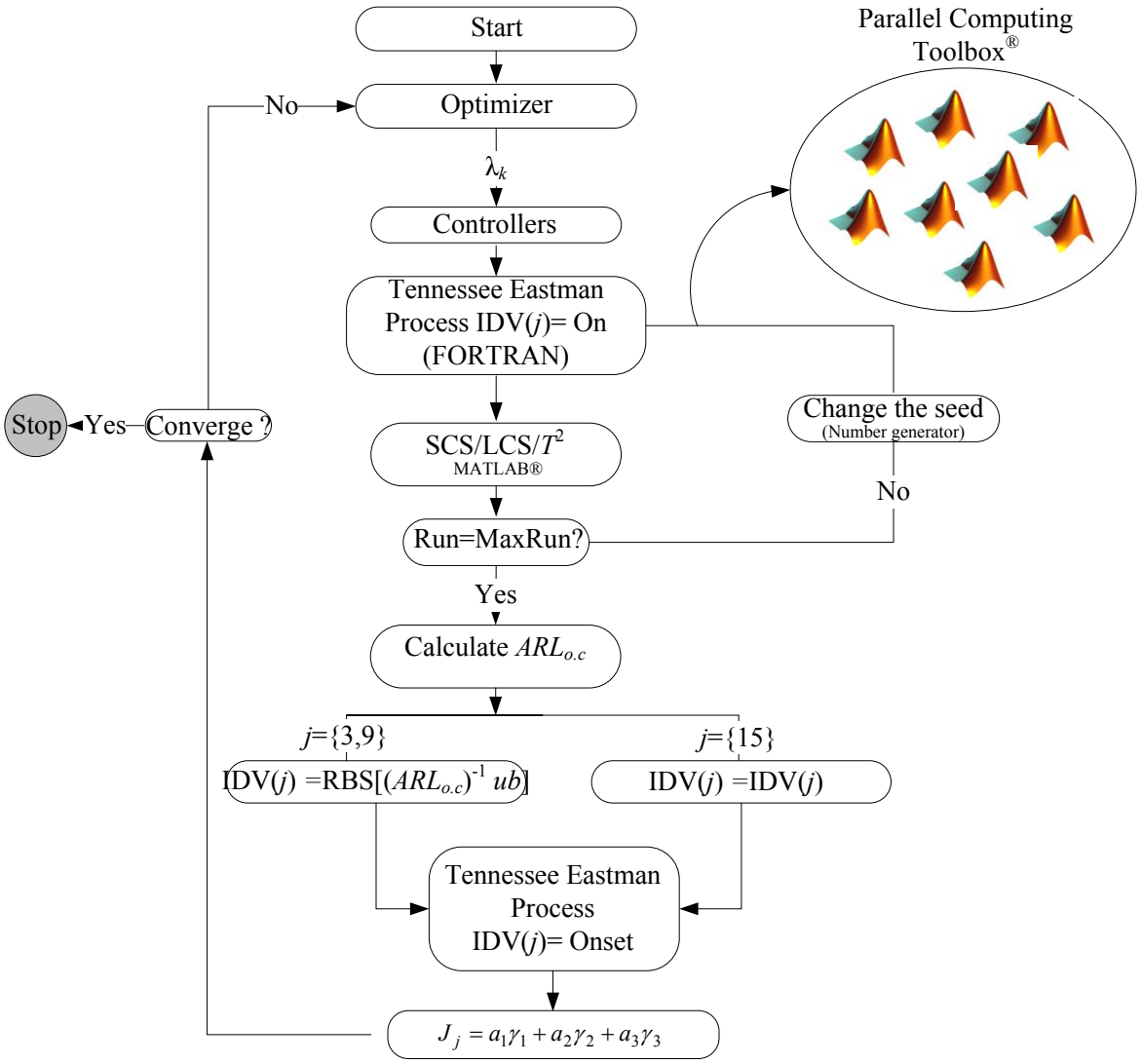


Fig. 5.9 The flow chart of solving the proposed dynamic optimization problem

5.6 Results and discussions

5.6.1 Individual faults case

First, the 3 faults are assumed to occur one at a time. For each one of these faults a single controller is optimized. This controller is selected as the one that involves the variable that is deemed most relevant for that particular fault as specified in Table 5.2. More specifically, the reactor cooling water valve (slave controller) is identified as the fault relevant controller for IDV (3) and IDV (9), while the condenser cooling water valve is identified for IDV (15), see Fig. 5.10. The algorithm explained above and depicted in Fig.5.9 is implemented for each of the three faults. For each fault, many different initial values have been assumed to avoid local optimal solutions. The results of the optimization for each fault in terms of the optimal tuning parameters and the value of the cost function at the optimum are given in Table 5.4. For these results, a_i in equation 5.7 are set to equal to one, for i=1, 2 and 3, respectively and a sampling frequency equal to (1/180) Hz is chosen. Also, in Table 5.4 the tuning parameters provided by Lyman and Georgakis, (1995) and the cost obtained with these parameters are presented for comparison purposes. As can be seen from the table, the $ARL_{o.c}$ has been significantly reduced for each of the three faults. For example, for IDV (15), stiction in the condenser's cooling water valve, and by using the proposed scheme, $ARL_{o.c}$ has been reduced from 89.65 hours to 40.59 hours and the associated cost, has been reduced from 7.34 to 0.87. On the other hand, for IDV (3), the proposed tuning schemed was able to shorten the $ARL_{o.c}$ by 2.69 hours and the cost by approximately 80% compared with the cost obtained with the tuning parameters proposed by Lyman and Georgakis, (1995). Similar trend has been obtained for IDV (9).

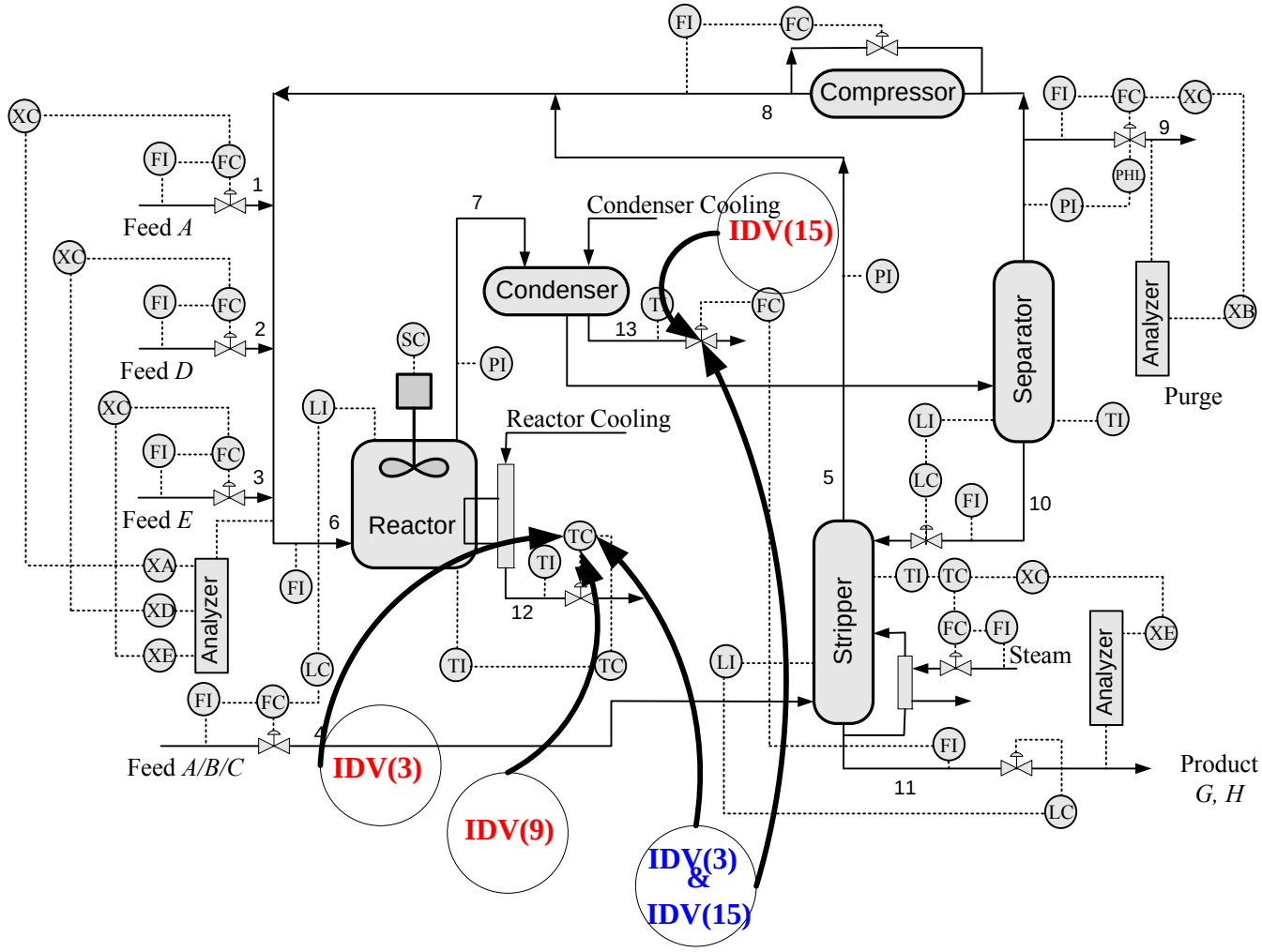


Fig.5.10 The fault relevant controllers, for IDV (3), IDV (9), IDV (15) and IDV (3) & IDV (15)

A close look at Table.5.4 reveals that the controllers are tuned less aggressively than the controllers used by Lyman and Georgakis (1995) as evidenced by the consistently smaller proportional gains resulting from the optimization. This is expected since the LCS and SCS operations, used to drive the CUSUM based T^2, need a relatively slow convergence of manipulated variables to their final steady state solution to achieve the necessary detection. By detuning the corresponding controllers, the manipulated variables converge more slowly to their steady state values thus providing the necessary variability for detection but at the expense of controller performance.

The proposed algorithm is implemented on 8 Intel® Xeon® processors, each of 2 GHz and total of 12.0 GB RAM. The parallel computing paradigm induced in the proposed algorithm exploits these 8 processors and significant reduction in the computation time is noticed. The CPU time required to solve each problem was approximately 3 hours on the server specified above where most of the computational burden was due to the $ARL_{o.c}$ estimation.

Table.5.4 The tuning parameters and the $ARL_{o.c}$ obtained from implementing the proposed algorithm (Fig.5.9)

Fault (#)	**Tuning scheme**	K_c **(Unit of CV)$^{-1}$**	τ_I **(hr)**	$ARL_{o.c}$ **(hr)**	**Cost**
IDV(3) (Reactor Controller)	Lyman and Georgakis,1995	-1.560	0.403	102.40	0.973
	Proposed Scheme	-0.265	2.437	**99.71**	**0.161**
IDV(9) (Reactor Controller)	Lyman and Georgakis,1995	-1.560	0.403	273.05	1.001
	Proposed Scheme	-0.269	0.711	**115.65**	**0.225**
IDV(15) (Condenser Controller)	Lyman and Georgakis,1995	1.090	0.722	89.65	7.343
	Proposed Scheme	0.236	0.914	**40.59**	**0.876**

5.6.2 Simultaneous faults case

Since simultaneous faults are common in industrial systems, the proposed methodology can be extended to account for such event. For illustration purposes, the simultaneous occurrence of IDV (3) and IDV (15) is considered. IDV (3) and IDV (15) are two faults of different nature and their combined effect is not necessarily the superposition of the two inputs, in particular when a nonlinear system is considered. For this case i.e. IDV (3) and IDV (15), the reactor cooling and the condenser cooling controllers are the two controllers that could be retuned to achieve an optimal trade-off between detection and control performance. Since there are two different controllers that involve the variables used for monitoring the faults, there are a total of 3 different options corresponding to the cases where either one of the two controllers are retuned or when the two controllers are retuned simultaneously.

For each one of these 3 options, two costs are considered as per equations 5.9 and 5.10 corresponding to the faults occurring either simultaneously or sequentially. The combination of 3 possible tuning options with two possible optimization costs results in a total of 6 cases as shown in Table 5.5. Comparisons of the costs obtained with the controllers' settings given by Lyman and Georgakis, (1995) are also presented in Table 5.5. It is clear from the Table 5.5 that the optimization procedure is leading to significant reductions in cost up to an order of magnitude as compared with the costs obtained with the controllers' settings of Lyman and Georgakis. For example for the case that the reactor cooling controller is tuned using equation 5.9, the $ARL_{o.c}$ is reduced by approximately 54.36% and the corresponding cost is reduced from 1.479 to 0.163. The second case is when both controllers, i.e. reactor cooling and the condenser cooling controllers are tuned simultaneously using equation 5.9. In the latter case, the $ARL_{o.c}$ is reduced form 41.30 to 12.93 hrs and the associated cost is decreased form 9.26 to 7.27. It should

be noticed that the cost was defined only in terms of the variables used for fault monitoring and control. Extending the proposed methodology to include the overall costs of the process is possible but it significantly increases the computational burden thus it has been left for future study.

Table.5.5 The tuning parameters and the $ARL_{o.c}$ results from implementing the proposed algorithm (Fig 5.9). The simultaneous faults case i.e. IDV (3) & IDV (15)

Cost function	**Controller(s)**	***Controller Parameters (λ)**	$ARL_{o.c}$ **(hr)**	**Cost**	**Cost (Leyman and Georgakis.,1995)** $ARL_{o.c}$**= 41.3 hr**
Eq.(9)	Reactor cooling water valve	[-0.269 , 0.679]	18.85	0.163	1.479
Eq.(9)	Condenser cooling water valve	[0.026 , 1.062]	110.99	0.190	2.007
Eq.(9)	Reactor & Condenser cooling water valves	[-1.117, 0.448 0.905 , 0.787]	12.93	7.268	9.259
Eq.(10)	Reactor cooling water valve	[-0.265 , 0.676]	97.64	0.326	3.685
Eq.(10)	Condenser cooling water valve	[0.235, 0.876]	118.43	1.574	3.639
Eq.(10)	Reactor & Condenser cooling water valves	[-1.138, 0.541 0.659, 0.881]	115.91	10.469	18.680

*λ is of size 2 when single PI controller is used and 4 when two PI controllers are considered. In the latter, $\lambda = [\lambda_1\ \lambda_2]$.

5.7 Conclusions

CUSUM based statistics combined with Hotelling's T^2 charting is used to detect faults in a chemical process. This method was successful for detecting three faults in the Tennessee Eastman problem that were impossible to observe with other previously applied methods. However, long detection delays, as quantified by the $ARL_{o.c}$ are noticed. To speedup detection, the controllers involving the variables used for detection are re-tuned to achieve an optimal trade-off between fault detection and control performance. For this purpose, an optimization problem is formulated. This optimization involves the minimization of costs due to the

occurrence of faults that cannot be detected and control variability with respect to the tuning parameters of controllers that involve variables used for fault detection. To quantify the limit of fault observability, the $ARL_{o.c}$ is used. Both the individual and the simultaneous occurrence of faults are considered. The proposed objective function minimizes the cost associated with the lack of detection while penalizing any excessive variability. It is shown that the proposed methodology results in significant reductions of costs when compared with the costs obtained with controllers' setting previously reported in the literature for the process under study.

Chapter 6

Concluding Remarks and Future Work

6.1 Overview

The increasing size and complexity of modern process plants has made automation essential for their successful operation. Automated fault detection and diagnosis strategies are a key subset of automation strategies of chemical processes due to their potential for providing safer and more profitable operations. As mentioned in the literature review, the fault detection and diagnosis algorithms are broadly classified according to the type of models that are used, namely first-principles, empirical or a combination of both. Empirical model based methods were adopted in this project. First principles based methods were not considered in this project. While they are very well suited for well-defined small systems for which it is relatively easy to formulate an accurate first-principles based model they are much more difficult to develop for larger scale systems like the Tennessee Eastman Process. Thus, empirical model based techniques were preferred for this project especially when dealing with large processes.

The problem of fault observability has been identified as a key factor for the prompt detection and the unambiguous diagnosis of faults when empirical based models are used. This project first addresses the problem of fault observability through two main approaches: (1) the interpretation of the notion of fault observability, including a tool to quantify it i.e. $ARL_{o.c}$, and the development of new fault detection and diagnosis strategies that permit the detection of faults which have not been possible to detect by other techniques. These strategies are illustrated through a case study involving the Tennessee Eastman Process which has been used in the past as a benchmark for comparing fault detection and diagnosis techniques (Downs and Vogel,

1993); (2) the development of an optimization based strategy that takes into consideration the key factors associated with the fault observability problem including the economical impact of lack of fault observability. The following is a summary of the contributions of this project and some recommendations for the future work.

6.2 Concluding Remarks

When a fault occurs, the objective is to detect and isolate it as promptly as possible. The notion of fault observability is interpreted in this work as the time elapsed between the onset of the fault and the time of its detection. Three faults of the Tennessee Eastman Process have been reported in the past to be impossible to detect or isolate by previously reported techniques. The difficulty in detecting these faults stems from the small signal to noise ratio associated with the process variables that are monitored to detect these faults. In the present work the lack of observability has been quantified using the $ARL_{o.c}$. Accordingly, small $ARL_{o.c}$ are associated with easy to detect faults, whereas faults that are difficult to detect, such as the 3 faults considered for the Tennessee Eastman Process, have very large $ARL_{o.c}$ associated with their detection. The careful analysis of these faults has lead to the development of a novel fault detection and diagnosis strategy referred to as the CUSUM based PCA detection and diagnosis algorithm. The ability of the CUSUM operation to incorporate the history of process measurements is combined with the capability of PCA to simultaneously handle the correlation and filter the noise associated with the process measurements. The CUSUM based PCA detector was successful in both detecting the faults in question, although with some delay.

For the purpose of isolating the root cause of a particular fault, a bank of CUSUM based models was developed to precisely diagnose the root cause of these faults.

It was also recognized in this work that there is an economic cost associated with the resulting lack of observability. For instance, if the relative magnitude of the fault is very small and its impact on operation is minimal, it may not be necessary to detect and diagnose it. On the other hand, if the impact of the fault on the economics of the plant is significant, then it is important to try to detect, diagnose and eventually eliminate its cause. At the same time, since the variables used for fault detection and diagnosis are either inputs or outputs in feedback loops, it was recognized that there is a tradeoff between fault observability and closed loop performance. In other words, the implemented feedback control scheme can provide good closed loop performance but at the same time it can disguise the fault's symptoms that are used by the algorithms for detection or isolation.

This lead to the second part of the work where these concepts have been incorporated within an optimization based strategy that can find controllers that achieve an optimal tradeoff between prompt detection (fault detection objective) and controller performance (feedback control objective) while minimizing the cost associated with faults that cannot be observed. The proposed optimization strategy was tested with different fault/signal to noise ratios, different fault frequencies and with different standard univariate and multivariate statistical detections schemes. The methodology proposed in chapter 4 is equally applicable to any standard statistical fault detectors (univariate or multivariate). Both individual and simultaneous occurrences of faults have also been addressed within the proposed approach.

The proposed optimization in chapter 4 has been developed using a single unit operation, namely, continuous stirred tank reactor (CSTR) whereas most of the industrial plants consist of many interconnected unit operations. Accordingly, it was necessary to evaluate the applicability of the propose methodology on a large scale industrial problem. This motivated the last part of

this project, where the proposed tuning method was extended to the Tennessee Eastman Process which involves several interconnected units. The proposed tuning method was successful in finding significant tradeoffs between detection and closed loop performance.

The need for detuning the controller to detect a fault at the expense of increase variability greatly depends on the situation at hand. If the process is affected by faults that can be easily observed, then identifying and eliminating the root cause resolves the problem in question and no further detuning of the fault relevant controller may be required. On the other hand, some faults are extremely difficult to observe but nevertheless they may be persistent and they may have a significant economic impact on the process. Moreover, this sort of unobservable faults may be recurrent, e.g. valve stiction is a recurring fault in chemical and petrochemical plants. For these types of scenarios, the proposed optimization methodology is justified.

6.3 Recommendations for future works

6.3.1 The effect of nonlinearity on fault observability

(1) In most chemical and biological systems, the cross-correlation among the process variables is of a nonlinear nature. However, the majority of the fault detection and diagnosis algorithms used in this work are linear.

For example, when a fault occurs; e.g. control valve stiction, the mathematical relation between the controller's process variable and the manipulated variable is highly nonlinear. It is then expected that algorithms that are based on the assumption of linearity would perform poorly in detecting these faults, hence resulting in a lack of observability. Two main reasons can be proposed in defense of linear techniques: (a) since plants are generally operated in closed-loop;

in the presence of a fault, process variables generally remain in a close proximity of a particular steady state operating condition. (b) There is a relative complexity associated with the training of the nonlinear fault detection and diagnosis algorithms. However, in cases where the process cannot be maintained close to a fixed point of operation or in cases where the nominal operating point is changed frequently, nonlinear techniques will be required. A new nonlinear PCA called Kernel Principal Component Analysis (KPCA) has been proposed (Scholkopf, 1998). In comparison with other nonlinear PCA based fault detection and diagnosis algorithms, KPCA has the advantage that it does not involve any kind of nonlinear optimization, e.g. such as the optimization needed for finding the weights in Artificial Neural Networks (ANN). In fact it only requires standard linear algebra techniques such as the standard static PCA. However, two main obstacles hinder the applicability of the KPCA for large scale problems. First, the complexity of Gram matrix calculation required by the KPCA algorithm when the size of the problem increases (Scholkopf, 1998). Second, the unavailability of an identification procedure of the variables relevant for monitoring, such as the contribution plot, when KPCA based detection is used. An appropriate treatment the aforementioned shortcomings can facilitate the use of KPCA.

(2) The proposed methodology in chapter 4 was based on identifying several linear closed loop transfer functions. Due to the nonlinear characteristic behavior of the CSTR, a nonlinear transfer function could be more appropriate. In that case a robust control approach could be used to tackle the nonlinearity. This can be done by incorporating the nonlinear behavior as uncertainty supplementing a nominal linear model. Tools from robust control theory can then be used to solve the resulted optimization problem.

6.3.2 Optimal designed experiments for empirical based methods

In contrast with first principle-model-based techniques, empirical model-based methods lack information about causal relations between input faults and the outputs. Accordingly, a more systematic way to incorporate prior knowledge about the system into the correlation models is required. One option could be to use a hybrid modeling approach that combines simple mechanistic models about the system under consideration with the empirical model based approach. When such mechanistic models are not available, an optimal designed experiment can be used instead. When multivariate statistical fault detection methods are used, the T^2 statistic can be thought of as the predicted response within normal limits of the score space, while Q corresponds to the lack of fit. The objective of the experimental design will be to enhance the quality of the normal operating data and therefore the estimated loading vectors based on the above statistics. This in turn can improve the observability characteristic of the monitored statistics.

6.3.3 The integration between plant design, feedback control and the fault detection and diagnosis modules

The importance of considering the dynamic and control aspects during the early stage of the plant design has been often recognized as leading to improved controllability and operability characteristics. During the course of this project, the tradeoffs between control and fault detection and diagnosis outcomes have been considered to achieve an acceptable performance. Incorporating plant design in the optimization in addition to fault detection and closed loop performance considerations could be advantageous to achieve an improved process operation management. For example, proper design of the chemical plant can reduce the burden in

diagnosing or searching for the root cause of a particular fault. Plant design criteria could be incorporated into the optimization in the form of a search for the optimal sensor location. The latter is important for maximizing the amount and quality of the information that will be ultimately used by the fault detection and diagnosis systems.

Permissions

(1) Computer and chemical Engineering

The author has the right to include the journal article in full or in part, in a thesis or dissertation.

Copyright guidelines available at:

http://www.elsevier.com/wps/find/authorsview.authors/copyright#whatrights

(2) Industrial and Engineering Chemistry research

Licence number: 2563801374886

Copyright guidelines available at:

http://pubs.acs.org/userimages/ContentEditor/1218205107465/dissertation.pdf

(3) DYCOPS conference

The author has the right to include the journal article in full or in part, in a thesis or dissertation after three month of the conference.

Copyright guidelines available at:

http://www.dycops2010.org/index.php?option)com_content&view)article&id)11&Itemid)12

(4) Chemical Engineering Science

This paper has not been accepted yet. However, the author has the right to include the journal article in full or in part, in a thesis or dissertation.

Copyright guidelines available at:

http://www.elsevier.com/wps/find/authorsview.authors/copyright#whatrights

References

Alt, F.B. (1984) "Multivariate quality control", *Encyclopedia of Statistical Sciences*, 110-122

Bakshi, B.R. (1998) "Multiscale PCA with application to multivariate statistical process monitoring", *AIChE Journal*, **44**, 1596-1610

Baughman, D.R., Liu, Y.A. (1995) "Neural networks in bioprocessing and chemical engineering" Academic press Inc., San Diego

Benjamin J. O., De La Pena, D. M., Davis, J.F., Christofides, P. D. (2008) "Enhancing data based fault isolation through nonlinear control", *AIChE Journal*, **54**, 223-241

Bhushan, B., Romagnoli, J.A. (2008) "Self-Organizing Self-Clustering Network: A Strategy for Unsupervised Pattern Classification with Its Application to Fault Diagnosis", *Industrial and Engineering Chemistry Research*, **47** (12), 4209–4219

Bin Shams, M., Budman, H., Duever, T. (2010a) "Finding a trade-off between observability and economics in the fault detection of chemical processes", *Computer and Chemical Engineering*, (In Press)

Bin Shams, M., Budman, H., Duever, T. (2010b) "Enhancing fault observability using feedback control", *Industrial and Engineering Chemistry Research*, (In Press)

Bin Shams, M., Budman, H., Duever, T. (2010c) "Detection, Identification and Diagnosis using CUSUM based PCA", *Chemical Engineering Science*, (Submitted)

Bin Shams, M., Budman, H., Duever, T. (2010d) "Fault detection using CUSUM based techniques with application to the Tennessee Eastman Process", *Proceedings of the 9th International Symposium on Dynamic and Control of Process systems* (*DYCOPS*), Leuven, Belgium

Bishop, C.M. (1995) "Neural Networks for Pattern Recognition", Oxford University Press, New York

Brook, D., Evans, D.A. (1972) "An approach to the probability distribution of CUSUM Run length", *Biometrika*, **59**, 539-549

Chen, J., Bandoni, J., Romagnoli, J.A. (1996) "Robust PCA and normal region in multivariate statistical process monitoring", *AIChE*, **42**, 3563-3566

Cheng, C.Y., Hsu, C.C., Chen, M.C. (2010) "Adaptive kernel Principal component analysis (KPCA) for monitoring small disturbances of nonlinear processes", *Industrial and Engineering Chemistry Research*, **49**, 2254-2262

Chiang, L.H., Braatz, R. (2003) "Process monitoring using causal map and multivariate statistics: fault detection and identification", *Chemometrics and Intelligent laboratory Systems*, **65**, 159-178

Chiang, L.H., Russel, E.L., Braatz, R.D. (2001) "Fault detection and diagnosis in industrial systems". Springer, London

Chiang, L.H., Kotanchek, M.E., Kordon, A.K. (2004) "Fault diagnosis based on Fisher discriminant analysis and support vector machines", *Computers and Chemical Engineering*, **28** (8), 1389-1401

Choudhury, M.A.A.S., Thornhill, N.F. and Shah, S.L. (2006) "Automatic Detection and Quantification of Control Valve Stiction", *Control Engineering Practice*, **14** (12), 1395-1412

Choudhury, M A.A.S., Shah, S. L., Thornhill, N. F. (2004) "Diagnosis of poor control loop performance using higher order statistics", *Automatica*, **40**, 1719–1728

Croiser R.B. (1988) "Multivariate generalization of cumulative sum quality control schemes", *Technometrics*, **30**, 291-303

Ding, S.X., Zhang, P., Naik, A., Ding, E.L., Huang, B. (2009) "Subspace method aided-driven design of fault detection and isolation system", *Journal of Process Control* **19**, 1496-1510

Dong, D., McAvoy, T.J. (1996) "Nonlinear principal component analysis-based principal curves and neural networks", *Computers and chemical engineering*, **20** (1), 65-78

Downs, J.J., Vogel, E.F. (1993) "A plantwide industrial process control problem", *Computer and Chemical Engineering*, **17** (3), 245-255

Dunia, R., Qin, J. (1998) "Joint diagnosis of process and sensor faults using principal component analysis", *Control Engineering Practice*, **6**, 457-469

Gertler, J. (1988) "Survey of model based failure detection and isolation in complex plants", *IEEE Control Systems Magazine*, 3-11

Gertler, J., Cao, J. (2004) "PCA-based fault diagnosis in the presence of control and dynamics", *AIChe Journal*, **50**, 388-402

Hastie, T., Stuetzle, W. (1989) "Principal curves", *Journal of the American Statistical Association*, **84**, 502 - 516

Hawkins, D.M., Olwell, D.H. (1998) "Cu-mulative Sum Charts and Charting for Quality", Springer-Verlag, New York

Himes, D., Storer, R., Georgakis, C.(1994) "Determination of the number of the Principal components for disturbance detection and isolation", *Proceedings of the American Control Conference*, Baltimore, Maryland

Himmelblau, D.M. (1978) "Fault Diagnosis in chemical and petrochemical processes", Elsevier Predd, Amsterdam

Horch, A. (1999) "A simple method for detection of stiction in control valves", *Control Engineering Practice*, **7**, 1221-1231

Hotelling, H. (1947) "Multivariate quality control-illustrated by the air testing of sample bombsights", *Techniques of Statistical Analysis*, Eisenhart C, Hastay MW,Wallis WA(eds.). McGraw Hill, Newyork, 111-184

Isermann, R. (2006) "Fault diagnosis systems", Springer, Berlin

Jackson J. (1991) "A User Guide to Principal Components", Wiley, New York

Jackson. J., Mudholkar, G. (1979) "Control Procedures for Residuals Associated with Principal Component Analysis", *Technometrics* **, 21**, 341-349

Jacobson, C.A. , Nett, C.N. (1991) "An integrated approach to controls and diagnosis using the four parameter controller", *IEEE Control Systems*, **11**(6), 22-29

Kramer, M.A. (1991)"Nonlinear principal component analysis using autoassociative neural networks", *AIChE Journal*, **37** (2) 233-343

Ku,W., Storer, R., Georgakis, C. (1995) "Disturbance detection and isolation by dynamic principal component analysis", *Chemometrics Intelligent Laboratory*, **30**, 179-196

Lee, J.M., Yoo, C., Lee, I. (2004) "Statistical monitoring of dynamic processes based on dynamic independent component analysis", *Chemical Engineering Science*, **59**, 2995-3006

Lee, J.M.,Yoo, C., Choi, S.W., Vanrolleghem, P.A., Lee, I.B.(2004) "Nonlinear process monitoring using kernel principal component analysis", *Chemical Engineering Science* ,**59** (1), 223-234

Ljung, L. (1987) "System Identification: Theory for the user", Prentice-Hall, Englewood Cliffs, New Jersey

Lou, S. J., Budman, H., Duever, T. A. (2003) "Comparison of fault detection techniques", *Journal of Process Control*, **13**, 451-464

Lowry C.A.,Woodal W.H., Champ C.W., Rigdon S.E.(1992) "A multivariate EWMA control chart", Technometrics, **34**, 46-53

Lucas, J.M., Crosier, R.B. (1982) "Fast initial response for CUSUM quality control schemes: give your CUSUM a head start", *Technometrics*, **24**, 199-205

Lyman, P.R., Georgakis, C. (1995) "Plantwide control of the Tennessee Eastman problem", *Computer and Chemical Engineering*, **19**, 321-331

Macgregor, J.F., Jaeckle, C., Kiparissides, C., Kotoudi, M. (1994) "Process monitoring and diagnosis by multiblock PLS methods", *AIChE Journal*, **40** (5), 826–838

MacGregor, J.F., Kourti, T. (1995) "Statistical process control of multivariate processes" *Control Engineering Practice*, **3**, 403-414,

Mahadevan, S., Shah, S.L. (2009) "Fault detection and diagnosis in process data using one-class support vector machines", *Journal of Process Control*, **19** (10), 1627-1639

Mahdi, S., Mehrdad, S. (2010) "A decentralized technique for robust simultaneous fault detection and control of uncertain system", *Proceedings of the American Control Conference*, Baltimore, Maryland

Maulud, A.,Wang, D., Romagnoli,J.(2006) "A multi-scale orthogonal nonlinear strategy for multi-variate statistical process monitoring", *Journal of Process Control*, **16**, 671-683

Mendel, J. M., (1991) "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications", *Proc. IEEE*, **79**, 278–305

Miller, P., Swanson, R.E., Heckler, C.F. (1998). "Contribution plots: the missing link in multivariate quality control", *Applied mathematics and computer science*, **8**, 775-792

Montgomery, D.C, (1997) "Introduction to Statistical Quality Control", 3rd edn, John Wiley, New York

Negiz, A., Cinar, A. (1997) "Statistical Monitoring of Multivariable Dynamic Processes with State Space Models", *AIChE Journal*., **43** (8) 2002-2020

Prakash, J., Patwardhan, S., Shah, S.L. (2010) "State estimation and nonlinear predictive control of autonomous hybrid system using derivative free State estimators", *Journal of Process Control*, **20** (7), 787-799

Qin, S.J. (2003). "Statistical process monitoring: basics and beyond", *Journal of Chemometrics* **54**, 480-502

Raghuraj, R., Bhushan, M., Rengaswamy, R. (1999) "Locating sensors in complex chemical plants based on fault diagnostic observability criteria", *AIChE Journal*, **45** (2), 310-322

Raich, A., Çinar, A. (1997) "Diagnosis of process disturbances by statistical distance and angle measures", *Computers and Chemical Engineering*, **21** (6), 661-673

Rawlings, J., Bakshi, B. (2006) "Particle filtering and moving horizon estimation", *Computers and Chemical Engineering*, **30**, 1529-1541

Riggs, J. (1999) "Chemical process control", Ferret Publishing, Texas

Romagnoli, J. A., Sanchez, M. C. (2000) "Data Processing and Reconciliation for Chemical Process Operations", Academic Press, San Diego

Ross, J.P. (1988) "Taguchi Techniques for Quality Engineering", McGraw-Hill, New York

Scholkopf, B., Smola, A.J., Muller, K. (1998) "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, **10**, 1299-1399

Thornhill, N. (2005) "Finding the source of nonlinearity in a process with plant wide oscillation", *IEEE transaction on control system technology*, **13** (3) 434-443

Tyler, M.L, Morari, M.(1994) "Optimal and robust design of integrated control and diagnosis modules", *Proceedings of the American Control Conference*, Baltimore, MD, 2060-2064

Vaidyanathan, R.,Venkatasubramanian,V. (1991) "Process fault detection and diagnosis using Neural Networks-II.Dynamic processes", *AIChe Annual meeting*, Los Angeles, CA

Vedam, H.,Venkatasubramanian,V. (1999) "PCA-SDG based process monitoring and fault diagnosis", *Control Engineering Practice* , **7** (7) 903-917

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, N. (2003a) "A review of process fault detection and diagnosis Part I: Quantitative model based methods", *Computer and Chemical Engineering*, **27** (3), 293-311

Venkatasubramanian, V., Rengaswamy, R., Kavuri, N. (2003b) "A review of process fault detection and diagnosis Part II: Qualitative models and search strategies", *Computer and Chemical Engineering*, **27** (3), 313-326

Venkatasubramanian, V., Rengaswamy, R., Kavuri, N., Yin, K.(2003c) "A review of process fault detection and diagnosis Part III: Process history based method", *Computer and Chemical Engineering*, **27** (3), 327-346

Wang, D., Romagnoli, J.A. (2005) "Robust multi-scale principal components analysis with applications to process monitoring", *Journal of Process Control*, **15** (8), 869-882

Willsky, A.S. (1976) "A survey of design methods for faultier detections in dynamic systems", *Automatica*, **12**,601-611

Woodall,W., Ncube, M. (1985) "Multivariate CUSUM quality control procedures", *Technometrics*, **27**, 285-292

Yoon, S., MacGregor, J. (2001) "Fault diagnosis with multivariate statistical models part I: using steady state fault signatures", *Journal of Process Control*, **11**, 387-400

Yoon, S., MacGregor, J.F. (2004) "Principal component analysis of multiscale data for process monitoring and fault diagnosis", *AIChE Journal*, **50**, 2891-2903

Zhang, Y. (2009) "Enhanced statistical analysis of nonlinear process using KPCA, KICA and SVM", *Chemical Engineering Science*, **64**,801-811

Zhang,Y., Zhang, Y. (2010) "Fault detection of non-Gaussian processes based on modified independent component analysis" ,*Chemical Engineering Science*, **65** (16), 4630-4639

Appendix A

Scale CUSUM (SCS) Parameters

The parameters of the Scale CUSUM (SCS) have been derived as follows. Let, $x_i \sim N(0, \sigma^2)$, $i=12\ldots n$ and $y_i=|x_i/\sigma|^{\lambda}$. The characteristic of the distribution of y_i are easily worked out from the standard normal distribution. That is

$$\Pr[y_i < c] = 2\Phi(c^{1/\lambda}) - 1 \tag{A1}$$

where Φ () denote the standard normal function. Furthermore, the k^{th} moment of y_i is as follows:

$$E[y_i^k] = \mu = 2^{0.5.\lambda.k}.\Gamma[0.5(\lambda.k+1)]/\sqrt{\pi} \tag{A2}$$

where Γ() is the gamma function. With λ=0.5, the transformed variate y_i has a distribution which is very close to normal. In particular, using (A2), the first and second moments are given as followings:

$$E[y_i^k] = E[y_i^1] = \mu = 2^{0.25}.\Gamma(3/4)/\sqrt{\pi} = 0.82218 \tag{A3}$$

$$V(y_i) = \sqrt{(2/\pi)}.\Gamma(1) - \mu^2 = (0.34915)^2 \tag{A4}$$

Appendix .B

Fault detection using CUSUM based techniques with application to the Tennessee Eastman Process

(*Adopted form Bin Shams et al., 2010d, Proceedings of the 9th International Symposium on Dynamic and Control of Process systems (DYCOPS), Leuven, Belgium*)

B.1 Overview

In this paper, a cumulative sum based statistical method is used to detect faults in the Tennessee Eastman Process (TEP). The methodology is focused on three particular faults that could not be observed with other fault detection methodologies previously reported. Hotelling's-T^2 charting based on the cumulative sums of the faults' relevant variables was successful in detecting these faults, however, with significant delay. The speed of detection is further enhanced by retuning the fault's relevant controller at the expense of closed loop performance.

B.2 Introduction

An important aspect for the safe operation of chemical process is the rapid detection and removal of faults. Different methods have been proposed in the literature for fault detection and fault diagnosis (Venkatasubramanian *et al*., 2003). These methods can be broadly categorized into three main classes: (1) Data driven methods; (2) Analytical methods, and (3) Knowledge based methods (Chiang *et al*., 2001). Each of these methods has its own advantages and disadvantages depending on the problem. A number of researchers suggest combining these methods to improve detection. For examples, (Chiang and Braatz, 2001; Lee *et al*., 2003) have observed that data driven analysis is enhanced if knowledge of the process is used to describe

fundamental causal relationships among variables. Analytical methods require the use of first-principle models, thus making them less attractive for large scale systems. Therefore, they are not considered in the current work. In the chemical industry, large amount of data are measured by a variety of sensors and subsequently stored. These data generally exhibit high correlation in time and cross-correlation among variables. On the other hand, most of the data driven monitoring techniques assumed that data are uncorrelated and normally distributed. Different approaches have been proposed to mitigate the violation of these assumptions such as time series analysis and projection to latent variables methods (e.g. Principal Components Analysis) (MacGregor and Kourti, 1995).

Most monitoring data driven techniques are based on the statistical hypothesis-testing principle. Two types of errors occur when performing hypothesis testing referred to as type I and type II errors. A type I error occurs when a control chart indicates a fault in the absence of it, whereas a type II error occurs when a control chart fails to declare the existence of a fault, although it has occurred (Montgomery, 1997).

This paper proposes the application of Cumulative-Sum (CUSUM) based models for the detection of faults in the Tennessee Eastman problem (TEP) (Downs and Vogel, 1993). More specifically, the paper will investigate the application of Location CUSUM (LCS) and Scale CUSUM (SCS) based models to detect three particular faults that have been found unobservable by other algorithms previously applied to the TEP (Ding *et al*, 2009; Zhang, 2009; Chiang *et al*, 2001; Chiang and Braatz, 2001; Ku *et al*, 1995). After demonstrating the detection capability of the CUSUM based methods for each one of the three faults, a Hotelling's T^2 chart based on a cumulative sum of the observations is proposed for the individual or simultaneous detection of these three faults. Then, to quantify the fault observability, a statistical measure that is related to

the speed of detection is defined. Finally, since the faults are observed from variables that are embedded within control loops, the effect of controllers' tuning parameters on the trade-offs between speed of fault detection versus process variability will be assessed. The Appendix is organised as follows: A description of the implemented CUSUM and Hotelling's T^2 statistics' and the metric used to gauge fault observability are given in section B.2. Section B.3 presents an overview of the faults considered in the Tennessee Eastman Process (TEP) and illustrates the use of the CUSUM based methods for the detection of the three abovementioned faults Then, using the statistical measure of observability presented in section B.2, the tradeoffs between fault observability to process variability are investigated.

B.3 CUSUM, Hotelling's T^2 and average run length (*ARLo.c*)

B.3.1 The Cumulative sum (CUSUM) based control charts

A key disadvantage of Shewhart like control charts often used for detection is that they only use current time-interval information while not accounting for the entire time history. Hence, those charts are relatively insensitive to small shifts in the process variables especially for small signal to noise ratio. These shortcomings motivate the use of other alternatives such as the univariate or the multivariate version of the CUSUM based charts (MacGregor and Kourti, 1995). Three types of statistical charts are used. Specifically, location cumulative sum (LCS), scale cumulative sum (SCS) and the Hotelling's T^2. The current study proposes the use of a combined version of the three algorithms as described in the following section. The LCS and SCS algorithms are examples of univariate statistics while the Hotelling's T^2 is a multivariate statistics. Both the LCS and SCS are performed using the following two statistics, corresponding to a two sided hypothesis test (Hawkins and Olwell, 1998):

$$C_i^+ = \max[0, C_{i-1}^+ + x_i - (\mu_{i.c} + k)] \quad \text{(B.1)}$$

$$C_i^- = \max[0, C_{i-1}^- + (\mu_{i.c} - k) - x_i] \quad \text{(B.2)}$$

$$C_0^+ = C_0^- = 0$$

; where k , $\mu_{i.c}$, C_i^+ and C_i^- are the slack variable, the *in control* mean , the upper and the lower CUSUM statistics, respectively. The role of the slack variables is to introduce robustness to the calculated statistics. At every new sample, the statistics' in equations (B.1) and (B.2) result in the accumulation of small deviations in the mean (LCS) or small changes in the variability (SCS). These accumulations are corrected using the slack variable and compared to zero using the (max) operation. When either one of the two statistics, in equations (B.1) and (B.2), exceed a threshold H, the process is considered to be out of control. Following their respective definitions, the LCS is especially effective for detecting changes in the average whereas the SCS is suitable for detecting changes in variability. Guidelines for the selection of k and H have been reported (Hawkins and Olwell, 1998; Montgomery, 1997). Typically k is selected to be half of the expected shift in either μ or σ. H is determined so that a prespecified $ARL_{o.c}$, to be defined in the following section, is achieved. It should be noticed that when using equations (B.1) and (B.2), the LCS uses the original raw data x_i, whereas the SCS uses the following standardized quantity:

$$x_i = \frac{\sqrt{|y_i|} - 0.822}{0.349} \quad \text{(B.3)}$$

; where y_i denotes the original raw data. A derivation of the quantities in equation B.3 is given in Appendix A. Although LCS and SCS can be applied to individual measurements, there are many situations in which a pooled representative statistic for more than one variable is necessary. This is especially important when it is desired to present the operators with compact

information to simplify the monitoring activities for the process. For that purpose, when the monitored variables are normally and statistically independent, the Hotelling's T^2 can be used. The Hotelling's T^2 statistics and the upper and lower control limits are given by:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \tag{B.4}$$

$$UCL = \frac{p(m+1)(m-1)}{(m^2 - mp)} . F_{\alpha, p, m-p} \tag{B.5}$$
$$LCL = 0$$

; where p is the number of monitored variables and m is the total number of samples. In the current work the cumulative sum statistics' are combined together into one statistic, namely, the Hotelling's T^2 as described in a later section.

B.3.2 The $ARL_{o.c}$ as an observability measure

Observability of a fault is referred to the ability to detect the fault from the chosen set of measurements. In the current work, the statistical measure used to gauge observability is the out-of-control Average Run Length ($ARL_{o.c}$). The subscript ($o.c$) stands for *out of control*. The $ARL_{o.c}$ is defined as the average number of points that must be sampled or plotted before the chart signals and it is a function of the probability of type II error (β), that is:

$$ARL_{o.c} = f(\beta) \tag{B.6}$$

Due to their integrating nature, cumulative sum based techniques require some time before a fault can be detected, especially if the changes are very small. Accordingly, the $ARL_{o.c}$ is a suitable metric to quantify this expected delay in detection.

For example, if in response to a certain fault, the calculated $ARL_{o.c}$= 1, the fault would be detected, on the average, after the first sample following the onset of the fault. On the other hand, an $ARL_{o.c}$= *infinity* or a very large number implies that the fault is unobservable or it takes

a long time to observe it. The value of the $ARL_{o.c}$ depends on the type of chart that is used for monitoring. Several analytical expressions are available for specific statistical charts (Montgomery, 1997). The above discussion showed the feasibility of using the $ARL_{o.c}$ as a fault observability index. Different approaches to estimate the $ARL_{o.c}$ based on the Markov chain approach appeared in the literature, (e.g. Brook and Evans, 1972) but in practice, the $ARL_{o.c}$ is usually estimated from simulations conducted with random realizations of the disturbances (Woodall and Ncube, 1985). The latter approach is adopted in the current study.

B.4 Tennessee Eastman Process (TEP) and the ``Unobservable`` faults

The Tennessee Eastman process has been widely used as a benchmark problem to compare various monitoring solutions (Chiang *et al*., 2001; Ku *et al*., 1995; Lee *et al*., 2004; Ding *et al*., 2009). The process is open loop unstable and consists of five major unit operations, as shown in (Fig. 1): reactor, condenser, compressor, separator and stripper. The process produces two liquid products (G and H) and one by-product (F) from four gaseous reactants (A, C, D, E) and an inert (B). Based on the required product mix and production rate, the plant can be operated according to six different modes of operation. The original open loop FORTRAN code was provided by Downs and Vogel, 1993. The simulations of the plant were done with the second decentralized control structure proposed in (Lyman and Georgakis, 1995). Different monitoring techniques have been tested and reported for the TEP. These techniques have shown different capabilities in detecting the majority of the 20 faults generally assumed for the process. However, all of these previously reported techniques have consistently failed in detecting the three particular faults described in Table B.1.

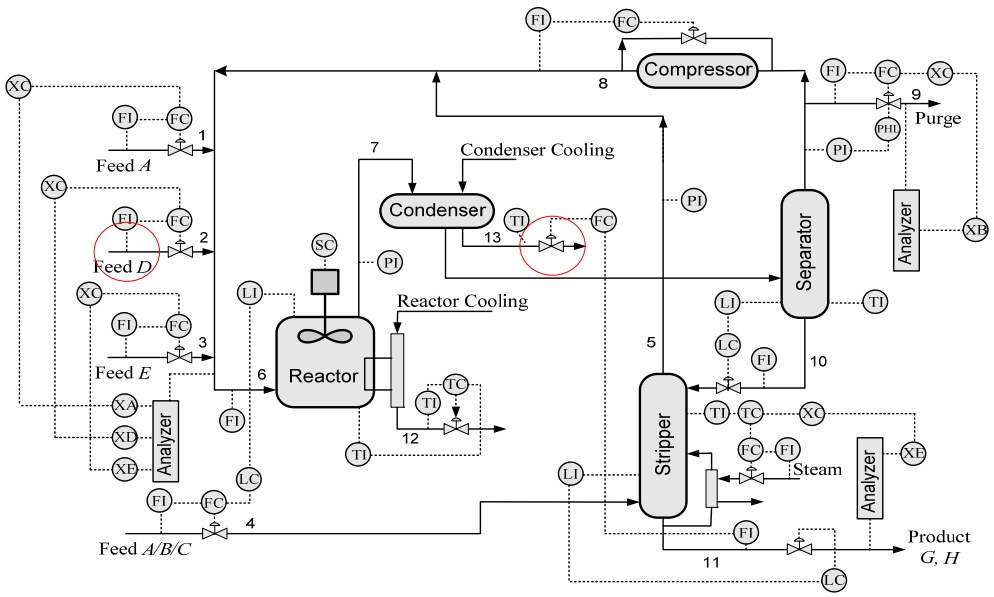


Fig.B.1. Tennessee Eastman process with the second control scheme described in (Lyman and Georgakis, 1995); the circles indicate the location of the faults described in Table B.1

Table.B.1 The unobservable faults of the TEP

Faults	Description	Nature
IDV(3)	D feed temp.	Step change
IDV(9)	D feed temp.	Random variation
IDV(15)	Condenser cooling water valve	Valve stiction

The resulted lack of observability when using specific techniques is attributed to the statistically insignificant changes, (i.e. changes in the process mean and/or variance), exhibited by the system when these faults occur. For fairness, it should be stressed that in most of the reported work, the detection was based on current time measurements thus the entire time histories of the measurements were not considered for detection as done in the CUSUM calculations proposed in the current study. However, the fact remains that these faults have not

been detected in previous studies while they may have an important economic or operational impact. Thus it is still very relevant to attempt to detect them. Later in the paper it will be shown that CUSUM based statistics are successful in observing these three faults after a certain period of time following the occurrence of the fault.

B.4.1 Previous attempts to tackle the TEP faults

Almost, all of the methods previously applied to the TEP were of multivariate nature. Among these techniques, for example, is the dynamic principal component analysis (DPCA) proposed by (Ku *el al*; 1995). Fig.B.2 shows the results of the application of DPCA for the TEP using the T^2. The statistic T^2 in Fig.B.2 is based on the sum of squares of the scores resulting from DPCA model. The DPCA has the advantage of taking into account information along several time intervals in contrast to the conventional static PCA which is based solely on data collected at the current time. Accordingly, DPCA is more suitable for dynamical systems. The bounds of normal operation corresponding to a 95% and a 99% confidence levels i.e. no fault has occurred, are shown as dotted lines in Fig.2 and are calculated by equation (B.5). The meaning of these bounds is that if the T^2 is above these bounds after the occurrence of the fault, then the fault is signalled. For the plots in Fig.B.2 the corresponding faults were introduced at time=160 samples. However, as shown in Fig.B.2, the T^2 statistics fails to surpass the thresholds after the onset of the 3 faults, i.e. IDV(3), IDV(9) and IDV(15). Hence, these faults cannot be detected by DPCA. It should be noticed that when a PCA/DPCA are used, p is replaced with a in equation (B.5), where a is the number of principal components retained in the PCA/DPCA model.

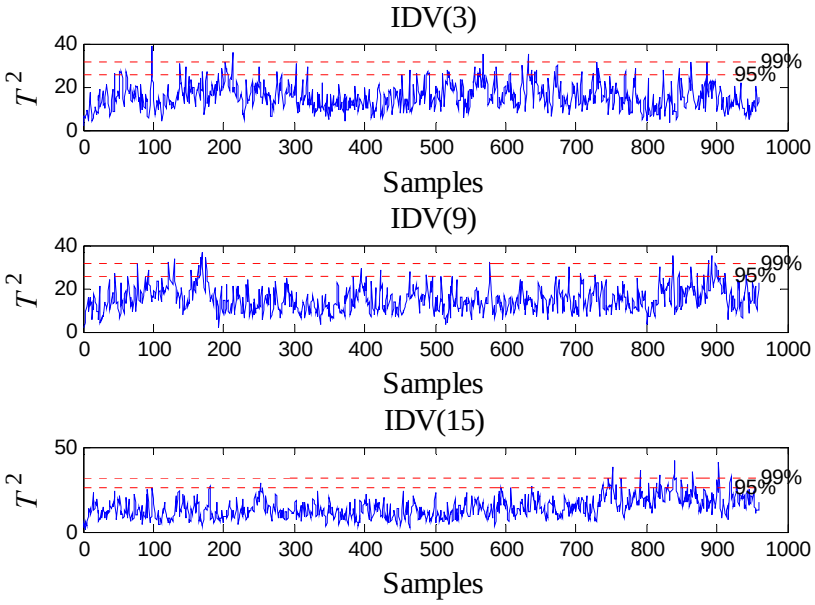


Fig.B.2. T^2 based on DPCA for the three unobservable faults of TEP, i.e. IDV (3), IDV (9) and IDV (15). The horizontal dotted lines represent the 95% and 99% confidence limits.

Chang *et al*, 2001 have performed a comparative study of multivariate techniques for detection of the faults in the TEP. They compared the PCA, DPCA and the subspace Canonical Variate Analysis (CVA) algorithms. Their performance index was the misclassification rate, i.e. the number of times the fault is not detected although the fault has occurred. Their conclusion was that the CVA results in the lowest misclassification rate in particular when monitoring the residual space with respect to an identified state space model. However, the three faults in Table 1 were excluded from the overall comparison, simply, because they can not be observed and high misclassification rates were associated with them. Other methods that were applied to the TEP problem are the Dynamic Independent Component Analysis (DICA) (Lee *et al*, 2003) and recently, a new subspace identification based method proposed by (Ding *et al*., 2009). All these methods excluded from their overall analysis the three faults given in Table B.1.

B.4.2 The CUSUM charting approach for the TEP unobservable faults

The inability of previous techniques to detect the 3 faults given in Table B.1 motivates the use of cumulative sum measures. Although Multivariate Cumulative Sum (MCUSUM) of all of the available TEP measurements could be used to detect the TEP faults (MacGregor and Kourti, 1995; Woodall and Ncube, 1985), this technique was still unable to detect these three faults. The latter has been tested by the authors; however, the results are not shown for brevity. Accordingly, it was decided to use a univariate Cusum on relevant variables as follows. Since the Cusum based statistics' are especially suitable to detect small changes in the process mean or small changes in process variability, it is important to identify the specific variables that exhibit these types of changes and apply the cusum operation on these variables. To find the variables for which the CUSUM operation should be applied, knowledge about the process was used. For example, it was observed that IDV (3) (Table B.1; small constant change in feed concentration) affects the steady state in the reactor. Since the reaction is highly exothermic and to keep the level of conversion at a desired level, manipulated variable XMV [10] must change to eliminate any changes in the mean of the steady-state reactor temperature. Then, the local cumulative sum of the manipulated variable XMV[10] is expected to provide detection of the corresponding fault after sufficient errors between the new steady state mean and the old steady state mean are integrated by the CUSUM operation. Based on similar arguments it is possible to find the individual relevant variables that are most sensitive to each fault and for which the Cusum operation should be applied to detect that particular fault. The faults and the corresponding variables used for detection are given in Table B.2.

Table.B.2 The unobservable faults/process variables pairing

Faults	Measurement*	Description
IDV(3)	XMV[10]	Reactor cooling water flow
IDV(9)	XMEAS[21]	Reactor cooling outlet temp.
IDV(15)	XMV[11]	Condenser cooling water flow

*The variable measurements as appeared in (Down and Vogel, 1993)

Also, since IDV(9) is a random disturbance around a mean and since IDV(15) results in cycling of the condenser cooling water flow due to valve stiction, the overall effect of these two faults is to increase the variance in their relevant variables as shown in Table B.2. Accordingly, the location CUSUM (LCS) is applied to monitor the effect of IDV(3) since it involves a shift in mean whereas the scale CUSUM (SCS) was used to monitor the effects of both, IDV (9) and IDV(15) , since they result in changes in variance. The sampling frequency for the CUSUM charts was (1/180) Hz (3 min. time intervals). In all the following simulations, the faults are introduced after 160 samples, that is, after 8 hours of a normal operation. Fig.B.3 shows the application of the LCS on XMV[10] when IDV (3) occurs. In this figure the fault was introduced at time=8 hours and was removed at time=700 hours to show whether the CUSUM statistics is able to predict both the occurrence and removal of the fault. The figure shows that the average time required for detection ($ARL_{o.c}$) is approximately 127.05 hrs. This time is calculated from the onset of the fault until the breaching of the threshold. An accurate $ARL_{o.c}$ requires averaging over a large number of noise realizations (Woodall and Ncube, 1985). It is also clear from Fig.B.3, that a very long time is required for detection due to the extremely small signal to noise ratio. This explains the inability of other previously used techniques to detect this fault. Also, the algorithm is able to detect the removal of the fault after some time. It is also clear from Fig.B.3 that the rate of change of the LCS statistic is higher when the fault is removed

than when the fault is introduced. This is expected because of the nature of the cusum algorithms given by equation (B.1) and (B.2) whereby when the rate of change of the (LSC) statistic is negative, the accumulator is reset to zero. This, in turn, accelerates the return to the statistical in control state. Different noise realizations were tested and used to calculate the average run lengths ($ARL_{o.c}$).

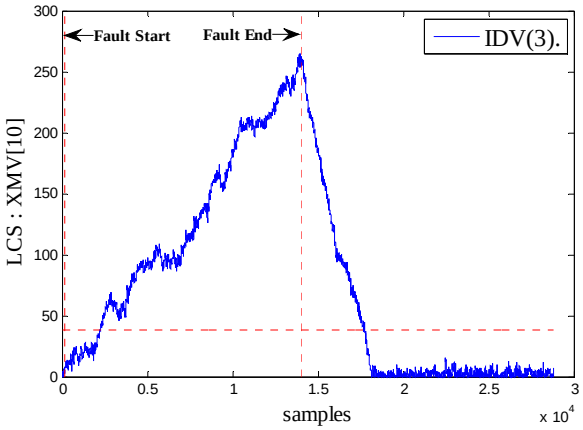


Fig.B.3 The LCS for IDV (3); horizontal line represent the statistical limit. The vertical lines represent the onset (after 8 hours) and the end of the fault (after 700 hours).

Fig.B.4 shows the detection of IDV(15) corresponding to valve stiction when using the SCS. The SCS for fault IDV(9) is not shown due to space limitation. In Fig.B.4 the fault has not been removed. The figures show that the SCS and LSC were successful in observing these two faults with $ARL_{o.c}$ values given in Table B.3. Thus, the CUSUM algorithms provide detection, but relatively long periods of time are required to detect the occurrence of the fault.

The immediate implication is that only faults that are of longer durations than the corresponding $ARL_{o.c}$ values can be detected using the Cusum based statistics'. Although three separate control charts could be used to monitor the 3 faults (Montgomery, 1997), it is often convenient for practical purposes to monitor the process with fewer.

Table.B.3 The estimated $ARL_{o.c}$ for the LSC, SCS and T^2

Fault	Statistics	$^{*}ARL_{o.c}$ (hr)
IDV(3)	LCS	127.05
IDV(9)	SCS	8.20
IDV(15)	SCS	41.00
IDV(3)	T^2	102.40
IDV(9)	T^2	276.05
IDV(15)	T^2	89.65
IDV(3) & IDV(15)	T^2	41.30

*All $ARL_{o.c}$ are calculated from after onset of the faults (i.e. after 8 hours)

In the current study, it is proposed to use the Hotelling-T^2 statistics to monitor the three faults with one single chart. For that purpose, the LCS algorithm is applied to XMV[10] whereas the SCS algorithm is applied , to XMEAS[21] and XMV[11], and then the corresponding cumulative sums are used to drive the Hotelling T^2 statistics defined in equation (B.4) ; where **x** is a vector sample composed of the 3 cumulative sums. To test for collinearity, the PCs of the covariance matrix were evaluated using a Scree plot (Chiang *et al*., 2001). Three PCs were found for all case studies. Fig.B.5, describes the Hotelling-T^2 results when only IDV(3) occurs while Fig.B.6 depicts the detection of IDV(9). In addition, Fig.B.7 illustrates the T^2 when both IDV(3) and IDV(15) occur simultaneously. In all cases T^2 based on the cusum statistics' were able to successfully detect the fault(s). Table.B.3 summarizes the relevant $ARL_{o.c}$ when the Hotelling's-T^2 charting based on the individual CUSUM statistics' was used.

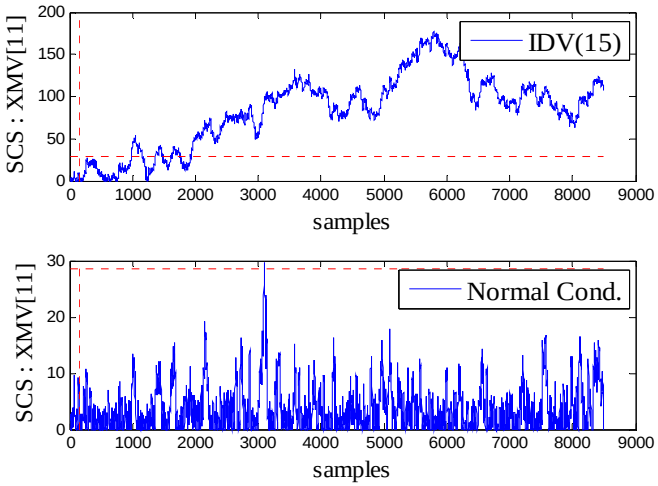


Fig.B.4 The SCS for IDV (15); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

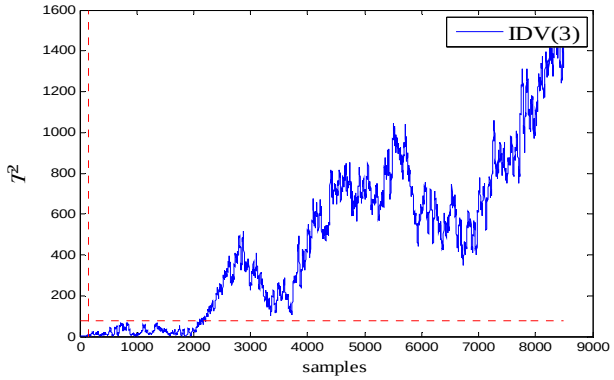


Fig.B.5 The Hotelling'T^2 for IDV (3); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

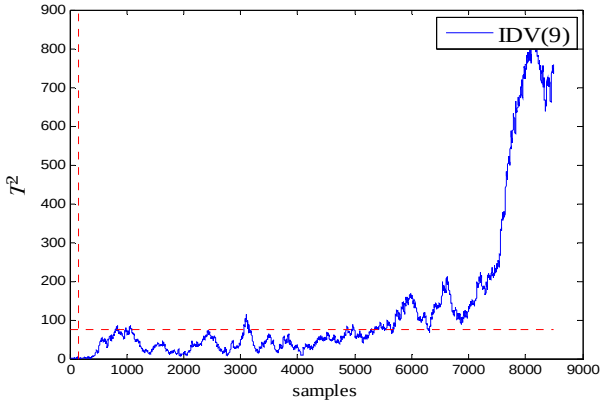


Fig.B.6 The Hotelling'T^2 for IDV (9); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

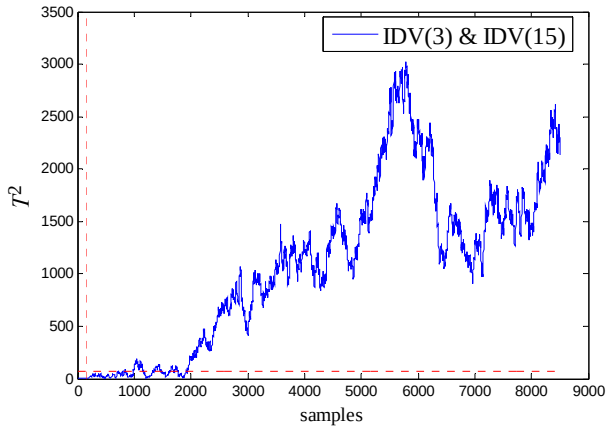


Fig.B.7. The Hotelling'T^2 for the simultaneous occurrence of IDV(3) and IDV(15); horizontal and vertical lines represent the statistical limit and the fault onset, respectively.

The $ARL_{o.c}$ of IDV (15) and the variability results as a function of the controller proportional gain (K) are shown in Fig.B.8. As can be seen from Fig.B.8, there is a significant interaction between the control and the detection scheme. The re-tuning of the controller significantly reduces the $ARL_{o.c}$ that would be required to observe IDV (15), but at the expense of significant degradation in performance as shown by the increased variability in the manipulated variable value. This variability may translate into significant wear of the corresponding valve. Thus, there is a motivation to seek for a trade-off between detection speed and closed loop performance provided that the related costs are available. The formulation of such optimization problem using the CUSUM based detection techniques is currently under investigation.

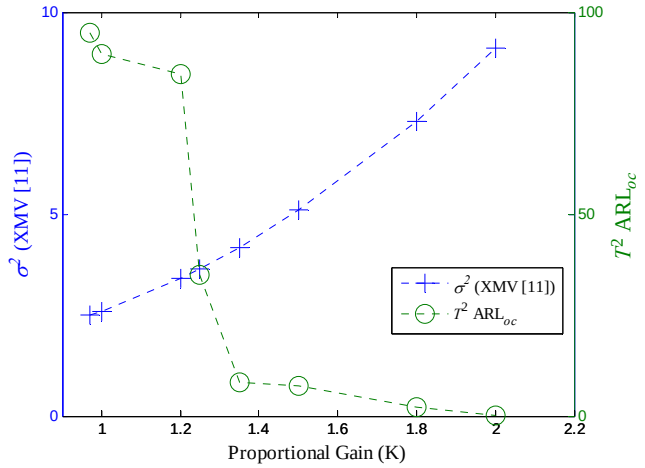


Fig.B.8 The change in variability and the T^2-ARL$_{o.c}$ as a function of the master controller's gain (XMV[11] is considered in the case).

B.5 Conclusions

A CUSUM based statistic combined with the Hotelling's T^2 charting is proposed. This method was successful in detecting three faults in the Tennessee Eastman problem that were impossible to observe with other previously applied methods. The 3 univariate CUSUMs were combined into one control chart by using Hotelling'T^2 statistics. Potential enhancements to the speed in detecting these faults, gauged by the $ARL_{o.c}$, can be achieved by formulating an optimization problem that explicitly considers the tradeoffs between detection and control performance.

Appendix C

MATLAB® and FORTRAN Codes

A copy of all programs developed during this project can be obtained from the CD available with Professor Hector M. Budman (hbudman@uwaterloo.ca) and Professor Thomas A. Duever (tduever@uwaterloo.ca) at the following addresses:

Department of Chemical Engineering

University of Waterloo

Waterloo, Ontario

Canada, N2L 3G1

Tel: +1 519 888 4567

Three folders titled; Chapter 1, Chapter 2 and Chapter 3, contain the relevant programmed codes for the corresponding chapter. All the programs with the extension "*.m" should be run with MATLAB® environment, while all programs with an extension "*.f" should be run with FORTRAN77 environment. All programs are started with few lines describing the purpose of the corresponding program.