STATISTICAL PROCESS CONTROL OF BATCH PROCESSES

By

PAUL NOMIKOS, B. ENG.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

McMaster University

STATISTICAL PROCESS CONTROL OF BATCH PROCESSES

DOCTOR OF PHILOSOPHY (1995) McMASTER UNIVERSITY

(Chemical Engineering) Hamilton, Ontario, Canada

TITLE: Statistical Process Control of Batch Processes

AUTHOR: Paul Nomikos B.ENG. (Aristotle University of Greece)

SUPERVISOR: Professor John F. MacGregor

NUMBER OF PAGES: xi, 146

ABSTRACT

Multivariate statistical procedures for the analysis and monitoring of batch and semi-batch processes are developed. The only information needed to exploit the procedures is a historical database of past successful batches. Projection methods based on principal component analysis and partial least squares are utilized to compress the information contained in the multivariate trajectory data and final product qualities, by projecting them onto low dimensional spaces. When additional information about the initial conditions and set-up of the batch process is available, multi-block approaches can be used to integrate the additional data into the proposed schemes.

The proposed methodology facilitates the analysis of operational and quality control problems in past batches, and allows for the development of simple multivariate statistical process control charts for on-line monitoring of the progress of new batches. Control limits for the proposed charts are developed using information from the historical reference distribution of past successful batches. The approach is capable of detecting subtle changes in the batch operation, and provides procedures for diagnosing assignable causes for the occurrence of observable upsets. The method's potential in analyzing past batches and tracking the progress of new batch runs, is illustrated through a simulation example and data collected from industrial polymerization reactors.

ACKNOWLEDGMENTS

First of all, I would like to thank my parents, Μαρία and Αρτέμιος Νομικός, whose sacrifices made everything possible. They have my everlasting love.

I am indebted to my supervisor, Dr. John F. MacGregor, for his guidance, encouragement, kindness, and help which made this project a truly enjoyable and stimulating experience.

I would also like to thank:

Dr. Roman Viveros for his active interest and valuable suggestions in the preparation of this thesis.

Dr. Svante Wold for his kind advise and criticism in the course of my investigations.

Michael Piovoso, Karlene Kosanovich, Mahmud Rahman, Ken Dahl (all of DuPont US), Kevin Deluzio, and Natalie Heisler (both of DuPont Canada), for their help in this work and for their suggestions and constructive comments.

Finally, special thanks to all the graduate students of McMaster, my friends in Hamilton, and my beloved Dora, who made this Ph.D. much more than what is contained in these pages.

Thank you all,

Paul Nomikos

TABLE OF CONTENTS

		Page
List of Figures		vii
List of Tables		xi
Chapter 1	Introduction	1
1.1	Monitoring Batch Processes	2
	1.1.1 State Estimation Approaches	4
	1.1.2 Knowledge Based Approaches	6
1.2	SPC in Batch Processes	7
Chapter 2	Analysis of Batch Data Using MPCA	12
2.1	Principal Components Analysis	12
2.2	Multi-way PCA in Batch Processes	13
2.3	Examples of MPCA in Batch Data	19
	2.3.1 Styrene-Butadiene Simulation Example	19
	2.3.2 Industrial Example	23
2.4	Discussion	27
Chapter 3	Building the Reference Statistical Model	31
3.1	Reference Distribution of Normal Batches	31
3.2	Selecting the Number of Principal Components	32
3.3	MPCA Reference Model	34
3.4	Testing and Diagnosing a New Batch	41
	3.4.1 Contribution Plots	41
Chapter 4	On-Line Monitoring of Batch Processes	47
4.1	On-Line Monitoring via MPCA	47
4.2	Anticipating the Future Observations in $\mathbf{X}_{NEW}$	51
4.3	Control Limits for the SPC Charts	56
	4.3.1 Control Limits on T-scores	57
	4.3.2 Control Limits on SPE	58
	4.3.3 Smoothing Window	61
	4.3.4 Overall Type I Error	63
4.4	Examples of On-Line Monitoring	64
4.5	On-Line Contribution Plots	69
4.6	Discussion	72

		Page
Chapter 5	MPLS and Multi-Block Approaches in Batch Processes	76
5.1	Partial Least Squares	76
	5.1.1 Approximate Confidence Intervals for the PLS Predictions	79
	5.1.2 MPLS Analysis of Batch Data and On-line Monitoring	82
	5.1.3 MPCA or MPLS	89
5.2	Multi-Block MPLS	92
	5.2.1 Scaling and Contribution of each $\mathbf{X}_k$ block	95
	5.2.2 Multi-Block MPLS in Industrial Data	97
Chapter 6	Comparing Covariance Matrices and Subspaces Developed by Projection Methods	104
6.1	Maximum Likelihood Ratio Test	105
6.2	A Geometric Test	107
6.3	A New Approach: Directional Variance Test	109
	6.3.1 Comparison of Two Industrial Parallel Batches	114
	6.3.2 Comparison of an MPCA and MPLS Model	117
6.4	Comparison of Projection Models Based on a Residuals Analysis	120
Chapter 7	Summary and Conclusions	122
Appendices		127
Appendix A	Styrene-Butadiene Simulation	127
Appendix B	Other MPCA Approaches for Monitoring Batch Processes	128
Appendix C	Control Limits for Quadratic Forms	133
Notation		134
Bibliography		138

LIST OF FIGURES

		Page
2.1	Arrangement of batch data in MPCA. Each horizontal slice contains the measurement trajectories from a single batch. Each vertical slice has the measurements of all the batches at a single time interval. MPCA extracts the variation in these vertical slices and studies how the measurements deviate for their mean trajectories.	15
2.2	MPCA and its equivalent PCA form for batch data. The three-way array $\underline{\mathbf{X}}(I \times J \times K)$ unfolds into a matrix $\mathbf{X}(I \times JK)$ where a normal PCA analysis is performed.	16
2.3	Measurement trajectories from the SBR example. The solid line is from a normal batch which will be used as a test batch in Section 4.4, the dashed line from the batch with the initial problem, and the dotted line from the batch with the problem halfway through its operation.	21
2.4	MPCA is able to discriminate the two abnormal batches (51 and 52) in the SBR example.	22
2.5	Results of MPCA analysis of the original 55 batches from the industrial example. Batches 50 through 55 are identified as abnormal because of their position in the reduced space, and batch 49 because of its residuals.	24
2.6	Measurements of batch 49 (dashed lines) are contrasted with the mean trajectories (solid lines) of the normal batches in the industrial example.	25
2.7	MPCA analysis of the first 48 batches from the industrial example. Batches 37, 39, and 43 through 48 can be identified as batches with unusual operation.	26
2.8	T-plot from the post analysis of an industrial dataset. Batches with similar faults clustered together in areas indicated by ellipses. The rectangular area at the center contains all the normal batches. Batch 36 had the same fault as batches 32 through 35 but in a smaller extent.	29
3.1	T, D_S and Q plots with their 95% and 99% confidence limits from the MPCA analyses of the reference databases for both the SBR example (left hand side plots) and the industrial example (right hand side plots). None of the batches exhibits any notably unusual behavior.	37
3.2	Cumulative percent of the total variation with respect to time and variables which is explained in each reference database. The plots in the left hand side are from the SBR example, and the plots in the right hand side are from the industrial example. Each line represents the cumulative percent of the total variance explained by the addition of each principal component.	39

		Page
3.3	Contribution plots for the batch with the initial problem in the SBR example (left hand side plots), and batch 49 in the industrial example (right hand side plots). The signs of the % contributions in the bottom two plots, indicate if the measurement variables were above (positive) or below (negative) their mean trajectories.	44
4.1	Graphical representation of a batch operation in the reduced space of MPCA. The stars on the reduced space are the t-scores which are the projections of the process measurements into the reduced space. The sum of squares of the residuals are the normal distances of the process measurements from the reduced space. The area inside the ellipse depicts the normal operational region.	50
4.2	Control limits (95% and 99%) for the SPE and t-scores of the SBR example, with the outermost values at each time interval for the three approaches of handling future deviations in $\mathbf{X}_{NEW}$. The upper plots are for the approach with zero, the middle plots for the approach with current deviations, and the bottom plots for the approach by projection.	52
4.3	Control limits (95% and 99%) for the SPE and t-scores of the industrial example, with the outermost values at each time interval for the three approaches of handling future deviations in $\mathbf{X}_{NEW}$. The upper plots are for the approach with zero, the middle plots for the approach with current deviations, and the bottom plots for the approach by projection.	53
4.4	Plots of estimated g and h values from the $g\chi^2_h$ distribution of the SPE. The left hand side plots are for the SBR example, and the right hand side plots are for the industrial example.	60
4.5	99% SPE control limits for the SBR (left hand side plot) and the industrial (right hand side plot) example. The dotted lines show the limits estimated without windowing (w=0). The solid lines show the limits estimated with a window width of 7 (w=3) and 5 (w=2) for the SBR and the industrial example respectively. The dashed lines show the limits estimated with a window width of 11 (w=5) and 9 (w=4) for the SBR and the industrial example respectively.	62
4.6	Monitoring charts with their 95% and 99% control limits for a new normal SBR batch.	65
4.7	Monitoring charts with their 95% and 99% control limits for the SBR batch with an initial impurity problem.	66
4.8	Monitoring charts with their 95% and 99% control limits for the SBR batch with an impurity problem halfway through its operation.	67

		Page
4.9	Monitoring charts with their 95% and 99% control limits for the industrial abnormal batch. Bars above observations denote truncated values for better graphical representation. The SPE values at time intervals 57 through 62 are ten times greater than what the graph shows.	68
4.10	Contribution plots for the abnormal batches. The upper left plot is for the SBR batch with initial problem. The upper right plot is for the industrial abnormal batch. The bottom plots are for the SBR batch with problem halfway through its operation.	70
5.1	T vs. U plot for the two MPLS components. Each point represents one of the 50 batches in the reference database. The t and u-observations, for both MPLS components, fall close to the diagonal line of the graph. This indicates that the MPLS latent variables in the x and y-space (t, u) are well correlated.	85
5.2	Monitoring charts for the SPE and t_1-scores with their 95% and 99% control limits (dashed and solid lines) for the normal SBR batch (left hand side plots) and for the abnormal SBR batch with problem halfway through its operation (right hand side plots). The abnormality in the bad batch is clearly flagged in the SPE chart after time interval 105.	88
5.3	On-line predictions, with their 95% and 99% confidence intervals (dashed and solid lines), for three of the five final product quality variables for the normal SBR batch (left hand side plots) and for the abnormal SBR batch with problem halfway through its operation (right hand side plots). The actual final product qualities are indicated by diamond marks. The PLS model for the bad batch is not valid after time interval 105 (absence of confidence intervals), and its predictions are not generally trustworthy.	90
5.4	Overall SPC monitoring scheme for batch processes based on Multi-block projection methods.	92
5.5	Multi-block MPLS analysis of the industrial example. T-score plots for each z and x-block are shown. The consensus t-score plot is given at the bottom of the figure.	100
6.1	Ellipses of two samples which have equal covariance determinants. The axis lengths of these ellipses are equal to one standard deviation. In the left hand side plot, the principal axes of sample B have been tilted by approximately 25°. In the right hand side, both samples have the same principal axes, but the variance in each direction is different.	106
6.2	T-plots from MPCA analyses of the reference batch databases of the two industrial claves. The upper t-plots are from MPCA analyses of each clave separately (left plot from clave A). The bottom left hand side t-plot is from an MPCA analysis of both claves mean centered and scaled together. The bottom right hand side t-plot is from an MPCA analysis of both claves mean centered and scaled separately. The first 36 batches in the bottom t-plots are from clave A.	115

Page

6.3 T-plots of the reference SBR database. The left hand side t-plot is from the MPCA analysis, and the right hand side plot is from the MPLS analysis. 118

B1 Unfolded **X** matrix and plots of p-loadings versus time. The left hand side plot shows the p-loadings of the first principal component for measurement variable 8 in the SBR example. The right hand side plot shows the p-loadings of the first principal component for measurement variable 10 in the industrial example. The rectangular areas in the later plot indicate periods of sudden changes in variable 10 where the p-loadings might have been set equal to zero. 130

LIST OF TABLES

		Page
3.1	Percentage of explained sum of squares (cumulative and for each principal component) from the MPCA analysis of both examples, and the results from the three criteria to determine the optimal number of principal components.	35
4.1	Overall Type I error for the control limits of the t-scores and SPE for the three approaches to handling the future observations in $\mathbf{X}_{NEW}$.	64
5.1	Sum of Squares due to Regression (SSR) and Sum of Squared Errors (SSE) along with their degrees of freedom for the PLS model.	81
5.2	Percentage sum of squares explained in **X** and **Y** and in each of the quality variables by the first two PLS components for the SBR reference database.	84
5.3	Scaling factors for the x-blocks.	99
5.4	Contribution factors of each block to the y-predictions, for each component and overall.	101
5.5	Percentage of explained sum of squares (cumulative) from the MB-MPLS analysis of the industrial example.	102
5.6	Percentage of explained sum of squares (cumulative) from the MPLS analyses between each x-block and **Y**.	102
A1	Range of perturbations from the base case for the creation of the reference normal database.	128

CHAPTER 1 Introduction

Batch and semi-batch processes are a significant class of processes in the chemical industry and play an important role in the production and processing of high quality, specialty materials and products. Examples include the production of polymers, pharmaceuticals, and biochemicals, the separation and transformation of materials by batch distillation and crystallization, and the processing of materials by injection molding. Monitoring these processes is very important to ensure their safe operation and to assure that they produce consistent high quality products. This work represents the first time multivariate Statistical Process Control (SPC) ideas which have been applied to dynamic batch trajectory data. It facilitates the analysis of operational and quality problems in past batches and allows for the development of multivariate SPC charts for on-line monitoring of the progress of new batches.

SPC is a key factor in industry for global competition. It is not only a tool, it is actually an attitude and a motivation for all individuals in a company to strive for continuous improvement in quality and productivity. Consistent quality production of any product has become imperative and one has to look no further than Japan to see the importance and success of SPC methods.

The objective of this research is to develop some new tools for monitoring and diagnosing problems in batch processes with the following desirable characteristics:

- Generic — Easy to be applied in every batch or semi-batch process.
- Informative — Easy to be used and interpreted by an operator and able to indicate clearly and quickly an abnormal operation.
- Simple — Their application does not require heavy, time consuming, or problematic computations.
- Ameliorative — To provide a continuous basis for improvement in the process.

To accomplish these goals, multivariate SPC methods are developed which utilize the information in the on-line measurements and make inferences about the operation of

the processes. It is common to hear from experienced operators: "When I see these variables behave in this manner, I know that something is wrong". From this expression, it is clear that there exists a need for a statistical method which can handle many variables and also take into account the correlation between them, with the aim to systematically and scientifically recognize significant deviations from the normal operating behavior of the process.

In batch processes, it is common to interchange the words control and monitoring since generally one uses the control method also for monitoring purposes. Control describes a method which will force a process to follow a desirable policy, while monitoring refers to a method which keeps track of the process progress and detect any faults. A fault is understood as any kind of malfunction in the process that leads to an unacceptable performance and in general to a product out of specifications. This work focuses only in monitoring batch processes.

1.1 Monitoring Batch Processes

Batch processes still pose an important challenge for the application of any control or monitoring scheme. General batch issues like scheduling, operation planning, optimization, qualitative control decisions, and developing operating policies have been discussed by MacGregor et al. (1984), Birewar and Grossmann (1989), Cuthrell and Biegler (1989), Crook et al. (1990), Stephanopoulos et al. (1990), and Rippin (1992). Nonlinear feedback control of product quality has been discussed by Kravaris et al. (1989), Kozub and MacGregor (1992a), and Peterson et al. (1992). The major difficulties limiting our ability to provide adequate control and monitoring include: the lack of on-line sensors for measuring product quality variables, the finite duration of batch processes, the presence of significant nonlinearities, the absence of steady state operation, and the difficulties in developing accurate mechanistic models that characterize all the chemistry, mixing and heat transfer phenomena occurring in these processes.

Batch processes generally exhibit some batch to batch variation arising from things, such as deviations of the process variables from their specified trajectories, errors in the charging of the recipe of materials, and disturbances originating from variations in impurities. Abnormal conditions which develop during a batch operation can lead to the production of at least one batch or a whole sequence of batches with poor quality product, if the problem is not detected and remedied. In spite of this, most industrial batch processes are run without any effective form of on-line monitoring to ensure that the batch is progressing in a manner that will lead to a high quality product.

Usually, batch processes operate in open-loop with respect to the product qualities, simply because few, if any, on-line sensors exist for tracking these variables. Upon completion of a batch, several quality measurements are made on a sample of the product in the quality laboratory. In some cases these measurements may be used to adjust the recipe or the operation of the next batch. In spite of the fact that a large number of process measurements on variables like reactor and jacket temperatures, pressures, densities, pH, flowrates, etc. are available almost on a continuous basis throughout the several hours of a batch run, there are almost no reported attempts in industry to utilize this information for monitoring the batch progress or to use it to detect and diagnose potential problems with the reactor operation.

Most of the existing industrial approaches, for achieving consistent and reproducible results from batch processes, are based on the precise sequencing and automation of all the stages in the batch operation. Monitoring is usually confined to checking that these sequences are followed, and that certain reactor variables, such as temperatures and reactant feedrates, are following acceptable trajectories. In some cases, on-line energy balances are used to keep track of the instantaneous reaction rate, and the conversion or the residual reactant concentrations in the reactor (Wu, 1985). Some effort has been made in industry at using relational database software to try to uncover particular attributes of the measurement trajectories, such as the timing of valve openings, or the

maximum temperature or pressure attained during an interval which appear to affect product quality, and then to monitor these attributes.

Current research approaches to monitoring batch processes have focused on the use of either fundamental mathematical models, or detailed knowledge based models (Frank, 1990). The first takes advantage of a mechanistic model to describe the process, and the monitoring procedure is based on state estimation methods. The second relies on the knowledge of the operators and engineers about the process to formulate artificial intelligence algorithms.

1.1.1 State Estimation Approaches

The approaches using fundamental models are usually based on state estimation methods (Jazwinski, 1970), which combine a fundamental model of the process with on-line measurements in order to provide on-line, recursive estimates of the underlying theoretical states of the batch process (Iserman, 1984; Schuler and De Haas, 1986). The fundamental model usually consists of a set of nonlinear differential equations which describe the deterministic part of the process.

The simplest approach to monitoring is to formulate a state estimator based on the deterministic model that should be satisfied during normal operation. Generally, observers or Kalman filters are used to reconstruct the states and the outputs of the system, and then tests on the output errors or innovations are used to detect faults (Willsky, 1976). If only a finite number of faults or events can occur, then several state estimators based on models incorporating different sets of plausible events can be seen in parallel. The most likely status of the process at any instant can then be evaluated using generalized likelihood ratio tests on the innovations (Basseville, 1988), or by computing the posterior probability of each model being valid. A high probability, of a particular state estimator being valid, would lead to an alarm and an indication of the probable cause. King (1986) presented an interesting use of this approach to detect hazardous batch reactor conditions, that could

lead to a runaway, by monitoring undesirable side reactions using temperature measurements and parallel Kalman filters.

In some industries, such as the aerospace industry, there often exist very good state models which not only describe the system, but also provide detailed representations of how model uncertainties, disturbances, and possible faults affect the system. In these situations robust Fault Detection and Isolation (FDI) methods have been developed (Patton et al., 1989), which transform the observer equations and decouple the system, so as to construct residuals which are affected only by the faults of interest. A bank of observers can then be used, where each observer is made sensitive to a different fault or group of faults, while being insensitive to common disturbances, modeling errors, and other faults. Tests to detect specific faults are then performed on these residuals.

An alternative approach that is well suited to chemical and biochemical batch reactors, which are subject to stochastic disturbances such as impurities and parameter variations, is to incorporate one's knowledge about these stochastic states into the state model and estimator. This step is also a key point in making the state estimator provide unbiased and robust estimates of the deterministic states (MacGregor et al., 1986). Monitoring the progress of batch processes then consists of tracking the development of important deterministic state variables (conversion, composition, particle size, molecular weight, etc.) with time, to check that they are following satisfactory trajectories. If unacceptable deviations in any of these states are detected, then not only can an alarm be given, but an assignable cause can usually be found in the behavior of the stochastic states (e.g. impurity concentrations have increased, or the heat transfer coefficient has dropped). Such an approach has been used by Kozub and MacGregor (1992b) to monitor polymer and latex property development in semi-batch emulsion polymerization. On-line energy balances have also been effectively implemented in this manner by using Kalman filters (MacGregor, 1986; De Valliere and Bonvin, 1989 and 1990; Bonvin et al., 1989; Schuler and Schmidt, 1992).

All these state estimation approaches to monitoring are "directional" in nature, in that they build into their models the possible faults or reasons for deviations from normal behavior. Although they are potentially very powerful approaches for monitoring batch processes, they present a number of problems in practice. Detailed theoretical models and on-line sensors related to product quality are necessary, if one wishes to track key quality states. Such detailed models and robust sensors are time consuming and difficult to develop, and the state estimation approaches are computationally intensive. The detection and diagnostic abilities of these estimators will also be highly dependent upon one's prior knowledge of the possible faults and disturbances that may occur, since these must be explicitly build into the estimator as part of the stochastic state vector, or included as plausible events in one of the parallel filters. Events or disturbances that are omitted from the model, may lead to biased estimates and faulty diagnosis, if they occur.

1.1.2 Knowledge Based Approaches

Knowledge based approaches use expert system and artificial intelligence methods to process the data. In rule based expert systems the process model is represented by a set of qualitative and quantitative governing descriptions based on the knowledge about the process available from operators and engineers. Associated with each behavioral description there is also a set of causality assumptions. These behavioral and causal descriptions are arranged in a hierarchical structure and diagnostic rules for each node in the hierarchy are generated from these descriptions (Ramesh et al., 1989; Birky and McAvoy, 1990).

Additional structure and quantitative analysis can be brought to these rule based expert systems through the use of probability theory or fuzzy logic, to represent the uncertainties associated with the models and the diagnostic hypotheses (Petti et al., 1990; Rojas and Kramer, 1992; Terpstra et al., 1992). An advantage of these approaches is that they do not require detailed theoretical models of the process. However, formulating the behavioral and causal descriptions, and the diagnostic hierarchy with its probabilistic or

fuzzy rules, may be just as difficult and time consuming as using the fundamental model approach. Recently, Fathi et al. (1993) incorporated state and parameter estimation modules within the diagnostic reasoning of a knowledge-based system, in order to overcome some of the deficiencies of both approaches and to increase the diagnostic ability of the system.

Another approach to the problem is through the use of artificial neural networks with various nonlinear functions (sigmoid, gaussian, wavelets, etc.) which have been demonstrated to be good pattern classifiers, and thus potentially capable of diagnosing faulty conditions (Venkatasubramanian and Chan, 1989; Himmelblau, 1992; Leonard and Kramer, 1992; Bakshi and Stephanopoulos, 1993). Their main drawback is that in order to develop such neural network classifiers, a training set must be available which contains an abundance of faults, something which is rarely available in real processes.

Most rule-based expert systems and neural network classifiers have been developed for continuous processes that are intended to operate at various steady states. In order to handle unsteady state processes, such as batch operations, a variety of methods has been developed recently for extracting information on temporal shapes or time profiles, and classifying process behavior based on this information using rule-based expert systems (Konstantinov and Yoshida, 1992; Holloway and Krogh, 1992; Cheung and Stephanopoulos 1992). They try to answer the question if the on-line observations received from the process up to the present time, are consistent with some acceptable dynamic behavior of the system. Their main drawbacks are their lack of a statistical basis for interpreting the data and classifying the results, and the complexity of the approaches when dealing with more than a few variables.

1.2 SPC in Batch Processes

The application of SPC charts to batch processes has been very limited. The established SPC charts like the Shewhart charts (Shewhart, 1931), CUSUM charts (Woodward and Goldsmith, 1964), and EWMA charts (Roberts, 1959) are inappropriate

for multivariate problems where the variables are correlated (Woodhall and Ncube, 1985; Jackson, 1980; Harris and Ross, 1991). Another major difficulty in applying SPC to batch processes arises from their dynamic nature. Most SPC methods utilize only the product quality measurements obtained at the end of each batch (e.g. Vander Wiel et al., 1992), and therefore monitor only the batch to batch variation. Hahn and Cockrum (1987) investigated the case where one has also a few quality measurements taken during the batch run. Marsh and Tucker (1991) recognized that the process variable measurements taken during a batch run, although transient in nature, do follow a certain dynamic pattern, and they proposed a simple SPC technique for monitoring a single measurement variable. Bonvin and Rippin (1990) have used target factor analysis to identify possible reaction stoichiometries from measured composition or thermal data, and to detect in real-time any changes in stoichiometry which may lead to a batch runaway (Prinz and Bonvin, 1992).

With on-line computers connected to most batch processes, massive amounts of data are being collected routinely during the batch on a large number of easily measured process variables such as temperatures, pressures, and flowrates. Although, it is not unusual to measure more than twenty variables around a batch process, this does not mean that twenty independent things are taking place. Only a few underlying events are driving a batch at any time, and all the measurements are just different manifestations of these same underlying events. The major difficulties are how to handle the large number of measured process variables, their time varying and highly correlated structure, and the nonlinear finite time nature of batch operations. Furthermore, not only is the relationship among all the variables at any one time important, but so is the entire past history of the trajectories of all these variables. To accommodate this kind of data, multivariate statistical projection methods based on Principal Components Analysis (PCA) and Partial Least Squares (PLS) are used to compress the data and to extract the information in them. The proposed schemes represent extensions of the multivariate procedures for monitoring continuous processes (Kresta et al., 1991; Slama, 1991; Skagerberg et al., 1992; MacGregor et al., 1994), to the nonlinear and finite duration batch processes. Moteki and Arai (1986) were

probably the first who used multivariate statistical analysis on measurements from a low-density polyethylene process to find new operating conditions.

The variation in the trajectories among a historical reference distribution of normal batches (common cause variation) is characterized by projecting the data into a low dimensional space that summarizes both the variables and their time histories during successful batches. The history of the process variable trajectories during a batch provide a "fingerprint" for each batch, and from these data an empirical model is built which characterize the operation of successful batch runs. The approach is based on the philosophy of Statistical Process Control (SPC), under which the behavior of the process is characterized using data obtained when the process is operating well and is in a state of control. The behavior of new batches is then compared against this in-control model and its statistical properties to test the null hypothesis:

H_0 : The on-line measurements of the process variable trajectories up to the current time in a new batch are consistent with normal batch operation as defined by the historical reference distribution.

This approach leads to the development of some new multivariate SPC control charts whose presentation and interpretation are no more difficult than conventional Shewhart charts, and yet they are much more powerful in their ability to detect even subtle changes in batch process operations.

The objective of the monitoring procedure is to detect and eliminate faults from future appearance, and thereby shrink the control limits and work towards a more consistent production of quality product. The approach is "non-directional" in that it will detect any deviation from normal behavior, and may not be as powerful as the state estimation and knowledge based approaches in detecting those specific faults that are built into their models. On the other hand, it is not sensitive to the assumptions of these "directional" approaches, and the only information needed to develop the monitoring procedure is a historical database of past successful batches. As with most "non-directional" SPC procedures no assignment as to the cause of the event is provided. Once

a significant deviation from normal operating performance has been detected, it is usually left to the engineers and operators to use their process knowledge to provide a quick diagnosis of possible causes, and to respond in an appropriate manner. However, multivariate statistical procedures such as PCA and PLS provide much more diagnostic information about an abnormality (Wise and Riker, 1991; Miller et al., 1993; MacGregor et al., 1994; Kourti and MacGregor, 1994).

As in all inferential approaches, the fundamental assumptions of "comparable" runs and "observable" events of interest must hold for the method to work. The first assumption states that the method is valid as long as the reference database is representative of the process operation. If something changes in the process (e.g. new catalyst), then one has to build a new database which embodies the change and reapply the method. The second assumption expresses the requirement that the events which one wishes to detect must be "observable" from the measurements that are being collected. No monitoring procedure can detect events that do not affect the measurements.

Both simulated and real industrial data are used to illustrate the proposed methods. Chapter 2 provides an introduction of how the on-line measurements from past batches are organized in a three-way array ($\underline{\mathbf{X}}$), and how Multi-way PCA (MPCA) discriminates between normal and abnormal batches in a post analysis of these data. In Chapter 3 the selection of a suitable historical reference distribution of past normal batches is illustrated and modeled via MPCA. New batches are contrasted against this reference MPCA model and their abnormalities are revealed. The on-line monitoring procedure is outlined in Chapter 4 along with its control charts. Chapter 5 covers how the proposed monitoring schemes can take into account any other information available for a batch run. Multi-way PLS (MPLS) is used to incorporate the end product quality data ($\mathbf{Y}$) which are available upon completion of the batch. Rather than focusing only in the variance in $\underline{\mathbf{X}}$, MPLS focuses more on the variance of $\underline{\mathbf{X}}$ that is more predictive for the product quality $\mathbf{Y}$. Monitoring with MPLS provides on-line predictions of the end product qualities. When additional information about the initial conditions and set-up of the batch process is

available, Multi-block MPCA or MPLS can be used to integrate these data ($\mathbf{Z}$) into the proposed monitoring schemes. Chapter 6 discusses the issue of comparing any two reduced spaces, and investigates the possibility if the same projection model can be used for similar designed batch reactors. Finally, Chapter 7 summarizes the conclusions and areas for further work.

CHAPTER 2 Analysis of Batch Data Using MPCA

In this chapter the application of MPCA to batch data is presented, and an understanding of the advantages and limitations of the method is illustrated through one simulated and one industrial example. It gives a brief description of Principal Components Analysis, and assumes that the reader has some basic knowledge both on projection methods and on statistics.

2.1 Principal Components Analysis

PCA is a well known multivariate statistical method (Mardia et al., 1989; Jackson, 1991), which has as its objective the explanation of the variance-covariance structure of a multivariate dataset, through a few linear combinations of the original variables with special properties in terms of variances. It decomposes a mean centered matrix $\mathbf{X}$ (I×M), where there are M measurement variables for I objects, into a series of R principal components ($\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r' = \mathbf{TP}'$), with each characterized by a loading vector ($\mathbf{p}_r$) and a score vector ($\mathbf{t}_r$). The principal components represent the selection of a new coordination system obtained by rotating the original variables. The objects are projected into the reduced space defined by the principal components where the information in the data is described adequately and in a simpler and more meaningful way. The principal components are ordered such that the first one describes the largest amount of variation in the data, the second one the second largest amount of variation, and so on. With highly correlated variables, one usually finds that only a few principal components are needed to explain most of the significant variation in the data.

The loading vectors ($\mathbf{p}_r$) are orthonormal ($\mathbf{P'P=I}$), and provide the directions with maximum variability. The t-scores ($\mathbf{t}_r$) give the coordinates for the objects in the reduced space. They are orthogonal ($\mathbf{T'T}$ is a diagonal matrix) and therefore are measuring different and uncorrelated underlying "latent structures" in the data. By plotting the t-

scores of one principal component versus another, one can easily see which of the objects have similarities in their measurements and form clusters, and which are isolated from the others and therefore are unusual objects or outliers.

The power of PCA arises from the fact that it provides a simpler and more parsimonious description of the data covariance structure than the original data by summarizing the data in terms of only a few (R) principal components. PCA has been applied to a broad spectrum of sciences, like biology, psychology, chemistry, quality and process control, and economics, revealing its robustness and potential strength to analyze large datasets. Its success in several different areas is partially related to the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Geladi and Kowalski, 1986; Wold et al., 1987a) for the calculation of the principal components. It is a simple, fast and effective algorithm to extract the principal components in a sequential manner (eigenvalues in descending order), and is a variant of the power method for calculating eigenvectors of a matrix (Goldberg, 1991). The loading vectors $\mathbf{p}_r$ are the normalized eigenvectors of the sample covariance matrix $\mathbf{X}'\mathbf{X}(n-1)^{-1}$, and $\mathbf{t}_r'\mathbf{t}_r(n-1)^{-1}$ are their corresponding eigenvalues. There are also other ways to compute the principal components such as the Singular Value Decomposition (Golub and VanLoan, 1989) or the QZ algorithm (Moler and Stewart, 1973), as well as modifications of the NIPALS for large matrices (Lindgren et al., 1993). In this work the NIPALS was chosen for its simplicity. This algorithm will also make easier to understand the development of the on-line monitoring in Chapter 4.

2.2 Multi-way PCA in Batch Processes

In many cases, especially in sciences like chemistry, psychometrics and image analysis, the data from an experimental study takes the form of three-way arrays. This is also the case in the batch monitoring problem. To understand the nature of the data available with which to monitor batch processes, consider a typical batch run in which j=1,2,...,J variables are measured at k=1,2,...,K time intervals throughout the batch. Similar data will exist on a number of such batch runs i=1,2,...,I. This vast amount of data

can be organized into a three-way array $\underline{\mathbf{X}}(I \times J \times K)$ as illustrated in Figure 2.1. In this thesis, bold underline capital symbols denote three-way arrays (see also Notation). The different batch runs have been arranged along the vertical axis, the measurement variables across the horizontal axis, and their time evolution occupies the third dimension. Each horizontal slice through this array is a $(J \times K)$ matrix containing the trajectories of all the variables from a single batch (i). Each vertical slice is a $(I \times J)$ matrix representing the values of all the variables for all the batches at a common time interval (k).

Several multi-dimensional statistical methods have been proposed for decomposing such data arrays into the sum of a few products of vectors and matrices, and to summarize the variation of the data in the reduced dimensions of these spaces. MPCA was introduced by Wold et al. (1987b) and has been successfully applied in image analysis (Geladi et al., 1989), and to some cases in chemometrics (Smilde and Doornbos, 1991). Other multi-way methods (Geladi, 1989; Smilde, 1992) such as the Tucker model, the PARAFAC model, the canonical decomposition, the three mode factor analysis (Zeng and Hopke, 1990) and the tensor rank (Sanchez and Kowalski, 1990) have been proposed for special situations. Although each of them has interesting mathematical aspects, the interpretation is complicated due to the mathematical principles used to develop them. MPCA because of its simplicity and well defined properties proved to be a very effective and easy to understand statistical method for analyzing batch data. In any case, neither MPCA nor any of the others methods has ever been used for a problem where one of the dimensions is a factor like time which gives in the variables a strong dynamic behavior.

MPCA is equivalent to unfolding the three dimensional array $\underline{\mathbf{X}}$ slice by slice (three possible ways), rearranging the slices into a large two dimensional matrix $\mathbf{X}$ (two possible ways), and then performing a regular PCA. Each of these six possible rearrangements (two are degenerate cases) of the data array $\underline{\mathbf{X}}$ into a large data matrix $\mathbf{X}$, followed by a PCA on the matrix $\mathbf{X}$, corresponds to looking at a different type of variability. For analyzing and monitoring batch processes, the most meaningful way of unfolding the array

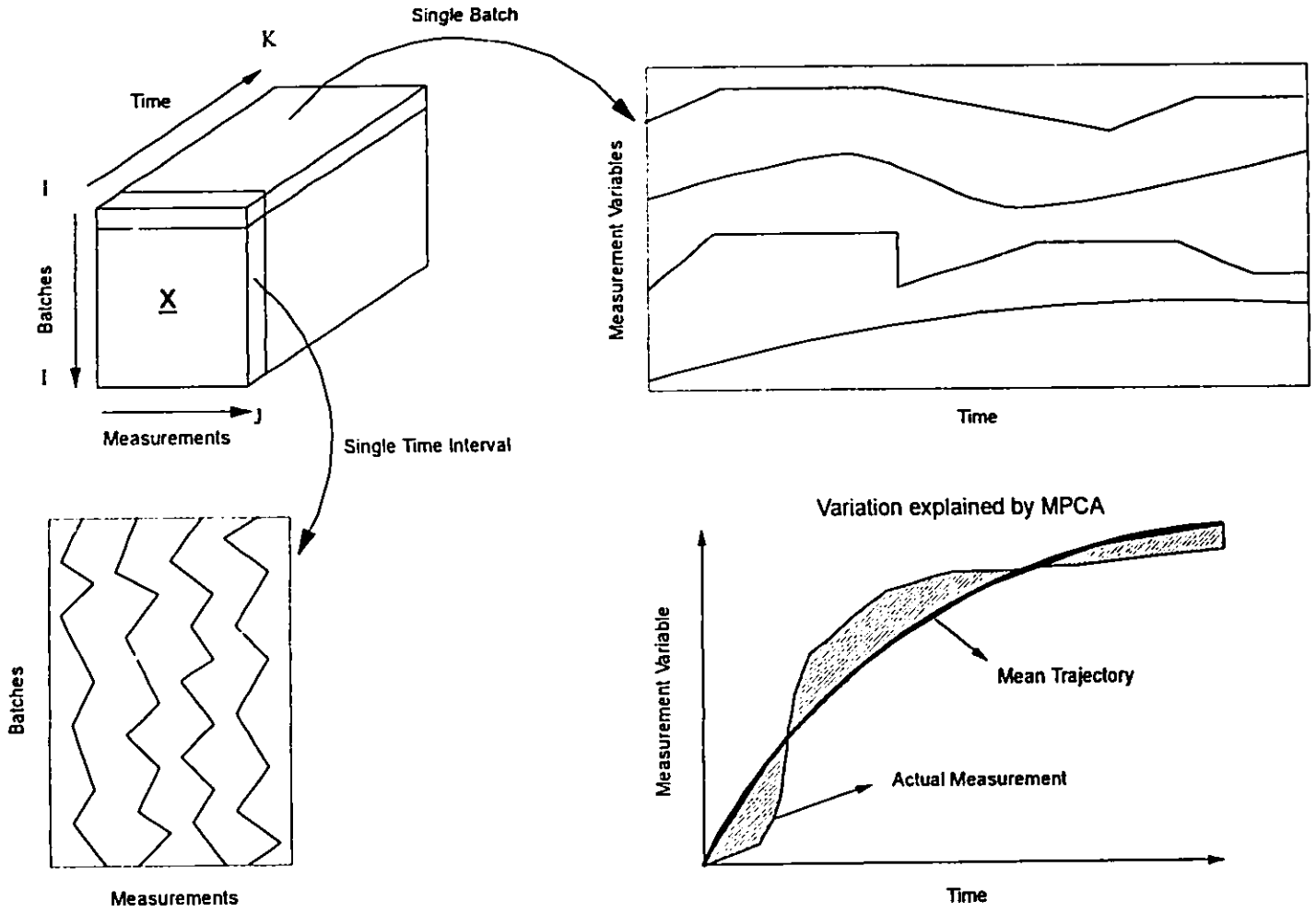

Figure 2.1 Arrangement of batch data in MPCA. Each horizontal slice contains the measurement trajectories from a single batch. Each vertical slice has the measurements of all the batches at a single time interval. MPCA extracts the variation in these vertical slices and study how the measurements deviate from their mean trajectories.

$\underline{X}$ is to arrange its vertical slices, corresponding to each point of time, side by side into a two dimensional matrix $X(I \times JK)$ with the vertical slice corresponding to the first time interval at the left hand side as it is shown in Figure 2.2. The data are then mean centered and scaled prior to performing a PCA. This unfolding is particularly meaningful because by subtracting the mean of each column of this matrix X, we are in effect subtracting the mean trajectory of each variable, thereby removing the main nonlinear and dynamic components 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 batches about their mean trajectories (Figure 2.1). This allows us to analyze the variability among the batches

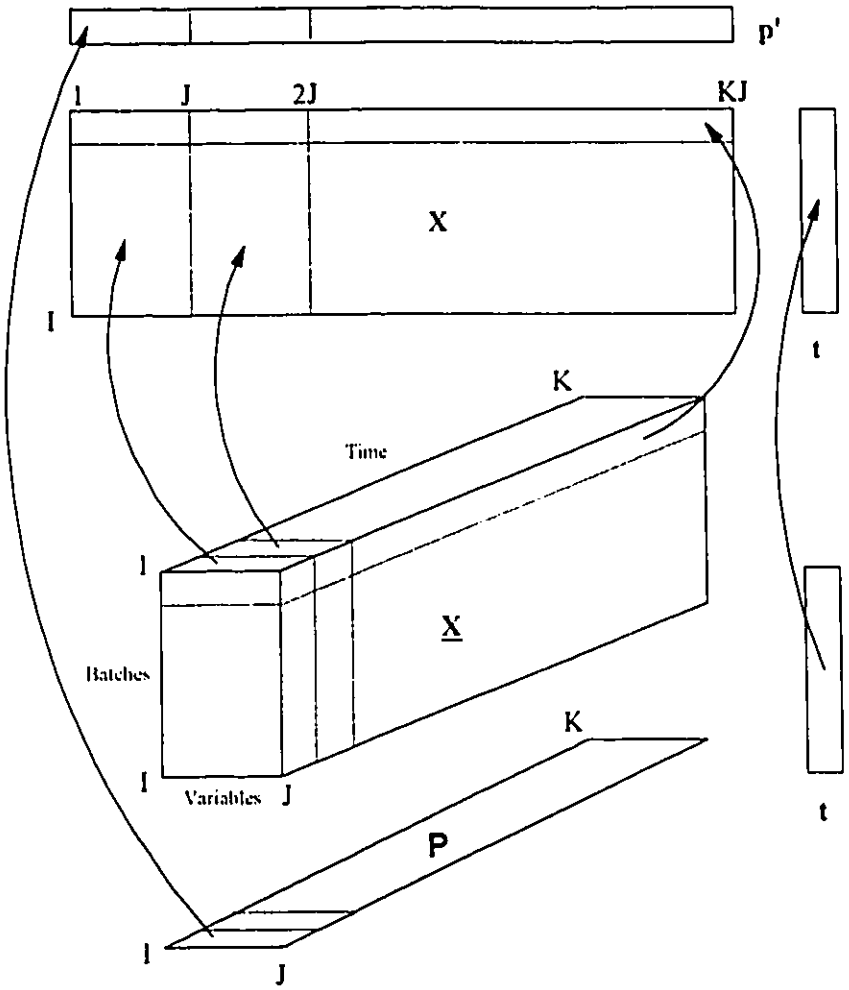

Figure 2.2 MPCA and its equivalent PCA form for batch data. The three-way array $\underline{X}(I \times J \times K)$ unfolds into a matrix $X(I \times JK)$ where a normal PCA analysis is performed.

in $\underline{\mathbf{X}}$ by summarizing the information in the data with respect both to variables and their time variation. The only other meaningful unfolding of $\underline{\mathbf{X}}$ is to arrange its horizontal slices, corresponding to each batch, one below the other into a two dimensional $\mathbf{X}(IK \times J)$ where the first K rows are the measurements from the first batch in the database. A PCA performed on this unfolded matrix is a study of the dynamic behavior of the process about the overall mean value for each variable. Although this variation might be of interest in some situations, it is not the type of variation of interest in SPC of batch processes.

The variables in each column of $\mathbf{X}$, after they are mean centered, are also scaled to unit variance by dividing by their standard deviation so to handle differences in the measurement units between variables and to give equal weight to each variable at each time interval. However, if one wishes to give greater or less weight to any particular variable, or to any particular period of time in the batch, these weights are easily changed. Another way of scaling which gives similar results to what we use in this paper, is to scale each variable at each time interval by its overall (throughout the batch duration) standard deviation. The benefit from such scaling is that periods with more variability with respect to other periods of the same measurement variable will be weighted more and will have a greater influence in the MPCA model. But if the variability in a particular period is very large, this will result in the rest of this variable's history being ignored in the MPCA model.

The proposed form of MPCA decomposes the data ($\underline{\mathbf{X}}$ or $\mathbf{X}$) into a summation of R products of score vectors ($\mathbf{t}_r$) and loading matrices ($\mathbf{P}_r$ or $\mathbf{p}_r$), plus some residuals ($\underline{\mathbf{E}}$ or $\mathbf{E}$) which are as small as possible in a least squares sense.

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{t}_r \otimes \mathbf{P}_r + \underline{\mathbf{E}} \qquad \text{or} \qquad \mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r' + \mathbf{E} = \mathbf{T}\mathbf{P}' + \mathbf{E}$$

where $\mathbf{P}_r(K \times J)$ is the folded matrix of the loading vector $\mathbf{p}_r(KJ \times 1)$, and the three-way matrix operation $\otimes$ is defined as $\underline{\mathbf{X}}(i,j,k) = \sum_{r=1}^{R} \mathbf{t}_r(i)\mathbf{P}_r(k,j)$. Figure 2.2 illustrates the correspondence between the scores and loadings of MPCA performed on the array $\underline{\mathbf{X}}$ and

those of a PCA performed on the equivalent unfolded matrix **X**. The NIPALS algorithm for sequential computing the dominant principal components is given below:

i. unfold $\underline{\mathbf{X}}(I \times J \times K)$ into $\mathbf{X}(I \times JK)$

ii. scale **X**

iii. choose a column of **X** as **t**

iv. **p=X't**

v. **p=p/|p|**

vi. **t=Xp**

vii. if **t** has converged then go to step viii, otherwise go to step iv

viii. **E=X-tp'**

ix. go to step iii with **X=E** to extract the next principal component

Usually, a small number of principal components can express most of the variability in the data when the variables are highly correlated. In the case of batch data, the measurement variables are highly cross-correlated with one another and highly auto-correlated over time. The principal components can point out any similarities and dissimilarities among batches. The batches can be compared with an MPCA analysis by plotting their t-scores and their sum of squared errors.

$$Q_i = \sum_{c=1}^{KJ} \mathbf{E}(i,c)^2$$

The t-plots represent the projection of each batch history onto the reduced plane defined by the principal components, while the Q-plot represents the squared distance of each batch perpendicular to this hyperplane. Each element of the t-vector corresponds to a single batch and depicts the overall variability of this batch with respect to the other batches in the database throughout the whole batch duration. The p-loading vectors provide the directions of maximum variability and give a simpler and more parsimonious description of the covariance structure of the data. Each loading vector, as one can see in Figure 2.2, summarizes the time variation of the measurement variables about their

average trajectories. Its elements are the weights applied to each variable at each time interval within a batch and when multiplied by the observed deviations of the variables from their mean trajectories for one particular batch give the t-score for that batch. The power of MPCA results from using the joint covariance matrix of the variable deviations from their mean trajectories over the entire batch history. Thus, it utilizes not just the magnitude of the deviation of each variable from its mean trajectory, but also the contemporaneous correlation among all the variables over the time history of the process.

2.3 Examples of MPCA in Batch Data

Two examples are considered to illustrate how MPCA can be used to perform a post-analysis on completed batch runs so to discriminate between similar and dissimilar batches. Such an analysis can be used to improve operating policies, and to gain an understanding of some of the major sources of batch to batch variation. The first example is from a simulation study of a semi-batch emulsion polymerization reactor, and the second from real industrial data. The simulation has been used as a framework, where we have perfect knowledge and command of the process operation, to develop the ideas of the present statistical procedure, and to test out its abilities to detect simulated operational problems. The industrial data illustrate a real application of the method and reveals both its strengths and its drawbacks.

2.3.1 Styrene-Butadiene Simulation Example

The simulated data are based on the semi-batch emulsion polymerization of Styrene-Butadiene to make a latex rubber (SBR). The simulations were performed using a detailed mechanistic model developed by Broadhead et al. (1985) and improved by Kozub (1989). The batch is initially charged with seed SBR particles and with all its initiator, chain transfer agent, emulsifier, water and a small amount of styrene and butadiene monomers. Styrene and butadiene monomers were then fed to the reactor, at an approximately constant rate for the rest of the batch procedure. Impurities in the initial

charge of the organic phase and in the butadiene feed to the reactor, were added to introduce meaningful disturbances for the purpose of this study. The feedrates of the monomers were simulated as first order autoregressive models. After completion of the simulation measurement noise was added to the temperature measurement of the feeds. The batch duration is 1000 minutes, and 9 (J) measurements are taken every 5 minutes (yielding 200 (K) time intervals) on the flowrates of styrene (1) and butadiene (2), on the temperature of the feeds (3), the reactor (4), the cooling water (5), and the reactor jacket (6), and on the density of the latex in the reactor (7). Also, estimates of the total conversion (8) and the instantaneous rate of energy release (9) were obtained from an on-line dynamic energy balance around the reactor and jacket (MacGregor, 1986).

A base recipe was chosen and 50 (I) batches were simulated to create a reference database of normal batches by introducing typical variations in the base case conditions (initial charge of seed latex, amount of chain transfer agent, level of impurities, etc.). Details about the base case simulation and the variations from it can be found in Appendix A. The resulting latex and polymer properties (composition, particle size, branching, crosslinking and polydispersity) of these fifty batches were consistent with the variations one might see during a sequence of industrial batch runs. These quality measurements define the acceptable quality region of the product, and a successful batch is taken to be one which falls under three standard deviations around the mean for each quality measurement.

Two additional batches with product just barely out of this specification region were simulated. One batch had an organic impurity contamination in the butadiene feed, 30% above that of the base case, right from the start of the batch; and the other had the same problem but this time the contamination, 50% above the normal level, started halfway through its batch operation. These unsuccessful batches, having the same cause for their abnormal operation, will help to investigate the ability of MPCA to detect a fault occurring at different times without confounding the results because of any differences in the cause of the abnormal operation.

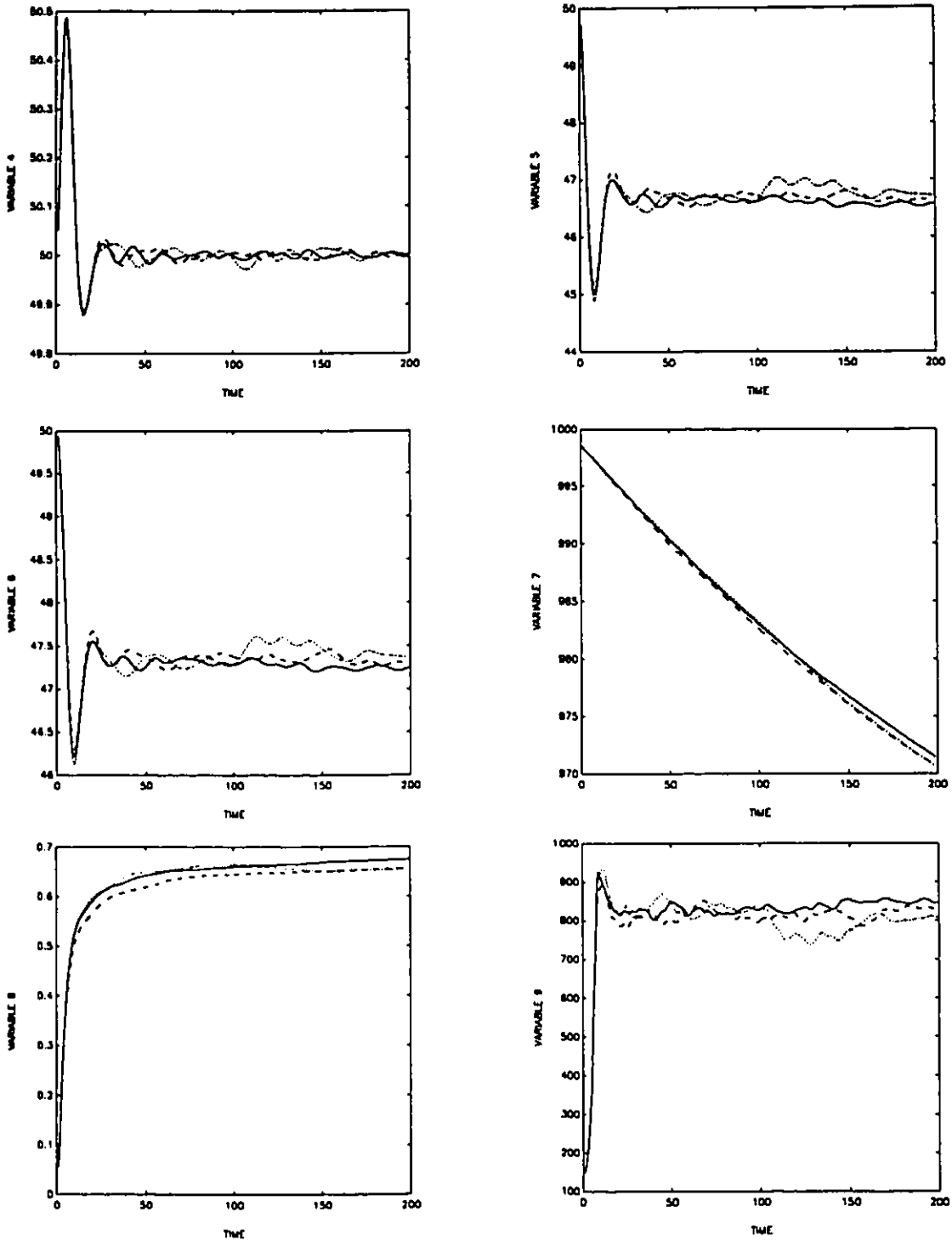

Figure 2.3 Measurement trajectories from the SBR example. The solid line is from a normal batch which will be used as a test batch in Section 4.4, the dashed line from the batch with the initial problem, and the dotted line from the batch with the problem halfway through its operation.

In both cases, there was not an abrupt fault which is often easily detectable by examining the plots of the individual variables. They were more incipient faults, typical in industry, where the abnormal operation is slowly developing and none of the individual measurements reveal clearly the fault. The trajectories for the individual variable measurements for the two bad batch runs and a normal one which will be used as a test batch in Section 4.4, are shown in Figure 2.3. One can see in this figure that there is not much observable difference among these runs. If all the trajectories from the database of the fifty good batches were plotted on this figure, any differences between these and the bad batches would be difficult to detect through a visual inspection of these individual plots.

By adding these two bad batches as the last (51 and 52) objects in the reference database, a MPCA was applied to the three-way array $\underline{\mathbf{X}}$ with dimensions 52×9×200. The projections of these 52 batches into the score plane of the first two principal components (t_1,t_2) are shown in Figure 2.4. This plot shows that the two faulty batches fall well outside the cluster of good batches, clearly indicating that their temporal development was different. All batches which have a similar history should cluster in the same region of the reduced space described by the principal components. The sum of squared errors (Q) in Figure 4 shows that there is no other batch with significant differences in its behavior.

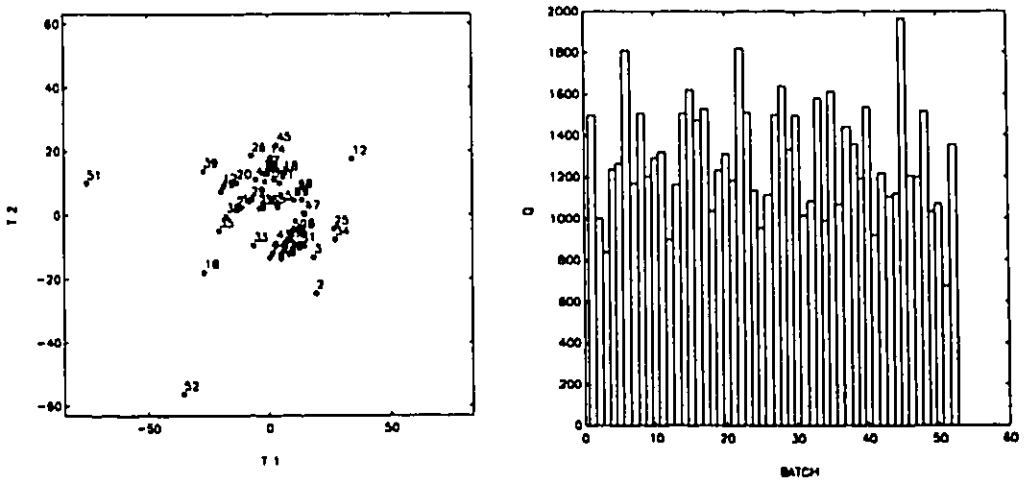

Figure 2.4 MPCA is able to discriminate the two abnormal batches (51 and 52) in the SBR example.

2.3.2 Industrial Example

Data supplied by DuPont US from an industrial batch polymerization reactor, are used to illustrate a real application of the proposed method. The cycle in the reactor consists of two stages and the time spent processing in both stages is approximately two hours. Ingredients are loaded into the reactor to begin the first stage. Reactor heating medium flows are adjusted to establish proper control of pressure and the rate of temperature change. The solvent used to convey ingredients to the reactor is vaporized and removed from the reactor vessel. The vaporization process is vigorous enough so that the contents of the vessel do not require stirring. After nearly one hour spent removing solvent, the second stage begins. During this stage the ingredients complete their reaction to yield the final product, a polymer. Once again, vessel pressures and the rate of temperature change are controlled during this processing stage. The batch finishes by pumping the polymer product from the vessel at the end of the second cycle.

The results of two critical property measurements are usually received twelve hours or more after each batch has finished, and therefore there is no time for recipe adjustments in the next batch. Furthermore, it is often difficult to establish when a batch is going wrong, and to diagnose what caused the properties to deviate from their targets. This makes the application of an SPC monitoring method attractive for this process. Failure to attain on-aim control of the critical property leads to increased manufacturing costs either through necessitating blending with other batches, or through downgrading the product to end uses that have a lower selling price.

A dataset of 55 successful and some unsuccessful batches was provided from the above process. Each batch had a duration of 100 time intervals (K), and 10 measurement variables (J) were monitored throughout the batch. Variables 1,2, and 3 are temperature measurements inside the reactor, whereas variables 6 and 7 are temperature measurements in the heating-cooling medium. Variables 4,8, and 9 are pressure measurements and the rest of the variables are flowrates of materials added to the reactor during its operation. Batches 40,41,42,50,51,53,54, and 55 had the two final quality measurements well outside

the acceptable limit, and batches 38,45,46,49, and 52 were above or very close to these limits.

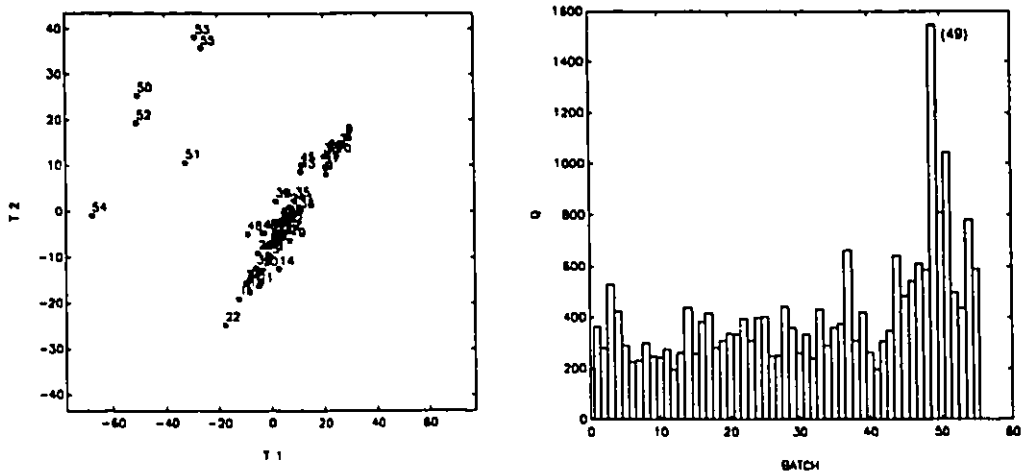

Figure 2.5 Results of MPCA analysis of the original 55 batches from the industrial example. Batches 50 through 55 are identified as abnormal because of their position in the reduced space, and batch 49 because of its residuals.

First, a preliminary MPCA analysis was conducted in order to identify if there was enough information in the process variable trajectory data to discriminate between normal and abnormal batches. Two principal components were extracted and Figure 5 shows that batches 50,51,52,53,54, and 55 are clearly identified from their position in the reduced space (t-plot) as being very different from the rest of the batches. In hindsight their different behavior can be seen by visual inspection of the measurement trajectories, if one superimposes these over the trajectories from normal batches on the same plot. However, due to their structural similarity (maxima, minima, points of inflection or discontinuity) with trajectories from normal batches, operators might easily see nothing different when they are displayed alone. Batch 49 also is identified as different due to its large residual (Q) in Figure 2.5. The quality of this batch was barely acceptable and its main difference compared to normal batches was during a time period (50-65 time intervals) when most of the measurement variables do not usually exhibit much variation (flat trajectories). The measurements of batch 49 are shown in Figure 2.6 where they are contrasted with the mean trajectories of the normal batches. The p-loadings of the first two principal compo-

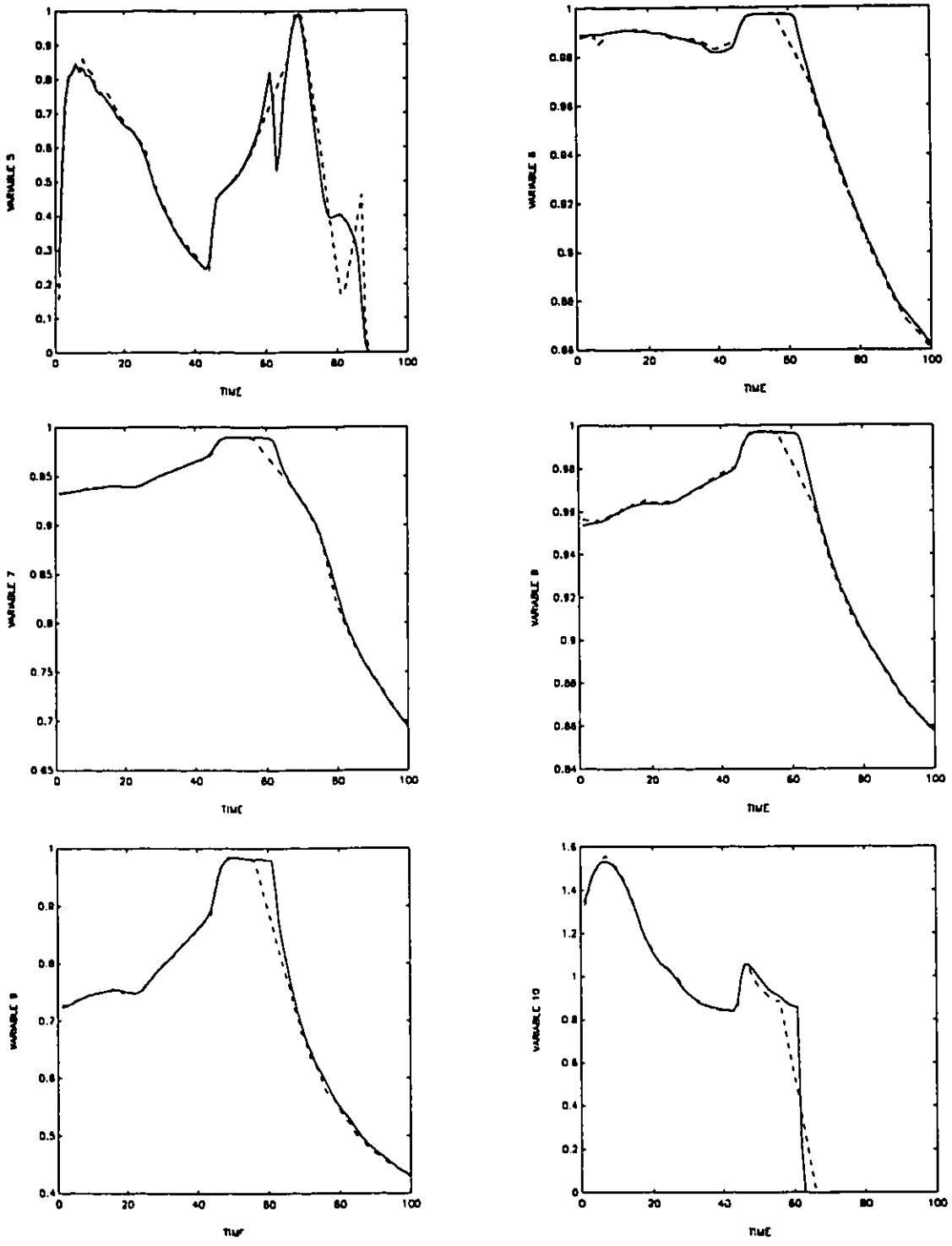

Figure 2.6 Measurements of batch 49 (dashed lines) are contrasted with the mean trajectories (solid lines) of the normal batches in the industrial example.

-nents during this time period are small and so any difference shows up in the residuals. This example with batch 49 points out the complementary nature of the t-scores and residuals. Any variation that is not explained by the current principal components, is contained in the residuals and it will be explained in one of the later components.

A second MPCA analysis was run, this time excluding these last six batches. The results from this analysis after three principal components are given in Figure 2.7. Batches 38,40,41, and 42 can not be identified as abnormal batches. There is nothing unusual in their trajectories to be captured by MPCA, since the cause of their unacceptable product does not seem to have any affect on the measured trajectories. This shows that there are cases where the measurements are typical of a successful batch and still the product may not meet the performance standards. The problems in batches like these may have come from poor quality materials, inadequate preprocessing, or from something that can be observed only if one measures other on-line variables as well. Again another group of batches clustered away from the main body of batches in the t_2-t_3 plot. Further investigation should be carried out to determine what went wrong with these batches since most of them are in a sequence (43 through 49) and two of these (45,46) gave unacceptable product. Note that the differences in the above batches could not be detected

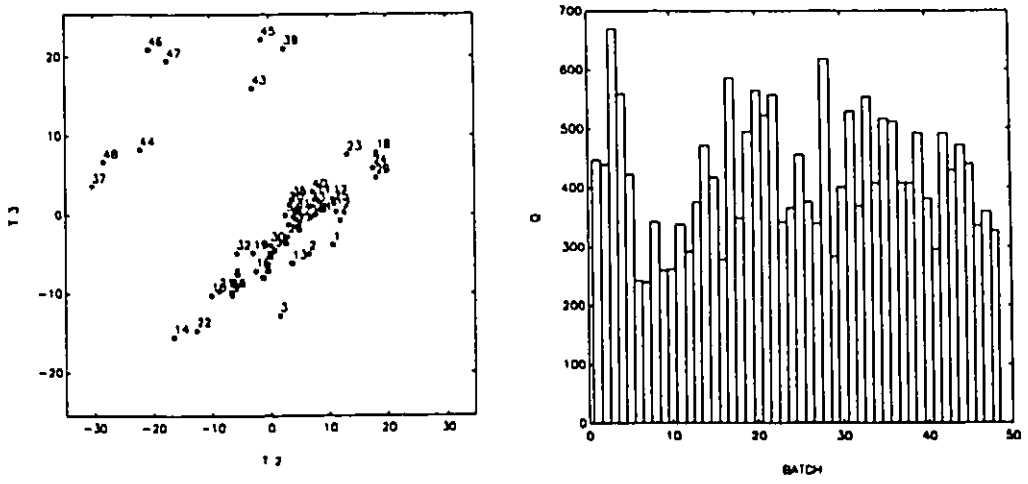

Figure 2.7 MPCA analysis of the first 48 batches from the industrial example. Batches 37, 39, and 43 through 48 can be identified as batches with unusual operational behavior.

by a simple visual inspection in the measurement variables. The residuals in Figure 2.7 suggest that there are no other major differences in the operational behavior of the batches.

2.4 Discussion

The power of the statistical approach presented here lies in the fact, that it utilizes the unsteady state trajectory data on all measured variables in a truly multivariate manner, as to account not only for the magnitude and trend of the deviations in each measured variable from its average trajectory, but also for the high degree of correlation in both time and among the deviations in all the variables. The two examples demonstrate the ability of MPCA to discriminate between normal and abnormal batch operation through the use of simple to interpret plots, and to detect any systematic variability in the measurements of a batch operation. The next chapter will present methods for identifying the time periods and the measurement variables which contribute most to the detected abnormality.

It is important to clarify here the use of the words normal and abnormal for a batch. MPCA characterizes a batch as normal if its behavior, as this is depicted in its measurements, is similar to that of a typical successful batch; i.e. a batch with acceptable product. An abnormal batch is characterized as one with measurements different from a typical operation. If there is a relationship between the on-line measurements and the final product quality, then MPCA should be able to discriminate between successful and unsuccessful batches. Consequently, MPCA answers the question if the on-line measurements contain sufficient information for classification of a batch as successful or unsuccessful. This is the most important question that must be answered first in any SPC scheme, and MPCA provides a simple and easy way to answer it.

There will be always cases where either one batch with normal behavior gave unacceptable product, or a batch with abnormal operation gave acceptable product. In the first case, the abnormality was not observable in the on-line measurements and one has to find new measurements for the abnormality to become identifiable. In the second case, one

has to investigate why there was no impact in the product quality which may lead to new designs and policies. Some variables at certain time periods during the batch may have little or no effect on product quality. However, an SPC scheme that detects all abnormal behaviors in the process variables may give an alarm for an incipient equipment failure such as an agitator or sensor deterioration. This way, one will have the opportunity to correct such process deteriorations which otherwise could lead to permanent malfunctions. MPCA provides a useful mean of augmenting knowledge gained from the final quality measurements for assessing whether or not a past batch is a normal one. This is a rather attractive means of characterization, if one considers how difficult it is to obtain quality measurements and the uncertainty which is involved in them.

Another interesting concept is the possibility of diagnosing the cause of a fault by utilizing the t-plots which show the position of a batch in the reduced space. Sometimes the type of fault that occurred in an abnormal batch is often not known. In other situations, such as in another industrial dataset from DuPont US, there was knowledge of what went wrong for each unsuccessful batch. The data were from an identical process which runs in parallel with the one described in Section 2.3.2. An MPCA analysis showed that unsuccessful batches with similar faults were clustered together in the t-plot of Figure 2.8. In such cases a diagnostic map may be constructed and areas identified which have a high probability of a particular fault. Then each time a new batch is projected into the reduced space, as will be shown in the next chapter, a plausible fault can be identified with the aid of statistical discriminant analysis (James, 1985). Although it is an appealing way of diagnosis, it is impossible to guarantee precise diagnosis of the fault. This comes from the fact that the t-scores are linear combinations of the measurements, and thus there is an infinite number of batch histories, although most of them are not physically feasible, which can place a batch at the same location in the reduced space. The SBR example illustrates this point where the two abnormal batches had the same cause of abnormality (impurity contamination in one of the feeds) but occurred at different periods of the batch process. These batches were placed apart in the reduced space (Figure 2.4), although they

had the same source of fault. MPCA discriminates between them simply because their measurement trajectories had different patterns arising from the different times at which the fault occurred. One had deviations right from the beginning, where the other had similar ones only after the completion of the first half of the batch. The same cause in different time periods characterized differently from MPCA. It is common in batch processes, the same cause of abnormality (such as impurities) occurring at different time periods to have different effects in the product quality. Thus, one shall characterize faults in batch processes based both on their cause and their behavioral process pattern.

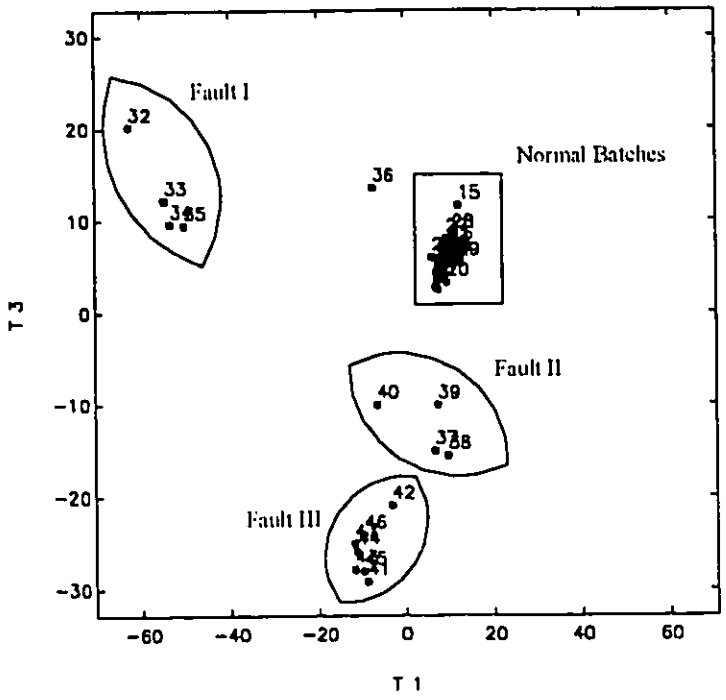

Figure 2.8 T-plot from the post analysis of an industrial dataset. Batches with similar faults clustered together in areas indicated by ellipses. The rectangular area at the center contains all the normal batches. Batch 36 had the same fault as batches 32 through 35 but in a smaller extent.

A difficulty with MPCA occurs when one encounters batch processes in which the duration of each batch or the timing for key events during each batch is different. An example of this occurs when various decisions during the batch are not automated, but left to the discretion of an operator. The batches are usually synchronized at time zero using a trigger variable whose change indicates the start of the batch. If the batch trajectory

shapes are similar from batch to batch and only the elapsed time required to achieve a given end point changes, then for post analysis one can often re-normalize the time scale so that all batches have the same duration. However, this is not feasible for on-line monitoring since the duration is not known a priori except in cases where one can easily develop rules to anticipate delays between different batch stages. One way to handle varying batch times in on-line monitoring is to replace time by another measured variable which progresses monotonically in time and has the same starting and ending value for each batch. Examples would be an on-line measure of conversion in a chemical reactor, or a measure of lance position in injection molding. Numerous other possibilities exist, but each is specific to the nature of the batch process. The industrial data was from a well automated process, and the only thing we had to do was to discard few observations in the original raw database prior to the start and after the end of the batch operation.

CHAPTER 3 Building the Reference Statistical Model

Chapter 3 discusses the selection of past batches which will compose the reference distribution of normal batches. The MPCA analysis of such a reference database provides the statistical model which describes the normal operation of a batch. New batches are contrasted against this MPCA model, and classified as normal or abnormal. Contribution plots reveal the time periods and variables primarily responsible for a detected abnormal behavior.

3.1 Reference Distribution of Normal Batches

Having established the observability of faults in the post analysis of past data, the next step is to build an MPCA model which summarizes the information contained in a database of good batches about the normal operation of the process, and to use this model as the statistical reference framework to classify a new batch as normal or abnormal. This classification will then be based on the similarity and statistical consistency of the trajectory measurements of a new batch with the historical reference distribution of trajectories from normal operation as summarized by this model.

The reference distribution is the history of past successful batch runs; that is those batches which produced good quality product, and exhibited no unusual faults or operating problems during their progress. In fact, one way of selecting the reference distribution is to take the history of previous batch runs, and omit all those batches that one would like the monitoring scheme to have pinpointed as being different. In general, the reference distribution should contain all those batches deemed to be subject only to common cause variation. The central idea is to use the statistics of this reference distribution to characterize the normal operation of the process, and to evaluate the behavior of new batches.

In the simulated SBR example, the fifty successful batches ($\underline{\mathbf{X}}(50\times9\times200)$) that will constitute the reference distribution of normal batches were determined by the

simulation. On the other hand in the industrial example, the preliminary MPCA analysis on Chapter 2 revealed that the first 36 batches ($\underline{\mathbf{X}}(36\times10\times100)$ should be selected as the reference distribution of normal batches. All the other batches should be excluded because either they had problematic operation, or gave unacceptable product. Problematic operation and unacceptable product usually coincide, but there are cases, as in this industrial example, where either a batch has some peculiar operation and still gives acceptable product, or the product is unacceptable without major deviations from a typical operation. Both cases should be excluded from the reference distribution. In the first case, one wants to detect any peculiar operation for investigation and eliminate it by implementing new design or operating policy improvements. In the second case, the reference database should not include any variation that may lead in a product out of specifications.

3.2 Selecting the Number of Principal Components

The number of principal components (R) needed to build an MPCA model which describes the major variation of all the variables about their mean trajectories and hence to provide a model which depicts adequately the normal behavior of a batch operation can be found with a number of criteria. These criteria range all the way from significance tests to graphical procedures (Jackson, 1991). One quick and dirty criterion is the broken stick rule (Jolliffe, 1986). This is based on the fact that if a line segment of unit length is randomly divided into z segments, the expected length of the rth longest segment is

$$G = 100\frac{1}{z}\sum_{i=r}^{z} 1/i$$

As long as the percentage of variance explained by each principal component is larger than the corresponding G, then one can retain the corresponding principal component. The number of segments is the maximum possible rank of $\mathbf{X}$, z=min(I,KJ), and the rule should be applied only to unit variance scaled matrices. This criterion is rather crude, but still is a

quick method to judge if a principal component adds any structural information about the variance in the data or explains only noise.

When the purpose of a PCA analysis is to construct a model which will be used to predict future observations, then the suggested criterion to obtain the optimum number of components is cross-validation (Stone, 1974; Efron, 1983 and 1986). Cross-validation shows how the prediction power of a PCA model increases as one adds more principal components. It is a simple, but computational lengthy, procedure similar to the jackknifing method. Given a database of I normal batches with J variables and K time intervals, the unfolded **X** matrix has dimensions I×JK. After scaling the matrix **X**, one batch is excluded from the database and a PCA model is built with the remaining (I-1) batches. This is done for all the batches in the database, and each time the sum of the squared prediction errors after each principal component is recorded for the batch not included in the model building. At the end these sum of the squared prediction errors corresponding to each principal component (r) are added for all the batches to give the $Press_r$.

One way to choose the model dimension is to select the one with minimum Press, but this has been shown to have poor statistical properties (Osten 1988). Wold (1978) and Krzanowski (1983, 1987) have proposed two criteria for choosing the optimal number of principal components. Wold checks the ratio

$$R = Press_r / RSS_{r-1}$$

where RSS_r is the residual sum of squares after the rth principal component based on the PCA model which is built using the whole database. This criterion compares the prediction power of a model based on r principal components with the squared differences between observed and estimated data using r-1 principal components. A value of R larger than unity suggests that the rth component did not improve the prediction power of the model and it is better to use only r-1 components. Krzanowski suggests the ratio

$$W = ((Press_{r-1} - Press_r) / D_m) \; / \; (Press_r / D_r)$$

$$D_m = I + JK - 2r$$

$$D_r = JK(I-1) - \sum_{i=1}^{r} I + JK - 2i$$

where D_m and D_r are numbers indicating the degrees of freedom required to fit the rth component and the degrees of freedom remaining after fitting the rth component respectively. This statistic is similar to the F-test for the inclusion of an additional variable in a linear regression model. It gives the ratio between the improvement in predictive power by adding the rth component and the predictive value using all r components. If W is larger than unity, then this criterion suggests that it is worthwhile including the rth component in the model.

There is no sound statistical test for the cross validation procedure. The main problem is not knowing how many degrees of freedom one starts with nor how many one extracts with each component (eg. Box et al., 1973). Thus, the number of principal components needed in a PCA model should be based on the overall picture that these criteria give.

3.3 MPCA Reference Model

Table 3.1 summarizes the results of extracting successive principal components from the reference database of the SBR and the industrial example. It gives the percent sum of squares explained and the three previously discussed criteria for selecting the required number of principal components. In both examples the broken stick rule (G) and the R statistic indicate that the first two principal components are significant and one may want to include the third principal component in the MPCA model as well. The W statistic suggests three components for both examples. Based on these results we chose to include three (R=3) principal components in both MPCA models as to take into account most of the predictable variation in the measurements during a normal operation. The fact that three principal components explain only about 30% in the SBR example and 55% in the industrial example of the total variability in the data, is not disappointing if one considers that these data describe the normal process operation and theoretically only random

variation should be in them. Also, one has to consider the large number of variables (JK, 1800 for the SBR and 1000 for the industrial example) in the unfolded matrix **X** from which he extracts the principal components. The rest of the variability that is not explained by the MPCA model is mainly due to measurement noise and to random variation in normal batch operation. Recall that in the SBR example three of the nine measurement variables (the two feed rates and the feed temperature) had mainly a stochastic character.

PC	%SSX	%SSX / PC	*G*	*R*	*W*
SBR example					
1	14.89	14.89	9.00	0.89	5.61
2	24.05	9.16	7.00	0.97	3.40
3	29.95	5.90	6.00	1.06	1.23
4	*35.08*	*5.13*	*5.33*	*1.13*	*0.73*
Industrial example					
1	38.55	38.55	11.59	0.65	17.58
2	50.22	11.68	8.82	0.91	5.35
3	56.75	6.53	7.43	1.04	2.24
4	*61.33*	*4.57*	*6.50*	*1.17*	*0.76*

Table 3.1 Percentage of explained sum of squares (cumulative and for each principal component) from the MPCA analysis of both examples, and the results from the three criteria to determine the optimal number of principal components.

Histograms, Lilliefors tests, chi-squared tests, and test for skewness and kurtosis (Lilliefors, 1967; D'Agostine and Stephens, 1986; Neave and Worthington, 1988; Horswell and Looney, 1992) showed that the t-scores of all principal components can be adequately approximated by a Multinormal distribution. This was to be expected since any linear combination of random variables, according to the Central Limit Theorem, should tend towards a Normal distribution. The formulation of the reference database contributes also to the normality of the t-scores. All these t-scores represent data from batches with similar behavior projected into the reduced space of the MPCA model. Their sample mean is 0 since they are linear combinations of mean centered variables, and depicts the position in the reduced space of the average normal operation. Thus, with the assumption that the t-scores follow a Multinormal distribution with population mean 0 and estimated

covariance matrix S (R×R) which is diagonal due to the orthogonality of the t-scores, one has the following Hotelling statistic (Tracy et al.,1992; Wierda, 1994) for the t-scores in the reference database:

$$D_S = t_R' S^{-1} t_R \; I / (I-1)^2 \quad \sim \quad B_{R/2,(I-R-1)/2}$$

where $t_R(R\times 1)$ is the vector containing the t-scores of a given batch from the R retained principal components. The critical values of the Beta variable at significance level α can be found from critical values of the F distribution by utilizing the relationship:

$$B_{R/2,(I-R-1)/2,\alpha} \cong (R/(I-R-1))F_{R,I-R-1,\alpha} \; / \; (I+(R/(I-R-1))F_{R,I-R-1,\alpha})$$

Accordingly, the 95% and 99% confidence ellipsoids in the t-plots are centered around zero and their axis lengths in the direction of the rth principal component are given by (Johnson and Wichern, 1988):

$$\pm \; (S(r,r)\, B_{1,(I-2-1)/2,\alpha}\, (I-1)^2 / I)^{1/2}$$

The D_S statistic gives a measure of the Mahalanobis distance in the reduced space between the position of a batch (t-scores) and the origin (0) which designates the point with the minimum variation in the batch process behavior.

Figure 3.1 shows the t-score plots with their 95% and 99% confidence ellipsoids in the space of the first two latent variables (t_1-t_2) for both examples. The score plots for the other two latent variables are similar. The scatter character of these plots indicates that all these batches belong to the same "normal" population, and none of them lies far away from the cluster, and thus exhibiting any unusual behavior. The area that this population occupies, defines the normal operational region in the reduced space; and the closer a batch is to the origin of this reduced space, the more similar is its operation to that of a typical batch run.

Batch 16 from the SBR example exhibits the highest variability in its measurements among all the batches in its database as it is shown in the t-plot and D_S-plot of Figure 3.1. Although its D_S statistic exceeds slightly the 99% confidence limits, it was included in the reference database. This batch had the highest acceptable level of organic impurities in its

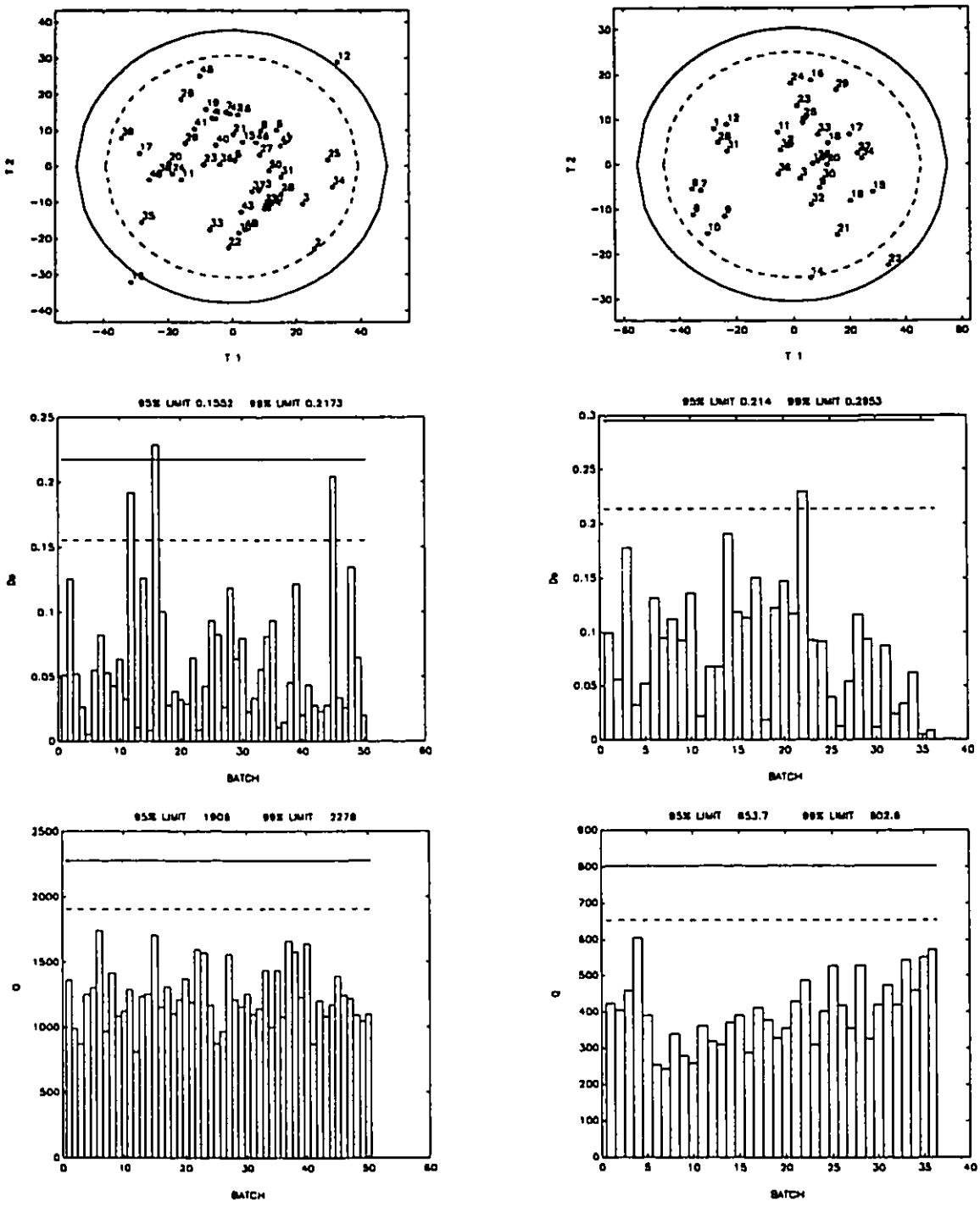

Figure 3.1 T, D_S, and Q plots with their 95% and 99% confidence limits from the MPCA analyses of the reference databases for both the SBR example (left hand side plots) and the industrial example (right hand side plots). None of the batches exhibits any notably unusual behavior.

butadiene feed in the reference database. The slight gap between two clusters in the t_1-t_2 plot of the industrial example (Figure 3.1) was not attributed to any significant difference, but simply to not having enough batches in this reference database to fill this gap and give a smooth variation across the first principal component.

The sum of squares of the residuals ($Q_i = \sum_{c=1}^{KJ} \mathbf{E}(i,c)^2$) from the MPCA models for both examples are given in bottom plots of Figure 3.1 with their 95% and 99% approximate confidence intervals. These Q_i values represent the squared perpendicular distance of the (J×K) dimensional point for the ith batch from the reduced space defined by the three principal components of the MPCA model. From Figure 3.1 it can be seen that all batches have been explained adequately, since none of them has a significantly large residual. The assumption behind these approximate confidence limits for Q is that the variables (JK) in the unfolded matrix **X** have a Multinormal distribution with population mean **0**. Again, this assumption is reasonable in our case since we have chosen a reference database of normal operating batches, we have avoided any outliers, and we have extracted the major predictable variation from the data in the retained three principal components.

Jensen and Solomon (1972), and Jackson and Mudholkar (1979) showed that the quadratic form Q=**e**'**e**, where **e** is an observation vector from a Multinormal population $N(\mathbf{0},\Sigma)$, has approximate upper confidence limits of significance level α:

$$\theta_1\left[1-\theta_2 h_\alpha(1-h_\alpha)/\theta_1^2 + z_\alpha(2\theta_2 h_\alpha^2)^{1/2}/\theta_1\right]^{1/h_\alpha}$$

$$\theta_1 = \sum\lambda_i \quad , \quad \theta_2 = \sum\lambda_i^2 \quad , \quad \theta_3 = \sum\lambda_i^3 \quad , \quad h_\alpha = 1 - 2\theta_1\theta_3/3\theta_2^2$$

where λ_i are the eigenvalues of Σ, and z_α is the critical value of the Normal variable at significance level α which has the same sign as h_α. In our case, we estimate the θ_i from the sample residual covariance matrix S_E=**E**'**E**/(I-1). Note that the matrices S_E(JK×JK) and **V**=**EE**'/(I-1) (I×I) have the same eigenvalues.

$$\hat{\theta}_1 = \text{trace}(\mathbf{V}) \quad , \quad \hat{\theta}_2 = \text{trace}(\mathbf{V}^2) \quad , \quad \hat{\theta}_3 = \text{trace}(\mathbf{V}^3)$$

The Q and D_S plots shown in Figure 3.1 appear to justify the normality assumptions that we have made and corroborate our choice of these databases to describe the normal batch operation in the two examples. These plots provide the diagnostics to test if we have included any unusual batch in our database, and if the MPCA model describes adequately the reference database. These "normal" statistical properties of the MPCA model are essential for the development of the on-line monitoring scheme in the next chapter.

It is informative to examine the percentage of the total variation in each variable and at each time, which is explained in each reference database by the principal components. The MPCA model attempts to explain all the predictable variation in the normal operating batch data. The left hand side of Figure 3.2 has the percentages of explained variance plotted on a cumulative basis against both time and variables for the SBR example, and the right hand side for the industrial example.

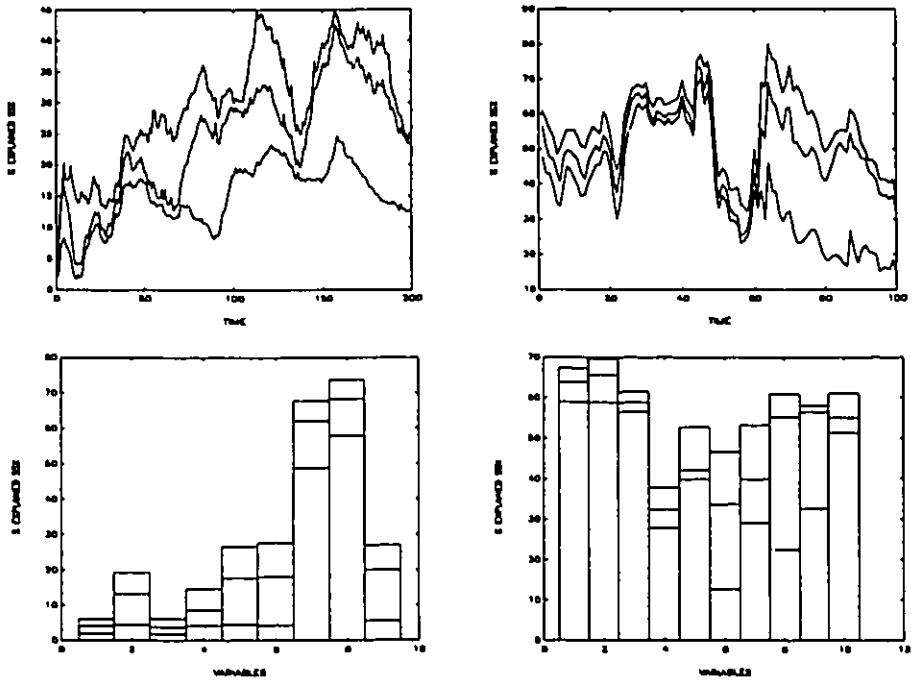

Figure 3.2 Cumulative percent of the total variation with respect to time and variables which is explained in each reference database. The plots in the left hand side are from the SBR example, and the plots in the right hand side are from the industrial example. Each line represents the cumulative percent of the total variance explained by the addition of each principal component.

In the SBR example one can see that the dominant variables in the first principal component are the density of the reactor's latex (7) and the total conversion (8). The temperatures around the reactor (5,6) and the rate of energy release (9) are explained mostly by the second and third principal component. Variables with mainly a stochastic character, such as the flowrate of the styrene (1), and the temperature of the feeds (3), are largely ignored by MPCA. Only the flowrate of the butadiene (2) contributes a little to the model, due to the fact that is carrying the incoming organic impurities which affect the behavior of all the other variables. The amount of variance accounted for, by each principal component versus time shows that on a relative basis, the second principal component concentrates, more than the other two principal components, on the variability at the last half part of the batch operation. The reason that the MPCA model does not explain a lot at the start of the batch operation, is because of the similar initial response of each batch in the database.

In the industrial example, the first principal component concentrates more on the first stage of the process where the solvent vaporization takes place and the variables associated with this period are 1, 2, 3, and 10. The second principal component captures most of the variability in the second stage of the process, where most of the polymerization takes place, and involves mainly variables 6, 8, and 9. When the process has two stages, as in this industrial example, it is common to see one principal component explain the first stage and another explain the second stage. MPCA does this because the correlation structure of the measurement variables is different in each stage, and thus a single principal component may not be able to explain both of them. None of the three principal components explains much of the variation during the transition period (50 through 65) since most of the variables during this time period have flat trajectories with few deviations from their mean values.

These plots of explained variation with respect to variables and time are very useful. They give insight into the process operation, and provide information for

understanding the MPCA model so that it can be used effectively to monitor new batches and detect faults.

3.4 Testing and Diagnosing a New Batch

The p-loading vectors from the MPCA analysis of the reference database contain most of the structural information about how the variable measurements deviate from their mean trajectories under normal operation. If a new batch is to be tested for any unusual process behavior, one can use these p-loading vectors to obtain the predicted t-scores ($\hat{\mathbf{t}}$) and residuals ($\hat{Q}$) for the new batch $\mathbf{X}_{NEW}$.

unfold and scale $\mathbf{X}_{NEW}(K \times J)$ to $\mathbf{x}_{NEW}$ $(JK \times 1)$, $\hat{t}_r = \mathbf{x}'_{NEW}\mathbf{p}_r$, $\mathbf{e} = \mathbf{x}_{NEW} - \sum_{r=1}^{R} \hat{t}_r \mathbf{p}_r$, $\hat{Q} = \mathbf{e}'\mathbf{e}$

The Hotelling statistic (D) to test if a new batch comes from the same population, is now given by (Tracy et al., 1992; Wierda, 1994):

$$D = \hat{\mathbf{t}}'\mathbf{S}^{-1}\hat{\mathbf{t}}\ I(I-R)/(R(I^2-1)) \sim F_{R,I-R}$$

where $\hat{\mathbf{t}}$ is the vector containing the predicted t-scores ($\hat{t}_r$) of all the retained principal components (R) in the MPCA model. This Hotelling statistic D for a new batch is similar to the Hotelling statistic D_S in the previous section for a batch in the reference database, and both are measuring the Mahalanobis distance in the reduced space between the position ($\hat{\mathbf{t}}$) of a batch and the origin (0). The D statistic has wider confidence limits than the D_S statistic since it refers to a new batch that was not used for building the MPCA model.

3.4.1 Contribution Plots

If the t-scores of the new batch are close to the origin and its residuals are small with respect to those in the reference database, then this indicates that its operation is similar to that in the reference database of normal batches. On the other hand, if the t-

scores or the residuals are large, one has to investigate what caused the abnormality in the batch behavior. By interrogating the underlying MPCA model, the contribution of each measurement variable to the deviations observed in the t-scores and the residuals can be plotted (Wise and Riker, 1991; Miller et al., 1993; MacGregor et al., 1994; Kourti and MacGregor, 1994). Although these plots may not provide an unequivocal diagnosis, they at least will clearly identify the time periods and the group of variables that are primarily responsible for the detected deviations.

The % contribution of all variables at time interval (k) to the sum of squares of the residuals of a batch ($\hat{Q} = \mathbf{e}'\mathbf{e}$) is given by:

$$\frac{100}{\hat{Q}} \sum_{j=1}^{J} e(jk)^2$$

the % contribution to $\hat{Q}$ of each measurement variable (j) over the entire batch history is:

$$\frac{100}{\hat{Q}} \sum_{k=1}^{K} e(jk)^2$$

and the % contribution to $\hat{Q}$ of variable (j) at time interval (k) is:

$$\frac{100}{\hat{Q}} e(jk)^2 \, \mathrm{sign}(x_{NEW}(jk))$$

The sign of the % contribution of variable (j) at time interval (k) in $\hat{Q}$ is determined by the sign that its deviation ($x_{NEW}(jk)$) had from its mean trajectory at that particular time (k). In this way one knows if this variable was above (positive) or below (negative) its mean trajectory at time interval (k), from its contribution sign.

Similarly, the % contribution of all variables at time interval (k) to the predicted t-scores ($\hat{t}_r = \mathbf{x}'_{NEW}\mathbf{p}_r$) is given by:

$$\frac{100}{\hat{t}_r} \sum_{j=1}^{J} x_{NEW}(jk)\, p(jk)$$

the % contribution to $\hat{t}_r$ of each measurement variable (j) is:

$$\frac{100}{\hat{t}_r}\sum_{k=1}^{K} x_{NEW}(jk)\,p(jk)$$

and the % contribution to $\hat{t}_r$ of variable (j) at time interval (k) is:

$$\frac{100}{\hat{t}_r}\,x_{NEW}(jk)\,p(jk)\,\mathrm{sign}(x_{NEW}(jk))$$

Since the t-scores have signs (in contrast with Q which is always positive) the above contributions to $\hat{t}_r$ can be positive or negative depending on whether or not the contribution of the particular variable or time interval had the same sign as $\hat{t}_r$. However, the contribution of variables or time intervals of interest (e.g. those with large deviations) almost always will have the same sign with $\hat{t}_r$, since these will contribute the most in $\hat{t}_r$ and therefore will determine its sign. In addition, the % contributions of variable (j) versus time will indicate by their signs if this variable was above or below its mean trajectory at particular time intervals of interest. In this manner, the t-score contribution plots of a variable versus time are easily interpreted, independently of the t-score ($\hat{t}_r$) sign.

As an illustration of the MPCA model and the contribution plots we shall test two batches. The first is the abnormal batch in the SBR example with the impurity problem right from the beginning of its operation. The second is batch 49 from the industrial example.

The predicted t-scores and residuals for the abnormal batch from the SBR example were found to be:

$\hat{t}_1 = -83.80$, $\hat{t}_2 = -9.16$, $\hat{t}_3 = -5.28$, $D=8.46$ ($F_{3,47,0.01}=4.25$), $\hat{Q} = 2181$ ($Q_{0.01}=2278$)

where $Q_{0.01}$ is the upper 99% confidence limit for Q in the reference database. From the MPCA plots of the reference database in Figure 3.1, it is clear that the abnormality of this batch is exhibited in $\hat{t}_1$. The contribution plots in the left hand side of Figure 3.3 show that the abnormality started early in the batch and continued throughout its whole operation. Variables 7 (density) and 8 (total conversion) were the major contributors to the deviation. Both were found to be always below their mean trajectories as it is shown in

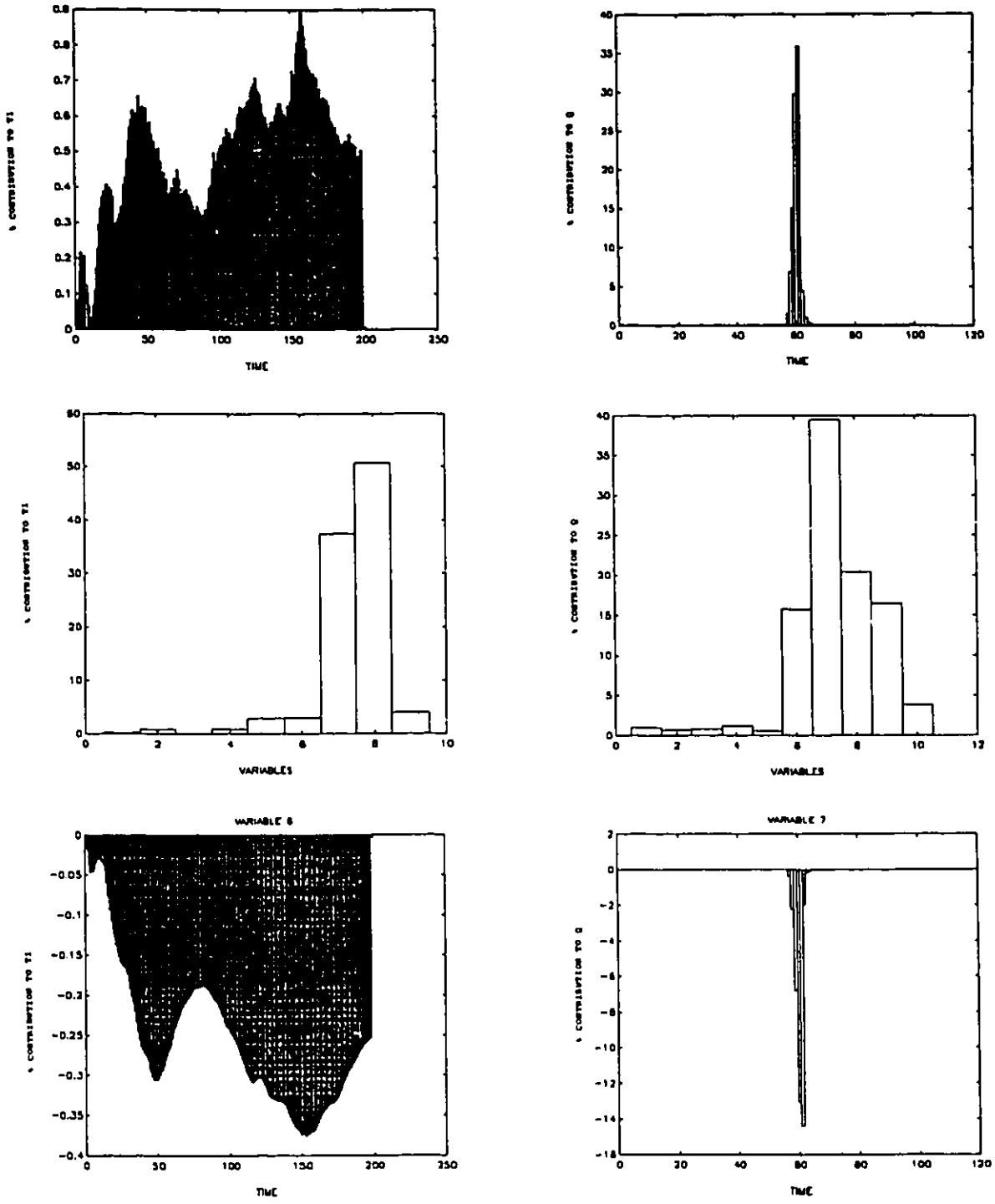

Figure 3.3 Contribution plots for the batch with the initial problem in the SBR example (left hand side plots), and batch 49 in the industrial example (right hand side plots). The signs of the % contributions in the bottom two plots, indicate if the measurement variables were above (positive) or below (negative) their mean trajectories.

the bottom plot of Figure 3.3 for variable 8. This suggest that something was constantly holding back the polymerization. Indeed this was the case. A high level of organic impurity was present in the butadiene feed during the entire batch run. The measurements for this batch are shown in Figure 2.3.

The results for batch 49 from the industrial example were:

$\hat{t}_1 = -5.93,\ \hat{t}_2 = 11.11,\ \hat{t}_3 = -5.66,\ D=0.53\ (F_{3,33,0.01}=4.45),\ \hat{Q} = 14424\ (Q_{0.01}=802)$

Clearly, the abnormality in this case is exhibited in the residuals. The contribution plots in the right hand side of Figure 3.3 reveal that the abnormality occurred between time intervals 57 and 65, and the major variables contributing in these deviations were mostly variable 7 along with variables 6, 8, and 9. Variables 6 and 7 are temperatures in the heating cooling medium, and variables 8 and 9 are pressure measurements. All these variables were found to be below their mean trajectories during the time period 57 through 65 as it is shown in the bottom plot in Figure 3.3 for variable 7. Figure 2.6 has the measurement trajectories for this batch, and reveals the power of MPCA to identify abnormal behavior. Almost everyone would have failed to identify what went wrong and this batch had borderline product quality. A visual inspection of Figure 2.6 will have probably suggested variable 5 as the variable with the most unusual behavior in time period 75 through 85. But, the MPCA model from the reference database had knowledge of such common cause variation in variable 5 at that particular time period, and integrated it as normal behavior of the batch operation. Upon closer examination of variables 6, 7, 8, and 9 in Figure 2.6, it can be seen that at time period 57 these four measurement variables did deviate in a systematic manner and returned at their mean trajectories around time period 65. The special event in this batch could be attributed to an operating problem with the cooling system, which was also the company's suggested explanation for this abnormal batch.

Here, we have to mention that the same contribution plots could be used when one is conducting MPCA analyses of both normal and abnormal batches together, as we did in

Chapter 2. Usually, the abnormalities from such analyses, as in the SBR example, will be detected in the t-scores since the MPCA model incorporates the abnormal behaviors in its p-loading vectors. In cases where there are few abnormal batches in the database under investigation, the contribution plots will be similar to those from the comparison of a new batch against the MPCA model of a reference database. In general, the contribution plots based on a reference MPCA model will give a clear picture of the abnormality since the MPCA model describes strictly the normal behavior of a batch process and an abnormality is revealed unconfounded from any other abnormal behavior that may be in a database of both normal and abnormal batches.

CHAPTER 4 On-Line Monitoring of Batch Processes

Chapter 3 discussed the selection of the reference database of normal batches. The MPCA analysis of such a database provides the statistical model which describes the normal operation of a batch. In this chapter we develop a method based on MPCA for monitoring the progress of a batch run in real-time. The development of the on-line control charts is illustrated through both the SBR and the industrial polymerization examples. Examples of on-line monitoring are presented along with their diagnostic plots.

4.1 On-Line Monitoring via MPCA

The p-loading vectors from the MPCA analysis of the reference database contain most of the structural information about how the variable measurements deviate from their mean trajectories under normal operation. The reduction in dimension is tremendous since most of the information in the reference database is captured in these few p-loading vectors which define the reduced space. As it is shown in Section 3.4, a new batch ($\mathbf{X}_{NEW}$) can be tested for any unusual process behavior by obtaining its predicted t-scores and residuals.

$$\text{unfold and scale } \mathbf{X}_{NEW}(K \times J) \text{ to } \mathbf{x}_{NEW}(JK \times 1) \ , \ \hat{t}_r = \mathbf{x}'_{NEW}\mathbf{p}_r \ , \ \mathbf{e} = \mathbf{x}_{NEW} - \sum_{r=1}^{R} \hat{t}_r \mathbf{p}_r$$

If the t-scores of a new batch are close to the origin and the residuals are small, then this indicates that the operation of this new batch is similar to these in the reference database of normal batches.

A problem arises when one wants to perform the test sequentially in time as the new batch evolves. In this situation the matrix $\mathbf{X}_{NEW}$ is not complete until the end of the batch operation. At each time interval during the batch operation, the matrix $\mathbf{X}_{NEW}$ has all the measurements only up to that time interval. The rest of the $\mathbf{X}_{NEW}$ matrix from the current time to the end of the batch is still undefined. The most valid way to overcome this

problem is to build K different MPCA models, one up to each time interval k using only the information available up to that time. This results in the need to store K loading vectors of dimension (Jk×1, k=1,2,..,K) for each principal component, and to apply the one appropriate for the current time k in order to calculate the scores and residual for that time. Although this is the most correct approach, the computational and storage requirements would be very large except for short duration (small K) batch processes having a relatively small number of on-line measurement variables (J). An alternative monitoring scheme, without using MPCA, is to make the assumption that the measured variables are Multinormally distributed around their mean trajectories, and at each time interval k to perform an F-test based on a Hotelling statistic. This will test if the Jk variables measured up to the current time (k) are too far away from the origin of the multivariate distribution estimated from the reference good database. This scheme is unattractive and for most practical purposes infeasible since one has to store and invert at each time interval k a large covariance matrix (Jk×Jk) which will usually be singular or at least ill-conditioned because of the highly correlated Jk variables.

Therefore, we propose several approximate methods for constructing sequential tests. All of these involve using the full loading vectors $\mathbf{p}_r(JK\times 1)$ obtained from the MPCA on the entire histories of the batches in the reference database, and then filling in the future observations in $\mathbf{X}_{NEW}$ in different ways. All of these approaches will give the same predicted t-scores and residuals at the end of the batch when the full $\mathbf{X}_{NEW}$ is known. To monitor the progress of a new batch, as new observations become available, one of the methods discussed in Section 4.2 is used to fill out the $\mathbf{X}_{NEW}$ matrix, and then the t-scores and residuals are calculated for each time interval. Thus, the following general procedure is used for monitoring a new batch:

ON-LINE MONITORING

i. take the new vector of measurements at time interval k: $\mathbf{X}_{NEW}(k,1:J)$

ii. subtract from $\mathbf{X}_{NEW}(k,1:J)$ the mean and divide by the standard deviation, which correspond at the kth time interval from the normal database, to get the vector with the current deviations from the mean trajectories: $\mathbf{x}_{NEW}((k-1)J+1:kJ)$

iii. fill in the rest of the data in $\mathbf{x}_{NEW}$

iv. $t_{r,k} = \mathbf{x}'_{NEW}\mathbf{p}_r \quad , \quad \mathbf{e} = \mathbf{x}_{NEW} - \sum_{r=1}^{R} t_{r,k}\mathbf{p}_r \quad , \quad SPE_k = \sum_{c=(k-1)J+1}^{kJ} \mathbf{e}(c)^2$

v. return to step i for the next time interval k+1

where the symbol "c:d" denotes the segment of elements between the c and d rows or columns of a matrix. The t-scores ($t_{r,k}$) represent the projection of $\mathbf{x}_{NEW}$ at time k onto the R-dimensional plane defined by the reference MPCA model. If a new batch is progressing in a manner that is consistent with the reference distribution of good batches, then it should stay close to the reduced space. Its perpendicular distance from the reduced space should be small (small residuals), and its t-scores values should continue to fall within the region of normal variation defined by the reference distribution.

There are two ways in which a new batch can exhibit deviations from the MPCA model. Its score values ($t_{r,k}$) can move outside the acceptable range of variation defined by a control region, or-and its residuals (e) could become large and the batch will be placed well outside, perpendicular to the reduced space. In the first case the model is still valid, and the new batch is still operating in the same way as the batches in the reference database, but it has a larger than normal variation in its measurements. This will show up clearly as large deviations of the t-scores from the origin (0) of the reduced space. In the second case the model is no longer valid, because a new event not in the reference set has occurred, and the new batch does not project onto the reduced space adequately. In this case, the residuals will become larger than a control limit defined by applying the model to the good batches in the historical database.

Thus, the residuals account for any variability which is not described sufficiently in the database of good batches. The best way to monitor the residuals is to use the Squared

Prediction Error ($SPE_k = \sum_{c=(k-1)J+1}^{kJ} e(c)^2$), which is the sum of squares of errors directly related with the latest on-line measurements at time interval k. The sum of the squared residuals over all time periods ($\hat{Q} = \mathbf{e}'\mathbf{e}$) 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 the errors associated with the filling in of future unknown observations in $\mathbf{x}_{NEW}$. We are interested to pinpoint the particular interval that something is going wrong and its duration, and the SPE is a very good indicator to accomplish this. A graphical representation of the meaning of each quantity that one has to track for monitoring on-line the operation of a new batch is given in Figure 4.1.

In Appendix B some other approaches can be found for on-line monitoring of a new batch. These approaches are based either in reformations of the batch data, or in reformations of MPCA. In general, none of them give better results than the approaches presented here, but they may be very effective in specific cases where one can take full advantage of them. Also in Appendix B, some suggestions are given of how to handle sudden changes in the measurement variables that one may have in a batch operation.

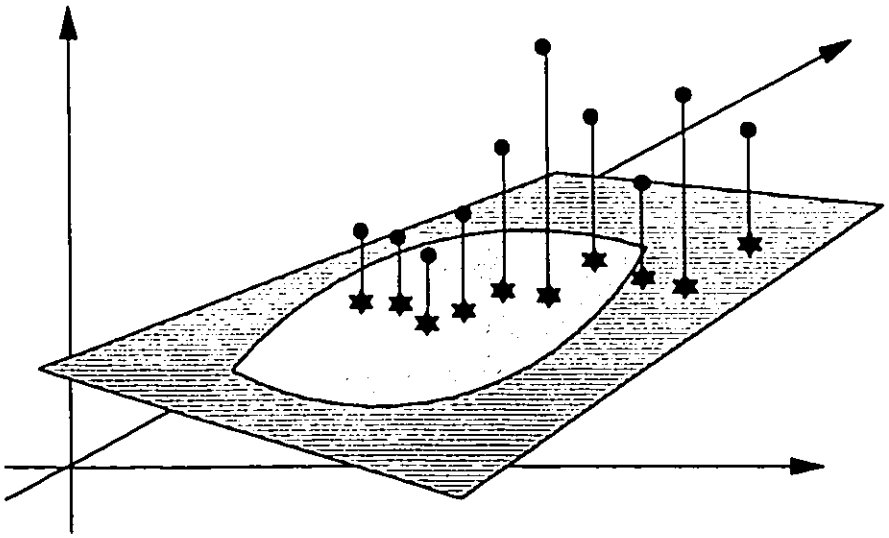

Figure 4.1 Graphical representation of a batch operation in the reduced space of MPCA. The stars on the plane are the t-scores which are the projections of the process measurements into the reduced space. The sum of squares of the residuals are the normal distances of the process measurements from the reduced space. The area inside the ellipse depicts the normal operation region.

4.2 Anticipating the Future Observations in X_{NEW}

Three methods are considered for filling in the unknown data in X_{NEW} between the current time interval k and the end of the batch. Recall that the x_{NEW} (unfolded and scaled X_{NEW}) contains the deviations of the measurements from their mean trajectories. The objective of all these approaches is to fill in the future values in the x_{NEW} in such a way, that the predicted t-scores at each time interval will be as close as possible to those that would be predicted, if one had the full x_{NEW}. Monitoring charts for the SPE and the first latent variable t_1 (similar charts are obtained for t_2 and t_3) for the SBR and the industrial example are shown in Figures 4.2 and 4.3 for each of these methods. Also in Figures 4.2 and 4.3 the approximate 95% and 99% control limits are shown with the outermost observations (five for the SPE and six for the t-scores) at each time interval from the reference normal database. The control limits for the t-scores and the SPE are developed in the next section.

APPROACHES FOR FILLING IN THE FUTURE DATA

i. The first approach to filling in the unknown observations in x_{NEW} is to assume that the future observations are in perfect accordance with their mean trajectories as calculated from the reference database. The assumption behind this approach is that the batch will operate normally for the rest of its duration with no deviations from its mean trajectories, and one has to fill the unknown part of x_{new} with zeros. This approach gives a nice graphical representation of the batch operation in the t-plots. In the upper part of Figures 4.2 and 4.3 one can see the cone shape of the control limits for the t-scores due to the assumption of future normal operation. A new batch always starts from the origin (**0**) of the t-scores in the reduced space and progressively moves out. The drawback of this approach is that the t-scores are less sensitive, especially at the start of the batch run, to detect an abnormal operation. The t-scores assert the overall performance of a batch since they take into consideration its whole operation: past, present, and future. Hence, they always have conservative (small) values because of the assumption that the batch for the

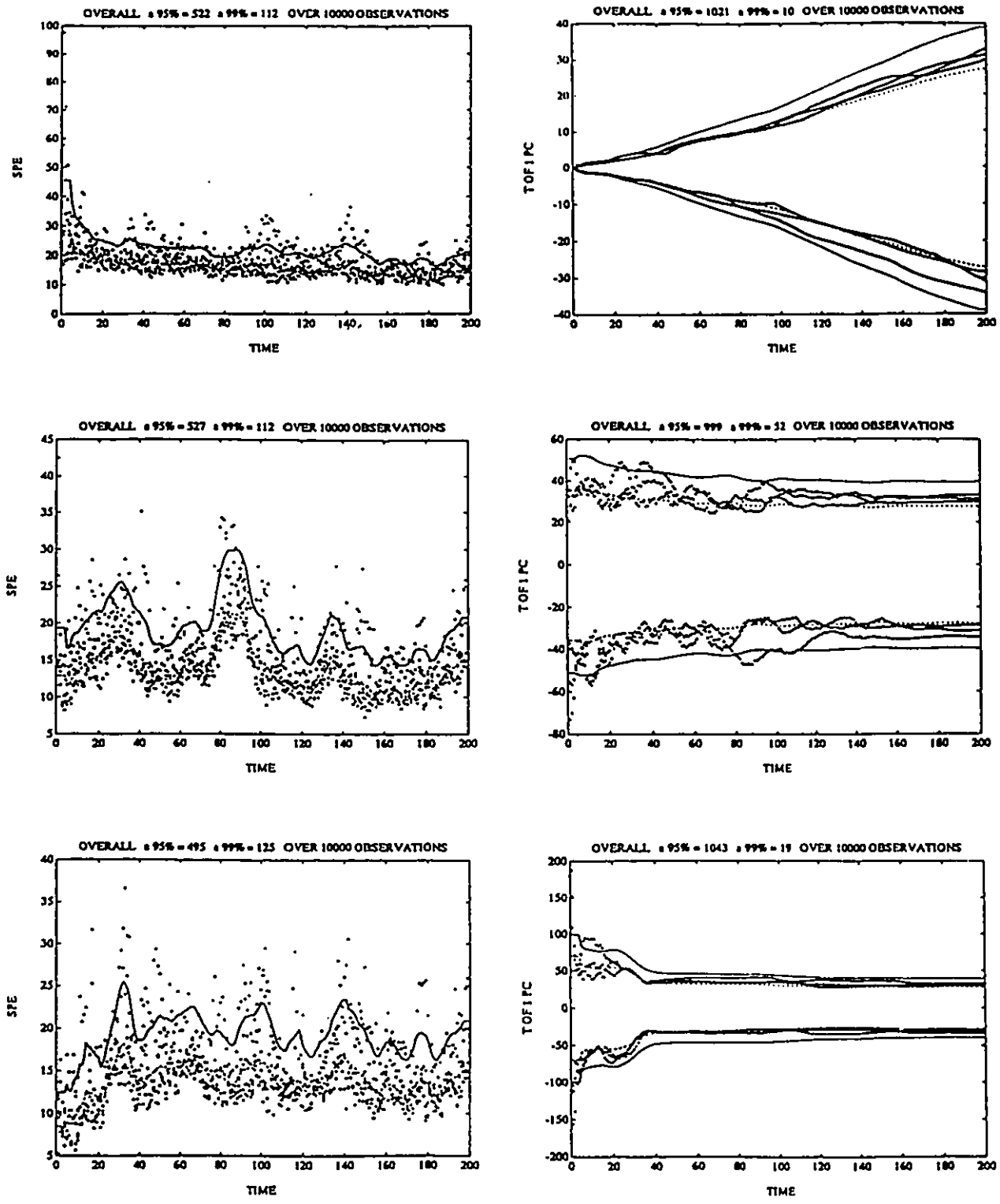

Figure 4.2 Control limits (95% and 99%) for the SPE and the t-scores of the SBR example, with the outermost values at each time interval for the three approaches of handling future deviations in x_{NEW}. The upper plots are for the approach with zero, the middle plots for the approach with current deviations, and the bottom plots for the approach by projection.

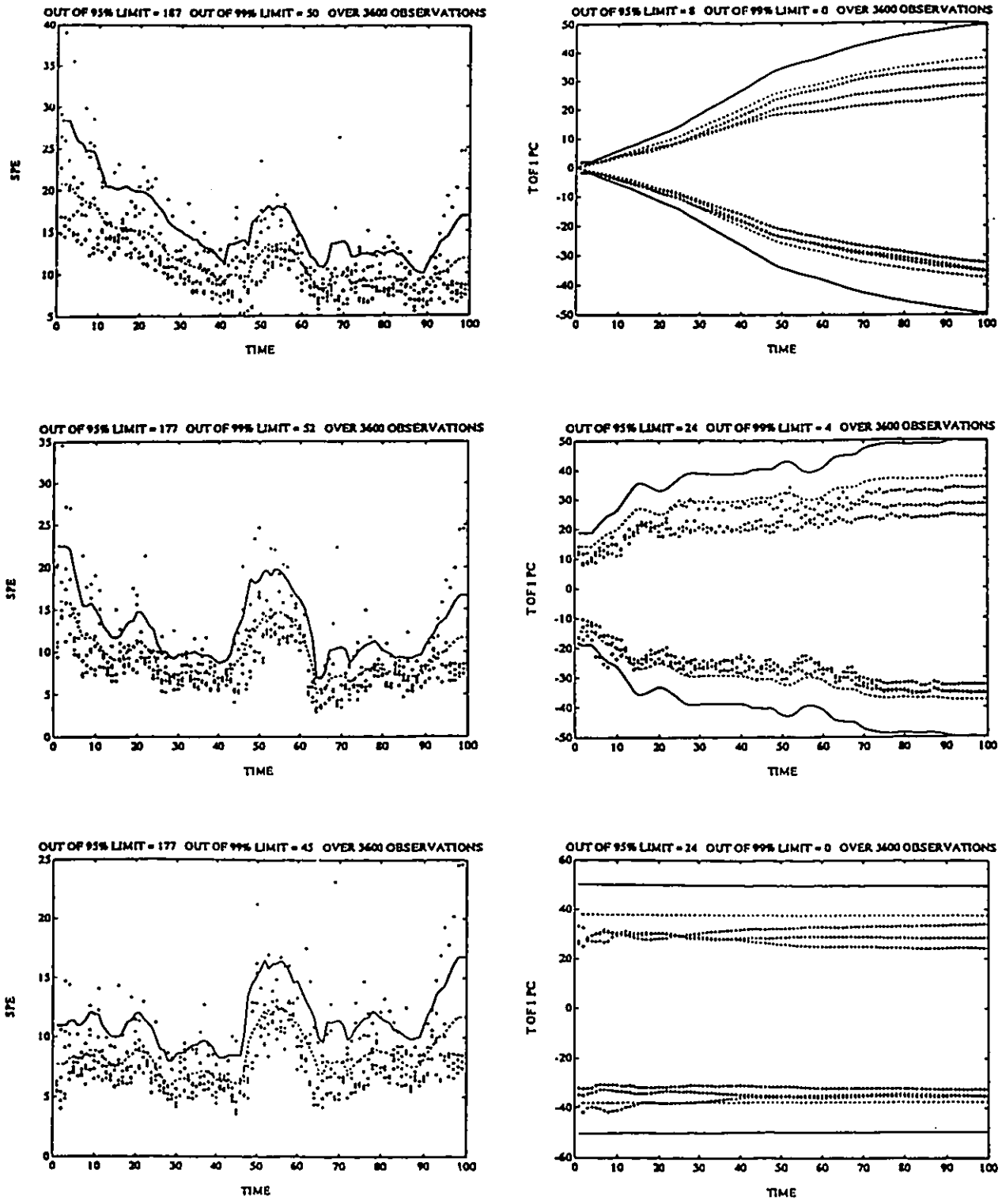

Figure 4.3 Control limits (95% and 99%) for the SPE and the t-scores of the industrial example, with the outermost values at each time interval for the three approaches of handling future deviations in x_{NEW}. The upper plots are for the approach with zero, the middle plots for the approach with current deviations, and the bottom plots for the approach by projection.

rest of its operation will not deviate at all from its mean trajectories. The advantage of this approach is the quick detection of an abnormality in the SPE control chart. Any deviation from the mean trajectories at any instant k will show up in the SPE since the unexplained part in $\mathbf{x}_{NEW}$ from the conservative t-scores will be large. Examples of on-line monitoring with this approach can be found in MacGregor and Nomikos (1992).

ii. A second approach is to assume that the future deviations from the mean trajectories of all measurement variables will remain for the rest of the batch duration at the current values observed at time interval k. In this case, the assumption is that the same errors will persist for the rest of the batch run. This is similar to the assumption made in model predictive control algorithms, such as the Dynamic Matrix Control (DMC) algorithms (Cutler and Ramaker, 1979), where the future values of the disturbances are assumed to remain constant at their current values over the prediction horizon considered. Under this assumption, the SPE chart is not as sensitive as in the first approach, but the t-scores pick up an abnormality more quickly. In general, this approach has shown good attributes, in several industrial examples that we have examined, for detecting a fault in a clear and quick manner. It will be used in Section 4.4 to monitor on-line the evolution of some batches from both the SBR and the industrial example. A compromise between the first two approaches, which shares their advantages and disadvantages, is to assume that the future deviations will decay linearly or exponentially from their current values to zero at the end of the batch run.

iii. The last approach uses the ability of PCA to handle missing data. The unknown future observations can be regarded as missing values from an object (batch) in MPCA. Hence, one can use the principal components of the reference database to predict these missing values by restricting them to be consistent with the already observed values up to time interval k and with the correlation structure of the measurement variables in the database as defined by the p-loading matrices of the MPCA model. MPCA can do this by projecting the already known observations ($\mathbf{x}_{NEW,k}(1:kJ)$) into the reduced space and calculating the t-scores at each time interval as:

$$\mathbf{t}_k = (\mathbf{P}_k'\mathbf{P}_k)^{-1}\,\mathbf{P}_k'\,\mathbf{x}_{NEW,k}$$

where $\mathbf{t}_k$(R×1) is the vector containing the predicted t-scores at time interval k from all R principal components, and $\mathbf{P}_k$(kJ×R) is a matrix having as columns all the elements of the p-loading vectors ($\mathbf{p}_r$) up to time interval k from all the principal components. This is the least squares solution to the problem where one searches for the t-scores ($\mathbf{t}_k$) which, along with the p-loadings from the MPCA model, will approximate the observations available up to time period k with the minimum error ($\mathbf{x}_{NEW,k} = \mathbf{P}_k\mathbf{t}_k + \mathbf{e}$). The matrix ($\mathbf{P}_k'\mathbf{P}_k$) is usually well conditioned even for the early time intervals because of the orthogonality property of the full (KJ×1) of the loading vectors ($\mathbf{p}_r$), and approaches the identity matrix as k approaches the final time interval K. This method appears to be superior to the other methods if at least 10% of the history of a new batch is known. It has the great advantage of giving t-scores very close to their actual final values, and thus their control limits have quite constant trajectories (Figure 4.2). Caution must be used at the beginning of a new batch where this method may give quite large and unexplainable t-scores since there is so little information to work with. This is clear in the control limits of the t-scores in the SBR example (Figure 4.2), where they are quite wide at the beginning of the batch. The control limits for the t-scores in the industrial example are almost constant throughout the batch operation because of the strong correlation among all the measurement variables in this example. Examples of on-line monitoring of this approach can be found in Nomikos and MacGregor (1995).

Which approach to use depends on the specific characteristics of the process under consideration. All approaches give similar results as the batch proceeds towards its end, and their main differences are exhibited during the first half of the batch operation. If the trajectories of the process measurements do not exhibit frequent discontinuities or early deviations, one may use the third approach since in this case the correlation among the measurement variables will be fairly constant. If there is knowledge that the disturbances in a given process are quite persistent, then it is better to use the second

approach which generally has worked well in most cases we have investigated. If a batch process does not exhibit persistent disturbances, or has variables with discontinuities in their trajectories, then it may be better to use the first approach. In general, one can use a combination of the above approaches, like starting with one approach and switching after some time to another one, and to build in this way some engineering knowledge in his monitoring scheme. Another way which is more time consuming, is to use time series models (Hamilton, 1994) to predict the future deviations from the mean trajectories.

4.3 Control Limits for the SPC Charts

Independently of which way one may choose to handle the future observations in a new batch run, the on-line monitoring will be based on control charts in which the control limits will be established that are appropriate for that approach. These charts will monitor the t-scores and the SPE of a new batch as it progresses. The control limits for these charts are calculated by passing each of the batches in the reference database through the monitoring procedure, as if they were new batches, and collecting their t-scores and SPE at each time interval k. These observations (50 for the SBR and 36 for the industrial example) for the t-scores and SPE at each time interval provide the external reference distribution (Box et al., 1978) upon which the control limits can be directly calculated. The assumption is that this external reference distribution is sufficient to capture the common cause variation in normal batch operations and that whatever mechanism gave rise to the observations in our reference database, is still operating in the same manner for the future batches.

The best way to obtain the control limits would be to use a second database of normal batches, not used in the model building step, and pass them through the monitoring procedure using the p-loadings obtained from the reference database. But usually, one has only 50 to 60 normal batches available, and thus has to use them all to build the MPCA model and capture most of the variation in the measurements.

4.3.1 Control Limits on T-Scores

The t-scores at each time interval k are linear combinations of the measurement variables and by the Central Limit Theorem should be approximately Normally distributed. Similar analysis of the t-scores, as in Section 3.3, revealed that they were again well approximated by a Multinormal distribution, except in the SBR example for the first few time intervals early in the beginning of a batch. Their distribution resembles more a Rectangular distribution than a Normal one. These early deviations from Normality result from the non-normal distribution of the initial conditions, which are distributed more like as in a factorial design, used in the SBR simulation (Appendix A).

Under the assumption of Normality the control limits at significance level α, for a new independent t-score, at any given time interval are given by (Chew, 1968; Hahn and Meeker, 1991):

$$\pm t_{n-1,\alpha/2}\, s_{ref}\, (1+1/n)^{1/2}$$

where n and s_{ref} are the number of observations and the estimated standard deviation of the t-score sample at a given time interval k (the sample mean is always zero since we subtract the mean trajectories from the raw data $\underline{X}$), and $t_{n-1,\alpha/2}$ is the critical value of the Studentized variable with n-1 degrees of freedom at significance level $\alpha/2$. As in Section 3.4, the Hotelling statistic for the t-vector of a new batch is:

$$D = t_k' S^{-1} t_k \; I(I-R)/R(I^2-1) \quad \sim \quad F_{R,I-R}$$

where $t_k(R\times 1)$ is the vector containing the predicted t-scores from all (R) the principal components at time interval k. The D statistic provides a measure at each time interval of the directed distance of the position of a new batch in the reduced space from the origin of normal operation. The axis lengths in the t-space of the confidence ellipsoids with significance level α in the direction of the rth principal component are given by (Johnson and Wichern, 1988)

$$\pm \; (S(r,r)F_{2,I-2,\alpha}\, 2(I^2-1)/I(I-2))^{1/2}$$

The area inside these ellipsoids define the normal operating region for a new batch in the reduced space.

As can be seen from the individual t-plots (Figures 4.2 and 4.3), the control limits change throughout the duration of the batch reflecting a greater degree of variation at certain times. Rather than trying to show how the joint elliptical contours change with time on the joint latent variable space plots (eg. t_1-t_2), we constantly display the control contours appropriate for the end of the batch (Figures 4.6 through 4.9). That is why we use, in defining the control ellipsoids, the estimated covariance matrix S of the t-scores from the analysis of the reference database in Chapter 3. The hypothesis tested by these control ellipsoids at each time interval is the classification of a batch as normal or abnormal, based on the measurements available until that time interval and the assumed future behavior of the batch. Similarly, the control limits for the D statistic are also based on that appropriate for the final time and have the same interpretation.

4.3.2 Control Limits on SPE

The SPE is a quadratic form of the errors associated with the latest observations at time interval k ($SPE_k = \sum_{c=(k-1)J+1}^{kJ} e(c)^2$). These errors (e(c)) were found to be well approximated by a Multinormal distribution $N(0,\Sigma)$ based on Normality tests as in Section 3.3. Box (1954), Jensen and Solomon (1972), and Jackson and Mudholkar (1979) have derived approximate distributions for such quadratic forms. Box showed that it is well approximated by a weighted Chi-squared distribution ($g\chi^2_h$) where the weight (g) and the degrees of freedom (h) are both functions of the eigenvalues of Σ (Appendix C). Jensen and Solomon, and Jackson and Mudholkar's approximate distribution (see Section 3.3) is very close to that given by Box (Appendix C) in cases where one has extracted the dominant principal components, as it is in our case.

We use the $g\chi^2_h$ approximation of Box for the distribution of the SPE to estimate the control limits at any point in time. Although the g and h can be estimated from the

eigenvalues of the estimated Σ, a simpler approach is used here based on matching moments between a $g\chi^2_h$ distribution and the reference distribution of SPE at any time interval k. The mean and variance of the $g\chi^2_h$ distribution (μ=gh, σ^2=2g^{2}h) are equated to the sample mean (b) and variance (v) of the SPE sample at each time k. Thus, g and h are estimated by:

$$\hat{g} = v / 2b \quad , \quad \hat{h} = 2b^2 / v$$

We chose to do this because we have to estimate a control limit at each time interval, and this way is faster than using the traces of powers of the residual covariance matrix (J×J) at each time interval. It is a quick way to estimate g and h reasonably well, provided that the number of SPE observations is sufficiently large and there are no outliers in the sample. Thus the upper control limit on the SPE at significance level α for time interval k are given by:

$$(v / 2b)\, \chi^2_{2b^2/v,\alpha}$$

where $\chi^2_{2b^2/v,\alpha}$ is the critical value of the Chi-squared variable with $2b^2/v$ degrees of freedom at significance level α.

Estimates of g and h are shown in Figure 4.4 for the monitoring scheme used in the middle plots of Figures 4.2 and 4.3 (i.e. filling in with the current deviations). These plots are similar to those obtained using the other two approaches to handling the future observations in $\mathbf{X}_{NEW}$, with differences occurring mainly in the first ten to fifteen time intervals where each approach has its own distinct character. These plots of estimated g and h provide information about the changing nature of the distribution of the residuals throughout the duration of the batch. Low values of the degrees of freedom (h) indicate that the distribution is dominated by large variability of only a few of the measurement variables about their mean trajectories. High values of h occur during more stable periods where deviations from most of the variables are contributing evenly to the SPE. The g is simply a scaling factor to enable one to match the moments.

In the SBR example (left hand side of Figure 4.4), one can see clearly in the h-plot

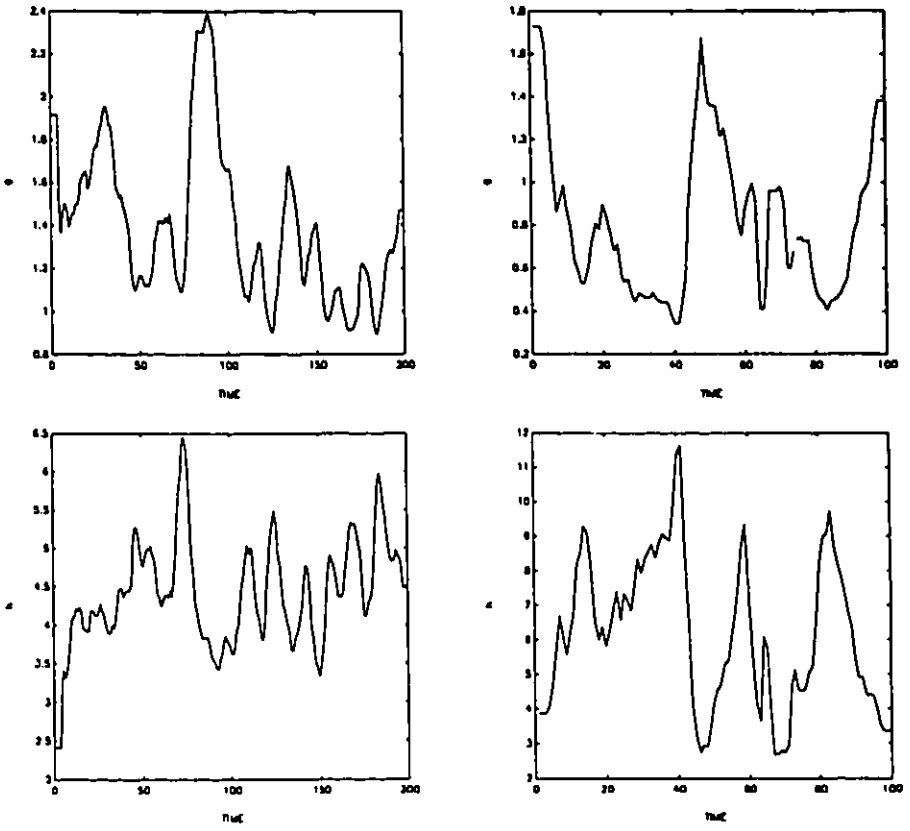

Figure 4.4 Plots of estimated g and h values from the $g\chi^2_h$ distribution of the SPE. The left hand side plots are for the SBR example, and the right hand side plots are for the industrial example.

that the equivalent of 4 to 5 measurement variables contribute constantly in the residuals throughout the batch operation. This indicates that there are no major changes in the operation of the process. Since h is fairly constant throughout the batch operation, the g-plot indicates that in the last part of the SBR operation (100-200 time intervals) there is little variation in the measurements variables which is not explained by the MPCA model. In the industrial example (right hand side of Figure 4.4), the batch period (10-40) represents a fairly smooth behavior during the well controlled vaporization stage where there are 7 to 10 degrees of freedom in the SPE. Period (45-50) represents the transition from the vaporization to polymerization stage where a few variables dominate the SPE, and the batch period (65-70) is the operation during polymerization where the degrees of freedom change constantly. In both these periods a few variables are changed rapidly.

Sudden changes made to the process at the start and at the end of the batch run, contribute to low degrees of freedom there.

This is a general methodology to get approximate control limits for the t-scores and the SPE. The assumptions behind these control limits are reasonable and have shown to hold well in several industrial examples that we have investigated. If one finds that the proposed distributions do not represent accurately the sample observations, then one has to use some other approximation. One way is to choose a general distribution like the Gamma distribution and estimate its parameters at each time interval. This is time consuming and has the extra work of finding control limits with the same significant level for all the time intervals by integrating the Gamma density function (Johnson and Kotz, 1970). Another way to estimate the control limits, especially when you have a large number (I) of batches in your reference database, is by using again the idea of external reference distribution. The sample observations at each time interval have to be arranged in ascending order, and the 0.8*I and 0.9*I observations to be taken as the 80% and 90% limits of the population. Then based on these two limits to estimate the 95% and 99% control limits by extrapolation, making the additional assumption that the upper tail of the distribution has the same shape of a Normal for the t-scores, or a Chi-squared distribution for the SPE. By using the ordered 0.8*I and 0.9*I observations as the 80% and 90% limits, one avoids being affected by spurious observations that may exist in the sample's upper tail.

4.3.3 Smoothing Window

Because the number of observations in the reference distribution at each time interval may not be very large (I=50 in the SBR, and I=36 in the industrial example), the control limits on the t-scores and SPE can be quite variable. Since most batch processes progress in a reasonably smooth manner and each time interval is closely related to its neighbours on either side, one might expect the control limits to change smoothly.

Therefore, we have used the idea of windowing taken from spectral analysis (Jenkins and Watts, 1968) for the results in all the figures in this chapter. The reference distribution at time interval k was composed of the observations at time intervals k-w through k+w. In the SBR example w was set to 3 and in the industrial example w was equal to 2. In effect, we have used a moving rectangular window that is 2w+1 time intervals wide to combine data for the estimation of the control limits on the t-scores and SPE at the center of the window. In our examples this provides n=(2w+1)I observations for the calculation of the limits at each time interval. In general, the width of the smoothing window will depend upon the number of batches in the reference distribution, and on the nature of the process itself. In the case where there are many batches in the reference dataset, very little smoothing will be necessary. If the sampling rate of data acquisition is fast with respect to the process dynamics, then a wider smoothing window can be possibly used. The use of a window helps small sample sizes, but relies upon the additional assumption that the variance of the statistics vary in a smooth manner with respect to time. In practice (as in spectral estimation), several window widths can be tried and one chosen which gives reasonably smooth control limits (low variance) but does not affect their main shapes (low bias). Figure 4.5 shows how the 99% SPE control limits for both examples change by using different window widths.

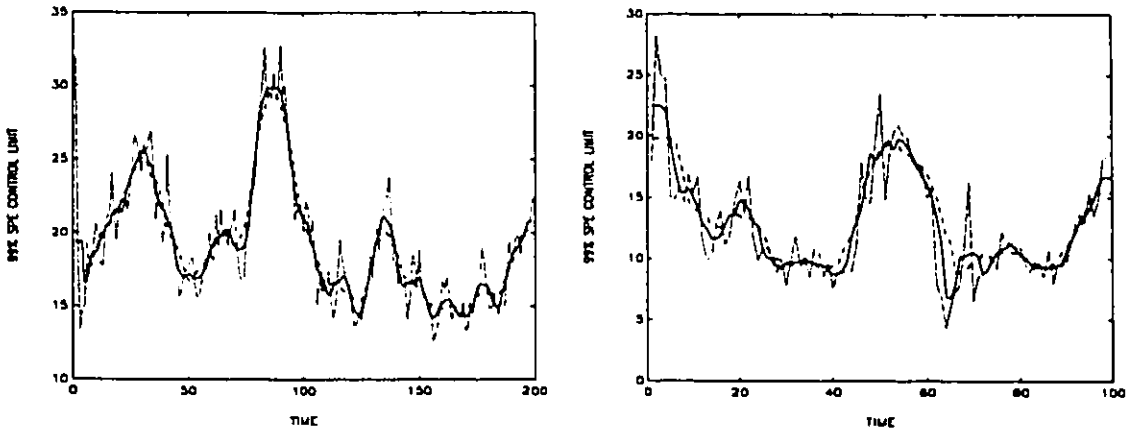

Figure 4.5 99% SPE control limits for the SBR (left hand side plot) and the industrial (right hand side plot) example. The dotted lines show the limits estimated without windowing (w=0). The solid lines show the limits estimated with a window width of 7 (w=3) and 5 (w=2) for the SBR and the industrial example respectively. The dashed lines show the limits estimated with a window width of 11 (w=5) and 9 (w=4) for the SBR and the industrial example respectively.

4.3.4 Overall Type I Error

The proceeding control limits for the t-scores and SPE were based on the approximate distributions of these statistics at any one point in time. Considering only that period of time, the α value in these tests would be the Type I error. Although this is a common procedure for setting control limits (Bauer and Hackl, 1980; John, 1990; Montgomery, 1991; MacGregor and Harris, 1993), it is not correct when one considers the sequential application of the procedure over the entire batch run. In general, the Type I error associated with monitoring the entire K time intervals will be different from the α value for the instantaneous test. If the statistics were independent over time then the overall Type I error would be given by $1-(1-\alpha)^K$ which is 0.86 for the SBR and 0.63 for the industrial example, for instantaneous α=0.01. If this was the case, it would create many false alarms. However, the t-scores and SPE values at successive times are not independent, and the overall Type I error could only be determined knowing the joint distribution of these statistics over all periods. To establish the approximate Type I error for the control limits each set of batch data in the reference database is passed through the monitoring procedure, and the number of values of each test statistic (t-scores and SPE) falling outside the control limits can be enumerated. Thus, an overall Type I error can be estimated for the control chart as the number of values of the test statistic outside the control limits in the reference database divided by the total number of observations (IK). These overall Type I errors for both examples are presented in Table 4.1 for the two values of instantaneous Type I errors (α=0.05 and α=0.01). The estimated overall Type I errors for the SPE test are quite close to the instantaneous α values. We have found this to be generally true in every dataset we have so far investigated. The overall Type I errors for the t-scores are generally close to the nominal α value for instantaneous α=0.05, but the approximation is poorer for instantaneous α=0.01.

Approach	Zeros		Current deviations		Projection	
SBR example						
t1	0.102	0.001	0.099	0.005	0.104	0.002
t2	0.079	0.004	0.090	0.016	0.070	0.012
t3	0.102	0.001	0.095	0.009	0.091	0.013
SPE	0.052	0.011	0.053	0.011	0.049	0.012
Industrial example						
t1	0.002	0.000	0.006	0.001	0.006	0.000
t2	0.052	0.011	0.041	0.003	0.086	0.009
t3	0.030	0.000	0.051	0.002	0.056	0.000
SPE	0.052	0.014	0.049	0.015	0.049	0.012
Instantaneous α	0.050	0.010	0.050	0.010	0.050	0.010

Table 4.1 Overall Type I error for the control limits of the t-scores and SPE for the three approaches to handling the future observations in X_{NEW}.

4.4 Examples of On-Line Monitoring

Four examples are given in Figures 4.6 through 4.9 for on-line monitoring: three from the SBR simulation and one from the industrial polymerization process. None of these test batches was included in the reference database of normal batches used to develop either of the MPCA models. The batch in Figure 4.6 is a new SBR batch which gave acceptable final polymer quality. Figures 4.7 and 4.8 show the monitoring charts for the two abnormal batches in the SBR example. The batch in Figure 4.9 is batch 49 from the industrial example. This later batch yielded a product of marginal quality, in that the quality measurement was right at the acceptable limit. The measurements of these batches can be seen in Figures 2.3 and 2.6.

Figures 4.6 through 4.9 show plots for the SPE and one of the t-scores along with their 95% and 99% control limits. Also shown are plots of the reduced space (eg. t_1-t_2) and the D statistic. As discussed in Section 4.3.1, the 95% and 99% control limits on these latter two charts are only approximate since they are based on the covariance matrix S of the t-scores from the post analysis of the reference database which has available the measurement variables for the whole batch duration. These two charts evaluate at each time the expected performance of the whole batch duration assuming that the future

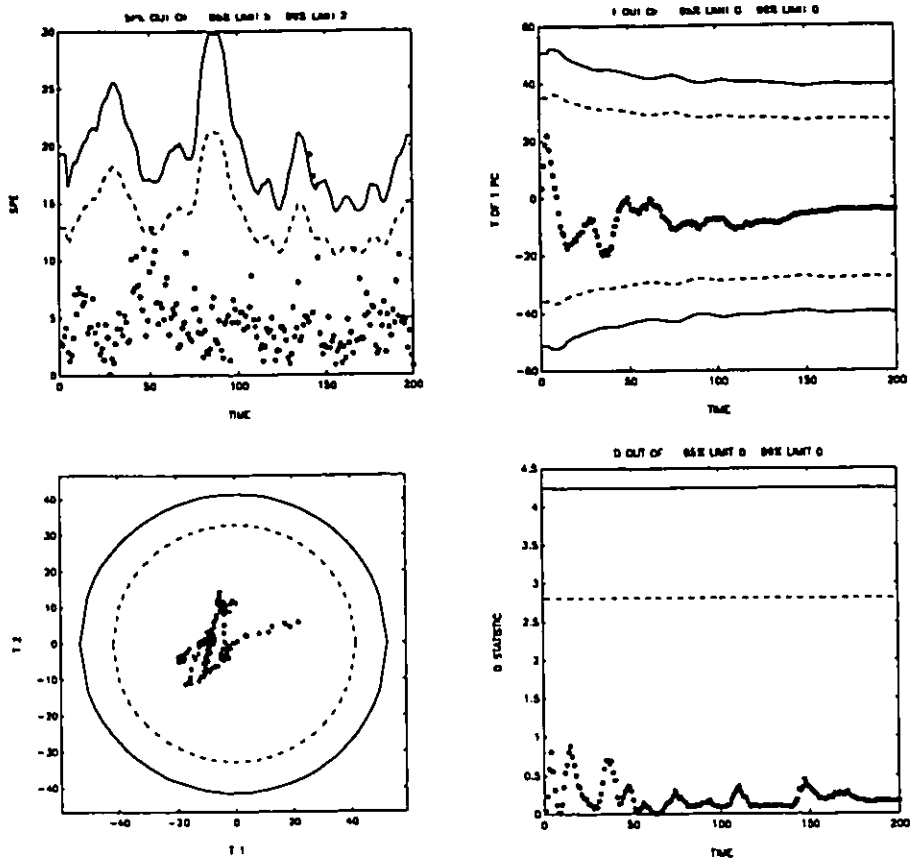

Figure 4.6 Monitoring charts with their 95% and 99% control limits for a new normal SBR batch.

behavior of the batch is well described by the approach which is used to fill in the unknown observations in X_{NEW}. To provide more precise control limits for the joint t-space and the D-statistic charts would require evaluating and storing the estimated joint covariance matrix (R×R) of the t-scores for each point in time. Note that the t-scores are truly orthogonal only at the final time corresponding at the end of the batch.

The interpretation of the monitoring charts is straightforward. The new normal batch in the SBR example (Figure 4.6) has all its latent vector plots and its SPE well within the control limits, implying that at no point during the batch operation there is any evidence that everything is not proceeding well. The batch with the impurity contamination in the butadiene feed starting right from time zero (Figure 4.7), is clearly flagged as a batch with problems within the first 15 time intervals. Several of its SPE

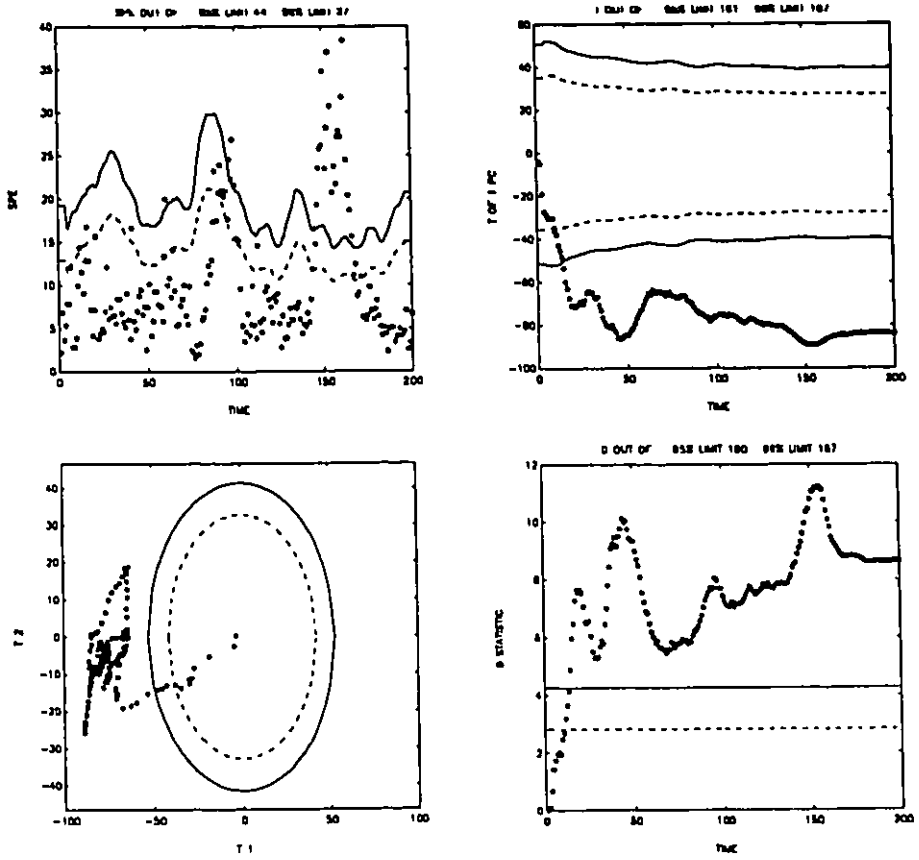

Figure 4.7 Monitoring charts with their 95% and 99% control limits for the SBR batch with an initial problem.

values exceed the 99% limit, but the clearest detection is provided by the charts associated with its t-scores. The reason that the problem shows up most clearly in the individual t_1-chart comes from the fact, that the major variables contributing to the first principal component (Figure 3.2) are the total conversion and the measured latex density, both of which show abnormally low values at the start of a batch when impurities are present. In the other abnormal batch with the impurity contamination in the butadiene feed at the mid-point of the batch operation (Figure 4.8), the problem is most clearly alarmed in the SPE plot. This indicates that a type of variation or fault has been encountered that was not present in the reference database, which is indeed the case here. None of the batches in the normal database had a sudden change in the level of the organic impurities in the feeds half-way through its operation. Most of them had small perturbations in the level of the

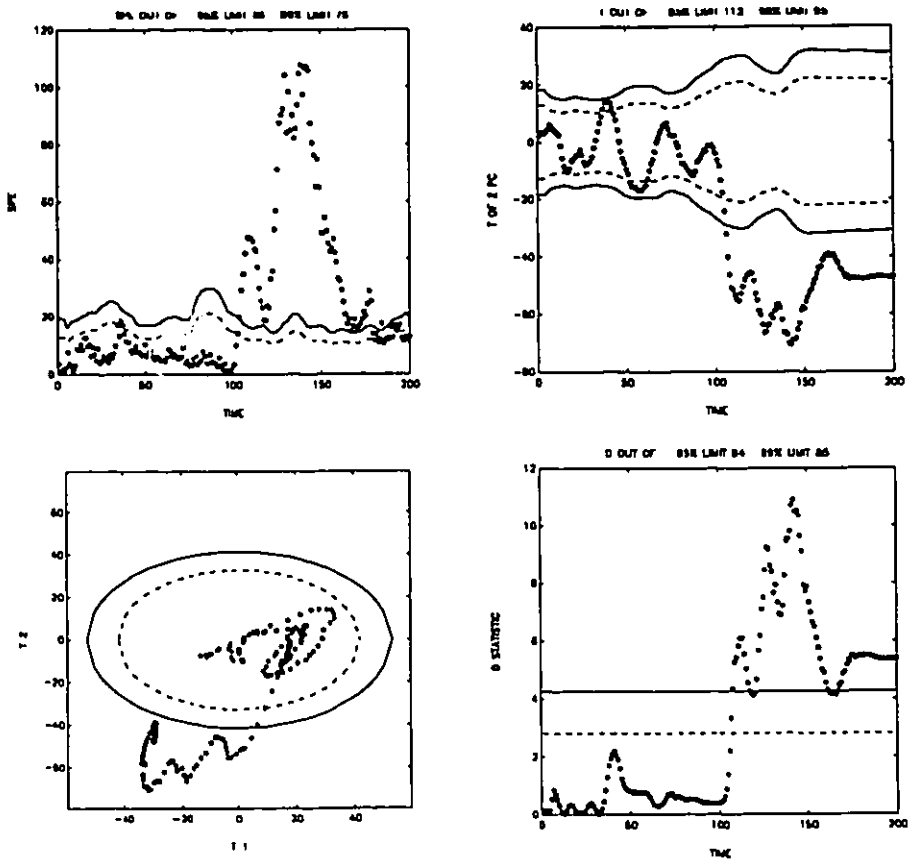

Figure 4.8 Monitoring charts with their 95% and 99% control limits for the SBR batch with an impurity problem halfway through its operation.

organic impurities right from the beginning of the batch. It is also interesting to see that in addition to the SPE, it is now the second principal component (t_2) that also clearly detects the problem. This is reasonable because as one can see from Figure 3.2, the second principal component is dominated by information over the last part of the batch operation, where the fault occurred. The SPE chart for batch 49 from the industrial example clearly signals that something is unusual between time intervals 57 and 65 (Figure 4.9). During this period the p-loadings are small which makes the t-scores slow to respond to the change. After the 65th time interval the measurements return to their normal trajectories, as do the t-scores because now the unusual previous behavior plays a less significant role on them since we are filling in the unknown part of x_{NEW} with the current deviations we have at the last time interval. Unlike the SPC charts for the abnormal batches investigated

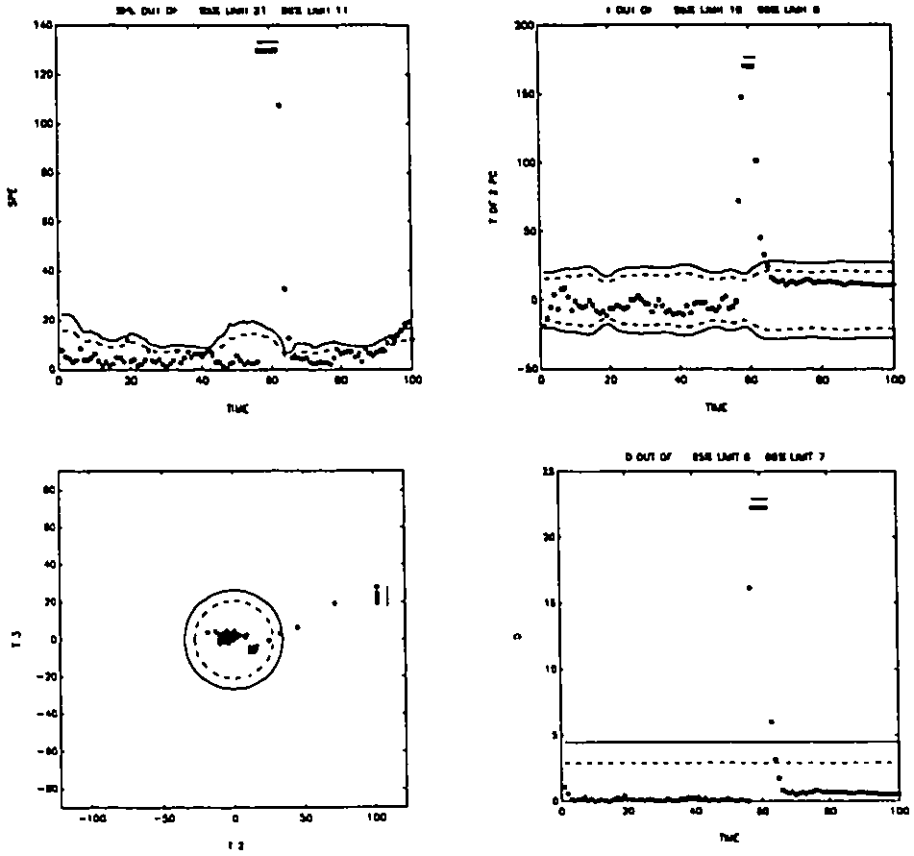

Figure 4.9 Monitoring charts with their 95% and 99% control limits for the industrial abnormal batch. Bars above observations denote truncated values for better graphical representation. The SPE values at time intervals 57 through 62 are ten times greater than what the graph shows.

in the SBR example, where once a problem was detected, the t and SPE values remained outside of the control limits for the remainder of the batch, in this industrial batch (49) the problem disappears shortly after time 65. In spite of its return inside the acceptable control region, this batch is characterized as abnormal because of the violation of the control charts during the period 57-65. Indeed, this deviation in the batch did result in a product with a borderline quality. Although its SPE returned below the control limits after time interval 65, its overall sum of squared residuals (Q) at the end of the batch is very large as already shown in Section 3.4. An example of on-line monitoring for a normal industrial batch can be found in Nomikos and MacGregor (1995).

4.5 On-Line Contribution Plots

Once a fault or special event has been detected, it is important to diagnose the event to find an assignable cause. For this aspect of SPC, the multivariate methods are much more useful than univariate methods. By interrogating the underlying MPCA model, the contribution of each measurement variable to the deviations observed in the SPE and in the t-scores can be displayed as discussed in Section 3.4.1. These contribution plots can be immediately displayed on-line by the operator as soon as the special event is detected. They will clearly indicate how much each of the measurement variables contributed to the t-score or SPE under investigation, and help one to hypothesize for an assignable cause for the fault detected.

The % contribution at time interval k of measurement variable j to $SPE_k = \sum_{c=(k-1)J+1}^{kJ} e(c)^2$ is given by:

$$\frac{100}{SPE_k} e(jk)^2 \, \text{sign}(x_{NEW}(jk))$$

The t-scores ($t_{r,k} = x'_{NEW}p_r$) at each time interval take into account all the measurements, both those that we already know and those that we fill in for the future observations. Therefore, the % contribution of variable j is evaluated throughout its whole history (past, present, and future) and its sign is solely determined by the sign of its present deviation at time interval k.

$$\frac{100}{t_{r,k}} \text{sign}(x_{NEW}(jk)) \sum_{k=1}^{K} x_{NEW}(jk) p_r(jk)$$

In this way, as discussed in Section 3.4.1, one can easily check by the sign of the % contribution to either the SPE or the t-score, if the variable was above (positive) or below (negative) its mean trajectory at the time interval under consideration.

The SBR batch with the initial impurity contamination, is clearly detected as abnormal in the t_1-chart at time interval 14 (Figure 4.7). The contribution plot for this t-score in Figure 4.10, reveals that variables 7 (density) and 8 (total conversion) were

primarily responsible. Both of them are below their mean trajectories which suggests that something is holding back the reaction. Since there is no indication that something is wrong with the cooling system, a possible explanation will be an impurity contamination which was in fact the case in this example. If this monitoring scheme had been in effect during the batch, then one might have stopped the batch to investigate the origin of the impurities. The contribution plot in Figure 4.10 is very similar to the contribution plot we got when we compared its whole operation against that of the reference database (Figure 3.3).The similarity in these plots is because we used the current deviations to fill in the future observations in $\mathbf{x}_{NEW}$, and in this case the impurity contamination continues to affect the batch in the same manner throughout its whole operation.

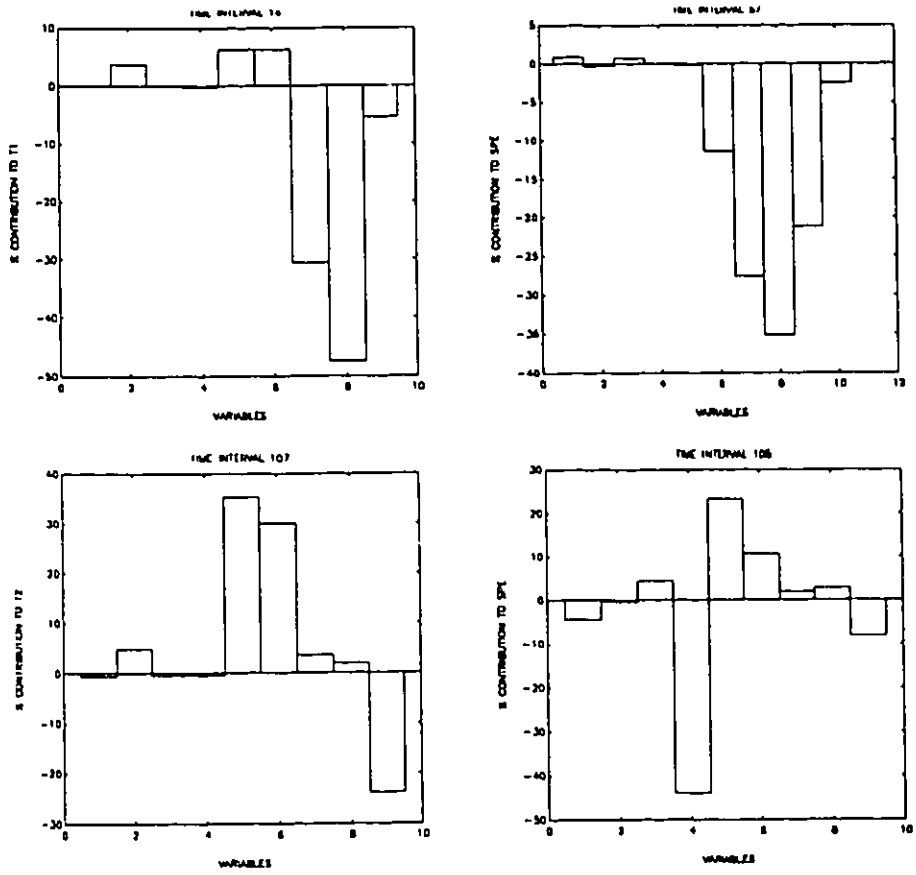

Figure 4.10 Contribution plots for the abnormal batches. The upper left plot is for the SBR batch with initial problem. The upper right plot is for the industrial abnormal batch. The bottom plots are for the SBR batch with problem halfway through its operation. The signs of the % contributions indicate if the variables were above (positive) or below (negative) their mean trajectories.

In the other SBR batch, with the impurity contamination halfway through its operation, the SPE detect the problem at time interval 105 and the t_2-score at time interval 107 (Figure 4.8). The contribution plots for both the SPE and t_2 are given in Figure 4.10. Both show that the reactor temperature (variable 4) is much lower than usual, and the cooling water temperature (variable 5) along with the cooling jacket temperature (variable 6) are higher than normal. This implies that the cooling system is working satisfactorily and tries to compensate a temperature drop in the reactor. Since the temperature of the monomer feeds (variable 3) is at normal level, and the t_2 contribution plot suggest that the instantaneous rate of energy release (variable 9) is below its mean trajectory, a possible explanation for this behavior can be attributed to something like impurities which are slowing down the rate of polymerization in the reactor. With these indications in hand, the batch might have been stopped to investigate the source of the problem, or a new monomer feed tank used, and-or the final batch product segregated until its final quality was determined. Note here that variable 4 (reactor temperature) appears solely in the SPE contribution plot since the MPCA model does not account much for this variable (Figure 3.2). Variables 7 (density) and 8 (total conversion) do not appear to contribute in this batch as in the previous batch with the problem right from its beginning. In this case, we are in the middle of the batch operation and these two variables do not change very fast. Eventually as the batch progresses and the impurities continue to enter with the butadiene feed, both will drop below their mean trajectories.

The abnormality in the industrial batch 49 is clearly detected in the SPE chart at time interval 57 (Figure 4.9). The contribution plot for the SPE is given in Figure 4.10. The major variables contributing are 6, 7 (temperatures in the heating cooling medium), 8, and 9 (pressures). All of them simultaneously deviate below their mean trajectories. As in Section 3.4.1 the cause of the abnormality in this batch could be attributed to a failure in the cooling system of the process. The SPE contribution plot is similar to what we had in Section 3.4.1 (Figure 3.3) since the abnormality continuously affects the process during the time period 57 through 65.

4.6 Discussion

The on-line monitoring examples demonstrate that the proposed charts preserve the SPC ideas of using easily displayed and interpreted charts that can detect quickly an abnormality. One has to monitor closely the SPE and D control charts, which provide complementary information about the process operation, and if something is going wrong, to use the t-score charts and the contribution plots to get a better understanding of the fault. In some cases, as discussed in Section 2.4, it may be possible to identify areas in the reduced space corresponding to a particular fault, and thus to construct an expert system for diagnosis. It should also be noted that a violation of the control charts does not mean that the product will be unacceptable. It only means that the operational behavior of the batch is unusual and this unusual behavior may lead to low quality product. MPCA has only information about the on-line measurements of a process. The objective of the monitoring procedure is to detect and eliminate faults from future appearance, and thereby shrink the control limits and work towards a more consistent production of quality product. In Chapter 5, it will be discussed how one can directly use the final product quality measurements to develop the monitoring scheme by using Multi-way Partial Least Squares (MPLS).

Given the ease with which these multivariate monitoring charts were able to detect the simulated faults, one might expect that the faults would also be apparent in the trajectories of the individual variables. But as discussed in Section 2.3 and illustrated in Figures 2.3 and 2.6, the faults, resulting in products that barely violate the specification regions, are not readily apparent in the individual trajectories. The problem with the one at a time inspection of each variable trajectory, is that one is only looking at the magnitude and possibly the trends of the deviations in that one variable. However, the true process is multivariable, and all the variables are highly correlated with one another. The power of the proposed monitoring scheme results from the fact, that MPCA uses the joint covariance matrix of the variable trajectory deviations. By doing this, it utilizes not just the magnitude and trends of the deviation of each variable from its mean trajectory, but also

the correlation among all of the deviations over the history of the batch. It is this correlation structure among the variables that appears to be most important in detecting faults. When a fault occurs, the relationship among the trajectory deviations often changes substantially, even though their individual magnitudes may not be large.

The only requirement of the proposed method is the availability of a good database on past batches, and the ability to access the same data in real time for new batches. Sometimes data collection systems and data historians on industrial batch processes are inadequate. Historical data on all the variable trajectories are often not saved, but rather summarized by only a few raw statistics. The sensors for measuring the on-line variables must also be well maintained on a regular basis.

The sampling rate must be adequate for capturing the important trajectory information in the process. If there exists prior knowledge that a particular period during the batch is very important for product quality, then the sampling rate should be increased over that period. By increasing the sample rate it will be possible to track faster any deviations from the average trajectories over that period, and it will also weight that period more heavily in the MPCA model since there will be more p-loadings corresponding to it. Another way to weight a particular period or variable more heavily is with proper scaling when the MPCA model is being built. The method can also handle different numbers of measurement variables during different stages of the batch operation. One can either substitute zeros for the deviations at the time intervals that these variables are not measured, or augment column-wise the unfolded matrix **X** at the appropriate time intervals where these variables are measured.

In common with all on-line monitoring methods, these multivariate SPC methods can only detect "observable" events, that is events which influence at least one or more of the measured variables. No monitoring procedure can detect events that do not affect the measurements. This is analogous to the requirement of "observability" in state estimation of mechanistic models (Kuo, 1987). Some events which lead to quality problems may pass undetected if they have no impact on the measurement variables. The only way of

improving this situation is to add new measurements which are responsive to these events. Indeed, in another industrial batch process which we investigated, an important quality problem related to the surface properties of the product was not able to be detected by these MPCA charts. However, this had not been unexpected since none of the on-line measurements available was related to surface chemistry.

The batch runs must be comparable for MPCA to be effective. By comparable runs, we mean batches which operate in reactors of similar design, with the same catalysts, the same operational program, etc. If the operating procedure of the process changes, a new MPCA model must be built to accommodate this change. This is a general requirement of any method based on an empirical reference model. In our industrial example, the reactor unit is removed from service every several hundred batches for routine maintenance and cleaning. This cleaning changes the heat transfer characteristics of the reactor for the first few batches after the reactor is placed in service again, and special precautions and control in the operation of these early batches is required. Our industrial batch database was from a seasoned unit which makes the resulting MPCA model unsuitable for monitoring these first batches after the cleaning.

Throughout this thesis, the 95% and 99% control limits at each time interval are used in all the control charts. One is free to set any significance level in these limits with the usual trade-off of Type I and Type II error. In general, the instantaneous 99% control limits for the t-scores are reasonable and will not create many false alarms as it is indicated from the overall Type I error in Table 4.1. This is because the t-scores capture specific directions of variability in the measurement variables defined by the p-loadings. Any other variation is accumulated in the residuals, which makes them more susceptible to large values. Thus, in accordance with the suggested significance levels in the Shewhart charts, we propose the use of 99% and 99.9% control limits in the SPE-chart.

As a final comment on the presented approach, we point out that the post analysis of a database of batches and the estimation of control limits for the monitoring charts have substantial computational requirements, although the calculations are very simple. In a

486-machine, it takes no more than an hour to get the control limits for the SBR example. By relying upon large databases, one simply has to process a large amount of data. On the other hand, once this off-line analysis has been completed, the computational requirements for the on-line monitoring algorithm are extremely light and simple. It seems feasible for one to be able to track easily a batch process with a sampling rate of 5 to 10 seconds for a new set of observations.

CHAPTER 5 MPLS and Multi-Block Approaches in Batch Processes

This chapter covers how the proposed monitoring schemes can take into account any other information available for a batch run. Multi-way Partial Least Squares (MPLS) is used in the SBR example to incorporate the end product quality data (**Y**) which are available upon completion of the batch. A brief description of PLS is given at the beginning of this chapter, and approximate confidence intervals for the PLS predictions are developed. Multi-block MPLS is used on an industrial example to integrate additional data (**Z**) about the initial conditions and set-up of the batch process.

5.1 Partial Least Squares

PLS is a projection method, for the linear modeling of the relationship between a set of response variables (**Y**) and a set of predictor variables (**X**). Dr. H. Wold in 1982 derived the PLS algorithm, which now has become a rather popular method in a broad spectrum of sciences, such as chemistry, biology, psychology, industrial process control, and quality control (Geladi and Kowalski, 1986; Kvalheim, 1988; Kresta et al., 1991; Skagerberg et al., 1992; Dayal et al. 1994; Kettaneh-Wold et al. 1994). In all of these areas, PLS has demonstrated its efficiency and robustness to analyze large datasets, where collinearity (the matrix **X'X** is ill-conditioned or singular) is evident.

The birth of PLS originated in geometric intuition, rather than in traditional statistical arguments. Its geometric interpretation is closely related to the geometry of PCA. PLS can be interpreted as performing PCA on the covariance of **X** and **Y** (**Y'X**), and thus the decomposition is not affected only by the variance of the **X** and **Y** matrices but also by the correlation between them. PLS tries to accomplish two tasks simultaneously. It searches to find the directions of maximum variability in the x and y-space, and at the same time it tilts these directions so the direction in the x-space has maximum correlation

with the direction in the y-space. The t and u latent variables, which are linear combinations of the x and y variables respectively, are the projections of the data onto these directions. The mixed relation among the latent variables (t, u) in the x and y-space provide the ability of the PLS model to give adequate predictions. The essential criteria for the predictability of a regression model, is the number of variables included in the model. PLS uses the t and u latent variables to address the regression problem, and in this sense gives the minimum number of variables that is necessary.

Several algorithms have been proposed to extract the PLS components (Lingren et al., 1993; Kaspar and Ray, 1993) but in this thesis the NIPALS algorithm will be used for its simplicity and for being consistent with the PCA algorithm used in Section 2.2.

i. scale **X** (I×J) and **Y** (I×M) (usually by subtracting the mean of each column and divide by its standard deviation)

ii. choose a column of **Y** as **u**

iii. **w=X'u**

iv. **w=w/|w|**

v. **t=Xw**

vi. **q=Y't/(t't)**

vii. **u=Yq/(q'q)**

viii. if u has converged then go to step ix, otherwise go to step iii

ix. **p=X't/(t't)** **E=X-tp'** **F=Y-tq'**

x. go to step ii with **X=E** and **Y=F** to extract the next PLS component

It is important to note that the PLS algorithm selects only one pair of latent variables (t, u) at a time, and then uses the residual matrices (**E**, **F**) for the calculation of the second pair. The PLS components are selected in such a way, that give maximal reduction in the covariance (**F'E**) of the data (**E** and **F** for the first PLS component are equal to **X** and **Y**). This is accomplished in a step-wise procedure, which maximizes the correlation between the latent variables (t, u) at each step.

A brief outline of the mathematical properties of PLS will be given in this paragraph. The proofs of these properties can be found in Lorber et al. 1987, Manne 1987, Huskuldson 1988, Helland 1988, and Phatak et al. 1992. PLS decomposes the **X** (I ×J) and **Y** (I×M) matrices into a summation of R score vectors (**t** (I×1)) and loading vectors (**p** (J×1), **q** (M×1), plus some residual matrices (**E** (I×J), **F** (I×M)):

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r' + \mathbf{E} \quad , \quad \mathbf{Y} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{q}_r' + \mathbf{F}$$

or if we combine the **t**, **p**, and **q** vectors into **T** (I×R), **P** (J×R), and **Q** (M×R) matrices

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \qquad \mathbf{Y} = \mathbf{TQ}' + \mathbf{F}$$

where **T**, **P**, and **Q** are given by: $\mathbf{T} = \mathbf{XW}(\mathbf{P}'\mathbf{W})^{-1}$, $\mathbf{P}=\mathbf{X}'\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}$, $\mathbf{Q}=\mathbf{Y}'\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}$

and the regression coefficients (**B**, **Y=XB**) are given by: $\mathbf{B}=\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{Q}'$

The w-vectors are orthonormal, the t-vectors are orthogonal, and the matrix (**P'W**) is upper triangular with ones as diagonal elements. The **w** is eigenvector of **E'FF'E** the **q** is eigenvector of **F'EE'F**, **t** is eigenvector of **EE'FF'**, and **u** is eigenvector of **FF'EE'**. All these eigenvectors correspond to the maximum eigenvalue of these matrices which is equal to (**u't**)(**q'q**)(**t't**). Keep in mind, that **E** and **F** are the residual matrices of the previous PLS component upon which the **w**, **t**, and **u** were calculated.

The number of PLS components (R) is usually determined by cross-validation as in PCA (Wold, 1978; Stahle and Wold, 1987). There is a balance between bias and variance in the estimators. If you get too few components, you have a large bias and small variance; if you get too many you are in the opposite situation. The *R* statistic given in Section 3.2 will be used to determine the number of PLS components that one has to keep in his model. In PLS the Press_r and RSS_r statistics in *R* are based on the y-residuals since the purpose of the PLS model is to predict adequately the response variables (y).

Predictions for a new set of observations $\mathbf{x}_{\text{NEW}}$ (J×1) are given by:

$$\hat{\mathbf{t}} = \mathbf{x}_{\text{NEW}}'\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \quad , \quad \hat{\mathbf{y}} = \hat{\mathbf{t}}\mathbf{Q}' \quad , \quad \mathbf{e}' = \mathbf{x}_{\text{NEW}}' - \hat{\mathbf{t}}\mathbf{P}'$$

A great advantage of PLS over other regression methods is that the x-residuals (**e**) can be used to measure the regression model validity. If the x-residuals are large (compared to

the x-residuals in the model database), then the correlation structure of the x-variables has been changed and the y-predictions are not trustworthy for this set of observations (x_{NEW}). Many studies have been contacted to test PLS against other regression methods (Kvalheim and Karstang, 1989; Stone and Brooks, 1990; Kowalski, 1990; Phatak et al., 1992; Frank and Friedman, 1993), and in most cases PLS performs well with respect to the other methods.

5.1.1 Approximate Confidence Intervals for the PLS Predictions

A major problem in the statistical analysis of PLS is the nonlinear extraction of the PLS components. PLS does not only look at the conditional distribution of the y-variables given the x-observations, but treats both x and y as random variables connected through the latent variables t and u. In the following we shall treat only the case with univariate **y**. For the multivariate case (**Y**), one has to treat each of the y-variables separately. The **X** and **Y** matrices are assumed to be mean centered. The statistical properties which we shall derive in this section, are based on the work of Searle (1982) for regression based on generalized inverses. At the beginning of this section, we look at some properties of generalized inverses in linear regression and how they are related to PLS, and then we derive approximate confidence limits for the PLS predicted responses.

The regression problem $\mathbf{y}=\mathbf{X}\beta$ can always get a solution in the following form:

$$\mathbf{b}=\mathbf{G}\mathbf{y}$$

where **G** is a generalized inverse of **X**.

PLS gives a right weak generalized inverse **G** of the PLS approximation of the **X** matrix $\hat{\mathbf{X}}=\mathbf{T}\mathbf{P}'$ ($\hat{\mathbf{X}}\mathbf{G}\hat{\mathbf{X}}=\hat{\mathbf{X}}$, $\mathbf{G}\hat{\mathbf{X}}\mathbf{G}=\mathbf{G}$, $\hat{\mathbf{X}}\mathbf{G}=(\hat{\mathbf{X}}\mathbf{G})'$) which is given by:

$$\mathbf{G}=\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'$$

Rao and Mitra (1971), and Boullion and Odell (1971) show that a right weak generalized inverse of $\hat{\mathbf{X}}$, which has the same rank as $\hat{\mathbf{X}}$, gives the least squares solution for the problem $\mathbf{y}=\hat{\mathbf{X}}\beta$, in which:

$$|\hat{\mathbf{X}}\mathbf{b}-\mathbf{y}| \leq |\hat{\mathbf{X}}\mathbf{g}-\mathbf{y}| \qquad \forall \mathbf{g}$$

Although this minimum is defined uniquely, there is an infinite number of right weak generalized inverses of $\hat{\mathbf{X}}$, and thus regression coefficients, which give the same minimum. PLS give a particular solution (b, G) to the problem. The property of invariance of generalized inverses, guarantees that the predicted $\hat{\mathbf{y}}$ has a unique value ($\mathbf{Xb}$) no matter what right weak generalized inverse of $\hat{\mathbf{X}}$ one chooses to use. Although, PLS does not provide the minimum norm solution (min|b| which is unique), its solution is close to this, since the matrix $\mathbf{W(P'W)^{-1}P'}$ is generally close to being symmetric. In the latter case, G would also be a left weak generalized inverse of $\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}\mathbf{G}\hat{\mathbf{X}}=\hat{\mathbf{X}}$, $\mathbf{G}\hat{\mathbf{X}}\mathbf{G}=\mathbf{G}$, $\mathbf{G}\hat{\mathbf{X}}=(\mathbf{G}\hat{\mathbf{X}})'$). PLS provides the Moore-Penrose generalized inverse of the original X for the full rank decomposition of X, and in the case where X has full column rank, PLS provides the ordinary least squares solution $\mathbf{b}=(\mathbf{X'X})^{-1}\mathbf{X'y}$.

To proceed with Searle's analysis, one needs a generalized inverse of $\hat{\mathbf{X}}'\hat{\mathbf{X}}$. PLS gives the following reflexive generalized inverse L for $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}'\hat{\mathbf{X}}=\hat{\mathbf{X}}'\hat{\mathbf{X}}$, $\mathbf{L}\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{L}=\mathbf{L}$) :

$$\mathbf{L}=\mathbf{W(P'W)^{-1}(T'T)^{-1}(W'P)^{-1}W'}$$

Define the indempotent matrix $\mathbf{H}=\mathbf{G}\hat{\mathbf{X}}'\hat{\mathbf{X}}=\mathbf{W(P'W)^{-1}P'}$ ($\mathbf{H}^2=\mathbf{H}$), which has rank equal to the number of PLS components we have extracted (rank(H)=R). Under the assumption that the y-variable is distributed Normally as $N(\mathbf{X}\beta, \sigma^2)$, and that L is independent of the y-variable, we get the following statistical analysis:

$\mathbf{b}=\mathbf{L}\hat{\mathbf{X}}'\mathbf{y}+(\mathbf{H}-\mathbf{I})\mathbf{g}$	for arbitrary g (from now on, assume $\mathbf{g}=0$)
$E(\mathbf{b})=\mathbf{H}\beta$	b is a biased estimator of β
$\text{var}(\mathbf{b})=\mathbf{L}\sigma^2$	

The statistical test which we can derive from the above results, in the usual regression notation, is summarized in Table 5.1.

$SSR=\hat{y}'\hat{y}$	df = R	MSR=SSR/R
$SSE=(y-\hat{y})'(y-\hat{y})$	df = I-R-1	MSE=SSE/(I-R-1)

Table 5.1 Sum of Squares due to Regression (SSR) and Sum of Squared Errors (SSE) along with their degrees of freedom for the PLS model.

The statistic MSR/MSE has an F distribution with R and I-R-1 degrees of freedom, and the expected value of MSE is σ^2 (Searle, 1982). The problem is that this statistic does not check for significant regression ($\beta \neq 0$). It tests for the null hypothesis ($\mathbf{X}\beta=0$. The only conclusion we can derive, if this test is significant, is that the PLS model accounts for a significant portion of the variation in the y-variable. The β is not an estimable function, since $\mathbf{b}$ is not invariant to the generalized inverse of $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ that is used for $\mathbf{G}$ ($\mathbf{b}$ has an infinite number of descriptions). The only estimable function is any quantity $\mathbf{c}'\beta$, in which $\mathbf{c}'\mathbf{H}=\mathbf{c}'$. This $\mathbf{c}'\beta$ has $\mathbf{c}'\mathbf{b}$ as its Best Linear Unbiased Estimator which is distributed Normally as :

$$\mathbf{c}'\mathbf{b} \sim N(\mathbf{c}'\beta, \mathbf{c}'\mathbf{L}\mathbf{c}'\sigma^2)$$

This shows that we are not able in general to test for the significance of each coefficient separately, but only certain linear combinations of them. In spite of this, PLS provides a way to derive confidence intervals for the predicted y-variable since a new set of observations x_{NEW} (N×1) can be decomposed as $\hat{\mathbf{x}}'_{NEW}=\hat{\mathbf{t}}\,\mathbf{P}'$, and $\hat{\mathbf{x}}'_{NEW}\beta$ is an estimable function ($\hat{\mathbf{x}}'_{NEW}\mathbf{H}=\hat{\mathbf{x}}'_{NEW}$). Thus, the $\hat{\mathbf{x}}'_{NEW}\mathbf{b}$ is Normaly distributed with mean $\hat{\mathbf{x}}'_{NEW}\beta$ and variance $\hat{\mathbf{x}}'_{NEW}\mathbf{L}\hat{\mathbf{x}}_{NEW}\sigma^2$ which is equal to $\hat{\mathbf{t}}\,(\mathbf{T}'\mathbf{T})^{-1}\hat{\mathbf{t}}'\sigma^2$. The confidence interval at significance level α for an individual y-response ($\hat{\mathbf{x}}'_{NEW}\mathbf{b}+\varepsilon$, where ε is the error that is Normally distributed $N(0, \sigma^2)$), is given by :

$$\hat{y} \pm t_{I-R-1,\alpha/2}(MSE)^{1/2}(1+\hat{\mathbf{t}}(\mathbf{T}'\mathbf{T})^{-1}\hat{\mathbf{t}}')^{1/2}$$

where $\mathbf{T}$ and MSE is the t-score matrix and the Mean Squared Error of the PLS analysis of the data upon the PLS model was built, and $t_{I-R-1,\alpha/2}$ is the critical value of the Studentized variable with I-R-1 degrees of freedom at significance level $\alpha/2$.

The motivation behind the above analysis was to develop a simple expression for approximate confidence intervals for the PLS predictions. The confidence intervals for the

y-predictions derived above, are general for any PLS study. If the (1) in the $(1 + \hat{\mathbf{t}}(\mathbf{T}'\mathbf{T})^{-1}\hat{\mathbf{t}}')$ term is dropped, one gets the equation for the confidence interval of the expected value of a y-response ($\hat{\mathbf{x}}'_{NEW}\mathbf{b}$). Of course, the assumption that $\mathbf{L}$ is not a function of the y-variable is incorrect, and the above confidence interval is only an approximate one. Phatak et al. (1993) recognized this, and for the case of univariate $\mathbf{y}$ did a first order linear approximation of $\mathbf{b}$ around a set of observations ($\mathbf{X}_O$, $\mathbf{y}_O$) to get improved confidence intervals for $\hat{y}$. Although his approach is more accurate than the zero order approximation used here, it is computationally much more time consuming.

5.1.2 MPLS Analysis of Batch Data and On-Line Monitoring

Most batch and semi-batch processes operate in open loop with respect to product quality variables, simply because few, if any, on-line sensors exist for tracking these variables. Upon completion of the batch a range of quality measurements are usually made on a sample of the product in the quality control laboratory. The MPCA in the proposed SPC schemes only makes use of the process variable trajectory measurements ($\underline{\mathbf{X}}$) taken throughout the duration of the batch. Measurements on product quality variables ($\mathbf{Y}$) taken at the end of each batch were used only to help classify a batch as successful or unsuccessful. However, such product quality data can be used in a much more direct fashion. Multi-way Partial Least Squares (MPLS) can be performed using both the process data ($\underline{\mathbf{X}}$) and the product quality data ($\mathbf{Y}$). Rather than focusing only on the variance of $\underline{\mathbf{X}}$, MPLS focuses more on the variance of $\underline{\mathbf{X}}$ that is more predictive for the product quality $\mathbf{Y}$.

As in MPCA, one has to unfold $\underline{\mathbf{X}}$ ($I \times J \times K$) into $\mathbf{X}$ ($I \times JK$) and then perform a normal PLS between $\mathbf{X}$ and $\mathbf{Y}$. The columns of $\mathbf{X}$ and $\mathbf{Y}$ are scaled by subtracting their mean and dividing by their standard deviation. In this fashion MPLS summarizes and compresses the data with respect to both x and y variables and time into low dimensional spaces that describe the operation of the process (measurement variation around their

mean trajectories) which is most relevant to final product quality. MPLS decomposes **X** (I×JK) and **Y** (I×M) as **X=TP'+E**, and **Y=TQ'+F**. Each row of the **T** (I×R) matrix corresponds to a single batch and depicts the overall variability of this batch with respect to the other batches in the database. The **W** (JK×R) matrix summarizes the time variation of the measurement variables about their average trajectories, and its elements give the weights applied to each variable at each time interval within a batch to give the t-scores for that batch. The **Q** (M×R) matrix relates the variability of the process measurements to the final product qualities. When there are product quality measurements available during the batch, a three-way array $\underline{\mathbf{Y}}$ (I×M×K_Y) can be formed. The number of quality measurements (K_Y) during a batch run may not be the same as the number of process measurements (K). Usually, K_Y is much smaller than K because of the difficulties in measuring on-line quality variables. In such cases, the simplest way to apply MPLS is to unfold $\underline{\mathbf{Y}}$ (I×M×K_Y) into **Y** (I×MK_Y) in the same way of unfolding $\underline{\mathbf{X}}$ (I×J×K) into **X** (I×JK).

As in MPCA, batches with unusual operation will appear in MPLS either as batches with large t-scores, or with large residuals in the x-space ($Q_X = \sum_{c=1}^{KJ} \mathbf{E}(i,c)^2$), or with both. Additionally in MPLS, if the residuals for a batch in the y-space ($Q_Y = \sum_{c=1}^{M} \mathbf{F}(i,c)^2$) are large, it means that its final product qualities are not well predicted by its process measurements through the MPLS model.

The SBR example described in Section 2.3.1 will be used to illustrate the MPLS method. The resulting latex and polymer properties of the product in this example were summarized at the end of the batch in five quality variables (**Y** (50×5)): (1) composition (% styrene), (2) particle size (Å), (3) branching (branches / reacted monomer units), (4) crosslinking (crosslinks / reacted monomer units), and (5) polydispersity.

The ability of MPLS to discriminate between batches with acceptable product and unsuccessful batches was tested through a post analysis of the 50 normal batches plus the

two batches with product quality barely outside the acceptable region. MPLS was able to detect clearly these abnormal batches by placing them in the reduced space (t-plots) away from the main central cluster formed by the 50 normal batches, as MPCA did in Figure 2.4. Having established the observability of faults, an MPLS model was built from the 50 "good" batches, which summarizes the information contained in them about the normal operation of the process. This model will be used as the statistical reference to classify new batches as normal or abnormal.

Two PLS components were needed based on cross-validation (Section 5.1), to capture the variation of the process variables about their average trajectories which is most predictive of the final product qualities. The cross-validation R statistic for the first three PLS components found to be 0.47, 1.08, and 1.21. The cumulative percentage sum of squares explained (%SS) by the two principal components of the **X** and **Y** matrices and of each quality variable separately, is given in Table 5.2. One should always use cross-validation to determine the number of components in the PLS model and to assert its predictability. To rely only on the percentage of explained **Y** is misleading because of the large number of predictor x-variables (9×200=1800 in the SBR example). Any regression model could have accounted for a large portion of the variability in the **Y**. For the industrial example in Section 2.3.2, cross-validation revealed that only the first PLS component may be significant for predictions. The R statistic in this industrial example for the first three PLS components was found to be 1.03, 1.84, and 3.28.

	X	Y	Y1	Y2	Y3	Y4	Y5
Component 1	14.82	57.10	52.87	7.93	91.21	91.23	42.24
Component 2	23.05	65.08	54.30	20.79	91.28	91.29	67.74
MSR/MSE ($F_{2,47,0.05}$ = 3.20)			27.92	6.17	245.97	246.21	49.93

Table 5.2 Percent sum of squares explained in **X** and **Y** and in each of the quality variables by the first two PLS components for the SBR reference database.

The last row in Table 5.2 is the regression statistic Mean Sum of squares due to Regression (MSR) over the Mean Squared Error (MSE) (Section 5.1.1) with its 95% critical value, which shows how well the x-data account for the variation in each y-

variable. These F-tests provide another way from a regression point of view to check how well each of the y-variables is explained by the MPLS model. As it can be seen from Table 1, quality variables 3 and 4 are explained very well from the MPLS model and only quality variable 2 (particle size) is poorly explained by the process measurements. This arises because the particle size is determined largely by the variation in the number of seeded particles charged initially in the reactor, and it is not influenced much by resulting process conditions.

The t_1 vs. t_2 and the residuals plots indicated, as in MPCA, that there are no unusual batches in the reference database. Plots of the latent vectors t vs. u are shown in Figure 5.1. When the latent variables (t, u) of the x and y-space are highly linearly correlated, then all the observations in the t vs. u plots fall close to the diagonal of the graph. The plots in Figure 5.1 show that the variation explained by the first two PLS components in the x-space is very well correlated with the corresponding variation in the y-space. The linear nature of these plots suggests that nonlinear PLS (Wold, 1992) would probably not be needed. Indeed, performing such a nonlinear PLS gave essentially identical results to the linear analysis. The particular unfolding of $\underline{X}$ that is being used and the subtraction of the average trajectories from the process measurements have apparently

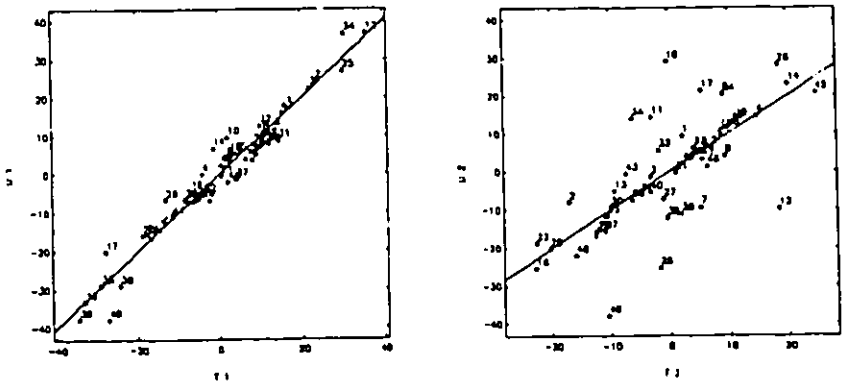

Figure 5.1 T vs. U plots for the two MPLS components. Each point represents one of the 50 batches in the reference SBR database. The t and u-observations, for both MPLS components, fall close to the diagonal line of the graph. This indicates that the MPLS latent variables in the x and y-space (t, u) are well correlated.

eliminated most nonlinear effects in the data. Plots of the percent explained in **X** with respect to time and variables in MPLS were similar to those obtained from MPCA in Figure 3.2.

The predicted t-scores ($\hat{\mathbf{t}}$ (1×R)), the predicted quality variables ($\hat{\mathbf{y}}$ (1×M)), and the residuals (**e** (1×KJ), **f** (1×M)) for a new batch $\mathbf{X}_{NEW}$ (K×J) are given by:

unfold and scale $\mathbf{X}_{NEW}$ (K×J) to $\mathbf{x}_{NEW}$ (KJ×1)

$$\hat{\mathbf{t}} = \mathbf{x}'_{NEW}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \quad , \quad \hat{\mathbf{y}} = \hat{\mathbf{t}}\mathbf{Q}' \quad , \quad \mathbf{e}' = \mathbf{x}'_{NEW} - \hat{\mathbf{t}}\mathbf{P}' \quad , \quad \mathbf{f} = \mathbf{y} - \hat{\mathbf{y}}$$

The problem which arises again in the on-line application of the above equations is that the $\mathbf{X}_{NEW}$ matrix is not complete until the end of the batch operation. The approaches suggested in Section 4.2 can be applied also in MPLS. The approach shown here uses the ability of PLS to handle missing data (Kresta et al., 1994; Nelson et al., 1995). PLS does this by projecting the already known observations up to time interval k ($\mathbf{x}_{NEW,k}$ (kJ×1)) into the reduced space defined by the **W** and **P** matrices in a sequential manner as following:

at each time interval k

$\mathbf{e}=\mathbf{x}_{NEW,k}$

for r = 1 to R

$\hat{\mathbf{t}}(1,r) = \mathbf{e}'\,\mathbf{W}(1{:}kJ,r) \,/\, (\mathbf{W}(1{:}kJ,r)'\mathbf{W}(1{:}kJ,r))$

$\mathbf{e}' = \mathbf{e}' - \hat{\mathbf{t}}(1,r)\mathbf{P}(1{:}kJ,r)'$

end

$$SPE_k = \sum_{c=(k-1)J+1}^{kJ} \mathbf{e}^2(c)$$

where the symbol (1:kJ,r) indicates the elements of the rth column from the first row up and to the kJth row. PLS, essentially, predicts these missing values by restricting them to be consistent with the already known values, and with the correlation structure of the process variables as defined by the **W** and **P** matrices. This approach gives t-scores very close to their final values as the $\mathbf{X}_{new}$ becomes complete, but during the first few time intervals may give poor estimates of the t-scores since there is so little information to

work with. However, in our experience with MPCA and MPLS on this and other examples (Nomikos and MacGregor, 1995; Kourti et al., 1995), this method works well by the time one has about 10% of the batch history. The reasons for this is that one is not building a PLS model based on large amounts of missing data, but using an already well established PLS model to predict the future behavior of a new batch. Furthermore, the early data are complete for all the measurement variables up to the current time interval, and are very good for predicting the future trajectory deviations which arise from variations in the initial batch charge conditions (ie. impurities, particle concentrations, etc.).

Now, one can calculate at each time interval the predicted t-scores, the predicted final quality variables, and the residuals. Note that there are no residuals (**f**) in the y-space during the on-line monitoring since the actual values of the quality variables will be known only at the end of the batch. Thus as in MPCA, one has to monitor the t-scores and the SPE for a new batch by using SPC charts which can be constructed as discussed in Chapter 4. If an abnormal situation is detected by either of these charts, one can diagnose the fault by interrogating the underlying MPLS model to find which process variables were primarily responsible for the detected deviations. This diagnostic information can be found by checking the contribution of each process variable to the deviations observed in the t-scores and residuals as described in Section 4.5.

The same multivariate SPC monitoring ideas that were developed using MPCA can be extended directly in MPLS. The SPC charts for the t-scores and SPE can be constructed exactly as described in Chapter 4. The additional information that one can get from MPLS is on-line inferences of the final quality of the product. MPLS gives, at each time interval, predictions of the final quality variables of the product. These predictions do not have anything to do with the actual values of the quality variables at a given time interval. They only refer to the values which the product quality variables will have upon completion of the batch. The assumption that the y-variables are distributed Normally as $N(\mathbf{X}\beta,\sigma^2)$ can be checked by plotting the y-residuals at each time interval for all the

batches in the reference database (Drapper and Smith, 1981). These plots for the SBR example showed no significant deviations from Normality.

Figure 5.2 shows the on-line monitoring charts, with their 95% and 99% control limits, for two batches. One batch is the normal batch used in Chapter 4, and it shows no abnormality in any of the monitoring charts. The other batch is the abnormal batch with the problem half-way through its operation, and is clearly flagged as abnormal in the SPE chart around time interval 105. After this time interval the observations from this batch move away from the reduced x-space. The MPLS model is not any longer valid, and one should treat the predicted t-scores with caution.

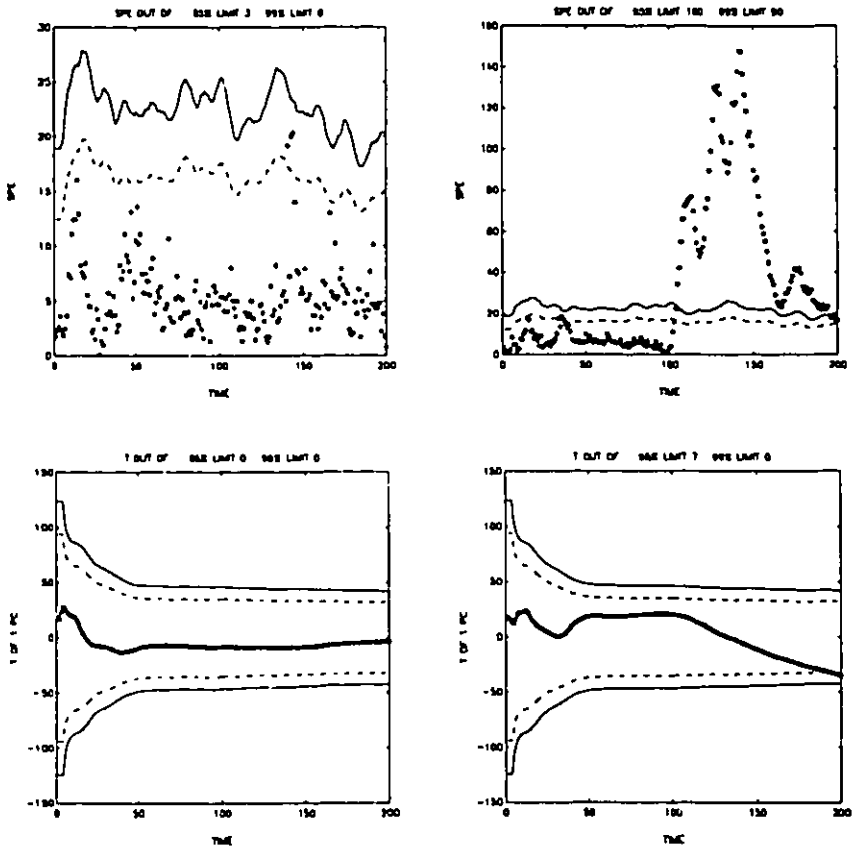

Figure 5.2 Monitoring charts for the SPE and t_1-scores with their 95% and 99% control limits (dashed and solid lines) for the normal SBR batch (left hand side plots) and for the abnormal SBR batch with problem halfway through its operation (right hand side plots). The abnormality in the bad batch is clearly flagged in the SPE chart after time interval 105.

Figure 5.3 shows the on-line predictions with their 95% and 99% confidence intervals, for three of the five quality variables for the two batches. The predictions for the normal batch match well the actual final quality values of the product. For the abnormal batch, the quality predictions after time interval 100 indicate its problem that the product qualities have started to change. Quality variable two, which was poorly explained in the MPLS model, has the poorest predictions for the abnormal batch. On the other hand, quality variable three, which was very well explained in the MPLS model, has very good predictions and its final prediction is very close to the final measured value for the abnormal batch. Since the MPLS model is no longer valid for the abnormal batch after time interval 105 (the SPE exceeds its control limits in Figure 5.2), the predictions after this time interval are not trustworthy. The confidence limits for the final product qualities are no longer valid, and they have not been plotted beyond this time interval in Figure 5.3. PLS models both x and y-spaces to give good predictions, and also provides a measure through its residuals in the x-space of how well the PLS model can be trusted. Although the predictions for the abnormal batch may not be accurate after time interval 105, the directions that the quality variables follow can be trusted in general, and this can help considerably in diagnosing the source of the abnormality.

5.1.3 MPCA or MPLS

The question that arises in monitoring batch processes is whether to use MPCA or MPLS. MPCA uses only the information about the process operational behavior ($\underline{\mathbf{X}}$) and its model describes how the on-line process measurements deviate from their average trajectories when the process operates in an "in-control" state. As a consequence, it will flag any abnormality in the process measurements even though it may be irrelevant to the quality of the product. As an example, a batch-run may have a slightly different agitator power profile because of a deterioration in its agitator mechanism. This event will cause an alarm in the MPCA monitoring. If the agitator power is not correlated with the final product qualities, the MPLS monitoring may not detect this deterioration in the agitator.

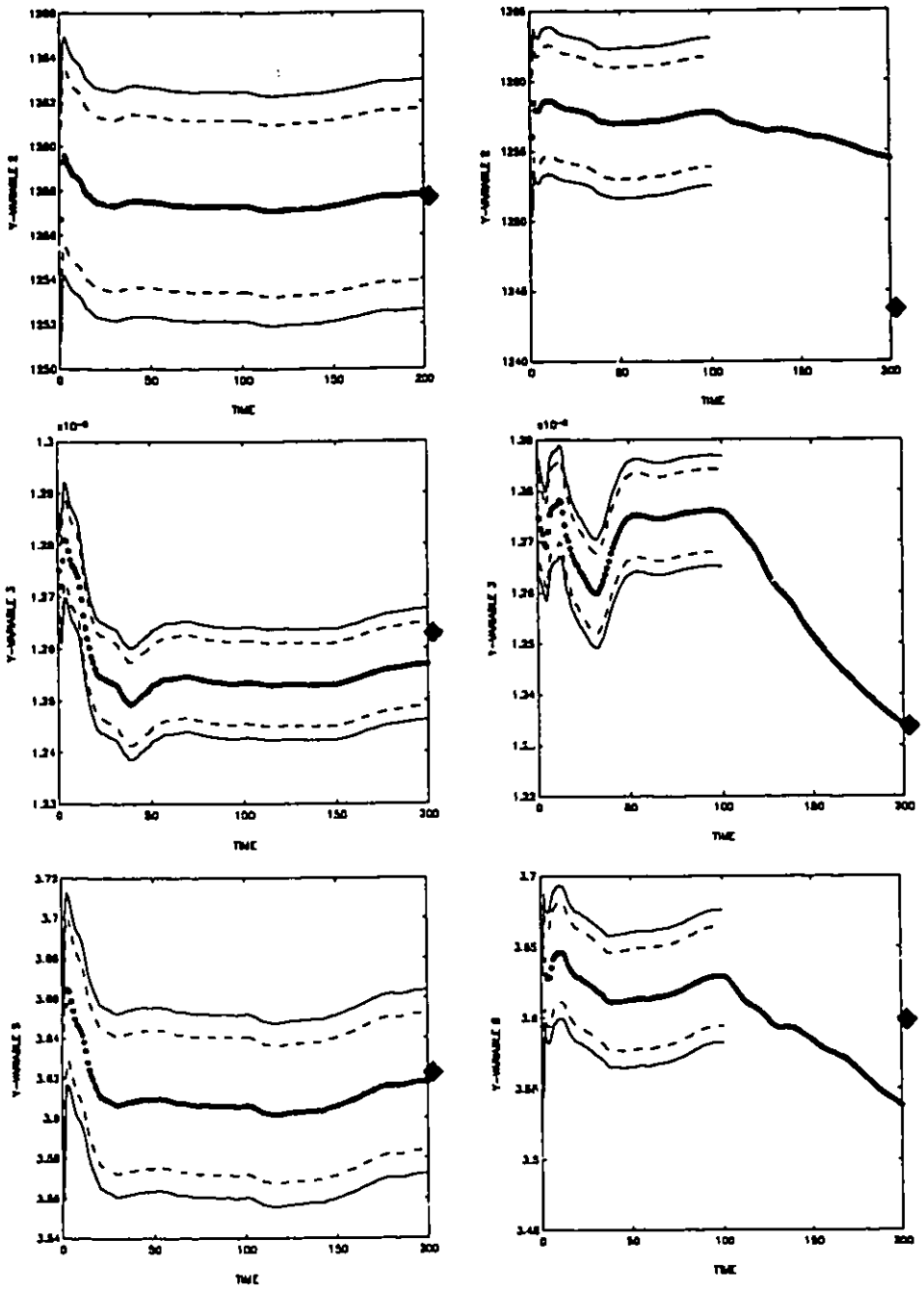

Figure 5.3 On-line predictions, with their 95% and 99% confidence intervals (dashed and solid lines), for three of the five final product quality variables for the normal SBR batch (left hand side plots) and for the abnormal SBR batch with the problem halfway through its operation (right hand side plots). The actual final product qualities are indicated by diamond marks. The PLS model for the abnormal batch is not valid after time interval 105 (absence of confidence intervals), and its predictions are not generally trustworthy.

Therefore, which approach one uses will depend upon whether or not one is primarily interested in events that will probably offset product quality or in any type of abnormal behavior. In general, it may be beneficial to try to detect all process deteriorations and correct them before they lead to permanent malfunctions.

A difficulty with using MPLS to analyze and monitor batch processes is having a sufficient number of quality variables which describe adequately the product quality. There are many types of batch quality variables. Typically, these are measurements of physical properties of the product, or variables which indicate if the product will have acceptable operation in the next stage, and sometimes variables from customer feedback. For an effective PLS, the **Y** matrix should have as columns, quality variables which are closely connected with the batch process, as to be well correlated with the process measurements. Also, these quality measurements should span a wide range of product properties because it is hard to believe that the whole batch operation can be reflected to a single quality measurement, or that one quality measurement can capture all the quality aspects of the final product. Another difficulty with the batch quality measurements is that they are usually susceptible to a significant amount of measurement error. In such cases, the uncertainty in the quality measurements can make the use of MPLS inappropriate.

A potential combination of MPCA and MPLS in monitoring batch processes is to build an MPCA model based on a reference database of normal batches, and an MPLS model based on a database of both successful (product within specifications) and unsuccessful batches. The MPCA model will be used for on-line monitoring as it has been described in Chapter 4. When an abnormality will be detected by the MPCA model, the MPLS model will be used to give on-line predictions of the final product quality as described in this chapter. In this case, the MPLS model will be valid over a wider range of product qualities, and its predictions will be in general more accurate than those based on an MPLS model built on only successful batches. Every time an unsuccessful batch will be detected, the MPLS model will be updated by augmenting its database with the new data from the unsuccessful batch.

5.2 Multi-Block MPLS

The batch monitoring schemes based on MPCA and MPLS proposed until now, can be extended to situations where the process can be naturally blocked into subsections. It is typical in batch processes to have a pre-processing stage (either in the same or in a different vessel) before the actual reaction stage. Also, extra information relevant to the batch process, usually is available in the form of a **Z** matrix. This **Z** matrix may contain information about feed-stock qualities, initial batch conditions of temperature and pressure, compositions of the initial charge, hold times in the reactor or in preprocessing vessels, and discrete conditions such as operator shifts or raw material suppliers. All these different blocks of information along with the end product qualities (**Y**) can be brought under a single SPC scheme (Figure 5.4) with the use of Multi-block projection methods

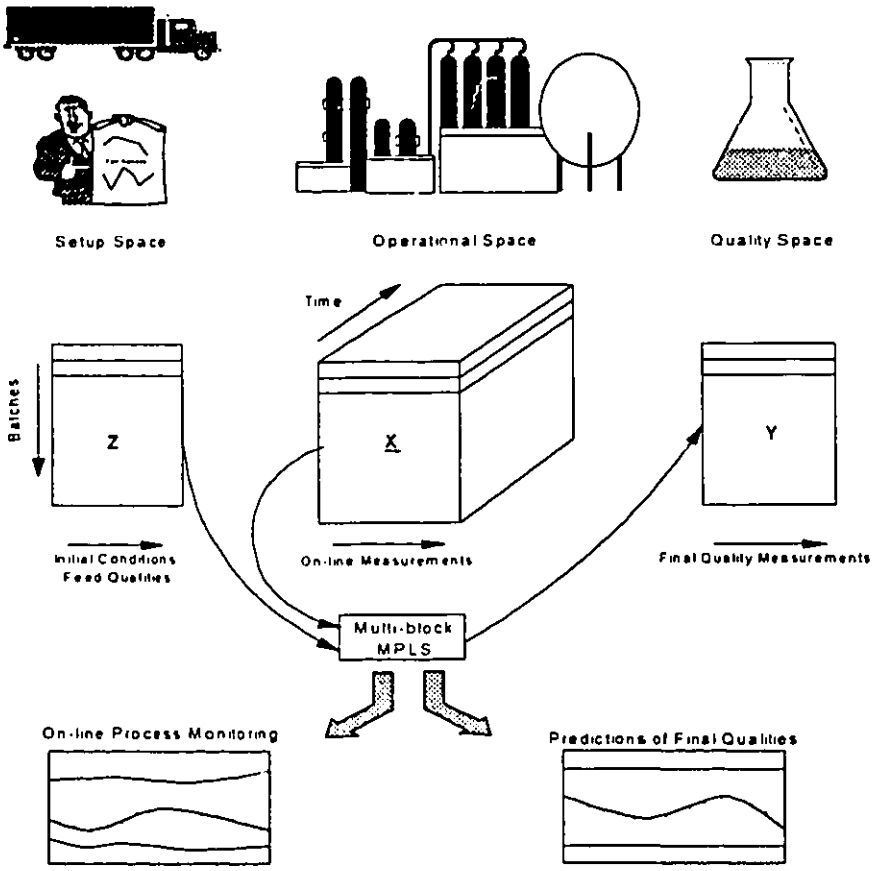

Figure 5.4 Overall SPC monitoring scheme for batch processes based on Multi-block projection methods.

(Wold, 1987b; Wangen and Kowalski, 1988). These methods allow one to establish monitoring charts for each block as well as for the entire process. Examples of Multi-block approaches in Chemical Engineering can be found in Slama (1991), MacGregor et al. (1994), and Kourti et al. (1995).

Although here the simple Multi-Block MPLS (MB-MPLS) algorithm will be presented and applied in an industrial example, the Multi-block MPCA algorithm is similar and can be found in Wangen and Kowalski (1988). Actually, the Wangen and Kowalski algorithm treats the more general problem where one has complex interblock relationships. Blocks that both predict and are predicted can be modeled through their algorithmic formulation.

Multi-block data analysis has its origins in the fields of sociology and econometrics. Multivariate projection methods based on the NIPALS algorithm, analyzing such block data are largely due to Herman Wold (1982) and Svante Wold (1987b). The MB-MPLS algorithm presented below is based on the work of Wangen and Kowalski (1988). Lets assume that one has A (a=1,2,..,A) $\mathbf{X}_a$ (I×M_a) predictor blocks and one explanatory $\mathbf{Y}$ (I×M) block.

MULTI-BLOCK MPLS ALGORITHM

i. unfold and scale $\mathbf{X}_a$ and $\mathbf{Y}$

ii. choose a column of $\mathbf{Y}$ as $\mathbf{u}$

iii. for a=1 to A

$\mathbf{w}_a=\mathbf{X}_a'\mathbf{u}$

$\mathbf{w}_a=\mathbf{w}_a/|\mathbf{w}_a|$

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

end

iv. $\mathbf{T}=[\mathbf{t}_1\ \mathbf{t}_2\ \ldots\ \mathbf{t}_A]$

v. $\mathbf{w}_c=\mathbf{T}'\mathbf{u}$

vi. $\mathbf{w}_c=\mathbf{w}_c/|\mathbf{w}_c|$

vii. $\mathbf{t_c=Tw_c}$

viii. $\mathbf{q=Y't_c/(t_c't_c)}$

ix. $\mathbf{u=Yq/(q'q)}$

x. if **u** has converged then go to step xi, otherwise go to step iii

xi. for a=1 to A

$\mathbf{p_a=X_a't_a/(t_a't_a)}$

$\mathbf{E_a=X_a-t_ap_a'}$

end

$\mathbf{F=Y-t_cq'}$

xii. go to step ii with $\mathbf{X_a=E_a}$ and $\mathbf{Y=F}$ to extract the next component

By extracting R components the $\mathbf{X_a}$ and $\mathbf{Y}$ matrices are decomposed as:

$$\mathbf{X_a} = \sum_{r=1}^{R} \mathbf{t_{a,r}} \mathbf{p'_{a,r}} + \mathbf{E_a} \quad , \quad \mathbf{Y} = \sum_{r=1}^{R} \mathbf{t_{c,r}} \mathbf{q'_r} + \mathbf{F}$$

The y-predictions $\hat{\mathbf{y}}$ (1×M) for a new set of observations $\mathbf{x_{NEW,a}}$ (1×Ma) are given by:

scale $\mathbf{x_{NEW,a}}$

$\mathbf{e_a=x_{NEW,a}}$, $\hat{\mathbf{y}}=\mathbf{0}$

for r=1 to R

for a=1 to A

$\mathbf{t_a=e_a'w_a}$

end

$\mathbf{t=[t_1\ t_2\ \ldots\ t_A]}$

$\mathbf{t_c=tw_c}$

$\mathbf{e_a'=e_a'-t_ap_a'}$

$\hat{\mathbf{y}}=\hat{\mathbf{y}}+\mathbf{t_cq'}$

end

The algorithm can be thought as an expansion of PLS where single measured variables in the x-space are replaced by blocks of measured variables. The order that the x-

blocks will be placed does not play any role in the algorithm. At each iteration, the algorithm tries to find the directions ($\mathbf{w}_a$) of maximum variability in the x-blocks. Then, it collects the t-scores ($\mathbf{t}_a$) from each block into a composite matrix **T** upon which a PLS iteration is performed to give a consensus t-score ($\mathbf{t}_c$) which is highly correlated with the y-block. Because of the deflation procedure in step xi, the t-scores for each block ($\mathbf{t}_a$) are orthogonal, but the consensus t-scores ($\mathbf{t}_c$) lack the orthogonality property.

How one will define the blocks is usually based on engineering judgment and on the objective of the study. General rules are given in Slama (1991) and MacGregor et al. (1994). Each block should define a process unit or data from highly correlated variables. The interaction between blocks should be minimal. Only variables which connect two blocks (such as feed streams) shall be present in more than one block. When the blocking has been done in a meaningful fashion where there is not much of interaction between blocks, the consensus t-scores ($\mathbf{t}_{c,r}$) are close to being orthogonal. A diagnostic test to check if the blocking was successful, is to compare the percent explained in the **Y** from the MB-MPLS model and that of MPLS models between each $\mathbf{X}_a$ and the **Y**. These should be comparable for the same number of components (R).

The advantage of putting different sections of a process or other set-up conditions into separate blocks ($\mathbf{X}_a$) is that one can establish monitoring charts and diagnostic plots for each block separately as described in Chapter 4. Additionally, the consensus t-scores ($\mathbf{t}_c$) can be used to overall monitor the process. By simply combining all the x-blocks into a single matrix $\mathbf{X}=[\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_A]$, one loses interpretation and the monitoring and diagnostic schemes can become cumbersome to manage and comprehend.

5.2.1 Scaling and Contribution of Each $\mathbf{X}_a$ Block

The scaling of each x-block is of great importance in MB-MPLS. By simply subtracting the mean of each column of $\mathbf{X}_a$ and **Y** and dividing by its standard deviation, one may give an undue importance to those x-blocks with the largest number of variables in them. An x-block with twice as many variables of another block, carries two times more

variance. Clearly, the x-block with the most variables will dominate the MB-MPLS model since the y-predictions will be heavily based only on the information that this x-block carries.

The scaling that will be used here is based on the assumption that each x-block is equally important to predict the **Y** matrix. Initially, all matrices ($\mathbf{X}_a$, **Y**) are scaled by subtracting the mean from each column and dividing by its standard deviation. Now, each $\mathbf{X}_a$ ($I \times M_a$) block has variance equal to M_a variance with it since the variance of each of its columns is 1. The total variance of all x-blocks is $\sum_{a=1}^{A} M_a$. The scaling factor (sf_a) for each $\mathbf{X}_a$ block which will give equal variance to all blocks, is given by:

$$sf_a = \sqrt{\frac{\sum_{a=1}^{A} M_a}{A\, M_a}}$$

By multiplying each x-block with this scaling factor, all x-blocks will have the same variance equal to $\sum_{a=1}^{A} M_a / A$. If one has knowledge of how important each x-block is for the y-predictions, one may multiply, on top of the proposed scaling, each x-block with a number that reflects its importance.

To find how much each x-block contributes to the y-predictions, one has to expand the equation for $\hat{\mathbf{Y}}$.

$$\hat{\mathbf{Y}} = \sum_{r=1}^{R} \mathbf{t}_{c,r}\mathbf{q}_r' = \mathbf{t}_{c,1}\mathbf{q}_1' + \mathbf{t}_{c,2}\mathbf{q}_2' + \ldots + \mathbf{t}_{c,R}\mathbf{q}_R' = \mathbf{T}_1\mathbf{w}_{c,1}\mathbf{q}_1' + \mathbf{T}_2\mathbf{w}_{c,2}\mathbf{q}_2' + \ldots + \mathbf{T}_R\mathbf{w}_{c,R}\mathbf{q}_R'$$

Each of these terms, which are the contributions of each component to the y-predictions, can be further expanded using the t-scores of each x-block ($t_{a,r}$) as:

$$\mathbf{T}_r\mathbf{w}_{c,r}\mathbf{q}_r' = t_{1,r}w_{c,r}(1)\mathbf{q}_r' + t_{2,r}w_{c,r}(2)\mathbf{q}_r' + \ldots + t_{A,r}w_{c,r}(A)\mathbf{q}_r'$$

Thus, a measure of the contribution of each $\mathbf{X}_a$ block ($cf_{a,r}$) to the y-predictions for a given component (r) is given by:

$$cf_{a,r} = (\mathbf{t}_{a,r}'\mathbf{t}_{a,r})\, w_{c,r}(a)\, (\mathbf{q}_r'\mathbf{q}_r)$$

The first term in the contribution factor depicts the amount of variance explained in each block. This amount of variance ($t_{a,r}'t_{a,r}$) is multiplied by its block weight ($w_{c,r}(a)$) to give the actual amount of variance that contributed to the y-predictions. The last term ($q_r'q_r$), which is common for all x-blocks for a given component (r), reflects the importance of each component to the predictions. This way by adding all the $cf_{a,r}$ factors for all components for a given x-block, one gets an overall measure of each x-block contribution to the y-predictions.

5.2.2 Multi-block MPLS in Industrial Data

Data supplied by DuPont Canada from an industrial polymerization process, similar to the one described in Section 2.3.2, will be used to illustrate the proposed MB-MPLS method. There are two distinct stages in the process. During the first stage, the solution of ingredients are loaded into an evaporator, where part of the solvent is vaporized and removed from the vessel under proper control of temperature and pressure. This stage takes approximately one hour. Upon finishing of this stage, the remaining reaction solution is transferred to the reactor where the polymerization takes place. Feedrate, temperature and pressure profiles are implemented with servo-controllers, and precise sequencing operations are produced with tools such as programmable logic controllers. After approximately 190 minutes, the finished product is expelled under pressure from the reactor vessel. Two critical property measurements related to the extent of the polymerization and to the product molecular weight distribution, are usually received 10 to 16 hours after the completion of the batch. These results cannot be used in a timely fashion to compensate for poor product quality. Furthermore, it is often difficult to establish what caused the quality variables to deviate from aim in the manufacture of an unsuccessful batch.

A dataset of 92 batches was provided from the above process. Unfortunately, all batches in this dataset were successful ones, and the company could not provide any unsuccessful batches. The purpose of the analysis will be to investigate if a Multi-block

SPC scheme is conceivable, and to identify major components of process variation. Four blocks of information were available for each batch. Two x-blocks consisted of the on-line measurements from the evaporator ($\underline{\mathbf{X}}_1$ (92×10×86)) and from the polymerization reactor ($\underline{\mathbf{X}}_2$ (92×21×231)). During the evaporator stage (86 time intervals), 10 on-line measurement variables are available around the evaporator and its heating system. Four of them are temperatures, four pressures, and two are flowrates. The polymerization stage has a duration of 231 time intervals, and 21 measurement variables are monitored. Eight variables are pressures, eleven temperatures, one flowrate, and one measurement from an energy balance.

For each batch there was also available information about the primary solution qualities and other pre-processing conditions. The first block of such information $\mathbf{Z}_1$ (92×9) consists of 9 quality variables of the primary solution of ingredients charged in the evaporator. These measurements are related to color, pH, temperature, and other physical properties of the solution. The primary solution is kept in tanks, and each tank supplies three to four batches. Thus, the same set of quality measurements is common for every three to four batches. The second block $\mathbf{Z}_2$ (92×8) contains any set-up and pre-processing information. Four operator shifts run all the batches and one indicator variable (1 or 0) has been assigned to each of them. The fifth variable in $\mathbf{Z}_2$ is the idle time before the start of the polymerization stage. It measures the waiting period for the reaction solution in the evaporator to be charged in the reactor vessel. The last three variables in $\mathbf{Z}_2$ are cycle times which measures the time for certain variables to achieve important temperatures or pressures during the polymerization stage.

First a preliminary MB-MPLS analysis on the whole database was conducted in order to identify any unusual batches. The process blocks ($\underline{\mathbf{X}}_1$, $\underline{\mathbf{X}}_2$) were unfolded to $\mathbf{X}_1$ (92×860) and $\mathbf{X}_2$ (92×4851) matrices. All matrices were mean centered and scaled to column unit variance, and the scaling factors (Section 6.2.1) used to give equal variance to each block are given in Table 5.3.

	Z_1	Z_2	X_1	X_2
sf	12.614	13.379	1.290	0.543

Table 5.3 Scaling factors for the x-blocks.

An MB-MPLS analysis of this database without using these scaling factors resulted in modeling only the variation of the huge X_2 matrix, ignoring any information from the other predictor blocks. This was evident in the t-score plots from this analysis where the t-score plot of X_2 was exactly the same as the plot of the consensus t-scores which measures the overall variability of each batch.

The t-plots from this preliminary MB-MPLS analysis are shown in Figure 5.5. The t-scores of three to four consecutive batches, as shown in the t-plot of Z_1, are the same since these batches had the same primary solution qualities. The second component, where its axis span (-60, 60) is double than that of the first component (-30, 30), indicates batches 54 through 57 to be different. Variable 6 was unusually high for these batches. This variable is a measure of water conductivity which has little effect on product quality, although such high values as these in batches 54 through 57 are alarming for the process operation. The t-plot of Z_2 discriminates batches 1 and 30 from the rest. The idle time for these batches had high values. Batch 30 stayed almost twice the average time in the evaporator, and its contents became quite hot. This had a major effect in the polymerization stage where this batch achieved temperature and pressure set points faster than any other batch in the database. This is why this batch is also flagged in the t-plot of the X_2 block. Its behavior had an impact in the product quality variables which had borderline values. Batch 1 stayed in the evaporator a considerably longer time than batch 30 (about ten times longer). Its contents were heated for the first part of its idle time, but then gradually it cooled off. It is suspected that some polymerization took place during the last part of its idle time, resulting in a peculiar behavior at the beginning of the polymerization stage. Most of its variables were below their mean trajectories, something which resulted in this batch being placed away from the central cluster in the t-plot of X_2. One of its quality variables had a somewhat high value. The fact that the product qualities

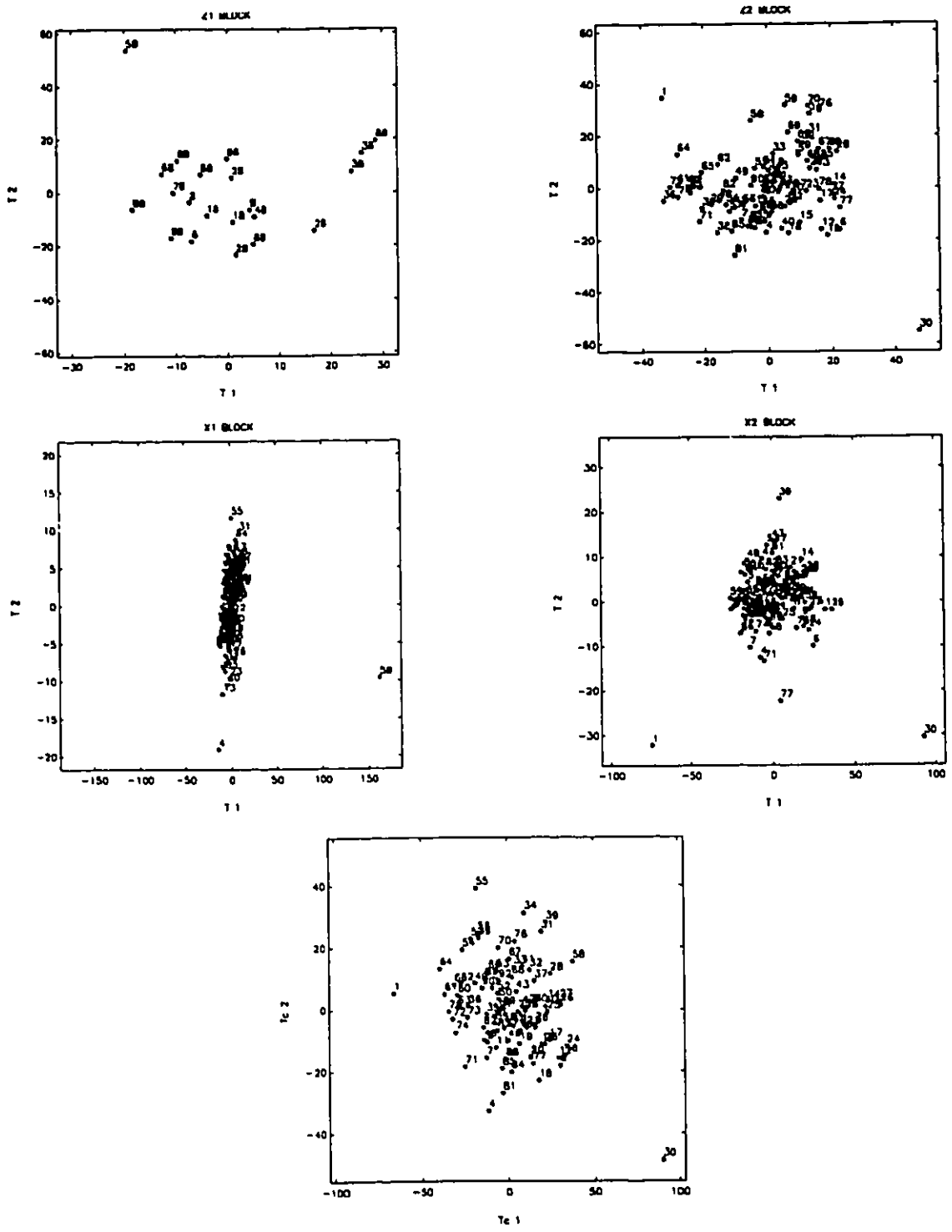

Figure 5.5 Multi-block MPLS analysis of the industrial example. T-score plots for each z and x-block. The consensus t-score plot is given at the bottom of the figure.

of batch 30 were more affected than those of batch 1 is reflected on the position of these batches in the t-plots of the $\mathbf{Z}_2$, $\mathbf{X}_2$, and in the consensus t-plot. In all these plots, batch 30 has been placed further away from the central cluster of normal batches than batch 1. The t-plot of $\mathbf{X}_1$ identifies batch 58 as very unusual. All of its trajectories were totally different from the mean trajectories in the evaporator stage. No explanation was given for its strange behavior, or why this behavior did not have any impact in the product qualities. Batches 1 and 30 are identified as different in the consensus t-plot. Only these batches had an unusual behavior ($\mathbf{Z}_2$, $\mathbf{X}_2$) which also affected their product qualities.

A second MB-MPLS analysis was run, this time excluding batches 1, 30, and 54 through 58. The scaling was the same as in the preliminary analysis. Two components were found to be significant based on cross-validation. The R statistic for the first three components was found to 0.89, 0.59, and 1.11. The t-plots for all the blocks and the consensus t-plot showed no other abnormal batches. The contribution of each block to the predicted $\hat{\mathbf{Y}}$ is given in Table 5.4, and the percent explained from each block is given in Table 5.5.

	$\mathbf{Z}_1$	$\mathbf{Z}_2$	$\mathbf{X}_1$	$\mathbf{X}_2$
Component 1	2.08	13.24	3.37	12.89
Component 2	7.61	15.35	2.05	00.92
Overall	9.69	28.59	5.42	13.81

Table 5.4 Contribution factors of each block to the y-predictions, for each component and overall.

Block $\mathbf{Z}_2$ is the most important block for the product quality because of its idle time and its cycle times which play a very significant role in the polymerization stage. The indicator variables for the crew shifts have a minimal effect in the predictions because of the very low values in their w-loadings. The next significant block is $\mathbf{X}_2$ where the polymerization operation is highly related with the idle and cycle times of block $\mathbf{Z}_2$. The least important block is the evaporator $\mathbf{X}_1$ which seems to have little effect on quality. Note that the percentage explained in the x-blocks (Table 5.5) are not good indicators of the importance of each x-block to predict $\mathbf{Y}$. It is clear from the above analysis that by

controlling closely the idle and cycle times of Z_2, the reactor operation (X_2) becomes more reproducible and should result in the production of consistently high quality product.

	Z_1	Z_2	X_1	X_2	Y
Component 1	19.17	12.75	35.05	22.44	11.81
Component 2	44.14	27.30	39.58	37.69	31.14

Table 5.5 Percentage of explained sum of squares (cumulative) from the MB-MPLS analysis of the industrial example

The block contributions to the product qualities coincide with what the company's engineers think about the importance of each block. The weighting factors of importance that they gave us were: 1 for Z_1, 0 for the crew shifts and 3 for the rest Z_2, 2 for X_1, and 4 for X_2. An MB-MPLS analysis of the normal database, using the company's scaling factors as described in Section 5.2.1, gave comparable percentage explained in the x-blocks and **Y** with those in Table 5.5 and the t-plots were very similar to those in Figure 5.5.

Although the percentage explained in **Y** by extracting two components is only about 30%, one has to consider the measurement error in the quality variables. The variance of the measurement error in the two quality variables is estimated to be 0.323e-4 and 0.087e-4 respectively. The variance of the two quality variables in the normal database was found to be 0.818e-4 and 0.667e-4 respectively. Therefore, only about 70% of the total variability in **Y** can be theoretically explained. From this perspective, the MB-MPLS model accounts for more than 50% of the explainable variability in the **Y**.

As discussed in Section 5.2, one way to check if the blocking was sensible and effective in predicting **Y**, is to compare the percent explained **Y** in the MB-MPLS model with that based on MPLS analyses between each x-block and **Y**. The results from such analyses can be found in Table 5.6. In these analyses, each matrix was mean centered and column unit variance scaled.

	Z_1	Y	Z_2	Y	X_1	Y	X_2	Y
Component 1	21.05	3.27	19.07	21.88	12.77	15.45	35.08	9.74
Component 2	31.55	9.23	43.04	23.96	26.81	30.57	39.15	27.71

Table 5.6 Percentage of explained sum of squares (cumulative) from MPLS analyses between each x-block and **Y**.

The above percents explained sum of squaresare comparable with those in Table 5.5. The MB-MPLS model outperforms any individual MPLS model. Note, that the percentages of **Y** explained in Table 5.6 are not characteristic of how well each x-block predicts **Y**. Cross-validation could have answered this question. The MB-MPLS model by hierarchically combining information from each x-block extracts variability directions in each x-block which are interrelated and most predictive for **Y**.

Based on the above MB-MPLS analysis, it should be reasonable to build a monitoring scheme according to the principles given in Chapter 4. The main advantage will be to identify potential unsuccessful batches early from the beginning of an operation. In other examples examined (Kourti et al., 1995) an MB-MPLS model clearly alarmed batches with problems indicated in feed qualities or in pre-processing conditions which resulted in product out of specifications.

CHAPTER 6 Comparing Covariance Matrices and Subspaces Developed by Projection Methods

Projection methods such as PCA and PLS are used in a broad spectrum of sciences including psychology, biology, chemistry, chemical engineering, and quality control. In many cases, one wishes to check if the projection model developed for one process or physical system is valid to be used on another with a similar configuration, and in the case that it is not valid, to identify the differences between the two systems. As an example from chemical engineering, consider two parallel units (eg. reactors or distillation columns) which have the same design and operating policy, and one wants to apply the PCA or PLS model developed based on data from one of the units to the other unit. The same question comes up also when one wishes to switch from a PCA to a PLS model, and vice versa. It is of interest to know how different are the two models in modeling the x-space and what are their main differences. Also, in cases where the two models have different number of components, it is of interest to check if the low dimensional space is contained in the space defined by the higher order model. Another important situation is to know when the correlation structure in the measurement variables has changed significantly, and thus, the projection model used to monitor or control the process has to be updated. This is a common situation in chemical engineering since most of the processes are improved or redesigned periodically.

Comparing subspaces based on projection methods is equivalent to a comparison of covariance matrices. This is an interesting problem in statistics, and in the first part of this chapter we shall explore traditional statistical methods indicating their weaknesses. A new statistical approach will be presented which has a nice geometrical interpretation. Its basic development will be given, along with suggestions for improving its distributional features. Two examples will be given for this approach. The first example is based on the industrial polymerization process of Section 2.3.2, where we shall test if the same MPCA

model can be used in two parallel batch reactors. In the second example, we shall examine if the MPCA (Section 3.3) and MPLS (Section 5.1.2) models of the SBR example define the same subspace in the x-space. The chapter will close with a discussion of the difficulties of tests based on residuals from projection models.

6.1 Maximum Likelihood Ratio Test

Let $\mathbf{X}_A$ ($n_A \times m$) and $\mathbf{X}_B$ ($n_B \times m$) be mean centered matrices of observations from Multinormal distributions $N(\mu_A,\Sigma_A)$ and $N(\mu_B,\Sigma_B)$ respectively. The maximium likelihood ratio test statistic for equivalence of covariance matrices (H_0: $\Sigma_A=\Sigma_B$, H_1: $\Sigma_A \neq \Sigma_B$) is given by (Seber, 1984; Mardia et al., 1989):

$$l = n_{AB} \ln(\det(S_{AB})) - n_A \ln(\det((n_A-1)S_A/n_A)) - n_B \ln(\det((n_B-1)S_B/n_B)) \sim \chi^2_{m(m+1)/2}$$

where: $n_{AB}=n_A+n_B$, $S_{AB}=((n_A-1)S_A+(n_B-1)S_B)/n_{AB}$, $S_A=\mathbf{X}_A'\mathbf{X}_A/(n_A-1)$, $S_B=\mathbf{X}_B'\mathbf{X}_B/(n_B-1)$, and $\ln(\cdot)$ and $\det(\cdot)$ symbolize the natural logarithm and the determinant respectively. It is a one-tailed test, and its critical level (p) is given by the probability of l exceeding its observed value under the null hypothesis H_0. The test becomes less accurate as the ratio n_A/n_B differs significantly from unity. Several approximations have been suggested (Box, 1949; Gnanadesikan and Lee, 1970), and their properties (power of the test, exact distribution, sensitivity to normality) have been studied for certain values of n_A, n_B, and m (Ito, 1969; Layard, 1974; Lee et al., 1977; Nagarsenker, 1978).

The maximum likelihod ratio test and all of its approximations are based on determinants of sample covariance matrices. This is the real weakness of the test because it checks mainly for equality of the overall variance ($\det(S_i)$) that each sample carries, and ignores differences in directionality making out this variance. This will be illustrated through an example shown in Figure 6.1. Part of the problem in the maximum likelihood ratio test is that it tries to reduce all the information from the sample covariance matrices into a single statistic (l) for characterizing the overall covariance similarity between two samples, and it does this by evaluating the likelihood function only at one point: its maximum.

The sample covariance matrix (S) of a mean centered observation matrix **X** (n×m) can be decomposed based on principal component analysis (Section 2.1) as:

$$\mathbf{S} = \mathbf{X'X}/(n-1) = (\mathbf{TP'})'(\mathbf{TP'})/(n-1) = \mathbf{P}(\mathbf{T'T})\mathbf{P'}/(n-1) = \mathbf{P}\Lambda\mathbf{P'}/(n-1)$$

The orthonormal loading matrix **P** has as column vectors the directions of maximum variability, and the diagonal matrix ($\Lambda/(n-1)$) has as elements the variance that each principal direction carries. The determinant of **S** is equal to the product of the elements of ($\Lambda/(n-1)$).

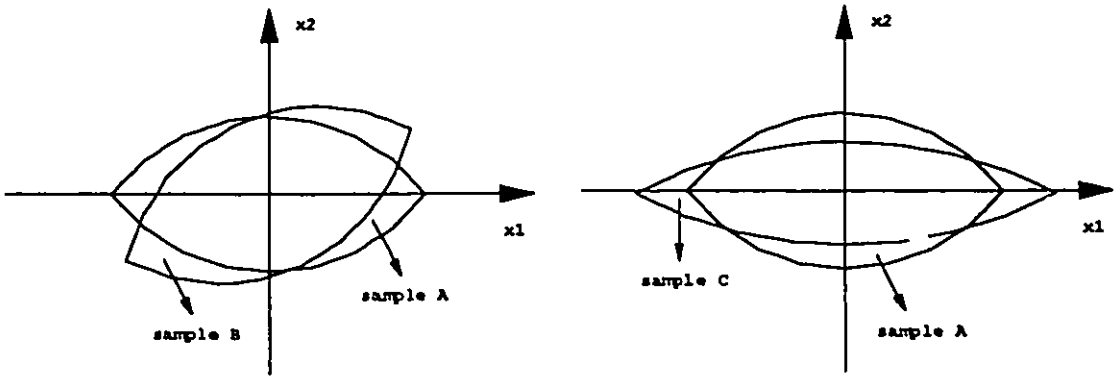

Figure 6.1 Ellipses of two samples which have equal sample covariance determinants. The axis lengths of these ellipses are equal to one standard deviation. In the left hand side plot, the principal axes of sample B have been tilted by approximately 25°. In the right hand side, both samples have the same principal axes, but the variance in each direction is different.

Figure 6.1 has two examples which will be discussed in this and in the following section. The left hand side plot in Figure 6.1 shows the distribution of two samples ($\mathbf{X}_A$ and $\mathbf{X}_B$). The number of variables (m) is 2, and for convenience the number of observations (n_A, n_B) is assumed to be equal to 31 for both samples. The two sample covariance matrix can be written as:

$$\mathbf{S}_A = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} = \frac{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 120 & 0 \\ 0 & 30 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}'}{31-1} \quad , \quad \det(\mathbf{S}_A) = 4$$

$$S_B = \begin{bmatrix} 3.4 & 1.2 \\ 1.2 & 1.6 \end{bmatrix} = \frac{\begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 120 & 0 \\ 0 & 30 \end{bmatrix} \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix}'}{31-1} \quad , \quad \det(S_B) = 4$$

Both sample covariance matrices have equal determinants, which results in ellipses of equal area in Figure 6.1. The difference in these two covariance matrices is that the principal axes of S_A have been tilted by approximately 25°. The maximum likelihood ratio test gives l=6.609 which has a critical level (χ^2_3) of p=0.087. This test is influenced largely by the determinants of the sample covariance matrices and ignores differences in the principal directions of variability.

The real problem with this test comes apparent as the number of variables increases and the variables become more correlated. In most examples encountered in practice, the sample covariance matrices are ill-conditioned or singular. As a result, their determinants are close or equal to zero, which makes this test and any of its proposed approximations totally inappropriate.

6.2 A Geometric Test

Krzanowski (1979) gave a test for comparison of covariance matrices based on the underlying geometry of multivariate observations. This test examines the angles formed between principal components from different samples. Let $S_A=P_A\Lambda_AP_A'/(n_A-1)$ and $S_B=P_B\Lambda_BP_B'/(n_B-1)$ be the two sample covariance matrices as described in Section 5.1. The angle cosine between the ith principal component of sample A and the jth principal component of sample B, is given by the (i,j) element of $P_A'P_B$. By looking at the diagonal elements of the $P_A'P_B$ matrix, one can judge if the principal components from different samples coincide or not. The cosine of the minimum angle between an arbitrary vector in the space defined by the principal components of sample A and the one most nearly parallel to it in the space of sample B, is given by $\sqrt{\lambda_1}$, where λ_1 is the largest eigenvalue of $P_A'P_BP_B'P_A$. A small minimum angle indicates that the two spaces nearly coincide. Also,

the trace of $\mathbf{P_A'P_BP_B'P_A}$ is between 0 and R, where R is the number of principal components retained in the sample covariance matrices. A value close to 0 indicates that the two spaces are orthogonal, and a value close to R indicates that the two spaces coincide. The trace of $\mathbf{P_A'P_BP_B'P_A}$ and the minimum angle provides a measure of the overall compatibility of the two covariance matrices. This approach is attractive because of its geometric interpretation, it can handle ill-conditioned or singular sample covariance matrices, and one needs only the p-loadings from a principal component analysis to apply it.

We now investigate how this approach performs on the example of Section 6.1 (left hand side plot of Figure 6.1). The angles between the principal components of the two samples are indicated in parenthesis.

$$\mathbf{P_A'P_B} = \begin{bmatrix} 2/\sqrt{5}\ (26.5^{\circ}) & 1/\sqrt{5}\ (63.4^{\circ}) \\ 1/\sqrt{5}\ (63.4^{\circ}) & -2/\sqrt{5}\ (26.5^{\circ}) \end{bmatrix} \quad , \quad \mathbf{P_A'P_BP_B'P_A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

It clearly shows that the principal components of the second sample have been tilted by approximately 25° from those of the first sample. The minimum angle is zero, and the trace of $\mathbf{P_A'P_BP_B'P_A}$ is 2 equal to the number of principal components in the covariance matrices. Of course in this example by using two principal components in the sample covariance matrices, we have filled the two dimensional space, and thus the conclusions from $\mathbf{P_A'P_BP_B'P_A}$ are meaningless. Although in this case it was trivial to interpret why the $\mathbf{P_A'P_BP_B'P_A}$ indicated that the two spaces coincide, in larger problems this becomes a drawback of the method. The same space can be spanned by two different sets of principal components, and this will become apparent only from the angle cosines in the $\mathbf{P_A'P_B}$ matrix. But, even the angles cosines do not give any information about any differences in the amount of variability that each principal direction carries. In addition, the interpretation of the angle cosines is not straightforward since there is not a measure (statistic with known distribution) of what angle magnitudes are considered significant for two covariance matrices to be different.

In the case where the principal directions are similar but the variability in these directions are different, the above approach will fail to detect any difference. This situation is illustrated in the right hand side plot of Figure 6.1. The two sample covariance matrices in this case are:

$$\mathbf{S}_A = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} = \frac{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 120 & 0 \\ 0 & 30 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}'}{31-1} \quad , \quad \det(\mathbf{S}_A) = 4$$

$$\mathbf{S}_C = \begin{bmatrix} 7.5 & 0 \\ 0 & 0.533 \end{bmatrix} = \frac{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 225 & 0 \\ 0 & 16 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}'}{31-1} \quad , \quad \det(\mathbf{S}_A) = 4$$

Both have the same principal directions, but each direction carries a different amount of variability. The $\mathbf{P}_A'\mathbf{P}_C$ and $\mathbf{P}_A'\mathbf{P}_C\mathbf{P}_C'\mathbf{P}_A$ matrices are both equal to the identity matrix, which shows a perfect covariance similarity between the two samples. The maximum likelihood ratio test for this example gave l=6.026 which has a critical level (χ^2_3) of p=0.111. Again its p-value is quite high because the determinants of the two sample covariance matrices are equal.

6.3 A New Approach: Directional Variance Test

In this section we shall develop a new approach for comparing covariance matrices. The approach is based on both the geometric ideas discussed in Section 6.2 and on the Union Intersection Test (UIT) principles. A UIT is not available for comparing two sample covariance matrices (Seber, 1984; Mardia et al., 1989). Such a test would have been able to test for overall equality of variance in every direction in the x-spaces defined by the principal components of the two sample covariance matrices. Schuurmann et al. (1973) have investigated approximate UIT for selected values of n_A, n_B, and m. The proposed statistical test assesses whether the variance in any single direction of a sample space A equals the variance that this direction has in a sample space B. In addition, it finds

the worst direction with variance dissimilarity in the two spaces, and uses this as an overall test.

Let $\mathbf{X}_A$ ($n_A \times m$) and $\mathbf{X}_B$ ($n_B \times m$) be mean centered matrices of n_A and n_B observations of m variables from Multinormal distributions $N(0,\Sigma_A)$ and $N(0,\Sigma_B)$ respectively. Let also $\hat{\mathbf{X}}_A = \mathbf{T}_A\mathbf{P}_A'$ and $\hat{\mathbf{X}}_B = \mathbf{T}_B\mathbf{P}_B'$ be their principal component approximations (Section 2.1) with $\mathbf{T}_i$ ($n_i \times R_i$), $\mathbf{P}_i$ ($m \times R_i$), and $m \geq R_i$ (i=A,B). The $\hat{\mathbf{X}}_A$ and $\hat{\mathbf{X}}_B$ matrices may have a different number (R_i) of principal components. In such cases, one wants to check if the low dimensional space is contained into the higher one. From Multinormal distribution theory (Mardia et. al. 1989), we know that:

$$\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A \sim W_m(\Sigma_A, n_A - 1) \quad , \quad \hat{\mathbf{X}}_B'\hat{\mathbf{X}}_B \sim W_m(\Sigma_B, n_B - 1)$$

where $W_m(\Sigma_i, n_i-1)$ denotes the Wishart distribution with scaling matrix Σ_i and n_i-1 degrees of freedom. For a fixed m-vector a such that $a'\Sigma_A a \neq 0$ and $a'\Sigma_B a \neq 0$, we have that (Mardia et al., 1989):

$$\frac{a'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A a}{a'\Sigma_A a} \sim \chi^2_{n_A-1} \quad , \quad \frac{a'\hat{\mathbf{X}}_B'\hat{\mathbf{X}}_B a}{a'\Sigma_B a} \sim \chi^2_{n_B-1}$$

The statistic which tests for variance equality between two spaces in the direction of a fixed vector a (H_0: $a'\Sigma_A a = a'\Sigma_B a$, H_1: $a'\Sigma_A a \neq a'\Sigma_B a$), is given by:

$$VD = \frac{a'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A a}{a'\hat{\mathbf{X}}_B'\hat{\mathbf{X}}_B a} \overset{H_0}{=} \frac{\dfrac{a'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A a}{a'\Sigma_A a}}{\dfrac{a'\hat{\mathbf{X}}_B'\hat{\mathbf{X}}_B a}{a'\Sigma_B a}} \sim \frac{n_A - 1}{n_B - 1}\frac{\chi^2_{n_A-1}/(n_A-1)}{\chi^2_{n_B-1}/(n_B-1)} \sim \frac{n_A - 1}{n_B - 1}F_{n_A-1,n_B-1}$$

The test statistic VD checks for variance equality along the projections of a in the spaces defined by the principal directions of the two sample covariance matrices. The restriction that the projection of a in either space is different from zero ($a'\Sigma_i a \neq 0$), works as a safety to prevent the VD statistic to become either zero or infinity. For the univariate case (m=1, a scalar, x_i ($n_i \times 1$) vector), the VD statistic gives the classical test for variance equality (H_0: $\sigma_A^2 = \sigma_B^2$, H_1: $\sigma_A^2 \neq \sigma_B^2$):

$$VD = \frac{\mathbf{a}'\mathbf{x}_A\mathbf{a}}{\mathbf{a}'\mathbf{x}_B\mathbf{a}} = \frac{s_A^2}{s_B^2} \sim F_{n_A-1,n_B-1}$$

where s_i^2 is the sample variance of sample i.

The proposed test is a two-tailed test and its critical level (p) is given by twice the probability of VD exceeding its observed value (or becoming smaller than its observed value in the case that is smaller than 1) under the null hypothesis H_0. An acceptable region for VD at significance level α is given by:

$$(F_{n_B-1,n_A-1,\alpha/2})^{-1} = F_{n_A-1,n_B-1,1-\alpha/2} \leq VD \leq F_{n_A-1,n_B-1,\alpha/2}$$

Lets assume that $\mathbf{a}$ belongs to the eigenspace of Σ_B of non-zero eigenvalues ($\mathbf{a}'\Sigma_B\mathbf{a}\neq 0$), and thus it can be written as a linear combination of the principal components of space B:

$$\mathbf{a} = \mathbf{P}_B\Lambda_B^{-1/2}\mathbf{c}$$

where $\mathbf{c}\neq\mathbf{0}$ is an arbitrary ($R_B\times 1$) vector. By restricting $\mathbf{a}$ to belong in the eigenspace of non-zero eigenvalues of the sample covariance matrix (S_B) in the denominator, we can find the minimum and maximum value of VD.

$$VD = \frac{\mathbf{a}'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A\mathbf{a}}{\mathbf{a}'\hat{\mathbf{X}}_B'\hat{\mathbf{X}}_B\mathbf{a}} = \frac{\mathbf{c}'\Lambda_B^{-1/2}\mathbf{P}_B'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A\mathbf{P}_B\Lambda_B^{-1/2}\mathbf{c}}{\mathbf{c}'\Lambda_B^{-1/2}\mathbf{P}_B'\mathbf{P}_B\Lambda_B\mathbf{P}_B'\mathbf{P}_B\Lambda_B^{-1/2}\mathbf{c}} = \frac{\mathbf{c}'\Lambda_B^{-1/2}\mathbf{P}_B'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A\mathbf{P}_B\Lambda_B^{-1/2}\mathbf{c}}{\mathbf{c}'\mathbf{c}}$$

This is a Rayleigh quotient (Goldberg, 1991) and its maximum and minimum values are given by:

$$\lambda_{R_B} \leq VD \leq \lambda_1$$

where λ_1 and λ_{R_B} are the maximum and minimum eigenvalues of $\mathbf{V}_D = \Lambda_B^{-1/2}\mathbf{P}_B'\hat{\mathbf{X}}_A'\hat{\mathbf{X}}_A\mathbf{P}_B\Lambda_B^{-1/2}$ ($R_B\times R_B$) respectively. The direction ($\mathbf{a}_w$) in space B which has the greatest difference in variance in the two spaces (worst direction) is given by:

$$\mathbf{a}_w = \mathbf{P}_B\Lambda_B^{-1/2}\mathbf{c}_w$$

where $\mathbf{c}_w$ is the eigenvector of $\mathbf{V}_D$ for which the VD statistic (eigenvalue of $\mathbf{V}_D$) has the smallest critical level (p). It is very important here to point out that by restricting $\mathbf{a}$ to

belong in the eigenspace of S_B, the VD statistic, which is a ratio, will give in general a different worst direction (a_w) than if we invert VD and restrict a to belong in the eigenspace of non-zero eigenvalues of S_A.

If we knew the distribution of the maximum and minimum eigenvalues of V_D, then we would be able to construct a Union Intersection Test. Their distribution apparently depends on the number of observations in the two samples, on the number of measurement variables, and on the degree of correlation between the measurement variables. It is complicated to derive the exact distribution of these eigenvalues since they come from the data themeselves, and probably only Monte Carlo simulations can give some insights for their distribution. The F_{n_A-1,n_B-1} distribution will be used to compare the magnitude of these eigenvalues, although the critical levels from such comparisons will be underestimated. When the VD statistic for the worst direction (a_w) has a large critical level, this is indicative of significant covariance similarity. In the case with a small critical level, one can conclude that there is a departure from covariance equivalence for at least certain directions.

Based on the above analysis, the methodology that we propose for testing covariance matrices or subspaces developed by projection methods, has two steps:

i. Check for variance equality, based on the VD statistic, in the direction of all principal components on both spaces. This will show if the principal directions on both spaces carry the same amount of variance. When two covariance matrices have significant differences in their principal directions with respect to variance, then this is a very good indication of their dissimilarity. Although in this case, as in the case of the eigenvalues of the V_D matrix, the directions under investigation come from the data themselves, the VD statistic is reasonably approximated by an F_{n_A-1,n_B-1} distribution since these directions are at least independent from the data forming the numerator (for the principal components of sample B) or the denominator (for the principal components of sample A) of the VD statistic.

ii. As an overall test, one can use the maximum and minimum eigenvalues of $\mathbf{V_D}$ to check for overall similarity, and identify the direction ($\mathbf{a_w}$) yielding the highest difference in variance.

This approach avoids the use of a single statistic to characterize the covariance similarity of two samples, as does the maximum likelihood ratio test. It has a nice geometric interpretation based on principal components, and identifies the worst direction ($\mathbf{a_w}$) of covariance dissimilarity between two samples. It is based on UIT principles since it uses the eigenvalues of $\mathbf{V_D}$ as an overall criterion, and as the other two methods examined earlier (Sections 6.1 and 6.2) it is invariant to rotation of the data. An important note is about the computational requirements in calculating the VD statistic and the $\mathbf{V_D}$ matrix. Even though the $\hat{\mathbf{X}}_A$ and $\hat{\mathbf{X}}_B$ matrices may be very large, especially when the number of measurement variables (m) is large, the calculations of VD and $\mathbf{V_D}$ are easily done if one computes them as shown below:

$$VD = (\mathbf{a}'\mathbf{P}_A)\Lambda_A(\mathbf{P}_A'\mathbf{a}) \,/\, (\mathbf{a}'\mathbf{P}_B)\Lambda_B(\mathbf{P}_B'\mathbf{a})$$

$$\mathbf{V_D} = \Lambda_B^{-1/2}(\mathbf{P}_B'\mathbf{P}_A)\Lambda_A(\mathbf{P}_A'\mathbf{P}_B)\Lambda_B^{-1/2}$$

We now investigate how the proposed approach performs in the two examples discussed in Sections 6.1 and 6.2. For the example in Section 6.1, the VD statistic values between S_A and S_B for all their principal directions, along with their critical levels ($F_{30,30}$), are found to be:

$$VD_{A1} = 1.176\ (p{=}0.660) \quad , \quad VD_{A2} = 0.625\ (p{=}0.204)$$

$$VD_{B1} = 0.850\ (p{=}0.660) \quad , \quad VD_{B2} = 1.600\ (p{=}0.204)$$

These results show that there are no significant differences between the principal directions in the two spaces. The directions of the second principal components show the least agreement. The $\mathbf{V_D}$ along with its eigenvalues and eigenvectors are given below:

$$\mathbf{V_D} = \begin{bmatrix} 0.85 & -0.6 \\ -0.6 & 1.6 \end{bmatrix} \quad \begin{matrix} \lambda_1 = 1.932\ (p = 0.074) & \mathbf{c}_{w1}' = [-0.485 \quad 0.874] \\ \lambda_2 = 0.517\ (p = 0.074) & \mathbf{c}_{w2}' = [0.874 \quad 0.485] \end{matrix}$$

The overall test is more sensitive (p=0.074) to the dissimilarity of the two sample covariance matrices than the maximum likelihood ratio test (p=0.087). The directions (since both eigenvalues have the same critical levels) which have the greatest difference in variance between the two sample covariance matrices are $\mathbf{a}_{w1}$'=[0.111 -0.123] and $\mathbf{a}_{w2}$'=[0.032 0.115]. For the example in Section 6.2, the same analysis for the sample covariance matrices S_A and S_C gave the following results:

$$VD_{A1} = 0.533\ (p=0.090)\ ,\quad VD_{A2} = 1.875\ (p=0.090)$$

$$VD_{C1} = 0.533\ (p=0.090)\ ,\quad VD_{C2} = 1.875\ (p=0.090)$$

$$\mathbf{V}_D = \begin{bmatrix} 0.533 & 0 \\ 0 & 1.875 \end{bmatrix} \quad \begin{matrix} \lambda_1 = 1.875\ (p = 0.090) & \mathbf{c}'_{1w} = [0\ \ 1] \\ \lambda_2 = 0.533\ (p = 0.090) & \mathbf{c}'_{w2} = [1\ \ 0] \end{matrix}$$

All the results indicate the same level of covariance dissimilarity (p=0.090), and the worst directions are the directions of their principal components ($\mathbf{a}_{w1}$'=[0 0.250], $\mathbf{a}_{w2}$'=[0.066 0]).

6.3.1 Comparison of Two Industrial Parallel Batches

The first example is from the DuPont US batch process (clave A) described in Section 2.3.2. Thirty six batches were selected (Section 3.1) as the reference distribution of normal batches upon which the MPCA model was built. From a parallel reactor (clave B) with identical configuration, 31 batches were characterized as normal batches from an MPCA analysis of this clave. The MPCA t-plots for both claves are shown in the upper plots of Figure 6.2. An MPCA analysis was conducted on both claves together ($\underline{\mathbf{X}}$ (67×10×100)) where the first 36 batches were from clave A, in order to examine the possibility of using a common MPCA model for both claves. The batches form two distinct ellipses in the bottom left t-plot of Figure 6.2, depending on the clave in which they were produced. Clearly, this t-plot indicates a shift in the means between the two claves since the centers of these ellipses are different. The parallel orientation of the ellipses suggests that the variance in both claves has similar principal directions, and similar variance in each of these directions since both ellipses have similar axis lengths. Upon closer examination of the mean variable trajectories between the two claves it

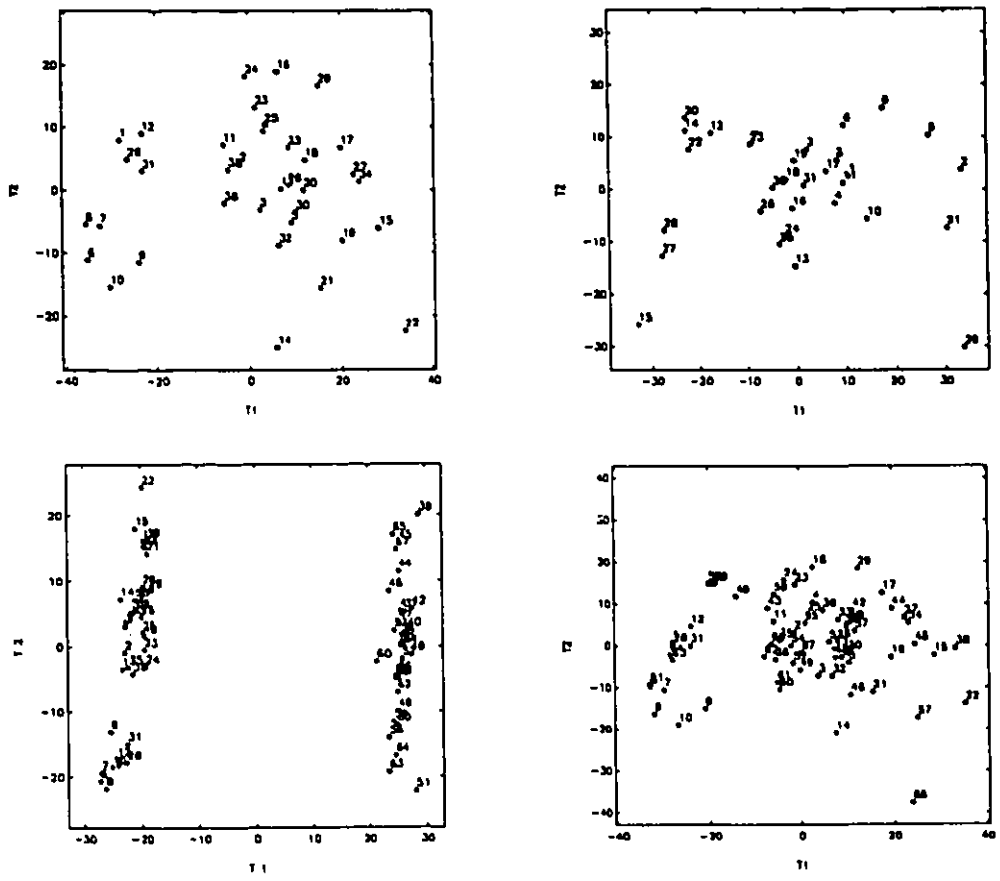

Figure 6.2 T-plots from MPCA analyses of the reference batch databases of the two industrial claves. The upper t-plots are from MPCA analyses of each clave separately (left plot from clave A). The bottom left hand side t-plot is from an MPCA analysis of both claves mean centered and scaled together. The bottom right hand side t-plot is from an MPCA analysis of both claves mean centered and scaled separately. The first 36 batches in the bottom t-plots are from clave A.

became apparent that there was a vertical shift in the trajectories of the second clave. This was attributed to slightly different sensor calibration, which explains why the claves were separated in two distinct ellipse in their t-plot. Based on these observations, a new MPCA analysis was carried out. This time the $\underline{\mathbf{X}}_A$ and $\underline{\mathbf{X}}_B$ matrices were first column mean centered and unit variance scaled separately, and then they were put together to form the $\underline{\mathbf{X}}$ matrix. Again the first 36 batches were from clave A. The t-plot from this analysis is shown in the bottom right plot of Figure 6.2. By comparing the t-plots of the individual claves and the t-plot of both claves scaled separately, one can see that the batches occupy

similar spots in these plots. This supports the hypothesis that the batches from the two claves have similar covariance structure with the only difference being a shift in their mean trajectories.

The VD values from comparing the two clave subspaces, one built upon the 36 batches of clave A ($\hat{\mathbf{X}}_A = \mathbf{T}_A\mathbf{P}_A'$) and the other built upon the 31 batches of clave B ($\hat{\mathbf{X}}_B = \mathbf{T}_B\mathbf{P}_B'$) where $\mathbf{X}_A$ and $\mathbf{X}_B$ were scaled separately and 3 principal components were extracted from both, are given below along with their critical levels ($F_{35,30}$).

$VD_{A1} = 2.117$ (p=0.039) , $VD_{A2} = 1.456$ (p=0.297) , $VD_{A3} = 5.088$ (p=0.000)

$VD_{B1} = 0.841$ (p=0.618) , $VD_{B2} = 0.756$ (p=0.424) , $VD_{B3} = 0.303$ (p=0.000)

$$\mathbf{V}_D = \begin{bmatrix} 0.841 & 0.320 & 0.339 \\ 0.320 & 0.756 & 0.432 \\ 0.339 & 0.432 & 0.303 \end{bmatrix} \quad \begin{array}{ll} \lambda_1 = 1.386 & (p = 0.365) \quad \mathbf{c}_{w1}' = [0.637 \quad 0.629 \quad 0.445] \\ \lambda_2 = 0.494 & (p = 0.046) \quad \mathbf{c}_{w2}' = [0.748 - 0.642 - 0.164] \\ \lambda_3 = 0.020 & (p = 0.000) \quad \mathbf{c}_{w3}' = [-0.182 - 0.438 \; 0.880] \end{array}$$

The VD statistics indicate that the third principal component (VD_{A3}, VD_{B3}, λ_3) is different in the two claves. The same conclusion is apparent from the worst direction ($\mathbf{c}_{w3}$) which is mainly focused on the direction of the third principal component of clave B. The first two principal components of the two claves seem compatible, and thus a new analysis was done based on only the first two principal components. The results are shown below:

$VD_{A1} = 2.154$ (p=0.035) , $VD_{A2} = 1.534$ (p=0.235)

$VD_{B1} = 0.833$ (p=0.599) , $VD_{B2} = 0.755$ (p=0.422)

$$\mathbf{V}_D = \begin{bmatrix} 0.833 & 0.324 \\ 0.324 & 0.755 \end{bmatrix} \quad \begin{array}{ll} \lambda_1 = 1.120 & (p = 0.756) \quad \mathbf{c}_{w1}' = [0.748 \quad 0.663] \\ \lambda_2 = 0.467 & (p = 0.031) \quad \mathbf{c}_{w2}' = [-0.663 \quad 0.748] \end{array}$$

Only the first component of clave A (VD_{A1}) seems a little different. The B space is well projected in the A space since all the VD_B values and the eigenvalues of $\mathbf{V}_D$ have reasonable critical levels. Here we have to point out that the largest and smallest eigenvalues of $\mathbf{V}_D$ are the maximum and minimum values of the VD statistic for any vector ($\mathbf{a}$) that belongs to space B. It is clear from the analyses of this example that the VD statistic can take larger or smaller values as we test the principal components of space A which do not belong necessarily in space B. From the above results, it is reasonable to

conclude that the first two principal components from the two claves define similar subspaces. A common MPCA model is reasonable to be used for monitoring both claves, but the batches have to be scaled differently based on which clave they were produced.

The geometric test proposed by Krzanowski (Section 6.2) gave the following angles (shown in parenthesis) between the 3 principal components of the two claves.

$$\mathbf{P}_A'\mathbf{P}_B = \begin{bmatrix} -0.770\ (39.5^\circ) & -0.273\ (74.1^\circ) & -0.199\ (78.5^\circ) \\ -0.300\ (72.5^\circ) & 0.678\ (47.3^\circ) & 0.228\ (76.7^\circ) \\ 0.187\ (79.1^\circ) & -0.054\ (86.9^\circ) & -0.176\ (79.8^\circ) \end{bmatrix}$$

The increasing angles between the principal components of the two samples (39.5°, 47.3°, 79 8°) show another deficiency of this test. If the angle between the first principal components of the two samples is large, then the angle between each of the subsequent components will be even larger. This is because principal components are orthogonal to each other due to their sequential nature. Therefore in this example, the angles between the second (47.3°) and the third (79.8°) principal components lack of interpretation. The other problem with the angles between principal components is that they ignore any information about the variance that each component carries, and thus, they do not check how well each subspace projects to the other. The minimum angle in this example was found to be 31.5° and the trace of $\mathbf{P}_A'\mathbf{P}_B\mathbf{P}_B'\mathbf{P}_A$ to be equal to 1.379 which is well below the number of principal components (3) in this case. By using only the first two principal components in each clave, the minimum angle was found to be 34.2° and the trace of $\mathbf{P}_A'\mathbf{P}_B\mathbf{P}_B'\mathbf{P}_A$ was 1.218. The minimum angle in this case increased by removing one principal component from each subspace, because the remaining subspaces coincide less. The absence of a measure of significance in this test, makes the results in this example inconclusive.

6.3.2 Comparison of an MPCA and an MPLS Model

The second case which we will investigate is from the SBR example described in Section 2.3.1. An MPCA model for the SBR reactor was built in Section 3.3, and in

Section 5.1.2 an MPLS model was built based on the same data. The t-plots from these analyses are shown in Figure 6.3, and they are similar especially with respect to the t_1-axis. An interesting feature of this example is that the MPCA model has three components, whereas the MPLS model has two. A comparison of these two models will be done in order to check whether the low dimensional space of MPLS is contained in the higher dimensional space of MPCA. In this example, the test based on the VD statistic is approximate since both $\hat{\mathbf{X}}_A$ and $\hat{\mathbf{X}}_B$ matrices are coming from the same data, and thus they are not independent. In applying the proposed procedure in comparing the two x-subspaces, the MPCA model has to go in the denominator of the VD statistic, since we know its principal directions and it has higher (3) dimensionality than the MPLS model (2). Remember that the MPLS p-loading matrix ($\mathbf{P}_A$) does not have the principal directions of $\mathbf{X}_A$. If we used the MPLS model in the denominator of the VD statistic, the overall test would not give the worst case since one has to know the principal directions of the covariance matrix in the denominator to formulate the V_D matrix. The directions in the MPLS p-loading matrix ($\mathbf{P}_A$) are major directions of variability and will be used to evaluate the VD_A statistics to show us if these directions carry similar variability in both spaces. One could have found the principal directions of the MPLS model, but they are not needed as long as one has the principal directions for one of the two models.

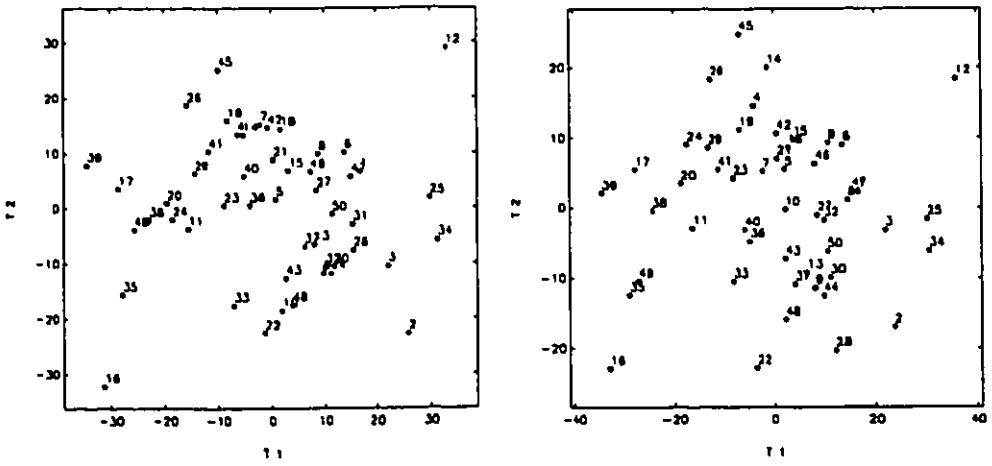

Figure 6.3 T-plots of the reference SBR database. The left hand side t-plot is from the MPCA analysis, and the right hand side t-plot is from the MPLS analysis.

The VD values from this comparison along with their critical levels ($F_{49,49}$) are given below:

$$VD_{A1} = 1.004\ (p{=}0.989)\ ,\ VD_{A2} = 1.189\ (p{=}0.547)$$

$$VD_{B1} = 0.968\ (p{=}0.910)\ ,\ VD_{B2} = 0.454\ (p{=}0.006)\ ,\ VD_{B3} = 0.091\ (p{=}0.000)$$

$$\mathbf{V}_D = \begin{bmatrix} 0.968 & 0.086 & 0.093 \\ 0.086 & 0.453 & 0.200 \\ 0.093 & 0.200 & 0.091 \end{bmatrix} \quad \begin{array}{ll} \lambda_1 = 1.000\ (p = 1.000) & \mathbf{c}'_{w1} = [-0.968 - 0.206 - 0.145] \\ \lambda_2 = 0.513\ (p = 0.021) & \mathbf{c}'_{w2} = [-0.246\ \ 0.896\ \ 0.370] \\ \lambda_3 = 0.000\ (p = 0.000) & \mathbf{c}'_{w3} = [-0.053 - 0.394\ \ 0.917] \end{array}$$

The VD_A values indicate clearly that the MPLS space is contained in the space defined by the MPCA model. The VD_A values along with the VD_{B1} and VD_{B2} values show a good similarity between the two first components of the two models. As it was expected, the third MPCA component, which has the largest weight in $\mathbf{c}_{w3}$, define a direction in the x-space which is almost perpendicular ($p \cong 0$) to the space that the MPLS model define. In this example one of the eigenvalues of V_D (λ_1) is approximately equal to 1. This means that the direction that the corresponding eigenvector ($\mathbf{c}_{w1}$) defines, belongs to the intersection of the two sample subspaces.

An analysis was done based on only the first two components of the MPCA model. The results are shown below:

$$VD_{A1} = 1.006\ (p{=}0.983)\ ,\ VD_{A2} = 1.278\ (p{=}0.393)$$

$$VD_{B1} = 0.968\ (p{=}0.909)\ ,\ VD_{B2} = 0.454\ (p{=}0.006)$$

$$\mathbf{V}_D = \begin{bmatrix} 0.968 & 0.086 \\ 0.086 & 0.454 \end{bmatrix} \quad \begin{array}{ll} \lambda_1 = 0.982\ (p = 0.949) & \mathbf{c}'_{w1} = [0.987\ \ 0.161] \\ \lambda_2 = 0.440\ (p = 0.005) & \mathbf{c}'_{w2} = [-0.161\ \ 0.987] \end{array}$$

The MPLS space is contained in the MPCA space (VD_{A1}, VD_{A2}), and the first MPLS component is very similar to the first MPCA component (VD_{A1} and VD_{B1} close to one). The worst direction of similarity is mainly concentrated on the direction of the second principal component of the MPCA model ($\mathbf{c}_{w2}$). This was expected, if one considers how MPLS constructs its space. The first principal component from an MPCA model when it carries a large amount of variability with respect to the other principal components, as in this case (Table 3.1), is very close to the first MPLS component. The second MPLS

component has been tilted with respect to the second MPCA component as to be well correlated with the y-variables.

In this example, if one wants to apply the geometric test proposed by Krzanowski (Section 6.2), one has first to find the principal components of the x-space defined by the MPLS model. The maximum likelihood ratio test is inapplicable in this example, as was in the two clave example (Section 6.3.1), because the sample covariance matrices are singular.

6.4 Comparison of Projection Models Based on a Residuals Analysis

Projection methods such as PCA and PLS decompose an observation matrix (**X**) into two parts (**X=TP'+E**): one deterministic (**TP'**), and one random (**E**). A new observation based on a projection model can also be decomposed in a deterministic part and an error. Here we explore how these errors can be used to compare two projection models. Since the projection models will be used for new observations, our statistics will be based on the prediction error.

Assume that we have two projection models A and B, built upon some observations (A and B). One can find for each model the Press statistic ($Press_A$, $Press_B$) described in Section 3.2, and the sum of squared errors from predicting the observations from model A by using model B ($SSE_{A/B}$) and vice versa ($SSE_{B/A}$). The ratio of these statistics will be approximately distributed as an F distribution under the hypothesis that the true underlying model spaces are identical, and can be used as a test of the similarity of the two models.

$$(SSE_{A/B}/df_1) / (Press_B/df_2) \sim F_{df1,df2}$$

$$(SSE_{B/A}/df_3) / (Press_A/df_4) \sim F_{df3,df4}$$

And as an overall test, one can use the combination of both:

$$((SSE_{A/B}+SSE_{B/A})/df_5) / ((Press_B+Press_A)/df_6) \sim F_{df5,df6}$$

These tests were suggested by Dr. Svante Wold (personal communication). The problem with the above statistics is that the degrees of freedom (df) are unknown. One

neither knows how many degrees of freedom to begin with, nor how many to deduct with each component. Estimating the initial degrees of freedom is the greatest problem. Box et al. (1973) exposed the problem and suggested ways to overcome it.

By assuming that the degrees of freedom in the above statistics are equal for the MPCA models based on clave A and B of the industrial example, the results given below were found for the three component models. The batches in clave A were scaled separately from the batches in clave B.

$$SSE_{A/B} / Press_B = 1.272 \quad , \quad SSE_{B/A} / Press_A = 1.111$$

$$(SSE_{A/B}+SSE_{B/A}) / (Press_B+Press_A) = 1.190$$

Since the statistics have values close to 1, a good similarity between the two claves is suggested. This test does not give any indication of what might be their differences and which components are more similar than others. The same statistics for the two component models were found to be:

$$SSE_{A/B} / Press_B = 1.173 \quad , \quad SSE_{B/A} / Press_A = 1.067$$

$$(SSE_{A/B}+SSE_{B/A}) / (Press_B+Press_A) = 1.121$$

A better similarity between the two components model is suggested since the statistics now have values closer to 1.

No comparison can be done based on residuals, between the SBR MPCA and MPLS model. The MPCA model by simply having one more component will have much smaller residuals than the MPLS model. Thus, the SSE and Press statistics can not be compared in this case without knowing their degrees of freedom. Even if one restricts the MPCA model to the first two components, still the residuals are not comparable if one does not know their degrees of freedom. An MPCA model always gives smaller residuals than an MPLS model for the same observations and number of components, because the MPCA model gives the global minimum in the sum of squared residuals among all linear decompositions. Still, if the degrees of freedom were known in this example, the proposed test would be approximate since both models have been built based on the same data.

CHAPTER 7 Summary and Conclusions

Recent trends in North American and in most industrialized countries have been towards the manufacture of higher value added specialty chemicals, that are produced mainly in batch reactors. Examples include specialty polymers, pharmaceuticals, and biochemicals. There are also many other batch type operations, such as crystallization and injection molding, which are very important to the chemical and manufacturing industries. A new multivariate statistical approach to monitoring the progress of batch and semi-batch processes has been developed. Rather than utilizing detailed engineering knowledge about the process, as in model-based and knowledge-based approaches, this approach utilizes only the information contained in the historical database of past batches. Such information is readily available for any computer monitored industrial batch process. Projection methods such as Multi-way Principal Component Analysis (MPCA) and Multi-way Partial Least Squares (MPLS) are used to extract the information in batch data, and to construct simple monitoring charts, consistent with the philosophy of Statistical Process Control (SPC), which are capable of tracking the progress of new batch runs and detecting the occurrence of observable upsets. A simulation of a styrene-butadiene batch reactor along with a couple of industrial examples have been used to develop the ideas and to test the abilities of the proposed approach.

MPCA is used to extract the information directly from the trajectories of all the measured process variables ($\underline{\mathbf{X}}$), and to project it onto a low dimensional space defined by the principal components. The data reduction is tremendous, since all the information from a database of batches is captured in a few matrices which define the reduced space. A post analysis of past batches enables one to classify normal and abnormal batches by examining their position onto this reduced space. New batches can be monitored in real-time, using a sound statistical framework, by tracking their progress in this reduced space.

The approach is based on the basic concepts of SPC, whereby the future behavior of a process is monitored by comparing it against that observed in the past when the

process was performing well, that is in a state of statistical control. Control limits for the monitoring charts are derived from the statistical properties of this historical reference distribution of past normal batches. The proposed monitoring charts are in accordance with the SPC requirements for charts that can be easily displayed, interpreted, and can quickly detect a fault. The power of the proposed statistical approach lies in the fact, that it utilizes the unsteady state trajectory data on all measured variables in a truly multivariate manner, as to account not only for the magnitude and trend of the deviations in each measured variable from its average trajectory, but also for the high degree of correlation in both time and among the deviations in all the variables. Once an abnormality has been detected, contribution plots can be immediately displayed and the measurement variables responsible for the alarm to be identified. This will help one to hypothesize assignable causes for the fault detected.

As in all inferential approaches, the fundamental assumptions of "comparable" runs and "observable" events of interest must hold for the method to work. The first assumption states that the method is valid as long as the reference database is representative of the process operation. If something changes in the process (eg. new catalyst), then one has to build a new database which embodies the change and reapply the method. The second assumption expresses the requirement that the events which one wishes to detect must be "observable" from the measurements that are being collected.

MPCA only makes use of the process variable trajectory measurements ($\underline{\mathbf{X}}$) taken throughout the duration of the batch. Measurements on product quality variables ($\mathbf{Y}$) taken at the end of each batch can be utilized in a direct fashion in the proposed monitoring procedure with the aid of MPLS. MPLS models both the $\underline{\mathbf{X}}$ and $\mathbf{Y}$ space and focuses more on the variance of $\underline{\mathbf{X}}$ that is most predictive of the product quality $\mathbf{Y}$. The multivariate SPC monitoring schemes developed for MPCA are extended directly to MPLS. The additional information that one gets from MPLS is on-line predictions of the final product quality. Approximate confidence intervals, which can be applied to any PLS study, were developed for the MPLS predictions.

When additional information about the initial conditions and set-up of the batch process is available, Multi-block methods based on MPCA and MPLS can incorporate this information into the proposed monitoring schemes. Such prior information can be organized in a new matrix (**Z**) which may have variables like feed quality measurements, initial amounts of initiator or emulsifier, preprocessing conditions such as preheat duration, position of the batch in the cleaning cycle, operator on duty, etc. Changes in these variables may often be as important to the product quality as variations in the process trajectories. In addition, Multi-block methods allow one to develop overall monitoring schemes for processes consisting of several units or processing stages, as in the industrial example investigated in this thesis (Section 5.2.2). Each unit or processing stage can be organized in a single block, and then monitoring charts can be established for each block as well as for the entire process.

The proposed monitoring schemes are generic as to be applied in any batch or semi-batch process, and they do not require any problematic computations. The monitoring charts along with their contribution plots are as easy to use and interprete as conventional Shewhart charts. Through simulated and industrial examples, it has beeen shown that the proposed methods can clearly and quickly detect an abnormal operation. The objective of the monitoring procedure, as any SPC method, is to detect faults, diagnose them, and eliminate their cause and thereby shrink the control limits and work towards a more consistent quality product.

There are two major benefits from the proposed methodology. First, it can be used as a tool for investigating huge databases of batch data, which otherwise remain stored and unexploited. Analysis of such databases can provide valuable knowledge about the batch process as we have seen in industrial cases. Out of specification products can be traced back to their process operation and in most cases identify their cause. Many times the same fault is responsible for most of the batches with product out of specifications. By identifying such a problem, one can redesign the reactor or change its control policy as to eliminate the cause from future appearances. The second major benefit from the

proposed methodology comes from the on-line monitoring of a batch process. Usually, except in extreme cases, there is no indication that something was wrong in a batch run with an incipient fault. Only after several hours or days a product is characterized as out of specifications when the quality results become available from the laboratory analysis. In most cases by that time, the product has been stored and mixed with other product, or further processed in subsequent units, or even shipped to the customer. In such cases, the cost for the company can be enormous. An on-line monitoring scheme based on the proposed methods will immediately detect batches with operational problems and questionable product quality. The product from all such batches will not be further processed or mixed with other product, until their quality measurements become available from the laboratory analysis. Such a strategy should significantly improve productivity and product quality.

The step-wise procedure of MPLS and Multi-block MPLS to explain the variability in the quality space $\mathbf{Y}$ using the variability of the $\underline{\mathbf{X}}$ space, forces these methods to model both spaces. This characteristic is rather beneficial since these projection models will try to describe the functional relationship of the process operation to the desired production properties. If one includes in the reference database batches of several product grades, then MPLS and Multi-block MPLS may be able to provide a good initial guess for the batch recipe and the operational trajectories that will yield a new grade with pre-specified quality measurements. This is a rather desired knowledge since customers sometimes know what property changes will make the product more suitable for their needs. Moteki and Arai (1986) were probably the first who introduced ideas like this, but their approach was not well documented and too specialized for their given process.

Many articles have been recently written on the topic of combining SPC and Automatic Process Control (APC) (Box and Kramer, 1992; Vander Wiel et al., 1992; Tucker et al., 1993). It is important to note that the multivariate SPC charts proposed in this thesis are meant to be applied to data collected while the underlying batch process is

being controlled by the feedforward and feedback APC schemes. If a significant disturbance occurs, the controllers may be able to compensate for it. Although its final product may be acceptable, the proposed methods will detect this batch as abnormal because of the unusual behavior caused by the excessive control actions.

When discussing these multivariate SPC monitoring schemes a very appealing feedback control idea has sometimes been proposed; namely, that whenever a fault is detected, the MPCA or MPLS model be inverted to solve for values of the adjustable variables that could be reset to bring the batch back into the in-control region of the reduced space. However, this suggestion is not reasonable because the MPCA and MPLS models are not cause and effect models, but rather only models of the correlation structure of the process variables under routine operating conditions. They cannot be used to predict the effects that independent changes in some of the measurement variables will have on the quality of the final product. By taking a corrective action based on the monitoring scheme, we are altering the very nature of the new batch data with respect to those in the reference database where no such control actions were present, thereby invalidating the model. Furthermore, once the on-line control charts have indicated the occurrence of a special event, by simply forcing the observations back into the control region will not imply that an acceptable product will result as illustrated in the industrial example with batch 49 in Section 4.4.

Of course, this does not mean that nothing should be done when a fault is detected. The nature of any corrective action will depend upon the underlying cause, and upon the time during the batch at which the fault occurred. Therefore, the best form of control action in the proposed SPC schemes is probably to use the on-line diagnostic tools to interrogate the underlying projection model for possible reasons for the fault, and respond accordingly using one's process knowledge, a cause and effect model, or an expert system. Even if the current batch cannot be saved, SPC philosophy dictates that an assignable cause be found and corrected so as not to affect any future batches.

APPENDICES

Appendix A Styrene-Butadiene Simulation

The program by Kozub (1989) simulates the semi-batch emulsion production of hot SBR latex. The initiator is potassium persulfate, the chain transfer agent is n-dodecyl mercaptan, the emulsifier is sodium lauryl sulphate, the organic impurity is 4-ter-butyl catechol, and the water impurity is dissolved oxygen.

The program simulates a reactor with volume 10 lt. The volume of the cooling jacket is 2.52 lt, the heat transfer area is 0.172 m^2, and the heat transfer coefficient is 30 Watt/m^2K. The batch duration is 1000 min and the sequence of the copolymerization used in this study as the base case is the following. The reactor is initially charged with seed SBR particles (2.0e+17 number of particles/lt of 0.04 mol styrene and 0.226 mol butadiene), a small amount of monomers (0.1 mol styrene and 0.1 mol butadiene), and all the initiator (0.03 mol), chain transfer agent (0.0275 mol), emulsifier (0.01 mol), and water (4.8 lt). The initial water and organic impurities in the reactor are 0 mol and 1.0e-5 mol respectively. Styrene and butadiene monomers are then fed to the reactor at a constant rate (0.00871 mol/min each) for the rest of the batch operation. Both feedrates have been simulated as first order autoregressive models (ϕ=0.8, σ_a^2=0.81e-8 mol/min) and their temperature is 50°C. Measurement noise (σ_a^2=0.1e-3°C) is added in the feedrate temperature after completion of the simulation. The temperature of the reactor is maintained at 50°C through out the operation with a PI controller (K=5, τ=20 min) which manipulates the temperature of the cooling water which has a flowrate of 0.3 lt/min. Water (0 mol/min) and organic impurities (3.0e-7 mol/min) are coming into the reactor along the styrene and butadiene feedrates respectively and are proportional to them.

With this recipe as base case, fifty batches were simulated to create the database of normal batches. For each batch, some conditions from the base case were perturbed. A list of these perturbations are given in the following table.

Conditions	Base Case	Range of Perturbations
Initial Styrene (mol)	0.1	±5%
Initial Butadiene (mol)	0.1	±5%
Initial Reacted Styrene (mol)	0.04	±10%
Initial Reacted Butadiene (mol)	0.226	±10%
Number of particles/lt	2.0e+17	±2%
Chain Transfer Agent (mol)	0.0275	±10%
Initial Water Impurities (mol)	0	0.2e-5 - 1.1e-5
Initial Organic Impurities (mol)	1.0e-5	±10%
Incoming Water Impurities (mol/min)	0	0.5e-7 - 1.5e-7
Incoming Organic Impurities (mol/min)	3.0e-7	±10%
Heat Transfer Coefficient (Watt/m^2K)	30	±3%

Table A1. Range of perturbations from the base case for the creation of the reference normal database.

Appendix B Other MPCA Approaches for Monitoring Batch Processes

A brief description of three other MPCA approaches for monitoring batch processes are given here along with their advantages and disadvantages. The major motivation in these approaches is to avoid in the monitoring scheme to have filling in the unknown part of a new batch ($\mathbf{X}_{NEW}$) at each time interval. None of them overcome the timing problem (see Chapter 2.4) of regular MPCA. In all three approaches, one has to subtract the mean and divide by the standard deviation which corresponds at each time interval, and apply the p-loadings at the correct time intervals.

At the end of this appendix, some suggestions are given of how to handle in MPCA or MPLS sudden changes in the measurement variables during the batch operation.

First Approach

The first approach unfolds and scales $\underline{\mathbf{X}}(I\times J\times K)$ to $\mathbf{X}(I\times JK)$ as in regular MPCA. Then it rearranges $\mathbf{X}$ into $\mathbf{X}_F(IK\times J)$ where its first I rows are the first J columns of $\mathbf{X}$, the I+1 through 2I rows of $\mathbf{X}_F$ are the J+1 through 2J columns of $\mathbf{X}$, etc. A normal PCA is performed in $\mathbf{X}_F$ which models now the overall variation of the measurements around their mean trajectories throughout the batch operation. The advantage of this approach is that

each principal component has a p-loading vector common to all time intervals. The major disadvantage is that there is no data reduction. The matrix $\mathbf{X}_F$, in the way that it was formed, has J independent columns and one needs J principal components to describe it adequately. It leads to almost the same result as plotting Shewhart charts for each measurement variable available. The columns of $\mathbf{X}_F$ are uncorrelated because the correlation structure of the measurement variables changes throughout the batch operation. Even a single change in the correlation structure of one measurement variable is possible to make the columns of $\mathbf{X}_F$ to be uncorrelated.

SECOND APPROACH

As it is shown in Figure B1, it is not unlike for one measurement variable to maintain the same correlation with the other variables for certain periods of time. In such time periods, the p-loadings will remain almost constant. The approach presented here explores this observation and tries to stabilize the p-loadings during such periods gaining a data reduction in the p-loading vectors. One has to unfold and scale the $\underline{\mathbf{X}}$ into $\mathbf{X}$ as in regular MPCA (Figure B1), and use the following algorithm proposed by Dr. Svante Wold (personal communication).

i. Let $\mathbf{X}_{ALL}=\mathbf{X}_1$ (the first J columns of $\mathbf{X}$)

ii. Do a PCA on $\mathbf{X}_{ALL}$ and get two (for example) principal components ($\mathbf{p}_1$, $\mathbf{p}_2$)

iii. Augment $\mathbf{X}_{ALL}=[\mathbf{X}_1\ \mathbf{X}_2]$, $\mathbf{p}_1'=[\mathbf{p}_1'\ \mathbf{p}_1']$, $\mathbf{p}_2'=[\mathbf{p}_2'\ \mathbf{p}_2']$

iv. $\mathbf{p}_1=\mathbf{p}_1/|\mathbf{p}_1|$, $\mathbf{p}_2=\mathbf{p}_2/|\mathbf{p}_2|$

v. Check if $\mathbf{p}_1$ and $\mathbf{p}_2$ describe adequately $\mathbf{X}_{ALL}$.

If they do, augment $\mathbf{X}_{ALL}$ and the p-loading vectors as in step iii (by repeating the p-loadings corresponding to the last time interval) and continue for the next time interval. In this case the correlation structure has not change and one can use the same p-loadings for the next time interval.

If the p-loading vectors do not describe adequately $\mathbf{X}_{ALL}$, get extra principal components based on the residuals of $\mathbf{X}_{ALL}$, until $\mathbf{X}_{ALL}$ is described adequately.

Augment $\mathbf{X}_{ALL}$ and all the p-loading vectors as in step iii and continue for the next time interval.

The main problem in this algorithm is how to choose when you have explained adequately $\mathbf{X}_{ALL}$. A quick way is to use the broken stick rule discussed in Section 3.2. Although this method seems appealing because of its data reduction in the p-loadings, there are a couple of problems with it. First of all, the t-scores are not orthogonal any more, and thus different latent variables may describe similar things in the data. The other major problem

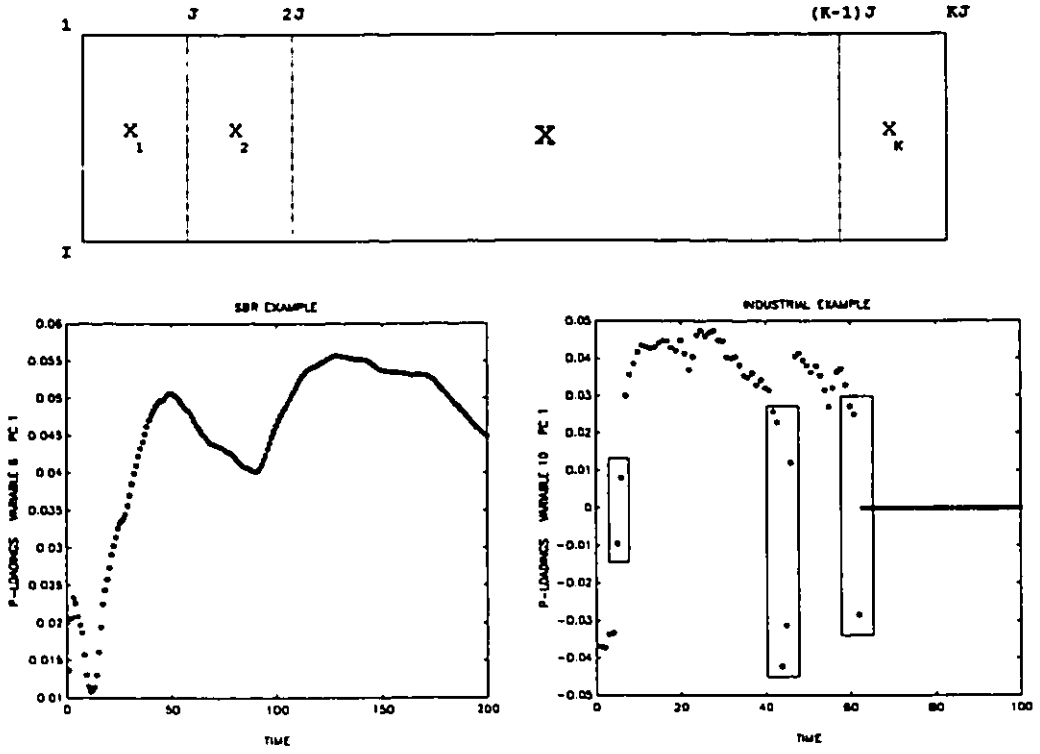

Figure B1 Unfolded X matrix and plots of p-loadings versus time. The left hand side plot shows the p-loadings of the first principal component for measurement variable 8 in the SBR example. The right hand side plot shows the p-loadings of the first principal component for measurement variable 10 in the industrial example. The rectangular areas in the later plot indicate periods of sudden changes in variable 10 where the p-loadings might have been set equal to zero.

comes when the correlation structure among the measurement variables changes frequently along the batch operation, or when there are sudden changes in the measurement variables. In such cases, the method will end up with too many principal components where probably some of them will try to cancel the effect of some others.

Several ideas to overcome this problem were considered (such as to identify from the beginning the time periods of constant correlation structure based on a regular MPCA, and build this knowledge in the algorithm through a multi-block approach), but none can resolve the loss of orthogonality in the t-scores.

THIRD APPROACH

Again, in this approach we try to explore the fact that the p-loadings of a measurement variable remain constant for certain periods of time as shown in Figure B1. The algorithm given below was developed in cooperation with Dr. Svante Wold, and it is an evolving formulation of PCA which progressively augment its p-loadings at each time interval.

i. unfold and scale $\underline{\mathbf{X}}$ into $\mathbf{X}$ as in regular MPCA (Figure B1)

ii. Pre-specify how many principal components you want in your model.
Do a PCA on $\mathbf{X}_1$ (first J columns of $\mathbf{X}$) and extract the principal components (eg. $\mathbf{p}_1$, $\mathbf{p}_2$).

iii. Set $\mathbf{X}_{OLD}$ and $\mathbf{p}_{OLD}$ to the $\mathbf{X}$ and $\mathbf{p}$ that you have already investigated.
$\mathbf{X}_N=\mathbf{X}_2$, $\mathbf{X}_{ALL}=[\mathbf{X}_{OLD}\ \mathbf{X}_N]$, $\mathbf{p}_N=\mathbf{p}_1$ ($\mathbf{p}_N$ is set equal to the p-loadings corresponding to the last time interval)

iv. $\mathbf{p}'=[\mathbf{p}_{OLD}'\ \mathbf{p}_N']$

v. $\mathbf{t}=\mathbf{X}_{ALL}\mathbf{p}$

vi. $\mathbf{t}=\mathbf{t}/|\mathbf{t}|$

vii. $\mathbf{p}_N=\mathbf{X}_N'\mathbf{t}$

viii. $\mathbf{w}=\mathbf{X}_{OLD}'\mathbf{t}$

ix. $f=|\mathbf{w}|/|\mathbf{p}_{OLD}|$

x. $\mathbf{p}'=[f\mathbf{p}_{OLD}'\ \mathbf{p}_N']$

xi. Check convergence on $\mathbf{p}$. If it has not converged go to step v.
If it has converged: $\mathbf{p}=\mathbf{p}/|\mathbf{p}|$, $\mathbf{t}=\mathbf{X}_{ALL}\mathbf{p}$, $\mathbf{X}_{ALL}=\mathbf{X}_{ALL}-\mathbf{t}\mathbf{p}'$, return to step iv for the next principal component.

It is an adaptive algorithm which augments the p-loading vectors to incorporate new variables in a PCA model, which works well in batch data when the correlation structure of the measurement variables does not change frequently. The t-scores are not orthogonal again in this algorithm. Several deflations were tried in step xi, but none was able to fix the orthogonality problem. Another problem with this algorithm is that one has to pre-specify the number of principal components that one wants to extract.

SUDDEN CHANGES IN THE MEASUREMENT VARIABLES

This is a suggestion of how to handle sudden changes that one may have in one or more measurement variables, like in variable 10 of the industrial example (Figure 2.6). When there is a small number of batches in the reference database, there is a lot of uncertainty of what is the time period that such sudden changes normally occur. In cases where there are many batches in the reference database, MPCA will resolve the problem by giving small p-loadings at the appropriate time intervals. Thus in monitoring a new batch, if a change occurs during the time period with the small p-loadings, this will have a small impact in the t-scores, indicating a normal operation. When there is a small number of batches in the reference database, one may not have covered the full time period that such changes may occur. In this case, it will be better to proceed as we suggest below to avoid false alarms in the monitoring scheme.

Do a regular MPCA in your data, and discuss with the engineers what is the acceptable time range that such changes may occur. Set the p-loadings which correspond to these time periods equal to zero, and use these modified p-loadings in setting up the control charts and for monitoring new batches. In this way, any changes in these variables during these particular periods will be ignored as normal variation. As an example, the p-loadings of variable 10 in the industrial example that could have been set equal to zero, are those inside the rectangular frames in Figure B1.

Appendix C Control Limits for Quadratic Forms

Let **x** be an observation vector from a Multinormal population $N(0,\Sigma)$, and λ_i the eigenvalues of Σ, then approximate upper control limits for the quadratic form $Q=x'x$ with significance level α are given by:

(Box, 1954) $\quad g\chi^2_{h,\alpha}$

(Jackson and Mudholkar, 1979) $\quad \theta_1\left[1-\theta_2 h_0(1-h_0)/\theta_1^2 + z_\alpha(2\theta_2 h_0^2)^{1/2}/\theta_1\right]^{1/h_0}$

where $\chi^2_{h,\alpha}$ is the critical value of the Chi-squared variable with h degrees of freedom at significance level α, and z_α is the critical value of the Normal variable at significance level α which has the same sign as h_0. The rest of the parameters are given below:

$$\theta_1 = \sum\lambda_i \quad , \quad \theta_2 = \sum\lambda_i^2 \quad , \quad \theta_3 = \sum\lambda_i^3$$

$$g = \theta_2/\theta_1 \quad , \quad h = \theta_1^2/\theta_2 \quad , \quad h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$$

The relationship between them becomes clear when one uses the Wilson-Hilferty approximation for the Chi-squared variable (Evans et al., 1993) and rewrites Box's equation as follows:

$$gh\left[1 - 2/9h + z_\alpha(2/9h)^{1/2}\right]^3$$

Every term in this equation approximates well the corresponding term in Jackson and Mudholkar's equation when one has extracted most of the significant principal components (λ_i with large values) and thus $\theta_2^2 \approx \theta_1\theta_3$.

Notation

normal symbols		= scalars
bold small symbols		= column vectors
bold capital symbols		= matrices
underline bold capital symbols		= three-way arrays
'	eg. $\mathbf{T}'$	= transpose
×	eg. $\mathbf{T}$ (I×R)	= matrix dimensions
,	eg. $\mathbf{T}(i,r)$	= matrix element
:	eg. $\mathbf{T}(c{:}d,f{:}g)$	= matrix segment of elements
^	eg. $\hat{t}$	= predicted or estimated value
||	eg. $\|t\|$	= vector norm
det(·)		= determinant
E(·)		= expected value
ln(·)		= natural logarithm
min(·)		= minimum value
var(·)		= variance

a, A	= number of blocks
$\mathbf{a}$	= fixed vector
$\mathbf{a}_w$	= worst direction of covariance similarity
b	= sample SPE mean
B	= Beta variable
$\mathbf{b}$, $\mathbf{B}$	= estimated regression coeficients
c	= index variable
$\mathbf{c}$	= random vector
cf	= contribution factor
df	= degrees of freedom

D	= Hotelling statistic for the t-scores of a new batch
D_S	= Hotelling statistic for the t-scores in a database
D_m	= cross-validation degrees of freedom
D_r	= cross-validation degrees of freedom
$\mathbf{e}$, $\mathbf{E}$	= x-residuals
F	= F distribution
$\mathbf{f}$, $\mathbf{F}$	= y-residuals
g	= weight for the SPE distribution
$\mathbf{g}$	= random vector
G	= statistic for broken stick rule
$\mathbf{G}$	= generalized inverse of $\hat{\mathbf{X}}$
h	= degrees of freedom for the SPE distribution
h_0	= distribution parameter of quadratic form
H	= statistical hypothesis
$\mathbf{H}$	= indempotent matrix
i, I	= number of batches
j, J	= number of measurement variables
k, K	= number of time intervals
l	= maximum likelihood ratio test
$\mathbf{L}$	= generalized inverse of $\hat{\mathbf{X}}'\hat{\mathbf{X}}$
m, M	= number of variables
MSE	= mean squared error
MSR	= mean squares due to regression
n	= number of observations
N	= Normal distribution
p	= critical level
$\mathbf{p}$, $\mathbf{P}$	= p-loadings
$\mathbf{P}$	= folded p-loading matrix

Press	= cross-validation prediction sum of squares
Q	= sum of squared residuals
q, Q	= q-loadings
r, R	= number of principal components
R	= cross-validation statistic
RSS	= cross-validation residual sum of squares
s_{ref}	= estimated t-score standard deviation
sf	= scaling factor
SPE	= on-line sum of squared prediction errors
SSE	= sum of squared errors
SSR	= sum of squares due to regression
S	= sample covariance matrix
t	= Studentized variable
t, T	= t-scores
u	= u-scores
VD	= statistic for covariance equivalence
v	= sample SPE variance
V	= matrix related to S_E
V_D	= matrix related to VD
w	= parameter defining the smoothing window width
W	= Wishart distribution
W	= cross validation statistic
w, W	= w-loadings
$\underline{\mathbf{X}}$	= database of batch process measurements
X	= unfolded and scaled $\underline{\mathbf{X}}$
$\mathbf{X}_{NEW}$	= process measurements of a new batch
$\mathbf{x}_{NEW}$	= unfolded and scaled $\mathbf{X}_{NEW}$
y, Y	= quality measurements

$\mathbf{Z}$	= initial setup conditions
z	= number of segments of a unit length line
α	= significance level
β	= regression coeficients
θ	= distribution parameter of quadratic form
λ	= distribution parameter of quadratic form
Λ	= diagonal matrix equal to $\mathbf{T'T}$
μ	= population mean
σ	= population standard deviation
Σ	= population covariance matrix
χ^2	= Chi-squared distribution

BIBLIOGRAPHY

Bakshi, B.R., and G. Stephanopoulos (1993), "Wave-Net: A Multiresolution, Hierarchical Neural Network with Localized Learning", *AIChE Journal*, **39**, 57-81.

Basseville M. (1988), "Detecting Changes in Signals and Systems. A Survey", *Automatica*, **24**, 309-326.

Bauer, P., and P. Hackl (1980), "An Extension of the MOSUM Technique for Quality Control", *Technometrics*, **22**, 1-8.

Birewar, D.B., and I.E. Grossmann (1989), "Incorporating Scheduling in the Optimal Design of Multiproduct Batch Plants", *Computers and Chemical Engineering*, **13**, 141-161.

Birky G. J., and T.J. McAvoy (1990), "A General Framework for Creating Expert Systems for Control System Design", *Computers and Chemical Engineering*, **14**, 713-728.

Bonvin, D., P. De Valliere, and D.W.T. Rippin (1989), "Application of Estimation Techniques to Batch Reactors I. Modeling Thermal Effects", *Computers and Chemical Engineering*, **13**, 1-9.

Bonvin, D., and D.W.T. Rippin (1990), "Target Factor Analysis for the Identification of Stoichiometric Models", *Chemical Engineering Science*, **45**, 3417-3426.

Boullion, T.L., and P.L. Odell (1971), *Generalized Inverse Matrices*, John Wiley & Sons - Interscience, New York.

Box, G.E.P. (1949), "A General Distribution Theory for a Class of Likelihood criteria", *Biometrika*, **36**, 317-346.

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.

Box, G.E.P., W.G. Hunter, J.F. MacGregor, and J. Erjavec (1973), "Some Problems Associated with the Analysis of Multiresponse Data", *Technometrics*, **15**, 33-51.

Box, G.E.P., W.G. Hunter, and J.S. Hunter (1978), *Statistics for Experiments*, John Wiley & Sons, New York.

Box, G.E.P., and Kramer, T. (1992), "Statistical Process Monitoring and Feedback Adjustment - A Discussion", *Technometrics*, **34**, 251-267.

Broadhead T.O., A.E. Hamielec, and J.F. MacGregor (1985), "Dynamic Modeling of the Batch, Semi-Batch and Continuous Production of Styrene-Butadiene Copolymers by Emulsion Polymerization", *Makromolekulare Chemie Supplement*, **10**, 105-128.

Cheung, J.T., and G. Stephanopoulos (1992), "Representation of Process Trends, Part I. A Formal Representation Framework", *Computers and Chemical Engineering*, **14**, 495-539.

Chew, V. (1968), "Simultaneous Prediction Intervals", *Technometrics*, **10**, 323-330.

Cuthrell, J.E., and L.T. Biegler (1989), "Simultaneous Optimization and Solution Methods for Batch Reactor Control Profiles", *Computers and Chemical Engineering*, **13**, 49-62.

Cutler, C.R., and B.L. Ramaker (1979), "Dynamic Matrix Control: A Computer Control Algorithm", AIChE 86th National Meeting, Paper No. 51b, Houston.

Crook, C.A., N. Shah, C. Pandelides, and A. Macchiettos (1990), "A Combined MILP and Logic Based Approach to the Synthesis of Operating Procedures for Batch Plants", AIChE Annual Meeting, Chicago.

D'Agostino, R.B., and M.A. Stephens (1986), *Goodness-of-fit Techniques*, Marcel Dekker, New York.

Dayal, B.S., J.F. MacGregor, P.A. Taylor, S. Marcikic, and R. Kidlaw (1994), "Application of Feedforward Neural Networks and Partial Least Squares Regression for Modeling Kappa Number in a Continuous Kamyer Digester", *Pulp and Paper Canada*, **95**, 26-32.

De Valliere, P., and D. Bonvin (1989), "Application of Estimation Techniques to Batch Reactors II. Experimental Studies in State and Parameter Estimation", *Computers and Chemical Engineering*, **13**, 11-20.

De Valliere, P., and D. Bonvin (1990), "Application of Estimation Techniques to Batch Reactors III. Modeling Refinements which Improve the Quality of State and Parameter Estimation", *Computers and Chemical Engineering*, **14**, 799-808.

Drapper, N., and H. Smith (1981), *Applied Regression Analysis*, John Wiley & Sons, New York.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation", *Journal of the American Statistical Association*, **78**, 316-331.

Efron, B. (1986), "How Biased is the Apparent Error rate of a Prediction Rule?", *Journal of the American Statistical Association*, **81**, 461-470.

Evans, M., N. Hastings, and B. Peacock (1993), *Statistical Distributions*, John Wiley & Sons, New York.

Fathi Z., W.F. Ramirez, and J. Korbicz (1993), "Analytical and Knowledge-Based Redundancy for Fault Diagnosis in Process Plants", *AIChE Journal*, **39**, 42-56.

Frank, P.M. (1990), "Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge Based Redundancy. A Survey and Some New Results", *Automatica*, **26**, 459-474.

Frank, I.E., and J.H. Freidman (1993), "A Statistical View of Some Chemometrics Regression Tools", *Technometrics*, **36**, 147-163.

Geladi, P., and B. Kowalski (1986), "Partial Least Squares Regression: A Tutorial", *Analytica Chimica Acta*, **185**, 1-17.

Geladi, P. (1989), "Analysis of Multi-way (Multi-mode) Data", *Chemometrics and Intelligent Laboratory Systems*, **7**, 11-30.

Geladi, P., H. Isaksson, L. Lindqvist, S. Wold, and K. Esbensen (1989), "Principal Component Analysis of Multivariate Images", *Chemometrics and Intelligent Laboratory Systems*, **5**, 209-220.

Gnanadesikan, R., and E.T. Lee (1970), "Graphical Techniques for Internal Comparisons Among Equal Degree of Freedom Groupings in Multiresponse Experiments", *Biometrika*, 57, 29-337.

Goldberg, J.L. (1991), *Matrix Theory with Applications*, McGraw Hill, New York.

Golub, G.H., and C.F. VanLoan (1989), *Matrix Computations*, John Hopkins University Press, Baltimore.

Hahn, G.J., and M.B. Cockrum (1987), "Adapting Control Charts to Meet Practical Needs: a Chemical Processing Application", *Journal of Applied Statistics*, **14**, 35-52.

Hahn, G.J, and W.Q. Meeker (1991), *Statistical Interval. A Guide for Practitioners*, John Wiley & Sons, New York.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, New Jersey.

Harris, T.J., and W.H. Ross (1991), "Statistical Process Control Procedures for Correlated Observations", *Canadian Journal of Chemical Engineering*, **69**, 48-57.

Helland, I.S. (1988), "On the Structure of Partial Least Squares Regression", *Communications in Statistical Simulations*, **17**, 581-607.

Himmelblau, D.M. (1992), "Use of Artificial Neural Networks to Monitor Faults and for Troubleshooting in the Process Industries", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

Holloway, L.E., and B.H. Krogh (1992), "On-Line Trajectory Encoding for Discrete-Observation Process Monitoring", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

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.

Iserman R. (1984), "Process Fault Detection Based on Modeling and Estimation Methods. A Survey", *Automatica*, **20**, 387-404.

Ito, K. (1969), "On the Effect of Heteroscedasticity and Nonnormality Upon Some Multivariate Test Procedures", In P.R. Krishnaiah (Ed.), *Multivariate Analysis*, Vol. II, 87-120, Academic Press, New York.

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

Jackson, J.E. (1980), "Principal Components and Factor Analysis", *Journal of Quality and Technology*, **12**, 201-213.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, John Wiley & Sons, New York.

James, M. (1985), *Classification Algorithms*, John Wiley & Sons, New York.

Jazwinski, A.A. (1970), *Stochastic Process and Filtering Theory*, Academic Press, New York.

Jenkins, G.M., and D.G. Watts (1968), *Spectral Analysis and its Applications*, Holden-Day, San Francisco.

Jensen, D.R., and H. Solomon (1972), "A Gaussian Approximation to the Distribution of a Definite Quadratic Form", *Journal of the American Statistical Association*, **67**, 898-902.

John, P.W. (1990), *Statistical Methods in Engineering Quality Assurance*, John Wiley & Sons, New York.

Johnson, N.L., and S. Kotz (1970), *Continuous Univariate Distributions*, Houghton & Miffin, Boston.

Johnson, R.A., and D.W. Wichern (1988), *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Jolliffe, I.T. (1986), *Principal Component Analysis*, Springer-Verlag, New York.

Kaspar, M.H., and W.H. Ray (1993), "Partial Least Squares Modeling as Successive Singular Value Decompositions", *Computers and Chemical Engineering*, **17**, 985-989.

Kettaneh-Wold, N., J.F. MacGregor, B. Dayal, and S. Wold (1994), "Multivariable Design of Process Experiments", *Journal of Chemometrics*, **23**, 39-50.

King, R. (1986), "Early Detection of Hazardous States in Chemical Reactors", Dynamics and Control of Chemical Reactors and Distillation Columns, IFAC DYCORD Symposium, Bournemouth UK, 8-10 December, Pergamon Press.

Konstantinov, K.B., and T. Yoshida (1992), "A Method for On-Line Reasoning about the Time-Profiles of Process Variables", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

Kourti, T., and J.F. MacGregor (1994), "Recent Developments in Multivariate SPC Methods for Monitoring and Diagnosing Process and Product Performance", submitted to *Journal of Quality Technology*.

Kourti, T., P. Nomikos, and J.F. MacGregor (1995), "Analysis, Monitoring and Fault Diagnosis of Batch Processes Using Multi-Block and Multi-Way PLS", submitted to *Journal of Process Control*.

Kowalski, K.G. (1990), "On the Predictive Performance of Biased Regression Methods and Multiple Linear Regression", *Chemometrics and Intelligent Laboratory Systems*, **9**, 177-184.

Kozub, D.J. (1989), *Multivariable Control Design - Robustness and Nonlinear Inferential Control for Semi Batch Polymerization Reactors*, Ph.D. Thesis, Department of Chemical Engineering, McMaster University, Canada.

Kozub, D.J., and J.F. MacGregor (1992a), "Feedback Control of Polymer Quality on Semi-Batch Copolymerization Reactors", *Chemical Engineering Science*, **47**, 929-942.

Kozub, D.J., and J.F. MacGregor (1992b), "State Estimation for Semi-Batch Polymerization Reactors", *Chemical Engineering Science*, **47**,1047-1062.

Kravaris, C., R.A. Wright, and J.F. Carrier (1989), "Nonlinear Controllers for Trajectory Tracking in Batch Processes", *Computers and Chemical Engineering*, **13**, 73-82.

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.

Kresta, J., T.E. Marlin, and J.F. MacGregor (1994), "Development of Inferential Process Models Using PLS", *Computers and Chemical Engineering*, **18**, 597-611.

Krzanowski, W.J. (1979), "Between Groups Comparison of Principal Components", *Journal of the American Statistical Association*, **74**, 703-707.

Krzanowski, W.J. (1983), "Cross-Validatory Choise in Principal Component Analysis; Some Sampling Results", *Journal of Statistical Computation and Simulation*, **18**, 299-314.

Krzanowski, W.J. (1987), "Cross-Validation in Principal Component Analysis", *Biometrics*, **43**, 575-584.

Kuo, B.C. (1987), *Automatic Control Systems*, Prentice Hall, New Jersey.

Kvalheim, O.M. (1988), "Interpretation of Direct Latent Variable Projection Methods and their Aims and Use in the Analysis of Multicomponent Spectroscopic and Chromatographic Data", *Chemometrics and Intelligent Laboratory Systems*, **4**, 11-25.

Kvalheim, O.M., and T.V. Karstang (1989), "Interpretation of Latent Variable Regression Models", *Chemometrics and Intelligent Laboratory Systems*, **7**, 39-51.

Layard, M.W.J. (1974), "A Monte Carlo Comparison Tests for Equality of Covariance Matrices", *Biometrika*, **16**, 461-465.

Lee, J.C., T.C. Chang, and P.R. Krishnaiah (1977), "Approximations to the Distributions of the Likelihood Ratio Statisticsfor Testing Certain Structures on the Covariance Matrices of Real Multivariate Populations", In P.R. Krishnaiah (Ed.), *Multivariate Analysis*, Vol. IV, 105-118, Academic Press, New York.

Leonard, J.A., and M.A. Kramer (1992), "A Decomposition Approach to solving Large-Scale Fault Diagnosis Problems with Modular Neural Networks", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

Lilliefors, H.W. (1967), "On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown", *Journal of the American Statistical Association*, **62**, 399-402.

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. Kowalski (1987), "A Theoretical Foundation for the PLS Algorithm", *Journal of Chemometrics*, **7**, 45-59.

MacGregor, J.F., A. Penlides, and A.E. Hamielec (1984), "Control of Polymerization Reactors: A Review", *Polymer Process Engineering*, **2**, 179-206.

MacGregor, J.F. (1986), "On-Line Reactor Energy Balances via Kalman Filter", 6th International IFAC Conference, 27-31, Ohio, October.

MacGregor, J.F., D.J. Kozub, A. Penlides, and A.E. Hamielec (1986), "State Estimation for Polymerization Reactors", Dynamics and Control of Chemical Reactors and

Distillation Columns, IFAC DYCORD Symposium, Bournemouth UK, 8-10 December, Pergamon Press.

MacGregor, J.F., and P. Nomikos (1992), "Monitoring Batch Processes", NATO Advanced Study Institute for Batch Processing Systems Engineering, Antalya, Turkey, May 29-June 7, Springer-Verlag, Heidelberg.

MacGregor, J.F., and T.J. Harris (1993), "The Exponentially Weighted Moving Average", *Journal of Quality Technology*, **25**, 106-118.

MacGregor, J.F., C. Jaeckle, C. Kiparissides, and M. Koutoudi (1994), "Monitoring and Diagnosis of Process Operating Performance by Multi-Block PLS Methods with an Application to Low Density Polyethylene Production", *AIChE Journal*, **40**, 826-838.

Mardia, K.V., J.T. Kent, and J.M. Bibby (1989), *Multivariate Analysis*, Academic Press, London.

Marsh, C.E., and T.W. Tucker (1991), "Application of SPC Techniques to Batch Units", *ISA Transactions*, ISSN 0019-0578, **30**, 39-47.

Miller, P., R.E. Swanson, and C.E. Heckler (1993), "Contribution Plots: The Missing Link in Multivariate Quality Control", submitted to *Journal of Quality Technology*.

Moler, C.B., and G.W. Stewart (1973), "An Algorithm for Generalized Matrix Eigenvalue Problems", *SIAM Journal of Numerical Analysis*, 10(2), 36-47.

Manne, R. (1987), "Analysis of Two Partial Least Squares Algorithms for Multivariate Calibration", *Chemometrics and Intelligent Laboratory Systems*, **2**, 187-197.

Montgomery, D.C. (1991), *Statistical Quality Control*, John Wiley & Sons, New York.

Moteki, Y., and Y. Arai (1986), "Operation Planning and Quality Design of a Polymer Process", IFAC DYCORD Symposium, Bournemouth UK, 8-10 December, Pergamon Press.

Nagarsenker, B.N. (1978), "Nonnull Distributions of Some Statistics Associated with Testing for the Equality of two Covariance Matrices", *Journal of Multivariate Analysis*, **8**, 396-404.

Neave, H.R., and P.L. Worthington (1988), *Distribution Free Tests*, Unwin Hyman Ltd, London.

Nelson, P., P.A. Taylor, and J.F. MacGregor (1995), "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations", submitted to *Chemometrics and Intelligent Laboratory Systems*.

Nomikos, P., and J.F. MacGregor (1994), "Monitoring Batch Processes Using Multiway Principal Component Analysis", *AIChE Journal*, **40**, 1361-1375.

Nomikos, P., and J.F. MacGregor (1995), "Multivariate SPC Charts for Monitoring Batch Processes", *Technometrics*, **37**, 41-59.

Osten, D.W. (1988), "Selection of Optimal Regression Models via Cross-Validation", *Journal of Chemometrics*, **2**, 39-48.

Patton, R.J., P.M. Frank, and K.N. Clark (1989), *Fault Diagnosis in Dynamic Systems. Theory and Applications*, Prentice-Hall, Englewood Cliffs, New Jersey.

Petterson, T., E. Hernandez, Y. Arkun, and F.J. Schork (1992), "A Nonlinear DMC Algorithm and its Application to a Semi-Batch Polymerization Reactor", *Chemical Engineering Science*, **47**, 737-753.

Petti, T.F., J. Khein, and P.S. Dhurjati (1990), "Diagnostic Model Processor Using Deep Knowledge for Process Fault Diagnosis", *AIChE Journal*, **36**, 565-575.

Phatak, A., P.M. Reilly, and A. Penlidis (1992), "The Geometry of 2-Block Partial Least Squares Regression", *Communications in Statistical Theory and Methodology*, **21**, 1517-1553.

Phatak, A., P.M. Reilly and A. Penlidis (1993), "An Approach to Interval Estimation in Partial Least Squares Regression", *Analytica Chimica Acta*, **277**, 495-501.

Prinz, O., and D. Bonvin (1992), "Monitoring Discontinuous Reactors Using Factor-Analytical Techniques", IFAC Symposium, DYCORD +92, College Park, Maryland.

Ramesh, T.S., J.F. Davis, and G.M. Schwenzer (1989), "CATCRACKER: An Expert System for Process and Malfunction Diagnosis in Fluid Catalytic Cracking Units", Proceedings AIChE Annual Meeting, San Francisco.

Rao, C.R., and S.K. Mitra (1971), *Generalized Inverse of Matrices and its Applications*, John Wiley & Sons, New York.

Rippin, D.W.T. (1992), "Current Status and Challenges of Batch Processing Systems Engineering", Proceedings of the NATO Advance Study Institute, Antalya-Turkey, May 29-June 7, Springer-Verlag, Heldelberg.

Roberts, S.W. (1959), "Control Chart Tests Based on Geometric Moving Averages", *Technometrics*, **1**, 239-250.

Rojas, C., and M.A. Kramer (1992), "Belief Networks for Knowledge Integration and Abductive Inference in Fault Diagnosis of Chemical Processes", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

Sanchez, E., and B.R. Kowalski (1990), "Tensorial Resolution: A Direct Trilinear Decomposition", *Journal of Chemometrics*, **4**, 29-45.

Schuler, H., and K. De Haas (1986), "Semi-Batch Reactor Dynamics and State Estimation", Dynamics and Control of Chemical Reactors and Distillation Columns, IFAC DYCORD Symposium, Bournemouth UK, 8-10 December, Pergamon Press.

Schuler, H., and C.U. Schmidt (1992), "Calorimetric-State Estimators for Chemical Reactor Diagnosis and Control: Review of Methods and Applications", *Chemical Engineering Science*, **47**, 899-915.

Schuurman, F.J., V.B. Waikar, and P.R. Krishnaiah (1973), "Percentage Points of teh Joint Distribution of the Extreme roots of the Random Matrix $(S_1+S_2)^{-1}$", *Journal of Statistical Computational Simulations*, **2**, 17-38.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, John Wiley & Sons, New York.

Seber, G.A.F. (1984), Multivariate Observations, John Wiley & Sons, New York.

Shewhart (1931), W.A, *Economic Control of Quality of Manufactured Product*, Von Nostrand, New Jersey.

Skagerberg, B., J.F. MacGregor, and C. Kiparissides (1992), "Multivariate Data Analysis Applied to Low Density Polyethylene Reactors", *Chemometrics and Intelligent Laboratory Systems*, **14**, 341-356.

Slama, C.F. (1991), *Multivariate Statistical Analysis of Data Obtained from an Industrial Fluidized Catalytic Process Using PCA and PLS*, Master Thesis, Department of Chemical Engineering, McMaster University, Canada.

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.

Smilde, A.K. (1992), "Three-way analyses. Problems and Prospects", *Chemometrics and Intelligent Laboratoty Systems*, **15**, 143-157.

Stahle, L., and S. Wold (1987), "Partia! Least Squares Analysis with Cross-Validation for the Two Class Problem: A Monte Carlo Study", *Journal of Chemometrics*, **1**, 185-196.

Stephanopoulos, G., G. Henning, and H. Leene (1990), "Model L.A. A Modeling Language for Process Engineering. Part I The Formal Framework", *Computers and Chemical Engineering*, **14**, 813-846.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of Royal Statistical Society*, **2**, 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 Royal Statistical Society*, **52**, 237-269.

Terpstra, V.J., H.B. Verbruggen, M.W. Hoogland, and R.A.E. Ficke (1992), "A Real Time, Fuzzy, Deep Knowledge Based Fault Diagnosis System for a CSTR", On-line Fault Detection and Supervision in the Chemical Process Industries, IFAC Symposium, Newark-Delaware, April 22-24.

Tracy, N.D., J.C. Young, and R.L. Mason (1992), "Multivariate Control Charts for Individual Observations", *Journal of Quality Technology*, **24**, 88-95.

Tucker, W.T., Faltin, F.W., and Vander Wiel, S.A. (1993), "Algorithmic Statistical Process Control: An Elaboration", *Technometrics*, **35**, 363-375.

Vander Wiel, S.A., W.T. Tucker, F.W. Faltin, and N. Doganaksoy (1992), "Algorithmic Statistical Process Control: Concepts and an Application", *Technometrics*, **34**, 286-297.

Venkatasubramanian, V., and K. Chan (1989), "A Neural Network Methodology for Process Fault Diagnosis", *AIChE Journal*, **35**, 1993-2002.

Wangen, L.E., and B.R. Kowalski (1988), "A Multi-block Partial Least Squares Algorithm for Investigating Complex Chemical Systems", *Journal of Chemometrics*, **3**, 3-20.

Willsky, A.S. (1976), "A Survey of Design Methods for Failure Detection in Dynamic Systems", *Automatica*, **12**, 601-611.

Wise, B.M., and N.L. Ricker (1991), "Recent Advances in Multivariate Statistical Process Control Improving Robustness and Sensitivity", IFAC ADCHEM'91, Toulouse, France, Pergamon Press, 125-130.

Wierda, S.J. (1994), "Multivariate Statistical Process Control - Recent Results and Directions for Future Research", *Statistica Neerlandica*, **48**, 147-168.

Wold, H. (1982), "Soft Modeling. The Basic Design and Some Extensions", In K.G. Joreskog and H. Wold (Ed.), *Systems Under Indirect Observation*, Chapter 1, Vol II, North Holland, Amsterdam.

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models", *Technometrics*, **20**, 397-405.

Wold, S., K. Esbensen, and P. Geladi (1987a), "Principal Component Analysis", *Chemometrics and Intelligent Laboratory Systems*, **2**, 37-52.

Wold, S., P. Geladi, K. Esbensen, and J. Ohman (1987b), "Multi-Way Principal Components and PLS Analysis", *Journal of Chemometrics*, **1**, 41-56.

Wold, S. (1992), "Non-Linear PLS Modeling II. Spline Inner Relation", *Chemometrics and Intelligent Laboratory Systems*, **14**, 71-84.

Woodhall, W.H., and M.M. Ncube (1985), "Multivariate CUSUM Quality Control Procedures", *Technometrics*, **27**, 285-292.

Woodward, R.H., and P.L. Goldsmith (1964), *Cumulative Sum Techniques*, Oliver & Boyd, London.

Wu, R.S.H. (1985), "Dynamic Thermal Analyzer for Monitoring Batch Processes", *Chemical Engineering Progress*, **81**, 57-61.

Zeng, Y., and Hopke, P.K. (1990), "Methodological Study Applying Three-Way Factor Analysis to Three-Way Chemical Datasets", *Chemometrics and Intelligent Laboratory Systems*, 7, 237-250.