Nonlinear Fault Detection and Diagnosis using Kernel based Techniques applied to a Pilot Distillation Column

by

David Phillpotts

A dissertation submitted in partial fulfillment
of the requirements for the degree

Master of Engineering (Control Engineering)

in the

Department of Chemical Engineering
Faculty of Engineering, the Built Environment and Information Technology

University of Pretoria
Pretoria

1st May 2007

Nonlinear Fault Detection and Diagnosis using Kernel based Techniques applied to a Pilot Distillation Column

David Phillpotts

Nonlinear Fault Detection and Diagnosis using Kernel based Techniques applied to a Pilot Distillation Column

Author:	David Phillpottts
Date	1st May 2007
Supervisor:	Professor P. L. de Vaal
Department:	Department of Chemical Engineering
	University of Pretoria
Degree:	Master of Engineering (Control Engineering)

Synopsis

Fault detection and diagnosis is an important problem in process engineering. In this dissertation, use of multivariate techniques for fault detection and diagnosis is explored in the context of statistical process control. Principal component analysis and its extension, kernel principal component analysis, are proposed to extract features from process data. Kernel based methods have the ability to model nonlinear processes by forming higher dimensional representations of the data.

Discriminant methods can be used to extend on feature extraction methods by increasing the isolation between different faults. This is shown to aid fault diagnosis. Linear and kernel discriminant analysis are proposed as fault diagnosis methods.

Data from a pilot scale distillation column were used to explore the performance of the techniques. The models were trained with normal and faulty operating data. The models were tested with unseen and/or novel fault data. All the techniques demonstrated at least some fault detection and diagnosis ability. Linear PCA was particularly successful. This was mainly due to the ease of the training and the ability to relate the scores back to the input data. The attributes of these multivariate statistical techniques were compared to the goals of statistical process control and the desirable attributes of fault detection and diagnosis systems.

Keywords: Kernel based methods; Fault detection; Fault diagnosis; Statistical process control

Acknowledgements

I would like to give my sincere thanks to Professor Philip de Vaal, my friends in the Process Modelling and Control group as well as my family for their support and guidance.

I am also grateful for Mintek's generous support of my university career.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Roman Symbols

$\mathbf{A}$	Matrix of all eigenvectors
a	Individual eigenvectors for PCA or semimajor axis length for an ellipse
B	Constant for estimating control limits of S charts
b	Semiminor axis length
D	Constants for estimating variances or ranges for control charts or depth region in bagging
d	Constants for estimating variances or ranges for control charts or V-mask lead distance for CuSum charts
F	Statistical distribution
g	Weight of the χ^2 distribution for SPE limits, alternatively: number random groups for the PRESS procedure, or number of classes for discriminant analysis
g_k	Expected variance of k^{th} group
h	The degrees of freedom of the χ^2 distribution
$\mathbf{I}$	Identity matrix of the appropriate dimensions
i	An index
j	An index
$\overline{\mathbf{k}}_t$	Centred test kernel matrix
$\overline{K}$	Centred kernel matrix
K	Kernel matrix

k	Parameter for the V-mask design, alternatively: number of groups in a weighted control chart or referring to PCA: number of variables
ℓ	Optimal weight matrix
$\mathbf{L}$	Scores or reduced projected set
$\tilde{l}$	Cumulative variance
l	Eigenvalues or variance explained by a PC
m	Number of samples available to determine the control limits or span of moving average
$\mathbf{m}$	Variable representing the mean in KPCA
n	Sample size
p	Number of quality characteristic or variables
$\mathbf{R}$	Sample correlation matrix
$\overline{R}$	Mean Range
R	Range
r	Element of the sample correlation matrix in PCA or length of a hexagon side for a hexagonally binned scatterplot
$\boldsymbol{S}$	Covariance Matrix
$\overline{S}$	Mean standard deviation
S	Standard Deviation (univariate) or sample covariance matrix (multivariate)
S^2	Variance
s	Sample standard deviation
T	Score Matrix
T^*	Depth median
T^2	Hotelling's Distribution
t	Scores for each of j PC directions
$\mathbf{v}$	Eigenvector for KPCA
$\mathbf{W}$	Matrix of all Eigenvectors (equivalently PCA directions)
$\mathbf{w}$	Principal component direction
W	Arbitrary Random Variable or Criterion for PRESS
$\mathbf{X}$	Matrix of data for all samples and for all variables

$\overline{\overline{x}}$	Mean of moving average or mean of the mean of all subgroups
$\overline{x}$	Mean of the data sample represented by x
x	Sampled Data
x_{new}	Score calculated using all the PCs
$\mathbf{Y}$	Matrix of data for all samples and for all variables belonging to a specific class
y	Univariate observations
Z	Standard normal random variable
$\mathbf{Z}$	Matrix of all PCA scores
z	z-score for PCA or exponentially weighted data for EWMA charts

Acronyms

AEM	Abnormal Situation Management
AIC	Akaike Information Criteria
ARL	Alternating Residual Least squares
ARL	Average Run Length
ASM	Abnormal Situation Management
CL	Centre Line
CuSum	Cumulative Sum
DCS	Distributed Control System
DKPCA	Dynamic Kernel Principal Component Analysis
EWMA	Exponentially Weighted Moving Average
GMA	Geometric Moving Average
HART	Highway Addressable Remote Transducer
KDA	Kernelised Discriminant Analysis
KFDA	Kernelised Fisher Discriminant Analysis
KLT	Karhunen-Loéve Transform
KPCA	Kernel Principal Component Analysis
LCL	Lower Control Limit
LDA	Linear Discriminant Analysis

LSL	Lower Specification Limit
MANOVA	One-way Multivariate Analysis of Variance
MKPCA	Multiway Kernel Principal Component Analysis
MSE	Mean Squared Error
MSPC	Multivariate Statistical Process Control
NIPLAS	Non-linear Iterative PArtial Least Squares
PC	Principal Component
PCA	Principal Component Analysis
PRESS	PREdiction Sum of Squares
QDA	Quadratic Discriminant Analysis
RMV	Raw Measurement Value
SPC	Statistical Process Control
SPE	Squared Prediction Error
SVM	Support Vector Machine
UCL	Upper Control Limit
USL	Upper Specification Limit
VI	Validity Index
VMV	Validated Measurement Value

Greek Symbols

α	The probability of describing normal variance as out of control (a type 1 error)
β	Probability of not detecting a V-mask shift
χ^2	χ^2 Statistic
Δ	Required detectable shift in mean for a V-mask
δ	Difference from desired value
λ	Constant for exponential weighting
λ	Eigenvalue for KPCA
Λ^{-1}	Diagonal matrix of the inverse of the eigenvalues
μ	Population mean

$\overline{\Phi}$	Mapped point
Φ	Nonlinear mapping function
$\hat{\Phi}$	Mapped point using all principal components
π	Denoting a class
Σ	Covariance Matrix
σ	Population standard deviation for general statistics, alternatively: kernel argument
$\hat{\sigma}$	Estimate of σ
Θ	Point belonging to a bivariate dataset
θ	V-mask angle

Subscripts

α	At α confidence
c	Centred data
d	Index for PCs
$data$	Referring to the sampled data
h	Index
L	Lower
$\mathbf{L}$	Denoting the first $\mathbf{L}$ Principal components
LCL	Lower Control Limit
max	Maximum of the data
min	Minimum of the data
$pooled$	Pooled unbiased estimate of Σ
T	Target
t	Referring to a time point
U	Upper
UCL	Upper Control Limit

Superscripts

h	Feature dimensionality
m	Input dimensionality
T	Matrix transpose

CHAPTER 1

Introduction

The monitoring of industrial processes for performance and fault detection is an essential part of the drive to improve process quality. The requirements of improved productivity, efficiency, safety and reduced levels of manning have led to increased investigation into fault detection and diagnosis.

With the increased use of instrumentation, huge quantities of dynamic plant data are available in real-time. Regulatory control actions are now routinely performed by computer systems with considerable success. The role of operators and control systems have shifted from being primarily focused on regulatory control to a broader supervisory role. The data available from a plant contain hidden redundancies together with important information about impending and current faults, as well as process performance. Unfortunately, much of this information is not used due in part to the complexity of extracting the complex relationships from within the data. Human reactions still serve as the primary response to detecting and diagnosing faults in chemical processes. Consequently considerable valuable knowledge is not utilised to identify, prevent and respond to faults and other undesirable operating conditions on the plant (Fourie (2000) and Jemwa & Aldrich (2006)).

Multivariable statistical approaches to process monitoring, fault detection and diagnosis are well accepted and rapidly developing fields. The techniques have an ability to extract useful relationships from massive data sets. These relationships can be monitored for faults and analysed and aid diagnosis of these faults (Choi et al., 2005).

The aim of this investigation is to firstly explore existing statistical methods for the detection and diagnosis of faults (definitions follow). The next step is to survey more modern nonlinear kernel based techniques for data feature extraction and classification. These techniques will be applied to data containing faults sourced from a lab scale distillation column with a commercial control system and instrumentation. The results from

these techniques will be compared to linear multivariate techniques. The attributes of these multivariate fault detection methods will be compared to the desirable attributes and of goals of a fault detection diagnosis.

1.1 Definitions of Fault Detection and Diagnosis

A fault is generally defined as a departure of an observed variable or calculated parameter from an accepted range (Himmelblau, 1978). More specifically, a fault is an unpermitted deviation of at least one characteristic property of a variable from an acceptable behaviour. This means that a fault may lead to the malfunction or failure of the system Iserman (2005). The underling cause of the of the fault is called the root cause or basic event.

Iserman (2005) identified time dependency in faults as in figure 1.1. Faults can appear abruptly - with the cause and/or effects continuing until corrected, or incipiently, where the effect on the process remains constant until corrected. Intermittent faults disappear and reappear in time.

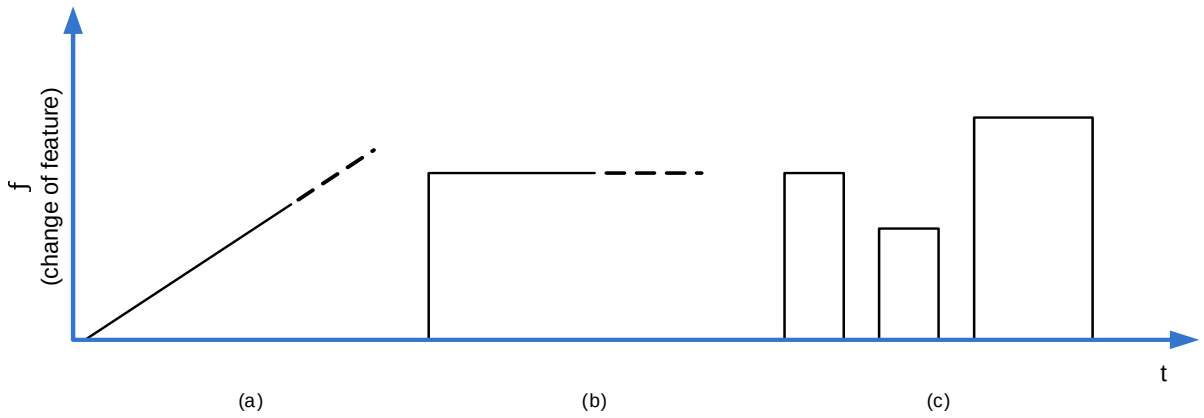

Figure 1.1: Time Dependancy of Faults: (a) Incipient; (b) Abrupt; (c) Intermittent.

Faults can be further classified as additive and multiplicative faults (see figure 1.2 Iserman (2005)). Additive faults influence a variable by the addition of fault features. These types of faults can appear as offsets in a process metric from a normal or desired value. Multiplicative faults affect the variable as a product and often manifest themselves as parameter changes with the process.

Faults may include (Venkatasubramanian et al., 2003a):

- Process unit failures
- Process unit degradation
- Control system failure
- Sensor failure

- Actuator failure
- Parameter drifts or gross changes
- Operation beyond normal regimes

A more detailed discussion follows:

Process Unit Failures and Degradation

Structural changes result in changes in the information flow between different variables. This would obviously affect any models used for control.

Control, Actuator or Sensor Failures

Gross errors usually occur with sensors and actuators. These types of fault often propagate rapidly though a process due to a control system. Problems with sensors include measurement noise. Many sensor and actuator faults can be readily detected and remedied by means of a data validation system. This ensures that the data feed to a fault detection system is valid.

Parameter Drifts or Gross Changes

These types of faults occur when an independent variable enters the system from the environment. This include unusual process disturbances. Parameter changes also affect any models (be they mathematical or expert knowledge) of the system. This may affect control or operator decisions.

Operation beyond Normal Regimes

This is closely related to parameter changes as operating in a different process regime will result in grossly different relationships between variables.

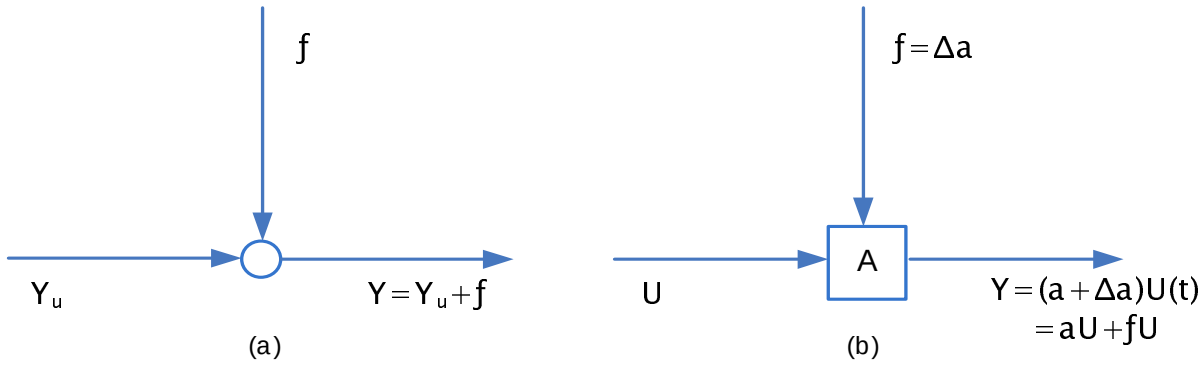

Figure 1.2: Basic Fault Models: (a) Additive and (b) Multiplicative Faults

1.1.1 Fault Detection as Part of a Plantwide Control System

Within the context of a plantwide control system, supervisory functions exist to indicate undesired or not-permitted process states and to take action in reaction to these indications. Iserman (2005) distinguishes between the following functions:

1. *Monitoring:* Variables are measured and compared to alarm limits. This information is presented to an operator.

2. *Automatic Protection:* Actions are taken in reaction to a dangerous state (includes interlocks).

3. *Fault Detection and Diagnosis:* Data is measured; features calculated; symptoms detected; diagnosis and decision on actions made.

The obvious advantages of the first two functions are simplicity and reliability. The disadvantage is the delay in reacting to sudden or gradually increasing faults. It is also not possible to perform in depth diagnosis. Therefore, a method having the features of the third function is necessary. This will allow deep insight into the process behaviour.

Fault detection and diagnosis systems can be viewed as a fault event classification system. See figure 1.3 (Brambley & Katipamula, 2005). The system will detect a fault, diagnose the case and then decide on a an appropriate action. The action will be evaluated in terms of the impact on the process. In this way, fault detection and diagnosis forms the first step of an Abnormal Situation Management (AEM) system (Fourie, 2000: 2-1–17). AEM involves the timely detection of an abnormal event, the diagnosis of its casual origins and the taking of appropriate steps to bring the process back to a safe operating state (Venkatasubramanian et al., 2003a).

Venkatasubramanian et al. (2003b) presents a diagram similar to figure 1.4, to demonstrate the integration of the fault detection and diagnosis system into a greater plantwide control system. Fault detection and diagnosis is probably too complex a task to be directly combined with regulatory control. Data points are acquired from the process. These data are then used directly for regulatory control and monitoring. Data trends and fault information are used to validate and reconcile the data. This information can then be used to improve process models used for fault detection, supervisory and regulatory control. Fault information is used by the supervisory control system. This information is also sent to the operator.

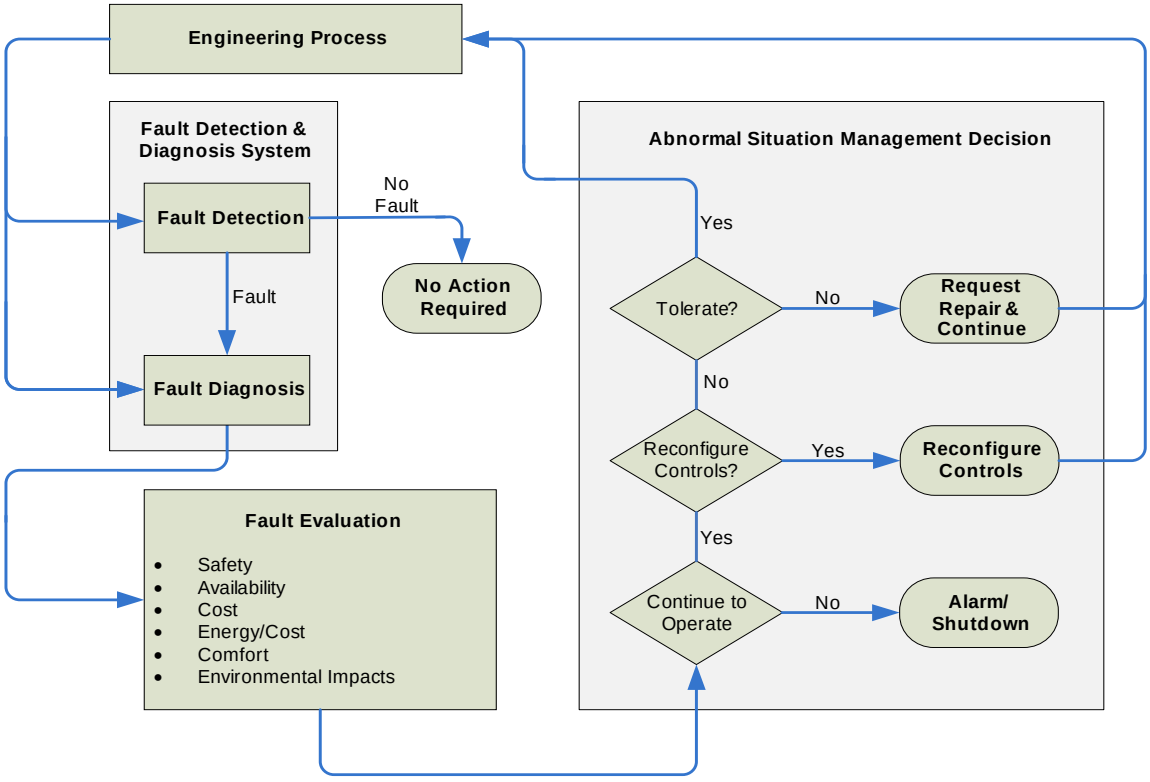

Figure 1.3: Role of Fault Detection and Diagnosis in an AEM System

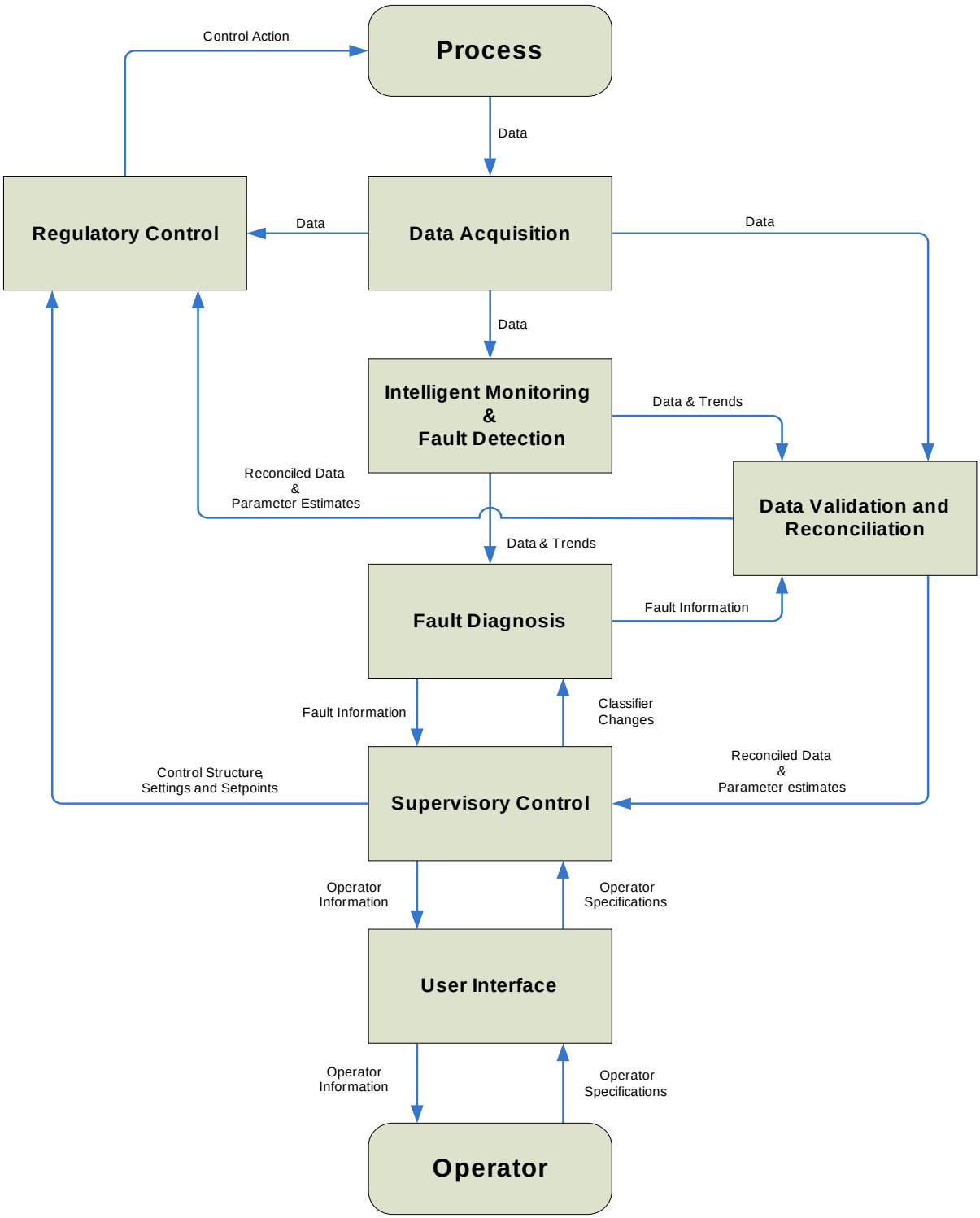

Figure 1.4: Plantwide Control System Integration Framework

1.2 Fault Detection as a Method of Continuous Improvement

Traditionally, many manufacturing organisations have operated on the basis of fault detection on the end product as shown in figure 1.5 (Owen, 1989). Despite the complex processes (including methods, people, materials, the environment and equipment), the action has been concentrated on the output. This is undesirable, because this method requires a high degree of product inspection (which is often infeasible; it is expensive (time and process capacity is lost because faults pass through the whole system before being corrected, reworked or rejected); furthermore it is demotivating.

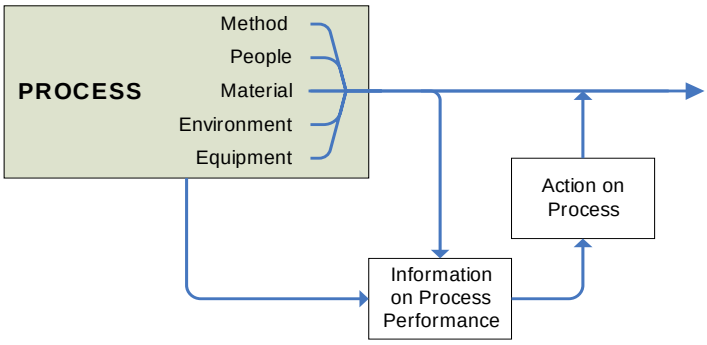

Figure 1.5: A Detection System

It is far better to use a prevention scheme as shown in figure 1.6. The emphasis in this system is the improvement of the process to minimise the effects of the fault or to prevent it from occurring again. A fault detection and diagnosis system is the key to either directly provide or to instigate this action.

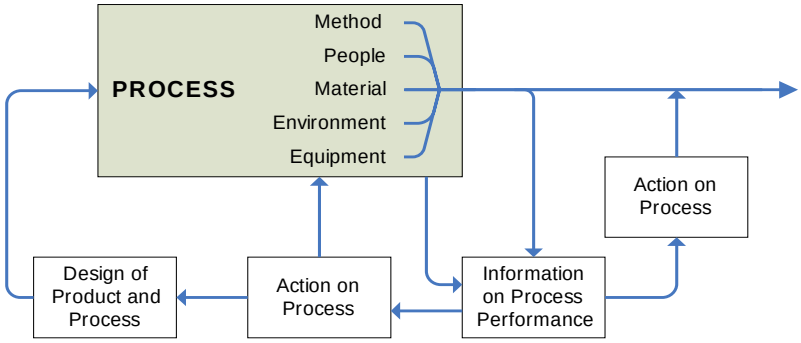

Figure 1.6: A Detection and Prevention System

When a process is important to the success of a company, it is important to continually improve. This may be a requirement due to the costs involved in ignoring quality (section 1.9) or because of a changing process. A concept similar to that of traditional regulatory control can be applied to quality improvement programmes. A target or benchmark is proposed or measured. The process is then measured and the data analysed for

quality performance in comparison to that benchmark. This is shown schematically in figure 1.7

A fault detection system would be able to flag if a fault has occurred and a diagnostic system would be able to suggest the cause (some expert knowledge or experimentation would be needed for a detection only system). This could be used to control either the output (as per figure 1.5) or the process itself (as per figure 1.6). Continuous improvement suggests a view to the long-term quality of the process. The process could then be improved by changes to the methods, people, material, environment or equipment. This cycle should be repeated to ensure maximum improvement is achieved. Also, as some quality improvements are made, other are likely to be revealed. A continuous improvement commitment will handle both changes to the process and the business and technical environment that the process is operated in.

It is possible that that process improvement analyses are preformed offline (see section 1.4.2) while the process regulation handles the correction of short-term faults.

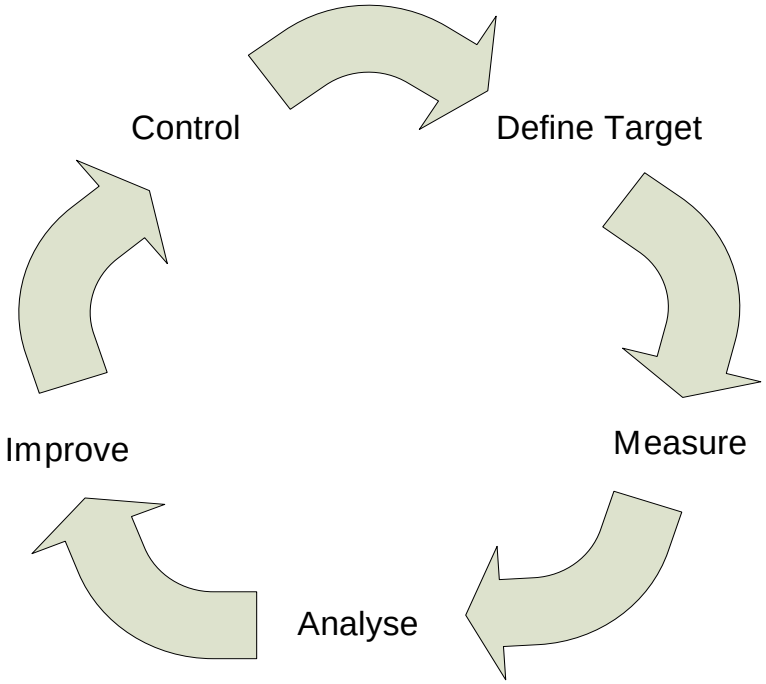

Figure 1.7: Continuous Improvement

1.3 Desirable Attributes of a Fault Detection and Diagnosis System

The following section describes the desirable attributes of a fault detection system. These aspects can be used to compare methodologies or to assess whether a fault detection and diagnosis system is successful.

This list is synthesised from both Dash & Venkatasubramanian (2000), Iserman (2005) and Venkatasubramanian et al. (2003a).

1. Real-time detection and diagnosis
2. Diagnosis in addition to detection
3. Fault isolation
4. Completeness
5. Early detection and diagnosis
6. Robustness and supervision of processes in transient states.
7. Novelty identification
8. Classification error estimate
9. Multiple fault identifiability
10. Adaptability
11. Reasonable modelling requirements
12. Reasonable storage and computational requirements
13. Detection of faults in closed loops.
14. Diagnosis of faults in actuators, sensors and other process components.

A more detailed discussion follows:

Real-time Prediction, Detection and Diagnosis

It is best if a system can give an indication of a fault while or before it has occurred. This allows corrective action to be taken as explained in section 1.2.

The alternative of analysing the data later, or after a fault has occurred, is still useful in indicating process quality or possibly the cause of any faults.

Diagnosis in Addition to Detection

A fault diagnosis system must, by definition, not only be able to flag instances in time when a fault has occurred, but also suggest (a set) of hypotheses for the cause of the fault. It is also desirable to provide explanation about how the fault propagated through the system. One would like a fault detection and diagnosis system to justify why certain hypotheses were proposed in addition to why certain others were not.

Diagnosis in addition to detection has several inherent tradeoffs that will be discussed directly below.

Isolability

Isolability refers to the ability of a fault detection and diagnosis system to differentiate between different failures. Ideally the system should return the single root cause of the fault while not returning any output responding to the faults that have not occurred. There is an important trade-off between isolation and the rejection of model uncertainties.

Isolability requires that the faults are orthogonal in the decision space. This would ensure that faults that have occurred will generate only the output corresponding to the faults that have occurred. However, a high degree of isolation will usually result in poor rejection of modelling uncertainties. This is due to the limited degree of freedom available for the fault diagnostic system design.

Completeness

In contrast to the requirements of isolation (which requires the generated fault set to be as minimal as possible), we also have the requirement of completeness - that is to require that the actual fault(s) are a subset of the proposed fault set. This trade-off will impact on the completeness or the accuracy of the predictions.

Early Detection and Diagnosis

Early detection and diagnosis is a highly desirable attribute. There will be a trade-off between the earliness of the detection and the isolation. Early detection will also generally result in increased incorrect detection and diagnosis (i.e. decreased robustness).

Robustness and Supervision of Processes in Transient States

Robustness with regard to model-process mismatch, noise and other uncertainties is desirable. The needs for robustness must be balanced with the need for performance. The ability to detect faults during transient states such as start-up and shut-down is also vital.

Novelty Identification

A fault detection and diagnosis system must be able to decide whether a process is running normally and, if not, diagnose the fault as a known malfunction or some other novel malfunction. Typically it is difficult to find historical process data for all possible faults. For this reason, it is highly desirable for a fault detection and diagnosis system to still detect novel faults while not classifying them as other kinds of known malfunctions or as normal operation.

Classification Error Estimate

Some indication of the reliability of the recommendation of a fault detection and diagnosis system is useful.

Multiple Fault Identifiability

The identification and diagnosis of multiple faults is difficult due to the interaction of processes and faults. This could be difficult for large interactive processes.

Adaptability

It is also desirable to have extendable systems. This would allow a fault detection and identification system to be expanded if more process information becomes available or if the process expands.

Reasonable Modelling Requirements

The amount and accuracy of modelling required is an important issue. Modelling will add to the complexity and time required to implement a fault detection and diagnosis system.

Reasonable Storage and Computational Requirements

Another trade-off exists between the ability to perform complex computations and the requirements of reasonable storage and computation. A fault detection system is ideally online and this would require real-time computation.

Detection of Faults in Closed Loops

The ideal fault detection and diagnosis system must be able to handle the interactions and dynamics added by closed loop regulatory systems.

Diagnosis of Faults in Actuators, Sensors and other Process Components

An ideal fault detection and diagnosis system would be able to identify faults from all origins.

1.4 Online versus Offline Fault Detection and Diagnosis

1.4.1 Online Fault Detection and Diagnosis

Online methods fall into two categories (Wetherill & Brown, 1991: 3):

- Screening methods
- Preventative methods

A short discussion follows:

Screening Methods

In this method, the *output* is screened, and if the quality does not meet the standards, the item is scrapped or reworked. This is an expensive method and is not ideal, as outlined in sections 1.2 and 1.3.

Preventative Methods

In this method, the *process* is screened and process control is used to prevent or reduce the source and effects of reduced quality. Examples include control charts (section 2) and continuous monitoring of process inputs and outputs.

1.4.2 Offline Fault Detection and Diagnosis

While medium-term fault detection can be done offline (for example to check if a large batch of chemicals was processed properly), offline fault detection is best used for the analysis of the causes of variability in the process with a eye to improving the process by making it less sensitive to these sources. These methods should be included from the beginning of production. Improving a process in this way requires insight and skill (Wetherill & Brown, 1991: 3).

1.5 Functions related to Fault Detection and Diagnosis

As discussed before, fault detection is a part of overall plantwide control strategy. There are several related methodologies.

Intelligent sensors and actuators have functions which remove some of the burden of fault detection from the fault detection system. The fault detection functions built-in

to such devices are typically much simpler than process wide detection systems. This is because the instruments are simple, well-defined with simple dynamics. The scope of the detection is also much smaller.

Expert systems can be used to complement some operator knowledge and to solve specialised problems, typically in narrow domains of expertise.

Data validation and reconciliation consists of adjusting and reconciling the process measurements to obtain more accurate measures of the process variables. Data validation may also be able to identify faulty sensors.

The use of these methodologies is addition to process fault detection allows the system to focus on the greater process fault detection problem while being able to assume to a greater degree that the instruments and actuators are behaving as intended, and that the data from the plant are correct. A discussion of these related methodologies follows:

1.5.1 Intelligent Sensors and Actuators

Sensors and regulators form the very core of a chemical process control system. Faults in sensors or actuators will propagate through the system due to the actions of the control system. Sensor faults include:

1. Bias
2. Drift
3. Precision
4. Degradation
5. Gross error
6. Complete failure

Traditionally, electric and pneumatic transmitters are used to generate signals from solid-state remote instruments (Nagy, 1992: 302). The primary element of the instrument may or may not be integrated with the transmitter. A solid-state sensor can easily be integrated with an amplifier and a analogue to digital converter. In the same way, intelligent sensors can be produced with integrated processing power. A comparison of an instrument with an intelligent sensor with an integrated sensor is shown in figure 1.8 (Nagy, 1992: 303). Intelligent sensors are essential for self-diagnosis in digital control systems (Nagy, 1992: 302). Intelligent sensors can automatically compensate for variations in process and ambient conditions. The use of a processing power allows the use of digitally stored data for precise ranging and various nonlinear correlations.

Modern instruments can self-diagnose all types of sensor faults listed above. The sensor creates a ‘Raw Measurement Value’ (RMV) from the process in the same way

that a conventional dumb instrument would. This RMV is then validated to a 'Validated Measurement Value' (VMV) and 'Validity Index' (VI). In the case of a fault, an abnormal RMV would be generated. If the validation routine worked, the validity index would be changed to show some fault status and the decreased confidence in the accuracy of the measured value. This process is shown in figure 1.9 as shown by Henry & Clarke (1991). The validity index may include some of the following information:

- Information about the validity of the measurement
- Information about the confidence in the the measurement
- Information about the signal
- Information about sensor settings
- Preventative maintenance information

This information is then provided to the 'Next Level Up' (NLU) in the control system (Henry & Clarke, 1991). This next level may be a controller, alarm or monitoring system.

Intelligent transmitters required new communication technologies. HART (Highway Addressable Remote Transducer), an open protocol, has been accepted by manufacturers such as Rosemount and Fischer. HART allows simultaneous digital and analog communication. There is also a widely used, all digital communication protocol named Fieldbus.

Many intelligent sensors have been implemented on commercial control systems. The isolation is robust. An additional advantage is that no training information is required.

Intelligent actuators operate in a similar way to intelligent sensors. An example of an intelligent sensor is a control valve which can generate an alert based on a discrepancy between the valve setpoint and the actual valve position. Actuators can also store information for maintenance purposes. An example for a control valve is the number of strokes that it has taken. This preventative maintenance information can be used to generate decisions such as to run until failure, run at reduced load or to repair immediately. This allows plant technicians to focus on faults rather than to endlessly check process equipment.

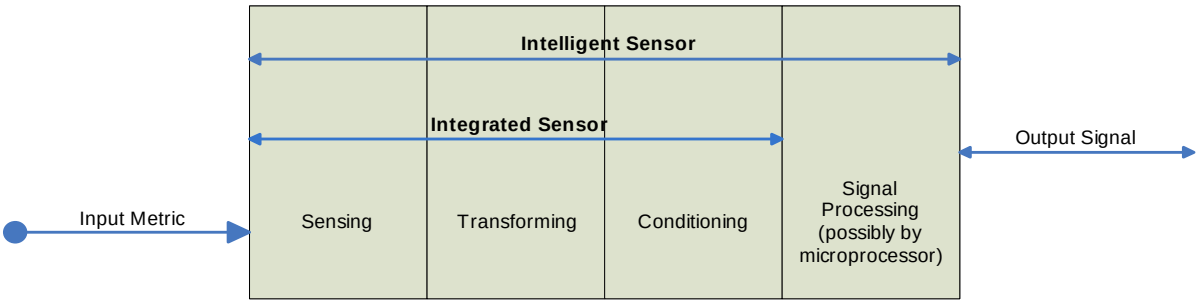

Figure 1.8: The Principle of the Intelligent Sensor

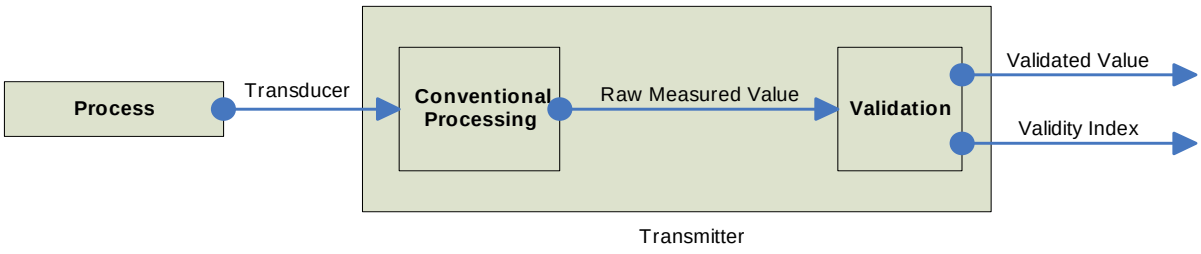

Figure 1.9: A Validating Sensor

1.5.2 Data Validation and Reconciliation

Large quantities of data are collected and stored from modern process plants. This data is subject to variations on quality (Romagnoli & Palazoglu, 2006: 429–448). Typical problems with raw data received from a plant are:

- Missing data
- Outliers
- Problems with precision
- Problems with accuracy
- Problems with resolution
- Saturated variables

To obtain higher quality data, the data should be reconciled. This process adjusts the process variables so that they are consistent with the material and energy balances. There are many methods for doing this. Model based methods such as those used by Weiss et al. (1996a), Weiss et al. (1996b) and Romagnoli & Palazoglu (2006: 437–448) are popular.

Neural Networks are often frequently used (Amand et al., 2001a). Fuzzy logic techniques are also in use (Weiss et al., 1996a). Estimators such as the Kalman filter can also be used (Bai et al., 2006).

Techniques (such as those used by Du et al. (1997)) are closely related to the fault detection methods used here, are also common. This can be conceptualised as fault detection being applied to data directly rather than to the information from a process. Du et al. (1997) suggests that fault detection and data validation be handled in a combined analysis. This has the disadvantage of being generally unable to separate process and measurement faults. Amand et al. (2001b) shows data reconciliation and fault detection by means of principal component analysis being done together. It also prevents the easy application of techniques that reconcile data of poor quality rather than just flagging it.

Crowe (1996) contains a useful summary of some of the most important techniques.

1.5.3 Expert Systems

Expert systems are a kind of artificial intelligent system (Murrill, 2005: 260–261). They are used to solve specialised problems (e.g. for a specific kind of process using a certain control and DCS system). The are called expert systems due to the contention of their ability to solve problems at the same level of an expert in the field.

Expert systems consist of the following components:

1. User interface

2. A representation of the problem state (often called the global database)
3. A knowledge base (called rules)
4. The control regime (also called the interface engine)

Many expert systems are only used offline to diagnose problems. They are often used for:

- Equipment condition monitoring
- Diagnosis
- Advising
- Scheduling
- Alarm handling
- Hardware diagnosis
- Tuning
- Configuration

In the future, expert systems may become a far more integral part of the real-time control system. This will be useful because of the great degree of uncertainty and variance in chemical processes.

1.6 Overview of General Fault Detection and Diagnosis Methods

The scope of fault detection and diagnosis is broad and there are a huge number of techniques that have been developed over the years. As discussed before, fault detection consists of some kind of detection of a change feature characteristic. There are many methods of characterising data and processes. Diagnosis usually involves some kind of classification, of which there are also many techniques covering many fields of science and mathematics.

Due to the vast range of fields, it would be impossible to give any meaningful discussion of all of the important methods. The primary fields of focus by researchers are summarised in figure 1.10. The papers by Venkatasubramanian et al. (2003a) include a valuable overview of model based methods while Venkatasubramanian et al. (2003b) review process data based methods. Patel (2000) also compares a large number of fault detection techniques.

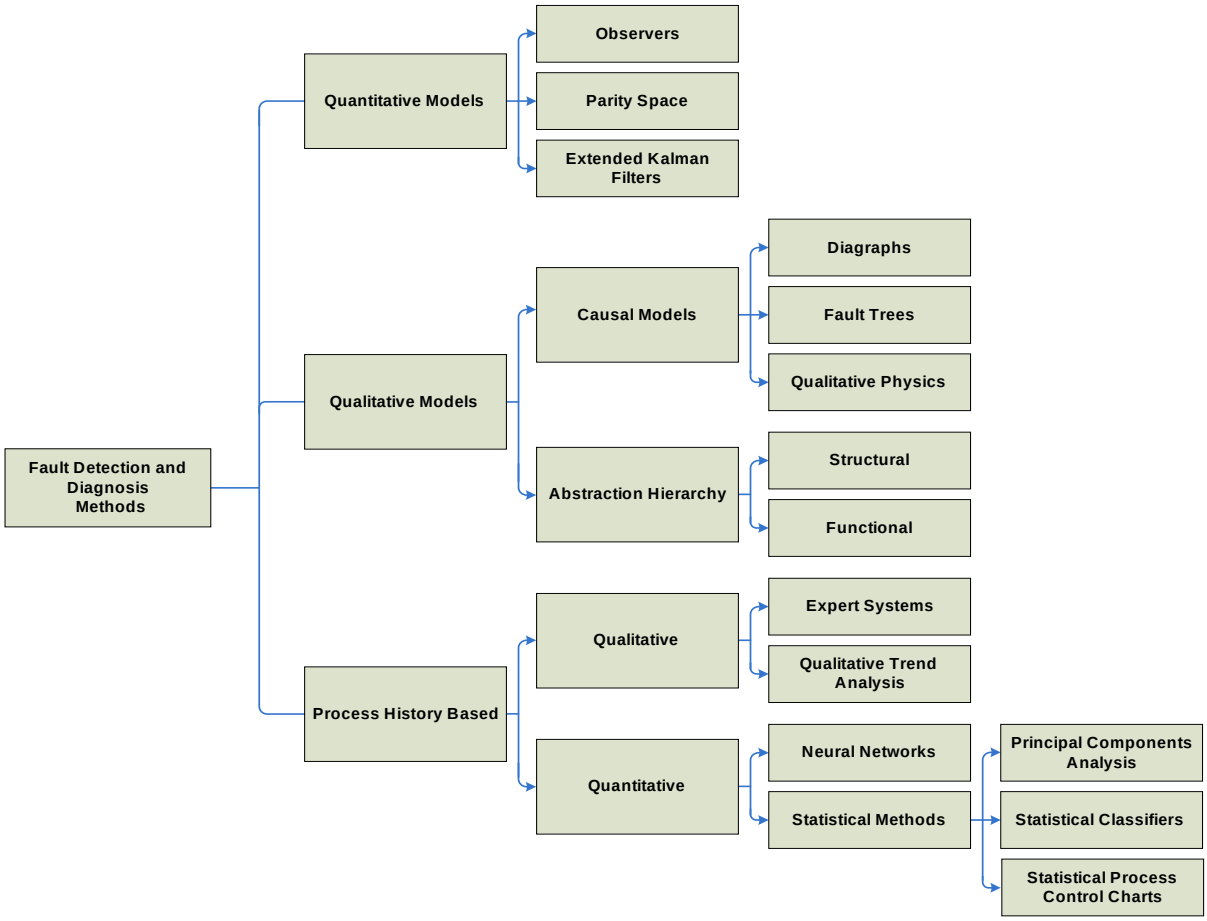

Figure 1.10: Fault Detection Methods

1.7 Model Based Methods

Many fault detection techniques detect faults based on differences between actual and expected behaviour. Quantitative models rely on analytical redundancy of explicit model of the system to diagnose the fault.

Observer based systems use a set of mathematical observers implemented in a model. Each observer is sensitive to a subset of faults while remaining insensitive to the remaining faults. This makes the diagnosis of multiple faults possible.

Parity space relations are generally rearranged variants of input-output models of a plant. Residuals of these relations are used to detect faults. The structure of the model can be used to diagnose or isolate the fault.

Kalman filters are used as state estimators. They are used to predict a current state from previous states. The current prediction is compared to the actual value. Kalman filters are successful in non-deterministic and/or noisy systems.

On the qualitative model side, models based on some fundamental understanding of the process are developed. This may be casual models or abstraction hierarchies. Casual models can be represented as digraphs, fault trees or as qualitative physics.

Abstraction hierarchies attempt to describe the behaviour of the system by means of the behaviour of its subsystems. The breakdown can be done according to the structure or the function of the subsystem.

1.8 Process Data Based Methods

In contrast to model based techniques, process data techniques only require large quantities of process data. This makes process data based methods ideally suited to processes that are well instrumented but complex to model. This is a characteristic of most modern chemical plants.

Extraction of features or characteristic of the process data can be done in one of two ways (Venkatasubramanian et al., 2003b). Qualitative methods include expert systems (section 1.5.3) and trend analysis.

Quantitative methods can be further classified into neural and statistical methods. Often the distinction between the two is blurred. An example of a hybrid of the techniques can be seen in Fourie (2000). Statistical methods have been widely used in the modelling and analysis of processes (Choi et al., 2004). Statistical methods include statistical process control charts (section 2.5), partial least squares (section 3.3.8) and principal component analysis (section 3.3) methods. Bersimis et al. (2005) gives a pleasant overview of multivariate statistical process control charts. These types of methods are well suited to real processes, stochastic in nature and subject to random disturbances. This makes the use of a system in a probabilistic setting reasonable.

1.9 Reasons for Implementing a Fault detection Programme

Faults can be regarded as an indicator of the quality of a process. Much has been said about quality with regard to the 'Japanese experience', 6-σ, quality circles etc. in the past decades. While there have been successes and failures, it seems there are some common motivations and factors affecting success (Owen, 1989: 317-328)

According to Owen (1989: 7-8) and Montgomery (1985: 7-8), there are five main reasons for requiring a fault detection programme. These are shown in figure 1.11.

External Pressure

While customer pressure is more likely to be present in industries where hard, discrete articles are required, there may be some amount of external pressure by customers of a continuous chemical process to increase quality by reducing variance and ensuring that product is delivered on specification and on time. Such external pressure may come from

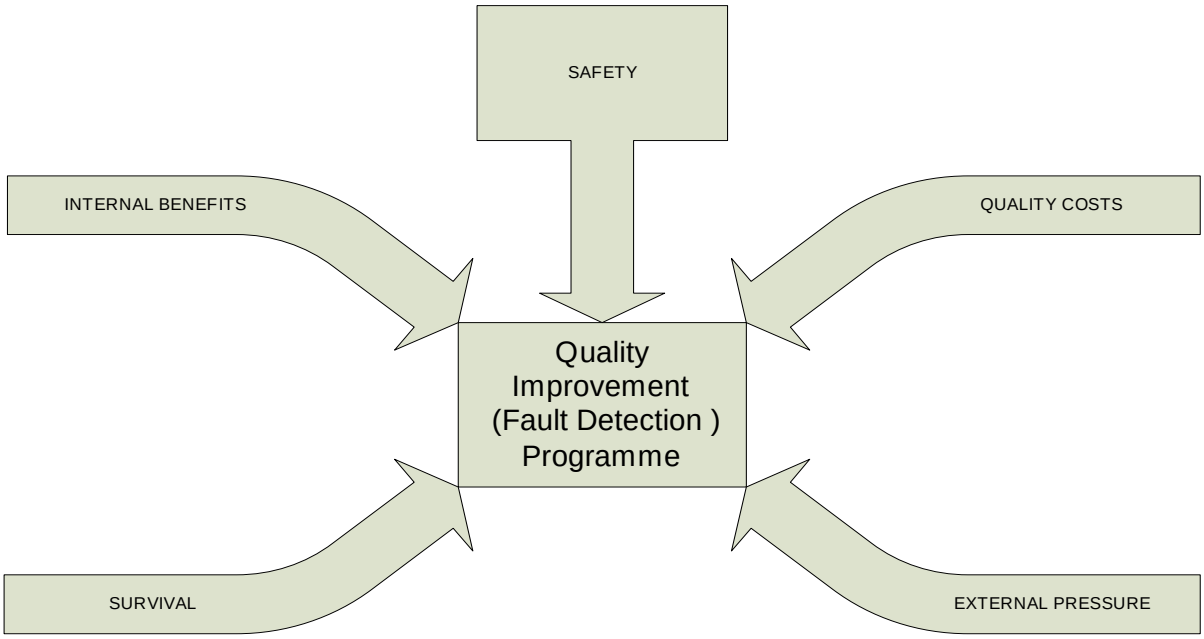

Figure 1.11: Reasons for Implementing a Fault Detection System

a down-stream process which requires fewer interruptions (possibly caused by faults) in a process stream. A downstream process may also request that variance in the product is reduced.

Survival

There is fierce competition in the chemical commodities market place. Organisations may be competing against other organisations on the other side of the globe. Processes are thus under pressure for management to perform well and profitably. The early detection and mitigation of faults, together with the eventual process improvements would be a critical part of guaranteeing process performance.

Also in situations where there is pressure to reduce waste and emissions, the early detection of faults that would cause emissions or large amount of scrap or flare would be advantageous.

Quality Costs

According to Montgomery (1985: 8), poor quality can cost a process in many ways. Faults can cause expensive damage or failure of equipment. Serious faults may result in liability for human or environment loss. A faulty process may be costlier to maintain.

Faults are also costly in terms of personnel and equipment required to manage and correct the faults.

Faults will also upset other operations (e.g. tuning or optimising) and generally reduce plant performance (with its associated costs).

Lack of quality may cause warranty, complaints, returns and liability from customers - possibly leading to the loss of external customers.

Internal failures will impact on profitability and capacity. The appraisal of quality is another cost.

It is likely (Owen, 1989: 9) that the quality costs will be higher than the costs of prevention of faults by some margin.

Safety

A fault could cause a potentially lethal and/or costly incident. Industrial statistics show that while major accidents are infrequent, minor accidents occur frequently, resulting in occupational illness and costing society billions of dollars every year (Bureau of Labor Statistics (1998) and Venkatasubramanian et al. (2003a)).

Fault detection and diagnosis by human operators is becoming increasingly difficult due to the broad scope of the activity. The majority of industrial incidents is caused by human error (Venkatasubramanian et al., 2003a). This would suggest the need for a dedicated fault detection or quality improvement programme.

Internal Benefits

According to Owen (1989: 8), internal benefits may include:

- More involved operators
- Fewer operators
- Increased process predictability
- Less downtime
- Increased capacity and/or performance
- Less scrap or waste
- Less maintenance

Owen (1989: 317-337) covers the establishment and pitfalls of a quality improvement programme in detail. This example is probably not completely relevant to modern systems.

The goals of this investigation are to investigate the ability of multivariate linear and kernel based techniques to detect and diagnose faults. The ideal fault detection and diagnosis system will have some or all of the attributes listed in section 1.3. The performance of the investigated techniques will be judged against these criteria.

Some definitions in summary:

1. *Fault* - Being the unpermitted departure of an observed variable or calculated parameter from acceptable behaviour. A fault will manifest as a change in some indication of the quality of the process.

2. *Detection* - Being the process of deciding if data should be flagged as faulty or not.

3. *Diagnosis* - Being the process of deciding on the cause (or set of causes) for the fault condition.

The remainder of the dissertation will follow the following structure:

Statistical Process Control: An overview of the key concepts of statistical process control. Traditional univariate control chart techniques are explained in detail in order to introduce the concept of SPC and the use of statistical measures to describe the quality of a process.

Multivariable Statistical Process Control: The reasons why univariate methods are not appropriate for multivariate correlated processes are given. A discussion of some multivariate data visualisation techniques is given in order to provide the reader with the background to interpret multivariate data sets and charts. Important techniques which allow the extraction of features that describe the data are derived and discussed. Their use for fault detection and diagnosis is covered in detail.

Classification and Discrimination: The difference between feature extraction and feature classification is discussed. The usefulness of data classification techniques for fault detection and diagnosis is explained.

Experimental Setup: The equipment used for the generation of data is described. The methodologies for the experiments to generate the normal and faulty data sets is also discussed. Addition explanations about the selection of variables and other experimental details are given.

Results: The results of the application of the linear and kernel based feature extraction and classification techniques to the faults sets are given.

Conclusions and Recommendations: Conclusions regarding the use, results and success of the various fault detection and diagnosis techniques are made. Recommendations regarding the directions for future work are also given.

CHAPTER 2

Statistical Process Control

Statistical process control (SPC) emerged in the late 1980's (Murrill, 2005: 255-260). At that time, the running of processes was mainly concerned with the adherence of a final product to specification. SPC techniques paved the way for the constant monitoring of process quality indicators, which in turn could lead to methodologies to control quality at every stage of the manufacturing process.

Initially, the principles of SPC were derived and applied to process involved with discrete manufacturing (such as hard article manufacture). In such processes, attributes directly related to the specifications of each article can easily be measured. Concepts such as scrap or zero defects can be used. This is clearly in contrast to continuous chemical processes where a quality indication typically has a continuous distribution. These chemical processes are often inherently complex, interconnected, nonlinear flow systems. Here, the concept of a variable has more meaning. Much of the control of continuous processes involves the systematic adjustment of the mass and energy balances.

SPC makes use of statistical measures of the quality of a process. These indicators can be used for fault detection and diagnosis.

Importantly, standard SPC assumes that there are no mechanisms causing auto-correlation between the variables being monitored. This will seldom be an accurate assumption in chemical processes which are fundamentally interactive. Additionally, the presence of control loops guarantees some measure of auto-correlation.

SPC is important in both the long and short term. On a short term basis, SPC can identify process problems. It is from this point of view that SPC and some of its techniques are covered in this dissertation. Any variance that is considered statistically normal can be accepted and any unexpected variance is probably due to some fault or change that can be identified by various techniques.

On the long term, SPC can be used to quantify improvements.

Conventional process control often assumes that process relationships are inherently deterministic. SPC assumes that the relationship is stochastic.

Macgregor & Kourti (1995) gives a good overview of some of the traditional applications of SPC for fault detection on chemical processes. These applications are discussed in more detail in the later sections.

2.1 Sources of Variance

Industrial processes are inherently non-stationary.

Typical sources are (Ott, 1975b: 3) and (Wetherill & Brown, 1991: 39):

- *Variational Noise* - This is the variation observed for measurements taken under the same conditions and specifications.
- *External Sources* - Variation from external sources will affect variance of the process.
- *Process Sources* - Any change in the condition (in the methods, people, materials and equipment) of the process itself.
- *Assignable Causes* - Known faults or changes, noise or disturbances.

2.2 States of Control

In traditional control, a process is said to be in control when the system is stable and the sensor values are close to the target values (often setpoints). In contrast, a process under statistical control is said to be in control when some sampled (and possibly transformed version of a) process measurement exhibits a certain (usually normal) distribution and is within some range or window (Murrill (2005: 255) and Box & Luceño (1997: 13)).

The 'in-control state' means that the output is varying in a stable manner about the mean. Any observed autocorrelation does not necessarily indicate that a process is not in a state of statistical control. This state of control can be referred to as an autocorrelated state of control

2.3 The Goals of Fault Detection by means of SPC

As stated by Fourie (2000: 7–12), the goals of any fault detection procedure that makes use of SPC principals are the following:

1. A "Yes"/"No" answer to the the question: "Is the process in control (in this case: not faulty)?"

2. A classification error estimate or probability of a type 1 error (the probability of the process being flagged as out of control incorrectly) should be specified.

3. Provide a procedure which takes into account the relationships between the variables.

4. A diagnosis of the fault.

With echniques that help attain goal 4, the amount of expert information required increases with the number of variables (Fourie, 2000: 7–12). The final goal is definitely the most difficult.

2.4 The Assumptions of Normality and Linearity

Many of the techniques discussed in this section make assumptions of linearity and/or normality. For a detailed discussion on the repercussions and robustness of the linear assumption as well as on methods for evaluating linearity, please refer to section 3.4.1.

Many statistical methods (particularly the univariate techniques) are generally very robust under violation of assumptions of normality (Harris, 1985: 331).However, there are exceptions. Little is known about the exact effects of non-normality on multivariate techniques (Jackson, 1991: 325–328). Evaluating this robustness is difficult to do theoretically. The only easy way is to perform empirical comparisons.

There are several sources of non-normality in chemical process data, including (Shunta, 1995):

- Outliers
- Sensors with sensitivity problems
- Variable constraints
- Nonlinearity
- Final control element problems like stiction

While many of these sources may indicate a fault, the source of non-normality may also be a part of a fault-free process.

If normality is shown not to be a viable assumption, the following are options (Johnson & Wichern, 1982: 151–165):

1. Proceed as normal - ignoring the non-normality.

2. Perform an initial analysis with the normal assumption, test it and decide if a modified analysis is necessary.

3. Transform the data points so that they are more normal (possibly using one of the methods described below).

Methods of transforming the data include taking their square roots, applying logarithmic functions or using Fisher correlations (Gifi, 1990). The principal component analysis generally increases the normality of datasets as it involves combinations of more than one variable (increasing the possibility of multivariate normality).

There are several tests for normality, including: the Kolmogorov-Smirnov, Lilliefors, Anderson-Darling, Shapiro-Wilk and Jarque-Bera tests (Gifi, 1990). More simple tests involve simply evaluating the skewness and kurtosis or comparing the mean, mode and median (which should be close together for normality). The normality index is also an estimate of the normality of the data. A simple graphical method to use is the probability plot method. These methods are discussed below.

2.4.1 Normality Index

The normality index gives the degree to which the data are normal (Jones, 2005).

It is calculated as follows:

1. Calculate the mean (μ) and standard deviation (s) of the data.
2. Calculate the distribution of the data - and bin the frequency.
3. Minimise the sum of the square errors between the fitted normal distribution (having mean $\hat{\sigma}$ and the binned data.

The means square error as well as the index:

$$\frac{|\mu_{data} - \hat{\mu}|}{s_{data}} \tag{2.1}$$

are indications of the normality of the data. A low index suggests normality.

2.4.2 Probability Plots

To test for normality, data can be plotted against a known theoretical distribution. This makes a normal probability plot a useful visualisation of the normality of the distribution of the data. An example is shown in figure 2.1. Here a linear relationship indicates that the data are strongly normal. Any curvature would suggest that the data would be better modelled by some other probability density function.

The theoretical probability (from the chosen distribution) is shown on the vertical axis, while the data are shown on the horizontal axis.

It may also be useful to compare the two distributions by graphing the quartiles (the division points of a ordered set of data divided into 4 parts containing the same

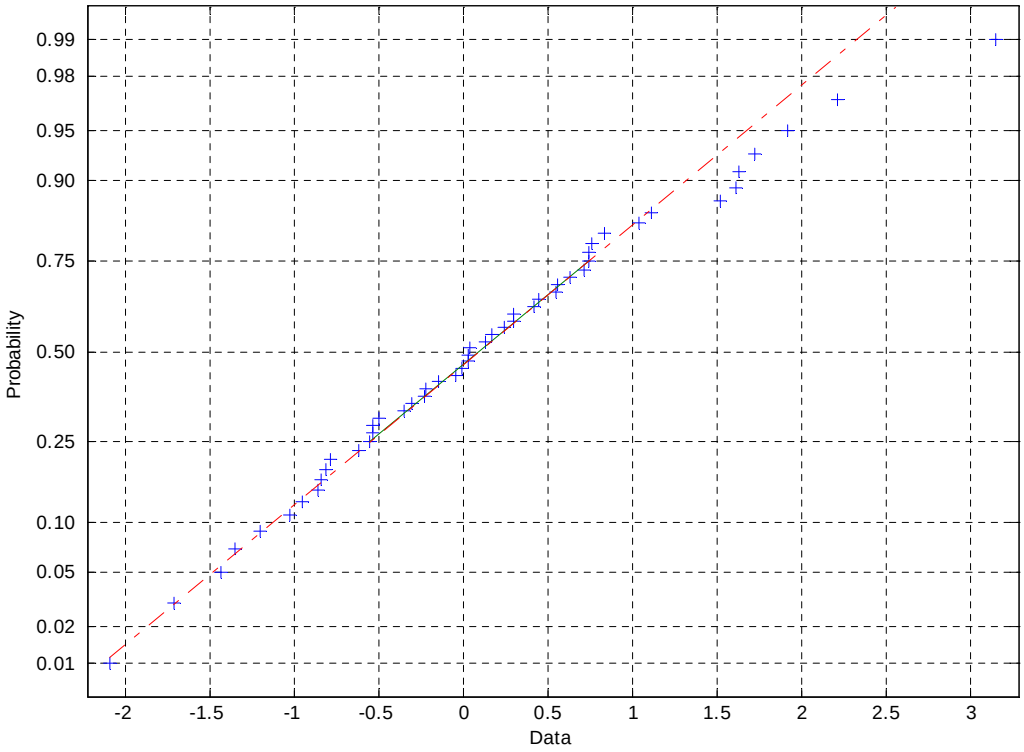

Figure 2.1: Probability Plot using a Normal Pseudo-Random Variable

number of data points) of one variable against the quartiles of the other. These are called quantile-quantile plots (Martinez & Martinez, 2004: 281).

2.5 Statistical Process Control Charts

Control charts have traditionally been used as a way to monitor process performance with a view to fault detection. They represent the first attempt at using statistics for monitoring and change detection (Venkatasubramanian et al., 2003b). A change in the statistical performance of a plant may signify a fault. Action can then be taken to identify and correct the fault.

Control charts should indicate:

1. Whether a process change has occurred.

2. Possible evidence to the cause of that change.

For these reasons, control charts are directly useful for fault detection.

2.5.1 Guidelines on Variable Selection and General use of Control Charts

At the outset of a fault detection program, it may be difficult to determine which characteristics should be included. Some guidelines are provided by Montgomery (1985):

1. At the outset of the control chart program, control charts should be applied to any process characteristic that could be important. It will soon become apparent which charts are important.

2. Re-evaluate the selection by removing any charts that are unnecessary and adding ones that operators find necessary.

3. The number and type of chart should be recorded. The number and type of chart will stabilise as the new process stabilises.

4. It is generally found that if control charts are being used effectively, the number of charts will increase as new knowledge about key process variables is gained.

5. At the beginning of the control chart program it is usual to have more attribute control monitoring the final products to see if they meet specification. As the program matures, these charts will be replaced with charts earlier in the process where the faults have been identified to occur. Ultimately, control charts could be implemented at the supplier or process input level.

6. Control charts should be implemented in an online manner. The results should be displayed as close to the section of the process being examined as is possible to ensure that feedback is rapid. Computer systems are obviously a huge advantage.

7. Operators must have responsibility for maintaining and interpreting the results so that the fault detection system is shaped by the process knowledge that these personnel have.

Here follows an outline of many of the most important control charts:

2.5.2 Time Series Plots

Time series plots are simply a representation of the process metric as it changes with time. This method forms the very basis of any any process monitoring. Ott (1975a) claims that to benefit from any data coming from a process, the important and basic rule is to *plot the data in a time series*. After this has been done, other general methods can be applied to the data.

While a simple time series may show important information about a process, many faults may not be apparent in most cases.

Time series plots are useful in showing (in a controlled system) the control error. The sudden appearance of a large control error (in the absence of any other changes to the process could indicate a control system fault.

Interpretation of Direct Time Series Plots

It is easy to directly see (or calculate) the appearance of

1. Gross error
2. Shift in average
3. Change in variability
4. Gradual change in average (the development of a trend)
5. Cycling or oscillatory behaviour

2.5.3 Shewhart Control Charts

The Shewart chart is also known as the $\overline{x}$ chart. The idea of the control chart was developed by Shewart in 1924. He suggested that a control chart having three objectives:

1. Define the standard which the process should strive to attain.
2. Aid in attaining the standard.
3. A basis for judging whether the standard has been achieved.

According to Wheeler (1990: iv) the four foundations of control charts are:

1. Shewhart's control charts will always use 3-σ limits
2. The standard deviation used to calculate the limits must be calculated from the average variance within the sub-groups. No other kind of estimate (example: Hartley's Conversion Constant) is acceptable.
3. The data will be collected in an organised manner. The organisation of the data must respect the context of the problem which the control chart is intended to address.
4. The process can and will utilise the information gained from the control charts

A Shewart chart is used to control the mean and the variance of a process (Murdoch, 1979: 36–55). The design of the Shewhart chart ensures that it is unlikely normal process variability will cause a false alarm. The Shewart chart is a chart of an average of the

process taken at a regular intervals. The statistical basis of the chart is (Montgomery, 1985: 171–193):

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2.2}$$

for a sample of size n. This is a simple average of a subgroup. An example of a Shewart chart is shown in figure 2.2. Here we can see that limits are applied to the data. Faulty samples (violating the limits) are marked. The data points are made up of the average of each sampled subgroup.

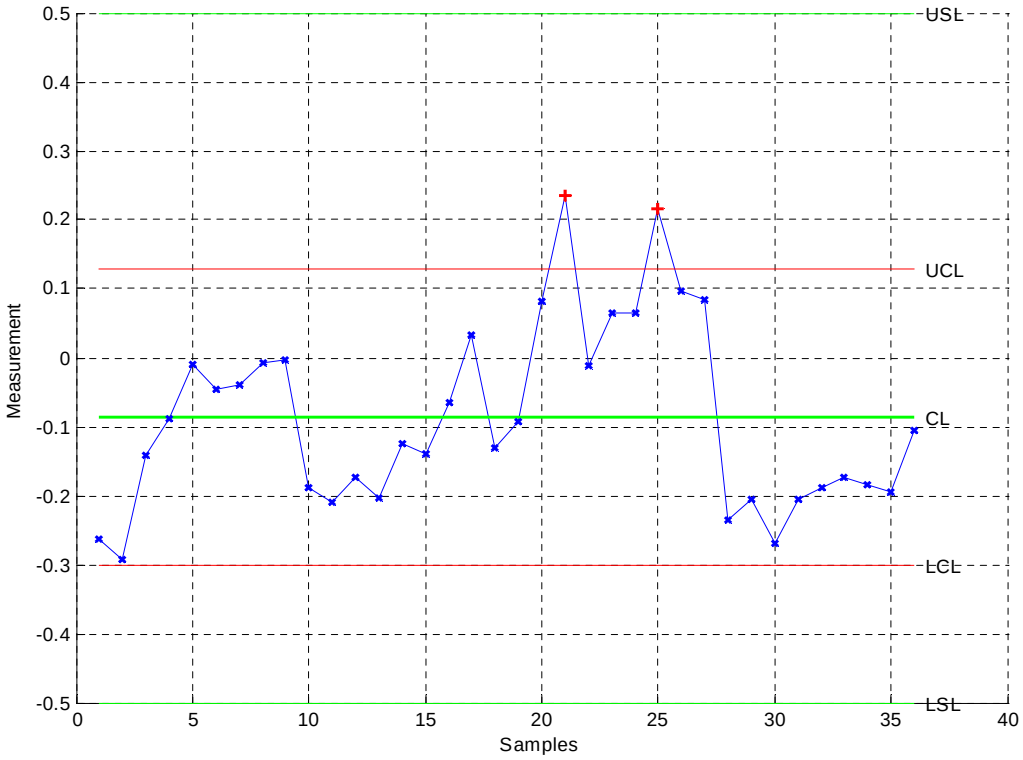

Figure 2.2: An Example of a Shewart Chart

Limits For Shewart Charts

Control Limits

In American literature, control limits are also referred to as action limits. These action limits should only be violated about every 200-1000 sample groups when the process is in statistical control (Bissel, 1994: 105). This prevents over control. Taking control action when the system is not faulty, and the true reason for overstepping the control limits can be attained to random variation, will often decrease performance. This is similar to

underdamped control loops

Often upper and lower specification limits are shown. It is often tempting to assume if an out of control subgroup still falls within product specification, that it is of no concern. This is often not the case as the specification limit generally applies to an individual measurement and not the average of a group of samples (particularly in manufacturing processes).

If the characteristic variable is normally distributed with a mean μ and a standard deviation σ, then the standard deviation of $\overline{x}$ is $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. The probably is $1 - \alpha$ that a sample mean will fall between:

$$\mu + Z_{\alpha/2}\sigma_{\overline{x}} = \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.3}$$

and

$$\mu - Z_{\alpha/2}\sigma_{\overline{x}} = \mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{2.4}$$

$$\tag{2.5}$$

These can then be used as control limits. In practise μ and σ are not known. The following methods can be used to select a value for μ:

- Estimation from preliminary samples (Montgomery, 1985: 173).
- Using an obvious aim value. These may include nominal values; some mean level comfortably away from a specification limit or some technologically sound optimum point (Bissel, 1994: 107–108).
- Some other estimate of the mean of the process.

σ may be found as follows:

- Estimation from preliminary samples (Montgomery, 1985: 173).
- Using the range of the data to select σ.
- Some other estimate of the variance of the process.

The above equations can be simplified by using standard values. This practice is more common in the US than in the UK. A common value used are 3-σ limits. This equates to an type 1 risk (the risk of a false out of control signal) of 0.27% A more common practise in the UK is to specify an average run length (ARL). ARL is defined as:

$$ARL = \frac{1}{\alpha} \tag{2.6}$$

with α the probability of describing normal variance as out of control (a type 1 error).

A commonly used value is for 1 in 370 samples resulting in a false alarm. This equates to limits at 3.09 $\sigma_x/\sqrt{n}$ distances from the centre line, which is very close to the US practise (Bissel, 1994: 116). According to the US practise, the parameters of the $\overline{x}$ chart are:

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}} \tag{2.7}$$

$$CL = \mu \tag{2.8}$$

$$UCL = \mu + 3\frac{\sigma}{\sqrt{n}} \tag{2.9}$$

with UCL, CL and LCL being the upper control, centre and lower control lines or limits respectively.

Warning Limits

A common practise in especially Europe is to make use of upper and lower warning limits (Bissel, 1994: 116). To prevent over control, action is not taken when these lines are crossed. Usually the operator will remain vigilant and recheck the sample (for manufacturing processes) or to wait vigilantly for the next set of data (in continuous processes).

Warning lines set at 1.96 $\sigma_x/\sqrt{n}$ away from the centre line are commonly used. These correspond to an ARL of 40 samples.

Control Chart Heuristics for Shewart Charts

The Nelson Rules (Nelson, 1984)

The Nelson Rules are a method to determine if variability in a control chart is caused by the random variability inherent to the process or if it is due to some special cause (possibly a fault).

If one of the following occur:

1. 1 point is more than 3 standard deviations from the mean in either direction.
2. More than 8 points in a row are on the same side of the mean.
3. More than 6 points in a row are monotonically increasing or decreasing.
4. The changes of more than 13 points in a row alternate in direction.
5. More than 1 out of 3 points in a row are more than 2 standard deviations from the mean in the same direction.

6. More than 3 out of 5 points in a row are more than 1 standard deviation from the mean in the same direction.
7. 15 points in a row are within 1 standard deviation of the either side of the mean.
8. 8 points in a row are all more than 1 standard deviation from either side of the mean.

then the following interpretation (for each rule) can respectively be made:

1. 1 sample is likely to be grossly out of control.
2. Some extended bias is likely to exist.
3. Some trend is likely to exist.
4. It is likely oscillation above random levels exists.
5. There is a medium tendency for the samples to be out of control to some degree.
6. There is a strong tendency for the samples to be slightly out of control.
7. Greater variation is expected
8. Jumping while missing the first standard deviation bands is unlikely.

The Western Electric Rules (Western Electric, 1954)

These rules are similar to the Nelson Rules. They also indicate if it is likely if a special cause of variability above that which can be expected of a statistically controlled process is present.

The rules are as follows:

1. 1 point is more than 3 standard deviation from mean in either direction.
2. More than 1 out of 3 points in a row are more than 2 standard deviations from the mean in the same direction.
3. More than 3 out of 5 points in a row are more than 1 standard deviation from the mean in the same direction.
4. More than 8 points in a row are on the same side of the mean.
5. 15 points in a row are within 1 standard deviation of the either side of the mean.
6. More than 6 points in a row are monotonically increasing or decreasing. This rule is only occasionally used.

The following interpretations can be made with respect to each of The Western Electric Rules:

1. 1 sample is likely to grossly out of control.
2. There is a medium tendency for the samples to be out of control to some degree.
3. There is a strong tendency for the samples to be slightly out of control.
4. Some extended bias exists.
5. Greater variation is expected.
6. Some trend is likely to exist.

Each of the Western Electric and the Nelson Rules have roughly the same probability of occurring (approximately 0.3%) with completely random data. For some processes it may be useful or necessary to omit one or more of the rules. For example, rule 5 of the Western Electric rules (or equivilently rule 7 of the Nelson Rules) would trigger a high number of false alarms in a process where the a major source of variance is intermittent.

The Owen Rules (Owen, 1989: 115–118)

The Owen Rules are similar to the Western Electric or Nelson Rules:

1. 1 point is more than 3 standard deviation from mean in either direction.
2. The Rules of Seven: Seven consecutive points either:
   - On the same side of the mean.
   - Monotonically increasing or decreasing
3. Unusual patterns or trends:
   - Bunching (for an example see figure 2.3)
   - Drifting (figure 2.4)
   - Jumps (figure 2.5)
   - Cyclic Patterns (figure 2.6)
   - Stratification (figure 2.7)
4. Middle Third Rule: The number of points in the middle third of the area between the 3 standard deviation control limits is much greater or much less than the total number of points present.

The following interpretations can be made with respect to each of The Owen Rules:

1. 1 sample is grossly out of control.
2. Some trend is likely to exist.
3. It is likely there is a non random source of variance.
4. It is likely that a non-normal distribution exists.

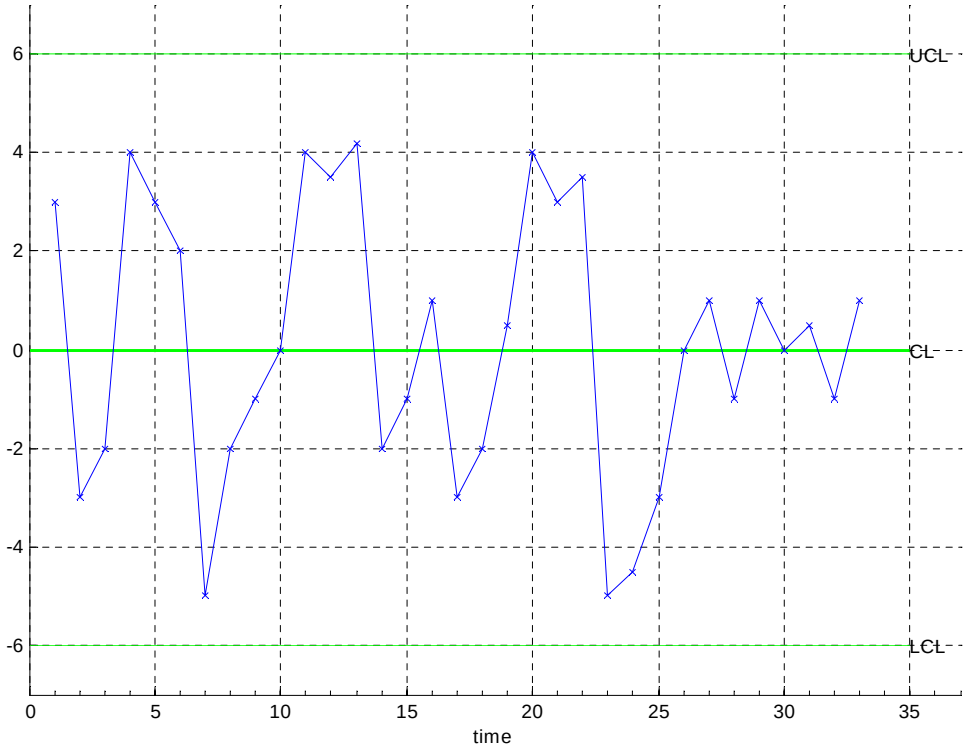

Figure 2.3: Owen Rule 3 - Bunching

The last of the Owen Rules are based on the properties of the normal distribution. It is known that according to this distribution that 68.3% of the data will fall in the middle third of the chart. If this is not the case there has been some shift in the process.

2.5.4 Modified Control Charts

One of the assumptions made in the design and application of the Shewart charts is that the tolerance of the process due to random variation is of a similar order of magnitude as the specification limits. This means that the process is assumed to operate near the limits attainable taking into account the variance inherent to the process.

There are many situations where the spread due to variance is significantly smaller than the spread in the specification limits (i.e. the difference in the upper and lower specification limits: $USL - LSL$). In these situations, a modified control chart can

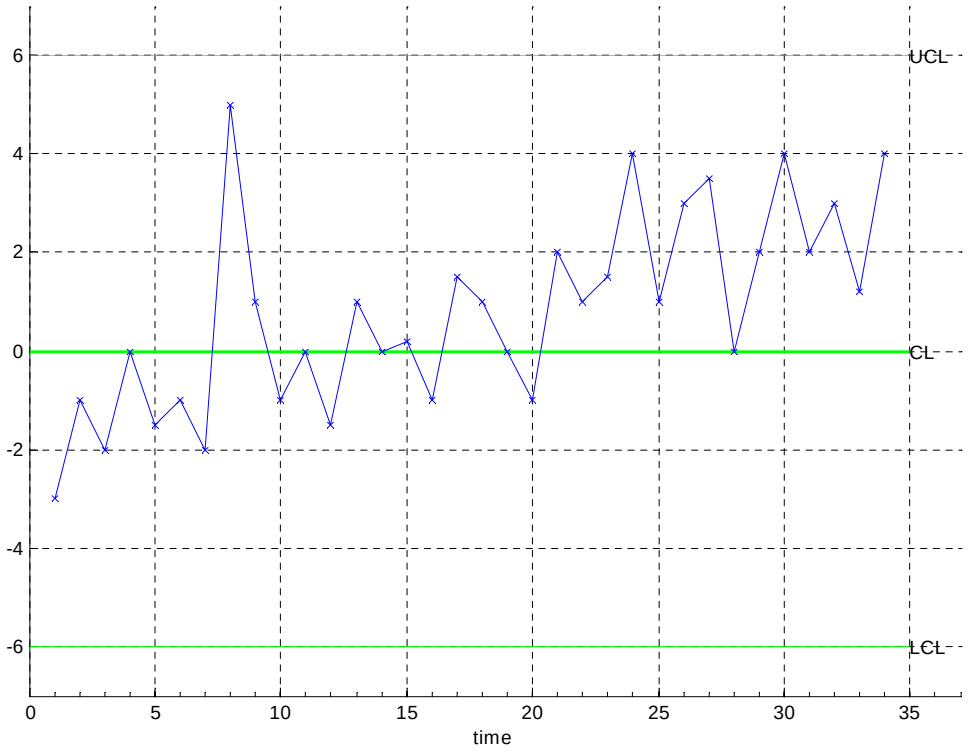

Figure 2.4: Owen Rule 3 - Drifting

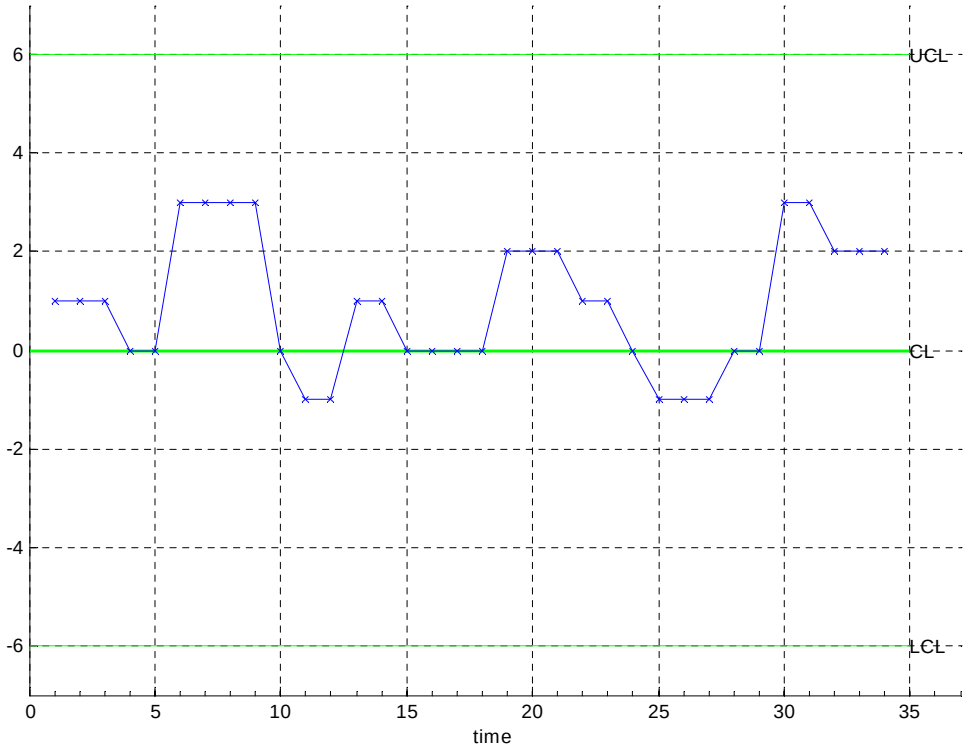

Figure 2.5: Owen Rule 3 - Jumps

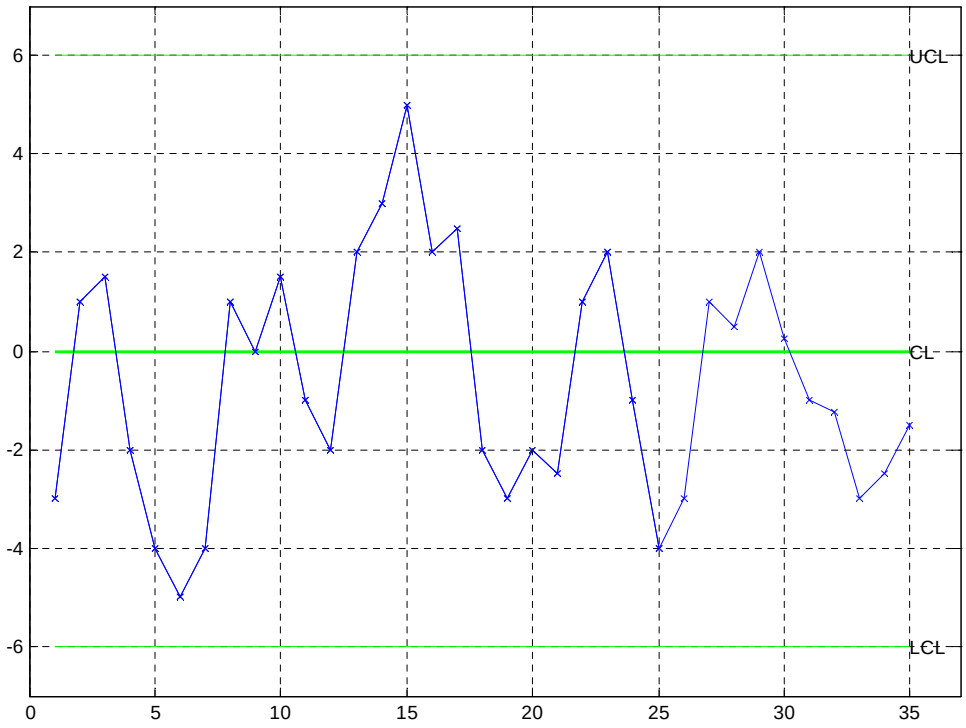

Figure 2.6: Owen Rule 3 - Cyclic Pattern

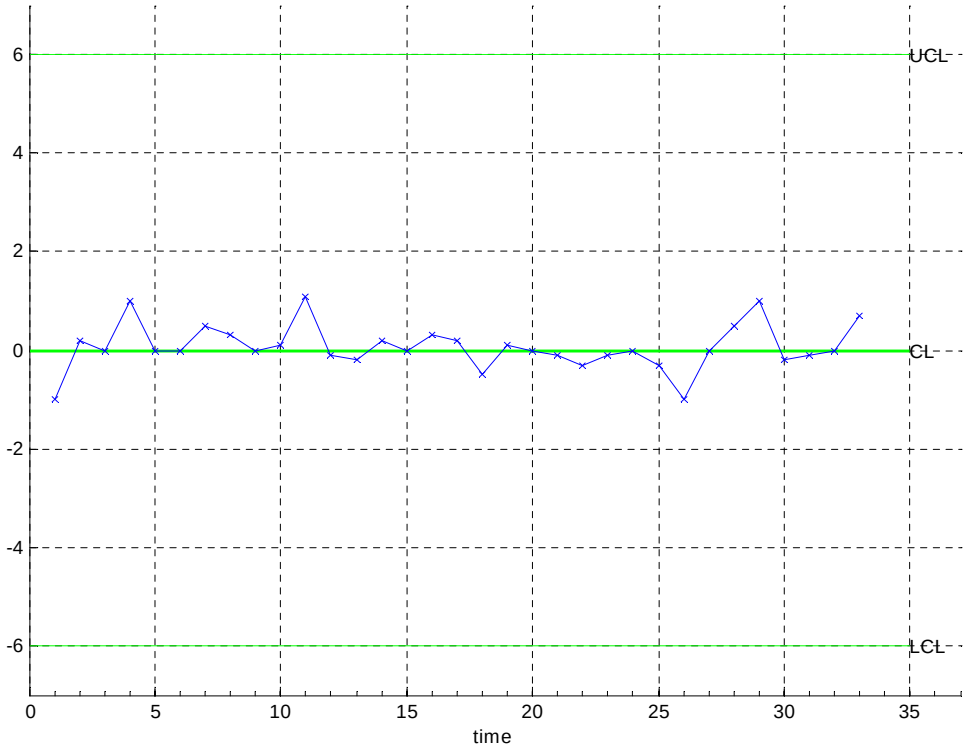

Figure 2.7: Owen Rule 3 - Stratification

be used (Montgomery, 1985: 221–225). The modified control chart is only concerned with detecting where the true process mean (μ) is situated so that it can be seen if the process is producing a faults exceeding a specified value of δ. The assumption made is the process variability (σ) is under control and that the data are normally distributed. The derivation of the chart is changed to assuming normal distributions exist either side of both the LSL and the USL.

For a specified δ, the true process mean must lie between μ_L and μ_U as shown in figure 2.8.

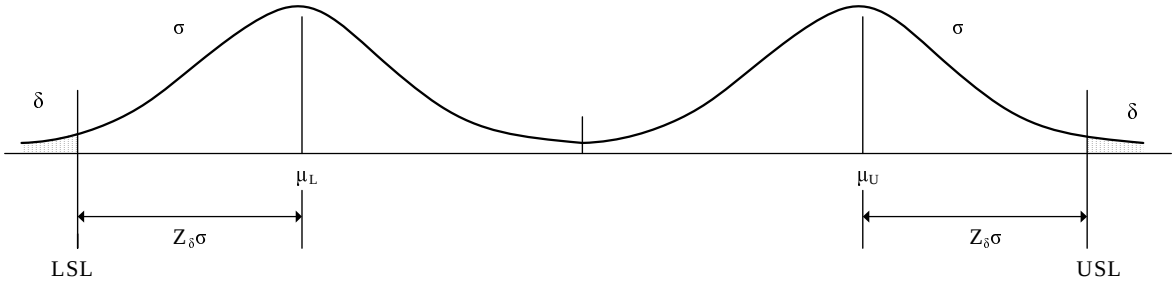

Figure 2.8: Distribution of Process Output

So:

$$\mu_L = LSL + Z_\delta \sigma \tag{2.10}$$

$$\mu_U = USL + Z_\delta \sigma \tag{2.11}$$

with Z_σ the upper $100 \cdot (100 - \delta)$ percentage point of the normal distribution. Clearly, an accurate estimate of σ must be available. As stated above, the process σ must be in control. It is thus wise to use an R or S chart (discussed below) in conjunction with this technique, both to get an estimate of the σ and to check that it is in control.

So, specifying a type 1 error corresponding to $3 - \sigma$, we get:

$$LCL = LSL + \left(Z_\alpha - \frac{3}{\sqrt{n}} \right) \sigma \tag{2.12}$$

$$UCL = USL - \left(Z_\alpha - \frac{3}{\sqrt{n}} \right) \sigma \tag{2.13}$$

for the control limits.

The modified control chart will look similar to a $\overline{x}$ chart with different control limits.

Some quality engineers recommend the use of 2-σ limits. This is rationalised by arguing that the tighter limits offer increased protection against process shifts with little increase in the type-1 risk (the risk of taking action due to misdiagnosis of random variance) (Montgomery, 1985: 224).

2.5.5 Moving-Range Charts

Moving range charts are an additional method to maintain control over both the process mean and the process variability. Moving range charts (referred to as R charts) are used as an indication of the variability or dispersion of a process (Montgomery, 1985: 171–196). The R chart is used more frequently than the the S chart (see section 2.5.6). This is because it gives a reasonable measure of the process variability while still being very easy to calculate.

The range of a process sample of size n is an elementary estimation of the process variability. The moving range is defined as:

$$R = x_{max} - x_{min} \tag{2.14}$$

So, with $R_1, \ldots, R_n$ as the range for n samples, the average range is:

$$\overline{R} = \frac{R_1 + R_2 + \cdots + R_n}{n} \tag{2.15}$$

The range can be used as an estimate of the variability as follows:
There exists a random variable W such that:

$$W = \frac{R}{\sigma} \tag{2.16}$$

so, an estimate of σ is:

$$\hat{\sigma} = \frac{\overline{R}}{d_2} \tag{2.17}$$

This method of estimation is particularly useful for low sample sizes (low n). For $n = 2$, it is 100 % efficient (as compared to the quadratic estimator S^2). It is still 85 % efficient with $n = 10$. It losses efficiency at high n because it ignores all information about the sample variance between x_{min} and x_{max}. It is, however much simpler to calculate. Table 2.1 shows some other values of the relative efficiency.

Table 2.1: Relative Efficiency for using the Sample Range as an Estimate of σ

n	Relative Efficiency
2	1.000
3	0.992
4	0.975
5	0.955
6	0.930
10	0.850

According to equation 2.16, $R = W\sigma$, since the standard deviation of R is:

$$\sigma_R = d_3\sigma \tag{2.18}$$

If σ is unknown, σ_R can be estimated by:

$$\hat{\sigma}_R = d_3\frac{\overline{R}}{d_2} \tag{2.19}$$

The 3-σ limits for an R-chart are:

$$UCL = \overline{R} + 3\hat{\sigma}_R = \overline{R} + 3d_3\frac{\overline{R}}{d_2} \tag{2.20}$$

$$CL = \overline{R} \tag{2.21}$$

$$LCL = \overline{R} - 3\hat{\sigma}_R = \overline{R} - 3d_3\frac{\overline{R}}{d_2} \tag{2.22}$$

Replacing:

$$D_3 = 1 - 3\frac{d_3}{d_2}$$
$$D_4 = 1 + 3\frac{d_3}{d_2}$$

the parameters for the control limits of the R chart with unknown σ become:

$$UCL = \overline{R}D_4 \tag{2.23}$$

$$CL = \overline{R} \tag{2.24}$$

$$LCL = \overline{R}D_3 \tag{2.25}$$

Values for D_3 and D_4 are shown in table 2.2.
Alternatively, if the variable standard deviation is known, the control limits become:

$$UCL = d_2\sigma + 3d_3\sigma \tag{2.26}$$

$$CL = \overline{R} = d_2\sigma \tag{2.27}$$

$$LCL = d_2\sigma - 3d_3\sigma \tag{2.28}$$

Replacing:

$$D_1 = d_2 - 3d_3$$
$$D_2 = d_2 + 3d_3$$

the parameters for the control limits of the R chart with known σ (also called the R

Table 2.2: *R*-Chart Control Limit Parameters with a known σ

Sample Size (n)	Factors for Control Limits	
	D_3	D_4
2	0	3.267
3	0	2.57
4	0	2.28
5	0	2.114
6	0	2.004
7	0.076	1.924
8	0.136	1.864
9	0.184	1.816
10	.223	1.777
11	0.256	1.744
12	0.283	1.717
13	0.307	1.693
14	.0328	1.672
15	0.347	1.653
16	0.363	1.637
17	0.378	1.622
18	0.391	1.608
19	0.403	1.597
20	0.415	1.585
21	0.425	1.5775
22	0.434	1.566
23	0.443	1.557
24	0.451	1.548
25	0.459	4.541

chart based on standard values) become:

$$UCL = D_2\sigma \tag{2.29}$$

$$CL = \overline{R} = d_2\sigma \tag{2.30}$$

$$LCL = D_1\sigma \tag{2.31}$$

Values for D_1,D_2 and d_2 are shown in table 2.3.

2.5.6 *S* Charts

With sample sizes of $n > 10$ or $n > 12$, the range method (see section 2.5.5), becomes less statistically sound. When this occurs, it is better to replace the R charts with S (occasionally referred to as a σ chart).

Table 2.3: R-Chart Control Limit Parameters with unknown σ

Sample Size (n)	Factors for Control Limits		
	D_1	D_2	d_2
2	0	3.686	1.128
3	0	4.358	1.693
4	0	4.698	2.059
5	0	4.918	2.326
6	0	5.078	2.534
7	0.204	5.204	2.704
8	0.388	5.306	2.847
9	0.547	5.393	2.970
10	0.687	5.469	3.078
11	0.811	5.535	3.173
12	0.922	5.594	3.258
13	1.025	5.647	3.336
14	1.118	5.696	3.407
15	1.203	5.741	3.472
16	1.282	5.782	3.532
17	1.356	5.820	3.588
18	1.424	5.856	3.640
19	1.487	5.891	3.689
20	1.549	5.921	3.735
21	1.605	5.951	3.778
22	1.659	5.979	3.819
23	1.710	6.006	3.585
24	1.759	6.031	3.895
25	1.806	6.056	3.931

S^2 is an unbiased estimate of the variance of the distribution (σ^2). S^2 is defined as follows (Montgomery, 1985: 197–202):

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2.32}$$

Control Limits for *S* Charts

If a standard value of σ is available, then the standard 3-σ limits are directly:

$$UCL = B_6\sigma \tag{2.33}$$
$$CL = c_4\sigma \tag{2.34}$$
$$LCL = B_5\sigma \tag{2.35}$$

B_6 and B_5 are customarily defined. Values for various sub-group sizes are given in in table 2.4.

for $n > 25$:

$$c_4 \approx \frac{4(n-1)}{4n-3} \tag{2.36}$$
$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}} \tag{2.37}$$
$$B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}} \tag{2.38}$$

These values can b obtained from table 2.4. Alternatively, if no σ is known, then σ must be estimated from previous data. With S_i as the standard deviation of the i^{th} sample:

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{2.39}$$

It can be shown that the control limits then become:

$$UCL = B_4\overline{S} \tag{2.40}$$

$$CL = \overline{S} \tag{2.41}$$

$$LCL = B_3\overline{S} \tag{2.42}$$

$$with:$$

$$B_3 = \frac{B_5}{c_4}$$

$$B_4 = \frac{B_6}{c_4}$$

The values for B_3 and B_4 are also shown in table 2.5.
for $n > 25$:

Table 2.4: S-Chart Control Limit Parameters with a known σ

Sample Size (n)	Limit Factors		
	B_5	B_6	c_4
2	0	2.606	0.7979
3	0	2.276	0.8862
4	0	2.088	0.9213
5	0	1.9764	0.9400
6	0.029	1.874	0.9515
7	0.113	1.806	0.9594
8	0.179	1.751	0.9650
9	0.232	1.7070	0.9693
10	0.276	1.669	0.9727
11	0.313	1.637	0.9754
12	0.346	1.610	0.9776
13	0.374	1.585	0.9794
14	0.399	1.563	0.9810
15	0.421	1.544	0.9823
16	0.440	1.526	0.9835
17	0.458	1.511	0.9845
18	0.475	1.496	0.9854
19	0.490	1.783	0.9862
20	0.504	1.470	0.9869
21	0.516	1.459	0.9876
22	0.528	1.448	0.9882
23	0.539	1.438	0.9887
24	0.549	1.429	0.9892
25	0.559	1.420	0.9896

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}} \tag{2.43}$$

$$B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}} \tag{2.44}$$

with c_4 defined as before.

2.5.7 Weighted Average based Control Charts

Weighted average plots represent some kind of average of the last k groups of size m (Wetherill & Brown, 1991: 123–124). They are useful because Shewart charts are relatively insensitive to small shifts in process mean (Montgomery, 1985).

Table 2.5: S-Chart Control Limit Parameters with a unknown σ

Sample Size (n)	Limit Factors	
	B_3	B_4
2	0	3.276
3	0	2.568
4	0	2.26
5	0	2.089
6	0.30	1.970
7	0.118	1.882
8	0.185	1.815
9	0.239	1.761
10	0.284	1.716
11	0.321	1.679
12	0.354	1.646
13	0.382	1.618
14	0.406	1.594
15	0.428	1.572
16	0.448	1.552
17	0.466	1.534
18	0.482	1.518
19	0.497	1.503
20	0.510	1.490
21	0.523	1.477
22	0.534	1.466
23	0.545	1.455
24	0.555	1.445
25	0.565	1.435

The Moving-Average Chart

Montgomery (1985: 235–239) discusses the concept of the moving-average control chart in some detail. Suppose samples (possibly of some size n), and let $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_t, \ldots$ denote the corresponding sample means. The moving average of span m at time t is defined as:

$$M_t = \frac{\overline{x}_t + \overline{x}_{t-1} + \cdots + \overline{x}_{t-m+1}}{m} \tag{2.45}$$

After time t, the oldest sample is dropped and newest on is added to the set. It is clear that each sample mean is weighted by $1/m$ while samples $\overline{x}_i$ with $i \leq t - m$ are weighted by zero. If $\overline{\overline{x}}$ is the centre of the chart, then the 3-sigma control limits for M_t are:

$$UCL = \overline{\overline{x}} + \frac{3\sigma}{\sqrt{nm}} \tag{2.46}$$

$$LCL = \overline{\overline{x}} - \frac{3\sigma}{\sqrt{nm}} \tag{2.47}$$

The control limits for M while $t < m$ are:

$$UCL = \overline{\overline{x}} + \frac{3\sigma}{\sqrt{nt}} \tag{2.48}$$

$$LCL = \overline{\overline{x}} - \frac{3\sigma}{\sqrt{nt}} \tag{2.49}$$

The alternative for using these changing limits while $t < m$ is to rely solely on the $\overline{x}$ until m sample means have been obtained.

The moving-average chart is more effective than the usual $\overline{x}$ chart in detecting small process shifts. Using both types of charts together can also yield good results. It is also possible to plot both lines on the same graph. Moving-average charts are also useful for plotting data when each sample consists of a single observation. For $m = 1$ and $n = 2$, the moving average chart simplifies into the moving-range (or R) chart (see section 2.5.5).

The Geometric Moving-Average Control Chart

The geometric moving-average chart (GMA) is also know as the exponentially weighted moving average chart (EWMA). Montgomery (1985: 239–243) discusses the following representation of the EWMA chart:

With λ (and with $0 < \lambda \geq 1$) some constant value and z_0 (at time $t = 1$) is $\overline{\overline{x}}$ (the starting value), we have:

$$z_t = \lambda \overline{x}_t + (1 - \lambda) z_{t-1} \tag{2.50}$$

Substituting recursively into 2.50 for z_{t-1}, $j = 1, 2, \ldots, t$, it can be shown:

$$z_t = \lambda \sum_{j=0}^{t-1} (1-\lambda)^j \overline{x}_{t-j} + (1-\lambda)^t z_0 \tag{2.51}$$

If $\overline{x}_i$ are independently random variables with variance σ^2/n, then the variance of z_t (Montgomery, 1985: 240–241) is:

$$\sigma^2_{z_t} = \frac{\sigma^2}{n} \sum \left(\frac{\lambda}{2-\lambda} \right) \left[1 - \left(1 - \lambda^{2t} \right) \right] \tag{2.52}$$

As t increases, it can be shown that $\sigma^2_{z_t}$ increases to a limiting value:

$$\sigma^2_z = \frac{\sigma^2}{n} \left(\frac{\lambda}{2-\lambda} \right) \tag{2.53}$$

Similarly to the derivation of the 3-σ Shewart control limits (see section 2.5.3), the upper and lower control limits for a moderately large t are:

$$UCL = \overline{\overline{x}} + 3\sigma \sqrt{\frac{\lambda}{(2-\lambda)n}} \tag{2.54}$$

$$UCL = \overline{\overline{x}} - 3\sigma \sqrt{\frac{\lambda}{(2-\lambda)n}} \tag{2.55}$$

While t is still small, the 3-σ control limits (based on equation 2.52) are:

$$UCL = \overline{\overline{x}} + 3\sigma \sqrt{\frac{\lambda}{(2-\lambda)n} \left[1 - (1-\lambda^{2t}) \right]} \tag{2.56}$$

$$UCL = \overline{\overline{x}} - 3\sigma \sqrt{\frac{\lambda}{(2-\lambda)n} \left[1 - (1-\lambda^{2t}) \right]} \tag{2.57}$$

The GMA chart is similar to the moving-average control chart. The control limits (for large t) are identical when $\lambda = 2/(m+1)$. However, the GMA chart is more effective than the moving average chart for detecting small process shifts.

2.5.8 Cumulative Sum Control Charts

While the Shewart chart is useful for process monitoring, it only uses process data at the current time point. The cumulative sum (or CuSum) control chart is an alternative to the Shewart chart. The CuSum chart plots the cumulative sums of the deviation of the sample values from a target value. This target value is generally set to the specification or

to the set-point. The choice of target point will affect both the position and slope of the chart. This will in turn affect the ability of the chart to detect process changes (Owen, 1989: 274).

Montgomery (1985: 225–235) discusses the CuSum chart as follows: With samples of size n (at some time point i) having average $\overline{x}_i$ and a process target of μ_0, the CuSum chart is formed by plotting:

$$S_m = \sum_{i=1}^{m} (\overline{x}_i - \mu_0) \tag{2.58}$$

against the sample number m.

Interpretation of the CuSum Chart

It should be clear that if the process remains in control at the target value, the cumulative sum defined in equation 2.58 will vary randomly around zero. If there is a small mean shift upward (so that $\mu_1 > \mu_0$) then a upward drift will develop in S_m. Conversely if $\mu_1 > \mu_0$, a downward drift will develop. The interpretation of any upward or downward trend will be that the process mean has shifted and some assignable cause (possibly a fault) should be sought.

The slope of the charted value added to μ_0 will give the current process mean (Owen, 1989: 272).

A formalisation of this visual procedure is the V-mask. A general V-mask construct is shown figure 2.9. A typical CuSum chart with a V-mask in operation is shown in figure 2.10.

The V-mask parameters are selected as as per the desired performance ((Montgomery, 1985: 228) and (NIST-Sematech, 2005)). The lead distance d and the angle θ is used to define the V-mask (as shown in figure 2.9. With σ_x as the standard deviation of $\overline{x}$; α as the probability of a false detection of process shift; β as the probability of failing to detect a shift in the process mean; Δ is the required detectable shift in the process mean, then the commonly used V-mask design is:

$$d = \left(\frac{2}{\delta^2}\right) \ln\left(\frac{1-\beta}{\alpha}\right) \tag{2.59}$$

$$\theta = \arctan\left(\frac{\Delta}{2k}\right) \tag{2.60}$$

$with:$

$$\delta = \frac{\Delta}{\sigma_x}$$

It is recommended by Montgomery (1985: 228) that k lie between σ_x and $2\sigma_x$ with

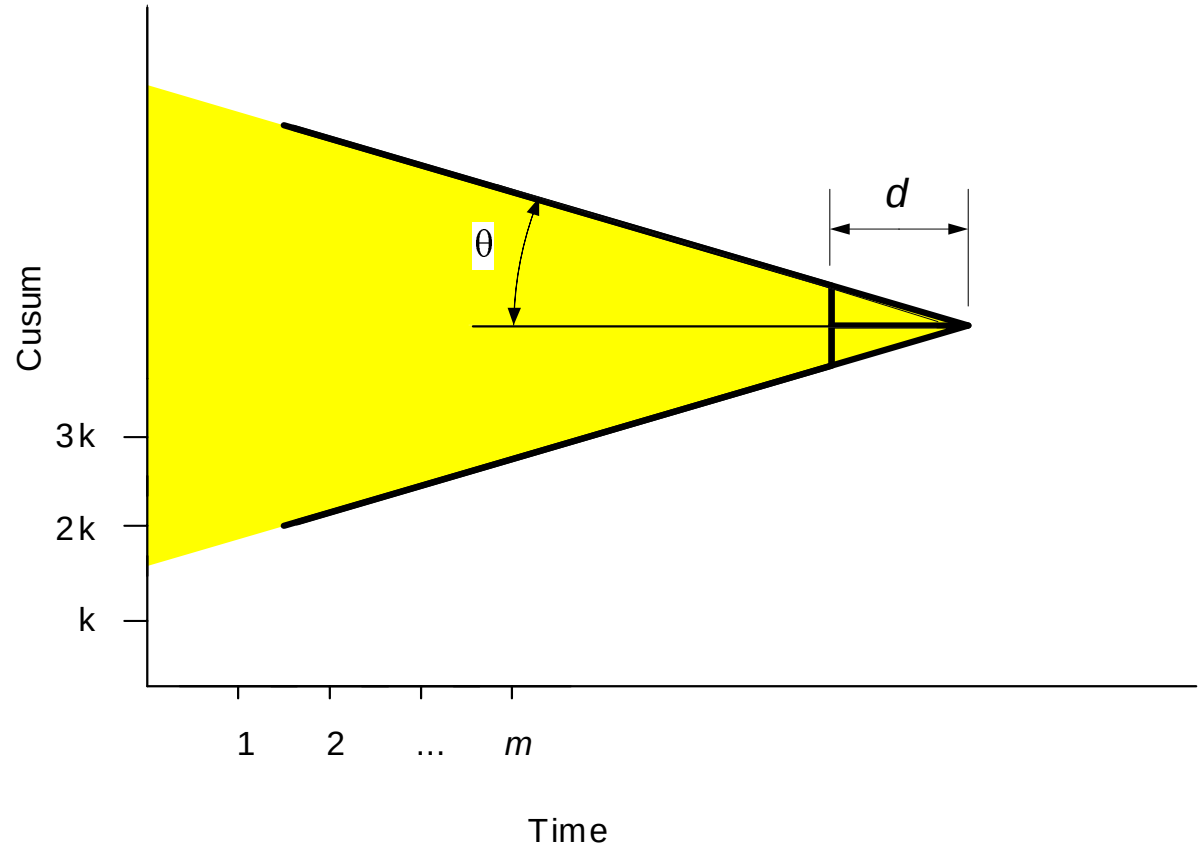

Figure 2.9: General Construction of the V-Mask

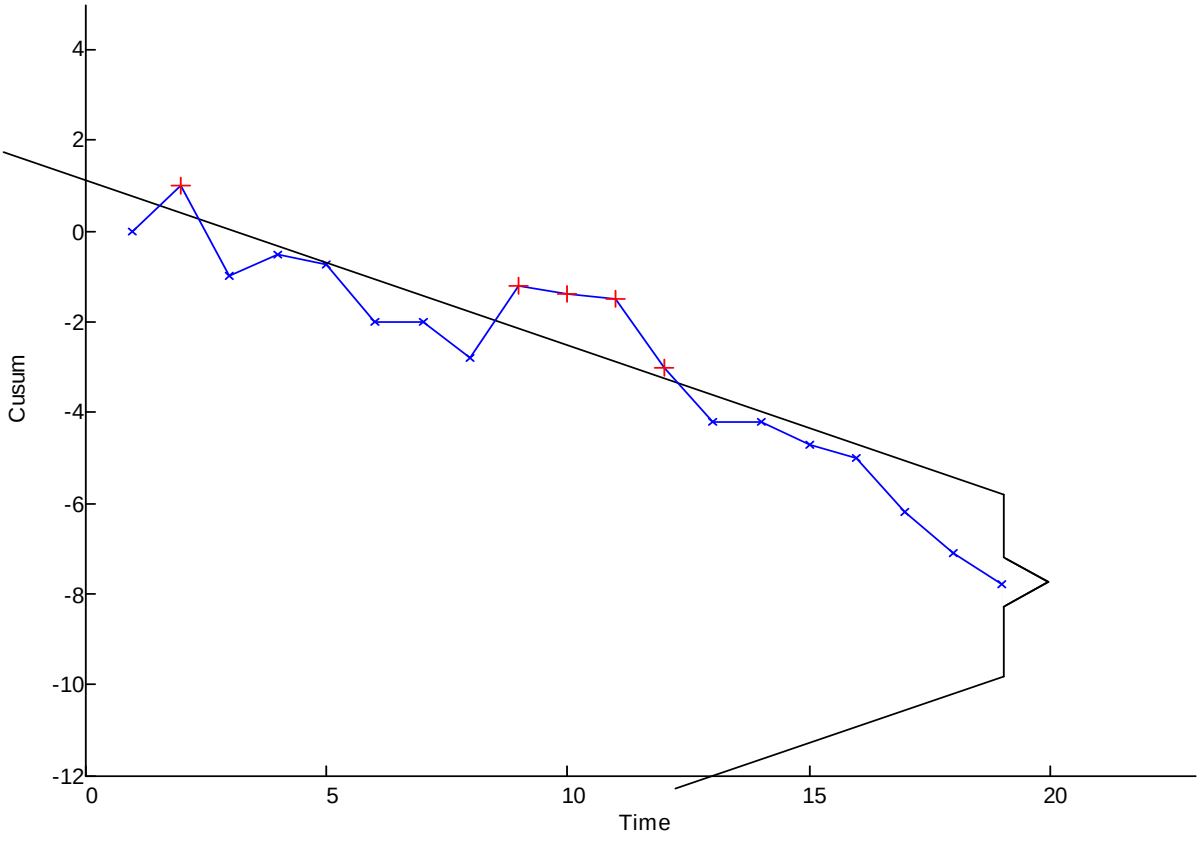

Figure 2.10: The CuSum Chart in Action

$k = 2\sigma_x$ being the preferred value. If β is small the equation 2.59 simplifies to:

$$d = -2\frac{\ln \alpha}{\delta^2} \tag{2.61}$$

The *current* sample of the process can be seen to be out of control (maybe due to a fault) when some points fall outside the arms of the V-mask (as shown in figure 2.10).

There are also alternative shapes for the V-mask. Improvements in the average run length of the CuSum chart can be made by using a semi-parabolic mask (Wetherill & Brown, 1991: 145). In practise, it is easier to make use of a snub-nosed mask (which is derived from superimposing two normal masks.

Alternatively, Bowker and Lieberman (1972) (quoted in Montgomery (1985: 230)) presents a method for choosing d and θ so a to minimise the average run length (ARL - see equation 2.6) at some δ (denoted $ARL(\delta)$). This is presented for various δ's (given in terms of standard deviations) and ARL's for in-control processes (denoted $ARL(0)$).

If equation 2.60 is simplifed, we can see that:

$$\frac{2k}{\sigma_x}\tan\theta = \frac{\delta}{2} \tag{2.62}$$

Illustrating the use of the Table 2.6: Suppose that a shift of $\pm\sigma_x$ needs to be detected, and that the average run length of the in control process process should be 500 samples (i.e. the operator will only very seldomly be disturbed by false alarms). We can read off the table that $(k/\sigma_x)\tan\theta = 1.0$ (similar to the result from 2.62) so θ can easily be calculated), and that $d = 2.7$. We can also see that $L(2.0)$ is only 3.4 so the operator will be notified very quickly.

Average Run Length of the Cusum chart with a V-Mask

Using the recommended parameters in equations 2.59 and 2.60, the average run length (ARL) is approximately 500 samples (Montgomery, 1985: 229).

The alternative method for V-mask design (as discussed in section 2.5.8) aims to obtain a particular ARL curve.

Advantages of CuSum Charts

Shewart charts are sometimes ineffective when the sample size is 1 ($n = 1$) because a momentary measurement error will immediately exceed the control limits. When $n > 1$, gross measurements errors can be damped by taking the mean of other samples. Cusum charts are particularly effective when $n = 1$. This makes the cusum chart an appropriate choice for fault detection.

Cusum charts are more effective than Shewart charts for detecting small shifts in process mean on the order of 0.5 σ_x to $2\sigma_x$. In this range , a CuSum chart will detect a

Table 2.6: Cusum Chart Design Parameters

		$ARL(0)$					
		50	100	200	300	400	500
$\frac{\delta}{\sigma_x} = 0.25$	$(k/\sigma_x)\tan\theta$	0.125			0.195		0.248
	d	47.6			46.2		37.4
	$ARL(0.25)$	28.3			74.0		94.0
$\frac{\delta}{\sigma_x} = 0.50$	$(k/\sigma_x)\tan\theta$	0.25	0.28	0.29	0.28	0.28	0.27
	d	17.5	18.2	21.4	24.7	27.3	29.6
	$ARL(0.50)$	15.8	19.0	24.0	26.7	29.0	30.0
$\frac{\delta}{\sigma_x} = 0.75$	$(k/\sigma_x)\tan\theta$	0.375	0.375	0.375	0.375	0.375	0.375
	d	9.2	11.3	13.8	15.	16.2	16.8
	$ARL(0.75)$	8.9	11.0	13.4	14.5	15.7	16.5
thu $\frac{\delta}{\sigma_x} = 1.0$	$(k/\sigma_x)\tan\theta$	0.50	0.50	0.50	0.5	0.50	0.5
	d	5.7	6.9	8.2	9.0	9.6	10.0
	$ARL(1.0)$	6.1	7.4	8.7	9.4	10.0	10.5
$\frac{\delta}{\sigma_x} = 1.5$	$(k/\sigma_x)\tan\theta$	0.75	0.75	0.75	0.75	0.75	0.75
	d	2.7	3.3	3.9	4.3	4.5	4.7
	$ARL(1.5)$	3.5	4.0	4.6	5.0	5.2	5.4
$\frac{\delta}{\sigma_x} = 2.0$	$(k/\sigma_x)\tan\theta$	1.0	1.0	1.0	1.0	1.0	1.0
	d	1.5	1.9	2.2	2.4	2.5	2.7
	$ARL(2.0)$	2.26	2.63	2.96	3.15	3.3	3.4

process shift approximately twice as fast as the corresponding $\overline{x}$ chart. The CuSum chart is also easier to visualise in terms of detecting the presence and onset of a process shift by observing a change in slope of the charted line.

Disadvantages of the CuSum Chart

The CuSum chart can be slow in detecting large process shifts. Lucas (1982) suggested a modification to the V-mask (using average run length parabolas) to improve performance with regard to large process shifts.

The CuSum chart is also not very effective for analysing past data or to bring a process to control. Diagnosis of patterns on the CuSum chart is difficult due to the assumption that the process is in statistical control (meaning an assumption is made that subsequent points are uncorrelated). Consequently, the CuSum chart often exhibits patterns as a result of this correlation. While this problem means that it may be difficult to use the CuSum chart for control or diagnosis, it will still be most useful for on-line fault detection (Owen, 1989: 290).

Other forms of the CuSum Chart (Montgomery, 1985: 232–235)

There are also tabular forms of the V-mask suited for computer implementation. They are similar to the Run-Sum charts discussed in section 2.5.9. There is also a technique

called the signed sequential-rank control chart, which is a non-parametric procedure. These techniques are very similar to the conventional CuSum chart and they will not be discussed further.

2.5.9 The Run-Sum Control Chart Technique

The Run-Sum control chart technique is a combination of the Shewart and the CuSum charts (Montgomery, 1985: 234). This technique has similar advantages and disadvantages (see sections 2.5.8) as the CuSum technique.

A Run-Sum chart is shown in figure 2.11. The Run-Sum chart is calculated on the Shewart chart. A conventional Shewart chart is drawn with some limits set some number of standard deviations away from the mean ($\overline{\overline{x}}$). Additionally, a score is made based on the position of each data point compared to the $\overline{\overline{x}}$ (the zero point or centre line). Each point will get this score and a running score is kept. This sum is called the Run Sum (denoted S). The score is assigned when the point is no more than 1 standard deviation from the control limits. Each point $\overline{x}_i$ will get a positive score when $\overline{x}_i > \overline{\overline{x}}$ and a negative score when $\overline{x}_i \leq \overline{\overline{x}}$. The value of the score is equivalent to a (rounded) number of standard deviations (from $\overline{\overline{x}}$). The score will be accumulated in S. S will reset to zero when the sign of the score changes or an assignable process shift occurs. Action should be taken (or a fault could be detected) when then value of S exceeds a certain value. Using the same data as the chart in figure 2.11:

$$Scores : -2; -0; -1; -2; -2; -0; +0; -0; +2; -0; -3; +0; +0; +0; +0$$
$$S : -2; -2; -3; -5; -7; -7; 0; 0; +2; -0; -3; +0; +0; +0; +0$$

So if the critical value was 6, then some fault (or other process shift) occurred at samples 5 and 6. Note that this process shift would not be detected with a Shewart chart with 3 σ_x limits.

2.5.10 Histograms

Histograms are useful for visualising the distribution of the data which can be used as a statistical measure of the quality of the process. For further details of the use of multivariate histograms, see section 3.2.5.

2.5.11 Other kinds of Charts

Wheeler (1990) refers to a chart for mean ranges and the chart for mean effects. These charts are extensions of the control chart technique, they are not widely used or described.

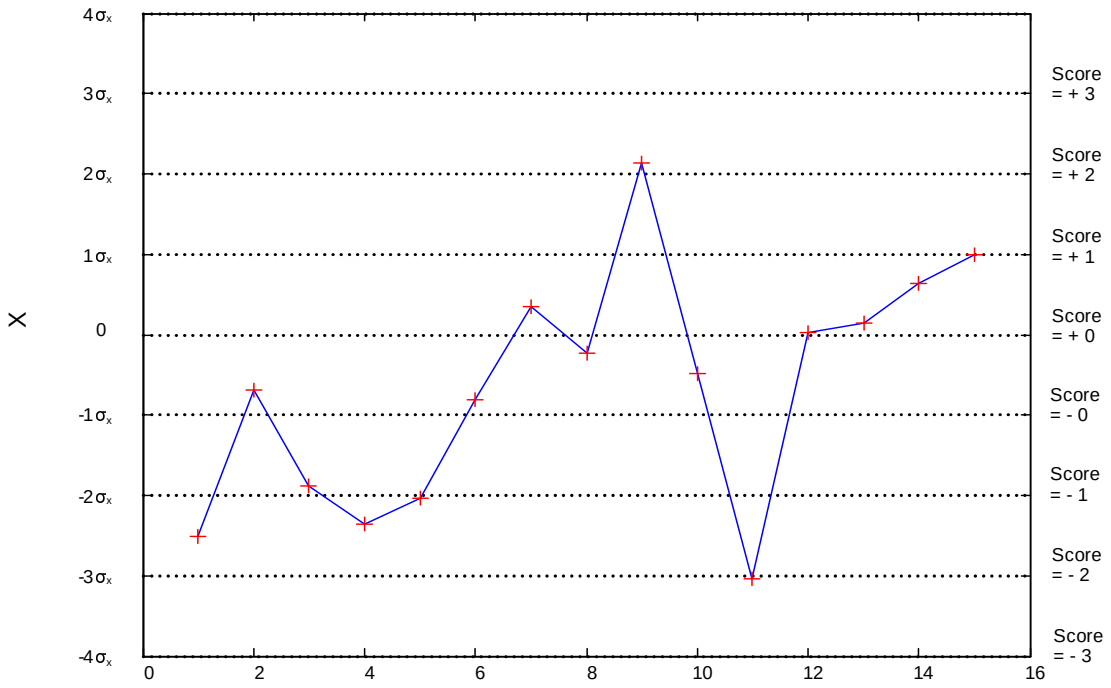

Figure 2.11: Example Shewart Chart used for a Run-Sum Calculation

CHAPTER 3

Multivariate Statistical Process Control

Traditionally, statistical process control (SPC) charts such as Shewart, CuSum and EWMA charts have been used to monitor process quality variables at the expense of monitoring process condition variables. These univariate control charts show poor performance when applied to multivariate processes (Lee et al., 2004a).

Chemical processes are inherently multivariate. When using conventional Shewart charts, a process (described by two variables x_1 and x_2) can only be considered to be in control when both $\overline{x}_1$ and $\overline{x}_2$ are both within their respective control limits. These plots could be combined as shown in figure 3.1

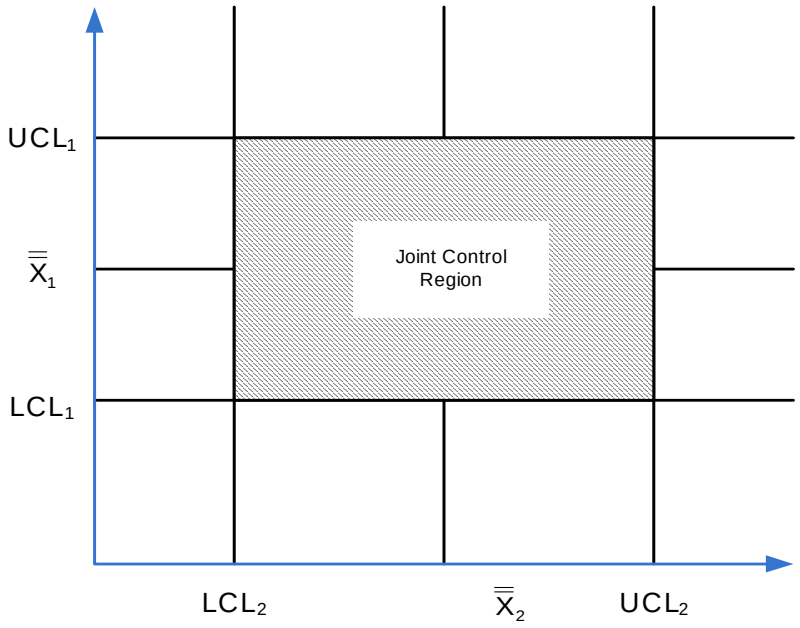

Figure 3.1: Control Region with Independent Limits for Two Variables

However, this approach is misleading. Using 3-σ control limits, the probability of

an in-control variable exceeding its control limits is 0.0027. Therefore, while using two variables (x_1 and x_2), the probability of both of the variables simultaneously exceeding their limits is $0.0027 \cdot 0.0027 = 0.000000729$, which is is significantly different from the probability of 0.0027 suggested by the use of 3-σ limits. This distortion increases exponentially with the number of process variables used. This means that the distortion can be severe for even moderate numbers of measured variables.

This problem is expounded if the variables are not independent (as is often the case with chemical processes).

What is ideally needed, are techniques that can reduce a high-dimensional multivariate dataset down to 2 or 3 composite metrics. These metrics can then be monitored in order benchmark performance, highlight problems and by so doing, provide a framework for continuous process improvement (Bersimis et al., 2005). Macgregor & Kourti (1995) gives a good overview of traditional MSPC techniques.

3.1 The Hotelling's T^2 Distribution

This section discusses the motivation for multivariate process monitoring by discussing the shortcomings of combining sets of univariate statistics.

Consider variables x_1 and x_2 (having nominal means $\overline{\overline{x}}_1$ and $\overline{\overline{x}}_1$) are jointly normally distributed. Let $\overline{x}_1$ and $\overline{x}_2$ be sample means (from a subgroup of size n). The covariance between the two variables is represented by S_{12}. In this way, the relationship between the variables is accounted for. The Hotelling's T^2 distribution with 2 and $n-1$ degrees of freedom is then as follows (Montgomery, 1985: 245–253):

$$T^2 = \frac{n}{S_1^2 S_2^2 - S_{12}^2}\left[S_1^2\left(\overline{x}_1 - \overline{\overline{x}}_1\right) + S_2^2\left(\overline{x}_2 - \overline{\overline{x}}_2\right)^2 - 2S_{12}\left(\overline{x}_1 - \overline{\overline{x}}_2\right)\left(\overline{x}_1 - \overline{\overline{x}}_2\right)\right] \tag{3.1}$$

If $T^2 > T^2_{\alpha,2,n-1}$ (the upper α percentage point of the distribution with 2 and $n-1$ degrees of freedom, then at least one of the characteristics is out of control. Therefore, for two parameters:

$$UCL = T^2_{\alpha,2,n-1} \tag{3.2}$$

This technique clearly has no centre lines or lower control limits.

Extending this to p quality characteristics, the characteristic can be represented by a $p \times 1$ vector as follows (as per Romagnoli & Palazoglu (2006: 452–453)):

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} \tag{3.3}$$

The test statistic is then:

$$T^2 = n\left(\overline{\mathbf{x}} - \overline{\overline{\mathbf{x}}}\right)^T \Sigma^{-1} \left(\overline{\mathbf{x}} - \overline{\overline{\mathbf{x}}}\right) \tag{3.4}$$

with Σ as the covariance matrix of the p quality characteristics and:

$$\overline{\overline{\mathbf{x}}} = \begin{bmatrix} \overline{\overline{x}}_1 \\ \overline{\overline{x}}_2 \\ \vdots \\ \overline{\overline{x}}_p \end{bmatrix} \tag{3.5}$$

The upper control limit can be determined for the statistical F distribution as follows:

$$UCL = T^2_{\alpha,p,n-1} = \frac{p(n-1)}{n-1} F_{\alpha,p,n-p} \tag{3.6}$$

Generally $\overline{\overline{\mathbf{x}}}$ and $\boldsymbol{S}$ are defined from preliminary data when the characteristics are assumed to be in control.

This process with m available samples of size n is summarised as follows:

$$\overline{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk} \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{3.7}$$

$$S^2_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{ijk} - \overline{x}_{jk}\right)^2 \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{3.8}$$

The covariance between the j^{th} and the h^{th} in the k^{th} sample is:

$$S_{jhk} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{ijk} - \overline{x}_{jk}\right)\left(x_{ihk} - \overline{x}_{hk}\right) \qquad \begin{cases} k = 1, 2, \ldots, m \\ j \neq h \end{cases} \tag{3.9}$$

$\overline{x}_{jk}$, S^2_{jk} and S_{jhk} are then averaged over the m samples to get:

$$\overline{\overline{x}}_j = \frac{1}{m}\sum_{k=1}^{m}\overline{x}_{jk} \qquad j = 1, 2, \ldots, p \tag{3.10}$$

$$S^2_j = \frac{1}{m}\sum_{k=1}^{m}S^2_{jk} \qquad j = 1, 2, \ldots, p \tag{3.11}$$

$$S_{jh} = \frac{1}{m}\sum_{k=1}^{m}S_{jhk} \qquad j \neq h \tag{3.12}$$

The covariance matrix $\boldsymbol{S}$ of size $p \times p$ is then formed as follows (Montgomery, 1985: 250):

$$\boldsymbol{S} = \begin{bmatrix} S^2_1 & S^2_{12} & S^2_{13} & \cdots & S^2_{1p} \\ & S^2_2 & S^2_{23} & \cdots & S^2_{2p} \\ & & \ddots & & \vdots \\ & & & & S^2_p \end{bmatrix} \tag{3.13}$$

Considering a simple case where $p = 2$, if the variables are independent, then $S_{12} = 0$. Equation 3.1 then reduces to a representation of an ellipse, centred at $(CL_1,\ CL_2) = (\overline{\overline{x}}_2,\ \overline{\overline{x}}_2)$. The principal axes of the ellipses will be parallel to the $\overline{x}_1$ and $\overline{x}_2$ axes. A representation of this is shown in figure 3.2.

If the characteristics are independent ($S_{12} \neq 0$), the principal axes of the elliptical control region formed are no longer parallel to the $\overline{x}_1$ and $\overline{x}_2$ axes. A representation of this can be found in figure 3.3.

The use of the elliptical control regions has the following disadvantages:

1. Time information is lost (unless an additional parameter such as colour is used).

2. It is difficult to plot for more than 2 characteristics (without a computer) and very difficult to visualise beyond 3 characteristics.

Also, for most applications when the number of variables is large, Σ in equation 3.4 will be almost singular due to the correlation with one another. A solution to this problem is dimensional reduction by means of principal component analysis (Romagnoli & Palazoglu, 2006: 453).

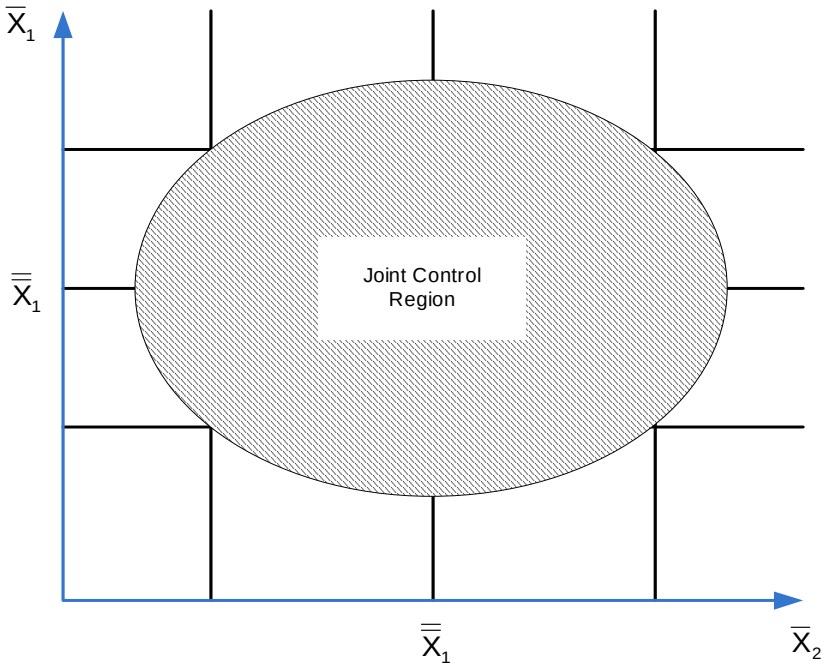

Figure 3.2: Control Region for Two Independent Limits

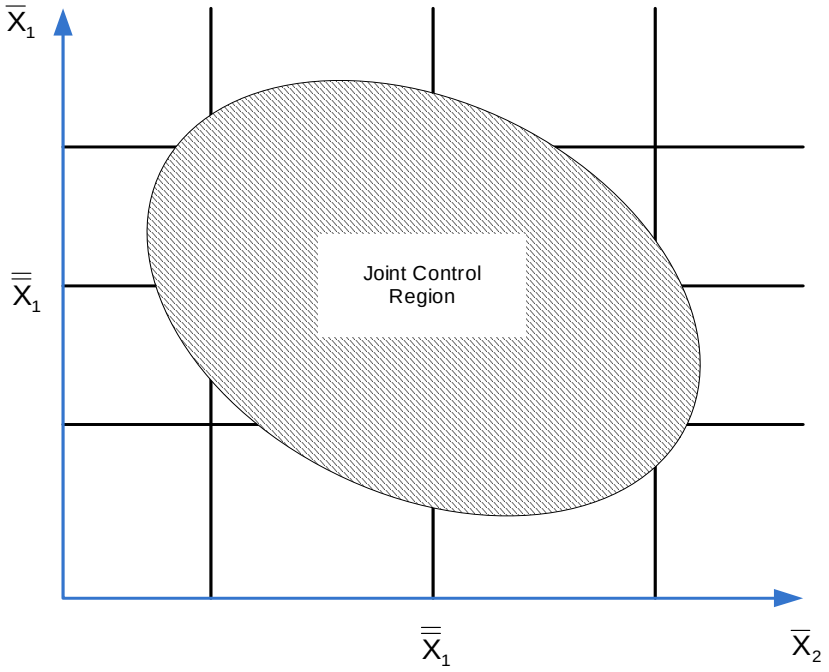

Figure 3.3: Control Region for Two Dependent Limits

3.2 Miscellaneous Multivariate Data Visualisation Techniques

Albazzaz et al. (2005) compares multidimensional data visualisation techniques with reference to multivariate statistical process control. As the monitoring process metrics is key to fault detection, several important methods that can be applied to fault detection and diagnosis are discussed here.

3.2.1 Parallel Coordinate Plots

A parallel coordinates plot is a tool for visualising multivariate data. Using a Cartesian coordinate system, the most dimensions that we can visualise is three (projected onto paper or a screen). This is because Cartesian axes are orthogonal. If the axes are instead drawn parallel to each other, many axes can be drawn on the same two dimensional display (Martinez & Martinez, 2004: 318–319). Each observation is represented by the sequence of its coordinate values plotted against their coordinate values. If the data are of different orders of magnitude, the values should be scaled in some way otherwise the values of variables with small ranges will be difficult to discriminate. The order of the variables is important as the adjacent axes provide some information about the relationship between consecutive variables. Each sample is shown as a separate line.

Generally observations are plotted with colours designating what group or class they fall into. The following can be determined from a parallel coordinates plot (Martinez & Martinez, 2004: 319):

1. Class separation in a given coordinate.
2. Correlation between pairs of variables.
3. Clustering or groups.

An example of a typical parallel coordinate plot is shown in figure 3.4. The diagram be can simplified by only showing the median for each class together with some quantiles as shown in figure 3.5. This shows example data from the continuous running of the distillation column (see section 5.1 for details of this column). We can easily see that the bottom temperature is generally higher than the top temperature. We can also see that the bottoms level is also usually much higher than the level at the top. Two classes, namely ‘Normal’ and ‘Faulty’, are plotted here. It can be seen that both temperature profiles are similar. The steam pressures are also similar, with the exception of two samples. The feed flowrates are dramatically different. The faulty set had many samples of low or no feed with three samples of higher than normal flowrate. While the top condenser level was similar for both groups, the bottom level of the faulty group showed

much higher variation (as confirmed in figure 3.5). As discussed in section 2.1, unusual variation may be as a result of a fault. There were also three outliers of lower than normal bottoms level. The faulty data set were taken from a period were the boiler had tripped and was restarted. This accounts for the steam pressure outliers in the faulty dataset. The low feed flowrates are explained by a vapour lock in the feed line. A hand operated choke valve was opened to relieve the airlock - accounting for the high feed flowrate values before the controller could correct the flow back to normal values. The same dataset is also analysed in section 3.3.8 using Andrew's plots.

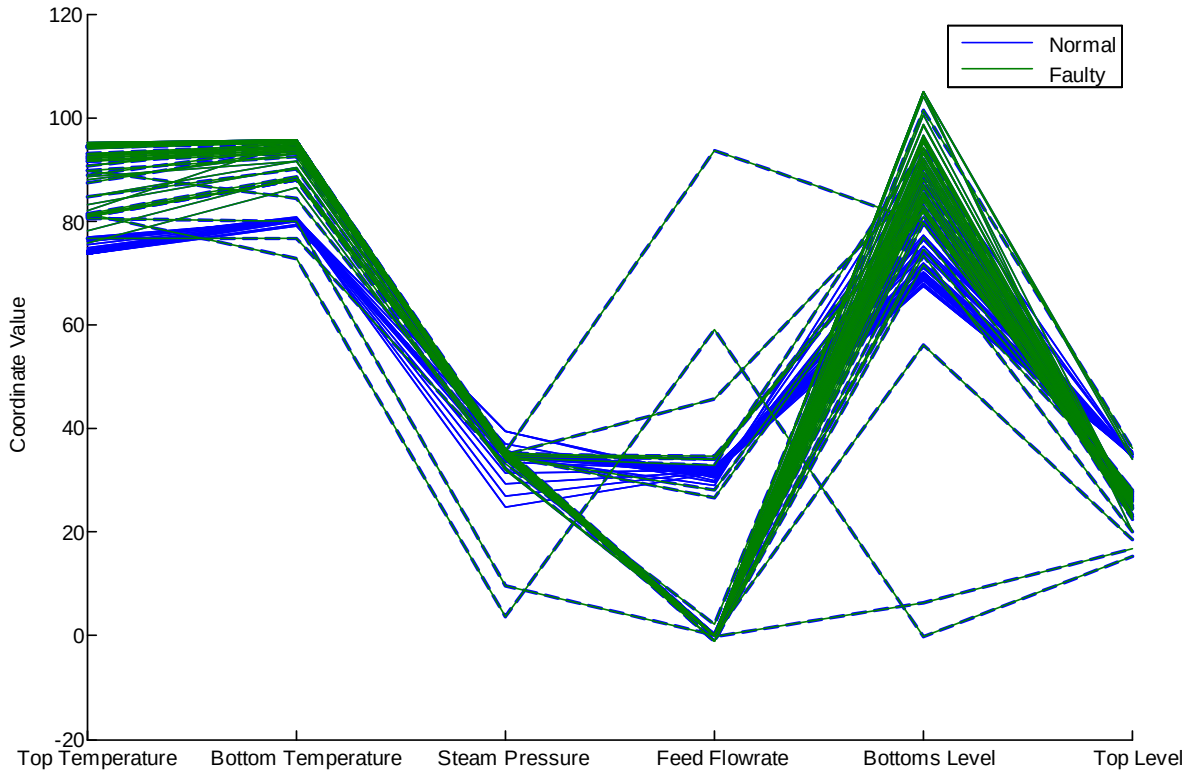

Figure 3.4: Example of a Parallel Coordinate Plot

The disadvantage of parallel coordinate plots is that, while they are useful summaries of data, they do not provide any means for removing redundancy in variables set. Any correlation between the variables is not exploited. Also it is not trivial to find a suitable order for the axes. It is possible that no general order which is suitable for revelling all possible information does not exist.

3.2.2 Boxplots

Boxplots are also called box and whisker diagrams. They are useful to summarise and compare many attributes of data sets quickly and easily. Horizontal lines are drawn at each of the quartiles (the division points of a ordered set of data divided into 4 parts containing the same number of samples) of the data. Vertical lines are drawn to join

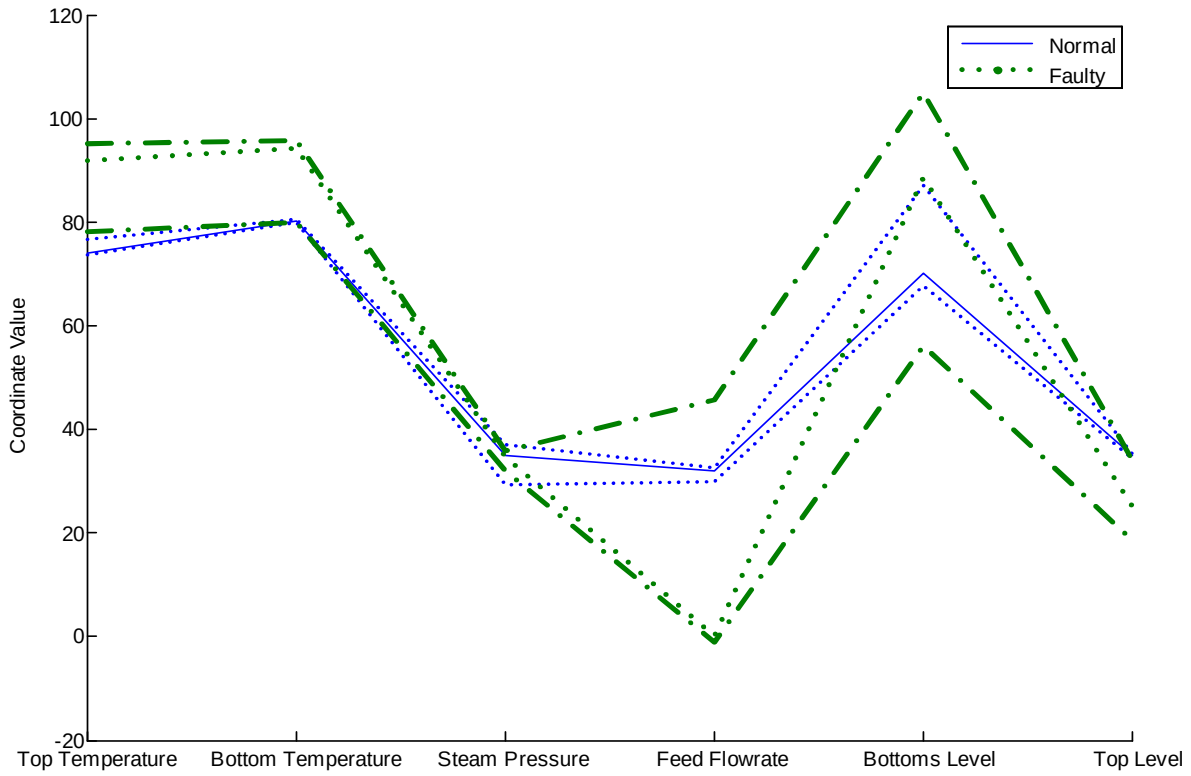

Figure 3.5: Example of a Parallel Coordinate Plot with Medians and 95% Quantiles

these lines into a box. Whiskers extend to some factor to the nearest adjacent values. Often this whisker is not allowed to extend beyond some factor (usually 1.5) of the quartile range. Outliers (points beyond the whiskers) are shown by some symbol. Extreme outliers (points more than three quartile ranges beyond a box edge) are shown with some other symbol (Montgomery et al., 2001: 35–36).

There are several variations of the boxplot. Often a notch is shown around the median line. The medians of the two groups differ at the 5% significance level if this notch does not overlap between groups. Other variants including boxplots using kernel density estimates and variable widths exist (Martinez & Martinez, 2004: 274–278).

An example of a set of data from a startup of the distillation column is given in figure 3.6. Here the temperatures of the top and the bottom plate are give as adjacent boxplots. Both datasets start from a low temperature. For the lower plate, this initial temperature is the end of the whisker. For the top plate it is marked as an outlier. Information about the location can be seen from the location of the median line. The notches of the two boxplots do not overlap. This means that the medians do differ significantly. The spread, skewness and longtailness of the sample can be seen from the lengths of the whiskers and boxes (relative to each other and relative to the median line). For example it can be seen that both plates reached approximately the same high temperature (during the heating period while the condenser drum filled). The median for the top plate was significantly lower than the bottom plate due to the time during

steady state conditions. This is in agreement with our expectation of a binary column. Faults could be detected by means of changes in the median, range and distribution of the data. Here outliers indicate that the column in the shut-down state is significantly different from the normal running data.

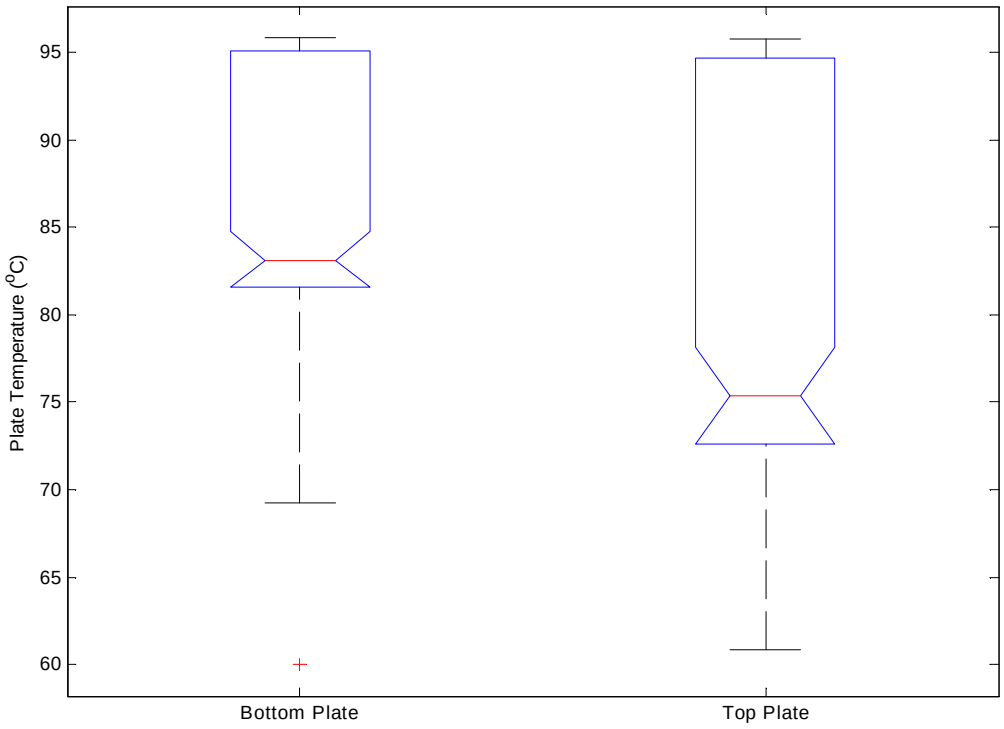

Figure 3.6: Boxplot of Example Startup Distillation Column Data

3.2.3 Bagplots and Alpha Bagging

The bagplot is a bivariate variation of the univariate boxplox (discussed immediately above). The bagplot was proposed by Rousseeuw et al. (1999). Bagplots show potential for use in conjunction with the results of principal component or discriminant analysis biplots (discussed later) to visualise multidimensional data containing different classes of observations. They can also be used for grouping with normal scatterplots.

The components of a bagplot are:

1. The depth median
2. The bag
3. The fence
4. Indication of any outliers.

The univariate boxplot features a vertical line at the median (see section 3.2.2). Correspondingly, the bagplot features a cross drawn at the depth median. The depth median is the point with the highest half space depth. A shaded bag indicates the bag which contains 50 % of the data points.

The procedure to calculate the bagplot is outlined by Rousseeuw et al. (1999), Wolf (2006), Martinez & Martinez (2004: 286–287) and Gardner et al. (2005). Firstly, the depth median (T^*) is found. The half-space location depth of point Θ belonging to a bivariate dataset is given by the smallest number of points contained in the closed half-plane whose boundary lines passes through Θ. A depth region D_k is then defined as the set of all Θ which have a half-space depth which is greater that or equal to k. The depth median is then usually the centre of gravity of the deepest region.

An interpolation procedure is then used to find the innermost 50 % of the points. Letting $\#D_k$ be the number of points in depth region D_k, a value of k for which:

$$\#D_k \leq \left\lfloor \frac{n}{2} \right\rfloor < \#D_{k-1} \tag{3.14}$$

is then found. The bag (B) is then found by interpolating between D_k and D_{k-1}.

It is possible to adapt the algorithm so that the bag contains the innermost α % points. This is refered to as alpha-bagging (Gardner(2001) quoted by Gardner et al. (2005)). Typical values of α are 90% or 95%. These points are then contained in a shaded bag. This is analoguous to the boxes of the univariate boxplot. The Fortran code supplied by Rousseeuw et al. (1999) was modified to use some arbitary α value. It was then recompiled and used with Matlab to create plots. Alternatively the matlab code of Gianferrari Pini (2004) could be used.

The location of the fence is found by expanding the location of the outside of the bag relative to the depth median (T^*) by some factor (usually 3 (Rousseeuw et al., 1999)). The area between the bag and the fence is often also shaded. The fence can be thought of as a convex hull of all the non-outliers (Martinez & Martinez, 2004: 287). In α-bagging, this area is not shaded or outlined. The original data points are superimposed on these shaded areas. The outliers (if displayed) may be highlighted in the plot. Highlighting these points is analogous to the whiskers of the univariate boxplot. If the data are tightly linear, the bagplot becomes an angled boxplot. An example of a bagplot using pseudo-random data is shown in figure 3.7. Here there are no outliers beyond the fence (the expansion of the bag by a factor of 3).

Because the half-space depth does not change with translational, rotational or other linear transformations, the bagplot is changed directly with these types of transformations(Rousseeuw et al., 1999). This has useful calculational implications with regard to linear transformations such as PCA or LDA (discussed later).

The bag is sill defined for cases where the dataset consists of more than 2 variables.

It is however computationally intensive (Rousseeuw et al., 1999). It is easy to draw a bagplot matrix for any dimension. This is simply a matrix of bagplots of each pair of variables. The diagonal elements are the bagplots of the variables against itself. This reduces to a boxplot (as discussed earlier) at a 45°angle.

Interpretation of the Bagplot

The bagplot enables us to visualises several characteristics of the data (Rousseeuw et al., 1999). The location of the the depth median cross indicates the position of the data. The spread of the data is indicated by the size of the bag and the fence. Any correlation between the variables can be seen in the orientation of the bag. The size of the bag is an indication of the spread of the data. Any skewness in the distribution can be seen by the shape of the bag and the location of the median within the bag. The tails of the distribution can be seen by the looking at the points near the fence as well as the highlighted outliers. In these ways, the bagplot give a useful indication of various aspects of process metrics. This can be used for fault detection. The bag can be used as a type of confidence interval around normal data. There is no set probability of data being in or outside the bag. This is because the percentage of points closest to the depth region is specified, instead of the probability of a point being in the region. The lack of a set

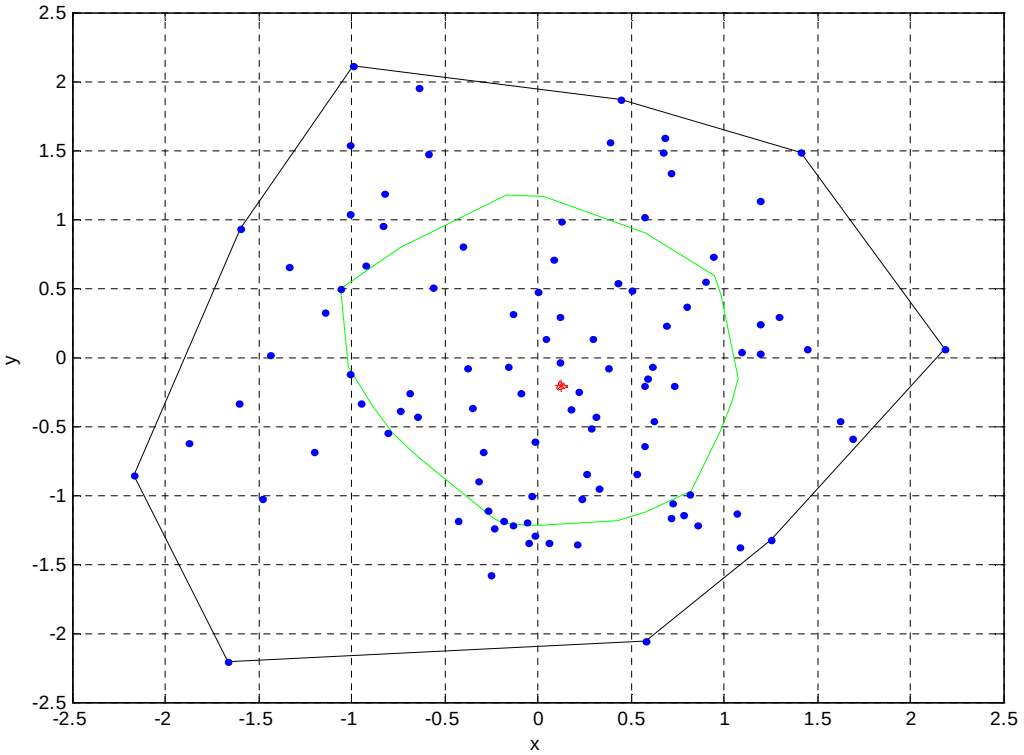

Figure 3.7: Example of a Bagplot

probability does not allow for a classification error estimate to be given. Alpha bags are superior to elliptical confidence limits in circling non-elliptical data regions.

3.2.4 Scatterplots

The scatterplot is a widely used multivariate visualisation technique (Martinez & Martinez, 2004: 294). It is useful to summarise the relationship between two variables. The scatterplot is analogous to the one dimensional time series plot (see section 2.5.2). For the 2 dimensional case, it is simply a plot of the (x_i, y_i) pairs against each other as points or other symbols. For three dimensions, the (x_i, y_i, z_i) triplets are plotted. A scatterplot is shown in figure 3.8. This represents data from some startup of the distillation column discussed in section 5.1. Here, a scatterplot of the pressure in the boiler against the valve opening of the valve controlling that pressure. Note that the setpoint also varied during this time. One can clearly see the column at its shut-down condition with the pressure at 0 kPa and the valve opening at 0 %. The setpoint was then set to some value around 35 kPa while the pressure was still building. One can see the controller opening the valve fully while the pressure rises until its setpoint. The setpoint was then varied several times.

Univariate histograms are shown next to the plot. This gives some representation of the distribution of each variable. This is often disguised in normal scatterplots when the points lie on or very close to one another. We can see that the majority of the points lie close to the setpoints. Wile a larger number of bins (relative to the tight setpoint tracking) are used for the histogram, the number is still to small to reveal whether the data are normally distributed around the setpoint.

The scatter plot is thus useful for the straight forward monitoring of data, observation of trends and information about groups within the data. This technique can be combined with alpha bags to find regions of high data density.

Scatterplot Matrices

Scatterplots become more difficult to visualise beyond 3 dimensions. The three dimensional case can also require some rotation to aid visualisation. Values can be difficult to read off quickly. Scatterplot matrices may then be useful when the dimensionality (p) is greater than two. The scatterplot matrix is merely a set of all the possible 2 dimensional scatterplots (Martinez & Martinez, 2004: 298). The diagonal elements are univariate histograms of each of the variables. This is useful in gauging the distribution of each variable. An example of a scatterplot matrix is shown in figure 3.9. This shows similar data to the data presented in the previous scatter plot. The plots in the upper triangle of the matrix are probably redundant as they are just rotations of those in the lower triangle.

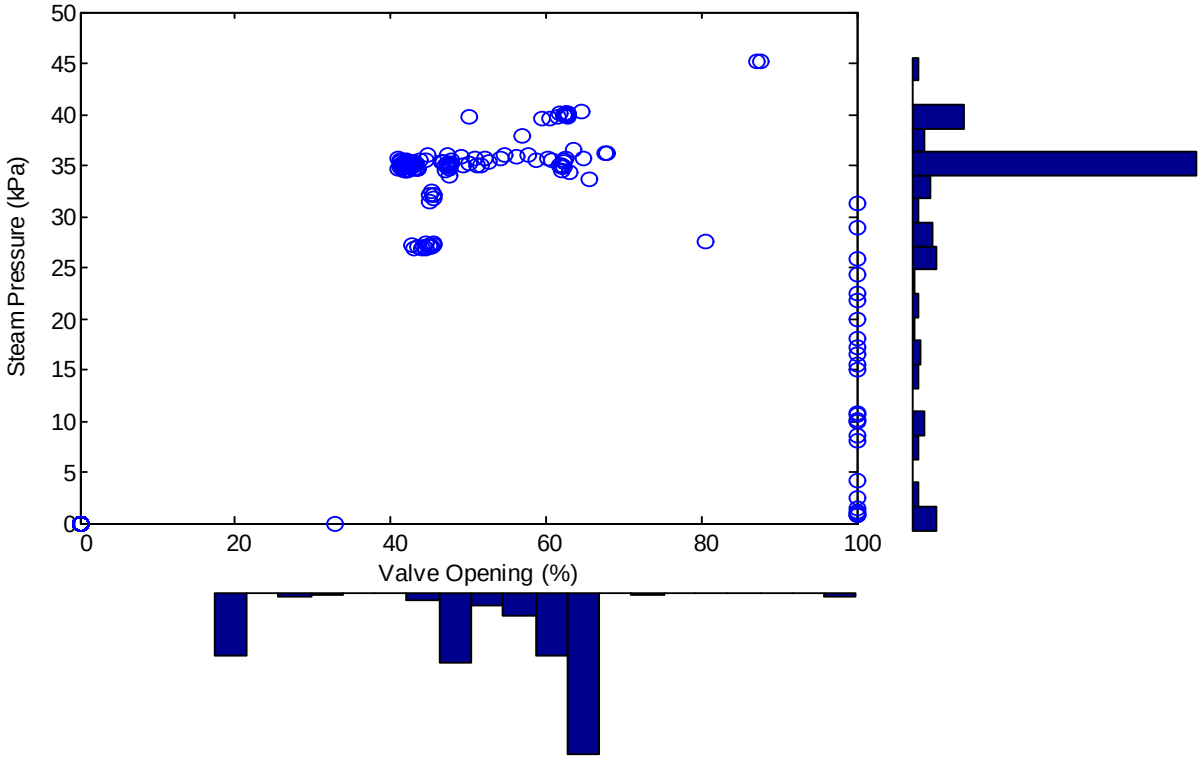

Figure 3.8: Scatterplot of Example Distillation Column Startup Data

Scatterplots with Hexagonal Binning

When the number of points (n) is large, there is a potential for a large amount of points coinciding or lying close together on the graph. The scatterplot with the adjacent his-

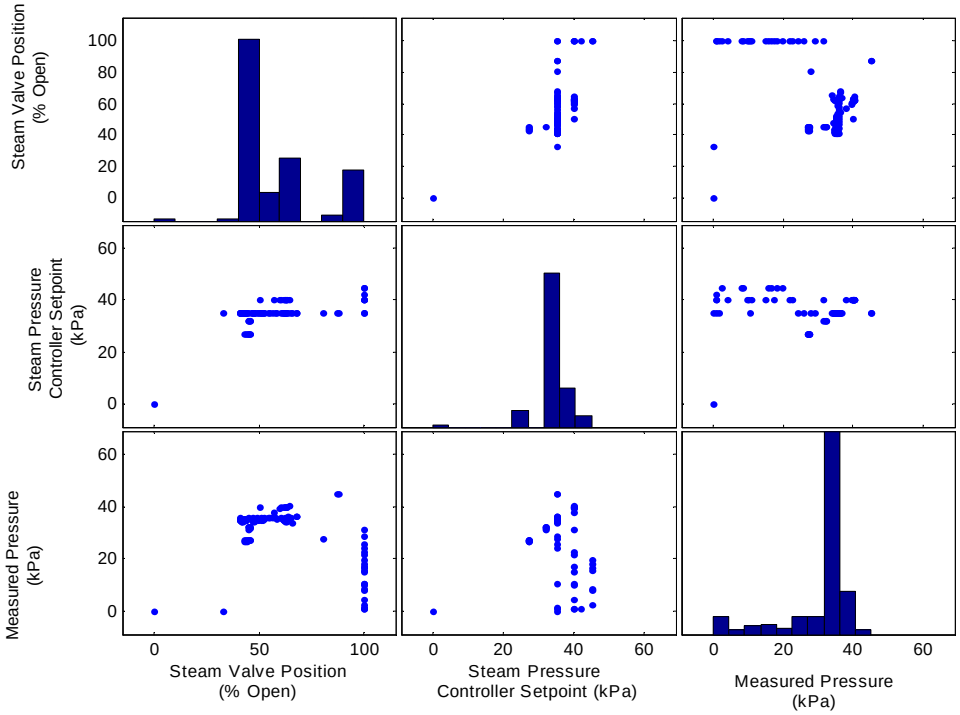

Figure 3.9: Scatterplot Matrix for Example Distillation Column Startup Data

tograms goes some way to revelling the presence of hidden points. Hexagonal binning can be used in order to combine the data points with some estimate of the data density (Martinez & Martinez, 2004: 300). This is a very similar concept to a bivariate histogram. The size of the data marker is used as the estimate of data density instead of bar height or colour. An example of this type of plot (using the same data as presented earlier) is shown in figure 3.10. A high number of bins were used in order to revel more data groups.

The procedure for drawing a hexagonally binned scatterplot is as follows (Martinez & Martinez, 2004: 300):

1. Find the length r of the side of the hexagon based on the number of bins and the range of the data.
2. Obtain the set of bins.
3. Bin the data
4. Scale the bins with no-zero bin counts to have an length of $1r$ and the hexagons with the lowest (non-zero) bin count to have a length of $0.1r$
5. Display the hexagons with side length r at the centre of each bin.

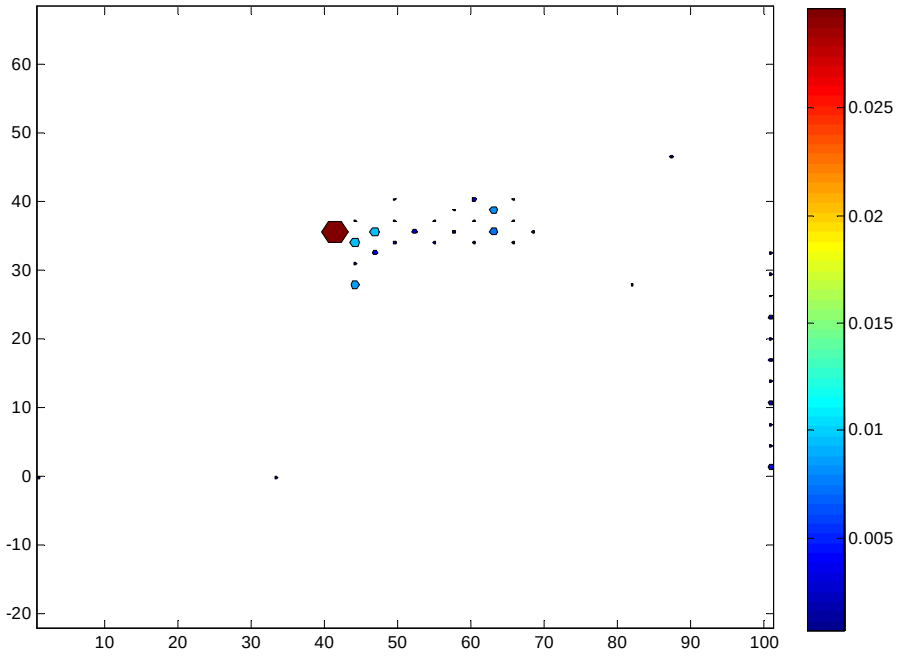

Figure 3.10: Scatterplot with Hexagonal Binning of the Example Distillation Column Startup Data

3.2.5 Histograms

Histograms are a graphical method that can be used to summarise a dataset. The distribution of the data described by vertical bars representing the the number of points falling into a bin. This bin count can also be described in other ways.

The number of bins affects to what degree the histogram is smoothed. There are various rules for the choice of the bin widths and, correspondingly, the number of bins (Martinez & Martinez, 2004: 264–266).

Bivariate histograms can be represented with the bin counts shown as bar heights (as in figure 3.11) or by means of a colour scale (as in figure 3.12). The first may have tall bins obscuring bins behind them. The second type of representation is often a good balance between conveying information and simplicity. The coloured histograms are often used in process monitoring (Groenewald et al., 2006).

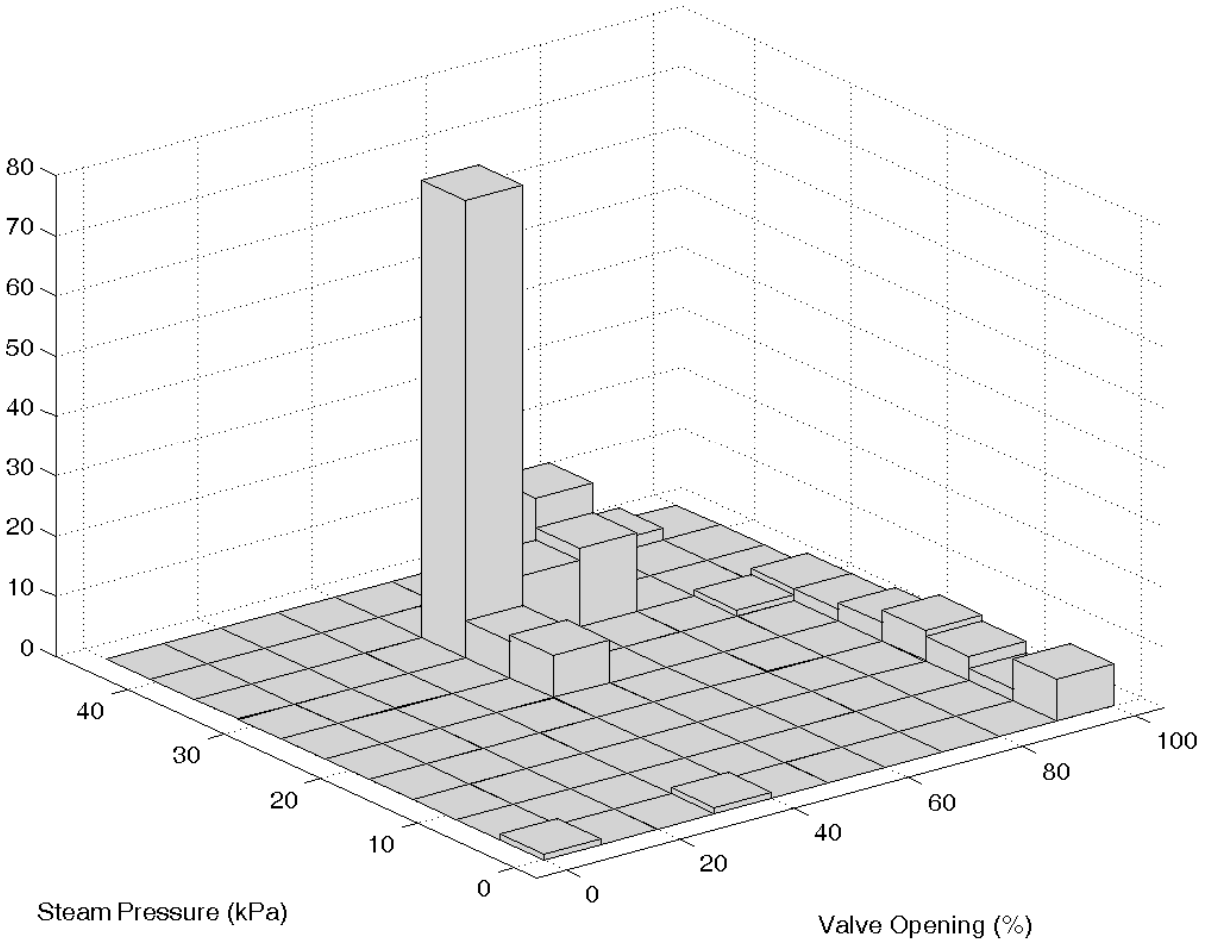

Figure 3.11: Histogram for Example Distillation Column Startup Data

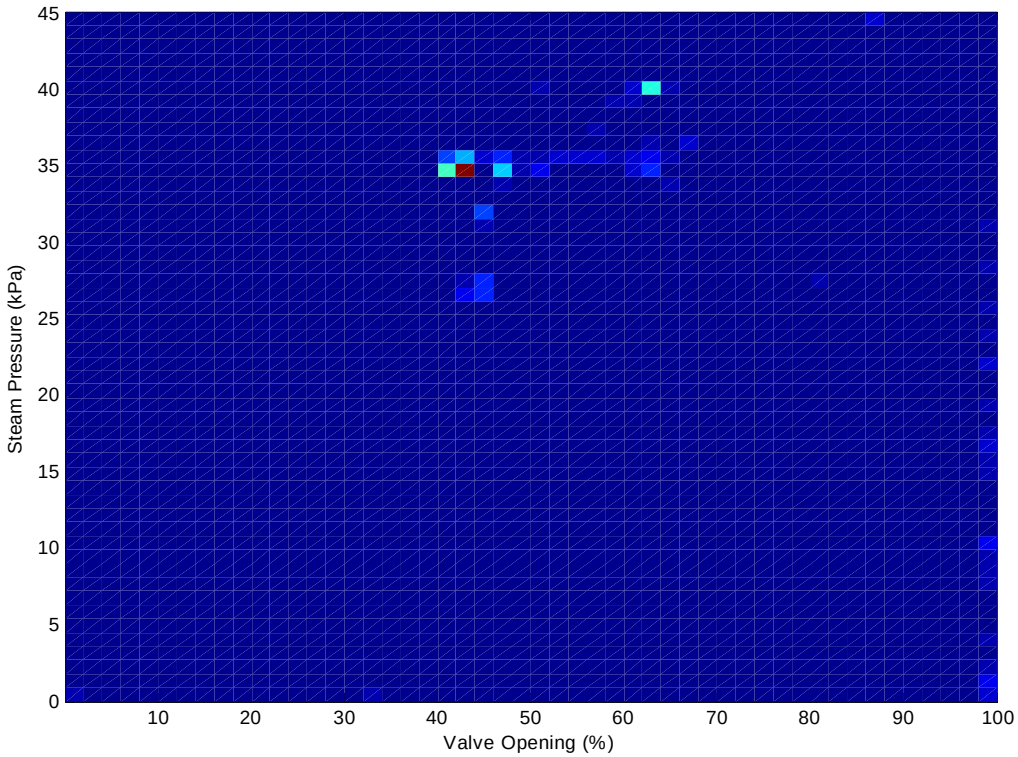

Figure 3.12: Coloured Histogram for Example Distillation Column Startup Data

3.2.6 Glyph Plots

High dimensional data can be represented by cartoon faces (Martinez & Martinez, 2004: 293–295). Each feature of the face represents a characteristic of the sample. Up to 18 characteristics can be represented using the conventional glyph faces. This means that samples with up to 18 characteristics can be visualised. The following are common characteristics:

1. Size of face
2. Forehead/jaw relative arc length
3. Shape of forehead
4. Shape of jaw
5. Width between eyes
6. Vertical position of eyes
7. Height of eyes
8. Width of eyes (this also affects eyebrow width)

9. Angle of eyes (this also affects eyebrow angle)
10. Vertical position of eyebrows
11. Width of eyebrows (relative to eyes)
12. Angle of eyebrows (relative to eyes)
13. Direction of pupils
14. Length of nose
15. Vertical position of mouth
16. Shape of mouth
17. Mouth arc length.

An example of a glyph plot with faces is shown in figure 3.13.

A star plot is a very similar type of diagram. The length of each spoke of the star corresponds to an observation in the sample. An example is shown in figure 3.14. This plot clearly shows dramatic differences between the means of data from a startup data to that of two different setpoint tracking periods.

This method give a more qualitative representation as compared to the other quantitative methods shown here. The choice of which variable to assign to each facial characteristic will also affect the plot dramatically. The method has the advantage of not being intimidating to non-technical users. It is also useful to quickly visually detect gross irregularities or changes (being indicative of a fault) between datasets.

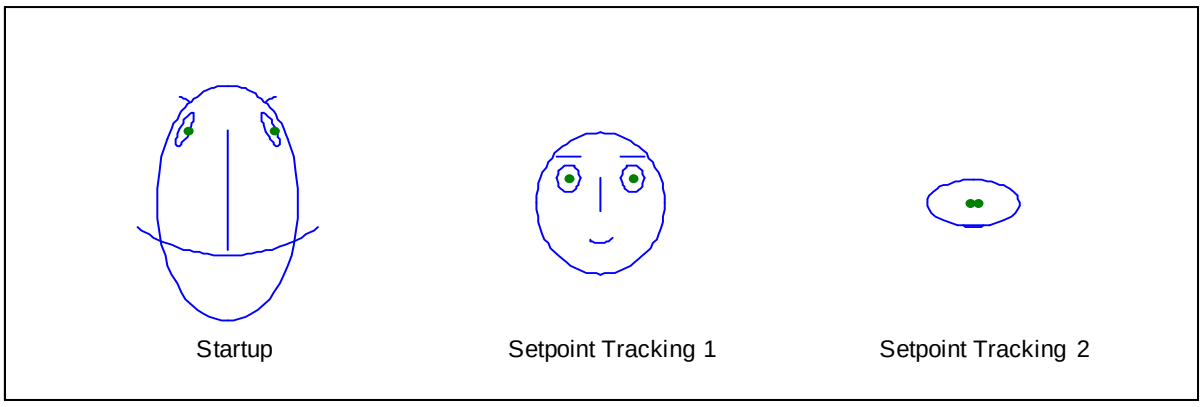

Figure 3.13: Glyph Face Plot of Example Distillation Column Data

For additional data visualisation techniques, see the competitors to principal component analysis - namely Andrew's function plots and partial least squares techniques (section 3.3.8). While these are dimensional reduction techniques, they are useful for visualisation.

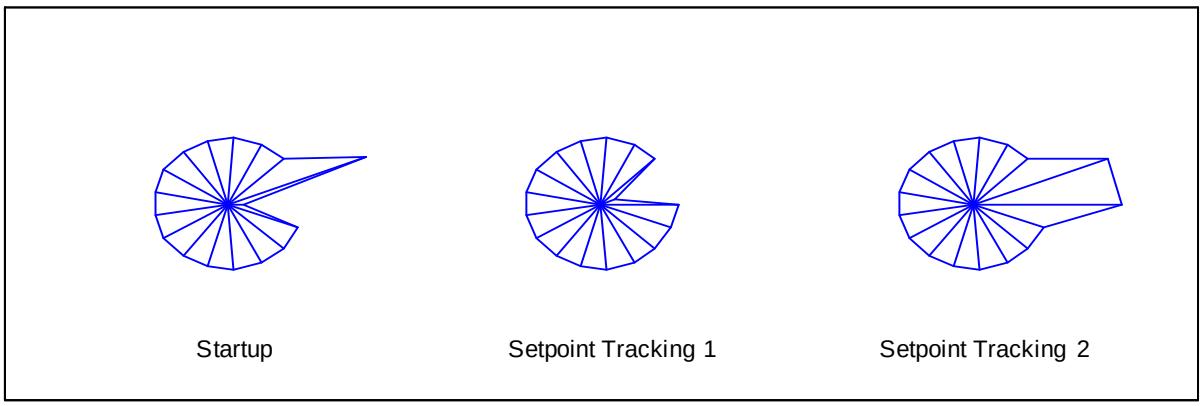

Figure 3.14: Glyph Star Plot of Example Distillation Column Data

3.3 Principal Component Analysis

As explained previously, multidimensional data is commonly attained from chemical plants. Often variables move together. This is due to the underlying driving forces that govern the process. Often data are redundant due to the excess of measurements compared to individual driving forces. Process control may also result in correlation between variables that are not normally related. The dimensionality of such a dataset could be reduced by replacing it with new variables that contain most (or all) of the important information. In statistical process control terms, important information is contained in *variance*.

Principal Component Analysis (PCA) is a linear technique for dimension reduction. PCA is also occasionally referred to as the Karhunen-Loéve Transform (KLT) or the Hotelling Transform (the reason for this is apparent from section 3.1). PCA is a linear transform. Data are projected for the original higher dimensional coordinate system to a lower order system with the first coordinate being the direction of any (linear) projection with the most variance. The second coordinate will then be in the direction of the projection with the second most variance and so forth. Dimensionality in the dataset can then be reduced by ignoring some of the resulting directions (typically ones that describe only small amounts of total variance). In this way, a model is obtained which captures as much variance as is desired by projecting the original variable linearly on to new coordinates. Importantly, these new variables are uncorrelated.

The PCA transform also has uses in creating models from images for computer vision (Smith, 2002).

3.3.1 Preliminary Data Preparation

Firstly the data must be validated as discussed in section 1.5.2. The PCA transform can be performed on raw data. This is usually only applicable where the variables have the same units.

The mean of each variable should be subtracted from that set of data. This is called centring (see equation 3.15). This is used with the covariance PCA method (see section 3.3.2).

The standard deviation is used to normalise the data column. Those data are then referred to as *standardised*. This can be represented as follows:

$$for \quad centring:$$
$$x_c = x - \overline{x} \tag{3.15}$$
$$for \quad normalising:$$
$$z = \frac{(x - \overline{x})}{s} \tag{3.16}$$

Equation 3.16 is known as the z-score. The s represents the standard deviation. This is generally used with the correlation method (see section 3.3.2).

The scaling as shown in equation 3.16 is particularly important when units differ (as is nearly always the case in chemical process data). Also the variance of different variables often differ widely. This would give undue weight to certain variables (Fourie, 2000: 7–11).

The variable z then will have mean of zero and a variance of one. It is important that, when comparing data, that the centring is done in a global manner to prevent misleading results (Martinez & Martinez, 2004: 22–23).

Scaling can also be done on the characteristic vectors. It is equivalent and probably conceptually more simple to directly -scale the data (Fourie, 2000: 7–10)

3.3.2 Derivation of the PCA Transform

With the centred data, the first principal component $\mathbf{w}_1$ of a dataset $\mathbf{x}$ is:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\mathbf{x}\right)^2\right\} \tag{3.17}$$

For the first $k-1$ components, the k^{th} component can be found by subtracting the first $k-1$ components from x as follows:

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\hat{\mathbf{x}}_{k-1}\right)^2\right\} \tag{3.18}$$
$$with$$
$$\hat{\mathbf{x}}_{k-1} = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i\mathbf{w}_i^T\mathbf{x} \tag{3.19}$$

The $\mathbf{w}$ matrix is equivalent to the a matrix of the eigenvectors of the covariance

matrix.

So the PCA transform is therefore equivalent to finding the singular value decomposition of a matrix of variables $\mathbf{X}$ (if the data matrix consists of N variables in columns of length M) :

$$\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^T \tag{3.20}$$

The reduced-dimensional data are then obtained by using the first $\mathbf{L}$ principal components ($\mathbf{W_L}$) as directions for $\mathbf{X}$ to be protected onto.

$$\mathbf{Y} = \mathbf{W_L}^T\mathbf{X} = \mathbf{\Sigma_L}\mathbf{V_L}^T \tag{3.21}$$

with $\mathbf{Y}$ a matrix of N column vectors and L rows. Each vector is a projection of the data vector from $\mathbf{X}$ onto the vectors (the principal components contained in $\mathbf{W}$).

Also the matrix $\mathbf{W}$ of singular vectors of $\mathbf{X}$ is equivalent to the eigenvectors of the matrix of observed covariances:

$$\mathbf{X}\mathbf{X^T} = \mathbf{W}\mathbf{\Sigma^2}\mathbf{W^T}. \tag{3.22}$$

PCA Using the Sample Covariance Matrix

The principal component analysis can be done using the sample covariance martix (Martinez & Martinez, 2004: 34). We start with a centred data matrix (see section 3.3.1 $\mathbf{X}_c$ of size $n \times p$). The sample covariance matrix is:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{x}_c^T\mathbf{X}_c \tag{3.23}$$

The eigenvectors are obtained by solving the following equations for $\mathbf{a_j}$:

$$\begin{aligned}(\mathbf{S} - l_j\mathbf{I})\,\mathbf{a}_j &= 0 \\ j &= 1, \ldots, p\end{aligned} \tag{3.24}$$

The eigenvalues and eigenvectors of $\mathbf{S}$ are then found. This is done by solving:

$$|\mathbf{S} - l\mathbf{I}| = 0 \tag{3.25}$$

This results in p eigenvectors which are orthogonal to each other. By convention, the eigenvalues (together with their associated vectors) will be stored in descending order.

The eigenvectors are simply the directions that the data will be projected onto. As shown in equation 3.26, this is a simple linear transformation. The elements of the

eigenvector can be thought of as the weight for each of the old variables. The eigenvectors are generally scaled to have unit length.

$$t_j = \mathbf{a}_j^T (\mathbf{x} - \overline{\mathbf{x}}) \tag{3.26}$$
$$j = 1, \ldots, p$$
$$or$$
$$\mathbf{Z} = \mathbf{X}_c \mathbf{A} \tag{3.27}$$

The matrix $\mathbf{Z}$ contains the principal component scores. These scores have zero mean because of the centred original data. If the transformation was done using $\mathbf{X}$ instead of $\mathbf{X}_c$, the scores would have some mean $\overline{\mathbf{z}}$.

We can also relate the scores back to the original variables as follows:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{A}\mathbf{z}. \tag{3.28}$$

PCA Using the Sample Correlation Matrix

The principal component analysis can also be done using the sample correlation matrix (Martinez & Martinez, 2004: 37–38). We again start with a centred data matrix (see section 3.3.1 $\mathbf{X}_c$ of size $n \times p$). The ij^{th} element of the sample correlation matrix $\mathbf{R}$ is:

$$r_{ij} = \frac{s_{ij}}{\sqrt{S_{ii}^2}\sqrt{S_{jj}^2}}. \tag{3.29}$$

S_{ij}^2 is the ij^{th} element of $\mathbf{S}$ as defined in equation 3.23. These represent the variance. The rest of the analysis is performed as with the covariance method.

Comparison of the Covariance and Correlation Method

These methods will not give identical results. The correlation method should be used when the variance between the original variables are very large. The normalisation of the variable prevents the principal components that represent the most variance from being dominated by the variables that have the most variance (Martinez & Martinez, 2004: 37). Using standard units also makes it possible to compare the different analysis results. This method will be used for the rest of the statistical algorithms used here onwards.

The covariance method is useful in that it is well described in general statistics literature. This adds some forms of statistical inference (Martinez & Martinez, 2004: 37).

There are other methods of deriving a PCA like transform. Van der Berg (2007) gives

details of the Non-linear Iterative PArtial Least Squares method (NIPALS algorithm) and Alternating least Squares Method (ARL algorithm) where the solution is (randomly) rotated in the factor space. These methods are much more computationally expensive and have little or no advantage over the covariance or correlation methods.

3.3.3 Dimensionality Reduction

It is generally found that after the PCA, the data can be adequately described using fewer factors than the original number of variables. These score variables are more generally normally distributed than the original variables (Groenewald et al., 2006).

The best PCA model would have the maximum number of components (corresponding to the number of variables). The dimensionality after the analysis to equation 3.27 is still p. However, as discussed previously, the PCA transform is needed as a dimensional reduction tool. As shown earlier, the sum of the eigenvalues is equal to the variance. The number of dimensions is reduced by neglecting some of the principal components. The obvious principal components to neglect are the ones that represent the least amount of the original variance. These correspond to he principal components with the smallest eigenvalues (Martinez & Martinez, 2004: 36).

The question now arises as to what degree the number of dimensions should be reduced to. There are some useful methods which can be used to give an indication if a a given number of principal components will represent the original data sufficiently. Discussion follows.

Cumulative Percentage of Variance Explained

This is a simple and popular technique (Martinez & Martinez, 2004: 38). The first k components that contribute to some cumulative percentage of the total variance (for all p components or original variables) are selected. The value for this cumulative total variance (t_d) ranges between 70% and 95 %.

The cumulative total variance is calculated using:

$$\tilde{l}_d = 100\frac{\sum_{i=1}^{k} l_i}{\sum_{j=1}^{p} l_j} \tag{3.30}$$

and for the correlation method:

$$\tilde{l}_d = \frac{100}{p}\sum_{i=1}^{k} l_j \tag{3.31}$$

Prediction Sum of Squares (PRESS)

The PRESS (**PRE**diction **S**um of **S**quares) procedure was introduced by Wold (1978).

As summarised by Fourie (2000: 7–13), the procedure is as follows:

If p is the number of variables and n is the number of (usually time dependent) samples, then the data can be represented in a $n \times p$ matrix $\mathbf{X}$. These data are then randomly divided time-wise in g groups. The first group is removed from the sample and the PCA is preformed on the remaining samples. The first principal component and then the first two (continuing until all p components are used) are then used to predict values of the deleted sample. For each predicted observation of the deleted sample, obtain the SPE statistic (explained in section 3.3.5).

After this has been completed, the removed sample should be replaced and another group should be removed. The calculation of the SPE statistic is then calculated as before again.

After this has been repeated for all g groups, the n SPE-statistics should then be summed for each k type of model (e.g. for the two component model). The PRESS statistic is then formed as follows:

$$PRESS_k = \frac{1}{np} \sum_{i=1}^{n} SPE_{ki} \tag{3.32}$$

Now, to check if the addition of the k^{th} principal component is warranted, W is calculated as follows:

$$W = \frac{(PRESS_{k-1} - PRESS_k) \cdot D_R}{D_M \cdot PRESS_k} \tag{3.33}$$

$$\begin{aligned} &with \\ &D_M = n + p - 2k \\ &D_R = p(n-1) - \sum_{i=1}^{k} (n + p - 2i) \end{aligned} \tag{3.34}$$

if $W > 1$, then the k^{th}principal component should then be retained and testing of the $(k+1)^{th}$ component should be tested in the same way. If $W < 1$ then that component need not be included. It is possible that after the first occurrence of $W < 1$ that later values of k will produce occurrences of $W > 1$. This may be due to outliers (Fourie, 2000: 7–15).

This technique is overly complex as compared to the other techniques. It does have the advantage of being quantitative, making it suited for computer calculation.

The Broken Stick Method

This methods makes use of the amount of variance explained by each component (Martinez & Martinez, 2004: 39).

If a line is *randomly* divided into p (corresponding to the maximum number of components or original variables) segments, then the expected length of the k^{th} longest piece is:

$$g_k = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i} \tag{3.35}$$

If the proportion of variance explained by the k^{th} component is greater than g_k, then the amount of variance that the component explains is greater than expected by pure chance. It would then be useful to keep this component.

The Size of Variance Technique

Using the correlation technique (see section 3.3.2), principal components with variance greater than 1 ($l_k \geq 1$) would be retained (Martinez & Martinez, 2004: 39). For the covariance technique, the component would be kept if its variance was greater that 70% of the average of all the variances, i.e.

$$l_k \geq 0.7\bar{l} \tag{3.36}$$

or

$$l_k \geq 1.0\bar{l} \tag{3.37}$$

$$\tag{3.38}$$

Occasionally, a cut of value of 100% rather than 70% is used. This method may be perferred due to its simplicity and ronbustness (Lee et al., 2004a). The justificiation is simply that the principal components contributing less than the average variance are probably insignificant.

The Scree Plot Method

The scree plot is a graphic method to gauge the amount of variance contributed by each component. It is a bar plot of l_k against k (the index of the component). The cumulative variance is also plotted as a line. To use the plot, the point where the line or the slope of a line between the blocks representing the value of l_k levels off. An example is shown in figure 3.15. In this example, between 2 and 4 principal components would be selected. A variant of this plot is the log-eigenvalue plot. This plot is used when the first few eigenvalues are much larger than the rest (Martinez & Martinez, 2004: 38).

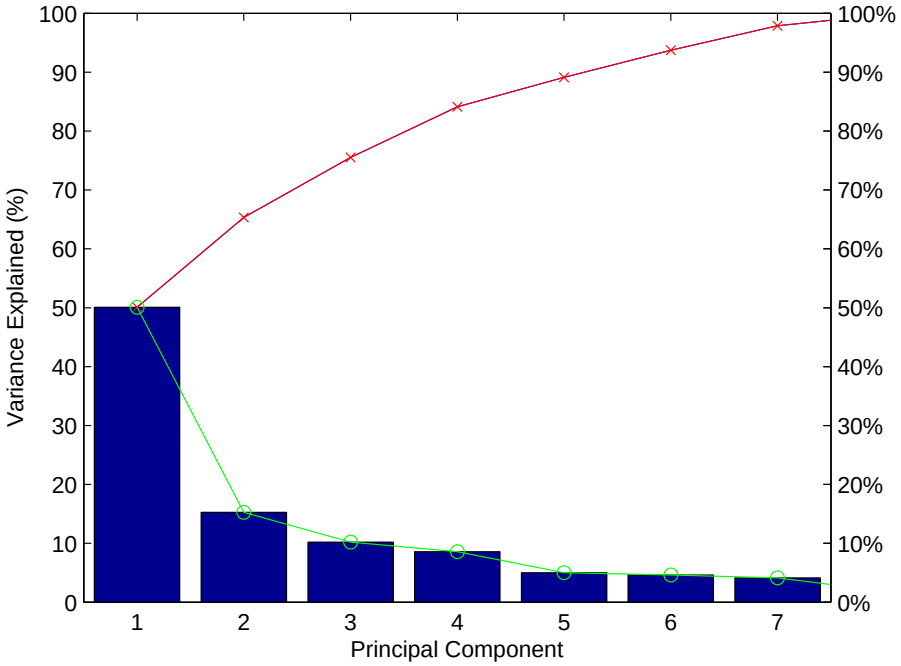

Figure 3.15: An Example of a Scree Plot

Other methods include the Akaike Information Criteria (AIC) mentioned in Lee et al. (2004a).

3.3.4 PCA compared to a Standalone Supervised Neural Network

Principal component analysis is frequently used together with neural network networks. The neural network is usually used for the discriminant or interpolation portion of the fault detection Rengaswamy et al. (2001). A supervised neural network can easily be used to preform a function similar to PCA directly (Crnkovic-Dodig, 2006).

There are several reason why PCA is preferable to a neural network:

- For large scale plant diagnosis, neural networks can become large and complex (Rengaswamy et al., 2001).
- Training becomes time consuming and difficult (Rengaswamy et al., 2001).
- It is computationally expensive.
- PCA is easily mathematically reproducible while a neural network is often more intangible and difficult to reproduce.
- A neural network has greater degrees of freedom, increasing the search space an adaption algorithm has to cover (Crnkovic-Dodig, 2006).

- A more complex search space results in a larger number of local minima (suboptimal solutions) for optimization to get stuck in.
- PCA is guaranteed to be the optimum linear feature extractor.

Neural networks are generally able to model nonlinear behaviour.

3.3.5 PCA for Fault Detection

PCA is the most widely used data-driven technique for process monitoring on account of its ability to handle high-dimensional, noisy and highly correlated data. Fault detection is done by monitoring biplots and statistical metrics resulting from the analysis.

Biplots

As shown in equation 3.27, the score vector X_{new} is created by the projection of the new data onto the PCA axes. The approach for plotting a biplot is simple. The scores are plotted with each principal component as an axis. Clearly for cases with more than 3 principal components, projections to 2 dimensional biplots must be made. Often each of the original process variables are shown as a vector on the same plot. This shows the contribution that each process variable has on the principal components. This also makes it possible to relate the score back to the original variables (critical for diagnosis using PCA). Faults can be detected using the biplot by training a normal operating PCA region with normal operating data. Any data outside this region has a greater than expected variance in the PCA space. Note that this approach will not detect all types of errors. It will only detect within model errors (see the sections on the T^2 and SPE statistics for an explanation of error types).

There are several ways of defining this region in biplots including the use of control ellipses, histograms (section 3.2.5), bagplots (section 3.2.3) along with kernel density estimates.

Using an ellipse method, an ellipse boundary is drawn around the normal operating region.If a new score is outside the boundary, it is faulty. An ellipse with origin $(0, 0)$is defined by:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{3.39}$$

a is the semimajor axis length and b is the semiminor axis length ($a > b$). These can be calculated from the covariance matrix. For the two-dimensional problem (Romagnoli & Palazoglu, 2006: 460–461):

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{21} \\ s_{12} & s_2^2 \end{bmatrix} \tag{3.40}$$

An ellipse that is rotated using linear algebra can also be drawn. This will give a better boundary for a skewed normal operating region.

Obviously $s_{12} = s_{21}$. Calculating this covariance matrix with the scores will allow the extraction of the diagonal elements as the variance. Now, an ellipse boundary of confidence α (for 3-σ limits, $\alpha = 3$) can be drawn, with a and b can be defined as:

$$a = \alpha\sqrt{s_1^2} \tag{3.41}$$

and

$$b = \alpha\sqrt{s_2^2} \tag{3.42}$$

Kernel density estimates are a method of finding a non-parametric, smoothed, nonlinear boundary. In essence, the algorithm finds a bandwidth and approximates the density of the data at the bandwidth intervals. The concept is similar to calculating a histogram by selecting the number/width of bins (the bandwidth) and counting the frequency of the data (the density). Fourie (2000) uses this method with biplots. Choosing the bandwidth can be difficult to do and the process is fairly computationally expensive compared to calculating a bagplot. Gardner et al. (2005) and Aldrich et al. (2004) discuss biplot methodology by making use of bagplots in a comprehensive manner. Groenewald et al. (2006) also uses histograms, with colour to indicate density.

The T^2 Statistic

The T^2 statistic is also known as the Mahalanobis distance (also see section 3.1 for a detailed discussion of the origins of this statistic).

It is possible to check a multivarible dataset of X being a $n \times 1$ vector of normally distributed variables with covariance matrix Σ to see if they meet a desired target of X_T by calculating the χ^2 statistic:

$$\chi^2 = (X - X_T)^T \Sigma^{-1} (X - X_T). \tag{3.43}$$

This statistic allows us to plot a control chart showing χ^2 against time. An upper control limit will be defined by χ^2_α, with α the level of significance for the test.

There are situations when the in-control covariance matrix is not known and it has to estimated. The estimation uses the sample of z past measurements as:

$$S_{kj} = \frac{1}{z-1} \sum_{i=1}^{z} \left(X_{ik} - \overline{x}_k\right) \left(X_{ij} - \overline{x}_j\right)^T. \tag{3.44}$$

Here $\overline{x}$ represents the estimate of the mean through a finite number of previous samples. The Hotelling T^2 statistic can then be represented as (Romagnoli & Palazoglu,

2006: 452):

$$T^2 = (X - X_T)^T S^{-1} (X - X_T) . \tag{3.45}$$

This can be further simplified by using score vector t_i of the first k principal components as follows:

$$T^2 = \sum_{i=1}^{k} \frac{t_i^2}{s_{t_1}^2} \tag{3.46}$$

with $s_{t_1}^2$ the estimated variance of t_i.

The upper control limit (UCL) is:

$$T_{UCL}^2 = \frac{(n-1)(n+1)k}{n(n-k)} F_{\alpha(k,n-k)}. \tag{3.47}$$

$F_{\alpha(k,n-k)}$ represents a $100\alpha\%$ upper critical point of the F distribution with k and $n-k$ degrees of freedom. The lower limit is clearly zero.

This statistic can be viewed as a type of within model distance metric (as discussed in section 3.1). It effectively measures from the mean of the data - similar to a univariate Shewart chart (section 2.5.3). If this distance is excessive as compared to the variance of the data, the process is not under statistical process control.

The Squared Prediction Error (SPE) Statistic

Monitoring the process via the T^2 statistic alone is not sufficient. This will only detect whether the variation of the first k principal components is greater than explained by the normal operating condition. If a new type of event or fault occurs (which is not in the data used to train the PCA model), then new principal components will appear. The new observation X_{new}, once projected, will move off the plane of the PCA model. We need to make use of a new statistic that will detect novel events by computing the squared prediction error (SPE) of the residuals of the new observation. This concept is shown diagrammatically in figure 3.16. The SPE statistic is also known as the Q statistic Romagnoli & Palazoglu (2006: 462).

The SPE statistic of a new observation is:

$$SPE_{new} = \sum_{j=1}^{n} (x_{new,j} - \hat{x}_{new,j})^2 . \tag{3.48}$$

The upper control limit for the SPE statistic is based on a χ^2 approximation as shown in Jackson (1991: 36–41) and Lee et al. (2004a):

$$SPE_{UCL} = g\chi_{h,\alpha}^2 \tag{3.49}$$

with $g = v/2m$ - the weight of the χ^2 distribution and $h = 2m^2/v$ - the degrees of

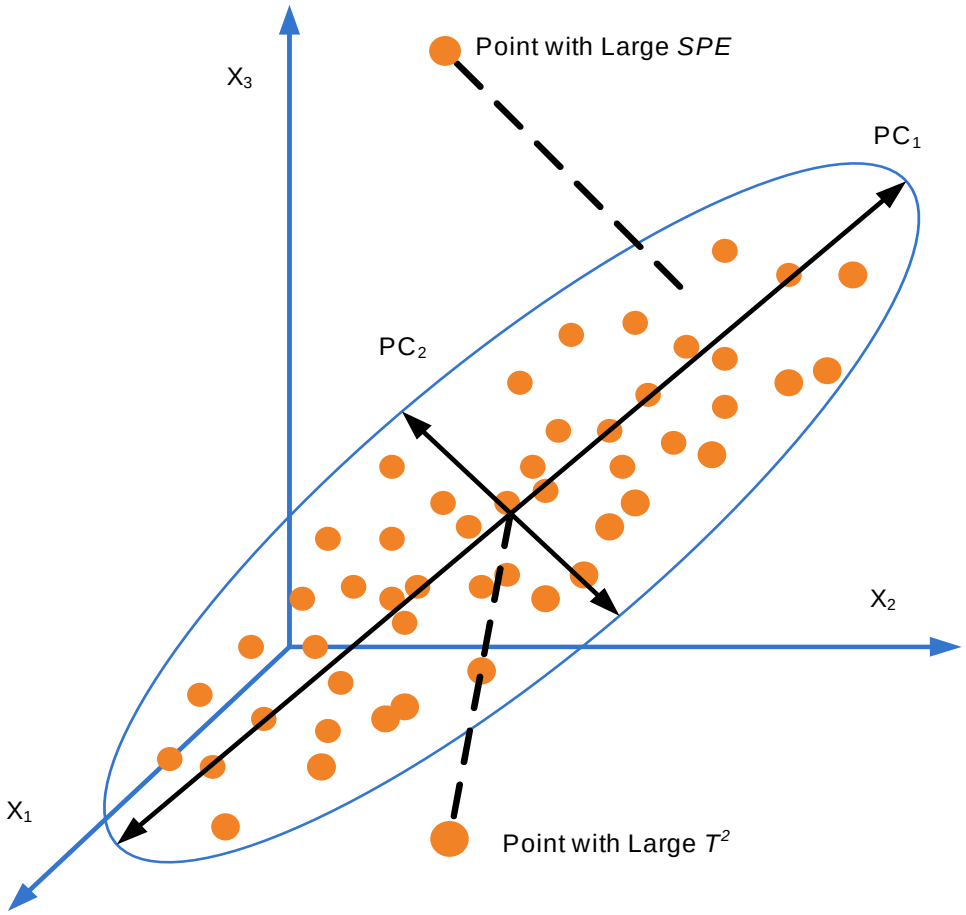

Figure 3.16: Comparison of the T^2 and SPE Statistics

freedom of the χ^2 distribution. m is the mean and v is the variance of the SPE normal operating region at significance level α.

The T^2 statistic can give false-negatives in fault detection due to the latent space sometimes being insensitive to changes in sensor arrays or to small process upsets (Romagnoli & Palazoglu, 2006: 462–463). This is because each variable is a combination of all process variables. A fault in one process variable may not have enough weight to trigger an out-of-control indication. On the other hand, the SPE measure is more sensitive to these types of changes than the T^2 or biplot. This is because any type of fault will propagate to model error.

All PCA estimated variables will be influenced by any type of disturbance in the input space. At the instant that the disturbance or fault occurs, it is more likely to manifest itself in the SPE than in the T^2 (Romagnoli & Palazoglu, 2006: 463). This corresponds to region I in figure 3.17. There may be significant faults that trigger the alarm in both measures (region II in figure 3.17). There may also be process upsets that remain undetected by the SPE statistic due to the extrapolating feature of the PCA model. In this case the latent space of the PCA model will capture the changes and no violation of SPE will be observed (region III in figure 3.17). If neither of the T^2 or SPE

limits are violated, then the process is probably operating normally.

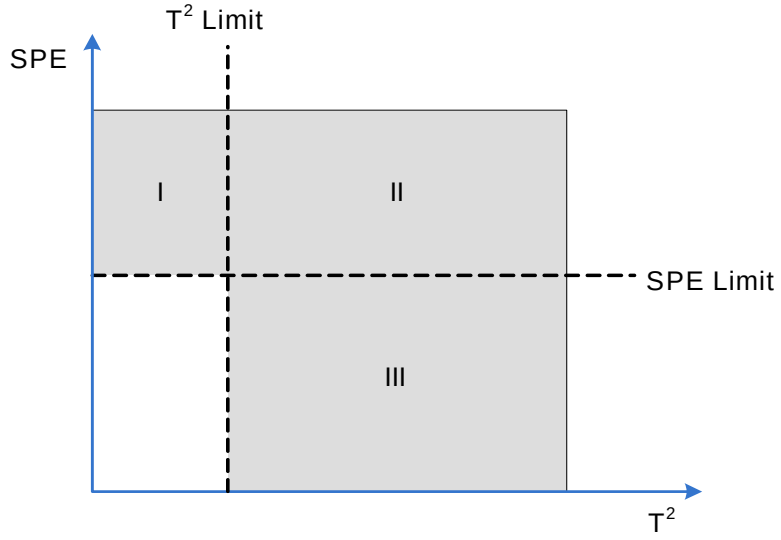

Figure 3.17: Use of both SPE and T^2 techniques for Fault Detection

Chen et al. (2004) advocates synthesising the T^2 and SPE statistics together into one index using a density estimate to obtain a joint distribution of the two statistics. This eliminates the need to have two charts and is more sensitive.

3.3.6 PCA for Fault Diagnosis

Once a fault has been detected, there are two main analytical options available to diagnose the cause of the fault:

1. Contribution plots
2. Classification by training regions.

Contribution Plots

Romagnoli & Palazoglu (2006: 463) discuss contribution plots. By interrogating the underlying PCA model at the point where the fault event has been detected to occur (usually the new sample exceeding the SPE or T^2 limits), one can extract a contribution plot that reveal the process variable or group of variables that the contribute to the deviation in the scores and SPE. This will probably not diagnose the cause of the fault unequivocally, but it will give insight into the possible cause and narrow the search for a diagnosis.

Classification by Training Region

Known faults from historical data can be projected on to the PCA model created with the normal operating data. With some luck, these will form distinct regions corresponding to each fault. Future fault can then be classified by which region they fall into. This discrimination between regions can be improved by techniques such as linear discriminant analysis (section 4). This method requires a comprehensive library of prior faults. Novel fault diagnosis is difficult or impossible to do directly and methods such as contribution plots must be used. As before, the contributions could be counted using the projections back on to the principal component axes. An additional rotation is induced by the LDA. This means that the data needs to be projected twice to get the contribution of the variables to each faulty region.

3.3.7 Fault Detection and Diagnosis Procedure

The use of a T^2 and a SPE chart is an effective combination for detecting faults with linear PCA (Kourti & MacGregor, 1995). Diagnosis is then performed by means of contribution plots and/or training regions. Macgregor & Kourti (1995), Kourti et al. (1996), Groenewald et al. (2006) and Kourti & MacGregor (1995) give good overviews of complete linear PCA fault detection and diagnosis systems. They generally follow the scheme shown in figure 3.18. The number of principal components to be used will be chosen by the size of variance technique (section 3.3.3). Note that the faulty and new data are normalised using the mean and variance of the normal dataset. Linear discriminant analysis can used to improve the discrimination between the regions.

3.3.8 Competitors to Principal Component Analysis

While principal component analysis is the optimum linear feature extractor, there are several other techniques that make use of a PCA transform as a step. There are also other methods of reducing multivariate dimensionality but these do not yield true principal components and thus give up some of the optimal properties associated with PCA (Jackson, 1991: 424–434). Examples of such techniques include image analysis, principal curves (Dong & McAvoy, 1996) and triangularisation techniques. Arbitrary components is another technique that could possibly be applied to well understood process requiring monitoring. Components are simply chosen subjectively to correlate well with variables thought to be important. A further alternative to this is to eliminate variables completely by:

1. Correlation methods

2. Backward iterative elimination using PCA

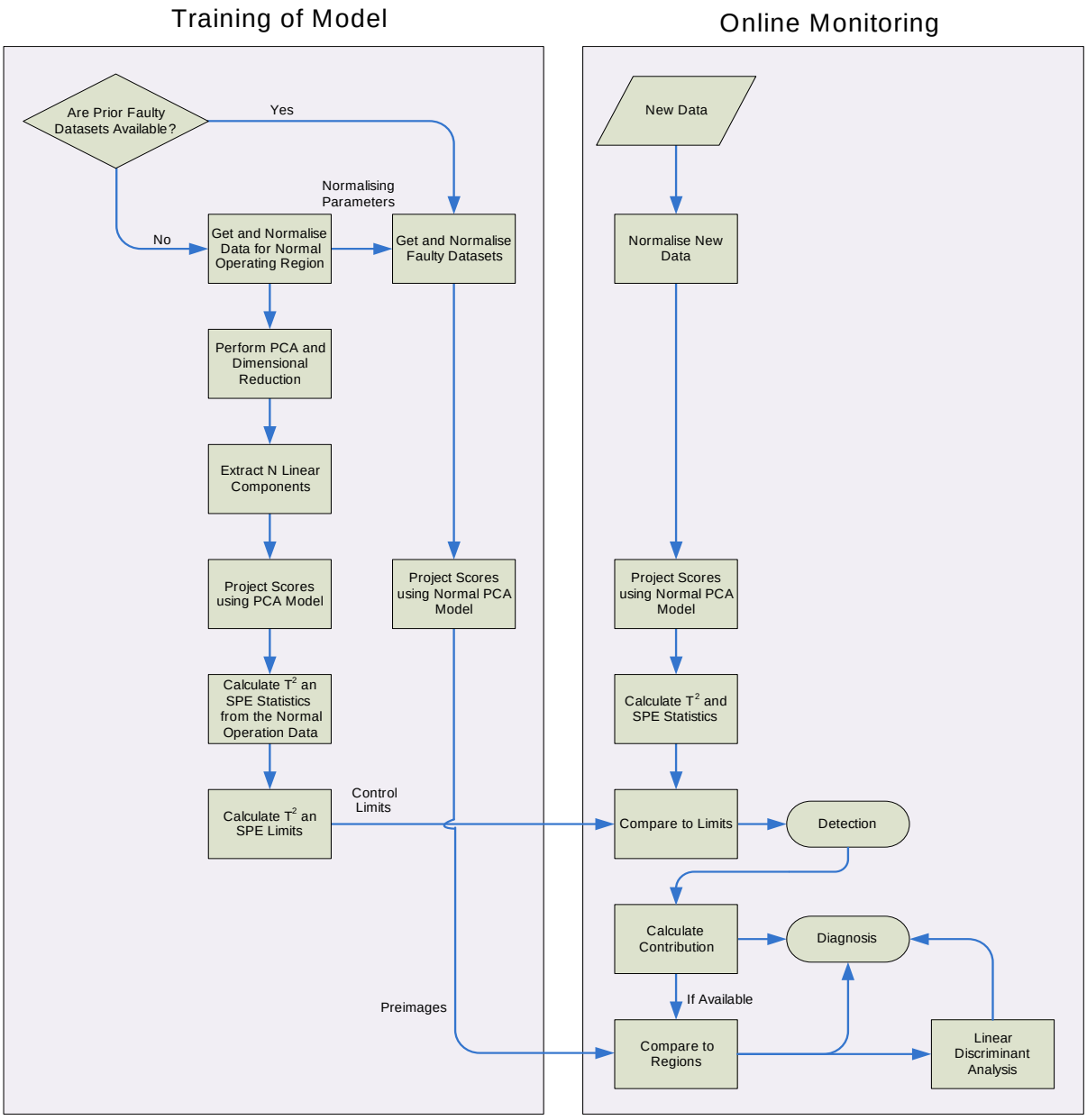

Figure 3.18: Linear PCA Fault Detection and Diagnosis Procedure

3. Hierarchical cluster methods.

Some directly reduced variable set can be then be used for analysis.

Andrew's Function Plots

Andrew's function plots also operate with a combination of the original variables (Jackson, 1991: 432–434). The data are transformed to a continuous function instead of new variables. The most commonly used function is:

$$f(t) = \frac{X_1}{\sqrt{2}} + X_2 \sin + 2\pi t X_2 \cos 2\pi t + \cdots \tag{3.50}$$

The curve is clearly influenced by the order of the variables. If the data are of different orders of magnitude, it should be rescaled. As a result of the use of the trigonometric functions, the following properties of the data are retained (Martinez & Martinez, 2004: 322):

1. Means
2. Distances (to some degree)
3. Variances

This means that data that are close together should result in Andrew's curves that are close together. This means that the curves can be used for clustering or classification.

An example of an Andrew's function plot is shown in figure 3.19 using the same dataset discussed in section 3.2.1. It is also easy to plot with just the median for each group and some confidence limit as shown in figure 3.20. As with the parallel coordinate plots, it is clear that the faulty data has a great variance. Any faults can be discerned as changes in the values of the data or the function. Diagnosis can be performed by observing a change from normal operating conditions to a different group. While the normal and faulty curves are largely similar, it is clear that some samples of the faulty dataset are markedly different. It is not possible to directly see (as it was with the parallel coordinate plots of section 3.2.1) which variables are to blame. While fault detection is easy, it is difficult to diagnose the fault.

Bersimis et al. (2005) refers to an algorithm for solving the problem of interpreting an out of control signal.

The advantage of the Andrew's plot is that it can transform a a large number of variables into a single function in a manner that is simple to calculate using a computer. The disadvantage is that it is quite difficult to interpret or relate back to the multivariate original values.

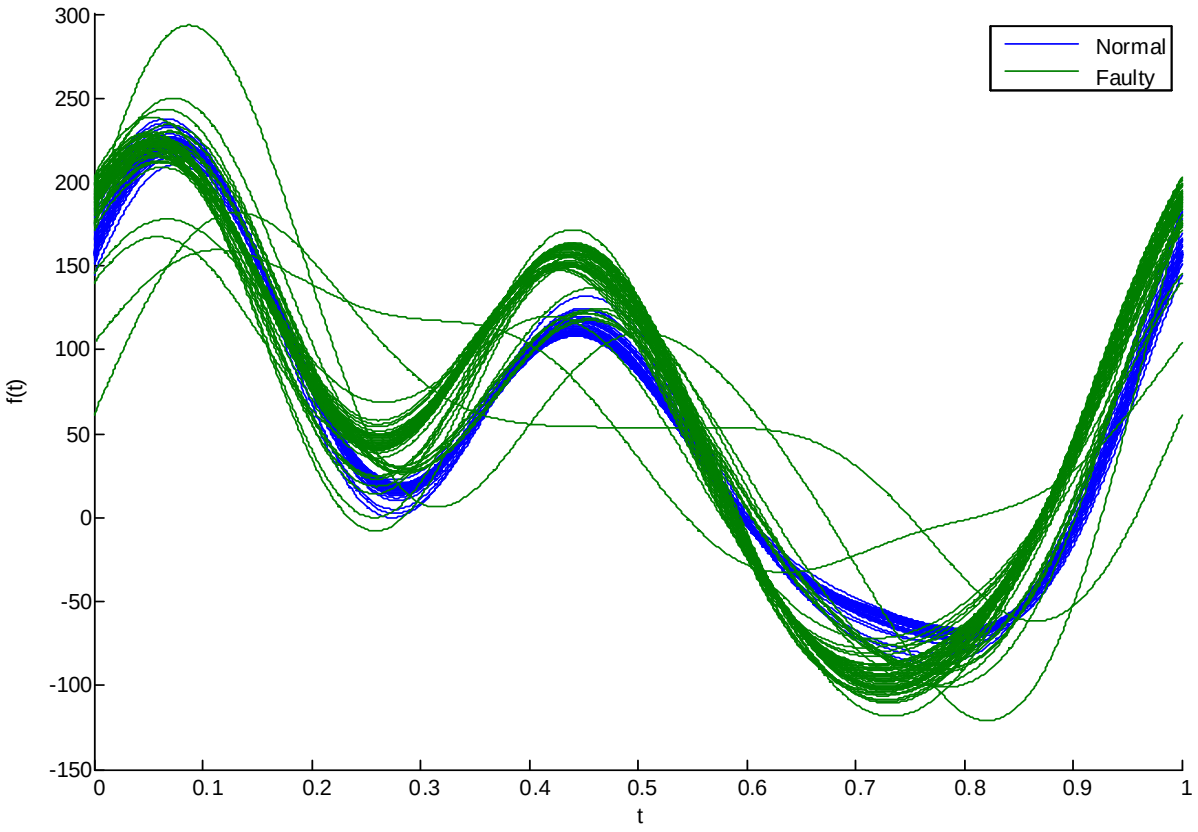

Figure 3.19: Example of an Andrew's Plot

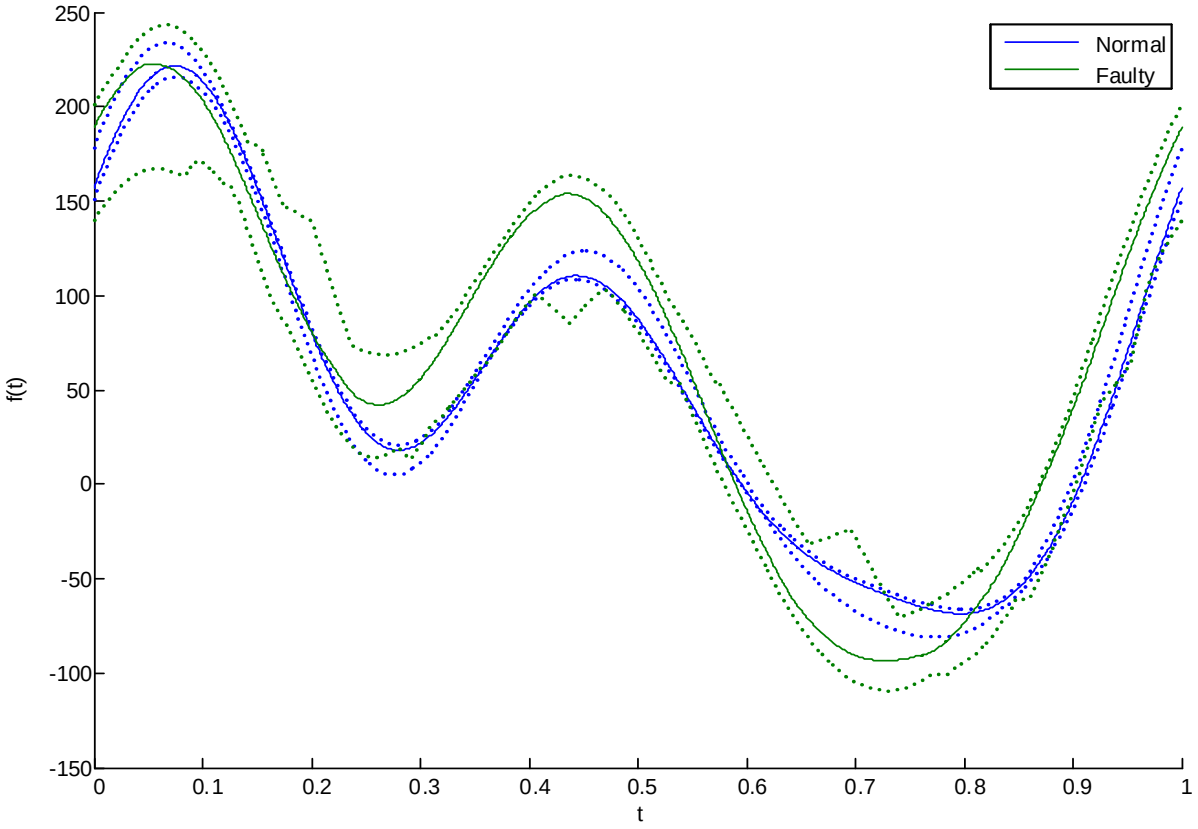

Figure 3.20: Example of an Andrew's Plot with Medians and 95% Quantiles

Partial Least Squares Techniques

Conceptually similar to principal component analysis, partial least squares is also a data modelling technique that can be used for dimensional reduction (Venkatasubramanian et al., 2003b). While PCA generally lumps process inputs and process variables together in an analysis, PLS is used to create a cause and effect type model. Latent variables that describe the variance in the process data while predicting the process quality are calculated. The model is used in a similar way to a PCA model - checking incoming data for mismatch. Goulding et al. (2000) shows an interesting case of the application of PCA and PLS. Macgregor & Kourti (1995) covers some traditional applications of PLS fault detection.

3.4 Kernel Principal Component Analysis

3.4.1 Principal Component Analysis as a Linear Method

Linearity is a ubiquitous assumption in multivariate statistics. Most of the above procedures assume linear relationships or apply linear limits. Few relationships in any physical system are linear over their entire range. The simplicity and elegance of these assumptions often outweigh the fact that it may be suboptimal.

When no existing information is available about the nature of the nonlinearity (for example a model), or the techniques to analyse the nonlinearity are too complex, the following options are available (Harris, 1985: 334–337):

1. Proceed with a linear analysis on the basis of convenience and ease of interpretation.
2. Perform an initial linear analysis, followed by a test for nonlinearity and then decide if a modified analysis is necessary.
3. Perform an analysis employing a polynomial model which is likely to provide closer approximation to any function, followed by the deletion of terms in the model which do not make a statistically significant solution.

The easiest tests to check for nonlinearity are simple plots and residual scores to straight lines. From these plots, it can be seen if there is a constantly linear relationship between the variables. Another obvious method is to plot the variable against the variable reconstructed through the model.

As discussed before, PCA finds an orthogonal transformation of the coordinate system in which the data were described. The principal components are linear combinations of the original variables. PCA is the optimal transform for linear data, however, for nonlinear processes, PCA performs poorly due to the assumption of linear characteristics (Dong & McAvoy (1996) and Lee et al. (2004a)). The effects of the assumption of

linearity must be questioned when using techniques such as principal component analysis and while using statistics such as the T^2 statistic. In the case of principal component analysis, the first components may not describe as much of the variance as expected. Gifi (1990: 102–104, 151–191) cover the problems of nonlinearity in principal component analysis from a theoretical point of view.

We can see in figure 3.21, linear PCA has easily found the linear relation in the data shown in the first row. The first principal component is clearly along the direction of the most variance with the second principal component orthogonal to it. In the second row (containing strongly nonlinear data), linear PCA has found a linear relationship that describes most of the variance, however, it can be clearly seen that it does not show the effect of the nonlinearity in the data. A linear PCA model of this process would be very poor indeed.

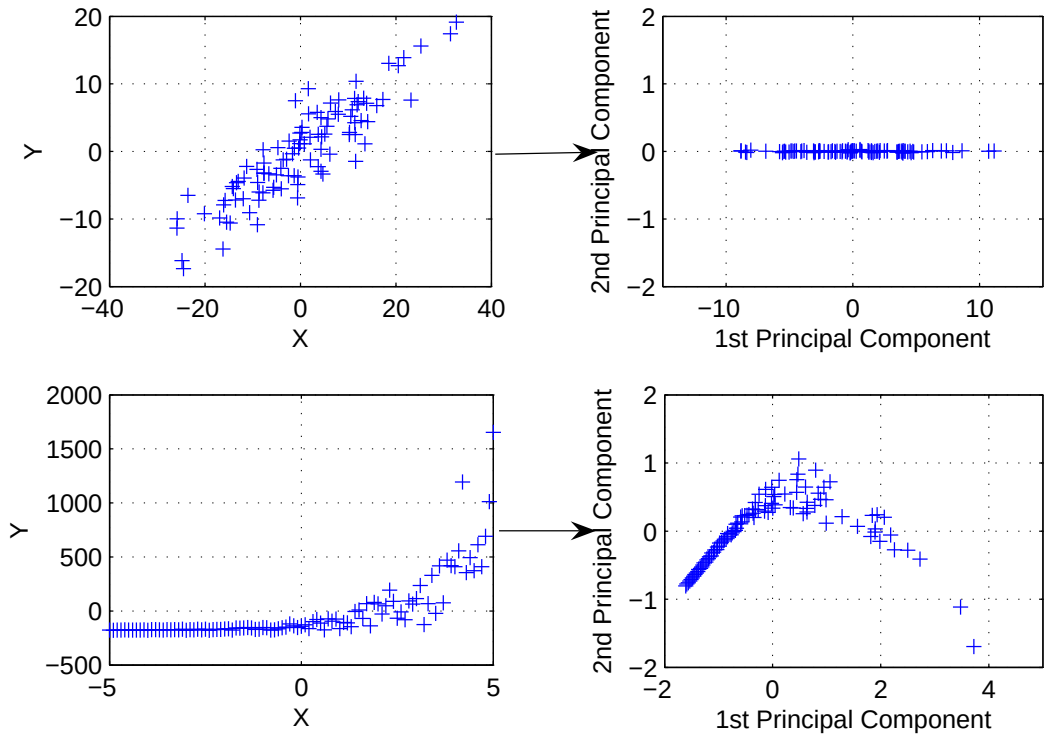

Figure 3.21: Principal Component Analysis on Linear (top row) and Nonlinear Data (bottom row)

3.4.2 Derivation of Kernel Principal Component Analysis

Kernel Principal Component Analysis (KPCA) was first put forward by Scholköpf et al. (1998). KPCA is by no means the only nonlinear principal component analysis technique. Dong & McAvoy (1996) proposed the use of principal curves. There are also several examples of the use of neural networks (e.g. Fourie (2000)) and support vector

machine (SVM) techniques. The key advantage of KPCA is that of reduction of computational requirements in that no nonlinear optimisation is required. Also the simple PCA calculation (section 3.3.2) can be used directly.

Conceptually, KPCA consists of two steps:

1. Mapping of data from the input space to a higher dimensional feature space.
2. PCA calculation in the feature space.

The derivation shown is according to Scholköpf et al. (1998) and Lee et al. (2004a) Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in R^m$ be the training observations for kernel PCA learning. We make use of a nonlinear mapping $\Phi :\in R^m \rightarrow z \in R^h$. The dimensionality of the feature space h can be large - possibly infinite.

The sample covariance in this feature space is then:

$$\mathbf{S}_\Phi = \frac{1}{n}\sum_{i=1}^{n}\left(\Phi(\mathbf{x}_i) - \mathbf{m}_\Phi\right)\left(\Phi(\mathbf{x}_i) - \mathbf{m}_\Phi\right)^T \tag{3.51}$$

where :

$$\mathbf{m}_\Phi = \sum_{i=1}^{n}\frac{\Phi(\mathbf{x}_i)}{n}. \tag{3.52}$$

The data are then centred after mapping. The mapped point is $\overline{\Phi}(x_i) = \Phi(x_i) - \mathbf{m_\Phi}$. A principal component $\mathbf{v}$ is then calculated by solving the eigenvalue problem (with $\lambda > 0$ and $\mathbf{v} \neq 0$):

$$\lambda\mathbf{v} = \mathbf{S}_\Phi\mathbf{v} = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{\Phi}(\mathbf{x}_i)^T\mathbf{v}\right)\overline{\Phi}(\mathbf{x}_i). \tag{3.53}$$

Then multiplying on both sides by $\overline{\Phi}(x_j)$, we get:

$$\lambda\left(\overline{\Phi}(\mathbf{x}_j)\cdot\mathbf{v}\right) = \overline{\Phi}(\mathbf{x}_j)\cdot\left(\Sigma_\Phi\mathbf{v}\right). \tag{3.54}$$

The mapping into a higher dimensional space can give rise to computational difficulties, since the size of the sample required increases exponentially with the dimension of the space (Jemwa & Aldrich, 2006). Key to the use of this technique for multivariate fault detection, is the concept of the kernel trick. If the algorithm to be computed can be expressed in terms of dot products in the feature space (R^h), the mapping can be done

implicitly by using kernel functions in the input space. Thus:

$$K_{ij} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \tag{3.55}$$

or for a centred kernel:

$$\overline{K}_{ij} = \overline{\Phi}(\mathbf{x}_i) \cdot \overline{\Phi}(\mathbf{x}_j), \tag{3.56}$$

which, crucially, is in $R^{n \times n}$ instead of the possibly infinite dimension h. This means that the mapping can have the same dimensionality as the input space. This means that the PCA calculation will probably be fairly easy (although not as easy as in the $R^{n \times 1}$ space for linear PCA). This allows the same results that would be prohibitive in the R^h space to be calculated quickly. This is the key concept behind the use of KPCA. This was the reason why the problem was formulated in such a way which only makes use of the values of dot products in the feature space. The choice of the kernel function then implicitly determines the mapping Φ and the feature space.

In summary, the kernel function allows us to compute the value of the dot product in the feature space without having to carry out the mapping Φ. This is critical as, for example, with a 16 dimensional input space and polynomial mapping (of order 5), the feature would have dimensionality of 10^{10}. It would clearly be difficult to perform PCA with this dimensionality. In contrast, using the kernel trick, the (input) dimensionality is very acceptable for fast PCA.

The problem can now be represented in the following simplified form:

$$\lambda \mathbf{v} = \frac{1}{n} \overline{\mathbf{K}} \mathbf{v}, \tag{3.57}$$

for all $\lambda > 0$ and with $\mathbf{v}$ the eigenvector and λ the eigenvalue. The centred kernel matrix for n observations is easily calculated as:

$$\overline{\mathbf{K}} = \mathbf{K} - \mathbf{I}_n \mathbf{K} - \mathbf{K} \mathbf{I}_n + \mathbf{I}_n \mathbf{K} \mathbf{I}_n \tag{3.58}$$

where:

$$\mathbf{I}_n = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \in R^{n \times n}. \tag{3.59}$$

see Scholköpf et al. (1998) for further details with regard to this matrix centring.

Note, that for normality of the principal component (i.e. $\|\mathbf{v}\|^2 = 1$), the calculated $\mathbf{v}$ must be calculated so that $\|\mathbf{v}\| = 1/n\lambda$.

After training the principal components in the feature space, the k^th projection of the centred value $\overline{\Phi}(\mathbf{x}_{new})$ is calculated by:

$$t_{new,k} = \left(\mathbf{v}_k \cdot \overline{\Phi}(\mathbf{x}_{new})\right) = \sum_{i=1}^{n} \mathbf{v}_i^k \left(\overline{\Phi}(\mathbf{x}_i) \cdot \overline{\Phi}(\mathbf{x}_{new})\right) \tag{3.60}$$

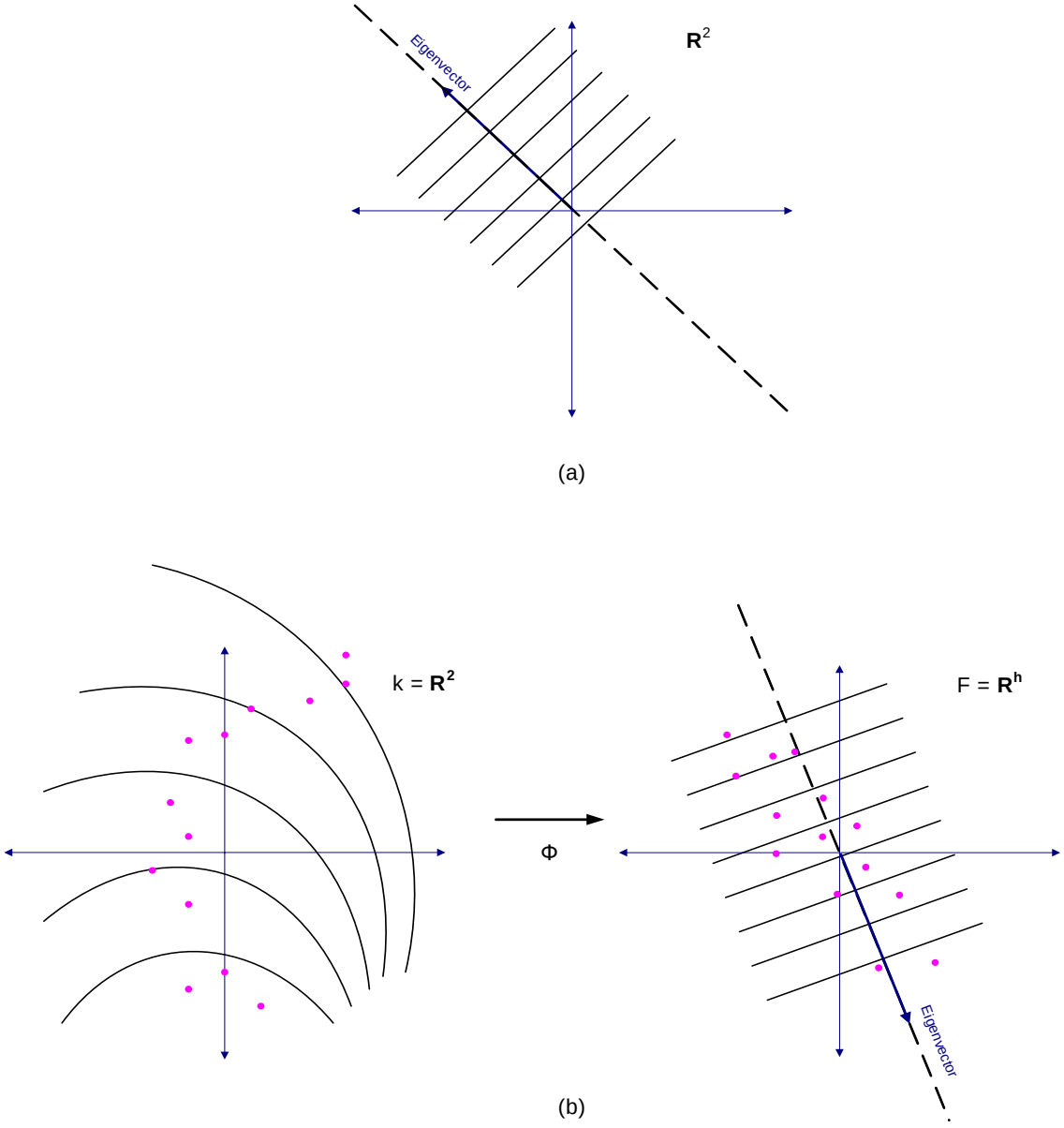

Figure 3.22: Linear PCA (a) compared to Kernel PCA (b)

Figure 3.22 shows how in the higher dimensional feature space, linear PCA is performed just like linear PCA is performed in the input space. The eigenvector in the feature space becomes nonlinear in the input space. It is very difficult, often impossible, to draw the preimage of the the eigenvector back into input space from feature space as

it may not even exist (Scholköpf et al., 1998).

New samples are scaled with the same parameters as the normal operating model. The kernel vector is then calculated as before. The test kernel $\overline{\mathbf{k}_t}$ is then centred as:

$$\tilde{\mathbf{k}_t} = \mathbf{k}_t - \mathbf{I}_t\mathbf{K} - \mathbf{k}_t\mathbf{I}_n + \mathbf{I}_t\mathbf{K}\mathbf{I}_n \tag{3.61}$$

with $\mathbf{K}$ and $\mathbf{I}_n$ as in the centring in feature space step (equation 3.58). The eigenvalue problem is then solved, and the scores calculated as before.

As in linear PCA before, the first components carry more variance than any other components. The principal components are uncorrelated and orthogonal.

In contrast to linear PCA, where one principal component is generated for every variable, KPCA has the ability to create a number of principal components that can exceed the input dimensionality. For most kernel choices, the number of principal components will equal the number of observations (or data points).

3.4.3 Kernel Functions

Examples of kernel functions include:

The Gaussian Radial Basis Functions:

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \tag{3.62}$$

Polynomial Functions:

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.63}$$

Sigmoidal Functions:

$$k(\mathbf{x}, \mathbf{y}) = \tanh(b_1\mathbf{x} \cdot \mathbf{y} + b_2) \tag{3.64}$$

$\mathbf{x}$ and $\mathbf{y}$ are given data set vectors to show the use of the dot product. Here, the parameters σ, b_1, b_2 and d are usually chosen by some kind of grid sampled model validation procedure. Empirically, experimentation shows that while using a Gaussian kernel, an initial guess of σ equal to the average standard deviation of the data works well. Clearly other kernels also satisfy $k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$. This criterion is called Mercer's Theorem (Lee et al., 2004a). Note that the sigmoidal kernel only satisfies the criteria for some values of b_1 and b_2. Also, note that $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})$ reduces to linear PCA. Guassian kernels are the only type of kernel applied to fault detection and diagnosis in literature. The type of kernel must be suitable to model the data.

3.4.4 Variants of Kernel Principal Component Analysis

There are many variants of KPCA. They are usually simply kernelised variants of extensions of linear PCA. These variants include the Dynamic KPCA (DKPCA) presented by (Choi & Lee, 2004). This technique extracts the relationships between past and current values of the measured variables by making use of a time lagged matrix.

Multiway KPCA (MKPCA) is a technique that can be applied to batch and batch-like (e.g. discontinuous start-up and shutdown) processes. It is also the kernelised equivalent of multiway linear PCA (Lee et al., 2004b). The MKPCA techniques are successful as batch processes are typically nonlinear. Each phase of the batch process is unfolded (by functional phase or time unit) these are compared to a metric or to previously calculated normal operating trajectories. A valuable and well presented tutorial of MPCA is given by Singh (May 2003).

Fourie (2000) and Fourie & de Vaal (2000) presented multiscale nonlinear PCA (using neural networks as opposed to kernels) computing the PCA of wavelet coefficients at each scale and then combining the results at relevant scales. Kernelised PCA could conceivably be applied for this technique.

3.4.5 Motivational Example of Kernel Principal Component Analysis

The example shown here is a modification of the ones presented by Scholköpf et al. (1998) and Lee et al. (2004a). This was done to validate the coded KPCA algorithm. The data points are generated using a simple second order polynomial with normal noise. A linear PCA model as well as a KPCA model is generated. This demonstrates the difference between the two techniques in their ability to model nonlinear data. The KPCA example uses a Guassian kernel with a kernel argument of $\sigma = 1$.

As can be seen in figure 3.23, lines of constant principal component value in the input space are shown as contours. For KPCA, there may be no empirical representation of these lines in the input space whereas for the linear PCA they are simple lines. Notice how the linear PCA merely found the axes of most variance through the 150 points (as expected). These directions are merely the horizontal and vertical directions. This model would explain none of the nonlinearity clearly seen in the input space. The principal components are represented in the plots on separate rows. The number of principal components (2) corresponds to the input dimensionality. The KPCA found relationships that describe the nonlinear parabolic function that was used to generate the data. Additionally, there are 150 principal component directions (corresponding to the number of points). Only the first three are shown. Only the first 2 principal components explain the underlying function directly. The others, explaining only small amounts of the variance, are mainly some description of the noise relationships. Therefore reducing the

dimensionality to 2 (in this case) would be adequate to provide a model that describes the nonlinearity well (but not all the noise). Notice how the first few eigenvectors of KPCA explain less of the variance than the first two PCA eigenvectors. This is because of the 150 eigenvalues in this example of KPCA. The first few find the most important relationships while the remainder explain less important relationships. This may affect the methods used to choose the number of dimensions to retain.

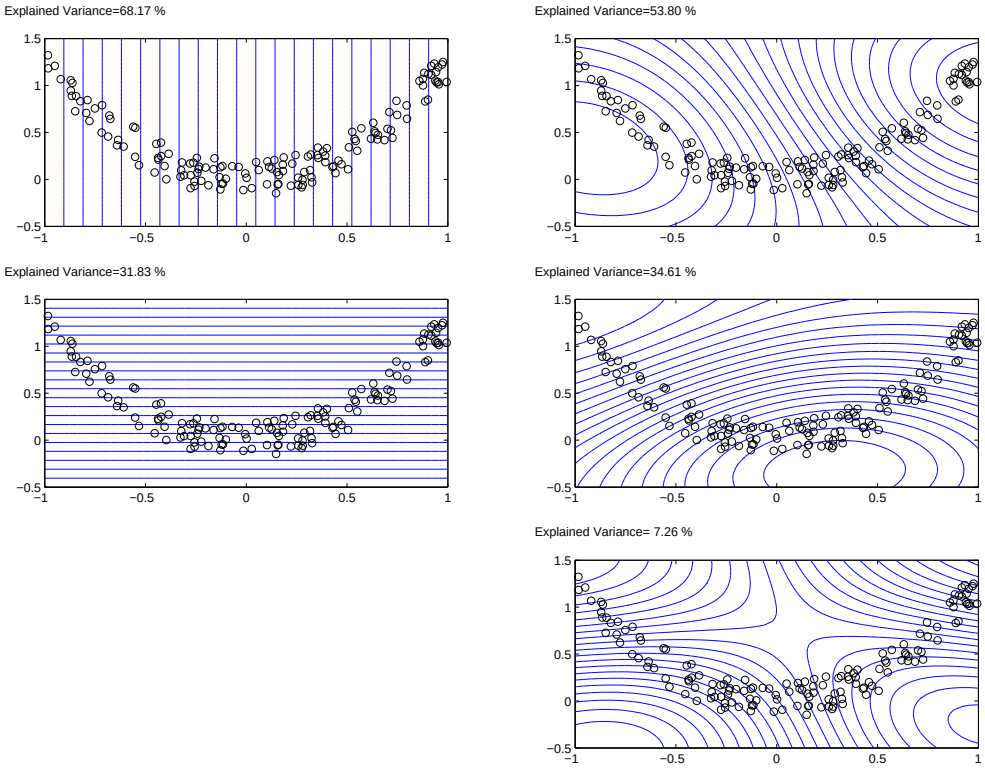

Figure 3.23: Motivational Example for KPCA (principal components in rows; left column: Linear PCA; right column: KPCA)

3.4.6 KPCA for Fault Detection

As with linear PCA, the T^2 and SPE statistics can be defined for KPCA. See section 3.3.5.

As before, the measure of variance with the KPCA model is given by the T^2 statistic. For KPCA is is given by (Lee et al., 2004a):

$$T^2 = T_i \Lambda^{-1} T_i^T \tag{3.65}$$

with T_i as the score matrix and Λ^{-1} as the diagonal matrix of the inverse of the eigenvalues of the retained principal components.

Again the upper control limit (UCL) is:

$$T^2_{UCL} \approx \frac{p(N-1)}{N-p} F_{p,N-p,\alpha} \tag{3.66}$$

with p the number of principal components and N the number of samples (Lee et al., 2004a).

The measure of goodness of fit of a sample to the KPCA model is the SPE statistic. As discussed above, there is no way of reconstructing all data from the feature space. Lee et al. (2004a) proposes a method of calculating SPE in the feature space. Now, we define:

$$SPE = \left\| \Phi(x) - \hat{\Phi}_p(x) \right\|^2 \tag{3.67}$$

after some manipulations, this reduces to:

$$SPE = \sum_{j=1}^{n} t_j^2 - 2\sum_{j=1}^{p} t_j^2 + \sum_{j=1}^{p} t_j^2 \tag{3.68}$$

$$= \sum_{j=1}^{n} t_j^2 - \sum_{j=1}^{p} t_j^2. \tag{3.69}$$

with n the number of non-zero eigenvalues and p the number of non-zero eigenvalues that are retained after the dimensional reduction.

As with linear PCA:

$$SPE_{UCL} = g\chi^2_{h,\alpha} \tag{3.70}$$

Choi et al. (2005) presents different statistics. They also advocate the use of a unified index.

Jemwa & Aldrich (2006) advocates the use of KPCA as a feature extractor only (see section 4. This is because the PCA statistics are defined for the input space variables and not for the feature space. This means to calculate the SPE and T^2 meaningfully, a mapping back to the input space is required. This is an ill posed problem as some of the points in feature space have no corresponding exact image in the input space. There are algorithms to find approximate preimages.

These use of these statistics are questionable as they are based in the feature space. Also as the number of observations used to train KPCA becomes larger, the control limits also become unrealistically large (Choi et al., 2005).

3.4.7 KPCA for Fault Diagnosis

As with linear PCA, regions of normal operating conditions and regions describing the various faults can be trained. The diagnosis of the fault would then correspond to the region that a newly project point on the biplot falls into. It is important to note that because no exact preimage of an eigenvector of the original process variables exists in an input space, the contributions to score of a new sample by process variables cannot be evaluated. This means that expert process knowledge is required to relate the fault region that new data points are projected into, to the variables on the process.

There are ways to get an approximate preimage empirically (as can be seen in the contour lines of figure 3.23). These are still not much use for diagnosis as they involve complex nonlinear optimisation. Cho (2007) applies this technique to fault detection and diagnosis on a nonlinear batch process. He shows improved fault diagnosis performance as compared to the linear methods.

While projecting data into regions is a valid approach, the separation between regions may not be enough to unambiguously diagnose a fault. Nonlinear discriminant analysis can be used to improve the classification.This will be discussed in more detail in chapter 4.

3.4.8 KPCA Fault Detection and Diagnosis Procedure

Bergh et al. (2005), (Choi et al., 2005) and Lee et al. (2004a) show complete KPCA based fault detection and diagnosis procedures.

A flow chart of the procedure for fault detection and diagnosis is shown in figure 3.24. Note that mean-centring of the non-model data refers to equation 3.61.

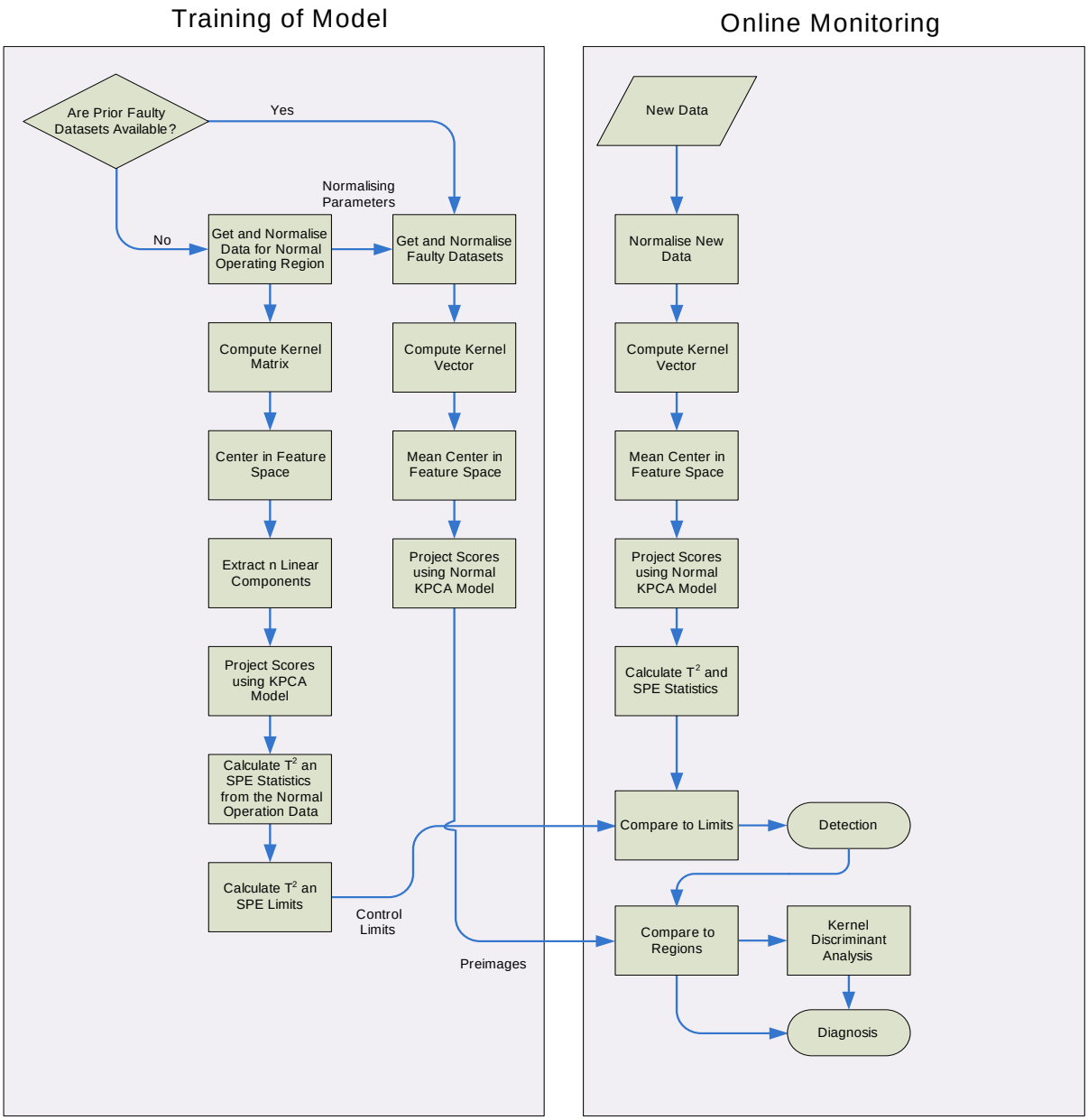

Figure 3.24: KPCA Fault Detection and Diagnosis Procedure

CHAPTER 4

Classification and Discrimination

Discriminant techniques can be applied to features extracted from data to aid classification between groups or classes within the data. These techniques can be applied to isolate fault classes. This may aid fault diagnosis.

The difference between feature extraction and feature classification will be covered in greater detail after the workings of linear discriminant analysis (being a simple classification and discrimination technique) are discussed.

4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised learning classification technique. This means that the algorithm will learn to separate data from different classes based on pre-classified training groups. A linear combination of features is found which will best separate the classes of event. This classification rule can then be used to classify new data samples. The classification of a data sample is useful in monitoring and diagnosis of plant data (Jemwa & Aldrich, 2005). The purpose of applying a discriminant analysis is to increase the isolation of faults. This may come at a cost of an decrease of robustness in the face of model uncertainty.

Fisher's linear discriminant technique is a form of the linear discriminant analysis (although the two are often confused). Fisher's technique does not make the assumption that linear discriminant analysis makes of normally distributed classes or equal class covariance. Fisher's technique is still a linear technique.

4.1.1 Derivation of Linear Discriminant Analysis

Derivation for Two Populations

Johnson & Wichern (1982: 462-471) gives an excellent discussion of the derivation of linear discriminant analysis for two populations. The following is a summary.

For a dataset $\mathbf{X}$ containing p random variables (i.e. $\mathbf{X}^T = [X_1, X_2, \ldots, X_p]$) and containing two classes labelled π_1 and π_2, we can define μ_1 and μ_2 (being the expected values an observation of each class). It is not necessary to assume that the two populations are multivariate normal.

We also make the assumption that each population's covariance matrix is the same:

$$\mathbf{\Sigma} = E\left(\mathbf{X} - \mu_i\right)\left(\mathbf{X} - \mu_i\right)^T, \qquad i = 1, 2 \tag{4.1}$$

This assumption is often violated in practise. However, the assumption still often leads to satisfactory results provided that the covariances of the original data are of the same order of magnitude.

The fundamental idea behind the technique is to transform the multivariate observations to univariate observations (y) by means of linear combinations (which are easily handled). We let μ_{1Y} and μ_{2Y} be the means of the Y's obtained from the $\mathbf{X}$ belonging to π_1 and π_2 respectively.

We are interested in the linear combination:

$$Y = \ell^T \mathbf{X}. \tag{4.2}$$

This means that:

$$\mu_{1Y} = \ell^T \mu_{\mathbf{1}} \tag{4.3}$$
$$\mu_{2Y} = \ell^T \mu_{\mathbf{2}} \tag{4.4}$$

while the variance for both classes is:

$$\sigma^2 = Var\left(\ell^T \mathbf{X}\right) = \ell^T Cov\left(\mathbf{X}\right)\ell = \ell^T \mathbf{\Sigma} \ell \tag{4.5}$$

The optimal weight matrix ℓ gives a direction which maximises the distance between the projected class means, while minimising the interclass variance (Jemwa & Aldrich, 2005). A diagrammatic representation of this is shown in figure 4.1. Here maximum separation of the two classes is achieved by maximising the difference between the means of the different classes while simultaneously minimising the class covariance matrix. This

best linear combination can be represented mathematically as follows:

$$\frac{\text{(Squared Distance between means of Y)}}{\text{(Variance of Y)}} = \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{\left(\ell^T\mu_1 - \ell^T\mu_2\right)^2}{\ell\boldsymbol{\Sigma}\ell} = \frac{\ell^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\ell}{\ell^T\boldsymbol{\Sigma}\ell} = \frac{\left(\ell^T\delta\right)^2}{\ell^T\boldsymbol{\Sigma}\ell} \tag{4.6}$$

This ratio is maximised by:

$$\ell = c\boldsymbol{\Sigma}^{-1}\delta = c\boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2), \qquad c \neq 0 \tag{4.7}$$

with $c = 1$ (for a linear combination)

$$Y = \ell^T\mathbf{X} = (\mu_1 - \mu_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{X}. \tag{4.8}$$

Fisher's linear combination coefficients ($\ell^T = [\ell_1, \ell_2, \ldots, \ell_p]$) are those that maximise equation 4.6.

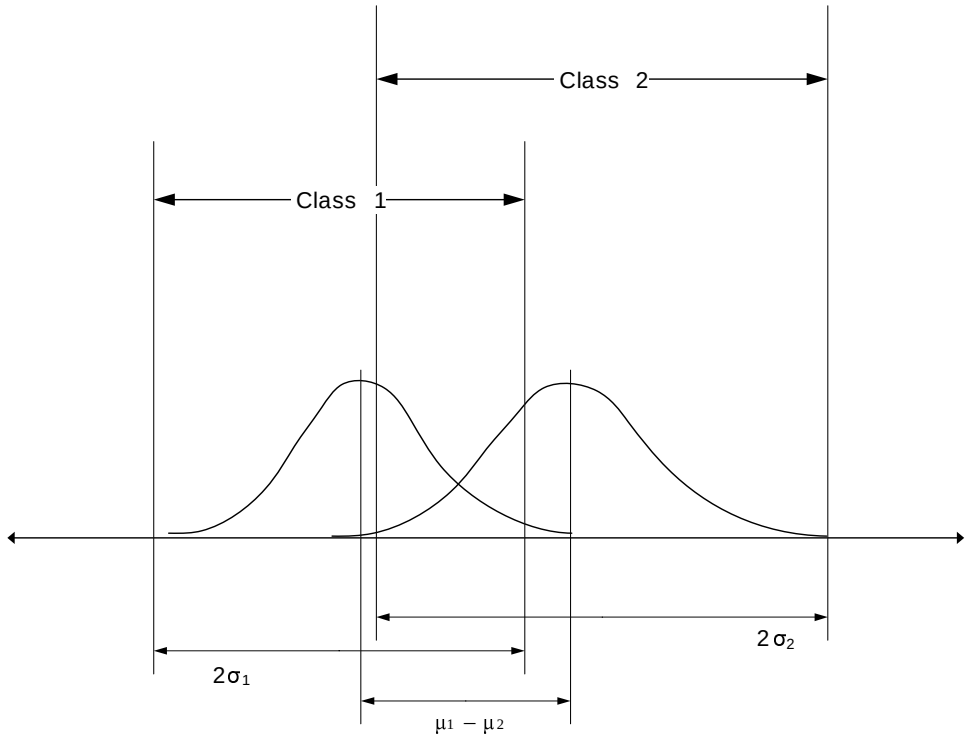

Figure 4.1: The Basic Idea of Linear Discriminant Analysis

Equation 4.8 can now be used for classification. Using x_0 as a new observation,

$$y_0 = (\mu_1 - \mu_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 \tag{4.9}$$

becomes the value of the discriminant function for this new observation. Additionally,

we let the midpoint between the two univariate means be:

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}(\ell^T\mu_1 + \ell^T\mu 2) = \frac{1}{2}(\mu_{1Y} + \mu 2Y)^T\mathbf{\Sigma}^{-1}(\mu_1 + \mu_2). \tag{4.10}$$

so:

$$E\left(Y_0|\pi_1\right) - m \geq 0 \tag{4.11}$$

and

$$E\left(Y_0|\pi_2\right) - m < 0 \tag{4.12}$$

This means that we can make a simple classification rule:

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_1 \text{ if } y_0 - m \geq 0 \tag{4.13}$$

and

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_2 \text{ if } y_0 - m < 0 \tag{4.14}$$

Usually the quantities μ_1, μ_2 and $\mathbf{\Sigma}$ are not known. The classifications rules in equations 4.11, 4.12, 4.13 and 4.14 cannot be implemented unless ℓ and m can be estimated from observations that have been classified (Johnson & Wichern, 1982: 465). For the two dimensional case, we will label them $\mathbf{X_1}$ and $\mathbf{X_2}$. The sample mean vectors ($\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$) and the covariance matrices ($\mathbf{S}_1$ and $\mathbf{S}_2$) can be determined as usual.

Since it is assumed that the parent populations have the same covariance matrix ($\mathbf{\Sigma}$), the sample covariance matrices ($\mathbf{S}_1$ and $\mathbf{S}_2$) are pooled to create a single unbiased estimate of $\mathbf{\Sigma}$:

$$\begin{aligned}\mathbf{S}_{pooled} &= \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \\ &= \frac{(n_1 - 1)\mathbf{S} + (n_2 - 1)\mathbf{S_2}}{(n_1 + n_2 - 2}\end{aligned} \tag{4.15}$$

The sample quantities $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$ ans $\mathbf{S}_{pooled}$ are substituted in for μ_1, μ_2 and $\mathbf{\Sigma}$ in equation 4.8 to obtain:

$$Y = \ell^T\mathbf{X} = (\mu_1 - \mu_2)^T\mathbf{\Sigma}^{-1}\mathbf{X}. \tag{4.16}$$

This means that the midpoint is then given by:

$$\hat{m} = \frac{1}{2}\left(\overline{y}_1 + \overline{y}_2\right) = \frac{1}{2}\left(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2\right)^T\mathbf{S}_{pooled}^{-1}\left(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2\right) \tag{4.17}$$

So the classification rule is now:

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_1 \text{ if}$$
$$y_0 - \hat{m} \geq 0 \tag{4.18}$$
$$\text{or allocate } \mathbf{x}_0 \text{ to } \pi_2 \text{ if}$$
$$y_0 - \hat{m} < 0. \tag{4.19}$$

An example of this classification being used is shown in figure 4.2.

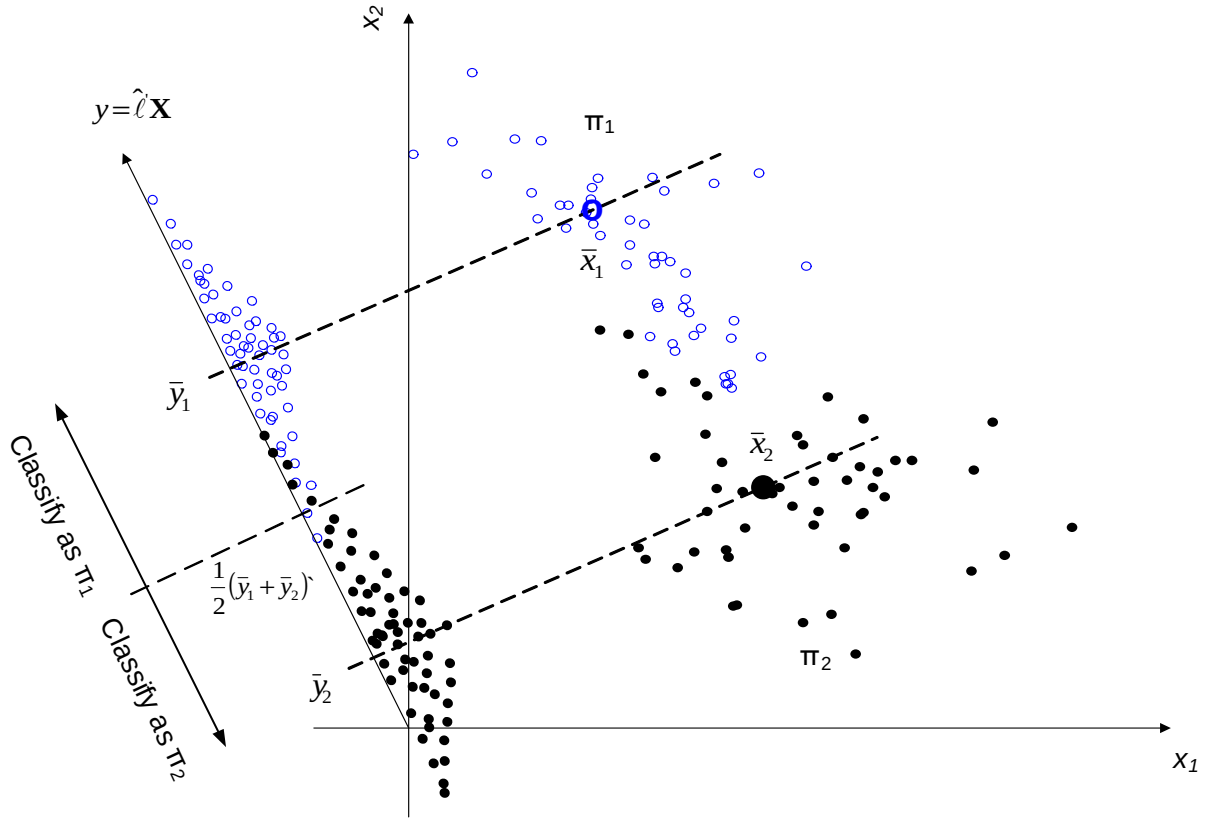

Figure 4.2: Representation of Fisher's Procedure with Two Groups

Derivation for Several Populations (Johnson & Wichern, 1982: 504-511)

The derivation for the multivartiable case is similar to the bivariate case (shown in the previous section). The primary difference being that the vector of coefficients ℓ maximises the ratio of the projected class means and the interclass variance. The problem them becomes an eigenvalue problem. The full derivation is presented in Johnson & Wichern (1982: 504-511).

$min(g-1, p)$ (with g being the number of class populations and p being the number of variables describing each population) eigenvalues are produced which correspond to the $min(g-1, p)$ discriminant functions. The most significant discriminant functions correspond to the largest eigenvalues. In this way, dimensional reduction is done in

the same step as discrimination. The number of dimensions chosen will usually depend on practical considerations (i.e 3 or fewer dimensions for easy visualisation), but the dimensionality can also be chosen in a similar way to the methods for PCA (shown in section 3.3.3).

4.1.2 Practical Considerations Regarding LDA

Classification can only be performed well if the classes are well separated. This can be tested before the analysis is performed by a MANOVA (testing for a significant level of difference between class mean vectors) (Johnson & Wichern, 1982: 519). If there is not a significant difference, the construction of classification rules will probably be futile.

There are other classifiers in addition to the simple linear classifier discussed here (Johnson & Wichern, 1982: 519-520). Another important classifier is the quadratic classifier (a more general form of the linear classifier). As the name suggests, QDA separates classes by means of a quadratic classifier. However, there are cases where a linear or (possibly even) a quadratic classifier would be inappropriate. An example is shown in figure 4.3. In this figure, strongly non-normal contours show how an assumption regarding the distribution of the data would result in poor classification.

As with principal component analysis, a (even linear) dataset will be incompletely described by using a subset of the discriminant functions. In some cases, this could provide insufficient information to perform discrimination adequately. The advantage of the linear combinations is that a combination of several variables is more likely to be nearly normal. This may aid discrimination.

As with PCA, the directions of the discriminants can be shown on a biplot. The contributions could be counted using the projections back on to the discriminant axes. If the LDA is performed on the PCA scores, the projection will need to be done again back to the principal component axes. The original contribution of process variables can be read off from there. In other words: an additional rotation is induced by the LDA. This means that the data needs to be projected twice to get the contribution of the variables to each faulty region.

An even more extreme example than that shown in figure 4.3 is a case where the means of data classes do not differ significantly. This would mean that the classes nearly lie on top of each other. To separate and classify the classes would require a technique that finds some other higher-order relation between the variables and uses that to form a classification rule.

Another practical consideration is the inclusion of variables that describe discrete (or categorical) states. These are frequently found in chemical processes (e.g. a controller mode variable). These will result in strongly non-normal distributions. It is unlikely that the mean of a class has any any meaning or that it will differ significantly from another

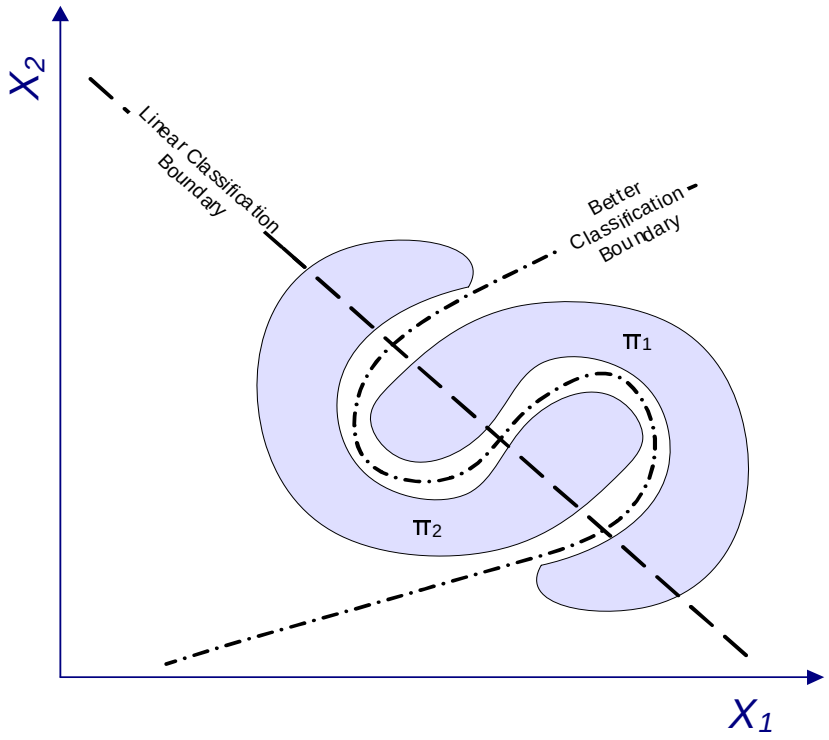

Figure 4.3: Non-normal Groups for which LDA is Inappropriate

class. Mixing continuous variables with discrete variables does not result in favourable conditions for LDA (Krzanowski (1977) quoted by Johnson & Wichern (1982: 519)).

4.2 Comparison of Feature Extraction and Feature Classification

Principal component analysis and kernel PCA are examples of feature extraction methods. They extract useful information from the data. In the case of PCA, this useful information is the linear relationship that explains the most variance in the input space. In the case of KPCA, it is the same relationship in the feature space.

From these features (in this case - models), we can characterise a normal feature set from normal operating data. A fault or process change will cause a change in the he features that we have extracted. Comparing the features of a new data sample to normal features allows us to detect the fault. Comparing the new features to the feature of known faults from prior experiences allows us to diagnose the fault. This is where the concept of classification begins.

Classification uses features extracted from the data to classify the data into classes. A classification method creates a rule that uses the features to decide into which class a data sample falls. In linear discriminant analysis, this rule was based on the distance to the means of the populations after some optimising of the ratio of inter- and intraclass

variances. The purpose of applying a discriminant analysis is to increase the isolation of faults. This may come at a cost of an decrease of robustness in the face of model uncertainty.

An example demonstrating the difference between classification and feature extraction is shown in figure 4.4. This example is similar to that of Franc & Hlaváč (June 24, 2004). Here two skewed Gausian mixtures are shown. There are close together (yet easily distinguishable in the input space) and have a similar relationship. PCA has found the direction of most variance. While this is a useful 1-dimensional model to describe the basic relation in the data, it can be seen that the resulting model projections in figure 4.5 does not allow for any classification between the groups. This will describe the basic $y = x$ relation of the data. In contrast, LDA has found a decision direction in 1-dimension. In figure 4.5, we can clearly see that LDA has found a clear classification model. While it did not find the useful relation in the data, it did find a way to classify the data into its two classes with good accuracy. An ideal method for fault classification would include the dimensional reduction modelling goodness of PCA or KPCA with the classification power of a discriminant method.

In summary, PCA and other feature extraction methods seek directions that are optimal for feature description and LDA and other discriminant techniques seek directions that are optimal for classification and discrimination (Cho, 2007).

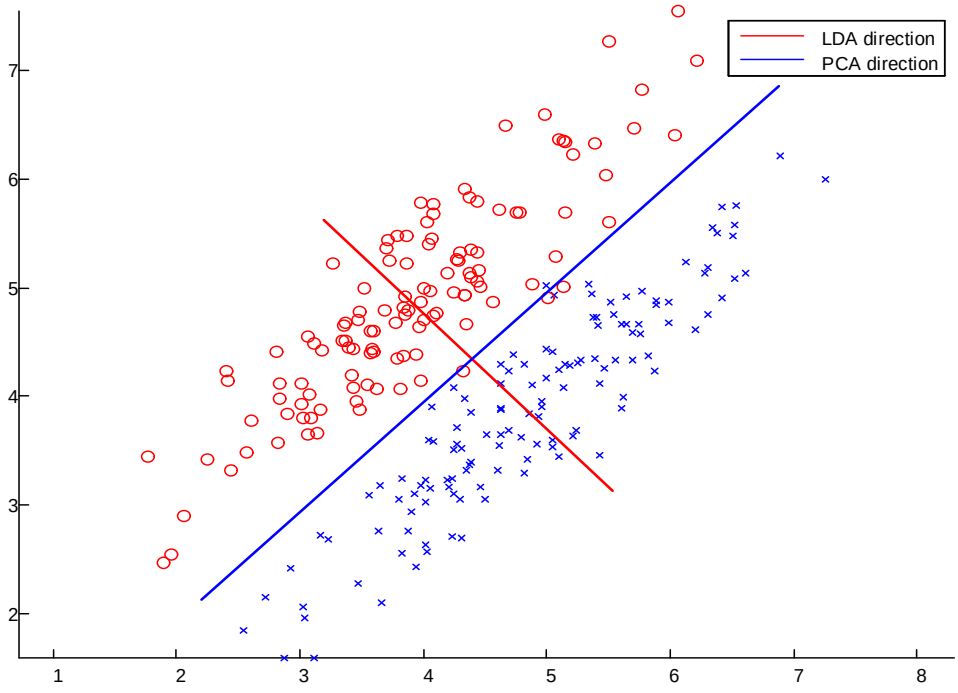

Figure 4.4: Comparison of the PCA and LDA Directions

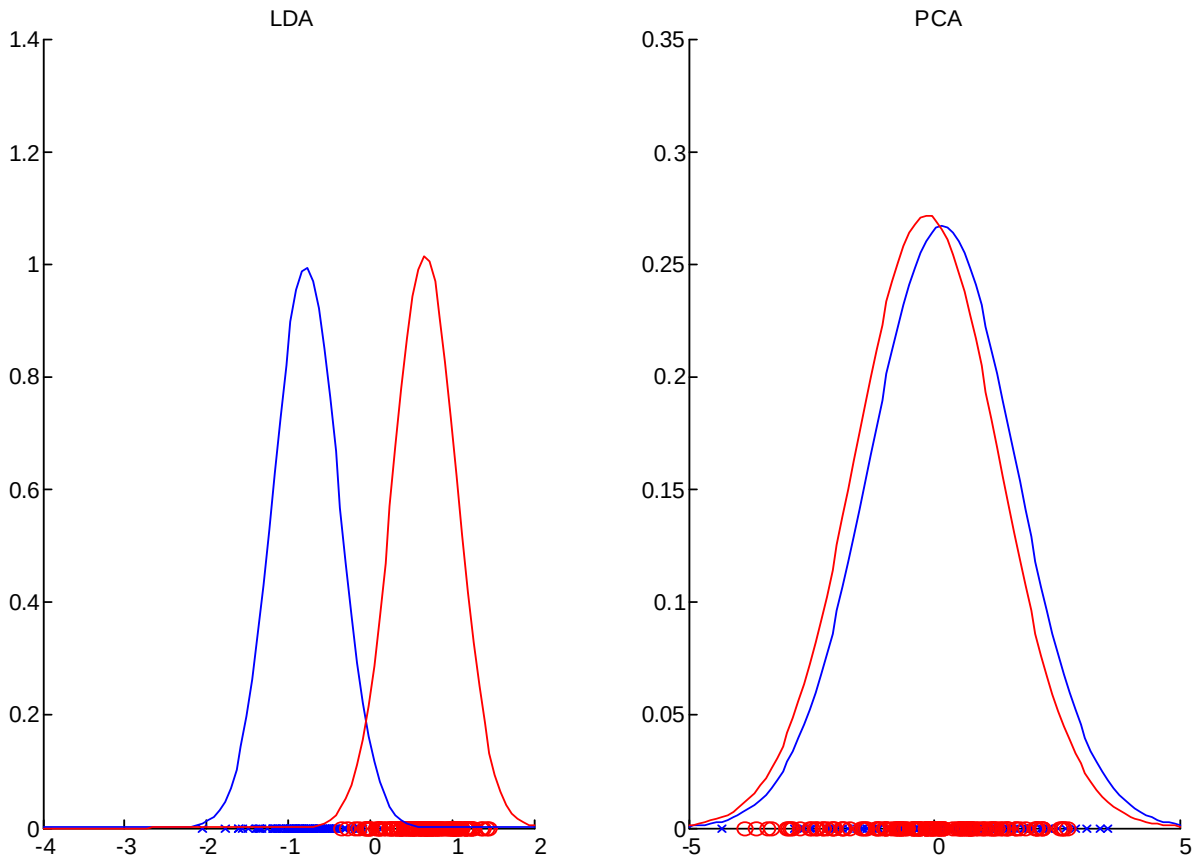

Figure 4.5: Comparison of the LDA and PCA Projections

4.3 Kernelised Discriminant Analysis

While linear discriminant analysis is a useful and powerful technique, it cannot represent nonlinear behaviour well. Kernelised discriminant analysis (KDA) proves to be more effective than LDA in various applications (Yang et al., 2004). The kernalised approach is important for the same reasons that it was important in KPCA. The linear discriminant analysis can then be performed on the reduced feature space to extract features for classification. KFDA seeks to solve the problem of linear discriminant analysis in the nonlinearly mapped feature space. This is analogous to KPCA performing linear PCA in the feature space. KDA is a far older method that KPCA (Hand, 1982: 11-15). KDA has the potential to increase fault isolation dramatically. However, it may also come at the cost of less general model for the data. This is because the model is over-fitted to the data. It is able to discriminant between the classes of trained data, but fails with unseen data.

Cho (2007) shows an example of Fisher discriminant analysis (could have equally being applied as linear discriminant analysis) being applied as a feature extraction and classification method. Yang et al. (2004) also covers the method in detail. Yang et al. (2004) reformulates the generalised discriminant analysis as a two step process namely:

1. Kernel principal component analysis

2. Linear discriminant analysis in feature space

This is equivalent to generalised discriminant analysis - just easier to understand.

Note that due to the lack of a mapping back to the input space, it is not directly possible to show what combination of variables are being used for discrimination.

4.3.1 Motivational Example for KDA

An example of the need for KDA (as opposed to LDA) is shown in figure 4.6. As in the KPCA example, we see that there are a maximum of two directions for LDA in this 2-dimensional input space. In contrast, KDA has found a discriminatory relationship for every point. Only the first three are shown here. The first LDA direction separates the cluster on the right from the two on the left. The second LDA direction will separate the upper left cluster from the lower left cluster - while splitting the right cluster. This means that it will be impossible to separate and discriminate between all three clusters at once. In contrast, the KDA's first two directions directions separate the 3 clusters. This means that with the same number of directions, it will probably be possible to discriminate between all 3 clusters. The third KDA direction is some relationship explaining the random noise in the upper left cluster.

Note that KDA is a similar but separate concept to support vector machines (SVM. An SVM classifier can be trained by (Franc & Hlaváč, June 24, 2004):

1. Applying a kernel projection.

2. Training a linear rule on these projected points.

3. Combining the kernel projection and the linear rule.

This creates a classifier where the boundary is nonlinear. It can also be used in a similar fashion to KDA presented here.

Hand (1982) covers kernel discriminant analysis in some detail.

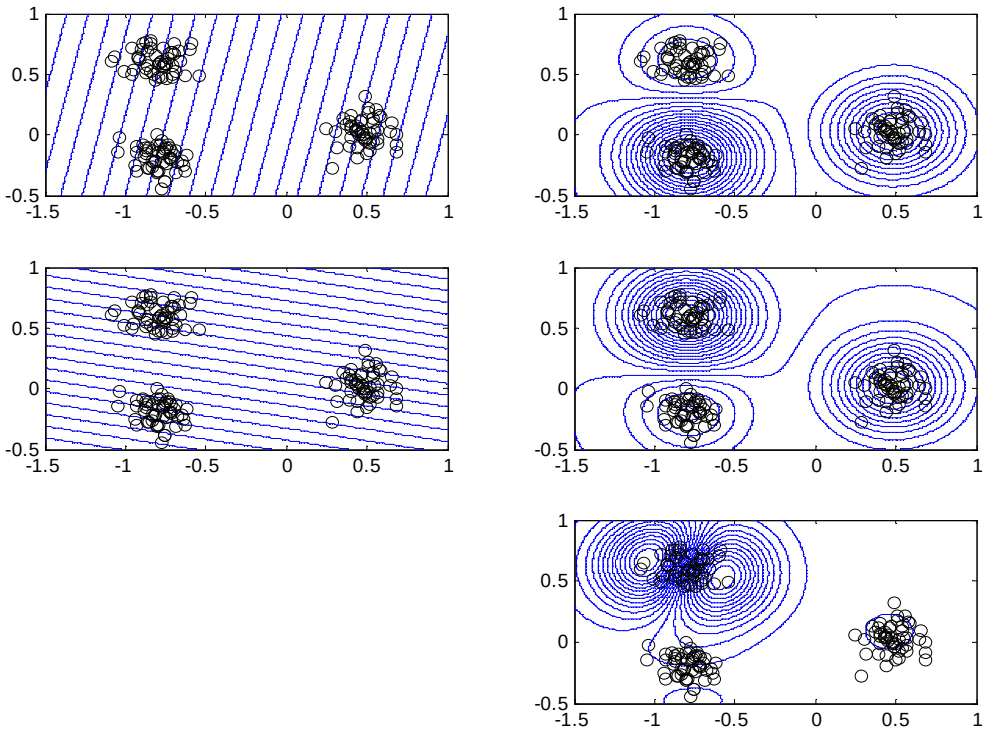

Figure 4.6: Motivational Example for KDA (left column: Linear LDA; right column: KDA)

4.4 Discriminant Analysis for Fault Detection and Diagnosis

The use of discriminant analysis for fault detection is far less direct than monitoring univariate control charts or the SPE and T^2 metrics from a PCA model. Here the difference between the feature extraction of PCA (or KPCA) and the classification of the discriminant analysis becomes apparent. The purpose here is to increase the isolation of the faults to aid diagnosis. The discriminant analysis provides a classification rule (to apply to features of the data) and not direct information about the features of the data. A T^2 or SPE statistic can be calculated easily (in the same manner for PCA). However, as the discriminant model is concerned with classification and not modelling the data, these statistics are not directly applicable to the detection of faults (defined as some change in the $features$ of the data).

The obvious way to make use of classification to detect faults is to create a normal operating class, together with data of known faults (each in their own class). A new data sample could then be checked to see if it should be classified into the normal operating region class or into one of the known fault classes. Novel fault detection can be performed by checking to see if the classification metric (used in the classification rule) is significantly

different from that of the metric of the normal operating region.

Checking which class the new data sample falls into would be a method of diagnosing the fault. Obviously the classes used to train the discriminant function would have to be representative of the types of fault likely to be encountered.

A discriminant analysis is clearly more useful in fault diagnosis than in fault detection. A fault detection and diagnosis system could use a feature extraction method (here PCA and KPCA are proposed) for fault detection, with a classification method (LDA and KDA are proposed) to aid diagnosis if a fault was detected.

CHAPTER 5

Experimental Setup

A well instrumented distillation column is the ideal test bed for a fault detection and diagnosis strategy. A distillation column combines both fast and slow dynamics at various levels of linearity. Due to the relatively small size of the column, it is also possible to have variables that are saturated. There is a large amount of data containing many variables which are strongly correlated.

Also due to the complexity of the process, there is a a strong probability of incidental failures or faults. This means that it will be possible to find representative faults without necessarily having to create them artificially.

5.1 Equipment

The column to be used is a ten plate glass distillation column. A process flow diagram for the column is shown in figure 5.1. The column is intended to run close to atmospheric pressure. The pressure in the column does rise (especially during startup where air must be expelled from the column) due to the small opening to the atmosphere in the reflux drum. The column separates a mixture of ethanol and water. The column is initially charged with an approximately 50% (mass basis) mixture of liquid. The top and bottom streams return back to a large feedtank. The size of the feed tank ensures that any changes to the composition in the top and bottom product streams are damped.

The boiler is fed with saturated steam of between 0 and 100 kPa from a supply of 1 MPa (gauge) saturated steam generated by means of an electric steam boiler. The throttling is done by a pneumatic control valve. The steam boilers are occasionally problematic.

The condenser is made of glass coils fed by cool municipal water. The level in the reflux drum is measured by means of differential pressure measurement of the liquid

head. This calculation depends on the pressure within the column and the density of liquid in the drum. Often when the pressure in the column becomes excessively high, the calculated level will be grossly incorrect as the column pressure and liquid head values are of different orders of magnitude.

The reflux is sent to plate 1. Airlocks frequently occur in the mass flow meters due to the small volume of the reflux stream as compared to the piping and control valves. Airlocks also occur in the top product stream.

The column is not insulated and consists of 10 plates (numbered from the top). The column diameter is approximately 15 cm.

The feed can be introduced on plates 3 and/or 7. For all cases presented here, the feed on plate 7 will be used. The feed is generally subcooled at a temperature of between 20 and 50°C. This depends on atmospheric conditions as well as on the volume of surplus feed in the feed tank (affecting the residence time for heat loss to the environment). The bottoms stream can be further cooled with the bottoms cooler, but this is seldom necessary. The pump, piping and control valve are large for the desired feed flowrate. This means that the control valve is typically only open less than 5% in conjunction with a almost closed hand valve. This small hand valve can then opened to allow more feed to flow. When this happens, the feed flow dynamics change dramatically - often leading to control instability. This hand valve often begins to clog during operation - eventually leading to complete flow failure. When this valve is changed, flow may be many times the desired value for short periods.

5.1.1 Instruments and Controls

Pneumatic control valves are placed on all the streams entering and leaving the column. The instrument labels are shown in figure 5.2. The following readings are available:

- Temperature from a Foundation field bus instrument on each plate.
- Temperature of the feeds, top and bottom products.
- Temperature of water entering and leaving the condenser.
- Steam pressure in the boiler after throttling.
- Levels in the boiler and in the reflux drum (calculated from heads pressures).
- Mass flowrates of the liquid returning to the boiler, bottoms product, reflux, cooling water, feed and top product streams.
- Pressure within the column.

The column is part of a DeltaV distributed control system. Data are stored during runs in a continuous historian.

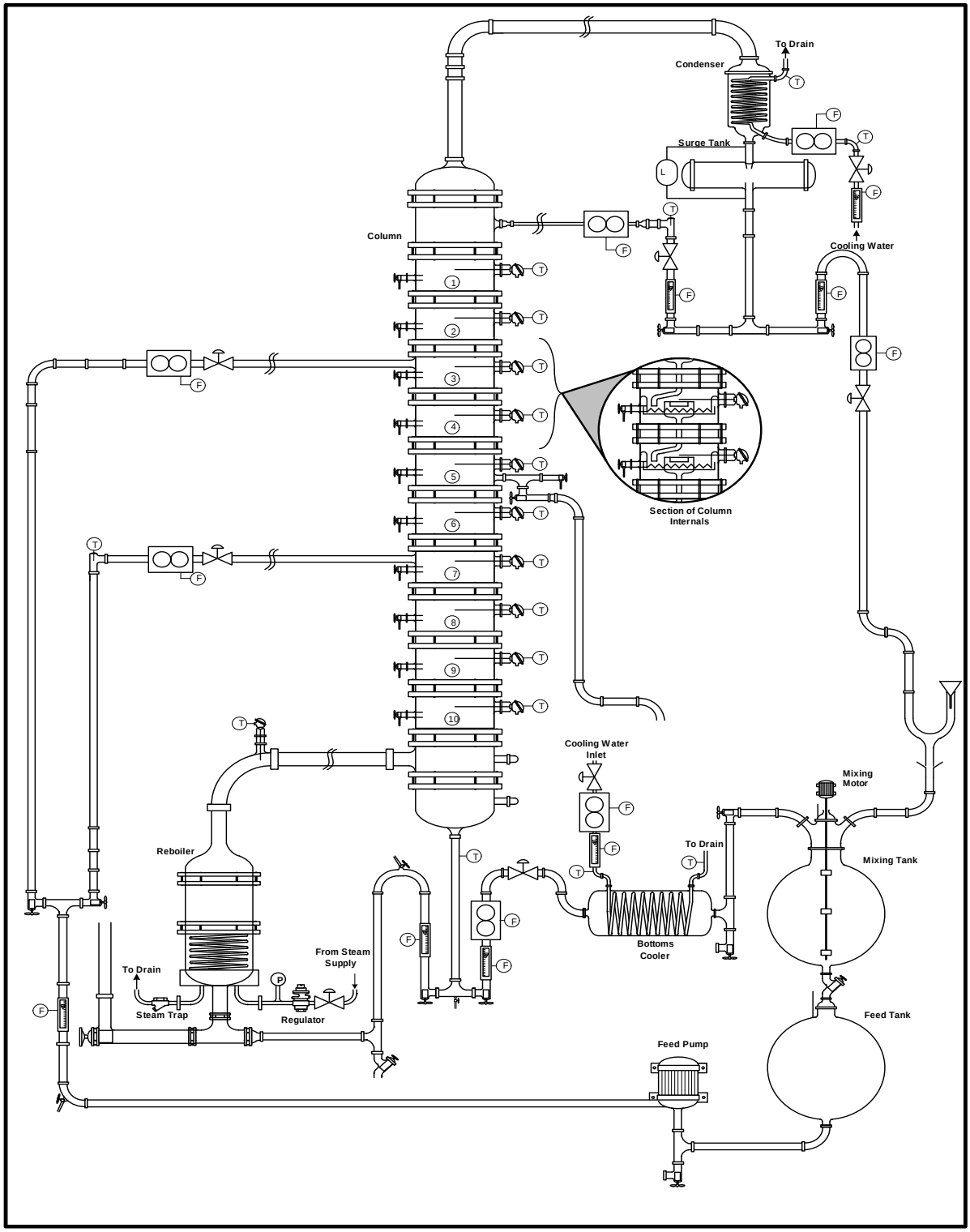

Figure 5.1: Detail Diagram of the Glass Distillation Column

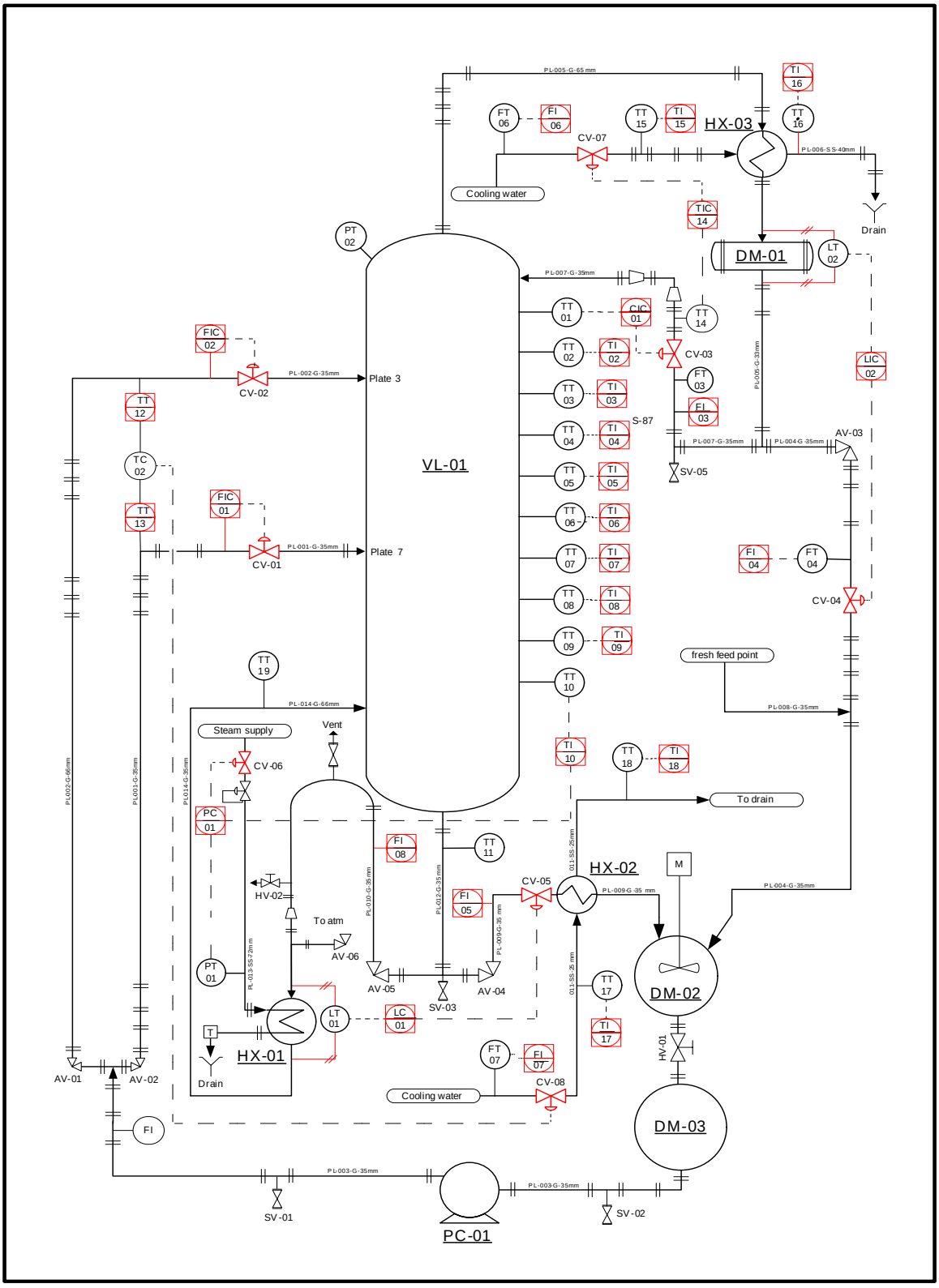

Figure 5.2: Piping and Instrumentation Diagram of the Glass Distillation Column

5.1.2 Control System

During operation, the distillation column is under computer control. A series of PID loops are used. The control loops are all tuned to give fast and stable responses. The control parameters are generally very well suited to the normal operating regimes.

A description of the control philosophy (as shown in figure 5.2) follows:

- The level in the boiler is maintained by LC-01 which manipulates the bottoms product flowrate by means of CV-05. An interlock exists that will stop the flow of steam to the boiler in the case of insufficient boiler level.
- The feed flowrate is maintained by FC-01 manipulating the flow directly by means of CV-01.
- The position of CV-07 on the cooling water flow is manually set at a value that will allow for sufficient cooling. An interlock exists that will cut the flow of steam to the boiler if this water flow is not measured to be above a certain value. The value of water flowrate usually saturates the instrument FT-06.
- An interlock on the pressure inside the column will cut the flow of steam to the boiler if the pressure rises above some safe value.
- The level in the reflux drum is maintained by LC-02 which manipulates the flow of top product (FT-04) by means of CV-04.
- TIC-01 manipulates the reflux flow (FT-03) by means of CV-03 to keep the top plate temperature (TT-01) on setpoint.
- PC-01 manipulates CV-06 to keep the steam pressure to the boiler (PT-01) on setpoint.

5.2 Experimental Design

Generally, the efficacy of a fault diagnosis system reflects the amount of process knowledge inherent in the representation method used to support fault identification (Frankowiak et al., 2005). For accurate fault detection it is important that the plant has been richly excited over the period that the training data has been captured (Goulding et al., 2000). As much of the normal operating region must be captured. It is important that the training data are also evenly distributed in order to give proper weight to the whole region Macgregor & Kourti (1995).

5.2.1 Generation of Training Data

The Normal Operating Training Dataset

The column is usually operated within a certain region where the column is known to perform adequately. The controllers are also tuned for this region.

The typical values are shown in table 5.1. The values shown are ones that are most indicative of the quality of the process and are usually watched closely during operation. Many other variables are monitored continuously (section 5.1). It would be futile to give an exhaustive list of all the variables.

Table 5.1: Typical Operating Values

Variable	Range	Comments
Steam Pressure	35-45 kPa	This usually depends on environmental conditions
Reflux Level	50-80 %	
Feed flowrate	35 kg/hr	The column is seldomly run at other feed flowrates
Cooling Water Flowrate	> 100 kg/hr	Excess cooling capacity very small temperature increase
Top Plate Temperature	75.5-78 °C	
Reflux Flowrate	5-10 kg/hr	
Top Product Flowrate	10-20 kg/hr	
Bottom Flowrate	15-40 kg/hr	High Variance
Liquid into the Boiler	0-30 kg/hr	High Variance
Pressure	0 kPa	Constant during normal operation Spikes during faults and startup
Boiler Level	50-70 %	

To create a dataset encompassing a normal operating region, the column was run so that the full range of these variables could be captured. As each variable was changed, the dynamics of the change as well as the operating point was captured in the data. The experiments shown in table 5.2 were done to create a normal operating region. These variables were selected as they are critical to the operation of the entire column. Even a small change in one of these variables will affect nearly all of the other variables in the column. Additional interaction is brought in due to the action of the controllers. An example: changing the temperature of the top plate involves controller FC-03 changing the reflux flowrate. This will in turn affect the level in the reflux that LC-02 is maintaining. This will in turn affect the flowrate of top product. Eventually, the increase in liquid flowing down the column (affecting the plate temperatures) will increase the level in the boiler.

The effects of changing the setpoint of the boiler level on the normal operating region was not investigated. This is because the level of the boiler is not normally manipulated

during the running of the column. It is regulated to ensure that the coils are not damaged by excessive heat due to insufficient level. Additionally the effect of changed reflux drum level setpoint was not investigated. This because it merely acts as a buffering capacity and the actual level does not matter much - only that there is sufficient level for reflux and top product flow during dynamic operation. Also the effect of the flowrate of cooling water was not investigated. The cooling water flowrate is chosen to be very high as a safety consideration. The rise in water temperature through the condenser is low. The cooling water could probably be reduced to less than 30 % of its usual value before the effect on the condensation rate (and column pressure) would be significant. The feed flowrate was also not changed from its normal operating point of 35 kg/hr. This is done intentionally to compare fault finding for a variable that has a small training region as compared to some of the other flows (e.g. the top product) which have large ranges. The feed composition was maintained at approximately 50 % ethanol (mass basis). Due to the recirculation, this would be difficult to maintain at another set value. There is no measurement of composition, so even recording a dynamic change would be very difficult. As usual, all feed was onto plate 7 and the bottoms cooler was not in operation.

Table 5.2: Experiments for the Normal Operating Region

Experiment	Feed	Steam Pressure	Top Plate Temperature
1	35 kg/hr	35 kPa	78 °C
2	↓	↓	76.5 °C
3			75 °C
4		40 kPa	78 °C
5		↓	76.5 °C
6			75 °C
7		45 kPa	78 °C
8		↓	76.5 °C
9			75 °C

The steady state values as well as the dynamics experienced while moving between the steady state values should provide an adequate representation of the non-faulty operating region. Enough time was given between setpoint changes for the system to reach the new steady state.

Air Failure Datasets

The purpose of this fault is to investigate the failure of a major utility. The air is required to hold all the valves open (all valves are air-to-open). Clearly the failure of the air supply will affect nearly all of the variables dramatically. This is clearly a major fault that should easily be detected by any method. This is an example of an abrupt fault (see section 1.1). Note that there is no direct measurement of the air pressure that is recorded. This means

that an air supply fault must be diagnosed by interpreting the effect the fault has on the measured variables. This fault is an example of a multiplicative fault.

To generate the data for the three training faults sets, the column was run at steady state. It was ensured that the operation was entirely fault free. The air supply was then cut completely. Once the steam pressure in the boiler had dropped to zero (due to the valve closing due to lack of air), the air supply was turned back on. The data was then recorded until the controllers brought the column back to normal operation. This was repeated three times from different fault-free operating points.

The fault testing data was generated in a similar way.

Steam Supply Failure Datasets

The electric boilers are often unreliable. The boilers will trip and the decrease in steam pressure is only noticed some time later (often once the column has lost much heat). It takes some time to return the column to normal operation. It would be useful to detect this fault early.

The fault training data was generated by slowly cutting off the steam supply until the steam pressure controller responds by opening the steam control valve all the way. The steam supply was then quickly resumed to simulate the boilers returning to normal operation. This was repeated twice from different operating points.

The fault testing data was taken from the data historian. The occasion used was when non-expert operators were using the column. The boilers had tripped and significant time passed before the fault was corrected. This fault had occurred approximately a year before the normal operating region experiments had been performed. The column was also under largely manual control. The passage of time after the testing fault occurred and the inexperienced operators will be useful to check the robustness of the fault detection and diagnosis techniques.

Feed Fault Datasets

The feed fault is another commonly experienced fault. As explained before in section 5.1, the feed control valve operates at low valve openings (less than 10 % - due to mis-sizing). A hand valve before the control valve is used to throttle the flow further. This valve is also nearly closed during normal operation. After the column has been operating for a while, the feed flow begins to reduce. To respond, the feed controller opens the valve (eventually to fully open). To correct the fault, the operator opens the hand valve a fraction. This causes a huge feed flow for a few seconds before the feed control valve can close again. The feed flow fault can be regarded as an incipient fault (see section 1.1).

To generate the two fed fault training data sets, the hand valve was slowly closed fractionally during normal operations. Once the feed control valve had opened fully, the

hand valve was returned to the normal position. The data was recorded until the feed flow rate had been returned to its normal value. This was repeated from a different operating point.

The fault testing data was taken from the data historian during a true incidental feed flow fault.

Top Product Flow Fault Datasets

Due to the large industrial control valve being used on a pilot scale distillation column, the streams with small flows (particularly the reflux and top product streams) frequently experience vapour locks. This may manifest as a complete flow failure, or as hysteresis on the flows. Usually, a 20 - 50 % valve opening provides enough flow for normal operation. Occasionally, due to vapour in the lines, valve openings of 80 % or more are required. This is shown in figure 5.3 on the right hand side plot. There appear to be two states, one where a small valve opening gives high flow and another which provides only a little flow. This hysteresis (changing between the two states often) is seldom noticed while monitoring univariate plots - even by experienced operators. This flow fault causes the top product flow to oscillate. As the level in the reflux drum is controlled by the top product flow, this level also oscillates severely. An example of this is shown in figure 5.4. The effects of this fault on the other variables (particularly the plate temperatures) is far more subtle. This should be contrasted to the air supply failure fault. The top product flow fault can be regarded as an intermittent fault (see section 1.1).

To generate the fault training data, the control valve positioning linkage was manually disconnected and manipulated to simulate severe hysteresis. The resulting data is shown on the right hand side of the figure 5.3. It is clear that this is a rather poor simulation. The testing fault data covered a significant amount of time. It is possible that the training fault did not have enough time to affect the rest of the column. It will be useful to see if this data is useful to detect real faults. This experiment was repeated from a different fault-free operating point.

The testing data is taken from a true incidental fault during real operation. The transition to and the return to normal operating conditions was not recorded.

Novel Fault Sets

Five different data sets are used as novel fault sets. Novel means that the algorithm has not trained with the data for fault regions (e.g. PCA and KPCA) or for classification (LDA and KDA). The novel data is used to test the ability of the techniques to detect unseen faults and to properly diagnose the cause of the fault.

The faults sets are:

1. *Water rich operation* - Instead of the usual charge of an approximately 50 % mix of

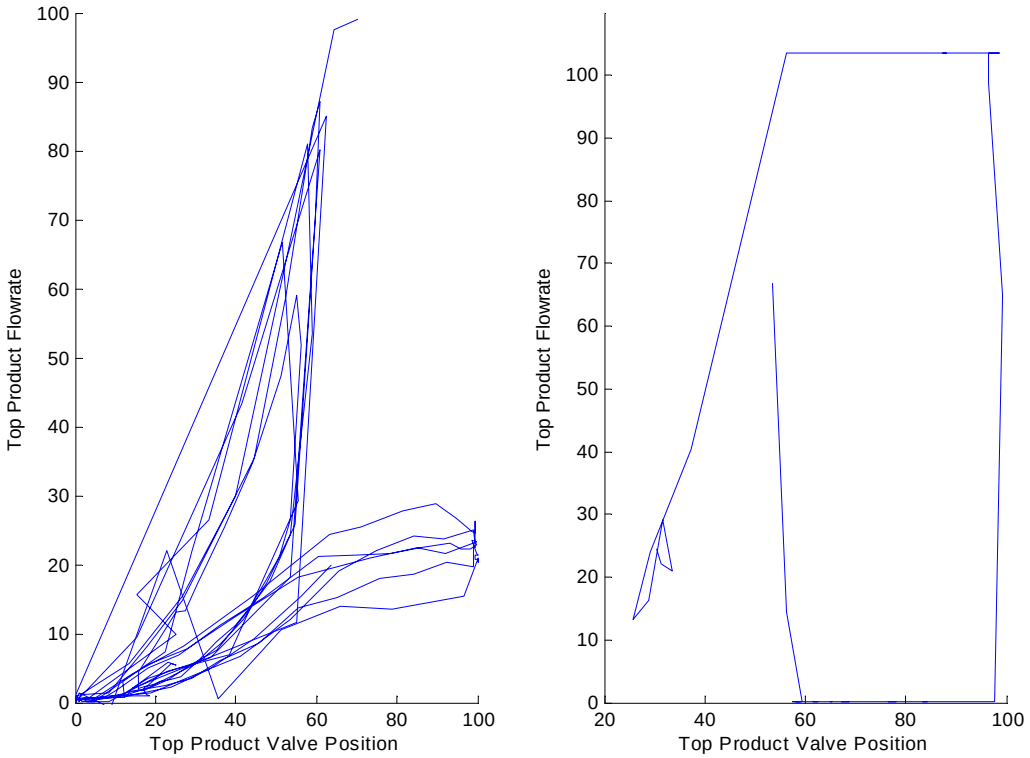

Figure 5.3: Top Flow Fault Data (Real on left hand side and artificial on the right)

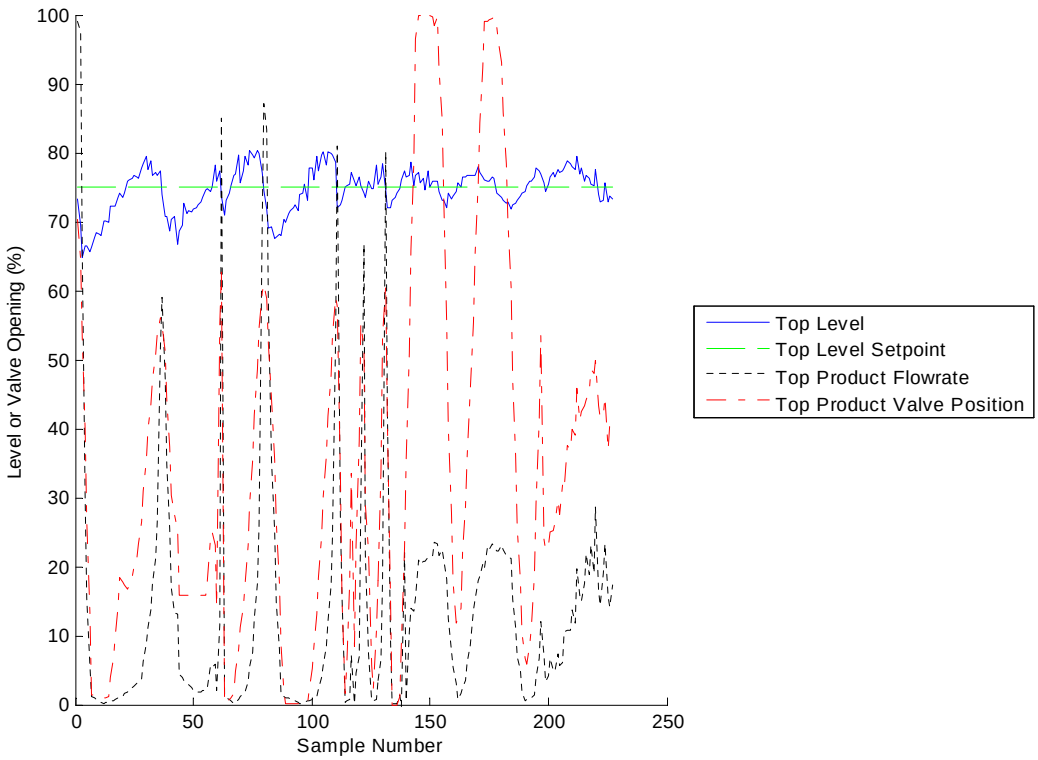

Figure 5.4: Explanation of the Top Flow Fault

ethanol and water, the column was completely drained and rinsed. The column was then filled with pure distilled water. The column plate temperatures were much higher than usual and more steam was required. The mass flow measurements were probably incorrect due to the dramatic change in density. The small amount of remaining alcohol collected at the top of the column. Some of the flow controllers were occasionaly unstable.

2. *Feed flooding operation* - The column was moved from normal operation to continuously running at a feed setpoint of 100 kg/hr (nearly triple the usual feed).

3. *Low ambient temperature* - The column was operated on a cold night after it had rained. The bay door separating the column from outside (usually closed) was opened fully. The data from this run was initially intended for use as part of the normal operating region, however, preliminary analysis showed it to be dramatically different from the other three data sets used for the normal operating region.

4. *Measurement fault* - The data for the temperature of plate 3 were changed to a value close to room temperature. This is intended to simulate a failed instrument. This is an example of an additive fault.

5. *Operation in a different regime* - The column was operated at its maximum rated boiler pressure of 100 kPa. The feed was set at 100 kg/hr (like the feed flooding operation). This is intended to provide data that, while at fault-free and at steady state, is completely different to the nominated steady states in section 5.2.1.

5.2.2 Selection of Variables for Analysis

Some details regarding the selection of variable were given in section 2.5.1. In this pilot process, there are few final quality measurements (for example top product composition). Here other process variables can be used as an indicator of the process quality. This is a problem common to many chemical processes and is handled well by the methods presented here.

It is not necessary to include all the measured variables in an analysis. Variable selection is an important issue and may be an evolving process. Several models may be made and experimented with. The expert knowledge of experience operators may be useful when selecting variables. Careful variable selection will enhance the robustness of the model (Kourti et al., 1996). On a real process, the variables to be used are often chosen by means of a structural or functional breakdown. An example of a functional breakdown being used to choose variables as well as to aid fault detection and diagnosis is shown in Groenewald et al. (2006).

The variables that will not be used are:

- *The cooling water flow-rate:* As discussed before.
- *Variables related to systems not in use:* For example: bottoms cooler variables. These variables should clearly not be a part of any plant model. Noise on these 'off' measurements may trigger false fault alarm.
- *Setpoints and parameters of controllers:* The fault detection and diagnosis will be focused on process faults as opposed to control faults. Also the position of setpoints may not necessarily have an impact on the process (for example if a setpoint is entered incorrectly and corrected before the controller changes the process). Setpoints also tend to be at discrete values. This may interfere with some of the analyses (particularly the LDA).

Valve positions of controlled valves will be included. The values from the output of the PID control loops will be used instead of the actual positions as the valves are reliable.

There is undoubtedly redundancy and correlation in the multiple temperature and flow measurements. This will be useful to demonstrate the dimension reducing modelling techniques. It will also demonstrate how most variables (excluding variables that are obviously problematic or contain less meaning) can be added without much knowledge of the process. A table containing the variables that will be used is shown in table 5.3.

5.2.3 Offline or Online Analysis

A decision has to be made whether the analyses are to be performed online or offline. As discussed in section 1.3, it is desirable to have a fault detection and diagnosis system running online. As confirmed in the results section, some of the models for the extraction of features as well as the classification models may be slow to train. This is not a serious hindrance to online analysis, the models only need to be trained once to set a basis line for the normal operating and fault regions. The model can be updated periodically.

Once the models are trained, the calculations for the PCA (or KPCA) scores projection, contributions (if applicable) and statistics can easily be performed online as data samples are read from the process.

The analyses presented in the following section have all been performed offline in a batch-wise manner in a Matlab environment. This is because running the calculations online would add no value to the way the results are presented here. The results are focused on exploring the detection and diagnosis of faults, rather than taking appropriate action to correct them.

Reference	Variable Name	Description
1	TT-01	Temperature on plate 1
2	TT-02	Temperature on plate 2
3	TT-03	Temperature on plate 3
4	TT-04	Temperature on plate 4
5	TT-05	Temperature on plate 5
6	TT-06	Temperature on plate 6
7	TT-07	Temperature on plate 7
8	TT-08	Temperature on plate 8
9	TT-09	Temperature on plate 9
10	TT-10	Temperature on plate 10
11	TT-11	Temperature of bottoms product stream
12	TT-13	Temperature of feed stream
13	TT-14	Temperature of reflux stream
14	TT-19	Temperature of the vapour entering the column from the boiler
15	FL-01	Mass flowrate of feed to plate 7
16	FL-03	Mass flowrate of reflux stream
17	FL-04	Mass flowrate of top product stream
18	FL-05	Mass flowrate of bottoms product returning to the feed drum
19	FL-08	Mass flowrate of bottoms product returning to the boiler
20	PI-01	Steam pressure within the boiler
21	PI-02	Column internal pressure
22	LI-01	Level with the boiler
23	LI-02	Level in the reflux drum
24	CV-01 VP	Valve position of the feed control valve
25	CV-03 VP	Valve position of the reflux control valve
26	CV-04 VP	Valve position of the top product control valve
27	CV-05 VP	Valve position of the bottoms product control valve
28	CV-06 VP	Valve position of the steam control valve

Table 5.3: Variables Selected for the Analysis

CHAPTER 6

Results

PCA and KPCA are applied to the data obtained from the pilot distillation column. Additionally, LDA and KDA were assessed for their ability to detect, isolate and diagnose faults. A detailed discussion follows:

6.1 The Normal Operating Region

As discussed in section 5.2.1, data are required that shows the full extent of the normal operating region. This should include all likely steady state values as well as the dynamics experienced while moving between these states.

The data used consists of three separate runs on different days. Each of these runs contained some of the experiments listed in table 5.2. Each of the experiments listed were performed at least twice. The runs were done on separate days to ensure that the data represent general operation. Also, the validity of the data can be ensured by checking if at least some section of each of the three runs overlap (corresponding to the repeated experiments).

The data were initially extracted from the continuous historian at 1 second intervals. These data were then averaged in to subgroups of 10 second intervals. This reduces the effect of noise in addition to reducing the computational burden. There were no missing data or outliers due to measurement error. Had there been any such problems, the data would have been manually reconciled. The normal operating region consists of approximately 2000 data samples.

The combined data from the experiments to obtain the normal operating region are shown in figure 6.1. For an explanation of the variables, please see table 5.3. It is clear that some variables have tight distributions (e.g. variable 2: the temperature on plate 2), while others have large variances (e.g. variable 19 - mass flowrate of liquid returning

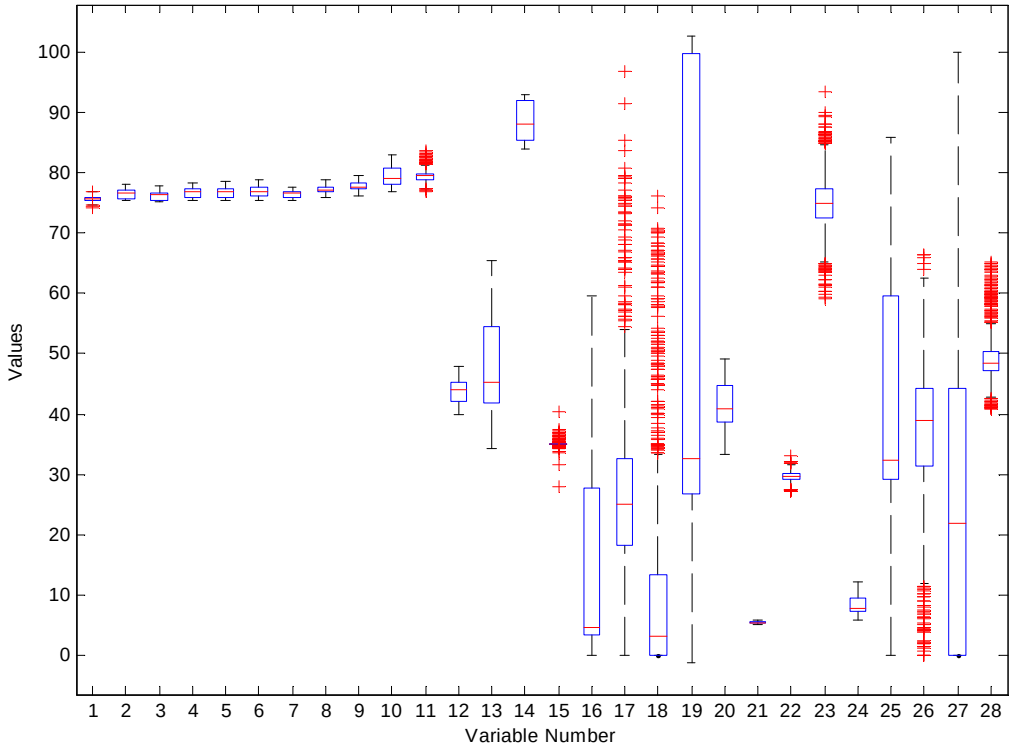

Figure 6.1: Boxplot of the Normal Operating Region Data

to the boiler). We can also see the presence of outliers (e.g. variable 18 - mass flowrate of bottoms product).

This section discusses the fault detection and diagnosis results for techniques that make use of models that attempt to model the features of the data. The modelled features of the data are then monitored for changes to detect faults. These metrics are compared to the metrics of known faults and the normal operating region to diagnose the fault.

The data are certainly not strongly linear. With 28 variables, it would be a lengthy process to investigate the relationship between all of them. In figure 6.2, we see a scatter plot matrix for some variables of interest (variables numbers 1, 10, 15, 20, 22 and 24). Even in the normal operating region, we can see that the data are not purely linear or normal. In some of the variables, it can be seen that some subgroups are linear and/or normal. These may correspond to individual experiments (shown in table 5.2). Also in the histograms (shown on the matrix diagonal), it can be seen that some variables are probably made of several groups of normal data (again possibly corresponding to the various steady states).

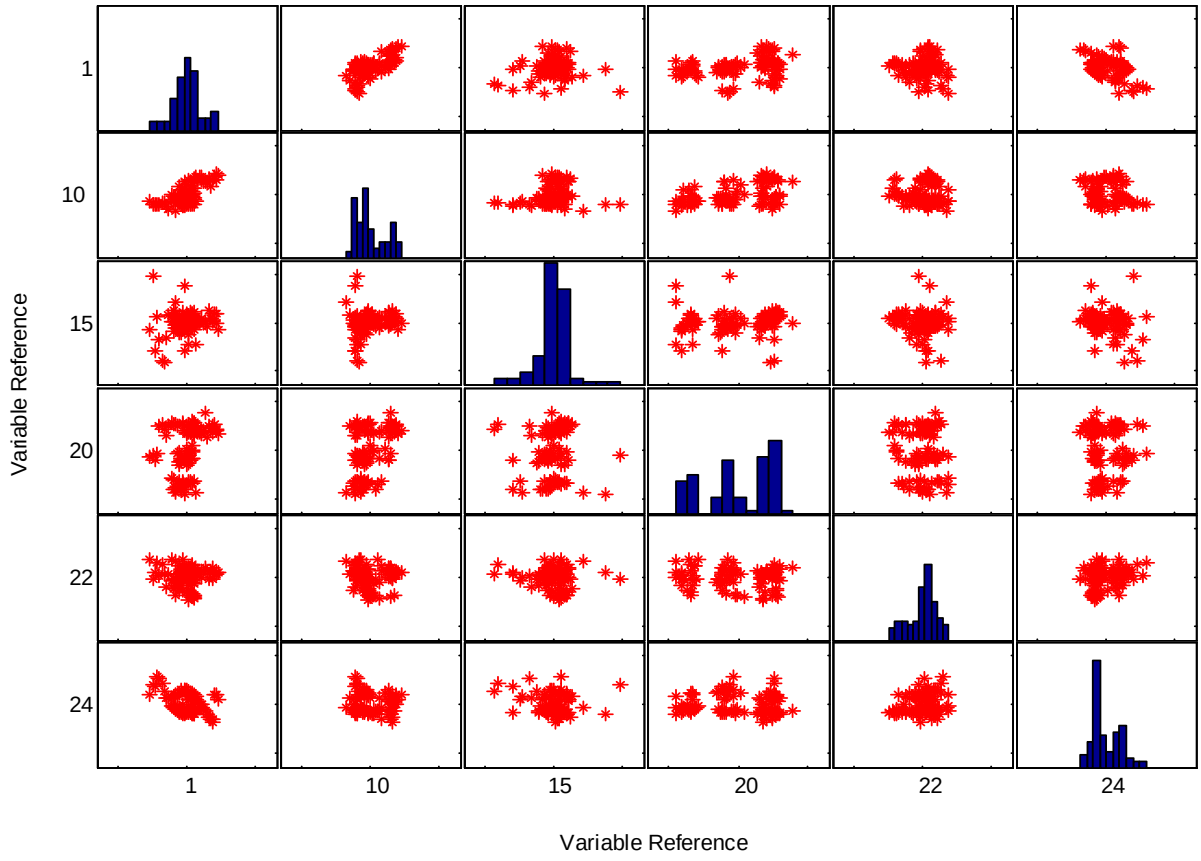

Figure 6.2: Scatterplot Matrix of some of the Normal Operating Region Data

6.2 Fault Detection and Diagnosis using Feature Extraction Methods

The results of the fault detection and diagnosis procedures implemented by means of principal component analysis and kernelised principal component analysis (using the procedures discussed in section 3.3.7 and 3.4.8) are discussed here.

6.2.1 Fault Detection and Diagnosis using Principal Component Analysis

Training of the PCA Model

The PCA model was trained using the normal operating data. The model took less than 10 seconds to train on a well equipped desktop computer. The Scree plot for the variance explained by the 28 principal components (corresponding to the 28 variables selected for the analysis) is shown in figure 6.3. About 42 % of the total variance is explained by the first principal component and about 15 % is explained by the second. The first two principal components together explain about 57 % of the total variance.

The dimensionality of the model must now be decided on. The cumulative percentage

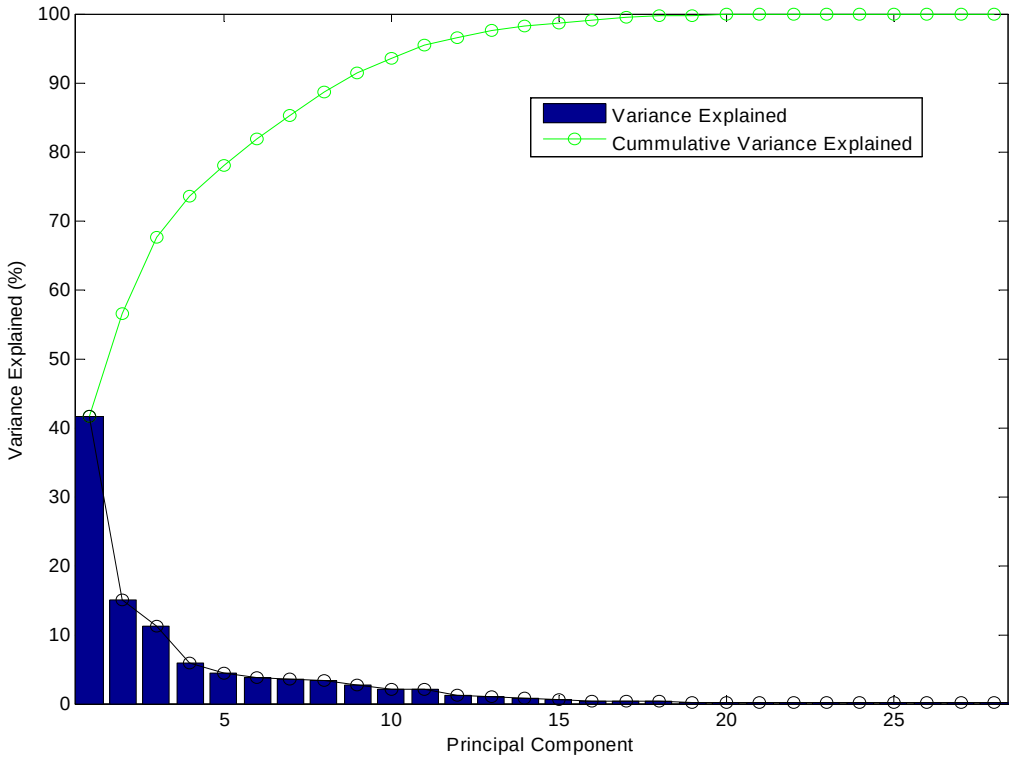

Figure 6.3: PCA: Variance Explained by the Principal Components

of variance, size of variance, broken stick and scree plot methods suggest the use of 3, 7, 14, 15 principal components (respectively). The size of variance, broken stick and scree plot methods do not work well here due to the large number of variables. This is because the less significant PCs (principal components) contribute so little to the analysis that it makes the average amount of variance explained very small. Note that the cumulative method suggests that a far lower number of variables is required. The number of variables used is purposefully large to demonstrate the dimensional reduction capabilities of the technique. Also, it shows how most variables (excluding the ones that are obviously problematic or contain less meaning) can be included without much knowledge about the process. All the variables that could possibly have any meaning were included (as discussed in section 5.2.2). These include many temperatures which clearly move together. For this reason, the suggested dimensionality given by these techniques is ignored. A model dimensionality of 2 is used. This will demonstrate that a severely reduced model can still provide metrics that can be successfully monitored for faults. The chosen dimensionality has the added advantage that it will be easy to represent the results in a simple biplot.

The contribution of each variable to the principal components is shown in figure 6.4.

The vector for each variable shows the relative contribution to the first and second principal component (shown on the x and y axis respectively). For example, variable 1

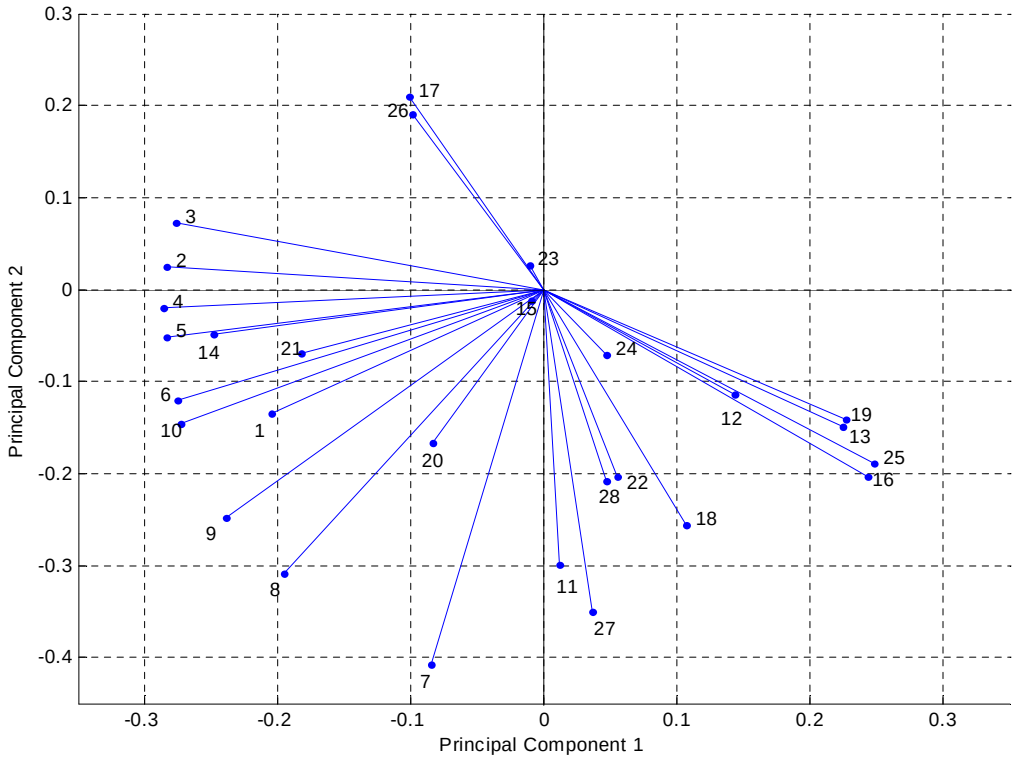

Figure 6.4: PCA: Variable Contributions to the Principal Components

(refer to table 5.3) has a negative overall contribution of about -0.2 on the first principal component and a contribution of about -1.4 on the second principal component. Variable 17 has a negative contribution to PC 1 and a positive contribution to PC 2. We can see that variables 17 and 26 (corresponding to the reflux flow and valve position) affect both the first and second PCs in almost an identical manner. This suggests that the two variables are strongly correlated, in keeping with their physical relationship. Variables 15 and 24 (the feed flow and valve position) both have only a small affect on the PCs. This is because the setpoint for the feed flowrate was not changed during the experiments for the normal operating region. It will be interesting to see how this affects the fault detection on the feed flow fault set. It is also evident that the second PC is affected in different directions by these two variables whose physical relationship should suggest that their data should be closely related. Any changes in the PCA scores as a result of changes in the feed values are more likely to manifest themselves in changes of the plate temperatures. The plate temperatures seem to dominate the model.

Figure 6.5 shows the 95 % bag around the normal training data. There are no significant outliers. Also, due to the shape of the normal operating region, we can see that an ellipse would not be as effective as the bag used here for the data encirclement. The median is almost centred within the bag, this suggests that a close to equal weight is given to the entire normal region. As discussed previously, the experiments in table 5.2

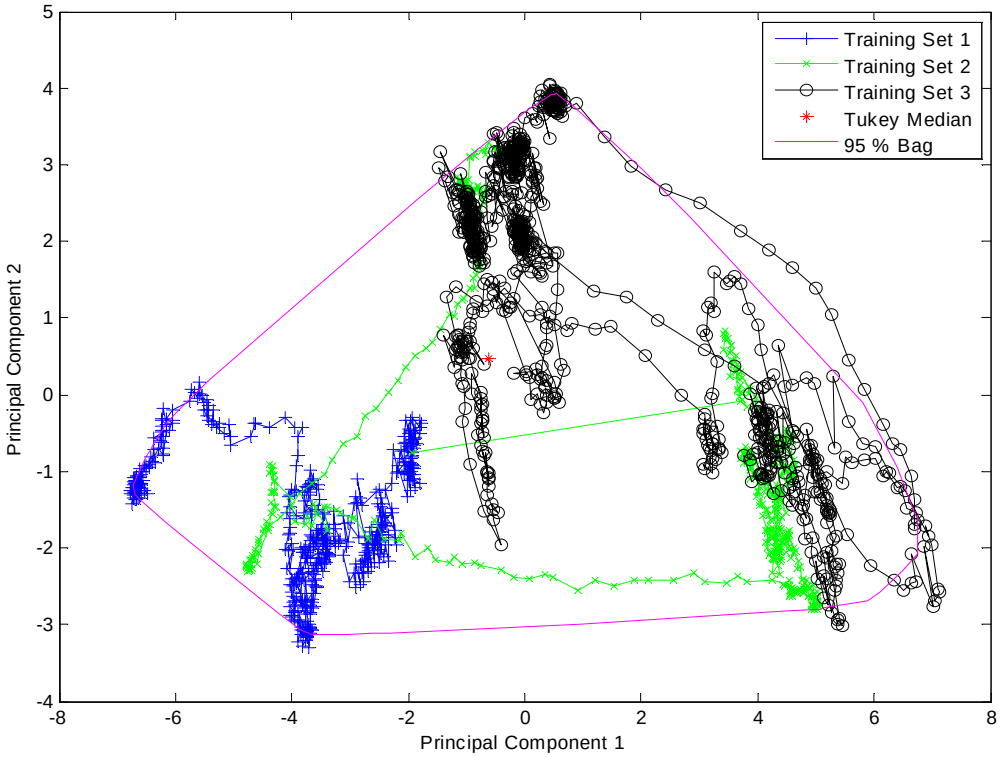

Figure 6.5: PCA: Biplot of the Normal Operating Region

were performed over 3 separate occasions of column operation. These are indicated in the figure. As some of the experiments were performed more than once on separate occasions, the data from these occasions do overlap (particulary the second training set). This suggests that - despite been run on different days and after different column startups - the data are consistent for the normal operating region.

The T^2 and the SPE statistics are calculated for the normal region to check the limits for the statistics. The normal operating region's data with the 95 % limits are shown in figure 6.6. The T^2 UCL is 6.0. The SPE UCL is 154.0. About 0.05 % of the T^2 and 3.4 % of the SPE data points are above the 95 % limits. For such a large sample, it is expected that these values would be closer to 5 %. This may suggest that the data are not normal. This may be due to the various steady states and the dynamics while moving between them. This breaks the data up into groups of possibly more normal data. There are a couple of points that exceed the SPE statistic by some margin in the normal operating region. Attempting to remove these points and to retrain the PCA model only resulted in other points exceeding the UCL in the same way. The phenomenon may be as a result of the nonlinearity in the process.

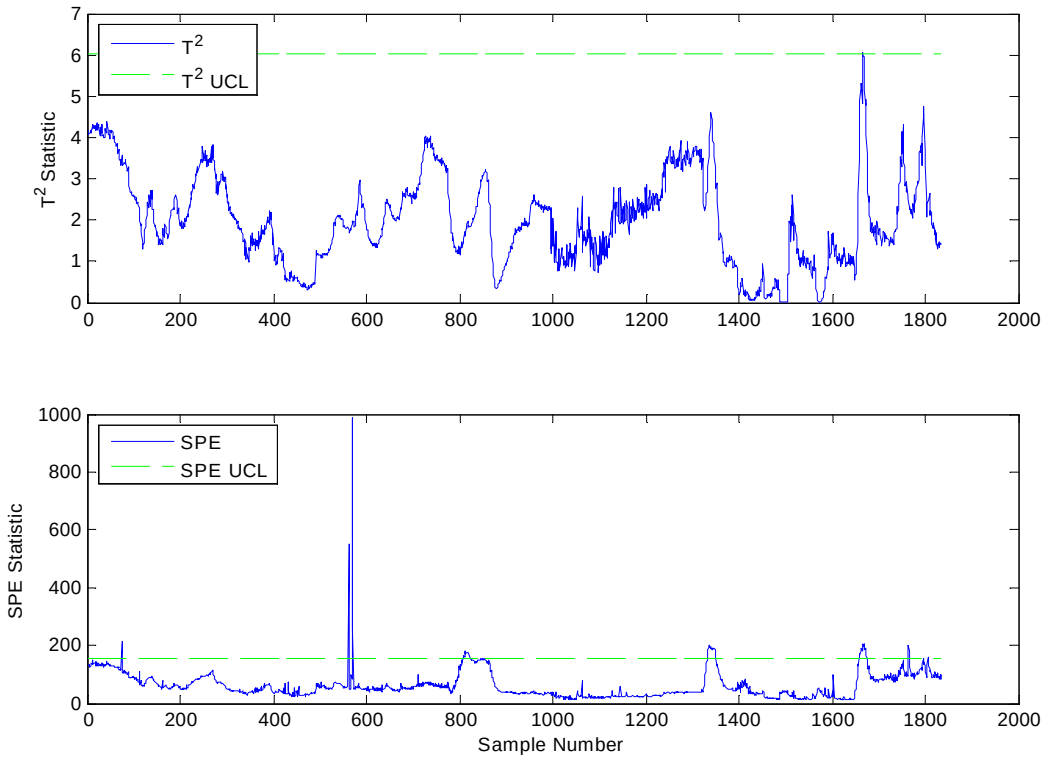

Figure 6.6: PCA: T^2 and SPE Statistics for the Normal Operating Region

The Air Failure Fault Set

This fault region was trained with 3 separately generated fault sets as discussed in section 5.2.1. The sets were each generated from starting points originating from slightly different areas within the normal operating region. The 95 % bag around the air fault scores is shown in the biplot in figure 6.7. It should be stressed that the fault region does not consist of a single point or a set of very similar points. This is because the column is still in dynamic (albeit faulty) operation.

The dataset chosen for the fault detection and diagnosis is fault set 1. In figure 6.8, we can see how the operation moves from within the normal operating region through the air failure fault bag. This can be used to diagnose a fault (the shift of data samples from the normal operating region) by checking to see if the new points fall into the air supply fault region. The controllers return operation back to within the normal operating region a while after the air utility is restored. The normal operating region and the air failure bags are well separated.

The plots of the T^2 and the SPE statistics together with the upper control limits (found from the normal operating data) are shown in figure 6.9. It can clearly be seen that the T^2 and the SPE limits were violated for the same 20 samples (at the same time) as the scores had moved into the air failure bag region. Both the T^2 and the SPE limits

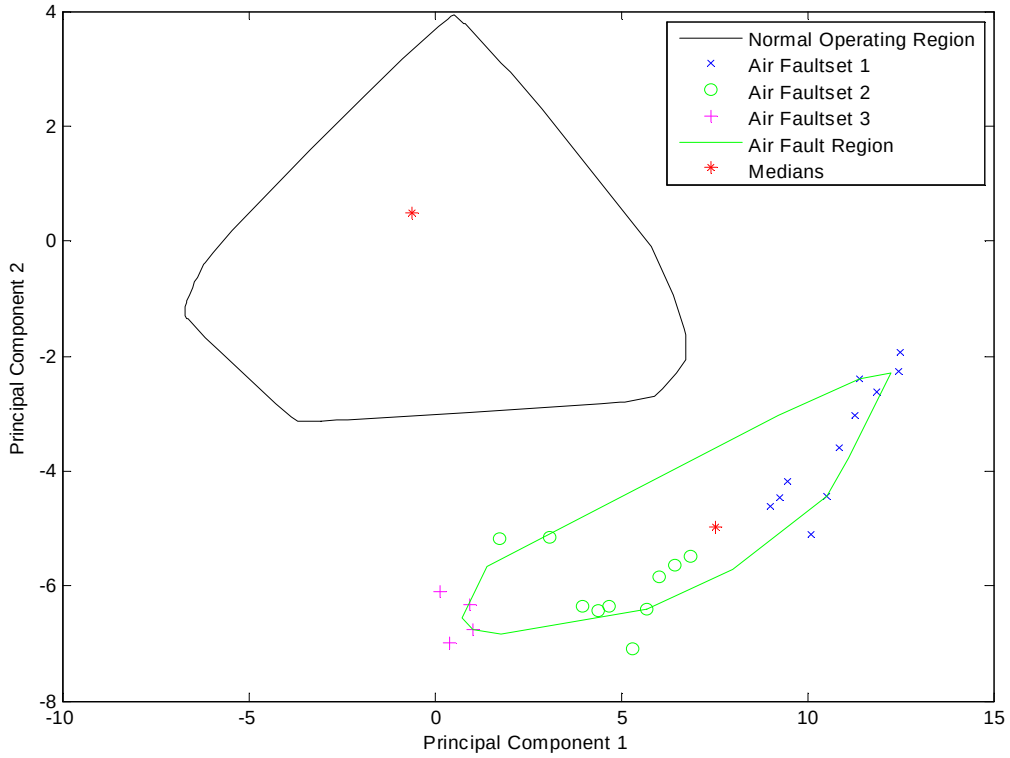

Figure 6.7: PCA: Training Sets for the Air Failure Fault

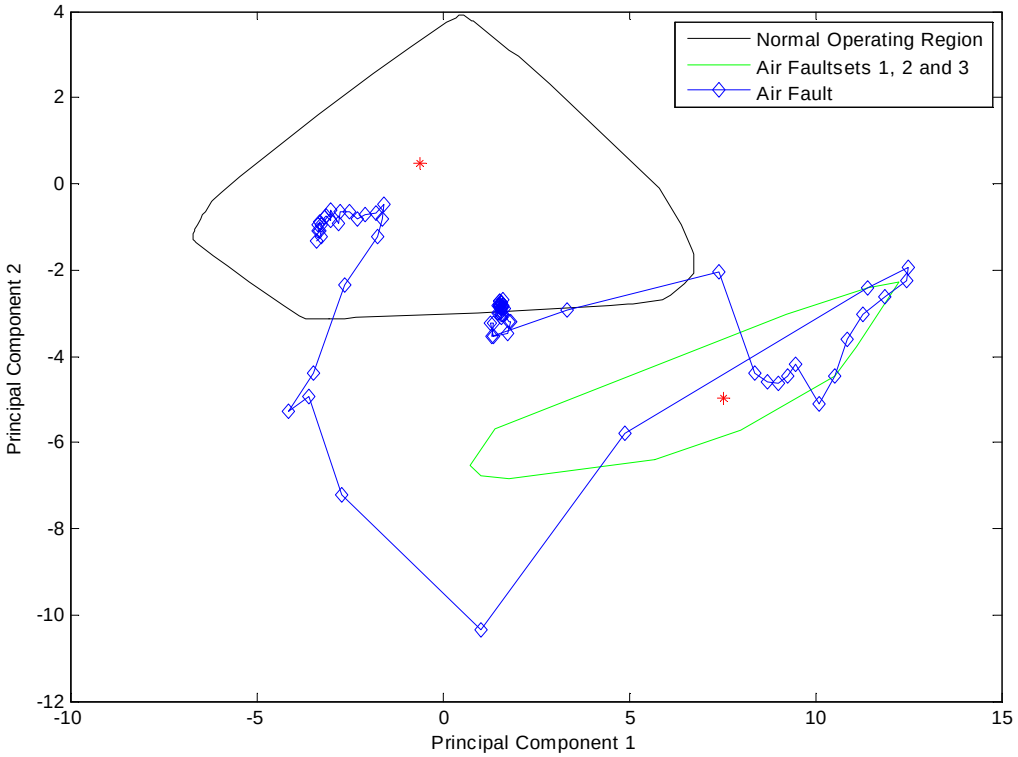

Figure 6.8: PCA: Air Failure Fault Transition

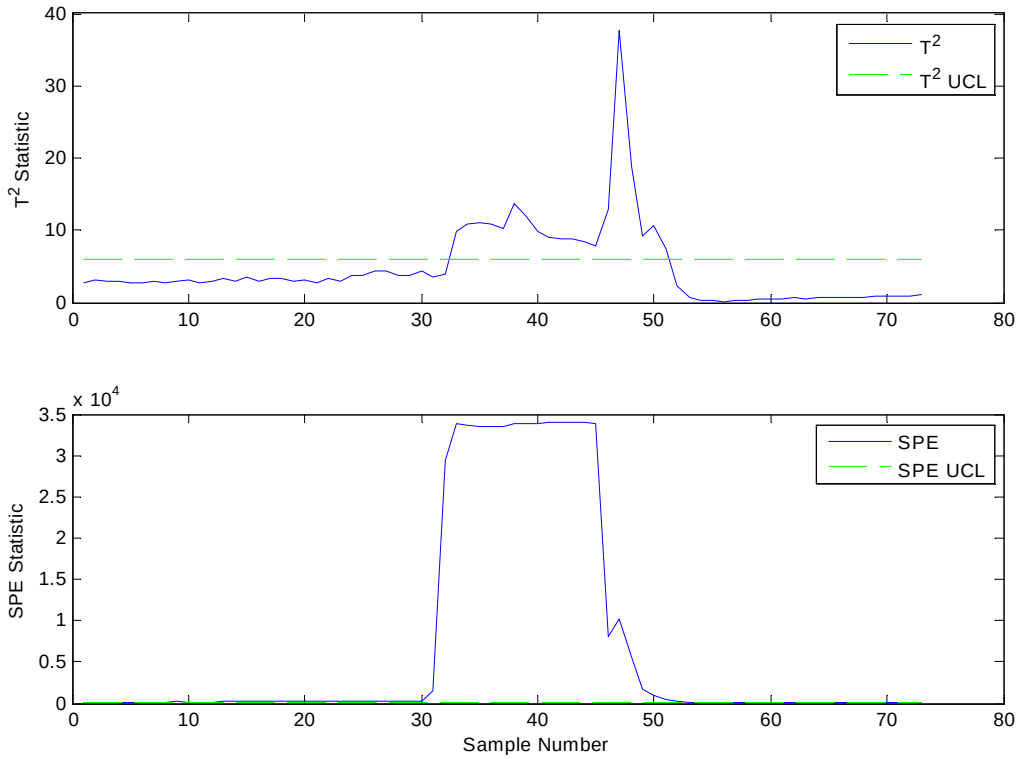

Figure 6.9: PCA: T^2 and SPE Statistics for the Air Failure Fault

are violated. This suggests both within and out of model error. The SPE limit has been surpassed by a huge margin. This suggests that the column is operating far from the normal operating region. This is expected due to the huge effect that a failure in the air supply would have on the valves. Using statistics derived from linear PCA, we have successfully detected a air failure fault.

In terms of diagnosis, a contribution plot for the point at which the T^2 statistic is a maximum is shown in figure 6.10. The contribution is the reverse calculation of the PCA rotation. The value of the score (in each PC direction) is multiplied by the contribution of variables to each principal component at the point of high T^2 or SPE. This is converted to a bar graph. Here we can see that the fault diagnosis did not lead to a definitive result (no air utility pressure was directly measured). The contributions were dominated by variables 21, 22, 1 and 20. These correspond to pressure within the column, boiler level, top plate temperature and steam pressure. The pressure within the boiler is likely to be affected strongly by the steam valve closing due to lack of air utility. The cooling and condensing of the vapour (due to lack of boil-up and condensing water) also manifests itself in changes in column pressure. Boiler level will also fall then rise due to the collapse of the bubbles then the lack of boil-up. The top plate may be strongly affected because of the reflux flow stopping. Thus the result of the contribution plot can easily be related back to physical phenomenon on the column. It does not give a definitive diagnosis,

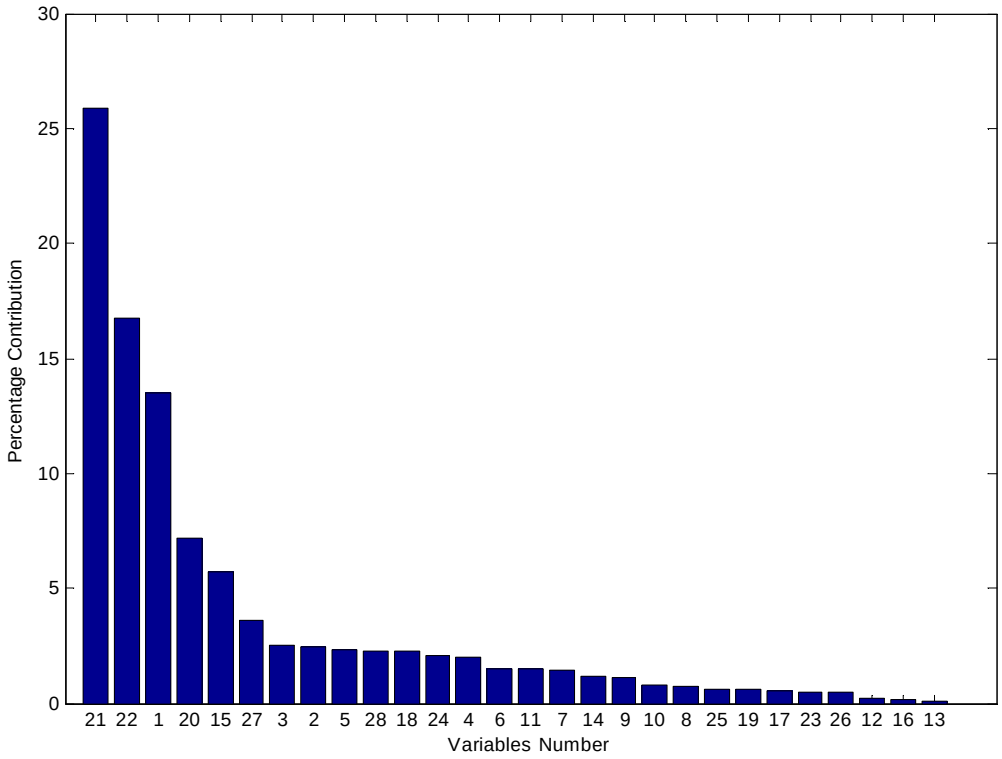

Figure 6.10: PCA: Contribution Plot for the Air Failure Fault

however, with interpretation, the search for the fault can be narrowed down.

The Steam Supply Failure Fault Set

The fault region for the steam supply dataset (discussed in section 5.2.1) was trained with two artificially generated steam faults. The biplot with a bag around the training data scores is shown in figure 6.11. The medians of the normal operating region and the fault bag are well separated. There is some overlap between the bags. The data of the two fault sets are very close together.

Instead of creating artificial training data by causing faults to occur on the running process, it may be possible to simulate training fault data for faults that are too rare to have occurred previously or faults that are too expensive or risky to trigger merely to get data. An example is shown in figure 6.12 of normal operating data with the steam pressure changed to replicate a failure in steam supply. None of the other variables were manipulated. This experiment completely ignores the important correlations that the steam pressure has on the other variables. Here, such a simple manipulation of the data is shown to be wholly inadequate. This artificial fault set does not lie near the more realistically generated steam fault sets. This strongly suggests that it is not useful. The steam supply failure region is clearly also affected by the other variables.

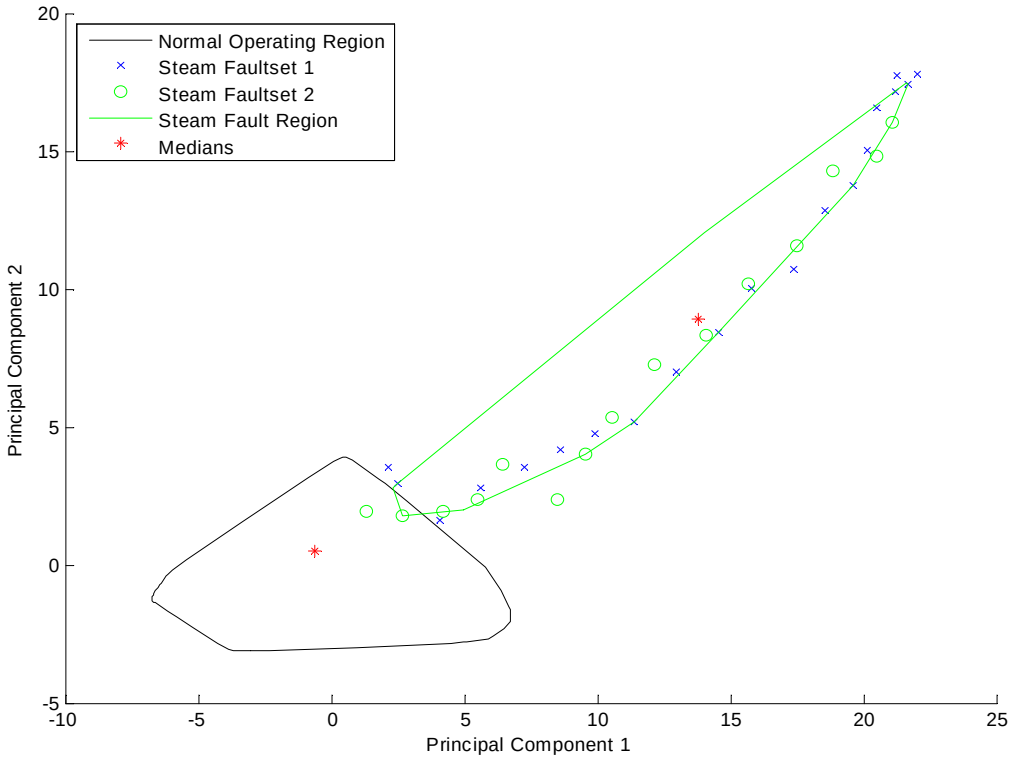

Figure 6.11: PCA: Training Sets for the Steam Supply Failure Fault

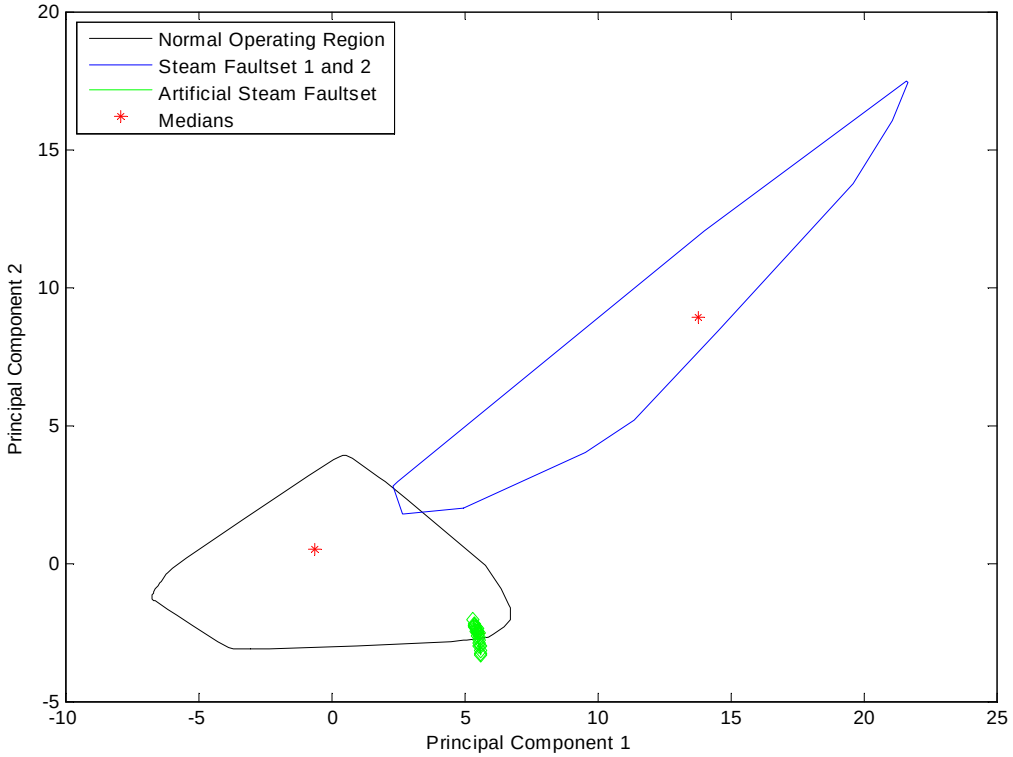

Figure 6.12: PCA: Manipulated Data used for Compared to the Steam Fault Training Sets

In figure 6.13, we can clearly see the process moving from normal operation into the steam supply failure bag. A fault could be diagnosed as being a steam supply fault by using this training region technique. Remember that the dataset for fault detection and diagnosis was recorded while being used by different operators more than a year before the data for the training faults and normal operating region was generated. This shows that the method (or the column itself) is robust despite the changes in operation and process. The recovery back to the normal operating region is not shown as the return of the boilers to working pressure was not recorded.

The plots of the T^2 and the SPE statistics together with their upper control limits are shown in figure 6.14. Both the T^2 and SPE statistics show peaks at the point where the steam had failed. Clearly this increase shows that the fault was detected by PCA model mismatch. It is interesting to see that the SPE statistic starts from a value significantly above the UCL (while it was supposed to be operating normally). This may be because of process changes between the (lengthy) time of generation of the fault set and the rest of the sets. The different operators may also be a factor as the SPE statistic indicates out of plane PCA model mismatch. PCA statistics were successful in detecting a steam supply fault.

In terms of fault diagnosis, the contribution plot for the steam supply fault (shown in

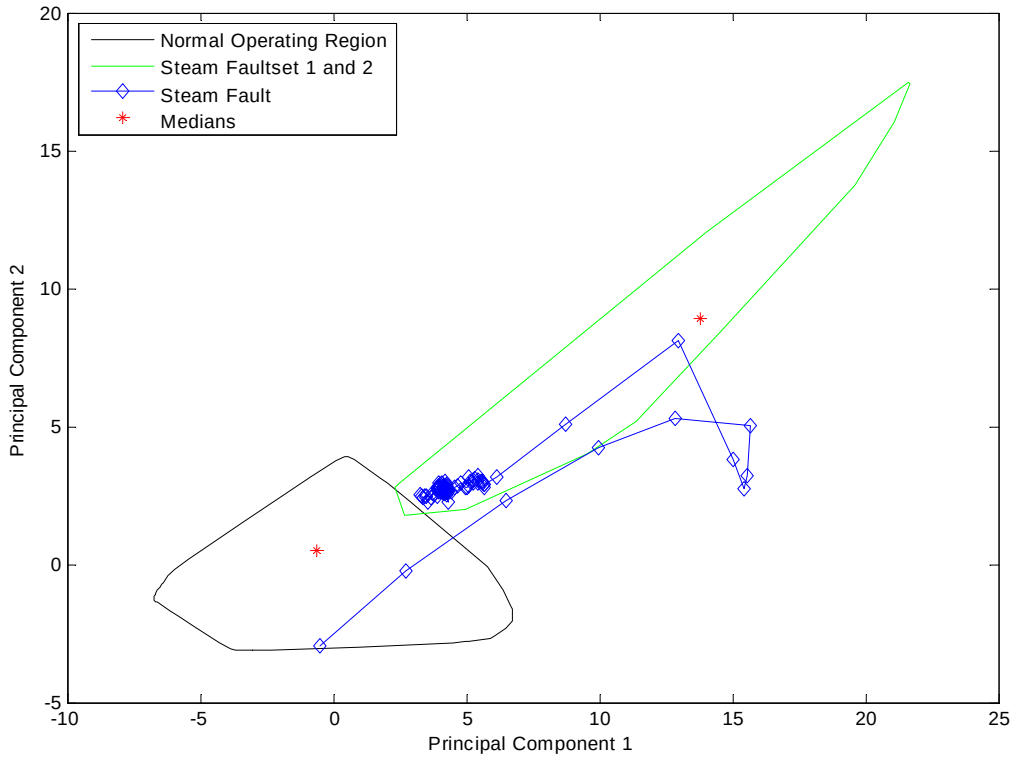

Figure 6.13: PCA: Steam Supply Failure Fault Transition

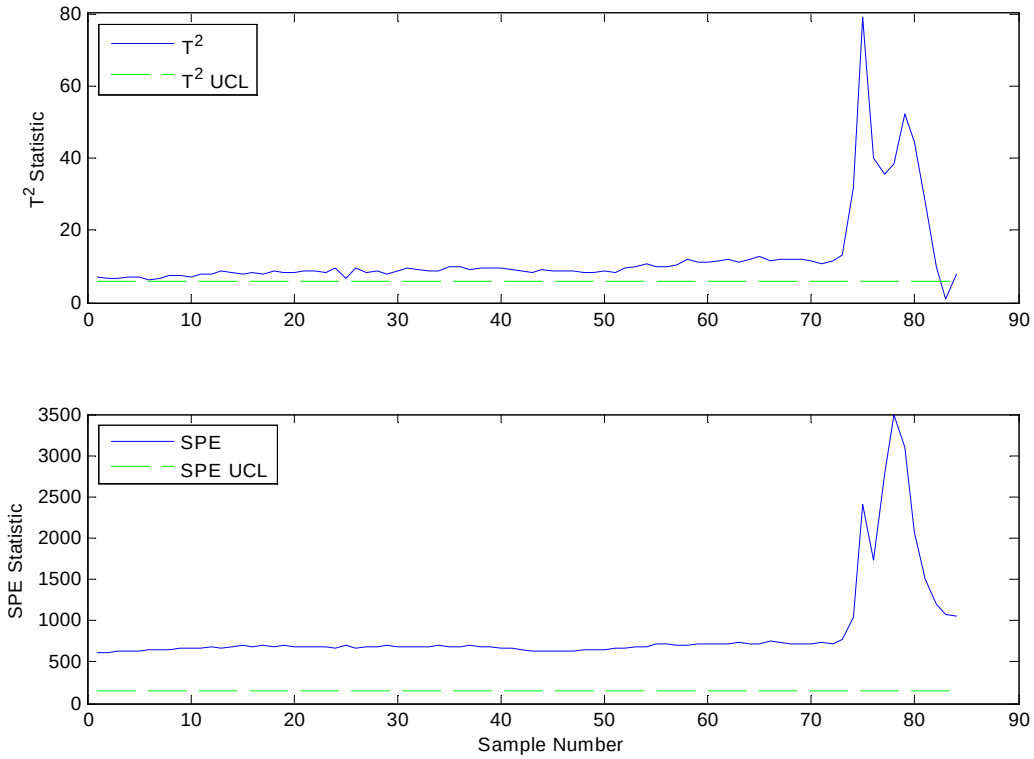

Figure 6.14: PCA: T^2 and SPE Statistics for the Steam Supply Failure Fault

figure 6.15) shows that variables contribute most to the scores that result in maximum T^2 values. Variables 21, 1, 7 and 22 contribute the most. Variables 21, 1 and 22 can be explained in the same way that they explained the contributions for the air supply fault. Variable 7 (the feed plate temperature) differs from the the air supply failure contribution plot. This can be explained by noting that the (cold) feed continued to flow during this time. The lack of boil-up warming the plate means that the cold feed had a large affect on the temperature of plate 7. Unfortunately the steam pressure variable (number 20) does not feature strongly. Possibly there was still some residual pressure (as opposed to the training sets where the pressure dropped to zero). The significantly lower pressure could still affect the rest of the column dramatically.

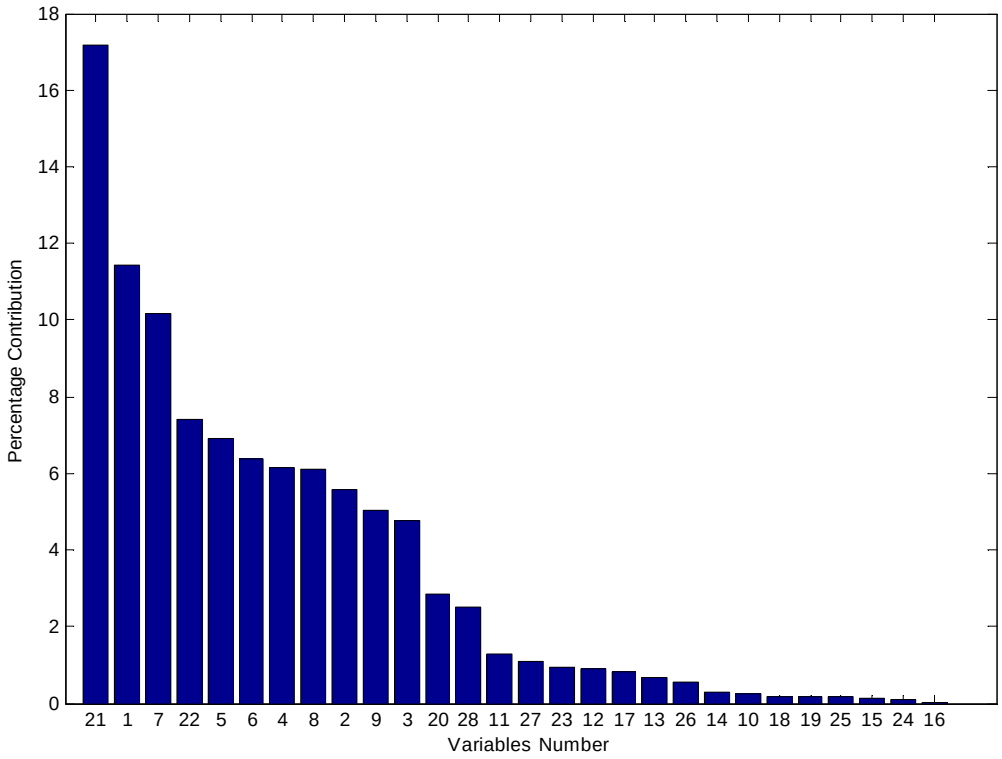

Figure 6.15: PCA: Contribution Plot for the Steam Supply Failure Fault

The Feed Flow Fault Set

The feed flow fault set (discussed in section 5.2.1) consists of 2 training data sets and one data set taken from actual operation for fault detection. The scores of the two training sets (together with the 95 % bag) is shown in figure 6.16. Both datasets lie fairly close together. The first set extends further than the second set. This is because the training data was generated from a different steady state. The feed fault bag and median is distinctly separated from the normal operation region.

In figure 6.17, the transition from normal operation to operation within the fault region can be seen. A fault could be diagnosed as being a feed fault by using this training region technique. Once the feed hand valve was corrected, the feed controller managed to return the column back to the normal operating region.

In figure 6.18, we can see the T^2 and SPE statistics for the feed fault set. The T^2 and SPE both spike at a similar time. Again the SPE statistic starts above the upper control limit. Clearly the SPE is sensitive to the initial conditions. Possibly the normal operating region used for training does not include all the relevant non-faulty operation conditions. A fault is still easy to detect using either of the statistics as both show marked increases at the time of the fault.

The contribution plot is shown in figure 6.19. Here we see that the variables con-

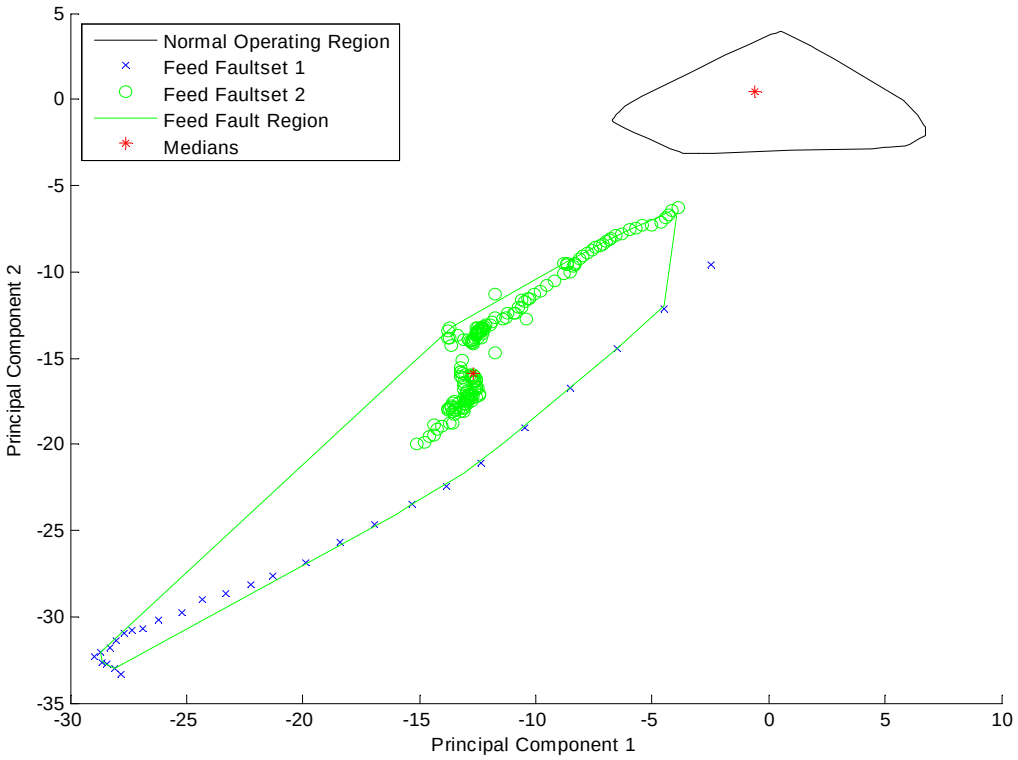

Figure 6.16: PCA: Training Sets for the Feed Fault

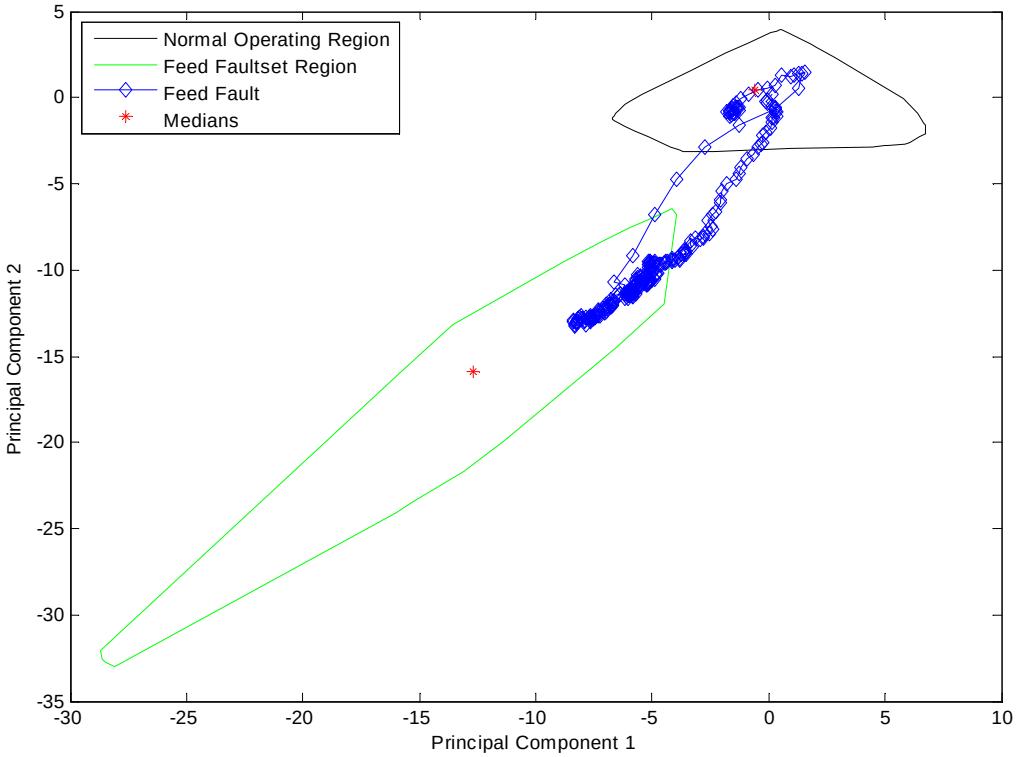

Figure 6.17: PCA: Feed Fault Transition

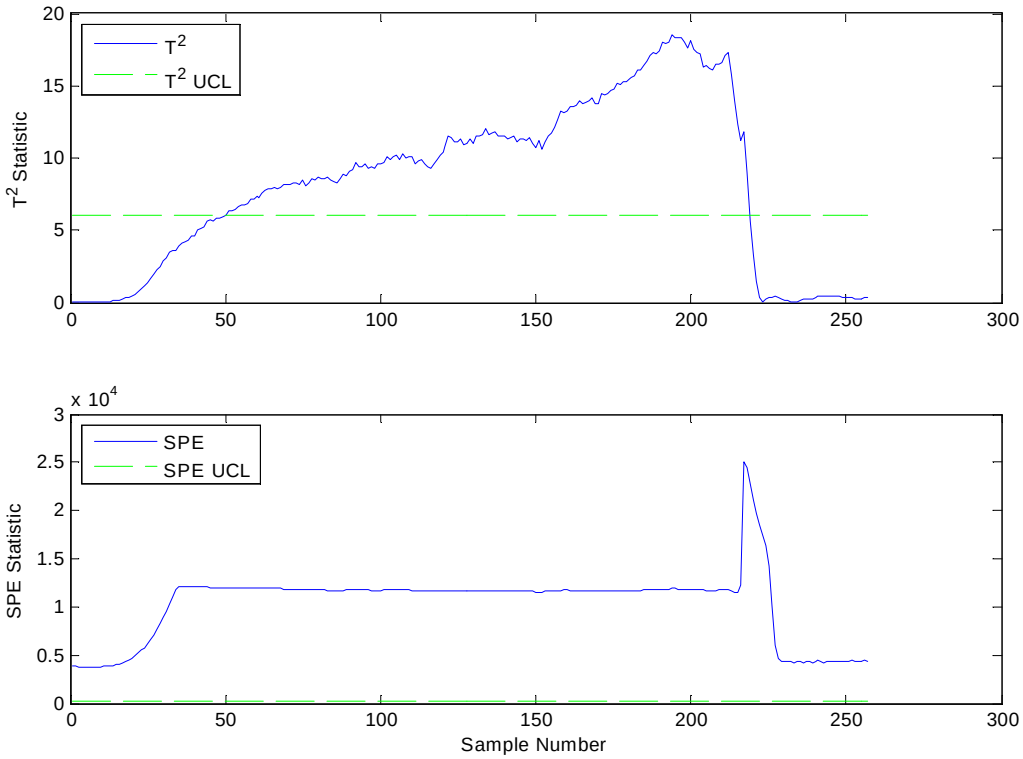

Figure 6.18: PCA: T^2 and SPE Statistics for the Feed Fault

tributing the most to the maximum T^2 are variables 9, 10, 8, 11 and 24. Variables 9, 10, 8 and 11 correspond to the temperatures on the plates below the feed plate. Possibly with the reduction of feed, these temperatures rose - contributing to the high T^2 and SPE values. Variable 24 is the feed valve position (CV-01). This points directly to the fault. It would have been ideal if this value had contributed most to the fault - making fault diagnosis easy. Unfortunately, as can be seen in figure 6.4, the contribution of the feed flow rate and feed valve position to the first two principal components is small. This is probably because the feed flowrate (and the feed valve position) was kept at a constant value during the training of the normal operating region. While the contribution plot did not directly point to the cause of the fault, some interpretation leads to the diagnosis.

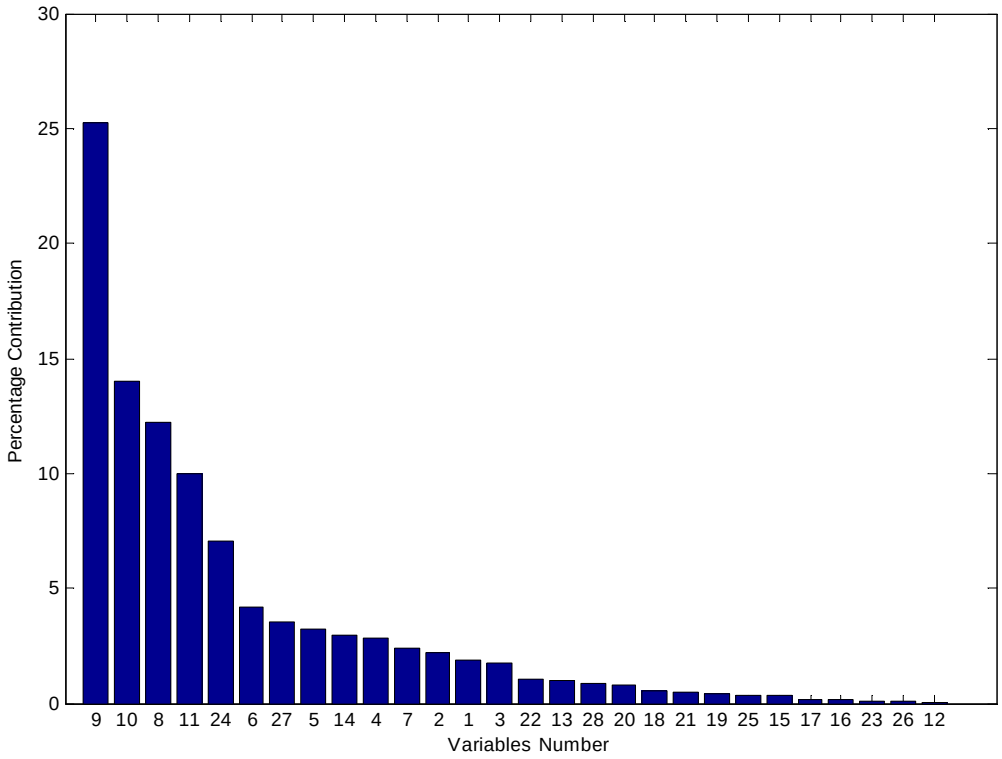

Figure 6.19: PCA: Contribution Plot for the Feed Fault

The Top Product Flow Fault Set

The top product flow fault set (discussed in section 5.2.1) consists of 2 training data sets and one data set taken from actual operation for fault detection. The scores of the two training sets (together with the 95 % bag) is shown in figure 6.20. Both datasets lie fairly close together. The feed fault bag and median lie completely within the normal operating region. This suggests that the PCA model considers the physical manipulation of the valve to simulate hysteresis in flow as part of the normal operating region. This may be because it did not simulate the fault as accurately as expected. Alternatively, the variables for the top product flow and valve position may not affect the PCA model enough.

In figure 6.21, operation while experiencing top product flow problems is shown. The transition from and back to the normal operating region is not shown. These samples lie close to the region for the feed flow fault (shown in figure 6.16) than the region of the faults used for training. As before, this may be because the artificial fault does not resemble the true fault accurately enough. Interesting, the shape of the actual fault region is similar to that of the training region. Fault diagnosis by observing which region data samples fall into (training regions) would not be successful here.

In figure 6.22, we can see the T^2 and SPE statistics for the top product flow fault set.

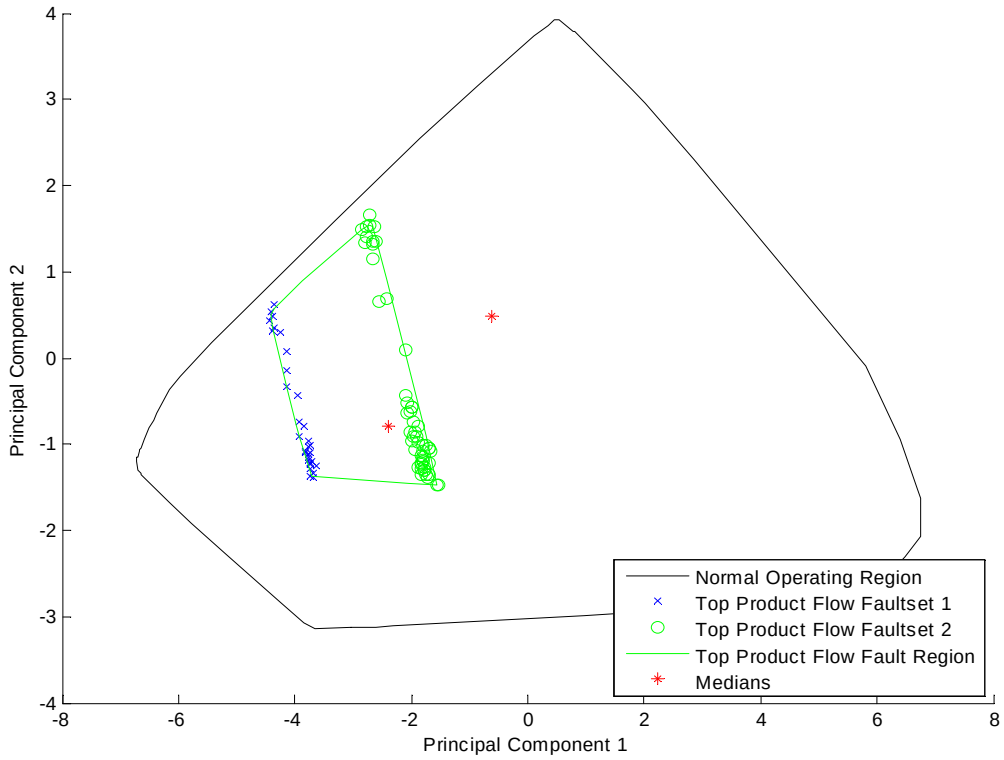

Figure 6.20: PCA: Training Sets for the Top Product Flow Fault

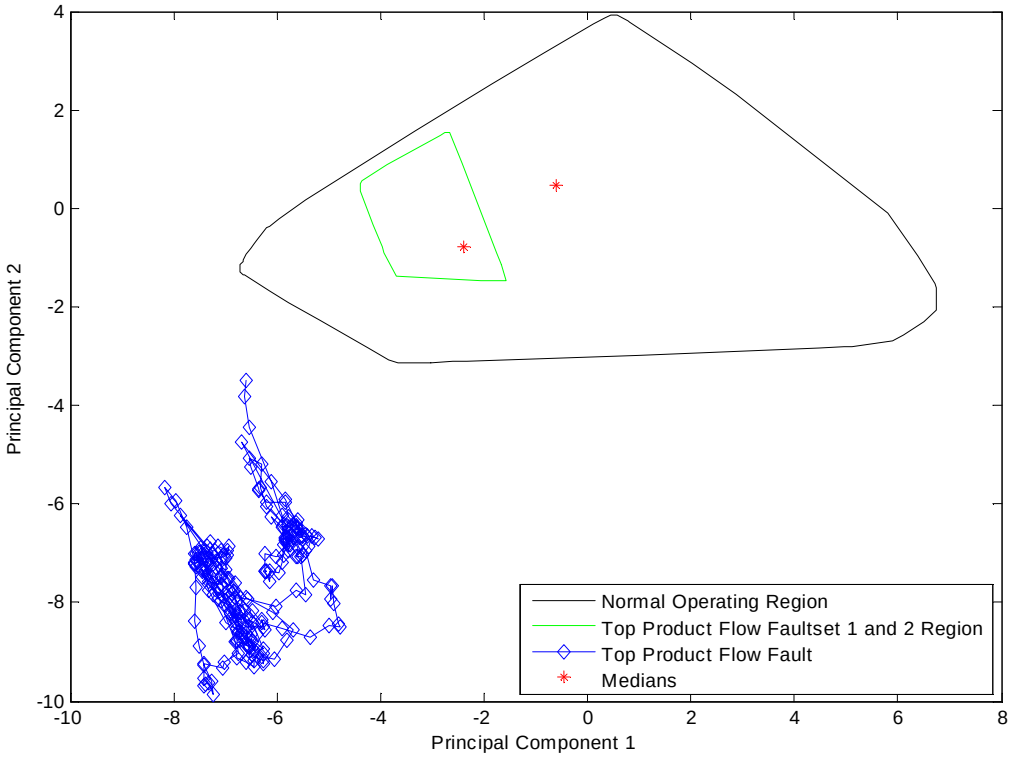

Figure 6.21: PCA: Operation with a Top Product Flow Fault

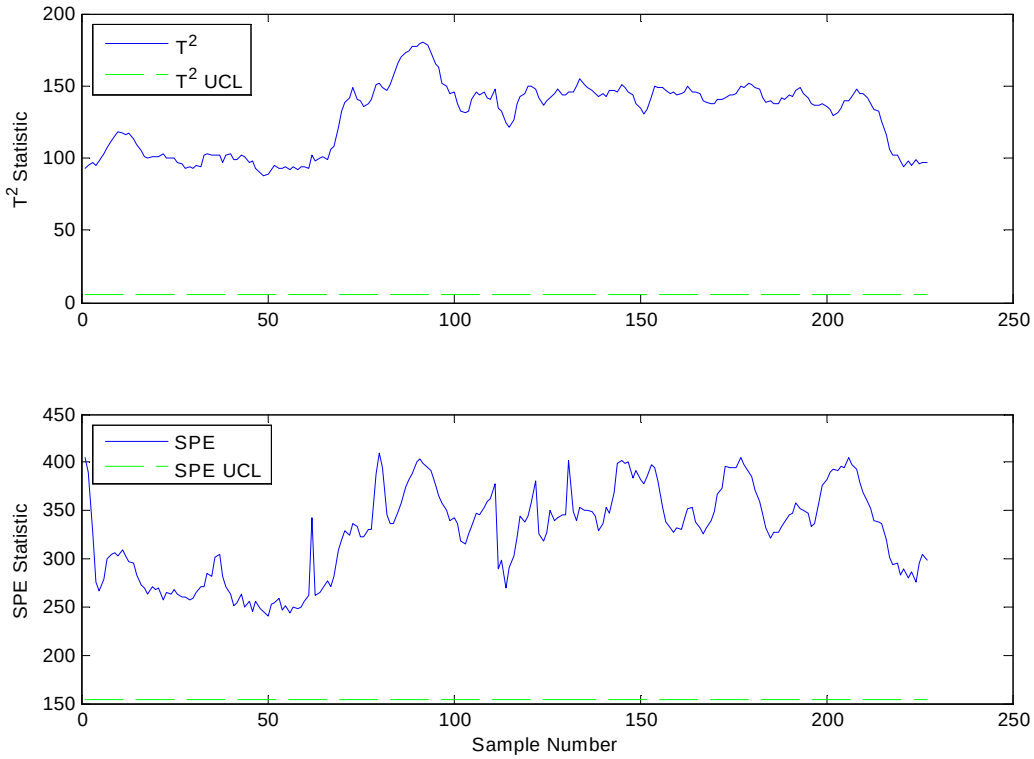

Figure 6.22: PCA: T^2 and SPE Statistics for the Top Product Flow Fault

Note, that all of the data are faulty and we expect that all of it to be above the upper control limit (as is the case). This suggests that if the column was operating normally and that a top product flow fault does develop, it will be possible to detect it as a violation of the UCL of both the T^2 and SPE statistics.

The contribution plot is shown in figure 6.23. Here we see that the variables contributing the most to the maximum T^2 are variables corresponding to the plate temperatures. The temperatures on the plates contribute more to the fault condition than the top product flow and valve position. As discussed in section, this may be because of the dominant role that the plate temperatures play in the model. Here, fault diagnosis using a contribution plot would not give useful results (even with some interpretation). As we saw earlier, fault diagnosis by checking to see which fault region the samples falls into was also not successful.

An overview of the fault regions discussed above is shown in figure 6.24. It can be seen that all the faults - excepting the top product flow fault - are well separated from each other and the normal operating region. Possibly more advanced nonlinear models and discriminant methods (discussed later) will be able to separate these regions more optimally. This may aid diagnosis by means of the training regions. This may lead to an improved top product flow fault region which can be discriminated from the normal operating region. This overview will also be useful in determining if the novel faults

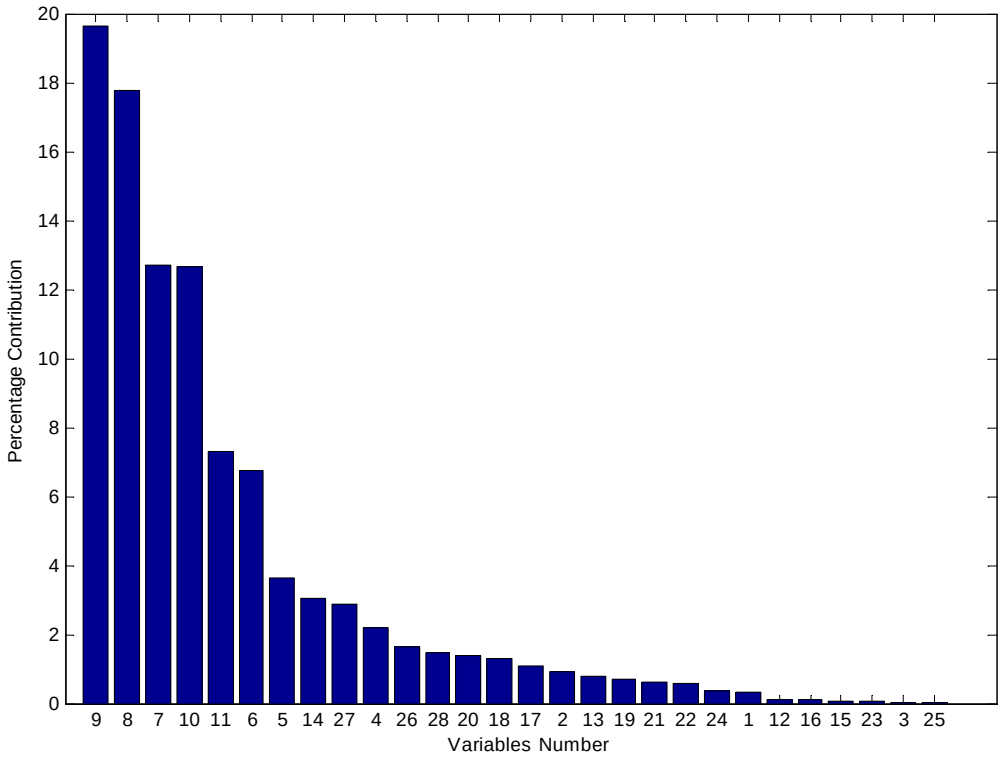

Figure 6.23: PCA: Contribution Plot for the Top Flow Fault

(discussed below - which have no training data), will be misdiagnosed with training regions as being one of the known faults.

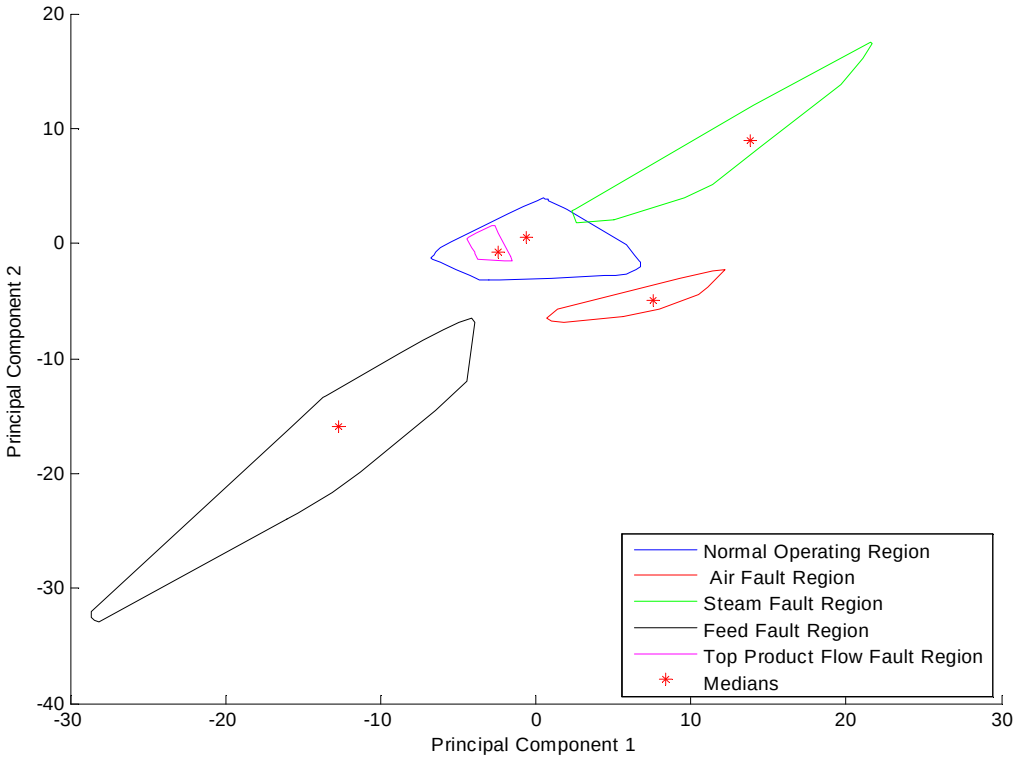

Figure 6.24: PCA: Overview of all Fault Regions

The Novel Fault Sets

The fault detection and diagnosis results for the novel fault sets (see section 5.2.1) using PCA are discussed here. As discussed before, these sets are intended to test the ability of the fault detection and diagnosis technique to handle faults that have not been seen before (therefore no training data are available).

The scores for all these fault sets are shown relative to the normal operating region in the biplot in figure 6.25.

For water rich operation, the data points lie at the bottom left of the biplot, beyond the extreme of the feed flow fault region. Clearly, if data samples are so far removed from the normal operating region, the operation can be considered faulty. One would be confident in diagnosing the fault as novel because it is so far from the training regions of the known faults in the biplot.

The feed flooding operation fault region lies closer to the feed flow fault region and the normal operating region. These point are in a very tight group near the normal operating region. It may be difficult to confidently flag these samples as faulty. This is because excessively low *and* high feed flows are part of the feed fault region. This is because of the rapid increase in feed when the feed choking is relieved (see section 5.2.1). Even if the fault was detected, it is possible that operation with excessive feed may be misdiagnosed

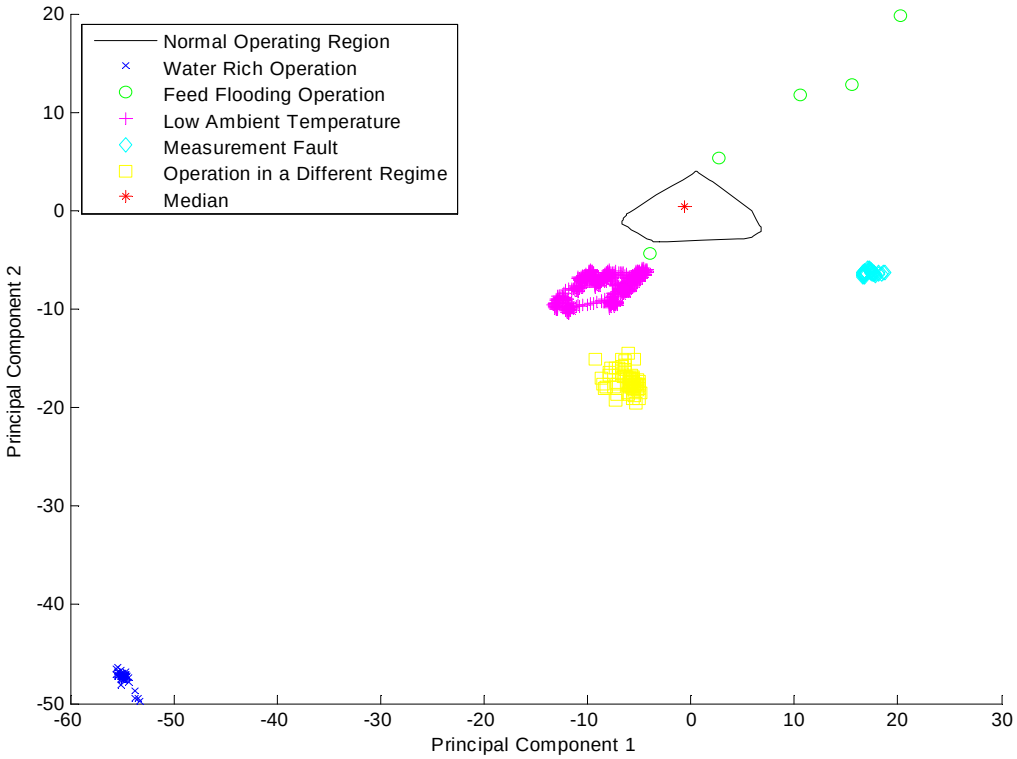

Figure 6.25: PCA: Novel Faults Overview

as a feed flow choking fault.

The scores for operation in low ambient temperature lie close to those of the feed fault region. It is possible that operation in low ambient temperature condition may be misdiagnosed as a feed flow fault.

The scores for the measurement fault lies in a tight group well apart from the normal operating region. Data samples this far from the normal operating region would be clearly identified as faulty. This group is fairly close to the region where air supply faults are known to fall. Confident diagnosis of this fault as novel (as opposed to being an air supply fault) may be difficult using a biplot with the training regions of the previous faults.

The scores from the operation in a different regime lie in a distinct group that is well separated from the normal operation region. One would be fairly confident in recognising this fault as novel from a biplot with all the training regions.

The T^2 and SPE statistics for the novel faults are shown in figure 6.26. Note that we expect all samples of these fault sets (except the feed flooding set) to violate the UCLs. This because the transition for these variables from the normal operating region was not possible (or recorded). The feed flooding operation statistics clearly show violation of the UCL when operation changed from normal operation to fault condition. This occurred when the feed setpoint was changed to a high value.

The SPE statistic was particularly useful (as expected) in detecting these faults. The

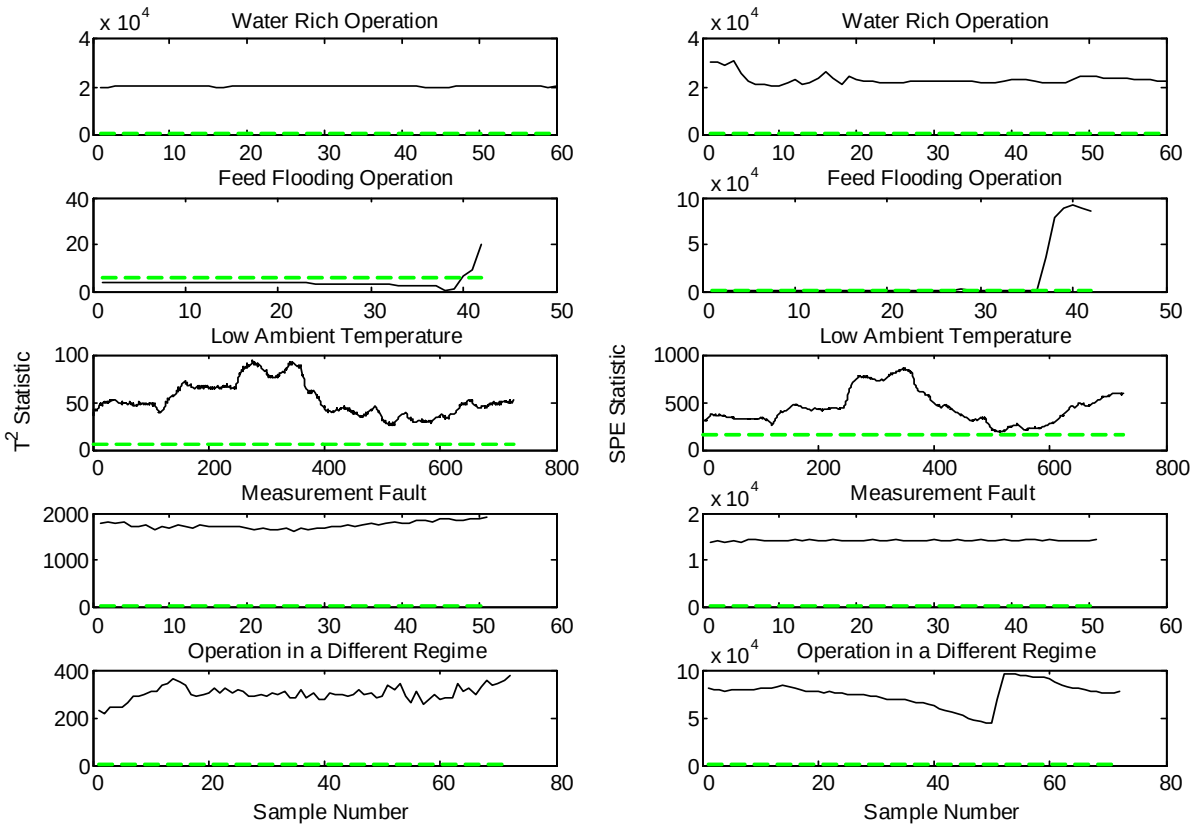

Figure 6.26: PCA: T^2 (left hand side column) and SPE (right column) Statistics for the Novel Faults

PCA monitoring statistic have been successful in detecting this set of novel faults.

The contribution plots for the novel faults is shown in figure 6.27.

The lower plate temperatures dominate the contribution plot for the water rich operation fault. This is due to the boiling point of water being higher than that of the ethanol-water mixtures that the normal operating region was trained on. The dominance of the plate temperatures may lead (correctly) to a diagnosis of a problem with the liquid on the plates.

For feed flooding operation, the first three main contributions are again the plate temperatures. This is due to the small influence that the feed variables have on the process (as discussed previously). Importantly, the fourth most important variable in the feed flow rate. With some understanding of the model, this fault could successfully be diagnosed as a feed fault. The feed valve position is not a significant factor in the contribution plot.

For low ambient temperature operation, the contributions are again dominated by the plate temperatures. This may be because of the heat losses due to the environment affecting the plate temperatures. This plot may lead to a successful fault diagnosis given some appropriate interpretation.

The contribution plot for the measurement fault unequivocally diagnoses the fault as

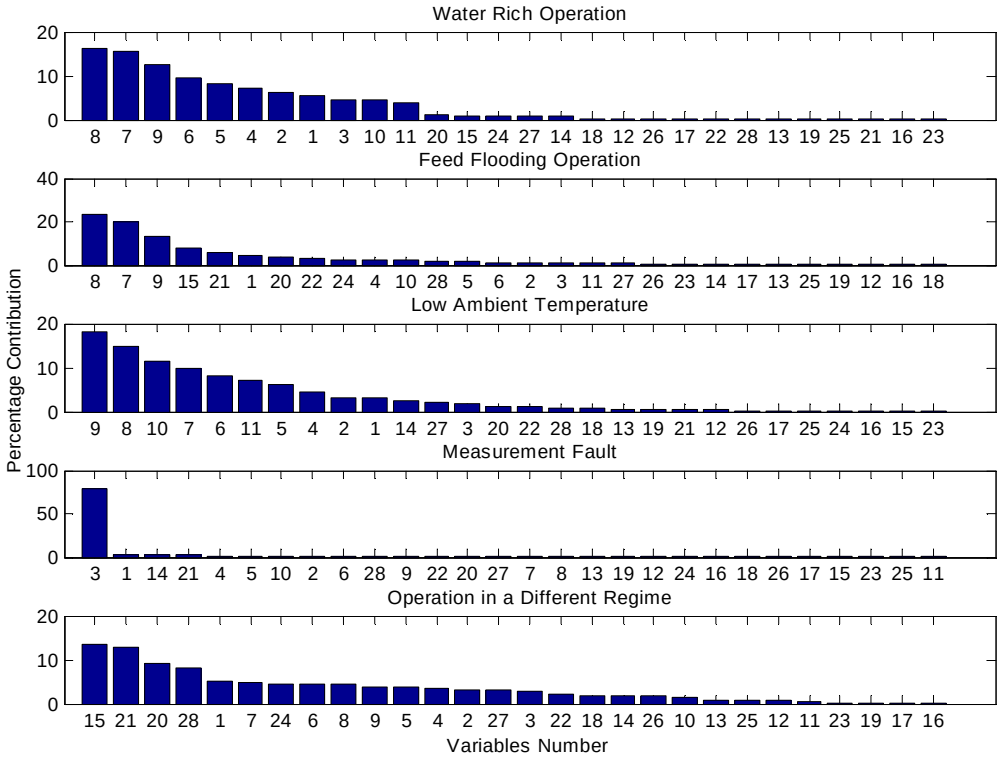

Figure 6.27: PCA: Contribution Plots for the Novel Faults

a measurement failure on the temperature of plate 3. This is an excellent example of a successful and unambiguous fault diagnosis.

Most of the variables contributed in some degree to the fault detection on the different operating regime fault set. The variables that dominate are the feed flowrate (which is nearly triple the normal value), the column pressure and the steam pressure and valve position (more than double the normal values). Faced with such a contribution plot, one would certainly suspect an unusual operation point due to the changes in these variables while the plate temperatures remained close to normal values.

6.2.2 Fault Detection and Diagnosis using Kernel Principal Component Analysis

The results for KPCA, the nonlinear extension of linear PCA, are discussed here.

Training of the KPCA Model

The KPCA model was trained using the normal operating data. The kernel argument for the Gaussian kernel was difficult to select Lee et al. (2004a) describes it as an crucial and open problem)f. Various values were experimented with. Parameters to judge the

success of model with a given kernel argument were:

- *CPU Time* - This gives an indication of the conditioning of the matrices. A lengthy calculation suggests that the conditioning is poor.
- MSE - An indication of the error of the mapped data. This is influenced by the percentage of variance represented by the first two PCs. Ideally, this should be high as we wish to describe as much of the features of the data as is possible with these two PCs.
- SPE - The error in predicting a few of the samples back to the input space. The score was calculated back to the input space and compared to the original sample. As this is not always possible (as discussed), only 10 points (which were it was found by trial and error to be possible to calculate back to the input space) were used. If the calculation was possible, it was lengthy (about three times longer than finding the model for all the points).

While this basic attempt at optimisation did not lead to any conclusive findings, it was decided that a kernel argument of 8 should be used. This value gave excellent classification with the KDA technique with comparatively little over fitting compared to other kernel argument choices. Clearly the choice of kernel argument needs more investigation. The model took about 5 minutes to train on a well equipped desktop computer.

Note that due to the fact that there are the same number of PC generated as there are data samples in the normal operating region (nearly 2000). A plot is shown only with the variance explained by each PC line in figure 6.29. We can see the first two principal components together represent 42.4% of the total variance. In terms of choosing an output dimensionality, we encounter the same problem as we did with the linear PCA. The average variance is so low due to the high number of principal components. The cummulative percentage of variance technique suggests 5 PCs should be used. This value is higher than expected because there are many less important PCs describing small amounts of noise. This causes the first PCs to describe less of the total variance. The broken stick and size of variance techniques suggest more than 50 PCs should be used. This is clearly not feasible for visualisation. As with linear PCA, an output dimensionality of two is used. This will offer the ability to make direct comparisons to the linear PCA results and to see if KPCA is more successful in providing a reduced model.

The biplot of the normal operating region is shown in figure 6.30. The 95 % contains the data well and there are no extreme outliers. The shape of the normal operating region is similar to that of the linear KPCA regions (albeit inverted).

As before, the T^2 and SPE statistics are calculated for the normal operating region. The 95 % control limits are found. None of the samples violate the T^2 limit and 6.6 %

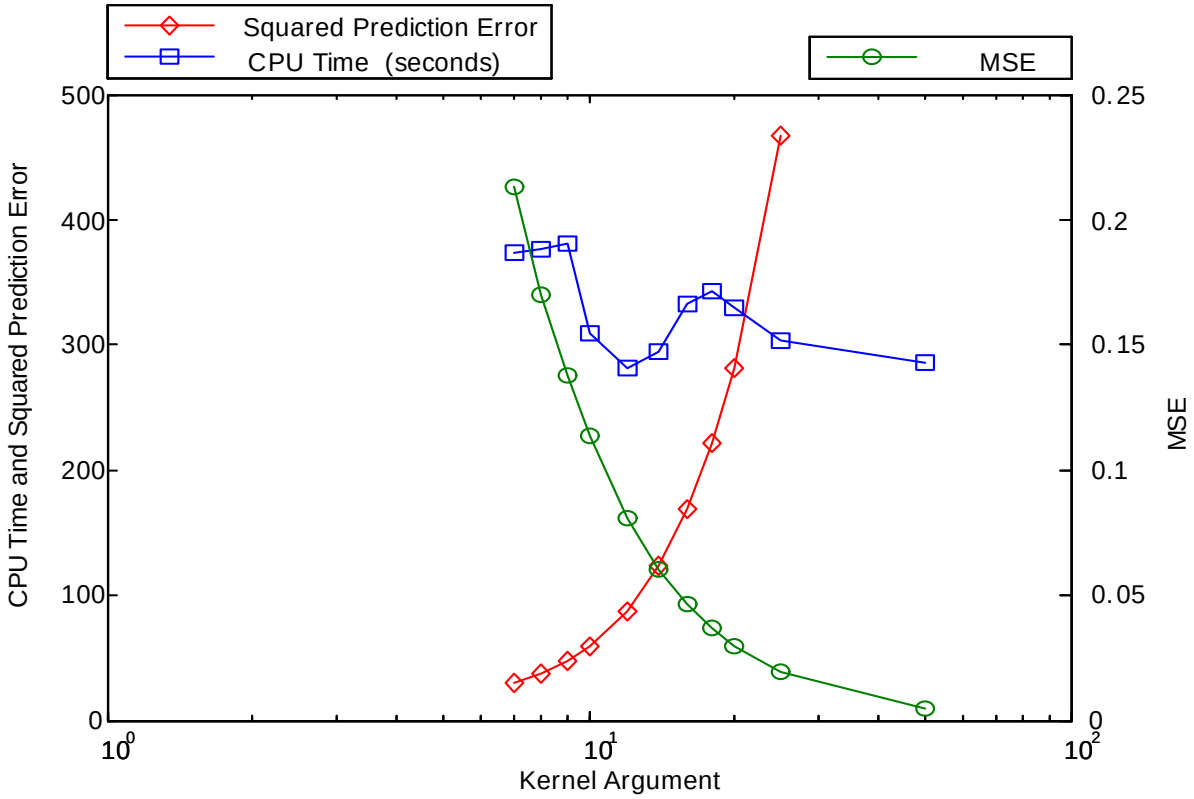

Figure 6.28: Attempted Optimisation of the Kernel Argument

of the samples are above the SPE limit. With a 95 % limit, we expect about 5 % of a large group of samples to violate the limit.

The Air Failure Fault Set

As before, with linear PCA, the air failure fault region was trained with 3 datasets. The fault region is shown in figure 6.31. Although there are many samples for each set, they all coincide at a single point in the upper part of the normal operating region.

The fault testing dataset which transitions from normal operation to a faulty condition and back to normal operation is shown in figure 6.32. The scores start and end in the normal region. They do move through the air supply failure fault region. From this biplot, it would be difficult to detect the fault. If the fault region is unique, and the condition is recognised as faulty, it would be clear that the data did move into the air failure fault training region.

The statistics suggested by Lee et al. (2004a) are shown in figure 6.33. These are not ideal for fault detection as the UCL is violated from the beginning. The faulty data is marked by a flat constant T^2 and SPE statistic. This corresponds to all the points in the (small) air supply fault region.

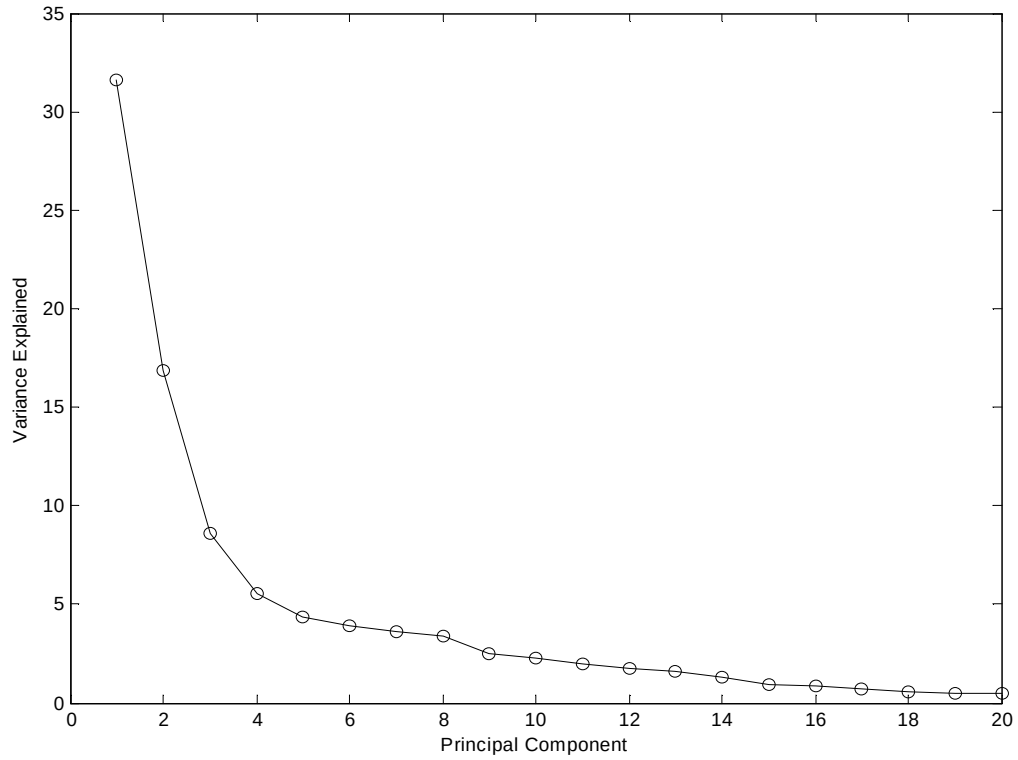

Figure 6.29: KPCA: Variance Explained by the Principal Components

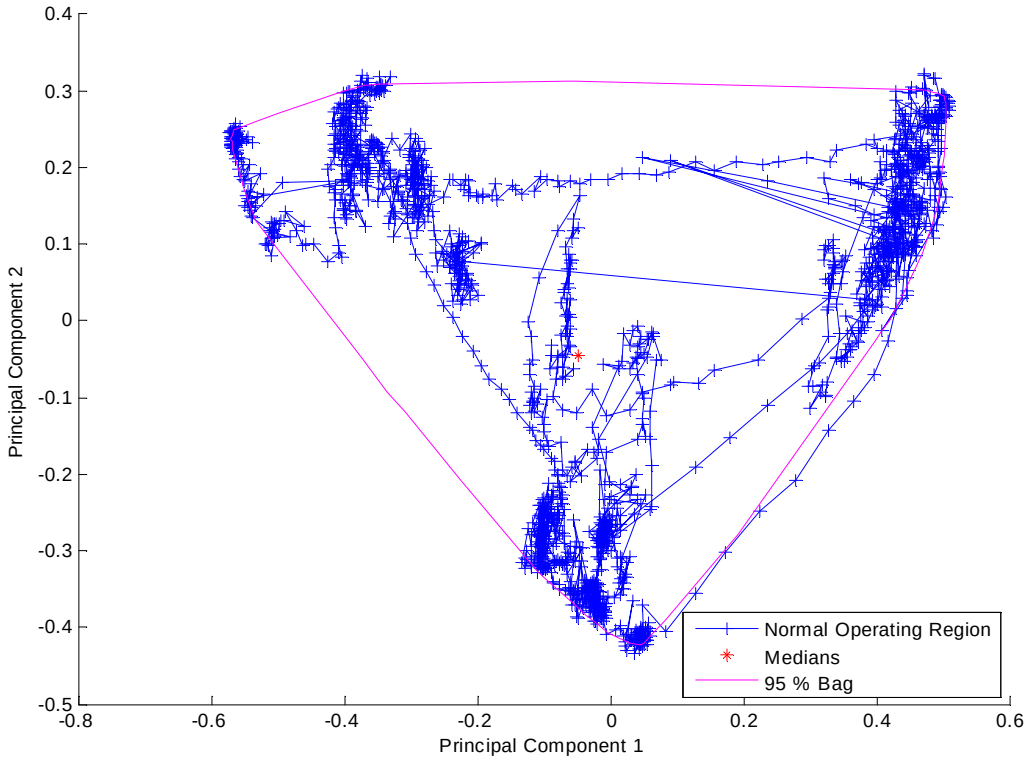

Figure 6.30: KPCA: Bagplot of the Normal Operating Region

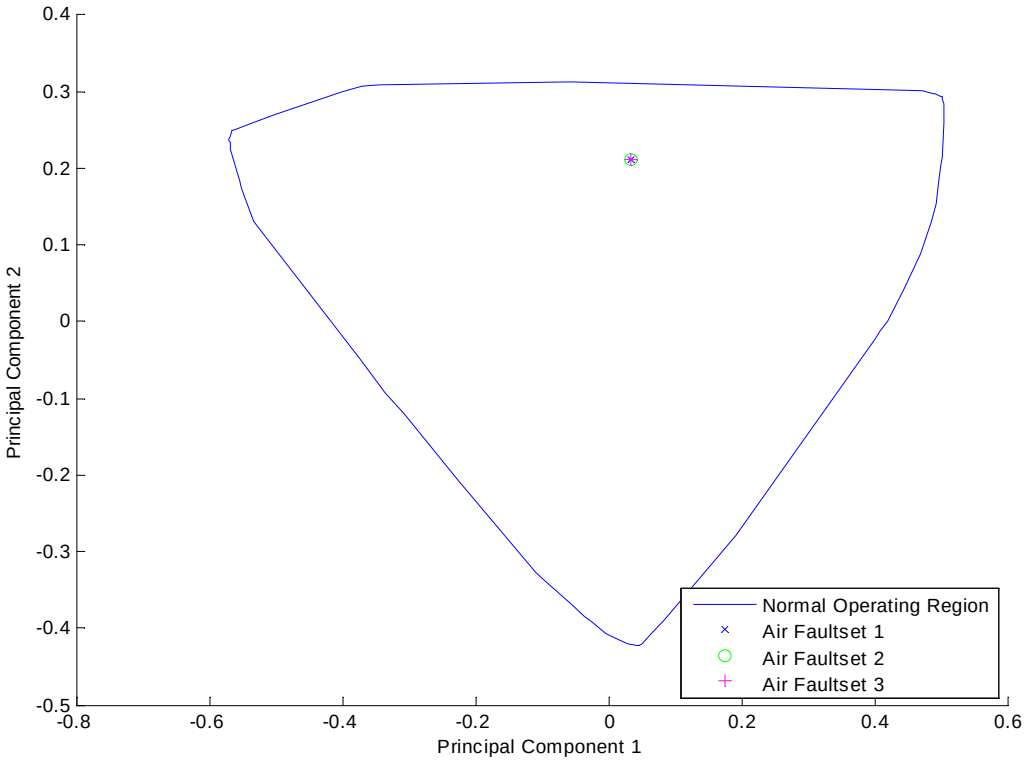

Figure 6.31: KPCA: Training Sets for the Air Failure Fault

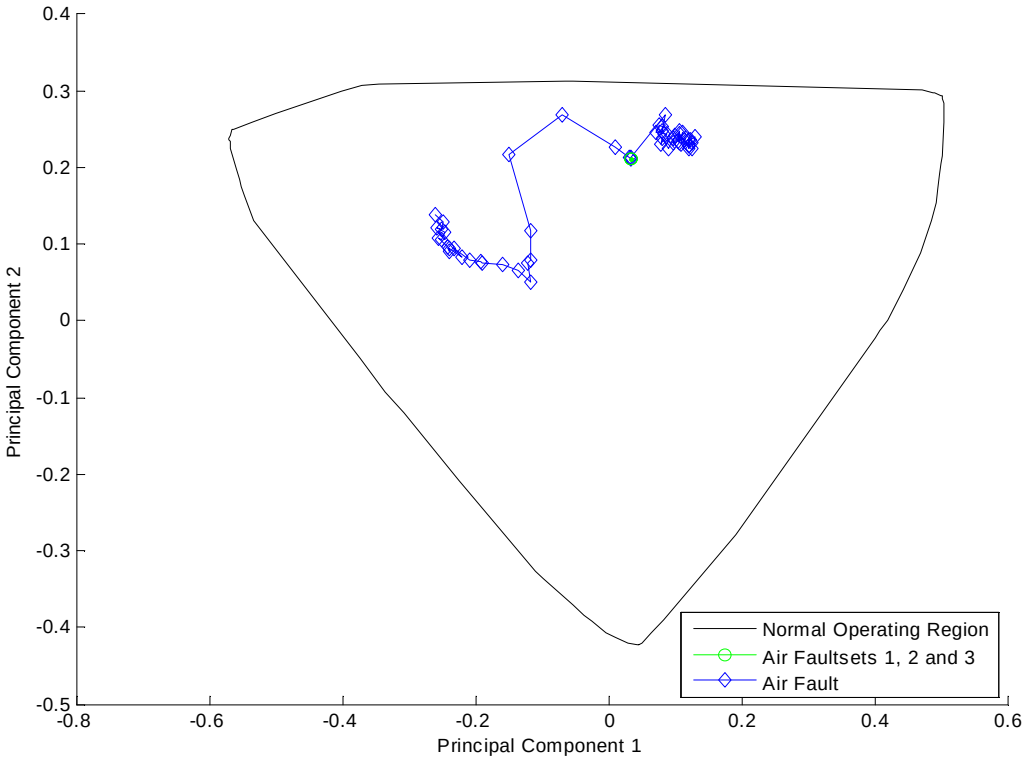

Figure 6.32: KPCA: Air Failure Fault Transition

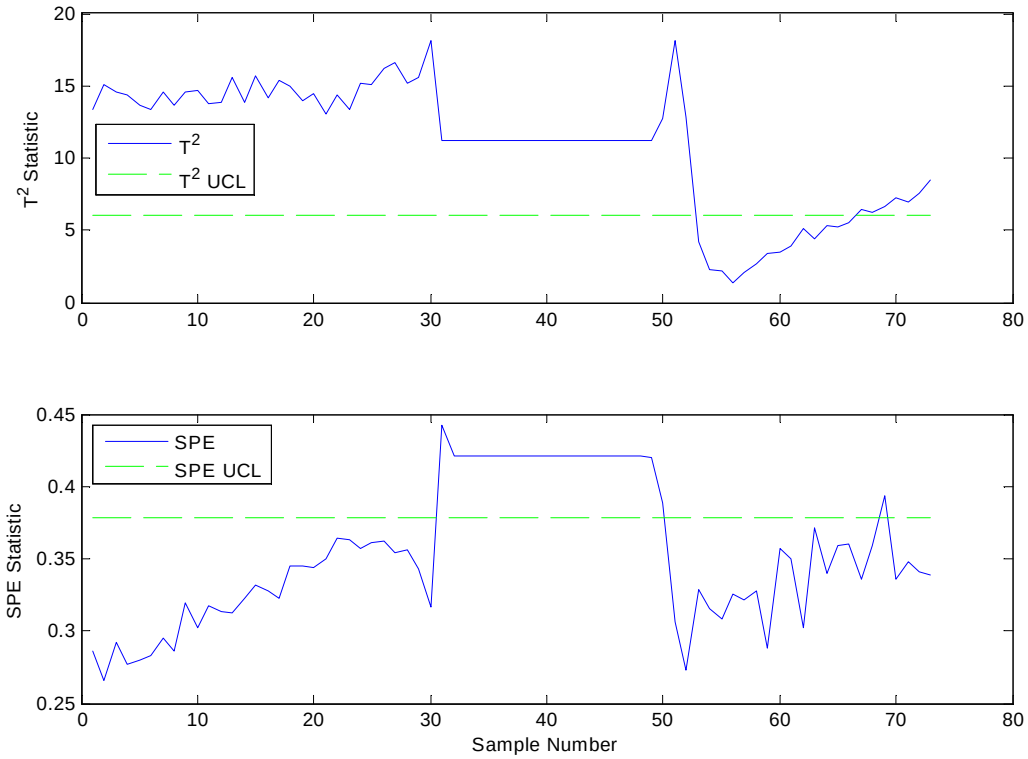

Figure 6.33: KPCA: T^2 and SPE Statistics for the Air Failure Fault

The Steam Supply Failure Fault Set

The two training sets for the steam supply failure fault form a tight region (when projected onto the biplot in figure 6.34), which coincides with the faults from the previously investigated air failure fault.

As with the air supply failure fault set, the fault testing data starts in the normal region (figure 6.35). The faulty data moves into the steam supply failure fault region. This would probably not be useful for fault detection. Unambiguous diagnosis is also not possible as the fault training region is not unique.

The T^2 and SPE for the steam supply fault is shown in figure 6.36. They are successful in detecting the fault. They both rise steadily from the normal operation values to a maximum for the faulty samples. Note that this set does not include the transition back to normal operation.

The Feed Flow Fault Set

As for the previous two fault sets, the feed flow fault training set forms a tight region when projected onto the biplot (figure 6.37). This region coincides with the fault regions of the air and steam supply faults.

All the feed fault testing data falls into the feed training fault region (figure 6.38). It

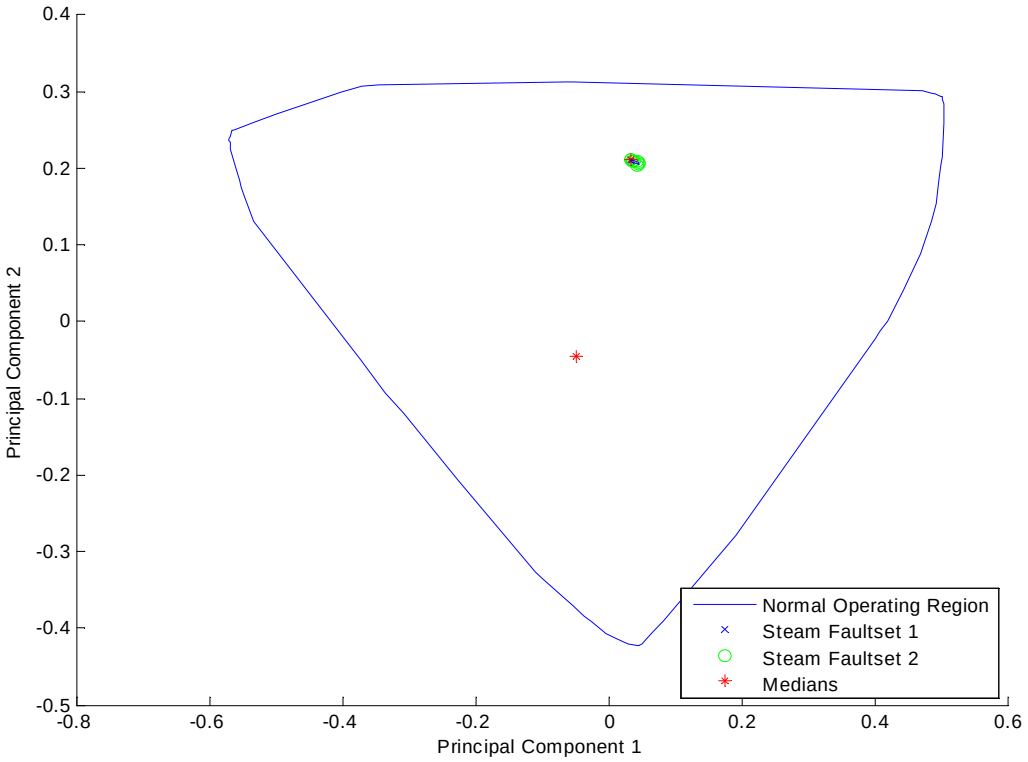

Figure 6.34: KPCA: Training Sets for the Steam Supply Failure Fault

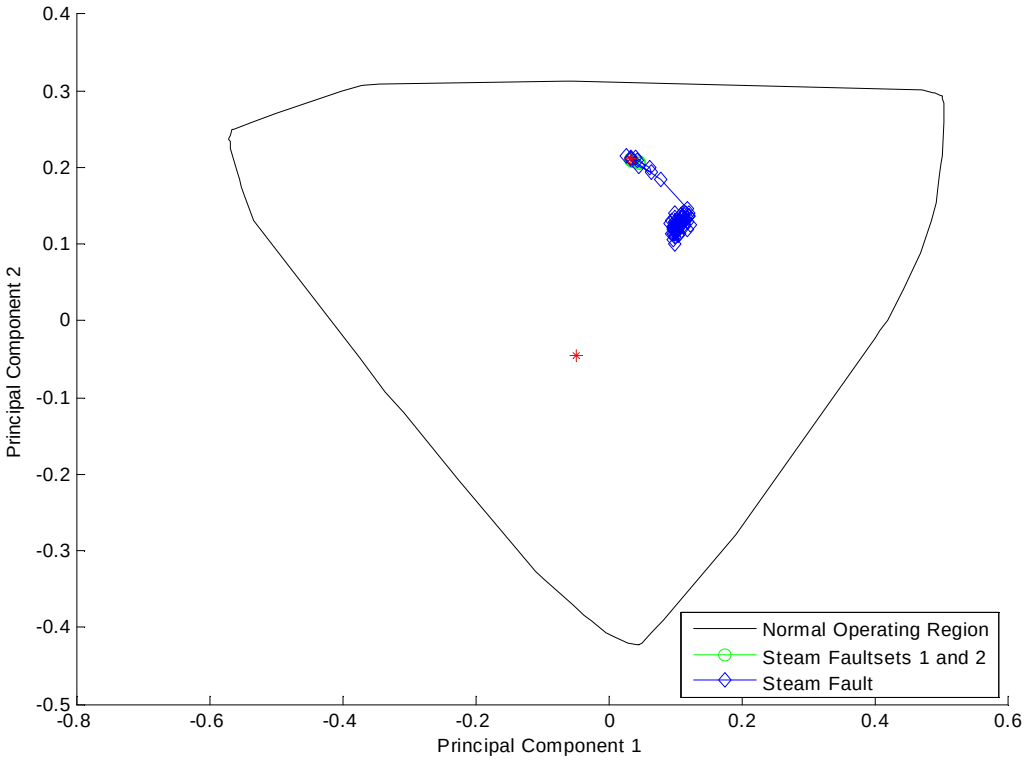

Figure 6.35: KPCA: Steam Supply Fault Transition

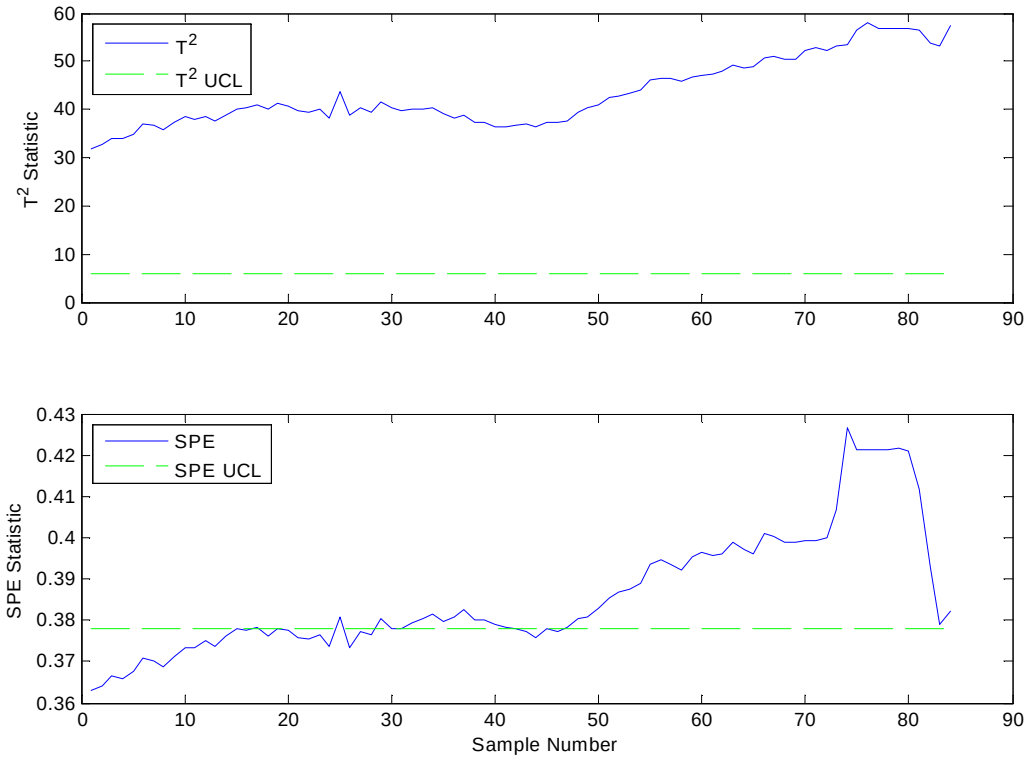

Figure 6.36: KPCA: T^2 and SPE Statistics for the Steam Supply Failure Fault

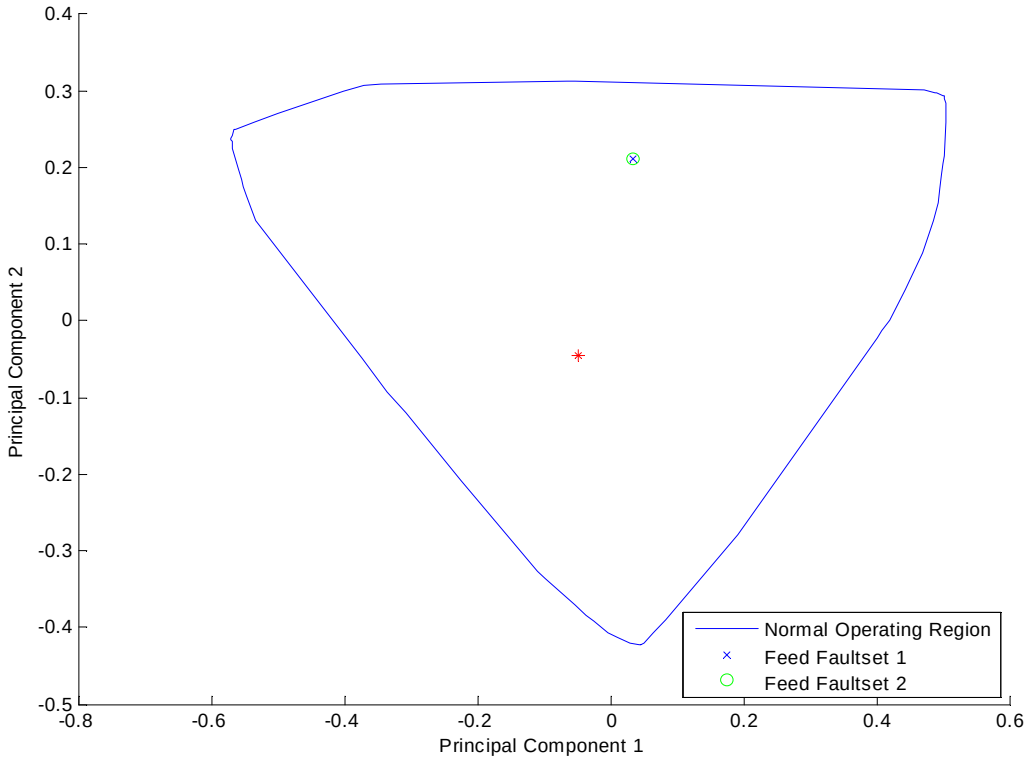

Figure 6.37: KPCA: Training Sets for the Feed Flow Fault

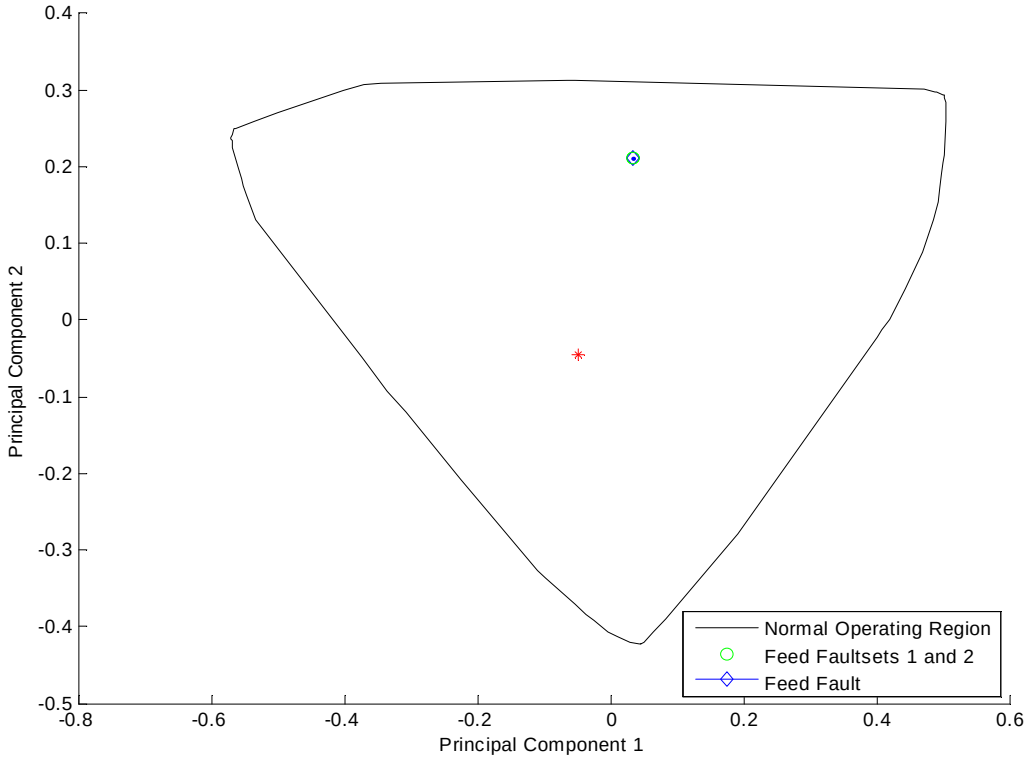

Figure 6.38: KPCA: Feed Flow Fault Transition

would not be possible to unambiguously diagnose the fault on the basis of fault region in this biplot.

Figure 6.39 shows a constant violation of the T^2 and SPE statistics. The T^2 statistic is violated by a huge margin. The T^2 plot would clearly detect this fault.

The Top Product Flow Fault Set

The top product flow fault training data is shown projected on to the biplot in figure 6.40. Interestingly, this data is not tightly grouped and does not coincide with the air, steam and feed fault regions. This training region is also within the normal operating region (similar to the PCA results for this fault and all the KPCA faults).

The fault testing data set is projected on to the biplot in figure 6.41. As with the linear PCA data, it does not fall within the training fault region. This may because of the way that the training data was generated (see section 5.2.1). It is interesting to note that this is the first fault testing set that can be detected as fault on the basis that it is outside of the normal operating region. The fault can not be diagnosed by examining which fault region the samples fall into (despite the training region being unique).

The T^2 statistic are violated for all the faulty samples (figure 6.42). The SPE statistic also rises strongly as the column cools due to the lack of steam. These statistics could

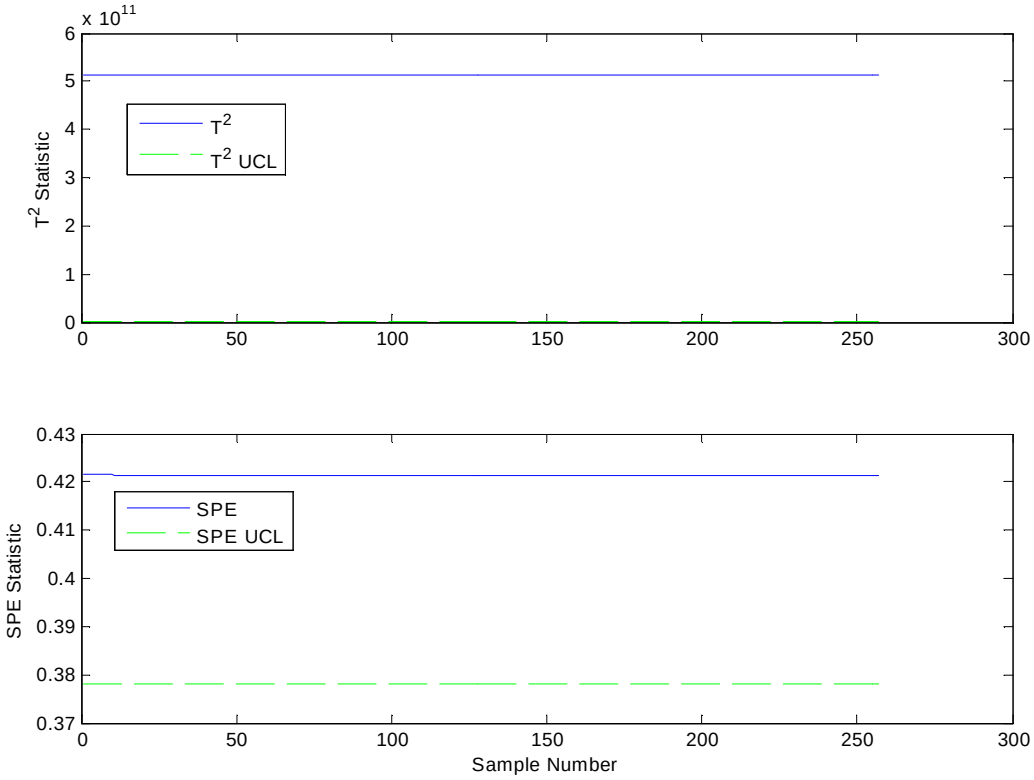

Figure 6.39: KPCA: T^2 and SPE Statistics for the Feed Flow Fault

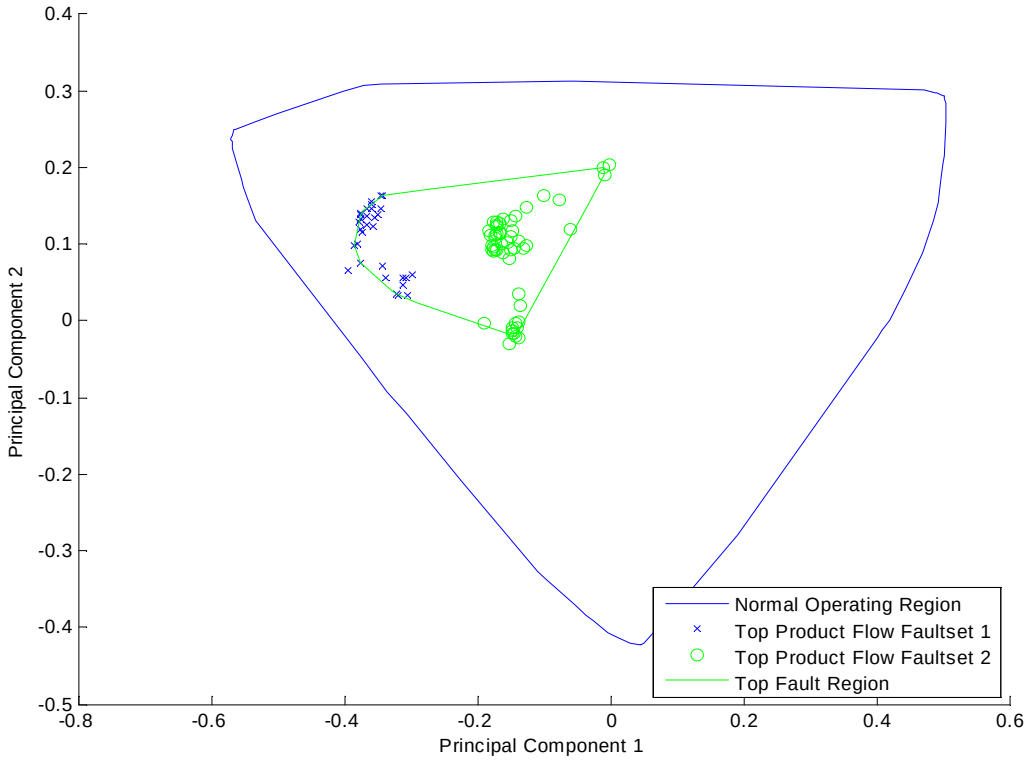

Figure 6.40: KPCA: Training Sets for the Top Product Flow Fault

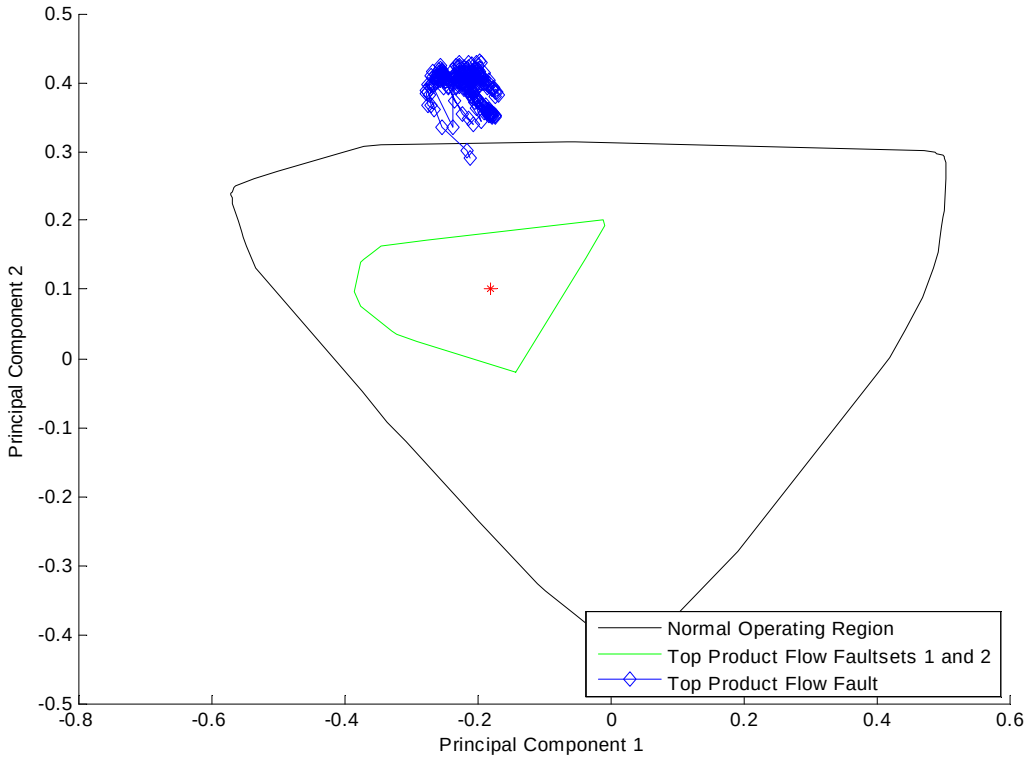

Figure 6.41: KPCA: Top Product Flow Fault Transition

be somewhat successfully applied to fault detection for this type of fault.

An overview of all the fault regions are shown in figure 6.43. Note that the air, steam and feed fault training regions are represented by circles as it is not possible to draw a bag around these regions due to conditioning problems (as the bagplot is implemented here). The air, steam, feed all coincide in a tight group. These groups are within the normal operating region. It will thus not be possible to detect a fault on the basis of samples being projected outside the normal operating region. It will also not be possible to use the fault regions on the KPCA biplot to diagnose the fault as they lie very close together. This is in contrast to the linear PCA results, where diagnosis was possible for all the fault sets excepting the top product flow fault testing set.

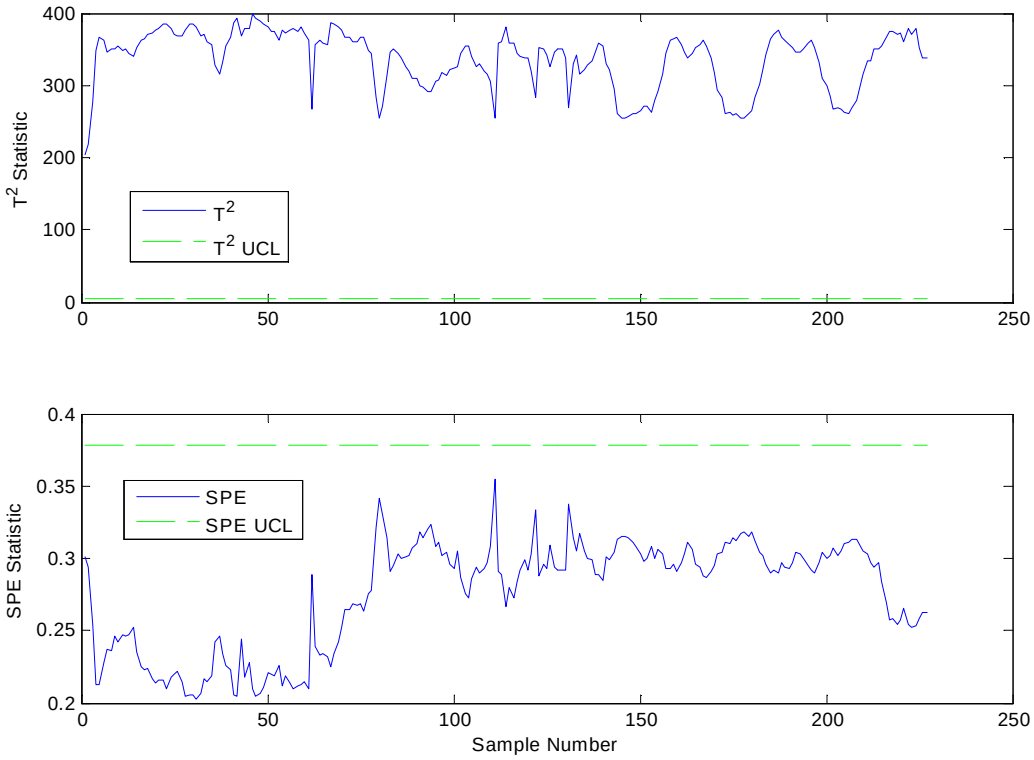

Figure 6.42: KPCA: T^2 and SPE Statistics for the Top Product Flow Fault

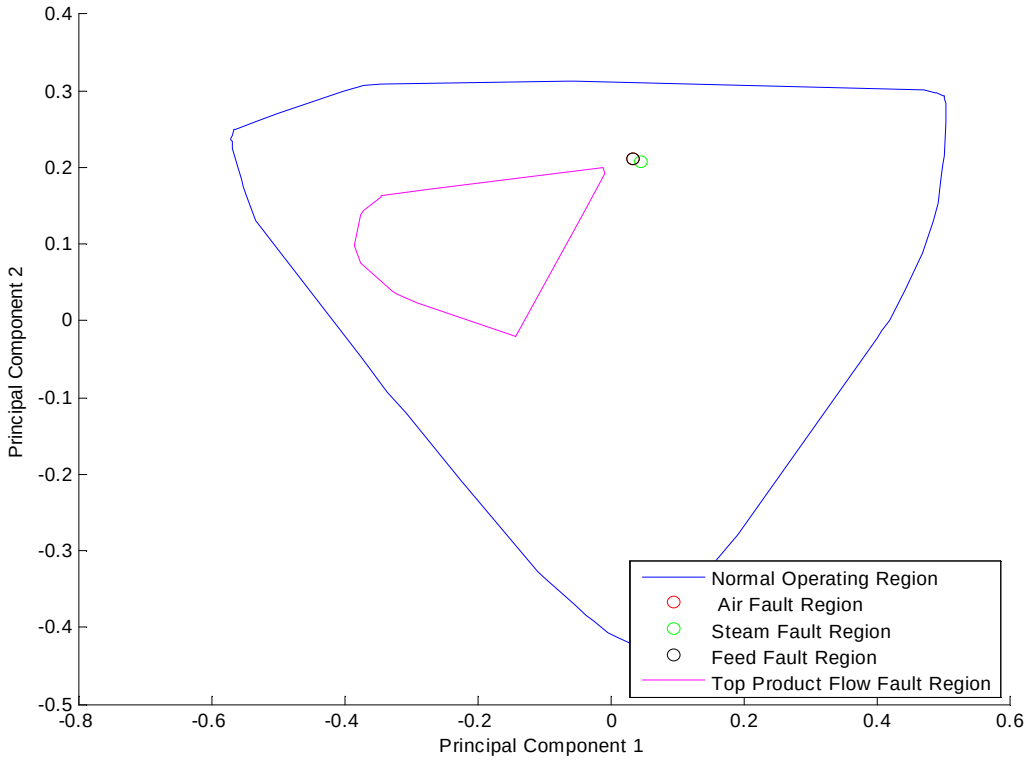

Figure 6.43: KPCA: Overview of all Fault Regions

The Novel Fault Sets

The projection of the scores for the novel data sets (discussed in section 5.2.1) is shown in figure 6.44. The faulty samples for the water rich operation, feed flooding, measurement and different operation faults fall into the exact same region as most of the training fault data (as summarised in figure 6.43). Clearly this region is useful for detecting many kinds of faults despite lying within the normal operating region. The scores for operation in low ambient conditions form a large group outside the normal operation. A KPCA biplot has been able to detect all the novel faults.

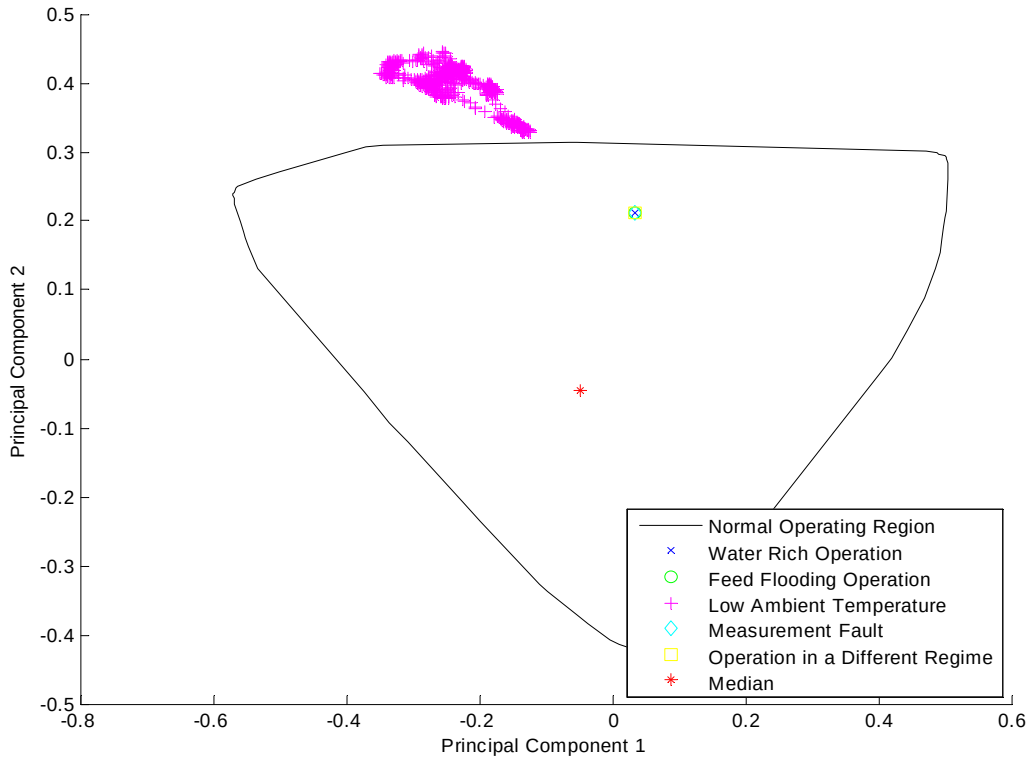

Figure 6.44: KPCA: Novel Faults Overview

The T^2 and SPE statistics are both able to detect the fault for all the novel fault sets (figure 6.45). In some cases, the T^2 and SPE limits are violated by huge margins.

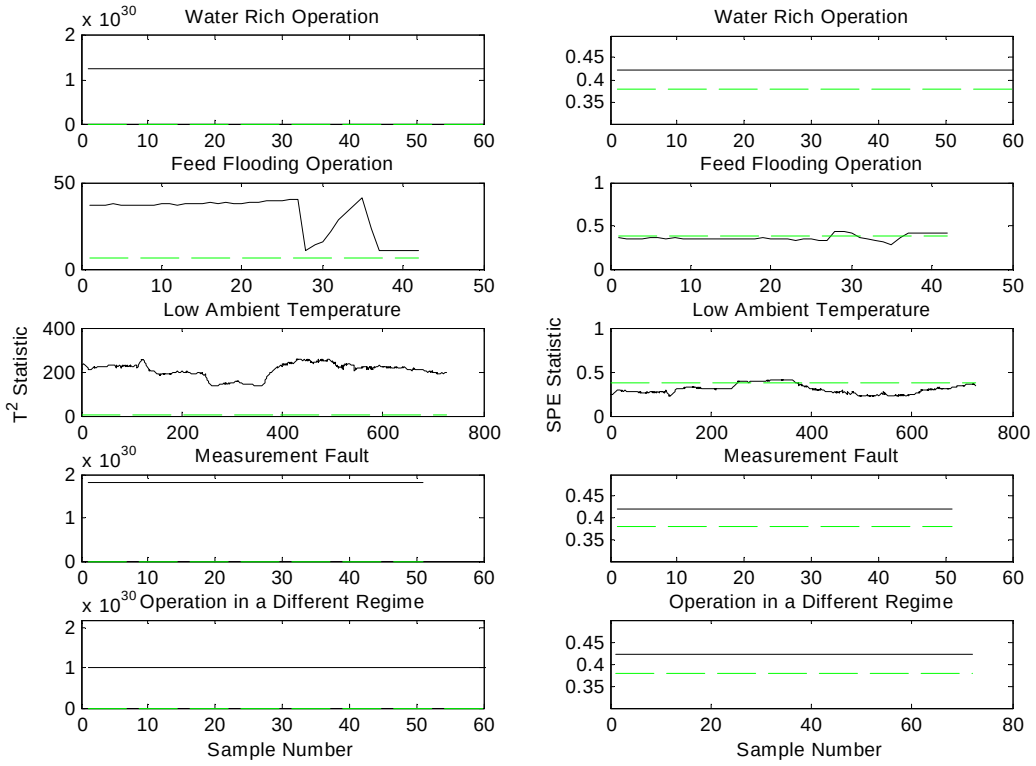

Figure 6.45: KPCA: T^2 (left hand side column) and SPE (right column) Statistics for the Novel Faults

6.3 Fault Detection and Diagnosis using Feature Classification Methods

The previous section showed results for methods that extract features from the data. Faults are recognised by the changes in these features. Faults were diagnosed by the contribution the variables had on the faulty feature or by seeing which training region the data fell into. The model's goal was to describe the features of the data. This section discusses the results for methods that create models that classify the data according to which fault set it belongs to. Features that classify (but that not necessarily describe well) the data are used to make classification rules. A new data sample can be detected as faulty if it falls into a group other than the normal operating group. Here the model's goal is not to describe the features of the data, but rather to use the data features to create a classification rule model. Faults are then only diagnosed by which class they fall into according to the classification model. Note that all the data (the normal as well as the faulty training data), tagged by class, is used to train the discriminant models

6.3.1 Fault Detection and Diagnosis using Linear Discriminant Analysis

Training of the LDA Model

The LDA model was trained using the normal operating operating region data together with the all the fault training data (separated by fault type). The model takes only a few seconds to train on a well equipped desktop computer. The contribution of the variables on each of the LDA directions is shown in figure 6.46. In contrast to the PCA model (section 6.2.1), the plate temperatures do not feature strongly. The model is influenced strongly by the feed flow rate and the feed valve (CV-01) position. It is interesting to note these two variables influence the first LDA direction in different directions. The steam valve position is also an important variable in the model.

The fault regions are shown in figure 6.47. Note how the groups are tighter than the PCA groups and separated by more distance (particularly the feed and air fault regions). The steam supply failure fault region also no longer overlaps the normal operating region as it did in the linear PCA analysis. It appears as if the feed fault training region has been broken into two more distinct groups. This may be because this faulty data was generated from different normal operating steady states. The air fault region is a tight group with the exception of two outliers. The feed flow fault region also has a distinctive extreme outlier. The origins of these outliers (which were not apparent in the PCA analysis) is not clear. The samples may be more indicative of the normal operating region than the fault, however the outlier for the feed fault region does not lie near the normal operating

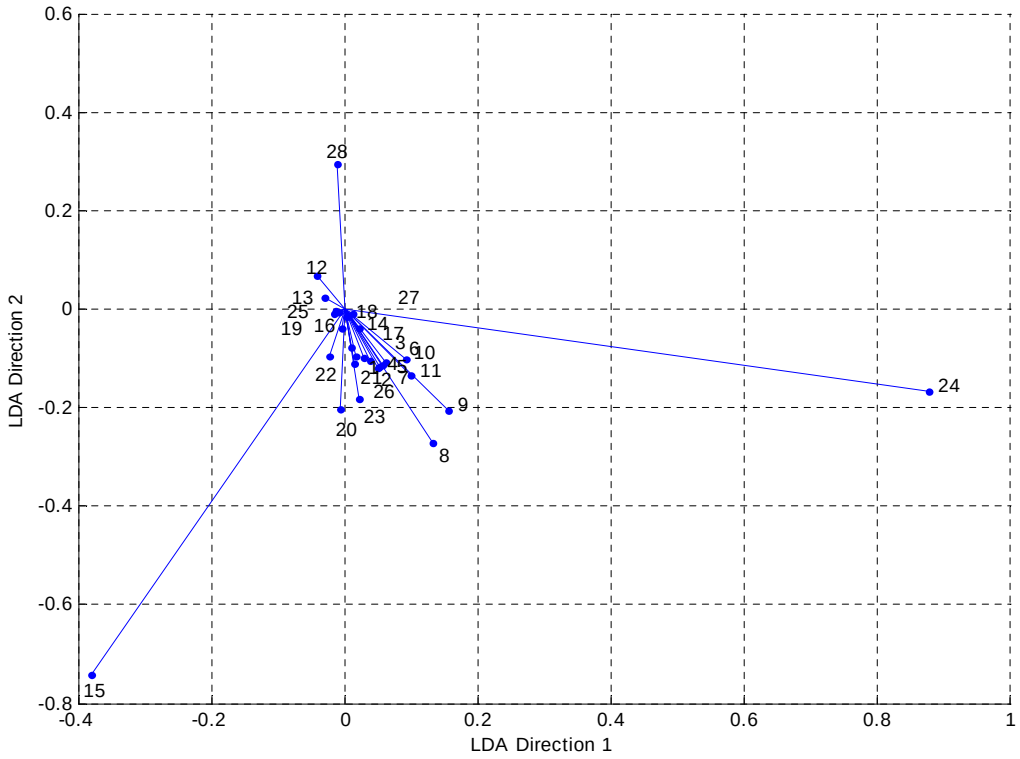

Figure 6.46: Contribution to the LDA Model

region. It is also interesting to note that the discriminant analysis (like linear PCA and KPCA) could not separate the top product flow fault training data from the normal operating data.

The Air Failure Fault Set

Figure 6.48 shows the scores for the fault set moving from the normal operating region into the the air fault bag. The data samples then return back to the normal operating region once the air supply has been restored. As with the linear PCA, one would have no difficulty detecting and diagnosing the fault by monitoring the biplot and the training regions.

The Steam Supply Failure Fault Set

Figure 6.49 shows the scores for the fault set moving from the normal operating region into the the steam fault bag. As before, one would have no difficulty detecting and diagnosing the fault by monitoring the biplot and the training regions.

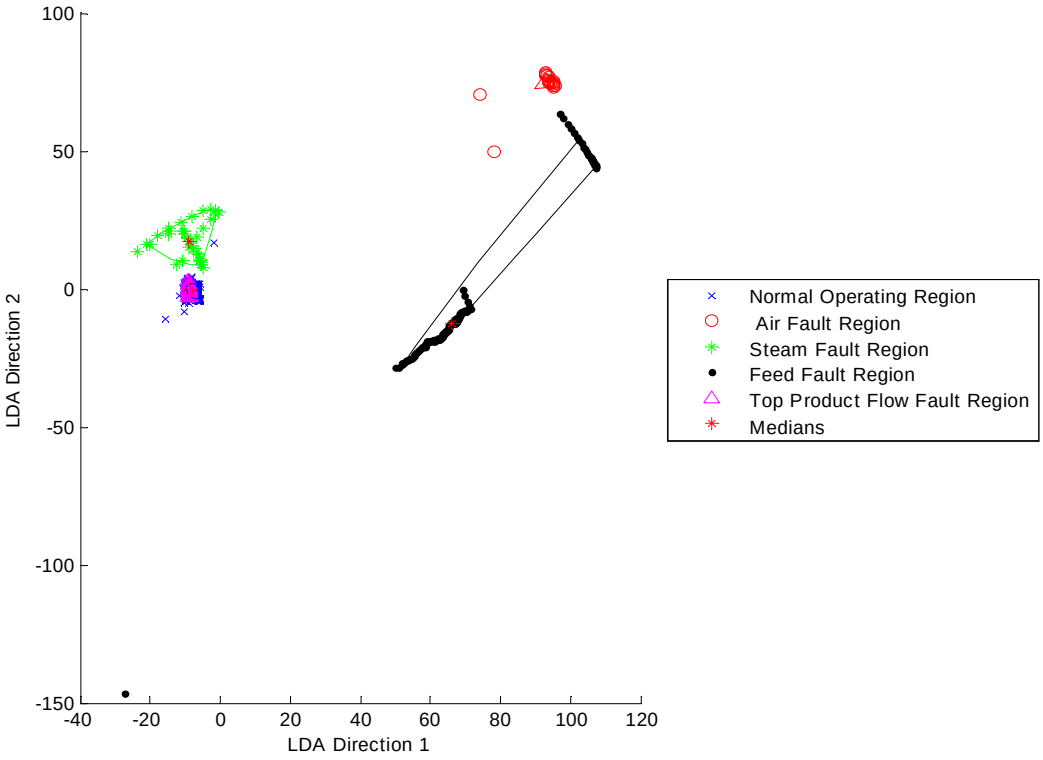

Figure 6.47: Biplot for LDA

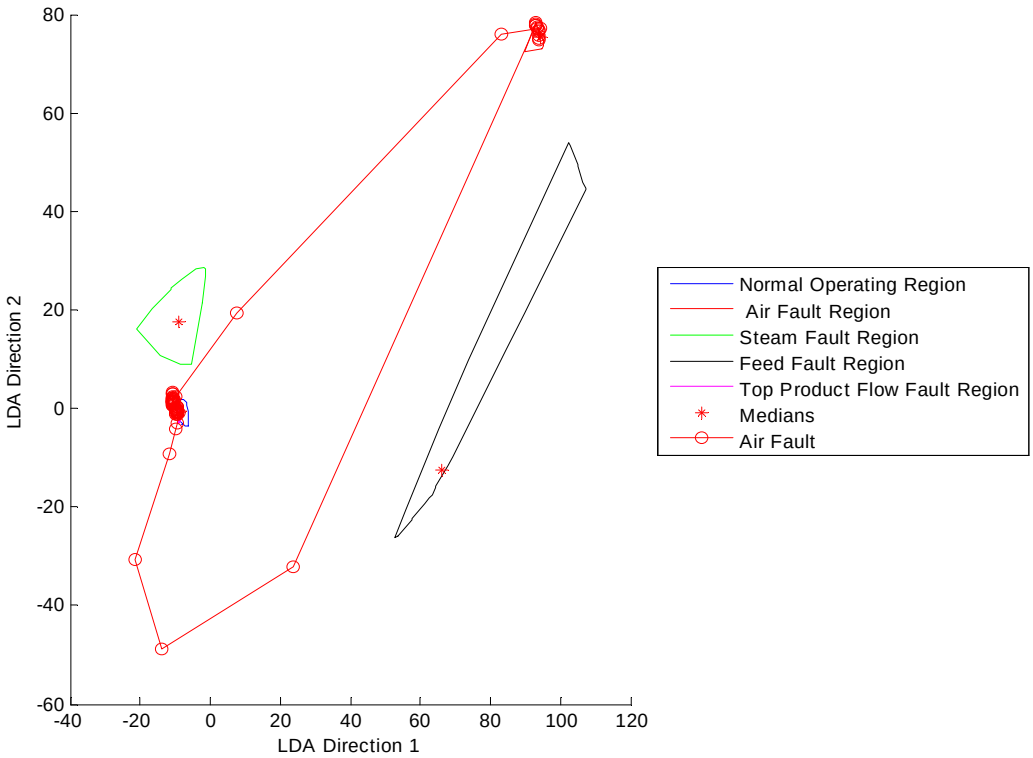

Figure 6.48: Air Supply Fault LDA Scores

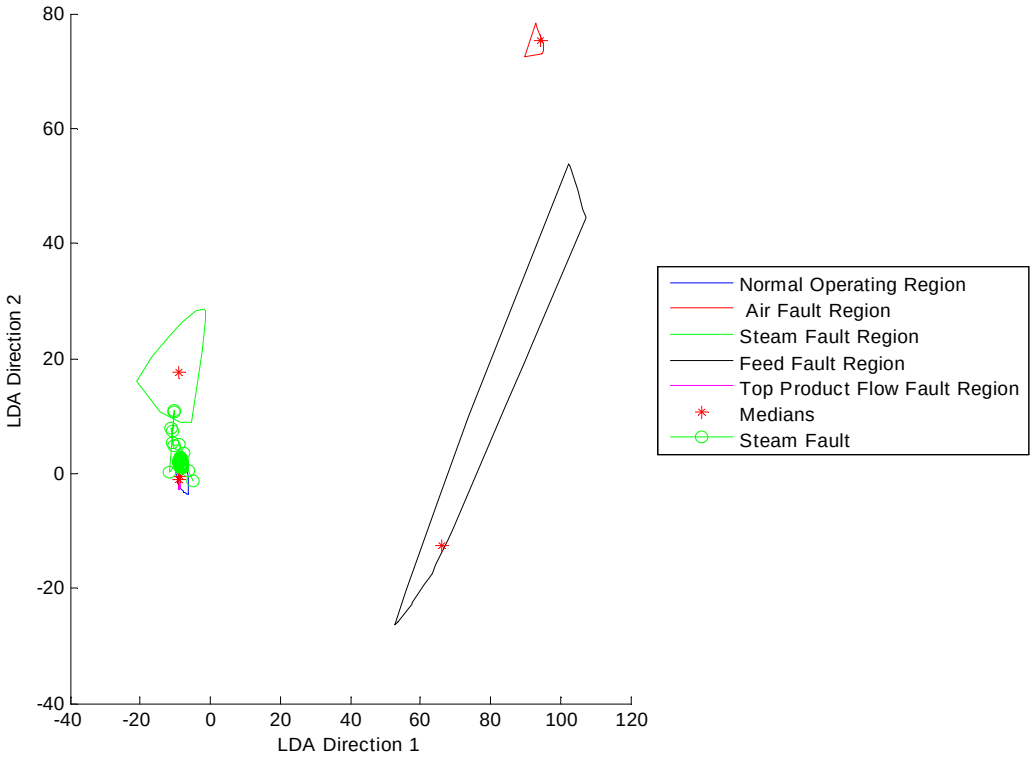

Figure 6.49: Steam Supply Fault LDA Scores

The Feed Flow Fault Set

Figure 6.50 shows the scores for the fault set moving into the the air fault bag. The process does not start or return to the trained normal operating region. The points move from below the normal operating region on the biplot. Note that this problem is not found with the PCA fault data. This shows that the LDA differs from the PCA model. Apart from the initial fault free operating point, one would have no difficulty detecting and diagnosing the fault by monitoring the biplot and the training regions.

The Top Product Flow Fault Set

Figure 6.51 shows the scores for the fault set. The data for the fault is still close to the normal operating region and the top product flow bag. It would be impossible to confidently detect or to diagnose the fault using LDA.

Note that it is not useful to discuss the results for the novel data sets here. The novel faults cannot be classified into the existing fault regions by LDA. The LDA analysis merely confirms that the data differs from the normal operating region. In this way it is useful for detection only.

The LDA has only improved the separation between the groups (compared to PCA) slightly. The improved groups were already distinct and fairly well separated in the

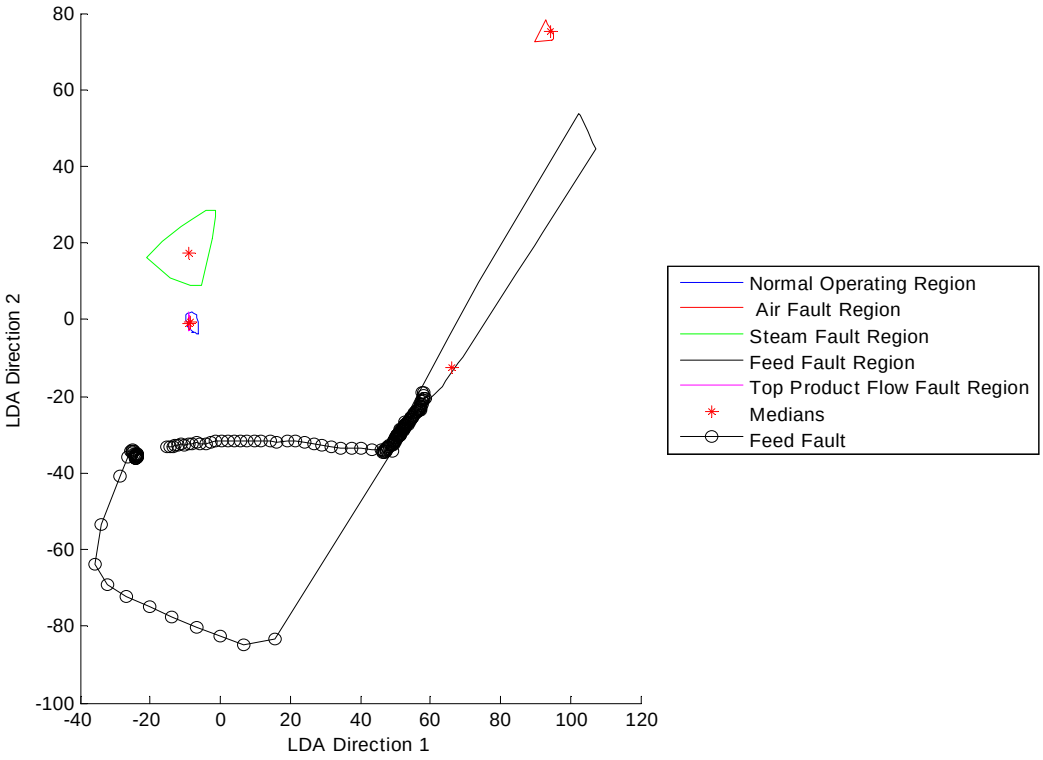

Figure 6.50: Feed Flow Fault LDA Scores

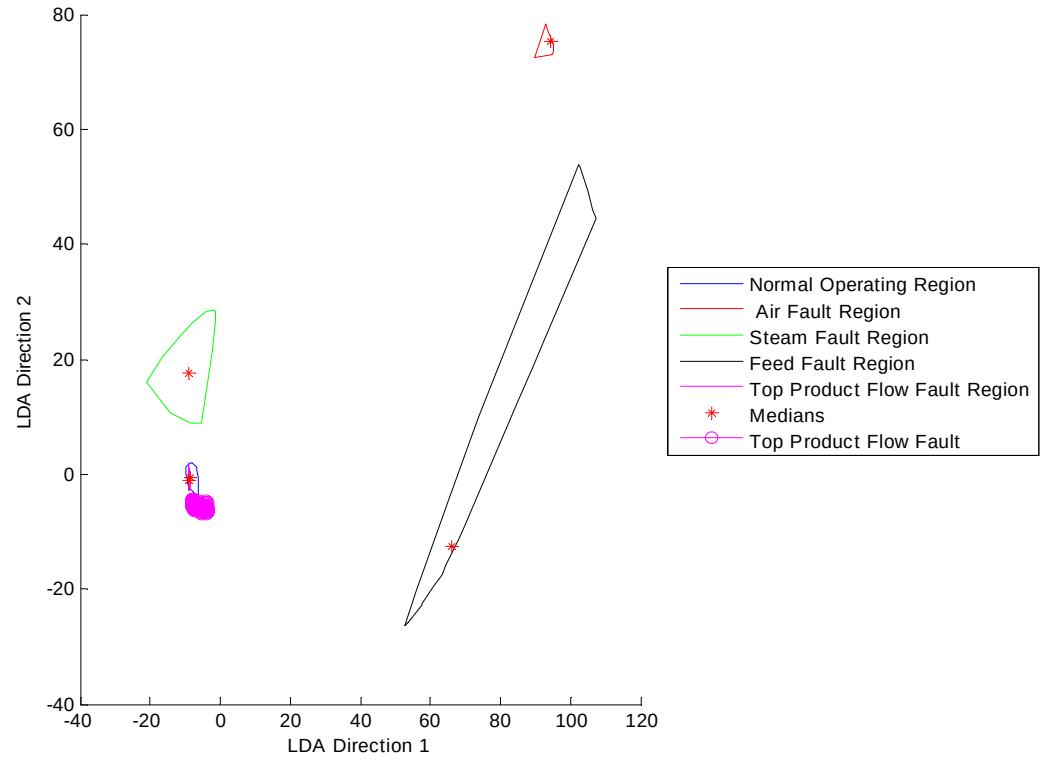

Figure 6.51: Top Product Flow Fault LDA Scores

biplots. This means that it should be easy to isolate the fault. The increase in separation between some of the groups has come at the cost of not having the PCA statistics and contribution plots (which proved useful for fault detection and diagnosis).

6.3.2 Fault Detection and Diagnosis using Kernel Discriminant Analysis

The results for KDA, the nonlinear kernel based extension to LDA, are discussed here.

Training of the KDA Model

The KDA model was trained using the same kernel argument or parameter (σ in equation 3.62) as used for the KCPA model. The model took about 8 minutes to train on a well equipped desktop computer. As with LDA, the KDA model was trained with the normal operating data as well as the faulty data. All these groups were divided up and tagged by group number for the algorithm. The projection of the scores, once the model had been trained, is trivially quick. As with the LDA model, the dimensionality was chosen to be two. Note that it is not possible to calculate the contribution that each of the variables make to the model.

In figure 6.52, we can see that the KDA gives us groups that are even tighter and better separated than any of the previous analysis results. This means that the isolation of the faults is improved. In contrast to LDA, the feed fault region appears as one group with no outliers. Similarly to the PCA and LDA analysis results, the top product flow region lies close to the normal operating region.

With a different choice of kernel argument, it is possible to get even better group separation. An example, using an kernel argument of 0.1 is shown in figure 6.53. The separation is even better, with the data samples of each group (consisting of many points differing in the input space) coinciding on a single point in each case. The group are so tightly distributed that the bagplot algorithm (as implemented) encounters conditioning problems when trying to bag these groups. The problem with this degree of separation is that the data is over fitted. The complex model in feature space has managed to find a complex enough relationship that each individual point has been 'specially' classified into a group. With unseen (non-training) data, this model will perform very poorly. The model will not be robust. Over fitting is already a concern with the more relaxed parameters used here.

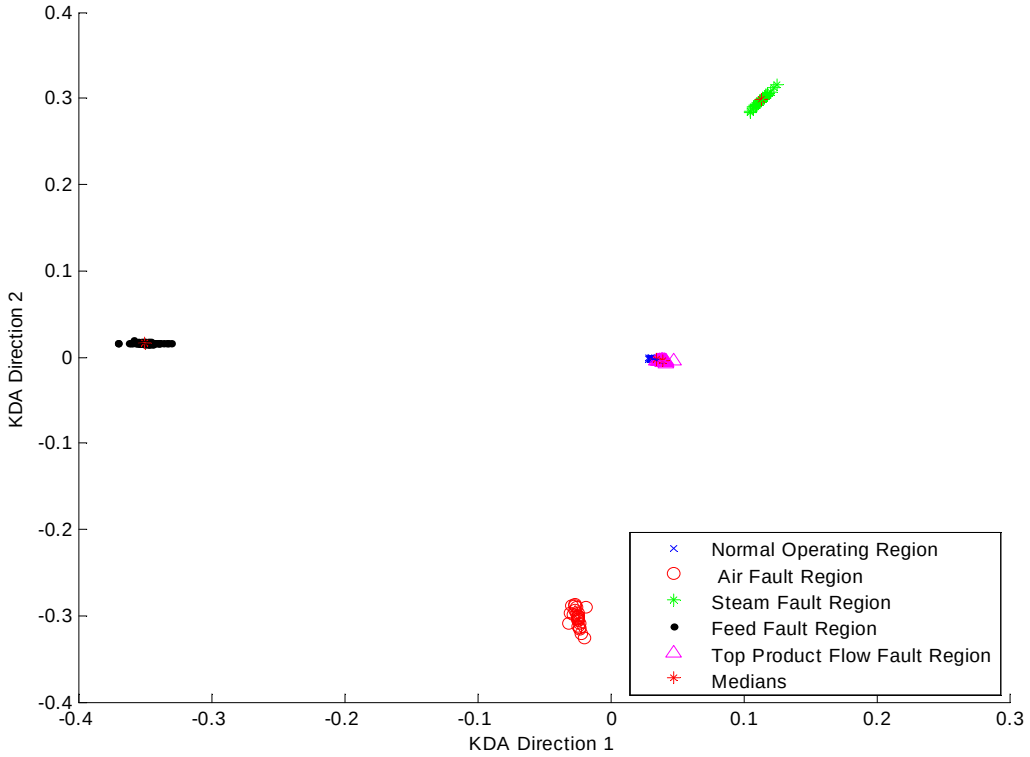

Figure 6.52: Biplot for KDA

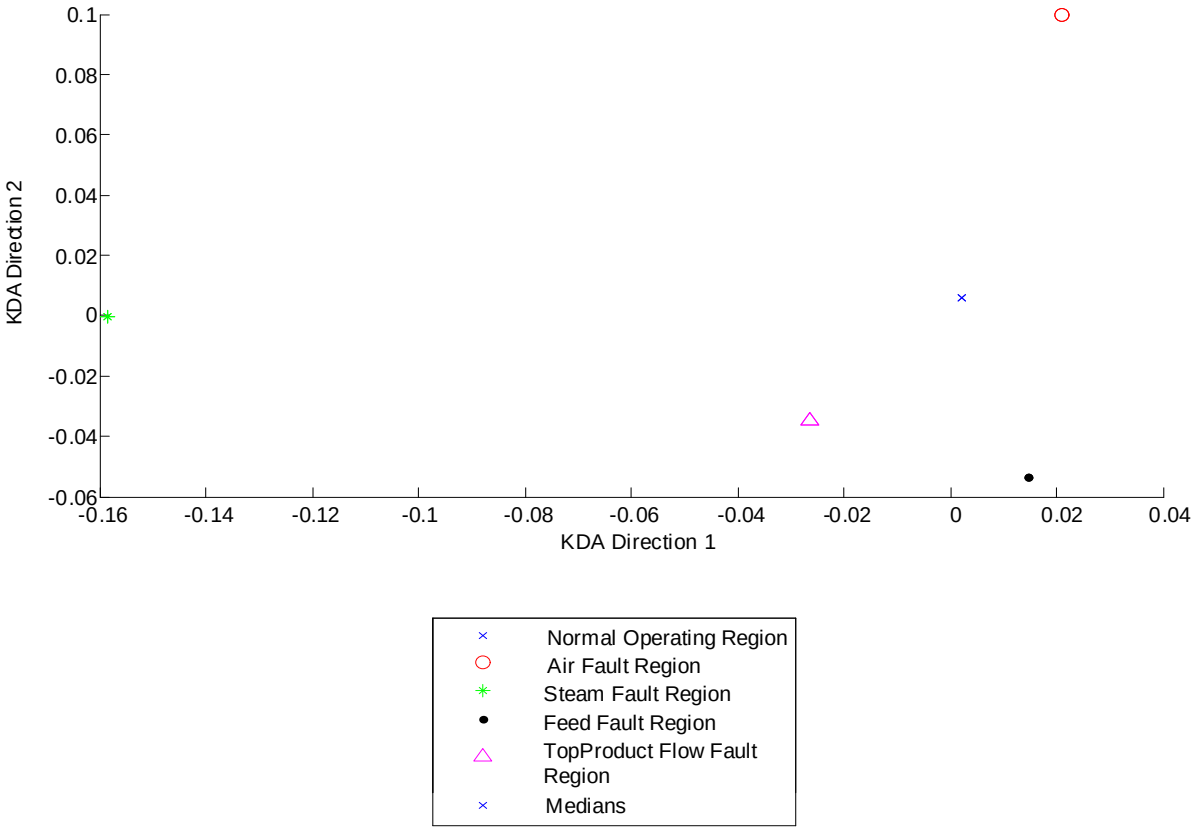

Figure 6.53: Overfitted KDA Data

The Air Failure Fault Set

Figure 6.54 shows the scores for the fault set moving from the normal operating region into the the air fault bag. The data samples then return back to the normal operating region once the air supply has been restored. As with the linear PCA, one would have no difficulty detecting and diagnosing the fault by monitoring the biplot and the training regions.

The Steam Supply Failure Fault Set

Figure 6.55 shows the scores for the steam supply failure fault set moving from an area close to the normal operating region slightly towards the steam fault bag. The data never comes close to the steam fault region. Here, KDA was unsuccessful in detecting or diagnosing the steam supply failure fault. This is in contrast to the LDA results, where the fault was detected and diagnosed successfully.

The Feed Flow Fault Set

Figure 6.56 shows the scores for the feed flow fault set moving from an area close to the normal operating region, strongly towards the steam fault bag. The data approaches but

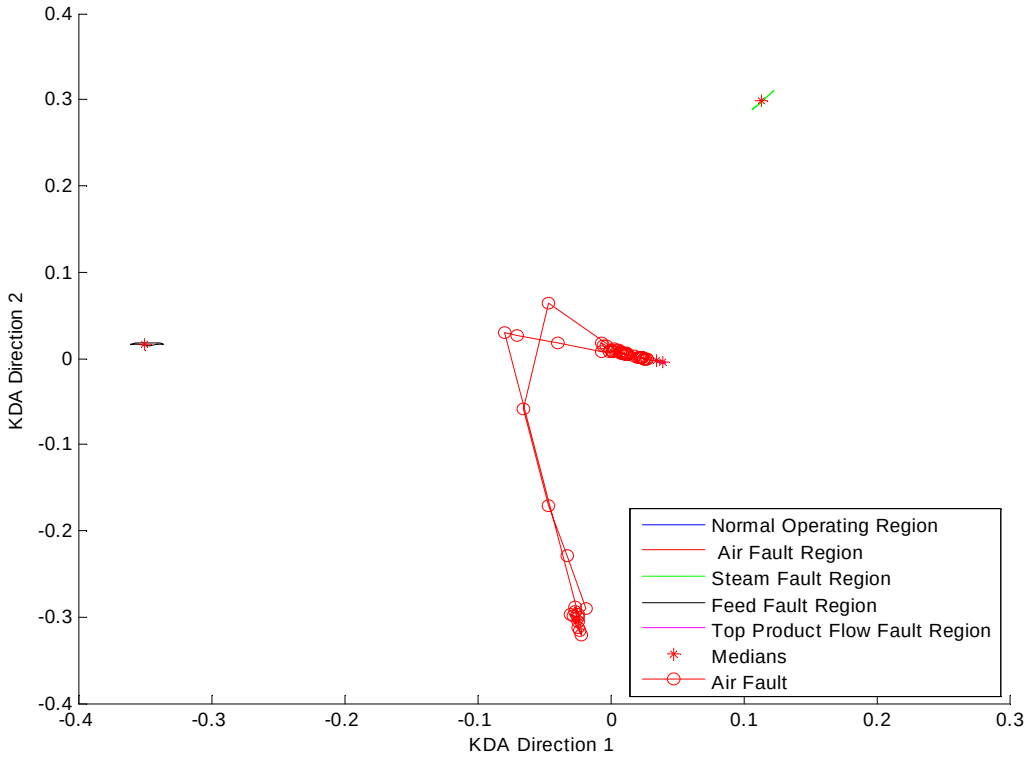

Figure 6.54: Air Failure Fault KDA Scores

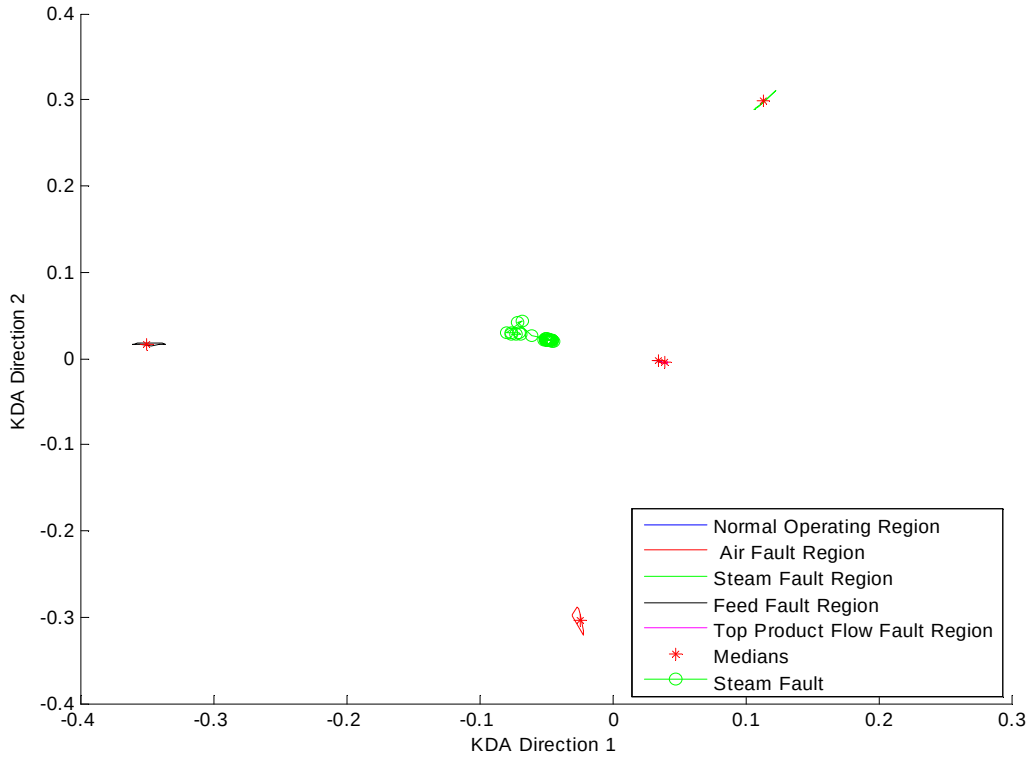

Figure 6.55: Steam Supply Failure Fault KDA Scores

does not enter the feed flow fault region. The KDA results also showed the data starting from a point significantly outside the normal operating region. In contrast, the LDA results show the data eventually entering the feed flow fault bag as the fault progressed. With some interpretation, these results could be used to successfully detect and diagnose the fault.

The Top Product Flow Fault Set

Figure 6.57 shows the scores for the top product flow fault set moving from an area fairly close to the normal operating region. The data moves in a direction that is the same as the direction of the top product flow fault bag relative to the normal operating region. Here it would be difficult (as with the other analysis results for this fault) to detect or diagnose the fault successfully.

The KDA gave more clear and tightly distributed groups compared to any of the other techniques demonstrated here. The model took significantly longer than the LDA or PCA model to calculate. The statistics and contribution plots that made PCA a successful technique are not easily defined for this method. One also loses any interpretation of the model relative to the input space (which the LDA model retained). There is also evidence of over fitting in the failure of all the fault training data to generalise to the (unseen)

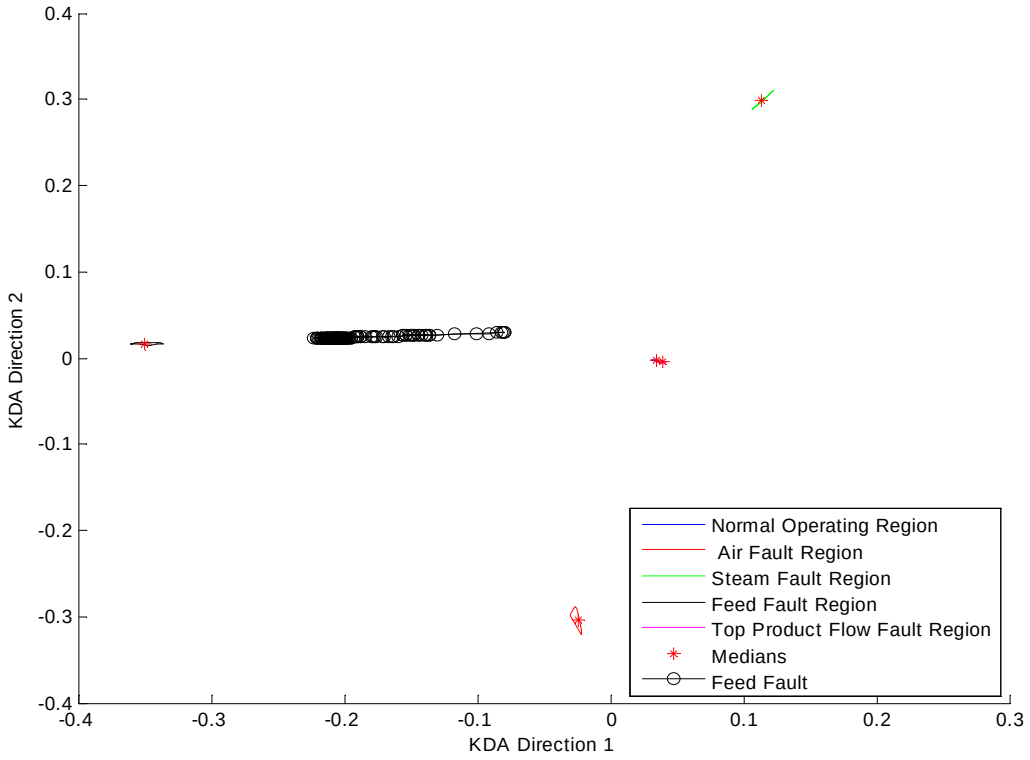

Figure 6.56: Feed Flow Fault KDA Scores

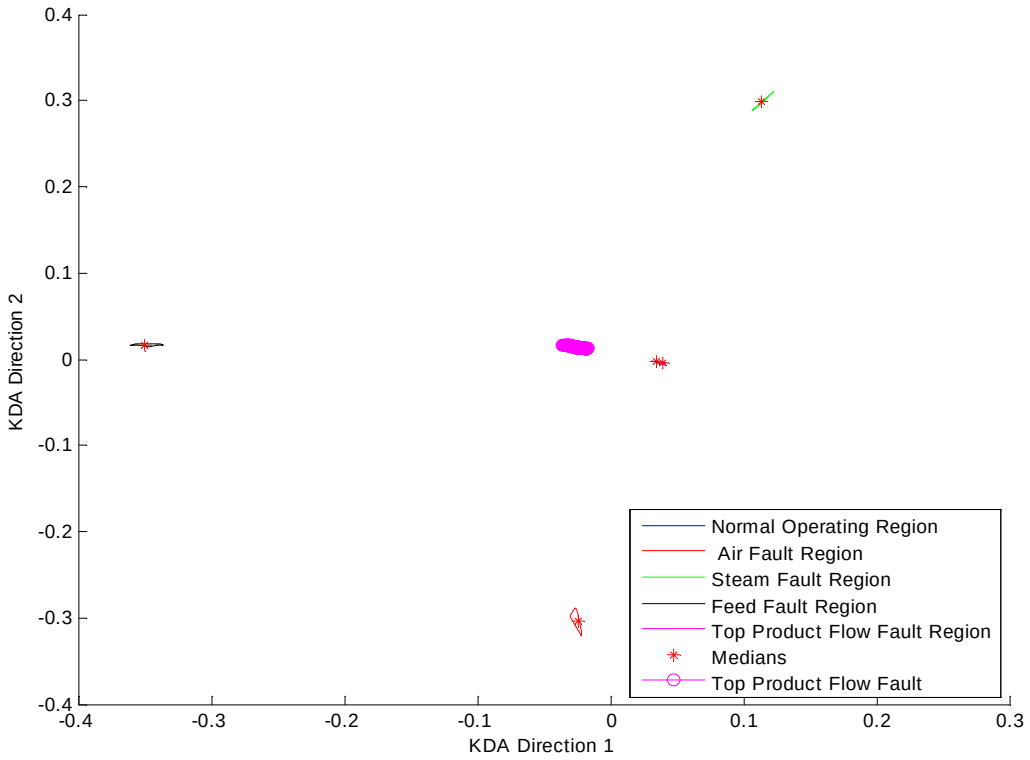

Figure 6.57: Top Product Flow Fault KDA Scores

data used to test the faults.

AS with LDA, it is not useful to discuss the results for the project of the novel fault set data onto the KDA biplot.

CHAPTER 7

Conclusions and Recommendations

The fault detection and diagnosis abilities of the following techniques on distillation column data was evaluated:

1. *Linear principal component analysis:* Fault detection by means of biplots, T^2 and SPE statistics. Fault diagnosis by means of fault regions and contribution plots.

2. *Kernel principal component analysis:* Fault detection by means of biplots and modified T^2 and SPE statistics. Fault diagnosis by means of fault regions.

3. *Linear discriminant analysis:* Fault detection by biplots. Fault diagnosis by classification into biplot regions.

4. *Kernel discriminant analysis:* Fault detection by biplots. Fault diagnosis by classification into biplot regions.

7.1 Performance of the Fault Detection and Diagnosis Techniques

7.1.1 Attaining the Goals of SPC

The goals of fault detection and diagnosis by means of SPC were discussed in section 2.3.

1. *In the process in control?*: All of the techniques (some more successfully than others) are able to provide a answer to whether the process is faulty or not.

2. *Specify a classification error estimate*: The ability to specify a classification error estimate (probability of incorrect flagging of non-faulty data) is possible with

these techniques. The distance to the normal operating region could be calculated and related to a probability. This will be difficult to calculate using the kernel based methods. This goal was not implemented here as the distance (and therefore probability of a type 1 error) can be judged visually.

3. *Account for the relationships between the variables*: All of the techniques take into account the relationships between the variables. The techniques have all exploited the correlation in the variables to reduce the dimensionality from 28 variables to 2 dimensional biplots. The T^2 and SPE statistics further reduce the features of the data to a single dimension.

4. *Fault Diagnosis*: All of the methods have the ability to diagnose faults. The techniques, excepting for KPCA, show some success in diagnosing the faults shown here.

Here, a single branch of SPC techniques have the ability to satisfy all the goals of SPC.

7.1.2 Desirable Attributes of the Techniques

The desirable attributes of a fault detection and diagnosis system were discussed in section 1.3. In summary, the statistical methods discussed in this dissertation have the following desirable attributes:

1. *Real time detection and diagnosis:* All of the techniques are easily capable of projecting scores of new data samples in real time. Contributions and T^2 and SPE statistics are also possible to implement in real time.

2. *Diagnosis in addition to detection:* All of the techniques have the ability to diagnose faults. KPCA did not perform well on the test fault data sets.

3. *Fault isolation:* The idea of using the discriminant methods is to improve the isolation between the different faults. This was investigated. PCA shows good isolation without the use of a discriminant technique to improve the isolation. KDA and LDA show the decrease in rejection of model uncertainty due to the the high degree of isolation for the training data.

4. *Completeness:* The contribution plots are (by definition) complete as they are open to interpretation. Using the trained fault images on the biplot is not likely to be complete (unless training data for every possible fault is found).

5. *Early Detection and Diagnosis:* All of the techniques have the ability to track the movement of a score from the normal operating region to another region in real

time. It is likely that the fault can be detected (and possibly diagnosed) before the effects of the fault are significant.

6. *Robustness and supervision of processes in transient states:* The techniques handle dynamic data well. The kernel based methods are not as robust as linear PCA due to possible over-fitting of the data.

7. *Novelty identification:* PCA and KPCA have the ability to detect novel faults directly. PCA successfully diagnosed the novelty fault sets presented here. The LDA and KDA methods will not be able to diagnose the fault as they require faulty training data. Detection may be possible by comparing the data to the normal operating data.

8. *Classification error estimate:* As discussed in the previous section.

9. *Multiple fault identifiability:* All of the techniques have demonstrated their ability to identify multiple faults. Their ability to detect multiple faults occurring simultaneously must still be tested.

10. *Adaptability:* Additional fault regions can be added (or the existing ones refined) if more information becomes available. The models will need to be re-trained (PCA and KPCA only need re-training if the normal operating operating data is extended).

11. *Reasonable modelling requirements:* Normal operating data can be sourced from everyday normal operation (provided the variables have been excited enough to represent the entire region). Fault training data can be sourced from incidental faults or from faults that were initiated with the intention to gather data. It may also be possible to use a model to create training data that simulates a fault.

12. *Reasonable storage and computational requirements:* Training the models was fairly quick. It will be possible to use the trained model online. Calculating the contribution to the kernel based methods was difficult or impossible.

13. *Detection of faults in closed loops:* The techniques all managed to detect faults in the presence of closed loops.

14. *Diagnosis of faults in actuators, sensors and other process components:* The techniques have managed to detect faults in actuators, sensors and other process components.

7.1.3 Performance on Faulty Data Sets

The data that these techniques are applied to is nonlinear, non-normal, multidimensional and correlated. The normal operating data appeared to be consistent for different operating points. This data was used to train the PCA and KPCA models. A two dimensional output space was chosen (despite this being lower than most guidelines recommended). A kernel argument of $\sigma = 8$ for KPCA (and KDA) was used. It was difficult to select an appropriate value and the chosen value may not be optimum. Fault regions were created on the biplots using training fault data. The fault detection ability was then tested using faults with the same underlying cause as the training data as well as novel faults. Fault detection for PCA and KPCA was done by means of monitoring the biplots and by means of the T^2 and SPE statistics. Fault diagnosis was done by monitoring which fault region the faulty scores moved into. PCA also allowed the calculation of contribution plots to show what variables were contributing to the faulty score.

The normal operating data combined with the fault data training sets were used to train the LDA and KDA models. Fault detection using the classification techniques made use of biplots. The position of the scores of a new data point was compared to that of the normal operating region. Fault diagnosis was done by checking which region the faulty score approached.

Table 7.1 summarises the results for the fault detection. The T^2 and SPE statistics were regarded as successful at fault detection if the statistics violated the control limit or showed a marked increase (if the statistic was already higher than the limit) for the faulty data samples. A biplot was regarded as successful at fault detection if the faulty scores moved away from the normal operating region. Note that the KPCA biplots were considered successful at fault detection because the faulty data moved into the small unified fault region (despite this region being contained within the normal operating region). Otherwise, the KPCA method would be considered unsuccessful at fault detection. The PCA and KPCA techniques were successful at detecting all faults presented here. The LDA and KDA techniques are intended more for increasing fault isolation to aid fault diagnosis.

The top product flow fault proved problematic as the training data did not represent the actual fault well. For all the methods, the training region was close to the normal operating region. This was true even for the discriminatory methods which increased isolation for the other faults. KDA could be used to separate the top product flow and the normal operating region. This came at a cost of strong over-fitting for the kernel arguments that were tested. When tested with the top flow testing set, the scores did not lie close to the training region (except for LDA). This shows that the training data is maybe a poor representation of the real fault. Also, it is possible that the training fault did not have enough time to affect the rest of the column as the testing data did.

KPCA managed to detect all the faults, not by observing the deviation of the scores from the normal operating region, but rather by noting the presence of a small region where all the faulty data (excepting the top product flow fault) falls. This area is interesting as it is within the normal operating region. The region is not necessarily a result of over-fitting, as unseen faulty and novel data also fell into this region. Normal operating data (from the beginnings of the transition to faulty operation) also fell outside this area. This region should also be further tested with unseen normal operating data. Linear PCA gave convincing results for all fault detection sets. The results of LDA and KDA were not as convincing.

The T^2 and SPE were also useful for detecting faults. While similar statistics can be calculated for KDA and LDA, they are not as directly meaningful. The SPE statistic for KPCA is only an approximation (due to the preimage problem) given by Lee et al. (2004a). For this reason, the statistics of KPCA are not as meaningful as the statistics of PCA.

Table 7.2 summarises the result for the fault diagnosis. The biplot method was considered a success if the scores moved strongly towards the correct region (e.g the KDA on the feed flow fault). PCA and LDA were successful in diagnosing all the faults for which training data was available. PCA additionally provided useful information regarding the diagnosis of the novel faults. Again the top product flow fault proved problematic for all techniques. Despite lacking the increased isolation of the discrimination methods, PCA was able to diagnose faults successfully.

Theoretically, KPCA and KDA should be able to at least match their linear counterparts with correct kernel argument choices. They should also be able to manage nonlinear relationships. In these results, we see that, with the increasing complexity of the methods, no useful improvement in fault detection or diagnosis was observed. In fact, the nonlinear kernel based methods generally performed worse than the simple linear techniques. With the kernel-based methods, useful attributes like the ability to quickly and

Table 7.1: Summary of Fault Detection Abilities

		Fault Sets				
Technique	Method	Air Supply	Steam Failure	Feed Flow	Top Flow	Novel
PCA	Biplots	√	√	√	√	√
	Statistics	√	√	√	√	√
KPCA	Biplot	√	√	√	√	√
	Statistics	√	√	√	√	√
LDA	Biplot	√	√	√	×	n.a.
KDA	Biplot	√	×	√	×	n.a.

Table 7.2: Summary of Fault Diagnosis Abilities

		Fault Sets				
Technique	Method	Air Supply	Steam Supply	Feed Flow	Top Flow	Novel
PCA	Biplots	√	√	√	×	n.a
	Contribution Plots	√	×	√	×	√
KPCA	Biplot	×	×	×	×	n.a
LDA	Biplot	√	√	√	×	n.a.
KDA	Biplot	√	×	√	×	n.a.

accurately relate the data in the PCA or LDA space back to the input space are lost. We see that the simple statistics and contributions plots of PCA proved very useful.

7.2 Recommendations

PCA and its kernel based derivatives are useful in creating models from multivariate data. The techniques can effectively create models which operate in low dimensional to aid calculation, visualisation and interpretation. Only operating data, and not in-depth knowledge of the process is required to create these models. Multivariate statistical techniques can be recommended for detecting and diagnosing faults on multivariate dynamic chemical processes.

PCA is simple and fast to train, use and interpret and is successful at detection and diagnosing faults. It is easy to relate data in the PCA space back to the input space. The technique appears to be robust in the face of data that is non-normal and nonlinear. PCA is probably useful in many cases except when the extent of nonlinearity is so severe that a nonlinear model is a necessity. If fault isolation with PCA proves difficult, a discriminatory method such as KDA or LDA can be considered for further analysis of the problem.

7.2.1 Future Work

To improve and extend on this work in fault detection and diagnosis, the following avenues are open for exploration:

1. Investigation of the effect of the choice of the kernel argument. The determination

of a method to select the optimum kernel argument based on the training data is needed to evaluate the kernel techniques more thoroughly.

2. Investigation of a more robust method to relate the feature space scores (from the kernel based methods) back to the input space. The use of optimisation methods may be useful.

3. Evaluation of the robustness of the statistical models and the training data sets (especially the top product flow fault set).

4. Comparison of the multivariate statistical methods to other methods. It may be interesting to compare data model and process model based methods.

5. The implementation of a PCA based fault detection and diagnosis online within the control system. The investigation of operator behaviour and appropriate responses to faults.

6. The comparison of the performance of the multivariate techniques to the monitoring of several univariate charts.

7. Investigate the fault detection and diagnosis performance with multiple simultaneous faults.

8. Quantify process performance improvement with the use of a online fault detection and diagnosis system.

BIBLIOGRAPHY

Albazzaz, H.; Wang, X. Z. and Marhoon, F. (2005) "Multidimensional Visualisation for Process Historical Data Analysis: A Comparative Study with Multivariate Statistical Process Control", *Journal of Process Control, 15*, 285–295.

Aldrich, C.; Gardner, S. and Le Roux, S. (2004) "Monitoring of Metallurgical Process Plants by using Biplots", *AiChE, 30*, 2167–2186.

Amand, T.; Heyen, G. and Kalitventzeff, B. (2001)a "Data Reconciliation for Simulated Flotation Process", *Computers & Chemical Engineering, 25*, 501–507.

Amand, T.; Heyen, G. and Kalitventzeff, B. (2001)b "Plant Monitoring and Fault Detection Synergy between Data Reconciliation and Principal Component Analysis", *Computers & Chemical Engineering, 25*, 501–507.

Bai, S.; Thibault, J. and McLean, D. D. (2006) "Dynamic Data Reconciliation: Alternative to Kalman Filters", *Journal of Process Control, 16*, 485–498.

Bergh, L.; Yianatos, J. B. and A., L. (2005) "Technical Note: Multivariate Projection Methods Applied to Flotation Columns", *Minerals Engineering, 18*, 721–723.

Bersimis, S.; Panaretos, J. and Psarakis, S. (2005) "Multivariate Statistical Process Control Charts and the Problem of Interpretation: A Short Overview and Some Applications in Industry", *WWW Document, http://www.stat-athens.aueb.gr/~jpan/papers/Panaretos-HERCMA2005ft.pdf*, 1–6.

Bissel, D. (1994) *Statistical Methods for SPC and TQM*, Chapman & Hall, London.

Box, G. and Luceño, A. (1997) *Statistical Control by Monitoring and Feedback Adjustment*, John Wiley & Sons, Canada.

Brambley, M. R. and Katipamula, S. (2005) "Methods for Fault Detection, Diagnostics and Prognostics for Building Systems A Review, Part I", *International Journal of HVAC&R, 11*.

Bureau of Labor Statistics (1998) *Occupational Injuries and Illnesses in the United States by Industry*, Government Printing Office, Washington DC.

Chen, Q.; Kruger, U. and Leung, A. Y. T. (2004) "Synthesis of T^2 and Q Statistics for Process Monitoring", *Control Engineering Practise, 12*, 745–755.

Cho, H.-W. (2007) "Identification of Contributing Variables using Kernel-based Discriminant Modeling and Reconstruction", *Expert Systems with Applications, 33*, 274–285.

Choi, S. W.; Lee, C. L.; Lee, J.-M.; Park, J. H. and Lee, I. B. (2005) "Fault Detection and Identification of Nonlinear Processes based on Kernel PCA", *Chemometrics and Intelligent Laboratory Systems, 75*, 55–67.

Choi, S. W. and Lee, I.-B. (2004) "Nonlinear Dynamic Process Monitoring based on Dynamic Kernel PCA", *Chemical Engineering Science, 59*, 1299–1319.

Choi, S. W.; Park, J. H. and Lee, I. B. (2004) "Process Monitoring using a Gaussian Mixture Model via Principal Component Analysis and Discriminant Analysis", *Computer and Chemical Engineering, 28*, 1377–1387.

Crnkovic-Dodig, L. (2006) "Principal Components Analysis", *WWW Document, http://blog.peltarion.com/2006/06/20/the-talented-drhebb-part-2-pca*, 6–8.

Crowe, C. M. (1996) "Data Reconciliation - Progress and Challenges", *Journal of Process Control, 6*, 89–98.

Dash, S. and Venkatasubramanian, V. (2000) "Challenges in the Industrial Applications of Fault Diagnostic Systems", *Computers and Chemical Engineering, 24*, 785–791.

Dong, D. and McAvoy, T. J. (1996) "Nonlinear Principal Component Analysis based on Principal Curves and Neural Networks", *Computers and Chemical Engineering, 20*, 65–78.

Du, Y. G.; Thibault, J. and Hodouin, D. (1997) "Plant Monitoring and Fault Detection Synergy between Data Reconciliation and Principal Component Analysis", *Artificial Intelligence in Engineering, 11*, 357–364.

Fourie, S. (2000) *Dissertation: Advanced Process Monitoring using Wavelets and Non-Linear Principal Component Analysis*, Department of Chemical Engineering, University of Pretoria.

Fourie, S. H. and de Vaal, P. (2000) "Advanced Process Monitoring using an on-line Non-Linear Multiscale Principal Component Analysis Methodology", *Computers in Chemical Engineering, 24*, 755–760.

Franc, V. and Hlaváč, V. "Statistical Pattern Recognition Toolbox for Matlab", Matlab Toolbox (June 24, 2004).

Frankowiak, M.; Grosvenor, R. and Pricket, P. (2005) "A Review of the Evolution of Microcontroller-based Machine and Process Monitoring", *International Journal of Machine Tools & Manufacture - Design, Research and Application, 45*, 573–582.

Gardner, S.; Le Roux, N. J. and Aldrich, C. (2005) "Process Data Visualisation with Biplots", *Minerals Engineering, 18*, 995–968.

Gianferrari Pini, L. "Misure di Profondit'a per lanalisi Statistica Nonparametrica Della Distribuzione Delle Vibrazioni Generate da una Macchina Rotante", Thesis - Politecnico di Milano (2004).

Gifi, A. (1990) *Nonlinear Multivariate Analysis*, Wiley, Chichester - England.

Goulding, P. R.; Lennox, B.; Sandoz, D. J.; Smith, K. J. and Marjanovic, O. (2000) "Fault Detection in Continuous Processes using Multivariate Statistical Methods", *WWW Document,*.

Groenewald, d. V.; Coetzer, L. P. and Aldrich, C. (2006) "Statistical Monitoring of a Grinding Circuit: An Industrial Case Study", *Minerals Engineering, 19*, 1138–1148.

Hand, D. (1982) *Kernel Discriminant Analysis*, Research Studies Press, Chichester - England.

Harris, R. J. (1985) *A Primer of Multivariate Statistics*, Academic Press, Florida - United States of America.

Henry, P. M. and Clarke, D. W. (1991) "A Standard Interface for Self-Validating Sensors", *IFAC Symosium - Safeprocess, 96*, 8b–013.

Himmelblau, D. M. (1978) *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*, Elvisar Press, Amsterdam.

Iserman, R. (2005) "Model-based Fault Detection and Diagnosis - Status and Applications", *Annual Reviews in Control, 29*, 71–85.

Jackson, J. E. (1991) *A User's Guide to Principal Components*, Wiley, New York.

Jemwa, G. T. and Aldrich, C. (2005) "Monitoring of an industrial liquid-liquid extraction system with kernel-based methods", *Hydrometallurgy, 78*, 41–51.

Jemwa, G. T. and Aldrich, C. (2006) "Kernel-based Fault Diagnosis on Mineral Processing Plants", *Minerals Engineering, 19*, 1149–1162.

Johnson, R. A. and Wichern, D. W. (1982) *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Jones, M. (2005) *Closed Loop Performance Monitoring*, Master Dissertation - University of Pretoria, South Africa.

Kourti, T.; Lee, J. and Macgregor, J. F. (1996) "Experience with Industrial Applications of Projection Methods for Multivariable Statistical Process Control", *Computers in Chemical Engineering, 20*, S745–S750.

Kourti, T. and MacGregor, J. F. (1995) "Process Analysis, Monitoring and Diagnosis, using Multivariate Projection Methods", *Chemometrics and Intelligent Laboratory Systems., 28*, 3–21.

Lee, J.-M.; Yoo, C. K.; Choi, S. W.; Vanrolleghem, P. A. and Lee, I.-B. (2004)a "Nonlinear Process Monitoring using Kernel Principal Component Analysis", *Chemical Engineering Science, 59*, 223–224.

Lee, J.-M.; Yoo, C. K. and Lee, I.-B. (2004)b "Fault Detection of Batch Processes using Multiway Kernel Principal Component analysis", *Computers and Chemical Engineering, 28*, 1837–1847.

Lucas, J. M. (1982) "Combined Shewart-CUSUM Quality Control Schemes", *Journal of Quality Technology, 14*, 785–791.

Macgregor, J. F. and Kourti, T. (1995) "Statistical Process Control of Multivariate Processes", *Control Engineering Practise, 3*, 403–414.

Martinez, L. and Martinez, A. (2004) *Exploratory Data Analysis with Matlab*, CRC Press, USA.

Montgomery, D. C. (1985) *Introduction to Statistical Process Control*, John Wiley & Sons, New York.

Montgomery, D.; Runger, G. C. and Hubele, N. F. (2001) *Engineering Statictics Second Edition*, Wiley, New York.

Murdoch, J. (1979) *Control Charts*, The MacMillan Press, London.

Murrill, P. W. (2005) *Fundamentals of Process Control Theory*, Instrument Society of America, United States of America.

Nagy, I. (1992) *Introduction to Chemical Process Instrumentation*, Elsevier, New York.

Nelson, L. S. (1984) "Rules for SPC Charts", *Journal of Quality Systems, 16 No.10*, 1075–1083.

NIST-Sematech (2005) "e-Handbook of Statistical Methods", *WWW Document, http://www.itl.nist.gov/div898/handbook/.*

Ott, E. R. (1975)a *Process Quality Control*, McGraw-Hill, New York.

Ott, O. E. (1975)b *Statistical Quality Control*, McGraw-Hill, New York.

Owen, M. (1989) *SPC and Continuous Improvement*, IFS Publications, UK.

Patel, B. (2000) *Investigation into Fault Detection and Diagnosis Techniques*, Master Dissertation - University of Cape Town, South Africa.

Rengaswamy, R.; Mylaraswamy, D.; Årzèn, K.-E. and Venkatasubramanian, V. (2001) "A Comparison of Model-Based and Neural Network-Based Diagnostic Methods", *Engineering Applications of Artificial Intelligence, 14*, 805–818.

Romagnoli, J. A. and Palazoglu, A. (2006) *Introduction to Process Control*, CRC Press, United States of America.

Rousseeuw, P. J.; Ruts, I. and Tukey, J. W. (1999) "The Bagplot: A Bivariate Boxplot", *The American Statistician, 53*, 382–388.

Scholköpf, B.; Smola, A. J. and Müller, K.-R. (1998) "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation, 10*, 1299–1319.

Shunta, J. P. (1995) *Achieving World Class Manufacturing through Process Control*, Prentice-Hall, New Jersey.

Singh, R. "Using Visualisation Techniques for Batch Conditioning", Web Tutorial - http://www.mathworks.com/company/newsletter/ (May 2003).

Smith, I. L. "A Tutorial on Principal Components Analysis", Tutorial (2002).

Van der Berg, F. (2007) "Introduction to Matlab and Mathematical Aspects of Bilinear Factor Models (PCA and PLS)", *WWW Document, http://www.models.ku.dk*, 77–79.

Venkatasubramanian, V.; Rengaswamy, R. and Yin, K. (2003)a "A Review of Process Fault Detection and Diagnosis - Part 1: Quantitative Model-Based Methods", *Computers and Chemical Engineering, 27*, 293–311.

Venkatasubramanian, V.; Rengaswamy, R. and Yin, K. (2003)b "A Review of Process Fault Detection and Diagnosis - Part 3: Process History Based Methods", *Computers and Chemical Engineering, 27*, 327–346.

Weiss, G. H.; Romangnoli, J. A. and Islam, K. A. (1996)a "Data Reconciliation - An Industrial Case Study", *Computers & Chemical Engineering, 20*, 1441–1449.

Weiss, G. H.; Romangnoli, J. A. and Islam, K. A. (1996)b "Data Reconciliation - An Industrial Case Study", *Computers & Chemical Engineering, 20*, 441–449.

Western Electric (1954) *Statistical Quality Control Handbook*, Western Electric Corporation, Indianapolis.

Wetherill, G. B. and Brown, D. W. (1991) *Statistical Process Control - Theory and Practice*, Chapman and Hall, London.

Wheeler, D. J. (1990) *Evaluating the Measurement Process*, Addison–Wesley Publishing Company, Avon.

Wold, S. (1978) "Exponentially Weighted Moving Principal Component Analysis and Projection to Latent Structures", *Chemical Intelligent Lab Systems, 23*, 149.

Wolf, P. (2006) "A Rough R Implementation of the Bagplot", *WWW Document, http://www.wiwi.uni-bielefeld.de/˜wolf/software/aplpack*, 382–388.

Yang, J.; Jin, Z.; Yang, J.; Zhang, D. and Frangi, A. (2004) "Essence of Kernel Fisher Discriminant: KPCA plus LDA", *Pattern Recognition, 37*, 2097–2100.