Multivariate Statistical Process Control of Chemical Processes

by

Michael Papazoglou Dipl. Chem. Eng.

A Thesis submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

Department of Chemical and Process Engineering

The University of Newcastle upon Tyne

1998

Abstract

The thesis describes the application of Multivariate Statistical Process Control (MSPC) to chemical processes for the task of process performance monitoring and fault detection and diagnosis. The applications considered are based upon polymerisation systems. The first part of the work establishes the appropriateness of MSPC methodologies for application to modern industrial chemical processes. The statistical projection techniques of Principal Component Analysis and Projection to Latent Structures are considered to be suitable for analysing the multivariate data sets obtained from chemical processes and are coupled with methods and techniques for implementing MSPC. A comprehensive derivation of these techniques are presented. The second part introduces the procedures that require to be followed for the appropriate implementation of MSPC-based schemes for process monitoring, fault detection and diagnosis. Extensions of the available projection techniques that can handle specific types of chemical processes, such as those that exhibit non-linear characteristics or comprise many distinct units are also presented. Moreover, the novel technique of Inverse Projection to Latent Structures that extends the application of MSPC-based schemes to processes where minimal process data is available is introduced. Finally, the proposed techniques and methodologies are illustrated by applications to a batch and a continuous polymerisation process.

Acknowledgements

There are a number of people who have contributed to the completion of this thesis and I would like to express my thanks to them all.

In particular, I would like to sincerely thank my supervisors, Professor A. Julian Morris, Department of Chemical and Process Engineering and Doctor Elaine B. Martin, Department of Engineering Mathematics, Directors of the Centre for Process Analytics and Control Technology (C.P.A.C.T), University of Newcastle upon Tyne, for their constant guidance, encouragement and support, throughout this project and during all of these years.

Many thanks to Professor C. Kiparissides, Director of the Laboratory of Polymer Reaction Engineering (L.P.R.E), Aristotle University of Thessaloniki and Chemical Process Engineering Research Institute (C.P.E.R.I), Greece, for supplying the simulation process models and data used in this work, for providing the guidance on polymerisation processes and for suggesting problems in polymerisation reactors to be solved.

Financial support is gratefully acknowledged from the European projects BRITE EURAM CT 93 0523 (INTELPOL) and ESPRIT PROJECT 22281 (PROGNOSIS), the Centre of Process Analysis, Chemometrics and Control, University of Newcastle and the Chemical Process Engineering Research Institute, Thessaloniki, Greece.

This work could not have been undertaken and completed without the constant love, support and encouragement of my wife Alexia and my two sons, Christos and Alexandros. It is impossible to express in writing my full gratitude to you all. Thank you so much. I hope your patience and sacrifices, during all of these years, will be well rewarded in the near future.

Table of Contents

4. Process Monitoring and Diagnostic Schemes Based on Multivariate Statistical Process Control 117

To the Family of my Fatherhood

Alexia, Christos and Alexandros

Chapter I

Introduction

1.1 General Overview

In today's competitive atmosphere, the chemical and process industries are being required to increase plant flexibility and to adapt to highly demanding situations where processes are subject to varying raw materials properties, changing technology and market conditions and fluctuating operating conditions. The expectation for improved product quality and the requirement to operate safely according to health, safety and environmental protection regulations have become imperative due to market and public demand. Successful operation in terms of high yield, better product quality and more consistent production at reduced operational costs and increased health and safety standards, can only be achieved when processes or plants are operated under well controlled conditions.

Efforts to manufacture a higher proportion of within specification product and to reduce the variability in the product quality, i.e. to produce more consistent product, has lead to an increase in the use of Statistical Process Control (SPC). SPC refers to a collection of statistical techniques and charting methods that have been found to be useful in ensuring consistent production and, consequently, in obtaining significant economic advantages. Traditional SPC can effectively detect or provide early warning of unusual events that can lead to off-specification production, process disturbances and faults, related to measurements of individual quality characteristics.

However, most modern industrial processes have available frequent on-line measurements on many process variables and, in some instances, on several properties of raw materials and final product. Furthermore, there are measurements of characteristics related to product quality that are usually measured infrequently off-line. Therefore, industrial quality problems are multivariate, since they involve measurements on a number of characteristics, rather than one single characteristic. As a result, univariate SPC methods and techniques provide little information about the interactions between characteristics and, therefore, it is not appropriate for modern day processes. Most of the limitations of univariate SPC can be addressed through the application of *Multivariate Statistical Process Control* (MSPC), which considers all the characteristics of interest simultaneously and can extract information on the behaviour of each characteristic relative to the others.

The major difficulty with multivariate data is that the variables being measured are almost never independent, but rather, they are highly correlated with one another at any given time. In trying to overcome these difficulties, a number of multivariate statistical projection methods can be applied, such as Principal Component Analysis (PCA) and Project to Latent Structures (PLS) or Partial Least Squares. These methods are particularly suited for the analysis of correlated data. They effectively project the data down onto low dimensional subspaces, that then contain all the relevant information relating to the process. Principal Component Analysis is one procedure that can be used to explain the variability in a single data set by defining a set of latent vectors that describe the direction of greatest variability. Projection to Latent Structures is similar to PCA, except that, PLS simultaneously reduces the dimensionality of both the process and quality variables spaces to calculate these latent vectors. In both PCA and PLS, the calculated latent vectors are uncorrelated and are a linear combination of the original correlated variables.

The key to successful process operation is efficient on-line performance monitoring. The primary aims of process monitoring are a reduction in off-specification production, the identification of important process disturbances and the early warning of process or plant faults. Where the early detection of process faults is followed by the location of their source, the efficiency and consistency of production and process/plant safety can be significantly improved. Consequently, on-line monitoring of process performance has become an extremely important part of any processing operation. Schemes for process monitoring, fault detection and diagnosis can then be used as intelligent supervisory process systems, which can support plant operators and process engineers in dealing with process deviations and help them in identifying the root cause of these deviations.

An MSPC monitoring scheme utilises a multivariate statistical model or representation that is constructed using either the statistical projection techniques of PCA or PLS. These techniques are only suitable for continuous processes that operate at steady state and, furthermore, they investigate the relationships between all the variables in the process in one single block. Moreover, all statistical projection techniques are data-oriented and, as a result, models for robust MSPC-based schemes can only be developed for processes where there is a wealth of data. Extensions of the projection techniques of PCA and PLS, namely Multi-Way and Multi-Block, can be used to construct statistical models for batch and semi-batch processes, and for complex processes comprising a number of distinct units. Additionally, a novel approach based upon the inversion of a PLS regression process model, which has the ability to generate the additional process data when there is only minimal plant data available for the development of a preliminary MSPC-based scheme, is also presented.

1.2 Outline of Thesis

Chapter 1 presents the motivation for the research work carried out, a general overview of the scientific areas covered and, finally, an outline of the main results obtained and innovations proposed.

Statistical Process Control (SPC) and its limitations, Multivariate Statistical Process Control (MSPC) and its advantages over SPC, when applied to industrial quality control problems are surveyed in Chapter 2. The charting techniques used by the methodologies of SPC and MSPC along with the approaches for constructing the appropriate control chart limits are also described.

The theoretical background underpinning the statistical projection methods of Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) is presented in Chapter 3 and the derivation of these methods using geometrical, mathematical and statistical considerations is also given. Finally, the relationship of the statistical projection methods of PCA and PLS with MSPC procedures is also presented.

Chapter 4 introduces the procedures that should be followed for the implementation of MSPC schemes for process monitoring, fault detection and diagnosis. Extensions to the projection techniques of PCA and PLS, namely Multi-Way and Multi-Block, are presented. Finally, a novel approach, Inverse Projection to Latent Structures (IPLS), for generating additional process data for the development of a MSPC monitoring scheme when minimal process data is available is presented. All issues associated with the implementation of MSPC monitoring schemes that utilise these statistical techniques are addressed.

Chapter 5 presents the application of MSPC-based schemes for process monitoring, fault detection and diagnosis to two example processes. Specifically, the polymerisation processes of Methyl-Methacrylate in a batch reactor and of Ethylene in a two-zone tubular reactor for the production of Low Density Poly-Ethylene, are considered. Comprehensive mathematical simulation programmes are utilised to represent the two processes and to generate the required process data sets for illustration of the proposed techniques. Inferential statistical models for the prediction of the final polymer properties and the estimation of the initial process conditions of the batch polymerisation process are developed. These models can be used in the general framework of MSPC-based schemes and address problems that are frequently encountered in industrial batch polymerisation processes. Additionally, the IPLS approach is illustrated by an example application to the batch reactor for an MSPC-based scheme. Finally, MSPC-based schemes for complex processes utilising Muti-Block projection techniques are illustrated with an application to the two-zone tubular reactor.

Conclusions and suggestions for further work are presented in Chapter 6 to complete the thesis.

Chapter II

Multivariate Statistical Process Control

2.1 Introduction

Maintaining product quality at a high level and ensuring product consistency is of widespread concern in the process and manufacturing industries. Efforts to manufacture a higher proportion of within specification product and to reduce the variability in the quality of a product has lead to an increase in the use of Statistical Quality Control (SQC) and Statistical Process Control (SPC). However, traditional SQC/SPC techniques cannot be applied effectively in multivariate quality control problems, which involve a vector of measurements of several characteristics rather than a single characteristic, typical of that encountered in industry today. Multivariate Statistical Process Control (MSPC), the multivariate extension of SPC, has been found to be particularly suited for many of the multivariate problems found in industrial quality control. MSPC is receiving significant attention because it is now recognised to have an important role to play in industry. It provides a diagnostic tool for the comprehensive on-line statistical monitoring of a process and the on-line detection and diagnosis of process malfunctions and it is applicable to both continuous and batch processes. This Chapter presents an extended overview of the univariate and multivariate Statistical Process Control methods and techniques that are currently found in industry today.

2.2 Traditional Statistical Process Control

Statistical Process Control (SPC) refers to a collection of statistical techniques and methods that have been found to be particularly useful in ensuring the consistent production of high quality products and, consequently, in obtaining significant economic advantages. These have been the major motivations for the extensive use and development of SPC methods during the past decades.

Statistical techniques for quality control, process improvement and sampling inspection trace their origins back to the early 1920's. In May 1924, Walter A. Shewhart of Bell Telephone Laboratories introduced the concept of the control chart, whilst seven years later in 1931, the initial theory of statistical quality control was developed (Shewhart, 1931). Work by him and others, including W. E. Deming, G. Tagushi, K. Ishikawa, J.M. Juran, G. E. P. Box, E.S. Page, S. Roberts, D. C. Montgomery, further refined and advanced the use of statistical quality and process control over the next seventy years. Traditional statistical quality and process control techniques reflect the nature of the discrete-event type of operations of the manufacturing industries, for which the techniques were initially developed. However, examination of these methods has shown that they can also be successfully applied to operations found in the process industries.

The primary objective of SPC is to control a process in a desired state with respect to a particular product specification (Chen, 1996). As a result, SPC tries to maintain the quality characteristics of products generated by a process, as close as possible to their desired target values by controlling and monitoring the performance of the process over time.

2.2.1 Statistical Quality Control and Statistical Process Control

Statistical Process Control (SPC) makes use of Statistical Quality Control (SQC) charting techniques, which have been well documented in traditional quality control textbooks (Juran, 1979; Ishikawa, 1986; Oakland and Followell, 1990; Banks, 1989; Wetherill and Brown, 1991; Montgomery, 1996). In the past, Statistical Quality Control encompassed both SQC and SPC, however, today there is a difference in the definition of these two terms, as a consequence of their underlying assumptions and philosophies (Alsup and Watson, 1993). In SQC, the quality of the product is assured by ensuring that the process is operating properly. On the other hand, SPC works under the assumption that if a process is operating properly, it will produce consistently high quality products. As a result, deviations from intended process operation will be responsible for products of poor quality. It can be seen that, both SPC and SQC act indirectly on the process, share the same tools and have the same objective, namely, quality improvement. However, SQC involves the application of a statistical methodology to the end product and it is associated with the product and its variations in quality, whilst SPC involves the application of a statistical methodology to the process parameters and it is associated with the process and focuses on process variability.

2.2.2 Sources of Process Variability

Process variability can be classified into two general types, based upon their source (Shewhart, 1931; Montgomery, 1996). In any process, there are many small, essentially unavoidable sources of variability that are inherent to the system itself. These are typically termed *chance variation, random cause variation* or *common* *cause variation* by many authors. Chance variation is predictable over time due to its randomness, but, it cannot be easily reduced or eliminated from the process. Examples of chance variation include variation due to temperature changes, raw materials variations, thermal and electrochemical noise, weather conditions, etc. A process that exhibits only chance variation is said to be *in statistical control.*

In addition, there are other sources of variation that may occasionally be present in a process. This type of variation forces an otherwise stable process to become unstable and unpredictable. Furthermore, it usually represents an unacceptable level of process performance and is termed *special cause variation* or *assignable cause variation* due to the fact that it can be readily assigned to an identifiable, particular cause or causes. Although, assignable cause variation is relatively large when compared to chance variation and is not predictable over time, it can typically be mitigated by applying appropriate corrective actions to the process. Example sources of assignable variation include different machine set-up conditions, change in shifts, different suppliers of raw materials, joining of different sub-assemblies etc. When a process exhibits only assignable cause variation, it is said to be *out-of-statistical-control.*

2.2.3 SPC Methodology

The behaviour of a process in a state of statistical control can be described by a statistical model by means of process average level and process spread. The model is built from data obtained when the process was operating well and only chance variation was present. SPC techniques monitor the performance of a process over time in order to verify that it remains in a state of statistical control. The occurrence of *unusual events* or *disturbances* can then be detected through the statistical analysis of the process variation, which involves the use of a statistical hypothesis testing procedure. This procedure is implemented by referencing the measured process behaviour, as described by data regularly collected during process operation, against the *in-control* model and its statistical properties. Having detected unusual events, SPC methods can then assist process operators in finding the assignable causes by investigating the process. Consequently, improvements in both the process and the quality of the products can be achieved by undertaking appropriate corrective action(s) that eliminate the causes before non-conforming product is produced. It can be seen that, the eventual goal of SPC is the elimination of all assignable causes of variation in the process, as stated by Montgomery (1996).

SPC can be considered as an activity designed to bring about process control and stability through the appropriate collection, analysis, interpretation and charting of numerical data. Furthermore, it is a philosophy of never ending quality improvement rather than a simple collection of statistical techniques and methods (Caulcutt, 1995).

2.3 Hypothesis-Testing in SQC and SPC

The occurrence of unusual events in a process can be detected by carrying out hypothesis-testing procedures based upon observed process data. A hypothesis is a statement about the state of a system (current or future, desirable or undesirable). Hypothesis-testing involves the evaluation of two hypotheses, namely, the *null* hypothesis, which is denoted as H_0 and expresses the current or assumed state of a system, and the *alternative* hypothesis, which is denoted as H_1 and expresses a future or desirable state. Having quantified the hypotheses, using knowledge about the system under study, one can reject or fail to reject the null hypothesis in favour of the alternative hypothesis and, therefore, can draw conclusions on the current the state of the system.

In statistics, a hypothesis is normally expressed in terms of the values of the parameters of a probability distribution. The value of the parameter specified in the null hypothesis is determined from past information or knowledge. The alternative hypothesis is interrogated by taking a random sample from the population under study and computing an appropriate test statistic. Depending upon the value of the test statistic, one can reject or fail to reject the null hypothesis in favour of the alternative hypothesis. The set of values that lead to the rejection of the null hypothesis is called the *rejection region.* In SQC/SPC, the hypothesis-testing problem may be summarised as follows :

H_0 : the process is operating under common cause variation

H_1 : the process is not operating under common cause variation

Thus, the null hypothesis (H_0) assumes that an unusual event is not present whilst the alternative hypothesis (H_1) constitutes a signal of the occurrence of an unusual event. The value of the parameter involved in H_0 is specified by past information that corresponds to a state of control, by a model of the process or by design considerations. The hypothesis-testing procedure involves periodic testing to investigate whether the value of the parameter has changed.

Chance variation is inherent to the process and, consequently, it is inherent to the sampling procedures. As a result, process average levels and variation, as calculated from random process samples may vary from sample to sample, even though the true process average level and spread remains constant. This results in the possibility of making one of two kind of errors when testing hypothesis :

type I error : reject the null hypothesis H_0 when it is true

type II error : fail to reject the null hypothesis H_0 when it is false

The probability of these two types of errors is denoted by :

$$\alpha : P\{\text{type I error}\} = P\{\text{reject } H_0 \mid H_0 \text{ is true}\} \tag{2.1}$$

$$\beta : P\{\text{type II error}\} = P\{\text{fail to reject } H_0 \mid H_0 \text{ is false}\} \tag{2.2}$$

Alternatively, it is more convenient to calculate the probability of correctly rejecting the null hypothesis as :

$$1 - \beta = P\{\text{reject } H_0 \mid H_0 \text{ is false}\} \tag{2.3}$$

The main role of statistical hypothesis-testing in SQC and SPC is to check the conformity of the process parameters or quality characteristics to their specified values and to assist in modifying the process until the desired values are achieved.

2.4 Control Charts

Control charts are the basic statistical tools used to monitor and control processes and systems. They can be easily constructed, visualised and interpreted. Furthermore, they have been shown to be very effective in practice. This is the main reason why they have been widely adopted and applied as a technique for effectively monitoring and controlling a process. Control charts were initially developed by Shewhart (1931) to help distinguish between variation exhibited in manufacturing processes that was inherent to the production system (common cause) and variation due to external factors (assignable cause).

The objective of a control chart is to monitor the performance of a process over time in order to verify that it remains in a state of statistical control. Typically, a control chart comprises a plot of a statistic over time, along with lines called *control limits*. The statistic is calculated using random process data. The control limits are selected so that if the process is in control, nearly all the calculated sample statistics will lie between them. However, when one or more of the sample statistics lies outside of the control limits or inside them in a systematic or non-random manner, then this event is interpreted as evidence that the process is out of control. A typical Shewhart-type control chart is shown in Figure 2.1.

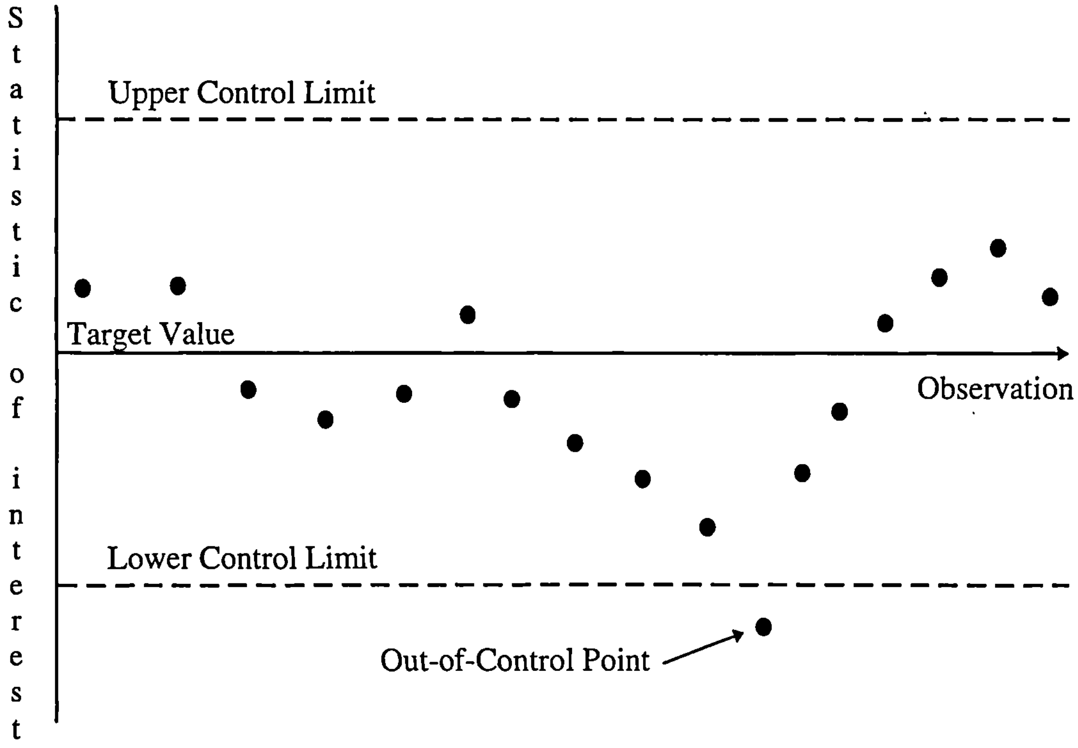


Figure 2.1. A Shewhart-type control chart

There are two distinct phases in constructing control charts, (Alt, 1985). The first stage (Phase I) involves testing whether the process was in control when the initial individual data were collected, and establishing appropriate control limits for monitoring purposes in the second stage (Phase II), in order to identify departures from the process standards, when future data is collected and monitored.

There is a close connection between control charts and hypothesis-testing. The hypothesis-testing procedure involved in traditional statistical quality and process control, is carried out on control charts on a constant basis, i.e. every sample. A sample statistic that lies within the control limits is equivalent to failing to reject the hypothesis of statistical control, whilst a sample statistic lying outside the control limits is equivalent to rejecting the hypothesis of statistical control. However, control charts go further than the hypothesis-testing framework. Control charts are usually used (a) to monitor a process, that is to detect the occurrence of unusual events that are departures from an assumed state of statistical control, (b) to assess process stability, that is to determine whether the process is still in control, and (c) to solve occurring problems by helping the investigator to identify the assignable causes of the problems. Control charts can be classified into two general types, namely, variable and attribute charts. The classification is based on whether the sample statistic is measured on a continuous scale (variable) or on a quantitative scale (attribute). In the design and construction of a control chart, there are many important issues including both the sensitivity and the ability of the control charts to perform their tasks. The most important issues are these of sample size and frequency of sampling. One approach to making a decision on these two issues is through the *average run length* (ARL) of the control chart. The ARL is the average number of points of the sample statistic that must be plotted before a point indicates the occurrence of an out-of control signal.

Description of the various types of control charts and their applications, along with a number of issues associated with them, can be found in standard statistical quality and process control textbooks (e.g. Banks, 1989; Wetherill, 1991; Montgomery, 1996). In the subsequent sections, the concepts and theory of the three most commonly applied control charts for the process mean, namely, the Shewhart, the Exponentially Weighted Moving Average and the Cumulative Sum control charts, are described. Information for control charts for the process variability and other statistics can be found in Wetherill and Brown (1991) and in Lowry et al. (1995).

2.4.1 Calculation of Control Limits

The specification of the control limits is the most critical decision that has to be made at the design stage (Phase I) of a control chart. Control limits are usually determined for the statistic being monitored and they define the boundary between the acceptance and the rejection region.

The fundamental assumption that underlies the calculation of control limits is that the process which generated the required data, was in a state of statistical control, i.e. the process data is independent and identically distributed. Violation of this assumption can lead to the misplacement of control limits and, therefore, to the misuse of the control charts (Alwan and Roberts, 1995). Another issue of importance is whether the calculation of the control limits should rely upon a distributional assumption. Exponents of the probabilistic approach argue that control limits are determined mathematically and the formulae used for their calculation is a direct application of Normal probability theory. On the other hand, exponents of the empirical approach argue that it is not necessary to assume a distribution or make any assumptions about the process or its data, since control charts are not based upon a distinct probability model. A description of these contrasting approaches can be found in Alwan and Roberts (1995) and in the discussion that follows their paper. It can be concluded that, although, the mathematical model that is used to calculate the control limits, is based mainly upon empirical evidence, however, the underlying assumption of normality should hold, that is the mathematical model should satisfy the underlying assumption of normality. The final conclusion is that, regardless of the approach one uses to calculate control limits, the control charts should work for the process under study.

The region on the control chart that the control limits mark out is called the c*ontrol region*. As the control region becomes wider, the risk of type I error decreases, but the risk of a type II error increases. Control limits are usually calculated by selecting the desired level of type I error probability. Usually, there are two control limits, namely, the *warning* and the *action* limits. Warning limits correspond to a 0.05 probability of type I error and provide an indication that the process may not be operating properly. Action limits corresponding to 0.01 probability of type I error, detect the occurrence of an unusual event, which may require corrective action to be taken.

2.4.2 Shewhart-type Control Charts

The most common type of control chart is that proposed by W.S. Shewhart in 1926. All control charts that are developed according to the general theory and principles proposed by Shewhart, are called *Shewhart-type* charts. They have been found to be appropriate for detecting large process shifts, but, are usually less sensitive in cases of small or slow shifts.

Suppose that, a statistic, which measures a characteristic of interest, is calculated for individual groups of samples randomly collected from a process, and that μ and σ denote the population mean and the population standard deviation of the statistic, respectively. A group of random process samples is called a *rational subgroup.* The values of the statistic of each rational subgroup can then be plotted against the subgroup number i. The control limits are then located at a distance from the population mean of the statistic (μ) that is L times the population standard deviation of the statistic (σ). This can be expressed mathematically as :

$$CL = \mu \pm L\sigma \tag{2.4}$$

The value of factor L is selected so that $100(1-\alpha)\%$ of the values of the statistic lie within the control region for a specific value of α, the probability of type I error. A typical example of a Shewhart-type control chart is illustrated in Figure 2.1.

2.4.3 Cumulative Sum Control Charts

The Cumulative Sum (Cusum) control chart is an alternative to the Shewhart-type chart, which can be used in the same context. It was first introduced by E.S. Page in

1954 and has been studied by a number of authors (Page, 1954; Page, 1961; Ewan, 1991; Gan, 1991; Woodall and Adams, 1993; Hawkins, 1993).

Cusum charts are generally used to detect small process shifts. Since they combine information from several samples, they are more effective than Shewhart-type charts, even in the case of subgroups of size n=1. They can detect process shifts of 0.5σ to 2σ in about half the time of a Shewhart chart with the same sample size, but they are slower in detecting large shifts (Montgomery, 1996).

A Cusum chart uses all the information in a sequence of values of a statistic by plotting the cumulative sums of their deviations from a target value. Suppose that rational subgroups of size n≥1 are collected from a process and that the average $\overline{x}_i$ of each rational subgroup is calculated. If μ_0 denotes the target for the process mean, then the Cusum control chart is formed by plotting the statistic :

$$C_i = \sum_{j=1}^{i} \left(\overline{x}_j - \mu_0 \right) \tag{2.5}$$

against the rational subgroup number i. A typical Cusum control chart is presented in Figure 2.2.

The control limits are usually calculated using the *V-mask* procedure (Barnard, 1959; Johnson, 1961). The out-of-control signal in a Cusum control scheme is given when the sample statistic C_i exceeds the control limits. Note that, re-initialisation of the Cusum statistic to target value is required after taking corrective action. A detailed discussion of the calculation of the ARL in Cusum control charts can be found in Montgomery (1996).

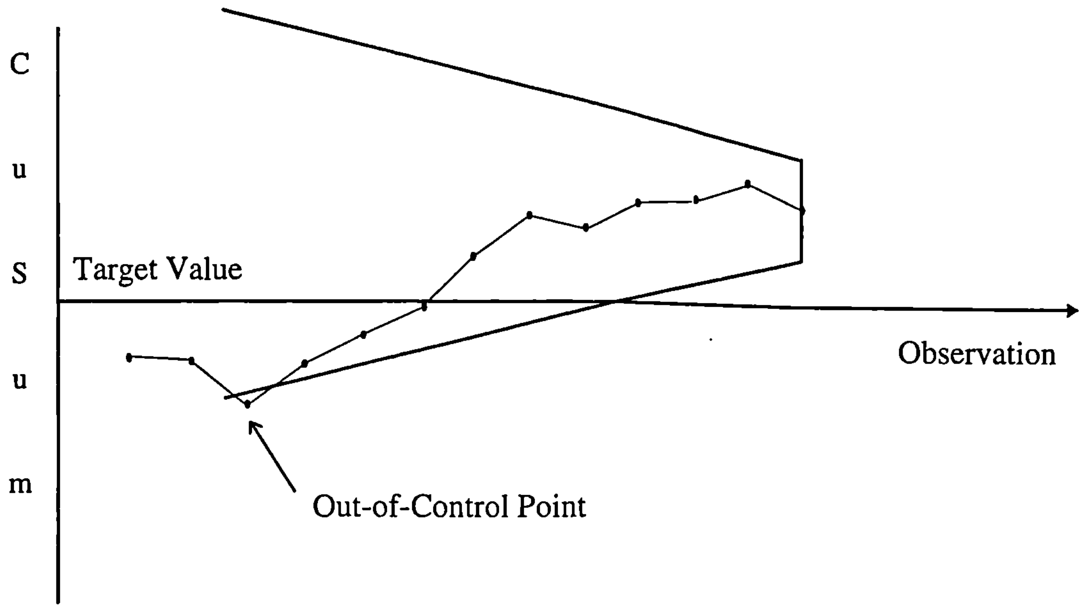


Figure 2.2. A Cumulative Sum control chart

2.4.3 Exponentially Weighted Moving Average Control Charts

An alternative to the Shewhart-type control chart, especially when one wants to detect small and moderately-sized sustained process shifts, is the Exponentially Weighted Moving Average (EWMA) control chart. It was introduced by S.W. Roberts in 1959. Comprehensive descriptions of EWMA are provided by many authors (Roberts, 1959; Crowder, 1989; Lucas and Saccussi, 1990; Davis and Woodall, 1994; Montgomery, 1996). The EWMA statistic is defined as :

$$z_i = \lambda \overline{x}_i + (1-\lambda) z_{i-1} \qquad (2.6)$$

or by recursive substitution as :

$$z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j \overline{x}_{i-j} + (1-\lambda)^i z_0 \qquad (2.7)$$

where $\overline{x}_i$ denotes the average of the i-th rational subgroup, λ is a weighting factor $(0 < \lambda \leq 1)$ and z_0 is the starting value of the statistic under study (first sample at i=1), which is usually taken to be equal to the population mean of the statistic (μ_0) :

$$z_0 = \mu_0 \tag{2.8}$$

The control limits for the EWMA control chart can be calculated based upon the assumption that the observations x_i that comprise the collected rational subgroup, are independent random variables :

$$CL_{EWMA} = \mu_0 \pm L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}\left[1-(1-\lambda)^{2i}\right]} \tag{2.9}$$

where L is a factor defining the width of the control limits and σ is the standard deviation of the sample under study.

EWMA can be viewed as a weighted average of all past and current observations. Specifically, a new moving average is formed each time a new sample is collected by calculating a weighted average of the new value and the previous moving average. A typical example of an EWMA control chart is illustrated in Figure 2.3. The performance of the EWMA control chart is approximately equivalent to that of the Cusum chart, although an EWMA chart is easier to set-up and operate. Furthermore, EWMA charts can be used to smooth the effects of known but uncontrollable noise in the data by appropriate choice of the weighting factor λ. Many chemical process with day-to-day fluctuations, fit into this category. Moreover, a modified EWMA control chart can be used for autocorrelated processes with a slowly drifting process mean (Mastrangelo and Montgomery, 1995). Issues including the ARL in EWMA control charts are discussed in Montgomery (1996).

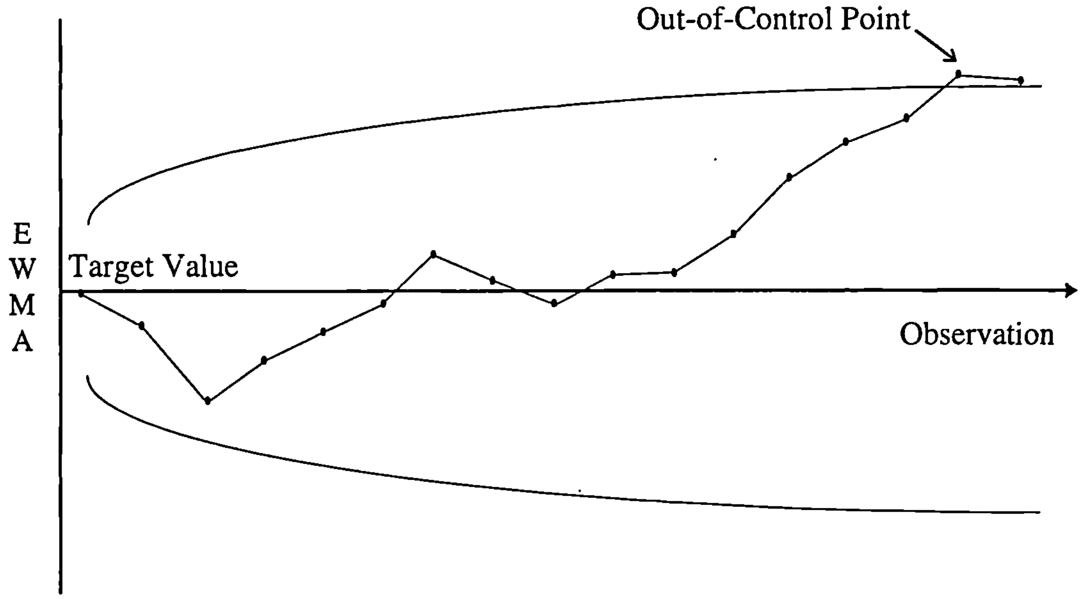


Figure 2.3. An EWMA control chart

2.5 Multivariate Statistical Process Control

Statistical Process Control (SPC) and control charts have evolved considerably since the first application of Shewhart charts. Over the past seventy years, SPC has grown and now can handle attribute data, moving averages and moving ranges, short-run applications and a variety of other exciting developments. However, the challenges in quality and process control continue to grow. As challenges grow, so procedures, methods and tools must also improve. The traditional SPC approach is not appropriate for modern day processes. Univariate SPC systems effectively only detect or provide early warning of unusual events that can lead to off-specification production, process disturbances and malfunctions, related to measurements of individual quality characteristics. However, most modern industrial processes are multivariate in nature. Consequently, industrial quality problems involve a vector of measurements on process and/or quality characteristics, rather than one single characteristic. As a result, univariate SPC methods and techniques provide little information about the interactions between characteristics, which are very important in complex processes, such as those found in the process and manufacturing industries. Most of the limitations of univariate SPC can be addressed through the application of *Multivariate Statistical Process Control* (MSPC), which considers all the characteristics of interest simultaneously and can extract information on the behaviour of each characteristic relative to the others. Therefore, companies willing to excel in the future should go beyond univariate SPC and focus upon MSPC.

Multivariate Statistical Quality Control (MSQC) and Multivariate Statistical Process Control (MSPC) was originally developed by Harold Hotelling in 1947. His work has been progressed by a number of researchers dealing with control procedures for more than one related variables (Jackson, 1956,1959,1985; Alt, 1985; Alt and Smith, 1988). However, in recent years MSPC has been recognised as having an important role to play in modern industry and a number of papers that extend traditional SQC/SPC techniques to the multivariate case, have been written.

There are four conditions that require to be satisfied by a multivariate statistical quality control or multivariate statistical process control procedure (Jackson, 1991).

1. The multivariate procedure should provide a single answer to the question of whether the process is in statistical control or not.

2. The overall type I error probability for the multivariate control procedure should be clearly specified.

3. The procedure to be followed should take into account all the relationships among the variables.

4. Procedures for finding assignable causes of unusual events occurring when the process is out of statistical control, should be available.

In the next section, the limitations of univariate SQC/SPC methods and the way which the multivariate approach addresses and handles these limitations and how it satisfies the above conditions is discussed.

2.5.1 The Limitations of Univariate SQC/SPC Methods

Two limitations are imposed when applying univariate SQC/SPC methods and techniques to multivariate control problems, that is the specification of the overall type I error probability and the construction of the control limits. These limitations originate from the conceptual underlying assumptions upon which the univariate control charting techniques are based and which are reflected in the approach used by these techniques to handle multivariate quality problems.

2.5.1.1 The Univariate Approach

In situations where more than one characteristics of interest, quality or process, is involved in a quality control problem, a separate univariate control chart for each characteristic, can be used to monitor the process. Although this approach readily provides a solution to the problem if the characteristics of interest are mutually uncorrelated, it can be misleading. Several authors (e.g. Alt, 1985 and 1988; Jackson, 1991; Montgomery, 1996) give clear examples illustrating that by using two separate control charts for two quality characteristics of a product, incorrect conclusion can be drawn. Consider that the quality of a product is described by two characteristics, namely, x_1 and x_2. Suppose now that, these characteristics are independent and normally distributed and, in addition, they are uncorrelated. The average level of the process can be monitored using two separate control charts for the means of the characteristics, $\bar{x}_j$ (j=1,2). The process is then considered to be in statistical control if and only if the means $\bar{x}_1$ and $\bar{x}_2$ of the two characteristics of interest for the rational subgroups collected from the process lie within their respective control limits. The use of separate control charts is equivalent to plotting the pair of means $(\bar{x}_1, \bar{x}_2)$ on a single chart, formed by superimposing one chart over the other, as shown in Figure 2.4.

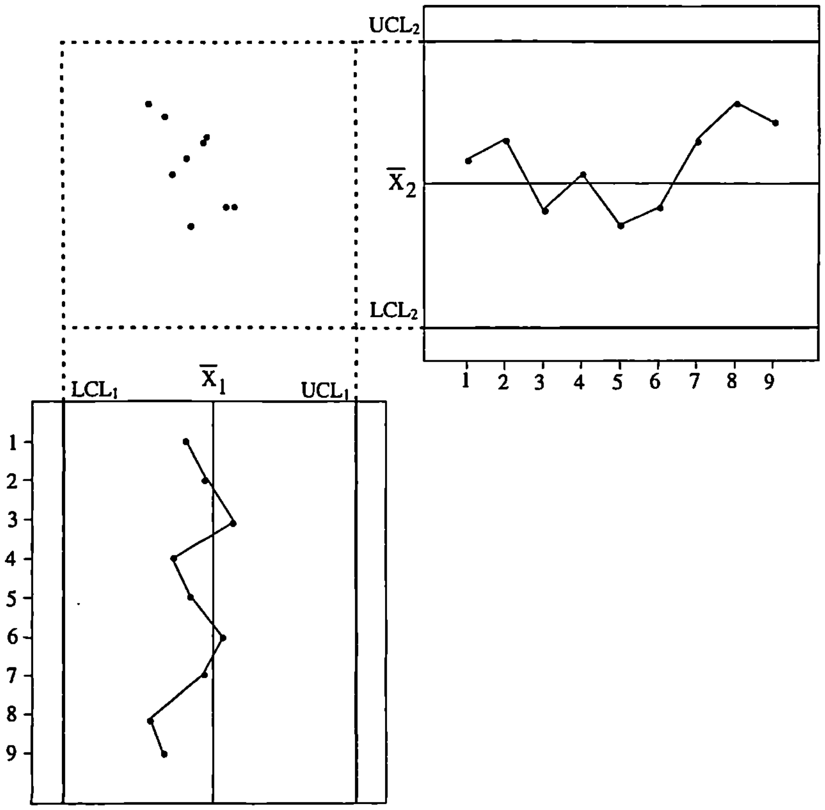


Figure 2.4. Rectangular univariate control chart

If the means of the two characteristics lie within the rectangular region of Figure 2.4, the process is considered to be in a state of statistical control. Suppose that for both characteristics, a type I error probability of 0.05 is specified. The joint probability that both of them would be in control is $(1-0.05)\cdot(1-0.05)=0.9025$, since these characteristics are uncorrelated. The overall type I error will be 0.0975. Therefore, the use of two independent control charts has distorted the joint control procedure in that the overall type I error is not equal to that described by the individual control charts. This distortion in the joint control procedure increases as the number of characteristics increases. In general (Montgomery, 1996), when the quality of a product is determined by m statistically independent characteristics and if m individual control charts for the mean values $\overline{x}_j$ (j=1,...,m) are used, with type I error probability of level α, then the overall type I error probability, denoted by α', for the joint control procedure is given by :

$$P\{\text{type I error}\}=\alpha'=1-(1-\alpha)^m \tag{2.10}$$

On the other hand, the probability that all means $\overline{x}_j$ (j=1,...,m) will simultaneously lie within their control limits when the process is in statistical control is :

$$P\{\text{all } \overline{x}_j \ (j=1,\ldots,m) \text{ within limits}\}=P\{\text{process in control}\}=(1-\alpha)^m \tag{2.11}$$

As an example, consider that one wants to simultaneously monitor nine statistically independent characteristics (m=9), with a type I error probability of 0.05 (Jackson, 1991). The overall type I error probability will be 0.37, that is at least more than one of these characteristics will indicate an out-of-control signal over one third of the time.

It was assumed that the two characteristics under study were uncorrelated. However, rarely are the process or the quality variables independent of one another. Usually, they are highly correlated, since only a few underlying events are driving the process at anyone time. As a result, measurements of different variables, process or quality, can be viewed as different reflections of the same underlying events. The problem previously described becomes more complicated when the characteristics are correlated. In the case they are perfectly correlated $(\rho = 1)$, the overall type I error would remain at α. Any kind of correlation less than perfect, involves a number of complex computations. It was shown that, when more than one characteristic determines the quality of a product, the overall type I error in a joint control procedure, can be incorrectly specified.

2.5.1.2 The Multivariate Approach

Consider now that, the two characteristics, x_1 and x_2, are jointly distributed according to the bivariate normal distribution and that rational subgroups of size n=1, for simplicity, are collected from the process. According to the Multivariate Normal Distribution Theory, when a (m×1) vector **x** of observations on m variables, follows an m-variate normal distribution with population mean (m×1) vector μ and square positive semi-definite (m×m) variance-covariance matrix Σ:

$$\mathbf{x} \sim N_m(\mu, \Sigma), \ |\Sigma| > 0 \tag{2.12}$$

then, the statistic :

$$\chi_0^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \tag{2.13}$$

is distributed as a chi-squared variate with m degrees of freedom (χ^2_m) :

$$\chi^2_0 = (x-\mu)^T \Sigma^{-1} (x-\mu) \sim \chi^2_m \tag{2.14}$$

and is called the *generalised distance* of **x** from μ. Furthermore, a probability of $(1-\alpha)$ is assigned to the *constant probability density contour* that is defined as :

$$\left\{x: (x-\mu)^T \Sigma^{-1} (x-\mu) \le \chi^2_{m,\alpha}\right\} \tag{2.15}$$

The contour of constant probability density is the surface of an ellipsoid centred at μ and with axes $\pm\chi_{m,\alpha}\sqrt{\lambda_j}e_j$, where (λ_j, e_j) is the j-th eigenvalue-eigenvector pair of Σ (j=1,...,m).

For the quality characteristics x_1 and x_2, that are jointly distributed as a normal bivariate, suppose that μ_1, μ_2 are the mean values and σ_{11}, σ_{22} are the variances of their population, respectively. The covariance between x_1 and x_2 is denoted by σ_{12}. All these statistics are assumed to be known :

$$\mu = |\mu_1 \;\; \mu_2|^T \qquad \text{and} \qquad \Sigma = \begin{vmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{vmatrix}, \quad (\sigma_{12} = \sigma_{21})$$

Since, it was assumed that the rational subgroups collected from the process are of size n=1 then, the means of the characteristics $(\bar{x}_1 \text{ and } \bar{x}_2)$ are equal to the values of the characteristics $(x_1 \text{ and } x_2\text{, respectively})$ for each subgroup, and they consist of a (2×1) vector $\bar{\mathbf{x}}$ (for each subgroup). The previous result is based upon the multivariate normal theory and can be reduced in the bivariate case to :

$$\chi^2_0 = (\bar{x}-\mu)^T \Sigma^{-1} (\bar{x}-\mu) \sim \chi^2_2 \tag{2.16}$$

or

$$\chi_0^2 = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \cdot \left(\begin{vmatrix} \bar{x}_1 \\ \bar{x}_2 \end{vmatrix} - \begin{vmatrix} \mu_1 \\ \mu_2 \end{vmatrix} \right)^T \cdot \begin{vmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{vmatrix} \cdot \left(\begin{vmatrix} \bar{x}_1 \\ \bar{x}_2 \end{vmatrix} - \begin{vmatrix} \mu_1 \\ \mu_2 \end{vmatrix} \right) \sim \chi_2^2 \qquad (2.17)$$

The statistic χ_0^2 can then be written as :

$$\chi_0^2 = \frac{1}{1-\rho_{12}^2} \cdot \left[\frac{(\bar{x}_1 - \mu_1)^2}{\sigma_{22}} + \frac{(\bar{x}_2 - \mu_2)^2}{\sigma_{11}} - 2\frac{\rho_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}(\bar{x}_1 - \mu_1)(\bar{x}_2 - \mu_2) \right] \qquad (2.18)$$

where ρ_{12} denotes the correlation coefficient between x_1 and x_2 that is defined as :

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} \qquad (2.19)$$

According to the multivariate normal probability theory and for a given probability of $(1-\alpha)$, equation (2.18) expresses an ellipse centred at (μ_1, μ_2), whose surface can be given by :

$$\frac{1}{1-\rho_{12}^2} \cdot \left[\frac{(\bar{x}_1 - \mu_1)^2}{\sigma_{22}} + \frac{(\bar{x}_2 - \mu_2)^2}{\sigma_{11}} - 2\frac{\rho_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}(\bar{x}_1 - \mu_1)(\bar{x}_2 - \mu_2) \right] \leq \chi_{2,\alpha}^2 \qquad (2.20)$$

This expression can be used as the basis of a control chart for the process average level, as it is expressed by the mean values μ_1, μ_2.Thus, for the two quality characteristics, a control region, whose boundary is an ellipse, can be constructed. This control region is often called a *control ellipse.* The degree of correlation between the characteristics affects the size and the shape of the control ellipse. The half length of the axes is given by $\chi_{2,\alpha}\sqrt{\sigma_{11} \pm \sigma_{12}}$, whilst their direction cosines are given by the eigenvectors of the variance-covariance matrix. In the case where the two characteristics are uncorrelated $(\sigma_{12} = 0 \text{ or } \rho_{12} = 0)$, the axes of the ellipse are parallel to the $\overline{x}_1$ and $\overline{x}_2$ axes. In the special case where both characteristics have the same variance $(\sigma_{11} = \sigma_{22})$, the ellipse is reduced to a circle centred at (μ_1, μ_2), as shown in Figure 2.5. However, in most cases, the characteristics under study are correlated $(\sigma_{12} \neq 0)$ and, therefore, the control ellipses take the form shown in Figure 2.6. It can be seen in Figure 2.6 that, if one uses the univariate rectangular control region, pairs of mean values $(\overline{x}_1, \overline{x}_2)$ that lie within region **A**, can lead to the wrong conclusion that the process is in a state of statistical control, whilst the process in practice is out of control.

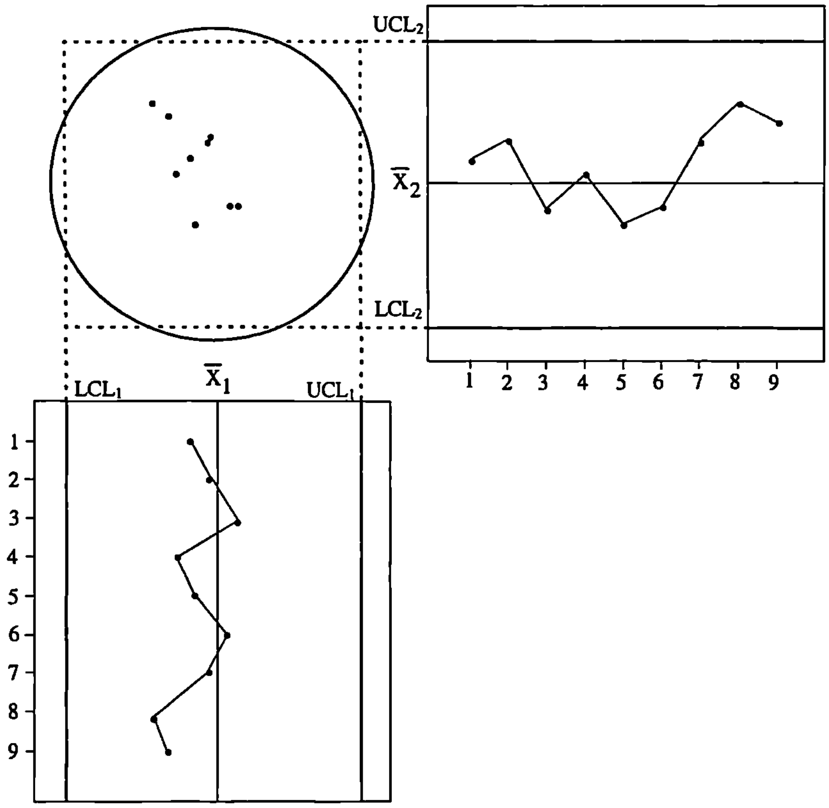


Figure 2.5. Control region for independent variables

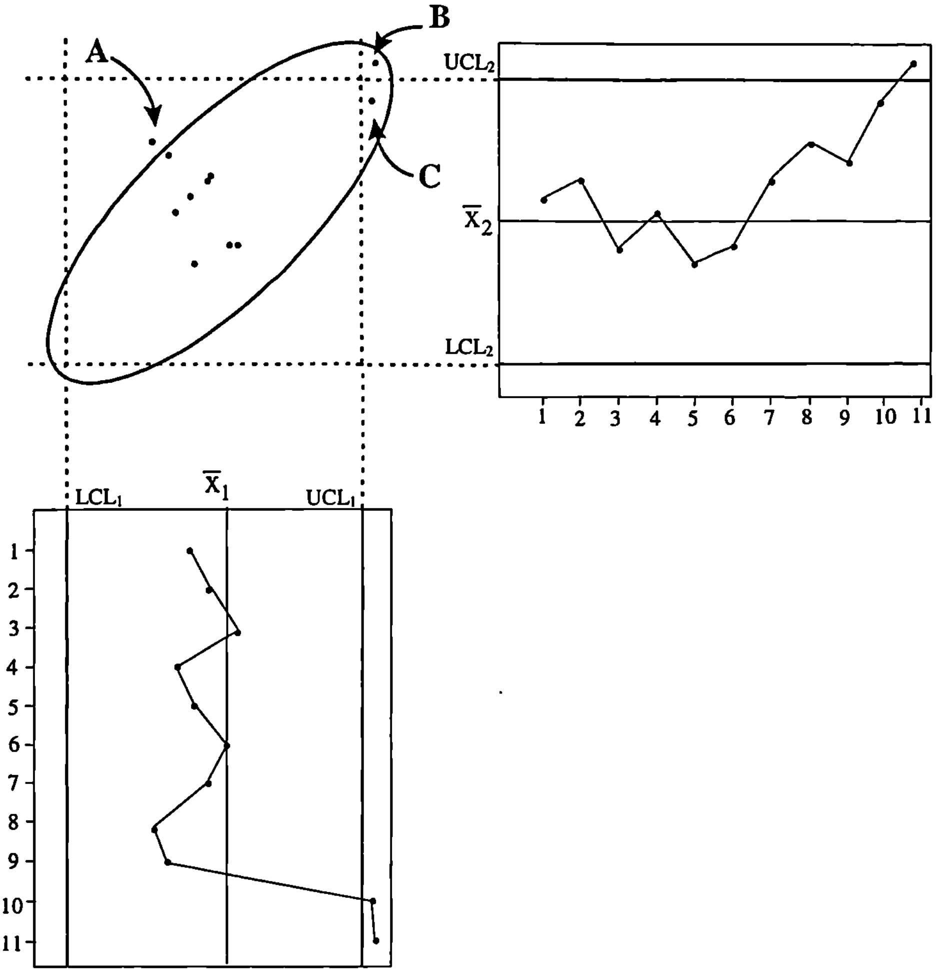


Figure 2.6. Control region for correlated variables.

Furthermore, pairs that lie within regions **B** and **C**, lead to the conclusion that the process is out-of-statistical-control, since one or both of the mean values violated their univariate control limits, whilst the process is in control.

A monitoring procedure for the multivariate case can be carried out as follows. Having defined a control region by the control ellipse, rational subgroups of the two characteristics can be collected from the process. If the pairs of the mean values $\left(\overline{x}_1, \overline{x}_2\right)$ of the subgroups, lie within the control region (Figure 2.6), then the process is considered to be in a state of statistical control. Two disadvantages are associated with the elliptic control region. The first is that the time sequence of the plotted points is lost. The second disadvantage is that it is difficult to construct the ellipse for more than two characteristics (Montgomery, 1996). Alternatively, the pairs $(\bar{x}_1, \bar{x}_2)$ can be used to calculate values of the χ_0^2 according to equation (2.18) and to plot them on a Shewhart-type control chart, termed a χ^2-chart. A typical χ^2-chart is presented in Figure 2.7 along with its control limit, which can be calculated for a given level of type I error probability, α. The concept of the χ^2-chart, provides the foundation for extending the univariate case to the multivariate situation.

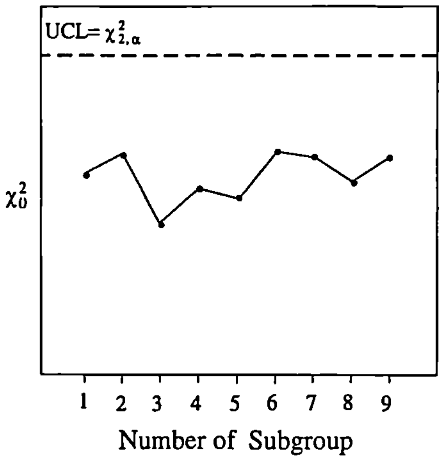


Figure 2.7. χ^2 - control chart

In the above multivariate examples, the population mean vector (μ) and variance-covariance matrix (Σ) were assumed to be known. In practice, however, they are unknown and, therefore, they need to be estimated from a preliminary sample of rational subgroups that were collected from the process when it was in control. Furthermore, when the size of the preliminary sample of rational subgroups is small then, instead of the χ^2 statistic, *Hotelling's T^2* can be used (Hotelling, 1947). This is presented in the following section. Discussion about the case where the size n of the rational subgroups collected from a process is greater than one, can be found in the literature (Montgomery, 1996).

2.5.2 Hotelling's T^2 Statistic

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a random sample of n vectors with observations on m variables, from a normal m-variate population with mean vector μ and covariance matrix Σ, then the maximum likelihood estimators of μ and Σ are :

$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{2.21}$$

$$\hat{\Sigma} = \frac{n-1}{n}S = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{2.22}$$

where $\bar{\mathbf{x}}$ and S are the sample mean vector and covariance matrix, respectively, which are statistically independent. According to the Central Limit Theorem, $\bar{\mathbf{x}}$ is distributed as a multivariate normal variate with mean μ and covariance Σ/n, and (n-1)S is distributed as a Wishart variate with n-1 degrees of freedom :

$$\bar{\mathbf{x}} \sim N_m(\mu, \Sigma/n) \tag{2.23}$$

$$(n-1)S \sim W_m(n-1, \Sigma) \tag{2.24}$$

The problem now is : "Is a specific vector μ_0 a plausible value for the population mean vector ? ". From the hypothesis-testing perspective, this problem can be stated as :

$$H_0{:}\mu = \mu_0 \qquad \text{versus} \qquad H_1{:}\mu \neq \mu_0$$

The appropriate statistic is Hotelling's T^2 (Hotelling, 1947), which is analogous to the Student's t-test in univariate statistical analysis :

$$T^2 = n(\bar{x}-\mu_0)^T S^{-1}(\bar{x}-\mu_0) \tag{2.25}$$

In cases where n-m is large (typically, greater than 30), T^2 is distributed as a chi-squared variate with m degrees of freedom (see section 2.5.1.2) :

$$T^2 = n(\bar{x}-\mu_0)^T S^{-1}(\bar{x}-\mu_0) \sim \chi^2_m \tag{2.26}$$

or

$$T_0^2 = (\bar{x}-\mu_0)^T S^{-1}(\bar{x}-\mu_0) \sim \frac{1}{n}\chi^2_m \tag{2.27}$$

where T_0^2 is the squared generalised distance from the sample mean to the test value and is also called the *Mahalanobis Distance* (Mahalanobis, 1936). The null hypothesis H_0 is rejected in favour of H_1 for a significance level α (out-of-control signal) if :

$$T^2 > \chi^2_{m,\alpha} \tag{2.28}$$

The 100(1-α)% joint confidence region of μ, in the space of **x**, is an ellipsoid :

$$T^2 \le \chi^2_{m,\alpha} \tag{2.29}$$

The centre of the ellipsoid is $\bar{x}$, and the lengths and directions of the axes are given by :

$$\pm\sqrt{\lambda_j \chi^2_{m,\alpha}}\,\mathbf{e}_j \tag{2.30}$$

where λ_j and $\mathbf{e}_j$ (j=1,...,m) are the eigenvalues and eigenvectors of the sample covariance matrix S, respectively.

In situations when the sample is small (n≤30), the T^2 is distributed as an F variate with m, and n-m degrees of freedom :

$$T^2 = n(\bar{\mathbf{x}} - \mu_0)^T S^{-1}(\bar{\mathbf{x}} - \mu_0) \sim \frac{(n-1)m}{n-m} F_{m,n-m} \tag{2.31}$$

or

$$T_0^2 = (\bar{\mathbf{x}} - \mu_0)^T S^{-1}(\bar{\mathbf{x}} - \mu_0) \sim \frac{(n-1)m}{(n-m)n} F_{m,n-m} \tag{2.32}$$

The null hypothesis H_0 is rejected in favour of H_1 for a significance level α (out-of-control signal) if :

$$T^2 > \frac{(n-1)m}{n-m} F_{m,n-m,\alpha} \tag{2.33}$$

or

$$T_0^2 > \frac{(n-1)m}{(n-m)n} F_{m,n-m,\alpha} \tag{2.34}$$

The 100(1-α)% joint confidence region of μ, in the space of **x**, is again an ellipsoid

$$T^2 \le \frac{(n-1)m}{n-m} F_{m,n-m,\alpha} \tag{2.35}$$

with centre $\bar{\mathbf{x}}$, and the lengths and directions of the axes are given by :

$$\pm\sqrt{\lambda_j \frac{(n-1)m}{n-m} F_{m,n-m,\alpha}}\, \mathbf{e}_j \tag{2.36}$$

A new approach to constructing T^2- based control limits has been presented by Tracy et al., (1992). In the start-up stage (Phase I), it is assumed that there is a sample of n vectors $\mathbf{x}_i$ with observations on m variables. A multivariate control chart can then be constructed, based upon Hotelling's T^2, using the statistic $T^2_{0,i}$ for each observation vector $\mathbf{x}_i$, which is the Mahalanobis Distance from the observation vector $\mathbf{x}_i$ to the mean vector of the sample of $\mathbf{x}_i$ vectors, $\overline{\mathbf{x}}$:

$$T^2_{0,i} = \left(\mathbf{x}_i - \overline{\mathbf{x}}\right)^T S^{-1}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right) \tag{2.38}$$

However, Tracy et al., (1992), suggested that, the assumption that $\overline{\mathbf{x}}$ and S are the estimates of the population values μ and Σ, does not hold true and, therefore, the $T^2_{0,i}$ statistic cannot be approximated by a chi-squared variate with m degrees of freedom. Furthermore, they proposed that the $T^2_{0,i}$ statistic is distributed as a Beta variate with m/2 and (n-m-1)/2 degrees of freedom (Gnanadesikan and Kettenring, 1972) :

$$T^2_{0,i} \sim \frac{(n-1)^2}{n} B_{m/2,(n-m-1)/2} \tag{2.39}$$

This distribution is applicable when the individual $\mathbf{x}_i$ vectors, collected during the start-up stage, lie within the control limits, e.g. the process was in statistical control. The control limit for a level of significance α (in-control signal) is given by :

$$T^2_{0,i} \leq \frac{(n-1)^2}{n} B_{m/2,(n-m-1)/2,\alpha/2} \tag{2.40}$$

Consider that a future individual observation vector $\mathbf{x}_f$ is collected in the second stage (Phase II), where one wants to test if the process is still in control. Assuming that $\mathbf{x}_f$, $\overline{\mathbf{x}}$ and S are statistically independent, then

$$\mathbf{x}_f - \overline{\mathbf{x}} \sim N_m\left(0, \frac{n+1}{n}\Sigma\right) \tag{2.41}$$

or

$$\sqrt{\frac{n}{n+1}}\left(\mathbf{x}_f-\bar{\mathbf{x}}\right) \sim N_m(\mathbf{0},\Sigma) \tag{2.42}$$

Hotelling's T^2 statistic for the $\mathbf{x}_f$ vector now becomes :

$$T^2=\left(\frac{n}{n+1}\right)\left(\mathbf{x}_f-\bar{\mathbf{x}}\right)^T S^{-1}\left(\mathbf{x}_f-\bar{\mathbf{x}}\right) \tag{2.43}$$

and is distributed as an F variate with m and (n-m) degrees of freedom :

$$T^2 \sim \frac{(n-1)m}{n-m}F_{m,n-m} \tag{2.44}$$

or

$$T_0^2 \sim \frac{(n+1)(n-1)m}{(n-m)n}F_{m,n-m} \tag{2.45}$$

The out-of-control signal at a significant level α is :

$$T_0^2 > \frac{(n+1)(n-1)m}{(n-m)n}F_{m,n-m,\alpha} \tag{2.46}$$

Similarly the joint confidence ellipsoid can be defined.

The approach to constructing control limits for Hotelling's T^2 statistic, proposed by Tracy et al., (1992), is statistically more correct and appropriate for the implementation of MSPC, since it discriminates between the two Phases of constructing control limits.

2.5.3 Important Issues In The Implementation of MSPC

As stated by Jackson (1991), there are four conditions that should be satisfied for the successful implementation of Multivariate Statistical Quality Control (MSQC) or Multivariate Statistical Process Control (MSPC) (section 2.5). In the previous sections, it was shown that the first two conditions, namely, the provision of a single answer to the question whether the process is in statistical control and the specification of the overall type I error probability, are satisfied by MSQC/MSPC. The third condition is partially satisfied. MSPC/MSQC takes into account the relationships between the characteristics of interest, but these relationships are not always clear, due to the fact that a large amount of correlated data can be collected from modern industrial processes and are, therefore, available to be used for statistical modelling. An extensively applied solution to this particular problem is the use of multivariate statistical projection methods.

The fourth condition, namely, the availability of procedures for the identification of assignable causes of unusual events, still constitutes a growing field of research in MSPC. The diagnosis of a multivariate control chart signal, that determines which of the monitored characteristics is responsible for the out-of-control signal, is not easy. Several diagnostic techniques that mainly involve the identification of the major contributors to the out-of-control signal, have been proposed in the literature. However, all of them have disadvantages and, as a result, the full implementation of MSPC schemes in industrial processes is still limited.

2.5.3.1 The Use of Multivariate Statistical Projection Methods

Every process has available frequent observations on many process variables, such as temperatures, pressures and concentrations and on several properties of raw materials and final products. Process observations are usually measured on-line while quality observations are typically measured off-line. Furthermore, there is available a history of past successful and some unsuccessful operation of the process. The major difficulty with this amount of multivariate data is that the variables being measured are almost never independent, but rather, they are autocorrelated in time and highly correlated with one another at any given time (collinear). This is due to the underlying relationships between the variables or to the place where the measurements were taken or due to the nature of the process. Therefore, there are only a few underlying events that drive the process at any time and measurements on all these variables are simply different reflections of the same underlying events. Finally, it is important to note that, not only are the relationships between the variables at any time, important, but, so is the entire past history. In trying to overcome these difficulties, the multivariate statistical methods of Principal Component Analysis (PCA) and Project to Latent Structures (PLS) or Partial Least Squares, have been applied. These methods are particularly suited to the analysis of correlated data by projecting the data down onto low dimensional subspaces, that contain all the relevant information about the process. Principal Component Analysis is a procedure used to explain the variability in a single data set by defining a set of latent vectors that describe the direction of greatest variability and that are uncorrelated. Projection to Latent Structures is similar to PCA, except that, PLS simultaneously reduces the dimensionality of both the process and quality variables spaces to find these latent vectors. In both PCA and PLS, the calculated latent vectors are uncorrelated and represent the original correlated variables. Therefore, when statistical projection methods are combined with the multivariate statistical process control, the overall type I error in the control charts can be directly computed. The statistical projection methods of PCA and PLS are described in Chapter III.

2.5.3.2 Interpretation of Multivariate Out-of-Control Signals

A practical disadvantage of most multivariate control charting techniques is that it is not possible to directly determine which of the characteristics being monitored, is responsible for the out-of-control signal. A number of methods have been proposed to address this problem for Hotelling's T^2 and χ^2 charts. In the case where a multivariate control chart indicates that the process is out of statistical control, the most obvious approach to use is to interrogate all the univariate control charts and to apply individual t-test on each characteristic in order to determine which variables are responsible for the out-of-control signal (Alt, 1985; Doganaksoy et al., 1991; Hayter and Tsui, 1994; Fuchs and Benjamini, 1994). However, two issues discussed in the previous sections, must be considered. First, it is difficult to determine the overall type I error probability when one uses simultaneous confidence intervals, and, secondly, univariate control charts can only detect an individual characteristic as responsible for an out-of-control signal and not a combination of them, since the correlation structure between the characteristics is not used in the construction of the univariate control charts.

An alternative is the decomposition of the T^2 statistic into two or more parts, based upon subsets of the characteristic of interest. Murphy, (1987), suggested partitioning Hotelling's T^2 into two parts. One part is then associated with the group of variables that are intuitively suspected (through engineering knowledge) to have caused the out-of-control signal $\left(T_j^2\right)$. The test statistic is the difference :

$$D = T^2 - T_j^2 \tag{2.47}$$

which is distributed as a chi-squared or F statistic, depending upon whether the population mean and variance-covariance are known or estimated, respectively. A hypothesis-test can then be used to select which possible combination of characteristics is responsible for the out-of-control signal. However, the number of possible combinations of characteristics increases as more characteristics are involved in the calculation of the T^2 values and, therefore, it is essential to use engineering knowledge.

A new unified approach on the interpretation of the T^2 decomposition proposed by Rencher, (1993), has been proposed by Mason et al., (1995 and 1997). When a future observation vector $\mathbf{x}_f$ on the m characteristics is collected, there are m! different possible ways to decompose the corresponding T^2 value or (m-1)! different possible ways to decompose it for each one of the m characteristics. In each case, T^2 is decomposed as the sum of the unconditional term T_p^2 for the p-th characteristic and the m-1 conditional terms for the remaining m-1 variables as :

$$T^2 = T_p^2 + \sum_{j=1}^{m-1} T_{j+1,1\ldots,j}^2 \tag{2.48}$$

where T_p^2 is the square of the univariate t statistic for the initial variable and $T^2_{j+1,1,...,j}$ is the conditional term of the (j+1)-th variable, which is the square of the (j+1)-th component of the vector $\mathbf{x}_f$ adjusted by the estimates of the mean and the standard deviation of the conditional distribution of the (j+1)-th variable given the first j variables. Each of the terms can be compared to an F distribution. However, this approach involves heavy computational effort as more characteristics are involved in the calculation of the T^2 values.

Finally, Healy (1987), Pignatiello and Runger (1990), and Hawkins (1993b), recommended a similar control statistic to detect a shift of the process mean in the direction of one of the process variables. Runger et al. (1996), proposed an approach to relate a shift of the process mean to the importance of a variable to it.

2.6 Multivariate Control Charts

Separate control charts for individual characteristics of interest are more easily interpretable, but are substantially less powerful, particularly in the presence of appreciable correlation between the characteristics under study. Therefore, the use of multivariate control charts is essential. However, multivariate control charts should retain the simplicity of the univariate charts concerning the graphical representation and the interpretation of results. Multivariate control charts can be constructed using the same concepts and under the same assumptions as univariate charts and, furthermore, are appropriate for multivariate data sets that exhibit less than full statistical rank, such as PCA and PLS (Palm et al., 1997). A review of multivariate control charts can be found in Lowry and Montgomery, (1995), while a comparison of them is presented by Harris and Ross, (1991), and Sparks, (1992).

2.6.1 Multivariate Shewhart-type Control Chart

In situations where a vector of characteristics are observed at each time period, Shewhart-type control charts are typically used. As in the univariate case, Shewhart-type control charts only use information from the current sample. Consequently, they are sensitive to large shifts in the value of the statistic that is being plotted. Hotelling's T^2 and χ^2 are the most common statistics used to construct multivariate Shewhart-type control charts.

2.6.2 Multivariate Cumulative Sum Control Chart

The Cumulative Sum (CUSUM) control chart is similar in principle to the univariate CUSUM chart and, consequently, it is used to detect shifts in the process mean (Healy, 1987; Crosier, 1988; Pignatiello and Runger, 1990). The CUSUM control chart is based upon a sequence of sequential probability ratio tests. Consider a vector of observations $\mathbf{x}_i$ on m variables obtained from a process that is distributed as $N_m(\mu, \Sigma)$. A Hotelling's T^2 value can be calculated at each point where a vector $\mathbf{x}_i$ is collected. The CUSUM of the scalar distance T^2 or its square root can be computed as :

$$C_i = \text{Max}\{0, C_{i-1} + T_i - k\} \tag{2.49}$$

with initial condition $C_0 \geq 0$. This CUSUM scheme signals an out-of-control situation when $C_i > h$.

Crosier (1988) considered replacing the scalar quantities of the univariate CUSUM by their vector counterparts so that :

$$C_i = \left\{ \left(r_{i-1} + x_i - \tau \right)^T \cdot \Sigma^{-1} \cdot \left(r_{i-1} + x_i - \tau \right) \right\}^{1/2} \tag{2.50}$$

$$r_i = \begin{cases} 0 & \text{if } C_i \le k \\ \left(r_{i-1} + x_i - \tau \right) \cdot \left(1 - k / C_i \right) & \text{if } C_i > k \end{cases} \tag{2.51}$$

where τ is the target for the process mean vector, Σ is the population variance-covariance matrix for the process vectors x_i, $r_0 = 0$ and for some chosen value of k>0. The out-of-control signal is given whenever :

$$\sqrt{r_i^T \Sigma^{-1} r_i} > h \tag{2.52}$$

These CUSUM charts use all the observations since the detection of the last special event rather than only the last observation vector as in the Shewhart-type charts. Their advantage over the latter charts is that their average run length is smaller for small shifts in the process mean. However, they have not been applied in MSPC schemes on chemical processes.

2.6.3 Multivariate Exponential Weighted Moving Average Control Chart

Multivariate EWMA (MEWMA) charts compute the exponentially weighted moving average of a vector of process observations x_i on m variables (Lowry et al., 1992; Prabhu and Runger, 1997). The MEWMA is a logical extension to the univariate EWMA and, consequently, they can be used to detect small and moderate shifts. It is defined as :

$$z_i = \Lambda x_i + (I - \Lambda) z_{i-1} \tag{2.53}$$

where $\mathbf{z}_0$ is a m-dimensional zero vector and $\Lambda = \text{diagonal}\left\{\lambda_{1,}\lambda_2,\ldots,\lambda_p\right\}$ with $0 < \lambda_j \le 1; j = 1,\ldots,m$ is a parameter that controls the magnitude of smoothing. Large values of λ_j result in greater smoothing and better detection of small shift. The quantity that is plotted on the control charts is :

$$Q_i^2 = \mathbf{z}_i^T \Sigma_{\mathbf{z}_i}{}^{-1} \mathbf{z}_i \tag{2.54}$$

where $\Sigma_{\mathbf{z}_i}$ is the covariance matrix of the $\mathbf{z}_i$ statistics. In case all λ_i are equal and Σ denotes the variance-covariance matrix of the population of $\mathbf{x}_i$, it is defined as :

$$\Sigma_{\mathbf{z}_i} = \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\Sigma \tag{2.55}$$

The MEWMA gives an out-of control signal when :

$$Q_i^2 > H \tag{2.56}$$

where the control limits H is chosen to achieve a specified in-control ARL.

The properties of the MEWMA chart are similar to those of the multivariate Cumulative Sum Charts. Lowry et al., (1992) gives guidance on the choice of the upper control limit for the MEWMA control chart. A design procedure for MEWMA charts that gives recommendations for parameter settings analogous to the results provided for univariate EWMA by Lucas and Saccucci (1990), can be found in Prabhou and Runger (1997).

2.7 Applications of MSPC

A number of applications of Multivariate Statistical Process Control can be found in the literature. Most of them include the application of multivariate statistical projection methods, such as PCA and PLS. Applications of MSPC of particular interest are those for on-line process monitoring, fault detection and fault diagnosis.

MSPC applications generally involve a procedure where process data are analysed and statistical models of the process are developed. Historical data sets collected from past successful process operations and operations under specific disturbances can be found in most industrial processes. The data sets from normal operation are used to construct a statistical model, which represents the situation where only common cause variation is present in the process. Multivariate Statistical process monitoring and control involves three activities :

1. Detection of abnormal behaviour (unusual event, out-of-control signals)

2. Identification of the variable(s) indicative of this unusual event

3. Diagnosis of the source responsible for this abnormal behaviour

Monitoring focuses on the detection and identification activities, whilst diagnosis provides the information for the intervention or control stage. The statistical model describing process behaviour under *normal operating conditions* (NOC) is used with new collected process data to decide whether the current operation is in control. In case of an out-of-control signal, further procedures are used to interpret this abnormal behaviour and to find the assignable cause(s) responsible for it. Overviews of SPC for multivariate processes and applications are presented by MacGregor, (1994), and MacGregor and Kourti, (1995).

A significant amount of work has been carried out in the field of MSPC of continuous processes. Industrial applications include catalytic cracking in petroleum refining (Slama, 1991), mineral processing (Tano et al., 1993), photographic paper manufacturing (Miller et al., 1995), a pulp digestion process (Dayal et al., 1994), a polymer solution and a chemical separation processes (Kosanovich and Piovoso, 1995), a ceramic melting process (Wise and Gallagher, 1996) and many others as summarised by Kourti and MacGregor, (1996). Furthermore, a number of simulation studies have been performed on an extractive distillation column and a fluidised bed reactor (Kresta et al., 1991), a LDPE tubular reactor (Skagerberg et al., 1992; Kourti and MacGregor, 1996; MacGregor et al., 1994), a CSTR (Zhang et al., 1996) and on the Tennessee Eastman process (Raich and Cinar, 1996).

In recent years, the application of MSPC has been extended from continuous to batch processes (Nomikos and MacGregor, 1995; Kourti et al., 1995). Examples application have been illustrated using simulated processes mainly on batch polymerisation processes (Nomikos and MacGregor, 1994; Nomikos and MacGregor, 1994b). Industrial applications include batch polymerisation reactors (Nomikos and MacGregor, 1995; Kosanovich et al., 1996) as well as a nuclear waste storage tank (Gallagher et al., 1996).

2.8 Summary

This Chapter has presented the concepts, the philosophy and the techniques for Multivariate Statistical Process Control (MSPC). Traditional Statistical Quality Control (SQC) and Statistical Process Control (SPC) methods and techniques have been successfully applied in industry. However, they are univariate and, therefore, their application to modern industrial control problems is limited by the nature of the modern industrial processes that comprise highly correlated variables. The introduction of Multivariate Statistical Process Control (MSPC) has successfully addressed most of these problems. Univariate SQC\SPC methods and control charting techniques, which form the basis of MSPC, along with the problems occurring in their application to multivariate control problems have been extensively discussed. The approach used by MSPC to address these problems along with the associated important issues and multivariate control charting techniques are also described in depth.

Chapter III

Multivariate Statistical Projection Techniques

3.1 Introduction

Multivariate statistical analysis is the visualisation and interpretation of a set of observations that describe a natural or physical phenomenon. Typically, the observed phenomena are complex and the resulting set of observations large. A particular set of techniques which effectively enable such a problem to be analysed are the multivariate statistical projection techniques. The objective of these techniques is to compress the data and, in doing so, summarise the information they contain. The most well known techniques are those of Factor Analysis (FA), Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Canonical Variation Analysis (CVA), Principal Component Regression (PCR) and Projection to Latent Structures (PLS). Recently, PCA and PLS have been applied to engineering problems, in the area of Multivariate Statistical Process Control (MSPC), since they are able to compress the large volumes of highly correlated data collected on a process and, furthermore, they satisfy the conditions imposed by MSPC problems (section 2.5, Jackson (1991)).

3.2 Multivariate Statistical Analysis of Data

All types of scientific data analysis have the collection of observations on a social or physical phenomenon as a common basis. Traditional statistical analysis is based upon the collection of observations on one *characteristic* or *variable* at a time. There is usually assumed to be no relationship between individual observations in time (*autocorrelation*). Furthermore, the statistical principle of random sampling ensures that different observations are uncorrelated, if not independent of one another. However, the complexities of most phenomena require observations on more than one variable to be collected and, therefore, most data bodies can be characterised as exhibiting multivariate behaviour. When more than one variable is observed, some form of correlation will exist between individual variables. Multivariate analysis simultaneously investigates all the variables to reveal the relationships between them, in order to interpret the data appropriately and to achieve the objectives of the analysis.

Multivariate data analysis consists of methods and techniques that represent the conjunction of concepts from linear and matrix algebra, multidimensional geometry and calculus with statistics. This is the main reason why it is called *multivariate statistical analysis*. Multivariate statistical analysis originated from the work of mathematicians of the last century. Nowadays, there are a number of books describing the various techniques (Mardia, et al., 1974; Kendall, 1980; Green, 1978; Anderson, 1984; Seber, 1984; Krzanowski, 1988; Everitt and Dunn, 1991; Johnson and Wichern, 1992; Krzanowski and Marriott, 1996; Gnanadesikan, 1997). Most authors have used a technique-oriented approach based upon mathematics and statistics to present the subject of multivariate statistical analysis. More recently, a number of authors have adopted a problem-oriented approach, which has enabled the use of these techniques by researchers with minimal mathematical and statistical knowledge.

The computational effort required to implement the various multivariate methods and techniques, rendered their application, during the early part of the century, to small data sets. However, the last twenty years has seen an explosion in computer power and, as a result, the use of multivariate techniques has become widespread and applications have extended to large data sets. It is difficult to establish a classification scheme for multivariate techniques, since they encompass a wide assortment of descriptive and inferential scientific tools. However, they are useful for two main scientific pursuits, namely:

1. *Explanatory research*, which includes the following objectives :
   a. *Analysis of variable dependence.* The data set is partitioned into two subsets. Relationships between variables of these sets must be determined in order to examine their dependence on one another. Techniques include Multivariate Analysis of Variance (MANOVA) and Multiple Linear Regression (MLR).

   b. *Analysis of variable interdependence.* The nature of the relationships between variables is of interest. Relationships can range from independence to collinearity. Techniques include Factor Analysis (FA), Principal Components Analysis (PCA) and other dimensionality-reduction or structural simplification methodologies.

   c. *Analysis of interobject similarity.* The nature of the relationships between objects is of interest. Relationships that force subsets of objects to fall into groups or clusters must be determined. Techniques include Cluster Analysis (CA) and other types of object-grouping techniques.

2. *Confirmatory research.* This is the testing of several alternative models of association between two or more variables or groups of objects. This may done to

validate assumptions or to reinforce previous convictions. Techniques include the F-test, the t-test and other statistical testing procedures.

It can be seen that the techniques of multivariate statistical analysis form a unified set of procedures that can be organised around a few original problems. However, they are not confined to a single discipline, but rather, they span a diverse range of scientific fields. Application areas can be drawn from the social, medical and physical sciences, engineering, applied economics and business management.

Many univariate statistical methods generalise quite naturally to higher dimensions, e.g. the Multivariate Analysis of Variance (MANOVA) is the multivariate generalisation of the univariate Analysis of Variance (ANOVA). Furthermore, most of the univariate continuous distributions have multivariate analogues with the property that all of their univariate marginal distributions belong to the same family, e.g. univariate and multivariate normal distributions. The heart of the univariate statistical analysis is the sample, which is a set of measurements for n objects on a single variable. Similarly, the heart of the multivariate statistical analysis is the multivariate sample. A multivariate sample arises whenever one takes random measurements on n objects for m variables that theoretically represent the process under study. The measurements on n objects for each variable of interest comprise a vector of dimension n. The vectors for all the variables of interest may or may not come from the same probability distribution but they are merged into a single common matrix, called the multivariate *data matrix* or multivariate *data set.*

The main problem in any multivariate statistical analysis of data is that of *data visualisation.* Any multivariate sample can be described in terms of two geometrical configurations. In the case where the objects are the focus of attention, each of the m variables can be associated with an axis in m-dimensional space, which are assumed to be mutually orthogonal. The m values of a particular object can then be taken as the co-ordinates of a point representing the particular object along the axes. Therefore, all the objects of a multivariate sample can be geometrically modelled as n points in an m-dimensional space, *object space*. On the other hand, *variable space* is defined by associating each of the n objects with an orthogonal axis in an n-dimension space. Similar to the previous configuration, the n values attached to a particular variable can be taken as the co-ordinates of a point in this space. The multivariate sample in this case is represented by a *geometrical model of m points in* an n-dimensional space. Most times, the object-oriented geometrical configuration is preferred. However, neither of the previous configurations allow an m- or an n-dimensional space to be transformed, to a two- or three-dimensional subspace without losing important information about the process. However, dimensionality-reduction techniques have the potential to transform high-dimensional space to a lower-dimensional subspace, without affecting the relative positions of the points that represent the multivariate sample and, furthermore, without losing the important information.

3.3 Reduction of Dimensionality

In addition to visualisation, a further problem associated with the statistical analysis of multivariate samples, is that of *interpretation*. Reducing the dimensionality of a problem by removing some of the variables, can lead to a reduction in the useful information and, thus, to the erroneous or deficient interpretation of the data.

Therefore, the issue when reducing the dimensionality in the multivariate statistical analysis of data is to ensure simplicity for visualisation, whilst retaining sufficient information for appropriate and relevant interpretation (Gnanadesikan, 1997).

Most of the techniques used to reduce the dimensionality of multivariate space use the concept of *latent variables.* A latent variable is a hypothetical variable constructed for the purpose of understanding a characteristic of interest that cannot be measured directly. The term was introduced in the social and behavioural sciences in order to describe particular concepts that are not directly observable, e.g. intelligence in psychology, economic expectation in economics. Although latent variables are not observable, they have a certain impact on the measured variables and, therefore, are subject to analysis. Latent variables are usually defined to be a linear combination of the measured variables.

In the following section, the two most commonly applied in MSPC latent variable techniques, Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) are described in detail. Furthermore, a number of other techniques, e.g. Factor Analysis, Canonical Correlation Analysis and Canonical Variation Analysis are discussed.

3.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is amongst the oldest and most widely used multivariate statistical technique for dimensionality reduction. The method of Principal Component Analysis dates back to Karl Pearson in 1901, who introduced it as a technique for fitting planes by orthogonal least squares. However, the general procedure, as it is known today, was developed by Harold Hotelling in 1933, who proposed PCA for analysing the covariance and correlation structures between a number of random variables. The development of the method was rather uneven in the ensuing years due to computational difficulties. The wider application of PCA occurred in the 1960s, during the "Quantitative Revolution" of the physical and social sciences, when the development of computers made it possible to apply multivariate statistical techniques to reasonably large-sized problems.

The objective of PCA is the explanation of the variance-covariance structure of a multivariate data sample containing significant redundancies, in terms of a set of uncorrelated latent variables, each of which is a particular linear combination of the original variables. The mathematical and statistical aspects underpinning PCA are well defined. Theoretical introductions to PCA can be found in a number of books on multivariate statistical analysis (Mardia, et al., 1974; Kendall, 1980; Anderson, 1984; Seber, 1984; Muirhead, 1982), while application-oriented introductions are presented by Krzanowski, (1988), and Johnson and Wichern, (1992). Books devoted to PCA include Jolliffe, (1986), and Jackson (1991). Furthermore, overviews on some of the concepts and properties that comprise the theoretical background of PCA are described in a number of papers (Wold, 1987; Geladi and Kowalski, 1986; Mackiewicz and Ratajczak, 1993).

As in any multivariate statistical technique, the starting point in Principal Component Analysis (PCA) is a multivariate sample of observations, which characterises n objects with respect to m random variables $x_1, x_2, ..., x_m$, and which is represented by a data matrix X of dimension (n×m). Each column vector $\mathbf{x}_j$ in matrix X contains observations on n objects for the j-th variable x_j (j=1,...,m), while each row vector $\mathbf{x}_i$ (i=1,...,n) contains observations on the i-th object for all the m variables. Furthermore, each row vector can be geometrically modelled as a point in the m-dimensional object space. PCA decomposes the multivariate data set X into a series of R principal components. Each principal component is characterised by a *score* vector ($\mathbf{t}_r$) and a *loading* vector ($\mathbf{p}_r$). Using this decomposition, the data set X can be written as a linear combination of the principal components :

$$X = \sum_{r=1}^{R} \mathbf{t}_r \cdot \mathbf{p}_r^T \tag{3.4.1}$$

or

$$X = T \cdot P^T \tag{3.4.2}$$

where T denotes the matrix of scores, whose columns are the score vectors ($\mathbf{t}_r$), and P^T denotes the matrix of loadings, whose rows are the loading vectors ($\mathbf{p}_r$). This procedure is illustrated in Figure 3.1.

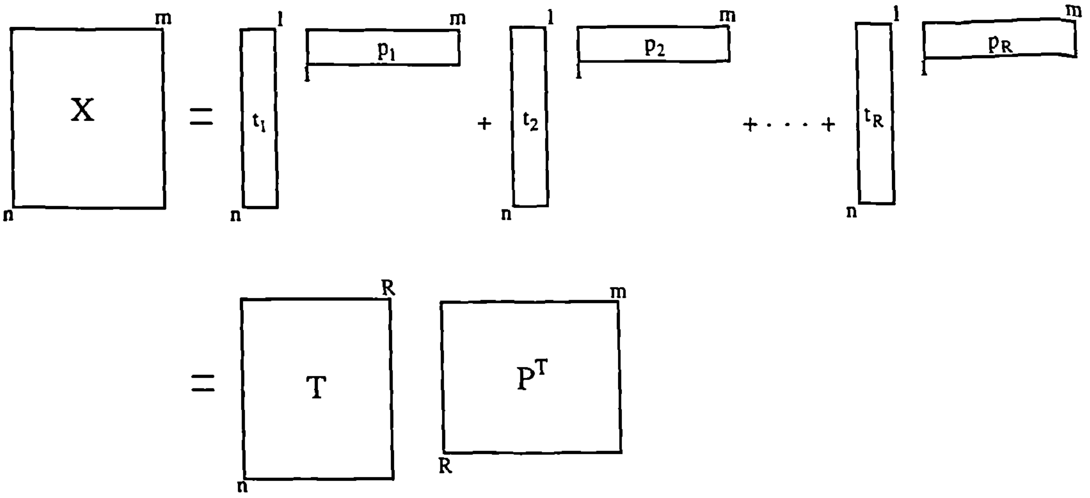


Figure 3.1. Decomposition of a data set X by PCA (Geladi and Kowalski, 1986).

3.4.1 Geometrical Interpretation and Mathematical Derivations of PCA

Within this section, the derivation of principal components using mathematical, geometrical and statistical considerations and the interpretation of the principal components is presented. Although mainly based upon the work of Krzanowski (1988), it summarises various topics described in a number of multivariate statistical analysis textbooks.

Consider an m-variate data sample of n objects. The sample can be graphically modelled as a swarm of n points in an m-dimensional object space by assigning each measured variable (j) to the unit vector ($\mathbf{u}_j$), which defines the j-th axis of the space (j=1,...,m). The co-ordinates of the n objects in the space are given by the data matrix X and, therefore, the corresponding n points are represented by a set of vectors $\mathbf{x}_i$ (i=1,...,n), the rows of X. Suppose that one wants to reduce the dimensionality of the m-dimensional object space, without losing important information about the process under study. One can consider a one-dimensional subspace formed by a new axis, whose one-dimensional unit vector v_1 is defined as :

$$v_1 = a_{11}\mathbf{u}_1 + a_{12}\mathbf{u}_2 + ... + a_{1m}\mathbf{u}_m \qquad (3.4.3)$$

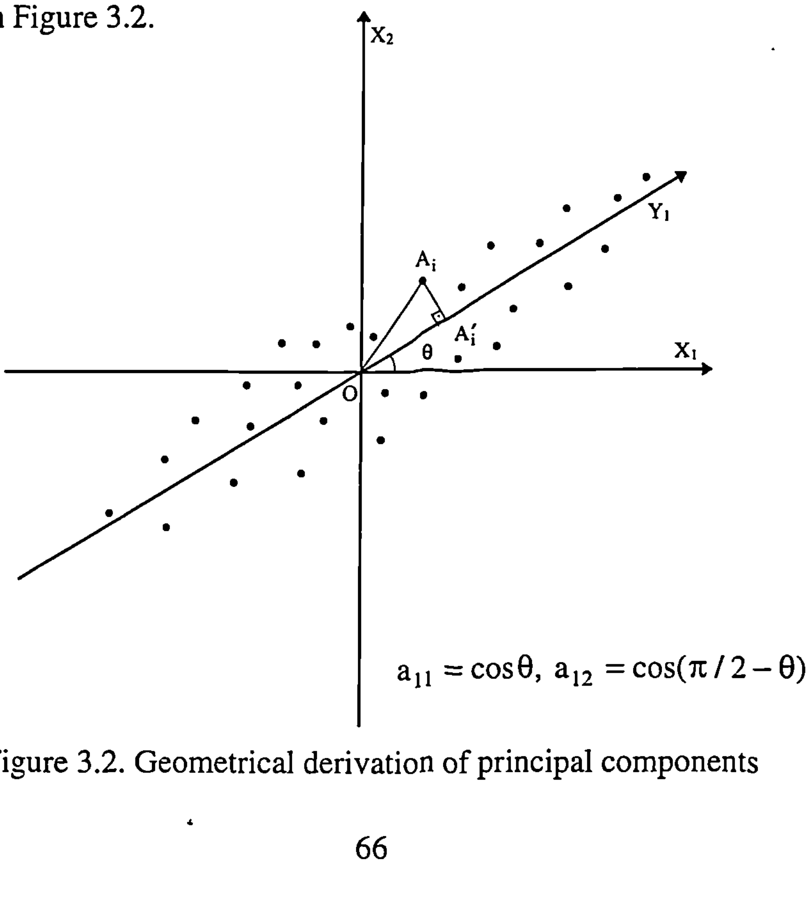
where a_{1j} is the direction cosine of v_1 relative to $\mathbf{u}_j$. Note that, the sum of the square direction cosines of a vector, originating at the centre of a space in $\Re^m$, is equal to unity :

$$\sum_{j=1}^{m} a_{1j} = 1 \Rightarrow \mathbf{a}_1^T\mathbf{a}_1 = 1 \qquad (3.4.4)$$

where **a** is an m-dimensional vector, whose elements are the direction cosines. The m-variate data sample can be approximated by the orthogonal projections of its n points onto the new axis, which are represented by a set of one-dimensional vectors y_{1i} (i=1,...,n) defined as :

$$y_{1i} = a_{11}x_{1i} + a_{12}x_{2i} + ... + a_{1m}x_{mi} = \mathbf{a}_1\mathbf{x}_i^T \tag{3.4.5}$$

The set of y_{1i} (i=1,...,n) can be represented by an n-dimensional vector $\mathbf{y}_1$. Since the projection of an object $\mathbf{x}_i$ onto the one-dimensional subspace is y_{1i}, then it can be seen that, the projection of the m-variate data set X is the vector $\mathbf{y}_1$. However, the projection leads to a displacement of these points from their original locations onto the new axis. The smaller the displacement of all the point, the better the m-variate data sample is approximated by its projections onto the new axis and, therefore, the better the m-dimensional space is compressed down onto the new one-dimensional subspace. Thus, a measure of *goodness-of-fit* for this compression can be defined as the sum of the squared perpendicular distances of the objects $\mathbf{x}_i$ (i=1,...,n) from the one-dimensional subspace. This procedure is illustrated for a bivariate sample (m=2) of n objects in Figure 3.2.



$$a_{11} = \cos\theta,\ a_{12} = \cos(\pi/2 - \theta)$$

Figure 3.2. Geometrical derivation of principal components

The centre of the two-dimensional space is denoted as O, the axes OX_1 and OX_2 correspond to the two variables $\mathbf{x}_1$ and $\mathbf{x}_2$, and OY_1 is the new defined axis. For a particular point A_i, its orthogonal projection onto OY_1 is A'_i. Applying the Pythagorean Theorem to the triangle $OA_iA'_i$, it can be seen that :

$$(OA_i)^2 = (OA'_i)^2 + (A_iA'_i)^2 \tag{3.4.6}$$

Summing over all n points, it follows that :

$$\sum_{i=1}^{n}(OA_i)^2 = \sum_{i=1}^{n}(OA'_i)^2 + \sum_{i=1}^{n}(A_iA'_i)^2 \tag{3.4.7}$$

Axis OY_1 can be optimally defined only when the sum of squares of the perpendicular displacements of all n points is minimised.

Similarly, in the m-dimensional object space, by applying the Pythagorean Theorem and summing over all the points representing the m-variate data sample, equation (3.4.7) can be written as :

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{n} e_i^2 \tag{3.4.8}$$

where d_i denotes the *Euclidean distance* of the i-th point ($\mathbf{x}_i$) from the centre of the m-dimensional space, f_i denotes the Euclidean distance of y_{1i}, which is the projection of $\mathbf{x}_i$ onto the new axis, from the centre of the one-dimensional subspace and, finally, e_i is the displacement of the i-th point caused by the orthogonal projection (see Figure 3.3). The placement of the new axis that minimises the sum of the squared displacements of all points can be found by optimally determining its direction cosines (**a**). The objective function can be formed by examining the individual terms of equation (3.4.8).

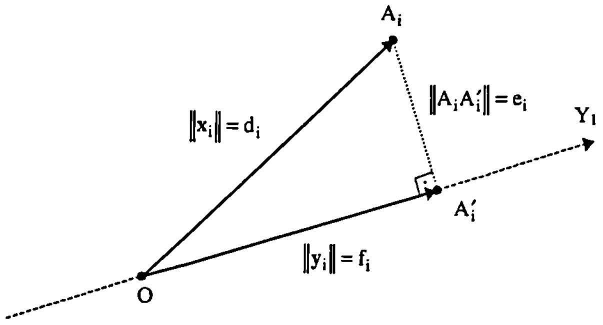


Figure 3.3. Projection of an object onto the new axis

Suppose now that, the centre of the m-dimensional object space is located at $\bar{\mathbf{x}}$, which is the mean vector of the data matrix X or the mean of the object vectors $\mathbf{x}_i$. It is an m-dimensional vector that contains the mean values of all objects, $\bar{x}_j$, for all the variables :

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)^T = \frac{1}{n} X^T \mathbf{1}_n \tag{3.4.9}$$

Therefore, each object in the m-dimensional space is represented by a vector $\mathbf{x}_i$, whose squared length is the squared Euclidean distance (d_i^2) of the corresponding point from the centre of the space :

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \tag{3.4.10}$$

The sum of the squared Euclidean distances for the n points is :

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_j)^2 = \sum_{i=1}^{n} \operatorname{diag}(X - \mathbf{1}_n \bar{\mathbf{x}}^T)(X - \mathbf{1}_n \bar{\mathbf{x}}^T)^T \tag{3.4.11}$$

On dividing by (n-1), the summation of equation (3.4.11) is equal to the total sample variance, Var(X), the sum of the diagonal elements, s_{ii}, or the trace of the sample variance-covariance matrix S, which is constant for a given sample (Johnson and Wichern, 1992) :

$$\frac{1}{n-1}\sum_{i=1}^{n}\mathrm{diag}\left[\left(X-\mathbf{1}_n\bar{x}^T\right)\left(X-\mathbf{1}_n\bar{x}^T\right)^T\right]=\sum_{i=1}^{n}\mathrm{diag}(S)=\sum_{i=1}^{n}s_{ii}=\mathrm{tr}(S)=\mathrm{Var}(X)$$

(3.4.12)

The projection of each object vector $\mathbf{x}_i$ onto the new axis is the one-dimensional vector y_{1i} as defined by equation (3.4.5). Similarly, as in the m-dimensional space, suppose that the centre of the one-dimensional subspace, is located at $\bar{y}_1$, the mean of the vector $\mathbf{y}_1$. The squared length of y_{1i} is the squared Euclidean distance of each projected point from the centre of the one-dimensional subspace :

$$f_i^2=\left(y_{1i}-\bar{y}_1\right)^2 \qquad (3.4.13)$$

which can be written as (equation 3.4.5) :

$$f_i^2=\left(y_{1i}-\bar{y}_1\right)^2=\left(\mathbf{a}_1^T\mathbf{x}_i-\mathbf{a}_1^T\bar{\mathbf{x}}\right)\left(\mathbf{a}_1^T\mathbf{x}_i-\mathbf{a}_1^T\bar{\mathbf{x}}\right)^T=\mathbf{a}_1^T\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T\mathbf{a}_1 \qquad (3.4.14)$$

Summing over all the objects and dividing by (n-1) it follows that, the term on the left-hand side is the total variance, $\mathrm{Var}\left(\mathbf{y}_1\right)$, of the vector $\mathbf{y}_1$, which is the projection of the m-variate data set X, onto the one-dimensional space :

$$\frac{1}{n-1}\sum_{i=1}^{n}f_i^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{1i}-\bar{y}_1\right)^2=\mathrm{Var}\left(y_1\right) \qquad (3.4.15)$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}\left[\mathbf{a}_1^T\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^T\mathbf{a}_1\right]$$

$$= \mathbf{a}_1^T \left[\frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \right] \mathbf{a}_1 = \mathbf{a}_1^T S \mathbf{a}_1 \tag{3.4.16}$$

where S is the sample variance-covariance matrix. Note that vector $\mathbf{a}_1$ is constant. Finally, the displacement of each object, e_i, is the residual of the projection of the i-th point from m-dimensional space ($\mathbf{x}_i$) onto the one-dimensional subspace (y_{1i}). Dividing equation (3.4.8) by (n-1) and replacing the corresponding terms of the sum of squared Euclidean distances by equations (3.4.12), (3.4.15) or (3.4.16), it follows that :

$$\frac{1}{n-1}\sum_{i=1}^{n} d_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i^2 + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \tag{3.4.17}$$

or

$$\sum_{i=1}^{n} s_{ii} = \mathbf{a}_1^T S \mathbf{a}_1 + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \tag{3.4.18}$$

that is

$$\mathrm{Var}(X) = \mathrm{Var}(y_1) + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \tag{3.4.19}$$

It can be seen that, for the purpose of determining the direction cosines $\mathbf{a}_1$ and, therefore, the placement of the new axis, the minimisation of the sum of squared displacements of the n objects is equivalent to the maximisation of the sum of squared lengths of their projections y_{1i}, since the total sample variance of X is constant. This problem can be stated as the maximisation of $\mathbf{a}_1^T S \mathbf{a}_1$ with respect to $\mathbf{a}_1$ and subject to $\mathbf{a}_1^T \mathbf{a} = 1$. It is a constrained optimisation problem and can be solved by applying the *Lagrange multiplier* technique (Krzanowski, 1988; Jolliffe, 1986; Basilevsky, 1994).

The Lagrangean expression can be written as :

$$\phi_1 = \mathbf{a}_1^T S \mathbf{a}_1 - \lambda_1(\mathbf{a}_1^T \mathbf{a}_1 - 1) \qquad (3.4.20)$$

where λ_1 denotes the undetermined Langrange multiplier. Differentiating with respect to $\mathbf{a}_1$ and setting the resultant equation to zero, it follows that :

$$\frac{\partial \phi_i}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0 \Rightarrow S\mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \Rightarrow (S - \lambda_1 I_m)\mathbf{a}_1 = 0 \qquad (3.4.21)$$

where I_m is an $(m \times m)$ identity matrix. Equation (3.4.21) is a set of m homogeneous equations with m unknowns. According to the theory of equations, a non-trivial solution $(\mathbf{a}_1 \neq 0)$ can be obtained when :

$$|S - \lambda_1 I_n| = 0 \qquad (3.4.22)$$

It therefore follows that λ_1 is an eigenvalue of the sample variance-covariance matrix S and the solution $\mathbf{a}_1$ is its corresponding normalised eigenvector ($\mathbf{a}_1^T \mathbf{a}_1 = 1$). There are m eigenvalues that provide the solution to the system of equations (3.4.22). However, by pre-multiplying equation (3.4.21) by $\mathbf{a}_1^T$, it follows that :

$$S\mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \Rightarrow \mathbf{a}_1^T S \mathbf{a}_1 = \lambda_1 \mathbf{a}_1^T \mathbf{a}_1 = \lambda_1 \underset{(3.4.16)}{\overset{(3.4.15)}{\Rightarrow}} \mathrm{Var}(y_1) = \lambda_1 \qquad (3.4.23)$$

that is, the eigenvalue λ_1 equals the total variance of the projections, which has to be maximised. Therefore, λ_1 must be chosen to be the largest eigenvalue of S. Consequently, the eigenvector $\mathbf{a}_1$ that contain the direction cosines of the new axis, can be determined and the one-dimensional subspace, upon which the multivariate sample is projected, can be defined.

Suppose now that, the compression of the m-dimensional space onto one dimension is not satisfactory due to the residuals, e_i, being large. One can then project the multivariate sample onto a two-dimensional subspace, that defines a new axis v_2, which is orthogonal to v_1 :

$$v_2 = a_{21}\mathbf{u}_1 + a_{22}\mathbf{u}_2 + ... + a_{2m}\mathbf{u}_m \tag{3.4.24}$$

where a_{2j} is the direction cosine of v_2 relatively to $\mathbf{u}_j$ and which satisfies the following conditions :

$$\sum_{j=1}^{m} a_{2j} = 1 \Rightarrow a_2^T a_2 = 1 \tag{3.4.25}$$

$$a_2^T a_1 = 0 \tag{3.4.26}$$

since the two axes are mutually orthogonal.

The m-variate data sample can the be approximated by the orthogonal projection of the n points onto v_2, which are represented by a set of one-dimensional vectors y_{2i} (i=1,...,n) :

$$y_{2i} = a_{21}x_{1i} + a_{22}x_{2i} + ... + a_{2m}x_{mi} = \mathbf{a}_2^T\mathbf{x}_i \tag{3.4.27}$$

Using a similar reasoning to that used when defining the first axis, one wants to determine the direction cosines $\mathbf{a}_2$ of the new axis, so that the displacement caused by the projection of the objects upon the new axis is minimised. This can be achieved by maximising the total variance of the vector $\mathbf{y}_2$, which contains the projections y_{2i} (i=1,...,n) of the objects $\mathbf{x}_i$ points onto the new one-dimensional space. The problem can be stated as the maximisation of $\mathbf{a}_2^T S \mathbf{a}_2$ with respect to $\mathbf{a}_2$, subject to $\mathbf{a}_2^T \mathbf{a}_2 = 1$ and $\mathbf{a}_2^T \mathbf{a}_1 = 0$. The Lagrangean expression under consideration is :

$$\phi_2 = \mathbf{a}_2^T S \mathbf{a}_2 - \lambda_2\left(\mathbf{a}_2^T \mathbf{a}_2 - 1\right) - \kappa\left(\mathbf{a}_2^T \mathbf{a}_1\right) \tag{3.4.28}$$

where λ_2, κ are the two unknown Lagrange multipliers. Differentiating with respect to $\mathbf{a}_2$ and setting the resultant equation to zero, it follows that :

$$\frac{\partial \phi_2}{\partial \mathbf{a}_2} = 2S\mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 - \kappa \mathbf{a}_1 = 0 \Rightarrow \left(S - \lambda_2 I_m\right)\mathbf{a}_2 = \frac{1}{2}\kappa \mathbf{a}_1 \tag{3.4.29}$$

Pre-multiplying (3.4.29) by $\mathbf{a}_1^T$, it can be seen that :

$$\mathbf{a}_1^T S \mathbf{a}_2 - \mathbf{a}_1^T \lambda_2 I_m \mathbf{a}_2 = \frac{1}{2}\kappa \mathbf{a}_1^T \mathbf{a}_1 \Rightarrow \mathbf{a}_1^T S \mathbf{a}_2 = \frac{1}{2}\kappa \tag{3.4.30}$$

Furthermore, pre-multiplying (3.4.21) by $\mathbf{a}_2^T$, it follows that :

$$\mathbf{a}_2^T S \mathbf{a}_1 = \lambda_1 \mathbf{a}_2^T \mathbf{a}_1 \Rightarrow \mathbf{a}_2^T S \mathbf{a}_1 = 0 \tag{3.4.31}$$

Since S is a square symmetric matrix, then

$$\mathbf{a}_2^T S \mathbf{a}_1 = \mathbf{a}_1^T S \mathbf{a}_2 = 0 \tag{3.4.32}$$

As a result, the second Lagrange multiplier in equations (3.4.30) and (3.4.29) is equal to zero ($\kappa = 0$), whilst the first Lagrange multiplier, λ_2, is again an eigenvalue of the sample variance-covariance matrix S and the solution $\mathbf{a}_2$ is its corresponding normalised eigenvector :

$$\left|S - \lambda_2 I_n\right| = 0 \tag{3.4.33}$$

Furthermore, λ_2 accounts for the maximum variation of $\mathbf{y}_2$:

$$\mathrm{Var}(\mathbf{y}_2) = \lambda_2 \tag{3.4.34}$$

and it corresponds to the second largest eigenvalue of S, since λ_1 accounts for the maximum variation of $\mathbf{y}_1$ and is the largest eigenvalue of S. The direction cosines $\mathbf{a}_2$ can consequently be calculated. The above procedure can be continued, up to the rank of the data matrix X. In the case when X is a full-rank matrix, m new axes can be defined and, thus, this procedure can be viewed as an orthogonal rotation of the original axes.

The procedure previously described is Principal Component Analysis (PCA). The new axes are termed *principal components*. The direction cosines $\mathbf{a}_r$ (r=1,...,R) are the *loading vectors* ($\mathbf{p}_r$), while the vectors of projections $\mathbf{y}_r$ of the m-variate data set X onto the new defined axes are the *score vectors* ($\mathbf{t}_r$). Usually a small number of principal components are extracted, since the primary objective of PCA is to compress the multivariate data set by projecting the m-dimensional space down onto a low-dimensional subspace.

A generalised model for Principal Component Analysis, can be described as follows. Consider that the mean $\overline{\mathbf{x}}$ of the object vectors $\mathbf{x}_i$ is equal to zero, since X has been mean-centred beforehand. Furthermore, the scalar quantity (n-1) defines the number of degrees of freedom of the multivariate sample and it can be omitted. Therefore the analysis can be based upon $X^T X$ instead of the sample variance-covariance matrix S. PCA can, thus, be considered as the *orthonormal projection* of the m-dimensional space onto a low-dimensional subspace and each principal component can be considered a linear combination of the original variables :

$$T = XP \tag{3.4.35}$$

where T is an $(n \times R)$ matrix whose columns are the score vectors $\mathbf{t}_r$ and P is an $(m \times R)$ matrix whose columns are the loading (direction cosines) vectors $\mathbf{p}_r$. A solution to the system of linear equations (3.4.35) must be found so that the variation of principal components is maximised :

$$\mathrm{Var}(T) = \mathrm{Var}(XP) = E\left(P^T X^T XP\right) = P^T E\left(X^T X\right) P = P^T X^T XP \tag{3.4.36}$$

Thus, the problem can be stated as the maximisation of $P^T X^T XP$ with respect to P and subject to the constraint of orthonormality, i.e. $P^T P = I_R$. This is equivalent to maximising the Lagrangean expression :

$$\Phi = P^T X^T XP - \Lambda\left(P^T P - I_R\right) \tag{3.4.37}$$

where Λ is a diagonal matrix of Lagrange multipliers. From the maximisation it follows that the Lagrange multipliers are the eigenvalues of $X^T X$ and P is a matrix, whose columns are the corresponding eigenvectors :

$$\left(X^T X - \Lambda\right) P = 0 \tag{3.4.38}$$

It can be stated that the eigenvalues of $X^T X$ are ordered in strictly decreasing order when $X^T X$ is non-singular $\left(\left|X^T X\right| \neq 0\right)$ (Basilevsky, 1994) :

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \tag{3.4.39}$$

Having obtained the principal components, the multivariate data set X can be reconstructed by means of equation (3.4.2) :

$$X = TP^T$$

The previous reconstruction holds only when all the principal components have been extracted (R=m). In any other case (R<m), the reconstruction leads to an estimate of the multivariate data set, denoted as $\hat{X}$:

$$\hat{X} = TP^T \tag{3.4.40}$$

which is linked to the original X by the reconstruction error or residual E as :

$$X = \hat{X} + E \tag{3.4.41}$$

or

$$E = X - \hat{X} = X - TP^T = X - XPP^T = X\left(I - PP^T\right) \tag{3.4.42}$$

Principal Components Analysis (PCA) has also been related to the Spectral Decomposition of the variance-covariance matrix S or $X^T X$ (Jolliffe, 1986; Flury, 1988). According to the spectral decomposition theorem, a square symmetric and positive definite matrix Γ can be decomposed as :

$$\Gamma = B\Lambda B^T \tag{3.4.43}$$

where Λ is a diagonal square matrix that contains the eigenvalues of Γ and B is an orthogonal matrix containing the associated normalised eigenvectors of Γ. Since PCA calculates the eigenvalues of S or X^TX and the associated eigenvectors P that are orthonormal ($P^TP = I_R$), then from equation (3.4.38) it follows that PCA is the spectral decomposition of X^TX (or S) :

$$(n-1)S = X^TX = P\Lambda P^T \tag{3.4.44}$$

Furthermore, PCA has been associated with the Singular Value Decomposition (SVD) of the data matrix X (Krzanowski, 1988; Jolliffe, 1986; Basilevsky, 1994). The SVD of a real matrix as the multivariate data set X is given by :

$$X = U\Sigma V \tag{3.4.45}$$

where U is an $(n \times R)$ matrix, which has the normalised eigenvectors of XX^T as its columns, V is an $(R \times n)$ matrix, which has the normalised eigenvectors of X^TX as its columns and Σ is an $(R \times R)$ diagonal matrix having the positive square roots of the ordered eigenvalues of X^TX as its diagonal elements $\left(s_r = \sqrt{\lambda_r}\right)$. Therefore, it can seen that both methods are related by matching equations (3.4.2) and (3.4.45). The loadings are the columns of the matrix V and the scores are the columns of the matrix $U\Sigma$ (Wold, 1987). Therefore, equation (3.4.45) can be written element by element as:

$$x_{ij} = \sum_{r=1}^{R} u_{ir} s_r v_{rj} \tag{3.4.46}$$

where u_{ir} and v_{rj} are elements of V and U matrices respectively

3.4.2. Properties of PCA and Other Considerations

A number of important properties of principal components can be extracted from their derivation. These properties can be summarised as follows :

1. The total variance which accounts for r-th principal component is the eigenvalue of the covariance matrix (3.4.23 and 3.4.34) :

$$\mathrm{Var}(\mathbf{t}_r) = \lambda_r \tag{3.4.47}$$

2. In the case of full decomposition of the multivariate data set X to principal components, the total variance of the R principal components is equal to the total variance of the original variables. When X is fully decomposed, the reconstruction error is equal to zero (E=0). Therefore, equation (3.4.19) can be rewritten as :

$$\mathrm{Var}(X) = \mathrm{Var}(T) + \mathrm{Var}(E) \underset{(3.4.47)}{\overset{(3.4.12)}{\Rightarrow}} \mathrm{tr}(S) = \sum_{r=1}^{R} \lambda_r \tag{3.4.48}$$

Consequently, the r-th principal components accounts for a proportion of the total variation of the original data set :

$$\pi = \frac{\lambda_r}{\mathrm{tr}(S)} \tag{3.4.49}$$

Furthermore, the first r_1 components account for a proportion of the total variation :

$$\Pi = \frac{\sum_{r=1}^{r_1} \lambda_r}{\mathrm{tr}(S)} \tag{3.4.50}$$

4. The score vectors of the principal components are orthogonal and measure different underlying latent structures in the data, whilst the loading vectors are orthonormal and define the directions of maximum variability :

$$\mathbf{t}_i^T \mathbf{t}_j = 0, \quad i \neq j \tag{3.4.51}$$

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \tag{3.4.52}$$

5. No standardised linear combination of $\underline{x}$ has a greater variance than $\mathbf{t}_1$ (3.4.39) :

$$\mathrm{Var}(\mathbf{t}_1) \leq \mathrm{Var}(\mathbf{t}_2) \leq \ldots \leq \mathrm{Var}(\mathbf{t}_R) \tag{3.4.53}$$

6. The principal components may vary depending upon the term of scaling.

Property (6) is of practical importance. Inappropriate scaling can affect the apparent relationships between the variables. The loading vectors provide the direction of maximum variability, which represents the variation that is common to all the objects. Therefore, by mean-centring the data, basic underlying variation is removed before the data has even been analysed using PCA. Furthermore, in the case where the original variables are measured in different units, the structure of the principal components depend upon the essentially arbitrary choice of measurements units and, therefore, variables with a large variance will tend to dominate the first principal component. The application of variance-scaling (normalisation to unit variance) in addition to mean-centring overcomes this difficulty. In this case, the principal components are extracted from the sample *correlation* matrix R, rather that the variance-covariance matrix S.

An important issue in Principal Components Analysis is the interpretation of the derived principal components, since this can help to determine the variables that have the greatest impact upon the variance of a particular component (Mardia, et al., 1974). The *coefficient of correlation* between the r-th principal component and the j-th variable is defined as :

$$\rho_{r,j} = \frac{p_{r,j} \cdot \sqrt{\lambda_j}}{\sqrt{s_{jj}}} \qquad (3.4.54)$$

and the *coefficient of determination* is given by :

$$\rho^2_{r,j} = \frac{p^2_{r,j} \cdot \lambda_r}{s_{jj}} \qquad (3.4.55)$$

where $p_{r,j}$ is the element of the eigenvector $\mathbf{p}_r$ for the j-th original variable and s_{jj} is the variance of the j-th variable. The coefficient of determination is the ratio of the estimated variance of the j-th variable to its actual variation.

The major advantage of modelling a multivariate data set in terms of principal components is the ease of visualisation of the multivariate data set. Specifically, the information contained in the original data set can be described in terms of two plots :

1. *Loading plot.* The loading vectors ($\mathbf{p}_r$) provide a picture of the relationships between the variables. One can infer the relative importance and influence of the original variables by observing the absolute values of the elements of the loadings. Furthermore, similarities between variables are evaluated in terms of the angle between the loadings and the sign of the co-ordinates on each component.

2. *Principal Component Scores plot.* These describe the relationships between the objects and the principal components. One can infer which variables are indicative of changes in the data set by observing the changes in the scores and knowing the relationships defined by the loadings. Furthermore, similarities in the scores plot are evaluated in terms of the angle between their object vectors ($\mathbf{x}_i$) and the distances between objects.

A further issue associated with the interpretation of the principal components is that of the approximation of each principal component (PC) as a linear combination of a subset of the original variables. The traditional practise is to select such subsets from their corresponding loading vectors and, therefore, from their correlation coefficients. For each PC, variables that have low loadings are discarded and the remaining subset is used to provide a linear combination, called a *truncated* principal component that approximates the original PC :

$$\mathbf{t}_{t,r} = X_t \mathbf{p}_t^T \tag{3.4.56}$$

where $t_{t,r}$ is the truncated PC that corresponds to the r-th original PC, X_t is the subset of the multivariate data set X that contains only observations for the retained variables and $\mathbf{p}_t$ the truncated vector of loadings that contains the loadings of the retained variables. Truncated principal components can be used to assess whether all variables or some subset of them provide meaning to a principal component, relevant to the problem under study. *Truncated PCA* can be usually applied to data sets comprising measurements on a great number of variables. Cadima and Jolliffe (1995) suggested that the loadings are not reliable for determining whether a subset of the m original variables are acceptable for defining a truncated principal component. The main reason for this is that it is both the loadings and the standard deviation of each variable, which determines the importance of a variable in the linear combination. As an alternative, they suggested regressing the principal components on the subsets of variables and using the multiple correlation coefficients as a criterion to whether or not to retain a particular variable. The multiple correlation coefficient r_m between the j-th principal component and a subset of *k* original variables is given by :

$$r_m = \sqrt{\lambda_j \left(\mathbf{p}_j^{kT} S_k^{-1} \mathbf{p}_j^k\right)} \qquad (3.4.57)$$

where λ_j is the eigenvalue of the j-th PC, $\mathbf{p}_j^k$ is a vector containing only those elements of the loading vector of the j-th PC that are associated with the k retained variables and S_k is a sub-matrix of the variance-covariance matrix of X which can be obtained by retaining only those columns of S that correspond to the retained variables. Furthermore, the correlation coefficient r_t between the j-th original PC and a truncated PC based on a subset *k* of original variables is given by :

$$r_t = \frac{\sqrt{\lambda_j} \mathbf{p}_j^{kT} \mathbf{p}_j^k}{\sqrt{\left(\mathbf{p}_j^{kT} S_k \mathbf{p}_j^k\right)}} \qquad (3.4.58)$$

The ratio r_t/r_m is the correlation between the truncated PC and the projection of the original PC onto the subset *k* of original variables. When $r_t/r_m = 1$ the truncated and the original principal components coincide and, therefore, the subset of variables may be retained while the remaining variables should be discarded in an approximation procedure.

From the analysis presented in the previous paragraphs, it can be seen that, Principal Components Analysis can be considered from three different perspectives as :

1. A technique to determine the principal axes that defines the direction of maximum variation, using the classical statistical approach (Hotelling, 1933).
2. A specific type of factor analysis (Harman, 1976; Cattell, 1978).
3. A technique for describing a data set under certain optimised algebraic and geometric criteria and, therefore, as a technique for data reduction (Pearson, 1901).

3.4.3 Calculation of Principal Components From a Multivariate Sample

The most popular method to calculate the principal components from a multivariate data set is the Non-linear Iterative Partial Least Squares algorithm (NIPALS) (Wold, 1987; Geladi and Kowalski, 1986; Martens and Naes, 1989)

The NIPALS algorithm does not simultaneously calculate all the principal components. It calculates the first principal component and then, it subtracts the outer product of its score and loading vectors from the data matrix X. The residual matrix is then used to calculate the second principal component and so on. The NIPALS algorithm is a fast and effective algorithm to extract the principal components in a sequential manner. It is also a variant of the power method for calculating the eigenvectors of a matrix (Goldberg, 1991). Using the NIPALS algorithm the score and the loading vectors are the eigenvectors of the $X \cdot X^T$ and $X^T \cdot X$ matrices, respectively (Geladi and Kowalski, 1986).

The NIPALS algorithm is as follows :

(1) h=0, $\mathbf{t}_h = 0$, $\mathbf{p}_h^T = 0$, $E_{h-1} = X$.

(2) h=h+1.

(3) The column vector $\mathbf{x}_j$ with the maximum variance is selected from the E_{h-1} matrix and defined to be by $\mathbf{t}_h$ (for the first iteration it does not matter which column vector is selected since every column has variance equal to unity).

(4) $\mathbf{p}_h^T = \mathbf{t}_h^T \cdot E_{h-1} / \left(\mathbf{t}_h^T \cdot \mathbf{t}_h\right)$

(5) Normalise $\mathbf{p}_h^T$ to length 1 : $\mathbf{p}_h^T = \mathbf{p}_h^T / \left\|\mathbf{p}_h^T\right\|$

(6) $\mathbf{t}_{h,new} = E_{h-1} \cdot \mathbf{p}_h / \left(\mathbf{p}_h^T \cdot \mathbf{p}_h\right)$

(7) If the score $\mathbf{t}_h$ of step (3) converges with that of step (6), then go to step (8), else go to step (4).

(8) $E_h = E_{h-1} - \mathbf{t}_h \cdot \mathbf{p}_h^T$

(9) Go to step 2.

As a convergence criterion, in step (7), the sum of squared differences is frequently used :

$$\sum_{h=1}^{n} \left(\mathbf{t}_{h,new,i} - \mathbf{t}_{h,i}\right)^2 \leq e \qquad (3.4.59)$$

3.4.3 Selection of the Optimal Number of Principal Components

An important issue associated with the application of Principal Component Analysis for the purpose of reducing the dimensionality in a multivariate data set, is to determine the optimal number of principal components (R) that are required to adequately account for the variation in the data set. With highly correlated variables, the first few principal components explain most of the variability present in the data.

The remaining principal components are not significant and typically explain the noise in the data. A number of criteria have been proposed in the literature to select the number of principal components, e.g. Jolliffe, (1986), Jackson, (1993), Ferre, (1995). A few of these approaches are based upon statistics, however, most of them are based upon heuristic approaches.

Heuristic Approaches

1. *Cumulative Percentage of Total Variation.* The simplest criterion is to retain only those principal components (R) that account for an arbitrary selected proportion (α) or a cumulative percentage ($100\alpha\%$) of the total variation in the multivariate data set. This is an *ad hoc* procedure and is unreliable. It is not recommended by many statisticians (Jolliffe, 1986; Jackson, 1991). Usually, a cumulative percentage of between 80% and 90% (i.e. $0.8 \le \alpha \le 0.9$) is defined as being optimal :

$$0.8 \le \frac{\sum_{r=1}^{R} \lambda_r}{\sum_{j=1}^{m} \lambda_j} \le 0.9 \qquad (3.4.60)$$

2. *Amount of Variance Explained by an Individual Principal Component.* If all original variables x_j are independent, then the principal components are the same as the original variables (e.g. PC_1=variable$_1$, PC_2=variable$_2$,...). Thus, principal components with variation (associated eigenvalue) less than or equal to one should be excluded, since they contain less information than one of the original variables. This is also an ad-hoc criterion and it is called *Kaiser's criterion* (Kaiser, 1960):

$$\lambda_R \le 1 \qquad (3.4.61)$$

An extension of Kaiser's criterion is the *Kaiser-Guttman* criterion, which suggests that a principal component should be retained only if its variation (λ_j) is greater than or equal to the average variation ($\overline{\lambda}$) of the principal components or a proportion (α) of it (Guttman, 1954; Cliff, 1988) :

$$\lambda_R \geq \alpha\overline{\lambda} \qquad ; 0.8 \leq \alpha \leq 0.9 \tag{3.4.62}$$

Recently, a modified Kaiser-Guttman criterion that does not ignore the error associated with the individual eigenvalues due to sampling has been proposed (Lambert, et al., 1990). This criterion involves the use of the *bootstrap* (Efron, 1979) to determine the confidence limits of the eigenvalues and testing whether the Kaiser-Guttman criterion lies within these limits, the *Bootstrap Kaiser-Guttman* criterion.

An alternative approach is the *Broken Stick* criterion (Frontier, 1976; Legendre, 1983). The idea behind is that, if one has a line of unit length, which is randomly divided into R segments then, it can be shown that, the expected length of the r-th longest segment is:

$$g_r = \frac{1}{R}\sum_{j=r}^{R}\frac{1}{j} \tag{3.4.63}$$

Thus, considering the line to be the total variation in the data set, the Broken Stick criterion suggests that the r-th principal component should only be retained if :

$$g_r \leq \lambda_r \tag{3.4.64}$$

This is a crude criterion and it must only be applied to unit variance-scaled matrices. Finally, another common method is the *Scree method.* This is where the value or the logarithm of the value of successive eigenvalues are plotted in rank order. The smallest eigenvalues typically represent random noise and tend to lie in a straight line, whilst the large eigenvalues move away from this line. Principal components up to those whose eigenvalue lie to the right of the point where the largest eigenvalue departs from the straight line, should be retained (Cattell, 1966; Cattell and Vogelmann, 1977).

Statistical Approaches

3. *Test of sphericity.* Bartlett's test of sphericity (Cooley and Lohnes, 1971) evaluates whether each consecutive eigenvalue is significantly different from the remaining eigenvalues. This test reveals the point where PCA summarises a spherical principal distribution of points. The test statistic follows a χ^2 distribution and the number of components R is selected so that :

$$(m-R)\ln\left[\sum_{r=R+1}^{m}\frac{\lambda_r}{m-R}\right]-\sum_{r=R+1}^{m}\lambda_r \le \frac{1}{(n-R)}\chi^2_{0.5(m-R-1)(m-R+2)} \qquad (3.4.65)$$

where m is the number of original variables, n is the number of objects.

4. *Tests for equality of the eigenvalues.* There are two tests that evaluate whether the first eigenvalue (Bartlett, 1954) or the second eigenvalue (Lawly, 1956 and 1963) of the correlation matrix is equal to the remaining set of eigenvalues. Both tests have limited application, since they only examine the significance of the first and the second eigenvalues. However, they do provide an assessment of the overall PCA.

5. *Partial Correlation.* Velicer (1976) suggested that the average of the squared partial correlations between the variables, given the values of the first r principal

components, may be used to determine the number of principal components that should be retained. It can only be applied when the sample correlation matrix has been used. The criterion is given by :

$$V_r = \sum_{\substack{i=1 \\ i \neq j}}^{r} \sum_{j=1}^{r} \frac{\left(\rho'_{ij}\right)^2}{m(m-1)} \tag{3.4.66}$$

where m is the number of variables, ρ'_{ij} is the partial correlation between the i-th and the j-th variable given the first r principal components, and is defined as the correlation between the residuals from the regression of the i-th variable on the first r principal components and the residuals from the regression of the j-th variable on the first r principal components. This criterion measures the strength of the linear relationship between the i-th and the j-th variable after removing the common effect of the first r principal components (Jolliffe, 1986). The optimum number of principal components R corresponds to the minimum V_r. This criterion has been applied successfully to select the number of factors to be retained in Factor Analysis, but it is inappropriate in PCA whenever a principal component is dominated by a single variable that is uncorrelated with the other variables.

6. *Cross-Validation.* The concept in cross-validation is that a subset of the multivariate data sample can be predicted satisfactorily by a statistical model that was built with it not included. Cross-validation methods have been suggested for PCA by Wold (1978) and Eastman and Krzanowski (1982). Both methods utilise the Prediction Sum of Squares (PRESS) proposed in regression by Allen (1974), which is the sum of squared differences between the predicted and the observed values of

the subset, but they differ in how a subset is chosen and how the PRESS is used for choosing the optimum number of principal components.

Wold (1978) suggested that an m-variate data sample of n objects can be divided into G subsets, X_g (g=1,...,G). Each individual subset X_g is then excluded :

$$X = X_g \cup X_g^- \tag{3.4.67}$$

and a PCA is performed on the remaining subset X_g^-. The resulting loadings P_g^- can then be used to calculate the scores (T_g) of the excluded subset, X_g, according to PCA :

$$T_g = X_g P_g^- \tag{3.4.68}$$

Predictions of X_g can then be obtained by retaining an increasing number (r=1,2,...) of principal components in the PCA model each time :

$$\hat{X}_{g,r} = \sum_{k=1}^{r} \mathbf{t}_{g,k} \mathbf{p}_{g,k}^{-T} \tag{3.4.69}$$

where $\mathbf{t}_{g,k}$ is the k-th column vector of the estimated matrix of scores T_g of the subset X_g and $\mathbf{p}_g^-$ is the k-th column vector of the matrix of loadings P_g^- of the subset X_g^-. The Squared Prediction Error (SPE) can be calculated for each number (r) of retained principal components in the model :

$$\mathrm{SPE}_{g,r} = \sum_{i=1}^{n_g} \sum_{j=1}^{m} \left(x_{ij} - \hat{x}_{ij} \right)^2 \tag{3.4.70}$$

where n_g is the number of objects included in the subset X_g and x_{ij}, $\hat{x}_{ij}$ denote the observed and predicted value of an element of the subset X_g, respectively. When the calculations for all the subsets are completed, the Prediction Sum of Squares can be calculated for each of number of retained principal components :

$$\mathrm{PRESS}_r = \sum_{g=1}^{G} \mathrm{SPE}_{g,r} \tag{3.4.71}$$

To decide whether the r-th principal component should be retained, Wold suggested examining the ratio :

$$R_W = \frac{\mathrm{PRESS}_r}{\mathrm{RSS}_{r-1}} \tag{3.4.72}$$

where RSS (Residual Sum of Squares) is the difference between the observed and predicted values of the complete data set, X, and can be calculated for each principal component (r) as :

$$\mathrm{RSS}_r = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} - \sum_{h=1}^{r} t_{ih} p_{hj} \right)^2 \tag{3.4.73}$$

where t_{ih} denotes the score of the i-th object on the h-th principal component and p_{hj} the loading of the j-th variable for the h-th principal component ($h \leq r$). The ratio R_W compares the predictive power of a model based upon r principal components with the squared difference between the observed and predicted data using (r-1) principal components. An R_W value greater than unity suggests that the predictive power of the model has not been improved by adding the r-th principal component and its better to retain (r-1) principal components. Wold suggested that the original data set should be divided into between 4 and 7 subsets, i.e. G=4 or 7 and that G must not be a divisor of the number of variables (m).

Eastman and Krzanowski, (1982), proposed that as much of the original data set X as possible should be used to predict each of the subsets, e.g, the size of the subset must be as small as possible. The smallest subset that can be excluded from an m-variate data sample is a single observation x_{ij} and, according to Eastman and Krzanowski, it should be predicted from all the data except the i-th object (row) and the j-th variable (column) of X. Suppose now that X^I denotes the subset where the i-th row has be excluded, and X^J denotes the data set where the j-th variable has been excluded. Applying a PCA to each of these data set and using the SVD method, it follows that :

$$X^I = U^I \Sigma^I V^I \qquad (3.4.74)$$

$$X^J = U^J \Sigma^J V^J \qquad (3.4.75)$$

Using SVD of the complete data set X, prediction for an element is given by means of equation (3.4.46) :

$$\hat{x}_{ij}^r = \sum_{k=1}^{r} u_{ik} s_k v_{kj} \qquad (3.4.76)$$

However, Eastman and Krzanowski suggested that the prediction of the part arising from U requires information on the i-th row and, therefore, U^I must be used, whilst prediction of the part arising from V requires information on the j-th column and, therefore, V^J must be used. For the central part Σ, it was suggested that information from both the i-th row and j-th columns is required and, therefore, a composite of the two should be used. Hence, the prediction of an element is given by :

$$\hat{x}_{ij}^{r} = \sum_{k=1}^{r}\left(u_{ik}^{I}\sqrt{s_{k}^{I}}\right)\left(\sqrt{s_{r}^{J}}\,v_{kj}^{J}\right) \tag{3.4.77}$$

where the sign of $\left(u_{ik}^{I}\sqrt{s_{k}^{I}}\right)\left(\sqrt{s_{r}^{J}}\,v_{kj}^{J}\right)$ is equal to the sign of $u_{ik}s_{k}v_{kj}$ from the decomposition of the complete data set for each principal component. The Prediction Sum of Squares can be calculated for each principal component as :

$$\mathrm{PRESS}_{r} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \hat{x}_{ij}^{r}\right) \tag{3.4.78}$$

The optimal number of principal components that should be retained is then determined by the statistic :

$$W = \frac{\left(\mathrm{PRESS}_{r-1} - \mathrm{PRESS}_{r}\right)/D_{m}}{\mathrm{PRESS}_{r}/D_{r}} \tag{3.4.79}$$

where D_m is a number indicating the degrees of freedom required to fit the r-th principal component and D_r is a number indicating the degrees of freedom remaining after fitting the r-th principal component :

$$D_{m} = n + m - 2r \tag{3.4.80}$$

$$D_{r} = m\cdot(n-1) - \sum_{i=1}^{r} i + m - 2r \tag{3.4.81}$$

The value of W gives the ratio between the improvement in the predictive power of the model achieved by adding the r-th principal component, to the predictive power of the model based upon r principal components. A value of W greater than unity suggests that the r-th principal component should be included in the model. This test is similar to an F-test for the inclusion of an additional variable in a linear regression model.

None of the criteria presented above provide a unique solution to the problem of selecting the optimal number of components that should be retained in a PCA model. Some of them are rules of thumb with no theoretical basis, some of them are more statistically acceptable and some of them are computationally intensive. Depending on the situation (i.e. size of data set, computational power and available time for analysis) the most appropriate criterion should be selected. Cross-validation is statistically more acceptable and, therefore, it should be performed in all situations. However, it is time-consuming and computationally intensive. On the other hand, Kaiser-Guttman criterion and Scree method are heuristic approaches but can quickly provide an assessment of the number of principal components to retain. However, the selection of the optimal number of principal components to retain into a PCA model, should be based upon the overall picture given by these criteria.

3.5 Projection to Latent Structures (PLS)

Projection to Latent Structures (PLS) or Partial Least Squares, has become a popular regression technique with a wide range of application for multivariate calibration problems. The power of PLS mainly comes from its ability to define independent latent variables from the covariance structure of given groups of highly correlated or collinear, real or observable variables. Thus, PLS can be viewed as a technique which can be used for both dimensionality reduction and modelling. PLS has its origins in the Non-linear Iterative Partial Least Squares (NIPALS) algorithm for general system-analysis models. It was proposed by Herman Wold (Wold, 1966; Wold, 1975). The basic mathematical and statistical background underpinning the method, has been described in a number of papers (Geladi and Kowalski, 1986; Martens and Naes, 1989; Wold, et al., 1984; Lorber, et al., 1987; Manne, 1987; Geladi, 1988; Helland, 1988; Hoskuldsson, 1988; Stone and Brooks, 1990; Garthwaite, 1994).

The objective of Projection to Latent Structures is to construct a linear relationship between two sets of data that contain observations from highly correlated variables. This is conceptually similar to Canonical Correlation Analysis (CCA). However, PLS selects linear combinations of the original variables in a way that eliminates redundancies in the data sets and defines a new set of variables, which are independent. PLS is, thus, similar to Principal Components Analysis (PCA), except that, PLS maximises the covariance of the two data sets, whilst PCA only maximises the variance of a single data set. PLS is based upon projecting the information contained in the high-dimensional space of the two data sets down onto low-dimensional subspaces, defined by the independent and latent variables. Therefore, the useful and relevant information contained in the large number of observable variables is summarised in terms of a small number of latent variables. The two data sets are typically denoted as the *predictor* (X) and the *response* (Y) or *independent* and *dependent* data sets, respectively.

PLS builds the regression relationship in a stepwise and sequential manner. There are several ways for achieving this, but the most common approach is the Non-linear Iterative Partial Least Squares (NIPALS) algorithm of Wold (1966). For each latent variable or dimension, the NIPALS algorithm calculates *two latent vectors*, $\mathbf{t}_i$ and $\mathbf{u}_i$, which are a linear combination of the predictor (X) and response (Y) data sets, respectively. These vectors define the latent dimension in each data set and are chosen such that the covariance between them is maximised. The NIPALS algorithm to perform PLS is as follows :

(1) Mean-centre and optionally variance scale the **X** and **Y** data sets.

(2) Set **u** equal to any column of the **Y** data set

(3) Regress the columns of **X** on **u** : $\mathbf{w}^T = \mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u}$

(4) Normalise the **w** vector to unit length

(5) Calculate the scores of **X** : $\mathbf{t} = \mathbf{Xw}/\mathbf{ww}^T$

(6) Regress the columns of **Y** on **t** : $\mathbf{q}^T = \mathbf{t}^T\mathbf{Y}/\mathbf{t}^T\mathbf{t}$

(7) Calculate the new scores of **Y** : $\mathbf{u} = \mathbf{Yq}/\mathbf{q}^T\mathbf{q}$

(8) If score **u** of step (7) converges with that of step (2), then go to step (9), else go to step (3)

(9) Calculate the loadings of **X** by regressing columns of **X** on **t** : $\mathbf{p}^T = \mathbf{t}^T\mathbf{X}/\mathbf{t}^T\mathbf{t}$

(10) Calculate the residual matrices **E** and **F** : $\mathbf{E} = \mathbf{X} - \mathbf{tp}^T$; $\mathbf{F} = \mathbf{Y} - \mathbf{tq}^T$

(11) To calculate an additional latent dimension, replace **X** and **Y** by **E** and **F** and repeat steps (2) - (10)

The sum of squared differences of the **u** vectors (equation 3.4.59) can be used as a convergence criterion.

As an alternative to the NIPALS algorithm, the maximum eigenvalue of the residual sample covariance matrix $\left(E_r^T F_r F_r^T E_r\right)$ or the successive Singular Value Decompositions of the cross-covariance matrix $\left(F_r^T E_r\right)$ of the residual data sets, can be used to perform the calculations of the latent dimensions (Hoskuldsson, 1988; Kaspar and Ray, 1993; Lindgren, et al., 1993; Wang, et al., 1994). Note that, in the beginning (r=0), residuals E_0 and F_0 do not exist and, therefore, are replaced by the original data sets X and Y, respectively. Using any of these methods, the predictor and the response data sets are decomposed as a series of latent variables, which can be written as a linear combination of the scores and loadings :

$$X = \sum_{r=1}^{R} \mathbf{t}_r \cdot \mathbf{p}_r^T + E \tag{3.5.1}$$

$$Y = \sum_{r=1}^{R} \mathbf{u}_r \cdot \mathbf{q}_r^T + F \quad \text{or} \quad Y = \sum_{r=1}^{R} b_r \cdot \mathbf{t}_r \cdot \mathbf{q}_r^T + F \tag{3.5.2}$$

Note that, each time the two data sets are decomposed in the score and loading vectors ($\mathbf{t}_r, \mathbf{p}_r$ and $\mathbf{u}_r, \mathbf{q}_r$ respectively) an inner relationship between the latent score vectors is built, whose coefficients are defined as:

$$b_r = \mathbf{u}_r^T \cdot \mathbf{t}_r \tag{3.5.4}$$

The final PLS regression model can then be written in terms of the latent vectors :

$$\hat{Y} = T \cdot Q^T \tag{3.5.5}$$

or alternatively, in terms of the original variables of the predictor data set :

$$\hat{Y} = X \cdot \hat{\beta} \tag{3.5.6}$$

where, $\hat{Y}$ are the estimated values of the response data set, T is a matrix whose columns are the score vectors ($\mathbf{t}_r$) of the predictor data set, P and Q are matrices whose columns are the loading vectors ($\mathbf{p}_r$ and $\mathbf{q}_r$) of the predictor and response data sets, respectively, W is a matrix whose columns are the weight vectors ($\mathbf{w}_r$) of the predictor data set, and $\hat{\beta}$ is the matrix of the linear regression coefficients :

$$\hat{\beta} = W \cdot (P^T \cdot W)^{-1} \cdot Q^T \tag{3.5.7}$$

Once a PLS regression model has been constructed and a new vector (**x**) of predictor data is available, predicted values ($\hat{\mathbf{y}}$) of the response variables can be obtained :

$$\hat{\mathbf{y}}^T = \mathbf{x}^T \cdot \hat{\beta} \tag{3.5.8}$$

3.5.1 Geometrical Interpretation and Mathematical Derivation of PLS

This section includes the derivation of the iterative procedure of the PLS technique, using mathematical, geometrical and statistical considerations and it describes the interpretation of the latent variables. Consider a multivariate data sample which contains n observations on (m+k) variables, which can partitioned into a predictor $(n \times m)$ data set X and a response $(n \times k)$ data set Y. The m-variate and k-variate data sets can be graphically modelled as swarms of n points in m-dimensional and k-dimensional object space, respectively. The co-ordinates of the n objects in these spaces are given by the data matrices X and Y, and the n corresponding points are represented by two sets of vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ (i=1,...,n), which are the rows of X and Y, respectively.

Suppose now that, one wants to find a relationship between the objects of the two data sets, that is to construct a regression model between the data set X and the data set Y. The statistical model can then be used both as a descriptive statistic and as a model for predicting future values of the response variables ($\hat{\mathbf{y}}$), when only values of the predictor variables are available ($\mathbf{x}_{new}$). The predictive relationships are assumed to be linear and, therefore, the regression model between the predictor and response data sets can be defined to be :

$$Y = X\beta + E \tag{3.5.9}$$

where β is an $(m \times k)$ matrix of regression coefficients. The predicted values of the future response vector $\hat{\mathbf{y}}$, given the corresponding predictor vector $\mathbf{x}_{new}$, are given by :

$$\hat{\mathbf{y}} = \beta^T \mathbf{x}_{new} \tag{3.5.10}$$

The most well known multivariate technique for calculating the matrix of regression coefficients is Multiple Linear Regression (MLR). However, a number of problems can be encountered when MLR is applied to data sets comprising highly correlated measurements. The derived regression coefficients β will typically have large variances and, hence, they will be unstable when small changes in the data occur. An extensive discussion of these problems can be found in Searle, (1977), Seber, (1977), Montgomery and Peck, (1992).

However, these problems can be overcome by eliminating the correlations, which exist between the original variables. This can be achieved by projecting the high-dimensional object spaces onto low-dimensional object subspace and retaining that part that is useful in identifying the relationships between the original variables. Assume initially that each high-dimensional space is projected down onto a one-dimensional subspace by the orthogonal transformation of their axes. The m-variate data set X can be approximated by the orthogonal projection of its n objects $\mathbf{x}_i$ (i=1,…,n) onto the one-dimensional subspace. Each object is then represented as a one-dimensional vector t_{1i} :

$$t_{1i} = w_{11}x_{1i} + w_{12}x_{2i} + \ldots + w_{1m}w_{mi} = \mathbf{x}_i^T\mathbf{w}_1 \tag{3.5.11}$$

The set of t_{1i} (i=1,…,n) comprises an n-dimensional vector $\mathbf{t}_1$ that is defined as :

$$\mathbf{t}_1 = X\mathbf{w}_1 \tag{3.5.12}$$

where $\mathbf{w}_1$ is an m-dimensional vector of the *weights* of the orthogonal transformation in X that are the direction cosines of $\mathbf{t}_1$ with respect to the orthogonal base of X. Similarly, the k-variate data set Y can be approximated by the orthogonal projections u_{1i} of its n objects onto the one-dimensional subspace :

$$u_{1i} = c_{11}y_{1i} + c_{12}y_{2i} + \ldots + c_{1k}y_{ki} = \mathbf{y}_i^T\mathbf{c}_1 \tag{3.5.13}$$

and, consequently, the vector $\mathbf{u}_1$ that contains the set of u_{1i} (i=1,…,n) is defined as :

$$\mathbf{u}_1 = Y\mathbf{c}_1 \tag{3.5.14}$$

where $\mathbf{c}_1$ is a k-dimensional vector containing the weights of the orthogonal transformation in Y that are the direction cosines of $\mathbf{u}_1$ with respect to the orthogonal base of Y. This procedure is illustrated in Figure 3.3, where a tri-variate data set X and a bivariate data set Y are projected down onto a one-dimensional subspace. For a particular object i, the corresponding one-dimensional projection vectors are t_{1i} and u_{1i}. In the case where the two vectors are parallel, they are linearly dependent and, therefore, each of them can be considered as a linear expansion or contraction of the other, such that :

$$u_{li} = b_{li} t_{li} \tag{3.5.15}$$

$$b_{li} = \frac{\|u_{li}\|}{\|t_{li}\|} \neq 0 \tag{3.5.16}$$

On the other hand, when these vectors are not parallel then, a *projection angle* θ is identified between t_{li} and u_{li}, so that each vector is linearly dependent with the projection of the other vector :

$$u^{p}_{li} = b_{li} t_{li} \tag{3.5.17}$$

where u^{p}_{li} denotes the projection of u_{li} onto t_{li} :

$$u^{p}_{li} = u_{li} \cos\theta \tag{3.5.18}$$

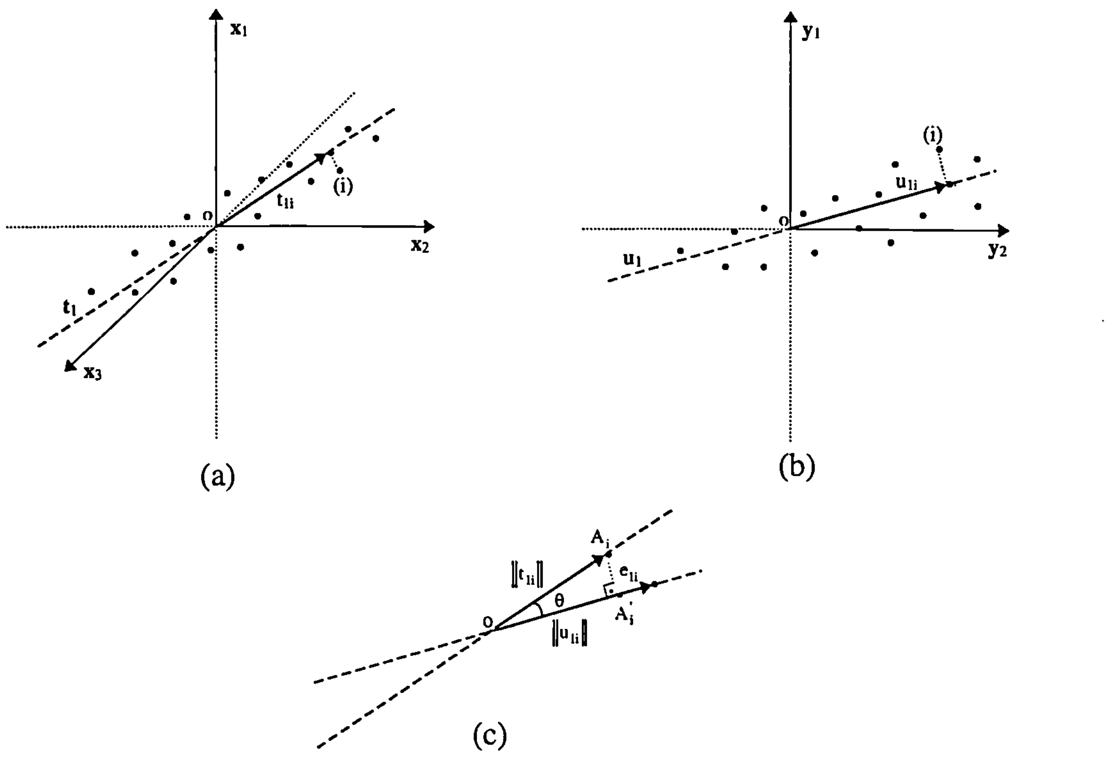


Figure 3.3. Geometrical interpretation of PLS

Equation (3.5.18) reduces to equation (3.5.15) when the cosine of the projection angle θ is equal to 1 or -1 or equivalently when θ is $\alpha\pi$ ($\alpha = 0,1,\ldots$). Applying the Pythagorean Theorem on triangle OA_iA_i', in Figure 3.3, it follows that :

$$\left\|t_{1i}\right\|^2 + \left\|u_{1i}\right\|^2 = e_{1i}^2 \tag{3.5.19}$$

where $\|\cdot\|$ denotes the length of a vector, and e_{1i} is the distance between t_{1i} and u_{1i}. Therefore, it can be concluded that, the best linear relationship between t_{1i} and u_{1i} can be obtained when the squared distance e_{1i}^2 is minimised. Using trigonometric rules, it can be seen that, the squared distance e_{1i}^2 is minimised when the squared cosine of projection angle θ is maximised :

$$e_{1i}^2 = \left\|t_{1i}\right\|^2 \sin^2\theta \Rightarrow \frac{e_{1i}^2}{\left\|t_{1i}\right\|^2} = 1 - \cos^2\theta \tag{3.5.20}$$

Since the projection angle is the same between all pairs of one-dimensional vectors (t_{1i}, u_{1i}), the previous conclusion can be generalised to the predictive relationship between the latent vectors $\mathbf{t}_1$ and $\mathbf{u}_1$. Furthermore, it can be seen that, equation (3.5.20) is invariant of the length of the latent vector. Since the latent vectors are orthogonal transformations of the original data sets then, one can select the set of weights $\mathbf{w}_1$ and $\mathbf{c}_1$ of the orthogonal transformations (3.5.12) and (3.5.14), respectively, so that the squared cosine of the projection angle is maximised. This is equivalent to rotating one of the latent vectors at an angle θ related to the other latent vector. The direction cosine of the projection angle is given by :

$$\cos\theta = \frac{\mathbf{t}_1^T \mathbf{u}_1}{\left\|\mathbf{t}_1\right\| \left\|\mathbf{u}_1\right\|} \tag{3.5.21}$$

It can be seen that, $\cos\theta$ is equal to the correlation between the latent vectors $\mathbf{t}_l$ and $\mathbf{u}_l$, since the inner product of these vectors is equal to their covariance and their lengths are equal to the square roots of their variances :

$$\operatorname{corr}(\mathbf{t}_l,\mathbf{u}_l)=\cos\theta=\frac{\mathbf{t}_l^T\mathbf{u}_l}{\|\mathbf{t}_l\|\,\|\mathbf{u}_l\|}=\frac{\operatorname{Cov}(\mathbf{t}_l,\mathbf{u}_l)}{\sqrt{\operatorname{Var}(\mathbf{t}_l)}\sqrt{\operatorname{Var}(\mathbf{u}_l)}} \qquad (3.5.22)$$

or equivalently :

$$\operatorname{corr}(\mathrm{X}\mathbf{w}_l,\mathrm{Y}\mathbf{c}_l)=\cos\theta=\frac{\mathbf{w}_l^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_l}{\|\mathrm{X}\mathbf{w}_l\|\,\|\mathrm{Y}\mathbf{c}_l\|}=\frac{\operatorname{Cov}(\mathrm{X}\mathbf{w}_l,\mathrm{Y}\mathbf{c}_l)}{\sqrt{\operatorname{Var}(\mathrm{X}\mathbf{w}_l)}\sqrt{\operatorname{Var}(\mathrm{Y}\mathbf{c}_l)}} \qquad (3.5.23)$$

The correlation between $\mathbf{t}_l$ and $\mathbf{u}_l$ is a measure of the linear relationship between $\mathbf{t}_l$ and $\mathbf{u}_l$ (Everitt and Dunn, 1991). Equation (3.5.22) is the Ordinary Least Squares (OLS) or Multiple Linear Regression (MLR) criterion for the predictive relationship between $\mathbf{t}_l$ and $\mathbf{u}_l$. Furthermore, equation (3.5.23) is the Canonical Correlation Analysis (CCA) criterion, which produces a sequence of uncorrelated linear combinations of the predictor variables $(\mathrm{X}\mathbf{w}_r^T)$ that maximally predict the corresponding linear combinations of the response variables $(\mathrm{Y}\mathbf{c}_r^T)$. It is known that, these criteria provide unbiased estimates of the regression coefficients and, furthermore, they are invariant to the scales of $\mathrm{X}\mathbf{w}_r^T$ and $\mathrm{Y}\mathbf{c}_r^T$. However, Frank and Friedman (1993) have shown that the criterion to be used should be biased away from orthogonal transformations of low data-spread directions in both X and Y spaces. This can be achieved by multiplying equation (3.5.22) and (3.5.23) by the variance of the orthogonal transformations :

$$\operatorname{Var}(\mathrm{X}\mathbf{w}_l)\operatorname{corr}^2(\mathrm{X}\mathbf{w}_l,\mathrm{Y}\mathbf{c}_l)\operatorname{Var}(\mathrm{Y}\mathbf{c}_l)=\operatorname{Cov}^2(\mathrm{X}\mathbf{w}_l,\mathrm{Y}\mathbf{c}_l) \qquad (3.5.24)$$

or equivalently :

$$\mathrm{Var}(\mathbf{t}_1)\,\mathrm{corr}^2(\mathbf{t}_1,\mathbf{u}_1)\,\mathrm{Var}(\mathbf{u}_1)=\mathrm{Cov}^2(\mathbf{t}_1,\mathbf{u}_1) \qquad (3.5.25)$$

The problem can be stated as maximising the squared covariance of the orthogonal transformations of the two spaces, $\mathrm{Cov}^2(\mathbf{t}_1,\mathbf{u}_1)$, with respect to $\mathbf{w}_1$ and $\mathbf{c}_1$ subject to the constraints $\mathbf{w}_1^T\mathbf{w}_1=1$ and $\mathbf{c}_1^T\mathbf{c}_1=1$. This is a constrained optimisation problem, which can be solved by applying Lagrange Multipliers. The equivalent Lagrangean expression to be maximised is :

$$\begin{aligned}\phi &= \left(\mathbf{t}_1^T\mathbf{u}_1\right)^2-\lambda\left(\mathbf{w}_1^T\mathbf{w}_1-1\right)-\kappa\left(\mathbf{c}_1^T\mathbf{c}_1-1\right)\\ &= \left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)^2-\lambda\left(\mathbf{w}_1^T\mathbf{w}_1-1\right)-\kappa\left(\mathbf{c}_1^T\mathbf{c}_1-1\right) \qquad (3.5.26)\end{aligned}$$

where λ, κ are the undetermined Lagrange multipliers. Differentiating with respect to and $\mathbf{w}_1$ and $\mathbf{c}_1$, and setting the resultant equation to zero, it follows :

$$\frac{\partial\phi}{\partial\mathbf{w}_1}=\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathrm{X}^T\mathrm{Y}\mathbf{c}_1-\lambda\mathbf{w}_1=0\Rightarrow\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathrm{X}^T\mathrm{Y}\mathbf{c}_1=\lambda\mathbf{w}_1 \qquad (3.5.27)$$

$$\frac{\partial\phi}{\partial\mathbf{c}_1}=\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathrm{Y}^T\mathrm{X}\mathbf{w}_1-\kappa\mathbf{c}_1=0\Rightarrow\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathrm{Y}^T\mathrm{X}\mathbf{w}_1=\kappa\mathbf{c}_1 \qquad (3.5.28)$$

Pre-multiplying equations (3.5.27) and (3.5.28) by $\mathbf{w}_1^T$ and $\mathbf{c}_1^T$, it follows that :

$$\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1=\lambda \qquad (3.5.29)$$

$$\left(\mathbf{w}_1^T\mathrm{X}^T\mathrm{Y}\mathbf{c}_1\right)\mathbf{c}_1^T\mathrm{Y}^T\mathrm{X}\mathbf{w}_1=\kappa \qquad (3.5.30)$$

and, thus, $\lambda = \kappa$, since the left-hand side quantities are scalar products and are the transpose of one another. Furthermore, the scalar quantity $\left(\mathbf{w}_1^T X^T Y \mathbf{c}_1\right)$, which is the covariance of the orthogonal transformations of the two spaces, $\mathrm{Cov}\left(\mathbf{t}_1, \mathbf{u}_1\right)$, is equal to the square root of the Lagrange multiplier λ or κ that will be denoted as λ_1 onwards:

$$\mathrm{Cov}\left(\mathbf{t}_1, \mathbf{u}_1\right) = \mathrm{Cov}\left(X\mathbf{w}_1, Y\mathbf{c}_1\right) = \left(\mathbf{w}_1^T X^T Y \mathbf{c}_1\right) = \sqrt{\lambda_1} \qquad (3.5.31)$$

Solving equation (3.5.28) with respect to $\mathbf{c}_1$ and replacing $\mathbf{c}_1$ in equation (3.5.27), it follows that :

$$\mathbf{c}_1 = \lambda_1^{-1}\left(\mathbf{w}_1^T X^T Y \mathbf{c}_1\right) Y^T X \mathbf{w}_1 \overset{(3.5.31)}{=} \frac{1}{\sqrt{\lambda_1}} Y^T X \mathbf{w}_1 \qquad (3.5.32)$$

and

$$\sqrt{\lambda_1} X^T Y \frac{1}{\sqrt{\lambda_1}} Y^T X \mathbf{w}_1 = \lambda_1 \mathbf{w}_1 \Rightarrow X^T Y Y^T X \mathbf{w}_1 = \lambda_1 \mathbf{w}_1 \qquad (3.5.33)$$

Therefore, λ_1 is an eigenvalue of the covariance of the cross-covariance matrix $Y^T X$ and the weight $\mathbf{w}_1$ corresponds to its normalised eigenvector ($\mathbf{w}_1^T \mathbf{w}_1 = 1$). Furthermore, it can be concluded from equation (3.5.31) that, λ_1 is the maximum eigenvalue, since it is equal to the squared covariance that has to be maximised. Similarly, solving equation (3.5.27) with respect to $\mathbf{w}_1$ and replacing it in equation (3.5.28), it follows that :

$$Y^T X X^T Y \mathbf{c}_1 = \lambda_1 \mathbf{c}_1 \qquad (3.5.34)$$

Therefore, λ_1 is an eigenvalue of the covariance of the cross-covariance matrix X^TY and the weight $\mathbf{c}_1$ corresponds to its normalised eigenvector $\mathbf{c}_1^T\mathbf{c}_1 = 1$. Furthermore, it can be seen that, $\mathbf{t}_1$ and $\mathbf{u}_1$ are eigenvectors of the matrices XX^TYY^T and YY^TXX^T. By multiplying equations (3.5.33) and (3.5.34) by X and Y, respectively, it follows that :

$$XX^TYY^T\mathbf{t}_1 = \lambda_1\mathbf{t}_1 \tag{3.5.35}$$

$$YY^TXX^T\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{3.5.36}$$

Having defined $\mathbf{t}_1$ and $\mathbf{u}_1$, one can calculate the corresponding coefficients (direction cosines) of the transformations of the high-dimensional predictor and response spaces X and Y, by regressing X and Y onto the one-dimensional subspaces:

$$\mathbf{p}_1 = \frac{X^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1} \tag{3.5.37}$$

$$\mathbf{q}_1 = \frac{Y^T\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{u}_1} \tag{3.5.38}$$

where $\mathbf{p}_1$ and $\mathbf{q}_1$ denote the vectors of coefficients of the transformations and are called loading vectors of the predictor and the response data sets, respectively. Note that, $\mathbf{p}_1$ and $\mathbf{q}_1$ have been calculated by regressing the high-dimensional spaces onto the one-dimensional subspaces, while the normalised weights $\mathbf{w}_1$ and $\mathbf{c}_1$ have been calculated by regressing the high-dimensional spaces X and Y one to another.

In order to strengthen the relationship between the predictor and response data sets, a second set of orthogonal transformations ($\mathbf{t}_2$ and $\mathbf{u}_2$) of high-dimensional spaces X and Y are calculated. However, it can be shown that the covariance of the second set of latent variables is less than the maximum covariance of the response-predictor cross-covariance matrix of the residuals X_1 and Y_1 that result from the application of the relationship between the first pair of orthogonal transformations (Hoskuldsson, 1988):

$$X_1 = X - \mathbf{t}_1\mathbf{p}_1 \qquad (3.5.39)$$

$$Y_1 = Y - \mathbf{u}_1\mathbf{c}_1 \qquad (3.5.40)$$

Therefore, one pair of orthogonal transformations ($\mathbf{t}_{r+1}, \mathbf{u}_{r+1}$) should be calculated at each iteration (r+1) by the residual matrices resulting from the previous iteration (r) :

$$X_r^T Y_r Y_r^T X_r \mathbf{w}_{r+1} = \lambda_{r+1}\mathbf{w}_{r+1} \qquad (3.5.41)$$

$$Y_r^T X_r X_r^T Y_r \mathbf{c}_{r+1} = \lambda_{r+1}\mathbf{c}_{r+1} \qquad (3.5.42)$$

$$X_r X_r^T Y_r Y_r^T \mathbf{t}_{r+1} = \lambda_{r+1}\mathbf{t}_{r+1} \qquad (3.5.43)$$

$$Y_r Y_r^T X_r X_r^T \mathbf{u}_{r+1} = \lambda_{r+1}\mathbf{u}_{r+1} \qquad (3.5.44)$$

This procedure is continued until a satisfactory predictive relationship between the predictor and response data sets is obtained. The number of latent dimensions (R) required to provide satisfactory prediction, without overfitting the data, is usually determined by cross-validation (Wold, 1978). The corresponding pairs of $\mathbf{p}_r$ and $\mathbf{q}_r$ can be calculated by equivalent equations to those of (3.5.37) and (3.5.38). It can be seen that, due to the rotation of the latent vectors $\mathbf{t}_r$, the orthogonality of the loading vectors $\mathbf{p}_r$ of the predictor (X) data set is lost. However, the loadings $\mathbf{q}_r$ of the response space (Y) are orthogonal and, therefore, are equal to the weights $\mathbf{c}_r$ of the orthogonal transformation in Y. The procedure described above is termed Projection to Latent Structures (PLS), and it defines the orthogonal transformation in high-dimensional predictor and response space in an iterative way, so that both the covariance of these transformations and the predictive relationship between the predictor and response data sets are simultaneously maximised.

3.5.2 Properties of PLS

The technique of Projection to Latent Structures (PLS) has been associated with Singular Value Decomposition (SVD) (Hoskuldsson, 1988) :

$$X_r^T Y_r = \mathbf{w}_r \sigma_r \mathbf{q}_r^T + \sum(\text{less significant terms}) \tag{3.5.45}$$

where $\mathbf{w}_r$, $\mathbf{q}_r$ are the first left and right singular vectors and σ_r is the largest singular value. Kaspar and Ray (1993) proposed an iterative procedure for PLS that considers the technique as a successive SVD of the residual cross-covariance matrix :

$$X_{r+1}^T Y_{r+1} = \left(I - \mathbf{p}_r \mathbf{w}_r^T\right) X_r^T Y_r \tag{3.5.46}$$

where

$$\mathbf{p}_r = \frac{X_r^T X_r \mathbf{w}_r}{\mathbf{w}_r^T X_r^T X_r \mathbf{w}_r} \tag{3.5.47}$$

$$\mathbf{q}_r^T = \frac{\mathbf{w}_r^T X_r^T Y_r}{\mathbf{w}_r^T X_r^T X_r \mathbf{w}_r} \tag{3.5.48}$$

$$b_r = \frac{w_r^T X_r^T Y_r q_r}{w_r^T X_r^T X_r w_r} \tag{3.5.49}$$

$$X_{r+1}^T X_{r+1} = (I - p_r w_r^T) X_r^T X_r \tag{3.5.49}$$

and latent vectors $\mathbf{t}_r$ and $\mathbf{u}_r$ are calculated according to NIPALS equations.

The basic properties of the latent vectors can be summarised as follows (Hoskuldsson, 1988) :

a. The latent score vectors as well as the coefficients of the orthogonal transformation of the predictor data are mutually orthogonal :

$$\mathbf{t}_i^T \mathbf{t}_j = 0$$

$$\mathbf{w}_i^T \mathbf{w}_j = 0 \quad \forall\, i \neq j \tag{3.5.50}$$

b. The weights **w** are orthogonal to the loading vectors **p** :

$$\mathbf{w}_i^T \mathbf{p}_j = 0 \quad \forall\, i \neq j \tag{3.5.51}$$

c. The loadings **p** are orthogonal in the kernel space of X :

$$\mathbf{p}_i^T (X^T X)^{-} \mathbf{p}_j = 0 \quad \forall\, i \neq j \tag{3.5.52}$$

Another interesting property of PLS is its ability to handle missing data in the predictor data set X (Kresta, et al., 1994). If any measurements, x_{ij}, are missing from a variable j then neither the weight nor the score, whose calculation involved them can be computed. However, since the NIPALS algorithm, in steps (3) and (5), can be viewed as the regression of the j-th variable on the latent score, then the weight and the score can be calculated as :

$$w_{r,j} = \frac{\sum_{i=1}^{n} x_{ij} u_{ij} \Delta_{ij}}{\sum_{i=1}^{n} u_{ij}^2 \Delta_{ij}} \tag{3.5.53}$$

$$t_{r,j} = \frac{\sum_{i=1}^{n} w_{ij} x_{ij} \Delta_{ij}}{\sum_{i=1}^{n} w_{ij}^2 \Delta_{ij}} \tag{3.5.54}$$

where $\Delta_{ij} = 0$ for a missing observation x_{ij} and $\Delta_{ij} = 1$ otherwise. After convergence, the loadings of the predictor data set are calculated as :

$$p_r = \frac{X u_r}{u_r^T u_r} \tag{3.5.55}$$

3.5.3 Selection of the Optimal Number of Latent Dimensions

An important issue associated with the application of Projection to Latent Structures, is to determine the optimal number of latent dimensions (R) that are required to provide a satisfactory predictive relationship between the predictor and response data sets, without overfitting. Usually, with highly correlated variables, the first few latent dimensions are significant, since the predictive relationship cannot be improved by retaining more latent dimensions. The most commonly used technique is Cross-Validation (Wold, 1978; Stone, 1974). The concept of cross-validation is that a subset of the response data set can be predicted by a PLS model that was built with it not included. A similar procedure to that for PCA is adopted (section 3.4.3).

Specifically, the m-variate predictor and the k-variate response data sets can be divided into G subsets X_g and Y_g, respectively (g=1,...,G). Each of pair of subsets (X_g and Y_g) is then excluded and a PLS model is built upon the remaining subsets X_g^- and Y_g^-. The resulting loadings P_g^- can be used to calculate the scores (T_g) of the excluded subset X_g according to the NIPALS algorithm:

$$T_g = X_g P_g^- \tag{3.5.56}$$

Predictions of Y_g can then be obtained by retaining an increasing number (r=1,2,...) of latent dimensions in the PLS model each time :

$$\hat{Y}_{g,r} = \sum_{j=1}^{r} b_{g,j} \mathbf{t}_{g,j} \mathbf{q}_{g,j}^{-T} \tag{3.5.57}$$

where $b_{g,j}$ is the regression coefficient, $\mathbf{t}_{g,j}$ is the j-th column vector of the matrix of scores, T_g, of the excluded subset, X_g, and $q_{g,j}^-$ is the j-th column vector of the matrix of loadings, Q_g^-, of the subset, X_g^-. For the excluded subset, Y_g, the Squared Prediction Error (SPE) can be calculated for each number (r) of retained latent variables in the model :

$$SPE_{g,r} = \sum_{i=1}^{n_g} \sum_{j=1}^{m} \left(y_{ij} - \hat{y}_{ij} \right)^2 \tag{3.5.58}$$

where n_g is the number of objects included in the subset X_g or Y_g, and y_{ij}, $\hat{y}_{ij}$ denote the observed predicted value of an element of the subset Y_g, respectively.

When the calculations for all the subsets are finished, the Prediction Sum of Squares (PRESS) can be calculated for each number of retained latent dimension :

$$\mathrm{PRESS}_r = \sum_{g=1}^{G} \mathrm{SPE}_{g,r} \tag{3.5.59}$$

An alternative to PRESS is the *Root-Mean-Square-Error of Cross-Validation* (RMSECV), which measures the ability of the model to predict the response values from new values. The RMSECV is related to PRESS as :

$$\mathrm{RMSECV}_r = \sqrt{\frac{\mathrm{PRESS}_r}{n}} \tag{3.5.60}$$

where n is the number of objects. The optimum number of latent dimensions that have to be retained corresponds to a minimum in the overall PRESS or RMSECV. Equivalently, a normalised form of the PRESS (NPRESS) that is divided by the sum of squares of the response data set, can be used (Kresta et al., 1991). This can be calculated for each response variable separately :

$$\mathrm{NPRESS}_{r,j} = \frac{\mathrm{PRESS}_r}{\sum_{h=1}^{r}\sum_{i=1}^{n}\left(b_h t_{i,h} q_{h,j}^T\right)^2} \tag{3.5.61}$$

and for the overall model :

$$\mathrm{NPRESS}_r = \frac{\mathrm{PRESS}_r}{\sum_{h=1}^{r}\sum_{i=1}^{n}\sum_{j=1}^{k} t_h q_h^T} \tag{3.5.62}$$

Other cross-validation methods are discussed by Hoskuldsson (1996).

3.6 Other Multivariate Projection Methods

Apart from PCA and PLS, there are a number of other multivariate projection methods that can reduce the dimensionality of one or two data sets, but from a different perspective.

3.6.1. Factor Analysis

Factor Analysis (FA) was developed initially by Charles Spearman in 1904 and its main field of application is the behavioural sciences and in particular psychology. Factor Analysis is concerned with whether the covariances or correlations between a set of observed variables (x_j, j=1,...,m) can be explained by a smaller number of latent variables (f_i, i=1,...,k) as :

$$x_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \ldots + \lambda_{jk} f_k \qquad (3.6.1)$$

or using matrix notation by :

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{u} \qquad (3.6.2)$$

where Λ is a matrix of fixed coefficients and **u** is a vector of random errors. Factor analysis is similar to PCA. PCA is an orthogonal transformation of the original variables, which does not depend upon an underlying model. On the other hand, FA is based upon a statistical model and it is more concerned with explaining the covariance structure of the variables, rather than with explaining the variances. Moreover, there are a number of assumptions made that are not always realistic but which have to be satisfied while setting up a FA model.

3.6.2 Principal Component Regression

Principal Component Regression (PCR) (Massy, 1965) is an extension of PCA that can be applied to model a response data set (Y) from a predictor data set (X). PCR comprises two steps :

1. A PCA is performed on the predictor data set and a set of principal components scores is obtained :

$$X = TP^T \tag{3.6.3}$$

where T and P are the matrices of scores and loadings, respectively.

2. The response data set is then regressed on the scores of the predictor data set :

$$Y = TQ^T + F \tag{3.6.4}$$

where Q are the loadings of the response data set and F is the reconstruction error of Y. The matrix of principal component regression coefficients is defined as :

$$B = PQ^T \tag{3.6.5}$$

It can be seen that, PCR defines a new set of uncorrelated latent vectors in the space of X that minimises only variance-covariance matrix $X^T X$ but which does not account for the relationship between X and Y.

3.6.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) can be viewed as a generalisation of PCA

(Hotelling, 1936). CCA seeks to account for the correlation structure between two sets of variables generated from the same multivariate data sample. Consider that a multivariate sample data set X is partitioned into two subsets X_1 and X_2, each containing measurements on n objects for p_1 and p_2 variables, respectively. The problem of canonical correlation is to find a linear combination in each subset :

$$u = a_1 x_{1,1} + a_2 x_{1,2} + \ldots + a_{p_1} x_{1,p_1} = a^T X_1 \tag{3.6.6}$$

$$v = b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_{p_2} x_{2,p_2} = b^T X_2 \tag{3.6.7}$$

so that the correlation between the two linear combinations, $\mathrm{corr}(u,v)$, is maximised, when they are standardised to unit variance, $\mathrm{Var}(u) = \mathrm{Var}(v) = 1$. Several modifications of CCA, such as *Redundancy Analysis* can be found in the literature (Basilevsky, 1994; Van den Wallenberg, 1977).

3.6.4 Canonical Variate Analysis

Canonical Variate Analysis (CVA) is a time series modification of Canonical Correlation Analysis (Akaike, 1976) that has been found to be suited for analysing sets of autocorrelated data. Applications to-date, have been restricted to the identification of complex systems that are typical of providing autocorrelated data (Larimore, 1983; Schaper, et al., 1994). Consider the identification of a dynamic process model given data comprising the inputs and outputs of a multivariate process. At any sample time, vectors containing past inputs and outputs and future inputs and outputs of the process can be formed. CVA seeks to find the optimal linear combination of the past vectors that allow for the prediction of future vectors. CVA is a very promising technique that can handle sets of autocorrelated data and, therefore, it may address the problems of applying MSPC to autocorrelated data, where PCA and PLS are inappropriate.

3.7 Multivariate Statistical Analysis of Data Using PCA and PLS

Multivariate statistical projection techniques, such as PCA and PLS, decompose data sets in an optimal way into two parts :

$$Z = \sum_{r=1}^{A} \gamma_r \delta_r^T = \sum_{r=1}^{R} \gamma_r \delta_r^T + \sum_{r=R+1}^{A} \gamma_r \delta_r^T = \textit{Systematic part} + \textit{Residual part} \qquad (3.7.1)$$

The systematic part represents the contribution to the data set due to the principal components or latent variables whilst the residual part represents the part which is unexplained by the statistical model and usually describes the noise associated with the data. According to Gnanadesikan and Kettenring (1972), multivariate analysis can be divided into :

A) The analysis of internal structure

B) The analysis of superimposed or extraneous structure

Therefore, the systematic part of the decomposition of the data sets expresses the internal structure, whilst the residual part expresses the superimposed structure.

3.8 Summary

Projection techniques comprise a set of multivariate statistical techniques that can be used for the purpose of efficiently analysing data sets containing highly correlated variables. Principal Components Analysis (PCA) can be used for the explanatory analysis of a data set. The dependencies between the original variables can be analysed. Furthermore, a new set of uncorrelated variables can be defined in terms of a linear combination of the original variables. This often leads to a significant reduction in the dimensionality of the problem and, thus, both visualisation and interpretation of the data can be more easily achieved. Projection to Latent Structures (PLS) optimises the relationship between two data sets. Both methods have been presented and derived using geometrical, mathematical and statistical approaches. Other techniques that can be used to reduce the dimensionality of a data set, have also been briefly described.

Chapter IV

Process Monitoring and Diagnostic Schemes Based on Multivariate Statistical Process Control

4.1 Introduction

The major aims of process monitoring are the reduction of off-specification production, the identification of important process disturbances and the early warning of process or plant faults. The early detection of process faults, followed by the location of their source, can lead to significant improvements in product quality and process/plant safety. Consequently, on-line monitoring of process performance has become an extremely important part of any processing operation, and a very fertile ground for the theoretical development and industrial deployment of intelligent process supervisory systems.

This chapter describes the procedures that require to be followed for the successful implementation of an MSPC-based scheme for process monitoring, fault detection and diagnosis. MSPC scheme utilises a multivariate statistical model or representation that is constructed using the statistical projection techniques of PCA or PLS. However, these techniques are only suitable for continuous processes that operate at steady state. Furthermore, these techniques investigate the relationships of all the variables simultaneously and, therefore, they do not take into account the topology of a complex process that comprises distinct units. Extensions of the projection techniques of PCA and PLS, namely Multi-Way and Multi-Block, that can be used to construct statistical models for processes exhibiting non-linear characteristics, such as batch and semi-batch processes, and for complex processes comprising a number of several distinct units, are described. Moreover, all statistical projection techniques are data-oriented and, as a result, models for robust MSPC-based schemes can be developed only for processes where there is a wealth of data. A novel approach for the generation of additional data is also presented. This approach is based on the inversion of a PLS regression process model, which has the ability to generate additional process data that is consistent with the minimal process plant data. Finally, all issues associated with the implementation of MSPC-based schemes that utilise the presented statistical techniques, are described.

4.2 On-Line Process Monitoring, Fault Detection and Diagnosis

In today's chemical and process industries, plants are becoming larger, more complex and heavily instrumented. Consequently, it is more difficult to locate the source of a fault. The requirements to manufacture product with minimal variation around a desired quality target and to operate safely according to health, safety and environmental protection regulations, has become essential due to market and public demand. As a result, consistent and safe production has been proven to be economically beneficial, whilst plant/process down time and the emission of pollutants have become even more expensive. Successful operation in terms of high yield, better product quality and more consistent production at reduced operational costs and improved health and safety standards, can only be achieved when processes or plants are operated under well controlled conditions.

The key to successful operation is efficient on-line process monitoring, which enables the early warning of process disturbances, process malfunctions or faults. Where early detection of such problems is followed by the location of their source, the efficiency and consistency of production can be significantly improved. Schemes for process monitoring, fault detection and diagnosis can then be used as intelligent supervisory process systems, which can support process operators and engineers in dealing with process deviations and identifying the root cause of these deviations. These schemes are based upon process models built from plant data.

Process models for on-line process performance monitoring and fault diagnosis can be divided into three general types, according to a number of authors (Himmelblau, 1978; Patton et al., 1989; Nomikos and MacGregor, 1994; Zhang et al., 1996), namely, *heuristic* models, *deterministic* models and *statistical* models.

A **heuristic** model is based upon the behavioural and casual description of certain specific phenomena in the process, using probability theory, fuzzy logic and neural networks. Faults can be detected and diagnosed by causally tracing symptoms back along their propagation paths or by comparing the predicted behaviour with the actual behaviour. Systems developed using this kind of models are known as *knowledge-based* or *expert* systems and their development, generally, demands considerable time and effort.

A **deterministic** model is based upon the underlying fundamental physical and chemical model of the process. Faults can possibly be detected and diagnosed under the assumption that they will cause changes to certain physical parameters, which, in turn, will lead to changes to some model parameters or states. Systems developed using deterministic models are known as *model-based* systems. Detailed deterministic models for complex process are difficult to develop. Furthermore, they have to be complemented with heuristic knowledge or models in order to efficiently implement the task of fault detection and diagnosis.

A **statistical** model is based upon the philosophy of Statistical Process Control (SPC), under which the behaviour of a process can be characterised using data obtained when the process is in a state of *statistical control*, that is, when the process is operating well. Faults can be detected and diagnosed by comparing the actual process behaviour to the *in-statistical-control* behaviour and its statistical properties. Systems developed using statistical models are called *MSPC-based systems* or *schemes.* A statistical model is not as powerful in detecting *built-in* faults as a heuristic or deterministic model. However, the only information needed to develop an MSPC-based scheme for on-line process monitoring and fault detection and diagnosis is a historical database of past successful process operations.

Approaches to develop on-line process monitoring and faults detection and diagnosis schemes using heuristic or deterministic models are *directional* in nature, that is, the reasons for deviations from the normal behaviour and for faults are built in the models. Although, they are very powerful approaches, their implementation is limited by the considerable amount of time and effort required to develop the models. On the other hand, the statistical approach is *in-directional* and, most of times, the diagnosis is left to the process operators and plant engineers, who diagnose the fault and take appropriate corrective actions, using their process knowledge. However, a statistical model can be easily developed.

Having developed a representative model of a process, a monitoring and diagnostic scheme can be defined, in its most abstract form, as a two-step model-based task (Stephanopoulos and Han, 1996) :

1. Compare the actual behaviour of the process, as given by the values of the process variables, against the behaviour predicted by the model and generate "residuals", which reflect the impact of the deviation or fault.

2. Evaluate these "residuals", identify the deviation or fault, that caused the observed behaviour, through a model-based inversion procedure and, furthermore, identify the process variables responsible for these faults.

4.3 MSPC-Based Schemes for On-Line Process Monitoring

There are three main steps involved in the development of an MSPC-based process monitoring scheme, namely, the *analysis of the historical process database*, the *development of the statistical model/representation* and the *testing/validation of the MSPC-based scheme.* Having developed the scheme, subsequent process performance is evaluated using the developed monitoring charts.

4.3.1 Analysis of the Historical Process Database

In any process, computers and data-acquisition systems are assigned the task of collecting on-line measurements on a number of process variables, on a frequent basis, *process data.* On the other hand, measurements that characterise the quality of the manufactured product may only be recorded infrequently after a laboratory analysis, *quality data.* Process data along with the corresponding quality data comprises the historical database. The historical database is first analysed, prior to model development, to check whether the data contains sufficient information to develop a model and also to detect the presence of non-conforming (abnormal) operation, i.e. data *pre-screening*. A number of important issues have to be taken into consideration in the *pre-processing* or *pre-screening* stage of the analysis of the historical database (Martin et al., 1997) :

Missing Data. In the majority of data sets, some measurements on variables will be unrecorded for some reason. The most common reason is the malfunction of the data-acquisition system. However, the standard statistical techniques require that the data matrix is complete prior to performing the analysis. Data can either be missing at random, for example due to a dropped test-tube, or not-missing at random, for example due to instrument failure. In situations where measurements are missing at random (MAR), the data matrix can be modified either by deleting partially observed objects or variables, or by in-filling with plausible values for the missing measurements, such as means, medians, last recorded value or, alternatively, a combination of them. In cases where measurements are missing in a non-random manner, the data matrix can be modified by estimating the missing values using time series reconstruction, multiple linear regression, principal component analysis, factor analysis, etc. (Martin et al. 1997).

Outliers. These are defined as measurements on variables that appear to be inconsistent with the rest of the data. Outliers can have a major effect on the statistical analysis. In particular, they can affect the direction of greatest variation and can impact the performance of the statistical model/representation. Robust statistical modelling requires the data to be free of outliers and, thus, the tasks of outlier identification and removal are of great importance.

Multivariate statistical projection techniques decompose data sets in an optimal way into a systematic and a residual part, which express the internal and the superimposed structure of the data, respectively (section 3.7). Outliers may be associated with each of these structures and it is important to keep their identities as distinct as possible. Hawkins (1980) refers to these as *Type A* and *Type B* outliers, respectively.

An outlier of Type A refers to an outlier from the assumed distributional form of the data. It will only be detected when the variation in the variables, in the reduced space of the retained principal components or latent variables, is greater than that which can be explained by common cause variation. A measure of common cause variation is given by Hotelling's T^2 statistic, which measures the squared distance of a point (process observation vector) from the centre of the reduced space (point of "zero" variation). The important consideration with Type A outliers is that they will be identified whether or not a projection technique is applied. However, the use of multivariate statistical projection methods usually enhances the chance of detecting them.

On the other hand, a Type B outlier refers to a point which differs from the internal structure of the data characterised by the statistical model. It will be detected when a totally new type of event occurs, which was not present in the internal structure, and it is an indication that a particular vector of observations cannot be characterised by the principal components or latent variables that define the reduced subspace. This result may occur because too few components were retained to produce a good statistical model or because the underlying covariance structure and its associated vector space, has changed with time, leading to a general lack-of-fit. The appropriate statistic for identifying this type of outliers is the squared perpendicular distance of the observation vector from the reduced space usually referred to as the *Q-statistic* (Jackson, 1991).

In general, outliers can be identified using the Mahalanobis Distance (MD), a measure of the distance of a point from the centre of the reduced space, coupled with the Squared Prediction Error (SPE) or Soft Independent Modelling of Class Analogy (SIMCA), measures of the squared perpendicular distance of a point from the reduced space. MD and SIMCA (or SPE) are complementary, since they measure the goodness-of-fit within and outside the model space, respectively. Alternatively one can look at the first principal component (Gnanadesikan and Kettenring, 1972) or at the minor principal component (Hawkins, 1974). A full description of the methods that can be used for outliers identification is given by Hawkins (1980).

Noise. The presence of noise in the data may obscure what is really happening within a process and, therefore, the removal of noise is an important task. Small amounts of noise usually can be removed by application of the statistical techniques of PCA and PLS. Significant amounts of noise, however, require the application of filtering techniques. A summary of suitable filtering techniques for process data are presented by Martin et al. (1997).

Data Transformation. Process data may need to be modified by applying a mathematical transformation, i.e. the substitution of the values of a variable with the values of a function of that variable. Typical examples of such mathematical functions include the square, the squared root, the inverse, the logarithm or the exponential function. There are two main reasons for applying mathematical transformations. The first is that transformations can reduce the non-linearity inherent within a system. The second reason is that sometimes transformed process data can produce a more suitable statistical model than the raw process data (Martin et al. 1997).

Scaling. Two main types of scaling can be applied to process data, namely, mean-centring and variance scaling. By mean-centring the data, the inherent common variation is removed prior to data analysis. Furthermore, in the situation where the original variables are measured in different units, the structure of the statistical model is dependent upon the essentially arbitrary choice of units of measurements and, therefore, the model will be biased to variables with large variance. The application of variance-scaling and mean-centring overcomes these kind of problems (section 3.4.2).

Variable Selection. The data should be checked for constant variables prior to model development, since variability is required in the data (Sharaf et al., 1986). Variables that do not exhibit variability can be detected by examining their standard deviation or the correlation matrix. They can be deleted from the data set or modified by adding to them an appropriate amount of noise, in order to ensure that they exhibit some kind of variability (Morris and Martin, 1997).

The outcome of the previous analysis is a data set of normal operation, where only *common cause* process variation is exhibited and observations exhibiting abnormal operation are clearly identified.

4.3.2 Development of Model and MSPC-based Scheme

Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) have been found to be particularly useful for analysing multivariate sets of highly correlated data, such as those found in historical databases of chemical processes, and for developing MSPC-based process monitoring and diagnosing schemes for two reasons :

1. PCA and PLS are dimensionality reduction techniques and, therefore, process data can be compressed, so that only the important and relevant information about the process is retained, while the rest, which usually explains the noise, can be discarded.

2. PCA and PLS define new latent vectors, which are uncorrelated linear combinations of the highly correlated original variables. Therefore, when these uncorrelated vectors are used to apply MSPC, then the overall type I error in the control charts can be directly computed.

A statistical model/representation of the process can be used as a basis of an on-line MSPC monitoring scheme. When new process data is collected, it can be evaluated against the nominal statistical representation and characterised as either *normal* or *abnormal*. Specifically, each time period when new data is collected, scores and quadratic residuals can be calculated by utilising the statistical process model. The monitoring procedure can then be implemented by constructing Shewhart-type control charts (Chapter II) in terms of time series plots of scores and residuals. This kind of control chart is called *process performance evaluation charts*, since they are used to continuously monitor and evaluate the performance of the process. Any efficient and flexible MSPC-based scheme should include plots of scores, which summarise the internal structure of the process and represent the common cause process variation, and plots of quadratic residuals, which depict new types of special event occurring and which are not present in the internal structure. When an unusual event is detected by these charts, it is possible to analyse it using the contribution of each variable to a high value of score or quadratic residual and, therefore, it is possible to find an assignable cause to it. The most commonly used monitoring charts are those of :

1. Plots of time series of individual process scores

2. Bivariate plots of time series of individual process scores $\left(t_i \text{ vs } t_j \, , \ \forall\, i \neq j\right)$.

3. Plots of the T^2 of the process scores or plots of the D-statistic time series (Nomikos and MacGregor, 1995):

$$D_s = \mathbf{t}_R^T S^{-1} \mathbf{t}_R \qquad . \qquad (4.3.1)$$

where $\mathbf{t}_R$ is a vector containing the scores of the R retained principal components or latent variables and S is the covariance matrix of the scores of the historical process data :

$$S = T^T T \qquad (4.3.2)$$

where T is a matrix $(n \times R)$ whose columns are the scores of the historical data on the R retained components or latent variables.

4. Time series plot of quadratic residuals, in terms of the Squared Prediction Error (SPE) :

$$\mathrm{SPE} = \sum_{j=1}^{m} \left(x_j - \hat{x}_j \right)^2 \tag{4.3.3}$$

where x_j and $\hat{x}_j$ are the actual and the model estimate of the value of the j-th variable, respectively.

The control limits for the previous charts are discussed in a following section. According to MSPC philosophy, when the quantities are plotted within their control limits, the process operates normally and exhibits only common cause variation.

4.3.3 Testing of the MSPC-based Scheme

Having developed the MSPC-based scheme, it should be tested in order to ensure that it can efficiently perform its task. Testing usually involves two steps. In the first step, the performance of the scheme is evaluated against the data contained in the historical database, whilst in the second the performance is evaluated against data sets that are known to belong to periods where unusual process events were detected. An effective scheme should be able to clearly identify data belonging to both normal and abnormal operations.

4.3.4 Control Limits for Process Performance Evaluation Charts

The control limits for the process performance evaluation charts are calculated based upon distributional assumptions. Control limits for **individual process scores charts** are constructed based upon the assumption of normality. The scores of the principal components or latent variables were found to follow a multivariate normal distribution (Horswell and Looney, 1992). This result arises from the fact that they are linear combinations of the process variables and according to the Central Limit Theorem, they should be approximately normally distributed. Under the assumption of normality, the control limits for the values of the r-th process score, at a level for significance α, can be calculated (Chew, 1968; Hahn and Meeker, 1991; Nomikos and MacGregor, 1995) :

$$\pm t_{n-1,\alpha/2} \cdot S_{ref,r} \cdot \left(1+\frac{1}{n}\right)^{\frac{1}{2}} \tag{4.3.4}$$

where n is the number of objects included in the nominal data set, $S_{ref,r}$ is the estimated standard deviation of the values of the r-th process score of the nominal data set (note that the mean is always 0) and $t_{n-1,\alpha/2}$ is the critical value of the student t-distribution with (n-1) degrees of freedom at a level of significance $\alpha/2$.

In cases where more than two principal components or latent variables are retained, then multiple plots of individual process scores make the monitoring procedure more complicated. However, a few bivariate plots of individual process scores (e.g. t_1 vs t_2, t_3 vs t_4) or process **Hotelling's T^2 statistic** or **D statistic** based upon all retained components or latent variables can simplify the procedure:

$$T^2 = D_s = \mathbf{t}_R^T S^{-1} \mathbf{t}_R = \sum_{r=1}^{R} \frac{t_r^2}{s_{t_r}^2} = \sum_{r=1}^{R} \frac{t_r^2}{\lambda_r} \tag{4.3.5}$$

where t_r and λ_r are the score and the variance of the i-th component or latent variable, respectively, $\mathbf{t}_R$ is a vector containing the scores on the R retained components or latent variables, S is the covariance matrix of the scores of the historical process data. It can be seen that, each term in equation (4.3.5) plays an equal role in the computation of T^2, irrespective of the amount of variance it explains, since each $\mathbf{t}_r^2$ has been scaled by the reciprocal of its variance. This illustrates one of the problems associated with T^2 when a large number of components or latent variables are retained and the original variables are highly correlated or when Σ is ill-conditioned. The lower order latent components explain very little of the variance and, generally, represent random noise. However, dividing $\mathbf{t}_r$ scores by their very small variances, even slight deviations, which have almost no effect on the data sets, will lead to an out-of-control signal in T^2. Under the assumption that the scores follow a multivariate distribution with population mean vector 0 and estimated covariance matrix S (R×R), which is diagonal due to the orthogonality of the scores, one can derive the D-statistic for Phase I of the construction of the control charts (Tracy et al., 1992), based upon Hotelling's T_0^2 :

$$D_s = \mathbf{t}_R^T S^{-1} \mathbf{t}_R \sim \frac{(n-1)^2}{n} B_{R/2,(n-R-1)/2} \qquad (4.3.6)$$

It can be shown that, the D statistic is distributed as a beta variate. Usually, one calculates the D statistic after having selected the optimum number of principal components or latent variables to be retained in the statistical model. The control limit, at a level of significance α , is given by :

$$D_s \leq \frac{(n-1)^2}{n} B_{R/2,(n-R-1)/2,\alpha} \qquad (4.3.7)$$

A joint confidence ellipsoid can, therefore, be defined. The centre of the ellipsoid is located at 0 and the length of each of its axes, along the direction of the r-th principal component or latent variable, is given by :

$$\pm\left[\frac{(n-1)^2}{n}S(r,r)B_{R/2,(n-R-1)/2,\alpha}\right]^{1/2} \qquad (4.3.8)$$

Note that, the eigenvalues of the diagonal covariance matrix S are its diagonal elements.

In Phase II, a D-statistic value for the score vectors of new process data can be calculated to test whether the process is still in control. The estimated covariance matrix is the one calculated in Phase I (S). According to Tracy et al. (1992), the D-statistic now is distributed as an F variate, as :

$$D_s = \mathbf{t}_R^T S^{-1} \mathbf{t}_R \sim \frac{(n+1)(n-1)R}{(n-R)n} F_{R,n-R} \qquad (4.3.9)$$

The control limit, for a level of significance α , is given by :

$$D_s \leq \frac{(n+1)(n-1)R}{(n-R)n} F_{R,n-R,\alpha} \qquad (4.3.10)$$

Similarly, a joint confidence ellipsoid can be defined. The centre of the ellipsoid is located at 0 and the length of each of its axes, along the direction of the r-th principal component or latent variable, is given by :

$$\pm\left[S(r,r)\frac{(n+1)(n-1)R}{(n-R)n} F_{R,n-R,\alpha}\right]^{1/2} \qquad (4.3.11)$$

The decomposition of T_0^2 proposed by Mason et al. (1995) is not useful in the case of the D-statistic since the covariance matrix S is diagonal. The D-statistic can be decomposed as :

$$D_s = \sum_{r=1}^{R} D_{s,r} = \sum_{r=1}^{R} \frac{t_r^2}{\lambda_r} \qquad (4.3.12)$$

where $D_{s,r}$ is the unconditional term, which accounts for the r-th principal component or latent variable. This result is similar to the decomposition already presented in equation (4.3.5). The unconditional term of the r-th principal component or latent variable is the squared value of its score, scaled by the reciprocal of its variance.

The residual term, in the decomposition of a data set, can be calculated as follows (3.7.1) :

$$Z = \sum_{r=1}^{R} \gamma_r \delta_r^T + \sum_{r=R+1}^{A} \gamma_r \delta_r^T = \hat{Z} + E \Rightarrow E = Z - \hat{Z} \qquad (4.3.13)$$

where $\hat{Z}$ is the estimate of the model or the systematic part of the data set Z and E is the residual part. The residual term can be tested by means of the **quadratic residuals.** For the data contained in the historical database, the quadratic residual is measured by the *Sum of Squares of Residuals* (SSR) or *Errors* (SSE) :

$$SSR_i = SSE_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T (\mathbf{x}_i - \hat{\mathbf{x}}_i) \qquad (4.3.14)$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ are the actual and the model estimates of the i-th process object vector, respectively. The quadratic residual for a new process vector $\mathbf{x}_{new}$ is measured by the *Squared Prediction Error* (SPE), as defined by equation (4.3.3) :

$$SPE = (\mathbf{x}_{new} - \hat{\mathbf{x}}_{new})^T (\mathbf{x}_{new} - \hat{\mathbf{x}}_{new}) = \sum_{j=1}^{m} (x_{new,j} - \hat{x}_{new,j})^2 \qquad (4.3.15)$$

Since the process data are mean-centred, the previous expressions represent the sum of squares of the perpendicular distance of a process object from the R-dimensional subspace that the statistical model defines and it can be viewed as a measure of the unstructured fluctuations (noise) that cannot be accounted for the model. The SSR or SSE are known as the *Q statistic* and it has been proposed mainly by J.E. Jackson (Jackson and Mudholkar, 1979; Jackson, 1991).

The confidence region for the quadratic residuals can be constructed by looking at their underlying distribution (Nomikos and MacGregor, 1995). Let **x** be an object vector from an m-variate normal population, $N_m(0,\Sigma)$, and λ_i be the eigenvalues of the population variance-covariance matrix Σ. Assume that Σ has full rank. Once the statistical model has been developed, the quadratic residual for each process vector will produce an overall fit of this vector to the model. Approximate control limits for a level of significance α for the quadratic residual are given by :

$$Q_\alpha = g \cdot \chi^2_{v,\alpha} \qquad \text{(Box, 1954)} \qquad (4.3.16)$$

$$Q_\alpha = \theta_1 \cdot \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{\frac{1}{h_0}} \quad \text{(Jackson and Mudholkar, 1979)} \ (4.3.17)$$

where $\chi^2_{v,\alpha}$ is a chi-squared variate with v degrees of freedom, and c_α is normal variate with the same sign as h_0. The remaining quantities are defined as follows :

$$\theta_1 = \sum_{i=r+1}^{a} \lambda_i \qquad (4.3.18)$$

$$\theta_2 = \sum_{i=r+1}^{a} \lambda_i^2 \tag{4.3.19}$$

$$\theta_3 = \sum_{i=r+1}^{a} \lambda_i^3 \tag{4.3.20}$$

$$g = \theta_2 / \theta_1 \tag{4.3.21}$$

$$v = \theta_1^2 / \theta_2 \tag{4.3.22}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \tag{4.3.23}$$

The relationship between the two proposed approximate control limits becomes clearer if one uses the Wilson-Hilferty approximation for the chi-squared variable (Evans et al., 1993) and rewrites Box's equation as follows :

$$Q_\alpha \cong gv \left(1 - \frac{2}{9v} + c_\alpha \left(\frac{2}{9v} \right)^{1/2} \right)^3 \tag{4.3.24}$$

Every term, apart from the second one, approximates the corresponding terms in equation (4.3.17) and thus :

$$\theta_2^2 \approx \theta_1 \cdot \theta_3 \tag{4.3.25}$$

A more convenient way to calculate the parameters θ_i, instead of calculating the eigenvalues λ_i of the unused or non-significant components, is to estimate them from the estimated residual covariance matrix (Nomikos and MacGregor, 1995) :

$$\Phi = \frac{E \cdot E^T}{n - 1} \tag{4.3.26}$$

as

$$\theta_1 = \text{trace}(\Phi) \tag{4.3.27}$$

$$\theta_2 = \text{trace}\left(\Phi^2\right) \tag{4.3.28}$$

$$\theta_3 = \text{trace}\left(\Phi^3\right) \tag{4.3.29}$$

4.4 Fault Detection and Diagnosis Using MSPC Process Representations

The essential requirement of any fault detection and diagnosis system is to readily detect abnormal events, to present possible root causes of these events, along with their possible consequences and recommended actions. However, all these requirements cannot be fulfilled completely by an empirical model. A wide variety of techniques and tools, such as pattern recognition, knowledge and rule-based expert systems, fuzzy logic, hypothesis testing and system identification techniques, usually have to be combined for an effective system.

There are two important questions that should be answered by any model used for monitoring and detection purposes. The first question is when is an event a fault. For statistical models, the answer is given by the MSPC philosophy : an event is a fault when it is statistically significant, that is, when the value of one or more of its statistics exceeds its confidence limits. The second question can be stated as whether the model is able to identify all possible faults. The philosophy of MSPC implies that a fault that cannot be observed with R components or latent variables, can be observed using at least (R+1) components or latent variables.

There are two ways in which a process operation can exhibit deviation from the statistical model. In the first case, the values of its scores can move outside the acceptable range of variation, which is defined by the control region. This type of deviation can be observed in any control chart associated with the scores and correspond to a Type A outlier. In this case, the model is still valid, but the magnitude of the process variation is too large. In the second case, the residuals can increase and the operation can be placed well outside, perpendicular to the reduced space. This type of deviation corresponds to Type B outlier and can be detected by the plots of the quadratic residuals. In this case, the model is no longer valid, because a new event not included in the reference set has occurred and the new process operation does not project onto the reduced space adequately. Although the multivariate monitoring procedures and their charts are very powerful ways for detecting deviations from normal operation, they do not indicate reasons for such deviations. Therefore, the development of diagnostic tools to identify the most likely combination of process variables responsible for abnormal operation is essential.

4.4.1 Fault Diagnosis : Isolating the Responsible Variables

A simple approach to diagnosis might be to develop an expert system based on the behaviour of the data projections in the principal components or latent variables space (Zhang et al., 1996). Certain types of faults can be characterised by the movement of the data projections into specific regions of the latent variable space, with or without specific directions. This would imply that an expert system could be developed from the behaviour of the latent variables from past faults.

More detailed information about possible causes can be obtained by closer interrogation of the underlying statistical model. Procedures have been proposed by MacGregor et al. (1994) and Miller et al. (1995). A review of procedures for the isolation of variables responsible for an out-of-control signal has been presented by Kourti and MacGregor (1996).

In the case where the unusual event is detected in the D-statistic (T^2) plot, the principal component(s) or latent variable(s) indicative of the out-of-control T^2 signal can be isolated by examining the normalised scores contributing to T^2 (Kourti and MacGregor, 1996). Since T^2, is a summation of the squared normalised process scores (4.3.5) :

$$T^2 = \sum_{r=1}^{R} \frac{t_r^2}{\lambda_r}$$

and scores are independent, then the normalised values $\left(t_r/\lambda_r\right)$ or their squares (Jackson, 1991) can be plotted. Those principal components or latent variables whose scores significantly contribute to the out-of control T^2 signal, can be isolated using *Bonferroni*-type control limits, as a rough guideline (Kourti and MacGregor, 1996). In the case when physical interpretation can be assigned to principal components or latent variables, it is possible to translate an unusual high value of a score into an assignable cause and, therefore, into corrective action. An alternative approach is to investigate which of the principal components or latent variables have unusual high values and to plot the contribution of each variable to them. Using this approach, the physical interpretation of the identified group of responsible variables is of interest rather than the physical interpretation of each principal component or latent variable (Miller, et al., 1995; MacGregor, 1994; Kourti and MacGregor, 1996).

Having detected an unusual event from a quadratic residual plot or process score(s) plot, the *contributions* of the individual variables can be examined and the variables indicative of the problem can be found by using *contribution plots*. The quadratic residual calculated for the purpose of detecting faults, is expressed in terms of the Squared Prediction Error (SPE). Consider the case where an unusual event is detected at time point k on an SPE plot. The Square Prediction Error of a particular process vector $\mathbf{x}_k$ is the sum of the squared prediction errors of all the m individual variables :

$$SPE_k = \sum_{j=1}^{m} \left(x_{j,k} - \hat{x}_{j,k}\right)^2 \tag{4.4.1}$$

where $x_{j,k}$ and $\hat{x}_{j,k}$ are the observed and predicted value of the j-th variable at the time point k. Therefore, the *individual contribution* of the j-th variable to the SPE value is its prediction error :

$$PE_{j,k} = x_{j,k} - \hat{x}_{j,k} \tag{4.4.2}$$

The predictions of each variable at any point in time, are given by either the PCA or PLS model :

$$\hat{x}_{j,k} = \sum_{r=1}^{R} t_{r,k} p_{r,j,k} \tag{4.4.3}$$

where $t_{r,k}$ is the predicted score of the r-th principal component or latent variable at time point k, and $p_{r,j,k}$ is an element of the loading vector of the r-th principal component or latent variable, corresponding to the j-th variable at time point k.

In the case where an unusual event is detected at a particular time point k on a process score plot, then the variable or group of variables that have a significant contribution, must be identified. Let the *importance* of a process variable (j) at a time point (k), to the r-th principal component or latent variable, be $\omega_{r,k,j}$. In PCA, $\omega_{r,k,j}$ is the element of the loading vector $\mathbf{p}_r$ of the r-th principal component for the j-th variable at time point k. Similarly, in PLS, $\omega_{r,k,j}$ is the element of the loading vector $\mathbf{w}_r$ of the r-th latent variable for the j-th variable at time point k. In ordinary PCA and PLS, the prediction of the scores of each principal component or latent variable, respectively, at each time point k, is the sum of the product of the current value of the process variables times their importance to the principal component or latent variable under consideration:

$$t_{r,k} = x_{k,1}\omega_{r,k,1} + \ldots + x_{k,m}\omega_{r,k,m} \qquad (4.4.4)$$

The individual contribution of the j-th variable to the score value is given by :

$$VC_{r,j,k} = x_{k,j} \cdot \omega_{r,k,j} \qquad (4.4.5)$$

and, therefore, the scores can be written as the sum of the contributions of each individual variable :

$$t_{r,k} = VC_{r,1,k} + VC_{r,2,k} + \ldots + VC_{r,m,k} \qquad (4.4.6)$$

In the previous summation, it can be seen that, only the individual variables contributions that have the same sign as the score are significant in driving the value of the score high and, therefore, only these contributions should be investigated (Kourti and MacGregor, 1996). However, in the case where the value of the score is squared, as in the interpretation of an out-of-control T^2 signal, then contributions from all the variables have to be taken into account. In the case where more than one score exhibits high values, it has been suggested by Kourti and MacGregor (1996) to calculate the total contribution of an individual variable, over all scores (G) with significant high values :

$$TVC_{j,k} = \sum_{g=1}^{G} VC_{g,j,k} \qquad (4.4.7)$$

Furthermore, the contribution of the j-th process variable to the change in the value of the r-th process score between two time points k_1 and k_2 can be computed as the difference between the individual variables contributions in those particular time points :

$$VC_{r,j,k_2} - VC_{r,j,k_1} \qquad (4.4.8)$$

Variables contributions to Prediction Errors (PE) or scores can be represented graphically, providing in this way diagnostic charts (Figures 4.1 and 4.2, respectively). Specifically, whenever an event is detected in an on-line monitoring chart, one can plot the contribution to Prediction Error or scores of the individual process variables to locate which variable or which group of variables are no longer consistent with normal operating conditions and have a significant contribution to the out-of-control signal and, therefore, are indicative of the event. This can be implemented by plotting the *instantaneous* contributions at the time point where the unusual event was detected (Figure 4.1), or *differential* contributions between two time points where the process deviated from normal operation (Figure 4.2). Differential contribution plots are more informative, since they reveal the driving force of the deviation.

It can be seen that, the diagnostic charts are qualitative diagnostic tools, since the statistical model used by the MSPC-based scheme, is not a cause and effect model. Although diagnostic charts do not clearly reveal the cause of the event, they can interrogate the underlying statistical model for possible reasons of faults and allow for corrective actions, that is to allow operators or engineers to response accordingly, using their process knowledge or even an expert system, to deduce possible assignable causes.

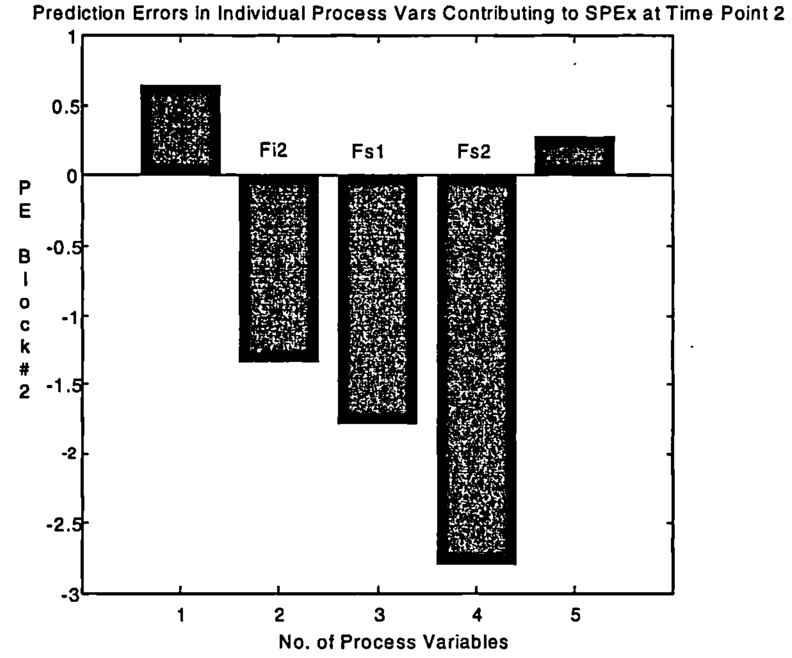


Figure 4.1. Typical instantaneous contributions plot

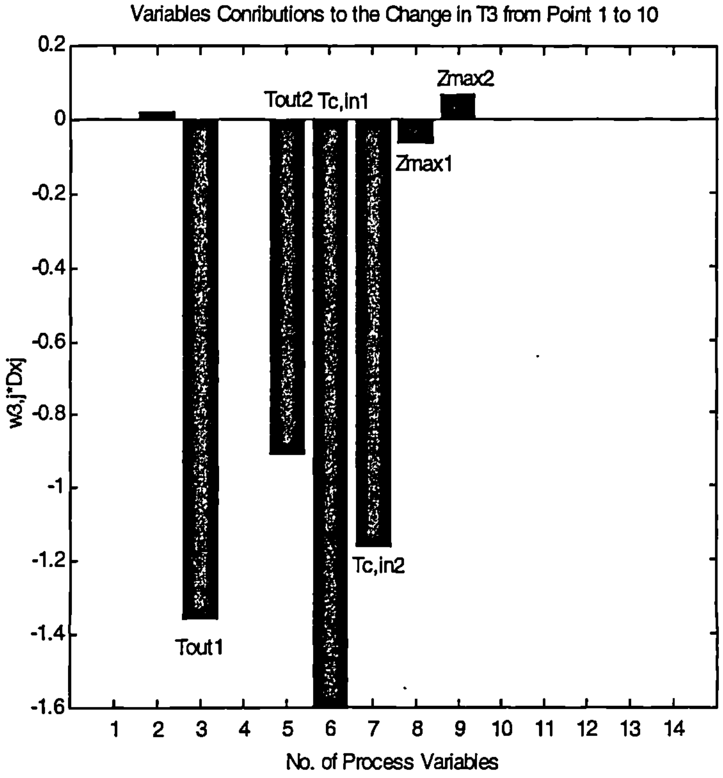


Figure 4.2. Typical differential contributions plot

4.5 Extensions of PCA and PLS Techniques

PCA and PLS techniques are more suitable for continuous processes that operate at steady state. In today's competitive atmosphere, the chemical and process industries are being forced to adapt to frequently changing technology and market conditions. This trend has lead to industry moving away from continuous operations into flexible batch and semi-batch modes of operation which focus on high quality, low volume production. Batch and semi-batch modes of operation cover a wide range of important chemical processes. Examples of batch processes include the production of polymers, separation and transformation processes such as distillation and crystallisation, fermentation, injection-molding processes and the manufacture of various specialty chemicals, biochemicals and pharmaceuticals. The main features, which characterise batch processes, are:

- Finite duration
- Highly non-linear behaviour and dynamic state operation
- Time variability (within batch variation)
- Inaccurate repeatability (batch to batch variation)
- Frequently changing process technology
- Flexibility in producing a variety of low volume and higher-added value products

Extensions of the projection techniques of PCA and PLS, namely, *Multi-Way PCA* (MPCA) and *Multi-Way PLS* (MPLS) have been proposed to construct statistical models for processes exhibiting non-linear characteristics, such as these found in batch and semi-batch processes (Nomikos and MacGregor, 1994; 1994b; 1995).

Additionally, industrial plants consist of a number of processes that are inter-connected. PCA and PLS provide a working approach to the modelling of these high-dimensional data sets, but the interpretation of the results is not always straightforward, since each latent variable contains contributions from many variables from different units and sections. Furthermore, these techniques investigate the relationships of all the variables simultaneously and, therefore, they do not take into account the topology of a complex process that comprises distinct units, which are not necessary related. In situations where there are many variables there is always a strong temptation to reduce the number of process variables. However, a reduction in the number of variables often removes information, makes the interpretation misleading and increases the risk of developing spurious statistical models. Alternatively, one can divide the variables into conceptually meaningful blocks and the appropriate extensions of the statistical projection techniques, namely, *Multi-*

Block PCA and *Multi-Block PLS* techniques can be applied. These are capable of analysing and modelling this kind of blocked data (MacGregor et al., 1994; Wangen and Kowalski, 1988; Wold et al., 1996).

MSPC-based schemes have typically been applied to industrial processes where large amounts of historical data have been collected, since the statistical projection techniques are data-oriented. However, difficulties can arise during the development of a robust monitoring scheme when there is only minimal plant data. There are many industrial situations where only a few data points are available from either an experimental design or initial product commissioning tests, which can be used to establish appropriate plant operating conditions. A major strategic challenge is therefore to build effective MSPC-based schemes based upon minimal 'design' process data. By utilising this 'design' data, sufficient new pseudo process data can be generated to establish an MSPC-based process monitoring scheme through Inverse Projection to Latent Structures (IPLS).

4.5.1 Multi-Way Extension of PCA and PLS

Data sets that form greater than two-way arrays are commonly encountered in experimental studies. Several multi-dimensional statistical techniques have been proposed for decomposing these multi-way arrays, such as canonical correlation, three-mode factor analysis, tensor rank and PARAFAC model, e.g. Zeng and Hopke (1990) Smilde and Doornbos (1991) Sanchez and Kowalski (1990), Smilde (1992). These techniques have been successfully applied in image analysis (Geladi et al., 1989) and in a few cases in the field of chemometrics (Smilde and Doornbos, 1991). Both Multi-Way Principal Components Analysis (MPCA) and Multi-Way Projection to Latent Structures (MPLS) were introduced by Wold et al. (1987) and they have been shown to be particularly useful for monitoring batch processes (Nomikos and MacGregor, 1994; 1994b). Multi-Way PCA (MPCA) and Multi-Way PLS (MPLS) are consistent in concept and algorithms to PCA and PLS, respectively, and, therefore, have the same goals and benefits. The relation between MPCA and PCA is that MPCA is equivalent to performing PCA on a large two-way array formed by unfolding the three-way data array. Similarly, a simple way to view MPLS is to consider unfolding the three-dimensional arrays X and Y into two dimensional arrays and performing PLS.

Data sets from batch processes form three-way arrays. Consider the case when there is available a historical database of I batches, where J process variables and L quality variables were measured over K and M time intervals, respectively, throughout the batch duration. All this information can be arranged into two three-way data sets **X** ($I \times J \times K$) and **Y** ($I \times L \times M$). Usually, the quality data set Y is two-dimensional, since quality measurements are available only at the end of the batch operation. There are three possible ways to unfold the three-way arrays. Usually, the array is unfolded in such a way as to put each of its vertical slices ($I \times J$) side by side to the right, starting with the one corresponding to the first time interval (Nomikos and MacGregor, 1994). This is the most meaningful approach since it allows the analysis of the variability between batches, i.e. summarising the information in the data with respect to both the variables and their time variation. Concerning the previous arrangement, it can be seen that different batches are organised along the vertical axis, variables along the horizontal axis and finally their time evolution occupies the third dimension of the arrays. Figure 4.3 illustrates the procedure of unfolding the three-way data sets in MPCA and MPLS.

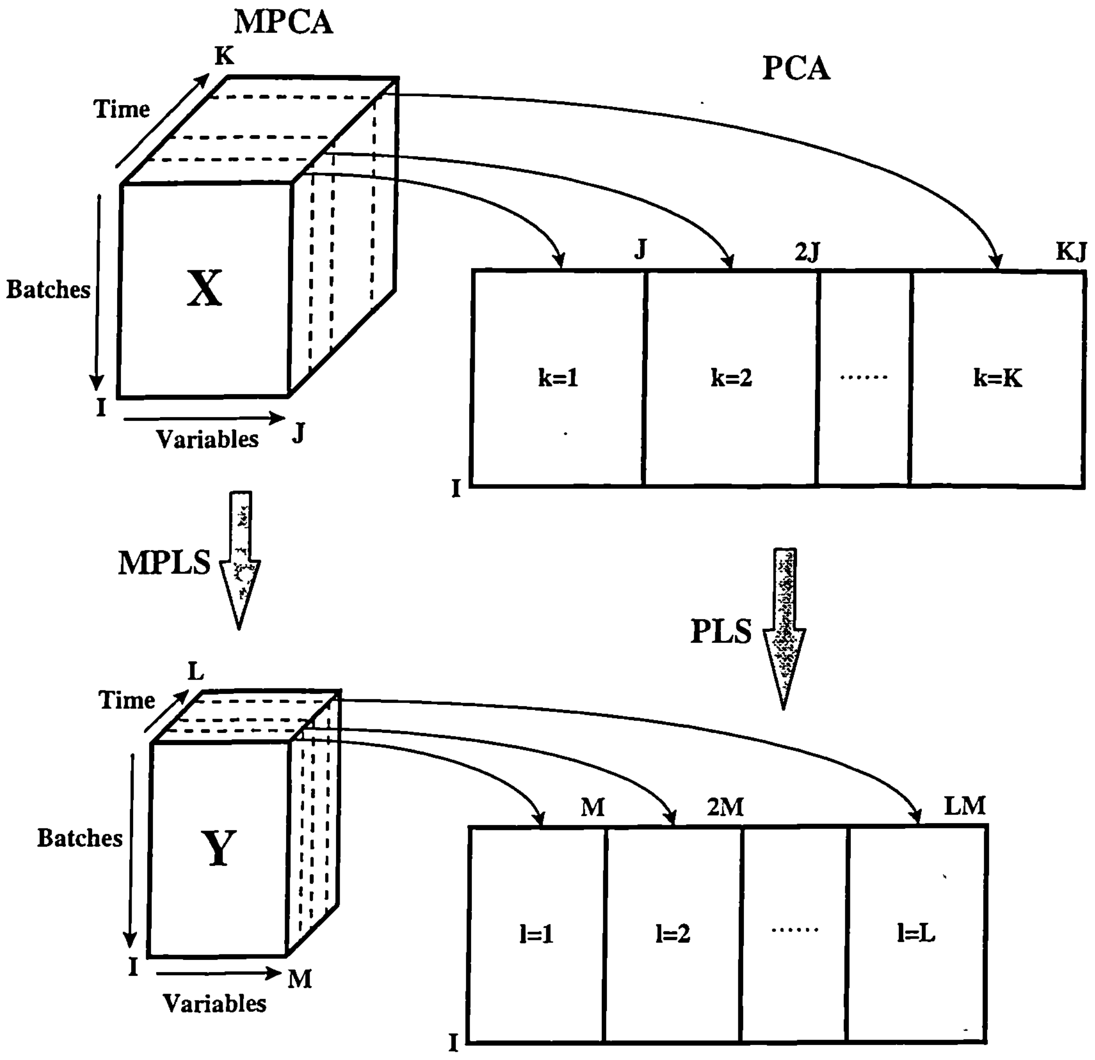


Figure 4.3 The procedure of unfolding three-way arrays

The objective of MPCA is to decompose the three-way array **X** into a series of principal components comprising score vectors ($\mathbf{t}_r$) and loading matrices (P_r), or unfolded as vectors ($\mathbf{p}_r$), plus a residual E, which is as small as possible, in a least squares sense :

$$\mathbf{X}=\sum_{r=1}^{R}\mathbf{t}_r\otimes \mathrm{P}_r+\mathrm{E} \quad \text{or} \quad \mathrm{X}=\sum_{r=1}^{R}\mathbf{t}_r\mathbf{p}_r^T+\mathrm{E} \tag{4.5.1}$$

This decomposition is in accordance with the principles of PCA. It separates the data block in an optimal way into two parts. The residual part (E) describes the noise added to the data and the systematic part ($\sum_{r=1}^{R}\mathbf{t}_r\mathbf{p}_r^T$), which expresses the data set as one fraction ($\mathbf{t}_r$) related only to batches and a second fraction ($\mathbf{p}_r$) related to variables and their time variation. In Multiway PLS (PLS), **X** and **Y** arrays are decomposed into a series of latent variables comprising score vectors and loading matrices, plus residual matrices E and F :

$$\mathbf{X}=\sum_{r=1}^{R}\mathbf{t}_r\otimes \mathrm{P}_r+\mathrm{E} \quad \text{or} \quad \mathrm{X}=\sum_{r=1}^{R}\mathbf{t}_r\mathbf{p}_r^T+\mathrm{E} \tag{4.5.2}$$

$$\mathrm{Y}=\sum_{r=1}^{R}\mathbf{t}_r\otimes \mathrm{Q}_r+\mathrm{F} \quad \text{or} \quad \mathrm{Y}=\sum_{r=1}^{R}\mathbf{t}_r\mathrm{q}_r^T+\mathrm{F} \tag{4.5.3}$$

Again, these decompositions are in accordance with the principles of PLS. The score vectors ($\mathbf{t}_r$) are orthogonal in both methods, whilst the loading (P) and weight (W) matrices or unfolded vectors ($\mathbf{p}_r$ and $\mathbf{w}_r$, respectively) are orthonormal. Each element of a score vector, corresponds to a single batch and describes the overall variability of this batch with respect to the other batches in the database, throughout the batch duration. Each loading vector or weight matrix, summarises the time variation of the measurement variables about their average trajectories and provides the direction of maximum variability and give a simpler description of the covariance structure of the data. The decomposition of data sets that is performed by PCA and PLS, are shown in Figure 4.4.

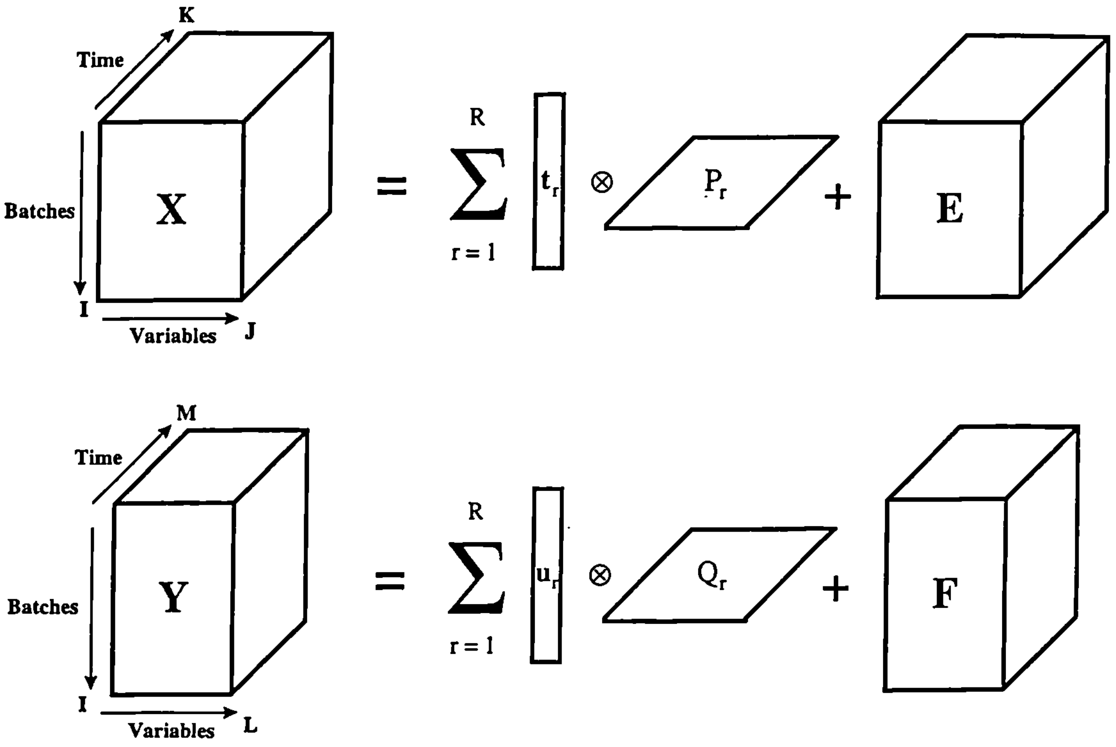


Figure 4.4 Arrangement and decomposition of the three-way arrays by MPCA and MPLS (Nomikos and MacGregor, 1994).

4.5.1.1 MSPC-based Schemes Using Multiway- PCA and PLS

In this section, MSPC-based procedures for the on-line process monitoring, fault detection and diagnosis of batch processes in real-time, using statistical models of MPCA and MPLS are presented. The philosophy is very similar to that of traditional SPC methods, where the future behaviour of the process is compared against a reference distribution based on past process history. The reference distribution is the history of past successful batches that have produced good quality product. The MPCA or MPLS model is built from the reference batches which characterise normal operation of the process and extracts all the information needed to monitor the behaviour of a new batch.

Mean-centring is usually applied to the data prior to performing a MPCA or MPLS analysis. The mean of each column of the **X** and **Y** data set is subtracted from each data element of this column. The way in which the **X** and **Y** matrices are unfolded combined with mean-centring is very important since it results in the subtraction of the mean trajectory of each variable and, thereby, in the removal of the main non-linear component in the data. A PCA performed on this mean corrected data is, therefore, a study of the variation in the time trajectories of all the variables in all the batches about their mean trajectories (Nomikos and MacGregor, 1995). Furthermore, by scaling the variables in each column of **X** and **Y** to unit variance, one can handle differences in the measurement units between variables and give equal weight to each variable at each time interval.

The loading matrix P_r (J×K) in MPCA or the loading matrices P_r and Q_r (L×M) and weight matrix W_r (J×K) in MPLS contain most of the structural information about how the variable measurements deviate from their average trajectories under normal operation. If a new batch is to be tested for unusual behaviour, one can use these matrices (or the resulting unfolded vectors) to check the hypothesis by obtaining the predicted scores and residuals for this batch. Consider that two data sets X_{new} (K×J) and Y_{new} (L×M), containing measurements on the process variables and quality variables, respectively, from a new batch are obtained.

The procedure is as follows :

1. Unfold the X_{new} data matrix to a row vector $\mathbf{x}_{new}$ of dimension J·K and the Y_{new} data matrix to a row vector $\mathbf{y}_{new}$ of dimension M·L.

2. Mean-centre and scale the X_{new} and Y_{new} data sets to unit variance, using the means and standard deviations of the reference database.

3. Predict the vector of scores for all the retained principal components (R) of the MPCA model as :

$$\hat{\mathbf{t}}^T = \mathbf{x}_{new} P \tag{4.5.4}$$

   or the vector of scores for all the retained latent dimensions (R) of the MPLS model as (section 4.5.3.1, equation (4.5.34)) :

$$\hat{\mathbf{t}}^T = \frac{\mathbf{x}_{new} W}{\left(P^T W\right)} \tag{4.5.5}$$

   where P and W are (J·K×R) matrices, whose columns are the loading $\mathbf{p}_r$ and weight $\mathbf{w}_r$ vectors, respectively.

4. Calculate the row vector (1×J·K) of the residuals of the new batch :

$$\mathbf{e} = \mathbf{x}_{new} - \hat{\mathbf{t}}^T P^T \tag{4.5.6}$$

5. In the case where an MPLS model is used, the row vector $\hat{\mathbf{y}}$ (1×M·L) of predictions of the M quality variables at each time point l (l=1,..,L) and the row vector of residuals (**f**) can be obtained as :

$$\hat{\mathbf{y}}_{new}^T = Q\hat{\mathbf{t}} \tag{4.5.7}$$

$$\mathbf{f} = \mathbf{y}_{new} - \hat{\mathbf{y}}_{new} \tag{4.5.8}$$

   where Q is a (M·L×R) matrix, whose columns are the loading $\mathbf{q}_r$ vectors.

If the scores of this new batch lie inside the normal operational region and the residuals are small, then it can be concluded that, its operation is similar to that of the reference database of normal batch operation.

A problem arises when one wants to perform the test sequentially in time, as the new batch evolves, that is to on-line monitor the batch operation (steps 1-5). In this situation, the data set X_{new} is not complete until the end of the batch. At each time interval (k) during the batch, the matrix X_{new} only has the measurements up to that time interval. The rest of the X_{new} block from the current time interval (k) to the end of the batch (K) is still undefined. Several ways to overcome this problem have been proposed by Nomikos and MacGregor (1995) :

1. *Zero Deviations Method* : This method assumes that the future measurements are in perfect accordance with their mean trajectories as calculated from the reference database. The assumption behind this method is that the batch will continue normally for the rest of its duration with no deviations. Recalling that the X_{new} data set after scaling contains the deviations of the measurements from their mean trajectories, one has to fill the unknown measurements with zeros.

2. *Current Deviations Method* : This method assumes that the future deviations from the mean trajectories will remain constant at the currently exhibited deviations at time interval k, for the rest of the duration of the batch. The assumption behind this method is that the same errors will persist for the rest of the batch. One, therefore, has to fill the unknown measurements at time intervals k+1,...K, with the values that the variables have at the time interval k.

3. *Projection Method* : The projection method does not fill in the unknown part of the data set X_{new}, but rather, uses the ability of PCA or PLS to handle missing data. Considering the unknown future measurements as missing values from an object in MPCA or MPLS, one can use the principal components of the MPCA model or the latent variables of the MPLS model to predict these missing values. However, the estimates of the missing values have to be consistent with the already observed values up to the time interval k and with the correlation structure of the measured variables in the database, as defined by the loading matrix (P) of the MPCA model or the matrix of weights (W) of the MPLS model. This can be done by projecting the already known measurements down onto the reduced space and calculating the scores at each time interval k using the MPCA model as :

$$\mathbf{t}_k = \left(P_k^T P_k\right)^{-1} P_k^T \mathbf{x}_{new,k} \tag{4.5.9}$$

or using the MPLS model as :

$$\mathbf{t}_k = \left(W_k^T W_k\right)^{-1} W_k^T \mathbf{x}_{new,k} \tag{4.5.10}$$

where $\mathbf{t}_k$ is a vector containing the scores of the retained principal components or latent variables at time interval k, $\mathbf{x}_{new,k}$ is a vector containing the measurements that are known up to the time interval k, P_k (k·M×R) and W_k (k·M×R) are matrices having as columns all the elements of the loading vectors $(\mathbf{p}_r)$ and weight vectors $(\mathbf{w}_r)$, respectively, up to the time interval k, from all the retained principal components or latent variables, respectively (r=1,...,R). The matrices $P_k^T P_k$ and

$W_k^T W_k$ are always well-conditioned because of the orthogonality property of the loading and weight vectors, respectively.

4. *Multi-Model Method* : This last method is the most valid way to address the previously defined problem. One can build different MPCA or MPLS models, each one up to the time interval k, using only the information available up to that time. The loading vectors ($\mathbf{p}_{r,k}$) for each principal component r and the weight vectors ($\mathbf{w}_{r,k}$) for latent variable, respectively, at each time interval k, should be stored. The scores at each time interval k can then be calculated by applying the corresponding loading or weight vectors in equations (4.5.4) or (4.5.5), respectively. This approach supposes that the appropriate number of principal components or latent variables of the overall MPCA or MPLS model, respectively, is also sufficient for each of these local-in-time and individual models.

To monitor the progress of a new batch, as new measurements become available, the t-scores can be calculated using any of the methods described above. The scores describe the overall performance of the batch. For the three first methods, the best way to track the particular instant that something behaves differently is to use the Squared Prediction Error associated with the latest on-line measurements at time interval k from the process :

$$SPE_k = \sum_{c=(k-1)m+1}^{kM} e(c)^2 \qquad (4.5.11)$$

The Sum of Squared Residuals (or *Q-Residuals*) over all time periods is not a good indicator since it does not represent the instantaneous perpendicular distance of a batch from the reduced space as does the SPE, and it is affected by errors associated with the in-filling or projection of the future unknown measurements in the data block X_{new}. However, sometimes, in order to avoid repetitive time consuming computations involved in the calculation of the control limit of the SPE at each time point of a process monitoring procedure, it is better to use Box's equation (4.3.16) and to try to approximate g and h by matching the moments of the $g\chi^2_v$ distribution (Nomikos and MacGregor, 1995). Let m and u be the estimated mean and variance of the quadratic residuals at a particular interval k, then g and v are approximated by :

$$g = \frac{u}{2m} \qquad (4.5.12)$$

$$v = \frac{2m^2}{u} \qquad (4.5.13)$$

and, therefore, the control limit of the quadratic residuals (SPE) at a level of significance and for each time interval k are given by :

$$SPE_\alpha = \frac{v}{2m} \cdot \chi^2_{2m^2/u,\alpha} \qquad (4.5.14)$$

The matching moments method is susceptible to error when there are outliers in the data or the number of quadratic residual values is small. However, outliers will have been removed during the pre-treatment of the historical process data and, furthermore, quadratic residual values used to estimate the control limit at each time point, can be fairly large, in this case a smoothing moving window procedure can then be applied (Nomikos and MacGregor, 1995).

If the Multi-Model method is used, then the Q-Residual is the k-th instantaneous perpendicular distance of the batch from the reduced space and, therefore, the SPE at the time interval k, should be based upon the Sum of Squares Residuals :

$$SPE_k = SSR_k = \sum_{c=1}^{kM} e(c)^2 \qquad (4.5.15)$$

where SSR_k is the Sum of Squared Residuals for the MPCA or MPLS model that corresponds to the k-th time point, e is the vector of residuals *for the batch currently* being monitored and M is the number of the process variables. Any unusual behaviour can be detected by the deviation of the process scores or the SPE from the normal operation as defined by its confidence limits.

Nomikos and MacGregor (1995) have discussed in detail the four methods. The Zero Deviations method has the advantage of a simple graphical representation of the operation of the batch in the score plots and rapid detection of an abnormality in the SPE plots. For a new batch, operation always starts from the origin of the scores in the reduced space, that is zero, and progressively moves out. The drawback of this method is that the scores are not very sensitive, especially at the start of the batch run to detect abnormal operation. Under the assumptions of the Current Deviations method, the SPE chart is not as sensitive as in the Zero Deviations method, but the scores identify the occurrence of an abnormality more quickly. The Projection method has the greatest advantage of giving scores very close to their actual final values if at least 10% of the history of a new batch is known, especially for normal operation. Caution must be used at the beginning of a new batch, where this method may give large and unexplained scores values, since there is little information to work with. The Multi-Model method is the most valid approach, but the computational efforts and storage requirements are extremely large except in situations where the data blocks are small, due to short duration of the batch process, relatively small numbers of process variables or relatively small numbers of batches included in the reference database. The shapes of the control limits for the process scores and the SPE, are presented in Figure 4.5, for the four methods. The method that should be used depends upon the specific characteristics of the process under consideration. It can be seen that, if the given batch process does not exhibit persistent disturbances or variables with discontinuities in their trajectories, then it is better to use the Zero Deviation method. If there is a *prior* knowledge that the disturbances in the given process are persistent, then it is better to use the Current Deviations method. The projection method should be used whenever the trajectories of the process measurements do not exhibit frequent discontinuities or early deviations, whilst the Multi-Model method should be used when excessive computations and storage requirements are not an issue. In general, as has been proposed by Nomikos and MacGregor (1995), one can use a combination of the above methods, switching after some time to another one method, and, in this way, to build in some engineering knowledge into the monitoring scheme.

Fault detection and diagnosis in MSPC-based monitoring schemes are similar to those implemented for PCA and PLS models The only difference is that prediction of the scores is based upon the overall process operation duration. Therefore, the scores are calculated as the sum of the contribution of each process variable on a cumulative basis up to the time point of interest (k) :

$$t_{r,k} = \sum_{j=1}^{M} \sum_{c=1}^{k} x_{c,j} \cdot \omega_{r,c,j} \tag{4.5.16}$$

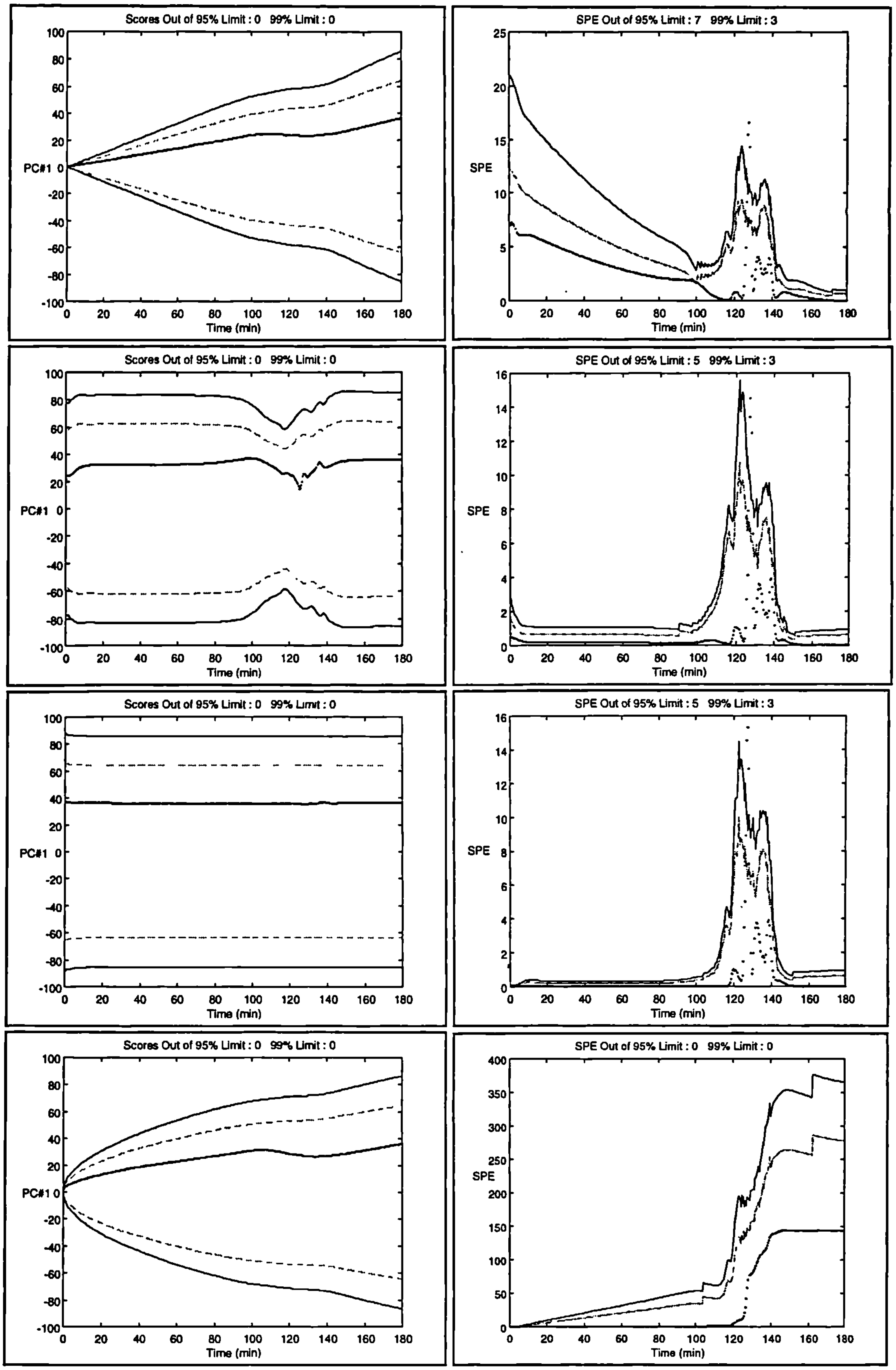


Figure 4.5 Control limits of scores and SPE for the four approaches (from top to bottom : zero deviations, current deviations, projection, multi-model)

The contribution of each variable is, therefore, given by :

$$VC_{j,k} = \sum_{c=1}^{k} x_{c,j} \cdot \omega_{r,c,j} \qquad (4.5.17)$$

where $\omega_{r,c,j}$ is the importance of the j-the variable on the r-th principal component or latent vector at time point c ($c \leq k$).

MPCA and MPLS techniques can be used to construct statistical models for any process or stage of a process where process data sets can take the form of a three-way arrays, such as batch and semi-batch processes.

4.5.2 Multi-Block Extensions of PCA and PLS

In the case where there is a large number of data items (objects or variables) to interpret or analyse, there are two commonly used approaches (Wold et al., 1996). The first approach, *sampling*, is where only a few data items are selected and looked at in detail, the remainders are neglected. The second approach is to divide data items into groups, blocks, categories or clusters and then to consider these groups as *super-items*. This approach is called *grouping* or *blocking*. Sampling is usually applied to cases when there are many variables, while blocking is applied when there are many objects.

Multiple Linear Regression (MLR) is the most commonly applied method for the statistical modelling of data. MLR requires more objects than variables in order to provide a well-conditioned matrix of input data. This has created the tendency to drastically reduce the number of predictor variables in a model. The reduction of variables in regression (sampling) is usually made by deleting those that have small blocked data set can be uniquely designated as a predictor or response set. On the other hand, Interconnected Multi-Block PLS (IMB-PLS) considers that there is more than one predictive relationships between different groups of predictor and response blocked data sets and, furthermore, that a blocked data set can be both a predictor and a response. However, in both approaches the variables are grouped according to their similarity or according to their origin or location in the process.

Multi-Block PCA (MBPCA) is an extension of PCA that can be derived in a similar way to that of Hierarchical Multi-Block PLS (HMB-PLS) (Cheng and McAvoy, 1996). It was originally proposed by Wold et al. (1987b). Using MBPCA, a data set X can be broken down into a set of A subsets X_a (a=1,...,A) by grouping the original process variables in a meaningful way. For each subset X_a, a score vector ($\mathbf{t}_{a,r}$) and a loading vector ($\mathbf{p}_{a,r}$) can be calculated for each principal component (r), according to the NIPALS algorithm. The scores from all the subsets are then collected into a composite matrix T_r and a *consensus* score vector ($\mathbf{t}_r$) and loading vector ($\mathbf{p}_r$) can be calculated by applying standard PCA to T_r. This procedure is repeated for the maximum number of principal components that can be extracted from the subsets X_a, that is equal to the minimum rank of the subsets X_a. Similarly, in the Hierarchical Multi-Block PLS (HMB-PLS), the two data sets X and Y are broken down into A X_a (a=1,...,A) subsets and B Y_b (b=1,...,B) subsets, respectively. For each latent dimension (r), the loading and score vectors of each subsets can be calculated. The scores $\mathbf{t}_{a,r}$ and $\mathbf{u}_{b,r}$ of the X_a and Y_b subsets, respectively, are collected into two composites matrices T_r and U_r. Consensus vectors of scores $\mathbf{t}_r$ and $\mathbf{u}_r$ can then be calculated by performing a NIPALS-PLS loop between these composite matrices. This procedure is illustrated for A=3 and B=2 in Figure 4.6 and it can be repeated for the maximum number of latent variables that can be extracted from the subsets X_a and Y_b, that is equal to the minimum of the ranks of the subsets X_a and Y_b.

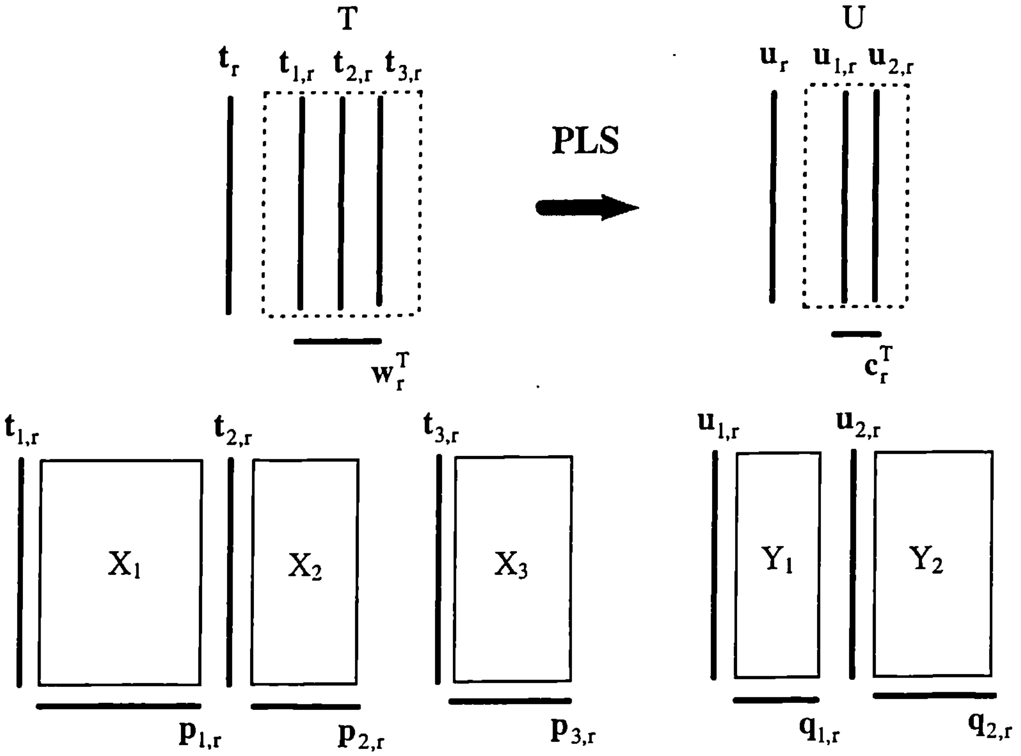


Figure 4.6 Hierarchical Multi-Block PLS (Wold et al., 1996).

Interconnected Multi-Block PLS (IMB-PLS) is especially suited for large complex systems, which consist of many distinct sections that are connected by a few variables (Wangen and Kowalski, 1988). In complex systems, the system can be broken into several blocks, each one corresponding to a distinct section of the system. Using the IMB-PLS technique, the data sets X and Y are pooled together in a data set Z, which is broken up into A subsets Z_a (a=1,...,A). Having calculated the loading and score vectors of each subset Z_a, a predictive relationship is defined every time a subset Z_g predicts or is predicted by one or more other subsets Z_h

($1 \leq g \leq A$; $1 \leq h \leq A$; $g \neq h$) and, therefore, more than one composite matrix T and U and sets of consensus score and loading vectors are defined. The advantage of IMB-PLS is mainly to allow easier interpretation of the data by looking at smaller more meaningful blocks and at predictive relationship between blocks. Figure 4.7 illustrates a typical structure of interconnected data blocks.

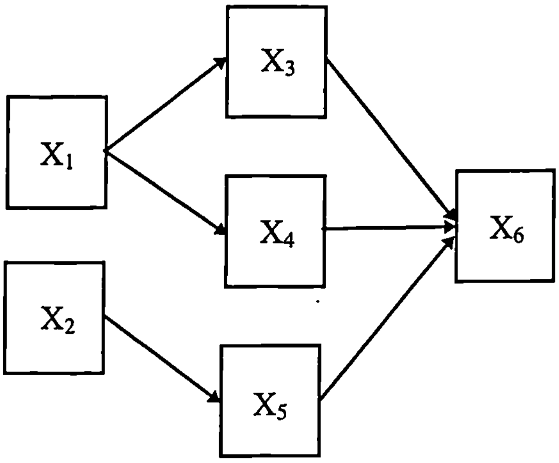


Figure 4.7 A typical Interconnected Multi-Block PLS structure

4.5.2.1 Interconnected Multi-Block PLS

Interconnected Multi-Block PLS (IMB-PLS) is similar to PLS. However, there are two main differences between the two techniques :

a. PLS models the predictive relationship between two blocks of data, whilst the IMB-PLS models the predictive relationship(s) between more than two data blocks.

b. In IMB-PLS, a block can predict more than one block and can be predicted by more than one block. In PLS, a block can only predict one block or can only be predicted by one block.

The second difference is of great importance and differentiates the NIPALS algorithm used in PLS, from the algorithm used in IMB-PLS. Having described the basic concepts of PLS in Chapter III, it is, now, possible to focus upon the structure of the NIPALS algorithm and, specifically, on those aspects of it which allow it to be extended to handle more than two blocks of data.

PLS - NIPALS algorithm

The objective of PLS is to build a linear relationship between a block (Y), which comprises measurements of the dependent variables, and a block (X), which contains measurements of the independent variables. The words *predictor* and *predictee* are used to define X and Y blocks, respectively. The basic steps of the NIPALS algorithm (section 3.5) for the calculation of each latent variable can be summarised as follows (NIPALS variation adopted from Hoskuldsson, 1988):

1. Initialisation

2. *Backward phase* - Calculation of scores (**t**) of the predictor block (X)

2a. Regression of predictee's score (**u**) on predictor. Predictor's weight (**w**) is calculated to be :

$$\mathbf{w}^T=\mathbf{u}^T X$$

2b. Regression of predictor's weight on predictor. Predictor's score (**t**) is calculated to be :

$$\mathbf{t}=X\mathbf{w}$$

3. *Forward phase* - Calculation of scores (**u**) of the predictee block (Y)

3a. Regression of predictor's score (**t**) on predictee. Predictee's weight (**c**) is calculated as :

$$\mathbf{c}^T=\mathbf{t}^T\mathrm{Y}$$

3b. Regression of predictee's weight on predictee. Predictee's score (u) is calculated as :

$$\mathbf{u}=\mathrm{Y}\mathbf{c}$$

4. Convergence check

5. Calculation of loadings for both blocks

6. Calculation of regression coefficients

7. Calculation of residual matrices

In the previous algorithm the most important steps are the backward and the forward phase where the scores are calculated (steps 2 and 3). In these steps :

- the original high-dimensional space is projected down onto new low-dimensional subspace by orthogonal projection
- the orthogonal properties of the latent vectors are derived.

The scores of each block are the projections of the block down onto the latent dimensions. Therefore, scores represent their corresponding blocks in the reduced subspace. Furthermore, it can be seen that :

- in the backward phase (step 2a), the **u**-scores, which are the representation of the predictee block (Y) in the subspace are regressed upon the predictor block (X)
- in the forward phase (step 3a), the **t**-scores, which are the representation of the predictor block (X) in the subspace are regressed upon the predictee block (Y)

Moreover, the NIPALS algorithm iterates from the right-end (predictee) block to the left-end (predictor) block during the backward phase and vice versa during the forward phase. The dimension of the vectors in the algorithm can be depicted as follows :

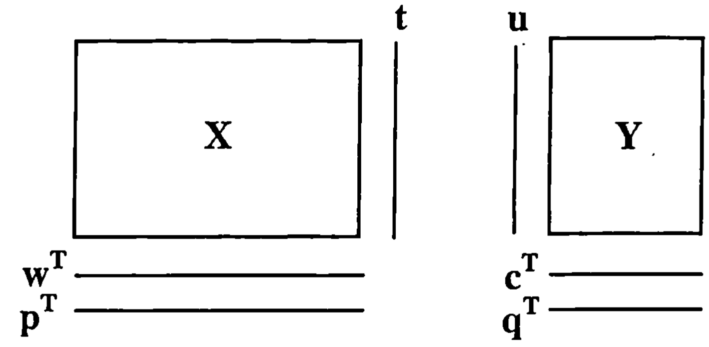


Figure 4.8 Dimension of vectors calculated by NIPALS (Hoskuldsson, 1988).

IMB-PLS - NIPALS algorithm

The most important difference between the IMB-PLS and the PLS methods, as previously stated, is that a block in IMB-PLS can predict or be predicted by more than one of the other blocks. The iterative algorithm used to calculate the latent variables in IMB-PLS is an extension of the NIPALS algorithm. The algorithm iterates through all the blocks from right to left and then from left to right, during the backward and the forward phases, respectively. Analogous to the PLS-NIPALS, the words predictor and predictee will indicate blocks that predict and blocks that are predicted, respectively. Furthermore, a "X" with the appropriate subscript, will denote any predictor or predictee data block.

The NIPALS algorithm models the relationship between two blocks. As a result, when a block is a predictor or a predictee of more than one block, then the algorithm needs to include additional steps to account for these additional blocks. In the case where a block is the predictor of only one block in the backward phase or the predictee of one block only in the forward case, then the algorithm proceeds as in PLS (steps 2a, 2b and 3a, 3b, respectively). In the case of a multiple predictor or predictee block (X_G), the NIPALS algorithm used in PLS is not applicable. In order to overcome this difficulty, all predictees or predictors blocks are compressed into one block. The most meaningful way to compress them is to utilise their scores, since they are representative of the blocks in the reduced space, and to combine them into a new composite block. This block is then used as the predictee or the predictor in the calculations. However, in order to apply NIPALS at this stage, a representation of the composite block in the reduced space needs to be calculated. Therefore, in the case of a multiple predictor or predictee block (X_G), the backward phase of the NIPALS algorithm has to be modified as follows :

2. *Backward phase* - Calculation of the scores (**t**) of a multiple predictor block X_G (predicts more than one block)

2a. Combine all the **u**-scores of the blocks that X_G predicts into a new composite matrix U. Define U as the predictee block.

2b. Calculate the scores of the predictee block - (i). Regress the predictor scores (**t**) on the predictee (U). The predictee weights ($\mathbf{c}_U$) can then be calculated by :

$$\mathbf{c}_U^T = \mathbf{t}^T U$$

2c. Calculate the scores of the predictee block - (ii). Regress the predictee weights ($\mathbf{c}_U$) on the predictee (U). The predictee scores ($\mathbf{u}_U$) can be calculated as:

$$\mathbf{u}_U = U\mathbf{c}_U$$

2d. Regress the predictee scores ($\mathbf{u}_U$) on the predictor. Calculate the predictor weights ($\mathbf{w}$) as :

$$\mathbf{w}^T = \mathbf{u}_U^T X_G$$

2e. Regress the predictor weights on the predictor. Calculate the predictor scores ($\mathbf{t}$) as :

$$\mathbf{t} = X_G\mathbf{w}$$

An example of a backward phase, where block X_G predicts blocks X_K and X_L, is illustrated in Figure 4.9.

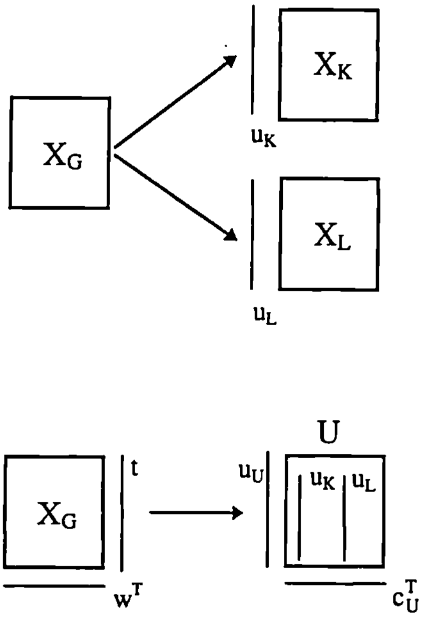


Figure 4.9 Adapted backward phase in NIPALS algorithm used in IMB-PLS.

Similarly, the forward phase of NIPALS algorithm has accordingly to be modified :

3. *Forward phase* - Calculation of scores ($\mathbf{u}$) of a multiple predictee block (X_G)

3a. Combine all the scores ($\mathbf{t}$) of the blocks that predict X_G into a new consensus matrix T. Now, define T as the predictor block.

3b. Calculate the scores of the predictor block - (i). Regress the predictee scores (**u**) on the predictor (T) and calculate the predictor weights ($\mathbf{w}_T$) as :

$$\mathbf{w}_T^T = \mathbf{u}^T T$$

3c. Calculate the scores of the predictor block - (ii). Regress the predictor weights ($\mathbf{w}_T$) on the predictor (T) and calculate the predictor scores ($\mathbf{t}_T$) as :

$$\mathbf{t}_T = T\mathbf{w}_T$$

3d. Regress the predictor scores ($\mathbf{t}_T$) on predictee. Calculate the predictee weights (**c**) as follows :

$$\mathbf{c}^T = \mathbf{t}^T X_G$$

3e. Regress the predictee weights on the predictee. Calculate the predictee's scores (u) as follows :

$$\mathbf{u} = X_G \mathbf{c}$$

An example of a backward phase, where block X_G is predicted by blocks X_K and X_L, is illustrated in Figure 4.10.

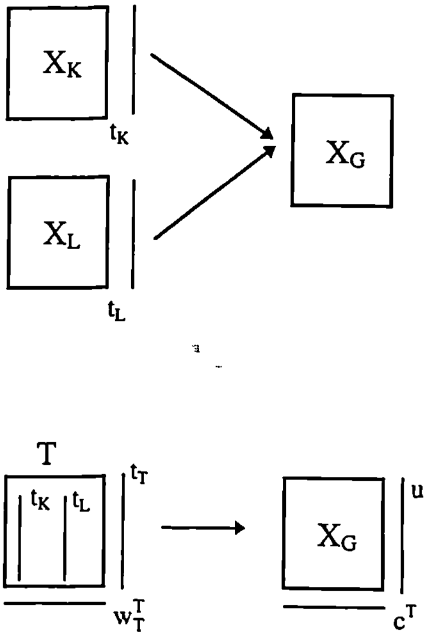


Figure 4.10 Adapted forward phase in NIPALS algorithm used in IMB-PLS.

4.5.2.1 Issues Concerning the Multi-Block Techniques

The major problem arising in the implementation of Multi-Block techniques is the selection of the structure of the system, that is the selection of variables to be included in the same block, since it is not something for which specific rules can easily be defined. The choice of blocks will, generally, depend upon engineering judgement, prior knowledge and the objectives of the study. There are two practical approaches to grouping variables. Using the first approach, blocks should correspond as closely as possible to distinct sections of the system, where there is maximum coupling between all variables within a block and minimum coupling between variables in different blocks. Variables associated with streams connecting two or more blocks, such as feed and recycle streams, should be included in all these blocks. Using the second approach, variables of the same type that measure the same physical quantity, such as temperatures or pressures, should be included in the same block. However, all possible blockings should be compared in order to select the most efficient, since poor blocking can lead to spurious models that are unable to perform their task.

Another important issue is the number of control charts that process operators have to monitor. Using an MSPC-based scheme that utilises a statistical model of the ordinary PLS technique, one should at least monitor 2-3 control charts (latent variables plots and SPE plot). In the case of an inter-connected process comprising several distinct processing units, an efficient approach is to build a PLS model for each separate unit. As a result, this increases the number of control charts that operators are required to monitor. However, by applying IMB-PLS the number of control charts is reduced. Specifically, the blocks that predict a multiple predictee block can be replaced by their composite matrix, since as will be shown in the subsequent chapter, abnormal events can possibly be detected in the scores plots of consensus matrices. This can drastically reduce the number of control charts depending upon the type of blocking that was applied to the process.

4.5.3 Inverse Projection to Latent Structures

MSPC-based schemes for process monitoring, fault detection and diagnosis have typically been applied to industrial processes where large amounts of historical data have been collected. However, difficulties can arise in the development of a robust MSPC-based scheme if only minimal plant data is available. There are many industrial situations where only small data sets are available from either an experimental design or initial product commissioning tests, which have been used to establish appropriate plant operating conditions. The IPLS methodology provides a novel approach based upon the inversion of a PLS regression model built upon a few process data. New process data are then constructed by interpolating from within a nominal region which is defined by the 'design' process data. The Inverse Projection to Latent Structures (IPLS) method proposed requires the derivation of a well-defined PLS model to estimate a predictor set of data X, which is consistent, in a statistical sense, with an "*a priori*" specified desired response set of data Y. The methodology developed is primarily based upon the Multiple Linear Regression (MLR) or Ordinary Least Squares (OLS) approaches, as presented by Seber, (1977).

Multiple Linear Regression (MLR) is the most commonly applied method for developing multivariate statistical regression models. However, a number of problems can be encountered when large data sets comprising highly correlated measurements, are presented to the technique. Typically, the derived coefficients will have large variances and, hence, they will be unstable when small changes in the data occur. This can result in major changes in the regression coefficients. On the other hand, the power of PLS lies in its ability to handle data of this type. Both MLR and PLS deal with the same generalised regression problem and, therefore, it is important to identify and understand the similarities and dissimilarities between the two approaches in order to obtain an appreciation of the mechanisms of the inverse PLS approach.

4.5.3.1 Derivation of Inverse PLS

The derivation of the statistical properties used in the methodological development paper are based upon the work of Searle (1984) for the generalised inverse approach to regression and upon the work of Nomikos and MacGregor (1994) for the multivariate regression modelling and PLS approaches. Assume that a statistical regression model between a predictor data set (X) and a response data set (Y) of the following form exists :

$$Y = X\,\beta \tag{4.5.18}$$

The linear regression coefficients (β) can then be estimated from the predictor and response data sets, X and Y, respectively :

$$\hat{\beta} = X^{+}\,Y \tag{4.5.19}$$

where X^+ is the generalised inverse of X. There are several approaches to determining the generalised inverse. The most frequently applied solution is based upon least squares estimation :

$$X^+ = \frac{X^T}{X^T X} \tag{4.5.20}$$

This approach can be mathematically inappropriate due to two of the most common problems associated with MLR. The theory requires that the number of objects (N) is greater than, or equal to, the number of predictor variables (M), i.e. $N \geq M$. In industrial situations this scenario is frequently not realisable. Furthermore, the input block matrix X can be ill-conditioned due to collinearity between the variables, i.e. one variable is approximately a linear combination of a number of the other process variables. This results in a problem with the inversion of $X^T X$, since it will have a determinant equal or close to zero and will therefore be singular. Alternatively, the generalised inverse X^+ can be calculated using the properties and relationships arising from PLS :

$$X^+ = \frac{W\,T^T}{\left(P^T W\right)\left(T^T T\right)} \tag{4.5.21}$$

In practice the PLS method only gives a right weak generalised inverse X^+ of the PLS approximation to the original predictor data set X, that is $\hat{X} = T\,P^T$. The definition of a right weak generalised inverse X^+ of $\hat{X}$ implies that :

$$\hat{X}\,X^+\,\hat{X} = \hat{X} \tag{4.5.22}$$

$$X^{+}\,\hat{X}\,X^{+} = X^{+} \tag{4.5.23}$$

$$\hat{X}\,X^{+} = \left(\hat{X}\,X^{+}\right)^{T} \tag{4.5.24}$$

According to Rao (1971), a right weak generalised inverse such as X^{+} can be termed the least squares generalised inverse, since it gives the least squares solution $\hat{\beta}$:

$$\left|\hat{X}\,\hat{\beta} - Y\right| \leq \left|\hat{X}\,\gamma - Y\right| \qquad \forall\gamma \tag{4.5.25}$$

For the modified regression problem :

$$Y = \hat{X}\,\hat{\beta} \tag{4.5.26}$$

The linear regression coefficients estimates, $\hat{\beta}$, can then be calculated using the PLS relationship :

$$\hat{\beta} = \frac{W\,T^{T}}{\left(P^{T}\,W\right)\left(T^{T}\,T\right)}\,Y \quad \Rightarrow \quad \hat{\beta} = \frac{W\,Q^{T}}{\left(P^{T}\,W\right)} \tag{4.5.27}$$

Proof :

$$\hat{X} = TP^{T} \Rightarrow T = \frac{\hat{X}P}{P^{T}P} \tag{4.5.28}$$

$$\Rightarrow P = \frac{\hat{X}^{T}T}{T^{T}T} \tag{4.5.29}$$

$$Y = TQ^{T} \tag{4.5.30}$$

$$T = \frac{XW}{W^{T}W} \tag{4.5.31}$$

$$Y = TQ^{T} \overset{(4.5.28)}{=} \frac{\hat{X}PQ^{T}}{P^{T}P} \overset{(4.5.26)}{=} \hat{X}\hat{\beta} \Rightarrow \hat{\beta} = \frac{PQ^{T}}{P^{T}P} \tag{4.5.32}$$

$$\hat{\beta} = \frac{PQ^{T}}{P^{T}P} \overset{(4.5.29)}{=} \frac{TQ^{T}}{P^{T}T} \overset{(4.5.31)}{\Rightarrow} \hat{\beta} = \frac{WQ^{T}}{P^{T}W} \tag{4.5.33}$$

Therefore, it follows (4.5.27) :

$$\hat{\beta} = \frac{WQ^{T}}{P^{T}W} \overset{(4.5.30)}{=} \frac{WT^{T}}{\left(P^{T}W\right)T^{T}T}Y$$

Furthermore :

$$\hat{\beta} = \frac{WQ^{T}}{P^{T}W} \Rightarrow \hat{X}\hat{\beta} = \frac{\hat{X}WQ^{T}}{P^{T}W} \underset{(4.5.30)}{\overset{(4.5.26)}{\Rightarrow}} T = \frac{XW}{P^{T}W} \tag{4.5.34}$$

Although PLS gives the least squares solution for the regression problem, equation (4.5.26), it does not uniquely define X^{+} and hence $\hat{\beta}$. However, for any choice of the generalised inverse X^{+}, the regression model, given by equation (4.5.26), generates a unique projection $\hat{y}$ on the space spanned by the linearly independent columns of $\hat{X}$ (Seber,1977). Furthermore, X^{+} can also be a left weak, or a minimum norm, generalised inverse of $\hat{X}$, and if the matrix $X^{+}\hat{X}$ is symmetric, X^{+} becomes the Moore-Penrose generalised inverse. It is known that PLS does not provide the minimum norm solution for the regression problem, equation (4.5.26) which is unique. However, its solution is as close as the PLS approximation $\hat{X}$ is to the original input matrix X, since it can be shown that $X^{+}\hat{X}$ is symmetric when X has full column rank. That is, if $X = \hat{X}$, then :

$$X^{+}\,\hat{X} = \frac{W\,T^{T}}{\left(P^{T}\,W\right)\left(T^{T}\,T\right)}\left(T\,P^{T}\right) = \frac{W\,P^{T}}{P^{T}\,W} = \frac{T\,W\,P^{T}}{X\,W} = \frac{\hat{X}}{X} = I \qquad (4.5.35)$$

Having developed a linear regression model, equation (4.5.18), with regression coefficients $\hat{\beta}$, suppose that interest is now focused upon predicting a new predictor data vector $\mathbf{x}_0$ from a vector of predefined values of the response variables, $\mathbf{y}_0$:

$$\mathbf{y}_0^T = \mathbf{x}_0^T\,\hat{\beta} \qquad (4.5.36)$$

By inverting the regression model, equation (4.5.27), and solving for $\mathbf{x}_0$, an equation system with three possible solutions, depending on the number of predictor (M) and response (K) variables, is obtained :

a) $K > M$ Model inversion corresponds to a projection from a high to a lower dimensional space. This is a standard least squares projection.

b) $K = M$ An exact inversion between the two dimensional spaces is possible.

c) $K < M$ A projection from a lower to higher dimensional space results. This is the most difficult but the most common outcome. In this situation the equation system is undetermined.

In the first two cases (a and b) there exists a unique least squares solution, whilst in the third case there are an infinite number of solutions. However, for the last case

($K < M$), a natural estimate, which is also the maximum likelihood estimate, $\dot{x}_0$, can be obtained through a least squares generalised inverse of $\hat{\beta}$, (Seber,1977) :

$$\dot{x}_0^T = y_0^T \frac{\hat{\beta}^T}{\hat{\beta}\,\hat{\beta}^T} \tag{4.5.37}$$

This equation is termed the classical estimator. However, the estimate $\dot{x}_0$ obtained is a biased estimator of x_0, since :

$$E\left(\dot{x}_0^T\right) \neq E\left(y_0^T\right) E\left(\frac{\hat{\beta}^T}{\hat{\beta}\,\hat{\beta}^T}\right) = E\left(x_0^T\right) \tag{4.5.38}$$

Although $\dot{x}_0$ is a biased and, thus, not a unique estimate of x_0, in a least squares sense, it gives the unique solution of y_0 to the regression problem, i.e. equation (4.5.27). The regression coefficients $\hat{\beta}$ in equation (4.5.27) have been calculated based upon the underlying PLS relationship. Consequently, they model the internal structure of the predictor data set (X) and response data set (Y) by maximising the covariance between the two data sets and, hence, summarising the internal relationship.

The inverse estimate $\dot{x}_0$ obtained from equation (4.5.37), i.e. the Inverse PLS estimate (IPLS), is not unique but it is a justifiable least squares estimate of the PLS approximation $\hat{x}_0$ of the new predictor data vector x_0, given the desired response data vector y_0. Therefore, depending upon how close the PLS approximation $\hat{x}_0$ is to the original data vector x_0, the classical estimator equation, equation (4.5.28), gives a realistic estimate of x_0 in terms of $\dot{x}_0$. The IPLS estimate lies in K-dimensional space, whereas the true dimensionality of the input variables space, M, is higher. Jaeckle and MacGregor, (1996), in a similar problem, but using principal components regression (PCR), proposed that a new component $\mathbf{z}_0$, lying in a space orthogonal to the generalised inverse of $\hat{\beta}$ and which spans the remaining $(\mathrm{M}-\mathrm{K})$ dimensions of the X space, should be added to $\hat{\mathrm{x}}_0$. Regardless of this, however, the IPLS estimate $\dot{\mathrm{x}}_0$ proposed here is theoretically justifiable. In practice, providing the initial PLS model is satisfactory, IPLS is capable of predicting values for its output variables close to their true values. The only requirement of this approach is the establishment of a good regression model using the PLS method. Inferential PLS models can then be built from process and quality measurements, initial process conditions etc., depending upon the availability of the data and the nature of the process and the problem.

4.5.3.2 Methodology For the Application of IPLS

In manufacturing processes where there is only a limited amount of process data available or a small number of completely recorded operations, techniques to generate more process data would be particularly advantageous in order to set up an initial MSPC-based monitoring scheme. In order to examine this, let us consider the situation either where a new process is being set up, or an existing process is being expanded into a new operating region. As a consequence, there are only a few process data measurements available from designed experiments to identify the nominal product quality and the associated operational conditions. The assumption is made that this data represents past successful operation and, therefore, it can be assumed to define the nominal operating range of desired production and, hence, the regression model. Having identified the set of data to be used to build the empirical representation, a PLS regression model is calculated to model the features of interest such as the initial process conditions and/or the final product quality from the process measurements (4.5.18) :

$$Y = X \beta$$

Having derived the underlying PLS model based upon minimal process data, an unknown new process measurement vector ($\mathbf{x}_0$) can be predicted from the corresponding predefined values of the features ($\mathbf{y}_0$), by inverting the PLS regression model to obtain the classical estimator, equation (4.5.37) :

$$\dot{\mathbf{x}}_0^T = \mathbf{y}_0^T \frac{\hat{\beta}^T}{\hat{\beta}\,\hat{\beta}^T}$$

In this way, using the process measurements computed from the IPLS model, a large number of process data within the nominal region of the regression model can be computed. These IPLS estimates, along with the existing plant measurements, can then be used to develop an initial MSPC-based monitoring scheme. As new process measurements become available from the production plant, itself, new and improved PLS and IPLS models can be built from the updated historical process database. Finally, when sufficient data from the actual manufacturing process becomes available, a robust MSPC scheme follows naturally based purely upon the plant data as is the current approach.

The above methodology offers an attractive and effective way to implement MSPC-based monitoring schemes, even in cases where there is initially limited process data from the plant. One of the key requirements of this approach is that a satisfactory PLS regression model can be built from the initial 'design' data. An additional assumption is that the process 'design' data define the nominal region of the PLS and IPLS models and that they are representative of the desired operating region for acceptable production. Thus, process data deemed to be outside the operating region needs to be excluded from the analysis.

Furthermore, quadratic residuals are unreliable measures of operating performance when IPLS-estimated process measurements have been used to develop a statistical representation/model. In general, the Sum of Squared Residuals (SSR or Q-statistic) of the IPLS estimated process data exhibit unusually large value. This result can be explained by looking at the composition of the residual sum of squares, Q, more closely :

$$Q = \sum_{i=1}^{\text{Objects}} \sum_{j=1}^{\text{Variables}} E_{i,j}^2 \qquad (4.5.39)$$

The Q-statistic is a metric based upon a measure of the deviation of the process measurements from the centre of the reduced space that the statistical model define, given by the residual, E. This is calculated for each individual object vector $\mathbf{x}_i$ as :

$$\mathbf{e}_i = \mathbf{x}_i - \mathbf{x}_i \mathrm{PP}^{\mathrm{T}} \qquad (4.5.40)$$

where $\mathbf{x}_i$ is the data block containing the process measurements of the i-th object vector of process data and P is the matrix of the loadings. The calculation of the control limits for the Q-statistic requires the calculation of the residuals for each object vector of the nominal database which in the case of the IPLS-based model can be written as :

$$\mathbf{e}_i = \dot{\mathbf{x}}_i - \dot{\mathbf{x}}_i \, \dot{\mathrm{P}} \, \dot{\mathrm{P}}^{\mathrm{T}} \tag{4.5.41}$$

where $\dot{\mathrm{P}}$ denotes the loadings array for the MPCA representation based upon the IPLS estimated process data and $\dot{\mathbf{x}}_i$ denotes the IPLS estimates of the process measurements for each object vector. These estimates contain an error associated with the approximation of the real process measurements by their IPLS estimates. This error cannot be calculated since the real measurements of the operations will in practice be unavailable and, therefore, it will be inherited to the control limits. However, when calculating the Q-statistic for a new object vector comprising real measurements, this error is not present (in the measurements) :

$$\mathbf{e}_{\text{new}} = \mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}} \dot{\mathrm{P}} \dot{\mathrm{P}}^{\mathrm{T}} \tag{4.5.42}$$

where $\mathbf{x}_{\text{new}}$ is the vector of the new real measurements. This results in the calculated values for the Q-statistic and the Squared Prediction Error (SPE) for each new real process data to potentially exceed the nominal control limits. As a result the residual sum of squares and the squared prediction error are unreliable measures of operating performance as they will contain an error which is not quantifiable and which will inflate the two metrics.

4.6 Applications

A number of applications of MSPC-based monitoring schemes has been presented in the literature. A 'process-oriented' literature survey has been presented in Chapter II. In this chapter a more 'technique-oriented' survey is presented.

Multi-Way PCA and Multi-Way PLS techniques have been introduced and successfully applied in batch processes mainly by the group of Professor J.F. MacGregor at McMaster University in Canada. Nomikos and MacGregor (1994 and 1995) have applied an MSPC-based scheme for a semi-batch emulsion styrene-butadiene rubber (SBR) polymerisation reactor and for an industrial polymerisation reactor, respectively, using the Multi-Way PCA technique. A similar scheme for the same SBR polymerisation reactor, but using the Multi-Way PLS technique has also been developed (Nomikos and MacGregor, 1994b). Other MSPC-based schemes that utilise Multi-Way PCA have been presented by Dong and McAvoy (1995) for a two-stage jacketed exothermic batch chemical reactor, by Gallagher et al. (1996) for a nuclear waste storage tank and by Kosanovich et al. (1996) for an industrial polymerisation reactor.

MSPC-based schemes for a two-zone LDPE tubular reactor has been developed using the Multi-Block PLS technique by MacGregor et al. (1994). Wold et al. (1996) applied the Hierarchical Multi-Block PLS technique to an industrial Residue Catalytic Cracker unit (RCCU). Finally, MSPC-based schemes that utilise a combined Multi-Way Multi-Block PLS technique, for an industrial polymerisation batch reactor has been presented by Kourti et al. (1995). Cheng and McAvoy (1996) have recently proposed the same combined Multi-Way Multi-Block approach for continuous dynamic processes.

4.7 Summary

Multivariate statistical projection methods, such as PCA and PLS, are known to be capable of establishing MSPC-based schemes for process monitoring, fault detection and diagnosis. The typical procedures that have to be followed and the issues associated with the implementation of an efficient MSPC scheme have been described. However, sometimes, these techniques fail to perform their task efficiently, since the techniques are only appropriate for continuous processes which do not exhibit non-linear behaviour and which only involve data from simple processing unit. These limitations can be resolved by applying more suitable statistical techniques. Specifically, Multi-Way PCA and PLS have been proposed to analyse data obtained from processes that exhibit dynamic character, such as batch and semi-batch processes. Furthermore, Multi-Block PCA and PLS have been proposed to handle industrial processes comprising interconnected sections and units. Finally, a novel approach for generating process data for MSPC schemes where there is only minimal process data, is proposed. The approach is based upon the inversion of a PLS regression model.

Chapter V

MSPC-Based Applications to Chemical Processes

5.1 Introduction

This chapter describes the application of multivariate statistical projection techniques for the development of MSPC-based monitoring schemes for two general types of processes that are commonly found in chemical industries, namely batch and continuous. Techniques and schemes for batch processes are illustrated through applications to a batch polymerisation reactor, whilst for continuous processes, to a continuous tubular polymerisation reactor.

5.2 MSPC-Based Applications to Batch Processes

The batch process considered was a pilot-scale batch polymerisation reactor for the production of polymer Methyl-Methacrylate (PMMA). The statistical projection techniques discussed in the previous chapter, were used to develop MSPC-based schemes to provide early warning of problems associated with product quality. Specifically, the first problem considered was the prediction of the final properties of the polymer product as early as possible in the batch, since in practise they are not known until the end of the operation. The second problem considered was the estimation of the initial conditions of the process, which can be typically related to a number of faults and malfunctions and which are of great importance for the successful operation of batch polymerisation reactors and for consistent polymer production (Kiparissides, 1995 and 1996). Finally, the application of the Inverse Projection to Latent Structures (IPLS) methodology to processes where minimal process data is available is presented. It is shown that, by applying the IPLS technique, the application of MSPC-based schemes can be extended to processes where there is minimal data for building a robust process representation.

5.2.1 Methyl-Methacrylate (MMA) Polymerisation Batch Reactor

The batch polymerisation reactor studied is a pilot-scale free-radical methyl-methacrylate (MMA) polymerisation reactor, developed and installed in the Laboratory of Polymer Reaction Engineering (LPRE), Department of Chemical Engineering, Aristotle University of Thessaloniki, Greece. The batch pilot-scale reactor is jacketed and provided with a stirrer for the thorough mixing of the reactants. Heating and cooling of the reaction mixture is achieved by circulating water at an appropriate temperature through the reactor jacket. The reactor temperature is controlled by a cascade control system consisting of a primary PID and two secondary PI controllers. The reactor temperature is fed back to the primary controller whose output is taken as the set-point of the two secondary controllers. The manipulated variables for the two secondary controllers are hot and cold water flow rates. The hot and cold water streams are mixed before entering the reactor jacket and provide heating or cooling for the reactor. The jacket output temperature is fed back to the secondary controllers. Figure 5.1 illustrates the pilot-scale batch reactor and its control system.

A detailed process simulation model, covering reaction kinetics, heat and mass balances, and automatic control, has been developed by LPRE and validated against the pilot-plant. Using this simulation, representative investigations of reactor operation and the effects of process malfunctions and faults can be studied. Process noise, typical of that found on the actual plant has been added to the on-line measurements (Kiparissides, 1996).

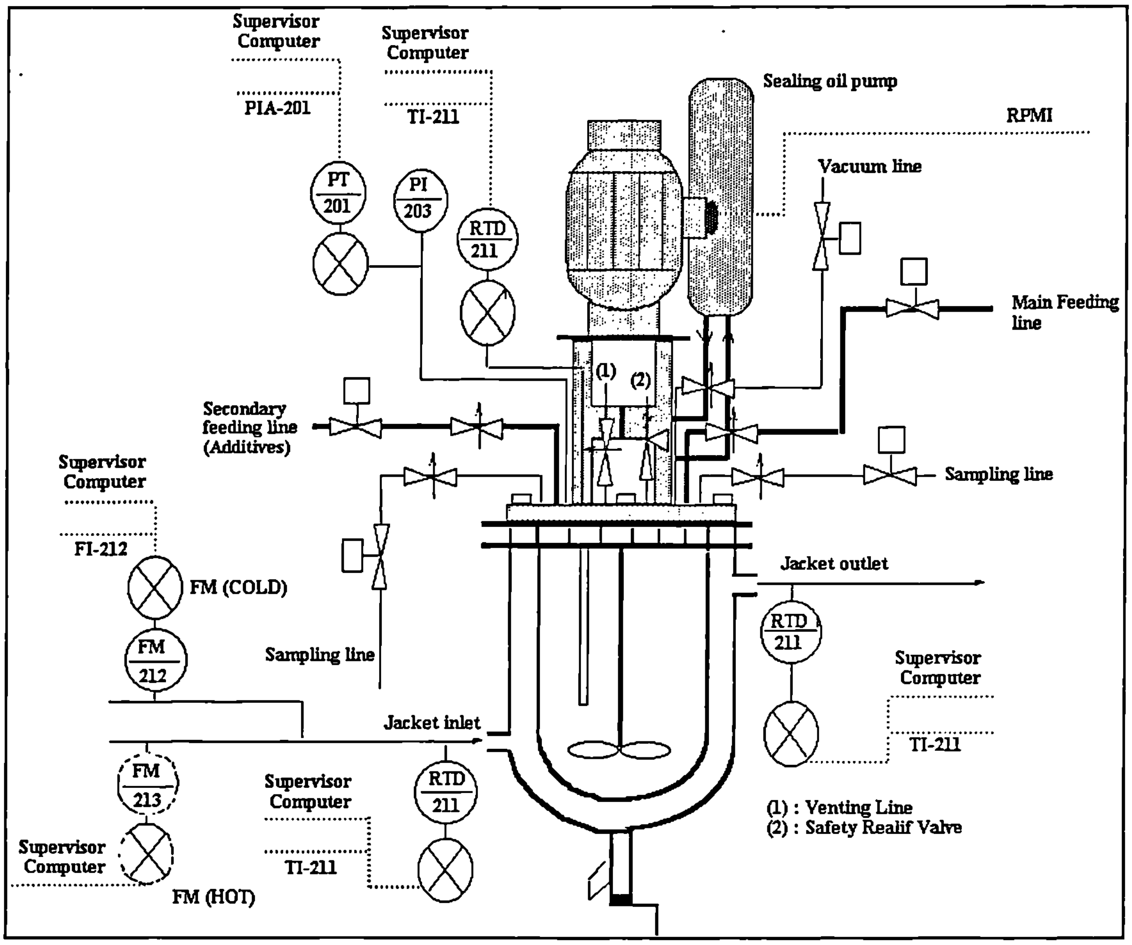


Figure 5.1 Flow diagram of the pilot-scale batch reactor (Kiparissides et al., 1997)

The initial process conditions of the polymer reactor studied here, include the reactor operating temperature (defined by the reactor temperature set-point, T_{sp}), the initial initiator weight (I_0) and the initial overall heat transfer coefficient (U_0). Other initial process parameters, such as environmental temperature and reaction mixture volume, are of less importance, since it has been found that, they do not affect the final polymer quality. The initial process conditions are listed in Table 5.1.

Initial Process Conditions	
T_{sp} (°K)	Reactor Temperature Set-Point
I_0 (gr)	Initial Initiator Weight
U_0 (Kcal/m^{2}min°K)	Initial Overall Heat Transfer Coefficient

Final Polymer Properties	
X_{MMA}	Final Conversion of MMA
M_N	Number Average M.W
M_W	Weight Average M.W

Initial Condition	Nominal Design Level	Level Variation
T_{sp}	345	±0.5
I_0	0.9 1.1 1.4	±0.1
U_0	0.05 0.08 0.10	±0.01

Table 5.1 Initial conditions, design levels and properties

The productivity variable of interest is the final conversion of monomer MMA (X_{MMA}). The molecular properties of interest are the weight average molecular weight (M_W) and number average molecular weight (M_N), Table 5.1. None of these properties are available on-line and are only measured infrequently, off-line, in the laboratory. During the polymerisation process, on-line measurements of conversion are available through the measurement of the density of the reaction mixture. Process measurements are collected on a one minute basis on the reactor temperature (T_r), on the inlet and outlet temperature of the coolant ($T_{c,in}$ and $T_{c,out}$, respectively), on the flow-rate of the coolant (F_j) and the conversion of monomer (Conv).

A number of polymer PMMA grades can be produced by an industrial reactor. The nominal experimental initial conditions were selected to represent realistic conditions of polymer PMMA production and are given in Table 5.1. These conditions correspond to the production of nine different grades of polymer product (Kiparissides, 1995). A set of seven batch simulations, for each polymer grade was generated through Monte Carlo variation of the selected initial conditions corresponding to a particular grade (Table 5.1). As a result, nine sets (e.g. 3^2 factorial design) of seven batch simulations were obtained, i.e. a total of 63 simulated batches were generated. Each set represents normal process operation when only common cause process variations are present and when only acceptable product quality was achieved. Five batch simulations from each grade were included in the training set (i.e. historical process database) from which the nominal statistical models were built. The remaining two batches formed the data sets upon which the models were validated. The training data set comprises process measurements from 45 batch simulations, whilst the validation set comprises measurements from 18 batch simulations. The trajectories of some of the process variables for a typical batch, that is included in the training set, are presented in Figures 5.2 and 5.3. From this point and on, a batch simulation will be called "batch" and the data produced by the batch reactor simulation program will be considered as "real" or "actual" data.

5.2.2 Prediction of Final Polymer Properties

Product quality is very important in polymerisation processes, since it affects the behaviour of the product in its final applications. Quality control in batch processes presents a challenging problem, since final product quality is not known until the end

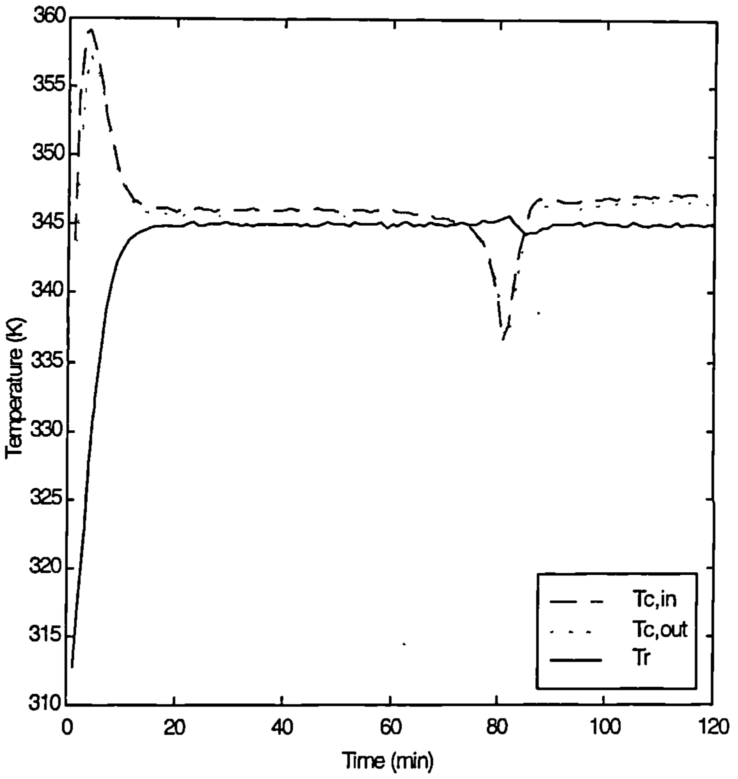


Figure 5.2 Temperatures around the batch reactor from a typical operation

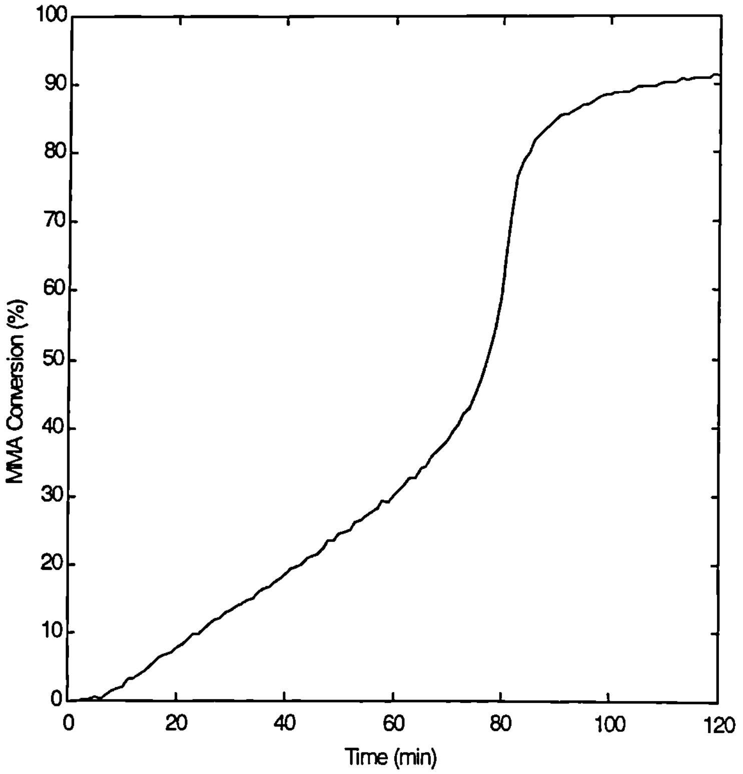


Figure 5.3 Conversion of MMA from a typical operation

of the batch. Furthermore, due to the lack of on-line instrumentation, quality variables are unmeasurable or only measured infrequently in the laboratory. In an attempt to overcome these difficulties, software sensors based upon statistical methods and neural networks, can be developed to infer the final quality from the available process data (Kiparissides and Morris, 1996). An empirical model based on the Projection to Latent Structures (PLS) technique, can be used to infer the polymer quality of PMMA using the initial process conditions of the batch (Papazoglou, et al., 1998). The reason for performing this study is to investigate whether an empirical model can provide reliable predictions of the final properties.

Model Development and Validation: The final polymer properties of interest for the batch MMA polymerisation reactor are final conversion of monomer MMA (X_{MMA}), number average molecular weight (M_N) and weight average molecular weight (M_W). These are captured in the Y data matrix of the PLS model, whilst the initial conditions of the polymer process form the X data matrix. Therefore, the initial process conditions need to be measured or estimated. These are the reactor temperature set-point (T_{sp}), the initial initiator weight (I_0) and the initial overall heat transfer coefficient (U_0). A linear PLS model of the structure shown in Figure 5.4, can then be developed based upon X and Y data sets. The initial process conditions are supposed to be measured. The model was assessed through cross-validation procedures which showed that all three latent variables should be retained. Table 5.2 summarises the amount of variability explained by the PLS model in each latent variable block and for each predicted variable. It can be seen that, the first latent variable describes the largest amount of variability in all the quality variables.

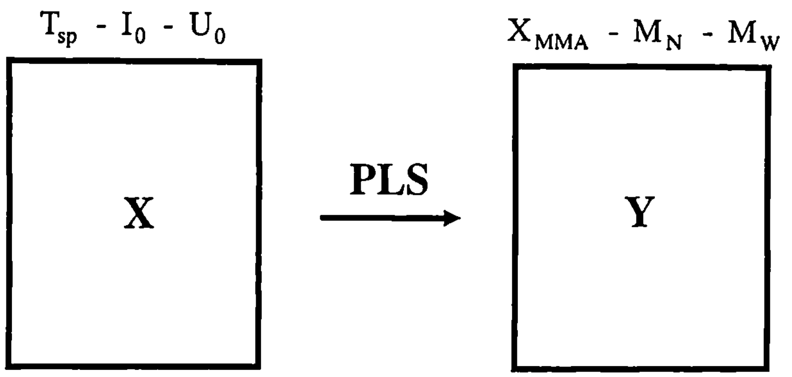


Figure 5.4 PLS model to infer final polymer properties

LV	% Variability Explained						
	X Block		Y Block		Quality Variables		
		Cumulative		Cumulative	X_{MMA}	M_N	M_W
1	33.97	33.97	89.21	89.21	60.20	87.34	91.38
2	32.83	66.80	10.66	99.87	99.77	97.57	97.83
3	33.20	100.00	0.05	99.92	~100.00	~100.00	~100.00

Table 5.2 Initial conditions, design levels and properties

Figure 5.5 shows the prediction of the final conversion, number and weight average molecular weights for the 45 batches included in the training set. The next aspect investigated was the ability of the model to provide satisfactory predictions of the final polymer properties for the batches included in the validation set. These eighteen, previously "unseen", batches of the validation set were drawn from the same population and their predictions are shown in Figure 5.6. These predictions are also quite satisfactory. It can be concluded that, PLS is able to model the strong relationship between the initial process conditions and the final properties of the polymer product.

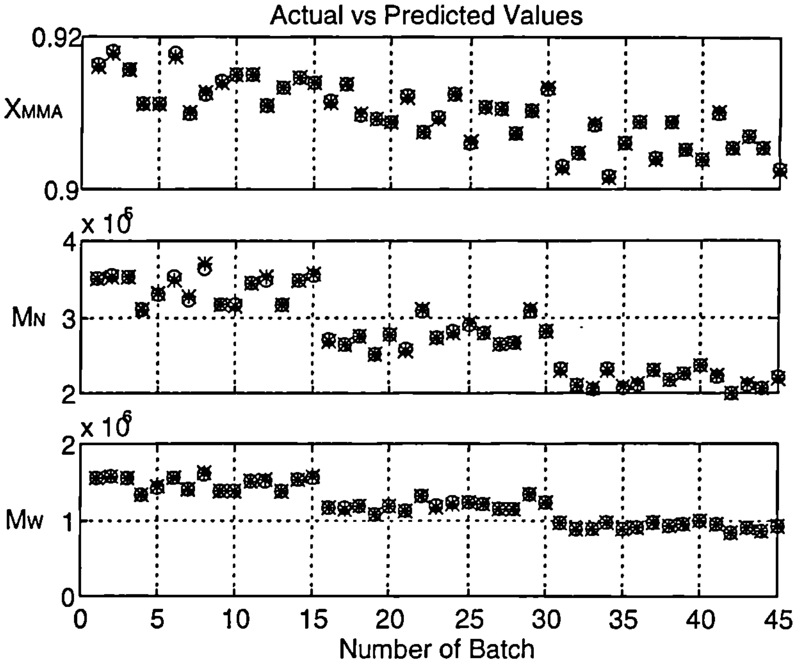


Figure 5.5 Predictions of final properties for the training set

(o: actual, *: predicted)

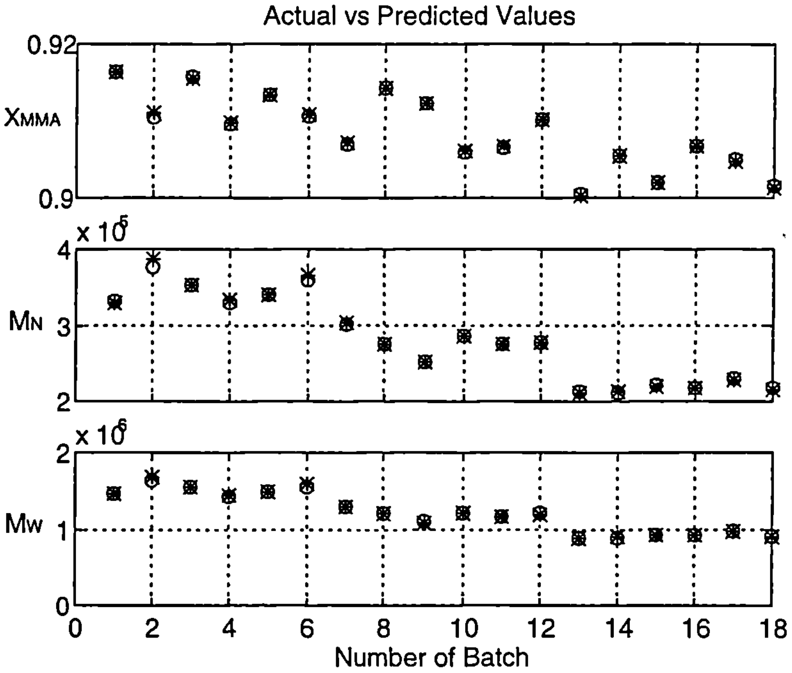


Figure 5.6 Predictions of final properties for the validation set

(o: actual, *: predicted)

A problem associated with the above procedure is related to fact that the PLS model has as its inputs the initial process conditions. The more accurate the initial conditions are, the more reliable the final properties predictions. However, in most situations, the initial conditions of the batch are not exactly known (Kiparissides, 1996), since some of them cannot be accurately measured or were unrecorded. Thus, estimates of them will be required. In the next section, a procedure to estimate the initial conditions using a statistical inferential model is presented.

5.2.3 Estimation of Initial Process Conditions

Initial process conditions in batch polymerisation reactors are known to influence the final properties of the polymer product. For the batch reactor of interest, the initial conditions (e.g. initiator concentration and overall heat transfer coefficient) are related to two commonly occurring problems, *reactive impurities* and *reactor fouling*. Both problems affect the polymerisation process and product quality. The presence of impurities is equivalent to a reduction in the initiator efficiency, whilst reactor fouling reduces the heat transfer capabilities of the reactor and, as a result, the reactor temperature control system becomes less effective. The detection and estimation of reactive impurities and reactor fouling are, therefore, of profound importance. Furthermore, detection should take place at an early stage of the polymerisation process, in order to allow for any possible corrective actions that will ensure the normal operation. Reactive impurities can be simulated by a decrease in the initial initiator weight and reactor fouling by a decrease in the initial overall heat transfer coefficient of the reactor. The amount of reactive impurities and the extent of reactor fouling can be determined by estimating the initial initiator weight and the initial overall heat transfer coefficient. In the next section, a linear multivariate statistical model is developed using Multi-Way Projection to Latent Structures (MPLS) to predict the initial process conditions.

Model Development: The analysis of the 45 batches of the training data set was performed using Projection to Latent Structures (PLS) and its multiway extension, MPLS. The predictor (X) data set contains the on-line process measurements, whilst the response (Y) data set comprises the initial process conditions, as presented in Table 5.1. The reactor temperature set-point (T_{sp}) has been included, since, as it will be shown in a subsequent section, it was found to improve the model predictions (see Figure 5.20). Nominal batch operation is usually achieved in two hours. However, since the objective is to estimate the initial process conditions at an early stage of polymerisation, only the part of the database covering the first sixty minutes of each batch was used in the analysis.

A number of other issues need to be addressed, including the identification of the sample time points in the batch which encapsulate sufficient information to enable a satisfactory model to be built. Specifically, this includes the selection of the time point at which data sampling starts, the selection of the sampling time interval and the identification of the minimum number of on-line process measurements (samples) to be included in the model. From a number of previous studies of the reactor data and its information content, the sampling time intervals, considered for model development, were selected as 1, 5 and 10 minutes. Linear PLS models were developed where on-line process samples at only one time point (k_1) formed the basis of the X block. For situations where on-line process samples at more than one time point (k_1 to k_2) were used to construct the X block, Multi-Way PLS models were developed, Figure 5.7.

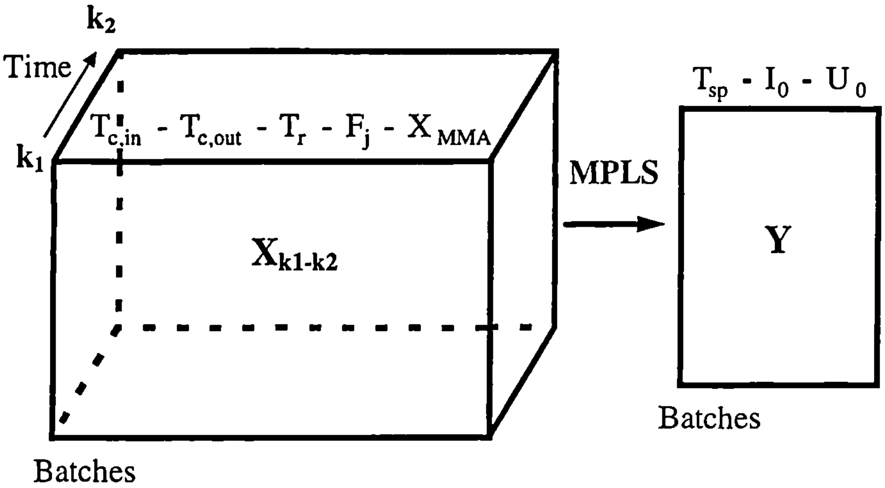


Figure 5.7 Structure of Multi-Way PLS Models

Concerning the construction of the Y block, two alternative scenarios were considered (Kiparissides, 1996). The first represents the ideal case, where the initial process conditions are accurately measured (called actual values). The second, and more realistic case, is where accurate initial process conditions are not available. In order to represent this later situation, the Y-block is in-filled with the initial condition values which correspond to that particular polymer grade being modelled (called theoretical values). For example consider a particular batch whose product can be classified as belonging to a polymer grade with initial conditions as defined by the set of theoretical values Set 1, Table 5.3. A set of seven batch simulations, for each polymer grade, was generated through Monte Carlo variation of the selected initial conditions corresponding to a particular grade (section 5.2.1). For the seven batch simulations belonging to the polymer grade of Set 1, the actual initial conditions are defined by the sets MC1-MC7, Table 5.3, whilst the theoretical initial conditions are

	T_{sp}	I_0	U_0
Set 1	345	1.1	0.100
MC1	345.04	1.078	0.094
MC2	345.13	1.132	0.104
MC3	344.78	1.163	0.097
MC4	344.81	1.002	0.092
MC5	345.27	1.044	0.091
MC6	344.66	1.145	0.096
MC7	344.98	1.091	0.102

Table 5.3 Actual (MC1-MC7)and Theoretical (Set 1) values

those of Set 1. In the analysis, it was assumed that actual plant initial conditions (e.g. MC1-MC7) were not available, thus, the initial conditions (Y data set) were defined by Set 1 (i.e. theoretical values).

Model Selection: On-line process measurements (samples) over the first sixty minutes of each batch are included in a training set, which forms the basis of the X data sets. The Y data set comprises the theoretical values for the initial process conditions of the corresponding batches included in the X data sets. It is now necessary to locate the most appropriate time point to start collecting measurements on the process variables and to determine the appropriate minimum number of samples required to develop a realistic model of the initial conditions. A comprehensive set of models was then developed, spanning a wide range of different operating scenarios. In each operating scenario, a *starting time of sampling* (k_1) is selected and samples are collected with a given sampling interval (i.e. 1,5,10 minutes) up to a *sampling end time* (k_2). For scenarios that include process models built upon more than one samples ($k_1 \neq k_2$), the MPLS technique was used (Figure 5.7). When the process model was built upon one sample ($k_1 = k_2$), the PLS technique was used. For each of the PLS and MPLS models, the number of latent variables required was determined through cross-validation. In order to select the most appropriate starting time point for sampling and the number of process samples required to estimate the initial conditions, the Sum of Square Errors of Calibration (SSEC), which is a measure of the fit of the model to the calibration (training) data, was employed :

$$\mathrm{SSEC} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{i,j} - \hat{y}_{i,j}\right)^2 \qquad (5.1)$$

where n is the number of batches included in the training set (45), m is the number of initial process conditions, $y_{i,j}$ and $\hat{y}_{i,j}$ are the actual and estimated value of the j-th initial condition of the i-th batch, respectively. The better the model, the lower the SSEC value.

Figures 5.8 - 5.11 illustrate the effect of altering the starting time of sampling (k_1) and the ending time of sampling (k_2) on the SSEC for fixed sampling intervals (e.g. 1, 5 and 10 minutes). It can be seen that, models built from samples that include the first ten minutes of batch operation, exhibit a high value of SSEC (Figures 5.8 and 5.9). By inspecting the process temperature trajectories during the first few minutes of polymerisation (Figure 5.2), the trajectories are seen to exhibit a highly non-linear behaviour, since this time period is the heat-up stage. Models covering this stage cannot estimate the initial process conditions, since no reaction has actually taken

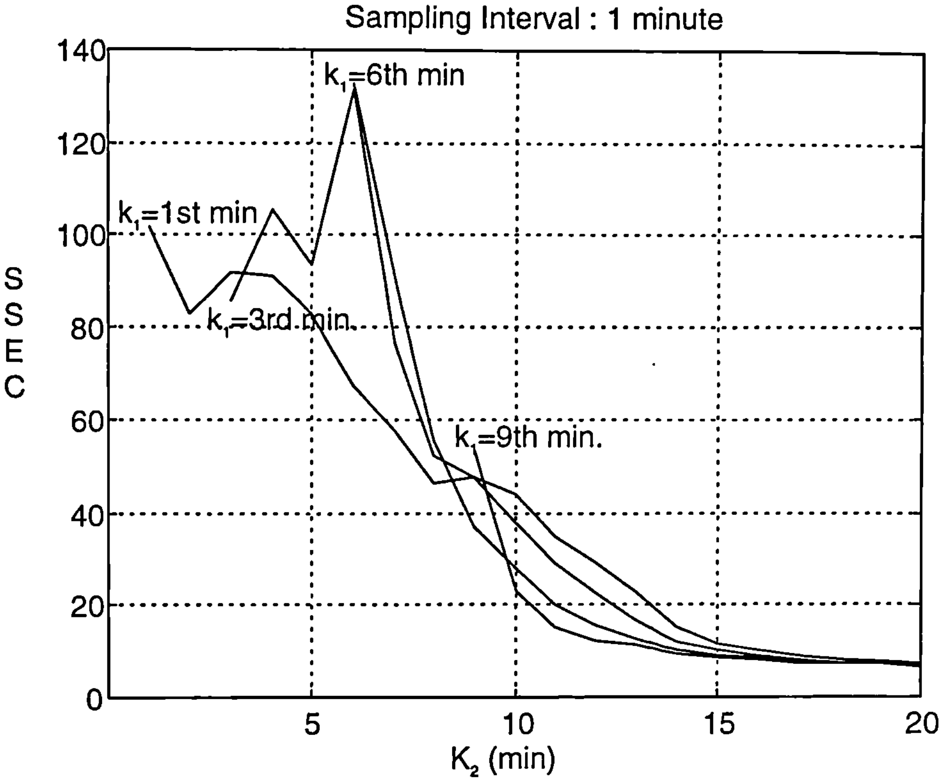


Figure 5.8 Effect of starting and ending time of sampling (interval 1 min)

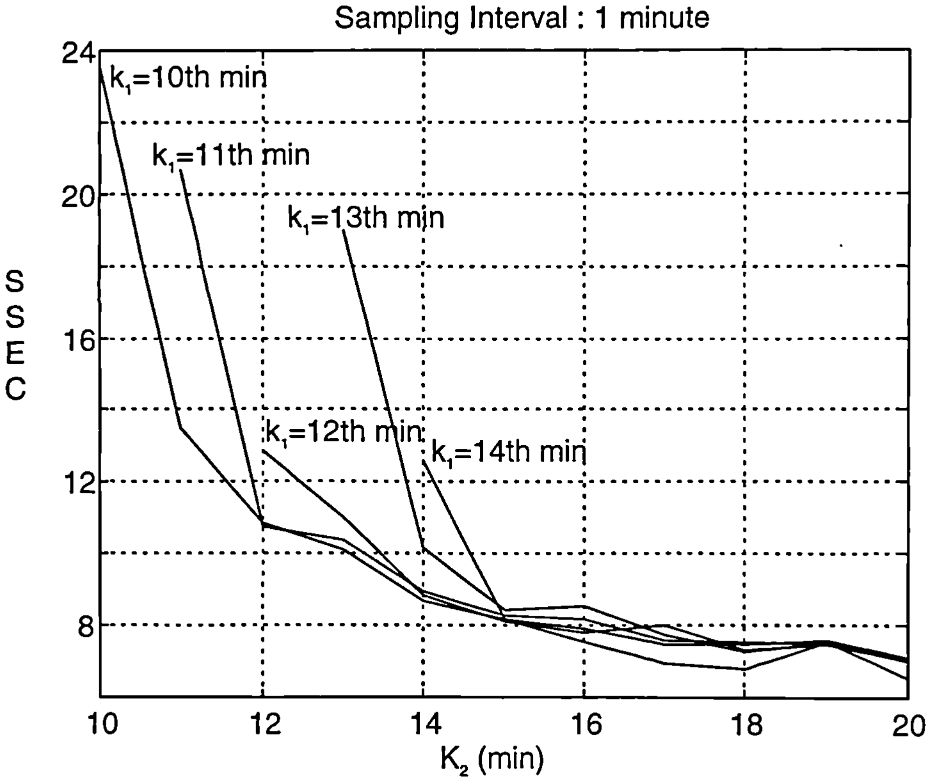


Figure 5.9 Effect of starting and ending time of sampling (interval 1 min)

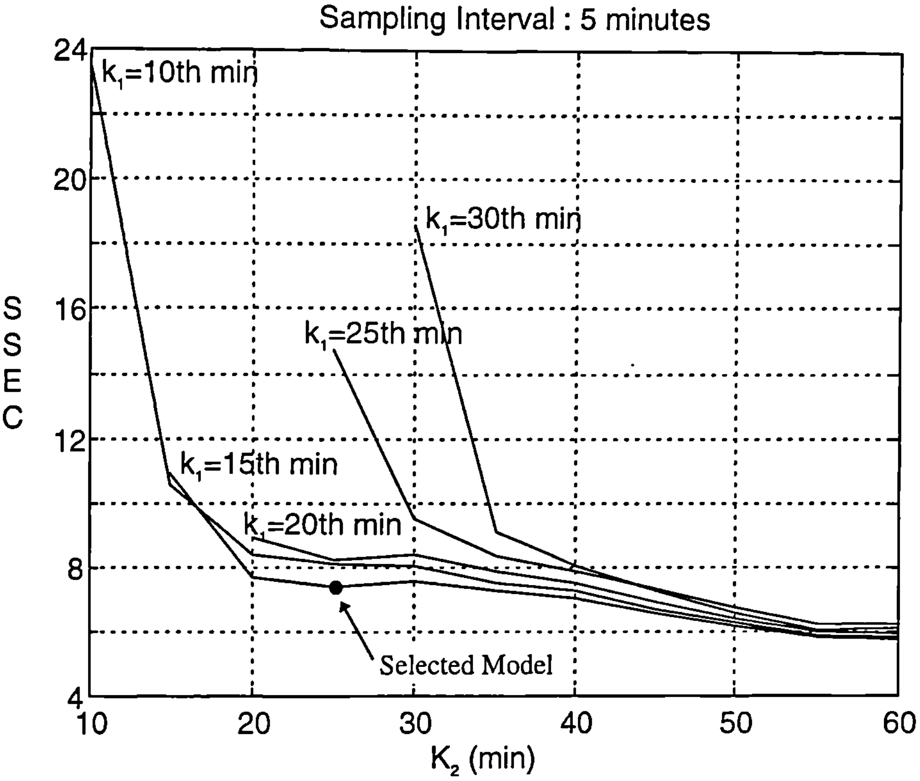


Figure 5.10 Effect of starting and ending time of sampling (interval 5 mins)

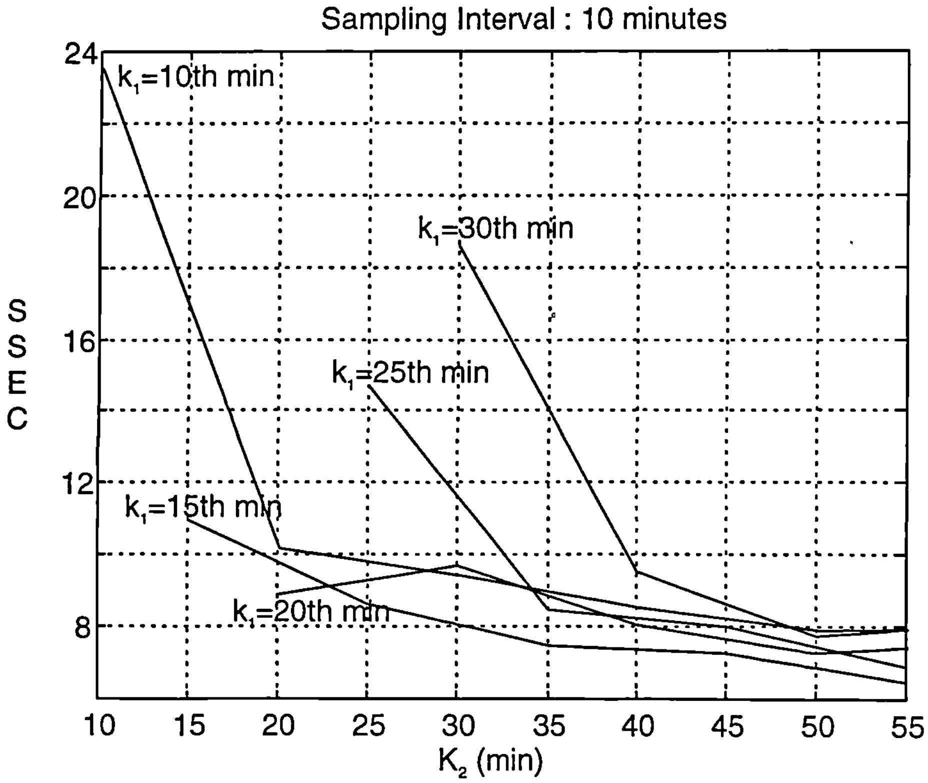


Figure 5.11 Effect of starting and ending time of sampling (interval 10 mins)

place at this time. As a result, the first nine minutes of batch operation have to be discarded, since they contained very little information on the process. It can be concluded that, starting at later time points, the information content increases and, as a result, the predictive capabilities of the model are improved (i.e. the value of SSE decreases). Figures 5.12-5.15 show the effect of including additional samples to the model for a specific value of the starting point of sampling (k_l). It can be concluded that, as the number of samples increases, the information content increases and, as a result, the predictive capabilities of the model are improved. By recalling that the overall objective is to estimate the initial process conditions, at an early stage of the polymerisation, it was concluded that the optimal scenario was to build a MPLS model from on on-line samples collected at the 15th, 20th and 25th minute (Figure 5.10).

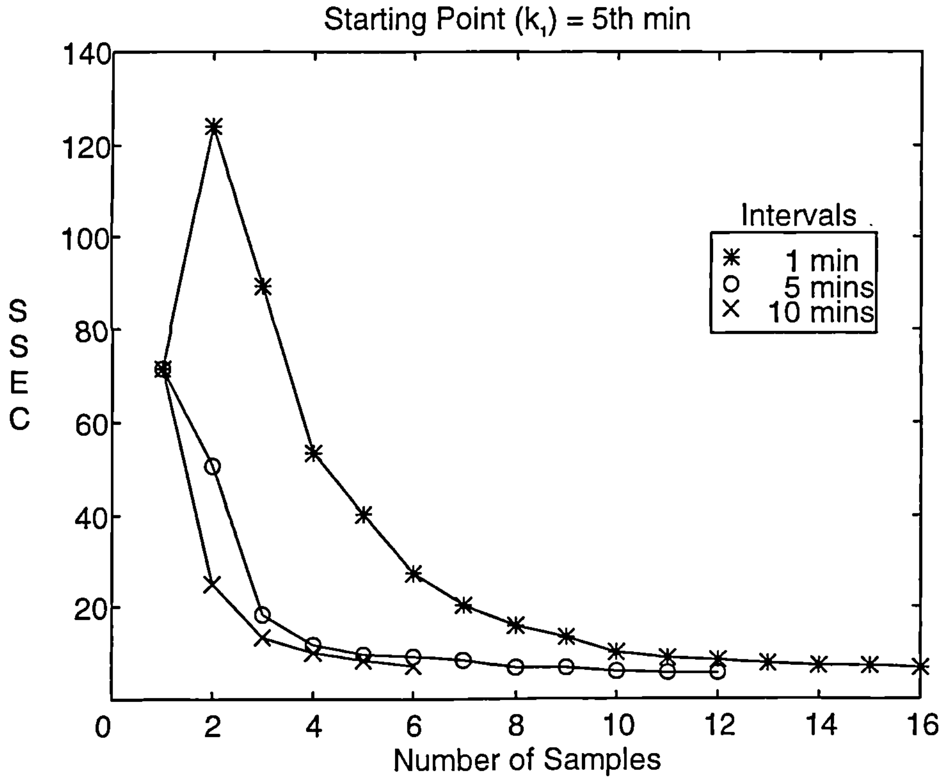


Figure 5.12 Effect of number of samples (starting point 5th min)

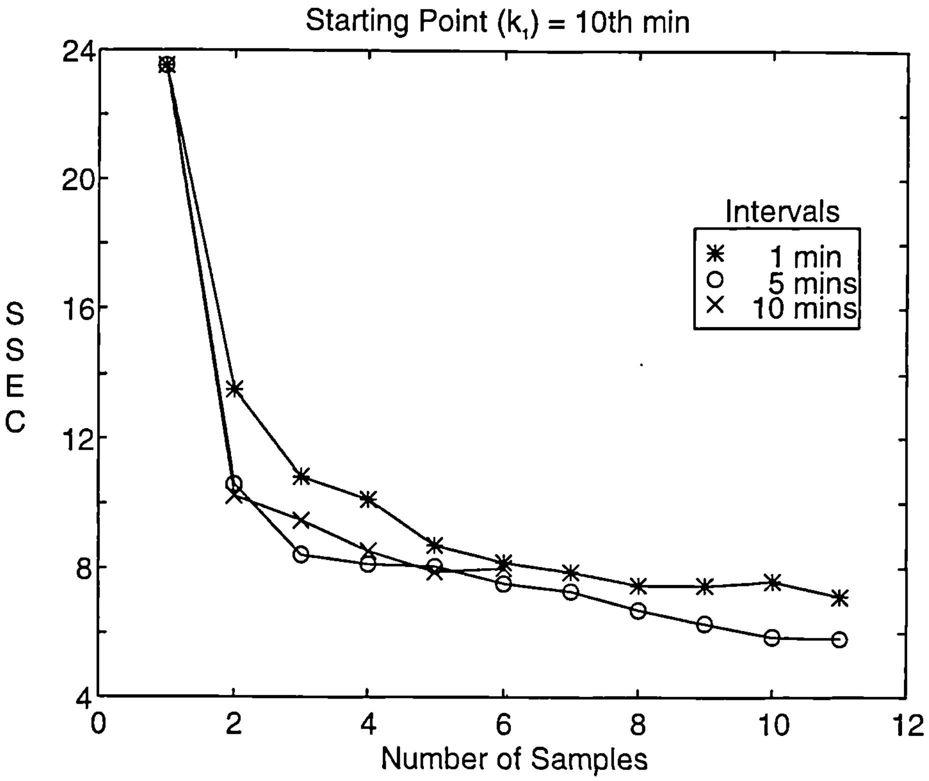


Figure 5.13 Effect of number of samples (starting point 10th min)

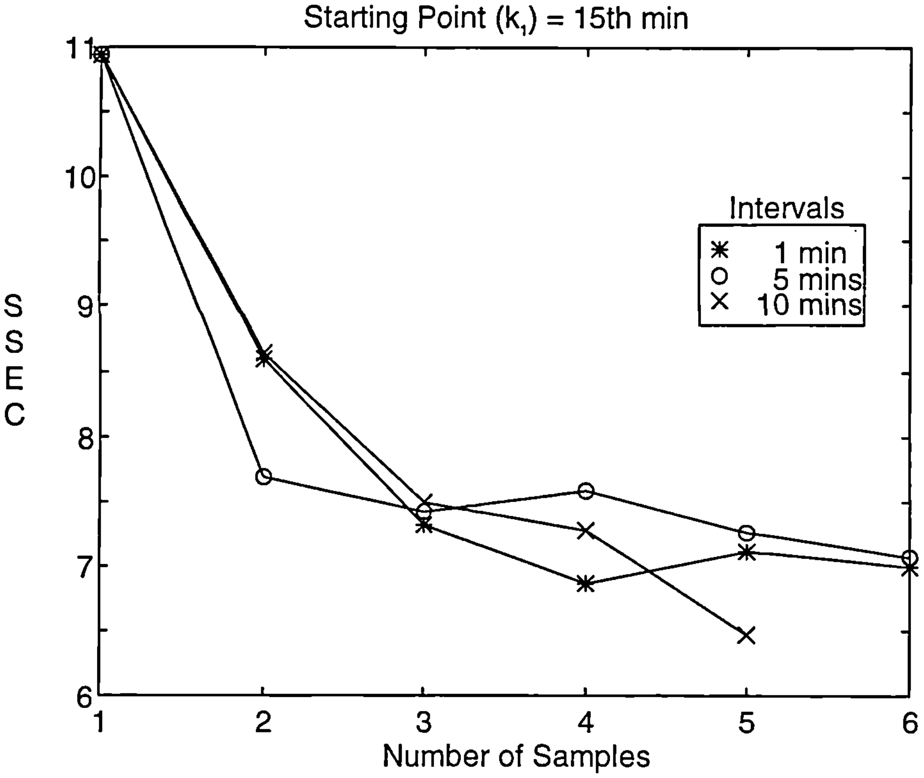


Figure 5.14 Effect of number of samples (starting point 15th min)

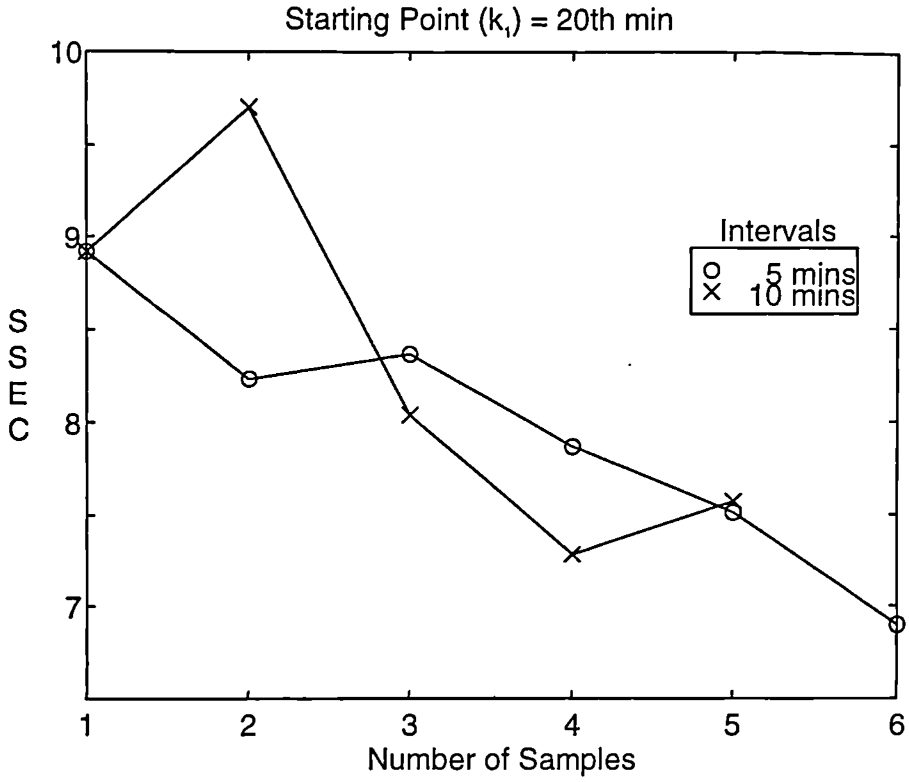


Figure 5.15 Effect of number of samples (starting point 20th min)

An MPLS analysis was carried out on the selected part of the process data (45 batches) of the training set. Six latent variables were selected to be included in the model using cross-validation. Table 5.4, summarises the variability explained in each block by the MPLS model. The projection of the batches included in the training set onto the reduced space of the first two latent variables does not exhibit any usual behaviour, since all process scores are lying inside the 95% confidence ellipsoid (Figure 5.16). Figure 5.17 illustrates the plot of the process scores ($\mathbf{t}_1$) versus the quality scores ($\mathbf{u}_1$) for the first latent variable, where it can be seen that, the assumption of a linear relationship between the predictor (X) and the response (Y) data sets is valid.

LV	X % Variability Explained		Y % Variability Explained	
		Cumulative		Cumulative
1	51.63	51.63	32.06	32.06
2	22.49	74.12	30.49	62.55
3	7.88	82.00	28.82	91.37
4	5.16	87.16	3.35	94.72
5	7.35	94.51	0.35	95.06
6	4.26	98.76	0.13	95.19

Table 5.4 Explained Variability by the MPLS model

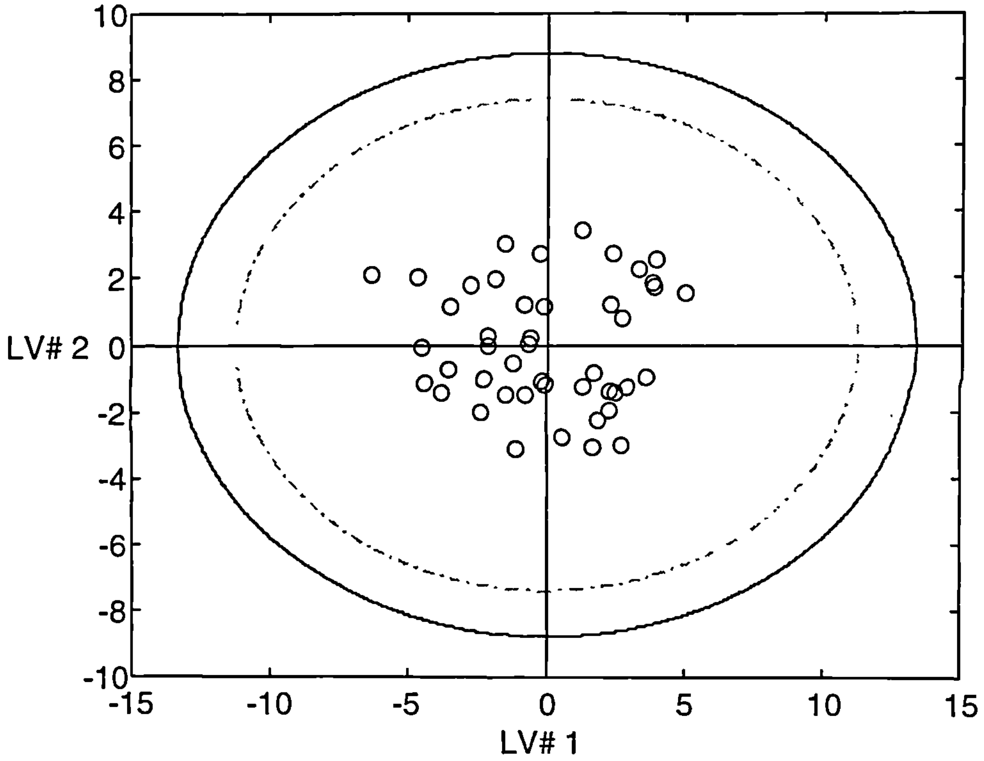


Figure 5.16 Process scores for the first versus the second latent dimensions

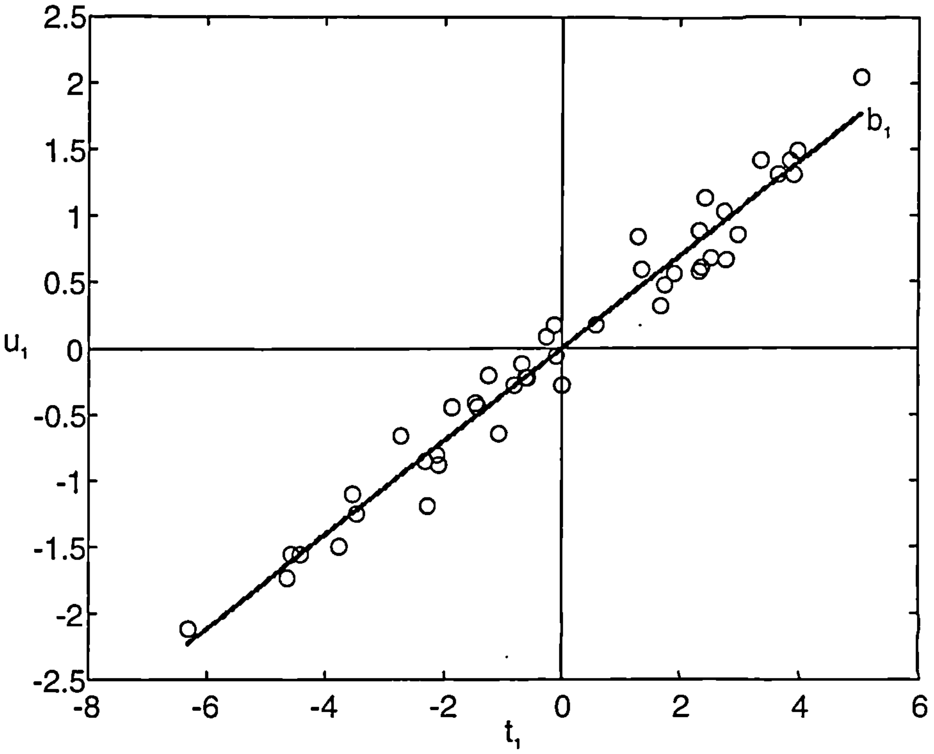


Figure 5.17 Process versus quality scores for the first latent dimension

Figures 5.18 and 5.19 show the estimated values of the initial initiator weight obtained from the PLS model and the initial heat transfer coefficient, respectively, for the 45 batches included in the training set, along with their theoretical and actual values. Similarly, Figure 5.20 present the estimated values for previously 'unseen', 18 batches of the validation set, along with their theoretical and actual values. It can be seen that, although the model has been trained with theoretical initial conditions, it is capable of providing satisfactory estimates of the 'unseen' initial conditions, very close to their actual values.

Other Issues. The data used to illustrate the estimation of initial process conditions was not pre-processed; it was used in its raw form and also the temperature set-point (T_{sp}) was included in the response (Y) data set, since it was found that it improves the predictive capabilities of the models. This is reasonable, since temperatures

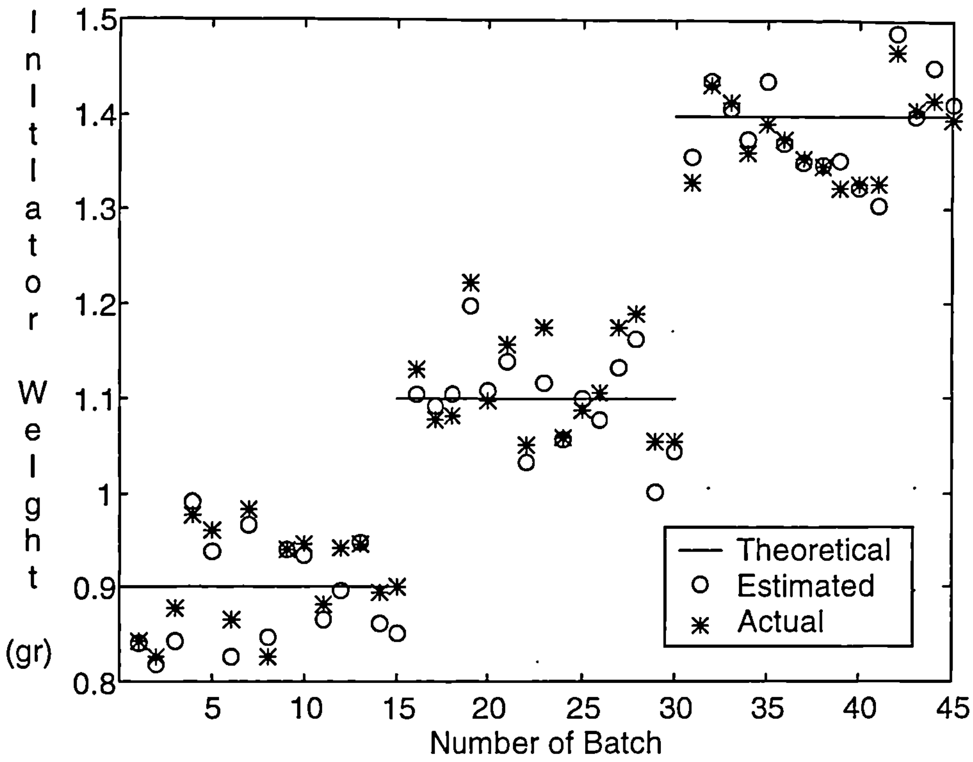


Figure 5.18 Estimated values of initial initiator weight for the training set

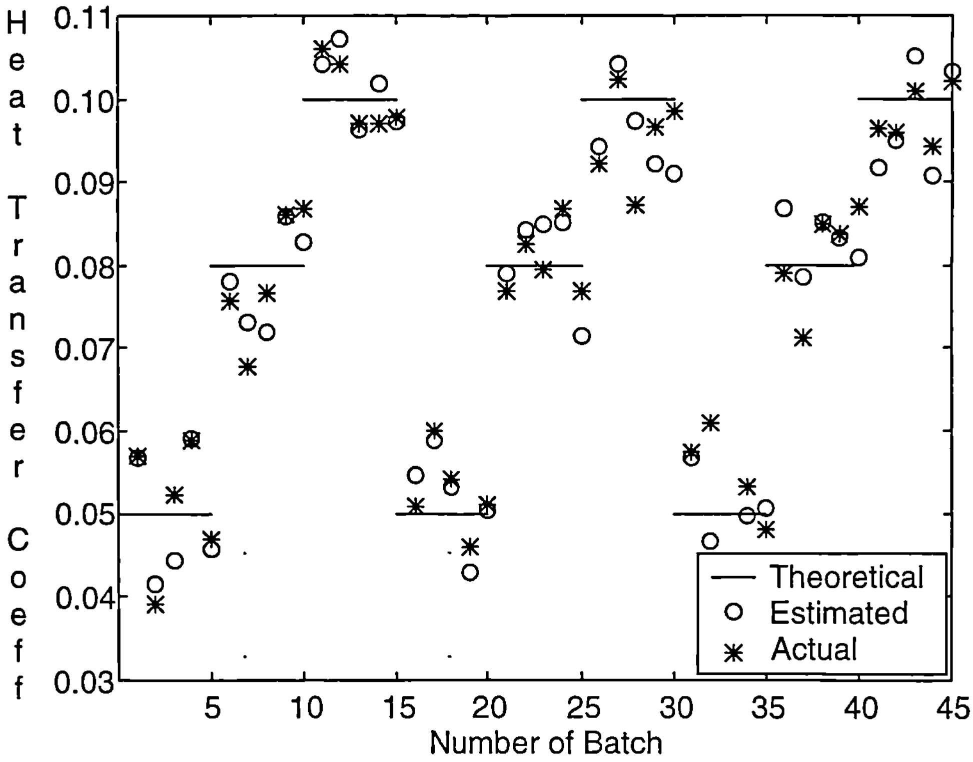


Figure 5.19 Estimated values of initial heat transfer coefficient for the training set

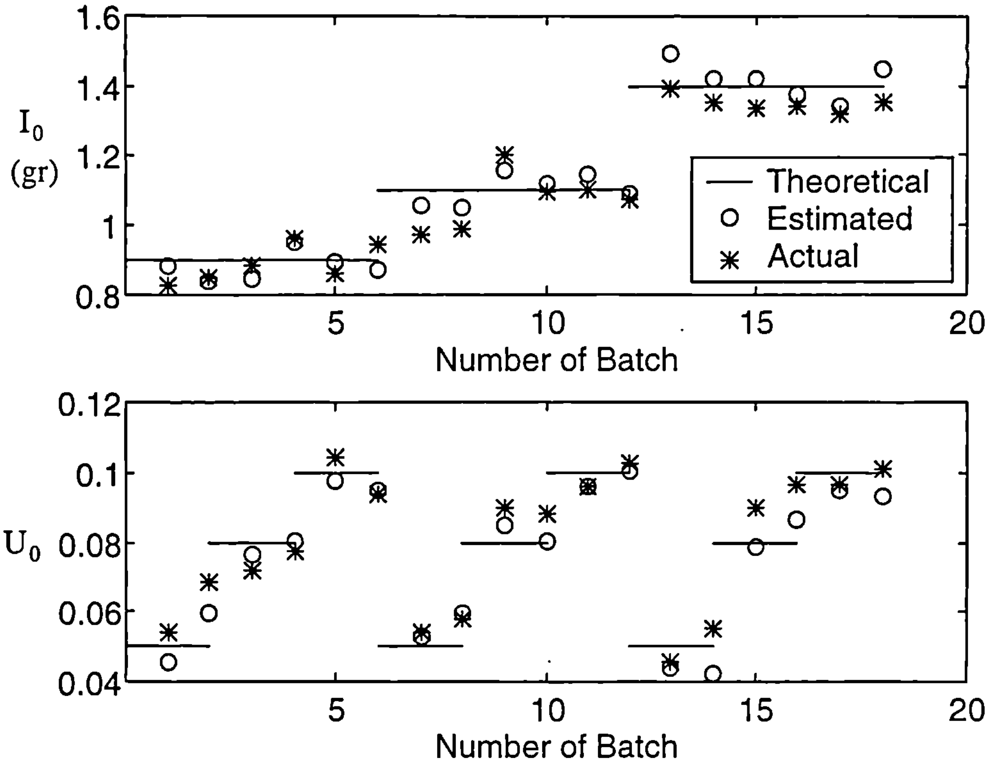


Figure 5.20 Estimated values of the initial conditions of the validation set

included in the predictor (X) data set depend upon the operating temperature and, therefore, are correlated. The performance of the model for a study of four cases presented in Table 5.5, in terms of the ratio of the Mean Square Residual (MSR) to the Mean of Square Error (MSE) for the two initial conditions of interest, is illustrated in Figure 5.21.

Noise - , T_{sp}+	Process Data filtered, Temperature set-point included
Noise + , T_{sp}+	Process Data *not* filtered, Temperature set-point included
Noise - , T_{sp} -	Process Data filtered, Temperature set-point *not* included
Noise + , T_{sp} -	Process Data *not* filtered, Temperature set-point *not* included

Table 5.5 Notation used in the study of the effect of filtering and temperature set-point.

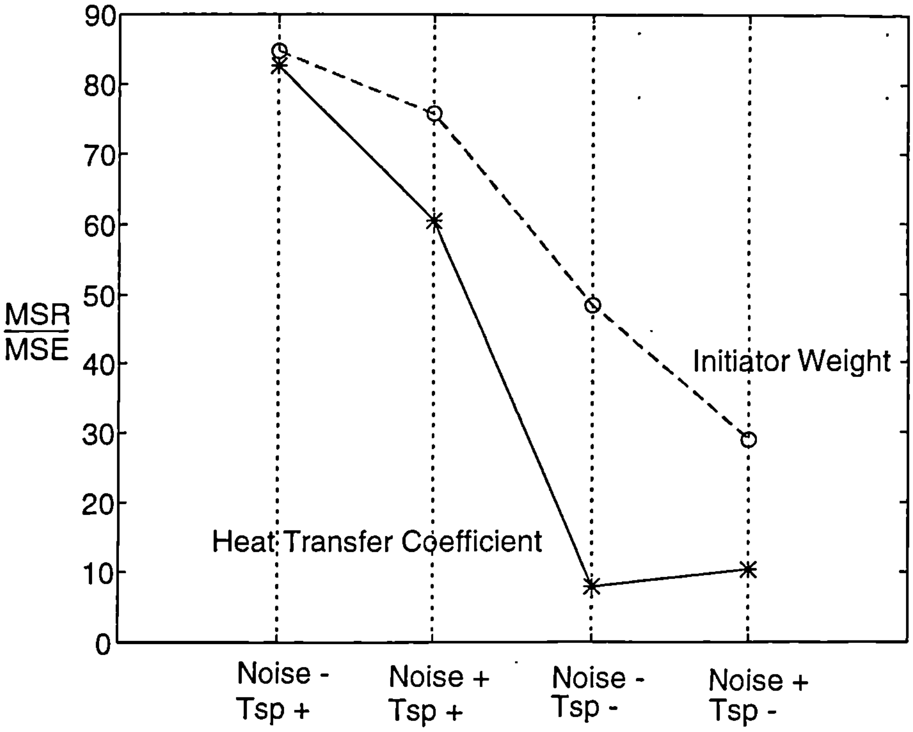


Figure 5.21 Effect of filtering and inclusion of the temperature set-point

The ratio MSR to MSE describes how well the process data account for the variability in each of the response variables (Nomikos and MacGregor, 1994b). It is distributed as an F variate with R and (n-R-1) degrees of freedom and it is defined as follows :

$$\left(\frac{\text{MSR}}{\text{MSE}}\right)_j = \frac{(n-R-1)\sum_{i=1}^{n}\hat{y}_{i,j}^2}{R\sum_{i=1}^{n}\left(y_{i,j}-\hat{y}_{i,j}\right)^2} \tag{5.2}$$

where R is the number of latent variables retained in the model, n is the number of batch included in the training set (45) and $y_{i,j}$ and $\hat{y}_{i,j}$ are the actual and estimated value of the j-th initial condition of the i-th batch, respectively. The better the predictor (X) set accounts for the response set (Y), the higher the value of the MSR/MSE ratio. It can be seen that, filtering of process data improves the predictive capability of the model, i.e. the MSR/MSE value increases. The fact that the

MSR/MSE value for the heat transfer coefficient decreases with filtering, when the temperature is not included in the model, can be explained by chance, since the amount of data upon which the model was built is very small. Furthermore, it can be concluded that, the inclusion of temperature set-point is necessary, since the predictive capability of the model is significantly improved when the temperature set-point is included in the Y block.

Another important issue that was considered, is the number of batches that should be included in the training set in order to develop a statistical model to infer the initial process conditions. The original training set comprises 45 batches (i.e. nine sets of five batches). Three additional training sets were created by extracting selected batches from the original training set. Specifically, they included 36, 27 and 18 batches (i.e. nine sets of four, three and two operations, respectively) were created. Figures 5.22 shows the effect of the number of batches included in the training data

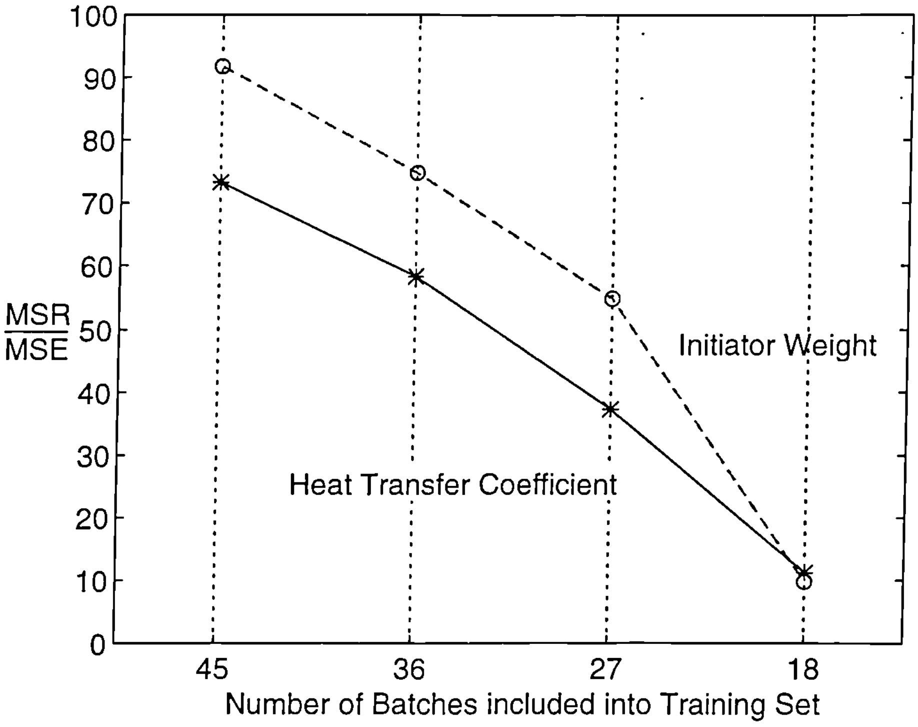


Figure 5.22 Effect of number of batches

set on the predictive ability of the model, in terms of MSR/MSE ratio for the actual initial conditions. It can be concluded that, as the number of batches included increases, the information content increases and, as it was expected, the predictive capability of the model improves, since PLS is a data-oriented technique.

5.2.4 Inferring The Quality of A Polymer Product Using Statistical Models

The final product quality of a batch polymerisation reactor can be predicted before or at an early stage of the operation by using a PLS model (section 5.2.2), which infers the final polymer properties from the initial process conditions. However, the initial conditions of a batch are not always known precisely and, furthermore, sometimes they are not even available. A solution to this problem is to use estimates of the initial conditions as model inputs. The Multi-Way PLS model, developed in section 5.2.3, was shown to provide reliable estimates of the initial conditions using only a few on-line measurements and, therefore, it can be used in conjunction with the PLS model to infer the final product quality. However, several other approaches that utilise statistical models to infer the quality of the polymer product can be used.

Consider the situation where the final polymer properties have to be predicted at an early stage of the polymerisation process. The entire database comprising on-line process measurements collected on a five minutes basis (X data set) and the final polymer properties (Y data set) using a Multi-Way PLS model could then be modelled (Figure 5.23). Note that, the process data set (X) comprises the on-line measurements of conversion of MMA, while the response (Y) data set comprises the

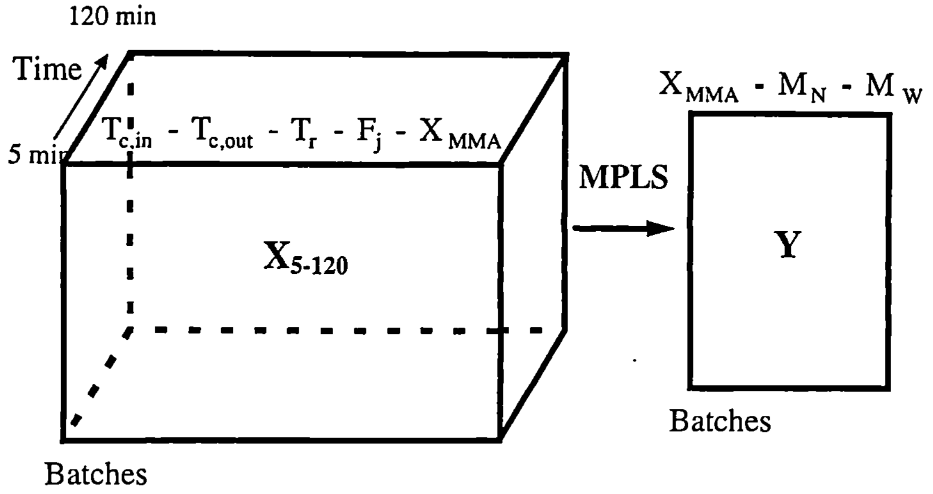
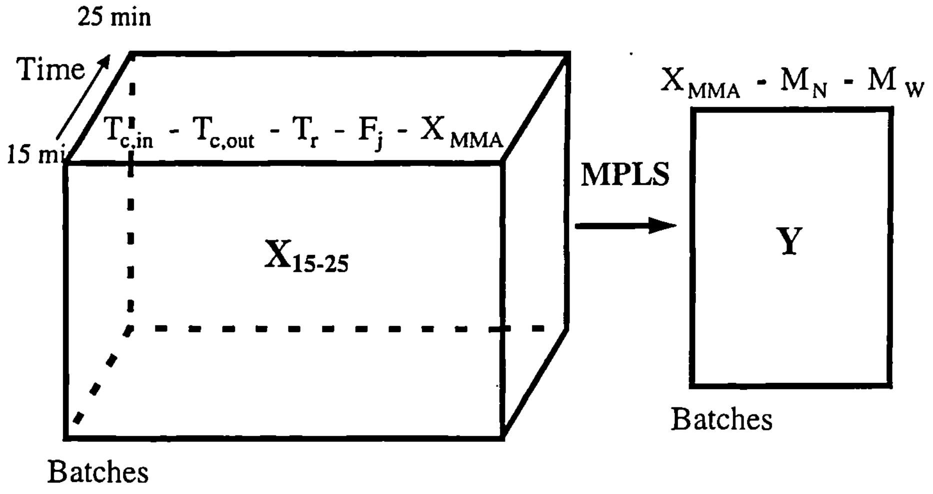

120 min
Time
5 min
$T_{c,in}$ - $T_{c,out}$ - T_r - F_j - X_{MMA}
$X_{5\text{-}120}$
Batches
MPLS
X_{MMA} - M_N - M_W
Y
Batches

Figure 5.23 MPLS model built upon the entire process database

final value MMA conversion, measured in the laboratory. This approach provides a model with excellent predictions, since it utilises all the on-line process data. However, it has some serious disadvantages. It is time consuming and it can only be applied at the end of the batch and, consequently, it is not practically applicable. Hence, it is presented as a reference point of a Multi-Way model's ultimate predictive ability.

A similar approach is that of Nomikos and MacGregor, (1994b). A Multi-Way PLS model for on-line monitoring, fault detection and diagnosis of the batch process, is based upon the entire process database. Although for each new batch all measurements are not available, at each time point, any of the four methods described in section 4.5.1.1, to in-fill the unknown process measurements can be used and predictions of the final polymer properties can be obtained. However, predictions of the polymer properties obtained at the 25th minute using any of these methods, are not satisfactory, since the unknown part of the vector of the process measurements (30th - 120th minute) is relatively large.

A more promising route is to build a MPLS from the on-line process samples collected at the same time points used in the estimation of the initial process conditions (section 5.3.2), i.e. the 15th, 20th and 25th minute (Figure 5.24). This selection, however, is arbitrary.



25 min
Time
15 min
$T_{c,in}$ - $T_{c,out}$ - T_r - F_j - X_{MMA}
$X_{15\text{-}25}$
Batches
MPLS
X_{MMA} - M_N - M_W
Y
Batches

Figure 5.24 MPLS model based on the 15th, 20th and 25th minute

A more appropriate selection can be made using a similar approach to that applied in section 5.3.3. Figure 5.25 shows the effect of altering the starting time of sampling (k_1) and the ending time of sampling (k_2) on the SSEC at a fixed *sampling interval of* 5 minutes. It can be seen that, (i) collecting samples at later time points, fewer samples are required to build a satisfactory model; (ii) as the information content increases, the predictive capabilities of the model improve; (iii) the model built upon the pre-selected process samples (15th, 20th and 25th minute) performs quite well. In order to improve the predictions obtained, the block of the initial process conditions can also be included, not as an interface between the process measurements (X data set) and the final polymer properties (Y data set), but as a block in a parallel branch. In this way, a Multi-Way - Multi-Block PLS model can be formed (Kourti et al., 1995). For the polymer process studied, however, structural limitations restrict the

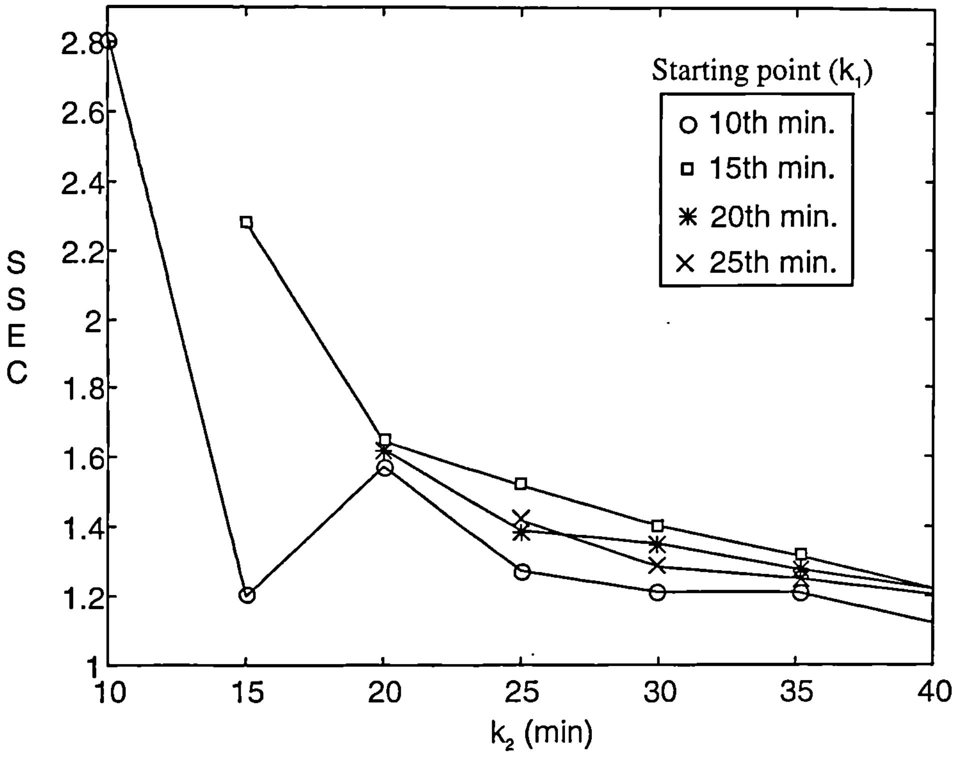


Figure 5.25 Effect of starting and ending time point of sampling

predictive ability of the model. Specifically, the small number of initial process conditions restricts the number of latent variables that can be retained in the model and, as a result, poor predictions obtained. The two approaches most suited are the proposed inferential approach that utilises the PLS model developed in section 5.2.2 and the Multi-Way PLS model built upon the pre-selected time points.

The different approaches presented can be compared in terms of the Root Mean Square Error of Calibration (RMSEC) and Root Mean Square Error of Prediction (RMSEP), respectively. These two quantities are defined as :

$$\mathrm{RMSEC} = \sum_{j=1}^{m} \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_{i,j} - y_{i,j}\right)^2}{n}} \qquad (5.3)$$

and

$$\mathrm{RMSEP} = \sum_{j=1}^{m} \sqrt{\frac{\sum_{i=1}^{k} \left(\hat{y}_{i,j} - y_{i,j}\right)^{2}}{k}} \qquad (5.4)$$

where n is the number of batches included in the training set (45), k is the number of batches included in the validation set (18), m is the number of final polymer properties (variables of Y data set, 3) and $y_{i,j}$ $\hat{y}_{i,j}$ are the actual and the predicted value of the j-th polymer property of the i-th batch. These measures estimate the average deviation of the model from the data and provide information about the fit of the model to the training and validation sets, respectively (Wise and Gallagher, 1996). The better the model fits the data, the lower the RMSEC and RMSEP values. Figure 5.26 illustrates the RMSEC values for the MPLS model built upon the entire process database (denoted as MPLS-A), the MPLS model built upon the process samples collected at the 15th, 20th, and 25th minute (denoted as MPLS-B), and the inferential PLS proposed in section 5.2.2. The PLS model outperforms the other models, including the MPLS model of the entire process database. This is reasonable, since the PLS model has been trained upon the actual values of the initial process conditions and does not use any process measurements, as MPLS-A and MPLS-B models do. However, as has been stated, the initial process conditions are not always known exactly. The proposed PLS model was validated against the actual values of initial conditions of the validation set and against the estimates of the initial conditions of both the training and validation set (Figure 5.27). The estimates were obtained by the MPLS model described in section 5.2.3. Predictions obtained from the estimated initial conditions of the training and the validation set are denoted as PLS(45) and PLS(18), respectively, while the prediction obtained from the actual

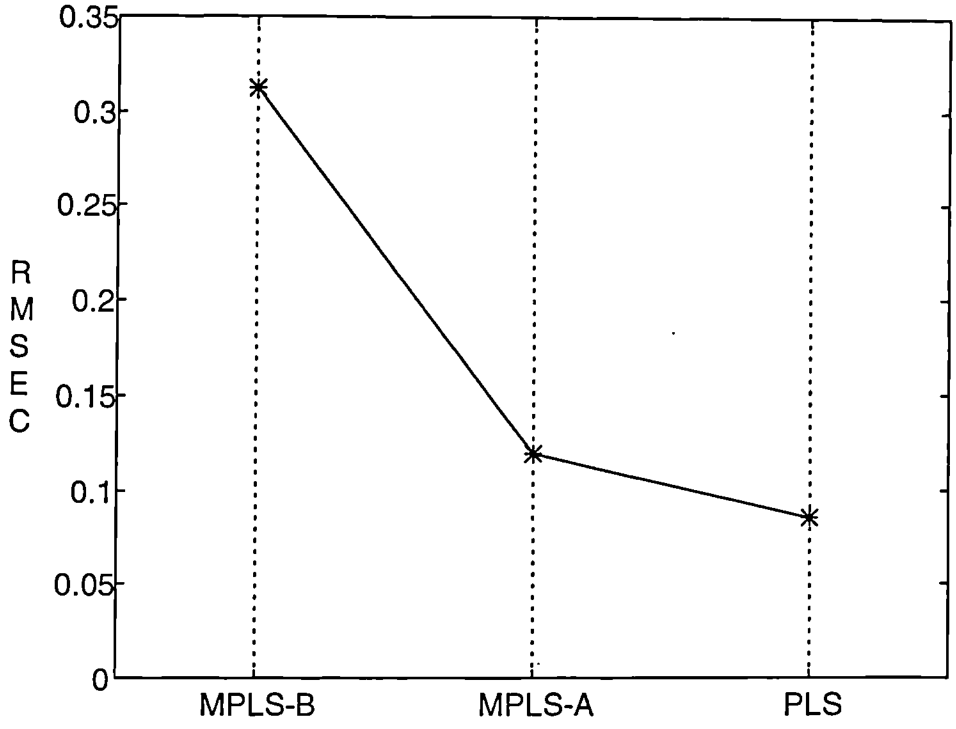


Figure 5.26 Comparison of the fit of models to the training set

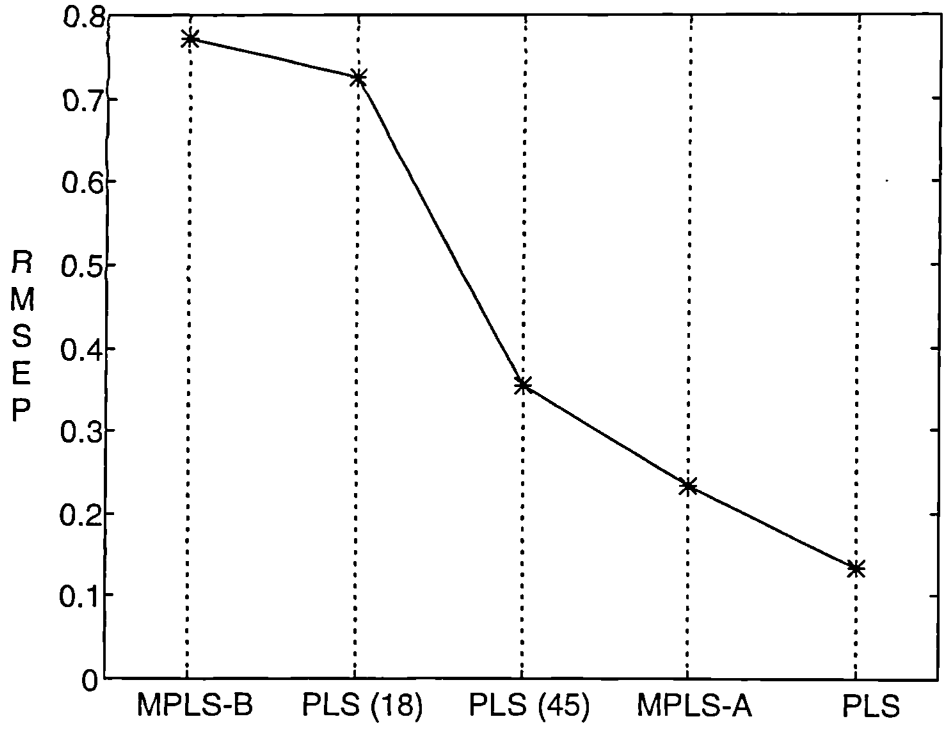


Figure 5.27 Predictive power of the proposed statistical models

initial conditions of the validation set is denoted as PLS. The rest of the models were validated against the validation set only. It can be seen, Figure 5.27, that, the inferential PLS model performs better when actual initial process conditions are available. However, in the case where estimates of initial conditions are utilised, the predictive power of the PLS model is decreased. This is an indication that the predictive power of the PLS model strongly depends on quality of the predictor (X) data set.

- The selection of the most efficient inferential statistical technique to use is difficult, since many factors, such as the nature of the process, the predictive power of the statistical models, and the availability of the data, have to be taken under consideration. When precise initial process conditions are available, the PLS model provides the best approach because of its predictive power and simplicity. This ideal situation is not always realisable under industrial conditions. Alternatively, estimates of the initial conditions can be used as inputs to the model.

5.2.5 Generating Additional Process Data For The Application of MSPC-Based Schemes

Robust MSPC-based monitoring schemes have been applied to chemical and manufacturing processes where large amounts of historical data is readily available. However, difficulties are encountered in situations where only minimal data is available from an experimental design or initial product commissioning. A major challenge is, therefore, to provide a technique that will allow the setting up of an effective monitoring scheme based upon minimal 'design' process data. In Chapter IV, the novel approach of Inverse Projection to Latent Structure (IPLS) was proposed to generate the additional process data for the development of MSPC-based schemes. The most important requirement for the application of the IPLS approach is that a satisfactory PLS regression model can be built from the initial "design" data. The IPLS methodology is illustrated by application to the pilot-scale batch methyl methacrylate polymerisation reactor. Two MSPC-based schemes are developed upon process data generated by the IPLS algorithm, (i) an inferential MPLS model for the estimation of initial process conditions and (ii) an MPCA-based scheme for on-line monitoring, fault detection and diagnosis. The performance of these schemes is compared with the usual approaches of building representations from large amounts of monitored process data.

5.2.5.1 Inverse PLS Model Development

Although a total of forty-five batches were originally generated from the pilot plant simulation, only a sub-set of these were used since the objective is to demonstrate the IPLS methodology for the development of MSPC-based schemes from limited process data. Six sets of training data, comprising five, seven, nine, eleven, thirteen and fifteen batches (N), were generated to investigate the effect of the number of batches on the IPLS model and to identify the minimum number of batches required for the development of an IPLS-based model. An initial set of five randomly selected batches formed the basis of all six data sets. For the set comprising seven batches, an additional two batches were selected from the remaining forty batches; for the set comprising nine, the previous seven formed the basis and an additional two batches were randomly selected; and so on. The complementary set of (45-N) batches were used for validation.

A Multi-Way PLS (MPLS) regression model (Nomikos and MacGregor, 1994b) was first built for each set of N experimental batches, i.e. the "design" process data. The predictor data set X (N×5×120) is defined by the measurements of the five process variables at each of the 120 time points (minutes) of the operation, whilst the corresponding three initial conditions define the response data set Y (N×3). The MPLS models were compared in terms of the Root Mean Square Error of Calibration (RMSEC) on the N batches included in the training data set. Figure 5.28 presents the RMSEC for the six MPLS models. As the number of batches included in the training data was increased, the performance of the MPLS model in fitting the calibration data improved, as expected. However, there was no significant improvement if more than

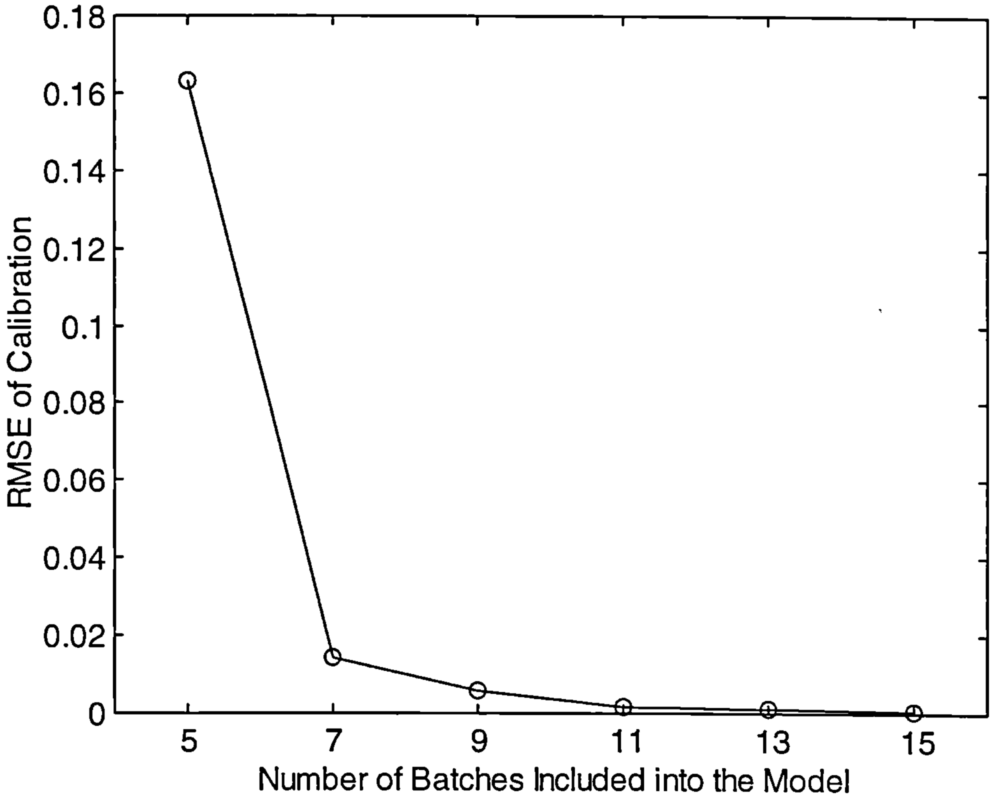


Figure 5.28 Effect of the number of batches included in the MPLS model

eleven batches were used to develop the model. The next step in selecting the number of batches to include in the nominal model, was to invert each of the six individual MPLS models to obtain an Inverted MPLS model. Now the initial conditions define the predictor data set X (N×3), whilst the measured process variables define the response data set Y (N×5×120). The initial process conditions of the remaining (45-N) batches were then used as the inputs in the IMPLS model and estimates of the corresponding process measurements were calculated. The Inverted MPLS models were then compared in terms of Root Mean Square Error of Prediction (RMSEP) of the remaining (45-N) batches. Figure 5.29 illustrates the RMSEP for the six inverted models. It can be seen that, the predictive power of the inverted model increases as

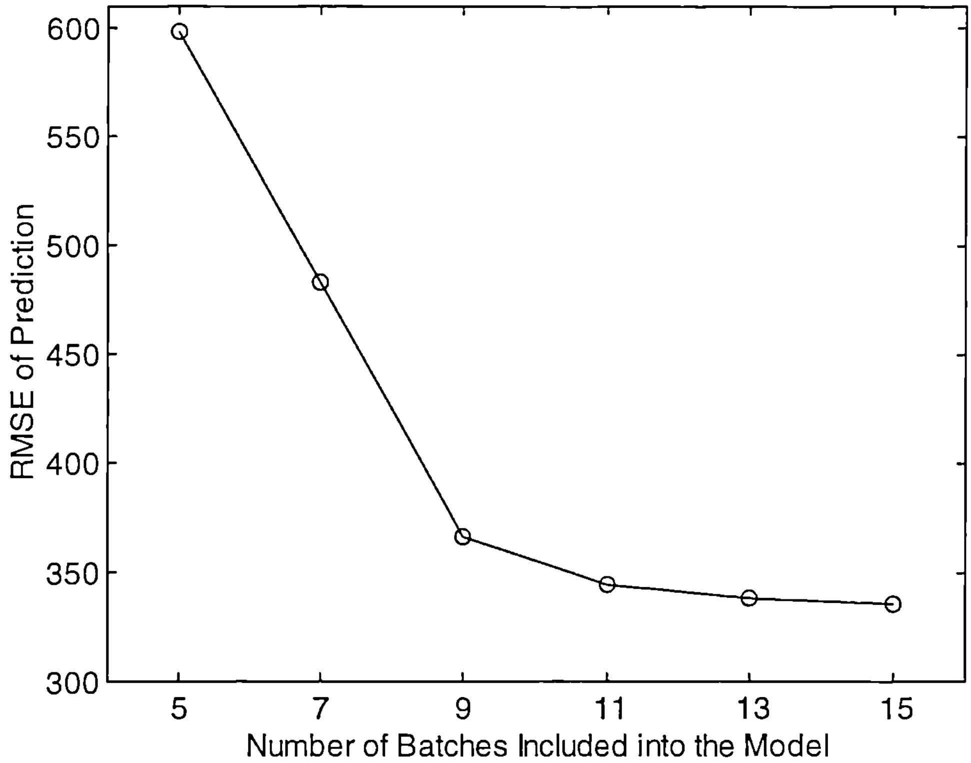


Figure 5.29 Effect of the number of batches included in the MPLS model to the predictive power of the Inverse MPLS model

the number of batches included in the training data set increases. However, no significant improvement in the inverted MPLS model is achieved if more than nine batches are included in the original MPLS model. Moreover, the RMSEP values converge to a minimum as more batches are included in the model. The error that is always be associated with any PLS-based model, is enhanced by an error associated with the IPLS estimates, since the IPLS-estimates are not unique nor the minimum norm, but they are the least squares estimates of the process trajectories.

The final selection of the number of batches that should be included in the MPLS regression model and, therefore, in the Inverted MPLS model, is based upon both the performance of the original and the inverted MPLS models. The number of batches finally selected was a balance between keeping the amount of available process data as small as possible and the performance of the IMPLS model as optimal as possible. Nine batches were selected to form the initial 'design' process data in the subsequent analysis of the methodology.

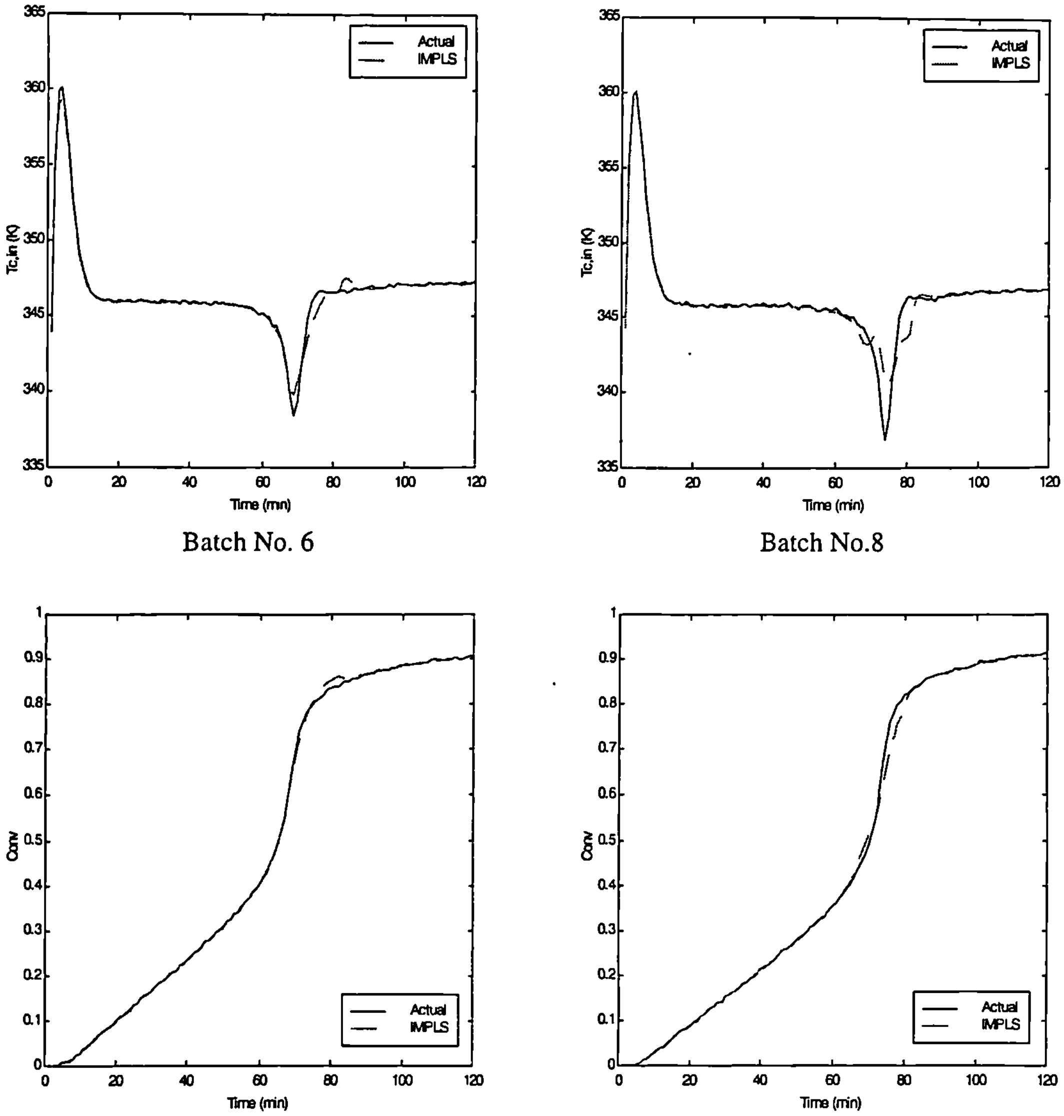
Having defined the desired number of batches to form the basis of the ensuing analysis, the corresponding Multi-Way PLS (MPLS) regression model based upon the nine 'design' batches, was selected. The number of latent variables required to provide a good prediction of the response Y data set was identified through cross-validation to be seven. A summary of the model is presented in Table 5.6. It can be seen that, the first latent variable primarily describes the variability in the temperature set point (T_{sp}), the second latent variable is dominated by initial initiator weight (I_0) and the initial fouling factor (U_0) is the focus of the third latent variable.

LV	% Variability Explained				
	X Block	Y Block	Quality Variables		
	Cumulative	Cumulative	T_{sp}	I_0	U_0
1	46.60	44.14	52.03	33.35	35.47
2	70.29	71.51	67.38	92.39	37.19
3	81.93	98.58	93.69	94.72	85.06
4	86.74	99.82	95.12	96.12	96.62
5	90.95	99.95	97.51	98.99	97.96
6	94.38	99.99	99.31	99.65	98.90
7	97.17	~100.00	99.78	99.84	99.98

Table 5.6 Explained variability by the MPLS model

The MPLS regression model was then inverted. To investigate the ability of the IPLS methodology to predict the trajectories of the process measurements, the set of initial conditions resulting from the nine ‘design’ batches were presented as the inputs to the generated IMPLS model and IMPLS *estimates*, $\dot{x}_0$, of the corresponding process measurements were then calculated.

Two typical process trajectories for the coolant inlet temperature and the monomer MMA conversion were selected from the nine ‘design’ batches, and examined more closely in Figure 5.30. Specifically, trajectories were selected from a batch with initial conditions lying in the middle of the operating region and a batch with initial conditions lying in the edge. These trajectories were compared with the corresponding IMPLS estimates. The IMPLS estimates exhibited greater oscillatory behaviour than the original trajectories. This may be a consequence of overfitting the



Batch No. 6

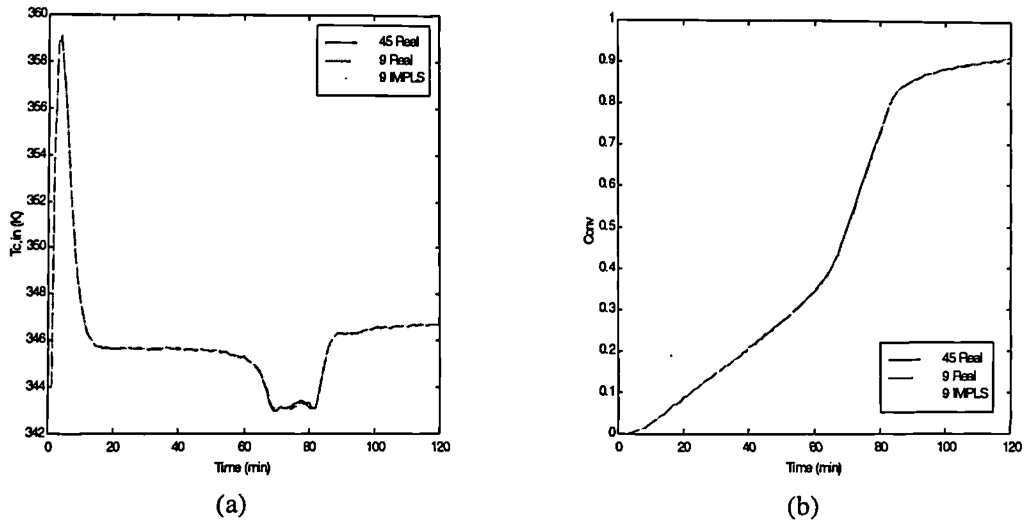
Batch No.8

Batch No.6

Batch No.8

Figure 5.30 Process trajectories of batches with initial conditions lying in the middle (Batch No.6) and in the edge (Batch No.8) of the operating region. Seven latent variables have been retained in the MPLS model.

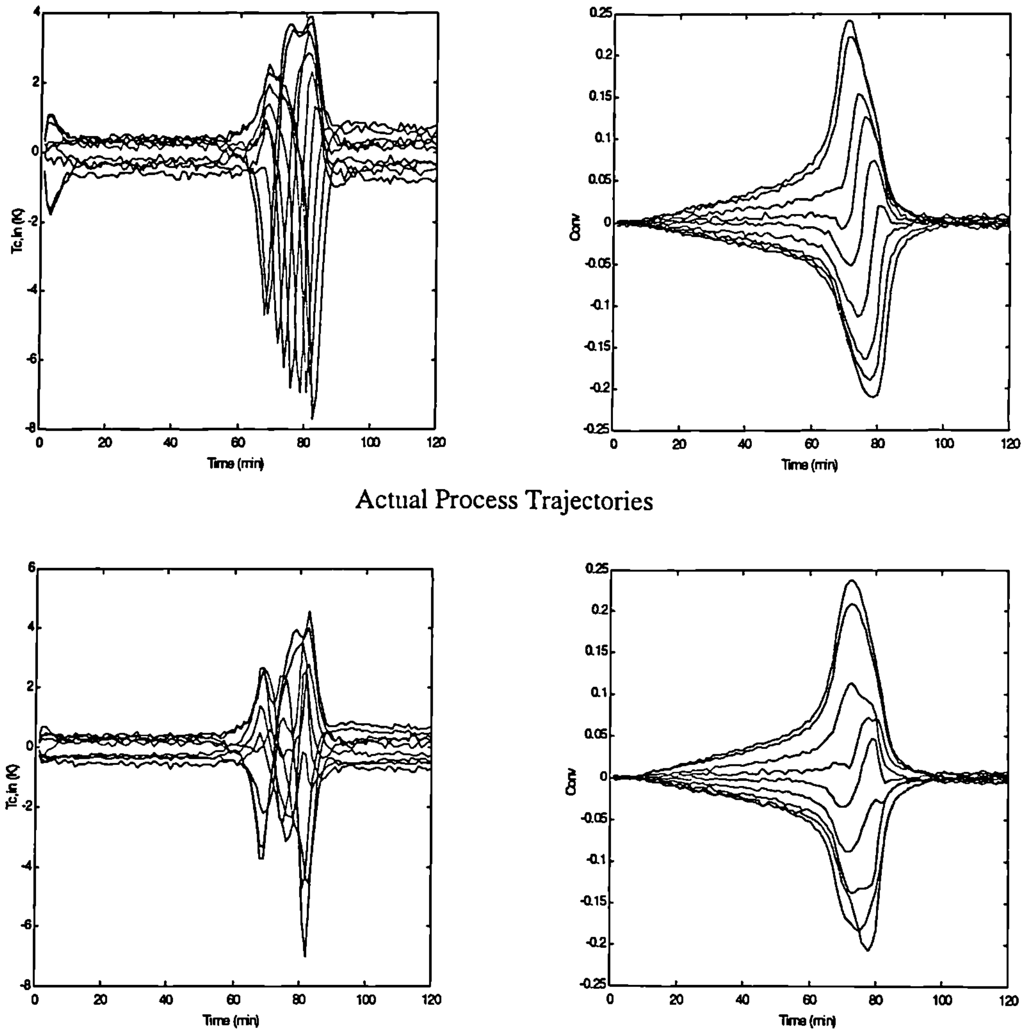
the IMPLS model and/or failing to sufficiently linearise the data. Batch processes are known to exhibit non-linear behaviour. This issue is typically addressed by subtracting the mean trajectory from the actual process trajectory and this should theoretically linearise the data. Figure 5.31 presents the mean trajectories of the coolant inlet temperature and the monomer MMA conversion calculated from the nine 'design' batches (9 REAL), their IMPLS estimates (9 IMPLS) and the training



(a)

(b)

Figure 5.31 Mean trajectories of (a) coolant inlet temperature and (b) monomer MMA conversion



Actual Process Trajectories

IMPLS-Estimated Process Trajectories

Figure 5.32 Deviations of process trajectories from their mean trajectory

set of 45 batches (45 REAL). It can be seen that, the mean trajectories coincide. Figure 5.32 presents the trajectories of the coolant inlet temperature and the monomer MMA conversion for the nine 'design' batches and their IMPLS estimates after having subtracting the corresponding mean trajectories. It can be seen that, the actual process trajectories are still non-linear and, as a result, there is an inherited non-linearity in the IMPLS estimates. One possible solution to this problem is to partition the batch into sections where the process trajectories are more linear in their behaviour and then develop a separate PLS regression model for each model.

The second issue examined was that of overfitting the IMPLS model by retaining seven latent variables in the MPLS regression model, as it was concluded by cross-validation. The possibility of overfitting was investigated by examining the predictive ability of the IMPLS model when different number of latent variables were retained in the MPLS model. The effect of retaining different number of latent variables in the MPLS regression model to the RMSE of Calibration and Prediction of the IMPLS model is presented in Figure 5.33. The RMSE of Calibration and Prediction for the IMPLS model were calculated in terms of equations (5.3) and (5.4), respectively :

$$\text{RMSEC} = \sum_{j=1}^{m} \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{y}_{i,j} - y_{i,j}\right)^2}{n}}$$

and

$$\text{RMSEP} = \sum_{j=1}^{m} \sqrt{\frac{\sum_{i=1}^{k} \left(\hat{y}_{i,j} - y_{i,j}\right)^2}{k}}$$

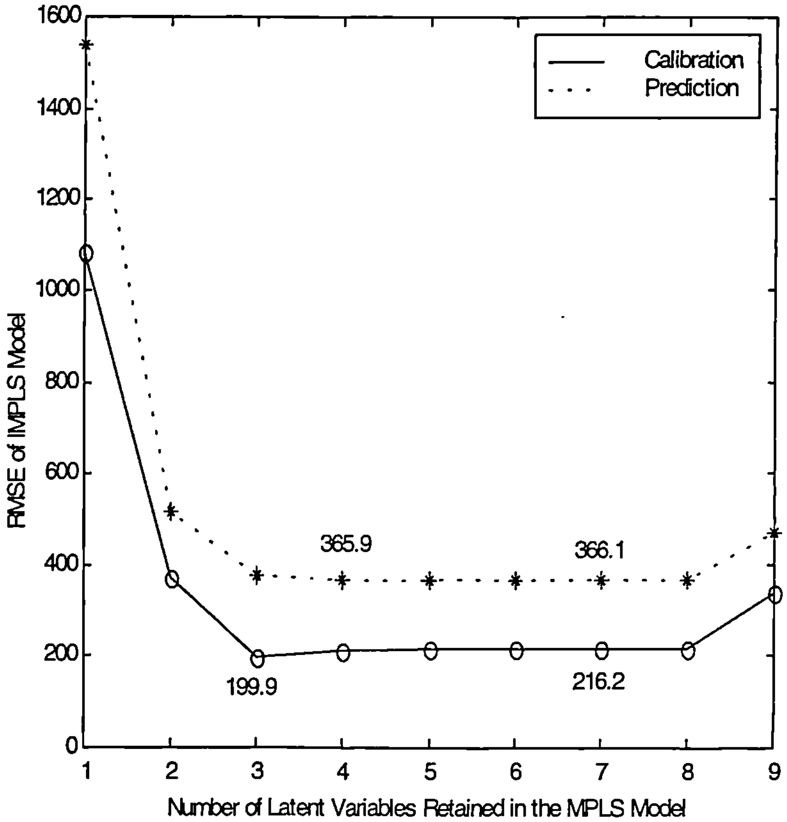
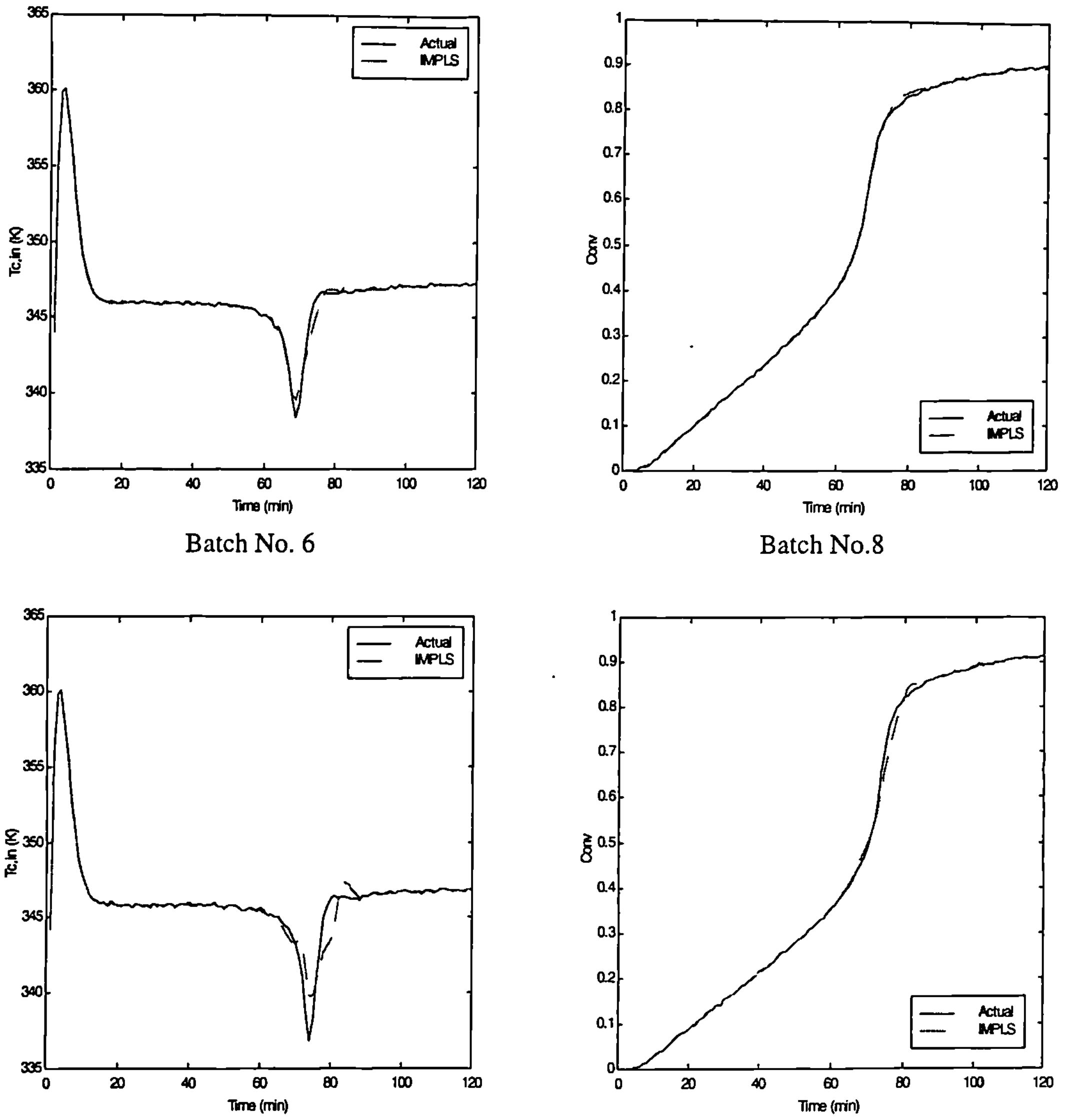


Figure 5.33 Effect of retaining different number of latent variables in the MPLS model to the RMSE of Calibration and Prediction of the IMPLS model

where n is the number of batches included in the training set (9), k is the number of batches included in the complementary set (45-n=9) which was used for validation, m is the number of process measurements that each batch comprises (600), which is equal to the product of the number of process variables (5) by the number of samples collected during the batches (120). It can be seen that, by retaining seven latent variables in the MPLS model, the predictions of IMPLS are overfitted for both the training and validation set. The minimum RMSEC value occurs when three latent variables were retained in the MPLS model, whilst the minimum RMSEP value occurs when four latent variables were retained. However, when applying the IPLS methodology in real processes, a validation set is not available. Therefore, it is preferable to retain in the MPLS model the number of latent variables that is suggested by the RMSEC (3) rather than that suggested by cross-validation (7), in order to minimise overfitting. Figure 5.34 presents the process trajectories illustrated in Figure 5.30, compared with the corresponding IMPLS estimates when three latent variables were retained in the MPLS model. It can be seen that, the oscillations have been reduced and the IMPLS estimates fit better the actual process trajectories. However, greater oscillatory behaviour is still exhibited by IMPLS estimates when the initial conditions of the batch lie in the edge of the operating region, since for extreme values of initial conditions the IMPLS model extrapolates.

The ability of the estimates of the process measurements calculated using the Inverse PLS methodology, to simulate the real process behaviour, was investigated through two application studies. The first application relates to the development of an inferential Multi-Way PLS model to estimate the initial process conditions at an early stage in the polymerisation process. The second application relates to the development of an MSPC-based scheme for monitoring, fault detection and diagnosis, based upon a Multi-Way Principal Component Analysis (MPCA) model. Both applications were implemented using both the estimates of the process measurements derived from the IMPLS model when three latent variables were retained in the MPLS model, and the corresponding process measurements obtained from the pilot plant simulation, which act as surrogate process data.



Batch No. 6

Batch No.8

Batch No.6

Batch No.8

Figure 5.34 Process trajectories of batches with initial conditions lying in the middle (Batch No.6) and in the edge (Batch No.8) of the operating region. Three latent variables have been retained in the MPLS model

5.2.5.2 Application 1 - Estimation of Initial Process Conditions

The estimation of initial conditions for batch polymerisation reactors at an early stage of the polymerisation process using an MPLS regression model was considered in section 5.2.3. It was concluded that the optimal scenario for the process under consideration was to build a Multi-Way PLS model from data collected at the 15th, 20th and 25th minute. Based upon this philosophy, an MPLS regression model was built upon the training data set that comprises the inverted MPLS-estimated process measurements for the thirty-six nominal batches (36E) and the process measurements for the nine 'design' nominal batches (9R), i.e. forty five batches in total. This model is termed the *mixed model* (9R+36E) and it relates the process measurements at the 15th, 20th and 25th minute of the polymerisation process to the initial conditions of interest, namely, the initial initiator weight (I_0), the initial heat transfer coefficient (U_0) and the reactor temperature set-point (T_{sp}). Concerning the construction of the response (Y) data set, two scenarios are possible. The first represents the ideal situation where the actual initial process conditions are known. The second, and more realistic case is where accurate initial process condition records are unavailable. In order to represent this latter situation, the Y data set, in the analysis of the mixed model, was in-filled with the theoretical values of the initial process conditions, which correspond to that particular polymer grade being modelled (Set 1, Table 5.3). The reactor temperature set-point is included since it has been found that it improves the predictive capabilities of the model.

Table 5.7 shows the amount of variability explained by the MPLS model for each block and for each of the initial conditions. The initial process conditions are fairly well estimated by four latent variables with 98.5% and 95% of the total variability being explained in the X and Y data sets, respectively. The number of latent variables to retained in the model (4) was selected using cross-validation. The first principal component primarily describes the variability in the temperature set point (T_{sp}), the

LV	% Variability Explained				
	X Block	Y Block	Quality Variables		
	Cumulative	Cumulative	T_{sp}	I_0	U_0
1	54.14	32.59	49.19	6.65	21.57
2	84.00	65.48	66.89	47.48	47.77
3	95.04	94.72	88.88	80.19	75.57
4	98.52	95.52	94.20	83.58	80.42

Table 5.7 Explained variability by the mixed model

second principal component is dominated by initial initiator weight (I_0) and the initial fouling factor (U_0) is the focus of the third principal component.

The second Multi-Way PLS model considered is that of section 5.2.3, which was built from the original forty-five batches of the training data set obtained from the pilot plant simulation. This model is termed *the original model* (45R) and, again, its objective was to estimate the initial process conditions. The model used the process measurements at the 15th, 20th and 25th minute of the polymerisation process to infer the actual, but seldom realisable values of the initial conditions (MC1-MC7, Table 5.3). The original model is presented as reference to the best possible predictions attainable by MPLS models.

Figure 5.35 shows the estimates of the initial conditions for the initiator weight and the heat transfer coefficient for an additional twelve previously "unseen" batches (Kiparissides, 1996), using both the mixed MPLS model (*) and the original MPLS model (+). The actual values of the initial conditions used in the simulation model to produce the trajectories are indicated by (o). As can be seen, the twelve batches simulate a sequence of operations where the reactor is subjected to fouling, cleaned and again subjected to fouling. The mixed MPLS model is seen to provide satisfactory estimates of the initial conditions, which are close to the actual values. This has been achieved in spite of the model being built from the less precise, but more realistic, set of initial conditions.

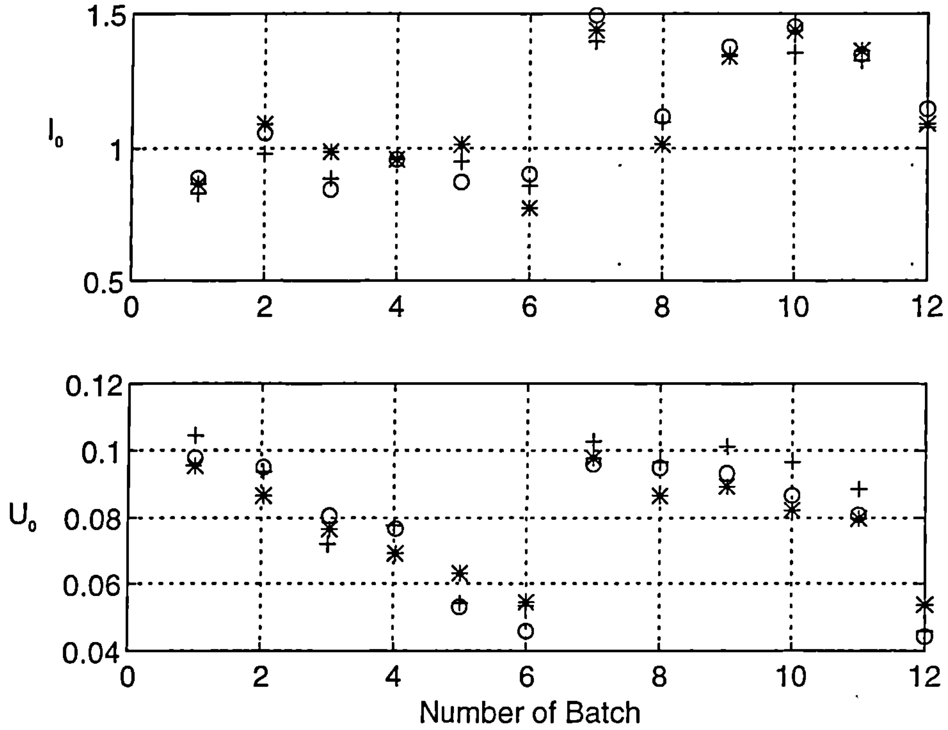


Figure 5.35 Estimated initial conditions for new batches using the mixed MPLS model (*) and the original MPLS model (+)

The performance of the mixed MPLS model (9R+36E) was then compared with three MPLS models built from different historical databases. For all approaches only those measurements recorded at the 15th, 20th and 25th minutes of the polymerisation process were used in the model development. The first model was built from the actual process measurements of the nine 'design' batches (9R), whilst the second MPLS model was based upon IMPLS-estimated process measurements of the thirty six complementary batches (36E). The final model was the original model, which was been built using the process measurements from the forty-five batches (45R) from the pilot plant simulation. Figure 5.36 illustrates the RMSE of Calibration for the N batches included in the training data set and the RMSE of Prediction on twelve "unseen" batches with respect to the theoretical initial conditions (Set 1, Table 5.3). Figure 5.37 shows the RMSE of Calibration for the N batches included in the training data set and the RMSE of Prediction on twelve "unseen" batches with respect to the actual initial conditions (MC1-MC7, Table 5.3). The temperature set-point (T_{sp}) has now been excluded from the RMSE calculations, since it is estimated fairly well by all the models and its contribution to RMSE values can be neglected. It can be seen in Figures 5.36 and 5.37 that, the performance of the mixed MPLS model in fitting both the training and the validation data sets, is quite similar to the optimal performance of the original MPLS model. The potential power of the mixed model (9R+36E) arises from the fact that it is a combination of a model built upon a few actual process measurements (9R) which provides sufficient "quality" of information about the process, with a model built upon IMPLS-estimated process measurements (36E), which provides sufficient "quantity" of information about the process. As can be seen from the performance of the (9R) and (36E) models, both sufficient "quality" and "quantity" are not enough to develop a robust model. Therefore, it was concluded that the mixed model can be used as an alternative to the original model when enough process data is not available to construct it.

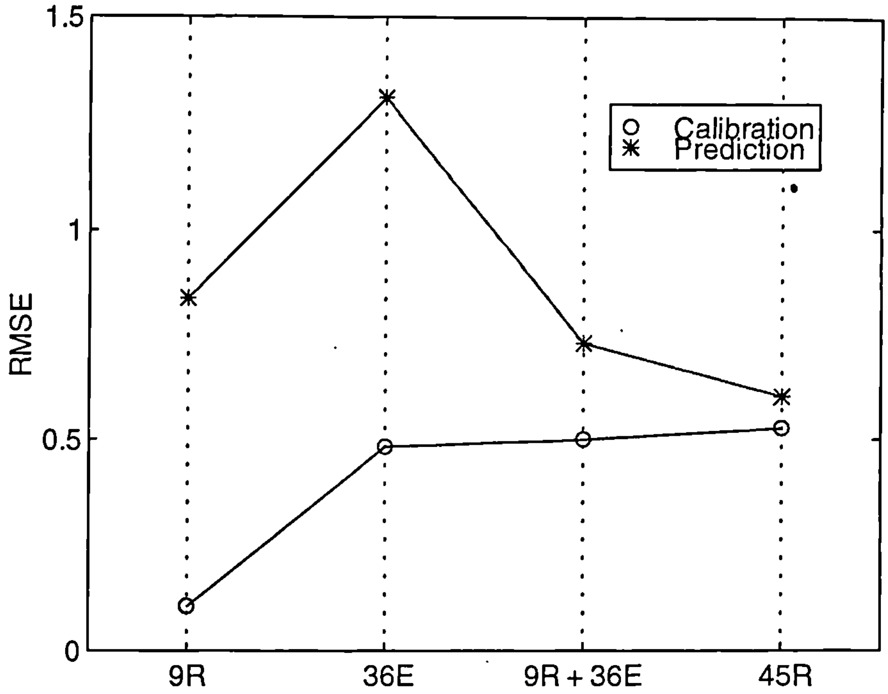


Figure 5.36 RMSE of Calibration and Prediction with respect to the theoretical initial conditions

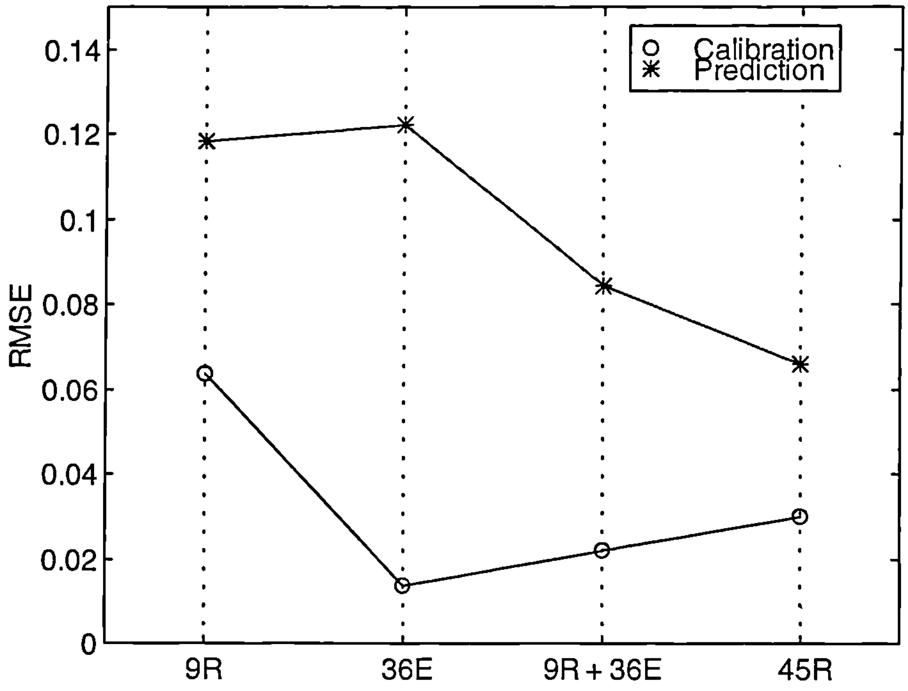


Figure 5.37 RMSE of Calibration and Prediction with respect to the actual initial conditions

section, the ability of the Inverse MPLS approach to establish a reliable MPCA monitoring scheme is investigated.

An MPCA model was built from a training data set comprising the inverted MPLS-estimated process measurements for the thirty-six nominal batches and the process measurements from the nine 'design' batches, i.e. forty five batches in total, i.e. *the mixed MPCA model.* A monitoring scheme based upon this set of process measurements was then developed. This approach was evaluated by comparing its performance to the performance of the MSPC scheme based upon an MPCA model built from the corresponding process measurements obtained from the pilot plant simulation for the forty-five batches, i.e. *the original MPCA model.*

Table 5.8 summarises the percentage of variability explained for the mixed MPCA model. Cross-validation showed that only three principal components were required to explain the majority of the variability in the X data set. Two additional batches were generated from the pilot plant MMA polymerisation simulation. The first batch (number 46) represents normal operation, since the initial conditions lying within the nominal ranges as defined in Table 5.1. The second batch (number 47) represents an

Principal Component	**% Variability Explained**	
		Cumulative
1	48.90	48.90
2	27.67	76.57
3	15.52	92.09

Table 5.8 Variability explained by the mixed MPCA model

example of a batch where there was an initiator problem. Although, the initial initiator weight is 25% below that of the nominal range, the resultant product quality only just lies outside the specification limits.

Figures 5.38 and 5.39 show the projection of the two new batches onto the reduced space of the first two principal components of the mixed calculated from the mixed and the original MPCA models, respectively. It can be seen that, both models clearly classify the first batch as normal and the second batch as abnormal. The abnormal behaviour was also identifiable in Figure 5.40 from the Residual Sum of Squares (RSS) or Q-statistic plot, since the resultant value was larger than the 99% control limit for the original MPCA model. However, for the mixed model both batches exceeded the 99% control limit (Figure 5.41). This latter result can be explained by looking at the composition of the Residual Sun of Squares (Q) for the MPCA model, more closely :

$$Q = \sum_{i=1}^{\text{Batches}} \sum_{k=1}^{\text{Time}} \sum_{j=1}^{\text{Variables}} E_{i,k,j}^2 \qquad (5.5)$$

The Q-statistic is a metric based upon a measure of the deviation of the process measurements from the MPCA representation, given by the residual matrix, E. This is calculated for each individual batch X_i :

$$E_i = X_i - X_i \, P \, P^T \qquad (5.6)$$

where X_i is a matrix containing the process measurements of the i-th batch and P is the three-dimensional array of the loadings. The calculation of the control limits for

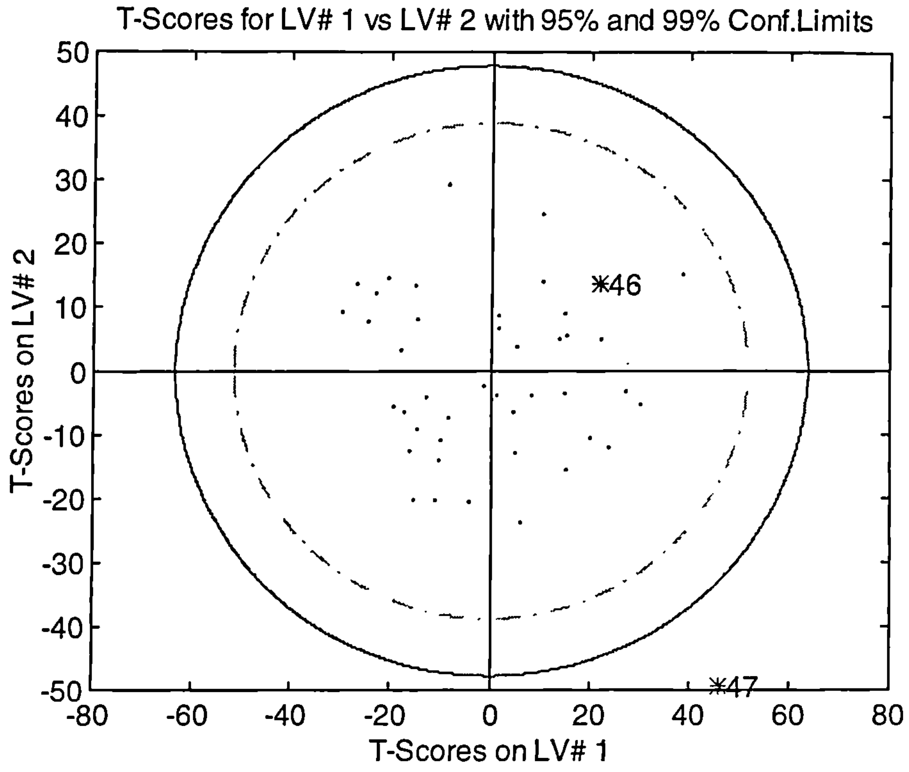


Figure 5.38 Projection of the new batches onto the reduced space define by the original MPCA model

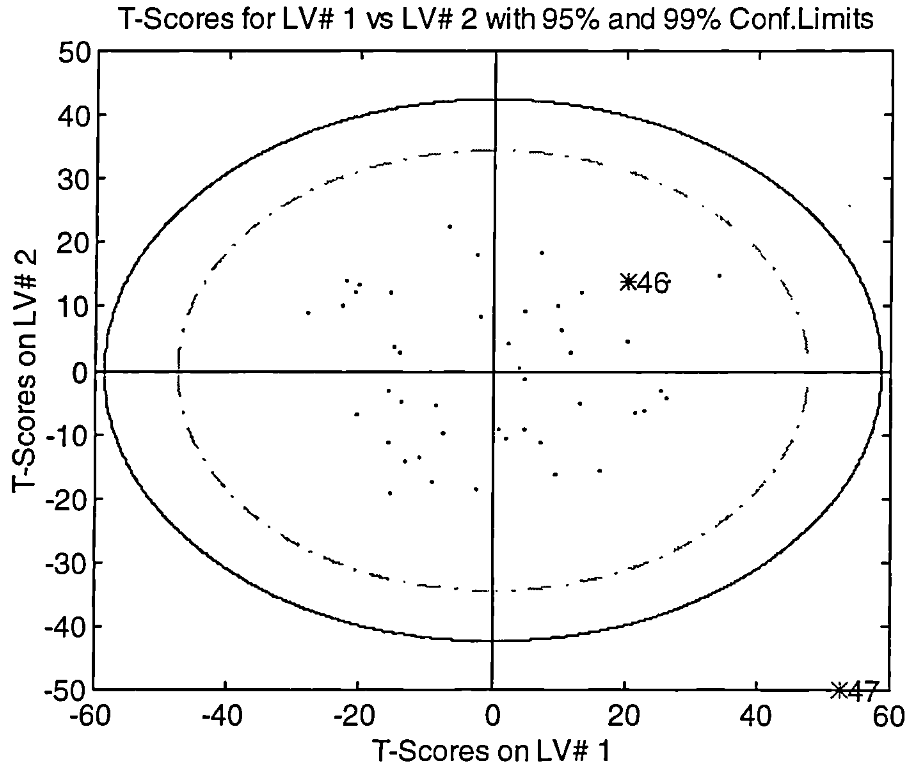


Figure 5.39 Projection of the new batches onto the reduced space define by the mixed MPCA model

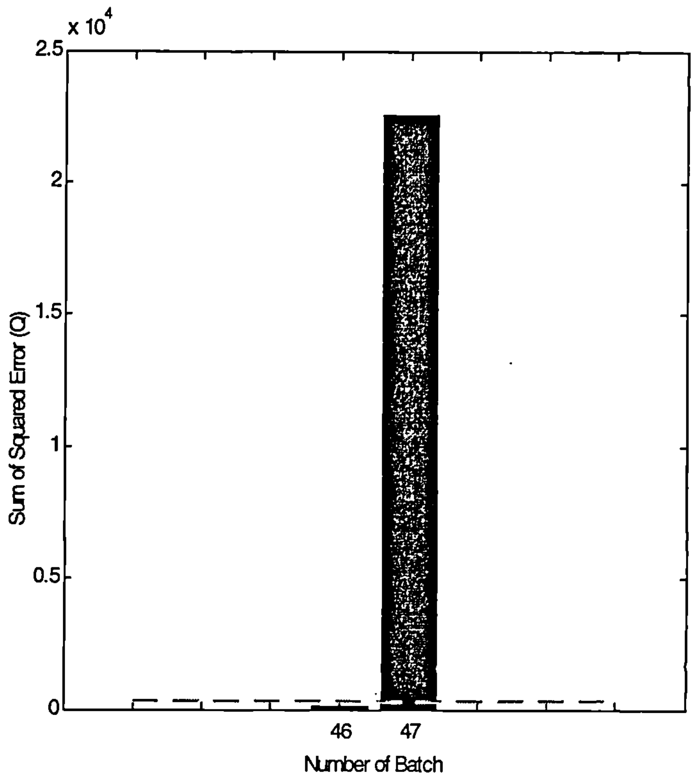


Figure 5.40 Q-statistic for the new batches using the original MPCA model

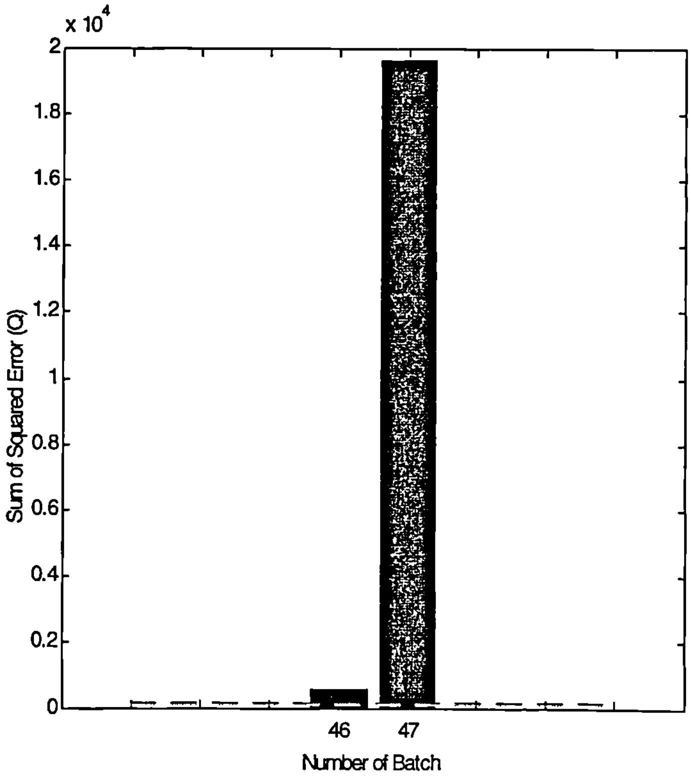


Figure 5.41 Q-statistic for the new batches using the mixed MPCA model

the Q-statistic requires the calculation of the residuals for each batch included in the nominal database, which in the case of the IMPLS-estimates, can be written as :

$$\dot{E}_i = \dot{X}_i - \dot{X}_i \, P \, P^T \tag{5.7}$$

where P denotes the loadings array for the mixed MPCA model and $\dot{X}_i$ denotes the IMPLS estimates of the process measurements for each batch. These estimates contain an error associated with the approximation of the real process measurements by their IMPLS estimates, which cannot be calculated since the real measurements of the batches will in practice be unavailable and which is inherited to the control limits. When calculating the Q-statistic for a new batch, this error will not be present in the process measurements :

$$E_{new} = \mathbf{x}_{new} - \mathbf{x}_{new} \, \mathbf{p} \, \mathbf{p}^T \tag{5.8}$$

where $\mathbf{x}_{new}$ is the unfolded vector of the new real measurements and $\mathbf{p}$ is the unfolded array of the mixed MPCA loadings. This results in the calculated values for the Q-statistic and the Squared Prediction Error (SPE) for each new real batch potentially exceeding the nominal control limits. As a result, the Residual Sum of Squares or Q-statistic and the Squared Prediction Error are unreliable measures of operating performance, since they will contain an error, which is not quantifiable and which will inflate these two metrics.

Following on from the development of the mixed MPCA model, the next question of interest is whether the MSPC monitoring scheme based upon the mixed MPCA model is able to identify abnormal operation and to differentiate between different assignable causes. The major problem associated with this approach is that the vector of process measurements ($\mathbf{x}_{\text{new}}$) is not complete until the end of the batch. The *projection method* as described in section 4.5.1.1, is adopted. This method considers the unknown future observations as missing values and uses the principal components of the MPCA model to predict these missing values by restricting them to be consistent with those values already observed up to time interval, k, and with correlation structure of the measurement variables in the database as defined by the loading matrices (P) of the MPCA model. MPCA does this by projecting the already known observations down onto the reduced space and calculating the scores at each time interval as :

$$\mathbf{t}_{r,k} = \left(P_k^T P_k\right)^{-1} P_k^T \mathbf{x}_{\text{new},k} \qquad k = 1,2,\ldots,K \qquad r = 1,2,\ldots,R \tag{5.9}$$

where $\mathbf{t}_{r,k}$ is a vector containing the scores of all the retained principal components up to time point k and P_k is a matrix whose columns are defined to be the elements of the unfolded three dimensional array of the MPCA loadings (P). This method has been found to be superior to the others proposed if at least ten percent of the measurements of a new batch are known (Nomikos and MacGregor, 1995).

Figures 5.42 and 5.43 and illustrate the on-line monitoring for the first score and SPE for batch number 46 using the mixed MPCA model, whilst Figures 5.44 and 5.45 present the on-line monitoring for the first score and SPE for the batch number 46 using the original MPCA model. It can be seen that, both MPCA models can successfully monitor the evolution of a normal batch in the score plots. However, only the original model can monitor successfully the batch in the SPE plot, since using the mixed model the SPE continuously exceeds its control limits, as it can be

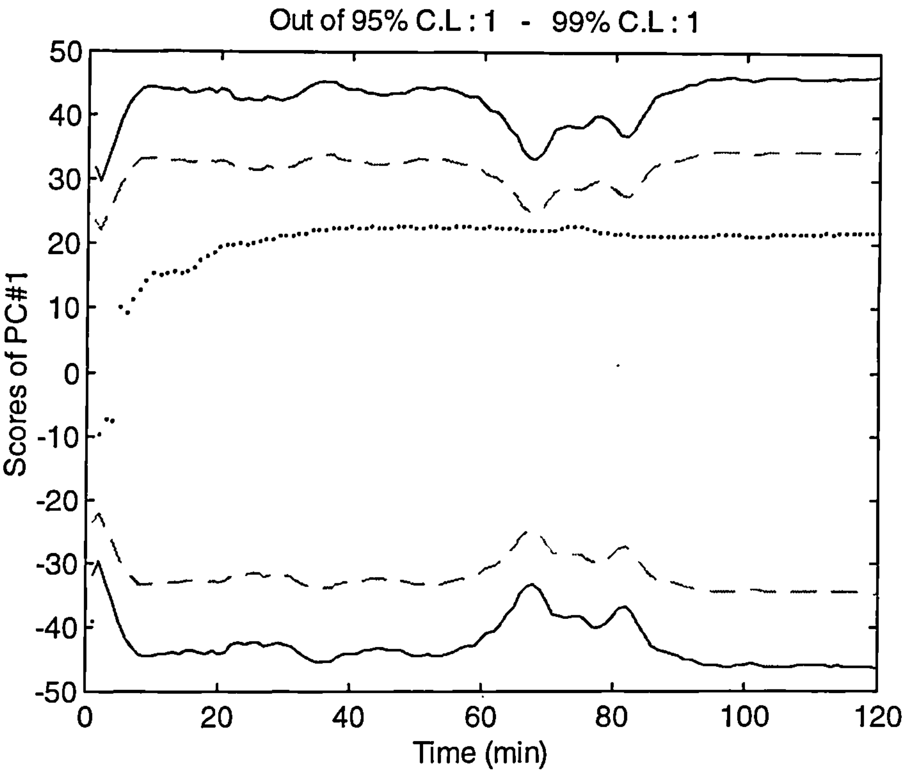


Figure 5.42 Monitoring of the first score of the mixed model for batch number 46

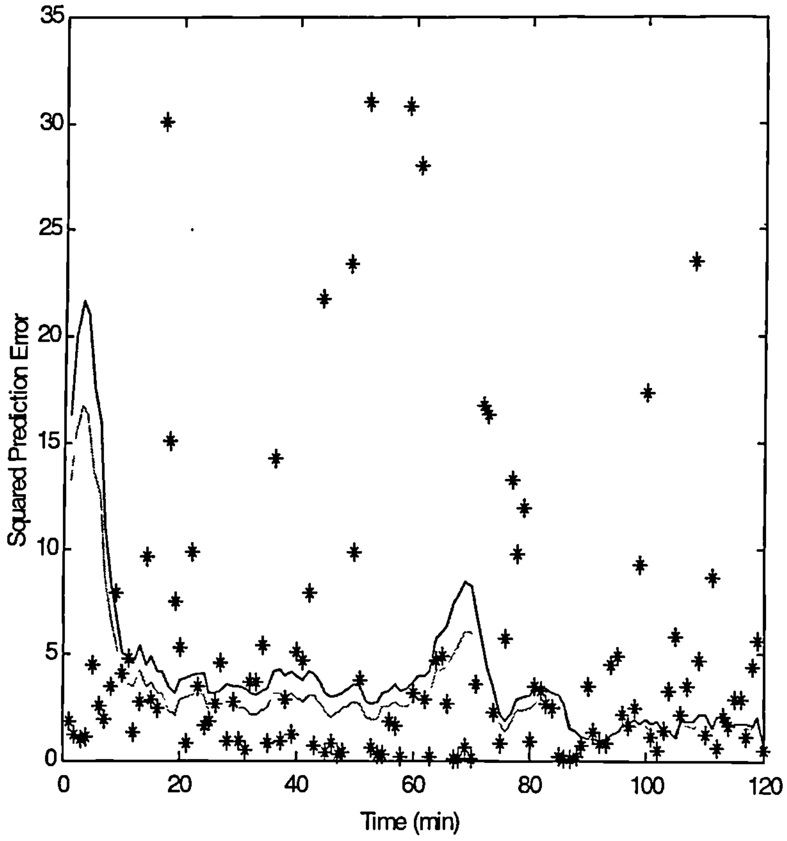


Figure 5.43 Monitoring of SPE of the mixed model for batch number 46

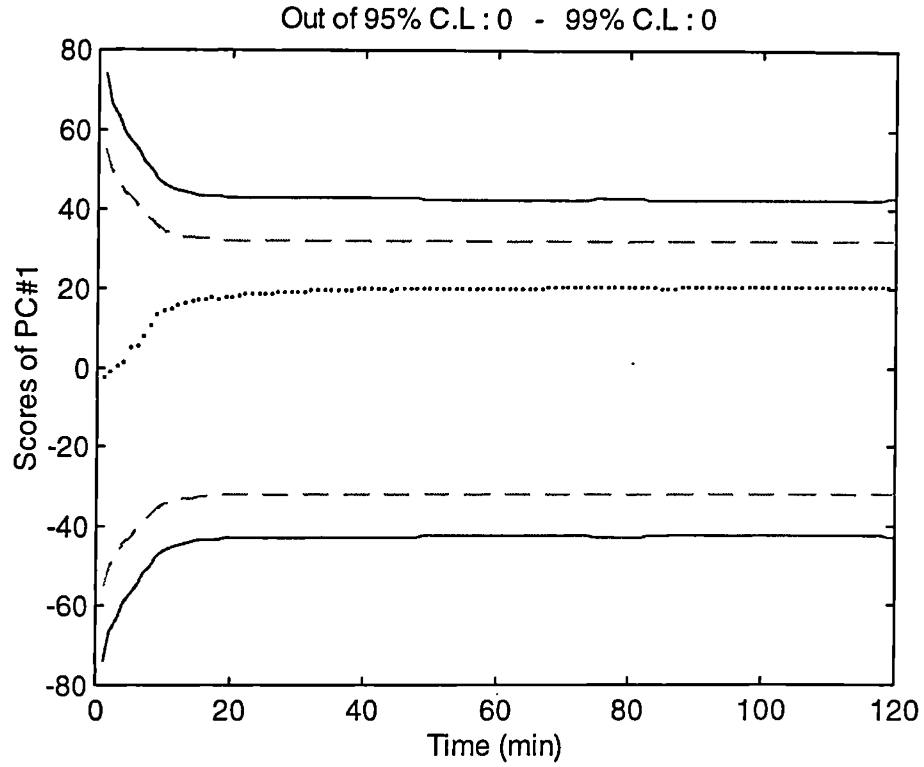


Figure 5.44 Monitoring of the first score of the original model for batch number 46

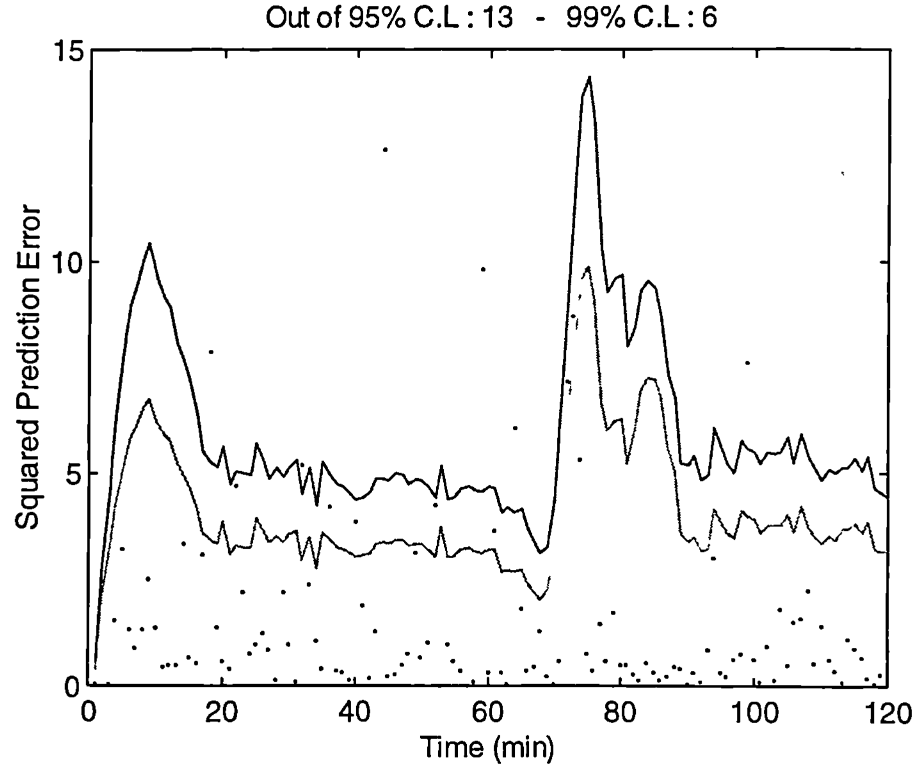


Figure 5.45 Monitoring of SPE of the original model for batch number 46

seen in Figure 5.43.

Figures 5.46 and 5.47 show the on-line monitoring score and SPE charts, respectively, for the abnormal batch number 47, using the original MPCA model. It can be seen that, the unusual event can be detected in both plots. The next task is to identify the cause of the problem by interrogating the underlying MPCA model. This is achievable by examining the contribution of the individual process variables to the SPE and to the scores value for the first principal component, at the time point where the fault occurred (MacGregor et al., 1994).

The SPE at each time point (k=1,...120) is the sum of the squared prediction errors for all the process variables (j=1,...,m) (equation 4.4.1) :

$$SPE_k = \sum_{j=1}^{m} \left(x_{k,j} - \hat{x}_{k,j} \right)^2$$

where the predictions $\hat{x}_{k,j}$ are calculated from the original MPCA model. Each of the terms $\left(x_{k,j} - \hat{x}_{k,j} \right)$ account for the contribution of the corresponding j-th process variable to the SPE at the k-th time point and it is denoted the Prediction Error. Similarly, the score of the r-th principal component at each time point k is the sum of the product of the current value of the process variable $\left(x_{k,j} \right)$ times their contribution to the principal component under consideration ($\omega_{r,k,j}$) (equation 4.4.4) :

$$t_{r,k} = x_{k,1}\, \omega_{r,k,1} + \ . \ . \ . \ + \ x_{k,J}\, \omega_{r,k,m}$$

However, the scores have been calculated by an MPCA projection (equation 5.9) and, therefore, each time point is the sum of the contributions of each individual process

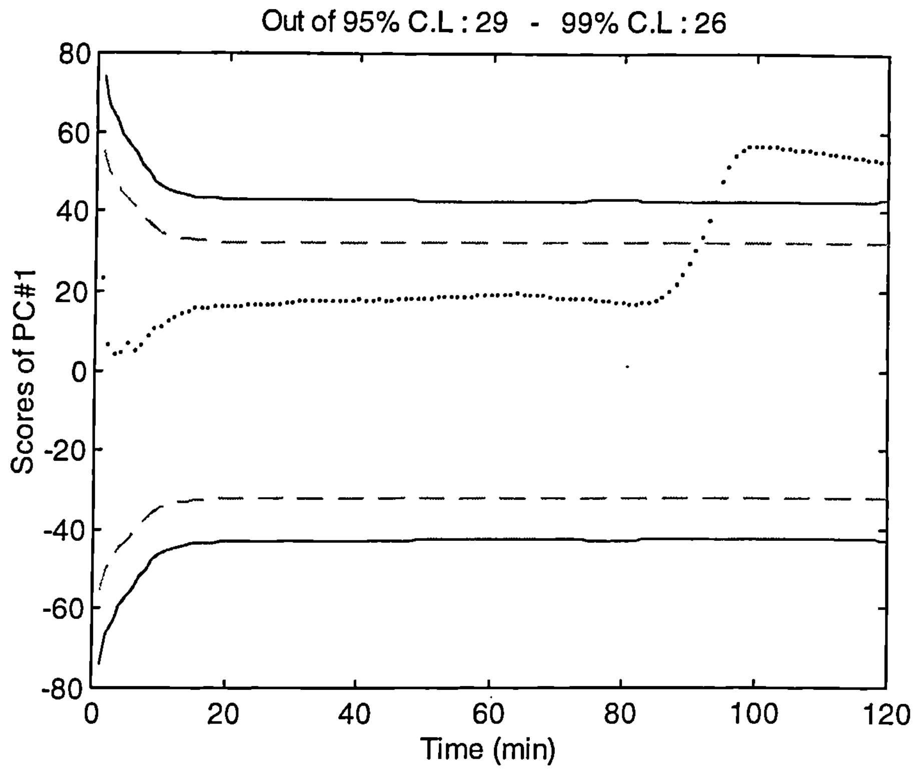


Figure 5.46 On-line monitoring of score of the original model for batch number 47

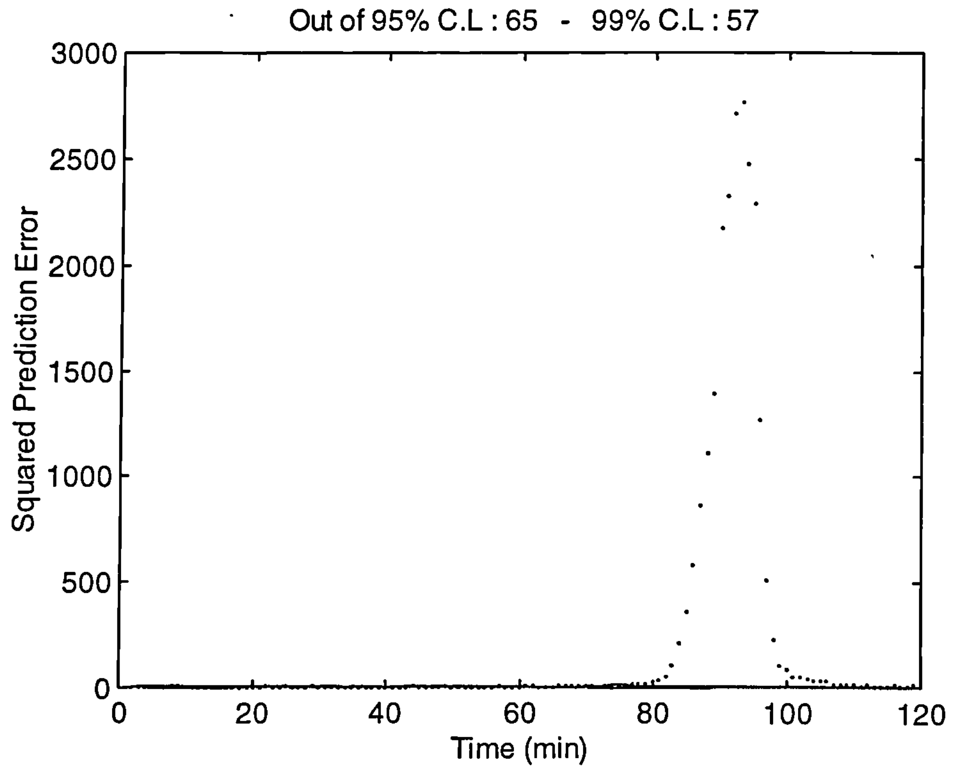


Figure 5.47 On-line monitoring of SPE of the original model for batch number 47

variables on a cumulative basis up to the time point of interest, k, :

$$t_{r,k} = \sum_{n=1}^{k} x_{n,1}\, \rho_{r,n\times 1} + \sum_{n=1}^{k} x_{n,2}\, \rho_{r,n\times 2} + \cdots + \sum_{n=1}^{k} x_{n,m}\, \rho_{r,n\times m} \qquad (5.10)$$

$\rho_{r,n\times j}$ is the element of the matrix $\left(P_k^T\, P_k\right)^{-1} P_k^T$ of dimension $(R \times k\ m)$ at each time point k (m is the number of process variables). The contribution of the individual process variables to the change in the value of the score between time point k_1 and time point k_2 can be calculated as :

$$t_{r,k} = \sum_{n=k_1}^{k_2} x_{n,1}\, \rho_{r,n\times 1} + \sum_{n=k_1}^{k_2} x_{n,2}\, \rho_{r,n\times 2} + \cdots + \sum_{n=k_1}^{k_2} x_{n,m}\, \rho_{r,n\times m} \qquad (5.11)$$

Closer examination of the differential contribution of each variable to the score for the first principal component at the point where the score lies outside the 99% control limits i.e. between the 90th and 96th time points, indicates that the variables contributing primarily to the problem are the jacket temperatures and the conversion of monomer MMA (Figure 5.48). However, at the 82nd time point, where the SPE initially moves outside the 99% control limit (Figure 5.49), the major instantaneous contribution comes from conversion of monomer (note that the reactor temperature is denoted as T_{reac}, while the conversion of monomer MMA is denoted as Conv). From *a priori* knowledge of the process (Kiparissides, 1996), it was concluded that, the main cause of the fault is a low amount of initiator. The increased value in the jacket temperatures can be explained by the underlying relationship between them and conversion of monomer MMA. However, the situation becomes clearer if we look at

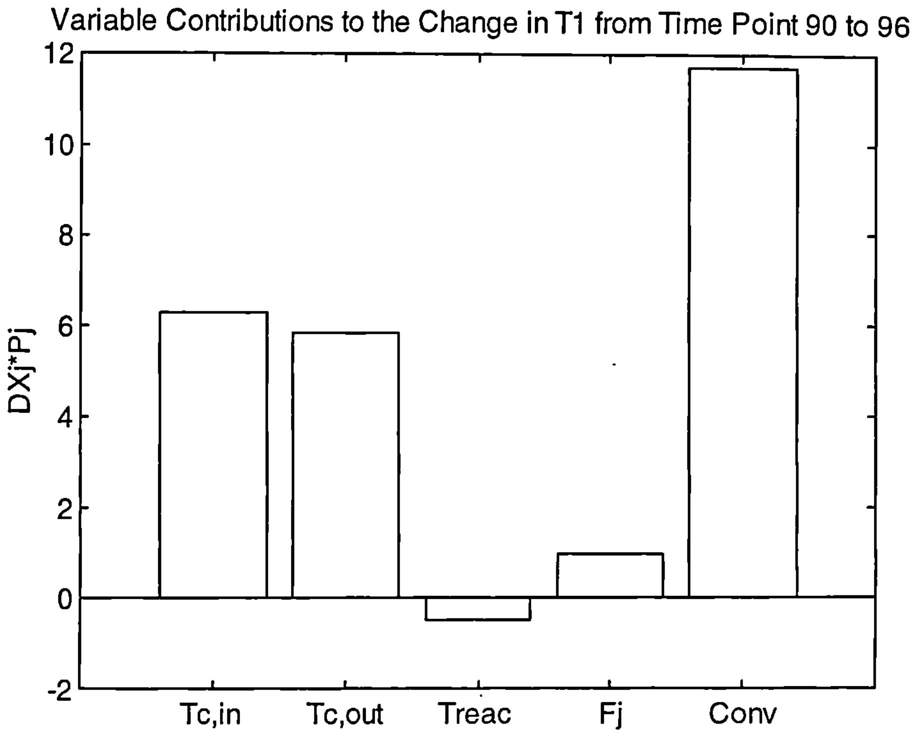


Figure 5.48 Differential contributions to the first score between the 90th and the 96th minute

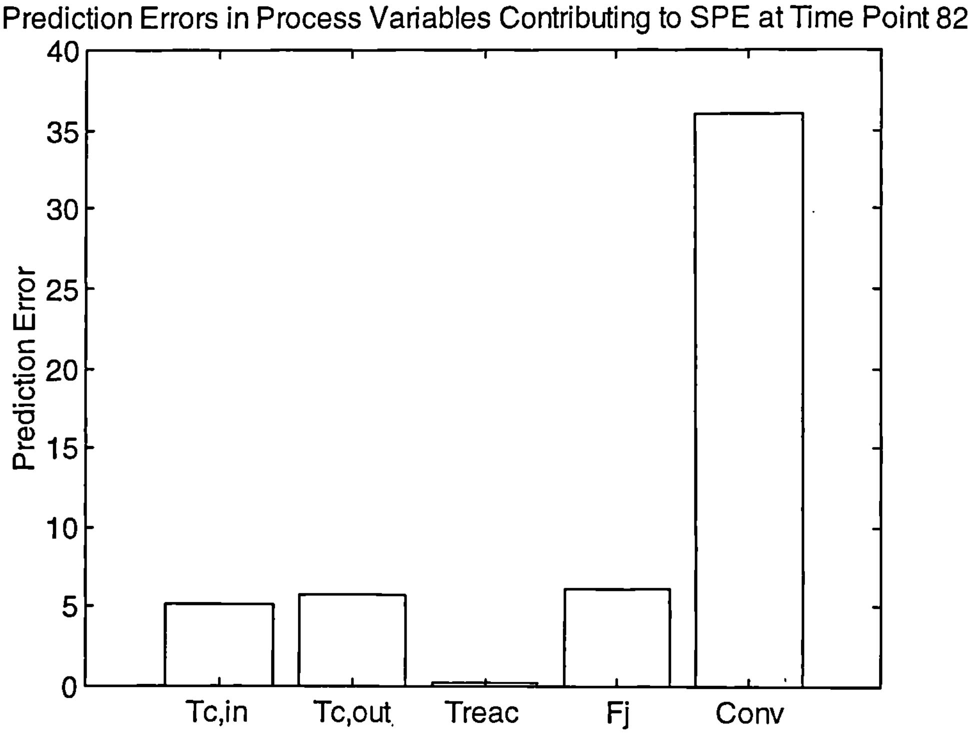


Figure 5.49 Instantaneous contributions to SPE at the 82nd min.

the continuous contributions plots of the first score and the SPE (Figure 5.50 and 5.51, respectively). These plots, which are differential contributions plots between the current time point k ($k_2 = k$) and the beginning of the batch ($k_1 = \text{1st min}$), clearly show that conversion of monomer MMA is mainly responsible for the deviations both in the scores for the first principal component and in the SPE plots, since its contribution rises before the contributions of the temperatures.

The monitoring procedure for the abnormal batch number 47 using the mixed MPCA model is illustrated in Figure 5.52. The unusual event, again, is detected by the model in the first score. The SPE plot is not utilised, since it was shown that it is an unreliable measure. The contributions of the process variables to the movement of the first score between the 85th and the 95th time points (Figure 5.53), again, show that conversion of monomer MMA is mainly responsible and, therefore, the amount of initiator injected into the reactor is identified as the main cause of the problem. The plot of the continuous contributions to the first score (Figure 5.54) confirms that conversion is clearly indicative of the fault and indicates that the fault started to be observed on both charts at the same time points (80th minute).

It is concluded that an MSPC scheme based upon an MPCA model, which has been built upon the IMPLS estimates of the real process measurements, is able to successfully monitor new batch and to identify faults. Although, it is not as reliable as the scheme based upon a model of the real process measurements, it can be improved as more new real process measurements, from completed normal batches, become available.

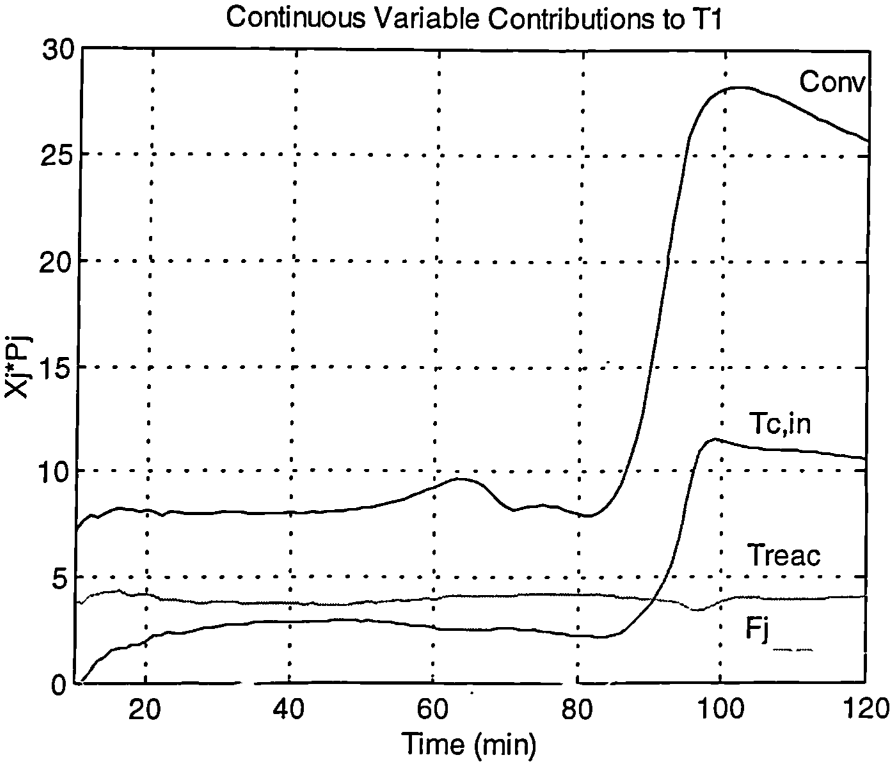


Figure 5.50 Continuous contributions to the first score for batch number 47

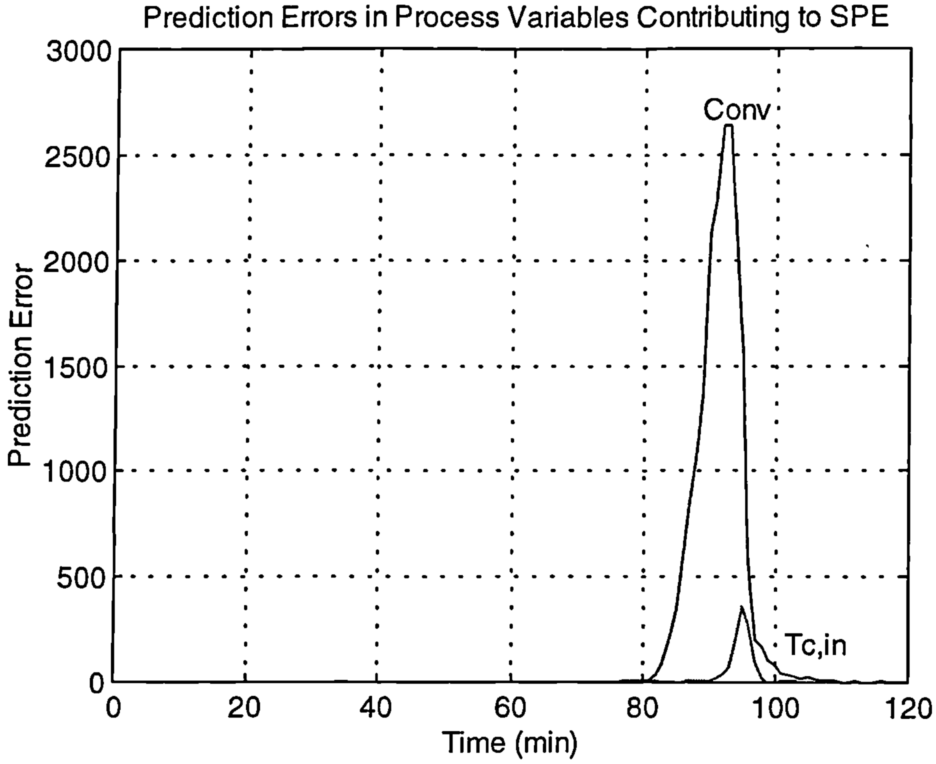


Figure 5.51 Continuous contributions to SPE for batch number 47

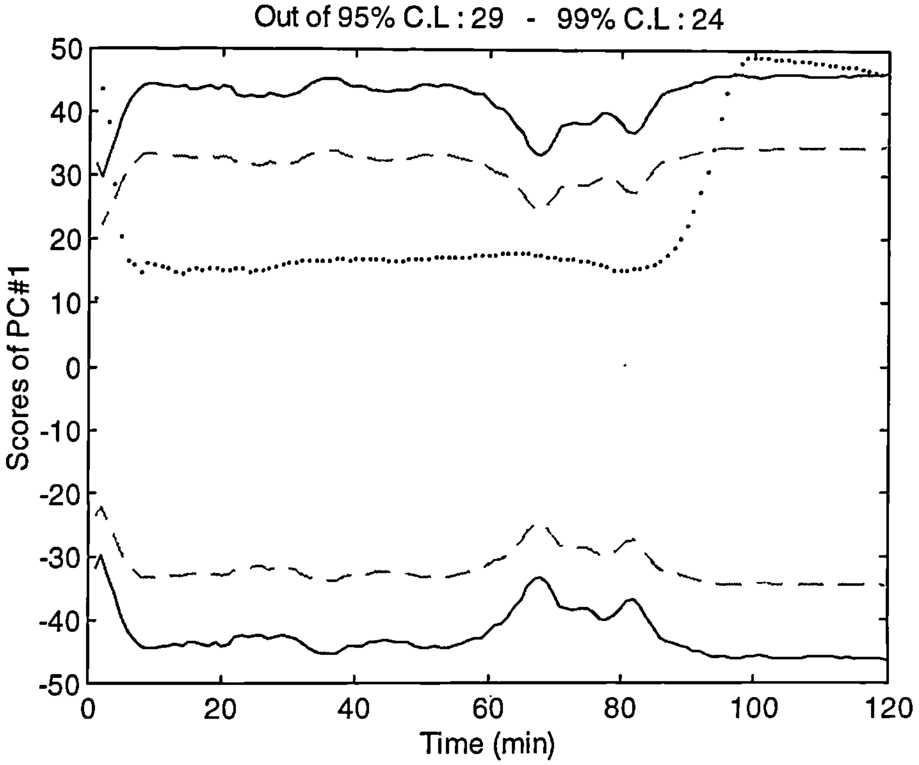


Figure 5.52 On-line monitoring of score of the mixed model for batch number 47

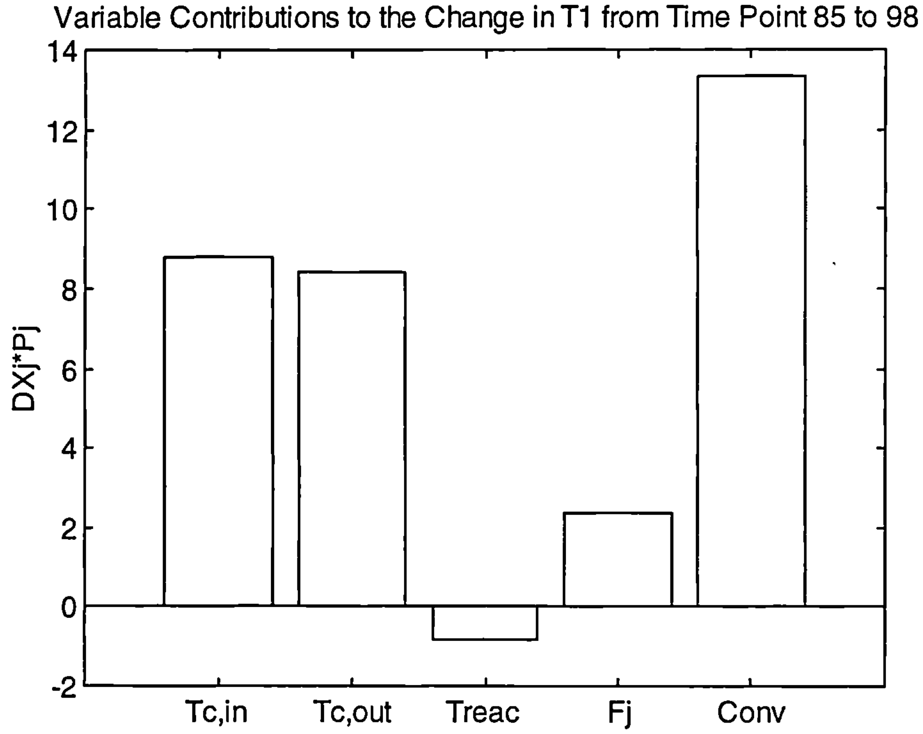


Figure 5.53 Differential contributions to the first score between the 85th and the 98th minute

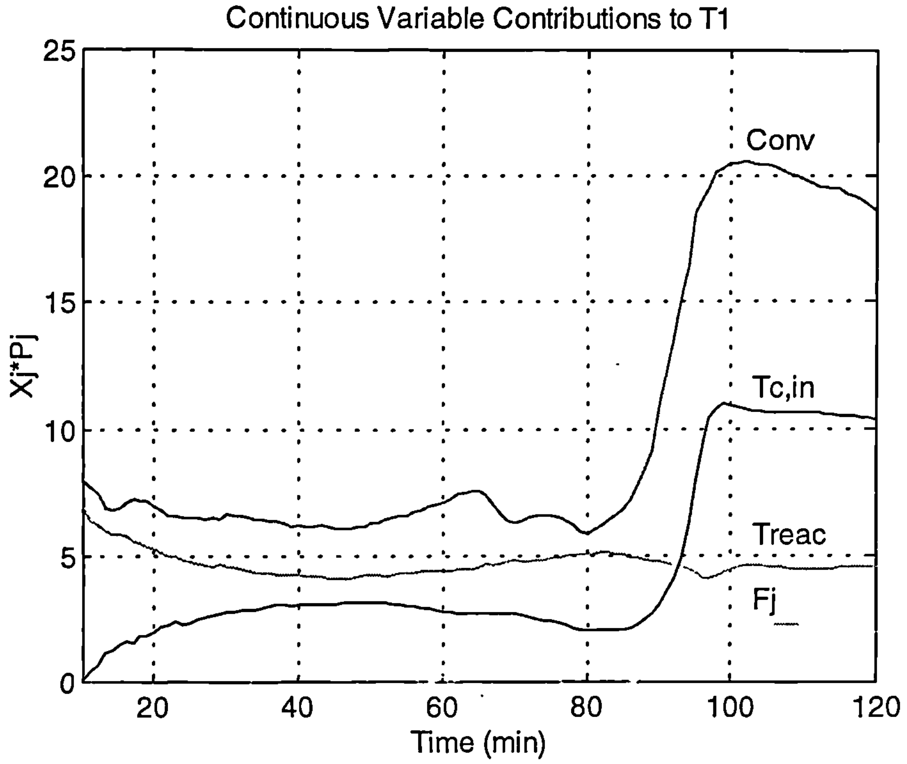


Figure 5.54 Continuous contributions to the first score for batch number 47

5.2.6 MSPC for Batches of Uneven Duration

In most MSPC applications found in the literature, it is assumed that, all batches included in the historical process database, are of the same duration. However, this is unrealistic, since consistent product quality using different initial conditions can be achieved at different operational times, the prescribed recipe is not identically tracked from operation to operation and there are several events that can drive the process away the normal operation, force the control systems to compensate for them and, therefore, delay the termination of process. As a result, batches found in a typical historical database, are of uneven duration.

There are two approaches to overcome this difficulty (Nomikos and MacGregor, 1994). The first approach proposed is to retain only those measurements belonging to the time period that is common for all batches. However, this can leave significantly important information about the process out of the statistical model and, therefore, can lead to the development of spurious models. Alternatively, it was suggested that, batches can be scaled using another process variable, instead of the process time (Rothwell, 1998).

For the case of the batch polymerisation reactor, the process variable used to scale the operation, was the on-line conversion (Conv). For the purposes of illustrating the previous approach, 24 batches were additionally simulated, with initial process conditions randomly selected from the nominal design levels (Table 5.1.). All the 24 batches were allowed to continue up to the time point where a value of conversion of 98% was achieved. Measurements on the remaining process variables were collected at each time point where the conversion was increased at 1%. Figure 5.55 presents the trajectories of the inlet temperature of the coolant ($T_{c,in}$) for the 24 additional batches, whilst Figure 5.56 presents the transformation achieved on these process trajectories when this kind scaling is applied. It can be seen that, the 24 uneven batches have been transformed to equal length.

An MSPC scheme for process monitoring that utilises an MPCA model, such as those described in section 5.2.5.3 can be developed using the transformed historical database of the 24 additional batches. Note that, the conversion (Conv) has been excluded from the process variables, and, *operational time* was included as the fifth variable. Figures 5.57 and 5.58 illustrate the monitoring procedure for another additionally simulated batch with initiator below its nominal design level. It can be seen that, the model is able to detect the occurrence of an abnormal event.

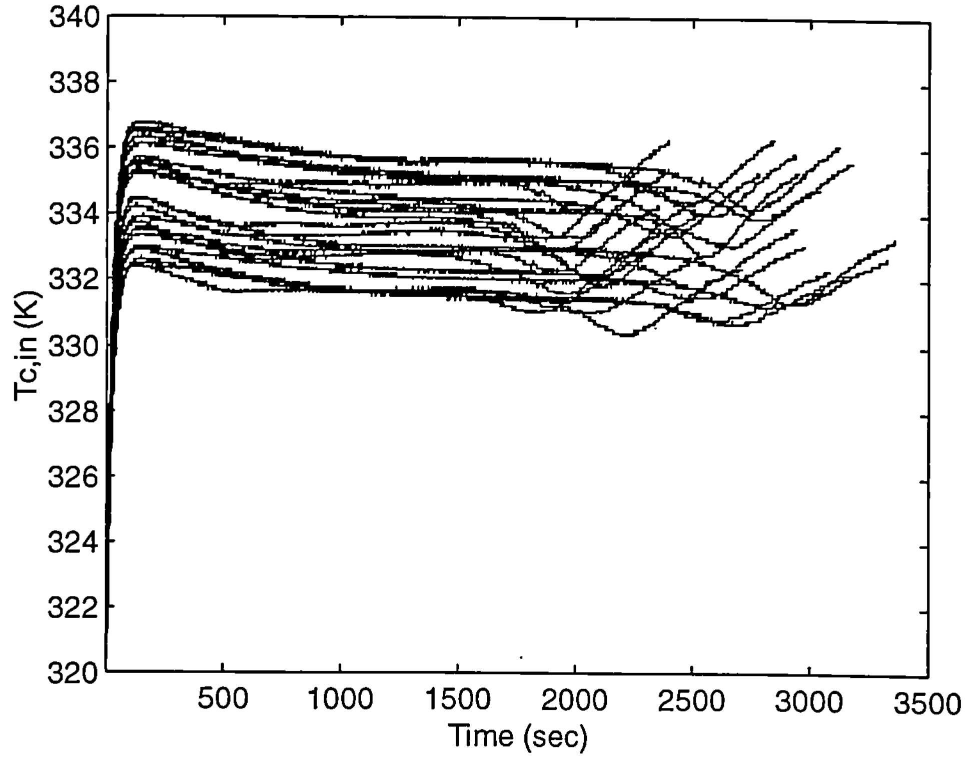


Figure 5.55 Process trajectories on a time scale

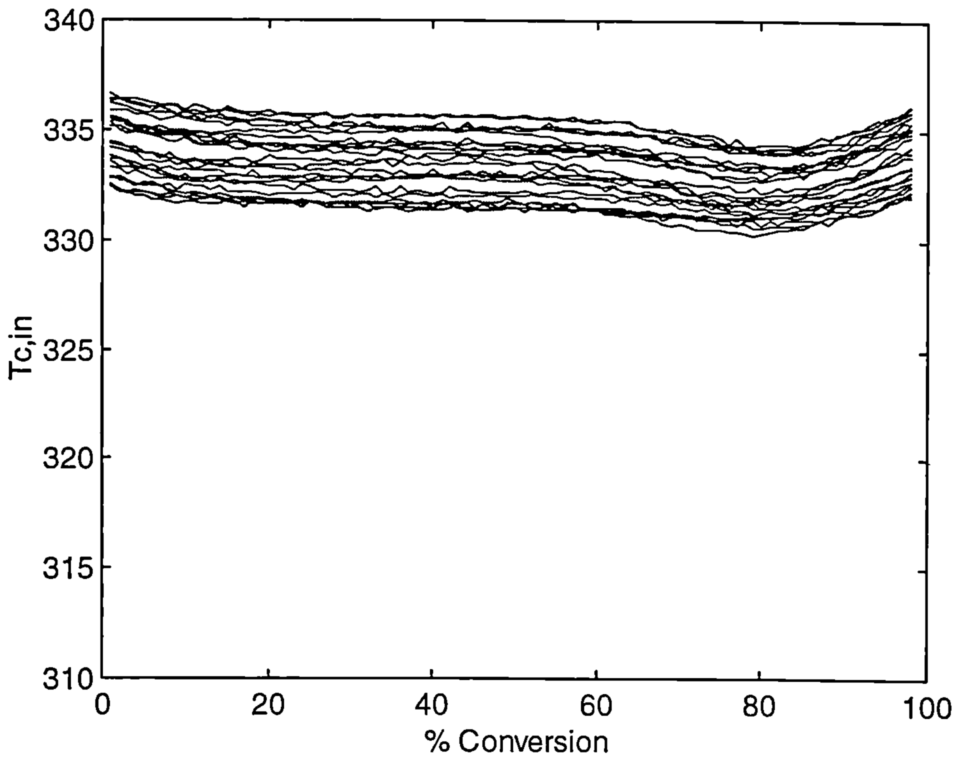


Figure 5.56 Process trajectories on a conversion scale

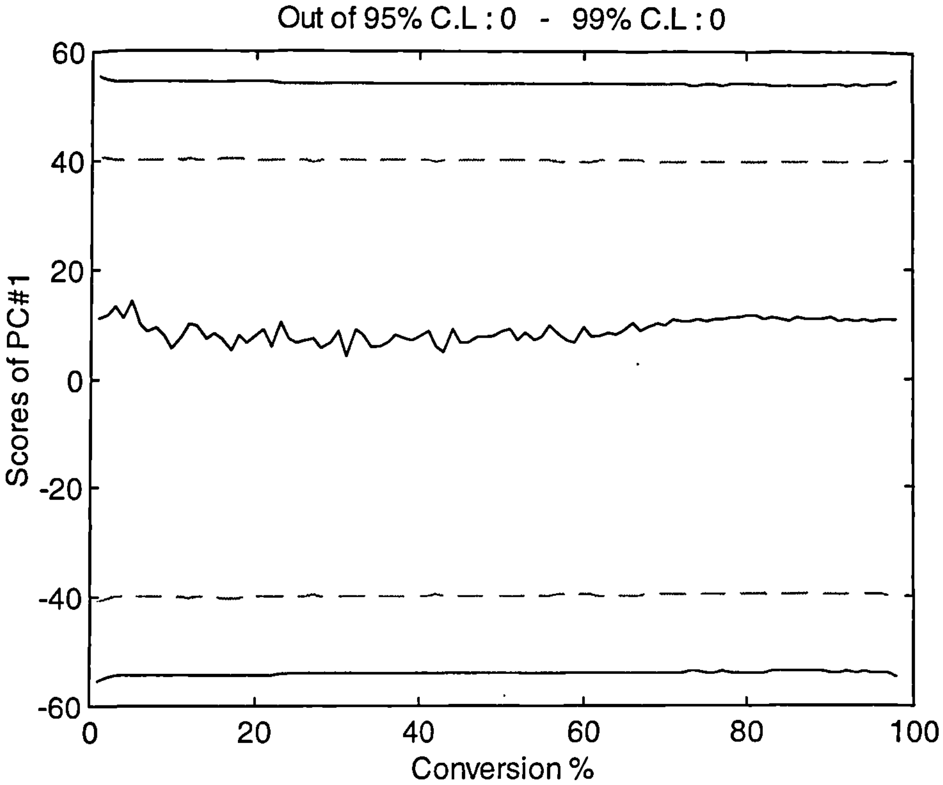


Figure 5.57 On-line monitoring of the first process score for the abnormal batch

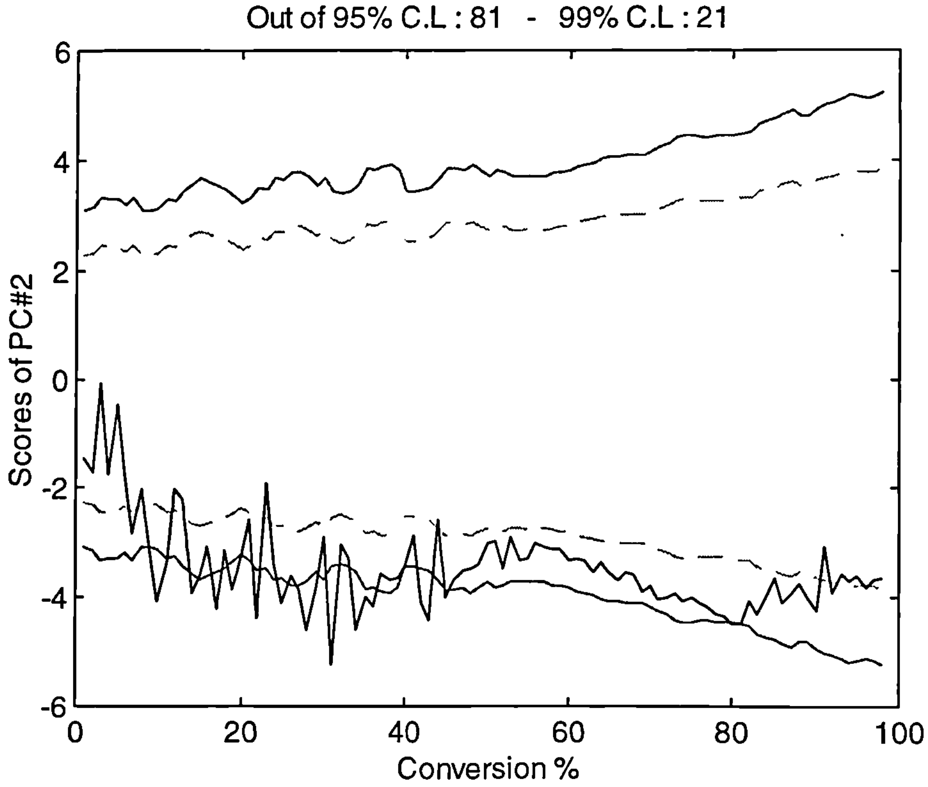


Figure 5.58 On-line monitoring of the second process score for the abnormal batch

However, the fault cannot be identified, since all the process variables contributed to the out-of-control signals of the second principal component score, between the 9% and 44% of conversion.

5.3 MSPC-Based Applications to Continuous Processes

In order to illustrate the advantages of the Multi-Block PLS method in the implementation of MSPC schemes for inter-connected processes, three different statistical models were developed for a two-zone Low Density Poly-Ethylene (LDPE) tubular reactor, using PLS and Multi-Block PLS. Specifically, the first model was developed using the classical PLS method, while the two other models were developed using the Interconnected Multi-Block PLS method, but applying different variable blocking procedures.

The application of Multi-Block PLS to the LDPE reactor has been already presented in MacGregor et al., 1994. However, the LDPE data have been analysed in order to illustrate both the proposed concepts and technique to inter-connected continuous processes.

5.3.1 Two-Zone LDPE Tubular Reactor

Low-density polyethylene (LDPE) is produced at high pressures in tubular and autoclave reactors. A detailed review of the literature, the reaction kinetics and the fundamental modelling of these LDPE processes is presented in Kiparissides et al. (1993). Based on this fundamental study, a steady-state process simulation programme has been developed by the Laboratory of Polymer Reaction Engineering (LPRE), Department of Chemical Engineering, Aristotle University of Thessaloniki, Greece. The simulation programme has been adjusted to match typical data produced by industrial processes. In this thesis only the first two-zones of an industrial tubular LDPE reactor, as depicted in Figure 5.59, are considered. These steady-state data might reasonably represent measurements collected from an industrial process at time intervals longer than the process time constants or averages of measurements taken over some time periods.

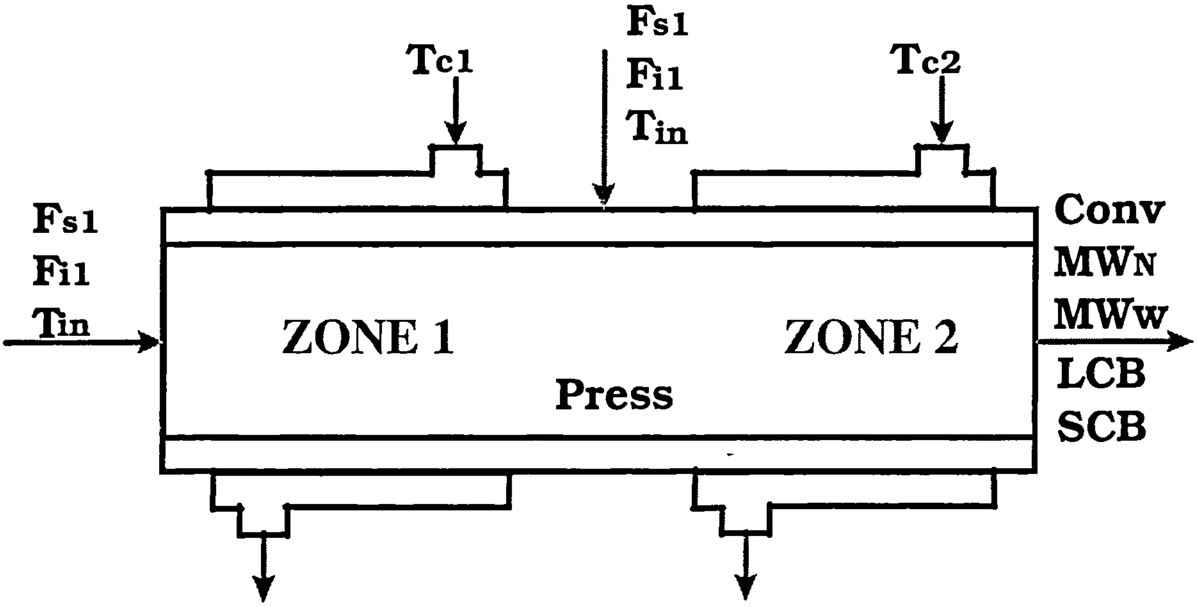


Figure 5.59 Two-zone LDPE tubular reactor

The major productivity variable of interest is the conversion per pass (CONV). The molecular properties of interest include the weight average molecular weight (MW_W), the number average molecular weights (MW_N), and the long-chain branching frequency (LCB) and short-chain branching frequency (SCB). None of these properties are available on-line and many of them are either not measured at all or are only measured infrequently. However, many on-line measurements such as the temperature profile down the reactor, the coolant temperature, and the solvent and initiator flow rates are available on a frequent basis. Although the entire temperature profile is available for each reactor section, the common industrial practice of summarising the profile in each section by its inlet (Tin), maximum (Tmax) and outlet temperatures (Tout), together with the position of the maximum (z) is adopted (MacGregor et al., 1994). Process and quality measurements assumed to be available are listed in Tables 5.9 and 5.10, respectively.

A number of grades of LDPE product can be produced by an industrial reactor, however, the production of only one grade was considered here. The nominal experimental initial conditions for this grade were selected to represent realistic

Process Variables	Definition
T_{IN}	Inlet temperature of the reaction mixture (K)
T_{MAX1}	Maximum temperature of the reaction mixture in the first zone (K)
T_{OUT1}	Outlet temperature of the reaction mixture in the first zone (K)
T_{MAX2}	Maximum temperature of the reaction mixture in the second zone (K)
T_{OUT2}	Outlet temperature of the reaction mixture in the second zone (K)
T_{CIN1}	Inlet temperature of the coolant in the first zone (K)
T_{CIN2}	Inlet temperature of the coolant in the second zone (K)
z_{MAX1}	Position of the reactor where Tmax1 appears (% of reactor length)
z_{MAX2}	Position of the reactor where Tmax2 appears (% of reactor length)
F_{I1}	Total inlet flow-rate of the initiators to the reactor (g/s)
F_{I2}	Total inlet flow-rate of the initiators in the intermediate feed-stream (g/s)
F_{S1}	Inlet flow of the solvent in the reactor (% of ethylene)
F_{S2}	Flow of the solvent in the intermediate feed (% of ethylene)
Press	Pressure of the reactor (atm)

Table 5.9 Process variables of the LDPE reactor (MacGregor et al., 1994)

Quality Variables	Definition
CONV	Cumulative conversion of monomer
MW_N	Number average molecular weight
MW_W	Weight average molecular weight
LCB	Long Chain Branching / 1000 atom C
SCB	Short Chain Branching / 1000 atom C

Table 5.10 Quality variables of the LDPE reactor (MacGregor et al., 1994)

conditions of polymer LDPE production and are given in Table 5.11. The training data set comprises 50 steady-state operations generated through Monte Carlo variation of the selected initial conditions and represents normal production of LDPE when only common causes variations were present and acceptable product quality was achieved. Additionally, two operations were simulated to represent two different

Variables	Range of Variation
F_{S1}	5.95 - 6.05 %
F_{S2}	5.95 - 6.05 %
Press	2,965 - 3,035 atm
T_{IN}	477 - 483 K
Fouling Factor Coefficient	15-25 cal/cm^2/s/K^{-1}
Impurities	15 - 35 % of initiator flow rates
Initiator flow-rate	0.408 - 0.510 g/s

Table 5.11 Process conditions for the reference set (MacGregor et al., 1994)

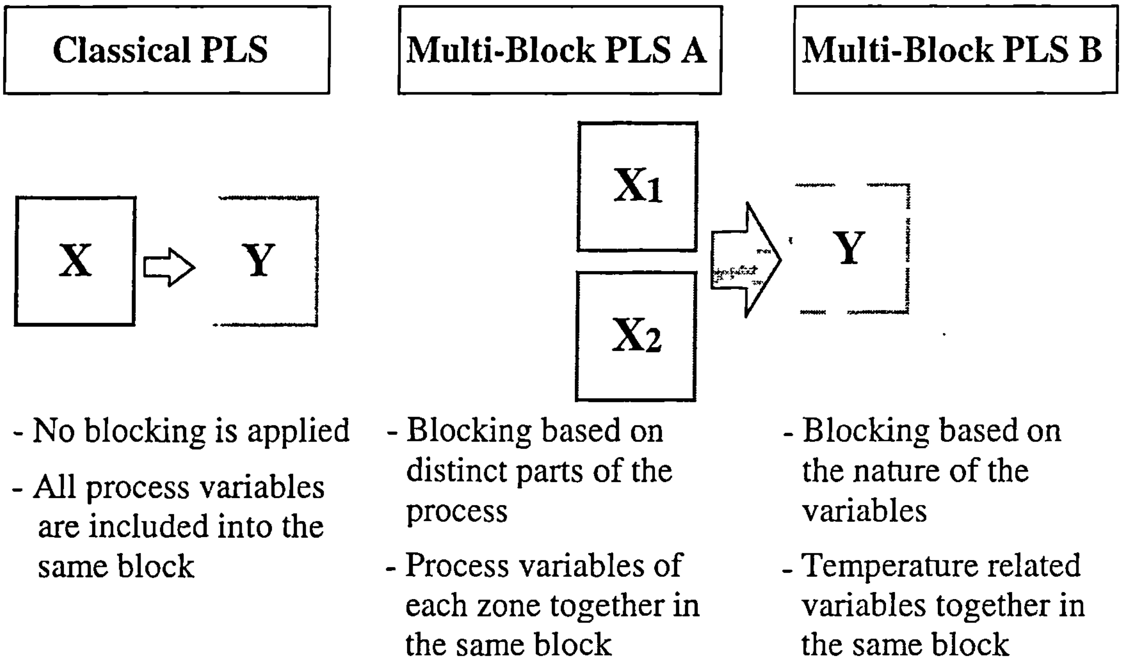
types of abnormal behaviour that can occur, namely, reactor fouling and changes in the amount of chain transfer agents entering with the solvent. More specifically, the first operation represented a reactor fouling problem occurring in the second zone, whilst the second operation described a problem relative to impurities entering with the feed of solvent in the first zone of the reactor (Kiparissides, 1997).

5.3.2 Description of PLS-based Models

Three statistical models were developed using Projection to Latent Structures (PLS) and the Interconnected Multi-Block PLS techniques, in order to establish MSPC-based schemes for the monitoring of the LDPE process.

The first model was build using the ordinary or *classical* PLS technique. All process variables (Table 5.9), were included in the predictor (X) data set, while the response (Y) data set consisted of the quality variables of interest (Table 5.10). The second model was build using the Interconnected Multi-Block PLS technique (MBPLS-A). Process variables were grouped according to their origin and location in the process and two different process data sets (X_1 and X_2) were created. Variables associated with the first zone and the second zone of the reactor were included in X_1 and X_2, respectively. Process variables common to both zones were included into both data sets (Pressure (Press) and temperature of reaction mixture leaving first zone (T_{OUT1}), which enters the second zone). Finally, the third statistical model was build using the Interconnected Multi-Block PLS method (MBPLS-B), but based upon another grouping approach. Variables were grouped according to their similarity and nature. Process variables relating to temperature were included in X_1, while the rest of the variables were placed in X_2.

Figure 5.60 summarises the three developed models and the variability explained by each of these models is presented in Table 5.12. It can be seen that, the first dimensions in each model need to summarise the variability of the process. Furthermore, PLS and MBPLS-B models explain the same amount of variability in Y block, this is an indication that poor blocking can lead to insufficient statistical modelling. It has to be stated that, the number of latent dimensions extracted in each model depends on the rank of the matrices involved in the development of the model. In all models, three latent dimensions, explaining almost 99% of the variability in Y block, were kept. Finally, only the plots indicative of the issues investigated are presented.



Classical PLS:
- No blocking is applied
- All process variables are included into the same block

Multi-Block PLS A:
- Blocking based on distinct parts of the process
- Process variables of each zone together in the same block

Multi-Block PLS B:
- Blocking based on the nature of the variables
- Temperature related variables together in the same block

Figure 5.60 Statistical models developed for the LDPE process

	PLS				MPLS-A						MPLS-B					
l.v	*X*	*total*	*Y*	*total*	*X_1*	*total*	*X_2*	*total*	*Y*	*total*	*X_1*	*total*	*X_2*	*total*	*Y*	*total*
1	25.4	25.4	72.5	72.5	39.4	39.4	38.1	38.1	72.1	72.1	29.4	29.4	25.5	25.5	72.5	72.5
2	20.1	45.5	22.3	94.8	16.6	56.0	17.1	55.2	21.5	93.6	27.3	56.7	14.8	40.3	22.3	94.8
3	13.2	58.8	4.0	98.8	21.5	77.5	19.9	75.1	4.9	98.5	18.1	74.8	19.4	59.7	4.1	98.9
4	9.8	68.6	1.1	99.9	11.5	89.0	13.4	88.5	1.4	99.9	14.6	89.4	23.8	83.5	1.0	99.9
5	10.4	79.0	0.1	100	2.2	91.2	9.7	98.2	0.1	100	9.5	98.9	16.5	100	0.1	100
6	8.8	87.8	0.0	100	8.7	99.9	1.7	99.9	0.0	100						
7	4.6	92.4	0.0	100	0.1	100	0.1	100	0.0	100						
8	4.4	96.8	0.0	100	0.0	100	0.0	100	0.0	100						
9	1.5	98.3	0.0	100												
10	1.6	99.9	0.0	100												
11	0.1	100	0.0	100												

Table 5.12 Variability explained by the statistical models

5.3.2.1 Classical PLS Model

The statistical representation of the process by the PLS model is presented in Figures 5.61 - 5.64. Figures 5.61 and 5.62 present the projection of the scores of the process to the reduced space defined by the first versus the second and the first versus the third latent dimension, respectively. Figure 5.63 shows the Square Prediction Error of the process variables and, finally, Figure 5.64 shows the linear internal relationship between the two blocks of variables. It can be seen that, the assumption of linear relationship between the X and Y blocks is valid.

The developed PLS model is now used to establish an MSPC-based scheme for process performance monitoring. The scheme is validated against the data sets that represent a reactor fouling problem occurring in the second reactor zone and a problem with the solvent feed in the first reaction zone. The fouling problem is detected in the latent subspace of first and the third latent variables (Figure 5.65) and specifically in the third latent variable. The differential contribution of the process variables to the third latent variable (Figure 5.66), at the particular time points where the fault was detected, indicate that the major contributing variables are related to temperature, exactly what one might expect in a fouled reactor (MacGregor et al., 1994; Kiparissides, 1997). However, the location of the fault cannot be identified, since the contributing variables belong to both reactor zones and, as a result, it can be concluded that, both reactor zones are subject to fouling. The solvent problem is detected in the SPE plot (Figure 5.67). It can be seen that, variables contributing in the increased prediction error are the solvent feed flow rates in both reactor zones (Figure 5.68) and, therefore, one may conclude that impurities have entered both reactor zones. As it was shown, an MSPC-based scheme developed using the PLS

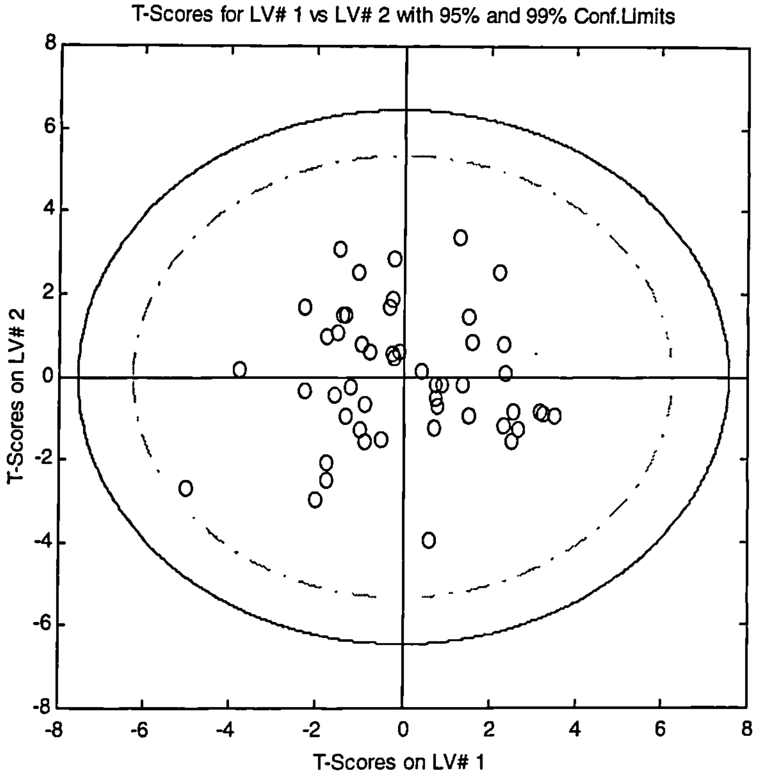


Figure 5.61 Scores plot of process variables on the first and the second latent dimension

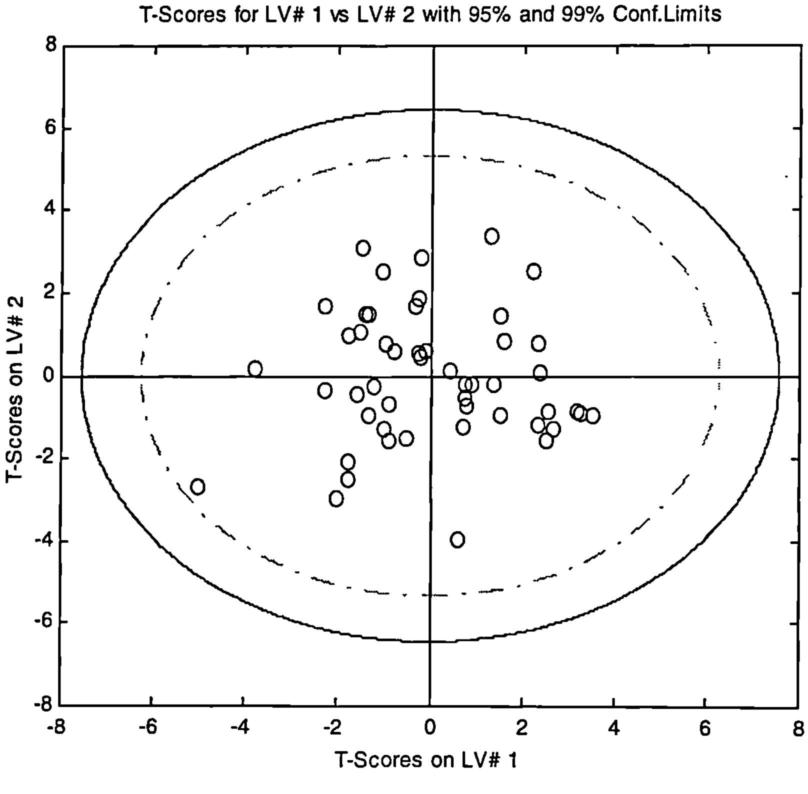


Figure 5.62 Scores plot of process variables on the first and the third latent dimension

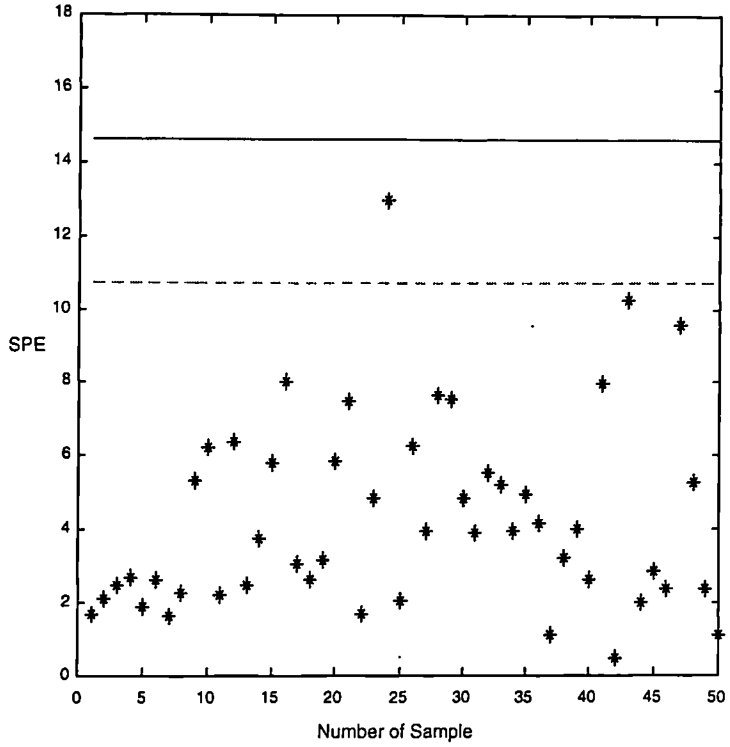


Figure 5.63 SPE plot of the process variables

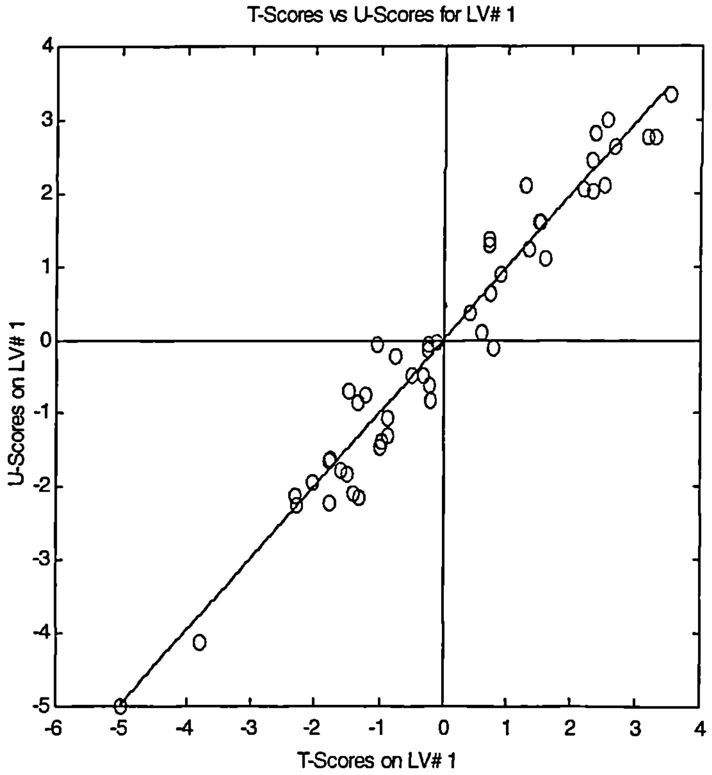


Figure 5.64 Linear internal relationship

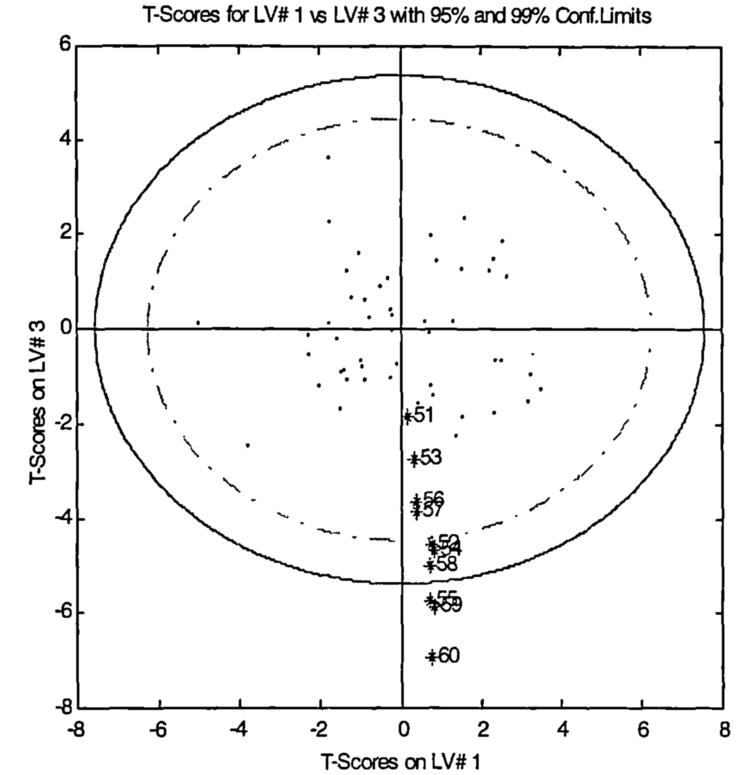


Figure 5.65 Bivariate scores plot detecting the fouling problem

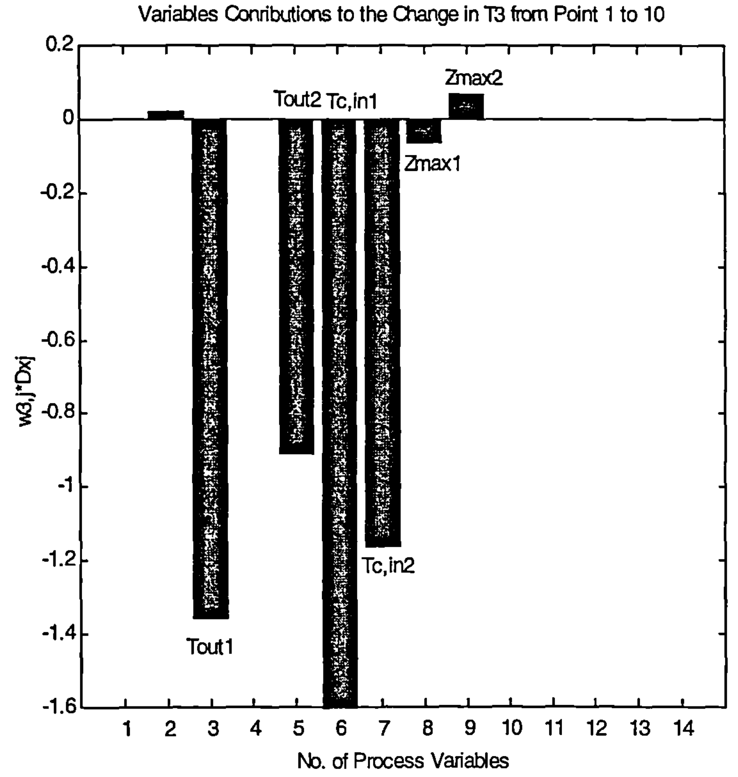


Figure 5.66 Plot of differential contributions to the third latent variable

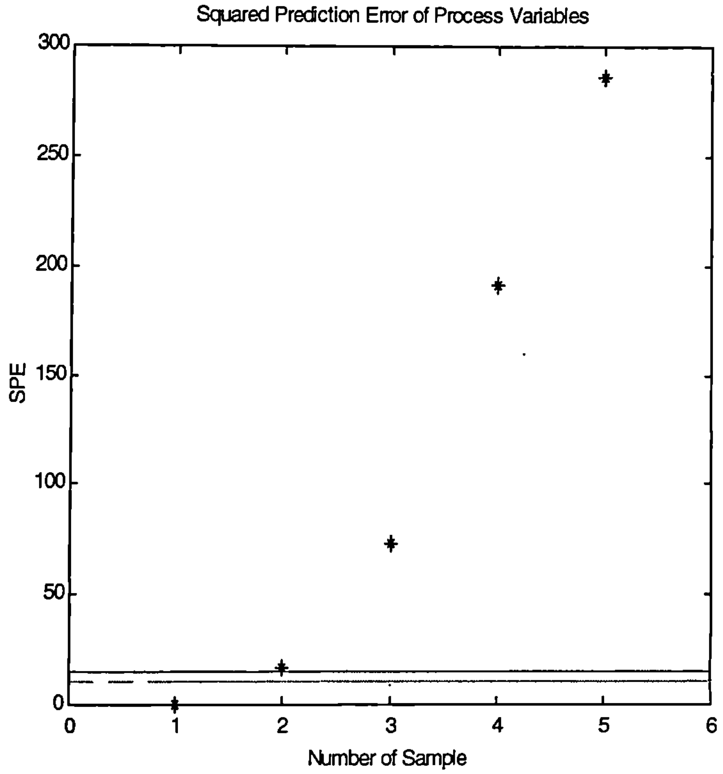


Figure 5.67 SPE plot detecting the problem in the solvent feed

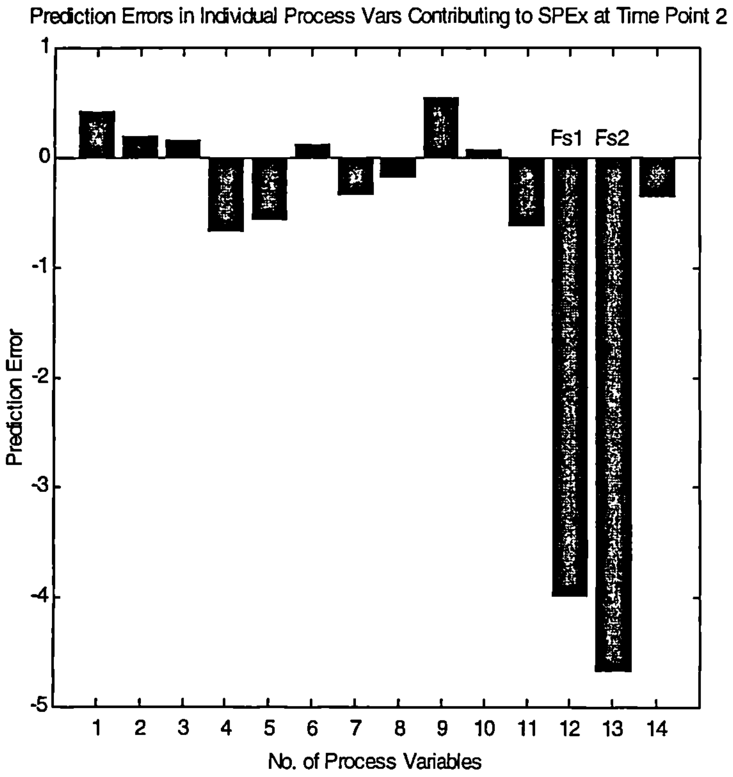


Figure 5.68 Plot of contributions to SPE

model, is able to monitor the process and to detect and diagnose faults. However, the PLS model of the LDPE process is not able to isolate the origin of the occurring faults. The variables responsible for the faults, identified in the contribution plots, belong to both reactor zones.

5.3.2.2 MBPLS model (A) - Zone blocking

The statistical representation of the process by the MBPLS-A model is presented in Figures 5.69 - 5.72. Figures 5.69 and 5.70 present the projection of the process onto the reduced space defined by the first and the second latent dimension for the two process blocks, respectively. Figure 5.71 shows the projection of the scores of the composite matrix (T) created by the two process blocks, onto the reduced space defined by the first two latent dimensions. Figure 5.72 presents the internal relationship between the scores of variables included in the two process blocks, as represented by the composite matrix (T), and the scores of the quality block (U). It can be seen that, the assumption of linear internal relationship is valid. Finally, Figures 5.73 and 5.74 illustrate the Square Prediction Error for each process block.

The developed MBPLS model was used to establish an MSPC-based monitoring scheme which was then validated against the two data sets comprising process faults. Figures 5.75 and 5.76 present the scores for the data set where a fouling problem is known to have occurred in the second zone. Figure 5.75 presents the scores plot of the process variables included in the first block, which corresponds to the first reactor zone. It can be seen that, although there is a trend in the plot of scores, the process is still well in-control in the first block. The fouling problem is only detected in the latent subspace of first and the third latent variables of the second block, which

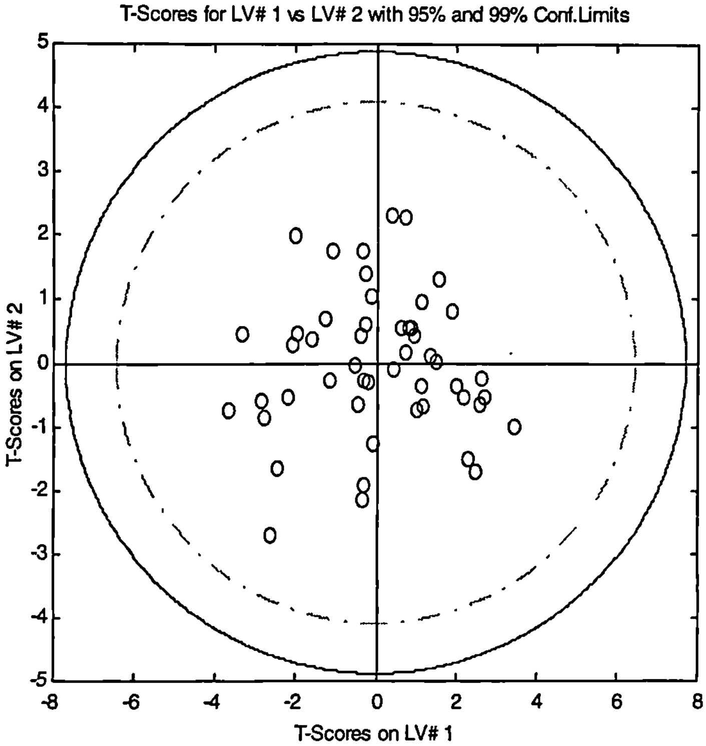


Figure 5.69 Scores plot of process variables of the first block projected down onto the first and the second latent dimension

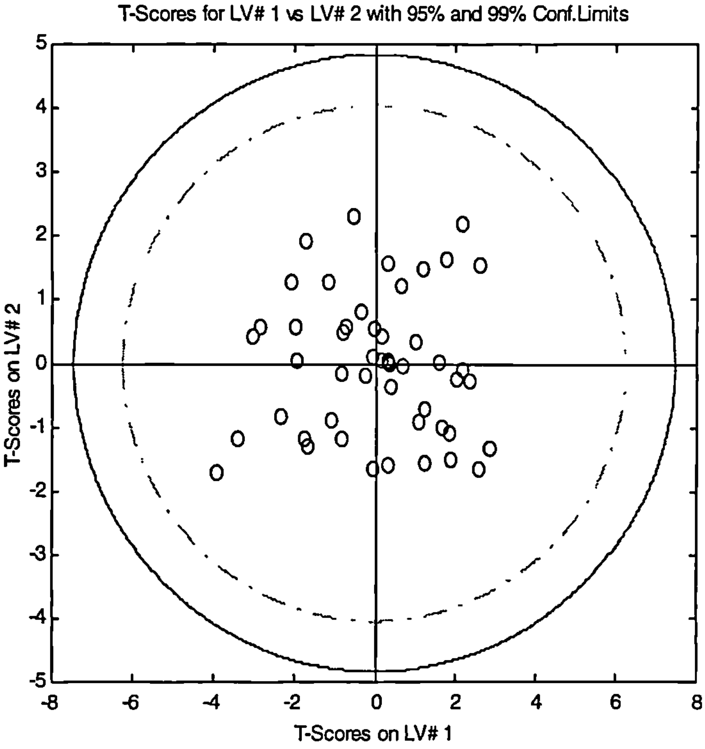


Figure 5.70 Scores plot of process variables of the second block projected down onto the first and the second latent dimension

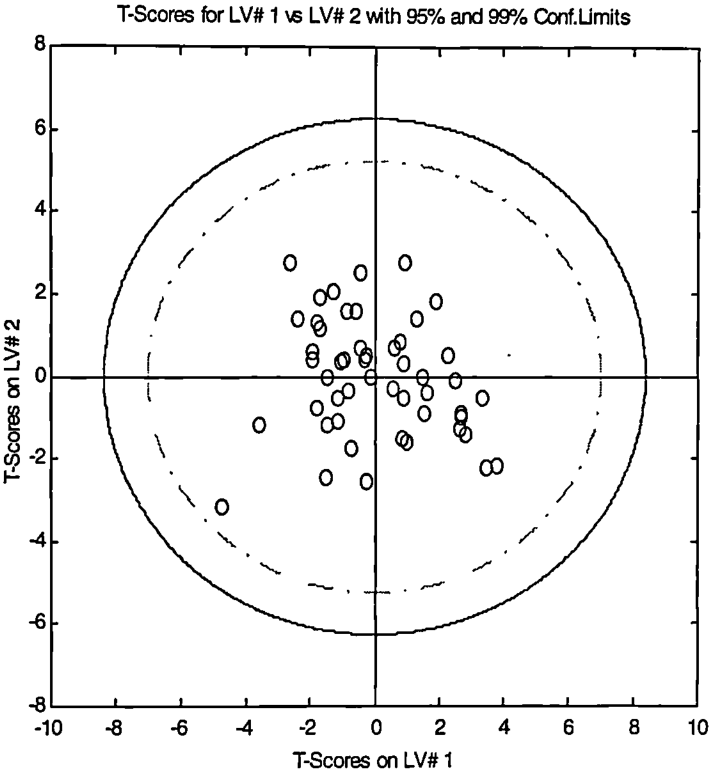


Figure 5.71 Scores plot of the composite matrix on the first two latent dimensions

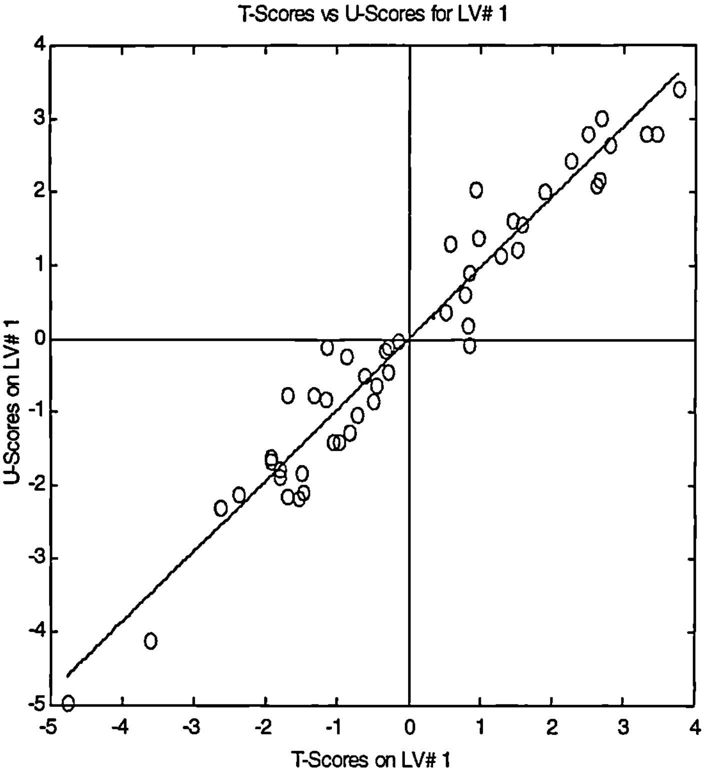


Figure 5.72 Linear internal relationship between the composite matrix (T) and the matrix of scores U

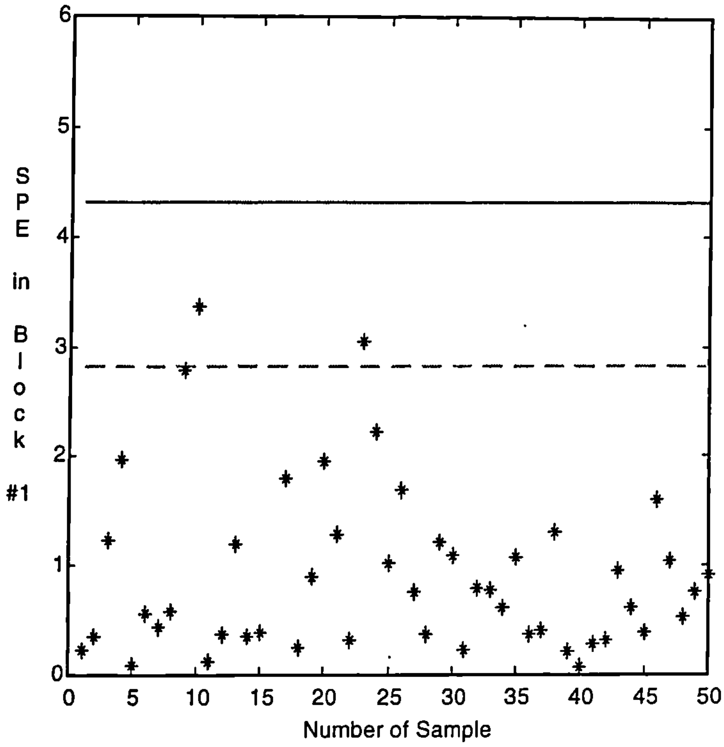


Figure 5.73 SPE of process variables included in the first block

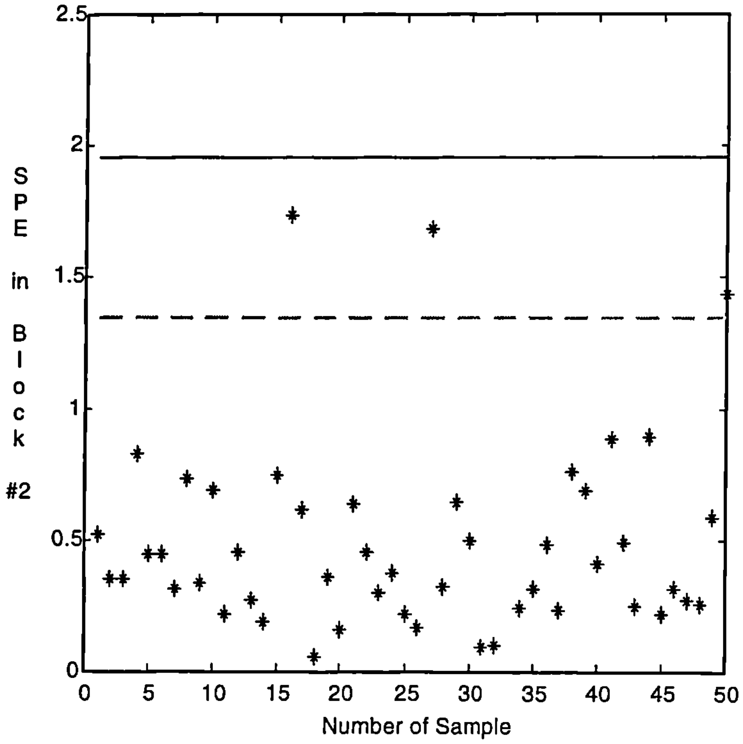


Figure 5.74 SPE of process variables included in the second block

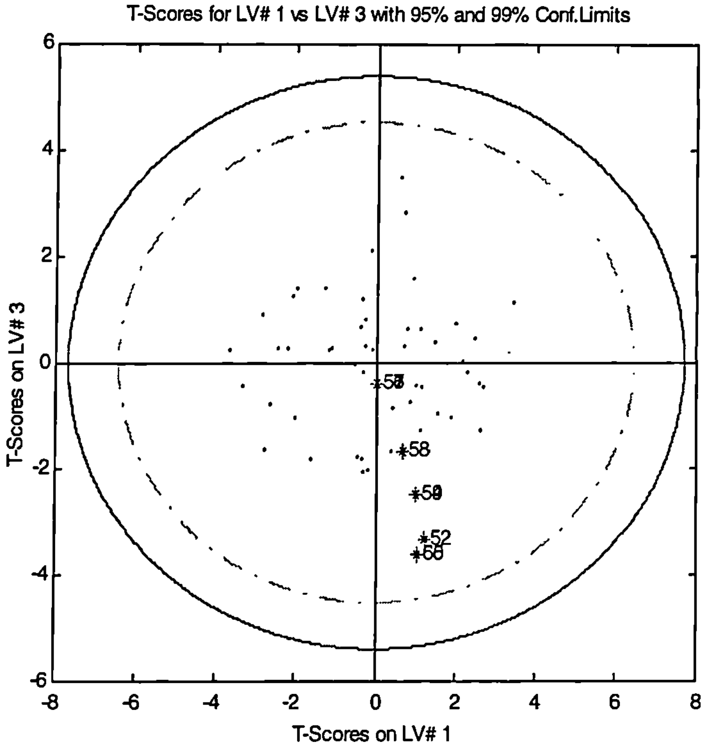


Figure 5.75 Scores plot of process variables included in the first block, on the first and the third latent dimension

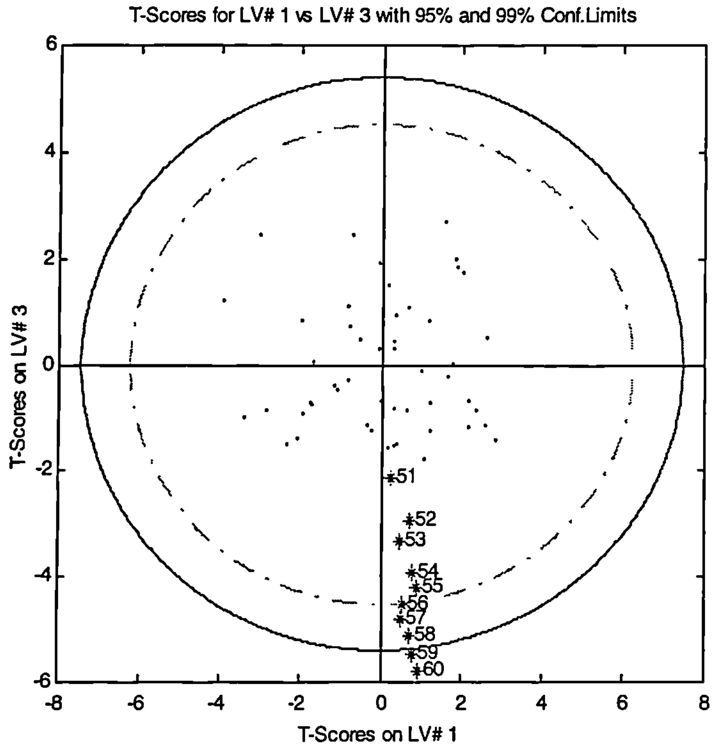


Figure 5.76 Scores plot of process variables included in the second block, on the first and the third latent dimension

corresponds to the second reactor zone (Figure 5.76). The differential contribution of the process variables of the second zone identify the variables which are indicative of the problem, i.e. T_{MAX1}, $T_{C,IN2}$, Z_{MAX2} (Figure 5.77). Finally, Figure 5.78 illustrates the scores plot of the composite matrix T. The fault is identified in this plot as well, since the composite matrix is the combined representation of the two process blocks. However, it is not so sensitive as the score plot of the second block, since the consensus scores are averages of the scores of the individual process blocks.

Figures 5.79 and 5.80 present the scores for the data set where the solvent problem occurred. The solvent problem is detected in the latent subspace of the first and the second latent dimension of the first block, which corresponds to the first reactor zone (Figure 5.79). Figure 5.80 presents the scores plot of the process variables included in the second block, which corresponds to the second reactor zone. It can be seen that, the process is still well in-control. The differential contribution of the process variables identify that the variable indicative of the problem is the inlet flow rate of the solvent in the first zone (Figure 5.81). Finally, Figure 5.82 illustrates the scores plot of the composite matrix. The fault cannot be clearly identified but there is a strong trend that will eventually force the process to move outside the in-control region of operation.

As it was shown, an MSPC-based scheme developed using the MBPLS-A model is able to assist process operators in detecting a fault and, furthermore, in identifying the origin and location of the problem.

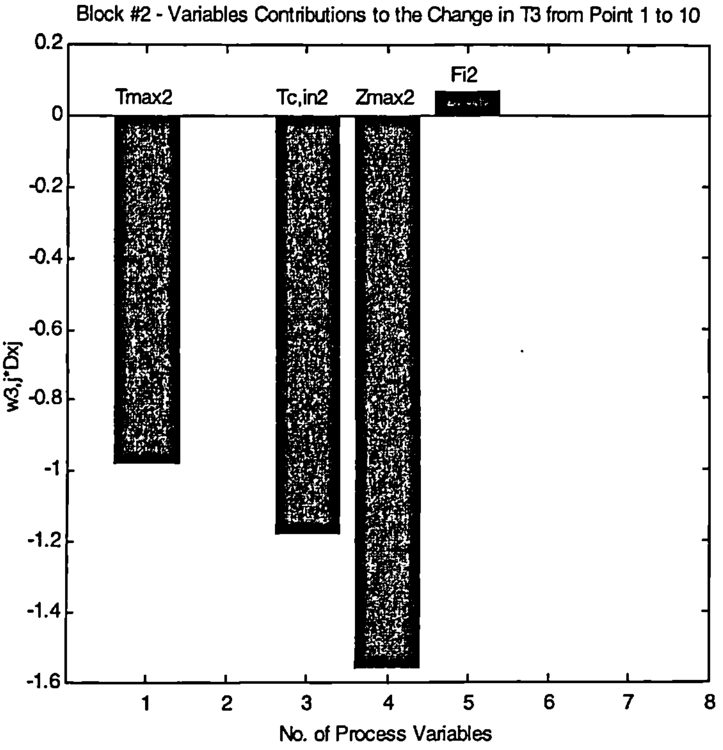


Figure 5.77 Contributions of process variables included in the second block to the third latent variable

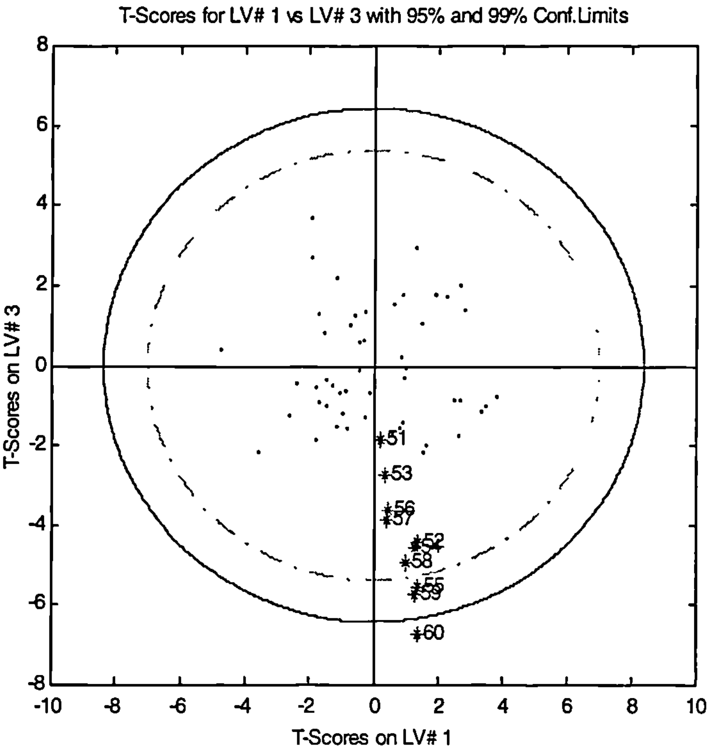


Figure 5.78 Scores plot of the composite matrix on the first and the third latent dimension

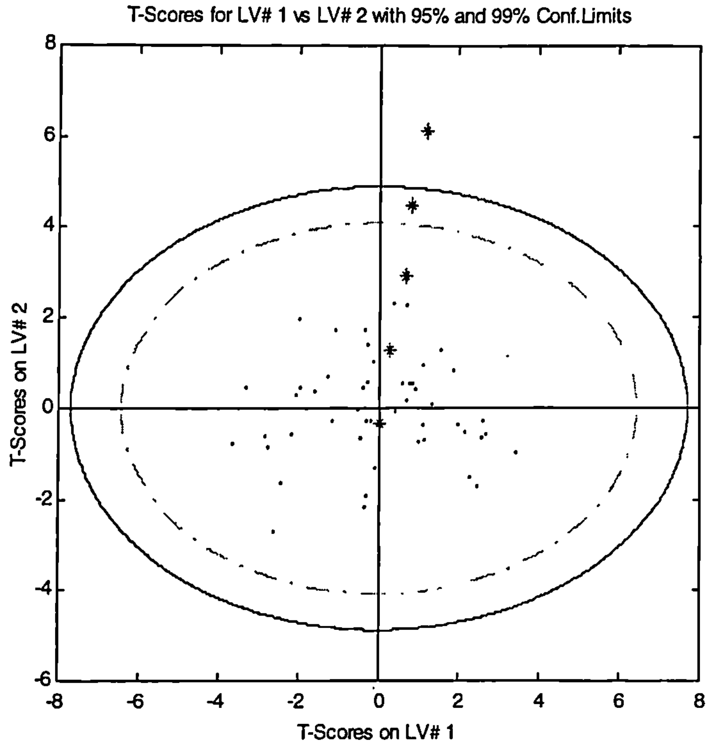


Figure 5.79 Scores plot of process variables included in the first block, on the first and the second latent dimension

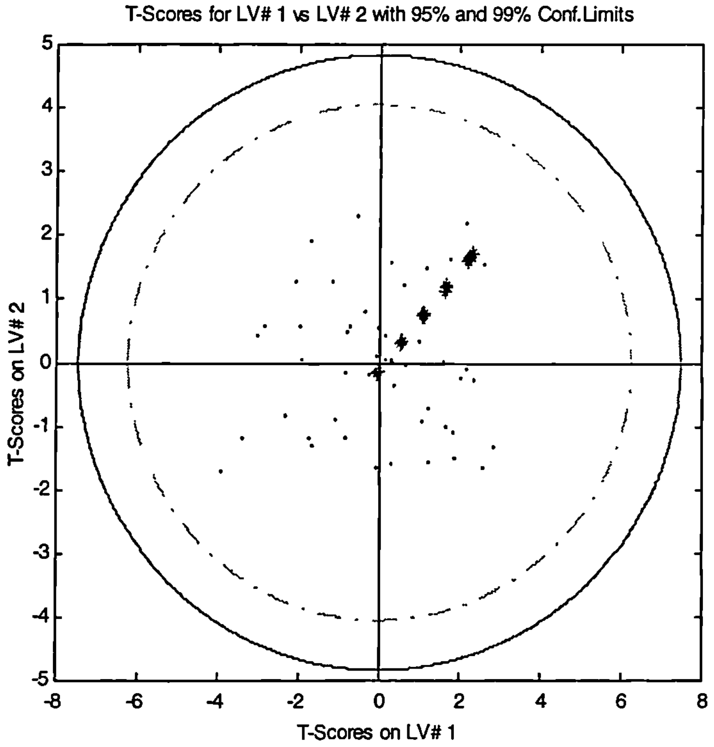


Figure 5.80 Scores plot of process variables included in the second block, on the first and the second latent dimension

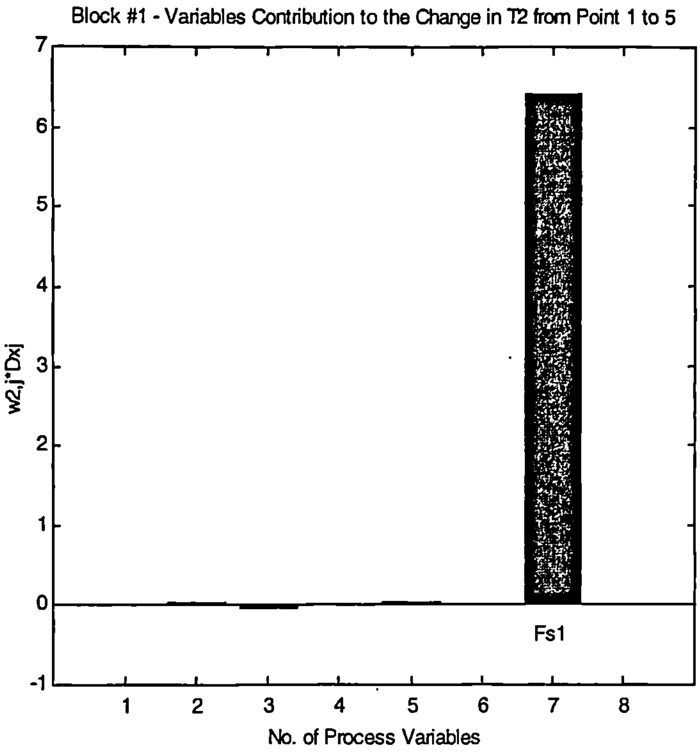


Figure 5.81 Contributions of process variables included in the first block to SPE

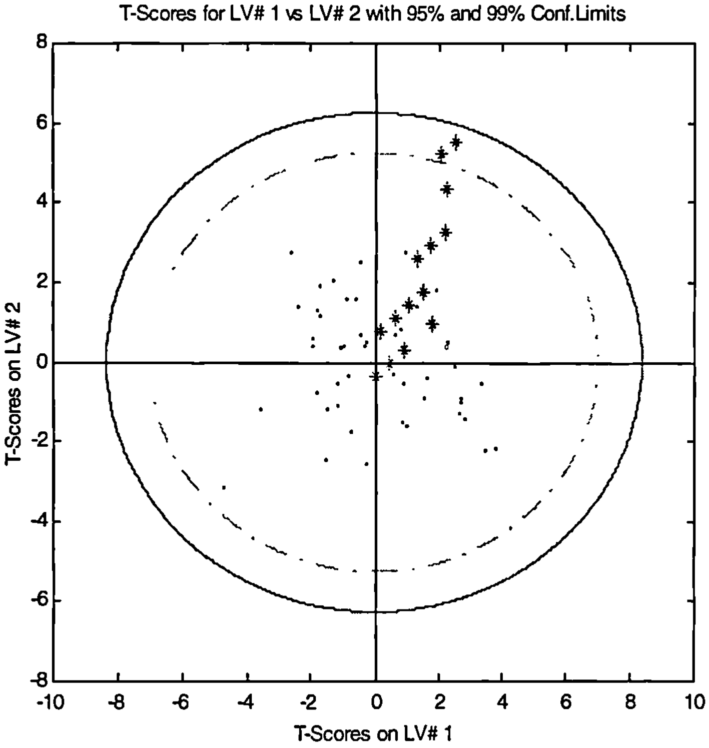


Figure 5.82 Scores plot of the composite matrix on the first and the second latent dimension

5.3.2.3 MBPLS model (B) - Similarity blocking

The statistical representation of the process using the MBPLS-B model is presented in Figures 5.83 - 5.88. Figures 5.83 and 5.84 present the projection of the process down onto the reduced space defined by the first two latent dimensions for the two process blocks, respectively. Figure 5.85 shows the projection of the scores of the composite matrix, created by the two process blocks, to the reduced space defined by the first and the second latent dimensions. Figure 5.86 presents the linear internal relationship between the two blocks of process variables as they represented by the scores of the composite matrix (T) and the scores of the quality variables (U). The assumption of linear internal relationship is valid. Finally, Figures 5.87 and 5.88 illustrate the Square Prediction Error included in the first and the second block, respectively.

The developed MBPLS model was used to establish an MSPC-based scheme, which was validated against the two data sets comprising process faults. The fouling problem in the second zone is detected in the score plot of the first dimension of the first process block (Figure 5.89), which corresponds to all temperature related process variables in both reactor zones. Figure 5.90 presents the scores plot of the process variables included in the second block, which corresponds to the rest of the process variables. It can be seen that the process is in-control. The location of the problem cannot be isolated, since the variables which exhibited greater changes than expected, belong in both reactor zones (Figure 5.91). Therefore, it can erroneously concluded that both reactor zones are subjected to fouling. Finally, Figure 5.92 illustrates the scores plot of the composite matrix (T). The fault is identified in this

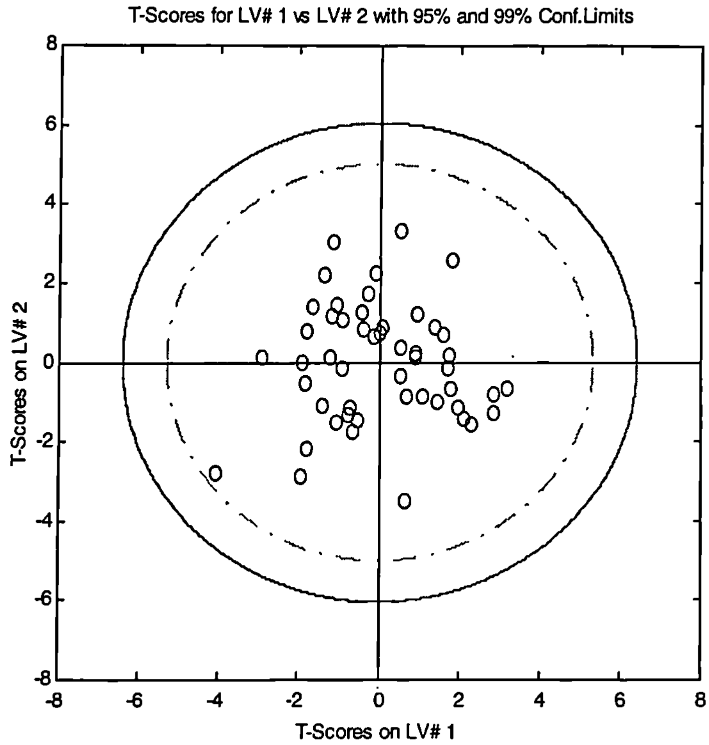


Figure 5.83 Scores plot of process variables of the first block, on the first and the second latent dimension

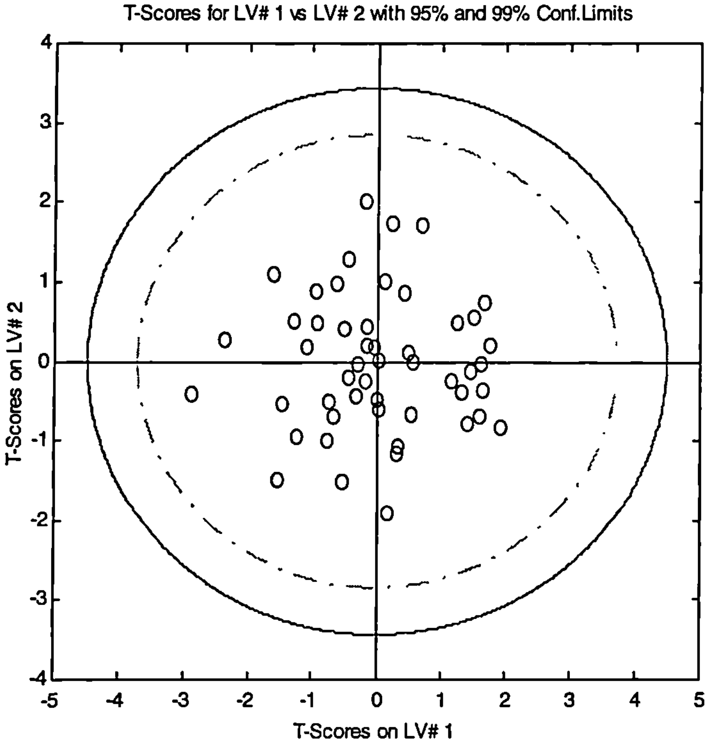


Figure 5.84 Scores plot of process variables of the second block, on the first and the second latent dimension

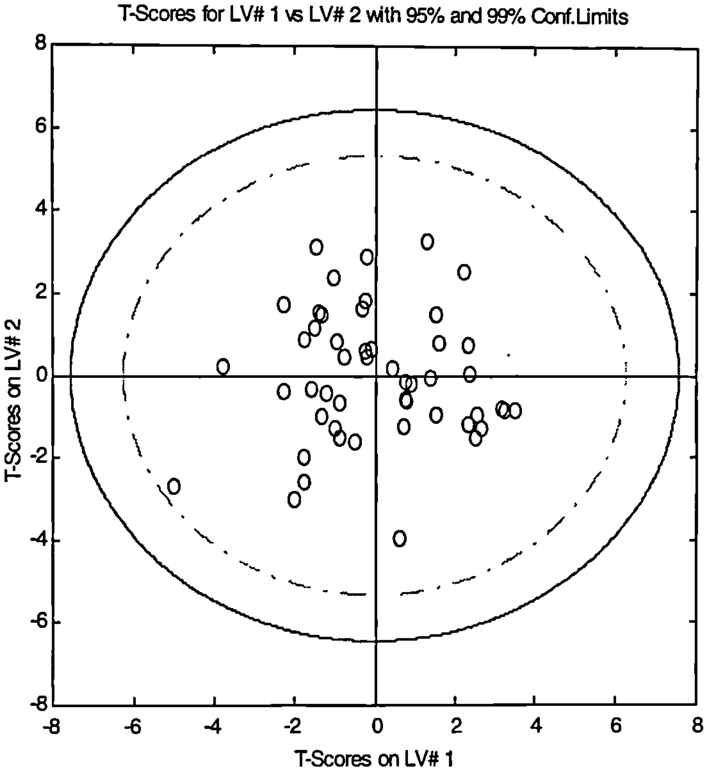


Figure 5.85 Scores plot of the composite matrix on the first two latent dimensions

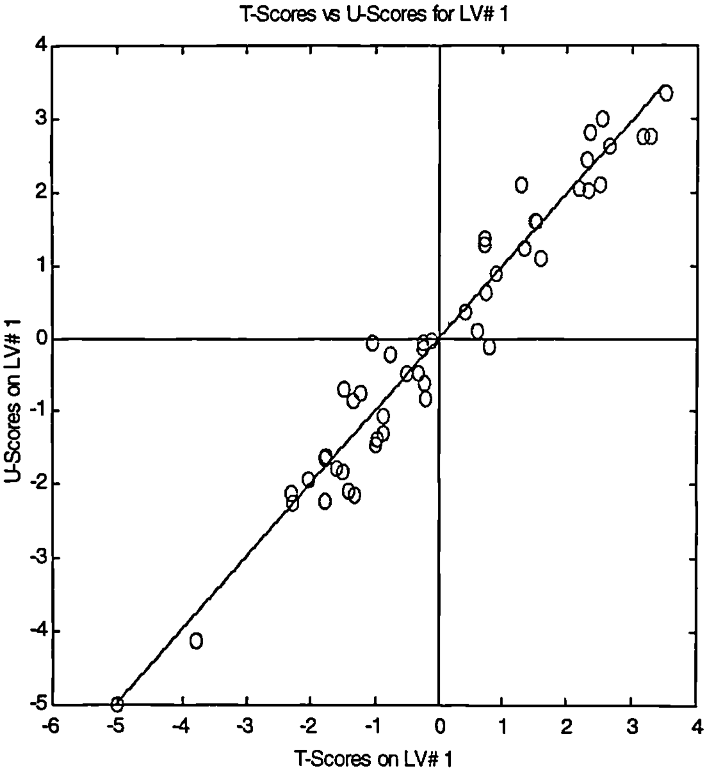


Figure 5.86 Linear internal relationship between the composite matrix (T) and the scores U

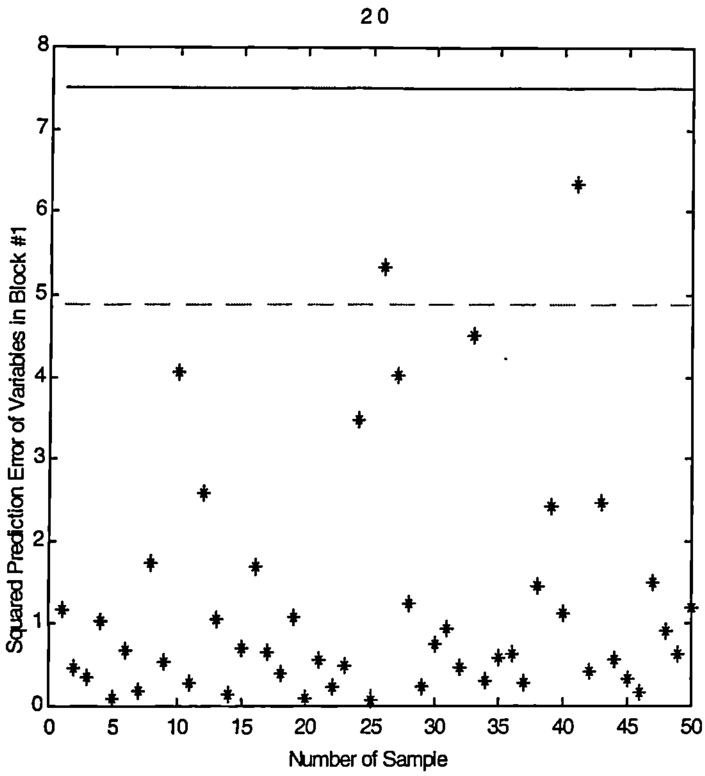


Figure 5.87 SPE of process variables included in the first block

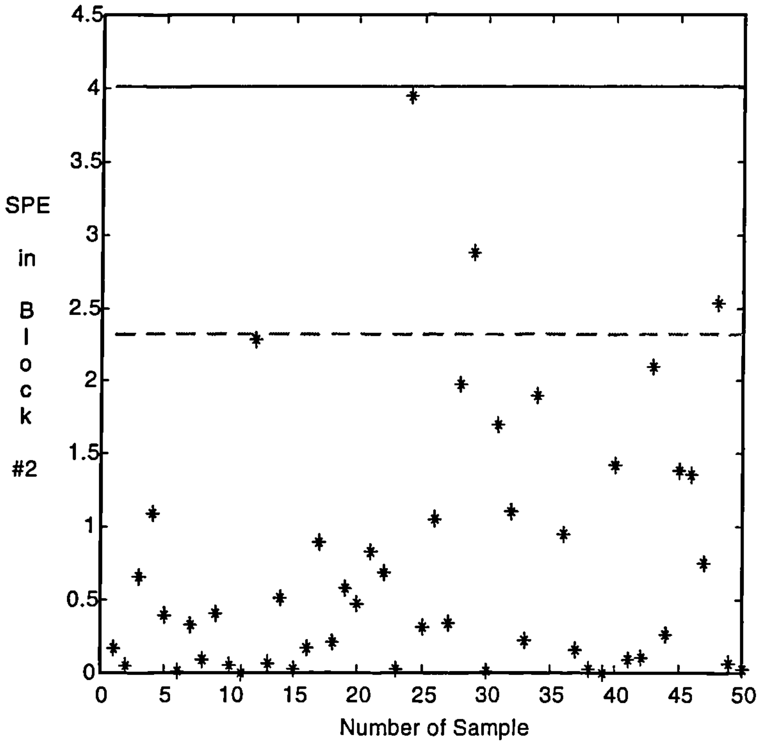


Figure 5.88 SPE of process variables included in the second block

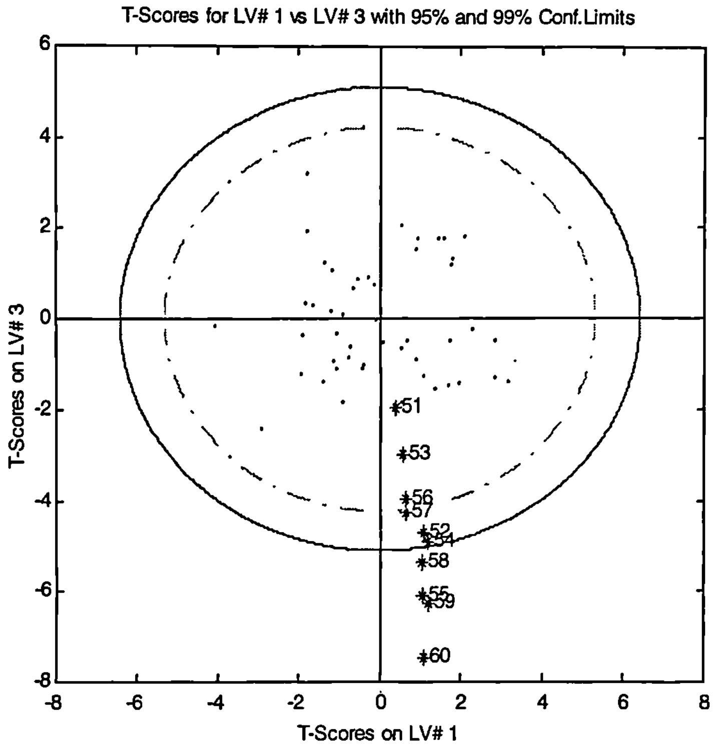


Figure 5.89 Scores plot of process variables included in the first block, on the first and the third latent dimension

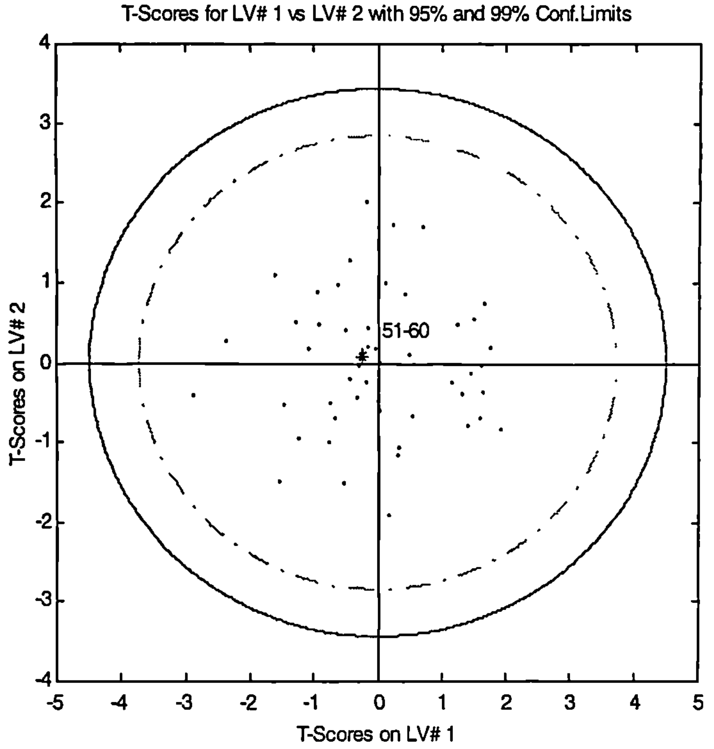


Figure 5.90 Scores plot of process variables included in the second block, on the first and the second latent dimension

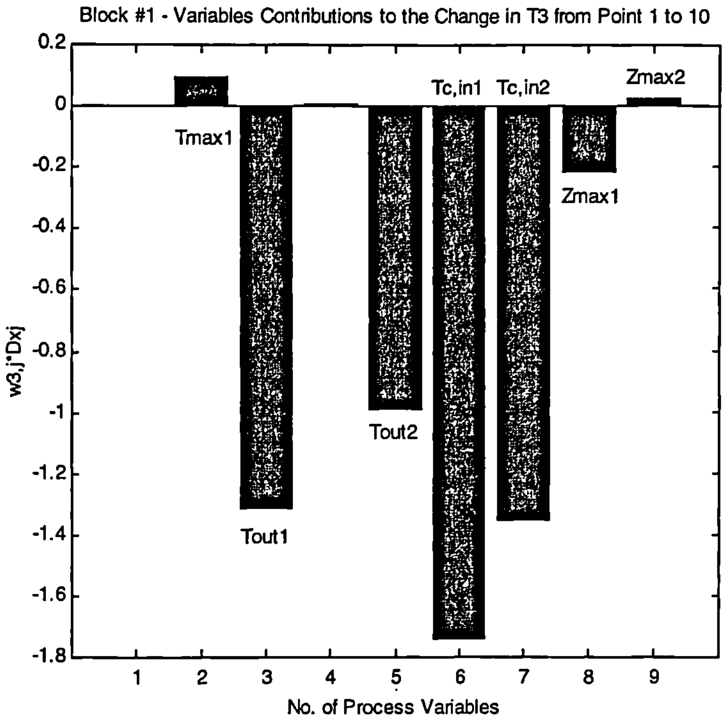


Figure 5.91 Contributions of process variables included in the first block

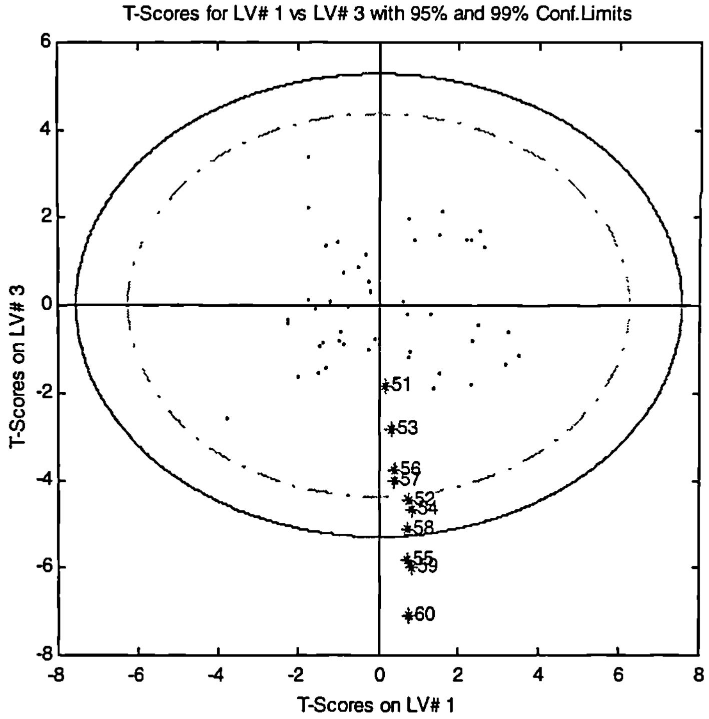


Figure 5.92 Scores plot of the composite matrix on the first and the third latent dimension

plot as well, but again the consensus scores plot is less sensitive than the scores plot of the first block.

The solvent problem in the first zone is detected in the SPE plot of the first block (Figure 5.93). The scores plot of the composite matrix for the first and the second latent dimension cannot detected the problem (Figure 5.94), as well as the scores plot of the individual blocks. The contributing process variables to the increased prediction error are the inlet flow rates of the solvent in the first and the second zones (Figure 5.95) and, as a result, the location of the fault cannot be correctly identified.

As it was shown, an MSPC-based scheme based on an MBPLS model developed using this particular variable blocking approach is able to monitor the process and to detect and diagnose faults. However, this model of the LDPE process is not able to isolate the origin of the occurring faults. The variables responsible for the faults, identified in the contribution plots, belong to both reactor zones. As a result, similarly to the PLS model, although the process operators will able to understand that the reactor is fouled or that impurities have entered the reactor, they will not able to isolate the fault and locate the zone where it occurred.

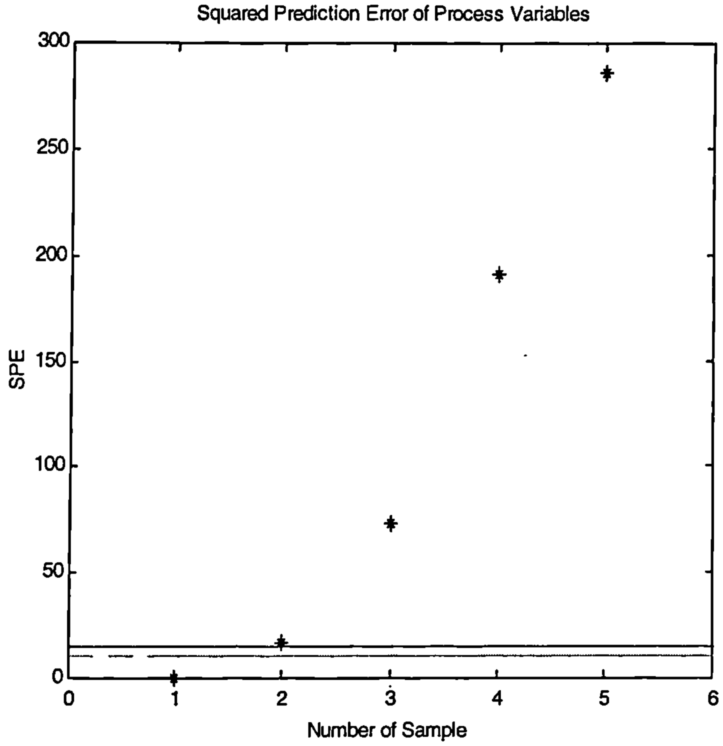


Figure 5.93 SPE of process variables included in the first block

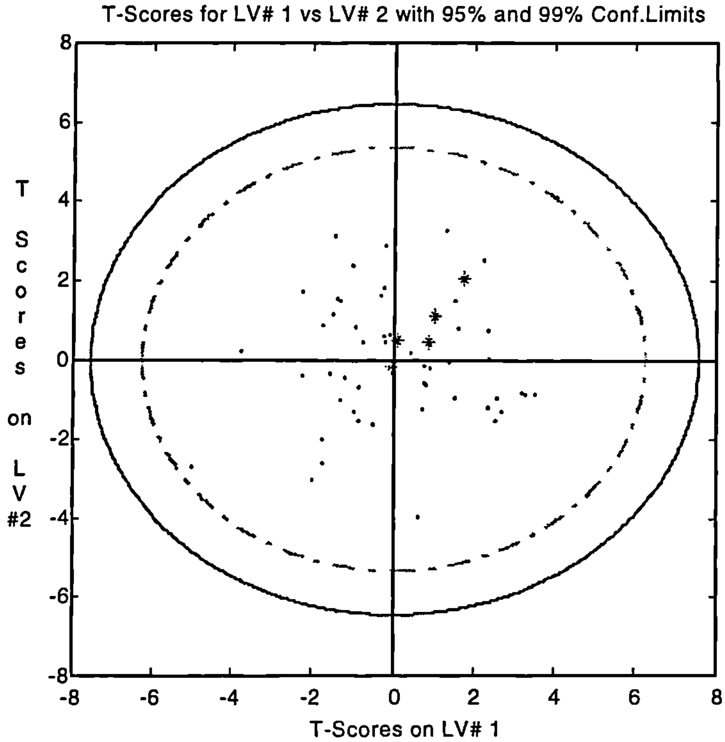


Figure 5.94 Scores plot of the composite matrix on the first and the second latent dimension

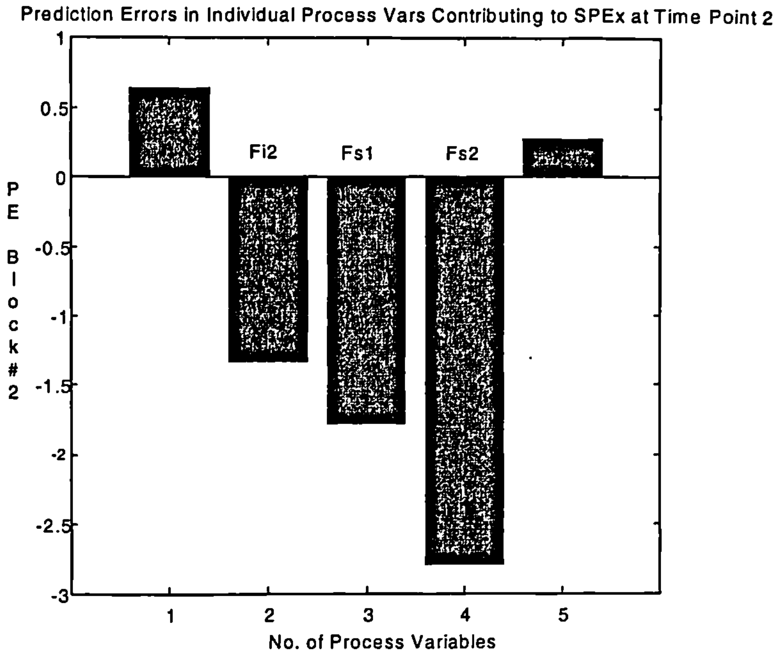


Figure 5.95 Contributions of process variables included in the first block

5.4 Summary

This chapter presented applications of the proposed multivariate statistical projection techniques to the development of MSPC-based schemes for process monitoring, fault detection and diagnosis. Schemes were developed for a batch and a continuous polymerisation process. Additionally, statistical models, that can be used in the general framework of MSPC schemes for these processes, have been presented and related issues were discussed.

Chapter VI

Discussion and Further Work

6.1 Summary and Discussion

The work presented in this thesis forms part of two on-going projects of the European Community: *Intelligent Manufacturing of Polymers* - BRITE EURAM CT 93 0523 (INTELPOL) and *Process Diagnostics for Plant Performance Enhancement* - ESPRIT PROJECT 22281 (PROGNOSIS). These projects are conducted by the Centre for Process Analytics and Control Technology (C.P.A.C.T), University of Newcastle, the Laboratory of Polymer Reaction Engineering (L.P.R.E), Chemical Process Engineering Research Institute (C.P.E.R.I) and Aristotle University of Thessaloniki, Greece, and a number of end-user companies from several European countries.

Chapter 2 introduced the methodologies of Statistical Process Control (SPC) and Statistical Quality Control (SQC). Specifically, the use of SPC and SQC in industrial quality control problems, the charting methods they use, their advantages and limitations. The chapter continued by introducing Multivariate Statistical Process Control (MSPC) and its advantages of MSPC over univariate SPC and SQC. The applicability of MSPC for modern industrial processes and the multivariate statistical and charting techniques applied in MSPC were also presented.

Chapter 3 introduced to the multivariate statistical analysis of data, the statistical projection techniques used for dimensionality reduction, Principal Component

Analysis (PCA) and Projection to Latent Structures (PLS) and their application in MSPC. Specifically, most bodies of data collected from modern industrial processes, are multivariate in nature and significant relationships exist between the measurements of several process variables. Statistical projection techniques compress the data down onto lower dimensional subspaces defined by latent variables which can then be used as the basis of MSPC schemes. The derivation of PCA and PLS using geometrical, mathematical and statistical interpretations was presented.

Chapter 4 is dedicated to the implementation of MSPC schemes for process monitoring, fault detection and diagnosis and includes the theoretical developments achieved in this thesis. The main steps that have to be followed for the appropriate implementation of an MSPC schemes are described. Furthermore, extensions of PCA and PLS that allow the implementation of MSPC to special types of process are presented along with the specific features of the corresponding MSPC schemes. Specifically, Multi-Way PCA and Multi-Way PLS are applicable to processes that exhibit non-linear characteristics, such as batch and semi-batch processes, whilst Multi-Block PCA and Multi-Block PLS techniques are appropriate for processes comprising many distinct units. Finally, the novel approach of Inverse Projection to Latent Structures (IPLS) for the generation of pseudo data required for implementing an MSPC scheme when minimal process data sets is available, was presented.

Finally, Chapter 5 presented applications of the proposed techniques and methodologies. Specifically, two example processes were considered : a batch polymerisation reactor of Methyl-Methacrylate and a two-zone tubular reactor for the production of Low Density Poly-Ethylane (LDPE). Two inferential statistical models were developed for the batch polymerisation reactor : a PLS regression model for the prediction of the final polymer properties from the initial process conditions of the batch reactor and a Multi-Way PLS regression model for the estimation of the initial process conditions at an early stage of the polymerisation process. The models can be used to handle problems concerning the operation of batch reactors and in the general framework of an MSPC schemes. All issues related to these models were discussed in detail. The IPLS methodology was also applied to the batch reactor. Specifically, having developed a robust Multi-Way PLS model from minimal process data, the IPLS methodology is then applied to generate the required amount of data to establish an MSPC scheme. The Multi-Block PLS technique is illustrated by application to a two-zone LDPE tubular reactor.

In conclusion, the theoretical aspects in this work include the derivation of Projection to Latent Structures using geometrical, mathematical and statistical interpretations, a detailed review of the NIPALS algorithm used to perform Multi-Block PLS and the derivation of the Inverse PLS approach to generate process data. On the other hand, the applications presented in this work, illustrated the proposed techniques and the effectiveness of the statistical approach for solving typical problems found in industrial processes. Both theoretical developments and applications contributed to several aims and tasks of the two projects (INTELPOL and PROGNOSIS).

However, there are some issues that have not been addressed in this thesis. Specifically, the topics of SPC and MSPC for the process standard deviation has not been considered, since most of the methods and techniques are dedicated to the process mean. Alternative statistical projection techniques, such as Factor Analysis

(FA), Principal Components Regression (PCR) have not been considered, since it has been found that PCA and PLS are currently the most effective techniques for the development and application of MSPC-based schemes in chemical processes. Finally, although not included, the Multi-Block PCA technique has not been illustrated in this thesis, however, preliminary work has shown that it performs for the two-zone LDPE tubular reactor similarly to Multi-Block PLS.

6.2 Suggestions for Future Work

During this work a number of topics have been covered and many issues and questions raised. Some of these have been answered, other remain as challenges for the future.

- The proposed methodology of IPLS has been successfully applied to generate pseudo process data and establish an MSPC scheme for the batch reactor simulation. However, the potential strength of this approach cannot be proven without testing it on a real process, where it is not always possible to develop a robust PLS regression model that has to be inverted according to the methodology proposed.
- The Multi-Block techniques have been proposed in this thesis as suitable for developing statistical representations of complex processes comprising several distinct units. However, the Multi-Block PLS technique have only been investigated and applied to a process of simple structure, a two-zone tubular LDPE reactor. Research has to be conducted on the theoretical aspects of the Multi-Block PCA and PLS technique and, furthermore, on the application of Multi-Block techniques to processes with more complex structure, comprising

more than two blocks, in order to investigate the ability of the proposed techniques to be used in plant-wide MSPC schemes.

- The statistical projection techniques of PCA and PLS that are used to develop the statistical representation of a process suffer from a significant disadvantage. Specifically, they are known to fail to provide sufficient statistical process models when autocorrelation is presented in the process variables. Alternatively, other techniques and methodologies have been proposed (Mastrangelo and Montgomery, 1995; Faltin et al., 1995; Larimore, 1983; Schaper, et al., 1994). More detail research has to be undertaken in this particular topic, since autocorrelated characteristics exist in many industrial processes. In this way, MSPC schemes suitable for processes with autocorrelated observations could be developed and implemented.

MSPC is a powerful methodology for process performance enhancement. Research is currently being undertaken in a number of research centres around world and the results from MSPC implementation in industrial processes appear very promising. However, the challenges will never stop to grow, since every process has its own characteristics that differentiates it from others, and companies requirements that are needing to be fulfilled by MSPC schemes, continuously increase due to frequently changing technology and market conditions.

Bibliography

Akaike, H. (1976) "Canonical Correlation Analysis of Time Series and Use of Information Criterion.", in R. K. Mehra and D. G. Lainoitis (eds) "System Identification : Advances and Case Studies", *Academic Press*, New York

Allen, D. M., (1974) "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction", Technometrics, **16**, 125-127

Alsup, F. and R.M. Watson (1993) "Practical Statistical Process Control - A Tool for Quality Manufacturing", *Von Nostrand-Reinhold,* New York

Alt, F. B. (1985) "Multivariate Quality Control", in S. K. a. N. L. Johnson (eds) "Encyclopaedia of Statistical Sciences", Vol. 110-122, *John Wiley*, New York

Alt, F. B. and N. D. Smith, Eds. (1988). "Multivariate Process Control", Handbook of Statistics, *Elsevier Science Publishers B.V.*, Amsterdam

Alwan, L. C. and H. V. Roberts (1995) "The Problem of Misplaced Control Limits", *Journal of the Royal Statistical Society*, **44**(3), 268-278

Anderson, T. W. (1984) "An Introduction To Multivariate Statistical Analysis", *John Wiley & Sons,* New York

Banks, J. (1989) "Principles of Quality Control" *John Wiley,* New York

Barnard, G. A. (1959) "Control Charts and Stochastic Processes", *Journal of the Royal Statistical Society - Series (B)*, **25**

Bartlett, M. S. (1954) "A Note on The Multiplying Factors for Various χ^2 Approximation", *Journal of The Royal Statistical Society - Series B*, **16**, 296-298

Basilevsky, A. (1994) "Statistical Factor Analysis and Related Methods", *John Wiley & Sons,* New York

Box, G. E. P. (1954) "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: Effect of Inequality of Variance in One-way Classification", *Annals of Mathematical Statistics*, **25**, 290-302

Cadima, J. and I. T. Jolliffe (1995) "Loadings and Correlations in the Interpretation of Principal Components", *Journal of Applied Statistics*, **22**(2), 203-214

Cattell, R. B. (1966) "The Scree Test for the Number of Factors", *Journal of Multivariate Behavioral Research*, **1**, 245-276

Cattell, R. B. (1978) "The Scientific Use of Factor Analysis in Behavioural and Life Science", *Plenum Press,* New York

Cattell, R. B. and S. Vogelmann (1977) "A Comprehensive Trial of the Scree and the KG Criteria for Determining the Number of Factors", *Journal of Multivariate Behavioral Research*, **12**, 289-325

Caulcutt, R. (1995) "The Rights and Wrongs of Control Charts", *Journal of the Royal Statistical Society - Series C*, **44**(3), 279-288

Chen, W. H. (1996) "The Effects of SPC on the Target of Process Quality Improvement", *Journal of Quality Technology*, **28**(2), 224-228

Cheng G. and T. J. McAvoy (1996) "Predictive On-line Monitoring of Continuous Procesees", *submitted to Computers and Chemical Engineering*

Chew, V. (1968) "Simultaneous Prediction Intervals", *Technometrics*, **10**

Cliff, N. (1988) "The Eigenvalues-Greater-Than-One Rule and the Reliability of Components", *Psychological Bulletin*, **103**, 267-279

Cooley, W. W. and P. R. Lohnes (1971) "Multivariate Data Analysis", *John Wiley & Sons*, New York

Crosier, R. B. (1988) "Multivariate Generalizations of Cumulative Sum Quality Control Schemes", *Technometrics*, **30**, 291-303

Crowder, S. Y. (1989) "Design of Exponentially Weighted Moving Average Schemes", *Journal of Quality Technology*, **21**(3), 155-162

Davis, R. E. and W. H. Woodall (1994) "A Study of Parabolic Control Limits for EWMA Control Chart", *Communications in Statistics - Simulation*, **23**(1), 17-26

Dayal, B., J. F. MacGregor, P. Taylor, R. Kildaw and S. Marcikic (1994) "Application of Feed-forward Neural Networks and Partial Least Squares Regression for Modelling Kappa Number in a Continuous Kamyr Digester", *Pulp and Paper Canada*, **95**(1), T7-T13

Doganaksoy, N., F. W. Faltin and W. T. Tucker (1991) "Identification of Out of Control Quality Characteristics in a Manufacturing Environment", *Communication in Statistics - Theory and Methods*, **20**, 2775-2790

Dong, D. and T. J. McAvoy (1995). "Multi-Stage Batch Process Monitoring" presented at *International Control Conference*, Seatlle, Washington, 1857-1861

Eastman, H. T. and W. J. Krzanowski (1982) "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis", *Technometrics*, **24**(1), 73-77

Efron, B. (1979) "Bootstrap Methods : Another Look at the Jackknife", *Annals of Statistics*, **7**, 1-26

Evans, M., N. Hastings and B. Peacock (1993) "Statistical Distributions" *John Wiley & Sons*, New York

Everitt, B. S. and G. Dunn (1991) "Applied Multivariate Data Analysis", *Edward Arnold*, London

Ewan, W. D. (1991) "When and How to Use Cu-Sum Charts", *Technometrics*, **5**

Ferre, L. (1995) "Selection of Components in Principal Component Analysis: A Comparison of Methods", *Computational Statistics and Data Analysis*, **19**, 669-682

Flury, B. (1988) "Common Principal Components and Related Multivariate Models", *John Wiley & Sons*, New York

Frank, I. E. and J. H. Friedman (1993) "A Statistical View of Some Chemometrics Regression Tools", *Technometrics*, **35**(2), 109-135

Frontier, S. (1976) “Etude de la Decroissance des Valuers Propres dans une Analyze en Composantes Principales: Comparison avec le Modele de Baton Brise”, *Journal of Experimental Marine Biology and Ecology*, **25**, 67-75

Fuchs, C. and Y. Benjamini (1994) "Multivariate Profile Charts for Statistical Process Control", *Technometrics*, **36**(2), 182-195

Gallagher, N. B., B. M. Wise and C. W. Stewart (1996) "Application of Multi-way Principal Components Analysis To Nuclear Waste Storage Tank Monitoring", *Computers and Chemical Engineering*, **20**(Supplementary), S739-S744

Gan, F. F. (1991) "An Optimal Design of CUSUM Control Charts", *Journal of Quality Technology*, **23**(4), 279-286

Garthwaite, P. H. (1994) "An Interpretation of Partial Least Squares", *Journal of the American Statistical Association - Theory and Methods*, **89**(425), 122-127

Geladi, P. (1988) "Notes on the History and Nature of Partial Least Squares (PLS) Modelling", *Journal of Chemometrics*, **2**, 231-246

Geladi, P. and B. R. Kowalski (1986) "Partial Least-Squares Regression : A Tutorial", *Chimica Analytica Acta*, **185**, 1-17

Geladi, P., H. Isaksson, L. Lindqvist, S. Wold and K. Esbenssen (1989) " Principal Component Analysis of Multivariate Images", *Chemometrics and Intelligent Laboratory Systems*, **5**, 209-220

Gnanadesikan, R. (1997) "Methods for Statistical Data Analysis of Multivariate Observations", *John Wiley & Sons*, New York

Gnanadesikan, R., and J. R. Kettenring (1972) "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data", *Biometrics*, **28**, 81-124

Goldberg, J. L. (1991) "Matrix Theory and Applications", *McGraw-Hill*, New York

Green, P. E. (1978) "Mathematical Tools for Applied Multivariate Analysis", *Academic Press*, New York

Guttman, L. (1954) "Some Necessary Conditions for Common Factor Analysis", *Psychometrika*, **19**, 149-161

Hahn, G. J. and W. Q. Meeker (1991) "Statistical Interval. A Guide for Practitioners", *John Wiley & Sons*, New York

Harman, H. H. (1976) "Modern factor Analysis", *Chicago University Press*, Chicago

Harris, T. J. and W. H. Ross (1991) "Statistical Process Control for Correlated Observations", *Canadian Journal of Chemical Engineering*, **69**, 48-57

Hawkins, D. M. (1974) "The Detection of Errors in Multivariate Data Using Principal Components", *Journal of the American Statistical Association*, **69**, 340-344

Hawkins, D. M. (1980) "Identification of Outliers", *Chapman and Hall,* London

Hawkins, D. M. (1993) "Cumulative Sum Control Charting: An Underutilized SPC Tool", *Quality Engineering,* **5**

Hawkins, D. M. (1993b) "Regression Adjustment for Variables in Multivariate Quality Control ", *Quality Engineering,* **25**, 170-182.

Hayter, A. J. and K. L. Tsui (1994) "Identification and Quantification in Multivariate Quality Control Problems", *Journal of Quality Engineering,* **26**(3), 197-207

Healy, J. D. (1987) "A Note on Multivariate CUSUM Procedures", *Technometrics,* **29**, 409-412

Helland, I. S. (1988) "On the Structure of Partial Least Squares", *Communications in Statistics - Simulations,* **17**(2), 581-607

Himmelblau, D. M. (1978) "Faults Detection and Diagnosis In Chemical and Petrochemical Processes", *Elsevier,* Amsterdam

Horswell, R. L. and S. W. Looney (1992) "A Comparison of Tests for Multivariate Normality That Are Based On Measures of Multivariate Skewness and Kurtosis", *Journal of Statistical Computation and Simulation,* **42**, 21-38

Hoskuldsson, A. (1988) "PLS Regression Methods", *Journal of Chemometrics,* **2**, 211-228

Hoskuldsson, A. (1996) "Dimension of Linear Models", *Chemometrics and Intelligent Laboratory Systems,* **32**, 37-55

Hotelling, H. (1933) "Analysis of a Complex of Statistical Variables into Principal Components", *Journal of Educational Psychology,* **24**, 417-441

Hotelling, H. (1936) "Relations Between Two Sets of Variates", *Biometrika*, **28**, 321-377

Hotelling, H. (1947) "Multivariate Quality Control", in Eisenhart, Hastay and Wallis (editors) "Techniques of Statistical Analysis", Vol. 111-184, *McGraw-Hill*, New York

Ishikawa, K. (1986) "Guide to Quality Control", *Asian Productivity Association*, Tokyo

Jackson, D. A. (1993) "Stopping Rules in principal Components Analysis: A Comparison of Heuristic and Statistical Approaches", *Ecology*, **74**(8), 2204-2214

Jackson, J. E. (1959) "Quality Control Methods For Several Related Variables" *Technometrics*, **1**

Jackson, J. E. (1985) "Multivariate Quality Control", *Communications in Statistics - Theory and Methods*, **14**(11), 2657-2688

Jackson, J. E. (1991) "A User's Guide to Principal Components", *John Wiley & Sons*, New York

Jackson, J. E. and G. S. Mudholkar (1979) "Control procedures for Residuals Associated With Principal Component Analysis", *Technometrics*, **21**(3), 341-349

Jaeckle, C. and J. F. MacGregor (1996) "Product Design Trough Multivariate Statistical Analysis of process Data", *Computers and Chemical Engineers*, **20**(Supplementary), S1047-S1052

Johnson, N. L. (1961) "A Simple Theoretical Approach to Cumulative Sum Control Charts", *Journal of the American Statistical Association*, **54**

Johnson, R. A. and D. W. Wichern (1992) "Applied Multivariate Statistical Analysis", *Prentice-Hall, Inc*, New Jersey

Jolliffe, I. T. (1986) "Principal Components Analysis", *Springer-Verlag,* New York

Juran, J. (1979) "Quality Control Handbook", *McGraw-Hill,* New York

Kaiser, H. F. (1960) "The Application of Electronic Computers in Factor Analysis", *Educational and Psychological Measurement*, **20**, 141-151

Kaspar, M. H., and W. H. Ray (1993) "Partial Least Squares Modelling as Successive Singular Value Decompositions", *Computers and Chemical Engineering,* **17**(10), 985-989

Kendall, M. G. (1980) "Multivariate Analysis", *Griffin,* London

Kiparissides, C., M. Papazoglou, A. Simoglou, G. Stavropoulos and V. Nasiopoulos (1997) "User Requirements and Functional Specifications for the A.U.T. Demonstrator Process", D1.1a, ESPRIT Project 22281, *L.P.R.E. - Aristotle University of Thessaloniki*

Kiparissides, C. (1997) "Personal Communication"

Kiparissides, C. (1996) "Personal Communication"

Kiparissides, C. (1995) "Personal Communication"

Kiparissides, C. and A. J. Morris (1996) "Intelligent Manufacturing of Polymers" *Computers and Chemical Engineering,* **20**(Supplementary)

Kiparissides, C., G. Verros and J. F. MacGregor (1993) "Mathematical Modeling, Optimization and Control of High-Pressure Ethylene Polymerization Reactors" *J. Macromol. Sci.-Rev. Macromol. Chem. Phys.*, **C33**(4), 437

Kosanovich, K. A. and M. J. Piovoso (1995). "Multivariate Statistical Methods Applied to Process Monitoring" presented at *EPRSC Innovative Manufacturing Initiative*, Newcastle Upon Tyne, U.K

Kosanovich, K. A., K. S. Dahl and M. J. Piovoso (1996) "Improved Process Understanding Using Multiway Principal Components Analysis", *Industrial Engineering Chemistry and Research*, **35**, 138-146

Kourti, T. and J. F. MacGregor (1996) "Multivariate SPC Methods for Process and Product Monitoring", *Journal of Quality Technology*, **28**(4), 409-428

Kourti, T., P. Nomikos and J. F. MacGregor (1995) "Analysis Monitoring and Fault Diagnosis of Batch Processes Using Multi-Block and Multiway PLS", *Journal of Process Control*, **5**(4), 277-284

Kresta, J. V., T. E. Marlin and J. F. MacGregor (1994) "Development of Inferential Process Models Using PLS", *Computers and Chemical Engineering*, **18**(7), 597-611

Kresta, J., J. F. MacGregor and T. E. Marlin (1991) "Multivariate Statistical Monitoring of Process Operating Performance", *Canadian Journal of Chemical Engineering*, **69**, 35-47

Krzanowski, W. J. (1988) "Principles of Multivariate Analysis - A User's Perspective", *Oxford University Press*, New York

Krzanowski, W. J. and F. H. C. Marriott (1996) "Multivariate Analysis Part 2 : Classification, Covariance Structures and Repeated Measurements", *Edward Arnold*, London

Lambert, Z. V., A. R. Wildt and R. M. Durand (1990) "Assessing Sampling Variation Relative to Number-of-Factors Criteria", *Educational and Psychological Measurement*, **50**, 33-49

Larimore, W. E. (1983). "System Identification, Reduced Order Filtering and Modeling Via Canonical Variate Analysis" presented at *Proceedings of the American Control Conference*

Lawly, D. N. (1956) "Tests of Significance for the Latent Roots of the Covariance and Correlation Matrices", *Biometrika*, **43**, 128-136

Lawly, D. N. (1963) "On Testing a Set of Correlation Coefficients for Equality", *Annals of Mathematical Statistics*, **34**, 128-136

Legendre, L. and P. Legendre (1983) “Numerical Ecology”, *Elsevier*, Amsterdam

Lindgren, F., P. Geladi and S. Wold (1993) "The Kernel Algorithm for PLS", *Journal of Chemometrics*, **7**, 45-59

Lorber, A., L. E. Wangen and B. R. Kowalski (1987) "A Theoretical Foundation for the PLS Algorithm" *Journal of Chemometrics*, **1**, 19-31

Lowry, C. A. and D. C. Montgomery (1995) "A Review of Multivariate Control Charts", *IIE Transactions*, **27**, 800-810

Lowry, C. A., C. W. Champ and W. H. Woodall (1995) "The Performance of Control Charts for Monitoring Process Variation", *Communications in Statistics - Simulation*, **24**, 409-437

Lowry, C. A., W. H. Woodall and C. W. Champ (1992) "Multivariate Exponentially Weighted Moving Average Control Chart", *Technometrics*, **32**, 1-12

Lucas, J. M. and S. M.S. (1990) "Exponentially Weighted Moving Average Control Schemes : Properties and Enhancements", *Technometrics*, **32**(1), 1-12

MacGregor, J. F. (1994). "Statistical Process Control of Multivariate Processes" presented at *ADCHEM 1994, IFAC*, Kyota, Japan

MacGregor, J. F. and T. Kourti (1995) "Statistical Process Control of Multivariate Processes", *Control Engineering Practice*, **3**, 403-414

MacGregor, J. F., C. Jaeckle, C. Kiparissides and M. Koutoudi (1994) "Process Monitoring and Diagnosis by Multi-Block PLS Methods", *Journal of the Americal Institution of Chemical Engineers*, **40**(5), 826-838

Mackiewicz, A. and W. Ratajczak (1993) "Principal Components Analysis (PCA)", *Computers & Geosciences*, **19**(3), 303-342

Mahalanobis, P. C. (1936) "On the Generalised Distance in Statistics", *Proceedings of National Academy of India*, **12**, 49-55

Manne, R. (1987) "Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration", *Chemometrics and Intelligent Laboratory Systms*, **2**, 187-197

Mardia, K. V., J. T. Kent and J. M. Bibby (1974) "Multivariate Analysis", *Academic Press*, London

Martens, H. and T. Naes (1989) "Multivariate Calibration", *John Wiley & Sons*, New York

Martin, E. B., A. Bettoni and A. J. Morris (1997) "Recommendations on Testing, Data Collection Procedures and Data Prescreening", Deliverable D1.4, *ESPRIT Project 22281*, C.P.A.C.C., University of Newcastle

Mason, R. L., N. D. Tracy and J. C. Young (1995) "Decomposition of T^2 For Multivariate Control Chart Interpretation", *Journal of Quality Technology*, **27**(1), 99-108

Mason, R. L., N. D. Tracy and J. C. Young (1997) "A Practical Approach For Interpreting Multivariate T^2 Control Chart Signals" *Journal of Quality Technology*, **29**(4), 396-406

Massy, W. F. (1965) "Principal Components Regression in Explanatory Statistical Research", *Journal of the American Statistical Association*, **60**, 234-246

Mastrangelo, C. M. and D. C. Montgomery (1995) "SPC with Correlated Observations for the Chemical and Process Industries", *Quality and Reliability Engineering International*, **11**

Miller, P., R. E. Swanson and C. F. Heckler (1995). "Contribution Plots: The Missing Link in Multivariate Quality Control" presented at *Multivariate Statistical Process Control and Plant Performance Monitoring Industrial Representatives Meeting*, University of Newcastle

Montgomery, D. C. (1996) "Introduction to Statistical Quality Control", *John Wiley*, New York

Montgomery, D. C. and E. A. Peck (1992) "Introduction to Linear Regression Analysis", *John Wiley & Sons*, New York

Morris, A. J. and E. B. Martin (1997) *Personal Communication*

Muirhead, R. J. (1982) "Aspects of Multivariate Statistical Theory", *John Wiley & Sons*, New York

Murphy, B. J. (1987) "Selecting Out of Control Variables with the T^2 Multivariate Quality Control Procedure", *The Statistician*, **36**, 571-583

Nomikos P. and J. F. MacGregor (1994) "Monitoring of Batch Processes Using Multi-way Principal Component Analysis", *Journal of the American Institution of Chemical Engineers*, **40**, 1361-1375

Nomikos, P. and J. F. MacGregor (1994b). "Multi-way Partial Least Squares in Monitoring of Batch Processes" presented at *First International Chemometrics Internet Conference InCINC 1994*

Nomikos, P. and J. F. MacGregor (1995) "Multivariate SPC Charts for Monitoring Batch Processes", *Technometrics*, **37**(1), 41-59

Oakland, J. S. and R. F. Followell (1990) "Statistical Process Control", *Heineman*, Oxford

Page, E. S. (1954) "Continuous Inspection Schemes", *Biometrics*, **41**

Page, E. S. (1961) "Cumulative Sum Control Charts", *Technometrics*, **3**

Palm, A. C., R. N. Rodriguez, F. A. Spiring and D. J. Wheeler (1997) "Some Perspectives and Challenges for Control Chart Methods", *Journal of Quality Technology*, **29**(2), 122-127

Papazoglou, M., E. B. Martin, A. J. Morris and C. Kiparissides (1998) "Monitoring and Quality Control of Batch Polymerisation Reactors by MSPC Methods" *to be presented in DYCOPS'98, Corfu, Greece*

Patton, R. J., P. M. Frank and R. N. Clark (1989) "Fault Diagnosis in Dynamic Systems: Theory and Applications", *Prentice-Hall*, London

Pearson (1901), K. "On Lines and Planes of Closest Fit to Systems of Points in Space", *Philosophical Magazine*, **2**, 559-572

Pignatiello, J. J. and G. C. Runger (1990) "Comparisons of Multivariate CUSUM Charts" *Journal of Quality Technology*, **22**(3), 173-186

Prabhu, S. and G. C. Runger (1997) "Designing a Multivariate EWMA Control Chart", *Journal of Quality Technology*, **29**(1), 8-15

Raich, A. and A. Cinar (1996) "Statistical Process Monitoring and Disturbance Isolation in Multivariate Continuous Processes", *Journal of the American Institution of Chemical Engineers*, **42**(4), 995-1009

Rao, C. R. (1971) "Calculus of Generalized Inverse of Matrices. Part I - General Theory", *Sankhya*, **26**(A)

Rencher, A. C. (1993) "The Contribution of Individual Variables to Hotelling's T^2, Wilk's Λ, and R^2" *Biometrics*, **49**, 479-489

Roberts, S. W. (1959) "Control Charts Tests Based on Geometric Moving Averages", *Technometrics*, **1**

Rothwell, S.G., E.B. Martin and A.J. Morris (1998) "Comparison of Methods for Dealing with Uneven Length Batches", to be presented in the *7th International Conference on Computer Applications in Biotechnology - CAB7*, Osaka, Japan

Runger, G. C. (1996) "Projections and the U2 Multivariate Control Chart", *Journal of Quality Technology*, **28**(3), 313-318

Runger, G. C., F. B. Alt and D. C. Montgomery (1996) "Contributors To a Multivariate Statistical Process Control Chart Signal", *Communications in Statistics-Theory and Methods*, **25**(10), 2203-2213

Sanchez, E. and B. R. Kowalski (1990) "Tensorial Resolution: A Direct Trilinear Decomposition", *Journal of Chemometrics*, **4**, 29-45

Schaper, C. D., W. E. Larimore, D. E. Seborg and D. A. Mellichamp (1994) "Identification of Chemical Processes Using Canonical Variate Analysis", *Computers and Chemical Engineering*, **18**(1), 55-69

Searle, S. R. (1977) "Linear Models", *John Wiley & Sons*, New York

Searle, S. R. (1984) "Matrix Algebra Useful For Statistics", *John Wiley & Sons*, New York

Seber, G. A. F. (1977) "Linear Regression Analysis", *John Wiley & Sons*, New York

Seber, G. A. F. (1984) "Multivariate Observations", *John Wiley & Sons*, New York

Sharaf, M. A., D. L. Illman and B. R. Kowalski (1986) "Chemometrics", *John Wiley & Sons*, New York

Shewhart, W. A. (1931) "Economic Control of Quality of Manufactured Product", *Von Nostrand-Reinhold*, Princeton

Skagerberg, B., J. F. MacGregor and C. Kiparissides (1992) "Multivariate Data Analysis Applied to Low Density Polyethylene Reactors", *Chemometrics and Intelligenct Laboratory Systems*, **14**, 341-356

Smilde, A. K. (1992) "Three-Way Analysers. Problems and Prospects", *Chemometrics and Intelligent Laboratory Systems*, **15**

Smilde, A. K. and D. A. Doornbos (1991) "Three-Way Methods for the Calibration of Chromatographic Systems: Comparing PARAFAC and Three-Way PLS", *Journal of Chemometrics*(5), 345-360

Sparks, R. S. (1992) "Quality Control with Multivariate Data", *Australian Journal of Statistics*, **34**(3), 375-390

Stephanopoulos G. and C. Han (1996) "Intelligent Systems In Process Engineering : A Review", *Computers and Chemical Engineering*, **20**(6/7), 743-791

Stone, M. (1974) "Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of the Royal Statistical Society*, **36**, 111-133

Stone, M. and R. J. Brooks (1990) "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression", *Journal of the Royal Statistical Society - Series B*, **52**(2), 237-269

Tano, K., P. O. Samskog, J. C. Garde and B. Skagerberg (1993) "Partial Least Squares Modelling of Process Data at LKAB: Predicting Chemical Assays in Iron Ore for Process Control" presented at *International Symposium on the Application of Computers and Operations Research in the Mineral Industries*, Montreal, Canada, Canadian Institution of Mining, Metallurgy and Petroleum

Tracy, N. D., J. C. Young and R. L. Mason (1992) "Multivariate Control Charts for Individual Observations", *Journal of Quality Technology*, **24**(2), 88-95

Van den Wallenberg, A. L. (1977) "Redundancy Analysis: An Alternative for Canonical Correlation Analysis", *Psychometrika*, **43**, 225-243

Velicer, W. T. (1976) "Determining the Number of Components From the Matrix of Partial Correlations", *Psychometrika*, **41**(3), 321-326

Wang, T. W., A. Khettry, M. Berry and J. Batra (1994). "SVDPLS : An Efficient Algorithm for Performing PLS" presented at *The First International Chemometrics InterNet Conference, (InCINC'94)*, http://www.emsl.pnl.gov:2080/docs/incinc/

Wangen, L. E. and B. R. Kowalski (1988) "A Multiblock Partial Least Squares Algorithm For Investigating Complex Chemical Systems", *Journal of Chemometrics*, **3**, 3-20

Wetherill, G. B. and D. W. Brown (1991) "Statistical Process Control - Theory and Practice", *Chapman and Hall*, London

Wise, B. M. and N. B. Gallagher (1996) "The Process Chemometrics Approach to Process Monitoring and Fault Detection", *Journal of Process Control*, **6**(6), 329-348

Wold, H. (1966) "Nonlinear Estimation by Iterative Least Squares Procedures" in F. N. David (eds) "Research papers in statistics: Festschrift for J. Neyman", *John Wiley & Sons*, New York

Wold, H. (1975) "Soft Modeling by Latent Variables : The Nonlinear Iterative Partial Least Squares Approach", in J. Gani (eds) "Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett", *Academic Press*, London

Wold, H. (1982) "Soft Modelling. The Basic Design and Some Extensions", in K. G. Joreskog and H. Wold (eds) "Systems Under Indirect Observations", Vol. II, *North Holland,* Amsterdam

Wold, S. (1978) "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models", *Technometrics*, **20**(4), 397-405

Wold, S., A. Ruhe, H, Wold and W. J. Dunn (1984) "The Collinearity Problem in Linear Regression the Partial Least Squares (PLS) Approach to Generalized Inverses", *SIAM - Journal of Scientific Statistical Computations*, **5**(3), 735-743

Wold , S. (1987) "Principal Component Analysis", *Chemometrics and Intelligent Laboratory Systems*, **2**, 37-52

Wold, S., P. Geladi, K. Esbensen and J. Ohman (1987) "Multi-Way Principal Components- And PLS-Analysis" *Journal of Chemometrics*, **1**, 41-56

Wold, S., S. Hellberg, T. Lundstedt, M. Sjostrom and H. Wold (1987b). "PLS Modeling with Latent Variables in Two or More Dimensions, Version 2.1" presented at *Frankfurt PLS Meeting*, Frankfurt, Germany

Wold, S., N. Kettaneh and K. Tjessem (1996) "Hierarchical Multiblock PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection", *Journal of Chemometrics*, **10**, 463-482

Woodall, W. H. and B. M. Adams (1993) "The Statistical Design of CUSUM Charts", *Quality Engineering*, **5**

Zeng, Y. and P. K. Hopke (1990) "Methodological Study Applying Three-Way Factor Analysis to Three-Way Chemical Databases", *Chemometrics and Intelligent Laboratory Systems*, **7**, 237-250

Zhang, J., E. B. Martin and A. J. Morris (1996) "Fault Detection and Diagnosis Using Multivariate Statistical Techniques", *Transactions of Institution of Chemical Engineers - Part A*, **74**, 89-96