IMPROVEMENT OF PROCESSES AND PRODUCT QUALITY THROUGH MULTIVARIATE DATA ANALYSIS

By

CARL DUCHESNE, B.A.Sc., M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

IMPROVING PRODUCT QUALITY THROUGH MULTIVARIATE DATA ANALYSIS

Doctor of Philosophy (2000)
(Chemical Engineering) | McMaster University
Hamilton, Ontario

TITLE: Improvement of Processes and Product Quality through Multivariate Data Analysis

AUTHOR: Carl Duchesne, B.A.Sc., M.Sc. (Laval University)

SUPERVISORS: Professor John F. MacGregor
Professor Theodora Kourti

NUMBER OF PAGES: xix, 203

Abstract

This thesis focuses on developing empirical methodologies for improving process operation and product quality, in four important chemical engineering problems, using multivariate projection methods, such as Principal Component Analysis (PCA) and Projection to Latent Structures (PLS). The four problems addressed in this work are concerned with (i) improving and optimizing the trajectories of manipulated variables in batch processes; (ii) improving the identification of non-parsimonious dynamic process models using the Jackknife and the Bootstrap methods; (iii) developing meaningful specification regions for raw materials entering a consumer's plant and, (iv) improving transition policies in start-ups, re-starts and grade changeovers that are routinely performed in multi-product plants.

The first problem addresses the situation where one desires to gain understanding of how and when, during the course of a batch, manipulated process variables have a significant effect on product quality. This amounts to estimating the sensitivity of product quality to manipulated process variables at various degrees of completion in a batch process. This information can be used in process development and in the optimization of already existing processes. The proposed approach involves adding designed experiments to batch policies currently used and then analyzing the resulting data bases using multi-way multi-block PLS. A new pathway PLS algorithm was developed for incorporating intermediate quality measurements collected during the course of each batch.

In the second problem, the identification of non-parsimonious dynamic process models is improved through a more judicious selection of the meta parameter in regularization methods (ridge regression) and latent variable methods. These methods are often used to overcome ill-conditioning frequently encountered in the identification of such over-parameterized models. A new criterion for selecting the meta parameter (ridge parameter in regularization methods and the number of components in latent variable methods) is proposed, based on Jackknife and Bootstrap statistics. It is shown that this criterion outperforms the use of cross-validation (default criterion) and leads to the identification of models that are closer to the true process behavior.

Developing an approach for defining multivariate specifications on incoming raw materials was important because there is a void in the quality control literature in this area. Specifications are usually defined in a univariate manner, based on past and often subjective experience. This work provides a sound, data-based approach for developing truly multivariate specification regions in a variety of industrial situations. The approach uses PLS methods to analyze historical data on the incoming raw materials, on the consumer's plant, and on the consumer's end product to define multivariate specification regions for the incoming raw material properties.

The last problem consists of improving the performance of process transitions (start-ups, re-starts and grade changeovers), using historical process data. In particular, this work addresses two questions: (i) how to improve transition policies to minimize transition time and amount of off-grade materials, while ensuring safe operating conditions and (ii) how to ensure that at the end of a transition, steady-state process conditions are such that good quality products are obtained, and are consistent with past periods of production.

Acknowledgements

This doctorate degree program was certainly the most difficult and challenging project I have undertaken in my life. However, it would have never been taken to an end without help, support and encouragement from the many people I wish to acknowledge here. But first, I am very thankful to the Department of Chemical Engineering at McMaster University and the Natural Sciences and Engineering Research Council (NSERC) of Canada for their financial support, which relieved me from a lot of worries.

To begin, I would like to thank Dr. John F. MacGregor for his guidance and constant support. I am also extremely grateful to him for his genuine human understanding, especially in my final year, when I decided to finish my work from home in Quebec City. I will never be able to put in words how great you have been to me, and I will always remember you as a mentor of the highest class.

My final year was also special, since Dr. Theodora Kourti joined the team. She has always been there when I needed (even when I was worried about taking the airplane). She found time to help me and support me even when she had not a single minute for herself. She has made my life so much better by sharing her expertise and knowledge. This thesis would have never been the same without her precious input. I also wish to mention the great contributions of Dr. Roman Viveros and Dr. John Vlachopoulos, who improved the quality of this work with their suggestions and ideas. Special thanks go to Dr. John Vlachopoulos for providing his commercial software on the film blowing process.

When I first came to Hamilton, I was alone, worried, and barely spoke any English. However, I quickly found a second family away from home. To my friends in the Department: François, Song, Thanassis, Seongkyu, San, Jose, Tracy, Ali, Christiane, Jason, Ivan, Chunliang, Manish, Steve, Laura, Jean-François, Ruijie, Jesus, Ed and Barb, I feel fortunate to have you as friends, and I am taking with me all the good memories of our intense discussions and the social and sport activities we enjoyed together.

To my closest friends, Tracy, Manish, Ken, François, Jean-François and Barb, your true and unconditional friendship will always remain in my heart. I thank you so much for the support you gave me. Special thanks go to Manish and Ken for hosting me several times in the course of my final year, and for all the printing jobs they did for me.

En terminant, je désire dédier cette thèse à mes parents, Rémi and Jocelyne. À travers votre amour inconditionnel, vos sacrifices fait sans compter, vous m'avez légué un précieux héritage, dont le sens des valeurs et le goût du travail bien fait font partie. Je vous serai toujours reconnaissant pour ce que vous m'avez donné et surtout, pour ce que vous m'avez permis de devenir. Le succès que je vie aujourd'hui est d'abord le vôtre. Cette thèse, je la dois aussi à ma future femme, Sonia. Sans ton amour, ton support et ta persévérance, ce travail n'aurait jamais vu le jour. Je te remercie de ta patience, ce travail est, bien-sûr, une victoire professionnelle, mais aussi le symbol de la victoire de notre amour.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Improving process operation and product quality is of paramount importance for the chemical processing industry. Processes are often required to operate over a wide range of conditions to satisfy consumer's demand for various products. The quality of these products should always meet high standards in spite of disturbances affecting the process operation, such as variations in raw material properties, environmental conditions, impurities and so forth. In addition, the profitability of chemical operations must be ensured through a good control of productions costs and productivity. To resolve this rather complicated problem, process models either built using fundamental principles or using process data (empirical models) are often developed to aid operators and engineers.

When good fundamental process models are available, a lot can be done to improve process operation and product quality. For example, these models can be used to optimize steady-state operation or transitions between different modes of operation. Robust control systems can be designed to reject disturbances and maintain stable operation. The sensitivity of process operability and product quality to variations in the properties of raw materials and other disturbances can also be assessed.

However, when no such detailed fundamental models exist and when they are too expensive and time consuming to develop, empirical models are a very useful alternative approach to achieve better process operation and products.

Empirical modelling approaches are also attractive since nowadays, process measurements are routinely collected using computers, and these data are readily available for modelling purposes. Often hundreds to thousands of highly correlated measurements are collected, with some missing values, due to sensor failure and other reasons. In the past, multivariate projection methods such as Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) have proven to be successful for analyzing these type of data bases in many applications, such as in the analysis of historical process data, process monitoring and fault detection (Kourti and MacGregor, 1995; Kourti *et al.*, 1996), soft sensors (Kresta *et al.*, 1994; Roney, 1998), dynamic model identification (MacGregor *et al.*, 1991; Dayal and MacGregor, 1996) and product design (Jaeckle and MacGregor, 1998; Jaeckle and MacGregor, 2000) etc. It would therefore seem logical to extend the application of these methods for solving other chemical engineering problems.

The objective of this thesis consists of developing sound, empirical methodologies based on multivariate projection methods, or modifications of them, to solve four important problems, frequently encountered in chemical engineering. These problems are: (i) improving and optimizing the process variable trajectories in batch processes; (ii) identifying better non-parsimonious dynamic models for process control purposes; (iii) developing multivariate specification regions for raw materials and, (iv) improving start-ups and grade changeovers and other process transitions in multi-product plants. A brief description of each problem as well as a discussion of the objectives and contributions of this work are presented below.

1.1 Multivariate Analysis and Optimization of Process Variable Trajectories for Batch Processes

In the development and optimization of batch processes, it is necessary to understand how and when during the course of a batch, manipulated process variables affect product quality. For example, in fermentation and polymerization, the batches often go through several different stages of operation, within which different physical phenomena occur. Some process variables strongly affecting quality early in the batch may have no effect beyond a certain degree of completion is reached, while some other variables may have a consistent impact on quality throughout the batch. Identifying what features of batch process trajectories affect final product quality is therefore important to improve the quality of already existing products and for developing new products.

A common solution to this problem, which is well documented in the literature, is to make use of a detailed fundamental model of the process in conjunction with some optimization algorithm. This allows one to compute the optimal trajectories to implement on the manipulated process variables to obtain the desired product. When such good models are not available, alternative empirical approaches include Evolutionary Operation (EVOP) and Response Surface Methods (RSM)(Box and Draper, 1969), however both methods are limited to the optimization of a few process variables only (steady-state optimization). Another empirical approach is aimed at classifying the outcome of a batch according to trajectory features, but this is complex; it involves filtering trajectories using wavelet functions and classification using decision trees (Bakshi and Stephanopoulos, 1994*a*, 1994*b*).

The primary objective of this work was to develop a simpler empirical method for identifying the features of process variable trajectories that are important for product quality. The proposed approach consists of superimposing some designed experiments to currently used batch policies (trajectories), and then extract the desired information using multi-way multi-block PLS projection methods. The methodology is illustrated using simulations of a Styrene-Butadiene Rubber (SBR) emulsion copolymerization process.

The contribution of this work is twofold. It provides a simple methodology to gain batch process understanding through identifying the features of batch trajectories that are important to final product quality. Second, in the course of this work, a new pathway PLS algorithm was developed to incorporate in the analysis, intermediate quality measurements that are collected during the course of each batch. This additional information allows to extract the desired features using a smaller number of batch runs (and designed experiments). This was the first time an algorithm was proposed for analyzing intermediate quality measurements in a physically meaningful fashion. A version of this chapter has been published in the *Chemometrics and Intelligent Laboratory Systems* journal (Duchesne and MacGregor, 2000a).

1.2 Jackknife and Bootstrap Methods in the Identification of Dynamic Models

Non-parsimonious dynamic models, such as finite impulse response (FIR) models and autoregressive with exogenous variables (ARX) models, are widely used in the industry for process identification. They provide a lot of flexibility with a minimal number of structural choices, and are readily incorporated in the design of multivariable model predictive controllers. However, parameter estimation of such non-parsimonious models is often ill-conditioned, due to the large number of lagged input variables used as predictors. Since it is well known that least squares and minimum prediction error estimates are sensitive to correlation among predictor variables, parameter estimation of non-parsimonious models is rather performed in practice using regularization methods (such as ridge regression) and latent variable methods (PCR, PLS, CCR). These methods can achieve lower mean square error in the parameter estimates (model is closer to true process) in ill-conditioned problems by obtaining a reduced variance in parameter estimates at the expense of a slight bias. The compromise between variance reduction and bias is chosen via a meta parameter, the ridge parameter in regularization methods and the number of components in latent variable methods. Whenever the latter methods are utilized, especially for latent variable methods, cross-validation is typically used to select the number of components (based on maximizing model predictive power), and this has been shown in past literature to provide poor results since too few components are generally selected. The objective of this work is therefore to develop a new criterion to aid in selecting non-parsimonious dynamic models (through a judicious choice of the meta parameter) that would not only lead to good model predictions, but that would also lead to capture the correct process structure (e.g. identify models that are closer to the true process).

The proposed criterion suggests to keep adding latent variables as long as the sum of squares of the model residuals is decreasing, but to stop when there is evidence that one starts overfitting. This evidence is provided by a measure of the total variance of the model parameter estimates (which rises rapidly when overfitting occurs), obtained using the Jackknife or the Bootstrap statistical procedures. Through a few simulation studies, it is shown that the proposed criterion outperforms the use of cross-validation in most cases (as to provide a better model, in the mean square error sense), and also shows when least squares can be used for estimating the model parameters without too much loss.

The contribution of this work is mainly in providing a new criterion for selecting the meta parameter of regularization methods and latent variables methods, for improving the identification of dynamic models. In particular, this work proposes a solution to the well known and unresolved problem of cross-validation selecting too few components to capture the correct process structure, when latent variable methods are utilized for parameter estimation. However, the proposed criterion is not limited to dynamic identification problems, but can be extended to other situations where the objective is not only to obtain good predictions of the model output, but also to capture the true process behavior as much as possible. In fact, the motivation for developing such a criterion arose in the course of the work on batch trajectory optimization. Better models of the batch process, in the sense that they were closer to the true process behavior, were obtained when using the number of latent variables suggested by this new criterion. A version of this chapter has been published in the *Journal of Process Control* (Duchesne and MacGregor, 2000b).

1.3 Defining Multivariate Specification Regions

Developing meaningful specification regions for selecting new incoming lots of raw materials in a customer's plant is crucial to ensure that the customer's desired final product quality is achievable, given the limits of operability of his own process. Defining such specification regions could be used, for example, to reduce the efforts made by operators and control systems in compensating for poor quality raw materials. This could be achieved by reducing the acceptance region for raw materials. However, in spite of their importance, no standard industrial practice seems to exist for defining raw material specifications. They are defined rather arbitrarily, based on past and often subjective experience. For example, one common practice is to set tight limits on important properties, to a range corresponding to the best supplier's quality. This may be due to the void in the quality control literature regarding methodologies for developing specifications. Extensive literature exists on univariate measures of quality (loss functions and desirability functions) and univariate measures of the ability of a process to meet certain specifications (capability indices), however these always assume that specifications are already defined.

Another issue encountered in practice (and in the literature) is that quality is frequently assumed to be a set of univariate properties, while in most process engineering problems, quality is a truly multivariate property. Material quality is often judged based on several highly correlated properties and so, applying univariate specifications on these properties would lead to a high probability of mis-judging quality. The objective of this research is therefore to propose a sound, empirical methodology for developing multivariate specification regions for incoming lots of raw materials entering a consumer's plant.

The proposed approach accounts for most industrial situations in which not only raw materials variations affect the final product quality, but the way the consumer's process is operated may also introduce significant variations. The method also accounts for the presence of feedforward and feedback control actions (from operators or control systems) compensating for variations in raw material properties. The main idea of this approach consists of defining a multivariate specification region that reflects the sensitivity of consumer's final product quality to various combinations of raw material properties, given specific process operational policies. The sensitivity is evaluated via the use of latent variable models, built over historical consumer's data bases (including all available measurements on raw materials, process manipulated variables and final product characteristics). The proposed approach is illustrated using simulations of a film blowing process, producing various types of polymer films.

This work greatly contributes to the field as it seems to be one of the very first attempts to develop a sound (empirical) methodology for defining multivariate specification regions for selecting new incoming lots of raw materials entering a consumer's plant.

1.4 Analysis of Start-Ups and Grade Transition Problems

Frequent start-ups or grade transitions are performed in multi-product processes, resulting in important production loss (lost of production time, production of off-grade materials, etc.). In order to ensure efficient transition operations, two problems need to be addressed. The first problem consists of finding the best transitions to implement during start-ups and grades changeovers, to miminize transition time and off-grade materials while maintaining safe operating conditions. Optimization based on fundamental process models is the most commonly encountered solution in the literature. However, when these models are not available, data collected from past transitions could provide useful information on how to improve transition performance, since in the past, some transitions were characterized by shorter duration, smaller amounts of off-specifications materials and safer operation. To reveal this information from historical data bases, this work proposes an efficient empirical technique to analyze transition data.

Practical transitional problems can be divided in two situations. One situation is when several distinct transition policies have been implemented in the past (and are present in the data bases), with some variations in their implementation. In this case, determining what transition policies have led to the most desirable performance is straightforward. However, since a wide range of policies may exist, one could use the methodology developed in the batch trajectory optimization work to gain understanding of the features that lead to the most desirable transition behavior. On the other hand, if only one transition policy has been implemented in the past (second situation), again with some variations in its implementation, then another approach needs to be taken for improving transitions (since the range of variation in the trajectories is expected to be smaller). It consists of developing a monitoring region for the transitions, based on the most successful past transitions. Methods that have been developed and successfully used for batch process monitoring can be used in this case (Nomikos and MacGregor, 1994*a*; Nomikos and MacGregor, 1994*b*; Nomikos and MacGregor, 1995; Kourti *et al.*, 1995). Monitoring transitions should allow one to reduce deviations from the most desirable transitions and therefore, should lead to an improvement in future transitions. This approach is illustrated in the thesis using an industrial example of a polymerization process.

The second problem is concerned with the final stage of transitions, the steady-state production of in-specification products. Since multi-product processes operate over a wide range of conditions, dictated by market demand, the question of how to guarantee that product grade currently produced is of high quality and is consistent with past production periods arises. This is the problem of defining "production readiness". The key issue is that generally, properties defining true product quality as seen by customers are never all measured, but only a subset of them are. This subset is used to judge "overall" product quality and to determine if it meets specifications. In flexible processes, often multiple operating conditions are capable of targeting the few quality variables in the desired region. However, the impact of operating differently on the unmeasured quality variables is unknown. It might happen that some material, thought to be on target, might in fact be out of specifications even if there is no evidence from the supplier viewpoint. The objective of this work is to propose an approach for defining "production readiness" regions. It uses already available process monitoring tools (PCA and PLS) and data from steady-state periods of production, for a specific product, that have led to good customer satisfaction in the past. The concepts are illustrated with a simulation study of a linear low density polyethylene reactor.

This research contributes to the field in proposing empirical approaches to extract the information contained in historical process data collected before, during and after transitions (start-ups, re-starts and grade changeovers), for solving operational and quality problems. In particular, the methods allow one to improve transition policies (reduce transition time and off-grade materials) and to ensure that at the end of transitions, steady-state processing conditions are such that good quality products, consistent with the past, are obtained. It seems that the first problem has never been approached from an empirical point of view, to the author's knowledge, and the second problem is rarely discussed in the literature.

1.5 Thesis Outline

This thesis consists of 7 chapters, the first one being the current introduction. Chapter 2 provides some background on multivariate projection methods and related modelling issues that is necessary for understanding this thesis. Chapters 3-6 form the core of this thesis. Batch trajectory improvement and optimization is discussed in chapter 3 and the use of Jackknife and Bootstrap methods in dynamic model identification is presented in chapter 4. Chapters 5 and 6 focus on the development of multivariate specification regions and on the multivariate analysis of start-up and grade transition problems, respectively. Finally the results obtained in this thesis are summarized in chapter 7, where some conclusions are drawn and future work is discussed.

Chapter 2

Background on Multivariate Projection Methods

This chapter provides a background on some multivariate projection methods and related modelling issues that are necessary for understanding the work presented in this thesis. The first section briefly describes the projection methods used throughout the thesis. Then, modelling issues such as scaling and selection of the number of latent variables are discussed.

2.1 Methods

Only basic descriptions of Principal Component Analysis (PCA), Principal Component Regression (PCR) and Projection to Latent Structures (PLS) are presented in this section, but for a more rigorous and complete discussion on these methods one should refer to Burnham *et al.* (1996).

2.1.1 Principal Component Analysis (PCA)

Principal Component Analysis is a classical multivariate data analysis approach and a tutorial with some chemical examples can be found in Wold *et al.* (1987). Figure 2.1 (a) shows a data table, **X**, that consists of I measurements taken on J different variables. PCA takes advantage of the correlation structure among the J variables to summarize **X** into a few principal components or latent variables. The number of principal components, A, is generally smaller than the number variables J and is often viewed as an estimate of the effective rank of **X**.

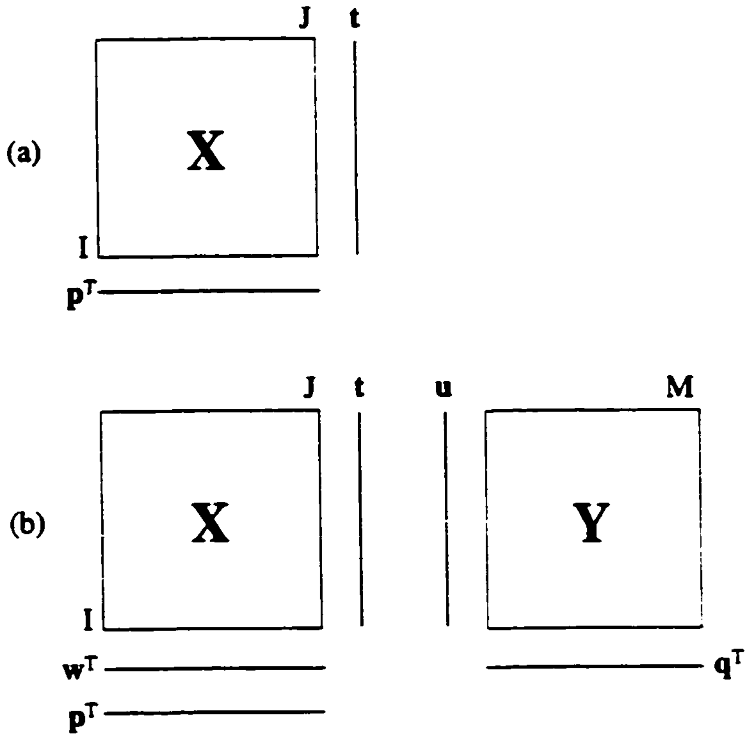


Figure 2.1: Data vectors and matrices involved in projection methods; (a) PCA; (b) PLS.

From a rigorous mathematical point of view, the PCA model building procedure starts with finding the direction in $\mathbf{X}$, or alternatively, finding a linear combination of $x-$variables $\mathbf{p_1}$ that explains the most variance in $\mathbf{X}$:

$$\max_{\mathbf{p_1}} \left\{ \mathbf{p}_1^\top \mathbf{X}^\top \mathbf{X} \mathbf{p}_1 \right\} \qquad \text{subject to} \qquad \mathbf{p}_1^\top \mathbf{p}_1 = 1.0 \tag{2.1}$$

A first summary variable or score, $\mathbf{t_1}$, is obtained simply by projecting $\mathbf{X}$ in the direction of $\mathbf{p_1}$, $\mathbf{t_1} = \mathbf{X}\ \mathbf{p_1}$. This high variance direction is then removed from $\mathbf{X}$, leaving a residual matrix $\mathbf{E_1} = \mathbf{X} - \mathbf{t_1}\ \mathbf{p}_1^\top$, containing the variance of $\mathbf{X}$ that is not explained by the first component. The construction of the PCA model can continue with the computation of a second linear combination $\mathbf{p_2}$, explaining the second highest amount of variance in $\mathbf{X}$. The objective in this case is the same as shown in Equation 2.1, but with replacing $\mathbf{p_1}$ by $\mathbf{p_2}$ and $\mathbf{X}$ by $\mathbf{E_1}$ and imposing the additional constraint that the second component be orthogonal to the first one (e.g. $\mathbf{p}_1^\top\ \mathbf{p}_2 = 0$). This procedure is repeated until the desired number of components is computed. The final structure of the model is $\mathbf{X} = \mathbf{T}\ \mathbf{P}^\top + \mathbf{E}$, which can be seen as an eigenvector decomposition of $\mathbf{X}$. In fact, the $\mathbf{p}$ vectors are just the eigenvectors of $\mathbf{X}^\top\ \mathbf{X}$ and the $\mathbf{t}$ vectors are the eigenvectors of $\mathbf{X}\ \mathbf{X}^\top$. When as many components are computed as there is variables (e.g. A=J), the decomposition of $\mathbf{X}$ is perfect and $\mathbf{E} = \mathbf{0}$.

An alternative approach for computing the $\mathbf{p}$ and $\mathbf{t}$ vectors sequentially is to use the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm. The starting point of this algorithm usually consists of mean-centering and scaling of matrix $\mathbf{X}$. Scaling is discussed in more details in Section 2.2.1. The following steps are outlined below:

1. Set $\mathbf{t}$ to be one column of $\mathbf{X}$

2. $\mathbf{p} = \mathbf{X}^\top\ \mathbf{t}/\mathbf{t}^\top\ \mathbf{t}$

3. $\mathbf{p} = \mathbf{p}/(\mathbf{p}^\top\ \mathbf{p})$

4. $\mathbf{t} = \mathbf{X}\,\mathbf{p}/(\mathbf{p}^{\top}\,\mathbf{p})$
5. Continue iterating between 2. and 4. until convegence on $\mathbf{t}$ or $\mathbf{p}$
6. Residual matrix: $\mathbf{E} = \mathbf{X} - \mathbf{t}\,\mathbf{p}^{\top}$
7. Store $\mathbf{p}$ and $\mathbf{t}$ in $\mathbf{P}$ and $\mathbf{T}$ respectively
8. Calculate next dimensions by returning to 1, using $\mathbf{E}$ as the new $\mathbf{X}$

After computing each latent variable, one needs to decide whether another dimension should be added to the PCA model. Criteria for selecting the number of components to keep in the model are discussed in Section 2.2.2.

2.1.2 Principal Component Regression (PCR)

Principal Component Regression is used when one wants to model the covariance structure in $\mathbf{X}$, as in PCA, but also desires to model the relationship between two blocks of data; a block of predictor variables $\mathbf{X}$ and a block of response variables $\mathbf{Y}$, as shown in Figure 2.1 (b). Regression coefficient estimates for PCR, $\hat{\beta}_{PCR}$, are computed in a very similar way as for Multiple Linear Regression (Draper and Smith, 1981):

$$\hat{\beta}_{PCR} = (\mathbf{T}^{\top}\mathbf{T})^{-1}\mathbf{T}^{\top}\mathbf{Y} \tag{2.2}$$

where $\mathbf{T}$ is the matrix of the scores obtained via a PCA model with a given number of components. In other words, PCR is just as MLR but instead of projecting $\mathbf{Y}$ on $\mathbf{X}$ directly (as does MLR), $\mathbf{Y}$ is projected in the reduced space of $\mathbf{X}$, $\mathbf{T}$. The structure of the PCR model is as follows:

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{T}\,\mathbf{P}^{\top} \\ \mathbf{Y} &= \mathbf{T}\,\hat{\beta}_{\mathbf{PCR}} + \mathbf{F} = \mathbf{X}\,(\mathbf{P}\,\hat{\beta}_{\mathbf{PCR}}) + \mathbf{F} \end{aligned} \tag{2.3}$$

This extention of MLR is particularly useful to reduce the large variance in regression coefficients obtained from MLR, resulting when $\mathbf{X}$ is ill-conditioned. It is even more useful when $\mathbf{X}$ is singular, since $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ does not exist in that case.

2.1.3 Projection to Latent Structures (PLS)

Projection to Latent Structures, or alternatively, Partial Least Squares is a truly multivariate latent variable regression method. PLS is used to model relationships both within and between two blocks of data, $\mathbf{X}$ and $\mathbf{Y}$. A tutorial on PLS is found in Geladi and Kowalski (1986) and a review of PLS history is available in Geladi (1988). Some mathematical and statistical properties of PLS were also addressed by Höskuldsson (1988) and by Burnham *et al.* (1996).

In PLS, the covariance structures of $\mathbf{X}$ and $\mathbf{Y}$ are modelled via a set of A latent variables, $\mathbf{t}$ and $\mathbf{u}$ respectively, as shown in Figure 2.1 (b). However, these latent variables are computed in such a way that the covariance between the two blocks $\mathbf{X}$ and $\mathbf{Y}$ is modelled as well. Mathematically, this is achieved by selecting a set of linear combinations of $\mathbf{X}$ variables, $\mathbf{w}_i$, $i = 1, \ldots, A$, that maximizes the covariance between the matrix of descriptors $\mathbf{X}$ and the matrix of responses $\mathbf{Y}$, under the following constraints:

$$\begin{aligned} \max_{\mathbf{w}_i}\left\{\mathbf{w}_i\, \mathbf{X}^{\mathsf{T}}\, \mathbf{Y}\mathbf{Y}^{\mathsf{T}}\, \mathbf{X}\, \mathbf{w}_i^{\mathsf{T}}\right\} \quad & \text{subject to} \quad \mathbf{w}_i^{\mathsf{T}}\, \mathbf{w}_i = 1 \\ & \text{subject to} \quad \mathbf{w}_i^{\mathsf{T}}\, \mathbf{w}_j = 0 \quad \text{for} \quad i \neq j \end{aligned} \tag{2.4}$$

The structure of the PLS model is shown below:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\,\mathbf{P}^{\mathsf{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T}\,\mathbf{Q}^{\mathsf{T}} + \mathbf{F} \\ \mathbf{T} &= \mathbf{X}\,\mathbf{W} \end{aligned} \tag{2.5}$$

where the A columns of $\mathbf{P}$ and $\mathbf{Q}$ define the linear combinations of the $\mathbf{X}$ and $\mathbf{Y}$

variables modelling their covariance structure. $\mathbf{E}$ and $\mathbf{F}$ are just model residuals. It has been shown by Höskuldsson (1988) that the vectors $\mathbf{w}$, $\mathbf{q}$, $\mathbf{t}$ and $\mathbf{u}$ are eigenvectors of $\mathbf{X}^\top\mathbf{YY}^\top\mathbf{X}$, $\mathbf{Y}^\top\mathbf{XX}^\top\mathbf{Y}$, $\mathbf{XX}^\top\mathbf{YY}^\top$ and $\mathbf{YY}^\top\mathbf{XX}^\top$, respectively.

Instead of computing all the latent variables at once, a version of the NIPALS algorithm was adapted for sequentially computing the PLS latent variables, one at a time. This algorithm is outlined below, with the starting point being mean-centering and scaling both $\mathbf{X}$ and $\mathbf{Y}$ matrices:

1. Set $\mathbf{u}$ to be one column of $\mathbf{Y}$
2. $\mathbf{w} = \mathbf{X}^\top\ \mathbf{u}/\mathbf{u}^\top\ \mathbf{u}$
3. $\mathbf{w} = \mathbf{w}/(\mathbf{w}^\top\ \mathbf{w})$
4. $\mathbf{t} = \mathbf{X}\ \mathbf{w}/(\mathbf{w}^\top\ \mathbf{w})$
5. $\mathbf{q} = \mathbf{Y}^\top\ \mathbf{t}/\mathbf{t}^\top\ \mathbf{t}$
6. $\mathbf{u} = \mathbf{Y}\ \mathbf{q}/(\mathbf{q}^\top\ \mathbf{q})$
7. Continue iterating between 2. and 6. until convegence on $\mathbf{t}$ or $\mathbf{u}$
8. Residual matrix: $\mathbf{E} = \mathbf{X} - \mathbf{t}\ \mathbf{p}^\top$, $\mathbf{F} = \mathbf{Y} - \mathbf{t}\ \mathbf{q}^\top$
9. Store $\mathbf{w}$, $\mathbf{p}$, $\mathbf{t}$ and $\mathbf{u}$ in $\mathbf{W}$, $\mathbf{P}$, $\mathbf{T}$ and $\mathbf{U}$ respectively
10. Calculate next dimensions by returning to 1, using $\mathbf{E}$ and $\mathbf{F}$ as the new $\mathbf{X}$ and $\mathbf{Y}$

Once the PLS model is built using a certain number of dimensions, the regression coefficient estimates can be computed using the following expression:

$$\hat{\beta}_{PLS} = \mathbf{W}\ (\mathbf{P}^\top\mathbf{W})\ \mathbf{Q}^\top \tag{2.6}$$

An interesting feature of PLS is that when all possible latent variables are computed ($A^{max} = \min\{J, I-1\}$), the PLS regression coefficient estimates, $\hat{\beta}_{PLS}$, and hence the predictions of that model, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}_{PLS}$, are exactly equal to MLR estimates.

2.1.4 Common Extensions to Projection Methods

Multi-Way data

Multi-way data is obtained when more than two arguments are necessary to describe a particular measurement, and so measurements are stored in arrays. For chemical process applications, 3-way arrays are the most commonly encountered type of multi-way data, as shown in Figure 2.2 (a).

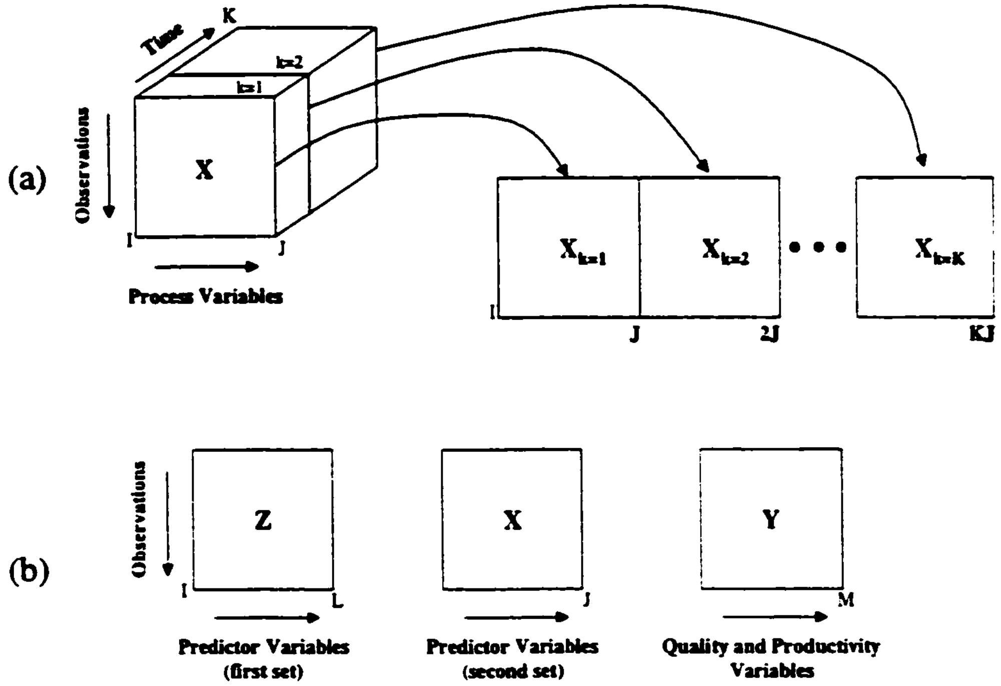


Figure 2.2: Nature of multi-way (a) and multi-block process data (b).

Arrays such as **X** mainly arise when trajectories of process variables are measured;

measurements on J process variables are collected at K sampling intervals in I instances. For example, this may consist of measurements taken on J process variables in a batch process, sampled at K intervals ($k = K$ being the end of each batch) and collected for I batch runs. Similarly, this could consist of transition policies measured during start-ups or grade changeovers.

To analyze 3-way arrays ($\mathbf{X}$) using the projection methods described in Section 2.1, one often unfolds $\mathbf{X}$ into a 2-way matrix. This can be done in six different ways, but for process applications (analysis and monitoring of batch process data), Nomikos and MacGregor (1994*a*, 1994*b*, 1995) and Kourti *et al.* (1995) suggest that the most meaningful way to unfold is to slice $\mathbf{X}$ along the time dimension and then juxtapose these submatrices to form an I by KJ matrix. This way of unfolding $\mathbf{X}$ is illustrated in Figure 2.2 (a). Unfolding $\mathbf{X}$ in this manner allows projection methods to focus on explaining variations around the average trajectories (since the unfolded $\mathbf{X}$ matrix is mean-centered) and also allows to readily incorporate productivity and quality data ($\mathbf{Y}$) into a multi-way PLS analysis.

A comparative study of various methods for analyzing 3-way batch process data (including projection methods) is found in Westerhuis *et al.* (1999). Discussions around other issues, such as unfolding, mean-centering, scaling, trajectory alignment and the use of the time variable in batch process data analysis are also found in the latter reference.

Multi-block data

Multi-block data bases are also frequently encountered in process data analysis. This type of data structure is illustrated in Figure 2.2 (b) and could arise when data are collected in different sections of the process (or in different successive stages). For example, $\mathbf{Z}$ could consist of measurements taken on the inital recipe (initial conditions) of a batch process, $\mathbf{X}$ could include the batch process trajectories (unfolded), and **Y** could characterize the resulting quality of each batch. One could also decide to block measurements that are similar in nature, such as several temperatures, pressures and flows. Multi-block algorithms have therefore been developed for analyzing the correlation structure among several data blocks (for example **Z** and **X**) using multi-block PCA. Correlation structure within these blocks and between a response block **Y** can also be analyzed using multi-block PLS, where both **Z** and **X** blocks would be predictor variables for **Y**.

Over the years, many multi-block (and hierarchical) PCA and PLS algorithms have been suggested to tackle various data analysis problems in chemistry, engineering and other disciplines. An exhaustive overview and a theoretical comparison of these different algorithms is found in Westerhuis *et al.* (1998). Among the theoretical findings published in this work, the equivalence between the standard PLS algorithm and the multi-block PLS algorithm proposed by Westerhuis and Coenegracht (1997) was demonstrated. It was shown that one could obtain the multi-block PLS results from standard PLS, when built using the same data sets, but with a particular scaling of the data. This result has been very useful in many instances to accomplish the work presented in this thesis.

2.2 Modelling Issues

Many issues exist in building empirical process models. However, this section focuses on two issues arising throughout the work of this thesis: the issues of scaling and of determining the number of components in projection methods. Some useful model diagnostic plots are also presented here.

2.2.1 Scaling

All projection methods decribed in section 2.1 are scale dependent and therefore, appropriate scaling needs to be performed on all measurements prior to analyzing them. When no prior process knowledge is available, one common practice is to scale all variables to unit variance, as this gives them equal importance in the model, with respect to one another. However, if prior knowledge exists, then scaling should be modified accordingly. For example, if it is common knowledge that a particular set of variables is roughly twice as important as another set, the most important set could be scaled to twice the variance of the less important set of variables.

Another scaling issue arises when multiple data blocks are used, each having a different number of variables, or when many similar variables are included in a block, but in different numbers (e.g. 10 temperatures, 5 flows and 2 pressures). When scaling all variables to unit variance in such situations, PCA or PLS models will focus more on the blocks of data having the most variables (since they exhibit more variance; 10 temperatures is 10 variance units while 5 flows is only five variance units). A way to re-establish a fair use of the variables in the model is to perform a block scaling, just after scaling all variables to unit variance. The block scaling consists of dividing each variable within a block by the square root of the number of variables in that block. The result of this is that each block of data has equal importance in the PCA or PLS models.

2.2.2 Number of Components

Another important issue in building empirical models with projection methods is to select the number of components to keep in the model in a meaningful way. The most widely used method for selecting the number of components in projection methods is cross-validation (Wold, 1978). This method suggests to keep adding latent variables to the model as long as they significantly improve the predictions of the model (PCA or PLS). Model predictive ability can be evaluated using the predictive multiple correlation coefficient $Q^2(a) = 1 - PRESS(a)/SS_r(a-1)$. $PRESS(a)$ is the total prediction error sum of squares obtained by cross-validating a model with a latent variables. This is performed by dividing the I observations included in the data base ($\mathbf{X}$ and/or $\mathbf{Y}$, see Figure 2.1 (b)) into g groups of size q ($I = gq$). Then, each group is deleted one at a time and a PCA or PLS model with a latent variables is built on the remaining $g-1$ groups. The prediction error sum of squares is then computed for the group not used to build the model. $PRESS(a)$ is the total of the prediction error sum of squares for all groups. $SS_r(a-1)$ is just the residual sum of squares of a model with $a-1$ latent variables. As long as Q^2 is greater than zero, the a^{th} dimension is improving the predictive power of the model. Therefore, one should keep adding latent variables until Q^2 is consistently lower than zero. Statistical hypothesis tests are sometimes used to verify if the a^{th} dimension has lead to a sufficient increase in Q^2 to be added to the model. Further details about this are provided in Chapter 4.

Another statistic that is often used for selecting the number of components is the fit multiple correlation coefficient $R^2(a)$, or alternatively, the explained variance for a model with a latent variables, $R^2(a) = 1 - SS_r(a)/SS_{tot}$. In this statistic, $SS_r(a)$ is the residual sum of squares of a model (PCA or PLS) with a latent variables, while SS_{tot} is the total sum of squares of the original data (before the model is built). As R^2 reaches values close to one, a very good fit of the data is obtained. The fit is poor when R^2 values are approaching zero.

Several other criteria have been developed for selecting an appropriate number of latent variables. A good overview of criteria available in the signal processing and chemometrics literature is found in Valle *et al.* (1999). A new criterion, useful for identifying process models that are closer to the true process, is also proposed in chapter 4 of this thesis.

2.2.3 Diagnostic Plots

Distance to the model plots and contribution plots are diagnostic plots frequently used in projection modelling methods. The former are used to verify how well each observation in the data set projects onto the reduced space of the model (plane or hyperplane). This is useful to identify outliers and the presence of a different correlation structure in specific observations. The latter plots provide the contribution of each variable to a move in the reduced space (projection) or to a change in the distance to the model. This is useful to gain insight from the data. Each of these plots is briefly discussed below.

Distance to the model plots

The distance of an observation (measurement on each variables) in the $\mathbf{X}$ space, $\mathbf{x}_i$, from the PCA or PLS model is given by the square prediction error of this observation, defined as $SPE_i = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^\top(\mathbf{x}_i - \hat{\mathbf{x}}_i)$, where $\hat{\mathbf{x}}_i$ is the projection of $\mathbf{x}_i$ onto the reduced space of the model (plane): $\hat{\mathbf{x}}_i = \mathbf{t}_i\mathbf{P}^\top$. SPE_i is therefore a measure of the perpendicular distance of the observation $\mathbf{x}_i$ from the plane defined by the PCA model with A latent variables. Note that we have chosen the distance in $\mathbf{X}$ space, but the equations are the same for the distance in the $\mathbf{Y}$ space, just by replacing $\mathbf{x}_i$ by $\mathbf{y}_i$ and $\hat{\mathbf{x}}_i = \mathbf{t}_i\mathbf{P}^\top$ by $\hat{\mathbf{y}}_i = \mathbf{t}_i\mathbf{Q}^\top$. When the distance is large, this indicates that observation $\mathbf{x}_i$ has a different correlation structure than what was normally seen in the historical data base used to build the PCA model, and so it is not well captured by this model. Statistical upper 95% or 99% limits on SPE are often used to determine if a particular observation should be removed from the model building procedure because of a too large distance to the model (outlier). Different approaches for defining these limits are discussed in Nomikos and MacGregor (1994*a*). One frequently used approach is to assume that SPE is approximately distributed as $g\chi^2(h)$, that is a multiple of a χ^2–distribution with h degrees of freedom. The two parameters g and h are estimated by matching the moments of the distribution, $E\left[g\chi^2(h)\right] = gh$ and $Var\left[g\chi^2(h)\right] = g^2(2h)$ using the historical data base. For the sample mean of the distribution, set $S\bar{P}E = m$, and for the sample variance of the distribution, set $\frac{1}{I-1}\sum_{i=1}^{I}(SPE_i - S\bar{P}E)^2 = v$. Solving for g and h, one obtains $g = (v/2m)$ and $h = 2m^2/v$. Therefore, SPE is distributed approximately as $\frac{v}{2m}\,\chi^2_{\alpha}(\frac{2m^2}{v})$, where $1-\alpha$ is the confidence level of the hypothesis test.

Another, but very similar measure of perpendicular distance to the model is the $DMOD$ statistic. This is just a normalized SPE. For the i^{th} observation in the $\mathbf{X}$ space, the distance is computed as $DMODX_i = \sqrt{SPE_i/(J-A)}$, where J is the number of variables in $\mathbf{X}$ and A is the number of latent variables. Note that $DMODY_i$ can be computed similarly. This measure of perpendicular distance to the model is shown here since it is used in the SIMCA-P 7.0 (Umetrics, 1998) software package, that has been often used throughout this thesis. To define an upper limit for $DMODX$, the following distance is used as a reference: $DMODX_o = \sqrt{\sum_{i=1}^{I} SPE_i/(I-A-1)(J-A)}$. The statistic $(DMODX_i/DMODX_o)^2$ is then assumed to be approximately F–distributed. One can therefore use this to define an hypothesis test to decide whether a particular observation has a too large distance to include in the model building procedure.

Contribution plots

Since the reduced space of PCA and PLS models (score space) is defined by linear combinations of the variables in $\mathbf{X}$, it is straightforward to identify the contribution of these variables to the difference between two specific score values. For example, the contribution of variable x_j to the difference between two values along t_1, Δt_1, is just $\Delta x_j\ p_{1,j}$ if PCA is used and $\Delta x_j\ w_{1,j}$ if PLS is used (note that x_j is the mean-centered and scaled value of the original measurement). When computing these contributions for each variable x_j, $j = 1, \ldots, J$, and plotting these contributions side by side, one obtains an insighful overview of what variables are strongly associated with the difference in score values. This information is then used in interpreting the modelling results. One could also obtain the overall average contribution per variable for a change in more than one score (Kourti and MacGregor, 1996). The contribution of each variable to SPE or $DMOD$ can also be computed in a similar fashion. Contribution plots are often readily incorporated in commercial software packages, such as SIMCA-P 7.0 (Umetrics, 1998).

Chapter 3

Multivariate Analysis and Optimization of Process Variable Trajectories for Batch Processes

3.1 Introduction

Batch and semi-batch processes are encountered everywhere in the processing industry and are usually used in the low volume production of high value products. Some examples of products obtained via batch operations include pharmaceuticals, biochemicals, some polymers and other specialty chemicals. In the development and optimization of such processes it is essential to understand the effect that process variable trajectories (histories) have on final product quality. In fermentation and polymerization, for instance, the batches often go through several stages of operation, each dominated by different physical and chemical phenomena, which may have a different influence on some aspects of product quality. For example, some process variables may have a strong impact on quality only during a particular stage, and none before or after. Other variables may have a consistent impact on quality, regardless of the degree of completion of the batch. It is therefore important to identify the sensitivity of final product quality to changes in the process variables at different degrees of completion within the course of a batch.

If a fundamental dynamic model of the batch process were available one could use it to optimize the final product quality through selection of process variable trajectories. Batch process optimization is well covered in the literature and a good overview is provided by Terwiesch *et al.* (1994). Past literature is mainly divided into off-line or *classical* optimization (Cawthon and Knaebel, 1989) and batch-to-batch optimization (Filippi-Bossy *et al.*, 1989; Zafiriou and Zhu, 1990; Dong *et al.*, 1996; Clarke-Pringle and MacGregor, 1998). The former requires very detailed mechanistic models and does not account for model mismatch, which renders the solutions suboptimal. On the other hand, batch-to-batch optimization was developed to overcome the difficulties with the classical approach, especially with regards to model mismatch. The main idea is essentially to use the results of the current batch run to either update the parameter estimates of a model or to use them as the forward integration step of the optimization procedure. Then, optimal process variable trajectories are computed for implementation in the next batch run. The optimization approaches are quite powerful, but most of the time require the use of mechanistic models, which are often not readily available.

Empirical approaches to batch trajectory analysis have been proposed. These involve the analysis of historical batch data. Bakshi and Stephanopoulos (1994*a*, 1994*b*) developed a formal methodology to analyze process signals to find temporal features relating to process performance. The methodology first involves filtering of the process signals, at different scales, using wavelet functions. Then, a special representation of the trends in terms of identifiable episodes is performed, and decision trees are built to relate these features to process performance. A fed batch fermentation case study is provided in Bakshi and Stephanopoulos (1994*b*). A total of 41 batches of normal operating data were available. Temporal features of 14 process trajectories were successfully identified for the classification of the fermentation yield into "bad", "okay" and "good" categories. This approach appears to be effective in feature extraction, especially qualitative features associated with trends or sequence of events.

Multivariate statistical methods have also been used very successfully for both the analysis of historical batch trajectory data and for the subsequent on-line monitoring of batch and semi-batch processes. Multi-Way Principal Component Analysis (MPCA) (Nomikos and MacGregor, 1994*a* and 1994*b*; Kourti *et al.*, 1996; Wold *et al.*, 1998), Projection to Latent Structures (MPLS) (Kourti *et al.*, 1996; Nomikos and MacGregor, 1995) and Multi-Block Multi-Way Projection to Latent Structures (MBPLS) (Kourti *et al.*, 1995) were used to model batch trajectory data and to identify operational problems. Multivariate SPC monitoring schemes were also proposed based on modelling the correlation structure of the data from the past "in-control" batch runs. Those monitoring techniques are applied to batch processes that are already developed and optimized. They are aimed at monitoring the process to ensure that the variables follow their priorly optimized trajectories as closely as possible.

This research is aimed at extending the use of Multi-Block Multi-Way PLS to the trajectory optimization stage. The term "optimization" is used in the sense of obtaining the sensitivities of final product quality to changes in the shape of process variable trajectories and subsequently using those for improving final quality. The proposed methodology is analogous to Response Surface Methods, although it is more complex since it involves optimizing the shape of the process variable trajectories instead of just the steady-state values of the variables. To make this optimization possible a wider range of variations in process trajectories than used for

SPC applications is required to compute the sensitivities. This may be obtained, for instance, by manipulating trajectories according to designed experiments or from multi-product processes, provided that trajectories corresponding to each product are different enough. The process understanding that can be gained from such an analysis is illustrated with several simulations of a Styrene-Butadiene Rubber (SBR) emulsion copolymerization.

Most applications of Multi-Block Multi-Way PLS for batch process analysis only consider the quality of the final product. However, process variables affecting quality only over a specific period of time are fairly difficult to identify with final quality measurements alone. The average effects over the entire batch history will tend to dominate the batch-to-batch variations in final quality. The collection of intermediate quality measurements obtained throughout the course of the batch should help to isolate local effects of variable trajectory changes. A new pathway algorithm, based on Multi-Block Multi-Way PLS has been developed to incorporate quality measurements collected during the course of a batch. The algorithm is applicable when the effects of changes in process variable trajectories on product quality can be assumed to be linear and additive. By allowing the use of all the intermediate quality measurements, the new algorithm can considerably reduce the number of runs necessary to identify time-varying relationships.

The contribution of this research is twofold. It illustrates a methodology to gain process understanding by analyzing how process variables affect product quality during the course of a batch. Second, it provides a new Multi-Block Multi-Way PLS algorithm that incorporates intermediate quality measurements. Its use has the potential for extension in other modelling applications where similar cause and effect paths prevail.

The chapter is structured as follows. The nature of semi-batch data is first presented, along with a discussion of the type of data required for the current modelling objectives. The new PLS pathway algorithm is then described. Selected simulations of the polymerization reactor are described followed by a discussion of the results.

3.2 Nature and Type of Batch Data

The data typically collected on a batch or semi-batch process can be classified into three blocks (**Z**, **X** and **Y**), as shown in Figure 3.1. The data included in each of these blocks has been discussed in the literature (Nomikos and MacGregor, 1994*a*, 1994*b* and 1995; Kourti *et al.*, 1995).

$\mathbf{Z}$ $(I \times L)$:

- Measurements taken on L variables, including initial charge recipe, type of catalyst, any process measurements taken before the batch starts and expected to have an influence on quality.

$\mathbf{X}$ $(I \times J \times K)$:

- Process variable trajectories. During a batch, measurements taken on J process variables, sampled K times, for I batches. This block of data includes manipulated and passively observed variables and any computed quantities from those measurements, such as instantaneous energy balances. Any trajectory feature that is expected to have an effect on quality should be included into that data block.

$\mathbf{Y_k}$ $(I \times M_k)$:

- Quality and productivity measurements collected at the end and at specified intermediate times (or degrees of completion) during the course of the batch. The number of quality variables per $\mathbf{Y_k}$ block, M_k, may differ as only a subset of the measured quality variables may be available during the course of the batch. For example, for a polymerization reactor, conversion may be available very frequently with an on-line densitometer, but average molecular weight may be available at the end only.

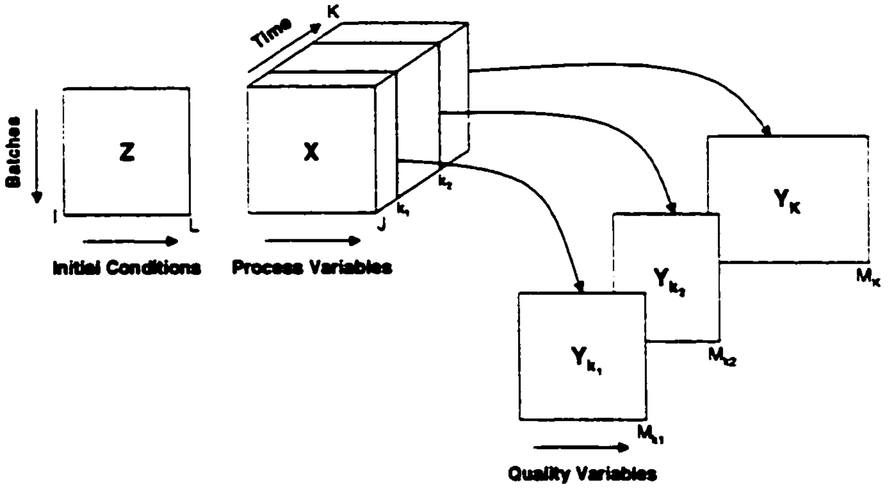


Figure 3.1: Nature of batch and semi-batch process data with intermediate quality measurements.

In this trajectory optimization problem one needs causal information and a much wider range of variation in the data than typically available from historical data. Superimposing designed experiments to nominal trajectories is preferred to ensure that the required type of data is obtained. The methodology presented in this paper is therefore well suited for pilot plants applications, where the goal is to look for operating trajectories capable of producing better quality materials. Data collected from existing multi-product plants may also cover a wide range of operating conditions, but would present higher risks of spurious results due to the more poorly designed nature of the data.

3.3 Modified Multi-Way Multi-Block PLS Algorithm

To analyze the data shown in Figure 3.1, they need to be reorganized to account for the causal pathway between the blocks. For simplicity, assume that for each batch run, 2 intermediate quality measurements are available, at time $k = k_1$ and $k = k_2$, in addition to the quality measurements collected at the end of the batch ($k = K$).

The three-way array, $\mathbf{X}$, can be dissociated in three parts, each corresponding to the period included in between intermediate quality measurements: from $k = 1$ to $k = k_1$, from $k = k_1 + 1$ to $k = k_2$ and finally from $k = k_2 + 1$ to $k = K$. Then, these arrays need to be unfolded in the time dimension (Nomikos and MacGregor, 1994*a*, 1994*b* and 1995; Kourti *et al.*, 1995; Westerhuis *et al.*, 1999) to analyse them with PLS. Denote the resulting 2-way matrices as $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_3}$ respectively. The complete set of blocks derived from the set of Figure 3.1 along with an indication of the causal paths between each block is provided in Figure 3.2.

Batch and semi-batch processes are integrating processes, such that any change in process variables made at a particular time during a batch will affect the behavior of the remainder of the batch. Therefore, $\mathbf{Y_{k_1}} = f(\mathbf{Z}, \mathbf{X_1})$, $\mathbf{Y_{k_2}} = f(\mathbf{Z}, \mathbf{X_1}, \mathbf{X_2})$ and $\mathbf{Y_K} = f(\mathbf{Z}, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3})$. The cause and effect path described in Figure 3.2 should be embedded into any model built on the batch data described above.

One approach to analyze the data from Figure 3.2 is to use only one $\mathbf{Y_k}$ block for each model as though each time interval between $\mathbf{Y}$ blocks were from a different process. Therefore, as many MBPLS models are built as there are intermediate and final quality measurement blocks as shown in Figure 3.3. The reader is referred to Westerhuis *et al.* (1998) for an overview of available MBPLS algorithms.

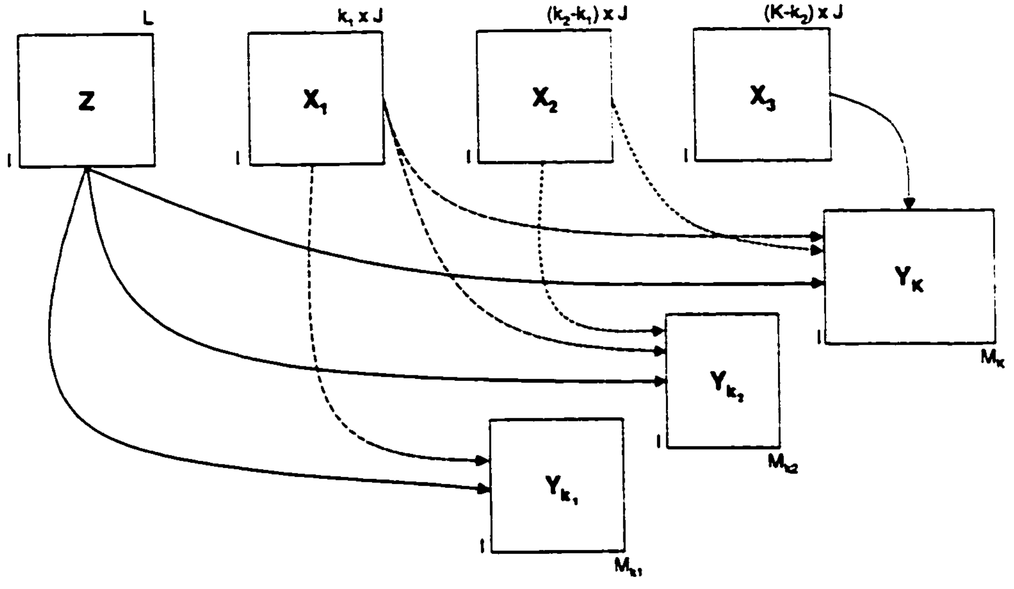


Figure 3.2: Causal pathway indicating the relationships between the recipe and each block of process variables to intermediate and final product quality.

The problem with this approach is that it does not account for the fact that all the data is from the same process, that is, successive $\mathbf{Y}_k$ blocks are highly related to previous ones, and that the $\mathbf{X}_k$ blocks in each row are the same $\mathbf{X}_k$ blocks. The integrating feature of the batch process has been neglected. Therefore, the weights ($\mathbf{w}$) corresponding to $\mathbf{X}_1$, for instance, are estimated three times in the case shown in Figure 3.3, obviously resulting in three different estimates. This does not provide a good idea of the "overall" effect that $\mathbf{X}_1$ has on product quality development. It would also be better to have only one model describing the impact of the process variable blocks on all the quality blocks.

To achieve a single model which respects the integrating nature of the batch

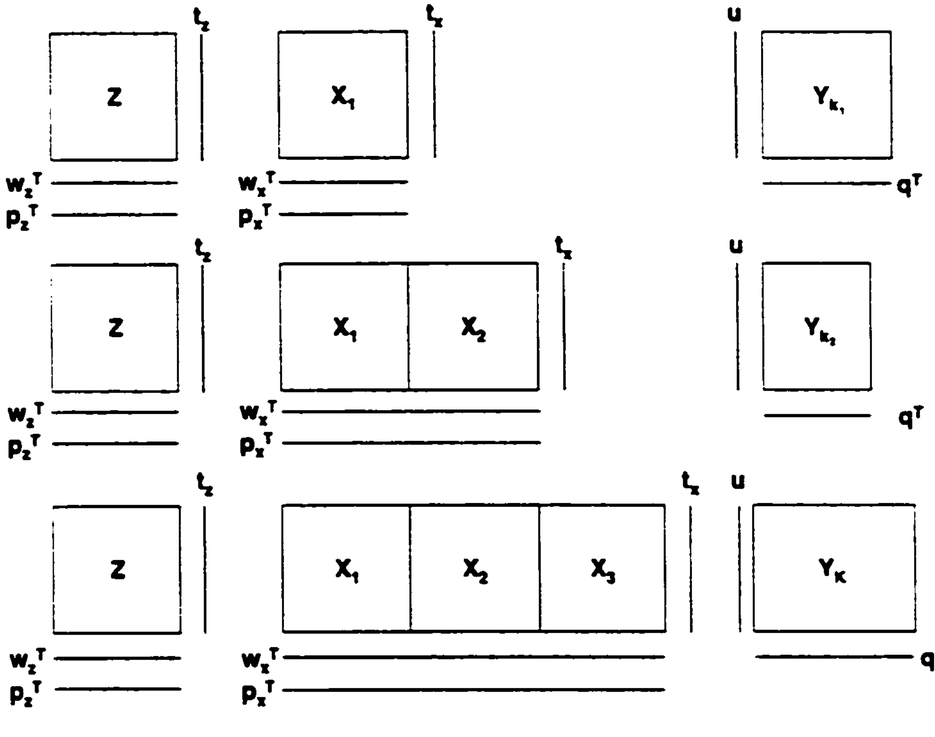


Figure 3.3: Separate Multi-Way Multi-Block PLS models, one for each block of intermediate quality measurements.

process, one could build jointly and simultaneously all MBPLS models shown in Figure 3.3, on single **Z**, **X** and **Y** blocks, arranged as in Figure 3.4. A single PLS model which relates the $\mathbf{Y}_k$'s at any time point to the prior $\mathbf{X}_k$'s and **Z** data can be estimated using the proposed pathway MBPLS algorithm given in Appendix A. It consists of some modifications to the commonly used NIPALS algorithm (Geladi and Kowalski, 1986). The causal path discussed in Figure 3.2 is now embedded into the resulting model as the new algorithm forces the estimates of **w**, **p** and **q** to be the same across all MBPLS models of Figure 3.3. However, constraining the values of **w**, **p** and **q** to be the same is valid only if the responses of quality variables to a change in a particular process variable are linear and additive (cumulative) and if the effects of $\mathbf{X}_k$ blocks in later operating periods do not depend upon the behavior in earlier periods. This is both the strength and the limitation of this new approach. Great advantages are obtained if the assumption is met, as a single model is obtained for the whole process, analyzing simultaneously and efficiently all the information contained in the measurements (**Z**, **X** and $\mathbf{Y_k}$ blocks). On the other hand, if the assumption is not valid the model can be poor. Examples of both situations will be presented as well as a procedure aimed at verifying if the pathway model structure is valid to analyze a particular set of data (test for linear and additive effects).

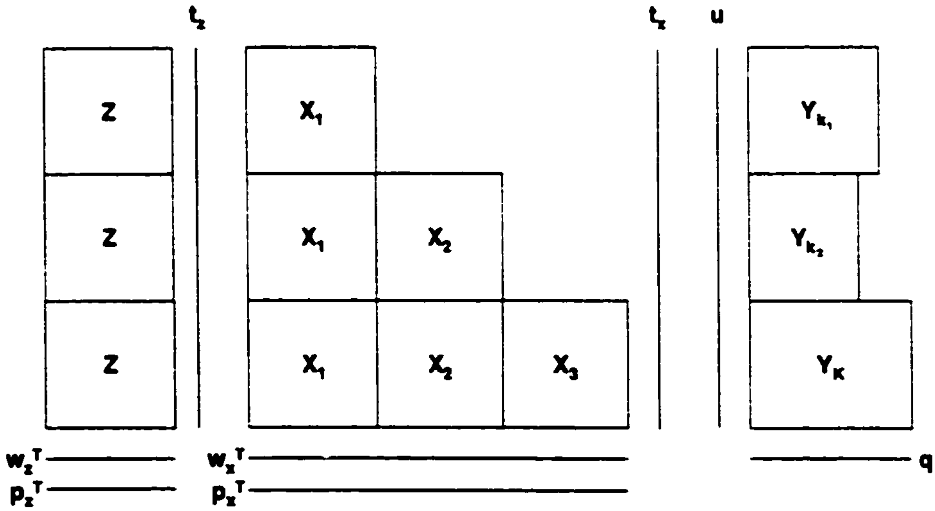


Figure 3.4: Modified Multi-Way Multi-Block PLS algorithm, incorporating intermediate quality measurements.

The proposed pathway model has several interesting features. It includes MBPLS as a special case, when quality is measured at the end of each batch only. It has potential for handling missing data as MBPLS does. The number of intermediate quality measurement blocks is not restricted. However, the quality measurements should be collected at corresponding times or degrees of completion for each batch. It accounts for situations in which only a subset of the quality variables measured at the end of a batch is available during the course of the batch. If the number of quality variables measured during the batch is always the same as at the end, then the new algorithm becomes simpler as shown in Appendix A. Finally, the algorithm could be extended to other problems sharing the same causal pathway. Steady-state modelling of several units in series from a continuous process is one possible application.

3.4 Simulation of a Styrene-Butadiene Emulsion Copolymerization

To illustrate the concepts discussed in this paper, simulations based on a detailed fundamental model (Broadhead, 1984; Broadhead *et al.*, 1985) of a reactor carrying out the free radical emulsion copolymerization of styrene and butadiene was used. This reactive system is known to consist of three different stages (Hamielec and Tobita, 1992), within which different physical phenomena take place. This process is therefore a good candidate for the application discussed in this chapter.

A schematic representation of the reactor is shown in Figure 3.5. The normal operating procedure for such a reactor starts with charging an initial recipe of different materials, as shown in Table 3.1. Then, the reaction mixture is heated up to its reaction temperature using hot water or steam flowing through the jacket. For all cases in this chapter, the temperature is controlled at 50°C. Once the reaction temperature is reached monomers, initiator and chain transfer agent start being fed according to some feed policies until the end of the batch. The batch duration is assumed to be 480 minutes for all cases studied.

Three process variable trajectories can be manipulated and are used in the analysis: the total monomer flow rate (F_M), the initiator flow rate (F_I) and the chain

Table 3.1: SBR recipes for initial charge.

Species (g)	Recipe 1	Recipe 2
Styrene	249.2	0.0
Butadiene	450.3	0.0
Water	5000.0	5000.0
Initiator	50.0	0.0
Soap	24.0	24.0
CTA	2.0	2.0

transfer agent flow rate (F_{CTA}). The flow rate of styrene and butadiene separately were always kept in the same ratio, such that only the sum of the two (F_M) is manipulated. It is further assumed that weight average molecular weight ($\bar{M}_w$), tri- and tetra-functional long chain branching ($\bar{B}_{n3}$ and $\bar{B}_{n4}$), total mass conversion (x) and polymer particle number and average diameter ($\bar{N}_p$ and $\bar{d}_p$) are measurable and are appropriate for characterizing polymer quality.

Designed experiments were added to nominal operating conditions in order to generate a wide range of variations in process trajectories, while maintaining initial charge recipes constant. The **Z** matrix therefore does not need to be included in the analysis for the cases treated in this chapter. Nominal conditions were set constant at $F_M = 5.67$, $F_I = 0.45$ and $F_{CTA} = 0.014$ g/min for the entire batch duration. The type and frequency of variations correspond to step changes of ±20% from nominal values, implemented every 40 minutes. Direction of the changes were decided according to a 2^3 factorial design, with the first 4 experiments repeated to give 12 changes per batch. The design is shown in Table 3.2. The frequency of changes should be chosen fast enough to provide a high resolution to identify how process trajectories affect quality. However, changes should also be maintained long enough for the effects to be observable in the measured quality. Periods of 40 minutes were considered adequate for the reacting system under study.

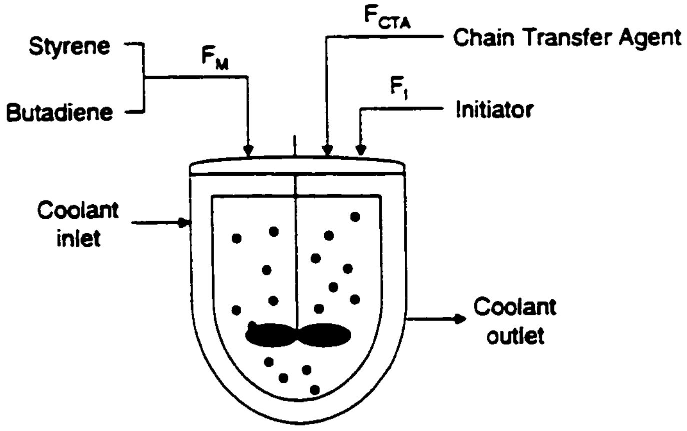


Figure 3.5: Reactor for semi-batch emulsion copolymerization of styrene-butadiene rubber.

In general, if prior knowledge were available about the number of stages, their sequence and duration, one could build a more appropriate design in which variations would be focused on each stage specifically. In that situation, a smaller number of changes and a smaller number of batch runs could be used for extracting the desired information.

For each case study presented in this chapter, 31 batch runs were generated by randomizing the order of appearance of each experiment of the design (rows in Table 3.2). An example of variations implemented on process variables for one such experiment is provided in Figure 3.6. These variations are designed to gather information about effects on product quality, taking place during specific periods of time

Table 3.2: Factorial design (2^3) with the first 4 experiments repeated.

Duration (min)	F_M	F_I	F_{CTA}
0-40	-	-	-
40-80	-	+	-
80-120	+	-	-
120-160	+	+	-
160-200	-	-	+
200-240	-	+	+
240-280	+	-	+
280-320	+	–	+
320-360	-	-	-
360-400	-	+	-
400-440	+	-	-
440-480	+	+	-

during a batch, or "time specific" effects. However, all 31 trajectories for each process variable have about the same average value. This means that these experiments do not generate information about effects on quality sustained throughout tne batch duration ("average" or "cumulative" effects). Varying the average of trajectories is also necessary and hence, 9 additional runs with constant flow rates were generated by designing the average trajectories using another 2^3 factorial design with a center point. In these additional runs, the values of process variables are maintained constant for the entire batch duration, as the nominal conditions, but were varied by ± 20% of those nominal values according to the design. The design center point consists of the nominal conditions. In summary, 40 batch runs were generated for each case study and those data bases contain the necessary information to identify both "time specific" and "average" effects of process variables on product quality.

Some random noise was added on each quality measurements. It was selected to be approximately 10% of the range of variation in each quality measurement.

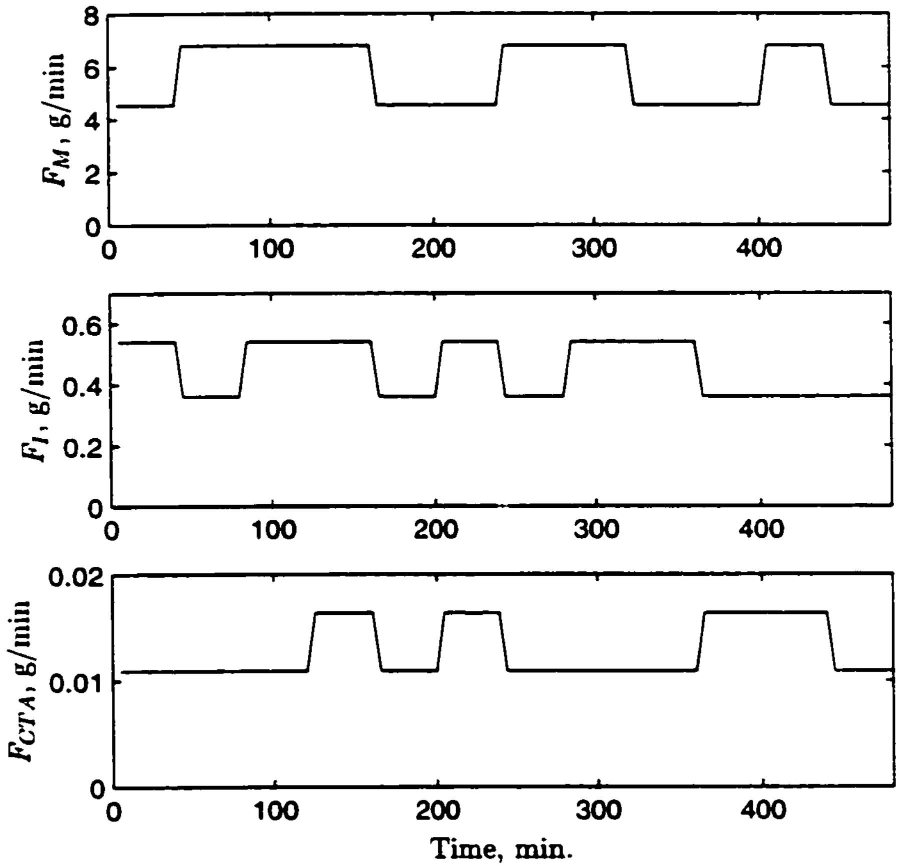


Figure 3.6: Designed variations on F_M, F_I and F_{CTA} for one batch, obtained by randomizing the order of appearance of each experiments from the 2^3 factorial design with the first 4 experiments repeated (Table 3.2).

3.5 Analysis and Discussion

3.5.1 Advantages for using the pathway model

A new criterion for selecting the number of latent variables is used, as described in chapter 4, which is more appropriate than cross-validation when trying to obtain good estimates of the model parameters. It essentially suggests to continue adding latent variables as long as the model parameter estimates are stable, and to stop when overfitting occurs. The model stability is evaluated using the delete-1 jackknife statistics, which also provide estimates of the uncertainty in the parameter estimates.

Modelling all quality variables (**Y**) in only one model (e.g. PLS2 vs PLS1) is not always an optimal choice. For instance, one could observe latent variables alternating between explaining each **Y** variable separately or alternating between explaining groups of **Y** variables. This may happen when those **Y** variables or groups of variables are not spanning the same spaces. This could easily be detected by looking at the explained variance (R_y^2) of each **Y** variable in each latent dimension, when all **Y** variables are modelled together. It therefore makes sense to block **Y** variables appearing to span similar spaces together. In this way a reduction in the number of latent variables for each model and better model performance are expected. In the case studies shown in this chapter, $\bar{M}_w$ and $\bar{N}_p$ were each blocked in a separate model and the other quality variables, $\bar{B}_{n3}$, $\bar{B}_{n4}$, x and $\bar{d}_p$ were blocked into a third model.

In this first study, initial charge recipe 1 from Table 3.1 is used and 40 batches designed as described in section 3.4 were run. The **X** array consists of the values of the three flow rates (F_M, F_I and F_{CTA}) for the 12 time periods and for the 40 batch runs. It is assumed that the pathway algorithm is valid for analyzing this database. A procedure to assess the validity of this new algorithm is presented in the next section. Note that in all case studies, mean-centering and scaling to unit variance has been applied to each of the $\mathbf{X_k}$ and $\mathbf{Y_k}$ data blocks prior to analysis. An overview of

Table 3.3: Overview of the pathway PLS models with 40 batch runs by means of cumulative sum of squares explained in **X** and **Y** blocks and the cumulative sum of squares predicted of **Y** after the specified number of dimensions. Model 1, 2 and 3 correspond to the following **Y** variable blocks: $\bar{M}_w$, $[\bar{B}_{n3}\ \bar{B}_{n4}\ x\ \bar{d}_p]$ and $\bar{N}_p$.

Model	No. PLS dim.	$R^2_{x,cum}$ (%)	$R^2_{y,cum}$ (%)	Q^2_{cum} (%)
1	6	38.93	96.38	89.27
2	8	50.81	88.42	78.75
3	9	50.25	90.03	85.92

the model performance is provided in Table 3.3. It shows three cumulative multiple correlation coefficients, $R^2_{x,cum}$, $R^2_{y,cum}$ and Q^2_{cum}, computed over a specified number of PLS components. The cumulative R^2 values give the percentages of the total sum of squares of **X** and **Y** that are explained by the fitted PLS models with the indicated number of dimensions. The cumulative Q^2 value is the percentage of the total sum of squares of **Y** that can be predicted with these models using a leave-one-out cross-validation procedure. Regression coefficient estimates for five of the six polymer properties are shown in Figure 3.7. Results for tri-functional branching are not shown, since they are essentially the same as those of tetra-functional branching. Each plot contains 36 regression coefficients, 12 for each of the three process variables. The first 12 coefficients provide the impact that F_M has on the final property of interest, over each of the 12 sequential periods of 40 minutes during a batch. Subsequently, one finds the 12 coefficients for the impact of F_I and F_{CTA}. In each plot, the coefficients corresponding to F_M, F_I and F_{CTA} are all separated with a dashed vertical line. Also shown are the one standard error estimates for each parameter obtained from the jackknifing procedure (refer to chapter 4 of this thesis).

Cumulative effects of process variables on final quality are characterized by large coefficients, which consistently have the same sign over the entire batch duration.

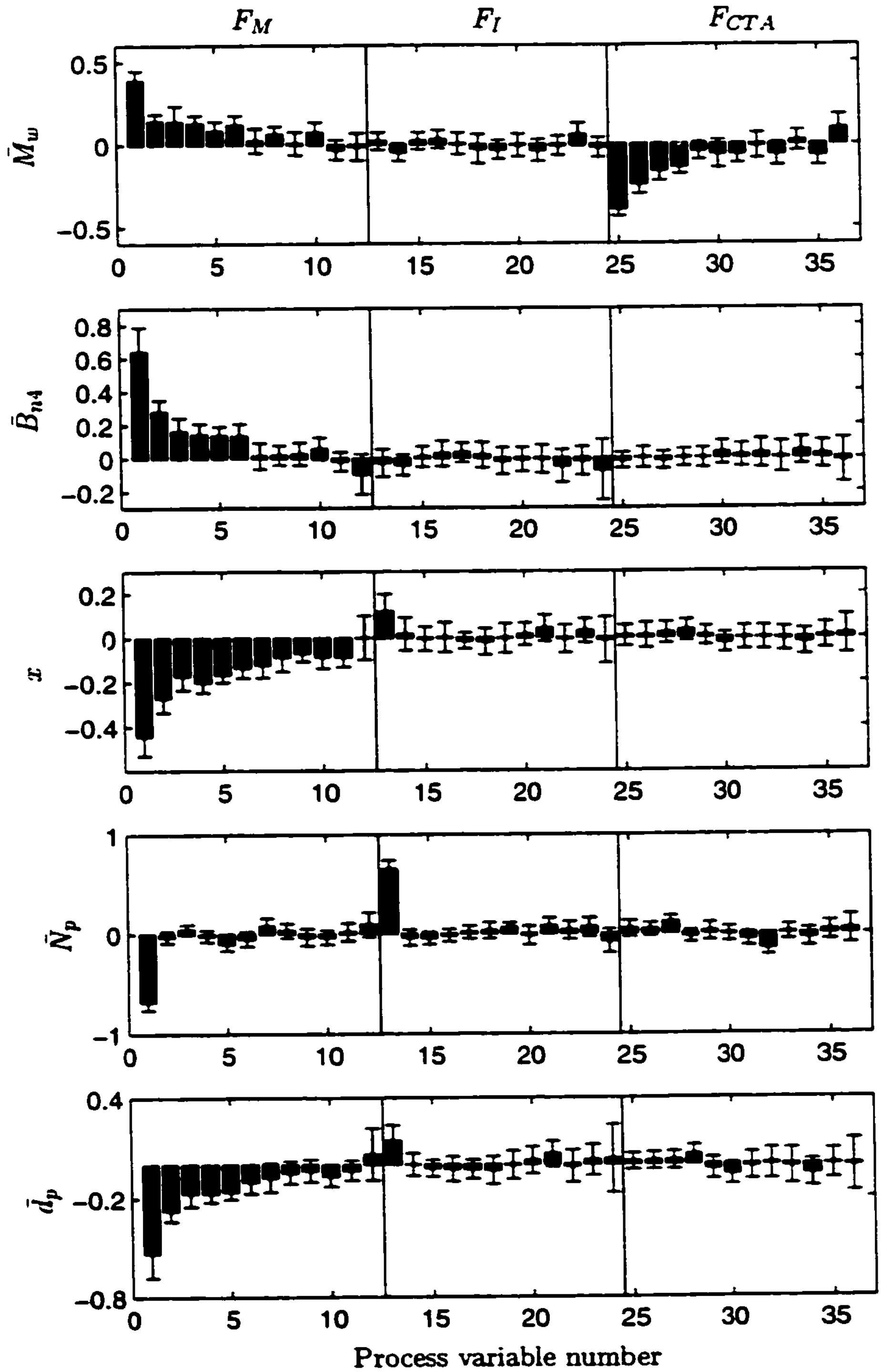


Figure 3.7: Regression results with the pathway algorithm, using 40 batch runs.

The effect of F_{CTA} on $\bar{M}_w$ and the effect of F_M on $\bar{M}_w$, $\bar{B}_{n4}$, x and $\bar{d}_p$ are clearly identified as cumulative effects. For example, it could be interpreted that the total amount of chain transfer agent and monomers fed during the first half of the batch seem to be responsible for molecular weight development. Also notice the exponential decay of those effects, saying that the same amount of material fed later in the batch has less impact on final quality. This is expected since it has less time to participate in reactions. On the other hand, there is obviously something different happening during the first 40 minutes of the batch as the corresponding regression coefficients are much stronger for all quality variables. In particular, it seems that the monomer and initiator feedrates F_M and F_I only affect the number of polymer particles ($\bar{N}_p$) during this first period within each batch and have no effect beyond this first period. This illustrates a "time specific" effect, where some aspects of quality can only be modified during a specific period of time during a batch.

The results in Figure 3.7 can be used to optimize the process to achieve modified quality variables. To obtain a higher number of polymer particles, a low flow rate of monomer and a high flow rate of initiator should be used *early* in the batch, since $\bar{N}_p$ can not be modified beyond about 40 minutes. This will also lead to a low $\bar{B}_{n4}$ and $\bar{M}_w$ and to a high x and $\bar{d}_p$ during that period. However, these can be modified in the remainder of the batch by appropriately selecting F_{CTA} and F_M respectively. The conclusions drawn from these empirical results are corroborated by polymerization principles. For example, polymer particles are only generated early in the batch, during a "nucleation" period and their number remains constant after this period is over (after about 10% conversion). A low F_M extends the nucleation period and a high F_I increases the rate of particle nucleation, both resulting in an increased $\bar{N}_p$. On the other hand, total mass conversion (x) is the ratio of the total mass of polymer to the sum of total mass of monomers and polymer at any time in the reactor. Increasing F_M therefore leads to operation at lower conversion levels and vice versa.

To illustrate the gain that is obtained by collecting intermediate quality measurements, the results of Figure 3.7 are compared with those obtained in Figure 3.8 using the same 40 batch runs but with only the final quality measurements available. Similar results are obtained for cumulative effects since variations in $\mathbf{Y}$ are dominated by them. However, exponential decay effects are not as appararent and the dependence of the number of polymer particles ($\bar{N}_p$) on F_M and F_I only during the first period is not as evident. This shows the advantage of using intermediate quality measurements as the "time specific" effects are more easily identified.

Another advantage of using intermediate quality measurements is the additional information brought into the model. One would therefore expect to be able to reduce the number of batch runs over that when only the final quality is available and still extract similar trajectory. To illustrate this, a subset of 16 runs were selected from the set of 40 described in section 3.4. Four runs with constant trajectories for the entire batch duration were used to identify average effects. The remaining 12 runs were randomly selected among the 31 runs on which designed variations around nominal conditions have been implemented. When this reduced data set is analyzed with the pathway algorithm, the regression results shown in Figure 3.9 are obtained. Although only 16 runs were used (highly fractionated design), the results were very similar to the case showed in Figure 3.8 with 40 runs and only the final quality measured.

3.5.2 Validation of the assumption behind the pathway model

The assumption behind the pathway algorithm (Figure 3.4) is that the effects of the $\mathbf{X_k}$ blocks are linear and additive. In other words the weights and loadings ($\mathbf{w}$'s and $\mathbf{p}$'s) associated with each $\mathbf{X_k}$ block are constant and independent of what happened in earlier blocks. In this section we test the validity of this assumption by building a

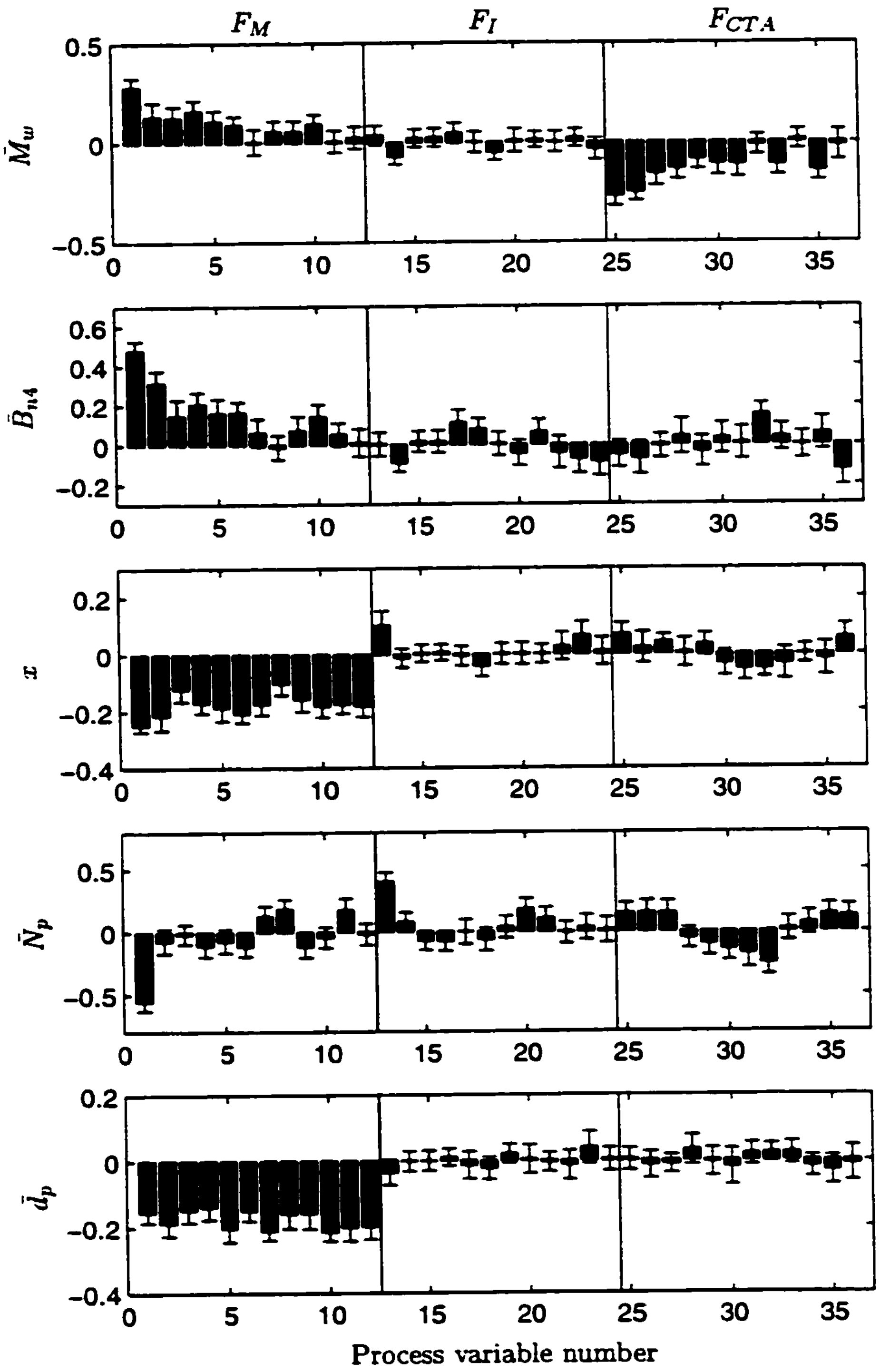


Figure 3.8: Regression results using 40 batch runs, when only final quality is available.

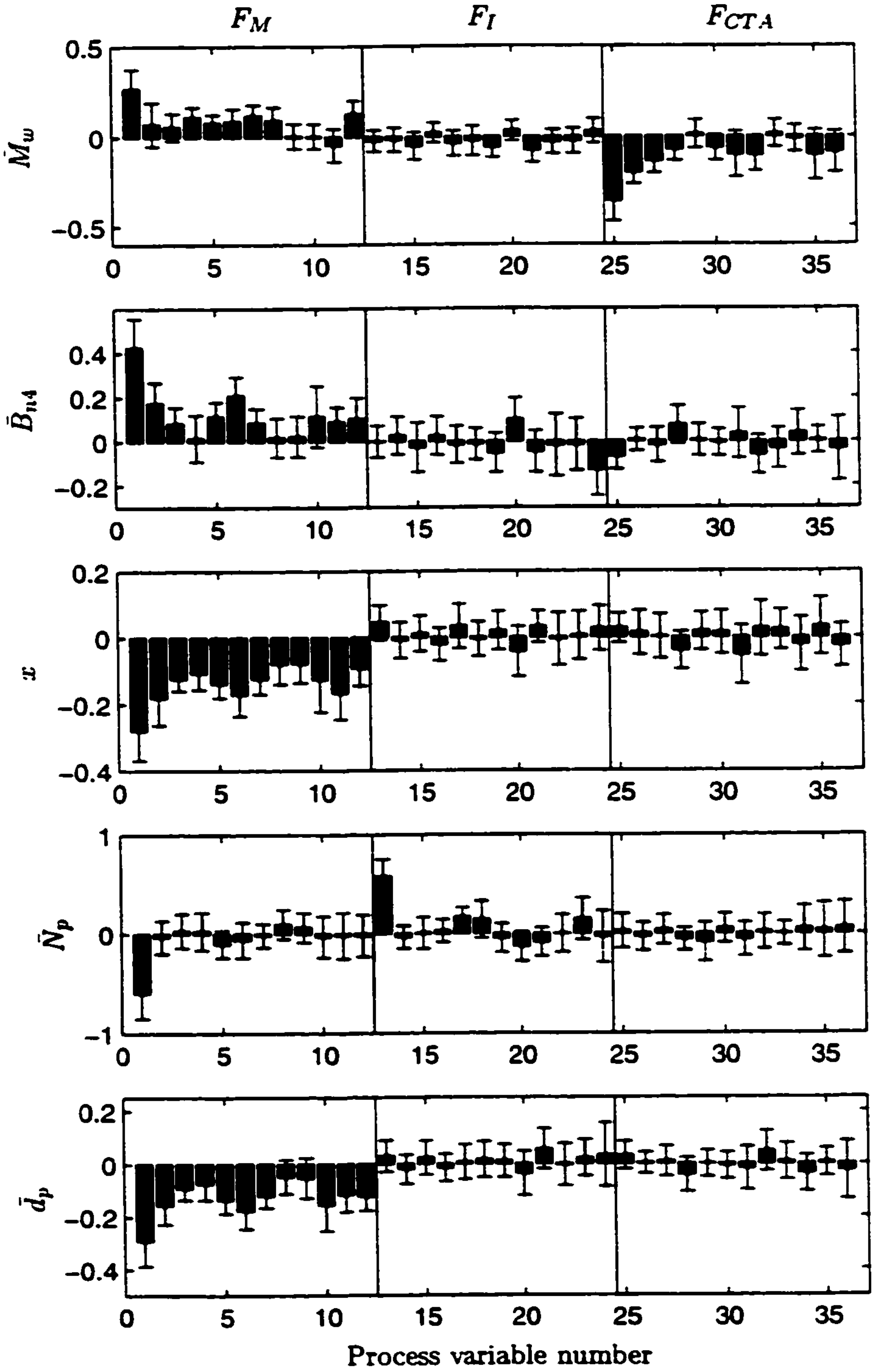


Figure 3.9: Regression results with the pathway algorithm, using 16 batch runs.

series of MBPLS models, one for each intermediate quality measurement block ($\mathbf{Y_k}$) as in Figure 3.3. If the assumption is valid then the estimated effects for each block in each of the multi-block models should stay nearly constant. To illustrate this, the case study of section 3.5.1 with 40 batch runs is used. For simplicity, consider the effects that F_M and F_I have on $\bar{B}_{n4}$ during the first 40 minutes of the batch (first regression coefficient for F_M and F_I). These effects are estimated 12 different times since they are involved in each MBPLS model. Their 12 estimates are plotted sequentially in Figure 3.10 along with their estimated standard error. The estimates of the effects of both of these variables on $\bar{B}_{n4}$ for the first period appear to be resonably consistent for the models built at the 12 different periods. Similar results were found using other response variables and different periods. This implies that the assumption behind the pathway model is not unreasonable in this case.

On the other hand, when batch data is generated using initial charge recipe 2 (Table 3.1) and using the same process variable trajectories as the previous study, much different results are obtained. Figure 3.11 shows again the estimates of the effect that F_M and F_I have on $\bar{B}_{n4}$ during the first 40 minutes, as estimated from the 12 MBPLS models built on this new data set. One can clearly see a consistent change in the estimates of these effects based on the $\mathbf{Y_k}$ data available at the different times. This suggests that this data should not be analyzed as a whole with the new pathway algorithm and therefore the only valid approach is to build multiple MBPLS models as in Figure 3.3, at least for the $\bar{B}_{n4}$ response.

Another but "overall" measure for validation of the pathway model structure is the total prediction error sum of squares $PRESS_T$. Table 3.4 shows the $PRESS$ for each quality variable, computed for all samples during the course of the each batch, for the pathway and the local models built on the data generated with recipe 1 and 2. The sum of these $PRESS$ values provide $PRESS_T$. When the pathway model structure is valid, the $PRESS_T$ is expected to be similar for both model structure,

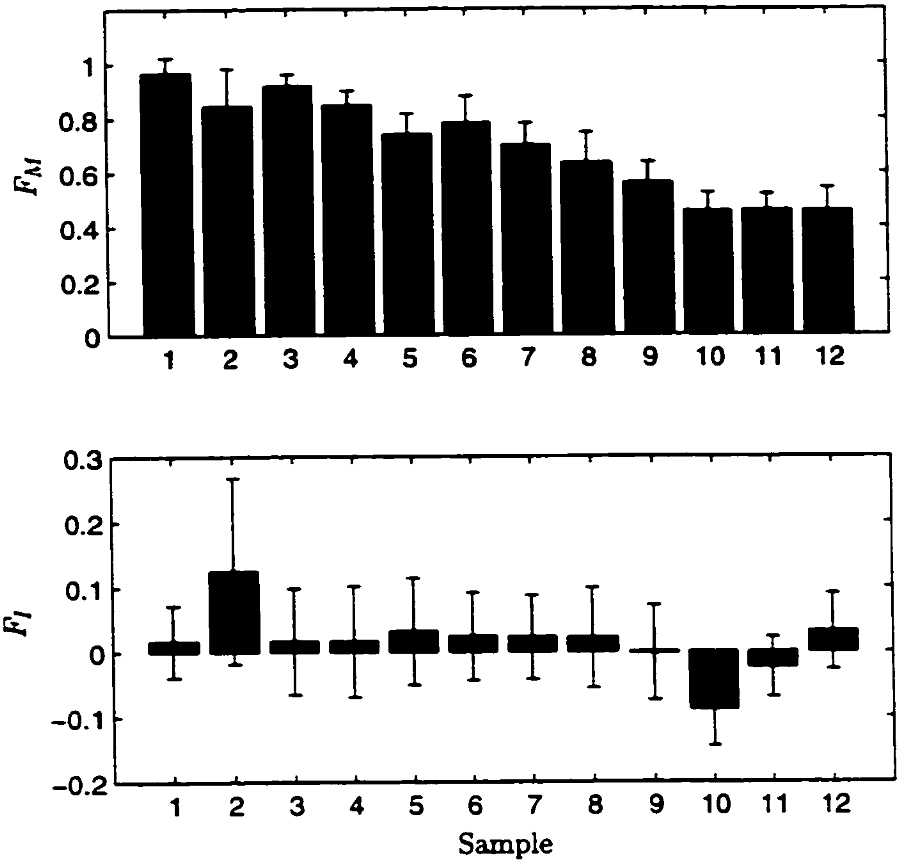


Figure 3.10: Effects of flow rates of monomers and initiator on tetrafunctional branching level during nucleation, estimated using local MBPLS models, when initial charge recipe 1 is used.

although it would always be smaller for the local models. However, if the pathway structure is not valid, then the $PRESS_T$ corresponding to the pathway model should have a much higher discrepancy with the value obtained for the local models. This situation is clearly shown in Table 3.4 where the $PRESS_T$ is a factor of 2 higher when using recipe 2, while it is only 35% higher with recipe 1.

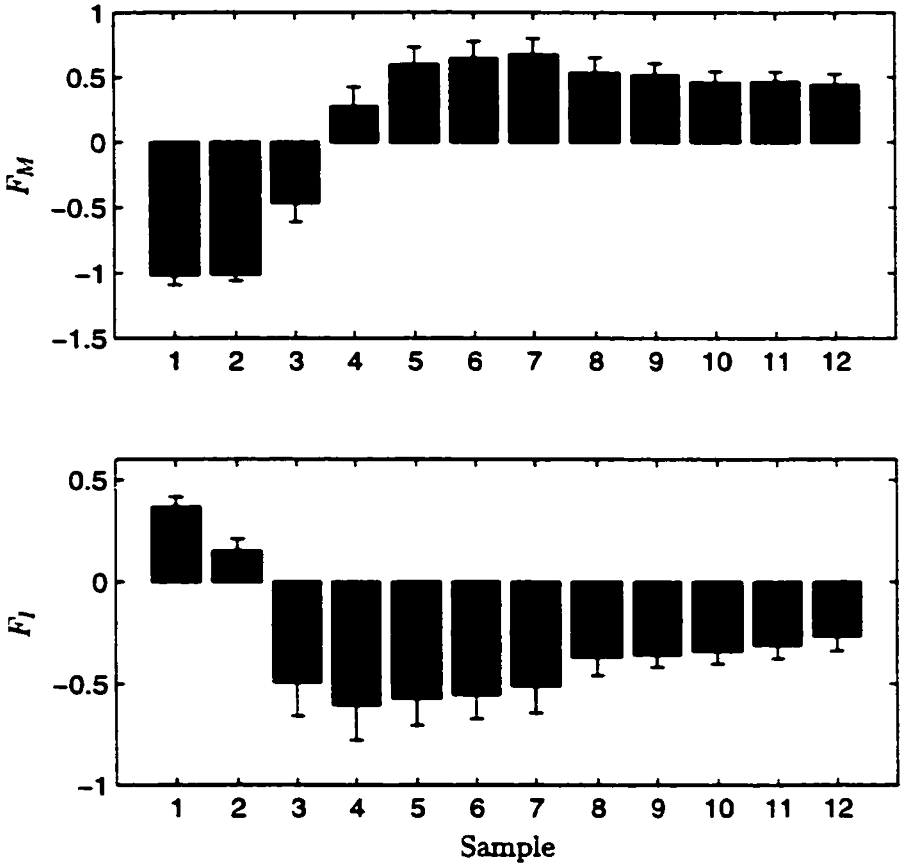


Figure 3.11: Effects of flow rates of monomers and initiator on tetra-functional branching level during nucleation, as estimated using local MBPLS models, when initial charge recipe 2 is used.

3.6 Conclusion

This chapter has developed an empirical methodology using designed experiments and analyses based on multi-block PLS algorithms for batch and semi-batch process improvement and optimization. The procedure is based on identifying the sensitivity of final product quality to the shape of process variable trajectories (histories). Understanding how and when during the course of a batch process variables have an impact on quality is essential for suggesting modifications to operating policies that

Table 3.4: Comparison of prediction error sum of squares per quality variable and total prediction error sum of squares for the pathway model and local PLS model structure, built using 40 runs and generated with recipe 1 and 2.

Recipe	Model Structure	M_w	B_{n4}	x	N_p	d_p	$PRESS_T$
1	local	30.81	48.44	23.00	78.35	91.73	272.33
	pathway	50.22	56.55	81.10	65.87	113.62	367.36
2	local	26.91	140.70	46.74	31.78	47.80	293.93
	pathway	131.04	332.90	71.43	23.10	45.62	604.09

may result in improved quality. Two types of effects were identified, namely cumulative and time-specific effects. Cumulative effects have an impact on product quality due to the accumulation of some species over the entire batch duration. On the other hand time-specific effects are those where quality is modified but only over a particular period of time during a batch. They are usually associated with stages of operation in which different physical phenomena dominate. Identification of trajectory features affecting quality was illustrated using a simulation study of styrene-butadiene rubber emulsion copolymerization. Useful insight about the process was gained through the proposed analysis.

A new pathway PLS algorithm was proposed, which allows one to make use of intermediate quality measurements. It consists of a modified multi-way multi-block PLS algorithm in which the pathway relationships between process variables and intermediate product quality is taken into account. The pathway model structure is valid under the assumption of linear additive effects between the process variables and quality. A procedure to verify this assumption was also provided.

Chapter 4

Jackknife and Bootstrap Methods in the Identification of Dynamic Models

4.1 Introduction

Non-parsimonious model structures such as finite impulse response (FIR) models and autoregressive with exogenous variables (ARX) models are often used in dynamic model identification. They provide a lot of flexibility to capture complex industrial process behavior while requiring a minimal number of structural choices. They are also readily incorporated in some multivariable controllers (e.g. DMC).

However, parameter estimation of non-parsimonious models is often ill-conditioned, due to the large number of lagged input variables used as predictors. In that situation, it is well known that least squares and minimum prediction error estimates are sensitive to correlation among the predictor variables and can lead to poor results. Regularization methods (such as ridge regression) and Latent Variable methods (Principal Component Regression, Partial Least Squares and Canonical Variates) have therefore been proposed as alternatives to multiple regression for parameter estimation of FIR and ARX models (Ricker, 1988; MacGregor *et al.*, 1991; Wise and

Ricker, 1992 and 1993; Dayal and MacGregor, 1996 and 1997; Shi and MacGregor, 2000). These methods can achieve lower mean square error in ill-conditioned problems by obtaining a reduced variance in the parameter estimates at the expense of a slight bias. To reach any given compromise, both regularized least squares regression and latent variable regression make use of a *meta parameter*, the ridge parameter and the number of latent variables respectively. The meta parameter is embedded into the regression procedure and needs to be selected. This introduces the model selection problem that is the motivation of this work.

Cross-validation (Wold, 1978) has been a commonly used technique for the selection of regularized least squares and latent variable models. It is based on maximizing the model predictive ability. Such a criterion for model selection is appropriate when prediction is the motivation for model building, as it is in inferential sensor development. However, the objective of process identification is to build models for purposes such as control system design and simulation. These models are usually inverted in designing controllers. Not only should the model provide good predictions, but most importantly it should also capture the correct process structure (e.g. dead-time). Ricker (1988) and MacGregor *et al.* (1991) showed that PCR and PLS models selected using cross-validation performed poorly in determining the process dead-time since too few latent variables were kept in the model. They concluded that adding extra latent variables is necessary. The importance of these extra latent variables (associated with smaller eigenvalues) when inverting the FIR model was demonstrated by Dayal and MacGregor (1996), who investigated the robust stability and control performance of FIR models identified with several different methods.

To illustrate the problem of using cross-validation in identification, consider a single-input single-output first order system with four periods of dead-time, corrupted by a random white noise. In this example, PLS is used to estimate the parameters (Wold, 1978; Geladi and Kowalski, 1986). Throughout this work, the SIMCA-P 7.0

(Umetrics, 1998) software package has been used to build PLS models selected using cross-validation. For each new latent variable added to the model ($a = 1, 2, \ldots, A$), a statistical hypothesis test is performed to verify if the predictive multiple correlation coefficient, $Q^2 = 1 - PRESS(a)/SS_r(a-1)$, has been increased sufficiently to include the new dimension a. $PRESS(a)$ is the prediction error sum of squares obtained by cross-validation for the a^{th} latent variable. The $PRESS$ is calculated by dividing the n observations into g groups of size q ($n = gq$). Then, each group is deleted one at a time, and a model is built using the remaining $(g - 1)$ groups. The prediction error sum of squares is computed for the group not used to build the model. The process is repeated leaving each group out once and only once. The $PRESS$ is then the total of the prediction error sum of squares for all groups. $SS_r(a - 1)$ is the sum of squares of the residuals of the PLS model containing $a - 1$ latent variables. Therefore Q^2 is a measure of the percent of the variance remaining after $(a - 1)$ latent variables that can be predicted by adding the a^{th} latent variable. Since Q^2 is not truly F−distributed, the test is based on a reference distribution obtained from simulation studies. The reader is referred to Ståhle and Wold (1987) and Wakeling and Morris (1993) for more details.

Two latent variables were found significant by cross-validation, when identifying the first order system. The results are shown in Figure 4.1. The dependence of the explained variance of the input and output signals (R_x^2 and R_y^2) and the dependence of the output prediction error sum of squares ($PRESS$) on the number of latent variables (a) is shown in Figure 4.1 a) - c). The PLS model with two latent variables has captured enough information to explain most of the variance of the output and to predict it fairly well. However, it used only a small amount of information about the input. Adding more latent variables leads to models having approximately the same explained variance and predictive power for the output (R_y^2 and $PRESS$), but it uses more information about the input, as R_x^2 increases. This means that several linear combinations of the lagged inputs provide equally good models for prediction. This is mainly attributable to correlation among predictor variables. It is therefore difficult to assess if the right process structure has been captured by the model when using a prediction criterion.

Figure 4.1 d) and e) show the true and estimated impulse and step responses of the first order system with two PLS dimensions. Clearly, the model based on two latent variables does not provide satisfactory estimates of the dead-time and dynamic response and when used for control, might lead to poor performance and robustness properties. The reason for poorly estimated dead-time with two latent variables is shown in Figure 4.1 f). It presents a comparison between the true input sequence and its explained value by the PLS model. The explained input sequence behaves as if the true input was highly filtered. The degree of filtering depends on the number of dimensions kept in the model. See Wise and Ricker (1992) for further details. The fact that the square edges of the true input sequence are not modelled is mainly responsible for the poor dead-time estimation. Substantial improvement of the FIR model structure could be achieved by adding more latent variables, which would use more information about the input sequence. In this example, cross-validation clearly leads to choosing too few latent variables to capture the FIR model structure.

The present work is concerned with developing a more appropriate objective criterion for selecting the meta parameter of latent variable regression methods, when the objective is to capture the structure of a process model. This is the case in dynamic model identification for process control. It is assumed that the model structure (FIR or ARX) is already chosen and is valid with respect to both the process and the type of disturbance affecting it. The proposed criterion is illustrated using PLS as the parameter estimation method, but it is not limited to PLS and can be applied to any latent variable method or any regularized least squares method. It is based

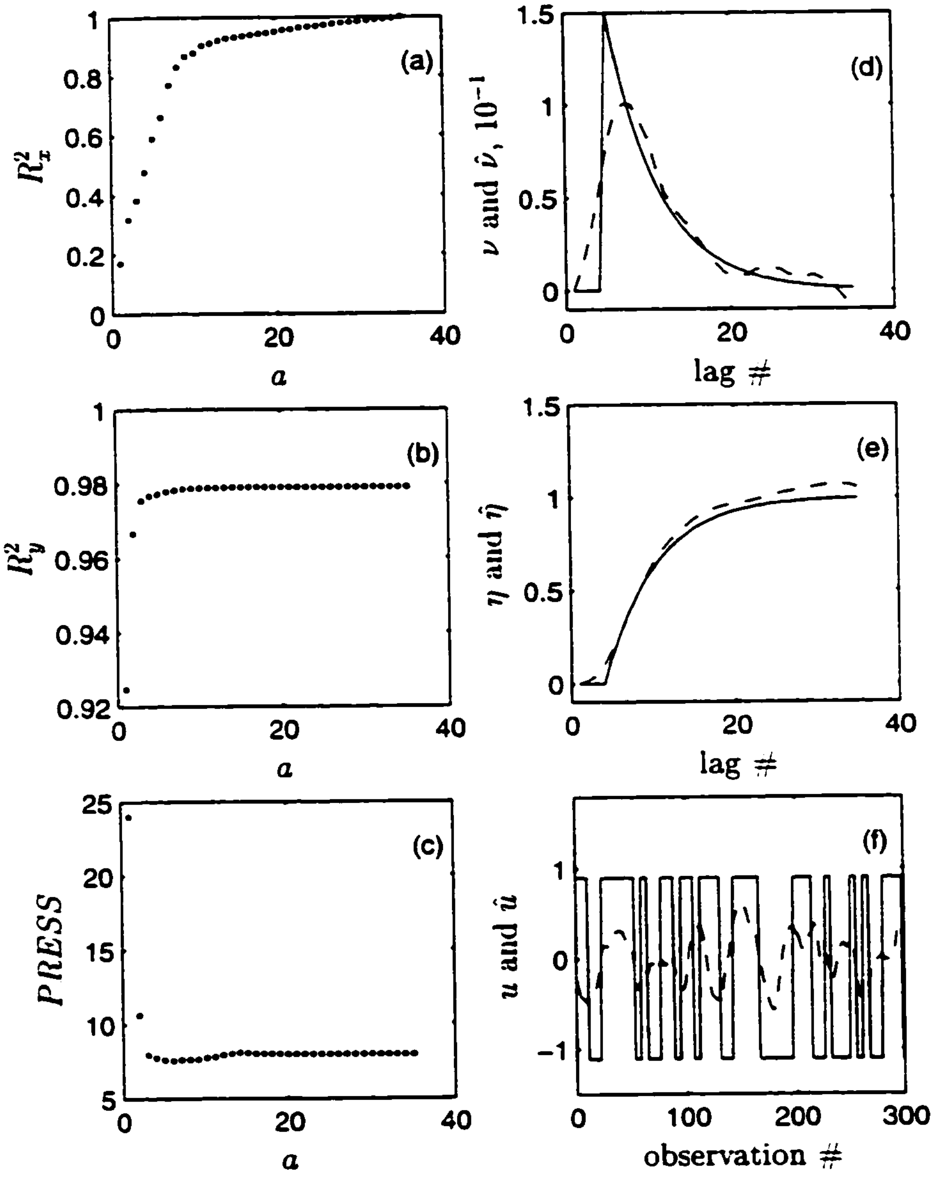


Figure 4.1: Typical FIR identification results using PLS: a) explained variance in the input space versus number of latent variables; b) explained variance in the output space versus number of latent variables; c) output prediction error sum of squares; d) impulse weights (solid: true; dashed: estimated); e) step weights (solid: true; dashed: estimated); f) PRBS input sequence (solid: true; dashed: estimated).

on the Jackknife or the Bootstrap statistics. It allows one to capture more information about the model structure and also provides an assessment of the uncertainty in parameter estimates. A more judicious choice of the number of latent variables is obtained, lying in between the number suggested by cross-validation and the maximum number of dimensions, corresponding to the least squares estimates. The concepts are illustrated with simulations on a system taken from MacGregor *et al.* (1991).

The chapter is organized as follows. The Jackknife and the Bootstrap statistical methods are first presented, leading to a description of alternative model selection criteria based on them. The simulation studies are then presented and their results analyzed. Then, some conclusions are drawn.

4.2 The Jackknife and the Bootstrap Statistics

The Jackknife (Quenouille, 1949; Tukey, 1958) and the Bootstrap (Efron, 1979) are statistical computer based methods used to assess the uncertainty in statistics computed from finite samples. They provide estimates of bias and standard error, which allow computing approximate confidence intervals with a minimal number of assumptions (Efron and Tibshirani, 1993). However, bias estimation is not covered in this section since uncertainties in estimating bias are often larger than for estimating standard errors (Efron and Tibshirani, 1993). Only information about the Jackknife and the Bootstrap that is relevant to the paper is presented here, but more information about existing methods for estimating standard errors is provided in Appendix B. Martens and Martens (1999) have recently used jackknifing methods in PLS regression to obtain approximate confidence intervals on various parameters of the PLS model. This is implemented in the latest version of Unscrambler (Camo-ASA, 1998). In this chapter we propose a procedure based on the jackknife method for selecting the number of components to use in PLS models. Approximate confidence intervals on the resulting model parameters are also obtained as a byproduct of this procedure. The reader is referred to (Efron and Tibshirani, 1993; Gray and Schucany, 1972; Efron, 1983) for more details on the Jackknife and Bootstrap.

4.2.1 The Jackknife

The delete-1 Jackknife (Efron and Tibshirani, 1993) is most commonly encountered in the literature and is conceptually very simple. Consider a data sample of size n, $\mathbf{x} = (x_1, x_2, \ldots, x_{n-1}, x_n)$, measured from a given process, and a parameter estimate $\hat{\theta} = f(\mathbf{x})$ estimated from it. The delete-1 Jackknife estimate of standard error of $\hat{\theta}$ is obtained by deleting each of the n observations, one at a time, and computing replicates of $\hat{\theta}$ using the remaining sampled data:

$$\hat{\theta}_i = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n-1}, x_n) \tag{4.1}$$

This process is repeated for $i = 1, 2, \ldots, n$, until each observation has been deleted once and only once, thereby generating n jackknife replications of $\hat{\theta}$. The jackknife estimate of standard error is given by Efron and Tibshirani (1993):

$$\widehat{ste}_{\text{jack},1} = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \bar{\hat{\theta}})^2\right]^{1/2} \tag{4.2}$$

where $\bar{\hat{\theta}}$ is the average of the n jackknife estimates $(\hat{\theta}_i)$. The factor $(n-1)/n$ has been chosen such that $(\widehat{ste}_{\text{jack},1})^2$ is an unbiased variance estimator of the sample mean (e.g. $\hat{\theta} = \bar{x}$). However, there is no guarantee that $\widehat{ste}_{\text{jack},1}$ is unbiased for other statistics.

For large data samples, using the delete-1 jackknife may be too computationally intensive and a grouped jackknife may be used. It consists of removing selected groups of observations instead of only one observation at the time. Again, each group of observations is deleted once and only once. Assume a sample of size n is divided into g groups of size q $(n = gq)$, the grouped jackknife standard error estimator of $\hat{\theta}$ is (Shao and Wu, 1989):

$$\widehat{ste}_{\text{jack},g} = \left[\frac{g-1}{g}\sum_{i=1}^{g}(\hat{\theta}_i - \bar{\hat{\theta}})^2\right]^{1/2} \tag{4.3}$$

Since the grouped jackknife includes the delete-1 jackknife as a special case, Equation 4.3 will be used throughout this chapter.

Although rarely discussed in the literature, some assumptions are required to ensure that $\widehat{ste}_{\text{jack}}$ is a valid estimator of the natural errors in $\hat{\theta}$. The main underlying assumptions are that the collected data set and the jackknife subsamples be representative of the population. The subsampling should also be balanced. The latter means that the jackknife subsamples should have approximately the same proportion of data collected from each region of the original sampling space. More details on balanced sampling are available in Miller (1974) and Hinkley (1977). The jackknife subsampling method breaks down for very small samples. It is also limited to statistics that are not too unsmooth nor too non-linear (Efron and Tibshirani, 1993).

4.2.2 The Bootstrap

The bootstrap is more recent and usually more computationally intensive than the jackknife. The basic idea, which is similar to Monte Carlo simulation, is to randomly resample $\mathbf{x}$, b times, *with replacement*. The b bootstrap subsamples have the same size as the original data set, $\mathbf{x}$, and allow one to compute b bootstrap replications of $\hat{\theta}$ $(\hat{\theta}_i)$. The bootstrap estimate of standard error for $\hat{\theta}$ is given by Efron and Tibshirani (1993):

$$\widehat{ste}_{\text{boot},b} = \left[\sum_{i=1}^{b}(\hat{\theta}_i - \bar{\hat{\theta}})^2/(b-1)\right]^{1/2} \tag{4.4}$$

The underlying assumptions of the jackknife also apply to the bootstrap. The number of bootstrap subsamples (b) is set by the user and is not limited to the number of observations in the original data set. A large b allows one to gather more information about the standard error in $\hat{\theta}$ and the bootstrap is more efficient with unsmooth or non-linear statistics (Efron and Tibshirani, 1993). However, resampling with replacement may lead to some unrealistic bootstrap samples. Generating a large number of subsamples is therefore recommended, but this greatly increases computation time.

4.3 A Jackknife Criterion for Selecting the Number of Latent Variables

The Jackknife criterion has been developed for improving the selection of non-parsimonious dynamic models. Before describing the criterion, the structure of the FIR and ARX models investigated in this paper are presented in turn.

The FIR model is of the form:

$$y_t = \sum_{i=1}^{m} \nu_i \left(z^{-1}\right) u_{i,t} + e_t \tag{4.5}$$

where m is the number of input variables. $\nu_i \left(z^{-1}\right)$ is a polynomial in the backward shift operator $\left(z^{-1}\right)$ containing the impulse weights, $\nu_{i,j}$, between the i^{th} input and the output variable:

$$\nu_i \left(z^{-1}\right) = \nu_{i,0} + \nu_{i,1}\, z^{-1} + \nu_{i,2}\, z^{-2} + \ldots + \nu_{i,r}\, z^{-r} \tag{4.6}$$

The polynomial order, r, should be chosen to be greater than the process settling time. Step weights are easily obtained as the cumulative sum of the corresponding impulse weights:

$$\eta_{i,j} = \sum_{k=0}^{j} \nu_{i,k} \tag{4.7}$$

where $\eta_{i,j}$ is the step weight of the i^{th} input at lag j.

ARX models are of the following form:

$$A\ (z^{-1})\ y_t = \sum_{i=1}^{m} B_i\ (z^{-1})\ u_{i,t} + e_t \tag{4.8}$$

They have more flexibility for modelling colored disturbances, but are also non-parsimonious models, except for fairly trivial cases. Consider a more parsimonious model of the form $y = (B/A)\ u_1 + (D/C)\ u_2 + (F/E)\ e$, where each input effect is modelled separately, and A to F are finite polynomials in z^{-1}. To convert this two input parsimonious model to ARX form would involve the following. Multiplying through by the inverse of the noise model gives $(E/F)\ y = (BE/AF)\ u_1 + (DE/CF)\ u_2 + e$. Expanding each rational polynomial term into a finite impulse response form yields the ARX form. Only in the single case where $A = C = E$ and $F = 1$ will the ARX model be parsimonious, i.e. $E\ y = B\ u_1 + D\ u_2 + e$.

Common approaches to estimating parameters in dynamic models are the Output Error Method (OEM) and the Prediction Error Method (PEM) (Ljung, 1999; Söderström and Stoica, 1989). Fitting the FIR model (Equation 4.5) by ordinary least squares (OLS) would correspond to OEM for e being an arbitrary disturbance, and to PEM if e were white noise. For ARX models (Equation 4.8) the PEM is simply OLS. An important point in the following section of this paper is that OLS corresponds exactly to PLS regression when all possible dimensions are used (i.e. when the number of PLS components is equal to the number of parameters to be estimated). This will generally lead to parameter estimates with large variance. These models often give poor predictions and controllers with poor performance or robustness (Dayal and MacGregor, 1996). It is for this reason that regularized least squares or latent variable methods such as PLS are generally preferred over PEM/OLS methods for fitting these non-parsimonious models.

The objective in model identification is not just to obtain a model which will give good predictions of the output, but to obtain a good approximation to the true underlying dynamic behavior of the process so that the controller design (involving inversion of the model structure) will result in good control of the process output or the simulations with different types of input variations will give reliable results. We will interpret this objective as obtaining the best estimates of the true impulse or step response of the process. If the model is parameterized as a finite impulse or step response function (see equations 4.6 and 4.7), this would imply that the objective is to obtain the best estimates of the model parameters $\underline{\nu}$ or $\underline{\eta}$. Past literature (Ricker, 1988; MacGregor *et al.*, 1991; Dayal and MacGregor, 1996) have shown that using cross-validation as a stopping criterion will usually lead to a poorly estimated impulse or step response of the process because too few latent variables are selected. Therefore for process identification it would seem logical to use the largest number of latent variables consistent with capturing the underlying model structure, but to stop when there is evidence that one is starting to fit noise. The sum of squares of the model residuals (SSE) will always decrease as one adds more latent variables and so it can not be used alone to select the appropriate model. The proposed jackknife criterion complements the SSE profile by revealing when the model is overfitting the data. The criterion is a measure of the total variance of the impulse or step response parameters, estimated by jackknifing or boostraping. The proposed model selection criterion therefore suggests that one continue to add latent variables as long as the SSE and the total variance of the parameter estimates are decreasing or stable, and to stop when adding more latent variables leads to a sustained increase in the parameters uncertainty as measured by the jackknife criterion. The jackknife criterion alone will not select the model having the best fit, but provides a measure of the model parameter stability and of the occurrence of overfitting. However, it will be shown by simulation studies that the joint use of the SSE profile and the jackknife stability criterion leads to the selection of models that coincide well with the minimum mean square error (MSE) between the true and estimated impulse and step weights.

The jackknife stability criterion is obtained by building a PLS model, with a latent variables, on each of the jackknife or bootstrap subsamples. For each dimension, $a = 1, 2, \ldots, A$, the total sum of squares of the estimated parameters, about their sample mean, is evaluated:

$$SS_{\hat{\theta}}(a) = \sum_{j=1}^{p} \sum_{i=1}^{b} \{\hat{\theta}_{i,j}(a) - \bar{\hat{\theta}}_j(a)\}^2 \tag{4.9}$$

where p and b are respectively the total number of parameter estimates ($\hat{\theta}$) and the number of jackknife (or bootstrap) subsamples. $\hat{\theta}_{i,j}(a)$ is the i^{th} replicate of the j^{th} parameter estimated using a latent variables and $\bar{\hat{\theta}}_j(a)$ is its sample mean over all b replicates.

When computing several latent variables, the result of the procedure is a stability profile, $SS_{\hat{\theta}}(a)$, that, when used in conjuction with the SSE profile, provides the basis for selecting the number of latent variables that should be kept in the model. When a significant portion of the noise is being fitted, $SS_{\hat{\theta}}(a)$ should rise rapidly with subsequent latent variables. A break should appear in the stability profile, suggesting that adding further dimensions is degrading the model. In other words, the proposed criterion suggests that one keep adding latent variables to the model as long as the residuals are small and the model is consistently estimated.

Figure 4.2 shows the results that are obtained for the first order process discussed in section 4.1, when using the proposed criterion. The model stability profile, shown in Figure 4.2 (a), suggests that it is possible to add up to 6 latent variables to the model before significant overfitting takes place, as $SS_{\hat{v}}(a)$ rises rapidly beyond 6 dimensions. Note that the OLS solution corresponds to using all 35 dimensions. The model with 6 dimensions captures more information about the input sequence and especially about the square edges, as can be observed in Figure 4.2 (b). Better estimates of the dead-time and the impulse and step response are obtained as shown in Figure 4.2 (c) and (d) (compare with Figure 4.1). The error sum of squares between the true and estimated impulse and step weights is decreased by a factor of two and a factor of four respectively. Further illustration of the improvements that can be achieved using the stability profile in identification are presented in section 4.5.

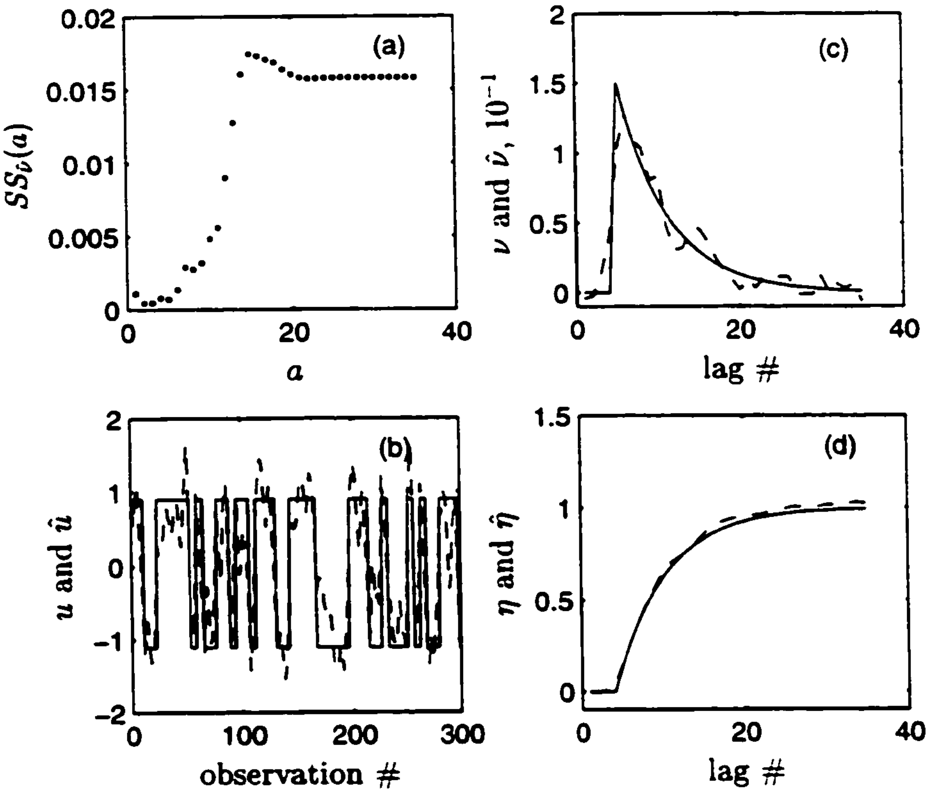


Figure 4.2: FIR identification results estimated with 6 latent variables as selected by the Jackknife criterion: a) model stability profile for 35 latent variables; b) PRBS input sequence (solid: true; dashed: estimated); c) impulse weights (solid: true; dashed: estimated); d) step weights (solid: true; dashed: estimated).

A parallel can be drawn between the procedure discussed in this section and the selection of the ridge parameter (k) in ridge regression using what is commonly called the ridge trace (Hoerl and Kennard, 1970a and 1970b). The ridge trace procedure corresponds to plotting the model parameter estimates as a function of increasing values of the ridge parameter. This also achieves a reduction in the variance of the parameter estimates at the expense of an increasing bias. The ridge parameter is then usually chosen to be that when the estimates appear to become stable. The proposed Jackknife stability profile can be used to select the ridge parameter in a more objective manner. As the ridge parameter (k) is increased the jackknifed sum of squares of the parameter estimates $SS_{\hat{\theta}}(k)$ will decrease sharply and then stabilize (plateau) when the variations in parameter estimates due to fitting noise has been reduced. The value of the ridge parameter at this stabilizing point should be used. Note that the stability traces $SS_{\hat{\theta}}(k)$ for ridge regression and $SS_{\hat{\theta}}(a)$ for PLS regression work in the same manner but in opposite directions since variance is reduced (and bias increased) for increasing values of k but for decreasing values of a.

An additional benefit from Jacknifing is that one can obtain estimates of the standard errors of the impulse and step response weights using equation 4.3. This allows one to plot approximate confidence intervals on the impulse and step responses.

4.4 Simulations

The ability of the proposed criterion to select better non-parsimonious dynamic models for control is illustrated through various simulation studies. A multi-input single-output dynamic system used in MacGregor *et al.* (1991) is also used in this study because it includes most of the common types of dynamic responses encountered in practice, namely processes with and without delay, first and higher order systems, inverse response and a variable with no effect.

The process under investigation is a five input, one output system corrupted by a noise, D_t:

$$y_t = \sum_{i=1}^{5} Z_{i,t} + D_t \tag{4.10}$$

where

$$Z_{1,t} = \frac{0.15\ z^{-4}}{1 - 0.85\ z^{-1}}\ u_{1,t} \tag{4.11}$$

Table 4.1: Disturbance variance characteristics.

Noise Structure	σ_a^2	σ_D^2
White	0.0104	0.0104
ARI	0.0016	0.2210

$$Z_{2,t} = \frac{0.045\ z^{-1}}{(1-0.85\ z^{-1})\ (1-0.7\ z^{-1})}\ u_{2,t} \tag{4.12}$$

$$Z_{3,t} = \left[\frac{-0.2\ z^{-1}}{1-0.8\ z^{-1}} + \frac{0.12\ z^{-5}}{(1-0.8\ z^{-1})\ (1-0.6\ z^{-1})}\right] u_{3,t} \tag{4.13}$$

$$Z_{4,t} = 0.0\ u_{4,t} \tag{4.14}$$

$$Z_{5,t} = \frac{0.3\ z^{-8}}{1-0.7\ z^{-1}}\ u_{5,t} \tag{4.15}$$

Dynamic elements for inputs 1, 2 and 5 have unit gain, while a gain of 0.5 and 0 are defined for input 3 and 4 respectively. Two different noise structures have been studied: a white noise sequence and a nearly non-stationary autoregressive (ARI) disturbance, respectively:

$$D_t = a_t \tag{4.16}$$

$$D_t = \frac{1}{(1-0.99\ z^{-1})\ (1-0.4\ z^{-1})}\ a_t \tag{4.17}$$

where a_t is a white noise $\mathcal{N}(0,\sigma_a^2)$ sequence. The variance characteristics of each noise sequence are given in Table 4.1.

Independent Pseudo-Random Binary Sequences (PRBS) were used for each process input. The magnitude and switching interval of the five PRBS sequences are found in Table 4.2. Note that the magnitude of the PRBS corresponding to input 2 is half of the gain of the second transfer function, while it is greater or equal to the gain of all other transfer functions. In addition, the switching interval on input 2 is rather slow. The second transfer function is therefore expected to be more difficult to identify and should have a greater uncertainty, due to lower signal to noise ratio.

Table 4.2: Characteristics of the designed PRBS input sequences.

Input Number	Magnitude	Switching Interval
1	1.00	6
2	0.43	10
3	1.00	6
4	0.75	1
5	1.20	3

In this work, FIR models with 35 lags of each input variable were enough to cover the five dynamic relationships (MacGregor *et al.*, 1991). A total of 175 impulse weights therefore need to be estimated.

The proposed criterion for non-parsimonious model selection is compared to cross-validation on the basis of the mean square error between the true and estimated impulse and step weights:

$$MSE_{\text{imp}} = \sum_{i=1}^{5} \sum_{j=0}^{34} (\hat{\nu}_{i,j} - \nu_{i,j})^2 \tag{4.18}$$

$$MSE_{\text{step}} = \sum_{i=1}^{5} \sum_{j=0}^{34} (\hat{\eta}_{i,j} - \eta_{i,j})^2 \tag{4.19}$$

The MSE measures the closeness of fit to the true models. MSE_{imp} mainly focuses on the process dynamics, while MSE_{step} focuses more on the steady-state gains.

In all cases, non-parsimonious model selection is performed using the grouped jackknife statistic whose estimates of standard error have been shown reasonable by a Monte Carlo simulation (see Appendix B). This choice is also motivated by its strong algorithmic similarity with cross-validation, that is already implemented in standard PLS software packages. The number of subgroups to use in the jackknife procedure was investigated using the simulation study with good input design described in section 4.5.1. The stability profile and the estimates of standard error on the parameters were computed using successively 10, 30, 60 and 100 jackknife subgroups. However, the corresponding stability profiles were found very similar and all suggested keeping about the same number of dimensions in the model. The proposed criterion is not very sensitive to the number of subgroups provided that the number of subgroups is not too small, and the data do not have small clusters in different regions. In this chapter, jackknifing with 30 subgroups was used since it provided a reasonable compromise between accuracy and computation time.

4.5 Results

The usefulness of the proposed non-parsimonious model selection criterion is illustrated through three case studies. The first two cases investigate a process corrupted by a white noise sequence, where a good and a poor experimental design (correlated inputs) have been implemented. In the third case, a process with a nearly non-stationary autoregressive disturbances is identified. A total of 335 observations were utilized for FIR model identification and computation of the $SS(a)$ profiles and the MSE's. This ensures that after initializing the calculations, 300 data points are available for the identification in all cases.

4.5.1 Identification of models with white noise

Good Input Design

In the base case study, the MISO process with white noise is identified assuming a FIR model. The five inputs are manipulated according to the design shown in Table 4.2. With a good design such as this, one would expect a very broad range of model dimensions (A) to give good results. The modelling results are presented in Figure 4.3 to 4.5.

The impulse and step response stability profiles, the sum of squares of the model residuals and the mean square error for the impulse and step responses for the first 50 PLS dimensions are shown in Figure 4.3. Cross-validation stopped the model building procedure at 4 latent variables. However, the stability profile for the impulse response ($SS_{\hat{\nu}}$) shows a very slowly rising profile up to about 12 latent variables, and a steeper rise beyond that, while that for the step response ($SS_{\hat{\eta}}$) shows a very flat profile in which increased variance does not appear to occur until beyond 20 to 30 latent variables. This implies that one could select between 6 to 12 dimensions and still obtain a stable model, but with smaller residual variance (SSE) than with cross-validation. As expected, a broad range of latent variable choices would appear to be acceptable. We have chosen 10 dimensions in this case as it accounts for most of the steep decrease in SSE (a factor of 4 smaller than the cross-validation results) while still giving stable impulse and step weight estimates.

Both MSE profiles confirm that one would obtain better models by adding more latent variables to the 4 dimensions suggested by cross-validation, as a minimum on the MSE profiles is not attained until 12 and 14 latent variables for the impulse and step responses respectively. Any choice between 8 and 15 latent variables would give low MSE's. The range of dimensions suggested the jackknife criterion covers this low MSE region relatively well. Note that the OLS (PEM in this case) solution would correspond to taking the maximum 175 components. As suggested in the plots of the jackknife statistics, and confirmed in the MSE_{imp} plot this would lead to a poorer model.

Figures 4.4 and 4.5 show the true and estimated impulse and step weights for the 5 inputs (plotted in a sequence one after the other), obtained with 4 and 10 latent variables (e.g. with cross-validation and the proposed criterion). Approximate 95% confidence intervals around the estimated impulse and step responses are also plotted. It consists of a two standard error limit, estimated through jackknifing. Improvements in each estimated transfer function are noticeable. The dead-time of the first transfer function is better estimated and significant reduction in the step weights uncertainty

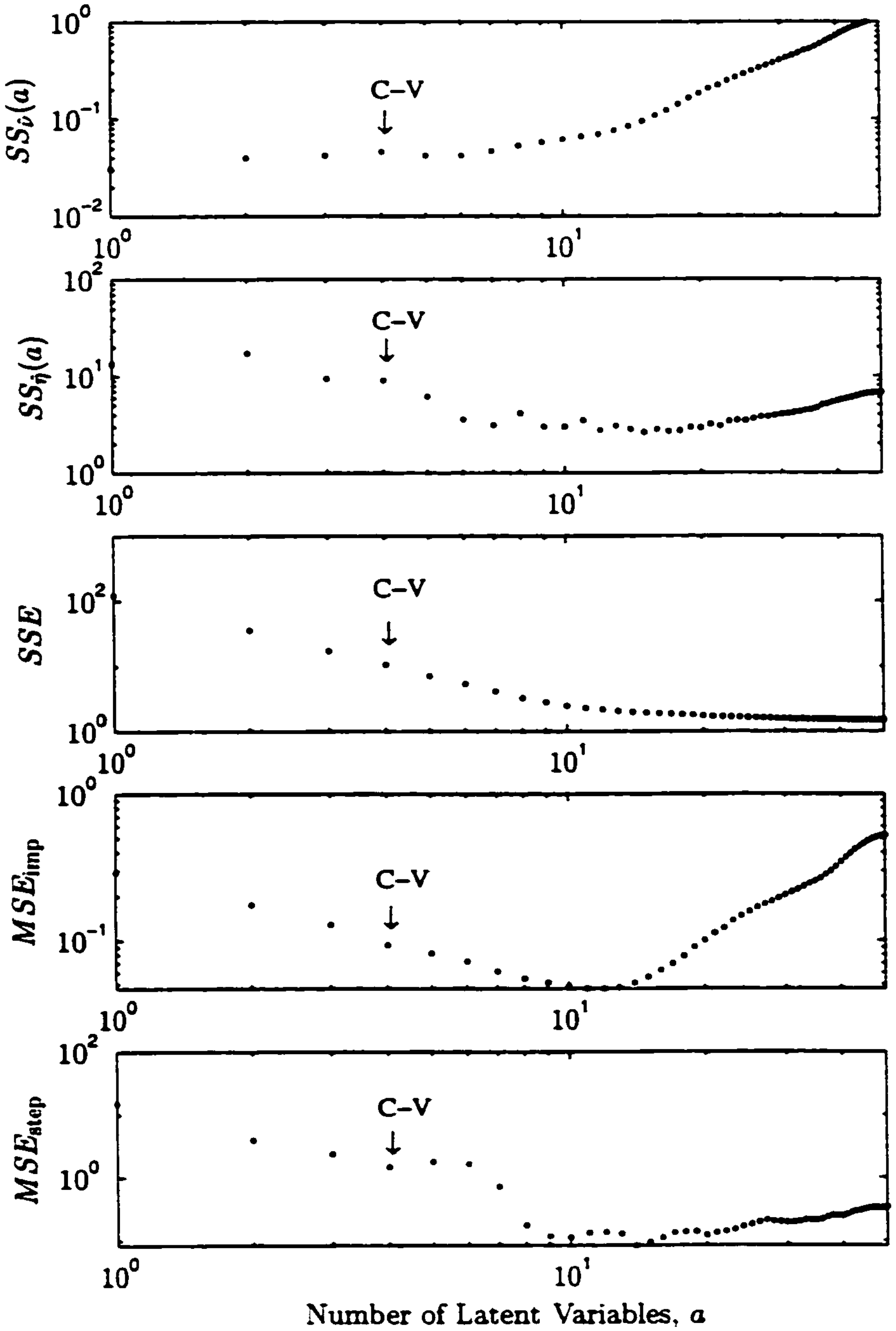
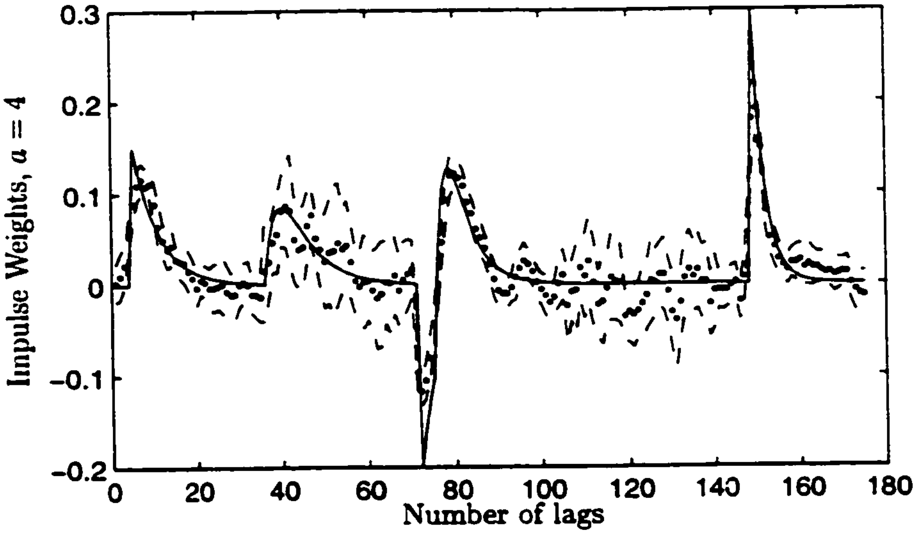


Figure 4.3: Process with white noise and good experimental design: model stability profiles for the impulse and step responses, sum of squares of the model residuals and mean square error for the impulse and step responses.



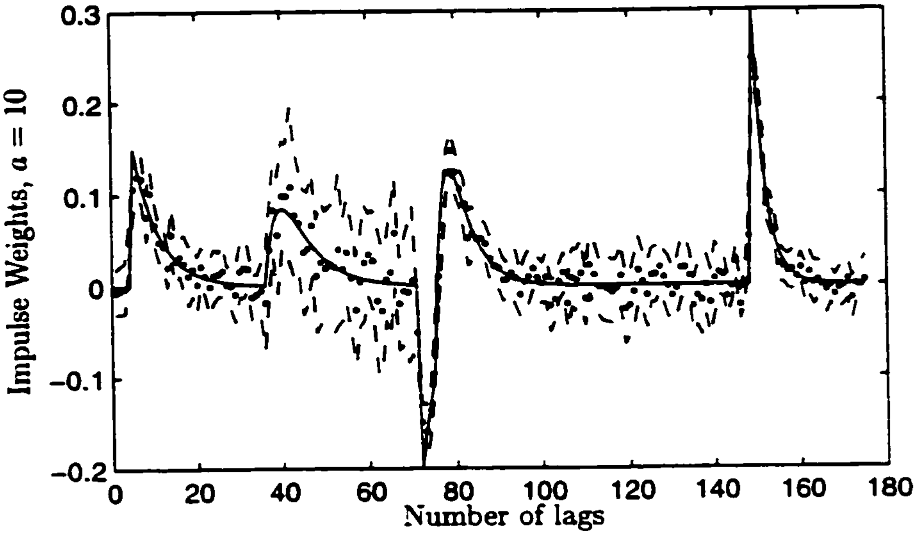
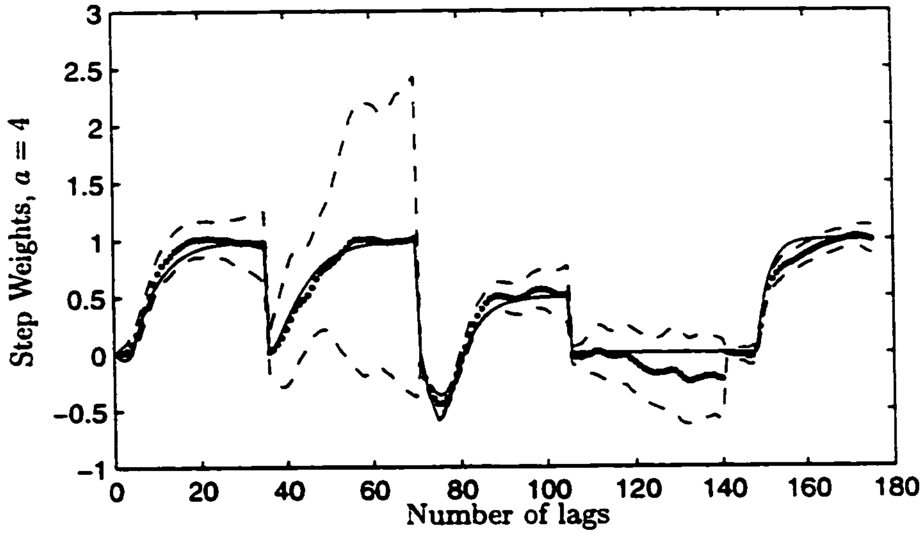


Figure 4.4: Impulse weights obtained with FIR model with white noise and good experimental design: selected with cross-validation ($a = 4$) and the proposed criterion ($a = 10$). True response—solid line, estimated response—dots, two-standard error limits based on jackknifed estimates of variance—dashed line.



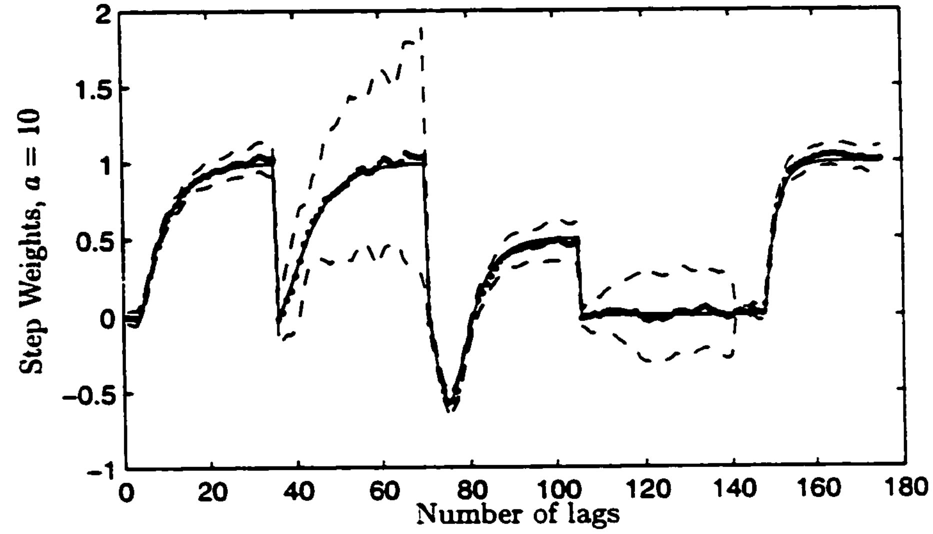


Figure 4.5: Step weights obtained with FIR model with white noise and good experimental design: selected with cross-validation (a = 4) and the proposed criterion (a = 10). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

for the second transfer function is achieved. The inverse response of the third transfer function is closer to the true response and so is the fourth transfer function. Finally, the estimated gains of all five transfer functions are also improved.

Using the proposed selection criterion only led to marginal improvements in this case because of the well designed nature of the experiments. The matrix of predictors (lags of inputs) is well conditioned, the signal to noise ratio in the data was high ($\hat{\sigma}_{y,\text{ noise free}}/\hat{\sigma}_{\text{noise}} \approx 14$) and the disturbance structure was a simple white noise. These all contribute to a straightforward identification problem. However, as the conditioning of the predictor matrix gets worse and when an autocorrelated disturbance is present, the identification problem gets more difficult and the proposed selection criterion will be much more useful, as shown in the following examples.

Correlated Inputs (Poorer Design)

When correlation between process inputs is introduced (poorer design), the predictor matrix gets more ill-conditioned. Cross-validation could be more misleading in selecting the right number of latent variables to capture the process structure. Indeed, the number of possible combinations of lagged inputs that can adequately predict the output should be greater.

To simulate a poorer design of the inputs, linear combinations of the PRBS sequences, shown in Table 4.2, have been taken to generate the following correlation structure among the inputs:

$$\left(\begin{array}{cccccc} & u_1 & u_2 & u_3 & u_4 & u_5 \\ u_1 & 1.0 & -0.325 & 0.216 & -0.421 & 0.324 \\ u_2 & \vdots & 1.0 & -0.132 & 0.946 & -0.079 \\ u_3 & \vdots & \vdots & 1.0 & -0.324 & -0.029 \\ u_4 & \vdots & \vdots & \vdots & 1.0 & -0.090 \\ u_5 & \vdots & \vdots & \vdots & \vdots & 1.0 \end{array}\right) \qquad (4.20)$$

Again, the white noise sequence with variance shown in Table 4.1 has been added to the process signal. The calculated signal to noise ratio is still large at about 13. Identification results using an FIR model are shown in Figure 4.6 to 4.8.

Only 4 dimensions were found significant by cross-validation. However, as shown in Figure 4.6, several additional dimensions could be added to the model without significantly degrading it. The total variability of the FIR model ($SS_{\hat{\nu}}$) is stable up until about 12 dimensions, beyond which overfitting becomes significant. The stability profile for the step response ($SS_{\hat{\eta}}$), on the other hand, goes through a minimum and remains stable until about 16 components. We have chosen 12 latent variables in this case, as it decreases the SSE by almost a factor of 6 compared to cross-validation results without leading to instabilities. The jackknife criterion again leads to better models than cross-validation as confirmed by the MSE profiles for the impulse and step responses. These reach minima with 14 and 12 dimensions respectively. Notice the similarity between the stability profiles and the MSE profiles beyond 12 latent variables. The same similarity was also evident in the first example (Figure 4.3).

The true and estimated impulse and step weights are shown in Figure 4.7 and 4.8. Clear improvements on dead-time and gain estimation have been achieved by using 12 latent variables suggested by the jackknife criterion as opposed to 4 suggested by cross-validation. Note that the approximate confidence intervals on the step responses (Figure 4.8) are also smaller using 12 latent variables, again reflecting the improved estimates. However, if one uses OLS for parameter estimation (which is equivalent to PLS with 175 components), much worse results are obtained as shown in Figure 4.9. All of the estimated step responses are very noisy and the approximate confidence limits (obtained again by jackknifing) are much larger than those for PLS obtained with 12 components. The estimated step response for input 2 and 4 are particularly poor. This shows the benefits obtained from using appropriately selected

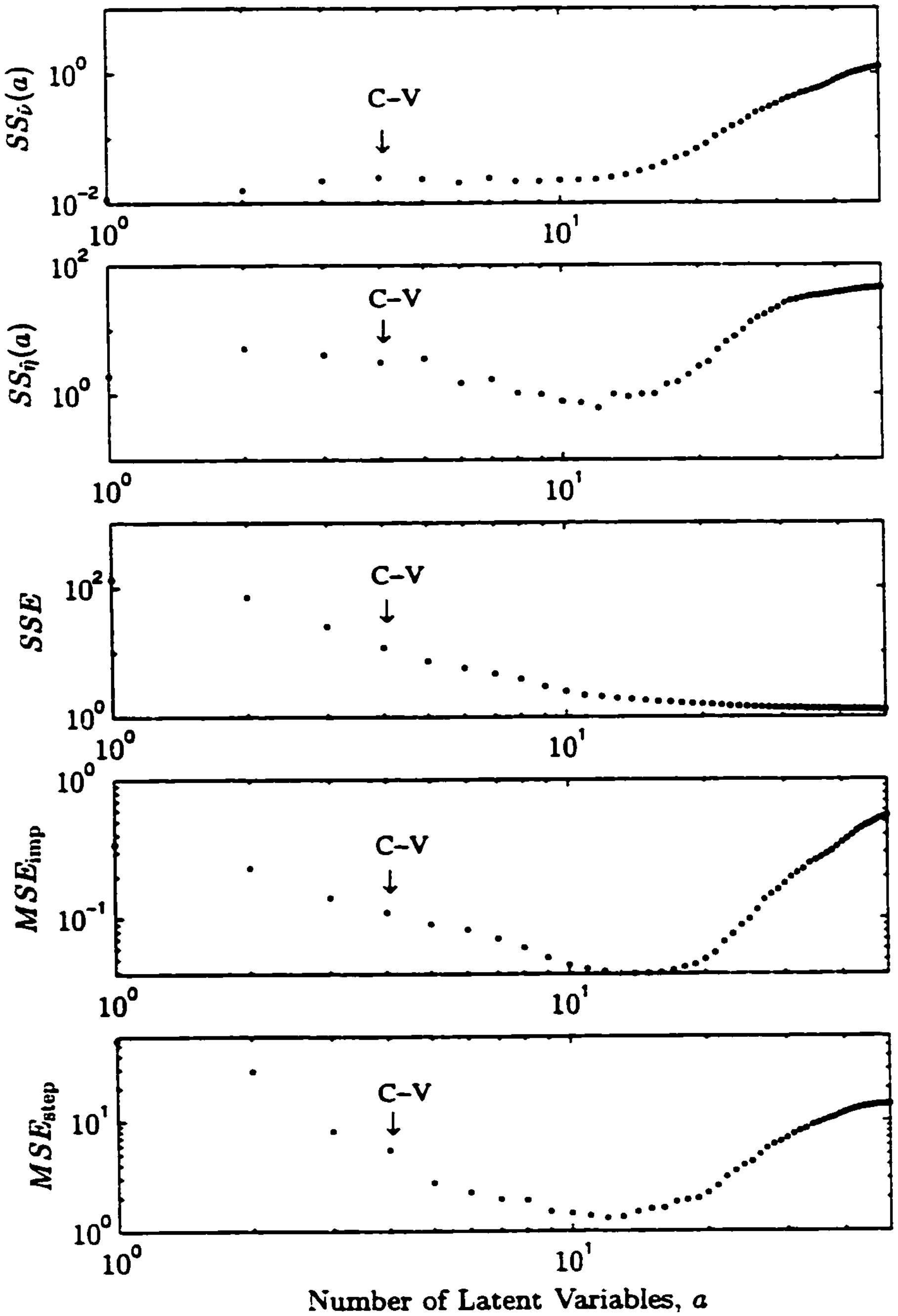


Figure 4.6: Process with white noise and poorer experimental design: model stability profiles for the impulse and step responses, sum of squares of the model residuals and mean square error for the impulse and step responses.

latent variable regression or regularized least squares methods.

4.5.2 Identification of models with structured disturbances

Identification of the MISO process with a nearly non-stationary autoregressive disturbance (Equation 4.17) is investigated in this section. The variance characteristics of the noise are given in Table 4.1 and the input PRBS design sequences are those proposed in Table 4.2. Both FIR and ARX model structures are used to identify the process dynamic relationships.

The high variance and low frequency content of the nearly non-stationary disturbance lead to a more difficult identification problem and poorer estimates of the gains. Fitting the first difference of the inputs and the output significantly improves the parameter estimates by removing the non-stationarity. Differencing is therefore used prior to identifying the FIR and ARX models. The signal to noise ratio with the differenced signals is about 4.

FIR model structure

The lower signal to noise ratio causes cross-validation to stop the model building procedure at even fewer dimensions, as observed by Kresta (1992). Only two latent variables are found significant for the prediction of the output. The stability, the SSE and the MSE profiles are shown in Figure 4.10 for the first 100 PLS dimensions. The behavior of the stability profiles ($SS_{\hat{\nu}}$ and $SS_{\hat{\eta}}$) are very different in this case. After an apparent increase from 6 to 12 latent variables, $SS_{\hat{\nu}}$ falls to a much lower minimum around 30 latent variables. On the other hand, $SS_{\hat{\eta}}$ continuously rises until 12 dimensions, after which it goes through a minimum around 28 latent variables before a final rise. Meanwhile the residual sum of squares continues to decrease. This suggests that one has to add more than 20 dimensions in order to obtain a stable model with a low SSE. We have chosen 30 dimensions in the following analysis

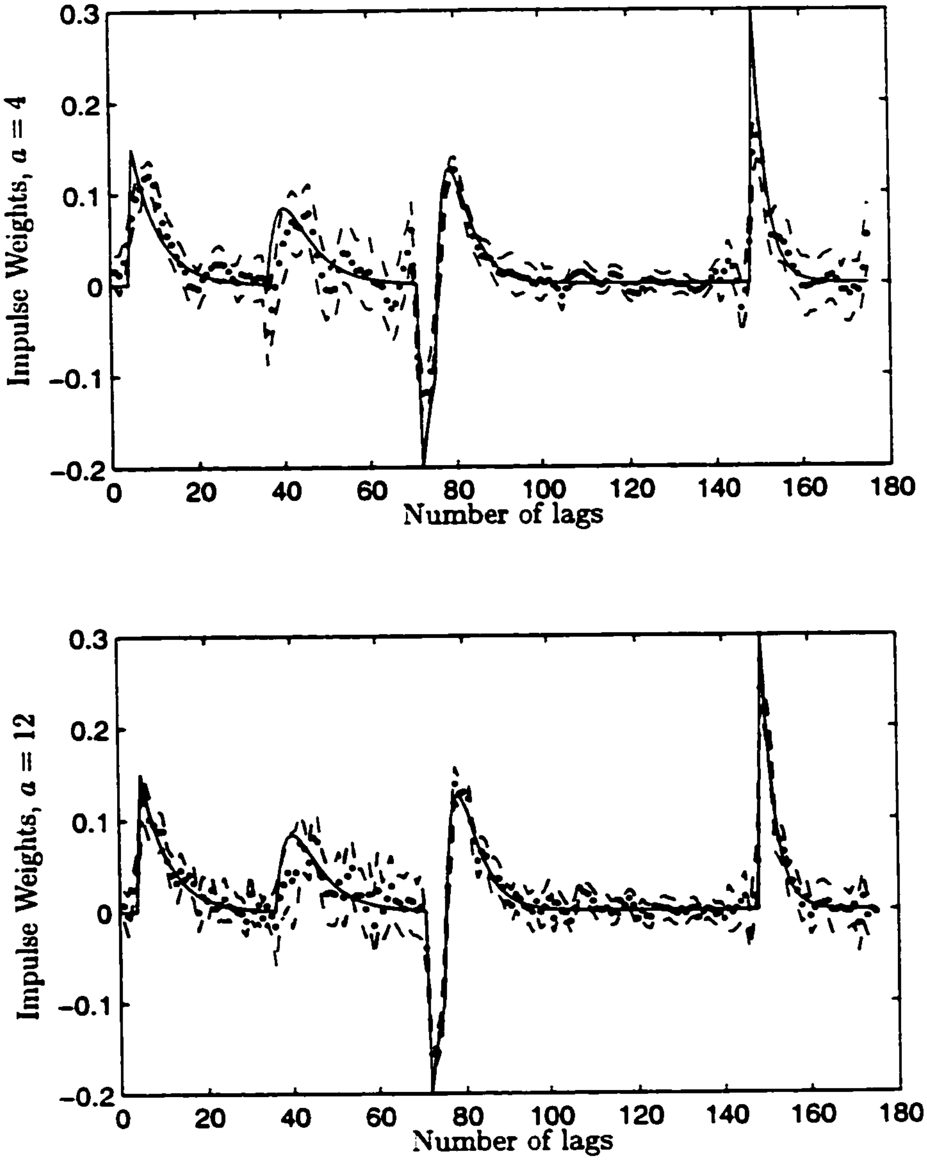


Figure 4.7: Impulse weights obtained with FIR model with white noise and poorer experimental design: selected with cross-validation ($a = 4$) and the proposed criterion ($a = 12$). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

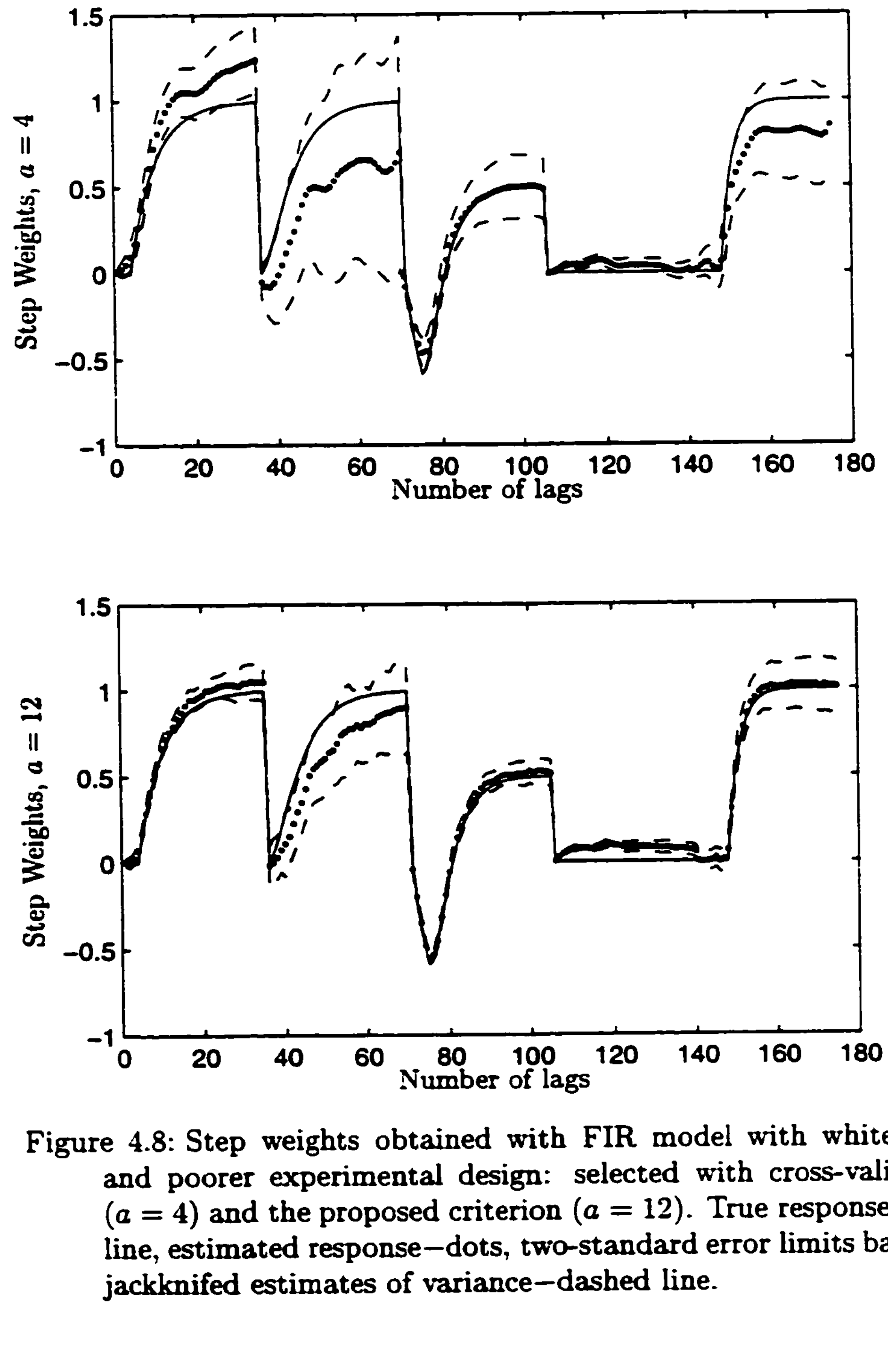


Figure 4.8: Step weights obtained with FIR model with white noise and poorer experimental design: selected with cross-validation ($a = 4$) and the proposed criterion ($a = 12$). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

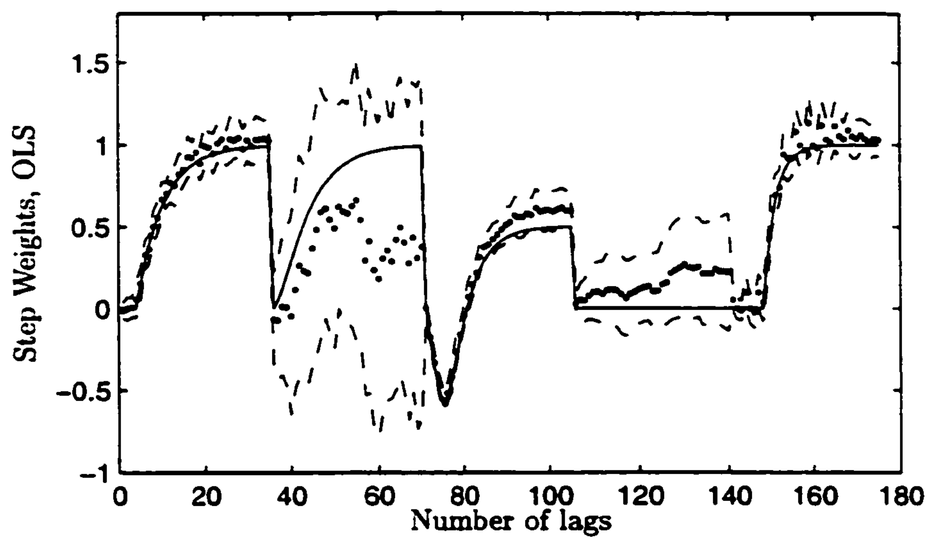


Figure 4.9: Step weights obtained with FIR model with white noise and poorer experimental design: identification using OLS (or PLS with 175 components). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

because it corresponds essentially to the minimum in both jackknife stability profiles. and to the point where the fit error (SSE) levels out at an effective minimum. It is worthwhile noting, however, that in this case OLS or PEM (equivalent to the use of PLS with 175 dimensions) could also be used without much loss since neither of the jackknife stability statistics exhibit too large of an increase by going to very high dimensions. In this sense the criterion can be used to verify whether OLS or PEM can safely be used on any particular data set.

The shape of the jackknife stability profiles in Figure 4.10 warrants some discussion. The small values of $SS_{\bar{\eta}}(a)$ at very low dimensions, and the local minimum in $SS_{\hat{\nu}}(a)$ around 5 coupled with the large residual sum of squares (SSE) in this region imply that although one is obtaining poor models (large SSE's), the parameters are being consistently estimated (low $SS_{\bar{\eta}}(a)$, $SS_{\hat{\nu}}(a)$). Figure 4.11 shows the step responses estimated using $a = 1, 8$ and 12 (dotted lines) and $a = 20, 30$ and 100

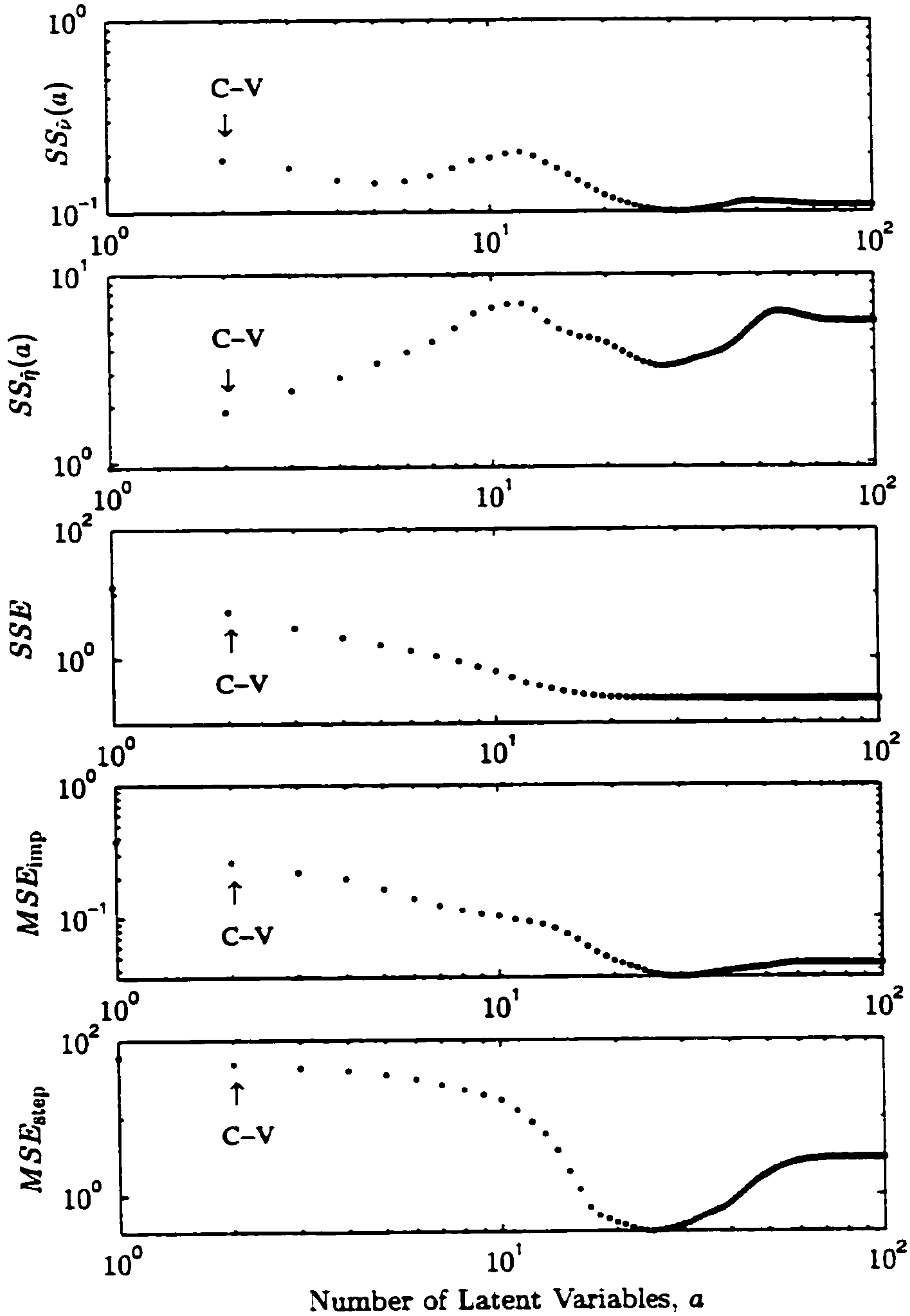


Figure 4.10: Process with non-stationary autoregressive disturbance and FIR model: model stability profiles for the impulse and step responses, sum of squares of the model residuals and mean square error for the impulse and step responses.

(dashed lines) latent variables. Clearly, PLS appears to be capturing very little of the steady-state gain information in the fitted responses until one adds more than 12 latent variables. Only at higher dimensions (20 to 30) is the complexity of the model sufficient to capture both the dynamic and steady-state response (see the MSE's in Figure 4.10 and step responses in Figure 4.11). The maxima in the jackknife statistics around 12 dimensions reflects the instability in the parameter estimates as one transitions between the two regions. Because of the nonstationary nature of the disturbance (Equation 4.17) the data were differenced prior to identification in order to remove the large low frequency component of the disturbance. This filtering operation also reduces the information on the steady-state process gains in the differenced data. Hence, the dominant early latent variables ($a = 1, \ldots, 12$) capture mainly the higher frequency dynamic response information, while the latter latent variables ($a > 12$) capture the smaller amount of information on the steady-state gains.

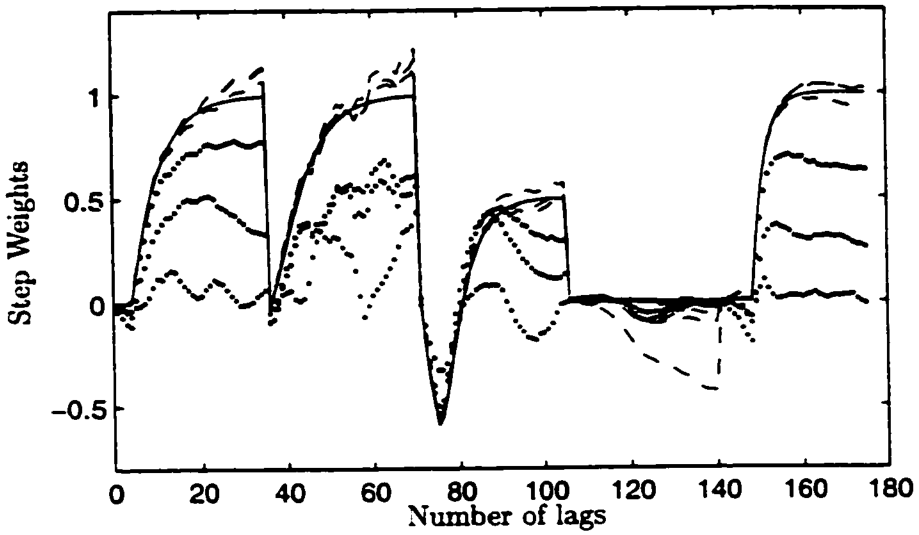


Figure 4.11: Step responses obtained with FIR model with non-stationary autoregressive disturbance: True response–solid line, estimated response with 1, 8 and 12 dimensions–dots, estimated response with 20, 30 and 100 dimensions–dashed line.

The MSE_{imp} and MSE_{step} profiles in Figure 4.10 clearly show that a model with 2 latent variables (cross-validation) is totally inadequate to capture the process structure, as 30 and 25 latent variables are necessary to reach the minimum values of MSE_{imp} and MSE_{step} respectively. The proposed jackknife criterion again leads to the selection of a model very close to that giving the minimum MSE's. Figure 4.12 and 4.13 shows the impulse and step weights obtained using cross-validation and the jackknife criterion (2 and 30 latent variables respectively). A very major improvement in all aspects of the estimated process dynamics is obtained.

ARX model structure

As discussed earlier the ARX model is an alternative non-parsimonious model whose advantage is the ability to accomodate colored disturbances more directly. For the simulated example in this paper the ARX model required almost as many parameters as the FIR model in order to capture the different dynamic responses. A total of 25 lags of each input and one lag of the output were found appropriate (126 parameters to be estimated). The model identification results for the ARX models using cross-validation and the jackknife criterion are almost identical to those obtained for the FIR models, as shown in Figure 4.14 to 4.16.

Cross-validation suggests that only 3 components are significant. However, this model has great instabilities and is also poorly estimated, as shown in Figure 4.14 by the means of the $SS_{\hat{\nu}}$, $SS_{\hat{\eta}}$ and the SSE profiles. To obtain a stable model with low residuals, one need to add more than 17 components. The model stability seems to be best in between 17 and about 30 latent variables, but OLS/PEM would not lead to important loss in this case either. We have chosen 30 components in this example. It is clear from Figure 4.14 that selecting more than 17 latent variables would lead to low MSE's. Figure 4.15 and 4.16 show the estimated impulse and step responses obtained with 3 and 30 latent variables (cross-validation and jackknife

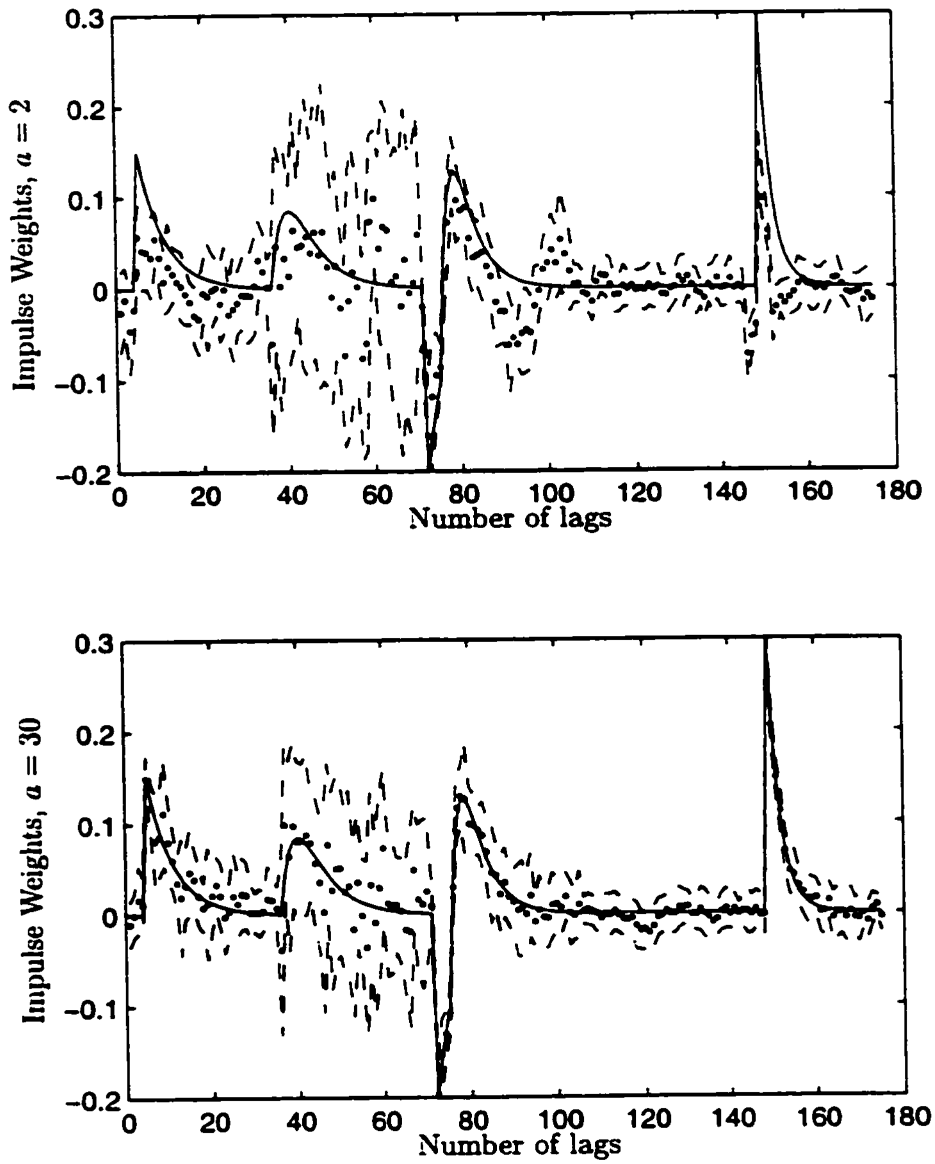
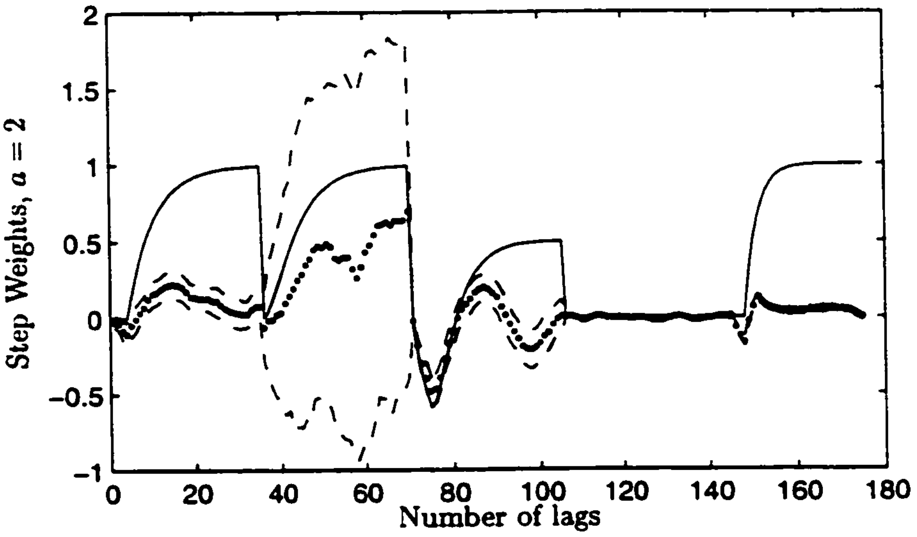


Figure 4.12: Impulse weights obtained with FIR model and ARI disturbance: selected with cross-validation ($a = 2$) and the proposed criterion ($a = 30$). True response—solid line, estimated response—dots, two-standard error limits based on jackknifed estimates of variance—dashed line.



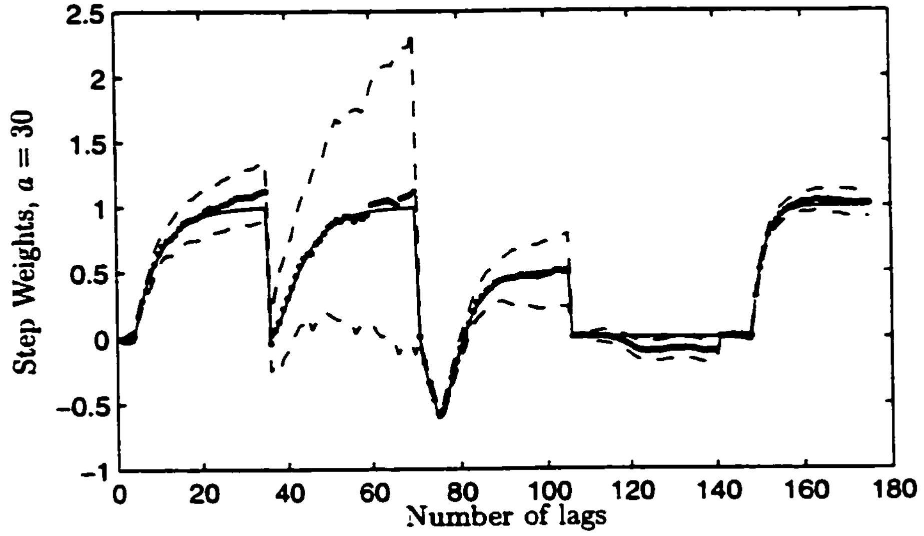


Figure 4.13: Step weights obtained with FIR model and ARI disturbance: selected with cross-validation ($a = 2$) and the proposed criterion ($a = 30$). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

criterion). As for the example with a FIR model, clear improvement in almost all aspects of the process dynamics is obtained when using the jackknife criterion to select the number of PLS components.

4.5.3 Alternative criteria

Alternative criteria for model selection have also been investigated. The Akaike Information Criterion (AIC) (Akaike; 1974), commonly used in parsimonious model identification was investigated. However, for the cases studied, the number of model parameters is so large that the AIC always suggested keeping only one latent variable in the PLS model. The AIC does not appear to be a viable approach for the non-parsimonious models considered in this chapter.

Another criterion is to use the lack of fit diagnostics proposed by Box and Jenkins (1970), originally developed for testing the structure of parsimonious transfer functions and noise models. Upon failure of these tests, modifications to model structures are suggested. Cross-correlation between inputs and residuals at different lags and autocorrelation of the residuals are used for that purpose. An appropriate transfer function model should not lead to significant cross-correlation coefficients between each input and the residuals. In addition, the residuals should not be significantly autocorrelated. The same idea could be applied for selecting the number of latent variables in PLS. When too few latent variables are included in the estimation procedure, the resulting lack of fit should be reflected in the above diagnostic tools, suggesting that further latent variables be added. The estimated cross-correlation coefficients $\hat{r}_{u_i,\hat{e}}$ between each input, $i = 1, \ldots, 5$, and the residuals, $\hat{e}$, and the autocorrelation coefficients of the residuals $\hat{r}_{\hat{e},\hat{e}}$ for the FIR model of section 4.5.2, selected using cross-validation are shown in Figure 4.17. Bartlett's approximate 95% confidence intervals for those coefficients are also provided.

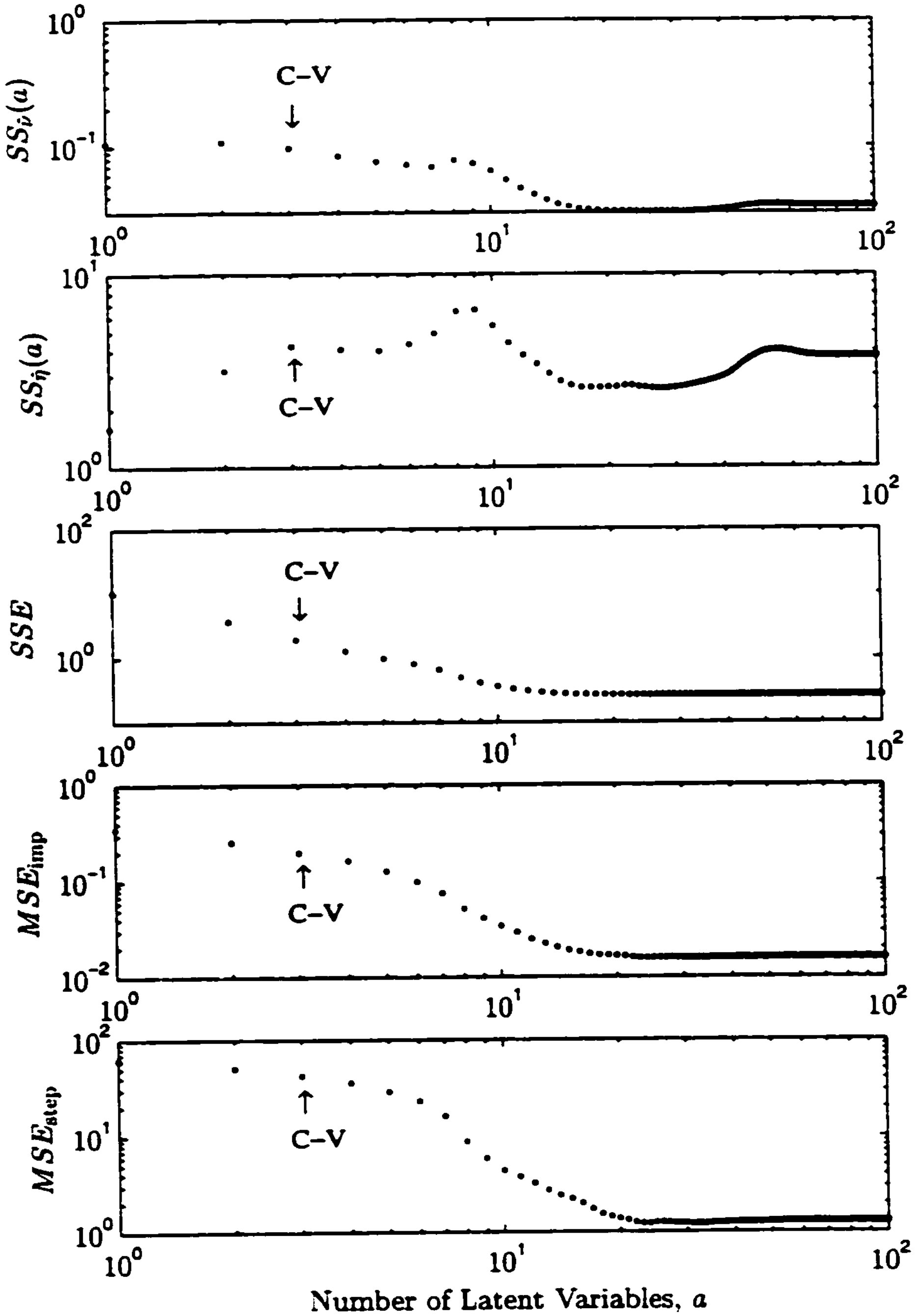


Figure 4.14: Process with non-stationary autoregressive disturbance and ARX model: model stability profiles for the impulse and step responses, sum of squares of the model residuals and mean square error for the impulse and step responses.

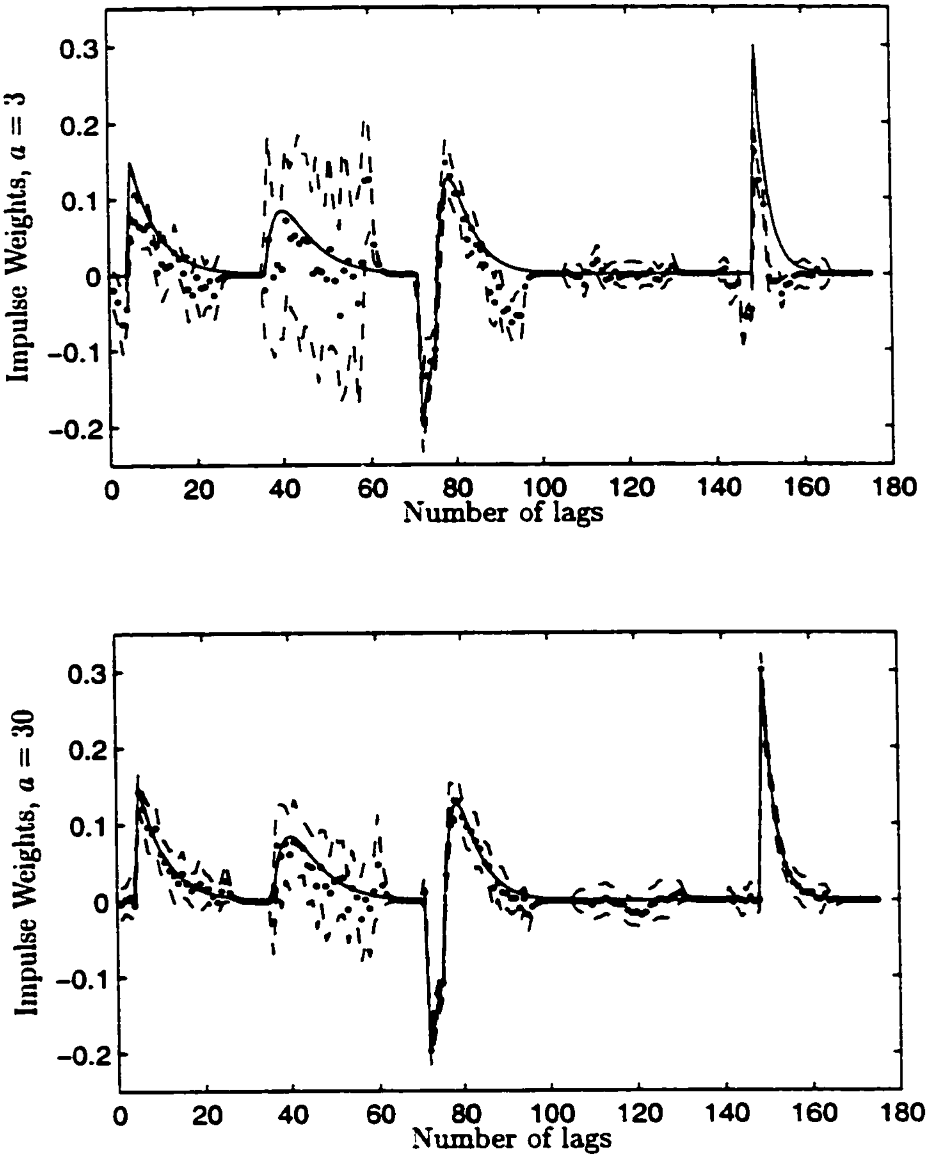


Figure 4.15: Impulse weights obtained with ARX model and ARI disturbance: selected with cross-validation ($a = 3$) and the proposed criterion ($a = 30$). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

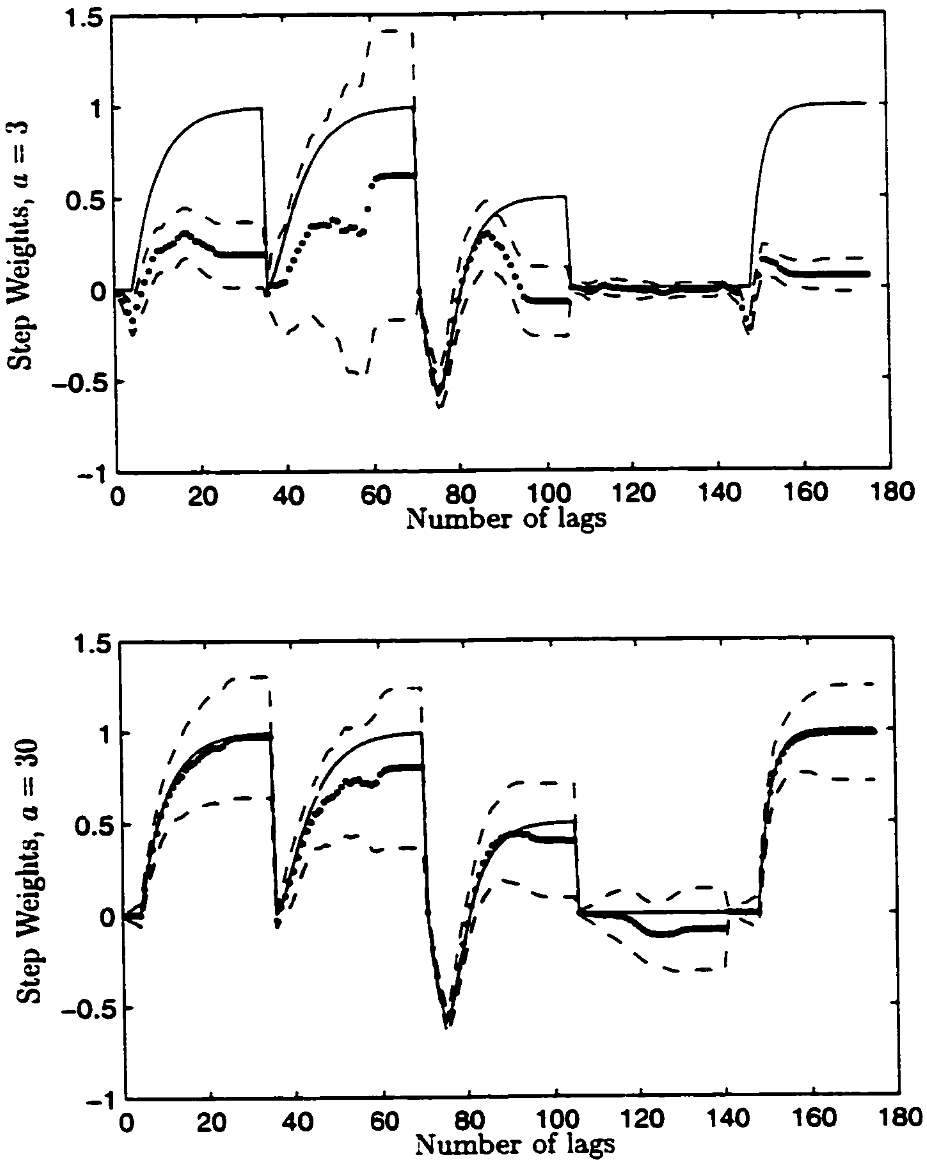


Figure 4.16: Step weights obtained with ARX model and ARI disturbance: selected with cross-validation ($a = 3$) and the proposed criterion ($a = 30$). True response–solid line, estimated response–dots, two-standard error limits based on jackknifed estimates of variance–dashed line.

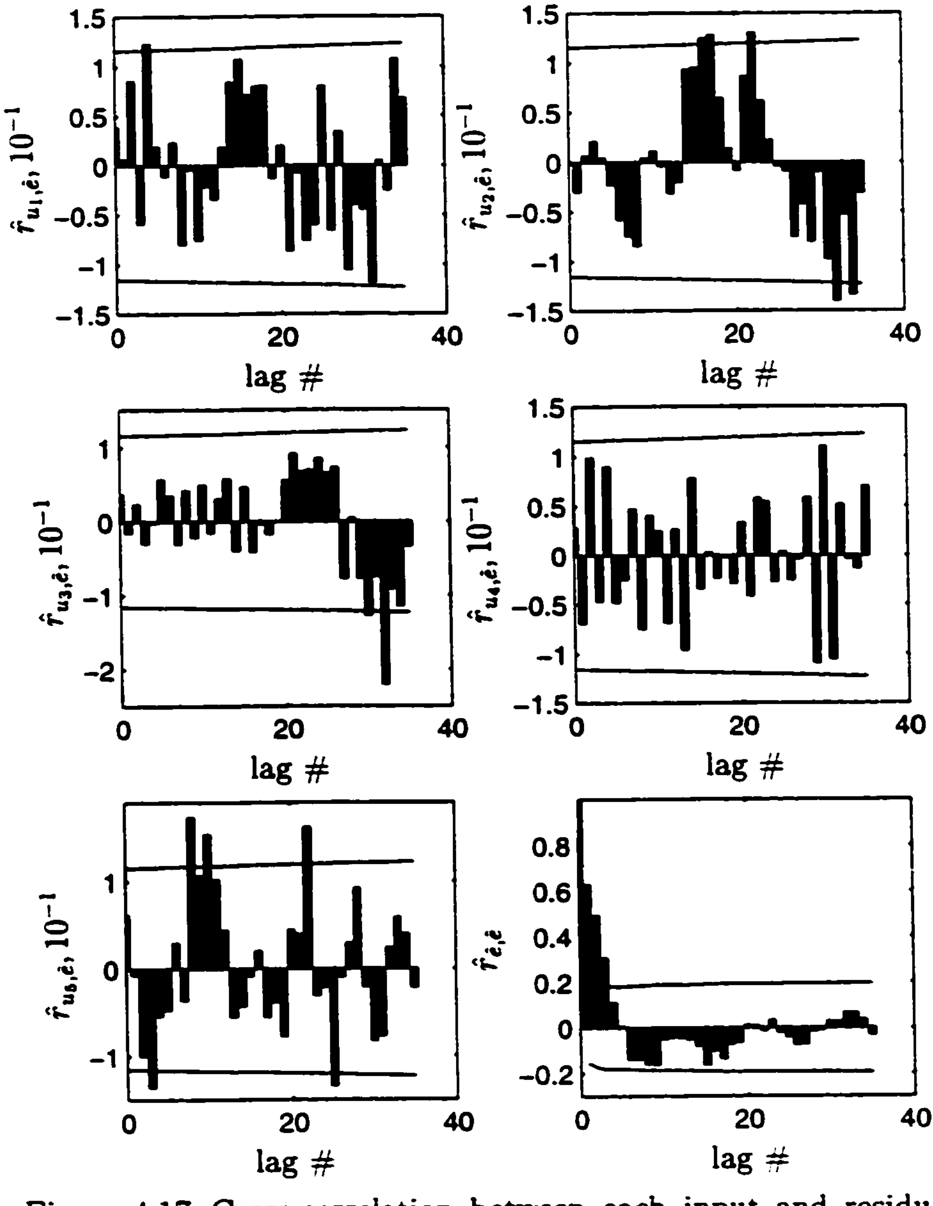


Figure 4.17: Cross-correlation between each input and residuals and auto-correlation of the residuals for the FIR model of section 4.5.2, estimated with 2 latent variables (cross-validation).

When selected with cross-validation ($a = 2$), the FIR model for transfer functions 1-3 and 5 show serious lack of fit (Figure 4.13). For such serious lack of fit one would have expected to see all the cross-correlation tests on the residuals clearly violated. However, only the cross-correlation plot for input 5 might lead one to suspect an inadequate model for this input. The reduced information content of the data (low signal to noise ratio and small number of observations) could be responsible for failure to detect the lack of fit. To check this, the same study was reproduced using 1950 observations, but similar results were obtained. The inability of the cross-correlation tests to detect even gross model inadequacies such as these appears to arise from the presence of so many adjustable parameters in the non-parsimonious models. Even though each input model is very inadequate the residuals are small, not highly autocorrelated, and not highly cross-correlated with any of the inputs. It seems that inadequate fits of some inputs can be compensated by the inadaquate fits of other inputs with the result that one still obtains small uncorrelated residuals. Therefore, the Box and Jenkins diagnostic tests can not be used reliably to select non-parsimonious dynamic models.

Using 1950 observations allowed us to assess the sensitivity of the Jackknife stability profile to the number of observations. The shape of the resulting profile was essentially the same as the one shown in Figure 4.10, and led to the selection of about the same number of latent variables.

4.6 Conclusion

A new criterion based on the Jackknife or the Bootstrap statistics has been developed in this work to improve the selection of non-parsimonious dynamic models, built for control system design or simulation purposes. The procedure is applicable for selecting the meta parameter of latent variable regression methods (number of latent variables) and regularized least squares regression methods (the ridge parameter). This methodology also has the advantage of providing estimates of the standard error and approximate confidence intervals for the impulse and step responses. The performance of the proposed criterion has been illustrated and compared to cross-validation through various identification case studies, in which PLS was used for parameter estimation.

Many papers have shown the inadequacy of cross-validation to guide in choosing a model capturing the right process structure. This is very important when the models are to be inverted as they are in controller designs or used in simulation with different input structures. The proposed criterion is based on selecting a model having both a low sum of squares of the residuals and stable parameter estimates. The procedure provides a stability profile for the model parameters, as a function of the number of latent variables. Several cases were simulated to illustrate the use of the stability profile. In essentially all cases investigated the parameter stability profile selected the number of latent variables to be very close to the number which minimized the mean square error for the impulse and step response models. This results in a model capturing most of the process structure. This approach outperformed the use of cross-validation in all cases and also gave better results than OLS/PEM. It also shows when the latter approach can be used without too much loss.

The Box and Jenkins auto and cross-correlation diagnostic tests were also investigated as a mean of selecting the number of latent variables. However, they were shown to be unreliable for providing evidence for lack of fit. This seems to be attributable to the large number of parameters estimated in the non-parsimonious models. As a result they appear to be of little use for model validation and for determining the number of latent variables when non-parsimonious models are being used. The proposed stability profile approach suggested in this chapter seems to be one of the few reliable methods for selecting the number of latent variables required to capture the model structure.

The use of the Jackknife is generally preferred to the Bootstrap since it is less computationally intensive and because of its algorithmic similarity with cross-validation, which is already implemented in most PLS software packages.

Chapter 5

Defining Multivariate Specification Regions

5.1 Introduction

Developing meaningful specification regions for selecting lots of raw materials entering a consumer's plant is essential to ensure that consumer's desired final product quality is achievable. Furthermore, the economic consequences of being able to establish meaningful specification regions are enormous. For the supplier of materials, if he can establish and meet specification regions that will consistently ensure consumer satisfaction, he can potentially gain a significant increase in market share. For the consumer, knowing what specifications he must place on incoming materials to ensure smooth operation of his process and to achieve the required quality of his own product will allow him to easily select suppliers who can meet these specifications or work with suppliers to achieve them. The resulting improvement in his product quality should also allow him to increase market share of his product. In addition, by more precisely defining raw material specifications needed for his process he may potentially be able to expand his supplier base and accept lower cost materials that are perfectly satisfactory for him. However, in spite of their importance, no standard industrial practice seems to exist for defining raw material specifications. They are rather defined arbitrarily based on past and often subjective experience. For instance, one practice is to set tight limits on important properties, to a range correponding to the best supplier's quality. Specifications are not only based on achievable final product quality but also account for process operating ability. For example, in polymer extrusion and film blowing, specification on raw polymer melt index MI (measure of viscosity) is largely based on the ability of the equipment to process the polymer easily.

Very few papers discussing issues around setting specifications on incoming raw materials have been found in the quality control literature. The importance of specifications for selecting good quality materials is always emphasized (Brinkley, 1991; Yarborough, 1995a and 1995b; Redlich, 1996), but methods for computing them are never discussed. The focus of these papers is, most of the time, on general quality management issues. Although literature exists for defining quality via univariate measures, using concepts such as loss functions (Taguchi, 1986; Fathi, 1990) and desirability functions (Harrington, 1965; Brown, 1990), these always assume that the product specifications are already available. Process capability indices (C_p, $C_{p,k}$, $C_{p,m}$) are also used to provide measures of the ability of a process to meet certain specifications. Therefore there is a need to develop a methodology for defining raw material specification regions on a sound analysis of supplier and consumer data bases.

In general, it is also assumed that raw material quality can be assessed univariately, using a one variable at a time approach. This is valid when the raw material properties of interest are independent from one another. In reality, the various properties one can measure on any given product are often highly correlated. Polymers, for instance, could be characterized by several viscosity measurements at different temperatures or shear stresses, or by melt flow indexes and density, which are all correlated measurements. In the manufacture of synthetic fibers quality measurements often consist of extentional measures under various loads, the elongation at break, and the load at break, all of which are point measurements along a stress \ strain curve. These measures are highly correlated with one another and with other quality measures such as denier, etc. In that situation, product quality is determined by the joint values of all measured properties. In other words quality is truly a multivariate property, requiring the correct combination of all the measured characteristics. When univariate specifications are used to select raw material having correlated properties, significant amounts of material may be misclassified. This can be discussed by reference to Figure 5.1.

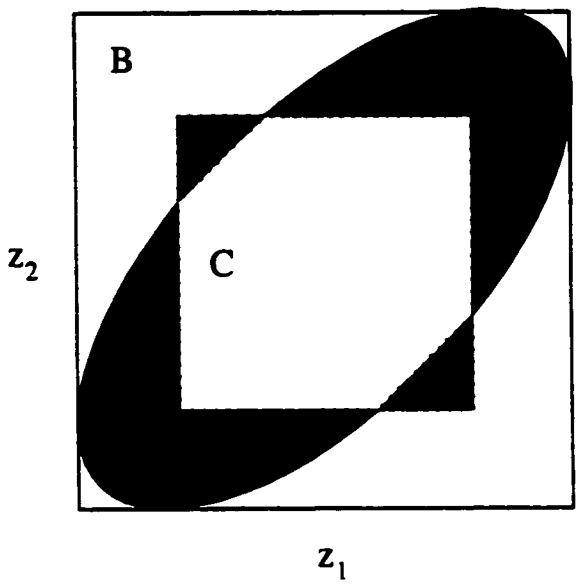


Figure 5.1: Problem of using univariate specifications on correlated raw material properties (z_1 and z_2).

Consider z_1 and z_2 to be two correlated raw material properties. Also assume the elliptical region "A" to be the true multivariate region within which raw materials are good quality. The square region "B" is a corresponding univariate specification region, accepting the same variance on each individual property as the multivariate region. Clearly, all raw materials falling into the region included in between the elliptical region "A" and the square region "B" correspond to accepting poor quality material. To improve the selection process, one could shrink region "B" to the square region "C". A lot less poor quality material would be accepted, but all the good quality material falling into areas of the ellipse located outside region "C" would be rejected. Indeed, the gray zones in Figure 5.1 all correspond to misclassified raw material quality.

To the author's knowledge, the only approach for defining multivariate specifications on raw materials has been proposed by De Smet (1993). Partial Least Squares regression (PLS) is first used to build a model between raw material properties and consumer's final product quality, available from historical data bases. In a second step, a region in the projection space of the model is defined to include most observations associated with good final product quality. This multivariate region can then be used to monitor and select new lots of raw materials entering the plant. The potential of this methodology has been illustrated with a simulation and an industrial example. However, several issues with this approach were not resolved. Indeed, this method assumes that the consumer's final product quality is out of control mainly due to poor raw material quality. However, a more general and realistic situation is that both supplier raw material variations and variable process operation are responsible for poor product quality. In addition, feedforward and \ or feedback control actions from operators or control systems are often implemented on such processes. When process variations and control actions have an important impact on final product quality, it becomes more difficult to build a satisfactory model for the effect of raw materials on product quality. The specification regions or incoming materials required by any consumer will depend very much on how he operates his own process. Uncontrolled variability in his process will inflate his product quality variations and require tighter specifications on the supplier raw material in order to achieve his final quality specifications. On the other hand, the use of good feedforward and feedback control in the consumer's process will compensate for some of the raw material variations and allow for wider specifications on incoming raw materials.

The objective of this research is to generalize the methodology proposed by De Smet (1993) to account for more industrial situations. The concepts are illustrated using a detailed simulator of a film blowing process, aimed at producing various kinds of polymer films.

The chapter is divided as follows. Data requirements for defining specifications are first discussed, followed by a description of the simulation studies. The proposed methodology is then described and illustrated.

5.2 Nature and Type of Data

The data necessary for developing multivariate specification regions involves three blocks, **Z**, **X** and **Y**, as shown in Figure 5.2. The type of measurements to be included within each block is discussed below.

Z $(I \times L)$:

- This block includes a total of L measurements characterizing the quality of each of the I lots of raw materials sent to the consumer. The set of measurements may not totally define the effect of raw materials on consumer's final product quality, but forms a currently accepted set of measurements for raw material quality. Specification regions will be developed for these measures.

$\mathbf{X}$ $(I \times J)$:

- Steady-state processing conditions used to process each of the I lots of raw materials should be summarized in this data block. It consists of averaged measurements collected on a total of J manipulated variables and independent disturbance variables in the consumer's plant. Any other type of process measurements should not be included in this block to avoid potential correlation to one or some raw material properties, which would lead to a misleading analysis. Only correlation allowed between $\mathbf{Z}$ and $\mathbf{X}$ is through feedback or feedforward control actions.

$\mathbf{Y}$ $(I \times M)$:

- It is assumed here that consumer's final product quality and process operability measures can be adequately assessed and characterized by a set of M measurements. These measures will be used to assess the results of variations in $\mathbf{Z}$ and $\mathbf{X}$.

Collecting the data base shown in Figure 5.2 implies that one is able to trace the processing conditions used for each lot of raw materials, and the final product quality that resulted from these process conditions and lots of raw materials.

In industrial data bases, variations in final product quality ($\mathbf{Y}$) due to raw materials may be overwhelmed by variations due to the process and disturbances ($\mathbf{X}$). In such a situation, it would be difficult or even impossible to develop specifications for raw material properties as their effect on final product quality would be masked by the effects of process variations and measurement noise. To overcome this typical situation, multivariate design of experiments (Wold *et al.*, 1986; Kettaneh-Wold *et al.*, 1994) could be used to supplement variations in $\mathbf{Z}$, and therefore identify its effect on $\mathbf{Y}$. Raw material properties can not be designed. However, suppliers often have some lots of raw materials not sent to consumer for some reason (thought as not appropriate for this particular consumer). The idea of multivariate design of experiments consists of selecting, among these lots of raw materials, a few that have a greater span of the raw material space than the historical data on raw materials already used by the consumer. To assist in the selection process, Principal Component Analysis (PCA) can be used. These concepts are discussed further and illustrated in section 5.4.2.

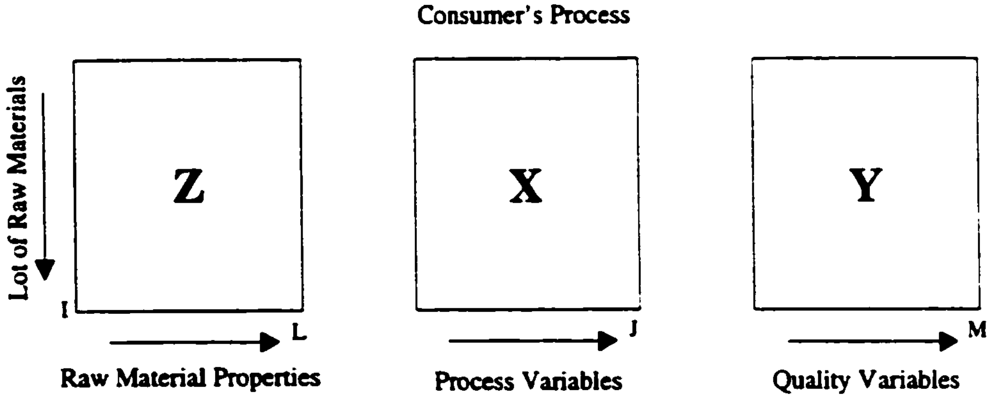


Figure 5.2: Data collected by the consumer.

5.3 Simulation of a Film Blowing Process

To illustrate the proposed methodology for setting multivariate specifications on raw materials, a first-principles based simulator of a film blowing process is used (Sidiropoulos, 1995). The simulator, B-Filmcad 1.0, was provided by Polydynamics Inc (Polydynamics, 1996). A schematic representation of the film blowing process is shown in Figure 5.3. A raw polymer is fed into an extruder for mixing and melting. After a given residence time, the polymer melt reaches a die shaped as an annuli. Often, multi-layer films with concentric annuli and different polymers in each are produced, but here we consider only a single layer film with one annulus. At the die exit, air is blown inside the extruded polymer to obtain a given inflation pressure. This allows the formation of a bubble-shaped film of a desired internal diameter. The film of polymer is then cooled while being continuously conveyed to cutters and other finishing devices. The cooling process is controlled by blowing ambient air on the outer surface of the film.

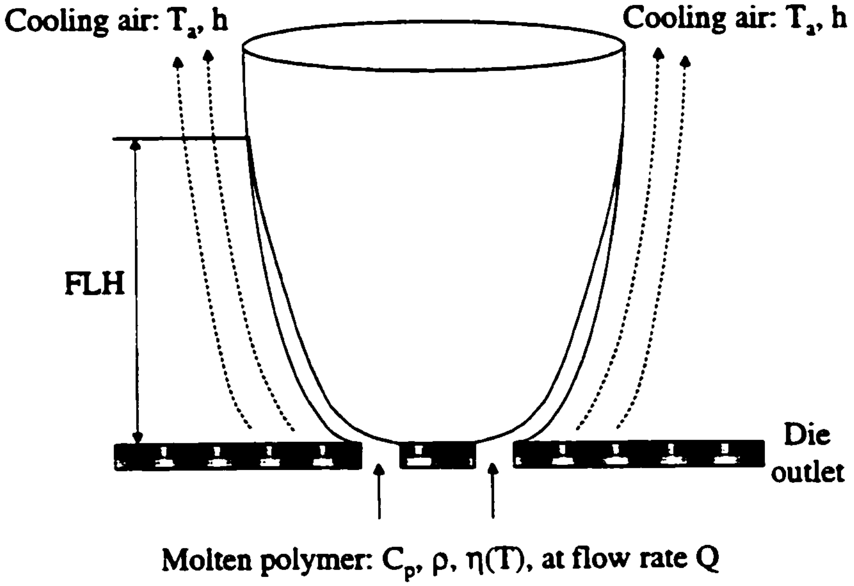


Figure 5.3: Typical film blowing process.

This is necessary to achieve desired film properties. After a certain distance from the die exit, called the frost line height (FLH), the final film properties are determined and remain constant. The data generated using the simulator consists of 12 raw polymer properties (**Z**), 3 process variables (**X**) and 2 film properties (**Y**). The heat capacity (C_p) and the density (ρ) of the polymer and the dependence of the raw polymer viscosity on temperature as given by Equation 5.1, are assumed to be adequate for characterizing the raw polymer quality.

$$\eta(T) = \eta_f \; exp\left[-a\ (T - T_f) + c\ \left\{\frac{1}{(T - T_s)^d} - \frac{1}{(T_f - T_s)^d}\right\}\right] \tag{5.1}$$

Parameters a, c and d are characteristics of the polymer (fitting parameters). η_f is the polymer viscosity at the reference temperature T_f, and T_s is its solidification temperature. To include the dependence of viscosity on temperature as an effect (**Z**), the viscosity at 10 different temperatures along the curve were collected. The effect of raw material variations on film quality is assumed to be captured by measuring C_p, ρ and 10 values of $\eta(T)$ for each lot of raw polymer.

Important process manipulated variables are the polymer flow rate (Q) and the air flow rate. However, the latter is replaced by the maximum local heat transfer coefficient (h_o) along the bubble, since the simulator does not allow modifications to the air flow rate. The last process variable considered is the ambient air temperature (T_a), which is a major process disturbance affecting cooling conditions and hence, film properties.

Film quality is assumed to be characterized by the full stress in the machine direction ($FMDS$), taken beyond the frost line, and the frost line height itself (FLH). The FLH is not a true film property, but is important to control since the film stresses are related to this variable.

A total of 3 cases are studied. Each consists of a data base similar to the one shown in Figure 5.2, including 50 lots of raw materials. To simulate different lots of raw materials, the values of the parameters of the viscosity function (η_f, a, c and d) and the values of polymer properties C_p and ρ were generated as random draws from normal distributions, with means equal to some nominal conditions and standard deviations as given in Table 5.1. The nominal values chosen for the raw material properties correspond to a LDPE polymer (Sidiropoulos, 1995). For example, the values for one lot of raw material generated in Section 5.4.1 are $[\eta_f \; a \; c \; d \; T_f \; T_s \; C_p \; \rho] =$ [19831 0.017871 6.06 0.992 180 95 2134 1042]. These values are used to obtain one row of **Z**. Part of the variations in processing conditions were obtained in a similar fashion, but since Q and h_o are manipulated in Sections 5.4.2 and 5.4.3 to control film properties, their corresponding standard deviations shown in Table 5.1 are the result of both random variations and control actions. More details concerning process variations are provided in Sections 5.4.1 to 5.4.3. For the parameters and properties of the lot of raw material discussed previously, the corresponding row of **X** (processing conditions) is $[Q \; T_a \; h_o] =$ [110 20 22]. Finally, the corresponding film properties (row of **Y**) is obtained by implementing the values of **Z** and **X** into the simulator.

For each viscosity curve, polymer viscosity was measured at the following 10 temperatures: 96, 97, 102, 103, 137, 138, 146, 147, 194, 195 °C. Raw material properties (**Z**) are the same for all cases studied unless otherwise stated. The process was constrained to always produce a film of a thickness of 0.088 mm and a blow-up ratio of 1.8 (ratio of bubble diameter to die diameter), using a die temperature of 195°C (i.e. process already has perfect control on these properties). Further discussion on variations implemented on parameters of Table 5.1 are specific to each case study, and are defered until they are presented in later sections of this chapter.

Random independent measurement noise were added to **Z** and **Y** variables. The standard deviation of the noise added on each variable is about 10% of the standard deviation of the noise-free variable. However, since viscosity measurements belong to the same curve, and are therefore highly correlated, the noise sequences added to these measurements are perfectly correlated, but still with standard deviations of about 10% of the standard deviation of noise-free viscosities. Such correlated noise structure would make sense, for instance, if all the errors were coming from a lab bias (different viscometer settings, bias in temperature sensors changing between

Table 5.1: Nominal conditions and standard deviation between lots of raw material properties and processing conditions in each of the three simulation case study.

		$\hat{\sigma}$		
Parameter	Nominal Value	Section 5.4.1	Section 5.4.2	Section 5.4.3
η_f (Pa-s)	16077	2429.7	2429.7	2429.7
a (-)	0.02113	0.00306	0.00306	0.00306
c (-)	5.71	0.86	0.86	0.86
d (-)	1.0	0.17	0.17	0.17
T_f (°C)	180	0.0	0.0	0.0
T_s (°C)	95	0.0	0.0	0.0
C_p (J/kg-°C)	2300	183.9	106.5	183.9
ρ (kg/m^3)	900	141.6	141.6	141.6
Q (kg/h)	110	0.00	8.34	2.34
T_a (°C)	20	0.00	6.70	6.63
h_o (W/m^2-K)	22	0.00	3.20	6.13

days, etc.). This is not entirely realistic, but this is used for illustration purposes only.

5.4 Methodology for Developing Specifications and Illustrations

A sound data-driven methodology to develop specifications on raw materials is to directly take into account their effect on consumer's final product quality. For raw material properties appearing as weakly or not related at all to final quality, no specifications are required. However, for those properties or combination of properties that seem strongly related to quality, tighter specifications are needed. Therefore, one needs a reliable mapping between quality space (**Y**) and raw material space (**Z**). This is not always obvious to obtain from historical data as variations in the consumer's process may also significantly affect final product quality.

Disturbances and operator actions may introduce process variations that lead to larger variations in product quality. On the other hand, operator or control systems actions may also be aimed at rejecting variations due to raw materials and other disturbances using feedback and feedforward control schemes, and may lead to reduced variations in product quality. To develop raw material specification properly, one needs to make an assumption on the type of process variations contained in the historical data base. In addition, an assumption on future process behavior is also required. Most practical situations can be classified in three categories according to the type of process variations that are present: (i) no significant process variations affecting quality; (ii) significant process variations affecting quality, but these variations are uncorrelated with raw materials; (iii) significant process variations affecting quality, and these are correlated with raw materials because of feedback and feedforward control. The proposed methodology for developing specifications in each of these cases is described below and illustrated using simulations of the film blowing process described in section 5.3.

5.4.1 Absence of Significant Process Variations Affecting Quality

This is the simplest situation in which the key assumption is that consumer's quality variations are due only to raw material variations. The objective is therefore to develop a specification region for raw material properties, such that most lots of raw materials leading to poor quality are rejected.

Consider the simple example depicted in Figure 5.4, in which specifications need to be developed for 2 raw material properties (z_1 and z_2) based on 2 quality characteristics (y_1 and y_2). In the $y_1 - y_2$ plot, dots identify good consumer's final product quality, while crosses correspond to poorer quality products. This would, in general, be addressed by quality control personnel. The general idea for developing the specification region for z_1 and z_2 consists of mapping the points corresponding to good quality from y-space (dots) into raw material property space (see the z_1 – z_2 plot in Figure 5.4). Then, an elliptical boundary, $(\mathbf{z}-\bar{\mathbf{z}})\,\hat{\sum}_z^{-1}\,(\mathbf{z}-\bar{\mathbf{z}})^{\mathrm{T}} = c_z$, is used to define the region in z-space where fall most of the raw material properties associated with good consumer's final product quality. In the elliptical region expression, $\hat{\sum}_z$ is the covariance matrix of raw material characteristics, estimated using only the raw material properties associated with good consumer final product quality (dots), and c_z corresponds to the size of the specification region. The latter can be selected to minimize the total number of misclassified points, n_t (e.g. sum of dots falling outside and crosses falling inside the elliptical region). When more than 2 z–variables are measured, the specification region can be defined by setting an upper limit on the Hotelling's T^2 (Anderson, 1984), $T^2 = (\mathbf{z}-\bar{\mathbf{z}})\hat{\sum}_z^{-1}(\mathbf{z}-\bar{\mathbf{z}})^{\mathrm{T}}$. Therefore, the limit set on T^2 is simply the value selected for c_z.

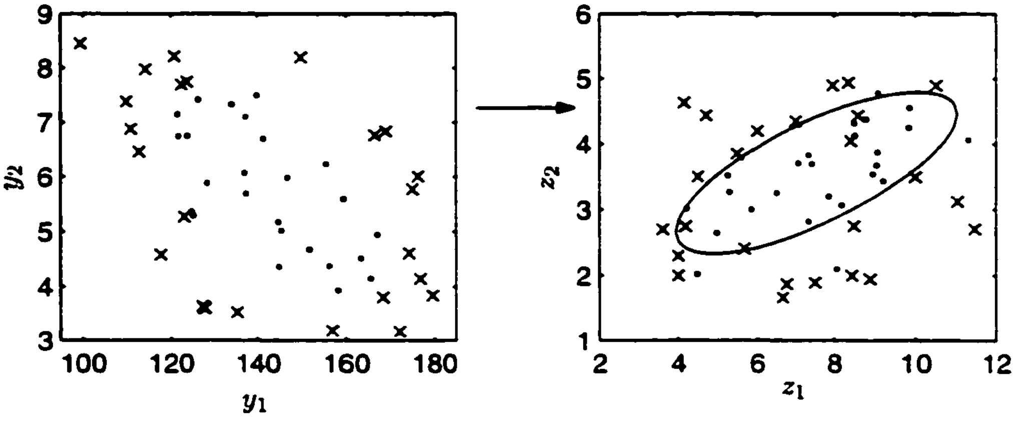


Figure 5.4: Simple example of developing specifications for 2 raw material properties z_1 and z_2, based on 2 quality variables y_1 and y_2.

Since a range of c_z values may exist to achieve nearly the same minimal value for n_t, another criterion can be used to select a more precise value for c_z within that range. It consists of choosing the value that best balances type I and type II errors, defined as follows. Type I error (or producer's error) consists of the proportion of truly good lots of raw materials that would be rejected by the consumer, under a given specification region, among all truly good lots of raw materials sent to the consumer. On the other hand, type II error (or consumer's error) is defined as the proportion of truly poor lots of raw materials that are accepted by the consumer, under a given specification region, among all the truly poor lots of raw materials sent to the consumer. It is important to note that both errors are competing with one another. Reducing the area, the volume or the hypervolume of the raw material specification region would achieve a lower type II error, at the expense of a higher type I error and vice-versa.

This simple approach, illustrated for the case where only 2 z and 2 y variables are available (Figure 5.4), becomes much more difficult when many raw material properties and many final consumer's product quality characteristics are measured (several z's and y's), and also when these spaces are not full rank. In such a situation, one needs a model to define the region in $\mathbf{Z}$ space that is important to quality, $\mathbf{Y}$. Latent variable regression methods such as Principal Component Regression (PCR), Canonical Coordinate Regression (CCR) and Projection to Latent Structures (PLS) can be used for that purpose. In particular, it would make sense to use PLS since it models high covariance directions within $\mathbf{Z}$ and $\mathbf{Y}$, as well as the covariance structure in between these data blocks, via a few latent variables $\mathbf{T}$. The structure of the PLS model is shown below:

$$\begin{aligned} \mathbf{Z} &= \mathbf{T}\,\mathbf{P}^{\mathsf{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T}\,\mathbf{Q}^{\mathsf{T}} + \mathbf{F} \\ \mathbf{T} &= \mathbf{Z}\,\mathbf{W} \end{aligned} \tag{5.2}$$

where the columns of $\mathbf{W}$ are just the linear combinations of raw material properties

(**Z**) that are important for final consumer's product quality (**Y**). The number of these combinations, or number of latent variables A, can be computed using the approaches described in the background section (chapter 2). These latent variables define a reduced space of dimension A, summarizing the raw material property space **Z**. The projection of **Z** onto that space is denoted as **T**. Covariance structures within **Z** and **Y** are defined by **P** and **Q** respectively. The reader is referred to the background chapter 2 for further details on the PLS model. One could therefore map good quality points from the **Y** space to the reduced space of the model (**T**), and then use the same methodology as discussed before to obtain the elliptical specification region. If the number of latent variables, A, is greater than 2, one could again define the specification region by setting an upper limit on the Hotelling's T^2, applied on the latent variables (e.g. $T^2 = \sum_{j=1}^{A} t_j^2/s_{t,j}^2$), where $s_{t,j}^2$ is the estimated variance of t_j.

Since specifications are defined in reduced space **T** instead of in the original raw materials space **Z**, one must also ensure that new incoming lots of raw materials are also well summarized by the PLS model. A valid specification in reduced space involves not only monitoring the projection of the properties from new incoming lots of materials (**T**), but also requires monitoring the distance of this projection from original measurements. For the i^{th} lot of raw material having measured properties $\mathbf{z}_i$, the distance of these measurements from the reduced space **T** is given by the square prediction error $SPE_i = (\mathbf{z}_i - \hat{\mathbf{z}}_i)^\top(\mathbf{z}_i - \hat{\mathbf{z}}_i)$, where $\hat{\mathbf{z}}_i$ is simply the projection of $\mathbf{z}_i$ onto the reduced space (**T**): $\mathbf{z}_i = \mathbf{t}_i\ \mathbf{P}^\top + \mathbf{e}_i = \hat{\mathbf{z}}_i + \mathbf{e}_i$. When this distance is higher than a given limit, the new lot becomes suspicious. This suggests that its properties reflect a different correlation structure than seen before in the historical data base. It is then impossible to predict the impact of this lot of raw materials on final product quality. It might be safer to reject it, even if the projection of its properties into reduced space of the model fall in the region of good quality raw materials. Upper limits on SPE are computed assuming that it is distributed approximately as $g\ \chi^2(h)$, that is a multiple of a χ^2–distribution with h degrees of freedom (Nomikos and MacGregor, 1994*a*). The two parameters g and h and estimated by matching the first two moments of the distribution using the historical data base.

In this work, a 95% limit was used to monitor the SPE of new incoming lots of raw materials. In the PLS model building stage, however, a higher limit (99%) is often used to avoid removing useful information (only extreme outlier points are removed). If such a high limit on SPE was also used for monitoring new incoming lots of raw material, consumers would have a higher chance to accept poor lots of raw materials and a higher type II error could result. Since there is no need to use the same limit on SPE for the model building stage and the acceptance stage, we will use a 95% limit for the specifications in order to reduce the chances of accepting poor materials.

Illustration

To illustrate the definition of raw material specification regions when raw materials are the major source of variation affecting quality, data was generated as described in Table 5.1. Processing conditions (**X**) were maintained constant at nominal conditions for all lots of raw materials and so, **X** is not required for this analysis. Two examples are shown in this section, both requiring latent variable models since a large number of highly correlated **Z** variables are measured. In the first example, all 12 raw material properties are assumed measurable and are measured. However, in the second example, variations in ρ are still implemented, but this property is assumed not measurable. This is to illustrate the practical problem that not all important properties of materials are measured or are measurable. The results of both examples are compared below.

A summary of the PLS modelling results built between raw material properties (**Z**) and the 2 quality variables (**Y**) for both examples is presented in Table 5.2. Shown

Table 5.2: Summary of the PLS modelling results, for the situation where variations in quality **Y** are due only to raw materials **Z**.

Example	A	$R^2_{Z,cum}$ (%)	$R^2_{Y,cum}$ (%)	$Q^2_{Y,cum}$ (%)
all properties	4	89.4	96.8	95.5
ρ unmeasured	2	76.1	75.4	72.2

are the number of components (A) kept in the model, as determined by leave-one-out cross-validation, as well as three cumulative multiple correlation coefficients, $R^2_{Z,cum}$, $R^2_{Y,cum}$ and $Q^2_{Y,cum}$. Cumulative R^2 values give the percentages of the total sum of squares of **Z** and **Y** that are explained by the fitted PLS models with the indicated number of dimensions. Cumulative Q^2_Y value is the percentage of the total sum of squares of **Y** that can be predicted with these models using a leave-one-out cross-validation procedure. Cross-validation was used throughout this chapter for selecting the number of PLS components and the procedure was performed using the SIMCA-P 7.0 software (Umetrics, 1998). For the situation where all **Z** properties are measured, a very good model is obtained with high amount of explained and predicted variance in **Y**. However, when ρ is not measured, only 2 latent variables are found significant (instead of 4) and a much poorer model is obtained, in terms of variance explained and predicted in **Y**. Since variance in **Y** due to ρ is left unexplained, the specification region for **Z** in this case should lead to a higher number of misclassified points and to higher type I and type II errors.

Quality measurements (**Y**) and the specification region for **Z** for both examples are shown in Figure 5.5 and 5.6 respectively. Assume that based on consumer's judgement (involving quality characteristics FLH and $FMDS$), 29 batches of final product were identified as good quality (dots), while 21 were identified as poorer quality (crosses). Since 4 latent variables were found significant to model variations in **Z** that are important for **Y**, the specification region in projection space (**T**) for the first example is based on the Hotelling's T^2. The region developed for the second example is shown by the means of an ellipse, since the PLS models has only 2 dimensions. To select the specification limit in both cases, different values of c_T were assumed and the corresponding number of misclassified points (n_t) and type I and type II errors were computed. The results are shown in Table 5.3. When all properties are measured, the total number of misclassified points is minimal for $c_T = 6.0$, while type I and type II errors are low and well balanced. This value was therefore imposed as the limit on T^2. On the other hand, a value for c_T in between 4.1−3.9 can be chosen for the situation where ρ is unmeasured, since these values are equivalent in terms of number of misclassified points and type I and type II errors. We have selected $c_T = 4.0$ in this case. Note that in both Figure 5.5 and 5.6, some observations appear to have higher distance to the model SPE_Z (observations 1, 23, 49 and 50 for the first example and observations 10 and 22 for the second example), and these would have been rejected if this was an on-line monitoring situation. Therefore, observations having a SPE value higher than the 95% limit were never used in computing $\hat{\sum}_T$, n_t and the type I and type II errors, and this holds for all case studies presented in this chapter.

When comparing both examples, the key result is that when ρ is not measured (but varies) a higher number of misclassified raw materials is obtained ($n_t = 7$ compared to 2), as well as larger type I (6.89% compared to 3.70%) and type II errors (26.31% compared to 5.26%). Indeed, poor final products quality (crosses) may be obtained because of too large or too low density (ρ), but with a correct combination of all the other raw material properties. If ρ is not measured, one would never have evidence that poor product quality could have been caused by raw materials.

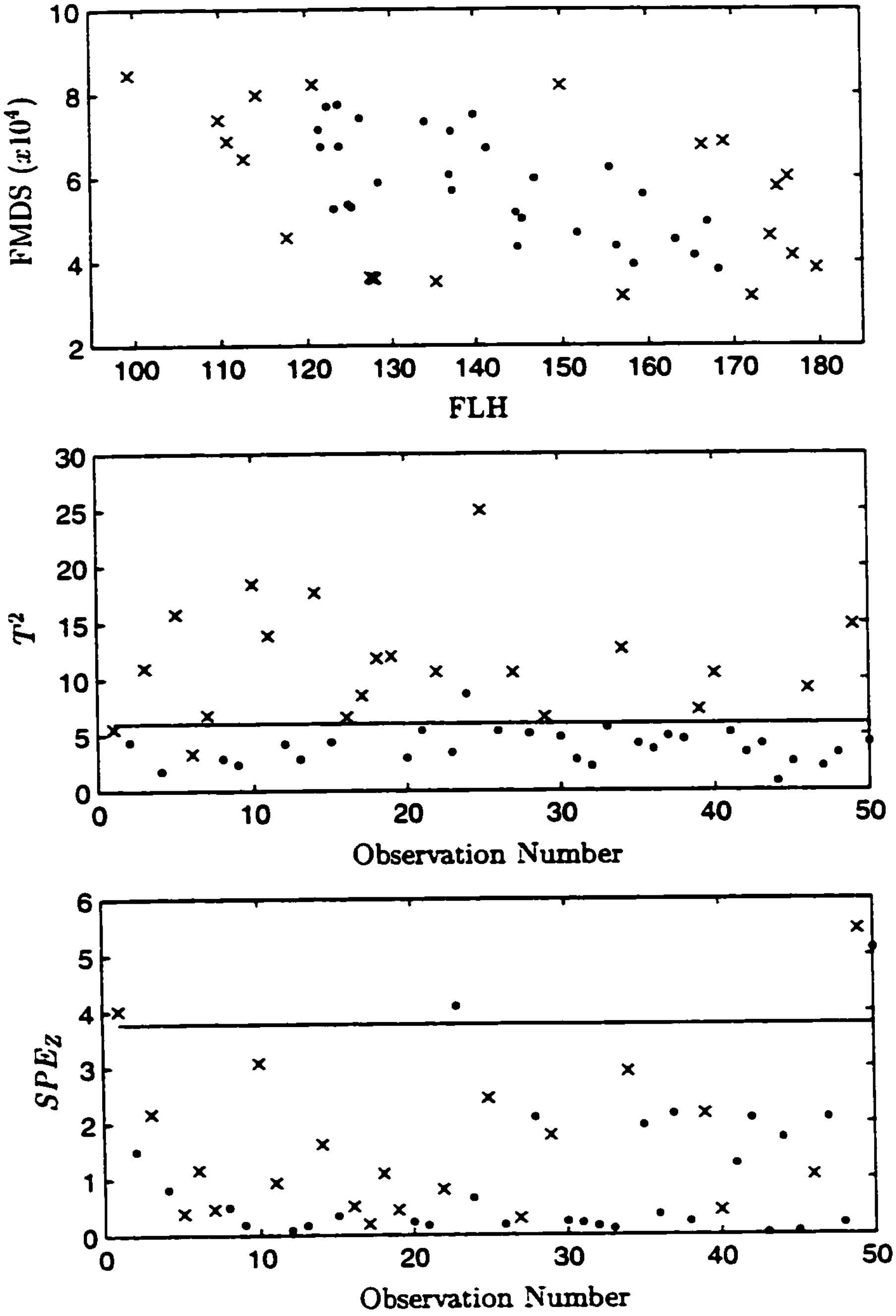


Figure 5.5: Specification region on raw material properties, for the situation where variations in $\mathbf{Y}$ are mainly due to raw materials and all 12 raw material properties are measured.

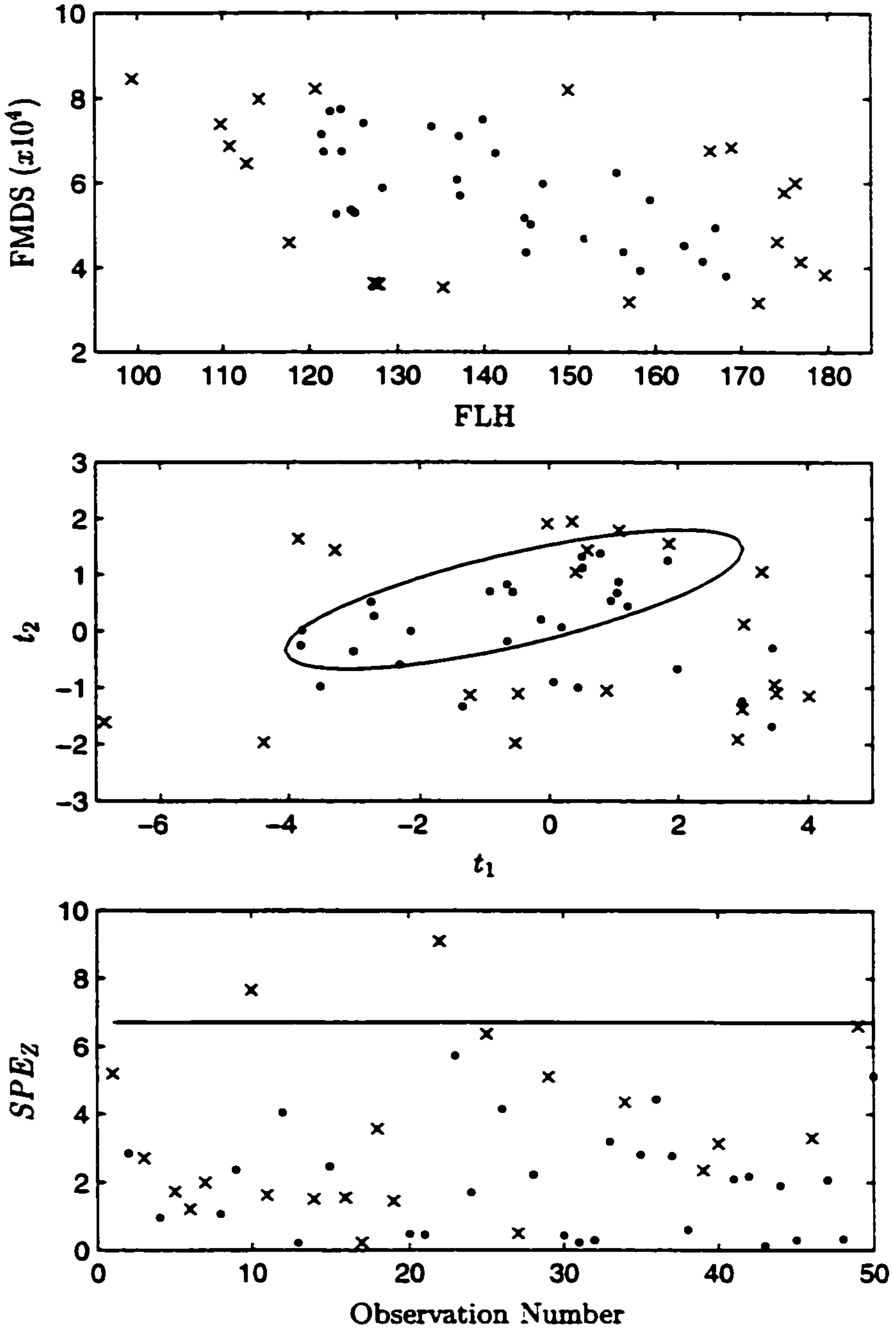


Figure 5.6: Specification region on raw material properties, for the situation where variations in **Y** are mainly due to raw materials, but ρ is unmeasured.

Table 5.3: Selection of the specification limit in projection space, c_T, in terms of total number of misclassified points, n_t, and type I and type II errors.

ρ measured				ρ not measured			
c_T	n_t	Type I (%)	Type II (%)	c_T	n_t	Type I (%)	Type II (%)
10.0	7	0.00	36.84	7.0	17	3.45	84.21
9.0	6	0.00	31.58	6.0	16	3.45	78.95
8.0	6	3.70	26.32	5.0	12	3.45	57.89
7.0	5	3.70	21.05	4.2	8	6.89	31.57
6.0	2	3.70	5.26	4.1	7	6.89	26.31
5.0	7	22.22	5.26	4.0	7	6.89	26.31
				3.9	7	6.89	26.31
				3.0	9	13.79	26.32
				2.0	12	34.50	10.53

5.4.2 Significant Process Variations Affecting Quality, but Independent of Raw Material Properties

This section is concerned with developing specification regions for raw materials (**Z**) when variations from the process (**X**) are also significantly affecting quality (**Y**). However, it is assumed here that process variations are independent of the variations in raw materials. In other words, this means that there is no feedback or feedforward control efforts made, either by operators or control systems, to compensate for variations in raw materials.

It is important to recognize here that whether or not process variations affecting quality can be removed in future operation will have an important impact on the specification regions for raw materials. In the next paragraphs, development of specification regions is addressed for the situation where process variations will remain in future operation, and the situation where these variations can be removed.

Process variations will remain in future

If one assumes that process variations can not or will not be removed in the future, the objective for developing specification regions for raw materials is to achieve desired final product quality most of the time, in spite of process variations that also affect quality. Under the assumption that future process variations will remain consistent with the past (e.g. as seen in the historical data base), these variations are ignored, and treated as an unexplained source of variance in quality, and specification regions can be developed just as described in section 5.4.1. This should, in general, lead to a greater number of misclassified points since poor final product quality can either be caused by raw materials or by processing conditions. For example, one could process a truly good lot of raw material, but according to the proposed methodology, this lot would be classified as poor whenever processing conditions are such that final product quality is poor. Tighter specification regions than when part or all process variations can be removed in the future will be obtained, since these specification regions are shrunk to meet quality requirements after process variations are added.

As the magnitude of quality variations due to the process increases relative those due to raw materials, it becomes more difficult to build a good model between **Z** and **Y**. This is a typical industrial situation where raw material variations are overwhelmed by process variations. To improve the model, one may need to generate more information about the effect of **Z** on **Y**. This can be obtained by processing a few additional lots of raw materials, selected using multivariate design of experiments. This is exemplified later in this section.

Process variations can be removed in future

If process variations affecting quality can and will be removed (partly or entirely) in the future through redesign of the process or changing operating procedures, then one can accept more variations in raw materials to meet the same quality requirements.

Specifications on raw materials are therefore developed based on the quality variations that would have been obtained had process variations been removed in the past ($\mathbf{Y}^*$). This can be estimated as follows:

$$\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}\,\mathbf{B_X} = \mathbf{Z}\,\mathbf{B_Z} + \mathbf{F} \tag{5.3}$$

where $\mathbf{B_X}$ and $\mathbf{B_Z}$ are regression coefficients estimated using any latent variable regression method. $\mathbf{F}$ contains modelling errors (unmeasured disturbances, measurement noise, non-linearities and so forth). When applying the same quality judgement to $\mathbf{Y}^*$ (than for $\mathbf{Y}$), one should find that a different set of final products meet quality requirements than when the judgement is based on $\mathbf{Y}$. For example, if a good quality raw material was processed under poor conditions with the result that final product quality is poor, removing the effect of the process from measured quality $\mathbf{Y}$ should move estimated quality $\mathbf{Y}^*$ back into desired region. Specification regions developed based on $\mathbf{Y}^*$ should lead to a smaller number of misclassified points and lower type I and type II errors, since process variations are accounted for and are removed from $\mathbf{Y}$. The reduced variability in $\mathbf{Y}^*$ should also lead to larger specification regions. Note that one could also remove only a fraction of the process variation, $\mathbf{X}\,\mathbf{B_X}$, from $\mathbf{Y}$ and still use the same methodology to develop the specification regions.

Illustration

In this section, it is desired to illustrate one common practical situation where variations in quality due to the process overwhelms those caused by raw materials. To create such a situation, the standard deviation of variations in C_p were decreased by almost half compared to the case of section 5.4.1, while all other raw material properties remain the same (see Table 5.1). This decreases the magnitude of variations in quality (mainly in FLH) due to raw materials relative to variations caused

Table 5.4: Summary of the PCA and PLS modelling results, for the situation where variations in quality **Y** are due to raw materials **Z** and to the process **X**, but these sources of variations are uncorrelated.

		R^2_{cum} (%)			Q^2_{cum} (%)	
Model	A	Z	X	Y	Z	Y
PLS **Z-Y**	3	81.5	-	54.9	-	27.9
PCA **Z**	3	83.1	-	-	58.1	-
PLS **Z-Y** with MDOE	3	78.7	-	62.9	-	43.6
PLS **Z-X-Y**	6	87.5	96.8	96.4	-	91.0

by the process. Process variations consist of two types: random variations implemented on Q, h_o and T_a and feedforward compensation for variations in T_a, using Q and h_o simultaneously. Random variations on Q and h_o (manipulated variables) are used to mimic operator variations in implementing the processing conditions and in responding to different events. Variations in T_a simulate a measured process disturbance. This disturbance is rejected by a feedforward controller to maintain cooling conditions. For example, when T_a is higher, h_o is increased and Q is decreased. This leads to a greater cooling rate and a reduced cooling load and therefore, this maintains FLH within certain limits. The magnitude of random variations implemented on Q are threefold compared to feedforward variations, based on standard deviation. Random variations in h_o are about the same level as the feedforward changes. In this mode of operation the variations in the process variables **X** (Q, h_o and T_a) are highly correlated among themselves, but are not correlated with the raw material variables (**Z**).

When process variations can not or will not be removed in future operation, then process measurements **X** are ignored and one could identify the effects of **Z** on **Y** directly using PLS. A summary of modelling results is given in Table 5.4. However, since process variations account for an important portion of variations in quality, a poor model is obtained ($R^2_{Y,cum}$ = 54.9% and $Q^2_{Y,cum}$ = 27.9%), and almost no variations in FLH is explained because it exhibits less variation, since the feedforward control has reduced the variability arising from the process disturbance (T_a). If one still wants to develop a specification region, then more information needs to be generated for characterizing the effect of **Z** on **Y**. This can be achieved through multivariate design of experiments (refer to section 5.2). To select additional lots of raw materials to process, a PCA analysis was performed on the joint properties of past lots of raw materials received by consumer and properties of 8 new lots not sent to the consumer. Two components were found significant and the results of this analysis are summarized in Table 5.4 and in Figure 5.7. Observations numbered 51-58 correspond to the lots of raw materials not sent to the consumer. The idea of multivariate design is to select a few additional lots, among 51-58, that span a wider range in the projection space of past lots of materials. We have selected lots 51, 53, 55 and 56 (circled) on that basis. Note that lot 53 also has a high distance to the projection space ($DMODX$) and therefore will bring some new information that has never been seen in the past. This should help to achieve a better definition of the specification region.

After processing the 4 additional lots of raw materials (51, 53, 55 and 56), using processing conditions that are consistent with past operation, the resulting quality measurements were collected. The new **Z** and **Y** data were added to the existing data and a new PLS model was built on the joint data set. This model shows significant improvement in predicting variations in quality caused by raw materials ($Q^2_{Y,cum}$ = 43.6%), as summarized in Table 5.4, especially for variations in FLH. The results are summarized in Table 5.5. The left half of Table 5.5 shows that limits of c_T = 5.0 and c_T = 6.0 achieve the same value for n_t, but c_T = 5.0 leads to a better balance between the type I and type II error. This limit was therefore selected and the specification region for raw materials is presented in Figure 5.8. Multivariate

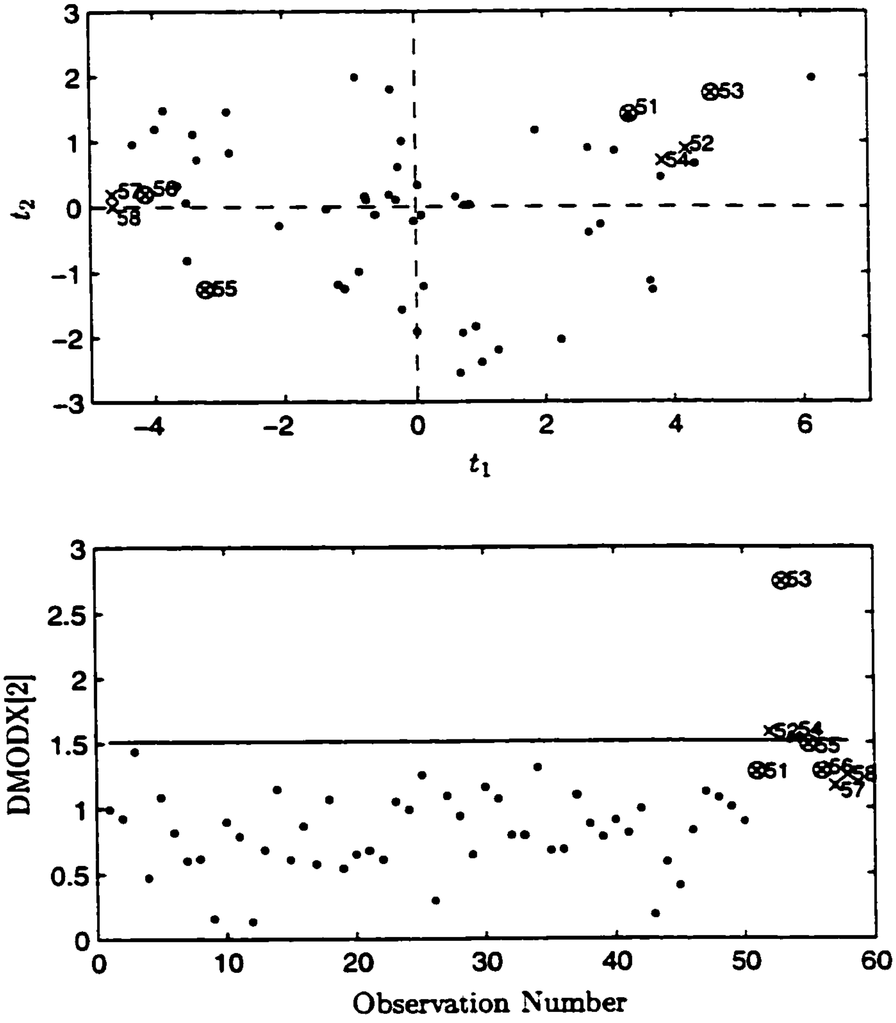


Figure 5.7: Results of a PCA analysis on the measured raw material properties, including some lots not sent by the supplier (51–58).

Table 5.5: Selection of the specification limit in projection space, c_T, in terms of total number of misclassified points, n_t, and type I and type II errors.

Process variations remain				Process variations removed			
c_T	n_t	Type I (%)	Type II (%)	c_T	n_t	Type I (%)	Type II (%)
8.0	12	0.00	60.00	13.0	9	0.00	50.00
7.0	10	0.00	50.00	12.0	8	0.00	44.44
6.0	9	3.45	40.00	11.0	7	3.22	33.32
5.0	9	6.90	35.00	10.4	6	3.22	27.70
4.0	13	27.59	25.00	10.0	5	3.22	22.21
3.0	15	41.38	15.00	9.3	6	6.45	22.21
				8.0	8	19.35	11.11

design points are circled in this figure.

If process variations could be completely removed in future operation, then specifications are based on estimated quality that would be obtained had process variations been removed in the past $\mathbf{Y}^*$ (instead of based on measured quality $\mathbf{Y}$). Estimates of $\mathbf{Y}^*$ are shown in Figure 5.9 (top plot), and were computed using Equation 5.3. Regression coefficients ($\mathbf{B_Z}$ and $\mathbf{B_X}$) were estimated using PLS with both $\mathbf{Z}$ and $\mathbf{X}$ merged in one single predictor matrix. Since $\mathbf{Z}$ and $\mathbf{X}$ have different number of variables, block scaling based on the square root of the number of variables in each block was used in order to give them equal importance in the model. A total of 6 components were found significant and modelling results are provided in Table 5.4. Predictions of this model ($Q^2_{Y,cum}$ in Table 5.4) are now much better, since it takes process variations into account. The variability in the corrected FLH^* shown in Figure 5.9 is clearly reduced compared to the original FLH measurements (Figure 5.8), since variations introduced by process variables (Q, h_o and T_a) are removed. However, the effect of these variables on $FMDS$ are not as strong as on FLH and so, variations in $FMDS^*$ (Fig. 5.9) are similar to variations measured in $FMDS$ (Fig. 5.8).

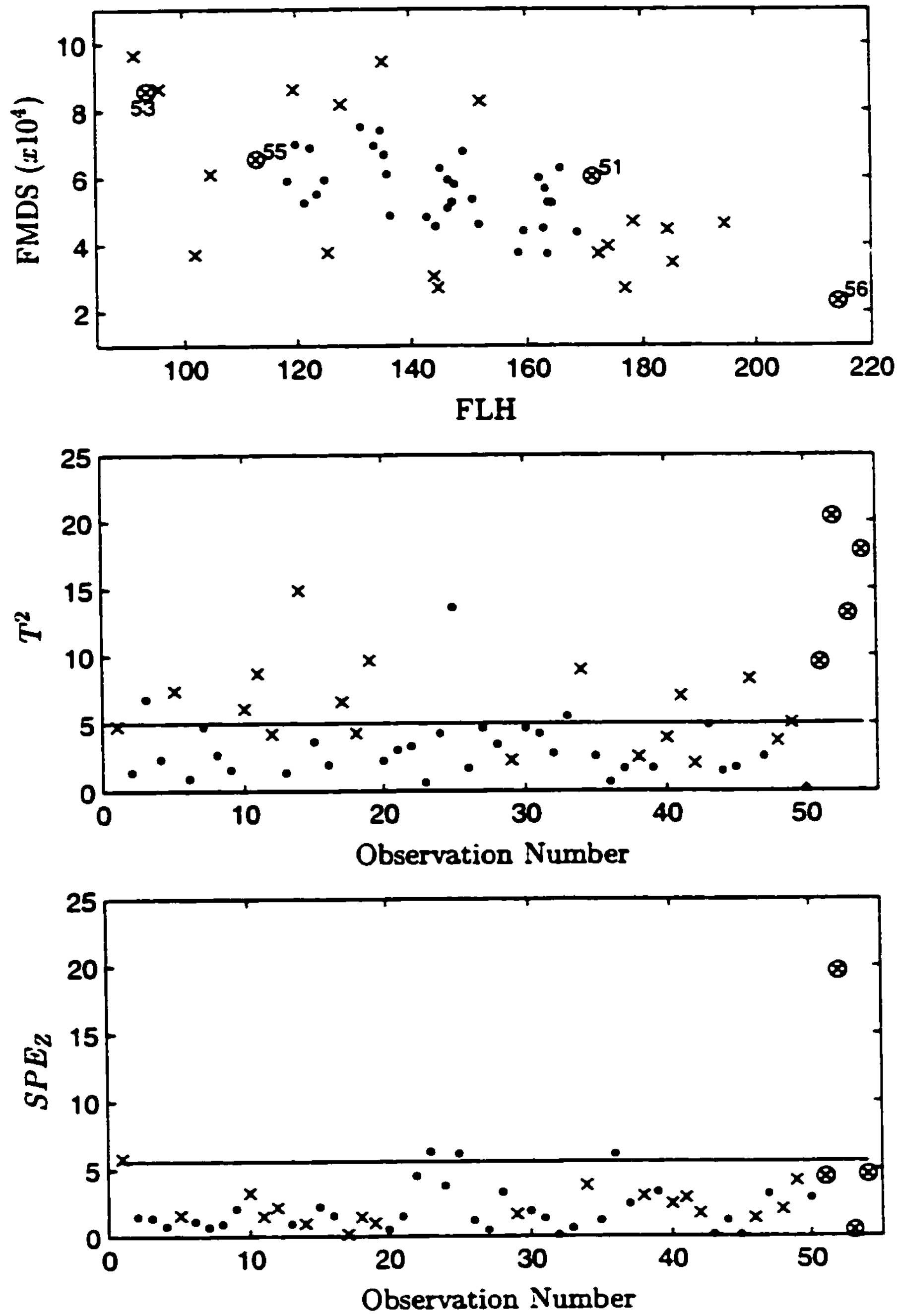


Figure 5.8: Specification region on raw material properties, based on the PLS model supplemented with designed experiments (observations 51, 53, 55 and 56).

To assess what product might have been good or bad, the same quality judgement was applied to $\mathbf{Y}^*$ as was previously done for $\mathbf{Y}$. Note that the same number of final products were identified as good quality (32), but when judging $\mathbf{Y}^*$ instead of $\mathbf{Y}$, the sets of final product meeting quality requirements are different. These new good quality points need to be mapped into the projection space of $\mathbf{Z}$, which is different than the projection space ($\mathbf{T}$) of the PLS model described above, since that model includes both the effect of $\mathbf{Z}$ and $\mathbf{X}$. One can separate the effects of the two blocks and obtain the projection space of $\mathbf{Z}$, $\mathbf{T_z}$, directly from the above PLS model using the work of Westerhuis *et al.* (1998). The reader is referred to the background chapter 2 of this thesis for further details. The right half of Table 5.5 shows the performance of the specification region for different limits. A value of $c_T = 10.0$ was chosen since it leads to only 5 misclassified points and to relatively small type I and type II errors (3.22% and 22.21% respectively). This specification region is also shown in Figure 5.9. Accounting for process variations and removing them from $\mathbf{Y}$ leads to a better classification of raw materials. This is shown by a lower minimum number of misclassified points, n_t (5 compared to 9), and lower type I and type II errors (3.22% and 22.22% compared to 6.90% and 35%).

5.4.3 Significant Process Variations Affecting Quality, Correlated with Raw Materials

A more realistic industrial situation is one where both raw material ($\mathbf{Z}$) and process ($\mathbf{X}$) variations have a significant effect on quality ($\mathbf{Y}$), but control actions are implemented by operators or control systems to compensate for some of the variations in raw materials and other disturbances. Control actions (feedback and \ or feedforward) imply that part of $\mathbf{X}$ is collinear with $\mathbf{Z}$. In this section, we look at the procedure for defining specification regions under three different scenarios: (i)

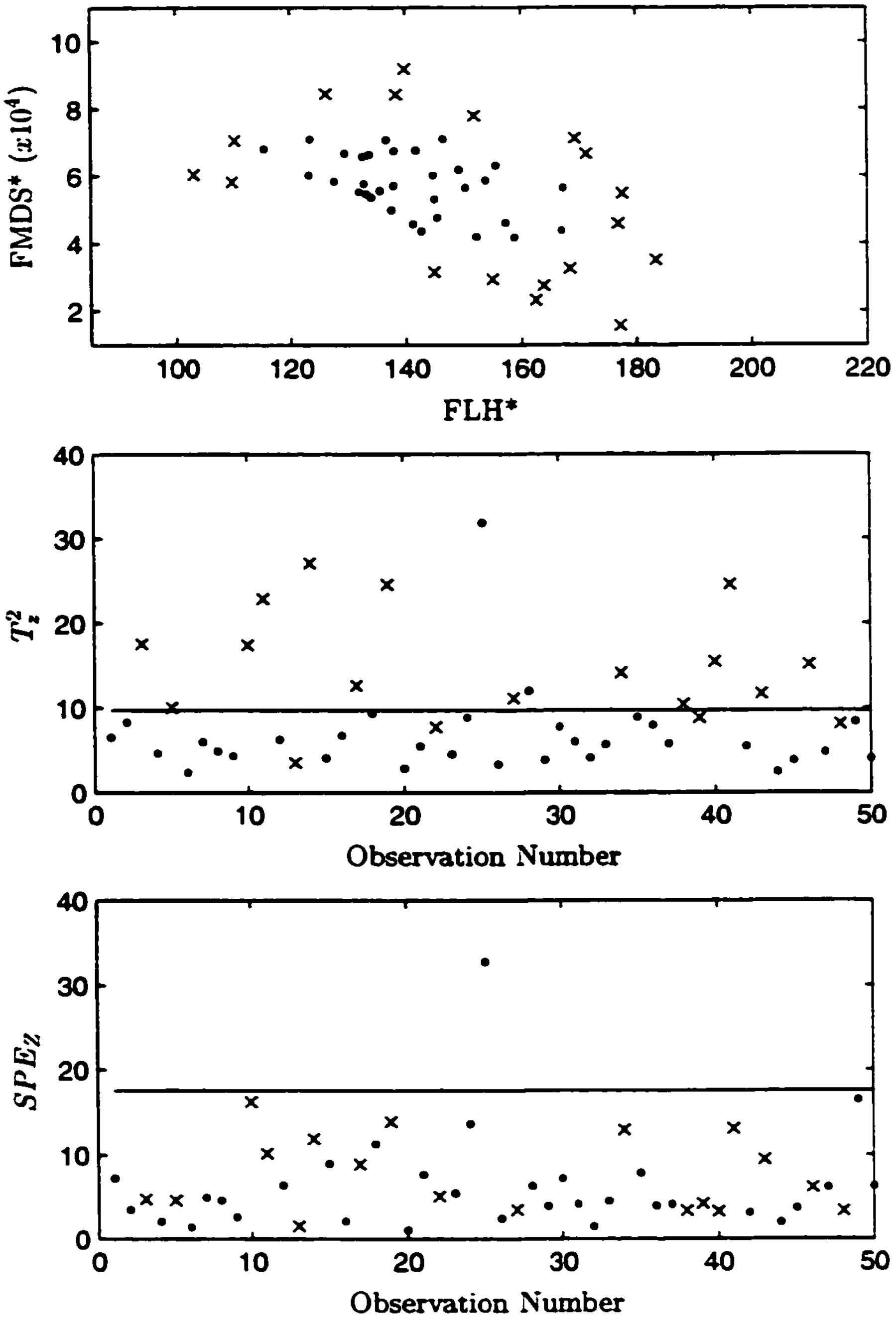


Figure 5.9: Specification region on raw material properties, after estimated quality variations due to the process have been removed.

assuming feedforward \ feedback control will remain in place thereby continuing to eliminate part of the variability arising from variations in the raw materials **Z**; (ii) assuming that the feedforward \ feedback control will be eliminated (i.e. stop fiddling with the process). This of course will lead to much tighter specification limits being required on incoming materials; (iii) assuming one actually were to improve the feedforward \ feedback control scheme to the best possible one. This would lead to the largest specification region on the raw materials and hence, would allow one to accept lower quality materials from larger number of suppliers and reduce costs. The procedure for each of these three scenarios is discussed in turn.

Feedforward \ feedback control remains in future

Although much of the raw material variations may be eliminated by the feedforward \ feedback control schemes, there is still a need to define specification regions to reject those more extreme variations in the raw materials that can not be compensated by the control system. A latent variable model can be built between **Z** and **Y**, to assess how much variance in **Y** (under control) is explained by **Z**. A poor model would result if the control system compensates for most of variations due to **Z** or if not enough information is available from the data base (such as missing measurements on raw materials). In the former case, one may need to augment the data with designed materials (MDOE points), provided that these materials introduce large enough variations in **Y** so that the control system can not eliminate them. If the model explains significant variations in **Y**, than one could define a specification region for raw materials using the methodology presented in section 5.4.1. Such a region would lead to rejection of those lots of raw materials for which control systems can not ensure good quality products.

Feedforward \ feedback control eliminated in future

If the objective is to eventually eliminate the control efforts made to compensate for poor raw material quality, then a new approach needs to be adopted. Instead of defining specifications based on actual quality measurements, $\mathbf{Y}$, one should rather based them on the quality that would have been obtained if no control had been done ($\mathbf{Y}^*$). Estimating $\mathbf{Y}^*$ involves decomposing process variations into a component that is collinear with raw materials, $\mathbf{X}^{\parallel}$, and a component that is orthogonal or independent from raw material variations, $\mathbf{X}^{\perp}$:

$$\mathbf{Y} = \mathbf{X}\ \mathbf{B_X} + \mathbf{Z}\ \mathbf{B_Z} + \mathbf{F} = (\mathbf{X}^{\parallel} + \mathbf{X}^{\perp})\ \mathbf{B_X} + \mathbf{Z}\ \mathbf{B_Z} + \mathbf{F} \tag{5.4}$$

It is assumed that $\mathbf{X}^{\parallel}$ arises from control systems (or operators) in compensating for raw material variations. It is obtained by projecting $\mathbf{X}$ onto $\mathbf{Z}$:

$$\mathbf{X}^{\parallel} = \mathbf{Z}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X} \tag{5.5}$$

If $\mathbf{Z}$ is singular or close to singular, as will usually be the case, then $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$ can be evaluated using any generalized inverse or using any latent variable regression method, such as PCR, CCR and PLS. The second component, $\mathbf{X}^{\perp}$, contains all other process variations that are uncorrelated with $\mathbf{Z}$, and is simply obtained by subtracting $\mathbf{X}^{\parallel}$ from $\mathbf{X}$:

$$\mathbf{X}^{\perp} = \mathbf{X} - \mathbf{X}^{\parallel} \tag{5.6}$$

The estimate of quality that would have been obtained if no control actions had been done to compensate for raw material variations is obtained using the following expression:

$$\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}^{\parallel}\ \mathbf{B_X} = \mathbf{X}^{\perp}\ \mathbf{B_X} + \mathbf{Z}\ \mathbf{B_Z} + \mathbf{F} \tag{5.7}$$

which simply adds to $\mathbf{Y}$ the variations in quality introduced by raw materials that have been eliminated by the operators and control systems ($-\mathbf{X}^{\parallel}\ \mathbf{B_X}$). The only assumption for $\mathbf{Y}^*$ to be a valid reconstitution of the quality had no control been performed is that $\mathbf{B_X}$ reflects the cause and effect relationship between $\mathbf{X}$ and $\mathbf{Y}$ (causal estimate). Estimating causal process effects, $\mathbf{B_X^c}$, from historical data may be possible only in the case where $\mathbf{X}^\perp$ is full rank and has enough variations to observe its effect in $\mathbf{Y}$. If these conditions are met, then one could obtain an estimate of $\mathbf{B_X^c}$ by regressing $\mathbf{Y}$ onto the joint matrix made of $\mathbf{X}^\perp$ and $\mathbf{Z}$. However, even if possible, such estimates of $\mathbf{B_X^c}$ from operating data will usually be very poor. Better estimates of the causal effects ($\mathbf{B_X^c}$) of $\mathbf{X}$ on $\mathbf{Y}$ can be obtained using a few designed experiments. On the other hand, if control systems are into place, estimates of the process gains may already be available. Process gains are causal estimates of $\mathbf{B_X}$ and could be used to compute $\mathbf{Y}^*$ directly from equation 5.7.

Once a valid estimate of $\mathbf{Y}^*$ is obtained, one could develop specifications on raw materials according to one of the following two situations. If it is assume that $\mathbf{X}^\perp$ will remain in the future, then the specifications are developed just as in section 5.4.1, considering $\mathbf{Y} = \mathbf{Y}^*$. On the other hand, if it is assumed that $\mathbf{X}^\perp$ can and will also be removed in the future, the specifications are computed as shown in section 5.4.2, where $\mathbf{Y} = \mathbf{Y}^*$ and $\mathbf{X} = \mathbf{X}^\perp$. In both cases, tighter specifications should be obtained than when control actions remain in future operation, since variations in $\mathbf{Y}^*$ are expanded compared to variations in $\mathbf{Y}$.

Defining specifications under improved feedforward \ feedback control

Another problem related to development of specification regions for raw materials is to assess how specification regions would change under better (or worse) control systems. For example, one might desire to accept raw materials from more suppliers in order to buy cheaper materials and to take advantage of a greater availability of these. To achieve this, improved control systems may be required to compensate for the increased variability of raw materials. The approach, however, again requires causal estimates of the process gains ($\mathbf{B}^{c}_{\mathbf{X}}$) and a simulation to compute new control actions and disturbances ($\mathbf{X}_{\mathbf{new}}$) resulting from the use of a modified control system.

The strategy first involves estimating the effects of raw materials on quality ($\mathbf{B_Z}$) and modelling errors ($\mathbf{F}$). This can be accomplished using a latent variable model relating $\hat{\mathbf{Y}}_{\mathbf{z}}$ to $\mathbf{Z}$ as follows:

$$\hat{\mathbf{Y}}_{\mathbf{z}} = \mathbf{Y} - \mathbf{X}\ \mathbf{B}^{c}_{\mathbf{X}} = \mathbf{Z}\ \mathbf{B_Z} + \mathbf{F} \tag{5.8}$$

Once these are estimated, one should make use of simulations to compute process variations under a modified control system, $\mathbf{X}_{\mathbf{new}}$, and use this to obtain an estimate of quality $\mathbf{Y}_{\mathbf{new}}$ that would be obtained under these conditions:

$$\mathbf{Y}_{\mathbf{new}} = \mathbf{X}_{\mathbf{new}}\ \mathbf{B}^{c}_{\mathbf{X}} + \mathbf{Z}\ \mathbf{B_Z} + \mathbf{F} \tag{5.9}$$

Finally, one could develop specification regions for the modified control system based on a latent variable model relating $\mathbf{Z}$ and $\mathbf{Y}_{\mathbf{new}}$, similarly as in the approach described in section 5.4.1.

Illustration

In this section, the results of developing specification regions for the first two scenarios are presented: feedforward \ feedback control remains in future and control eliminated in future. When generating the data, correlation between $\mathbf{Z}$ and $\mathbf{X}$ is introduced via a feedforward control that corrects for some of the variations in C_p using the polymer flow rate Q. Only h_o is then used as a feedforward variable to control cooling conditions, which are still affected by the disturbance, T_a. Variations in T_a are the same as these used in section 5.4.2. However, the standard deviation of random variations added on Q and h_o are only about a quarter of their control variations. These random variations mimic operator variations in implementing operating conditions, but could also be interpreted as designed experiments aimed at allowing a direct identification of process gains $\mathbf{B}^{c}_{\mathbf{X}}$ from the data.

Table 5.6: Summary of the PLS modelling results, for the situation where variations in quality **Y** are due to raw materials **Z** and to the process **X**, but these sources of variations are correlated.

Model	A	$R^2_{Z,cum}$ (%)	$R^2_{Y,cum}$ (%)	Q^2_{cum} (%)
PLS **Z**-**Y**, $FMDS$ only	2	67.4	82.6	77.2
PLS **Z**-**Y***, FLH^* and $FMDS^*$	4	89.3	84.8	78.4

Figure 5.10 presents the specification region on incoming raw materials when feedforward \ feedback remain in future operation. Poor final product quality in this case is mainly characterized by too large or to small values for $FMDS$, since FLH is well controlled (feedforward actions removing variations in T_a and C_p). As a result, building a latent variable model between **Z** and **Y** leads to almost no explained variations and predictions of FLH. One therefore only needs to judge quality based on $FMDS$. For purpose of illustration, a total of 33 observations were considered as good quality. The specification region shown in Figure 5.10 was obtained using a latent variable model built with $FMDS$ only. A summary of modelling results is provided in Table 5.6. The performance of the specification region developed for raw materials when control actions remain in future operation is shown in Table 5.7 for different values of c_T. A limit of $c_T = 4.6$ was chosen.

On the other hand, if raw material specifications need to be developed for the situation where the control policy was to be eliminated, then the resulting region is presented in Figure 5.11. The key result is that quality variations are now clearly expanded as compared to the previous case study (see Figure 5.10). This expansion corresponds to an estimate of the additional variations in quality, introduced by raw materials (C_p in this case), that will no longer be removed by the feedforward control system. Values for FLH^* and $FMDS^*$ are therefore estimates of the variations in quality that would have been observed if no control efforts had been done to compensate for poorer raw materials. Figure 5.12 shows a comparison of the reconstituted

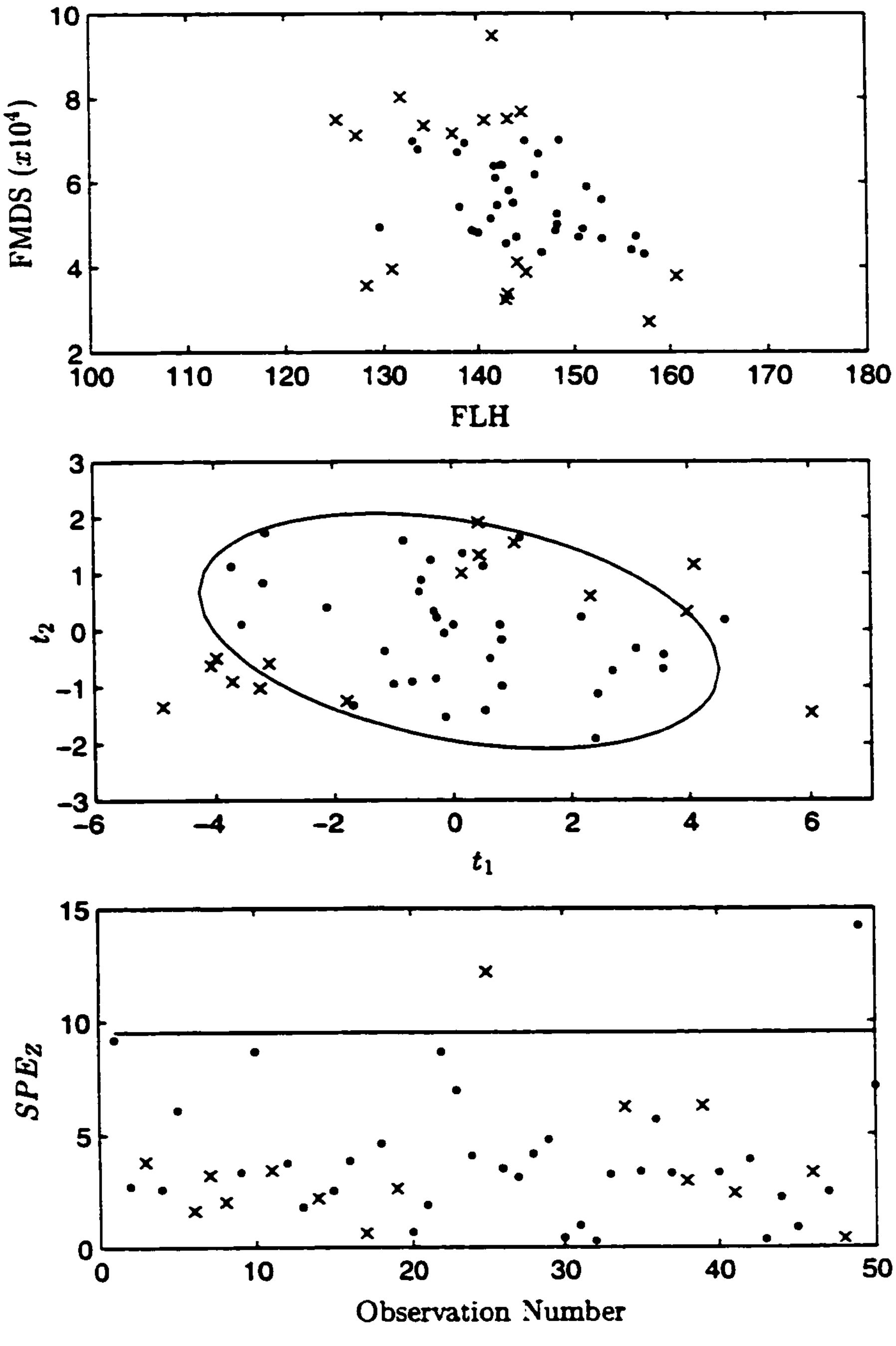


Figure 5.10: Specification region on raw material properties, assuming that process variations will remain in the future. Specifications based on how properties related to polymer viscosity affect FMDS.

Table 5.7: Selection of the specification limit in projection space, c_T, in terms of total number of misclassified points, n_t, and type I and type II errors.

Control actions remain				Control actions removed			
c_T	n_t	Type I (%)	Type II (%)	c_T	n_t	Type I (%)	Type II (%)
8.0	15	0.00	93.75	10.0	17	0.00	70.83
7.0	14	0.00	87.50	9.0	14	0.00	58.33
6.0	12	0.00	75.00	8.0	15	4.54	58.33
5.0	11	3.13	62.50	7.0	13	4.54	50.00
4.6	9	3.13	50.00	6.0	11	9.09	37.50
4.0	12	12.50	50.00	5.9	10	9.09	33.33
3.0	12	21.88	31.25	5.8	10	13.64	29.16
2.0	15	40.63	12.50	4.0	12	36.26	16.66
				3.0	16	63.60	8.33

variance in quality $\mathbf{Y}^*$ (dots) to its true value (plain line). Measured quality variations under control $\mathbf{Y}$ is also shown (crosses). In this example, it was possible to obtain reasonable causal estimates of process gains $\mathbf{B}^{c}_{\mathbf{X}}$ directly from the collected data, because of the presence of independent random variations implemented Q and h_o. If no such independent variations were included in the data, process gains would have had to be obtained using an independent identification study.

The raw material specification region (middle and bottom plots of Figure 5.11) was developed based on a latent variable model between $\mathbf{Z}$ and $\mathbf{Y}^*$. The results are provided in Table 5.6. Since variations removed by the control systems have been reconstituted, the hypothetical quality ($\mathbf{Y}^*$) needed to be judged using both FLH^* and $FMDS^*$. A total of 24 batches of final product was considered as good quality, which is significantly less than when control actions remain in the future. However, this was expected due to the expanded variance of $\mathbf{Y}^*$ compared to $\mathbf{Y}$. The right half of Table 5.7 provides the information necessary for the selection of the limit to set on T^2. Both $c_T = 5.9$ and $c_T = 5.8$ achieve the same minimal number of misclassified

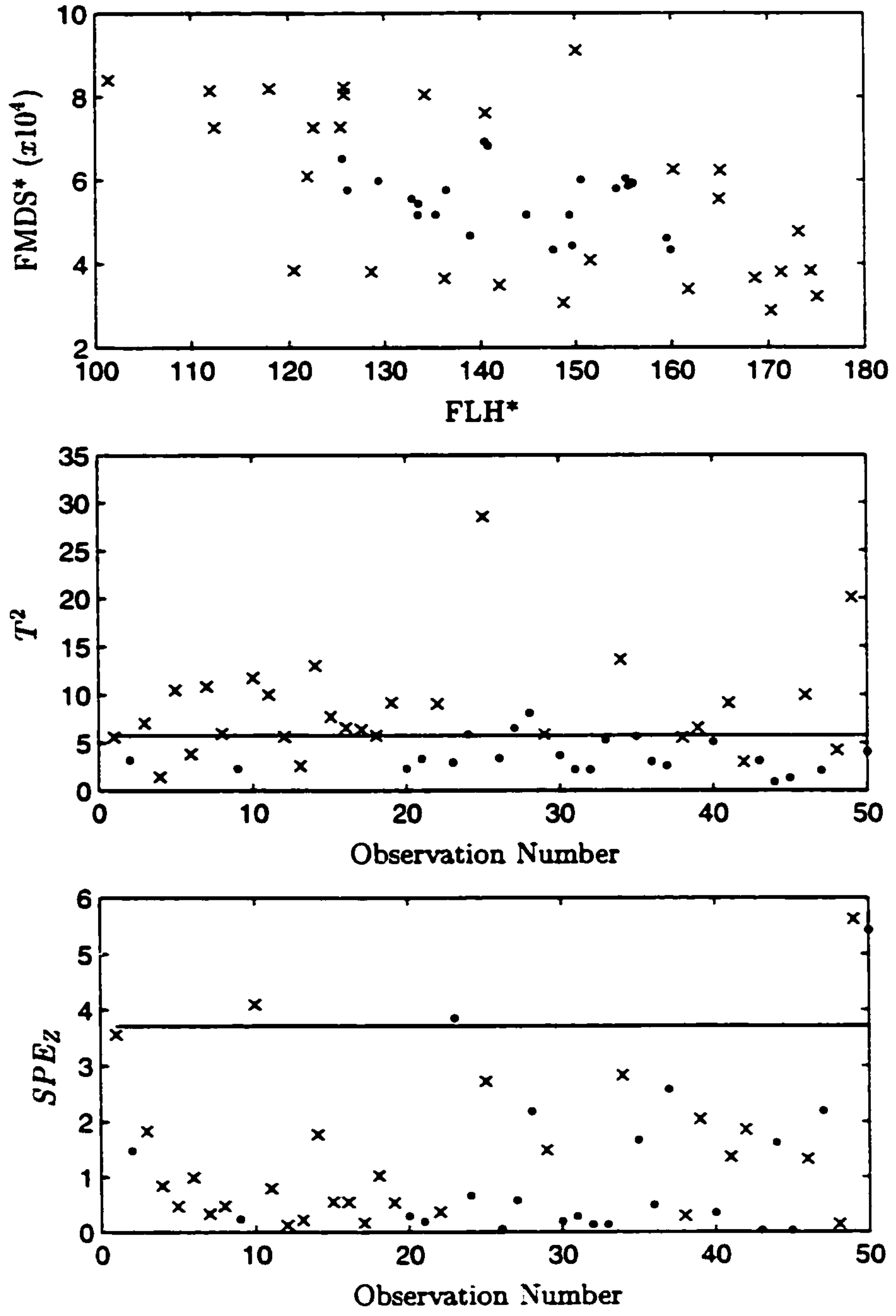


Figure 5.11: Specification region on raw material properties after partly restoring the variability in product quality, as if no control had been done (other process variations ignored).

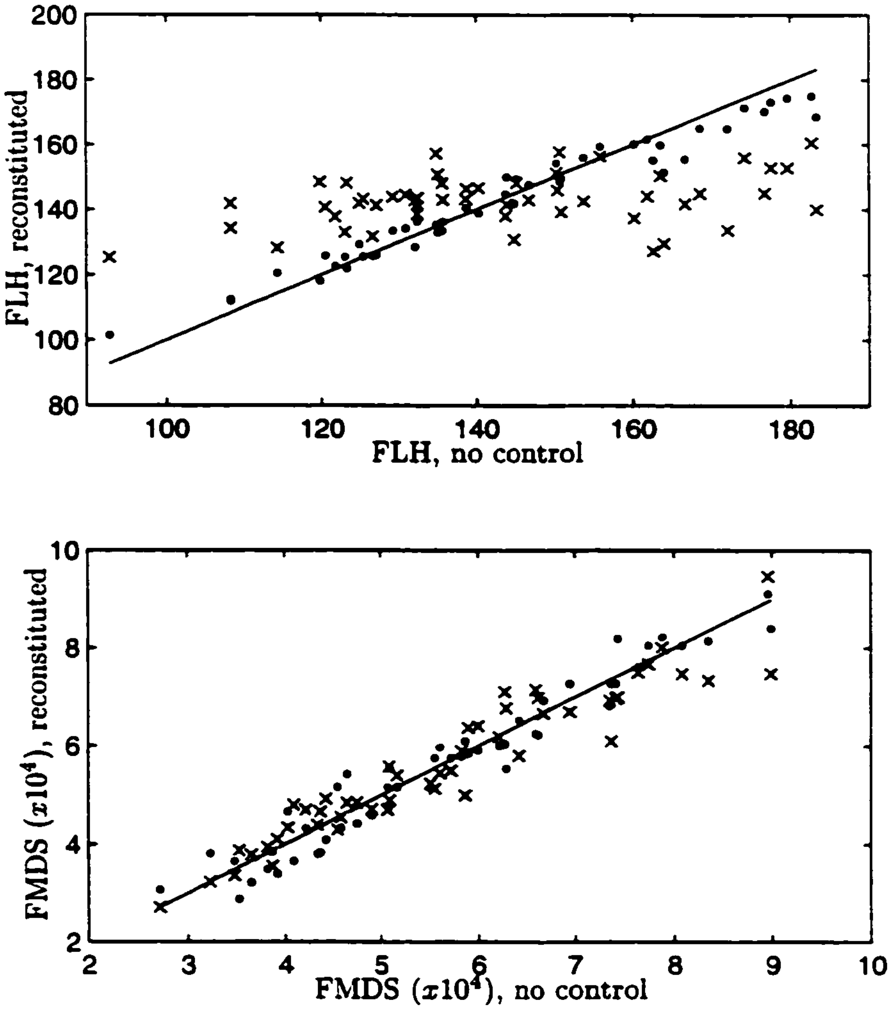


Figure 5.12: Reconstituted variance in product quality as if no control had been done. (Crosses indicated data under control; dots indicate reconstituted data; plain line defines the locus of perfect reconstitution of variance under no control actions.

points, but the value of 5.8 is preferred since it leads to a better balance of type I and type II errors.

Specification regions developed for the two scenarios considered in this section are not easily comparable since latent variable spaces are different, but clearly that region with control eliminated will require much smaller raw material variations. Issues involved in comparing the size of these regions are discussed in the next section.

5.5 Other Related Issues

5.5.1 One-sided specifications

One-sided specifications may arise in different situations. For example, some raw material properties (**Z**) may have an adverse effect on quality or productivity only when they attain values that are too low. Limitations in process operating ability may also require one-sided specifications (in polymer extrusion, can not process too high viscosity polymers etc.). Under several one-sided specifications, the region of good quality raw materials defined in the original measurement space (**Z**) or in the reduced space of a latent variable model (**T**) may also need to be open a one end. However, since the empirical approach proposed in this work is based on a finite **Z** data set, it may be difficult to identify one-sided specifications without prior knowledge. This is one limitation of the proposed approach.

5.5.2 Comparing multivariate specification regions

In developing multivariate specification regions, one may desire to compare the size of the regions obtained under different situations or assumptions. One approach for comparing the size of elliptical region is to compute its volume or hypervolume. When no latent variable model is used, the hypervolume ν of an elliptical region in **Z** space,

$(\mathbf{z} - \bar{\mathbf{z}})\ \hat{\sum}_z^{-1}\ (\mathbf{z} - \bar{\mathbf{z}})^\top = c_z$, is obtained as follows:

$$\nu = \prod\nolimits_{i=1}^{L} \sqrt{\frac{c_z}{\lambda_i}} \tag{5.10}$$

which is simply the product of all principal axes of this elliptical region (Ortega, 1987), and λ_i is the i^{th} eigenvalue of $\hat{\sum}_z^{-1}$. In this expression, $\mathbf{Z}$ remains scaled (to unit variance in this work) to remove the units, otherwise the size of the elliptical regions would be dominated by variables in $\mathbf{Z}$ having large units. However, computing hypervolumes is limited to cases where $\mathbf{Z}$ is not singular or nearly singular, as the hypervolume ν in such a situation would tend to infinity.

If a latent variable model is used to develop the specification region, then the region in the projection space of the model ($\mathbf{T}$) must be mapped into the scaled raw material property space $\mathbf{Z}$ (different models may have projection spaces of different dimensions and therefore, $\mathbf{T}$ space is not a good basis for comparison). The mapping of the specification region from the projection space to the raw material space is done on the basis that $\mathbf{Z}$ can be approximated by $\mathbf{Z} = \mathbf{T}\ \mathbf{P}^\top$ or, alternatively, $\mathbf{T} = \mathbf{Z}\ \mathbf{P}$. Substituting the expression for $\mathbf{T}$, the elliptical region in $\mathbf{Z}$ space becomes $(\mathbf{z} - \bar{\mathbf{z}})\ \mathbf{P}\ \hat{\sum}_T^{-1}\ \mathbf{P}^\top\ (\mathbf{z} - \bar{\mathbf{z}})^\top = c_T$. Then, Equation 5.10 is used replacing c_T for c_z, λ_i becomes the i^{th} eigenvalue of $\mathbf{P}\ \hat{\sum}_T^{-1}\ \mathbf{P}^\top$, and the product is computed from $i = 1$ to A only. It is important to note that comparing specification regions based on latent variable models also requires the models to have a similar structure (e.g. the same method, the same variables and scaling). However, one problem with this approach is that the number of dimensions in latent variable models is often lower that the number of original variables (e.g. $A < L$) and in such a situation, $\mathbf{P}\ \hat{\sum}_T^{-1}\ \mathbf{P}^\top$ is of rank A (only has A non-zero eigenvalues). This implies that no specification limits are required in the remaining $(L - A)$ dimensional space (i.e. infinite limits in these dimensions and therefore, infinite hypervolume). Defining a meaningful comparison between specification regions developed based on latent variable models still requires more investigation.

5.5.3 Several raw materials in the same process requiring specifications

Consider the case where 2 different raw materials are entering the same process as depicted in Figure 5.13, either from the same supplier or different suppliers, where $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the measured properties on each material. The key issue when developing specifications in such a situation, is to decide if specifications should be defined independently for each supplier material or jointly. To address that, one could build a multi-block latent variable model using the data base shown in Figure 5.13 (a) and verify if the projection vector (scores) corresponding to $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are significantly correlated to one another. If they are not, then specifications could be developed independently for each supplier, using the same methodology as presented in this chapter. On the other hand, specifications on $\mathbf{Z_1}$ and $\mathbf{Z_2}$ should be developed jointly (as shown in Figure 5.13 (b)) when there is a synergy between the properties of both raw materials (e.g. projection vectors are highly correlated). For example, if raw material # 1 has certain properties very low, then this material can be accepted as long as certain properties of raw material # 2 are quite high, etc.

5.5.4 Process Capability Indexes

Process capability indexes are used to estimate how likely is a given supplier of materials to meet consumer's requirements for these materials. It is therefore often used as a criterion for selecting suppliers. When only one measurement is monitored, then a univariate capability index is appropriate. Methods for computing various univariate capability indexes are well covered in the literature (Juran *et al.*, 1974; Taguchi, 1986; Taguchi *et al.*, 1989). Consider Figure 5.14 (a), where μ is the targetted value for the property of interest, x_{LCL} and x_{UCL} are the lower and upper supplier control

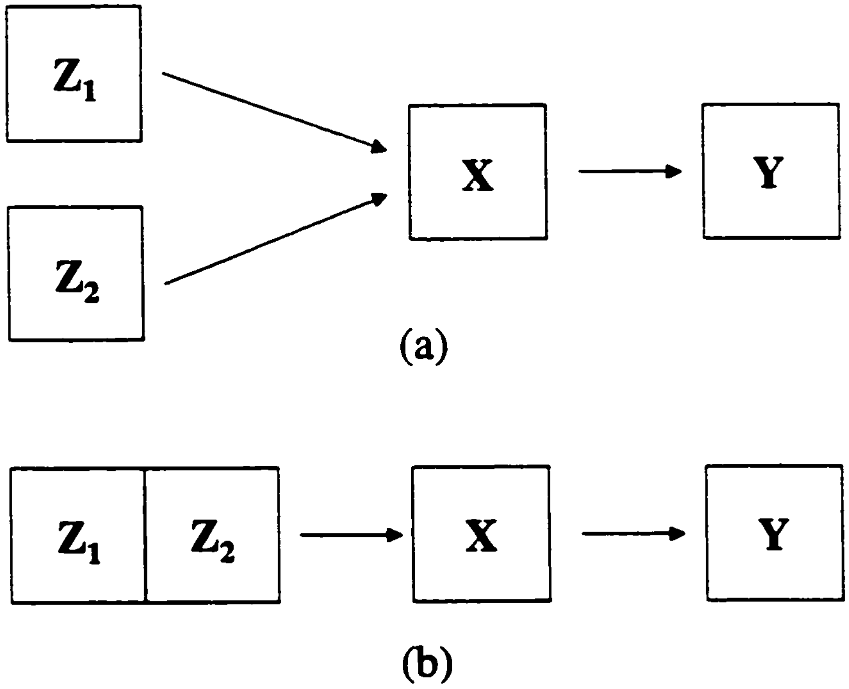


Figure 5.13: Developing specifications on raw materials bought from more than one supplier.

limits, usually defined as three standard deviations, based on supplier's process common cause variations. x_{LSL} and x_{USL} are the lower and upper consumer specification limits. One common univariate capability index is the $C_{p,k}$:

$$C_{p,k} = \min\left\{\frac{x_{USL}-\mu}{3s}, \frac{\mu - x_{LSL}}{3s}\right\} = \min\left\{\frac{x_{USL}-\mu}{x_{UCL}-\mu}, \frac{\mu - x_{LSL}}{\mu - x_{LCL}}\right\} \qquad (5.11)$$

where s is the estimated standard deviation from the supplier's process (common cause variations).

When more than one material property needs to be monitored, then a multivariate capability index (MC_p) should be used. A good review of available multivariate capability indexes is provided by Wierda (1994). According to Wierda, there exist three approaches for computing MC_p's. In one approach (Wierda, 1993), the

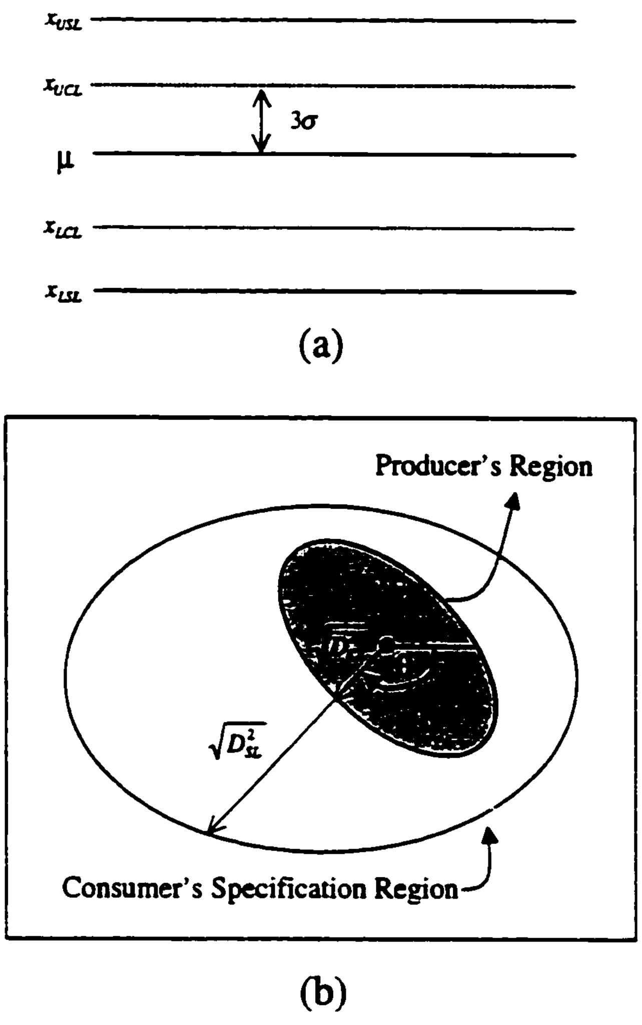


Figure 5.14: Process capability indexes: (a) univariate; (b) multivariate.

consumer specification limits consist of a L-dimensional box. The MC_p is just a function of the probability that products from a given supplier fall into this L-dimensional rectangular region (multivariate normal distribution is assumed). This is an extension of the univariate $C_{p,k}$ (Equation 5.11). To account for correlation among the properties, one could use the same definition for $C_{p,k}$ and compute the probability that products from a given supplier fall into an elliptical specification region instead of an L-dimensional box. A second line of approach, proposed by Kotz and Johnson (1993), makes use of loss functions instead of specification areas.

The last approach for defining multivariate capability indexes is based on elliptical specification regions, $(\mathbf{Z}-\tau)\ \mathbf{A}^{-1}\ (\mathbf{Z}-\tau)^{\mathrm{T}} = c^2$. Such specification regions are recommended since they take into account correlation among monitored properties, and this is the thesis of this chapter. Figure 5.14 (b) shows a multivariate example in two dimensions. The larger elliptical region is the specification area provided by the consumer. The smaller region is one where most of producer's data fall into. In general, the location, the size and the shape of these two elliptical regions may be different. Chan *et al.* (1991) proposed the following index:

$$C_{p,m} = \sqrt{\frac{L}{E\left[(\mathbf{Z}-\tau)\ \mathbf{A}^{-1}\ (\mathbf{Z}-\tau)^{\mathrm{T}}\right]}} \tag{5.12}$$

$C_{p,m}$ is essentially inversely proportional to the average Mahalanobis distance of the supplier data from the consumer's target. Another MC_p is introduced by Pearn *et al.* (1992), and explicitely takes into account the proportion of non-conforming items:

$$C_p^2 = \frac{c^2}{c_\nu^2} \tag{5.13}$$

where c_ν^2 is chosen to achieve a proportion of 0.0027 non-conforming items, which corresponds to the 0.9973 quantile of a chi squared distribution with ν degrees of freedom (assuming that $\mathbf{Z}$ is distributed according to a multivariate normal distribution):

$$Pr\left[(\mathbf{Z}-\tau)\ \mathbf{A}^{-1}\ (\mathbf{Z}-\tau)^{\mathrm{T}} \leq c_\nu^2\right] = 0.9973 \tag{5.14}$$

This amounts to finding an elliptical region, concentric to the consumer's specification region, that would include 99.73% of producer's data. Computation of C_p^2 is straightforward when both supplier and consumer elliptical regions share the same location and shape, but only differ in size. However, in the general case where this does not hold, the authors recognized that it would be rather complicated to compute C_p^2.

We therefore suggest a simpler approach for computing a MC_p index, based on a direct extension of univariate $C_{p,k}$ (see Eq. 5.11). The first step involves rewriting the univariate $C_{p,k}$ expression in terms of the Mahalanobis distance:

$$C_{p,k} = \min\left\{\frac{\sqrt{\frac{(x_{USL}-\mu)^2}{s^2}}}{\sqrt{\frac{(x_{UCL}-\mu)^2}{s^2}}}, \frac{\sqrt{\frac{(\mu-x_{LSL})^2}{s^2}}}{\sqrt{\frac{(\mu-x_{LCL})^2}{s^2}}}\right\} = \min\left\{\sqrt{\frac{D_{USL}^2}{D_{UCL}^2}}, \sqrt{\frac{D_{LSL}^2}{D_{LCL}^2}}\right\} \tag{5.15}$$

Similarly, the $MC_{p,k}$ could be defined as:

$$MC_{p,k} = \min\left\{\sqrt{\frac{D_{SL}^2}{D_{CL}^2}}\right\}_\theta \tag{5.16}$$

where D_{SL} and D_{CL} are the Mahalanobis distances from the location of supplier's region, to the consumer's specification limit and supplier's limit, respectively, taken along the same direction (angle θ). This is also shown schematically in Figure 5.14 (b). One could start from any angle and then compute D_{SL}, D_{CL} and the term within brackets in Equation 5.16 at desired angle steps until 360 degrees is reached. Finally, one would take the minimum ratio as the measure of $MC_{p,k}$. More investigation is required in this area.

5.6 Conclusion

This chapter looks at the problem of developing a methodology to define meaningful specification regions for raw materials entering a consumer's plant. Such specification regions are essential for selecting good quality raw materials that are easily processed and that allow consumers to meet quality requirements for their own products. In spite of the importance of specification regions for the industry and although, extensive literature exists in the quality control area, no sound methodology for defining such regions have been found. Specifications seems to be set arbitrarily, based on past and often subjective experiences, usually using univariate measures. However, measurements collected for characterizing raw materials are often correlated and therefore, quality should be assessed multivariately, through correct combinations of properties. An approach for developing multivariate specification regions is therefore proposed in this chapter based on a sound, empirical analysis of historical data bases.

In the proposed approach, specifications are placed on combinations of raw material properties that appear to strongly affect final consumer's product quality. It is also recognized that the way raw materials are processed in the consumer's plant will have an effect of final product quality. The approach therefore covers most practical situations, which can be classified in three categories, according to the type of process variations that are present: (i) no significant process variations affecting quality; (ii) significant process variations affecting quality, but these variations are uncorrelated with raw materials; (iii) and significant process variations affecting quality, and these are correlated with raw materials because of feedback and feedforward control. The appropriate methodologies for developing specification regions under each of these scenarios are illustrated using simulation studies of a film blowing process.

Chapter 6

Analysis of Start-Up and Grade Transition Problems

6.1 Introduction

This chapter examines the applicability of multivariate statistical methods to process analysis and monitoring during transitions. By transition we define continuous process transitions from grade to grade, the start-up of a continuous process or the re-start of a continuous process that went on hold (e.g. flow rates stopped) due to a technical problem. In multi-product continuous processes, transitions from grade to grade account for a significant portion of process time. As an example, it has been reported that gas-phase linear low-density polyethylene (LLDPE) reactors can cycle between as many as 50 different polymer grades (Xie *et al.*, 1994), to satisfy specific customer demands. Another common type of transition consists of process start-ups, either from empty units or a re-start, after a process hold caused by disruption from steady-state production. The former occurs when process is shut down for scheduled maintenance or when physical process modifications need to be done for producing another grade of material. Re-starts, however, are necessary after disruptions such as sudden electrical and equipment failures, uncontrolled operation and so forth. Although necessary, all of these transitions lead to important loss of production time and amounts of off-grade materials. When a transition reaches its final stage, an additional issue consists of deciding when the transition is over, or alternatively, when is the process ready for the production of in-specification products. The objective of this work is therefore twofold. Part of this chapter is dedicated to developing transition policies or to monitoring existing policies to achieve a reduction in transition time and amounts of off-grade materials. The second section focuses on the problem of defining "production readiness" and "start-up readiness" regions. Solution to both problems could be relatively straightforward when detailed mechanistic process models are available, but since this is often not the case, data base approaches are developed.

6.1.1 Improving Transition Policies

Grade transitions, start-ups and re-starts all share the same three common stages: the initial conditions, the transition and the final steady-state. This is depicted in Figure 6.1. Each dot in this figure correspond to a summary of all process measurements, averaged over a specific steady-state production period, and variables t_1 and t_2 are summary variables. Just before a transition begins, the state of the process defines the initial conditions. This corresponds to "state 1" in Figure 6.1. Initial conditions for a transition from a grade A to a grade B are the steady-state conditions currently observed for production of grade A. For a fresh start-up, initial conditions could include, for example, the initial charge of the process, or recipe, the supplier of ingredients or the conditions of initial feed to a reactor. Initial conditions for re-starts, on the other hand, correspond to process conditions developed as a result of the hold, such as increase in temperature, uncontrolled polymerization and so forth. The type of data base collected during the three stages of transitions is shown in Figure 6.2.

Initial conditions ($\mathbf{Z_I}$) are characterized by a collection of L measurements obtained for I transitions, and should fall within a certain range of conditions to guarantee successful transitions.

The second stage consists of the actual transition from one state to another, as illustrated by the means of paths numbered I to IV in Figure 6.1. The path followed by each transition should be one that minimizes transition time and amount of off-grade materials, while ensuring safe operation.

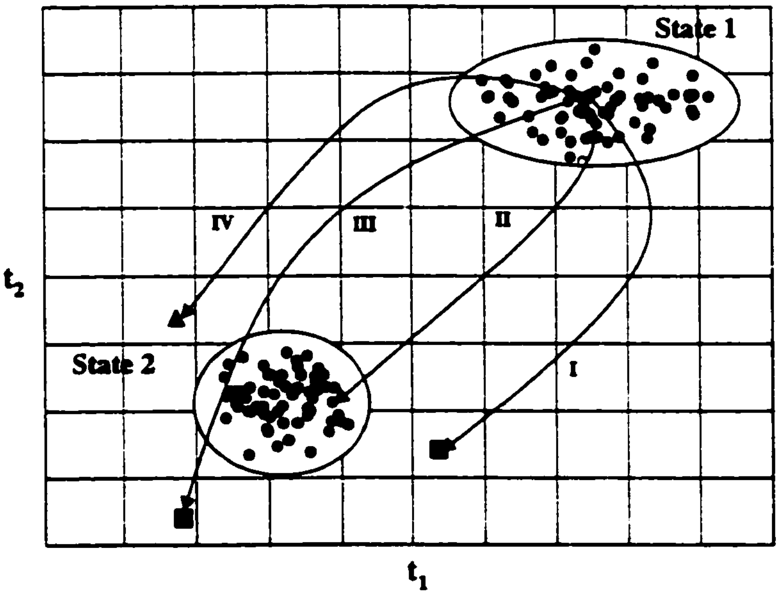


Figure 6.1: Summary the three common stages followed by any transition. Initial conditions: "state 1"; transition: paths I-IV; final steady-state: "state 2" .

Such transition policies are typically developed using detailed mechanistic process models, in conjunction with optimization algorithms (Sargent and Sullivan, 1979; Farber and Laurence, 1986; McAuley and MacGregor, 1992; Xie *et al.*, 1994; Ohshima *et al.*, 1994; Flender *et al.*, 1996; Wang *et al.*, 2000). Sub-optimal policies based on first principles models are also encountered in the literature. A discussion of the relative merits of a few, sub-optimal grade transition policies for olefin polymerization is provided by Debling *et al.* (1994), while Verwijs *et al.* (1995) proposes qualitative rules for achieving successful start-ups of continuously operated chemical reactors. Mechanistic models are, however, often not readily available or difficult and time consuming to develop. In these situations, a data base approach to improve transition policies would be very useful.

For each transition, process measurements are gathered on J process variables, sampled at K intervals, and are stored in a three-way array $\mathbf{X}$, as shown in Figure 6.2.

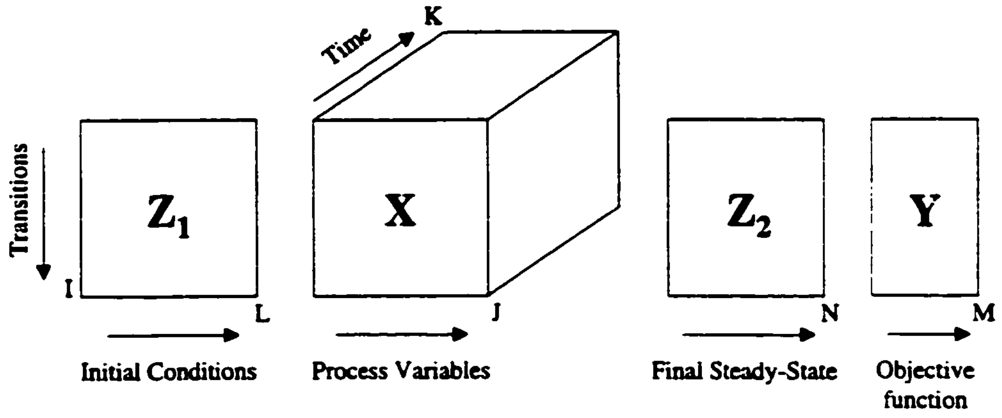


Figure 6.2: Nature of the data collected for the transition problem.

Data block $\mathbf{Z_2}$ consists of measurements collected on N process variables during the final steady-state achieved after each transition (third stage). This data block will be discussed in more details in section 6.1.2. Finally, the last data block, $\mathbf{Y}$, contains M measurements characterizing transition performance. It typically consists of transition time, amounts of off-grade materials, overall performance classification using "good"/"bad" categories and so forth. Data block $\mathbf{Y}$ should be viewed as the *objective function* used for improving transition policies and therefore, any important measurements defining "desirable" transitions or any combination or function of them should be included in $\mathbf{Y}$.

It should be recognized that data bases collected for transitions (Figure 6.2) are similar to those obtained from batch processes. Improvement of batch process operation has already been addressed through process monitoring and diagnosis (Nomikos and MacGregor, 1994*a*, 1994*b* and 1995; Kourti *et al.*, 1995), and through analysis of trajectory features for batch process optimization (Duchesne and MacGregor, 2000a). An important section of this chapter is dedicated to extending the use of these methods for improving the performance of grade transitions, start-ups and re-starts.

6.1.2 Defining "Production Readiness"

The final steady-state operation ("state 2" region in Figure 6.1) achieved after a transition is another important issue in the transition analysis. When transitioning from a grade A to a grade B, the final steady-state consists of a window of conditions for production of grade B, while for a fresh start-up or re-start, it consists of a window of operating conditions known to produce acceptable products.

Developing regions to assess production readiness is essential to ensure that after a transition is over, steady-state conditions are such that high quality products are obtained, and are consistent with past periods of production. Defining production readiness is not trivial, as it should be based on the "overall" product quality (experienced by customers), while only a few properties are typically measured and used to target final steady-state operation. The problem is that the operating conditions may sometimes be fairly different from what is normally used to produce acceptable product for the consumer. This may be due to several reasons, for example when some process variables are left at the settings corresponding to a previous grade or another operating mode, preferences of different teams of operators \ engineers, economics, etc. However, multi-product processes are generally flexible enough to achieve specifications on the few measured quality properties, in spite of these differences. This situation is illustrated in Figure 6.1, where the final steady-state conditions for transitions I, III and IV are different and fall outside the operating region that is normally used ("state 2" ellipse). These processing conditions may still achieve the desired quality on measured properties, but the impact that a different combination of operating conditions may have on the overall product quality as seen by customers is unknown. For example, consider a linear low density polyethylene (LLDPE) fluidized bed reactor transitioning between two grades of polymer. Grade quality of LLDPE copolymers is typically assessed using measureable melt flow index (MI) and density (ρ), to infer the full compositional and molecular weight distributions, CCD and MWD (McAuley and MacGregor, 1992). However, since MI and ρ only measure the location (mean) of the MWD and the CCD, respectively, they do not account for variations in the shape of these distributions, which is very important for customers. When buying two polymer products with the same MI and ρ, but produced under different conditions, a customer may find that they behave differently (i.e. one is more difficult to process than the other). The supplier could receive complaints regarding the quality of the one polymer product, even if MI and ρ were both within specifications.

Very few works related to this problem have been reported in the literature.

McAuley and MacGregor (1992) briefly discussed that using only a few quality variables to define production readiness in a gas-phase LLDPE reactor, in their case MI and ρ, does not ensure that the overall polymer quality (MWD and CCD) is on target. Another contribution is provided by Wurl *et al.* (1999), but is presented later since it is more related to the "start-up readiness" problem.

A logical solution to the production readiness problem is to target processing conditions as well as measured quality properties into a desirable region. This region should be based on steady-state conditions that have led, in the past, to good customer satisfaction. This can only be addressed based on a joint collaboration between suppliers and customers, as only customer feedback makes this problem observable. The data base available for defining a production readiness region is shown in Figure 6.3. The matrix containing steady-state conditions, $\mathbf{Z_2}$, is characterized by N process measurements, collected for I steady-state periods. A second block of data, $\mathbf{Y_{ss}}$, may also be available and could help in identifying the successful steady-state production periods. It may consist of quality variables, other than those used as product specifications, or customer feedback on product quality obtained during all or some steady-state periods. Feedback could be in terms of a "good" or "poor" classification of quality. Once desirable steady-state periods are identified, multivariate projection methods, such as Principal Component Analysis (PCA) and Projection to Latent Structures (PLS), are used to develop production readiness regions. This is developed similarly as for multivariate statistical process control charts for continuous processes (Kourti *et al.*, 1996; Zullo, 1996).

Definition of "start-up readiness" is anothe1 practical problem that could be solved in a similar fashion as for the "production readiness" problem. An example of start-up readiness issues encountered in the film coating industry is described here; it involves synchronizing two sections of the process before the start-up is initiated

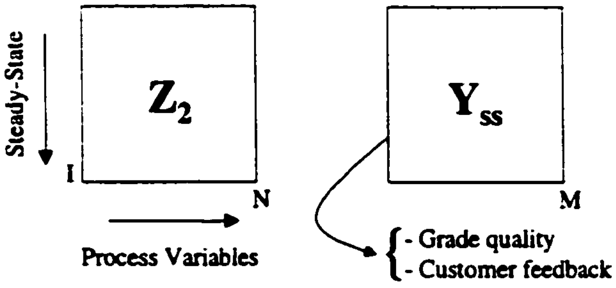


Figure 6.3: Nature of the data collected for defining production readiness.

(coating operation begins). One section is aimed at bringing the materials and solutions involved in the coating operation to specified temperatures, viscosities, etc., and is also responsible for delivering these solution to the coating devices. Process lines delivering these solutions need to be purged before the start-up is initiated. The second section of the process involves equipement for conveying the film to coating devices and also requires some adjustments prior to the start-up. However, coating solutions are generally aging and need to be used within a certain time window. Clearly, if the decision to initiate the start-up happens too early or too late, this may cause poorer transient performance and may even cause the failure of the start-up procedure. Therefore, a start-up readiness region could be developed to help synchronizing the processing conditions in both sections of the process and hence, ensure desirable start-up performance. This region would be based on operating conditions prevailing just before initiating the start-up procedures that led, in the past, to desirable performance. Again, this would involve building PCA or PLS models for defining the good operating region.

Such an approach to solve a similar problem is found in Wurl *et al.* (1999), who proposed to use existing methods for multivariate process monitoring to reduce the start-up time of a batch filament extrusion process. Their approach consists of building a Projection to Latent Structure (PLS) model on good production data and use this model to monitor processing conditions during the start-up. The model allows one to detect operational inconsistencies with good production. Then, contribution plots are used to provide a diagnosis, identifying what variables are associated with such inconsistent operation. In the final step of their method, the authors interpret the results of this diagnosis to modify process operation to achieve a shorter start-up procedure. However, care should be used when interpreting the results of contribution plots in this situation. Since the model was built on production data (no design of experiments), it does not necessarily reflect the true cause and effect process relationships. Interpreting contribution plots as indicators of cause and effect would be incorrect in this situation.

In the next sections of this chapter, methodologies for improving transition policies and for defining production readiness are presented and illustrated in turn. An industrial polymerization case study is used to illustrate the former. A LLDPE reactor simulation study is used for the latter.

6.2 Methodology and Illustrations

6.2.1 Analysis of Transition Policies

The first step of the procedure for improving transition policies involves a post-analysis of all the available data, as shown in Figure 6.2 (except for $\mathbf{Z_2}$). That is use the full set of data in a multi-block way and detect if from the projection space one could distinguish between good and bad cases. This is essential to verify if known problems are observable from the collected process data and furthermore, once observability is established, to determine if the problems can be diagnosed. Multi-way multi-block PCA and PLS can be used to provide such evidence. The nature of the transition data can be of two types and leads to two different problems and solution procedures. Consider Figure 6.4, showing score plots describing the two typical situations that one could encounter with transitions. Each dot in the figure represents a summary of the entire history of a transition. The first situation represents cases where a few distinct transition policies have been used in past operation. This is illustrated by the upper plot of Figure 6.4, where very distinct clusters are obtained and each correspond to a specific transition policy (4 policies are shown in this example; scatter within a cluster is due to random variations of a specific policy). This type of data would result, for instance, when different policies are deliberately implemented for the purpose of improving future operation, as in a designed set of experiments. This data could also result from different teams of operators \ engineers using different ways of implementing the transitions. When such a range of very distinct transitions is available from historical data, it is then straightforward to identify which policy is the best, in terms of transition time, amount of off-specification material and so forth. However, this data could allow one to gain significant process insight in identifying trajectory features that are associated with better start-up or grade changeover performance. The process knowledge gained through such an analysis represents a potential for improving and optimizing transitions. Identification of trajectory features for transient optimization has already been developed and presented in chapter 3 of this thesis. Discussion on the multiple policy problem are therefore not pursued further here.

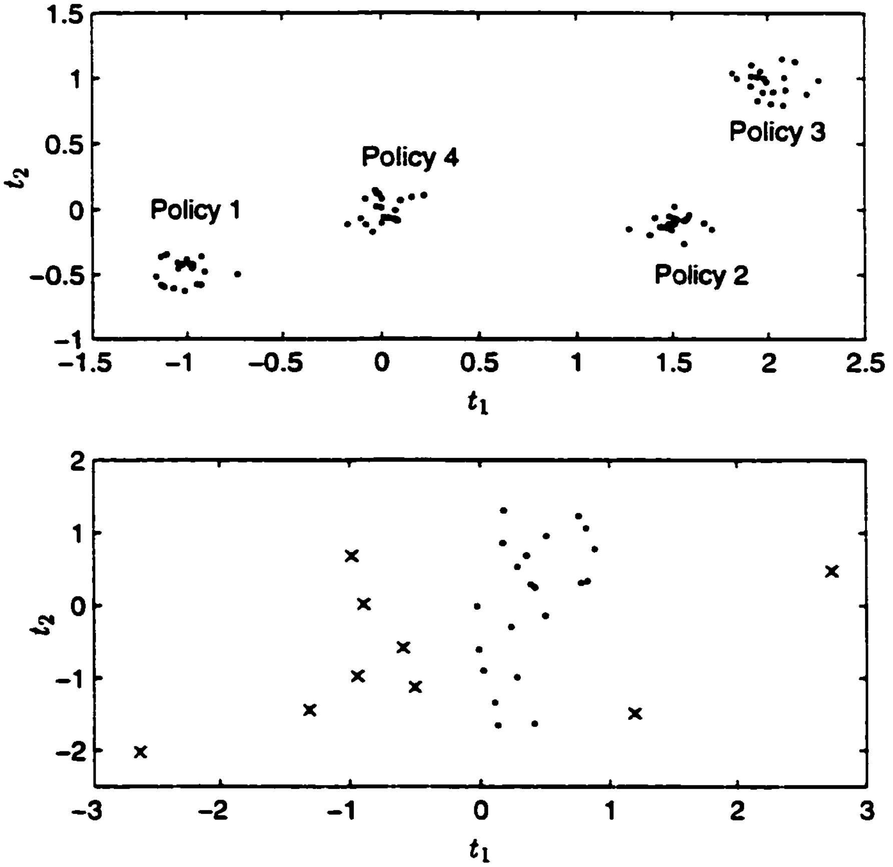


Figure 6.4: Summary of transitions history; upper plot correspond to the situation where several policies are used, lower plot correspond to situation were only one policy is used, but with variations in its implementation (dots: good transition performance; crosses: bad transition performance).

Alternatively, the post-analysis results can rather be similar to those shown in the lower plot of Figure 6.4. This second situation exists when there is really only one transition policy, but operators \ engineers are responsible for variations in its implementation. Some of these variations lead to slightly different lengths of transitions, but still acceptable performance (e.g. dots in this figure). These variations are not systematic and result from different problems in process that eventually affect the transition. On the other hand, other variations may lead to much longer transitions and significantly degrade transition performance (systematic variations), but typically, the problem at the source of these variations is not the same every time (this is shown by the crosses). When such a situation prevail, one could use past transitions that have led to shorter transition times, smaller amounts of off-specification materials and safer operation, and develop a monitoring scheme for the transitions based on the best ones (dots), to ensure that future transitions are consistent with the most desirable ones. Of course, this involves defining what a desirable transition means. For example minimizing off-specification material may lead to greater savings than mimizing transition time. Once the best or most desirable transitions are identified, one could use the multivariate statistical process control techniques, developed for batch processes (Nomikos and MacGregor, 1994*a*, 1994*b* and 1995; Kourti *et al.*, 1995), to build a monitoring region for the transitions. By reducing systematic variations around the desired policy through such a monitoring scheme, one would also expect consistent transitions and associated savings. An industrial example of the situation considered here is presented later in this section.

Another key issue that needs to be addressed is the alignment of trajectories collected during transitions (**X**). Since variations exist in implementing these transitions, the time necessary to reach steady-state operation also varies and hence, the length of process variable trajectories are different from transition to transition. However, multivariate statistical techniques developed for batch processes (MPCA, MPLS) require trajectories of the same length and so, these need to be aligned. The issue of alignment is to remove the time duration factors leaving only the trajectory shape factors to be analyzed by the MPCA or MPLS methods. The different time durations of various sections of the transition or the total time can be including into the **Z** or **Y** data blocks. Two approaches already exist for aligning trajectories: the indicator variable approach (Nomikos and MacGregor, 1994*b*; Kourti *et al.*, 1996) and the warping approach (Kassidas *et al.*, 1998). A more complete review of methods for aligning trajectory data is also provided by Westerhuis *et al.* (1999). The simplest method is to find an appropriate indicator variable, reflecting the progress of the transition, and then plot the trajectories according to this variable. The behavior of indicator variables must be monotonic (increasing or decreasing) and should always start and end to similar values for all transitions. When no such variable exist, the alternative is to use methods such as dynamic time warping (Kassidas *et al.*, 1998) to synchronize the trajectories. This is accomplished by a compression or expansion of specific sections of the trajectories (nonlinear warping). Although these alignment approaches are different, they have the common requirement that the basic shape of the trajectories must be similar. For example, a data base containing multiple distinct transition policies, as implied in the top plot of Figure 6.4, could not be aligned as a whole since the basic shape of the trajectories would be different. Alignment would only make sense for each policy, separately.

Industrial example

Improvement of transient operation is illustrated using data from a continuous industrial polymerization process. The problem is one of recovering from disruption of steady-state production. When operating in steady-state, interruption of process operation may occur for many different reasons. When this happens, the process remains on hold until the problem is fixed, after which the process is re-started. This means that flows to and from the process are stopped. During the hold period, the content of the different process units remains within the units and so, reactions continue to progress. The re-start operation therefore consists of recovering steady-state production regardless of the state that the process reached during the hold, until the cause of disruption is fixed and the process is ready for re-start.

Available for this analysis are data sets from 24 re-starts for the production of a specific polymer. The data is organized into 4 blocks: the steady-state operation prior to the hold $\mathbf{Z}_{ss}$ (24 × 26); the conditions during the process hold just prior to re-start $\mathbf{Z}_{\mathbf{hold}}$ (24 × 6); the start-up trajectories $\mathbf{X}$ (24 × 13 × 100) and the start-up performance criteria $\mathbf{Y}$ (24 × 4). In this case, two data blocks ($\mathbf{Z}_{ss}$ and $\mathbf{Z}_{\mathbf{hold}}$) are used to describe initial conditions prior to the re-start (e.g. $\mathbf{Z}_1$ in Figure 6.2). The steady-state conditions prior to hold ($\mathbf{Z}_{ss}$) describe the state of the process and the conditions in the reactor just prior to when the flows had to be stopped. The steady-state data consists of averaged conditions over a period covering one process residence time. On the other hand, the conditions $\mathbf{Z}_{\mathbf{hold}}$ describe the state that the reactor content reached during the hold.

Re-start performance criteria includes the amounts of two types of off-specification materials, the start-up duration and an "overall" classification of transition performance into "good", "acceptable" and "poor" categories. This judgement was made by process engineers and takes into account all aspects of re-start performance that are important for the company. The main objective of the company is to minimize the amounts of off-grade materials produced during the re-start, and so minimizing transition duration was not really as important for them. In addition, reducing one type of off-grade material is more important relative to the other. This should be reflected by some kind of weighting between the two amounts of off-grade, which is rather subjective. It was therefore decided to use the "overall" classification of re-start performance as a single criterion. This is also subjective, but it is more general and this judgement includes the most important aspects for the company. Since it was found impossible to discriminate between data sets falling into the "good" and "acceptable" categories using the available data, these two groups were merged. The performance criterion ($\mathbf{Y}$) is therefore a simple dummy variable, taking a value of $+1$ for "good" or "acceptable" re-starts and a value of -1 for poorer re-starts.

Prior to any analysis, it was decided to align the trajectory data ($\mathbf{X}$) using the indicator variable approach. The indicator variable is the cumulative mass flow rate of a key reactant divided by the estimated weight of the reactor content (which is full at all times). This indicates the number of residence times that have passed since the begining of the re-start procedure. It always starts at zero, but needs to be truncated at a point where final steady-state is reached. In this example, truncating after 6 residence times was found appropriate as most of the material, that remained within the units during the hold, is discharged and a stable steady-state behavior is obtained after this period of time. As an example of alignment, consider Figure 6.5 showing the trajectories of two process variables before and after being aligned.

Multi-block multi-way discriminant PLS was used to perform a post analysis of the 24 re-start data sets, and the results are summarized in Table 6.1. The multi-block PLS algorithm used is the one published by Westerhuis and Coenegracht (1997). Each block was scaled to unit variance such that all the blocks are treated with equal importance in the model. Satisfactory discrimination between "good"–"acceptable" and "poor" re-start data sets was obtained using two PLS components. The first component accounts for most of the discrimination as it explains 40% of the variance in $\mathbf{Y}$, while about 27% is explained by the second component. The percentage of explained variance for each predictor block indicate how much variations are used in modelling $\mathbf{Y}$ and therefore, R^2 values depend on the number of variables included in each block but not related to $\mathbf{Y}$. With only 21% explained variance, $\mathbf{Z}_{ss}$ and $\mathbf{X}$ have several variables that can not be associated with "good" or "poor" quality. Nevertheless, it is important to keep these variables for process monitoring as, in future operation, they could indicate faults that were not present in the historical data base.

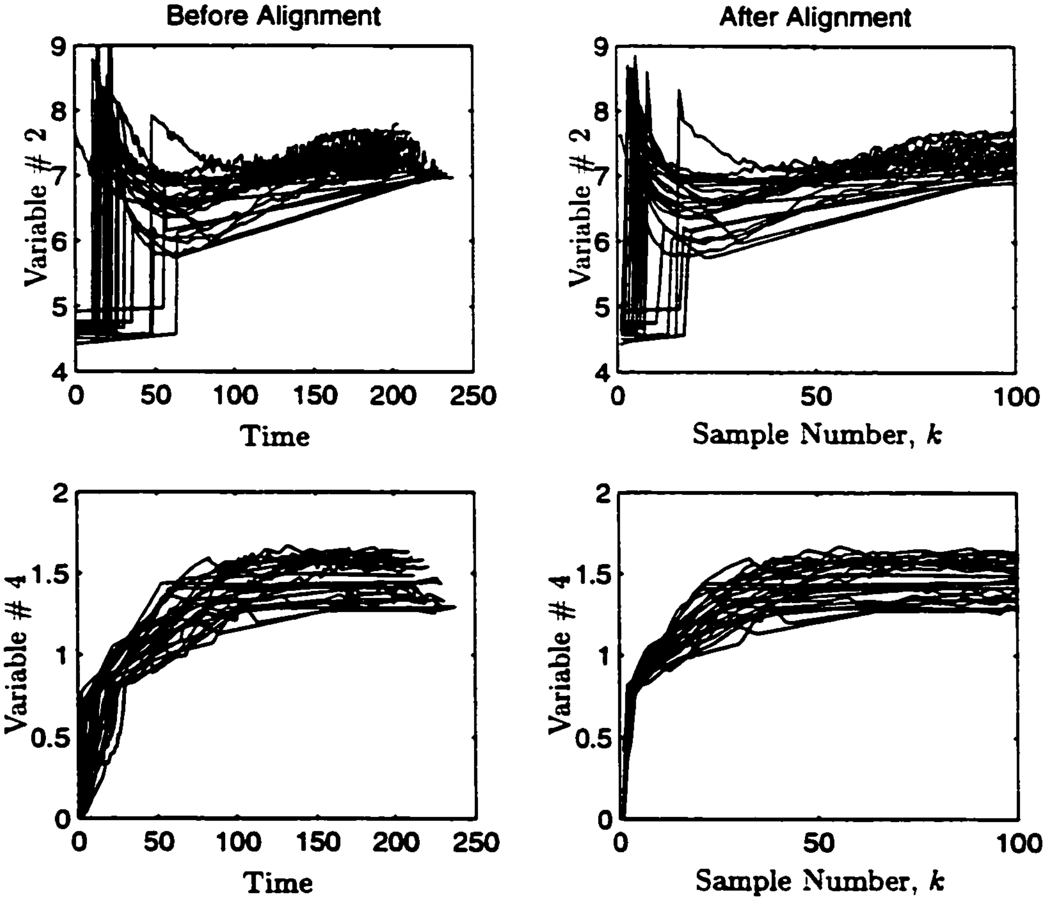


Figure 6.5: Alignment of trajectories collected on variable # 2 and # 4 within each of the 24 re-starts (left hand side: raw trajectories; right hand side: aligned trajectories)

The importance of each predictor block (% Weight) is also shown in Table 6.1. It is computed as the percentage that each super weight squared accounts for in the norm of the super weight vector. Super weights are obtained from the MPLS procedure, when the scores of each predictor block are put altogether into one matrix, and a PLS model is built using this new predictor matrix and **Y** (Westerhuis and Coenegracht, 1997). Super weights give the relative importance of each predictor block in explaining variations in **Y**. The first component focuses more on the hold data block and the second, on the steady-state and re-start trajectories. All three

Table 6.1: Summary of the multi-block multi-way discriminant PLS analysis. R^2_{cum} is the accumulated percentage of variance explained by the model, and % weight gives the relative importance of each block to that component.

	R^2_{cum} (%)				% Weight		
Component	$\mathbf{Z}_{ss}$	$\mathbf{Z}_{hold}$	**X**	**Y**	$\mathbf{Z}_{ss}$	$\mathbf{Z}_{hold}$	**X**
1	13.02	43.95	12.36	40.27	17.88	51.52	30.60
2	21.21	45.67	21.01	67.00	27.94	14.60	57.46

predictor blocks appear important in explaining **Y** (discrimination).

Figure 6.6 shows the score plots corresponding to each of the three blocks of predictors and the super level, which can be seen as an overall result. In each of these plots, dots are used to identify "good" or "acceptable" re-starts and crosses are used for "poor" re-starts. No very distinct clustering pattern (discrimination) is obtained in the steady-state data block. This suggests that steady-state data by itself can not explain "poor" start-up performance, but contributes to explain some aspects of it. t_1 essentially focuses on variables related to production rate. Data sets corresponding to lower production rates fall on the positive side of t_1 and higher production rates, on negative side of t_1. On the other hand, t_2 captures extreme variations in a few process variables. Since data sets number 10 and 15 had a much different value than usual for these variables, they appear as outliers on t_2.

The data collected during the hold, however, reveal a very important aspect of the start-up problem. Two clusters appear, one including observations 17-18 plus 20-24, and the second including the remaining observations. Since re-starts numbered 17-24 had much longer hold periods than usual, this leads one to suspect that hold duration has a significant impact on re-start performance. While the process is on hold, it acts as a batch process, such that concentrations of some reactants and heat accumulate. When holds have extended durations, the accumulation is such that

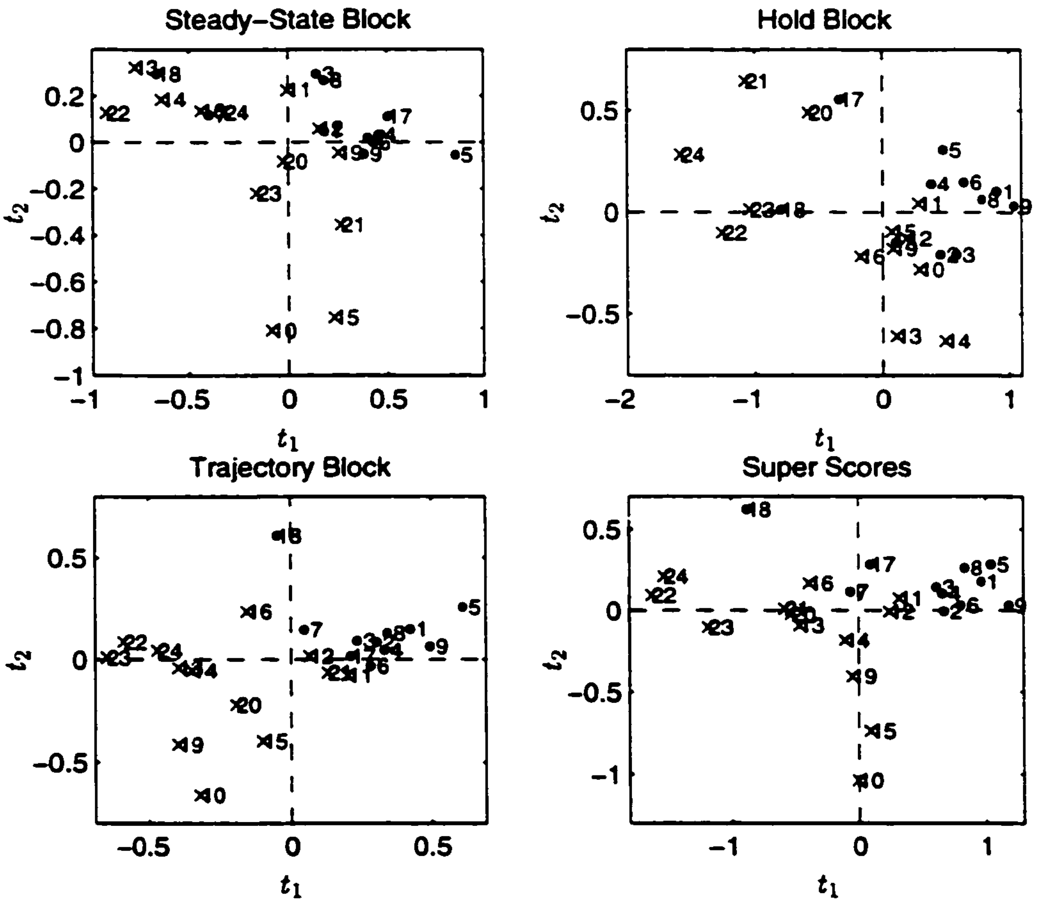


Figure 6.6: Multi-block discriminant PLS analysis of the 24 available re-start data sets (dots: "good" start-ups; crosses: "poor" start-ups.

reactions bring some of the key process conditions to a very different region, and this explains the two clusters in the scores of the hold block. The scaled values taken by one such key variable are shown in Figure 6.7, where the two operating regions are separated using a horizontal dashed line. The fact that most data sets in the range of 17-24 fall in the "poor" category suggests that once the longer hold region is reached (i.e. region above the dashed line in Figure 6.7), it becomes difficult or even impossible to recover from these holds with "good" re-start performance. Polymer reaction engineering knowledge also supports this empirical evidence. Therefore, when this region is reached, one could empty part or all content of the process and then start-up with fresh materials. This would avoid the costs associated with processing materials that are bound to produced large amounts of off-grade materials.

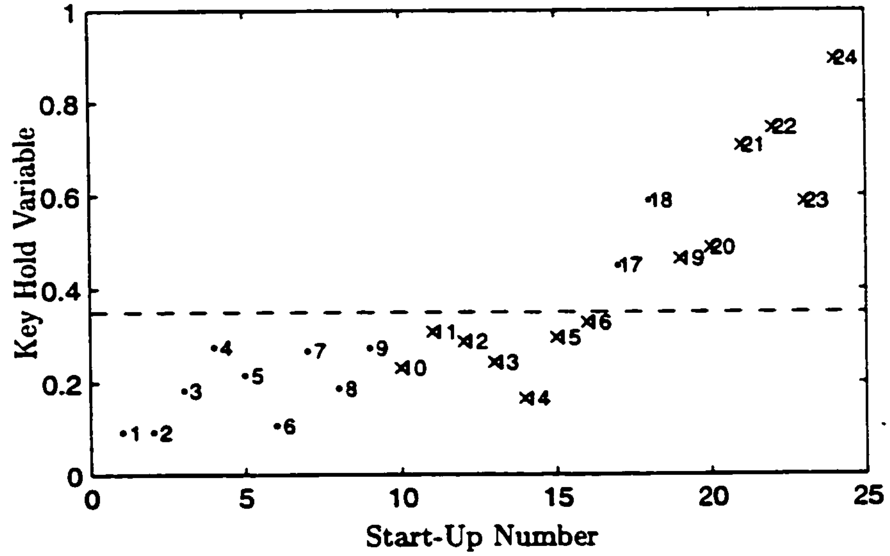


Figure 6.7: Values reached by a key process variable during the 24 hold periods. (dots: "good" re-starts; crosses: "poor" re-starts).

The separation between "good"-"acceptable" and "poor" re-starts is best shown in the score plot corresponding to the re-start trajectories. Most of the "good" and "acceptable" data sets cluster very closely to one another, while the "poor" data sets are projected in different areas. The reason for such a spread among the "poor" sets is attributable to the wide variety of problems at the source of "poor" performance. A relatively similar clustering appears at the super score level, suggesting that both the hold data and trajectories have a great contribution in explaining the production of large amounts of off-grade materials. This emphasizes the importance of monitoring the hold conditions and the start-up trajectories to minimize off-grade products. However, steady-state conditions should also be monitored to detect any abnormal situation and appropriate remedial actions should be implemented to avoid adverse effects on re-start performance. Note that observations 11 and 12 could not be discriminated by any of the predictor blocks. The problem causing them to fall into the "poor" category is unobservable from the collected data. In addition, start-up number 18 appears as an outlier in the score space of the trajectories and at the super score level. This is essentially due to a much higher production rate during the re-start procedure.

The proposed solution to this re-start problem involves developing a monitoring procedure for the re-start trajectories ($\mathbf{X}$), allowing one to recover steady-state operation with good performance, but for a limited range of variations in the steady-state conditions (before hold) and in the hold conditions (key hold variable below the dashed line in Figure 6.7). This procedure assumes that a monitoring scheme for the steady-state conditions ($\mathbf{Z}_{ss}$) is already into place.

A statistical process control (SPC) monitoring scheme has therefore been developed for the trajectories. It involves building limits around the projection space (scores) and residuals, as the re-start progresses, using only the data sets corresponding to the desired performance ("good" and "acceptable"). The procedure for computing these limits have been developed elsewhere (Nomikos and MacGregor, 1994*a*, 1994*b* and 1995). Data sets numbered 17-24 are not used in this analysis, since they appear to correspond to a different class of operating conditions (longer holds), and only 2 "good" or "acceptable" data sets are available in this region. The SPC limits are therefore built using only 9 re-starts. It should be emphasized here that building SPC limits with such a small number of data sets can be misleading, as the range of variation correponding to desired transient behavior may not be entirely spanned. However, the analysis shown next is a good illustration for the methodology proposed in this chapter.

The limits were developed based on a multi-way principal component analysis (MPCA). The number of components is typically chosen using a cross-validation procedure but with only 9 start-ups, cross-validation results were inconsistent. However, 2 components were found sufficient to detect abnormalities in "poor" data sets when these are monitored using the limits. We have therefore chosen 2 components on that basis, for illustration purpose, but more data should be collected and the MPCA model should be updated as new re-start data become available. Figure 6.8 and 6.9 show the results that one would have obtained if re-starts 10 and 13 would have been monitored. These figures show multivariate monitoring charts for the instantaneous square prediction error (SPE_{inst}) and for the scores (t_1 and t_2). The dashed and plain lines on these plots correspond respectively to 95% and 99% limits. Contribution to deviation in SPE_{inst} are also shown in Figure 6.8 and 6.9, at different sampling intervals (k).

For both re-start 10 and 13, the alarm on SPE_{inst} early in the transition ($k = 8$) is due to variable number 2. The behavior of variable 2 during these re-starts is shown in Figure 6.10 and is compared to its normal or desired behavior, defined the by "good" and "acceptable" sets (1-9). Clearly, in both cases, variable 2 remained at a high value for a much longer period of time. In addition, for re-start 10, variable 2 in the early part of the transition also had a much higher value than what is normally observed. This explains the alarm on the scores as well. Variable number 2 has a strong influence on the heat balance of the process and hence, on the reactions. When this variable reaches higher values for a longer period of time, especially early in the re-start, and when no corrective actions are implemented, then variable 5 and 6 generally take a longer time to settle to desired steady-state values. This is shown in the contribution plots for re-start 10 ($k = 60$) and re-start 13 ($k = 18$), where variables 5 and 6 have higher values than usual. Polymerization knowledge again supports this empirical evidence and suggests that this behavior on variable 5 and 6 is strongly associated with production of more off-grade material.

The slow drift in the SPE_{inst} for re-start 13, from about sampling interval

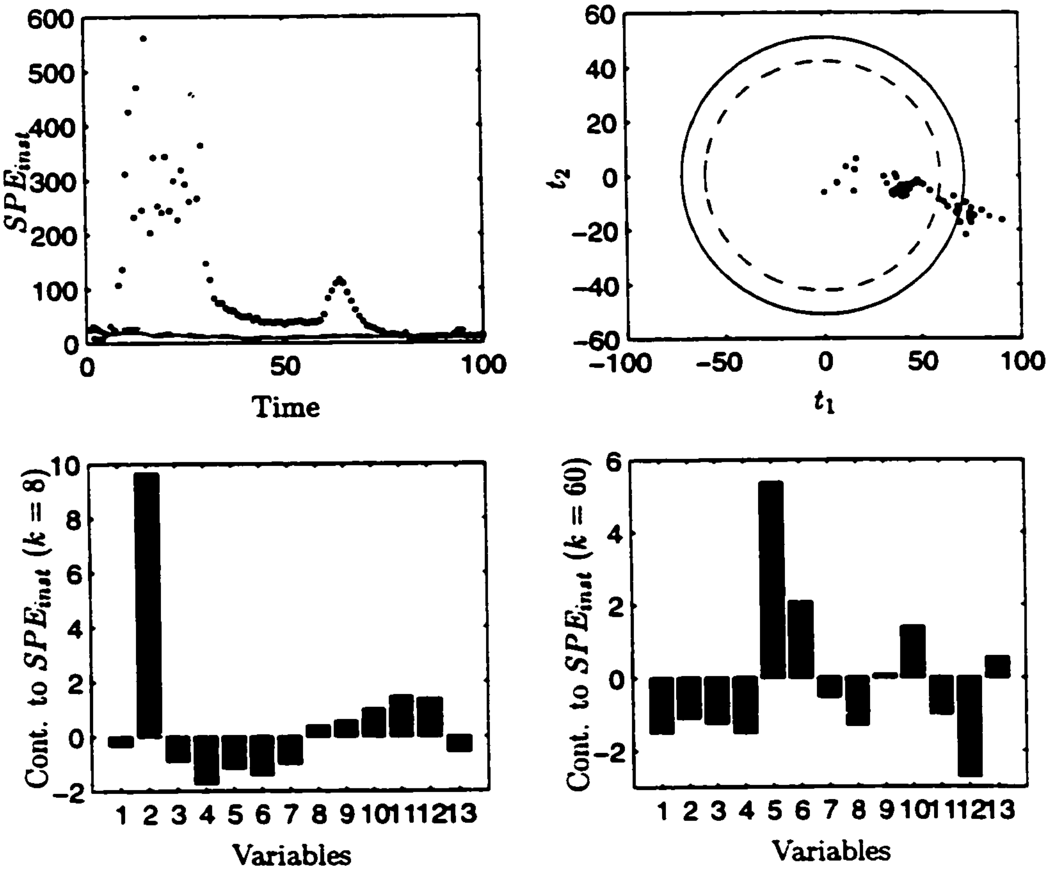


Figure 6.8: Multivariate statistical process monitoring of re-start number 10.

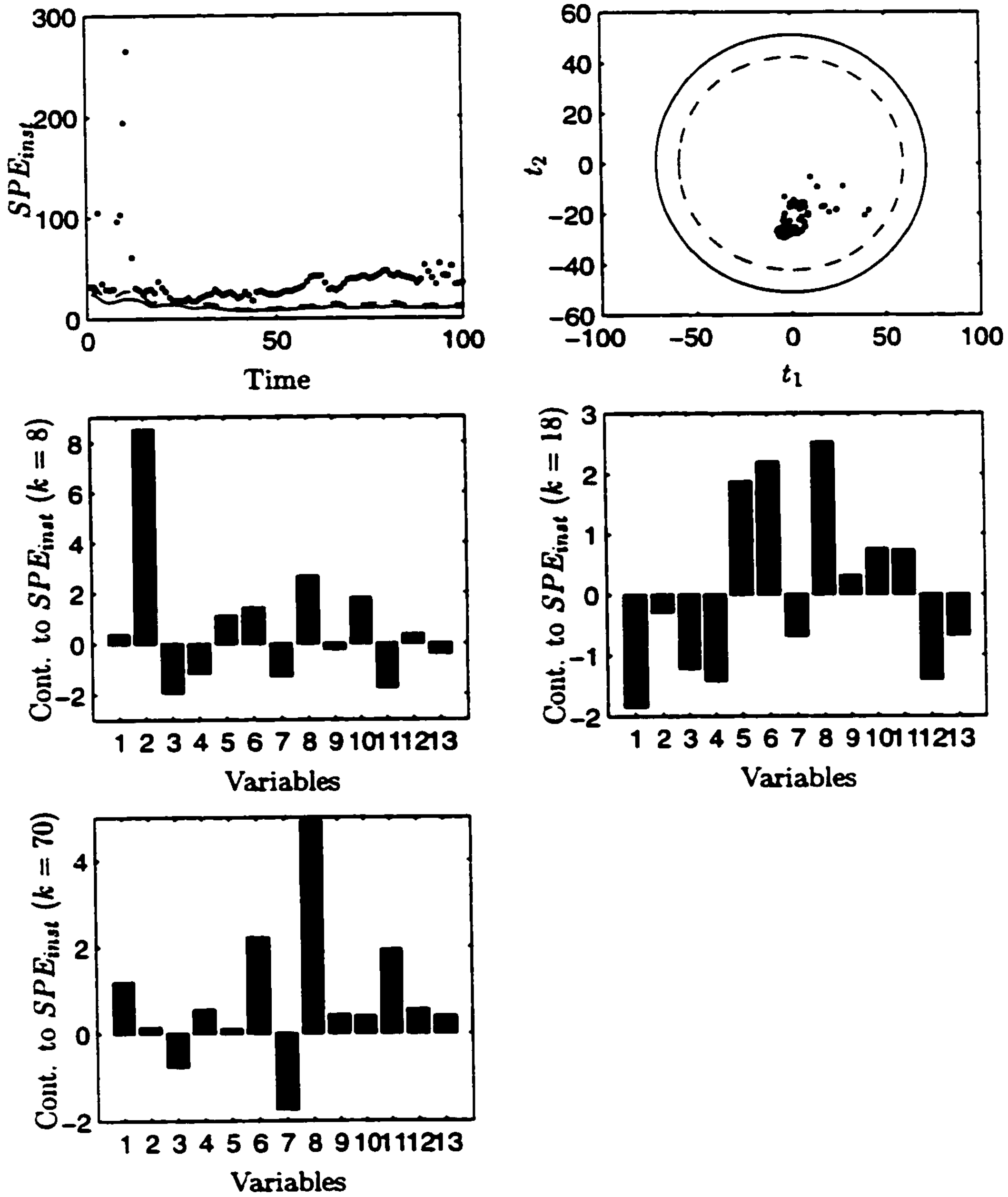


Figure 6.9: Multivariate statistical process monitoring of re-start number 13.

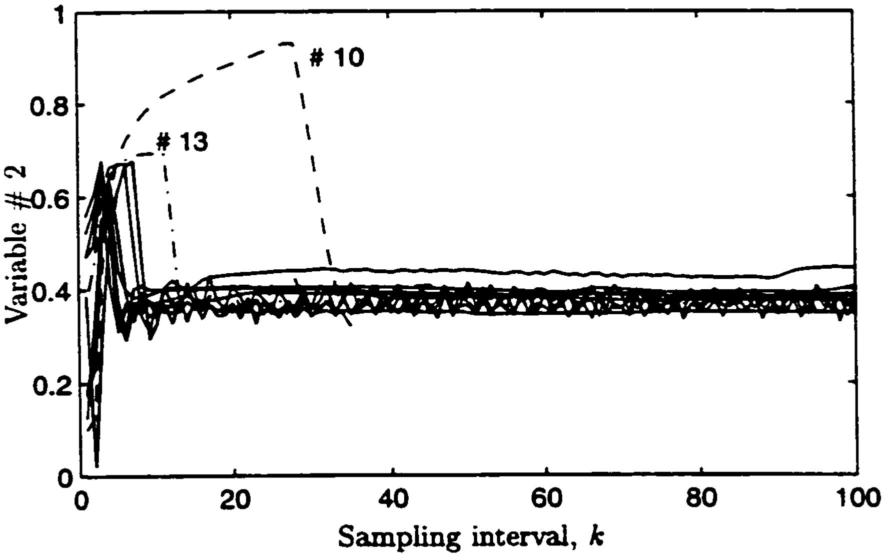


Figure 6.10: Behavior of variable # 2 during start-ups (plain line: "good"-"acceptable" data sets # 1-9; dashed line: "poor" start-up # 10; dashed dotted line: "poor" start-up # 13).

$k = 50$, is caused by a slowly increasing production rate. This is indicated by large contributions of variables 7 and 8, shown in the contribution plots for sampling interval $k = 70$. The combination of high values for variable 5 and 6 and higher production rate should normally lead to even larger amounts of off-grade materials.

The faults for the other "poor" re-start data sets were also identified, but not shown here. The procedure described in this chapter has demonstrated (even with crude limits obtained from only 9 re-starts) that had an on-line monitoring scheme been in place, these problems in the process operation would have been detected early enough to alert operators and prevent off-specification material from forming.

To conclude this section, Table 6.2 summarizes the four steps involved in the proposed methodology. This procedure can be further refined to modify the limits based on new information (data) collected from holds. In the present case, the limits were developed for process re-start trajectories, but only from a certain class of holds (i.e. hold variables were within a range). As more data become available, the limits

Table 6.2: Summary of the proposed methodology for improving transitions.

Step	Description
1	trajectory alignment (descriptor variable or DTW)
2	post analysis - discriminant PLS: find blocks responsible
3	set MSPC limits for monitoring trajectories
4	test if limits would have detected poor transitions

could be modified to account for different ranges of hold variables.

6.2.2 Production Readiness Analysis

The final stage of a transition is the steady-state production of in-specification products. To meet customer requirements, one needs to ensure high quality products, consistent with past periods of production. A logical way to achieve this is to target processing conditions and measured quality properties to regions known to produce good products. It is assumed here that multivariate monitoring charts are already developed and used for targeting measured quality properties and therefore, this section focuses on defining a successful region for processing conditions, as explained in section 6.1.2. This region is based on a close collaboration between a supplier and his customers, who should provide feedback on product quality made during each steady-state period of interest. In this chapter, customer feedback on quality is also assumed already available, but this is a common assumption for any statistical process control (SPC) related problems. The production readiness region should warn operators \ engineers whenever processing conditions (approaching steady-state) are different than those that normally lead to good product. This is valid, unless prior knowledge suggests that specific differences in operating conditions have no effect on "overall" quality. Projection methods such as PCA and PLS have been shown successful for fault detection and process monitoring (MacGregor and Kourti, 1995) and are used for developing production readiness regions.

To define such a region, one should first project the processing conditions from common cause variations around target ($\mathbf{Z}_{2,cc}$), onto the reduced space or score space ($\mathbf{T}$) obtained through a PCA analysis. If direct customer feedback on final product quality is available, in terms of "good" or "poor", or if additional quality variables are measured (other than targetted ones), this information should be included in a second data block, $\mathbf{Y}_{ss}$, and the reduced space should be computed using PLS. Then, an elliptical region of the form of $(\mathbf{t}_{cc} - \bar{\mathbf{t}}_{cc})\ \hat{\Sigma}_{t_{cc}}^{-1}\ (\mathbf{t}_{cc} - \bar{\mathbf{t}}_{cc})^{\top} = T^2$ is used to define production readiness. Here $\hat{\Sigma}_{t_{cc}}^{-1}$ is the estimated covariance matrix of $\mathbf{T}_{cc}$ and T^2 defines the size of the elliptical region. The latter is estimated using the Hotelling's T^2 distribution, $T^2 \sim \frac{(I-1)(I+1)A}{I(I-A)} F_{\alpha}(A, I-A)$, at critical level α (Anderson, 1984). I and A are respectively the number of observations in $\mathbf{Z}_{2,cc}$ and hence, in $\mathbf{T}_{cc}$, and A is the number of components used in the PCA or PLS analysis. Since this production readiness region is defined in reduced space, the distance of each observation from this region must also be monitored. For the i^{th} observation, the distance from the region (or residuals) can be measured by the square prediction error, $SPE_i = (\mathbf{z}_{2,i} - \hat{\mathbf{z}}_{2,i})^{\top}(\mathbf{z}_{2,i} - \hat{\mathbf{z}}_{2,i})$, where $\hat{\mathbf{z}}_{2,i}$ is the estimate of $\mathbf{z}_{2,i}$ obtained by PCA or PLS: $\mathbf{z}_{2,i} = \mathbf{t}_i\ \mathbf{P}^{\top} + \mathbf{E} = \hat{\mathbf{z}}_{2,i} + \mathbf{E}$. Approaches for computing upper limits on SPE are discussed in Nomikos and MacGregor (1994*b*). A valid production readiness region therefore consists of two monitoring charts, one for the projection space (score space) and one for the distance from this projection space (residuals). Final steady-state operation has reached this region when projected conditions fall within the limits of the projection space and when the residuals are below an upper limit. Only when both processing conditions and measured quality have reached their successful regions should the transition be considered terminated and product sold as in-specification.

Simulation example

The concepts related to definition of production readiness are now illustrated using a detailed mechanistic model of a gas-phase LLDPE reactor (McAuley *et al.*, 1990; McAuley, 1992). A schematic diagram of the LLDPE reactor is shown in Figure 6.11. A fluidized bed reactor is used to carry out copolymerization of ethylene and butene (copolymer) by the means of a Ziegler-Natta catalyst. Such a catalyst has multiple sites, but the model approximates this using only 2 sites. The fresh feeds of the system consists of ethylene (F_{Et}), butene (F_{But}), hydrogen as a chain transfer agent (F_H) and inerts (nitrogen). These gases maintain the fluidization of the bed, within which polymer particles are growing. Feed rates of ethylene and inerts are also used to control reactor pressure (P). The catalyst (F_{cat}) is fed directly into the reaction zone. The bed level or weight (B_w) is controlled by manipulating the polymer outflow rate ($F_{poly,out}$). Since the conversion per pass is low, unreacted monomers coming out of the reactor are recycled.

A bleed stream is therefore necessary to prevent build up of materials and impurities. Bleed rate (F_{bleed}) is adjusted using the bleed valve position (ν_p). Since the reaction is exothermic, heat is removed from the recycled gas stream before returning into the reactor, using a water-cooled heat exchanger. Reactor temperature (T_r) is controlled using a cascade control system, adjusting the cooling water temperature feeding the exchanger ($T_{w,in}$). To stabilize the polymerization reaction, ethylene concentration in the recycle stream is also controlled at a specific level using the rate of fresh ethylene feed. A nonlinear property control scheme is implemented on the process to maintain the two measured polymer quality variables, instantaneous MI and ρ, close to their set-points (McAuley and MacGregor, 1993). "Overall" polymer quality can be assessed by the overall compositional and molecular weight distributions (CCD and MWD), computed using Stockmayer's bivariate distribution for a two monomer system (McAuley *et al.*, 1990).

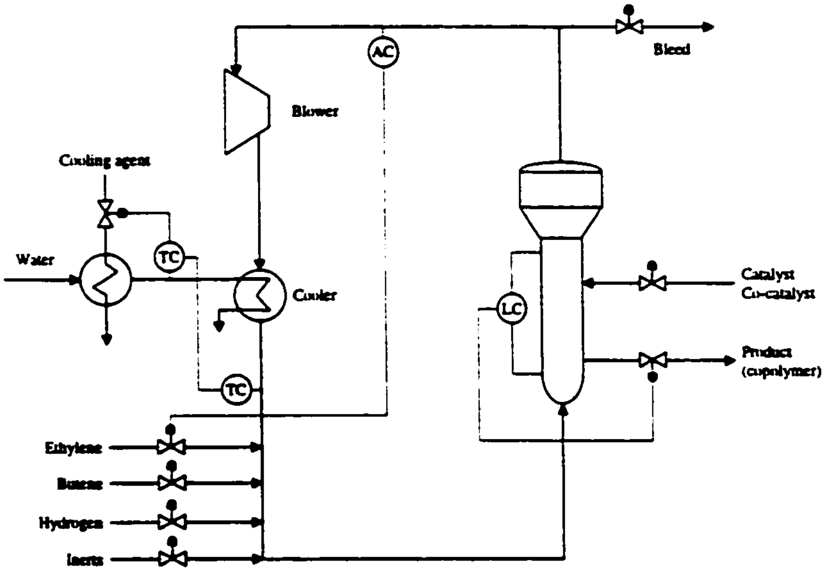


Figure 6.11: Gas-phase Linear Low Density Polyethylene (LLDPE) reactor.

To simulate a common cause variation region for process operation, 20 steady-state periods for the production of a specific polymer grade (MI = 12 g/min, $\rho = 920$ g/L) were generated, by randomly varying the reactor temperature set-point (T_{sp}), the bed level set-point ($B_{w,sp}$), the feed of catalyst (F_{cat}) and the bleed valve position (ν_p), around nominal conditions. The values of these variables were maintained constant during the entire steady-state periods, but were varied between them. Nearly non-stationary auto-regressive disturbances on the amount of impurities and on the relative amounts of each type of catalyst active sites as well as random measurement error on MI and ρ were also added to the simulations. For each production period, the nonlinear controller maintained MI and ρ near their specifications in spite of disturbances and variations in operating conditions. Table 6.3 shows the nominal values

Table 6.3: Process nominal conditions.

	T_{sp} (K)	$B_{w,sp}$ (tonnes)	F_{cat} (kg/h)	ν_p (%)
Nominal value	360.00	65.0	5.00	0.30
Std	1.19	0.9	0.11	0.02

and sample standard deviation for each of the 4 process variables. To introduce some collinearity within the 4 manipulated process variables, $B_{w,sp}$ and F_{cat} were varied in a correlated manner with a sample correlation coefficient of 0.84.

Three additional sets of operating conditions were also simulated, but including systematic differences not seen in the common cause data. However, even for these conditions, similar MI and ρ were obtained due to compensation from the non-linear controller. Steady-state periods number 21 and 22 were generated similarly as for common cause variation data, but reactor temperature set-point was left at other settings in both cases. Values of 350.1 K and 369.5 K were implemented on T_{sp} for runs number 21 and 22 respectively. Steady-state period number 23 was also generated as common cause data, but $B_{w,sp}$ and F_{cat} were varied in negatively correlated fashion. Figure 6.12 shows the measurements collected on MI and ρ every 18 minutes (upper plot), for each of the 23 steady-state production periods. Averaged MI and ρ values over these periods are also shown in Figure 6.12 (bottom plot). Clearly, the apparent polymer quality as measured by MI and ρ still appears to be good for operating periods 21 to 23. This demonstrates that the process has enough flexibility to achieve the same measured polymer quality in spite of all sources of variations.

If polymer quality could have been completely characterized by measuring the overall CCD and MWD for all 23 production periods, the results shown in Figure 6.13 would have been obtained. Average steady-state process conditions were used for computing the CCD's and MWD's. Distributions associated with periods 1-20 and 23 are very similar and only vary due to common cause sources of variations.

However, operation in periods 21 and 22 clearly lead to a very different product, even if MI and ρ were within specification. It is expected that supplier would have received complaints for these products, without having evidence for such poorer quality. It is for this reason that we assume that a the supplier would use steady-state periods 1-20 and 23 as an initial basis for developing the production readiness region.

A total of 11 process variables ($\mathbf{Z_2}$) are currently measured on the fluidized bed reactor: F_{Et}, F_{But}, F_H, F_{cat}, F_{bleed}, $F_{poly,out}$, T_r, $T_{w,in}$, P, B_w and ν_p. These measurements were averaged over their corresponding steady-state periods and PCA was therefore chosen for this analysis. The results are shown in Figure 6.14. Note that mean centering and scaling to unit variance was applied to $\mathbf{Z_2}$ prior to the PCA analysis, which was computed using the SIMCA-P 7.0 software (Umetrics, 1998). Three components were found significant by leave-one-out cross-validation and so, the Hotelling's T^2 statistic is used to show the projection space. $DMODX$ is the measure of perpendicular distance from the projection plane used in the SIMCA-P 7.0 software, but this is just a scaled SPE with similar distribution properties.

Figure 6.14 shows that steady-state operation for periods 1-20 and 23 is fairly consistent, as these sets of measurements project in a similar area (similar T^2 values) and also have a relatively small distance to the model ($DMODX$), except for period 23 that is clearly detected as an outlier. High $DMODX$ value for this production period illustrates a situation where the correlation structure among operating conditions is different from what is normally used ($B_{w,sp}$ and F_{cat} were varied in a negatively correlated fashion as opposed to positively). Although no major customer complaints were obtained for materials produced under these conditions, not enough data is available to confirm that this different way of operating the process consistently leads to high quality. It is therefore safer to remove such outlier observations before developing production readiness regions, unless prior knowledge suggests keeping it.

A new PCA analysis was carried out using only production periods 1-20.

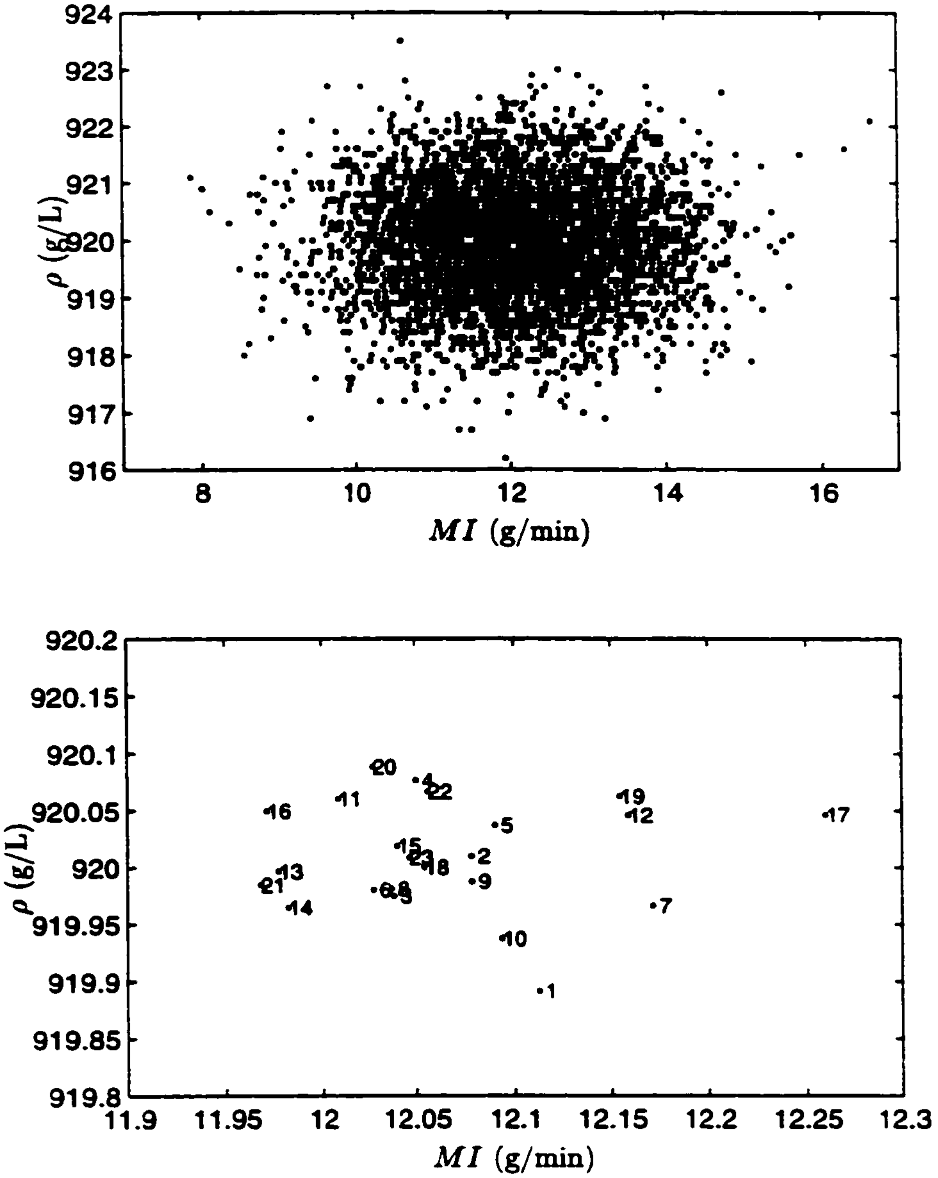


Figure 6.12: Measured melt index and density for all 23 steady-state periods (top plot: actual measurements; bottom plot: averaged measurements).

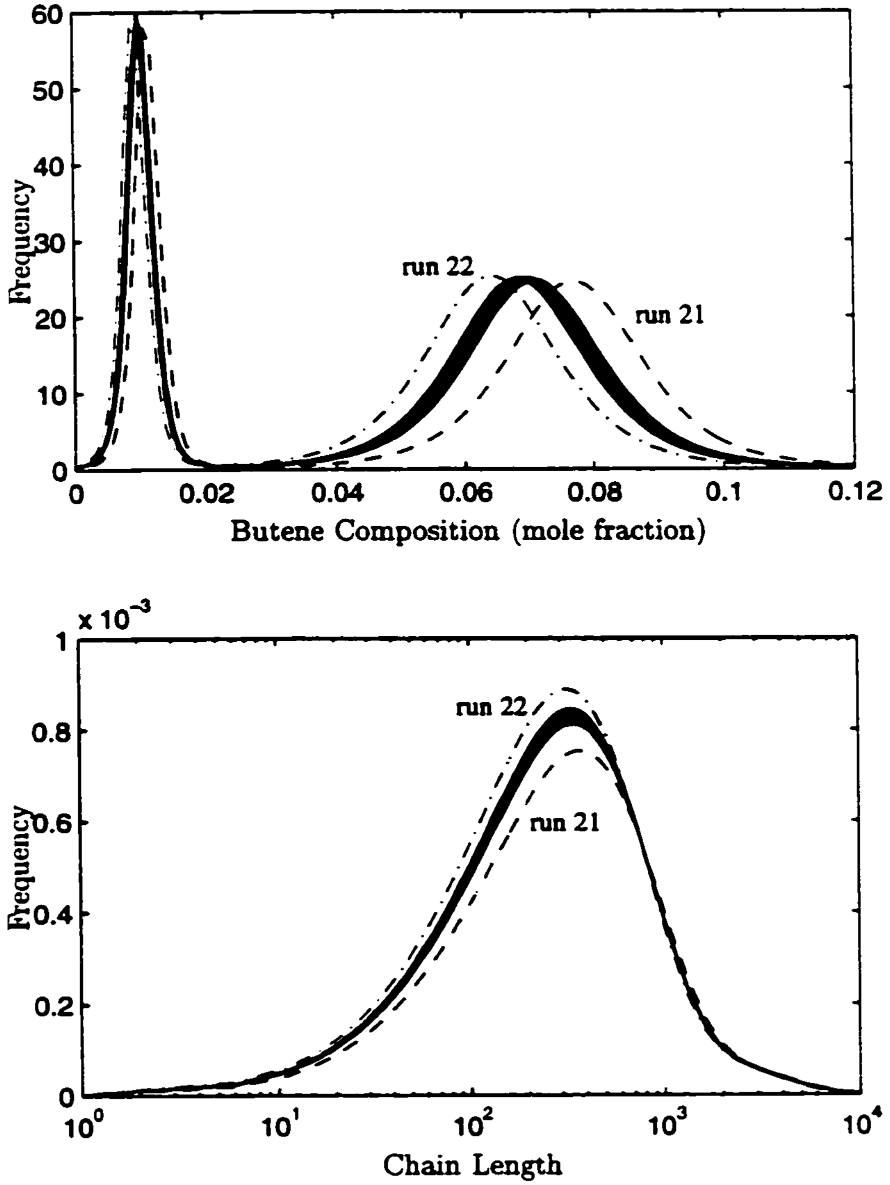


Figure 6.13: Polymer overall quality for the 23 runs, as described by the overall composition and chain length distributions (solid lines: runs 1-20 and 23, dashed line: run 21, dashed-dotted line: run 22).

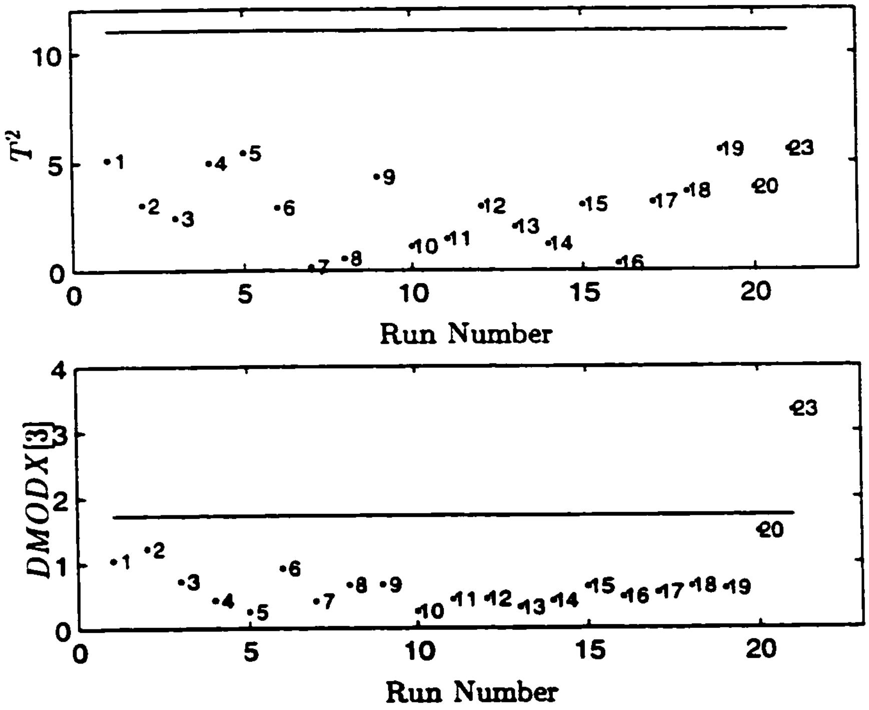


Figure 6.14: Results of the PCA model built on steady-state production periods 1-20 and 23: Hotelling's T^2 and distance to the model computed with 3 significant components.

Again, three components were found significant and a summary of modelling results is given in Table 6.4. The R^2 value gives the percentage of the total sum of squares of $\mathbf{Z_2}$ that is explained by each component ($A = 1,\ 2,\ 3$) of the fitted PCA model. The Q^2 value gives the percentage of the total sum of squares of $\mathbf{Z_2}$ that can be predicted with this model using a leave-one-out cross-validation procedure. Cumulative R^2 and Q^2 statistics are also provided. The production readiness region is developed in the score space ($\mathbf{T}$) of this new PCA model. Figure 6.15 shows the region defined for the first two components, since the problem is clearly shown in two dimensions. It consists of 95% limits for both the $t_1 - t_2$ space and for the $DMODX[2]$. However, the region is also shown for the complete model (3 components), using the Hotelling's T^2 and $DMODX[3]$.

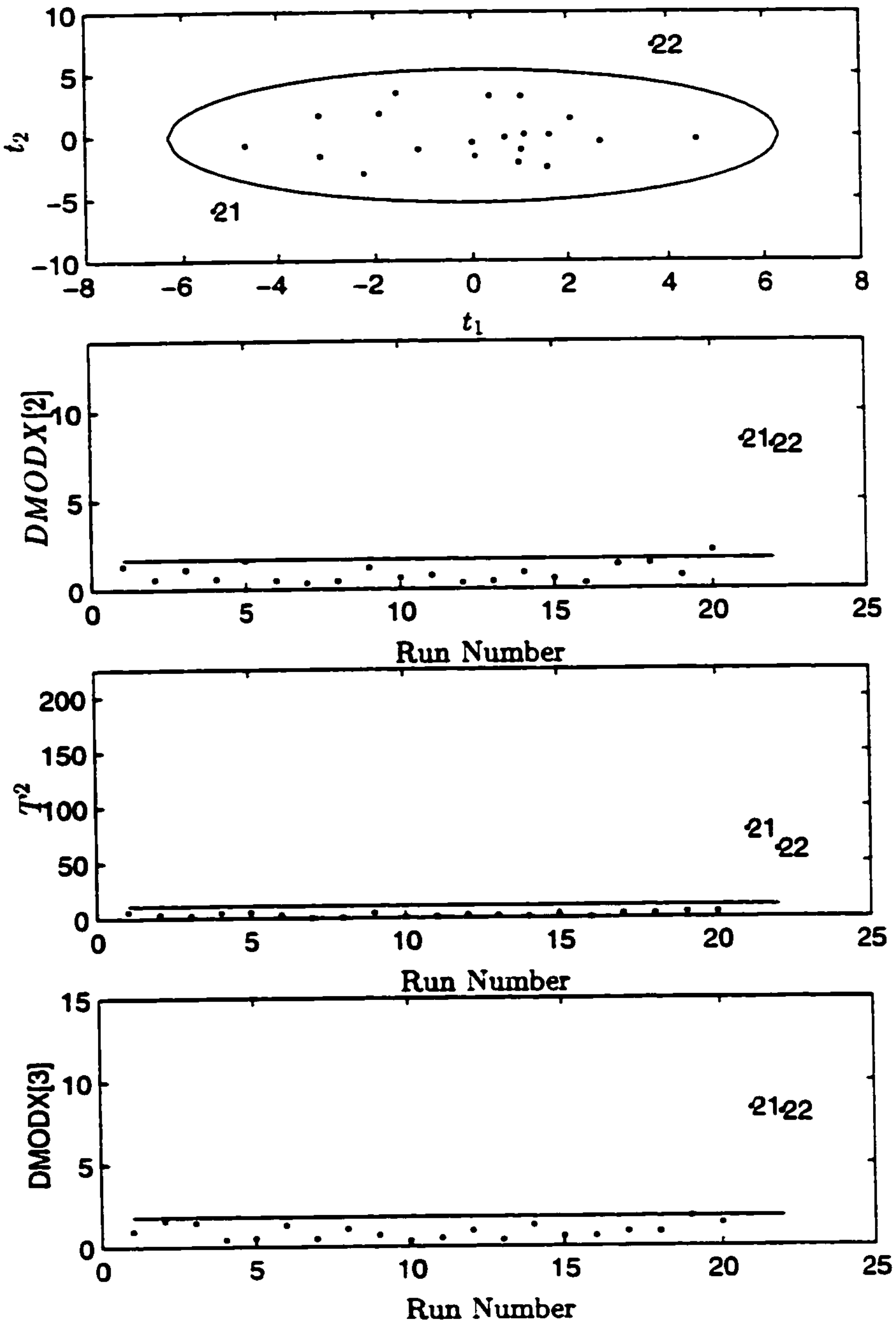


Figure 6.15: Results of the PCA model built on the 20 steady-state production periods within common cause variations: score plot and distance to the model for the first 2 components, as well as Hotelling's T^2 and distance to the model for all 3 components.

Table 6.4: Summary of the PCA model, by the means of the percent explained and percent predicted variance of $\mathbf{Z_2}$, $R^2_{Z_2}$ and $Q^2_{Z_2}$ respectively.

A	$R^2_{Z_2}$ (%)	$R^2_{Z_2,cum}$ (%)	$Q^2_{Z_2}$ (%)	$Q^2_{Z_2,cum}$ (%)
1	0.460	0.460	0.283	0.283
2	0.328	0.788	0.481	0.628
3	0.183	0.971	0.802	0.926

Steady-state periods 21 and 22 were not considered for developing the production readiness region, since customer complaints were obtained for materials produced during these periods. The averaged steady-state conditions for periods 21 and 22 were projected onto the reduced space of the PCA model, as if they would be monitored on-line, and their projection results are also shown in Figure 6.15. Clearly, these sets of operating conditions are different, both in the projection space ($t_1 - t_2$ or T^2) and in the distance from this space ($DMODX[2]$ and $DMODX[3]$), due to unusual temperature set-points. Had a production readiness region been implemented on-line, reactor temperature would have caused an alarm when approaching steady-state. This would have warned operators and engineers about the problem and immediate remedial actions could have been taken. Production of large amounts of poor quality materials could have therefore been avoided.

The results shown in Figure 6.13 warrant some additional explanations. Operation during steady-state periods 21 and 22 were at different reactor temperatures. A higher temperature increases the incorporation of ethylene monomer relative to butene and is therefore responsible for the shift of CCD to lower butene composition. Higher temperature also favors chain transfer reactions relative to propagation reactions, and leads to production of shorter polymer chains. The proportion of shorter chains in MWD is therefore greater.

6.3 Conclusion

Empirical methodologies for improving transition policies (grade changeover, start-ups and re-starts) and for defining production readiness have been presented in this chapter. These can be used for improving process and product quality when mechanistic models are not readily available. Improvement and optimization of trajectories during transitions would result in reduced amounts of off-specification materials and transition time, and this could also lead to faster reponse to market demand. On the other hand, a better definition of production readiness should ensure that a specific product grade is of high quality and is consistent with past production.

Improving start-up or grade transition from an empirical point of view, is based on the fact that in the past, some transitions were better than others, in terms of amounts of off-grade materials, transition duration and safety conditions. Developing a monitoring procedure for transient periods, based on the best transitions achieved in the past, seems to be a logical solution. This involves the use of multivariate statistical techniques such as multi-way principal component analysis (MPCA) and multi-block multi-way projection to latent structures (MBPLS). The concepts were illustrated using re-start data from an industrial polymerization process. Insightful empirical evidence was obtained by analyzing such data, on how to achieve a reduction in amounts of off-specification materials.

In flexible multi-product plants, many combinations of steady-state operating conditions may lead to achieve desired product quality specifications. However. product quality is never completely characterized and specifications are rather based only on a few measured quality variables. The impact of using different operating conditions on the unmeasured aspects of product quality is unknown. A customer buying such a product may find that it behaves differently or is more difficult to process, even if the few quality specifications are met. In this work, it was proposed to alleviate this problem by targeting processing conditions as well as the few measured quality variables to a desirable region, known to achieve good quality. Principal component analysis (PCA) was used for defining the desirable steady-state operation region. It was shown, by simulation studies of a linear low density polyethylene (LLDPE) reactor, that targeting steady-state process operation to such a desirable region improves product quality and consistency, regardless of the operating region one is transitioning from.

Chapter 7

Summary and Conclusions

The general objective of this thesis was to develop sound, empirical methodologies to solve a few important chemical engineering problems related to improvement of process operation and product quality. The following four problems were addressed in this thesis: (i) empirical optimization of batch process trajectories; (ii) improvement in the identification of non-parsimonious dynamic models using the Jackknife or the Bootstrap statistics; (iii) development of multivariate specifications for incoming raw materials to a consumer's plant, and (iv) improvement of start-ups and grade transition policies in multi-product plant. In the following sections, the work done in each area is summarized, the contributions to the field are outlined and some conclusions are drawn.

7.1 Multivariate Analysis and Optimization of Process Variable Trajectories for Batch Processes

The problem that was addressed in chapter 3 is that of obtaining the sensitivities of the final product quality in batch processes to manipulated process variables at various degrees of completion during the course of each batch. This information can then be used for process and product development as well as for improving and optimizing the policies of existing processes. The main contribution of this work is in providing a simpler, empirical approach to batch process optimization.

The proposed approach is based on adding designed experiments to currently used batch policies to generate the information required for the optimization. The type of designed experiments necessary to extract both time varying relationships, between the manipulated process variables and final product quality, and relationships that are consistent within the course of a batch was discussed. The collected data is then analyzed using the multi-way multi-block PLS method, and the regression coefficients derived from this method are used as the sensitivities. Various simulation studies of a Styrene Butadiene Rubber (SBR) emulsion copolymerization reactor were used to illustrate the proposed approach. Useful process insight for improving and optimizating this process was obtained when applying the empirical method developed in this work. It was also shown that only a few batch runs were necessary to extract relationships that are consistent during the course of a batch, but a much greater number of runs to extract time varying relationships were needed.

To help reducing the number of runs (and designed experiments), a new pathway PLS algorithm was also developed in this work, which incorporates intermediate product quality measurements collected during the course of each batch. This algorithm is valid under the assumption of linear and additive effects and a test was proposed to assess this assumption. In the simulations considered in this research, the number of runs was reduced by half, but the new algorithm still allowed the extraction of similar information. The new pathway algorithm is another important contribution to the field, as this seems to be one of the first times that intermediate quality measurements have been embedded into a single, multivariate analysis of batch data. The use of this algorithm can potentially be extended to other problems, when similar pathway relationships between blocks of data hold. A good example of another application would be the modelling and analysis of steady-state data (process variables and quality variables) collected from different process units in series.

7.2 Jackknife and Bootstrap Methods in the Identification of Dynamic Models

The problem of selecting the meta parameter in regularization methods (ridge parameter) and latent variable methods (number of latent variables) for parameter estimation of non-parsimonious dynamic models is addressed in chapter 4. Cross-validation is the default criterion used for selecting meta parameters, based on maximizing the model predictive ability, but it has been shown in the literature that this criterion suggests often keeps too few latent variables to capture the underlying process structure. It is crucial that such dynamic models capture the process structure as well as possible when these models are used for designing controllers or for process simulation under different conditions. This work contributes to the field by proposing a new criterion for selecting meta parameters that not only achieves good model predictions, but that also captures the correct process behavior (e.g. the model is closer to the true process).

The proposed criterion jointly uses two profiles, computed for a range of latent variables: (i) the model residual sum of squares (SSE) and (ii), the total variance of the estimated impulse ($SS_{\hat{\nu}}$) of step responses ($SS_{\hat{\eta}}$), obtained via the Jackknife or the Bootstrap procedures. The former profile measures how well the model fits the data with a given number of latent variables, but never gives an indication of when overfitting of the data occurs. Evidence for overfitting is given by the latter profile. which rises rapidly when noise and disturbances are being fitted by the model. The proposed criterion therefore suggest to keep adding latent variables as long as the

SSE and the total variance of the parameter estimates are decreasing or stable, and to stop when adding more latent variables leads to a sustained increase in parameter uncertainty.

Simulation studies were used to illustrate the new criterion. In all cases studied, using PLS for parameter estimation, the proposed criterion led to choosing better models (in the mean square error sense) than when these models are selected via cross-validation. It was also shown to give models that correspond closely to those having the lowest MSE deviations from the true process. It also indicates when least squares can be used for estimating the parameters without to much overfitting of the data. The new approach also provide approximate confidence intervals for the estimated impulse and step responses.

7.3 Defining Multivariate Specification Regions

Defining meaningful specification regions for incoming lots of raw materials entering a consumer's plant is essential for ensuring that consumer's meet their final product quality requirements, given the way they operate their process and their operability limits. However, in spite of their importance, there is void in the quality control literature on how to define these specification regions. This may explain why there seems to be no standard industrial practice for developing specifications; they appear to be defined in an arbitrary fashion, based on past and often subjective experiences. In addition, quality is often assessed using univariate measures while quality is, most of the time, a truly multivariate property. This research provides the first approach to developing meaningful multivariate specification regions.

The proposed approach is based on a sound, empirical analysis of historical data bases (including measurements on raw material properties, consumer's process operating conditions and final product quality characteristics). The method covers most industrial situations, in which not only the raw materials, but also the way the consumer's operate is own process may affect the final product quality. The situation where feedforward and feedback control actions are implemented (either by operators or control systems) to compensate for raw material variations is also considered. The idea consists of obtaining the latent variable space of raw material properties that affect final consumer's product quality and then, mapping the regions of raw material properties that yield good quality, given an assumption on the way the process will operate in the future. This can be done using latent variable modelling techniques. Then, the lots of final product judged by the consumer's as good quality are mapped into the raw material property space (or its approximation obtained with latent variable models) and an elliptical boundary is defined to include most good lots of raw materials while excluding most of the poorer ones. Consumer's process operability limits can also be accounted for in developing the specifications in a very similar empirical fashion. The key result of this approach is that large specification limits are placed on combinations of raw material properties that appear to have weak or no significant effect on final product quality, and tighter specifications on those combinations of properties having a strong impact on final product quality. Simulations studies of a film blowing process, used to produce various kinds of polymer films, was used to illustrate the proposed approach, in different industrial situations.

The work presented in this chapter is by far the most open ended problem treated in this thesis. It is one of the first attempts to develop a sound approach to define multivariate specification regions. This should provide a basis for future research in this area. In particular, further investigation is required in the following areas: (i) how to better define the boundary of the specification region; (ii) how to compare the sizes of specification regions developed under different assumptions, or built using different historical data sets and (iii), how to define meaningful multivariate capability indices to compare suppliers of raw materials.

7.4 Analysis of Start-Ups and Grade Transition Problems

Chapter 6 of this thesis focused on the analysis of problems related to transitions in multi-product plants (start-up, re-start, grade changeover). In particular, the following two issue were addressed: (i) finding the best transition policies to achieve a reduction in transition time and off-grade materials while ensuring safe operation and (ii), ensuring that products obtained after a transition are of high quality and are consistent with past periods of production. A common solution to the first problem is to optimize a detailed fundamental model. However, since such a good model is not always available, there was a need for developing an empirical approach for improving transition performance. On the other hand, the second problem is rarely discussed in the literature and so, a systematic study of this problem was required. The contribution of this work is twofold; it provides a sound empirical method to improve transition policies, and an approach to help ensuring high and consistent product quality.

Two solutions are proposed for improving transition performance, depending on the number of transition policies that have been implemented in the past. When multiple distinct policies have been implemented, it is straightforward to identify which policy is the best, but one could take advantage of the larger range of variations to gain insight on the features of transition policies associated with desired performance. This is achieved using the method proposed in this thesis for batch optimization (chapter 3). On the other hand, when only one transition policy has been used in past, but with variations in its implementation, then one could build a monitoring scheme based on the most desirable transitions achieved in the past. The monitoring scheme would allow for removing variations around the desired policy and a more consistent operation should result, as well as associated savings. The second solution (only one transition policy) was illustrated using an industrial polymerization example.

The problem with ensuring high and consistent product quality after a transition arises from the fact that multi-product plants may often be flexible enough to achieve similar measured product quality characteristics using different sets of steady-state operating conditions. However, since these quality measurements are almost never sufficient to completely characterize "overall" product quality, the impact of using various sets of steady-state conditions on the unmeasured quality variables is unknown. It is therefore suggested to always target process operation into the same region (defined for a specific product). This region is developed using already existing monitoring techniques (based on PCA and PLS) and past successful steady-state production periods. The concepts related to the definition of "production readiness" regions were illustrated using simulations of a linear low density polyethylene reactor.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **AC-19**, 716–723.

Anderson, T.W. (1984). *An introduction to multivariate statistical analysis.* second ed.. Wiley and Sons, Inc.. New-York.

Bakshi, B.R. and G. Stephanopoulos (1994*a*). Representation of process trends–iii. multiscale extraction of trends from process data. *Computers and Chemical Engineering* **18**, 267–302.

Bakshi, B.R. and G. Stephanopoulos (1994*b*). Representation of process trends–iv. induction of real-time patterns from operating data. *Computers and Chemical Engineering* **18**, 303–332.

Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis, Forecasting and Control.* Holden-Day. Oakland, Ca.

Box, G.E.P. and N.R. Draper (1969). *Evolutionary Operation: A Statistical Method for Process Improvement.* Wiley. New-York.

Brinkley, S. (1991). The team way to quality raw materials. *ASQC Quality Congress Transactions* **Milwaukee**, 497–500.

Broadhead, T.O. (1984). Dynamic modelling of the emulsion copolymerization of styrene/butadiene. M.Eng. thesis. McMaster University, Hamilton, Ontario, Canada. Department of Chemical Engineering.

Broadhead, T.O., A.E. Hamielec and J.F. MacGregor (1985). Dynamic modelling of the batch, semi-batch and continuous production of styrene/butadiene copolymers by emulsion polymerization. *Makromol. Chem. Suppl.* **10/11**, 105–128.

Brown, R.H. (1990). The desirability function: A process optimization tool. Technical report. Eastman Chemical Company.

Burnham, A.J., R. Viveros and J.F. MacGregor (1996). Frameworks for latent variable multivariate regression. *Journal of Chemometrics* **10**, 31–45.

Camo-ASA (1998). The Unscrambler, Box 1405 Vika, N-0115 Oslo, Norway.

Cawthon, G.D. and K.S. Knaebel (1989). Optimization of semibatch polymerization reactions. *Computers Chemical Engineering* **13**, 63–72.

Chan, L.K.S., W. Cheng and F.A. Spiring (1991). A multivariate measure of process capability. *International Journal of Modelling and Simulations* **11**, 1–6.

Clarke-Pringle, T.L. and J.F. MacGregor (1998). Optimization of molecular-weight distribution using batch-to-batch adjustments. *Industrial Engineering and Chemistry Research* **37**, 3660–3669.

Dayal, B.S. and J.F. MacGregor (1996). Identification of finite impulse response models: Methods and robustness issues. *Ind. Eng. Chem. Res.* **35**, 4078–4090.

Dayal, B.S. and J.F. MacGregor (1997). Multi-output process identification. *Journal of Process Control* **7**, 269–282.

De Smet, J.A. (1993). Development of multivariate specification limits using partial least squares regression. M. Eng. thesis. University McMaster, Hamilton, Ontario, Canada. Department of Chemical Engineering.

Debling, J.A., G.C. Han, F. Kuijpers, J. VerBurg, J. Zacca and W.H Ray (1994). Dynamic modeling of product grade transitions for olefin polymerization processes. *AIChE Journal* **40**, 506–520.

Dong, D., T. McAvoy and E. Zafiriou (1996). Batch to batch optimization using neural networks models. In: *IFAC 13th Triennial World Congress, San Francisco, CA*. pp. 253–258.

Draper, N.R. and H. Smith (1981). *Applied Regression Analysis*. second ed.. John Wiley and Sons. New-York.

Duchesne, C. and J.F. MacGregor (2000a). Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics and Intelligent Laboratory Systems* **51**, 125–137.

Duchesne, C. and J.F. MacGregor (2000b). Jackknife and bootstrap methods in the identification of dynamic models. *Journal of Process Control*. in press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. (1983). *The jackknife, the bootstrap, and other resampling plans*. CBMS-NSF regional conference series in applied mathematics. 38. Society for Industrial and Applied Mathematics. Philadelphia, Pa.

Efron, B. and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*. Monographs on statistics and applied probability. 57. Chapman and Hall. New-York.

Faber, K. and B.R. Kowalski (1997). Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *Journal of Chemometrics* **11**, 181–238.

Farber, J.N. and R.L. Laurence (1986). Optimization of continuous polymerization reactors: Start-up and change of specification. *Macromol. Chem., Macromol. Symp.* **2**, 193–207.

Fathi, Y. (1990). Producer–consumer tolerances. *Journal of Quality Technology* **22**, 138–145.

Filippi-Bossy, C., J. Bordet, J. Villermaux, S. Marchal-Brassely and C. Georgakis (1989). Batch reactor optimization by use of tendency models. *Computers and Chemical Engineering* **13**, 35–47.

Flender, M., G. Fieg and G. Wozny (1996). Classification of new product changeover strategy (nps) for different application of distillation columns. *Computers and Chemical Engineering* **20, Suppl.**, S1131–S1136.

Geladi, P. (1988). Notes on the history and nature of partial least squares (pls) modelling. *Journal of Chemometrics* **2**, 231–246.

Geladi, P. and B.R. Kowalski (1986). Partial least-squares regression: A tutorial. *Analytica Chemica Acta* **185**, 1–17.

Gray, H.L. and R.R. Schucany (1972). *The generalized jackknife statistic.* Statistics: textbooks and monographs; v.1. Marcel Dekker. New-York.

Hamielec, A.E. and H. Tobita (1992). Polymerization processes. In: *Ullman's Encyclopedia of Industrial Chemistry.* Vol. A21. VCH Verlagsgesellschaft mbH, Weinheim.

Harrington, E.C. (1965). The desirability function. *Industrial Quality Control Journal* **April**, 494–498.

Hinkley, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19**(3), 285–292.

Hoerl, A.E. and R.W. Kennard (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hoerl, A.E. and R.W. Kennard (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69–82.

Höskuldsson, A. (1988). Pls regression methods. *Journal of Chemometrics* **2**, 211–228.

Jaeckle, C.M. and J.F. MacGregor (1998). Product design through multivariate statistical analysis of process data. *AIChE Journal* **44**, 1105–1118.

Jaeckle, C.M. and J.F. MacGregor (2000). Industrial applications of product design through the inversion of latent variable models. *Chemometrics and Intelligent Laboratory Systems* **50**, 199–210.

Juran, J.M., F.M. Gryna and R.S. Binham (1974). *Quality Control Handbook.* McGraw-Hill. New-York.

Kassidas, A., J.F. MacGregor and P.A. Taylor (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE Journal* **44**, 864–875.

Kettaneh-Wold, N., J.F. MacGregor, B. Dayal and S. Wold (1994). Multivariate design of process experiments (m-dope). *Chemometrics and Intelligent Laboratory Systems* **23**, 39–50.

Kotz, S. and N.L. Johnson (1993). *Process capability indices.* Chapman and Hall. London.

Kourti, T. and J.F. MacGregor (1995). Process analysis, monitoring and diagnosis using multivariate projection methods - a tutorial. *Chemometrics and Intelligent Laboratory Systems* **28**, 3–21.

Kourti, T. and J.F. MacGregor (1996). Multivariate spc methods for process and product monitoring. *Journal of Quality Technology* **28**, 409–428.

Kourti, T., J. Lee and J.F. MacGregor (1996). Experiences with industrial applications of projection methods for multivariate statistical process control. *Computers in Chemical Engineering* **20 Suppl.**, S745–S750.

Kourti, T., P. Nomikos and J.F. MacGregor (1995). Analysis, monitoring and fault detection of batch processes using multiblock and multiway pls. *Journal of Process Control* **5**, 277–284.

Kresta, J.V. (1992). The Application of Partial Least Squares to Problems in Chemical Engineering. Ph.D. thesis. University McMaster, Hamilton, Ontario, Canada. Department of Chemical Engineering.

Kresta, J.V., T.E. Marlin and J.F. MacGregor (1994). Development of inferential process models using pls. *Computers in Chemical Engineering* **18**, 597–611.

Ljung, L. (1999). *System Identification: Theory for the User.* second ed.. Prentice Hall Press, Englewood Cliffs. N.J.

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

MacGregor, J.F., T. Kourti and J.V. Kresta (1991). Multivariate identification: A study of several methods. *IFAC ADCHEM Conference Proceedings, Toulouse, France.*

Martens, H. and M. Martens (1999). Modified jack-knife estimation of parameter uncertainty in bilinear modelling (plsr). *Accepted for publication in Journal of Food Quality and Preference.*

McAuley, K.B. (1992). Modelling, Estimation and Control of Product Properties in a Gas Phase Polyethylene Reactor. Ph.D. thesis. University McMaster, Hamilton, Ontario, Canada. Department of Chemical Engineering.

McAuley, K.B. and J.F. MacGregor (1992). Optimal grade transitions in a gas phase polyethylene reactor. *AIChE Journal* **38**, 1564–1576.

McAuley, K.B. and J.F. MacGregor (1993). Nonlinear product property control in industrial gas-phase polyethylene reactor. *AIChE Journal* **39**, 855–866.

McAuley, K.B., J.F. MacGregor and A.E. Hamielec (1990). A kinetic model for industrial gas-phase ethylene copolymerization. *AIChE Journal* **36**, 837–850.

Miller, R.G. (1974). An unbalanced jackknife. *Ann. Statist.* **2**, 880–891.

Nomikos, P. and J.F. MacGregor (1994*a*). Monitoring of batch processes using multiway principal component analysis. *AIChE Journal* **40**, 1361–1375.

Nomikos, P. and J.F. MacGregor (1994*b*). Multivariate spc charts for monitoring batch processes. *Technometrics* **37**, 1361–1375.

Nomikos, P. and J.F. MacGregor (1995). Multiway partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* **30**, 97–108.

Ohshima, M., I. Hashimoto, T. Yoneyama, M. Takeda and F. Gotoh (1994). Grade transition control for an impact copolymerization reactor. In: *IFAC Symposium, ADCHEM'94, Kyoto, Japan, May 25-27.* pp. 507–512.

Ortega, J.M. (1987). *Matrix theory: a second course.* Plenum Press. New-York.

Pearn, W.L., S. Kotz and N.L. Johnson (1992). Distributional and inferential properties of process capability indices. *Journal of Quality Technology* **24**, 216–231.

Phatak, A. (1993). Evaluation of Some Multivariate Methods and their Applications in Chemical Engineering. Ph.D. thesis. University of Waterloo, Waterloo, Ontario, Canada. Department of Chemical Engineering.

Phatak, A., P.M. Reilly and A. Penlidis (1993). An approach to interval estimation in partial least squares regression. *Analytica Chemica Acta* **277**, 495–501.

Polydynamics (1996). B-Filmcad 1.0, Polydynamics Inc., 1685 Main St. West, Suite 305, Hamilton, Ontario, Canada L8S 1G5.

Quenouille, M. (1949). Approximate tests of correlation in time series. *J. Royal. Statist. Soc.* **B11**, 18–44.

Redlich, H.A. (1996). Hands-on guide for controlling quality of raw materials. *The American Ceramic Society Bulletin* **75**, 71–72.

Ricker, N.L. (1988). The use of biased least-squares estimators for parameters in discrete-time pulse-response models. *Ind. Eng. Chem. Res.* **27**, 343–350.

Roney, S.D. (1998). Development of inferential sensors for chemical processes using partial least squares (pls). M.Eng. thesis. McMaster University, Hamilton, Ontario, Canada. Department of Chemical Engineering.

Sargent, R.W.H. and G.R. Sullivan (1979). Development of feed changeover policies for refinery distillation units. *Industrial Engineering Chemistry and Process Design and Development* **18**, 113–124.

Shao, J. and C.F.J. Wu (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17**(3), 1176–1197.

Shi, R. and J.F. MacGregor (2000). Modeling of dynamic systems using latent variable and subspace methods. *Journal of Chemometrics.* in press.

Sidiropoulos, V. (1995). Comparison of experiments with a model of the blown film process. M. Eng. thesis. University McMaster, Hamilton, Ontario, Canada. Department of Chemical Engineering.

Söderström, T. and P. Stoica (1989). *System Identification.* Prentice-Hall International. U.K.

Ståhle, L. and S. Wold (1987). Partial least squares analysis with cross-validation for the two-class problem: A monte carlo study. *Journal of Chemometrics* **1**, 185–196.

Taguchi, G. (1986). *Introduction to Quality Engineering.* White Plains. New-York.

Taguchi, G., E.A. Elsayed and T. Hsiang (1989). *Quality Engineering in Production Systems.* McGraw-Hill. New-York.

Terwiesch, P., M. Agarwal and D.W.T. Rippin (1994). Batch unit optimization with imperfect modelling: A survey. *Journal of Process Control* **4**, 238–258.

Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Statist.* **29**, 614.

Umetrics (1998). SIMCA-P 7.0, Umetrics Inc., 371 Highland Ave., Winchester, MA01890, U.S.A.

Valle, S., W. Li and S.J. Qin (1999). Selection of the number of principal components: The variance of reconstruction error criterion with a comparison to other methods. *Industrial and Engineering Chemistry Research* **38**, 4389–4401.

Verwijs, J.W., P.H. Kösters, H. van den Berg and K.R. Westerterp (1995). Reactor operating procedures for startup of continuously-operated chemical plants. *AIChE Journal* **41**, 148–158.

Wakeling, I.N. and J.J. Morris (1993). A test of significance for partial least squares regression. *Journal of Chemometrics* **7**, 291–304.

Wang, Y., H. Seki, S. Ooyama, K. Akamatsu, M. Ogawa and M. Ohshima (2000). A nonlinear predictive control for optimal grade transitions of polymerization reactors. In: *IFAC ADCHEM Preprints Vol.* **II**, *Pisa, Italy*. pp. 725–730.

Westerhuis, J.A. and P.M.J. Coenegracht (1997). Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics* **11**, 379–392.

Westerhuis, J.A., T. Kourti and J.F. MacGregor (1998). Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics* **12**, 301–321.

Westerhuis, J.A., T. Kourti and J.F. MacGregor (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics* **13**, 397–413.

Wierda, S.J. (1993). A multivariate process capability index. In: *ASQC Quality Congress Transactions–Boston*. pp. 342–348.

Wierda, S.J. (1994). Multivariate statistical process control - recent results and directions for future research. *Statistica Neerlandica* **48**, 147–168.

Wise, B.M. and N.L. Ricker (1992). Identification of finite impulse response models by principal components regression: Frequency-reponse properties. *Process Control and Quality* **4**, 77–86.

Wise, B.M. and N.L. Ricker (1993). Identification of finite impulse response models with continuum regression. *Journal of Chemometrics* **7**, 1–14.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**(4), 397–405.

Wold, S., K. Esbensen and P. Geladi (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 37–52.

Wold, S., M. Sjöström, R. Carlson, T. Lundstedt, S. Hellberg, B. Skagerberg, C. Wiksström and J. Öhman (1986). Multivariate design. *Analytica Chemica Acta* **191**, 17–32.

Wold, S., N. Kettaneh, H. Friden and A. Holmberg (1998). Modelling and diagnosis of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory Systems* **44**, 331–340.

Wurl, R.C., S.L. Albin and I.J. Shiffer (1999). Multivariate monitoring of batch process startup. In: *Fall Technical Conference, Houston, October.*

Xie, T., K.B. McAuley, J.C.C. Hsu and D.W. Bacon (1994). Gas phase ethylene polymerization: Production processes, polymer properties, and reactor modeling. *Industrial and Engineering Chemistry Research* **33**, 449–479.

Yarborough, M.J. (1995a). Raw material specifications. *The American Ceramic Society Bulletin* **74**, 49–50.

Yarborough, M.J. (1995b). Raw material specifications. *Ceramic Engineering and Science Proceedings* **16**, 298–301.

Zafiriou, E. and J. Zhu (1990). Optimal control of semi-batch processes in presence of modelling errors. In: *American Control Conference, San Diego, CA.*

Zullo, L. (1996). Validation and verification of continuous plants operating modes using multivariate statistical methods. *Computers and Chemical Engineering* **20, Suppl.**, S683–S688.

Appendix A

Pathway-PLS Algorithm

In this section only the new PLS pathway algorithm is described. However, one could also incorporate in the analysis a **Z** data matrix as a separate block in a straightforward fashion. The multi-block pathway algorithm would consist of any multi-block algorithm (Westerhuis *et al.*, 1998) in which the part corresponding to **X** is replaced by the pathway algorithm described below.

For simplicity, the "stair" matrix consisting of $\mathbf{X_k}$ blocks, $k = 1, 2, 3$ (Figure 3.4), is designated as $\mathbf{X_s}$. Similarly, the long matrix filled with $\mathbf{Y_k}$ blocks is designated as $\mathbf{Y_s}$. Note that **Y** variables collected in each intermediate samples ($\mathbf{Y_k}$ blocks) should be located in the first columns of $\mathbf{Y_s}$. It is also assumed that the data ($\mathbf{X_s}$, $\mathbf{Y_s}$) is mean-centered and appropriately scaled prior to be used in the algorithm. Each $\mathbf{X_k}$ and $\mathbf{Y_k}$ blocks are mean-centered and scaled independently prior to be arranged as in Figure 3.4. Auto-scaling has been used throughout this work. The modified NIPALS algorithm is shown below as a pseudo-Matlab code, where the ":" operator as in "$a_1 : a_2$" means from a_1 to a_2. In addition, the following indexes are defined: f is the number of **Y** blocks, I is the total number of batches, J is the number of process variable trajectories, M_b is the number of quality variables in block $\mathbf{Y_b}$ and M_{max} is the maximum number of quality variables among all $\mathbf{Y_k}$ blocks.

```
1. Start with $\mathbf{u}$ set to a full length column of $\mathbf{Y_s}$
2. For $b = 1$ to $f$
        $a_1 = 1 + (b-1)J$, $a_2 = J + (b-1)J$, $a_3 = 1 + (b-1)I$
        $\mathbf{w}(a_1 : a_2, 1) = \mathbf{X_s}(a_3 : fI, a_1 : a_2)^T\ \mathbf{u}(a_3 : fI, 1)\ /\ \mathbf{u}(a_3 : fI, 1)^T\ \mathbf{u}(a_3 : fI, 1)$
   end
3. Scale $\mathbf{w}$ to unit length
4. For $b = 1$ to $f$
        $a_1 = 1 + (b-1)I$, $a_2 = I + (b-1)I$, $a_3 = 1 + (b-1)J$, $a_4 = J + (b-1)J$
        $\mathbf{t}(a_1 : a_2, 1) = \mathbf{X_s}(a_1 : a_2, 1 : bJ)\ \mathbf{w}(1 : bJ, 1)\ /\ \mathbf{w}(1 : bJ, 1)^T\ \mathbf{w}(1 : bJ, 1)$
   end
5. For $m = 1$ to $M_{max}$
        $c$ equals all observation numbers from $\mathbf{Y_s}$ for which measurements on variable $m$ are available.
        $\mathbf{q}(1, m) = \mathbf{Y_s}(c, m)^T\ \mathbf{t}(c, 1)\ /\ \mathbf{t}(c, 1)^T\ \mathbf{t}(c, 1)$
   end
6. For $b = 1$ to $f$
        $a_1 = 1 + (b-1)I$, $a_2 = I + (b-1)I$
        $\mathbf{u}(a_1 : a_2, 1) = \mathbf{Y_s}(a_1 : a_2, 1 : M_b)\ \mathbf{q}(1 : M_b, 1)\ /\ \mathbf{q}(1 : M_b, 1)^T\ \mathbf{q}(1 : M_b, 1)$
   end
```

```
7. Check convergence of $\mathbf{u}$. If no convergence, go to 2

8. For $b = 1$ to $f$
        $a_1 = 1 + (b-1)J$, $a_2 = J + (b-1)J$, $a_3 = 1 + (b-1)I$
        $\mathbf{p}(a_1 : a_2, 1) = \mathbf{X}_s(a_3 : fI, a_1 : bJ)^T \; \mathbf{t}(a_3 : fI, 1) \; / \; \mathbf{t}(a_3 : fI, 1)^T \; \mathbf{t}(a_3 : fI, 1)$
   end

9. Compute residual matrix for $\mathbf{X}$ block:
   For $b = 1$ to $f$
        $a_1 = 1 + (b-1)I$, $a_2 = 1 + (b-1)J$, $a_3 = J + (b-1)J$
        $\mathbf{E}(a_1 : fI, a_2 : bJ) = \mathbf{X}_s(a_1 : fI, a_2 : bJ) - \mathbf{t}(a_1 : fI, 1) \; \mathbf{p}(a_2 : a_3, 1)^T$
   end

10. Compute residual matrix for $\mathbf{Y}$ block:
    For $b = 1$ to $f$
        $a_1 = 1 + (b-1)I$, $a_2 = I + (b-1)I$
        $\mathbf{F}(a_1 : a_2, 1 : M_b) = \mathbf{Y}_s(a_1 : a_2, 1 : M_b) - \mathbf{t}(a_1 : a_2, 1) \; \mathbf{q}(1 : M_b, 1)^T$
    end

11. Store $\mathbf{w}$, $\mathbf{p}$, $\mathbf{t}$, $\mathbf{u}$ and $\mathbf{q}$ in $\mathbf{W}$, $\mathbf{P}$, $\mathbf{T}$, $\mathbf{U}$, $\mathbf{Q}$ respectively

12. Compute next dimension by returning to 1, using $\mathbf{E}$ and $\mathbf{F}$ as the new $\mathbf{X}_s$ and $\mathbf{Y}_s$
```

Note that if the same quality variables are measured in all intermediate samples (all M_b's are equal), then steps 5, 6 and 10 become the same as the standard NIPALS algorithm:

5. $\mathbf{q} = \mathbf{Y_s}^T \mathbf{t} \;/\; \mathbf{t}^T \mathbf{t}$

6. $\mathbf{u} = \mathbf{Y_s} \mathbf{q} \;/\; \mathbf{q}^T \mathbf{q}$

10. $\mathbf{F} = \mathbf{Y_s} - \mathbf{t}\, \mathbf{q}^T$

Appendix B

Comments on the Evaluation of Standard Errors

Other methods are available and specifically developed for the estimation of standard errors in PLS parameter estimates. These methods are also limited to cases where the predictor matrix, **X**, is full rank. Instead of using a Monte Carlo type of resampling procedure, these alternative methods are rather based on a local linearization of the PLS estimators. Those estimators are slightly non-linear statistics and therefore the linearization, using a Taylor expansion around the measured data (**X** and **y**), is only an approximation. Höskuldsson (1988) proposed a zeroth order expansion, which assumes that regression coefficients estimated using PLS with a dimensions ($\hat{\beta}^a_{PLS}$) is independent of **y**. Later, Phatak *et al.* (1993) refined this expression with a first order linearization. A closed form expression is obtained, but requires the evaluation of a Jacobian matrix (derivatives of $\hat{\beta}^a_{PLS}$ with respect to **y**) that is relatively computationally intensive to obtain. Phatak's expression also suffers from numerical instability when the number of latent variables gets large. More recently, Faber and Kowalski (1997) adopted a Error-In-Variable (EIV) approach to standard error estimation in PCR and PLS. They derived a new expression for the situation where both $\mathbf{X}$ and $\mathbf{y}$ variables are corrupted by a noise. This expression includes the one proposed by Phatak *et al.* (1993) as a special case, when most of the noise comes from $\mathbf{y}$.

To justify the use of the jackknife in this work, its performance is compared to the bootstrap and to Phatak's approach. In identification, it is reasonable to assume that most of the noise comes from $\mathbf{y}$, so Faber and Kowalski's expression would lead to the same results as Phatak's, and is therefore not used here. The process under investigation is the one corrupted by an ARI disturbance, identified using a FIR structure using 300 observations (section 4.5.2). The basis for comparison of the above three methods is their closeness to the true standard error of each of the 175 estimated impulse weights. The true standard errors are estimated via a Monte Carlo simulation. A total of 500 data sets, each containing 300 data points, were generated using the MISO process model described in section 4.4. On each data set, a PLS model is built including 30 latent variables. The standard error on each individual parameter was computed with all 500 replicates, which is assumed to be precise enough to be representative of the truth.

Results of the comparison are shown in Figure B.1. Two variations of the approach of Phatak *et al.* are shown, which differ in the way the natural error in $\mathbf{y}$ (σ_y^2) is estimated: using the PLS residuals with modified degrees of freedom ($\widehat{ste}_{Phatak,1}$), and with the residuals of standard Least-Squares ($\widehat{ste}_{Phatak,2}$). Both the jackknife and the bootstrap standard error estimates ($\widehat{ste}_{jack}$ and $\widehat{ste}_{boot}$) are computed using 30 groups. The plain line in each plot represents the locus of perfect prediction of the true standard error (ste) on each parameter estimate. The dots correspond to the estimated standard error on each impulse weight, using the three considered methods. In this example, the jackknife, the bootstrap and Phatak's approaches have been used on only 5 randomly selected data sets from the 500 generated by the Monte Carlo study. Performing the three estimation methods on the 500 data sets would be extremely computationally intensive and is beyond the scope of this work. A few data sets provide a rough idea about the variance of each method in estimating standard error.

Using Phatak's approach as published (Phatak *et al.*, 1993) leads to estimates of standard error that are approximately two orders of magnitude greater than the true values. This is attributed to the inverse of the matrix, that the authors called **M**, which is numerically unstable. It involves a Krylov matrix that is very poorly conditioned in this case due to the large number of latent variables ($a = 30$). Conditioning problems with this expression have already been recognized by the authors (Phatak, 1993). A pseudo-inverse has therefore been used in this case to overcome the ill-conditioning problem, but both variations of Phatak's approach grossly underestimated the standard errors. This method works extremely well in some cases, as shown in Phatak *et al.* (1993), but ill-conditioning problems do limit its ability to provide precise standard error estimates, especially when several latent variables are kept in the model.

On the other hand, both the jackknife and the bootstrap give satisfactory results. The jackknife seems slightly better, but the bootstrap could perform equally well if more samples, say 200 or more, would be generated. This emphasizes the need to use a large number of subsamples in bootstrapping, which makes it more computationally intensive. Both resampling methods tend to slightly overestimate the standard error. Although unbalanced sampling could explain overestimation, it does not seem to be the case here. The fact that expressions 4.3 and 4.4 are not guaranteed to be unbiased is a more likely explanation. In addition, the slight nonlinear behavior of the PLS estimates could potentially cause overestimation as well. The impact that non-linear statistics have on jackknife estimates of standard errors is discussed in Efron and Tibshirani (1993). However, the inflation of the standard error estimates is not severe enough to invalidate the use of the jackknife.

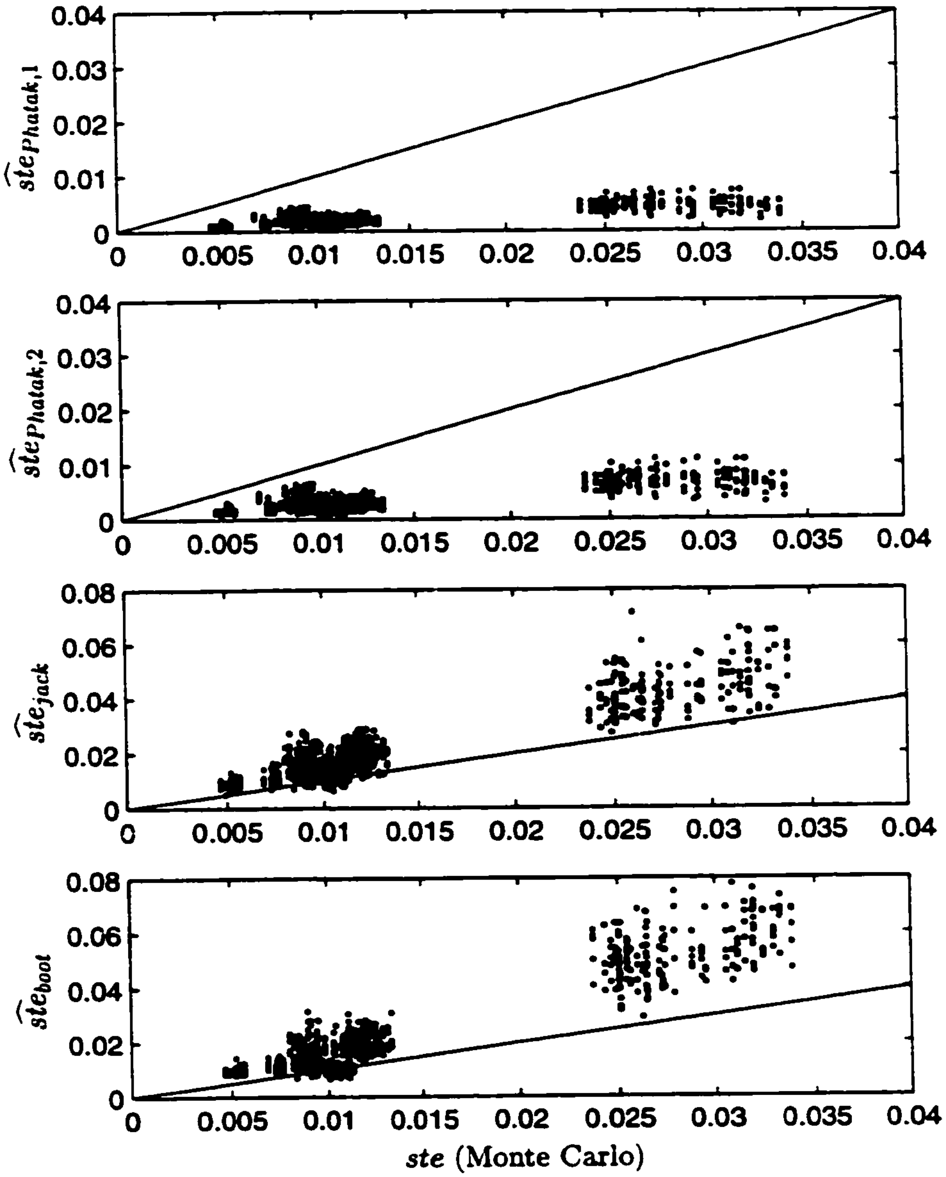


Figure B.1: Comparison of true (plain line) and estimated (dots) standard error of PLS parameter estimates using various methods: Phatak's with least-squares and PLS estimates of error variance ($\widehat{ste}_{Phatak,1}$ and $\widehat{ste}_{Phatak,2}$), jackknife and bootstrap.

In addition to being less computationally intensive, the jackknife algorithm shares a high degree of similarity with cross-validation, which is already implemented in most multivariate statistical software packages. The main difference consists in storing the estimated parameters after each cross-validation round. This is another advantage for using the jackknife which seems to be a good candidate for use in identification.