THE APPLICATION OF PARTIAL LEAST SQUARES
TO
PROBLEMS IN CHEMICAL ENGINEERING

By

JAMES VACLAV KRESTA, M. Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree of
Doctor of Philosophy

McMaster University
(c) Copyright by James Vaclav Kresta, April 1992
APPLICATIONS OF PARTIAL LEAST SQUARES
DOCTOR OF PHILOSOPHY (1992)
(Chemical Engineering)

McMASTER UNIVERSITY
Hamilton, Ontario, Canada

TITLE: Application of Partial Least Squares to Problems in Chemical Engineering

AUTHOR: James Vaclav Kresta,
B.Sc.Eng. (University of New Brunswick)
M.Sc. (University of Leeds, UK)

ADVISOR: Professor John F. MacGregor
Professor Thomas E. Marlin

NUMBER OF PAGES: 195
Abstract

The objectives of this work were to investigate the applicability of using Partial Least Squares (PLS) in three areas of chemical engineering: 1) for performance monitoring, 2) for inferential model development and 3) in dynamic model identification. One further goal of this work was to gain understanding of the properties of the PLS method in the context of chemical engineering. Properties investigated included: 1) effects of scaling, 2) the objective function in the algorithm and 3) the configuration of the X and Y data blocks. The first application of PLS was in the area of performance monitoring of a multivariate processes. A multivariate Statistical Process Control (SPC) procedure has been proposed for handling large numbers of process and quality variables. PLS is used to reduce the dimensionality of these large and highly correlated data sets down to a few latent variables which contain most of the information about the process behaviour under normal operating conditions. By compressing all the information on the process down to low dimensional spaces, and using simple plots of the data in these spaces, together with meaningful control limits, the essential idea and philosophy of Shewart's (1931) SPC methods have been preserved and extended to handle the large number of variables collected in most process industries today. PLS has also been shown to be a very powerful approach to inferential model building when large numbers of highly correlated measured variables are available. By retaining all the measurements without overfitting the data PLS is able to utilize all the information obtained from the process measurements. Further, PLS models are extremely robust to
missing data and sensor failures, an important feature of any inferential control scheme. The importance of obtaining a representative reference data set, (ie. one that represents the process relationships for the expected range of process conditions, and the structure of the process and control configuration under which the model is to be used), when using empirical models has also been demonstrated. The comparison of PLS, Ridge Regression (RR) and Least Squares identification of non-parsimonious (Finite Impulse Response (FIR) and Autoregressive exogeneous (ARX)) models for multiple input single output (MISO) systems shows that the biased estimates obtained from PLS and RR provide smoother estimates of the impulse and step weights. The difficulty in characterizing structured noise when using these non-parsimonious models is discussed. Finally, the use of PLS for identification of multiple input multiple output (MIMO) processes was investigated.
Acknowledgements

There are two groups of people I would like to acknowledge for their significant contributions to making this Ph.D. possible.

The first are my parents; who, in 1970, left a comfortable life, family, and almost all material wealth in Czechoslovakia in order that their children could have the opportunities not available in our native country at that time. They followed up this commitment by teaching us the values of hard work, a love of knowledge and to take advantage of the opportunities available to us. Finally they provided the support, both financial and emotional, for us to achieve whatever we desired. Thank you very much, I appreciate all your sacrifices more than you will ever know.

The second group are all the people who helped me learn. Special thanks go to: Dr. MacGregor and Dr. Marlin who supervised this work, added insightful comments and kept me going when I thought I wasn’t getting anywhere; and the graduate students of McMaster who made this Ph.D. much more than what is contained on these pages.

Finally, I would like to especially thank Suzanne with whom the past ten years of learning, exploring the world and meeting interesting people has been so much fun that I can’t wait for the next stage to begin.
Table of Contents

Chapter 1. Introduction ................................................................. 1

Chapter 2. Background, Theory and Application Issues ...................... 5
  2.1 Introduction ........................................................................ 5
  2.2 Notation, Conventions and Definitions .................................. 9
  2.3 Conceptual Formulation ....................................................... 11
  2.4 Mathematical Formulation .................................................. 16
    2.4.1 Principal Component Analysis (PCA) ........................... 17
    2.4.2 PLS and the NIPALS Algorithm ................................ 19
    2.4.3 Comparison of PLS, MLR and PCR .............................. 24
  2.5 Application questions as related to the problems studied .......... 32
  2.6 Summary ........................................................................... 36
  2.7 Notation ............................................................................ 38
  2.8 References ....................................................................... 40

Chapter 3. Application of PCA and PLS to Process Monitoring .......... 43
  3.1 Introduction .................................................................... 43
  3.2 Monitoring via multivariate SPC plots .................................. 44
  3.3 Monitoring a fluidized bed reactor ....................................... 51
  3.4 Monitoring a Methanol-Acetone-Water Distillation Column ...... 58
  3.5 Discussion of Present Results ............................................ 65
  3.6 Discussion of other applications of this method ..................... 67
  3.7 Conclusions ..................................................................... 68
  3.8 Notation .......................................................................... 69
  3.9 References ..................................................................... 70

Chapter 4. Inferential variables ...................................................... 72
  4.1 The Rationale for Inferential Control ................................... 73
  4.2 Current Approaches to Model Building ............................... 75
  4.3 Determining the Inferential Model Via PLS ......................... 80
  4.4 The Reference Data Set ...................................................... 87
  4.5 Distillation Case Studies ..................................................... 88
  4.6 Results of Steady-State Case Studies ................................... 94
  4.7 Scaling .......................................................................... 103
  4.8 Results of Dynamic Case Studies ....................................... 108
  4.9 Conclusions ................................................................... 118
  4.10 Notation ...................................................................... 120
  4.11 References .................................................................. 122
# Table of Contents

Chapter 5. Dynamic Model Identification .................................................. 124  
5.1 Introduction ................................................................................. 124  
5.2 Model Structures ....................................................................... 128  
5.3 Identification Methods ............................................................... 132  
  5.3.1 Properties of Least Squares Estimates ........................................ 132  
  5.3.2 Constrained Least Squares Approach .......................................... 137  
5.4 Comparison of Identification Methods (MISO) ............................ 140  
5.5 MIMO: Simulation and Identification ........................................... 176  
5.6 Conclusions ................................................................................ 184  
5.7 Nomenclature ............................................................................. 186  
5.8 References .................................................................................. 188

Chapter 6. Summary and Conclusions ................................................. 190

Appendix .......................................................................................... 199
Chapter 1. Introduction

Process computers routinely collect hundreds to thousands of pieces of data from a multitude of plant sensors. This has caused a "data overload" and very little is currently being done to utilize this wealth of information. Further, modern processes are becoming increasingly complex and integrated. Unfortunately, the most common approaches to process analysis and monitoring, model building and control have been univariate, i.e. considering the problem in terms of independent sets of single input and single output relationships or best as multiple input and single output relationships. The simple volume of the data collected and the interrelated (correlated) nature of the measurements collected make this approach increasingly difficult to apply. The data is clearly multivariate in scope, thus efficient analysis and use of the data requires a multivariate approach. In this work the application of multivariate statistical techniques, especially Partial Least Squares (PLS), is investigated. The specific areas considered are process analysis and performance monitoring, inferential model building and dynamic model identification for multivariate control (eg. Dynamic Matrix Control (DMC)).

Whatever the application, the data collected has certain features which make it appropriate to apply these techniques. Given the current process control computer systems, on-stream analyzers, and automated quality control labs, it is not uncommon to measure hundreds of process variables on line every few seconds or minutes, and tens of product variables every few minutes or hours. Although a large number of variables may be measured, they are almost never independent; rather, they are
usually very highly correlated with one another. The true dimension of the space in which the process moves is almost always much lower than the number of measurements. In some situations this is due to underlying fundamental relationships among the variables. For example, in the hypothetical reaction of \( A + B \rightarrow C \) where \( A \) and \( B \) are fed to the reactor in a specified ratio, although the concentrations of \( A \), \( B \) and \( C \) are being measured (three dimensional measurement space), the actual problem is univariate (the stoichiometric relationship and the fixed feed ratio each eliminate a degree of freedom). In other situations the placement of the measurements and the nature of the process make the measurements highly correlated. Consider a distillation column where only three variables change independently, reflux, reboil and feed composition. Initially, measurements are being made of the temperature profile using every fourth tray temperature. If the number of measurements is increased to every tray temperature to obtain a more detailed temperature profile, the dimensionality of the measured variable space has been greatly increased, but, the actual dimensionality of the problem has not changed. As a second illustration consider another situation where measurements on many different variables are made, but the nature of the process and the disturbances are such that they only allow the variables to move in a much lower dimensional space. For example, in the manufacture of synthetic fibers it is not uncommon to measure more than ten quality variables such as denier, elongation under different loads, breaking strength, dye depths, etc.. The physical meaning of these measurements guarantees that the process is only capable of making fibers with certain combinations of properties, and disturbances to the process will affect many of these variables in a
highly correlated manner. For example, fibers with very small deniers (weight/unit length) cannot be made with very high breaking strengths, and disturbances which lead to a reduction in denier lead to a reduction in breaking strength.

Two additional factors make even the classical multivariate regression techniques such as Multiple Linear Regression (MLR) inappropriate. First, it cannot be assumed that the process (X) variables, the regresands, are error free. Second, extracting and presenting the information contained in data sets of the size and nature being collected is a daunting task. Particularly for process monitoring, it is important to be able to visually present the information contained in these reams of data, while for model building the goal is to obtain prediction equations relating process variables to the quality variables and provide accurate predictions. In this work it will be shown that PLS is an appropriate tool for both of these applications.

Typically the data collected can be summarized in terms of two matrices or data blocks. The first, which will be termed the X block, is generally associated with process variables (i.e., those measurements around the process which describe the operation of the equipment). The second, termed the Y block, can generally be classified as those measurements describing the product quality. The columns of these matrices represent the measured variables while the rows represent the observations made. In order to extract information from these data blocks Principal Component Analysis (PCA) and PLS define new variables, which are linear combinations of the original variables and summarize the major variation in the original data. These new variables are defined in such a way that they are independent and if the original variables were correlated allow the dimensionality of the problem to be drastically reduced, thereby accommodating further analysis. PCA is a single
block method which attempts to summarize the variance structure within one block of data. It can be applied to either the X or Y block and is useful for both process analysis and monitoring. PLS is a two-block technique which summarizes the variance structure within each block and develops a predictive relationship between the X and Y blocks. This is the more versatile technique and can be used in all the applications mentioned.

To summarize, the objectives of this work were to investigate the applicability of Partial Least Squares (PLS) when implementing procedures for performance monitoring, inferential control scheme development and dynamic model identification. For each of these applications PLS is integrated with presently used techniques, in order to improve the overall procedure. PLS is an interesting multivariate statistical method which provides many advantages for these applications; however, the method is relatively new and its statistical and computational properties are still being vigorously discussed in the literature. Bearing this in mind, Chapter 2 provides an introduction to the PLS method, draws comparisons between PLS and other well established statistical methods such as multiple linear regression (MLR), Principal Component Analysis (PCA), and Principal Component Regression (PCR) and presents some important properties of the method. The goal is to provide the reader with a physical understanding of the PLS algorithm and the properties which have a direct bearing on the implementation of the method. Each of the subsequent chapters (3-5) focuses on one application of PLS and contains: 1) a discussion of why multivariate procedures are appropriate; 2) a comparison of PLS and other multivariate techniques which can be applied; and finally 3) a discussion of the results and the advantages, difficulties and pitfalls of using PLS.
Chapter 2. Background, Theory and Application Issues

Partial Least Squares (PLS) is a powerful tool for multivariate analysis and model building but as with all tools it must be used correctly to be most effective. Chapters 3 to 5 discuss the application of the method to three areas of chemical engineering: performance monitoring, inferential model development, and dynamic model identification. In this chapter the PLS algorithm is presented, and an understanding of the advantages and limitations of the PLS method is developed through a discussion of the historical development of PLS, a comparison to related multivariate methods and a discussion of application issues particularly relevant to chemical engineering.

2.1 Introduction

PLS, based on the NIPALS (Non-linear Iterative Partial Least Squares) algorithm, was developed by H. Wold for model building based on abundant data but little mechanistic knowledge. H. Wold used the method primarily for soft modelling of socioeconomic trends (H. Wold, 1982), building empirical models using measured input and output data and little or no other information. The method was improved and popularized by the work of two research groups: by S. Wold and co-workers at the University of Umeå, Sweden and by B. Kowalski and co-workers at the University of Washington. A general history of the development of PLS along with some early applications can be found in a review by Geladi (1988) and more recent developments are discussed by Kowalski and Seasholtz (1991). PLS has found success and wide ranging applicability for many multivariate problems in
chemistry, especially in the quantitative analysis of ultraviolet, near-infrared and chromatographic data. This application of PLS has been such an improvement over present methods that many instrument manufacturers now include PLS software for the analysis of the data. PLS has become one of the cornerstones of Chemometrics; the study of chemistry through statistics and mathematics. In the late 1980's the growth in this area reached the point that two new journals devoted to the subject began publishing: Chemometrics and Intelligent Laboratory Systems and the Journal of Chemometrics; the first book on the subject appeared in 1989 (Martens and Naes, 1989). A general description of the algorithm can be found in a tutorial article by Geladi and Kowalski (1986), more detailed mathematical discussions are also available (Hoskuldsson, 1988; Lorber et al., 1987; Wold et al., 1984b and 1987). The relationship of PLS to other methods has also been discussed by Naes and Marten (1985), Helland (1988), Stone and Brooks (1990) and Phatak et al. (1991). The problems faced in chemical engineering have many of the same characteristics as those in chemistry and thus the properties of PLS which make it so attractive in chemometrics also account for the recent interest in the method in chemical engineering. Although relatively few applications of PLS in chemical engineering have been published (eg. Ricker, 1988; Kresta, et al., 1991), work in the area has been presented at many conferences since 1988.

The interest in this method arises out of a general increased awareness of the advantages of applying statistical approaches to a wider range of chemical engineering problems. PLS is of particular interest because it is suitable for process analysis and model building in situations containing correlated and noisy measurements. PLS is a new method which promises advantages over current methods
under these conditions but other well established statistical methods are also used throughout this work as a means of comparison and validation of results. Of particular interest are Principal Component Analysis (PCA) (Chatfield and Collins; 1986) and Multiple Linear Regression (MLR) (Draper and Smith, 1981); these are well understood and the conceptual ideas behind these methods are similar to those of PLS and will be used in future discussions to facilitate a better understanding of PLS.

PLS is applied in this work with two different final objectives. In the first application the objective is to take advantage of the dimension reduction ability of PLS to provide insight into the process and to use this information to build performance monitoring charts. When applied in this fashion PLS behaves similarly to PCA and a comparison of the two methods provides insight into the PLS algorithm. In the second and third applications the objective is to build empirical models using correlated input variables to provide predictions of future values of the output or quality variables. Again by introducing PLS in the context of other multivariate regression techniques, valuable insights into the method, its advantages and limitations are obtained.

The key to understanding the PLS method and its success lies in understanding the nature of the data for which it is appropriate. All of the advantages derived by using PLS depend on the ability to reduce the dimensionality of the problem by summarizing the correlation among the variables. Modern chemical processes contain many sensors which provide complementary information about the same underlying changes in the process. In many processes no one sensor can provide all the relevant information; however, all of the sensors available around
the process provide a plethora of data not easily analyzed. Further, if all the measurements are used as independent variables in model building, then the problem is over-parameterized. By summarizing the information from these measurements, PLS is able to overcome the difficulties faced by MLR in model building and provides concise representation of the information for process analysis. If the measurements are independent (all measuring different changes in the process) PLS is unable to reduce the dimension of the problem and reduces to a computationally inefficient form of MLR.

Further, in order to achieve a successful implementation the purpose of the application must be clearly defined at the start of the problem. One of the considerations which must be kept in mind is the fact that PLS is an empirical modelling technique and like all other empirical techniques it models correlation not causality. PLS is an empirical modelling method and therefore relies exclusively on the information content of the reference data. (i.e. Variation which is to be included in the model must be included in the reference data.) This means that PLS predictions can only be used over the range of operation spanned by the reference set and can not be used if the underlying system (process, instrumentation, control and so on) changes between the time when the reference set was collected and the time when the predictions are to be used. For example, even if the temperature is an independent variable in the process operation, if the temperature variation is not included in the data used to build the model then the temperature will not be included in the model. The PLS (or PCA or any empirical) model is only as good as the information contained in the reference set.
2.2 Notation, Conventions and Definitions

In this thesis PLS is applied in three distinct areas and each area has its own well established nomenclature. For this reason each chapter contains a nomenclature (and reference) section. A common PLS notation is used throughout the thesis and is developed in this chapter. The notation used follows that of Geladi and Kowalski (1986); unfortunately, the notation surrounding PLS has not been standardized, nor is it always compatible with notation used in chemical engineering. Thus the notation found in the literature often depends on the subject area and the background of the author.

Throughout this work the discussion will focus on two blocks of data: the first, \( X \), is an \((n \times k)\) matrix of the input variables (observations of the process, inputs to the system, etc.) and the second, \( Y \), is an \((m \times k)\) matrix of the output variables (these are often the quality measurements associated with the system). The measurements included in these matrices will vary depending on the application, in some situations a measurement could be considered either an input or output variable depending on the other measurements available. For example, in the fluidized bed reactor modelled in Chapter 3, the reactor bed temperature at the outlet is often used as an indicator of the 'quality' of the reaction. In the simulation presented the actual output concentrations were measured so the reactor bed temperature was considered an input variable describing the state of the process. The configuration of the individual \( X \) and \( Y \) blocks is at the discretion of the user, but the way in which they are treated remains the same. The columns of these matrices are the variables measured and the rows represent the samples collected. Other vector and matrix
quantities will be denoted by boldface letters; vectors by lowercase letters (a, b, ...) and matrices by uppercase letters (A, B, ...). Scalar quantities will be indicated by regular lettering (A, B, ..., α, β,...).

The common application of PLS is in situations where the variables within the X and Y blocks are interrelated, correlated or collinear. By these terms we mean that the variation measured by these variables is not independent. When the variation in a data set is collinear then this variation can be summarized by one variable (t) which can be expressed as a 'linear combination' of the original variables (ie. \( t = c_1x_1 + c_2x_2 + ... \) or \( t = X^T c \)). When the data set is not collinear (but is highly correlated), the percentage of the variation which can be explained by this linear combination is related to the extent of the correlation.

PLS is a stage-wise operation and in each stage PLS defines new variables which summarize the variation in the X and Y blocks. The new variable is a linear combination of the original variables. The result of the linear combination for each observation, the value of the new variable, is called the 'score' and the weighting of each of the original variables in the linear combination is called the 'loading' for the variable. Each X and Y block contains more than one variable (although Y blocks of only one variable are possible) and more than one observation; therefore, in each stage vectors of scores and loadings are calculated. In the PLS terminology the new variables and vectors are called the 'latent variables' or the 'latent vectors' (ie. the hidden or underlying variables or vectors which describe the variation).
2.3 Conceptual Formulation

The mechanics of the PLS method can be demonstrated using a geometric representation or from the algorithm and its relationship to the other regression methods. In this section the relationship between the data in the reference set and latent vectors calculated by PLS is illustrated geometrically using a simplified three dimensional example. The way in which principal components are chosen in Principal Component Analysis of a single data block is very similar to the way in which latent vectors are chosen by PLS in a multiblock problem. The principles behind the PLS method will be explained by showing how principal components are calculated in PCA and then the differences between PCA and PLS will be described.

The data for the simplified three dimensional example used in this discussion has been plotted in Figure 2.1. This data represents a single data block used in PCA. For PLS this plot would be representative of the input block data and a similar plot would be shown for the Y block data. Often the measured variables \( x_1, x_2 \) and \( x_3 \) are not independent of one another and/or the variation in the data is constrained to specific directions by the physical processes occurring. Thus, the number of measurements does not necessarily represent the true dimensionality of the problem. This can be explained geometrically, by stating that the variation in the data is constrained to a subspace of the original measurement space; or mathematically, by stating that the data matrix, \( X \), is nearly singular. Although three variables are measured in this example, the data represents a problem with only two independent changes. The variation in the data, due to the independent changes, is in only two dimensions and any variation in the third dimension represents measurement noise.
Figure 2.1 Plot of a typical data set for PCA or PLS. Although, the measurement space dimension is three; $x_1$, $x_2$, and $x_3$ represent correlate variables and the dimension of dominant variation is two.

The dimension of the measurement space can be adjusted to be the same as the underlying dimension of the problem by deleting one of the measurements. However, this is not the best approach because some of the information is lost by removing the additional measurement. A better approach is to use PCA to define new orthogonal variables which summarize this variation and reduce the dimensionality of the problem from 3 to 2 without losing any significant information. While this small reduction may not seem impressive, in the BTX case study presented in Chapter 4, PLS was used to reduce the dimensionality from 45 in the measurement space to 3 in the latent variable space.

In Figure 2.1, any two vectors in the plane could be used to define the plane of the major variation, thus additional constraints are needed to uniquely define these variables. Since we are interested in the variation in the data, it would be logical to choose the direction of greatest variance as one of the variables and to choose the second variable as the direction of the greatest remaining variance but orthogonal to the first variable in order to maintain independence of the new variables (see
Figure 2.2a). A third variable could also be chosen to describe the remaining variation. This procedure for identifying the principal directions of variation in a single data block is called Principal Component Analysis (Anderson, 1984a; Mardia et al., 1982) and the new variables are called principal components.

Often PCA is used to define a new orthogonal basis for the data; however, for the applications considered in this work the goal is to summarize the major variation using a limited number of new variables. In this example the major variation of interest falls in the plane of the first two principal components, and any subsequent analysis would use the information summarized by these two variables. Choosing the correct number of variables to describe the true dimensionality of the data is central to successful implementation and is described in detail later.

Each principal component has two quantities associated with it, which relate the data shown in the plane of the principal components to the original observations. The principal components can be expressed as a linear combination of the original variables. The result of the linear combination for each observation, the value of the new variable, is called the 'score' and the weighting of each of the original variables in the linear combination is called the 'loading' for the variable. Each X block contains more than one variable and more than one observation; therefore, for each principal component, a vector of scores and a vector loadings is calculated. The 'loading vectors' (pₖ) define the location of the plane in terms of the original variables, and the 'scores' (tᵢ) locate each observation (xᵢ, a row in X) on this plane. The score is defined as the distance from the origin of the plane (x̄) along each principal component and can be calculated as the product of the loading vector and observation, tᵢ=Xp. If the first two principal components form a model for the
Figure 2.2 Plot of a typical data set for PCA or PLS, showing the first two principal components explaining the majority of the variation (a). The plane of the principal components provides a window on the process; each observation is located on this plane via its score \((t_1, t_2)\) (b). Further, a loading plot (c) can be used to show the relationship between the original variables and the principal components.

variation in the data, then the perpendicular distance from each original observation to the plane is a measure of the residual variance for that observation. In this way the information contained in any data block can be described using two 2-D plots: (1) the loading plot, showing the relationship between the original variables and the principal components (Figure 2.2b) and (2) the score plot, showing the relationship between the observations and the principal components (Figure 2.2c). One of the
most common uses for PCA is to use the score plots to discriminate between different natural phenomena, e.g. from separating bacteria (Sorensen et al., 1985) or normal plant operation and abnormal operation (Jackson and Mudholkar, 1979).

When using PCA the object is to find the principal component space which explains the greatest amount of variability in a single matrix of data (X). Often, we can identify a second set of variables (Y) which are of greater importance - e.g. product quality or productivity variables, which we would like to include in the analysis. One way to do this is to perform PCA on the new block. However, by analyzing each block separately no information is obtained concerning the relationship between the variation in X and the variation in Y and this information is often vital to the analysis of the problem. By using PLS, which performs a simultaneous analysis of both the X and Y blocks, the information regarding the relationship between the two blocks is included in the resulting model. This additional ability is the key advantage of using the more complex PLS method over the well understood PCA method.

The result of the PLS calculation is similar to that obtained from PCA. The variation in X (and in Y) is summarized in new orthogonal variables called latent vectors. The same notation is used for the loading vectors and the scores for the X block (p_i and t_i, respectively), the additional notation for the loading vectors of the Y block is q_i and u_i for the score vectors. The difference in the two methods is that the latent vectors for X are not just the directions of the greatest variance in X, but a combination of the directions of greatest variance and the directions having the highest correlation with Y. This difference between PCA and PLS has been explained by saying that PLS can be seen as 'tilting' the plane of the principal components to
account for the relationship between $X$ and $Y$. The difference in the principal components and the latent vectors may not always be large but in situations where minor directions of variation in $X$ have a strong effect on the variation in $Y$ the difference is significant and can be crucial to the proper understanding of the process.

Another advantage of PLS over PCA, which is not obvious from the geometric representation, is that when using PLS a regression model is obtained which relates $X$ and $Y$. This model can be used to obtain predictions of future $Y$ values from future $X$ values or used to check if the relationship between $X$ and $Y$ has changed.

2.4 Mathematical Formulation

As in the previous section it is beneficial to present the mathematical formulation of PLS against the background of the more familiar methods (MLR, PCR, and PCA). This provides additional background for the method by showing the commonality and the differences in all these analyses. The mathematical and statistical properties of PLS have been discussed in detail in by a number of workers (S. Wold et al., 1984b and 1987; Lorber et al., 1987; Naes and Martens, 1985; Manne, 1987; Hoskuldsson, 1988; Helland, 1988; Phatak et al., 1991). Many of these papers were written during the development of PLS, and they report several variations of the basic algorithm using different notations. As has been mentioned, the notation used here is the same as used by Geladi and Kowalski (1986). These authors generally concentrate on PLS as a multivariate regression technique; however, part of the advantage of using the PLS algorithm is the ability to reduce the dimensionality of the problem. While this is an integral part of the regression, it also has value in itself
as a tool for obtaining process insight and in performance monitoring. This aspect of the algorithm will be discussed first, with analogies drawn from application of PCA to the same problem.

2.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (Chatfield and Collins, 1980, S. Wold et al., 1987 or Jackson, 1980) is used to explain the structure of the variance in a single data matrix. As discussed in the previous section, PCA calculates a vector, called the first principal component, which describes the direction of greatest variability (PC1 in Figure 2.2). It is calculated as the 'orthogonal regression' of a line through the data in the space spanned by X, i.e. minimizing the sum of the perpendicular distances from the points to the line (Anderson, 1984b), or as that linear combination of the columns of X (i.e. Xp1), given by the eigenvector (p1) of XTX associated with the largest eigenvalue. The second principal component is orthogonal to the first principal component and explains the greatest amount of the remaining variability. It is obtained by fitting a line through the residual variance of X after fitting the first principal component, or as the linear combination (Xp2) of the columns of X, given by the eigenvector (p2) of XTX associated with the next largest eigenvalue. The result of the PCA calculation can be expressed as the decomposition of the data matrix into the following bilinear form:

\[ X = \sum_{a=1}^{A} t_{a}p_{a}^{T} + E_{A} = TP^{T} + E \] (2.1)

The traditional definition of PCA in the statistical literature concerns the transformation of the X matrix to a new orthogonal coordinate system. In the notation of equation (2.1) the dimensions of T and P would equal the rank of X. In applications
similar to the ones in this work PCA has taken on a slightly different meaning; in this context PCA refers to an 'approximation of the matrix $X$ by a model with a relatively small number of columns in $T$ and $P$' (Wold et al. 1987). Since the object of the present PCA application is not to determine a new orthogonal basis for $X$ but rather to model the variation in a few principal components it is necessary to choose the dimension $A$. Ideally $A$ is chosen such that there is no significant process information left in $E$; rather $E$ should represent random error, and adding an $(A+1)$ principal component would only be fitting some of this random error, and thereby increasing the prediction error of the principal components model in Equation (2.1). There are several ways for selecting this optimal model dimension, $A$. One can proceed until the percent of the variation explained by adding additional principal components is small, although this often leads to overfitting. A better procedure is to use cross-validation (S. Wold, 1978).

The definition of PCA presented by Wold derives from the fact that for correlated data sets the dimension of the major variation in the data is much smaller than the dimension of $X$. This indicates that calculating the entire $T$ and $P$ matrices using any of the conventionally used computational methods would be extremely inefficient. An alternate approach is to calculate the principal components one at a time and stopping when the major variation in the data is explained. The NIPALS algorithm for calculating principal components (S. Wold et al., 1987) is an efficient way to perform this calculation. This algorithm, as used in the PLS calculation is presented in the next subsection.
Using PCA the \( Y \) matrix can be similarly decomposed into

\[
Y = \sum_{\lambda} u_{\lambda} q_{\lambda}^T + F_A - UQ^T + F
\]

(2.2)

where \( U \) and \( Q \) model the major variation in this data block.

### 2.4.2 PLS and the NIPALS Algorithm

The PLS method was developed based on the NIPALS algorithm for application on problems where the dimension of the measurement space exceeds the true dimension of the problem, as characterized by the dimension of the variation in the data. PLS can also be used in the same way as PCA, for data compression and process analysis (e.g. discrimination of events). It performs a dimension reduction and decomposition similar to PCA; however, the matrices \( T, P, U \) and \( Q \) are different from those obtained by performing PCA on \( X \) and \( Y \) separately. One of the most useful analogies for understanding PLS is to interpret PLS as performing PCA on the covariance of \( X \) and \( Y \) (i.e. \( Y^T X \)), thus the decomposition is not effected solely by the variance in each block but is also effected by the correlation between the blocks. As with PCA, the PLS vectors can be calculated using several computational methods (Hoskuldsson, 1988); most of these techniques calculate the entire \( T, P, U \) and \( Q \) matrices and are not considered further. The appropriate PLS application assumes that the measurement space exceeds the true dimension; therefore, only the first few vectors are of interest in this type of application, and methods which compute all the vectors would be extremely inefficient. By using the NIPALS algorithm to calculate the PLS vectors only the relevant information is used; conversely for the situations where all the variables are independent using PLS and the NIPALS algorithm is an inefficient approach.
The NIPALS algorithm performs the dimension reduction in a stagewise manner, one PLS dimension at a time. Before calculating a new PLS dimension tests are performed to determine whether or not a new dimension should be added. The calculation at each stage can be described by the following algorithm (S. Wold et al., 1987).

0. Mean center and scale X and Y
1. Set u equal to a column of Y
2. \( w^T = u^TX/u^Tu \) (regress columns of X on u)
3. Normalize w to unit length
4. \( t = Xw/w^Tw \) (calculate the scores)
5. \( q^T = t^TY/t^Tt \) (regress columns of Y on t)
6. Normalize q to unit length
7. \( u = Yq/q^Tq \) (calculate new u vector)
8. Check convergence on u: if YES to 9, if NO to 2
9. X loadings: \( p = X^Tt/t^Tt \)
10. Regression: \( b = u^Tt/t^Tt \)
11. Calculate residual matrices: \( E = X - tp^T \) and \( F = Y - btq^T \)
12. To calculate the next set of latent vectors replace X & Y by E and F and go to step 1.

The NIPALS algorithm is shown schematically in Figure 2.3, each step in the algorithm is represented by an arrow on the diagram. This clearly shows the integrated reduction of X and Y matrices at each PLS dimension.

Before embarking on a discussion of the relationships between the results of this algorithm and the results from other statistical methods it is useful to discuss the algorithm itself.

The algorithm presented here is for the PLS2 case, i.e. the application where the Y block can contain one or more variables. For the PLS1 application, only one Y variable, the algorithm can be modified to be more efficient both computationally and in the use of memory. PLS1 is useful in calibration problems and these modifications are discussed by Martens and Naes (1989). One common modification of
the PLS2 algorithm is not to normalize \( q \). When this modification is used step 10 is redundant; the length of the vector \( q \) is equal to the variable \( b \). Using this form of the algorithm removes the ability to handle missing data in the \( Y \) block.

One of the objectives of developing the NIPALS algorithm was to be able to handle missing data. This capability of the algorithm can be explained with reference to the \( X \) block in the following way. If any element in the column of \( x_j \) is missing, that element, \( x_{ij} \), and the corresponding element in \( u \) (i.e., \( u_i \)) are set to zero while the \( w_j \) vector is calculated. Furthermore, in calculating \( t \) in step 4 the \( i^{th} \) element in \( w \) (\( w_i \)) will set to zero in calculating the denominator \( w^T w \). Note that \( w^T w \) has been included in step 4 for this reason, even though it would normally be unity by step 3. Finally, when calculating the residual variance (\( E = X - tp^T \)) at each dimension the \( ij^{th} \) element is maintained to be zero. Using this formulation of the NIPALS algorithm effectively replaces any missing data points by their estimates \( \hat{x}_{ij} \) on the PLS plane. In this way they have no effect in determining the model. Missing data
in Y are handled similarly.

As with the number of principal components in the model for one block, the number of PLS dimensions (A) required to provide good predictions, and not overfit the data, is an important consideration. Since the object of this calculation is to find a model for the prediction of future values of the Y matrix, one way to check the correct dimension of the model is to hold back part of the data (test set) and test to find which dimension has the lowest prediction error. This is an inefficient use of the data and methods have been developed where all the available data is included in the reference set and this kind of calculation is still accomplished (Kowalski and Seasholtz, 1991). The method used in this work is cross-validation (S. Wold, 1978). In cross-validation, the reference set being used to build the model is divided into several subsections, 4 to 7 subsections are a reasonable compromise between accuracy and computational time. Using all but one of these sections, a new model with one additional PLS dimension is calculated. This model is used to predict the Y values in the remaining subsection and the prediction error sum of squares (PRESS) for this subsection is calculated. This procedure is repeated until each subsection has been deleted only once; then the PRESS values for each subsection are summed to obtain an overall PRESS. If this overall PRESS value indicates a benefit in adding this additional PLS dimension to the model, then the entire data set is used to recalculate the final values of the latent variables. The optimal order corresponds to a minimum in the overall PRESS. Equivalently, a normalized form of the overall PRESS can be used. The normalized form used in this work is the PRESS divided by the sum of squares of the Y block, before adding the last PLS dimension. This value has been labeled the CSV and can be calculated for each Y variable separately.
and for the overall model. The cutoff value when using this variable in the cross-validation is 1.0, indicating the point where the additional latent variables do not explain any additional variance. Further discussion of the problems associated with overfitting is presented in Chapter 4.

Once the model of the correct order has been developed from a reference data set where both \( X \) and \( Y \) are available, its main use will be for the prediction or inference of future \( Y \)'s given new vectors of process variable measurements. (For performance monitoring predictions of the quality measurements are still calculated and compared with the measured values to provide an indication of disturbances that change the relationship between \( X \) and \( Y \). Suppose a new vector \( x^T \) of process measurements becomes available, which is mean centered and scaled to be consistent with the reference set. Predicted values of the quality variables \( \hat{Y} \) can be obtained from either of the following forms. In terms of the PLS latent vectors one obtains

\[
\hat{Y} = TBQ^T
\]

(2.3)

Alternatively, the PLS model can be rewritten as a linear regression model in terms of the original \( X \) variables as

\[
\hat{Y} = X\hat{\alpha}_{PLS}
\]

(2.4)

where

\[
\hat{\alpha}_{PLS} = \sum_{i=1}^{A} b_i \left( \prod_{j=1}^{i-1} (I - w_j p_j^T) \right) w_i q_i^T
\]

(2.5)

The regression form has the advantage that the relative importance of each \( X \) variable in the regression is immediately evident, but the PLS formulation produces more
robustness to missing values in the X block used in future predictions. From the regression form the predictions are easily calculated as

$$\hat{y} = x^T \hat{a}_{PLS}$$  

(2.6)

In order to obtain the robustness of the PLS latent vector form the following calculations are necessary:

(i) calculate the scores for each PLS dimension \((a=1,2,\ldots,A)\)

$$t_a = e_{a-1}^T w_a / w_a^T w_a$$  

(2.7)

$$e_a^T = e_{a-1}^T - t_a p_a^T$$ \((e_0 = x)\)  

(2.8)

(ii) calculate the new predictions as

$$\hat{y} = \sum_{a=1}^{A} t_a b_a q_a^T$$  

(2.9)

The advantage of using the PLS latent vector formulation for prediction is that it accounts for missing data. While this is important in all the applications, it can be critical for the inferential models discussed in Chapter 4.

2.4.3 Comparison of PLS, MLR and PCR

Before continuing the discussion of the PLS as a multivariate regression technique it is worthwhile to examine the more familiar techniques of MLR and PCR. It has been shown that the three methods MLR, PCR and PLS represent a continuum of regression methods with MLR and PCR the extremes (Stone and Brooks, 1990; Lorber et al., 1987). While analogies between these methods have appeared previously (eg. Geladi and Kowalski, 1986) it is beneficial to go through the arguments again as they help to visualize the PLS objective function and indicate how variable scaling affects the method. In the following discussion each of the
three methods will be formulated in several different ways in order to facilitate comparison. The development presented here follows that presented by Phatak et al. (1991) with additional explanations as appropriate to this work.

Before presenting the comparison of these methods it is important to emphasize that MLR and PCR use only one Y variable in the analysis, while the general PLS application can consider numerous Y variables. In order to facilitate comparisons the discussion regarding PLS will be restricted to PLS1, containing only one Y variable. This approach is often used when discussing the properties of PLS because analytical solutions are available; unfortunately some authors do not make it explicit that the results are only applicable to PLS1. Once the initial comparisons are presented, extensions to PLS2 will be discussed.

The least squares estimate obtained using MLR is given by

$$\hat{\alpha}_{MLR} = (X'X)^{-1}X'y$$  \hspace{1cm} (2.10)

and the new predictions are

$$\hat{y} = X\hat{\alpha}$$  \hspace{1cm} (2.11)

or

$$\hat{y} = X(X'X)^{-1}X'y$$  \hspace{1cm} (2.12)

This is the standard result for MLR, but the problem can also be expressed as finding the linear combination, $Xc$, whose squared coefficient of correlation with $y$ is a maximum. Thus we maximize

$$R^2 = \frac{c'X'y'yXc}{(c'X'Xc)(y'y)}$$  \hspace{1cm} (2.13)
subject to the constraint $c'c = 1$. Thus, MLR can also be formulated as the eigenproblem

$$(X'X)^{-1}X'y'y'Xc = \theta c$$  (2.14)

where $\theta = (c'X'y'y'X)/(c'X'Xc)$ (Phatak et al., 1991).

Two difficulties arise in the application of MLR to the problems studied in this work. The first concerns the correlated variables in $X$. If the correlation is very severe then the matrix $X^TX$ is nearly singular and $(X^TX)^{-1}$ is ill-conditioned; even if $X^TX$ is not singular the problem is over-parameterized leading to very imprecise parameter estimates and poor predictions (Draper and Smith, 1981). Therefore, in many applications, subsets of the process variables ($X$) are chosen in order to eliminate problems due to dimensionality and correlation (this is a commonly used technique when building inferential models); however, this necessarily discards information contained in the eliminated measurements. An alternate method which can be used to stabilize the regression coefficients and thus provide more accurate predictions is Ridge Regression (Hoerl and Kennard, 1970; Smith and Campbell, 1980; Draper and Smith, 1981); unfortunately this method does not reduce the dimension of the problem and becomes cumbersome for high dimensionality problems. (A variation on this method is used in a comparison with PLS for dynamic model identification in Chapter 5.) The second difficulty in applying this technique is the underlying assumption that the measurements in the $X$ block are error free. In practice, where process variables are measured, this is often a very poor assumption. Reilly and Patino-Leal (1981) have shown how errors in the $X$ measurements affect the results and must be included in the analysis. Both of these difficulties, correlation and error in $X$, can be partially addressed if PCA is performed
on $X$ and then MLR is performed using the principal components of $X$ predict $y$. This method is called Principal Component Regression (PCR) (Jolliffe, 1986). Unfortunately, the space defined by the principal components of $X$ is only that space exhibiting the greatest variation in the $x$'s and in not necessarily the space that is most predictive of $Y$ (S. Wold et al., 1984a; Geladi and Kowalski, 1986; Lorber et al., 1987). Furthermore, PCR treats each $y$ variable individually, where in fact the $Y$ space often consists of many highly correlated variables. This can lead to inconclusive results if the $y$ variables have little or no information individually but are highly informative as a group. Although PLS is ideal for this case, another approach is to use the principle components of $Y$ and the principle components of $X$ in the regression.

When applying PCR to the regression problem the $X$ matrix is replaced in the calculation by the matrix of principal components (In order to differentiate principal components from latent vectors of PLS we will use the notation of Phatak: principal components $e_i = X_y i, i = 1, 2, 3, ...$). The weights $y_i$ are the solution to the eigenproblem

$$X'X y_i = \lambda y_i$$

(2.15)

Once the matrix of non-zero principal components, $\Xi$, is determined then it can replace $X$ in the standard MLR formulation.

$$\hat{\alpha}_{PCR} = (\Xi'\Xi)^{-1}\Xi'y$$

(2.16)
PCR addresses the difficulties caused by correlation and measurement error in \(X\) by replacing the inversion of the poorly conditioned \(X'X\) by the well conditioned \(\Sigma'\Sigma\).

The difficulties associated with inverting a poorly conditioned \(X'TX\) matrix and noisy \(X\) variables are overcome by PLS in a manner similar to PCR; by defining a new set of variables which are well conditioned (mutually orthogonal) and using these in the regression. The advantage of PLS over PCR is in the definition of the new variables, which take into account not only variables in \(X\) but also correlation in \(Y\). This means that the variation in \(Y\) can be explained in fewer latent variables than principal components (Phatak et al., 1991).

The PLS1 problem can be formulated as finding the linear combination \(Xw\) whose covariance with \(y\) is a maximum, again subject to the constraint \(w'w = 1\). For the first dimension, the PLS1 objective function can be written as the minimization of the Lagragian equation

\[
w'X'y y'Xw - \mu(w'w - 1)
\]

(2.17)
equivalently the problem can be posed as an eigenproblem

\[
X'y y'Xw = \mu w
\]

(2.18)

For PLS1 the NIPALS algorithm is non-iterative and the solution for the first dimension can be determined explicitly as

\[
w_1 = \frac{X'y}{\|X'y\|}
\]

(2.19)

Thus the first PLS latent variable, \(t_1\), is defined as

\[
t_1 = Xw_1 = \frac{XX'y}{\|X'y\|}
\]

(2.20)
In the stepwise PLS algorithm the residual variance of \(X\), \(E\), is now calculated as
\[
E_1 = X - t_1p_1^T
\] (2.21)
and the procedure is repeated to obtain the second latent variable \(t_2\)
\[
t_2 = E_1p_2 = \frac{E_1^TE_1'y}{\|E_1'y\|}
\] (2.22)

Phatak et al. (1991) demonstrate the relationship between PCR and PLS very nicely by showing that the latent variable \(t_i\) is simply a linear combination of the principal components used in PCR. (The principal components and the scores are both linear combinations of the original variables; therefore, the scores can be expressed as a linear combination of the principal components). The relationship between weights in this linear combination and the parameters which define the variance and covariance between \(X\) and \(Y\) is presented here and gives an interesting insight into how scaling affects the PLS result. This relationship can be shown by rewriting equation 2.20 as
\[
\|X'y\|t_1 = XX'y = \Xi\Xi'y
\] (2.23)
and substituting \(\Xi\Xi' = XX'\), where \(\Xi\) is the matrix of principal components. Now
\[
\|X'y\|t_1 = \sum_{i=1}^{p} \varepsilon_i \varepsilon_i'y = \Xi\Xi'y
\] (2.24)
Notice that \(\varepsilon_i'y\) is the covariance between \(y\) and the \(i^{th}\) principal component and can be rewritten as
\[
\varepsilon_i'y = \|y\| r_i \lambda_i^{1/2}
\] (2.25)
where \(r_i\) is the coefficient of correlation between the \(i^{th}\) principal component and \(y\); \(\lambda_i\) is the variance of the corresponding principal component. Substituting 2.25 into
\[
\begin{align*}
\frac{\|X'y\|}{\|y\|} t_1 &= \sum_{i=1}^{c} g_i e_i \\
\text{where} \\
g_i &= r_i \lambda_i^{1/2}
\end{align*}
\]  
(2.26)

Thus, $t_1$ is a weighted sum of the principal components of $X$ where the weights are proportional to the product of the correlation coefficient between $y$ and the principal component and the standard deviation of the principal component.

This is only one of the many analogies that has been shown between PLS1 and other methods, alternately it has been shown by Manne (1987) to be related to the conjugate gradient method used in optimization, although it is not the conjugate gradient method in the standard form.

Showing comparisons between the general PLS2 algorithm and the other methods has proven difficult because the method becomes iterative and the algebra daunting. However, some general results concerning PLS2 have been shown, of particular interest are analogies stated by Hoskuldsson (1988). Using the power method for obtaining eigenvalues it can be shown that any PLS dimension is simply solving the eigenproblem

\[
E_{i-1} F_{i-1} F_{i-1}^T E_{i-1} w_i = \mu_i w_i \\
\text{where} \\
E_s = X, F_s = Y
\]  
(2.27)

The analogy to the Power Method for calculating eigenvalues clearly explains the references to the rate of convergence of the NIPALS algorithm made in the literature. Generally the rate of convergence depends on the difference in size of $\mu_i$ and $\mu_{i+1}$. Once $w$ is known all other PLS variables can be explicitly determined. Alternately, eigenvalue problems can be set to calculate each PLS variable separately. Hoskuldsson also showed some interesting relationships between PCA and PLS. PCA works on the $X$ matrix, and the loading vectors calculated using NIPALS are the
eigenvectors of the variance matrix \((X^T X)\). PLS on the other hand, works on both \(X\) and \(Y\), the loading vectors calculated in this case are the eigenvectors of the matrix \(X^T Y Y^T X\). This last matrix can be thought of in two ways: (1) the variance of \(X\) has been scaled using the "size" of the \(Y\) matrix \((YY^T)\), or (2) PLS is simply PCA performed on the covariance matrix of \(X\) and \(Y\) \((Y^TX)\). Hoskulsson also indicates different eigenvalue formulations which determine \(w\) as the solution to alternate objective functions to the one currently used by PLS. These may be more appropriate for specific applications; unfortunately it is not clear how these would be translated to an efficient computational algorithm like NIPALS.

Although these alternate formulations of PLS provide insight into the method, they are not efficient computationally. If only a few latent vectors are being calculated then it is more efficient to use the NIPALS algorithm; however, if all dimensions are exhausted, then all \(X\) variables were nearly independent and the PLS solution degenerates to the MLR solution.

PLS is also similar to canonical correlation (Chatfield and Collins, 1986); however, canonical correlation is restricted in the number of latent variables which can be computed by the number of output variables. For example if only one \(Y\) variable is being considered then only one latent variable can be calculated. In the same situation PLS is only limited by the information content of \(X\). Due to this restriction canonical correlation is not used in this work.
2.5 Application questions as related to the problems studied

There are several application questions which must be addressed when using PLS. The ones particularly important to chemical engineering applications are discussed in this section.

One of the most important concepts in the application of PLS is that the method can only use the information contained in the reference set. Unlike some other methods (e.g., stepwise regression) it is not easy to add heuristic rules at intermediate steps in the PLS model building procedure. Once the X and Y blocks are defined and the algorithm is started the user can only decide when the appropriate number of latent vectors have been calculated. This could be viewed as a drawback. However, since the formulation of the X and Y blocks is very flexible and it forces the user to clearly define the problem before starting the calculations, we see this as an asset. Further, this allows the procedures which use PLS to be automated as long as the problem is well defined and specified at the outset. Throughout the rest of this work the central importance of the reference set in obtaining a successful implementation is emphasized.

The PLS method, as used in this work, is a linear method which can only use the information in the X and Y blocks; however, non-linear problems can still be adequately addressed. This can be accomplished in one of three ways:

i) If the non-linearity is mild to moderate and the signal to noise ratio allows sufficient latent vectors to be calculated, then this aspect of the data structure will be captured by higher order latent vectors. (i.e. The first latent vector will provide the best straight line estimate of the variation and the residual will be
the non-linearity in the data; additional PLS vectors will then be calculated to
to model the non-linearity. The composite of all the latent vectors will account
for the non-linearity.

ii) Since over-parameterization is not a concern when using PLS, additional
variables can be added to the $X$ and $Y$ blocks which will account for the
non-linearity. These new variables can be product or cross terms or other
transformations of the original variables. The PLS calculation will give high
weighting to the variables which account for the variation in the data the best.
The model can retain all the terms or the terms with low weighting can be
removed and the model parameters re-estimated.

iii) Finally, it is possible to reformulate PLS as a non linear method (S. Wold
et al., 1989) in which the inner relationship $u_i = f(t_i, b)$ is modelled as a
polynomial or spline function.

Variable scaling also has a significant affect on the results obtained when
using PLS (or PCA). Comparison of equation 2.13 and the expression which is
maximized in the PLS regression $X'YY'X$ clearly shows why MLR is scale inde-
pendent and PLS is dependent on the scaling of both $X$ and $Y$. Alternately this can
be explained by remembering that the objective of PLS is to summarize the
information in large data sets by focussing on the combination of the direction of
the greatest variance in $X$ and the highest correlation with $Y$. The directions of
greatest variability can be the result of the real variation in the process or they can
be caused by the scaling used. An example of this might be a process where the
temperature variation in °C ($\sigma^2 = 1.0$) and the concentration in ppm ($\sigma^2 = 100$) are
included in the same $X$ matrix. Further suppose that the temperature variance is
more important that the concentration variance. Unfortunately the major direction in the unscaled data is the concentration not the temperature. The directions chosen by PLS are a result of the units in which the variables are measured, rather than actual properties of the system under consideration, then the results obtained will not show the desired properties of the system. By examining equation

$$\frac{\|X'y\|}{\|y\|} t_i = \sum_{i=1}^{r} g_i e_i \quad \text{where} \quad g_i = r_i \lambda_i^{1/2} \quad (2.26)$$

the effects of scaling $X$, $Y$ or both can be examined. Using PLS the latent variable ($t_i$) will be determined as a weighted sum of the principal components of $X$. The weights are dependent on the correlation coefficient between the principal component and $y$ and on the standard deviation of the principal component and while $r_i$ is restricted to values between $\pm 1$, $\lambda_i$ can become very large. This explains the observation that it is sometimes necessary to calculate several PLS dimensions before obtaining one which explains significant quantities of the variance of $y$. It can be said that the first latent variables are swamped by the variance of $X$ which is not predictive of $Y$ and must be removed before discriminating the latent variables which are predictive of $y$. Note that this does not effect the prediction of $y$ since the latent variables in the first PLS dimensions will have little impact on the prediction because the corresponding $b$ is small. Phatak produced supporting simulations which show the effects of $r$ and $\lambda$ on the percent of $y$ explained.

Equation 2.26 applies only for PLS1 but by performing a simple thought experiment the effect of scaling individual $y$ variables can also be demonstrated. Consider a situation containing several $y$'s but the variation in only one dimension. Then, as an approximation, this problem can be solved by PLS1; where the $Y$ is
replaced by the principal component which summarizes the variation. Different scaling within the Y block will effect this principal component and the correlation coefficient between the principal components of X and the variation in Y. This same result has been observed by the author when different scaling was used in the distillation column simulations. The largest effects on the results were produced by scaling the X block, smaller effects were seen as a result of the scaling used with individual Y variables. Scaling the entire X block versus the entire Y block has no effect on the results.

Several scaling approaches have been proposed in an attempt to remove this difficulty in the application of the method. The most common approach is to scale all variables to unit variance; however, this scaling disregards any physical information about the process. Alternately, it has been suggested (Johnston & Barton, 1987; Bonvin & Mellichamp, 1987) that scaling based on the physical understanding of the process should reflect the individual importance of each variable to the process. Both of these extremes have their drawbacks. Unit scaling of noisy variables and nearly constant variables in the reference set may unduly raise their importance in the model. The suggested scaling methods based on physical knowledge require such an extensive understanding of the process that they are impractical. Martens and Naes (1989) suggest a very sensible alternate scaling approach. They suggest that rather than scaling to unit variance \( W_E = 1/S_D \) one should scale by the reciprocal of the sum of the standard deviation of the variable \( S_D \) and an estimate of the standard deviation of the noise \( S_E \) (ie. \( W_E = 1/(S_D + S_E) \)). In this way noisy variables are discounted in the scaling procedure. The estimate of the standard deviation of the noise can be obtained from previous information about the system or as Martens
and Naes suggest by performing a PLS calculation on the data and obtaining an estimate of the noise from the unexplained variation in each variable. In this work no one method is recommended; rather, each of these methods were used depending on which was appropriate.

The PLS algorithm was developed as a data compression and multivariate modelling method for use with steady-state data. Using the method in this mode provides valuable insight into many chemical engineering problems; however, in many other applications the dynamic characteristics of the process are as important as the state-state information. By incorporating the ideas of time series and expanding the X matrix to include past values of the measurements, the PLS method can be adapted to model dynamic characteristics of the process. The application of this form of the PLS model building is investigated for the methanol-acetone-water distillation column in Chapter 4 and for dynamic model identification in Chapter 5.

2.6 Summary

The preceding discussion was not meant as an extensive review of the theory surrounding PLS but rather as a summary, highlighting those aspects which are particularly useful in obtaining a general understanding of the procedure and preparing the groundwork for further discussions related to the application of this method.

Much of the literature concerning the mathematical and statistical properties of PLS, became available during the course of this work. At the conception of the project the two most pressing application questions were to find a concise expression for the objective function used by PLS and an expression for the effects of scaling.
As these two issues are closely related they were examined together through numerical experiments. Different scaling approaches were investigated and it was discovered that no one scaling approach was ideal in all cases. While equation 2.26 presents a concise explanation for the effects of scaling, the only expression for the objective function of PLS2 is that it maximizes the covariance between X and Y at each PLS dimension. This expression for the objective function is not as easy to understand as the objective functions for other regression methods but it does provide some understanding of what results to expect for the method.

Some other important issues concerning the application of PLS to chemical engineering problems which were addressed in this chapter or will be addressed in the rest of the work include: 1) The understanding that the ability of PLS to reduce the dimension of the data is an important property of the method but also that this dimension reduction can only be accomplished if the variables in the data set are correlated. Further, this dimensional reduction facilitates not only the analysis but also the presentation of the results (SPC charts in Chapter 3) and prevents overfitting when used in model building (Chapter 4). 2) The understanding that PLS is an empirical modeling technique and that the quality of the predictions is dependent on the reference set used to build the model. The importance of this is shown and expanded upon in Chapter 4. 3) The understanding that the distribution of the measured variables into the two data blocks, X and Y, is dependent on the measurements available and the purpose of the application, and is at the discretion of the user. However, the way in which these blocks are treated by PLS remains the same. 4) The understanding that PLS is a linear method which can handle limited
non linearities through variable transformations, higher order latent variables or through the modification of the inner relationship; and finally 5) the difference between PLS1 and PLS2.

2.7 Notation

A - the optimal number of latent vectors

$b_a$ - the regression coefficient for the $a^{th}$ latent vectors of $X$ and $Y$ measurements

$B$ - the diagonal matrix of regression coefficients

CSV - ratio of the PRESS for the $i^{th}$ PLS dimension and the sum of squares of the $Y$ variables in the $i-1^{th}$ PLS dimension

$E$ - the residual matrix of $X$

$F$ - the residual matrix of $Y$

$I$ - the identity matrix

$k$ - the number of process/operational/input measurements

$m$ - the number of quality/output measurements

$n$ - the number of observations

$p_a$ - the loading vector for the $a^{th}$ latent vector of $X$ calculated by PCA or PLS

$P$ - the loading matrix for $X$ calculated by PCA or PLS

$p_i$ - the loading for the $i^{th}$ $X$ variable

$q_a$ - loading vector for $a^{th}$ latent vector of $Y$

$Q$ - loading matrix for $Y$

$r_i$ - coefficient of correlation
$S_p$ - estimate of the standard deviation of the measurement

$S_E$ - estimate of the standard deviation of the noise

$t_a$ - scores associated with the $a^{th}$ latent vector of $X$

$T$ - the matrix of scores associated with $X$

$u_a$ - scores associated with the $a^{th}$ latent vector of $Y$

$U$ - the matrix of scores associated with $Y$

$w_a$ - the loading vector used for prediction, calculated from PLS, for the $a^{th}$ latent vector

$W$ - the matrix of prediction loadings for $X$

$w_i$ - the prediction loading for the $i^{th}$ $X$ variable

$W_E$ - weights used in scaling

$X$ - $(n \times k)$ matrix of the process / operational / input measurements

$\bar{x} & \bar{X}$ - the mean of a variable of $X$ and the vector of means

$x_i$ - the value of the $i^{th}$ $X$ variable for one measurement

$\hat{x}_i$ - the predicted value of the $i^{th}$ $X$ variable for one measurement

$X_B$ - the composition of the bottoms product from the EDC (Table 3)

$X_D$ - the composition of the distillate product from the EDC (Table 3)

$Y$ - $(n \times m)$ matrix of the process performance / product quality / output measurements

$y_i$ - the value of the $i^{th}$ $Y$ variable for one measurement

$\hat{y}_i$ - the predicted value of the $i^{th}$ $Y$ variable for one measurement

$\hat{\alpha}$ - the regression coefficients

$e_i$ - the $i^{th}$ principal component of the $X$ matrix
\( \Xi \) - the matrix of the principal components of \( X \)

Superscript

\( T \) - the transpose of the matrix

2.8 References


Chapter 3. Application of PCA and PLS to Process Monitoring

Chapter 2 discussed the abilities of Principal Component Analysis (PCA) and Partial Least Squares (PLS) to summarize the information content in high dimensional correlated data sets. One area where this type of technology is needed is for monitoring the performance of highly integrated multivariate processes. A very interesting example of the application of PCA to the monitoring of a desulpherization process is presented by Denney et al. (1985). That example provided some of the motivation and a starting point for the work presented in this chapter.

3.1 Introduction

Monitoring performance and detecting faults is an integral part of the successful operation of any process. Performance can be monitored by comparing the actual results to the predictions from a mechanistic model, or by using statistical process control (SPC) charts (e.g. Shewart charts (Shewart, 1931), CUSUM charts (Woodward and Goldsmith, 1964) or EWMA charts (Hunter, 1986)) to compare the current state of the process against "normal operating conditions". The challenge currently facing process monitoring is the enormous amount of correlated data being collected from a multitude of sensors every few seconds, minutes or hours. This "data overload" and the lack of appropriate analytical tools has meant that very little is being done to utilize this wealth of information.

A mechanistic model could be used to monitor the performance of any process; however, the biggest drawback to using a mechanistic model is the need for a detailed model. Even if a detailed mechanistic model is available its parameters
are uncertain, and often they need to be updated in real-time. Using the established statistical control charting methods has the advantage that they need no model, but rather use the operational data directly; they are also easily applied and interpreted. The major drawback to these control charts is that they were developed for monitoring univariate problems or sets of independent variables, and the expansion to handle the case of many highly correlated variables is difficult (Woodhall and Mcube, 1985). These methods are still being used on a one variable at a time basis for multivariate processes, either formally using Shewart charts or informally by operators monitoring key variables. Generally, this approach has been adequate, although extremely inefficient; however, if the variables are correlated this approach can lead to erroneous results (Jackson, 1980).

This chapter presents a different and more efficient approach to process monitoring based on multivariable statistical methods of Principal Component Analysis and Partial Least Squares. These methods are particularly suited to analyzing large sets of correlated data. While, the analysis makes much more efficient use of the data and the dimensionality of the problem is greatly reduced, the results can still be interpreted and used in the same way as originally outlined by Shewart (1931).

3.2 Monitoring via multivariate SPC plots

Statistical process control charts such as the Shewart chart, CUSUM plot and EWMA chart are well established statistical procedures for monitoring stable univariate processes. The assumption behind them is that a process subject only to its natural ("common cause") variability will remain in a state of statistical control
unless a special event occurs. The control charts represent several statistical hypothesis testing procedures aimed at detecting the occurrence of a special event as quickly as possible. Upon detecting such an event the action called for would be to diagnose the problem, find an assignable cause for the deviation, and then correct the process by removing this assignable cause. In practice, such an approach of continually monitoring the process, detecting events, and removing their causes leads to long term improvements in the process.

A Shewhart chart consists of plotting the observations sequentially on a graph which also contains the target value and upper and lower control limits as shown in Figure 3.1. If an observation exceeds the control limits a statistically significant deviation from normal operation is deemed to have occurred, which triggers the search for an assignable cause. The control limits are usually determined by analyzing the variability in a reference set of process data collected when only normal or "common cause" variability is present and acceptable operation is achieved. The limits are then usually set at plus and minus three standard deviations about the target.

The philosophy applied here in developing multivariate SPC procedures is the same as that used for the univariate systems. As with the Shewhart Charts, an appropriate reference set is chosen which defines the "normal operating conditions" for a particular process. The choice of this reference set is critical to the successful application of the procedure and is discussed in more detail when describing the case studies.

The major difference in multivariate SPC is that we are usually facing the following challenges:
Figure 3.1 A typical Shewart Chart: * - indicates in control; o - indicates out of control.

1. The method must be able to deal with correlated data of high dimension, in both the independent (X) and the dependent (Y) variables.

2. The method must reduce the dimension of the problem substantially and allow simple graphical interpretations of the results.

3. If both process (X) and product quality (Y) variables are present, it must be able to provide good predictions of Y.

The first challenge is addressed through the choice of the methods used; both PCA and PLS are specifically designed to deal with highly correlated data, as discussed in Chapter 2. The second challenge can not be met through the appropriate choice of the analysis method alone. The ability to meet the second challenge rests both on the properties of the process and on the analysis method. The solution relies on the underlying premise that many of the measurements are highly correlated and simply represent different manifestations of the same disturbances or other events occurring in the process. It is also assumed that the underlying dimensionality of the process, when it is operating normally, is quite low, and in these circumstances
we can represent the most important elements of the behaviour on low dimensional plots defined by the dominant latent vectors obtained via PCA or PLS. In effect the low dimensional planes defined by these latent vectors provide a low dimensional window on the behaviour of the very high dimensional process. The third challenge, the ability to provide good predictions of Y, depends on many factors; however, by providing accurate predictions the ability of the monitoring procedure to distinguish abnormal behaviour is enhanced.

Applying PCA or PLS to the monitoring problem, a chart of the following form can be developed. For a plant with an underlying dimension of two, the axes of the score plot (T_1 and T_2) form the first two axes of the monitoring chart and each observation is located on this plot via its score. Using the regression equation developed for the two latent vectors (using PLS), the predicted y's are calculated for each new observation. These values are then used to calculate the Squared Prediction Error (SPE) \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) for each observation and this variable could form the third dimension of the monitoring chart (Figure 3.2a), or could be plotted against time as in standard Range charts (Figure 3.2c). If, as is often the case, the Y variables are measured only infrequently, and hence are not usually available at each monitoring interval, then one should plot the Squared Prediction Error for X, i.e. \( \sum_{i=1}^{k} (x_i - \bar{x}_i)^2 \). This latter value represents the squared distance of the new x vector from the prediction plane (T_1,T_2). Note that in the latter case one might want to weight each x_i element by its modelling power for y, i.e. \( \sum_{i=1}^{k} w_i (x_i - \bar{x}_i)^2 \).

It is worthwhile noting that the score plot (T_1,T_2) and its control limit contour is very similar to the Shewhart chart for the mean of the variable, (\( \bar{x} \)), while the SPE plot is similar to the range or standard deviation plot for the same variable. The
Figure 3.2 The multivariate monitoring chart. The axes are defined by T1, T2 and SPE: a) 3-dimensional representation, b) 2-dimensional representation, and c) SPE plotted vs time.

Observations included in the reference set are plotted on this chart; the limits of acceptable performance are set such that they encompass a certain percentage of the variation in the reference set (analogous to the control limits of the Shewart chart). It has been shown for PCA (Jackson, 1980; Jackson and Mudholm, 1979), that under normality assumptions and specific scaling, the control limits on the T1-T2 plane form an ellipse and the proper size of the control ellipse can be calculated; further, the Q-statistic can be used to establish the control limits for the residual or
SPE. Alternately, rather than basing the control limits on any hypothesized underlying distribution; they can be based on the reference distribution of the historical data contained in the reference set (Box et al., 1978). This is the method used in this work. The "normal operating conditions" should reflect well behaved plant conditions and will need to be reset if plant conditions change.

Once the chart has been set up it can be used to monitor the process. Using latent variable relationships developed for the reference set, the scores for each new observation vector can be located on the plane \((t_1 = xw_1, t_2 = (x - t_1p^T_1)w_2)\) and its SPE calculated \((\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)\) or \(\sum_{i=1}^{n}w_i(x_i - \hat{x}_i)^2\). The structure of these plots reflects the two ways in which abnormalities can enter the system and provides powerful diagnostic capability to determine the cause of the abnormality. If the abnormality is caused by a larger than normal change in one or more of the process variables but the basic relationship between the process and quality variables does not change, then the abnormality will manifest itself as a shift in the \(T_1-T_2\) plane, and the SPE will remain at an acceptable level below its action limits (the * symbols in Figure 3.2a and 3.2b). If on the other hand the abnormality enters through a new event not captured in the reference set, it will change the nature and possibly the dimension of the relationship between the process and quality variables; then, the SPE will increase (o in Figure 3.2a and 3.2b). Acceptable performance, i.e. performance which matches the reference set, would fall within the envelope of normal operation (Δ in Figure 3.2a and 3.2b).

One difference between the standard univariate SPC charts and the multivariate SPC chart is that there is no convenient time axis to measure the age of the observations. The passage of time can be indicated in the \(T_1-T_2\) plane by the markers.
used. This can be done in several ways including: (1) by marking the passage of
time using shading intensity - the most recent being the brightest and past values
fading so that only the past \( p \) observations are visible; or (2) by using time indexed
markers (such as \( s1, s2, s3, \ldots \)), but still showing only the past \( p \) observations on the
current display. Alternately, separate univariate charts for \( T_1 \) and \( T_2 \) could be plotted
against time in a manner similar to a Shewart chart. This approach is better than
plotting the original variables individually, since there are many fewer \( T' \)'s and they
are orthogonal; however, there is still a danger that an observation will lie within
the individual limits but be clearly abnormal when viewed on a bivariate chart (see
Figure 3.5, Case 1). The SPE would usually be plotted against time similar to the
standard Range or S-chart (Figure 3.2c).

The procedure described above and demonstrated in the simulations is for
the situation where both \( X \) and \( Y \) measurements are available at regular intervals;
or where the \( Y \) measurements are available at least in the reference set defining the
NOC used in the PLS analysis. If, however, the quality measurements (\( Y \)) were too
infrequent or totally unavailable, then the procedure could be reformulated. Two
possibilities exits: (1) if certain variables in \( X \) could be considered good inferential
variables for \( Y \), then \( X \) could be partitioned into \( X_1 \) and \( X_2 \), where \( X_2 \) would contain
the inferential variables and act as the \( Y \) in the previous discussion; or (2) if no
inferential variables were available, then the procedure would be based on PCA
instead of PLS and the principal component model, Equation (1), for \( X \) would be
used to form an SPE (\( \sum_{i=1}^{t} p_i(x_i - \bar{x}_i)^2 \)) in order to detect changes which disturb the
relationships among the \( X \) variables. Both of these methods have limitations. The
method using PCA is less sensitive because it may include variation which is not
directly related to quality variables. This method is also unable to detect disturbances which affect the relationship between the process and quality variables but are not included in X. The method using inferential variables is more sensitive to the specific disturbances which affect the particular inferential variables used. It can detect changes in the relationship between X and the inferential variables but may miss disturbances which do not affect these but affect the other unmeasured quality variables.

In order to retain the ease of interpretation the system must be chosen such that the bulk of the variation can be easily presented to the user. This means that normally the monitoring procedure would be restricted to three latent vectors which can be presented on a 3-dimensional plot and a separate SPE plot. For very complex systems, where three latent vectors may not explain the bulk of the variation then, the system should be divided into logical modular sections which can be monitored separately. If this is impractical then two further possibilities exist: (1) Use only three latent vectors and an SPE vs time plot. Although the three latent vectors will not explain all the major variation, the variation not captured by them will still be show up in the SPE plot; (2) if it is felt that more than three latent vectors would be useful, then the number of plots can be increased, although this would quickly become cumbersome.

3.3 Monitoring a fluidized bed reactor

The multivariate monitoring procedure outlined in the last section was applied to a fluidized bed reactor (FBR). Although only simulations were used in this study, the non-linear model used was one recently identified from experimental
runs on pilot plant (Kelly, 1991). The reaction in the FBR is the hydrogenolysis of n-butane on 10% nickel on silica catalyst (Shaw et al., 1972; Kelly, 1991). Due to the small vapour holdup and large thermal holdup, the dynamics of the concentrations can be considered at steady state with respect to temperature. Although this process is non-linear, especially when the extremes of operating conditions are included, this did not create significant problems for the monitoring method, since under normal operating conditions the process is roughly linear. This process proved to be ideal for a preliminary study of the technique because although numerous measurements were available, the underlying process proved to be essentially two dimensional. A simplified schematic to this process is presented in Figure 3.3 along with the measured variables.

![Diagram](image)

Figure 3.3  A schematic of the fluidized bed reactor (FBR) including measured variables.
The first step in the procedure is to choose which variables will be considered as process variables and which will be considered as indicators of product quality. An on-line gas chromatograph (GC) was available for the analysis of the product stream, and the concentrations of the reaction products (transformed to product selectivities and reactant conversions) were chosen as the product space ($Y$). The remaining measurements were included in $X$. If the GC measurements were not available, one alternative would be to include the outlet reactor temperature in the product space, as an inferential variable for the concentrations. The division of measurements into process variable ($X$) and product variable ($Y$) spaces is highly dependent on the measurements available, and the objectives of the monitoring scheme.

The choice of the reference set and the proper scaling are critical to the successful application of the procedure. The reference set used to develop the monitoring chart will determine the variations considered to be part of normal operation and ideally includes all variation leading to acceptable performance. If the reference set variation is too small the procedure will alarm too often, and if too great the sensitivity of the procedure for indicating abnormal operation will be impaired. In an industrial application the reference set would be chosen from past data. In the simulated example, the reference set was generated from a normal distribution about a nominal point. The variations and the nominal point chosen for this simulation correspond to the conditions used in the experimental work used to develop the model (Kelly, 1991). The four case studies were used to test the procedure: (1) a change in the Hydrogen to Butane ration in the feed (from 10 to 14);
(2) a ramped increase in the cooling oil temperature (2.5K per sampling interval);
(3) a baseline drift in the chromatograph reading (2% per sampling interval), and
(4) a change in the catalyst activity (from 1.06 to 2.12).

Once a reference set has been established, the scaling of the variables must be chosen. Scaling affects the relationships between variables. Any variable with large amounts of variation will dominate the first latent vector. Unless economic or physical considerations indicate that certain variables are more important than others, the variance for each measurement should be approximately the same. Following these considerations autoscaling (i.e., mean center the data and scale to unit variance) was successfully applied to the FBR case study.

Once the reference set and the scaling were defined, the latent vectors of the PLS model were calculated. Table 3.1 presents the loading vectors (which define the latent vectors in terms on the original variables) and the cumulative percent of the variance explained after each latent variable calculation for each variable. Only the first two latent vectors were used in the monitoring procedure since they explained 86.5% of the variation in Y, and because the reduction in PRESS due to the third latent vector was small and did not justify the added complexity of the monitoring charts (it explained only an additional 3.8% of the variance of Y). The scores plotted in the monitoring chart are calculated as the product of the measurement and the loading vector \( t_1 = X_{obs}w_1 \); \( t_2 = (X_{obs} - t_1p_1^T)w_2 \); the prediction of y is calculated as
Table 3.1 Results of the PLS calculation for the fluidized bed reactor showing the loadings for each latent vector (LV) and the cumulative percent of variation explained.

<table>
<thead>
<tr>
<th>Variables</th>
<th>First LV</th>
<th></th>
<th>Second LV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Butane concentration</td>
<td>0.3002</td>
<td>29.5</td>
<td>-0.461</td>
<td>80.5</td>
</tr>
<tr>
<td>Hydrogen concentration</td>
<td>-0.2957</td>
<td>39.2</td>
<td>0.4696</td>
<td>81.36</td>
</tr>
<tr>
<td>Ratio</td>
<td>-0.3054</td>
<td>32.4</td>
<td>0.4951</td>
<td>88.34</td>
</tr>
<tr>
<td>Volumetric Flowrate</td>
<td>-0.1008</td>
<td>0</td>
<td>0.085</td>
<td>4.5</td>
</tr>
<tr>
<td>T - Inlet</td>
<td>0.1099</td>
<td>9.0</td>
<td>-0.1893</td>
<td>12.0</td>
</tr>
<tr>
<td>T - Cooling Oil</td>
<td>0.586</td>
<td>56.0</td>
<td>0.4385</td>
<td>99.05</td>
</tr>
<tr>
<td>T - Reactor</td>
<td>0.6416</td>
<td>76.5</td>
<td>0.3131</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Total % Explained of X 23.7 52.7

| Methane selectivity   | 0.4507   | 83.7         | 0.4353    | 91.0         |
| Ethane selectivity    | -0.4046  | 65.3         | -0.6862   | 82.6         |
| Propane selectivity   | -0.4594  | 83.2         | -0.4127   | 89.8         |
| Butane conversion     | 0.4471   | 78.74        | 0.255     | 81.0         |
| Hydrogen conversion   | 0.4617   | 84.0         | -0.3229   | 88.0         |

Total % Explained of Y 78.8 86.5

Regression Coefficient (b) 1.265 0.4188

\[
\hat{y} = \mathbf{T}_B \mathbf{Q}_r
\]  

or

\[
\hat{y} = \sum_{s=1}^{4} t_s b_s q_r^T
\]

In Figure 3.4 the points from the reference set are plotted on the monitoring chart and the regions of acceptable performance are defined (the shaded areas). The four
case studies described earlier were used to evaluate the ability of the procedure to detect upsets or faults; the results are presented on the monitoring chart in Figure 3.5.

![Monitoring Chart]

**Figure 3.4** The monitoring chart for the fluidized bed reactor showing the points used in the reference set. The cross-hatched area indicates the space of normal operation.

The results clearly show the different types of abnormalities which can arise. In the first case, the shift in the hydrogen to butane ratio, the disturbance is measured by the process variables and is indicated by a shift of the points on the $T_1$-$T_2$ plane. Since, small variations in this ratio are included in the reference set, the PLS model includes this type of variation and the acceptable SPE shows that the prediction obtained is still accurate. The second case, the increase in the cooling oil temperature,
Figure 3.5 The monitoring chart for the fluidized bed reactor showing the results of the simulated case studies used to evaluate the method.

indicates the same type of abnormality. Further, if the process were truly linear, the SPE would remain within the limits set by the reference set; however, due to the non-linearities in the system the SPE increases with increasing temperature until finally it exceeds the acceptable limit. The last two cases, the base line drift in the GC and the change in the catalyst activity, have no effect on the measured process variables, but they have an extreme effect on the prediction equation. These effects cause drastic increases in the SPE indicating that the relationship between X and Y has been altered. The data presented here shows the method working on a system with variation but no sensor noise. When sensor noise was added to the system
white noise equivalent to one-half of the standard deviation of the signal from each variable was added to each variable), the same changes were detectable but less prominently.

3.4 Monitoring a Methanol-Acetone-Water Distillation Column

The second simulation used to demonstrate the proposed monitoring procedure is an extractive distillation column separating an azeotropic mixture of methanol, acetone and water (MAW) (Figure 3.6). The results presented are from simulation runs using a fundamental tray-by-tray model developed at McMaster University (Chin, 1989). Example input and output data sets for both this simulation and the FBR simulation used in the previous example are described in the Appendix. Only steady-state monitoring was performed; however, the possibilities of dynamic monitoring are discussed below.

As with the FBR the measurements around the distillation column must be divided into the process and product variables. The configuration chosen is described in Table 3.2. One question that immediately comes to mind is whether the concentrations of all the components in the product streams should be included in the product variable set or only those which will ultimately be used for control. The answer to this question is case specific. However, the following rule should be followed: use only those variables in Y which are of greatest interest from a monitoring point of view. Extraneous variables in the product space do not normally affect the PLS analysis; they do, however, impair the resolution of the monitoring procedure by increasing the limits on the SPE and/or by increasing the number of significant latent vectors. Significant changes in the residuals of the important
variables can be hidden by the variance of the extraneous variables which are included in the model. For this work all the information about the product streams was included in Y. When this was done nearly all the variation was still explained in three latent vectors and therefore in no way inhibited the performance of the monitoring procedure.

Figure 3.6 A schematic of the Methanol-Acetone-Water (MAW) distillation column including the measured variables.

Autoscaling was not used in this problem, as was done for the FBR, because the relationships between the tray temperatures in the temperature profile should be kept intact. This is done because more physical interpretation can be given to the first latent vectors if the data is unscaled. Several different scaling approaches were investigated, ranging from unit scaling all variables, except those in the temperature profile, to no scaling at all, to specifically increasing the variance of certain variables.
Table 3.2  The process and product variables measured for the MAW distillation column.

<table>
<thead>
<tr>
<th>Process Measurements</th>
<th>Product Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solvent Flow</td>
<td>Flowrate of Distillate</td>
</tr>
<tr>
<td>Solvent Temperature *</td>
<td>Acetone Concentration in Distillate</td>
</tr>
<tr>
<td>Feed Flowrate</td>
<td>Methanol Concentration in Distillate</td>
</tr>
<tr>
<td>Feed Temperature *</td>
<td>Water Concentration in Distillate</td>
</tr>
<tr>
<td>Pressure</td>
<td>Flowrate of Bottoms</td>
</tr>
<tr>
<td>15 Tray Temperatures</td>
<td>Acetone Concentration in Bottoms</td>
</tr>
<tr>
<td>Condenser Duty *</td>
<td>Methanol Concentration in Bottoms</td>
</tr>
<tr>
<td>Reboiler Duty</td>
<td>Water Concentration in Bottoms</td>
</tr>
</tbody>
</table>

Note: * - Indicates variables held constant during simulation.

The following observations were made. By scaling selected variables to high variances compared to the other variables, the latent vectors used by PLS model were altered. When the scaled variable belongs to $X$, then this variable usually dominates the first latent vector. When the scaled variable belongs to $Y$, then the variables which were highly predictive of this measurement were prominent in the first latent vectors. Scaling the entire $X$ block versus the entire $Y$ block did not affect the results. When the variances are almost the same size, then small adjustments in the scaling have an insignificant effect on the results. A final recommendation, if the interrelationships between a group of variables must be maintained (and they are measured in the same units), then the same scaling factor should be used for the entire group.
As with the FBR, the reference set for the MAW column was generated from a normal distribution around a nominal point. Five case studies were used with this example, including: (1) an increase in the acetone to methanol ratio with a decrease in the water content in the feed; (2) no change in the acetone to methanol ratio, but an increase in the water content in the feed; (3) a decrease in the acetone to methanol ratio accompanied by an increase in the water content; (4) a 33% increase in the solvent flowrate, and finally (5) a 10 K increase in the steam temperature to the reboiler. Calculations similar to those done for the FBR were performed and the results plotted on the monitoring charts. The monitoring chart had to be modified from the one used for the FBR because three latent vectors were necessary to model the relevant information. The scores from the three latent vectors form the axes of the monitoring chart (Figures 3.7 and 3.8) and the SPE is shown on a separate chart (Figure 3.9). The SPE would generally be plotted sequentially in time, but can be plotted against any of the t's, if this is more convenient for the application. Since the data generated for this example was not time dependent, SPE was plotted against $T_D$, which gave the best resolution of the events.
Figure 3.7 The monitoring chart for the extractive distillation column showing the points in the reference set and the area of normal operation.

The loading vectors for the three latent vectors used in the MAW monitoring chart are presented in Table 3.3. Also presented in this table are the cumulative percent of variance explained for each variable. The loading vectors are useful in interpreting the monitoring chart, and help assign causes to abnormalities. The first loading vector indicates that the first latent vector is dominated by changes in the reboiler duty and the accompanying shifts in the temperature profile. Further evidence of this and the value of these loadings in the interpretation of the monitoring chart is seen in Figure 3.8, where the change in the steam temperature to the reboiler are clearly seen as a shift in the $T_1$ direction. The second and third vectors are concerned with changes in solvent and feedstock flowrates; for the second vector these changes are in opposite directions, for the third they are in the same direction.
Table 3.3  Results of the PLS calculation using the reference set for the extractive distillation column showing the loadings for the latent vectors (LV) and the cumulative percent variation explained after each vector.

<table>
<thead>
<tr>
<th>Variables</th>
<th>First LV</th>
<th></th>
<th>Second LV</th>
<th></th>
<th>Third LV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Solvent Flow</td>
<td>0.06</td>
<td>0.0</td>
<td>-0.68</td>
<td>44.0</td>
<td>0.68</td>
<td>99.8</td>
</tr>
<tr>
<td>Feed Flow</td>
<td>0.06</td>
<td>0.0</td>
<td>0.52</td>
<td>50.0</td>
<td>0.65</td>
<td>99.1</td>
</tr>
<tr>
<td>Tray Temp 1</td>
<td>0.37</td>
<td>86.4</td>
<td>-0.31</td>
<td>95.0</td>
<td>-0.09</td>
<td>99.6</td>
</tr>
<tr>
<td>Tray Temp 2</td>
<td>0.15</td>
<td>86.4</td>
<td>-0.11</td>
<td>94.0</td>
<td>0.03</td>
<td>96.1</td>
</tr>
<tr>
<td>Tray Temp 3</td>
<td>0.004</td>
<td>6.0</td>
<td>-0.397</td>
<td>23.0</td>
<td>0.002</td>
<td>43.7</td>
</tr>
<tr>
<td>Tray Temp 4</td>
<td>0.06</td>
<td>68.0</td>
<td>-0.106</td>
<td>92.0</td>
<td>0.02</td>
<td>92.4</td>
</tr>
<tr>
<td>Tray Temp 5</td>
<td>0.03</td>
<td>43.0</td>
<td>-0.11</td>
<td>86.0</td>
<td>0.03</td>
<td>90.3</td>
</tr>
<tr>
<td>Tray Temp 6</td>
<td>0.15</td>
<td>73.0</td>
<td>0.16</td>
<td>89.0</td>
<td>0.03</td>
<td>90.7</td>
</tr>
<tr>
<td>Tray Temp 7</td>
<td>0.15</td>
<td>64.7</td>
<td>0.21</td>
<td>90.0</td>
<td>0.03</td>
<td>92.0</td>
</tr>
<tr>
<td>Tray Temp 8</td>
<td>0.11</td>
<td>60.5</td>
<td>0.16</td>
<td>90.0</td>
<td>0.02</td>
<td>91.8</td>
</tr>
<tr>
<td>Tray Temp 9</td>
<td>0.06</td>
<td>59.5</td>
<td>0.09</td>
<td>89.0</td>
<td>0.02</td>
<td>91.5</td>
</tr>
<tr>
<td>Tray Temp 10</td>
<td>0.03</td>
<td>60.9</td>
<td>0.039</td>
<td>84.0</td>
<td>0.03</td>
<td>91.3</td>
</tr>
<tr>
<td>Tray Temp 11</td>
<td>0.03</td>
<td>62.8</td>
<td>0.028</td>
<td>84.0</td>
<td>0.03</td>
<td>97.8</td>
</tr>
<tr>
<td>Tray Temp 12</td>
<td>0.02</td>
<td>60.4</td>
<td>0.18</td>
<td>75.0</td>
<td>0.03</td>
<td>96.6</td>
</tr>
<tr>
<td>Tray Temp 13</td>
<td>0.01</td>
<td>54.6</td>
<td>0.008</td>
<td>63.0</td>
<td>0.03</td>
<td>94.6</td>
</tr>
<tr>
<td>Tray Temp 14</td>
<td>0.006</td>
<td>30.5</td>
<td>-0.004</td>
<td>29.6</td>
<td>0.03</td>
<td>86.3</td>
</tr>
<tr>
<td>Tray Temp 15</td>
<td>0.144</td>
<td>57.9</td>
<td>0.19</td>
<td>86.0</td>
<td>0.136</td>
<td>96.3</td>
</tr>
<tr>
<td>Steam Temp</td>
<td>0.874</td>
<td>99.8</td>
<td>-0.04</td>
<td>99.8</td>
<td>0.08</td>
<td>99.9</td>
</tr>
<tr>
<td>Total % Explained of X</td>
<td></td>
<td>74.9</td>
<td>87.1</td>
<td></td>
<td>98.9</td>
<td></td>
</tr>
<tr>
<td>Distillate flow</td>
<td>0.42</td>
<td>57.7</td>
<td>0.31</td>
<td>86.1</td>
<td>0.26</td>
<td>96.2</td>
</tr>
<tr>
<td>X_D Acetone</td>
<td>-0.4</td>
<td>54.5</td>
<td>-0.38</td>
<td>97.3</td>
<td>-0.06</td>
<td>97.8</td>
</tr>
<tr>
<td>X_D Methanol</td>
<td>0.3</td>
<td>30.1</td>
<td>0.43</td>
<td>86.2</td>
<td>-0.24</td>
<td>95.1</td>
</tr>
<tr>
<td>X_D Water</td>
<td>0.41</td>
<td>51.3</td>
<td>0.15</td>
<td>56.6</td>
<td>0.51</td>
<td>92.3</td>
</tr>
<tr>
<td>Bottoms Flow</td>
<td>-0.17</td>
<td>8.0</td>
<td>-0.26</td>
<td>26.0</td>
<td>0.72</td>
<td>99.1</td>
</tr>
<tr>
<td>X_B Acetone</td>
<td>-0.51</td>
<td>84.5</td>
<td>0.2</td>
<td>96.0</td>
<td>0.16</td>
<td>99.5</td>
</tr>
<tr>
<td>X_B Methanol</td>
<td>0.06</td>
<td>0.2</td>
<td>0.51</td>
<td>77.8</td>
<td>-0.24</td>
<td>86.2</td>
</tr>
<tr>
<td>X_B Water</td>
<td>0.33</td>
<td>37.6</td>
<td>-0.425</td>
<td>94.5</td>
<td>0.2</td>
<td>94.8</td>
</tr>
<tr>
<td>Total % Explained of Y</td>
<td></td>
<td>40.4</td>
<td>77.0</td>
<td></td>
<td>95.1</td>
<td></td>
</tr>
<tr>
<td>Regression coefficient</td>
<td>0.6325</td>
<td>1.485</td>
<td></td>
<td></td>
<td>1.128</td>
<td></td>
</tr>
</tbody>
</table>
Again this interpretation is confirmed by the events plotted in Figure 3.8; a positive change in the solvent flowrate is seen as a shift in the plane of $T_2$ and $T_3$, in the positive direction for $T_2$ and in the negative for $T_3$. One interesting thing should be noted in this simulation; although feed composition variations were included in the reference set, they are not explicitly accounted for in the loading vectors because they were not measured. They are accounted for in the temperature profile, but are masked by the large changes caused by changes in the reboiler duty. This means that feed composition changes are difficult to detect solely from the information in the process measurements; however, they are easily detected once the plot of SPE is included (Figure 3.9).

![Graph with legend]

Figure 3.8 The monitoring chart for the MAW distillation column showing the results of the simulated case studies used to evaluate the method.
3.5 Discussion of Present Results

To obtain the best performance when using the monitoring charts the following guidelines should be observed. The product space should be restricted to those variables of interest in the monitoring procedure. Extraneous variables in the process measurement space do not inhibit the procedure, and all possibly relevant
measurements can be included. Scaling should be performed in such a way that the variances of the measurements reflect their relative importance (the use of instrument ranges and engineering knowledge of the process should be very effective). Care must be exercised when using unit scaling not to disturb inter-relationships among the variables. Reference sets must be designed to reflect the purpose of the monitoring procedure. This means that the procedure is case specific, and each application should be tested to see if it performs according to the specifications.

The loading vectors contain information relating the original variables to the latent vectors. This information used in conjunction with past experience and pattern recognition can be instrumental in assigning possible causes to the abnormalities.

The work presented here has been restricted to steady-state monitoring. By incorporating time lags (in the time series sense) it should be possible to extend this procedure to dynamic systems.

Finally, the work presented here was for the situation where regular quality measurements are available; if these are not available, then one of the previously mentioned modifications needs to be implemented. Although these methods have some additional limitations they can still be effective monitoring tools; however, the details of their implementation must still be investigated.

A major area for future research pertains to the appropriate procedure for maintaining and updating the monitoring procedure under changing operating conditions.
3.6 Discussion of other applications of this method

The monitoring procedure presented in this chapter has been published, Kresta et al. (1991) This work was well received and has been applied by other workers to industrial problems. R. Swanson presented results of a successful application of this procedure at Eastman Kodak (Swanson, 1990). Another industrial implementation of this technology was carried out on a solution area of a Du Pont Plant (Piovoso et al., 1991). Concurrently with this work Wise et al. (1989) also investigated the application of PCA and PLS in performance monitoring. The procedure proposed by Wise and co-workers concentrates on analyzing the residual variance after removing the variation from known disturbances. In this way they concentrate their efforts on detecting two types of sensor faults: 1) a change in the bias and 2) increased variance (noise) in the sensor measurements. Further by using normality assumptions they are able to calculate upper and lower bounds for the monitoring charts using the Q statistic. While the focus of this work and that done by Wise et al. is quite different the results are complimentary, as is shown in the implementation discussed by Piovoso et al. (1991). The object of the application was to provide a monitoring tool to alert the operators of a shift in the operation of the process. In this application the ideas from both Wise et al. (1989) and Kresta et al. (1991) were integrated by using the monitoring charts discussed here and adding statistical confidence intervals.

The monitoring procedure proposed here has also been applied to a low density polyethylene reactor (LPDE) (MacGregor et al., 1991). Further work in this area has also been done at McMaster University by Carol Slama (Slama, 1992) in conjunction with Shell Canada Ltd. Ms. Slama investigated the application of the
procedures described in this chapter to a fluidized catalytic cracking unit located at a Shell Canada Ltd. petroleum refinery. Initially, the reference data from this process contained over 300 variables representing hourly averages from three and a half months of operation. Due to computational restrictions this was reduced to 142 variables and 1400 observations. Through the application of PCA and PLS it was possible to obtain lower dimensional windows on this process. Although valuable insights into the process operation were obtained the shear size and complexity of this process created some difficulties in the implementation of the monitoring procedure. The normal operation of this process varied significantly and it was difficult to discriminate between normal shifts in plant operation and upset conditions.

3.7 Conclusions

A multivariate SPC procedure has been proposed for handling large numbers of process and quality variables. Multivariate statistical procedures are used to reduce the dimensionality of these large and highly correlated data sets down to a few latent variables which contain most of the information about the process behaviour under normal operating conditions. By plotting the projections (rows) of new process observations over time on this low dimensional plane one is able to detect larger than normal process variations, and by also plotting the squared prediction errors, (ie. the perpendicular distances from the plane) one is also able to detect major changes in the behaviour of the process caused by new events.

By compressing all the information on the process down to low dimensional spaces, and using simple plots of the data in these spaces, together with meaningful control limits, the essential idea and philosophy of Shewart's (1931) SPC methods
have been preserved and extended to handle the large number of variables collected in most process industries today. The tools necessary to establish the multivariate charts (PCA and PLS) may be more complex than usually used in univariate SPC, but from the user’s point of view, the presentation of the data, and the interpretation of the results is almost as simple.

3.8 Notation

$k$ - the number of process/operational/input measurements

$m$ - the number of quality/output measurements

$n$ - the number of observations

$p_a$ - the loading vector for the $a^{th}$ latent vector of $X$ calculated by PCA or PLS

$P$ - the loading matrix for $X$ calculated by PCA or PLS

$p_i$ - the loading for the $i^{th}$ $X$ variable

$q_a$ - loading vector for $a^{th}$ latent vector of $Y$

$Q$ - loading matrix for $Y$

$t_a$ - scores associated with the $a^{th}$ latent vector of $X$

$T$ - the matrix of scores associated with $X$

$T_i$ - the axis of the monitoring chart

$u_a$ - scores associated with the $a^{th}$ latent vector of $Y$

$U$ - the matrix of scores associated with $Y$

$w_a$ - the loading vector used for prediction, calculated from PLS, for the $a^{th}$ latent vector

$W$ - the matrix of prediction loadings for $X$
$w_i$ - the prediction loading for the $i^{th}$ X variable

$X$ - $(n \times k)$ matrix of the process / operational / input measurements

$\bar{x}$ - the mean of a variable of $X$ and the vector of means

$x_i$ - the value of the $i^{th}$ X variable for one measurement

$\hat{x}_i$ - the predicted value of the $i^{th}$ X variable for one measurement

$X_B$ - the composition of the bottoms product from the EDC (Table 3)

$X_D$ - the composition of the distillate product from the EDC (Table 3)

$Y$ - $(n \times m)$ matrix of the process performance / product quality / output measurements

$y_i$ - the value of the $i^{th}$ Y variable for one measurement

$\hat{y}_i$ - the predicted value of the $i^{th}$ Y variable for one measurement

Superscript

$T$ - the transpose of the matrix

3.9 References


Chin, A., Private Communication, Department of Chemical Engineering, McMaster University (1989).


Chapter 4. Inferential variables

Chapter 3 discussed the application of PLS to process monitoring. It used steady-state data and concentrated on monitoring changes in the inter-relationships, changes in the values of the X block variables, and changes in the relationship between the X and Y blocks. In that application no effort was made to use the PLS regression equation to predict the quality measurements (Y); even though these measurements are often available much less frequently than the process measurements (X).

A natural application of PLS is to build empirical models so that quality measurements can be predicted from process measurements. This model building can be achieved using classical regression techniques for situations where truly independent process variables are chosen to predict the quality variables. The opportunity for PLS lies in applications with many correlated process variables. PLS allows the user to extract the useful predictive information in all the variables without encountering many of the problems associated with using classical regression methods. One such process is a distillation column where tray temperatures are often used to predict product compositions. Two distillation case studies will be used here to demonstrate the development of inferential control schemes based on PLS modelling.

In this chapter PLS-developed models are shown to be good predictors of the quality variables. The application of PLS to the development of inferential
models in this chapter leads to a clearer understanding of the underlying assumptions and possible pitfalls when applying any of the empirical modelling techniques in inferential control schemes, and it points out the specific advantages of using PLS.

4.1 The Rationale for Inferential Control

An inferential model is often used in process control when a measurement of the true variable being controlled is not available in real time. Reasons for the lack of real-time measurement include high cost, low reliability, and long analysis times or long dead times for sensors located far downstream. In these cases, an inferential model provides an estimate of the process variable which can be used in the design of a feedback controller to provide approximate regulation of the true variable. The methodology used in this work is independent of the control which will be used in the final application; thus, the methods presented here can be used with essentially any control technology, from FiD to model predictive control, in single loop and multivariable designs. However, the model obtained relies heavily on the control and process structure, and it is important that these be consistent between the model building and implementation stages.

By inferential model, we mean a model employing easily measured variables to predict the true variable. For a distillation column, an inferential model for the distillate composition would be of the general form

$$\log(\hat{X}_D) = f(T_1, T_2, ..., F_1, ...)$$

(4.1)

where the $^\wedge$ indicates the predicted value. Generally, the model is process specific and must be developed or tuned for each application. The source of the inferential model could be a fundamental or short-cut process model, and if such a model is available and provides sufficient accuracy with reasonable computation, this
approach is favored. Often, no fundamental model appropriate for real-time use exists, and the structure of the model must be determined by an analysis of process data. We shall term this an empirical model, and developing empirical inferential models is the topic of this chapter. The data used in developing the empirical model could come directly from the process for operating plants or from a detailed simulation for design analysis.

The general goals for an inferential model are to provide an accurate prediction of the true controlled variable under all the normal changes in plant operation. We emphasize that the application of the model is in closed-loop control, not open-loop monitoring; therefore, the model should be applicable under feedback control, an important point emphasized later in this chapter.

Three main challenges exist in developing empirical inferential models. The first is to establish the model structure, which should be addressed in as general a way as possible without the requirement to restrict the model to one or a few measured variables. The second challenge involves the design and selection of reference data used in building the model. This data must represent the process relationships for the expected range of process conditions and the structure of the process and control configuration under which the model is to be used. The third challenge is that the model predictions must be robust to sensor failure or missing process measurements. All these challenges are addressed in this chapter.

In the next section, some currently used methods for developing inferential control schemes for distillation columns are reviewed. Shortcomings of methods are that they are limited to a few inferential variables and they are often only applicable to a specific process. The application of the PLS method to this problem
is discussed, first heuristically and then in more detail, including the important issues concerning the reference data set. The PLS is then employed to develop empirical inferential models for two distillation towers. While some of the discussion and all examples are related to distillation, no restrictive assumptions are made, and the generality of the method and its easy application to other processes should be apparent.

4.2 Current Approaches to Model Building

While many models have been developed for inferential control of specific unit operations, few general approaches to model development have been proposed. In distillation, a common practice is to select one tray temperature, which gives a simple model of the form

$$\log(\hat{x}_D) = \alpha_0 + \alpha_1 T$$  \hspace{1cm} (4.2)

The model has two parameters, the slope, $\alpha_1$, is determined in the model development phase and is unchanged thereafter. The correct composition is achieved in real time by adjusting the temperature set point via a cascade controller or periodic manual adjustments based on laboratory data. This feedback is equivalent to adjusting the intercept, $\alpha_0$.

A method used by practitioners to select an appropriate tray, as presented in Tolliver and McCune (1980), is based on sensitivity of temperature to changes in feed composition, reboiler duty and reflux flow. The tray temperature which exhibits the "best" sensitivity is selected. Selection is performed by graphical analysis of
steady-state simulations, with good sensitivity having the characteristics of strong correlation with the product composition for all disturbances. This method has the disadvantage of limiting the inferential model to one tray temperature.

Another method presented by Moore et al. (1986) applies principal component analysis to simulated data. Their method selects the tray temperature which has the largest variance in a data set containing changes in the reboiler duty and reflux flow. Again, this method suffers from the restriction of using only one temperature. Choosing the temperature with the largest variance does not necessarily select the inferential variable (here, temperature) with the best predictive power, although an adequate model may still result. For example, a tray temperature which experienced much larger changes than other tray temperatures in the reference set, for example ± 5°C, would not be a good inferential variable, if the product quality increased in both cases. Thus, this method does not appear to be either generally applicable or designed to satisfy the goals of inferential control stated above.

To develop a model with multiple measured variables, one might attempt to develop an inferential model, using linear (or non-linear) regression, to fit a model of the following form

$$\log(\hat{x}_D) = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_n F_1 + \ldots$$  \hspace{1cm} (4.3)

This method can encounter difficulties because of the highly correlated nature of process data. The cause of the correlation among the measured variables can be seen by considering the response of a distillation column to an increase in the distillate flow, with all other input variables held constant. In this case, the concentration profile changes in a correlated manner, with more heavy components on each tray, and all tray temperatures increasing accordingly. Thus, all the measured tray tem-
peratures will change; however, they all change in response to the single independent disturbance that occurred. Attempting to fit such highly correlated data by linear regression may lead to numerical errors and singularity. Even if a non-singular solution is reached, the results may not be useful for inferential control. Hypothetical results reached by "overfitting" data with a high dimensional model are given in Figure 4.1. (Data specific to the BTX column is presented in Table 4.1). The sum of squared errors between the fitted model and the reference data set decreases monotonically as the dimension of the model (ie. the number of fitted parameters) is increased. However, the sum of squared errors when the model is applied to new data not used in the regression, termed the prediction error, experiences a minimum where the model best represents the actual relationships in the process. Increasing the dimension beyond this minimum results in overfitting which adapts the model to the noise in the reference data set and results in poorer future predictions (Draper and Smith, 1981; Martens and Naes, 1989).

Table 4.1 Comparison of the fitted error and the prediction error obtained from the MLR and the PLS models for the distillate composition of the BTX tower. The results show the inflation in the prediction error for the MLR due to overfitting.

<table>
<thead>
<tr>
<th></th>
<th>MLR Model</th>
<th>PLS Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted Error</td>
<td>0.0327</td>
<td>0.0339</td>
</tr>
<tr>
<td>(Reference set; 76 points)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction Error</td>
<td>0.0897</td>
<td>0.0245</td>
</tr>
<tr>
<td>(Test set; 12 points)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.1 Comparison of the prediction error and the error in fitting the reference set as the model dimension increases, showing that, although error in fitting decreases monotonically with the model dimension, there is an optimal model dimension which minimizes prediction error.

Joseph and Brosilow (1978) used linear regression to fit inferential models but included a method for selecting a few temperatures from all of the measured tray temperatures. Their selection method, based on the condition number of the data set, prevented overfitting but again limited the model to a few measurements.

Many additional approaches have been proposed for using a few distillation tray temperatures for inferential control (Yu and Luyben, 1984; Fuentes and Luyben, 1983; Luyben, 1969). These methods require considerable insight into the tower responses and can provide good results when applied by an experienced practitioner. However, they cannot be extended to other processes, nor do they indicate the potential improvement available by using models with additional variables.

In conclusion, several methods are available to select one or a few measurements for use in building an inferential model. This chapter introduces a method for potentially including all measurements, thereby ensuring that all process information is included in the model, and furthermore obtaining an averaging of the noise arising
from the process measurements. Some of these ideas have been presented previously by Kresta et al. (1990a,b) and by Mejdell and Skogestad (1990). Recently this work of Mejdell and Skogestad (1991a) and two follow up papers have been published (Mejdell and Skogestad, 1991 b,c). The first paper (1991a) compares the estimates obtained from inferential models based on a Kalman-Bucy Filter, Brosilow's Inferential Control Method (Joseph and Brosilow, 1978) and a static Principal Component Regression model for a simulated linear distillation column using \( \mu \) analysis. The result of the comparison was that the Kalman-Bucy filter and the PCR model provide estimates of similar accuracy, the PCR model was recommended for its simplicity. The second paper investigates a number of application issues in the development of PLS inferential models using simulated distillation columns including: non-linearity; presence of measurement noise; scaling; and multicomponent systems. Among the results Mejdell and Skogestad show that linearization of the temperature profile produces a more significant improvement in the quality of the estimate than linearization of the composition using \( \log(X_p) \). The linearization of the temperature profile is relatively easy for binary distillation but becomes more difficult for multicomponent systems. They also investigate the effects of three different scaling calculations in the presence of measurement noise. Their results are similar to some results obtained in this work and are discussed in section 4.7.

The final paper shows the application of these ideas to a pilot plant distillation column, comparing models developed from simulation and designed experiments. The work of Mejdell and Skogestad is very interesting and presents information which is complementary to that presented here. This work concentrates on the importance of the reference set and the effect of feedback in the system on the
inferential model developed and further, shows that PLS-developed models can be implemented in a way which makes the model extremely robust to known sensor failure.

For all of the methods discussed here, the success of the inferential model is evaluated on the ability to predict the steady-state outlet composition under a variety of conditions and by feedback control performance. An alternate approach has been recently proposed by Lee and Morari (1991) for choosing the tray temperatures based on the robustness of the prediction.

4.3 Determining the Inferential Model Via PLS

Inferential model development is based on a reference data set which can be separated into two blocks; X associated with the process measurements and Y associated with quality measurements. In the final application, the Y measurements are not normally available and are inferred; however, they must be available as part of the original reference data set. The objective is to develop an inferential model such that it is possible to predict future values of the quality variables using future measurements of the process variables.

Clearly, the key issue in the model fitting is the high correlation among the measured variables. As shown in Figure 4.1, over-parameterized models fit the original data set well but lead to poor predictions. PLS has the ability to develop a good predictive model from data sets containing measurements on a large number of highly correlated variables by performing the regression using a small number of orthogonal latent variables which are linear combinations of the original variables. It is generally accepted that using only a limited number of new variables results in a biased estimator, but one with well-conditioned parameter estimates (i.e. the ones
with smaller confidence intervals) and good predictions (Kowalski and Seasholtz, 1991). Since the purpose of inferential models is to predict future relationships, rather than to fit past data, the PLS-developed model should perform better than models developed from multiple linear regression methods.

The models in this chapter were developed for both $X_0$ and $X_B$ simultaneously using PLS2 but for clarity the discussion concentrates on predictions of $X_D$. Since no assumptions about the process are used, extensions to other processes are straightforward. The model structure considered is given below.

$$\log(\hat{X}_D) = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \ldots + \alpha_n F_1 + \ldots \quad (4.4)$$

The model is linear in the coefficients and the predicted variable is the logarithm of the heavy key composition in the distillate product. It is well known that this transformation improves the linearity of this type of model over an extended range and is used throughout the work (Joseph and Brosilow, 1978). Other approaches to linearization are discussed by Mejdell and Skogestad (1991b).

The PLS model maintains low dimensionality while still retaining all measured variables (the tray temperatures, the flows of reflux, distillate, bottoms, and the reboiler steam temperature) by summarizing the information contained in these variables in a small number of independent "latent" variables which are used in the regression. The physical justification for expecting a low dimension in a distillation process is that only a few input variables are changing independently, e.g., feed composition, reflux flow, and reboiler duty. Thus, all measured variables are changing in a highly correlated manner to these changes, and a substantial reduction in dimension is possible. It is important to recognize that the ability to fit a model
using a small number of latent variables is the result of the process relationships, not the PLS method. If a reduction in dimension is not possible, then the PLS method will yield the same result as linear regression.

The following is a schematic of the reduction in dimensionality achieved with PLS. The dimensions in parentheses are from the BTX distillation example presented in section 4.5.

\[
\begin{align*}
\begin{bmatrix}
T_1 \\
T_2 \\
T_3 \\
F_1 \\
\vdots
\end{bmatrix}
\quad \Rightarrow
\begin{bmatrix}
t_1 = p_1T_1 + p_2T_2 + \ldots \\
t_2 = \ldots \\
t_3 = \ldots \\
\vdots
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
\log(\tilde{X}_D) = f_1(t_1, t_2, t_3) = \alpha_0 + \alpha_1T_1 + \alpha_2T_2 + \ldots + \alpha_nF_1 + \ldots \\
\log(\tilde{X}_G) = f_2(t_1, t_2, t_3) = \gamma_0 + \gamma_1T_1 + \gamma_2T_2 + \ldots + \gamma_nF_1 + \ldots
\end{bmatrix}
\end{align*}
\]

(4.5)

The dimensionality is reduced from the total number of measurements, 45, to the small number of latent variables, 3, with each latent variable expressed as a linear combination of all measured variables. The dimensionality of the latent variables is selected to be near the minimum in the prediction error for the reference data set as shown in Figure 4.1 by cross validation (discussed in Chapter 2). In most distillation columns, the dimensionality of three or four would be typical, but a few additional latent variables may be needed to explain the non-linear nature of some
distillation towers. Table 4.2 shows the typical information used to evaluate the benefits of additional latent variables for the MAW case study with noise. Based on this information an inferential model with 5 PLS dimensions was selected. The true controlled variable(s) can be expressed as a function of the latent variables and through a transformation, as a function of all measured variables. The result of this procedure is a model which uses essentially all the useful predictive information in the measurements but which prevents overfitting by limiting the dimension of the model.

Table 4.2  Cross-validation information used to select proper PLS dimension for the MAW open-loop steady-state inferential model (with white noise).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>PLS</th>
<th>CSV</th>
<th>PRESS</th>
<th>CSV_{individual}</th>
<th>Percent variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>6.709</td>
<td>0.1879</td>
<td>0.7118 0.5343</td>
<td>61.2 37.3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6.130</td>
<td>0.1930</td>
<td>0.7143 0.1326</td>
<td>83.3 63.8</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7.697</td>
<td>0.0735</td>
<td>0.7876 0.3226</td>
<td>93.5 73.4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.733</td>
<td>0.0328</td>
<td>0.4698 0.6837</td>
<td>95.4 87.7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>9.710</td>
<td>0.0307</td>
<td>0.9705 0.9931</td>
<td>95.6 90.0</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.246</td>
<td>0.0315</td>
<td>1.2533 1.0392</td>
<td>95.4 90.2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1.236</td>
<td>0.0302</td>
<td>1.2408 1.0973</td>
<td>95.5 90.1</td>
</tr>
</tbody>
</table>

The measurements which have the largest and smallest effects on the predictions are easily inferred from an inferential model having the form of equation (4.4). Variables with very small α's will have very little effect on the predictions. This type of evaluation, along with a good understanding of the physical process, can provide some insight into the mechanisms underlying the inferential relationships.
Once the model has been developed from a reference data set where both \( X \) and \( Y \) are available, its main use will be for the prediction or inference of future \( y \)'s given new vectors of process variable measurements. Suppose a new vector \( x^T \) of process measurements becomes available, which is mean centered and scaled to be consistent with the reference set. Predicted values of the controlled variables (\( \hat{y} \)) can be obtained, as discussed in Chapter 2, either from the linear regression form of the model as

\[
\hat{y} = x^T \alpha
\]

or from the PLS latent vector form as follows:

(i) calculate the scores for each PLS dimension \( a=1,2,...,A \)

\[
t_a = e_{a-1}^T w_a / w_a^T w_a
\]

\[
e_a^T = e_{a-1}^T - t_a p_a^T \quad (e_0 = x)
\]

(ii) calculate the new predictions as

\[
\hat{y} = \sum_{a=1}^{A} t_a b_a q_a^T
\]

A major concern with any inferential scheme is how it can handle failure of process sensors when this is recognized by, for example, checking the range of the signal. Thus, a failure results in an unknown value for the process measurement. Obviously inferential methods that depend upon only a few sensors are extremely sensitive to such failures. The sensitivity of the PLS developed models will depend upon which form is used. Since a large number of process sensor measurements are used in the linear regression model form (4.6), if a sensor fails and its reading is replaced by its mean or some other default value it will usually have a much smaller
effect on the predictions than models relying upon only a few sensors. On the other
hand, by using more sensors the probability of a failure is higher. Therefore, it is
recommended that the latent vector form of the PLS model (Equations 4.7 to 4.9)
be used along with the missing data feature of the NIPALS algorithm described in
Chapter 2. This will make the predictions almost completely insensitive to the
recognized failure of only a few sensors. The missing sensor data will not bias the
prediction, but only increases its variance slightly due to the reduced information.

The effects of losing the critical bottom tray temperature measurement in
the methanol-acetone-water (MAW) column is illustrated in Figure 4.2, for both the
regression equation form of the model and the latent variable form with its missing
data feature. For the regression equation form the average of the measurement is
used (ie. zero after mean centering), for the latent variable form the value is simply
designated as missing. The predictions from the latter form are almost unaffected
by the critical sensor failure since the other working sensors were able to provide
sufficient information. However, the predictions from the regression equation form
were greatly affected by simply replacing the missing temperature measurement by
its average value. Since all measurements are mean centered for the PLS calculations,
this is equivalent to discarding the information from the faulty sensor. Similar studies
on the benzene-toluene-xylene (BTX) column show smaller deterioration in the
regression equation form because the model predictions are based on many more
temperatures and none of them is dominant (Figure 4.3).
Figure 4.2 Comparison of the predictions of the key bottoms composition for the MAW column obtained using two different formulations of the model. These results show the effect of missing data (i.e., sensor failure) on the prediction. The missing data is for temperature 1, a key variable in the model.

Figure 4.3 Comparison of the predictions of the key distillate composition for the BTX column obtained using two different formulations of the model. These results show the effect of missing data (i.e., sensor failure) on the prediction. The missing data is for tray temperature 12.

As discussed in Chapter 2, several methods for variable scaling are available and the choice of the scaling procedure often depends on the application and all of
the methods mentioned could be used. In most of the models developed in this chapter the variables are measured in the same units, thus it was felt that the best approach to the scaling problem would be to use no scaling and to design the disturbances to the process in such a way that it reflected their relative importance in the actual plant. This approach worked very well for the steady-state predictions.

4.4 The Reference Data Set

The success in building an empirical model for inferential control will depend critically on the reference data collected. There are several important points that will be illustrated in the examples to follow. The first is that the reference data must contain information on all of the important manipulated variables and process disturbances, and the measurement noise in each sensor must be similar to that which will be encountered in the final application of the model. Some methods that have been proposed for building inferential models (eg. Moore et al., 1986) have collected "data" from a detailed simulation of the process and considered perturbations only in the manipulated variables. If real data is collected from a representative period of operation of the actual process, responses to all relevant input variable changes would be included.

The second, and most important point, is to remember that when building an empirical model from plant data by any method, one will only obtain a model for the correlation structure that was present when the data was collected. The model can be used to accurately predict the future behavior of the process only as long as the process continues to operate in a similar fashion. The introduction of feedback may alter this mode of operation drastically, and so, if one wishes to obtain an empirical model for inferential control, it must be done using data collected under
the methods mentioned could be used. In most of the models developed in this chapter the variables are measured in the same units, thus it was felt that the best approach to the scaling problem would be to use no scaling and to design the disturbances to the process in such a way that it reflected their relative importance in the actual plant. This approach worked very well for the steady-state predictions.

4.4 The Reference Data Set

The success in building an empirical model for inferential control will depend critically on the reference data collected. There are several important points that will be illustrated in the examples to follow. The first is that the reference data must contain information on all of the important manipulated variables and process disturbances, and the measurement noise in each sensor must be similar to that which will be encountered in the final application of the model. Some methods that have been proposed for building inferential models (e.g., Moore et al., 1986) have collected "data" from a detailed simulation of the process and considered perturbations only in the manipulated variables. If real data is collected from a representative period of operation of the actual process, responses to all relevant input variable changes would be included.

The second, and most important point, is to remember that when building an empirical model from plant data by any method, one will only obtain a model for the correlation structure that was present when the data was collected. The model can be used to accurately predict the future behavior of the process only as long as the process continues to operate in a similar fashion. The introduction of feedback may alter this mode of operation drastically, and so, if one wishes to obtain an empirical model for inferential control, it must be done using data collected under
a feedback control scheme that resembles, as closely as possible, that in which it will be used. A mechanistic or causal model would not suffer from such difficulties. These points will be illustrated in the examples to follow.

4.5 Distillation Case Studies

Two distillation column cases studies are presented in this chapter, with each tower used to demonstrate specific facets of empirical model building. One is a straightforward benzene-toluene-xylene (BTX) separation, and the second is an extractive distillation of an azeotropic methanol-acetone-water (MAW) mixture. They were simulated with rigorous tray-to-tray models: the BTX column with the SimSci PROCESS™ package (1986) and the MAW column with a steady-state and dynamic models developed in previous research (Chin, 1990). Details of these simulations are presented in Tables 4.3 and 4.4; example input and output data files are described in the Appendix. Schematics of the two processes along with their key manipulated and disturbance variables are given in Figures 4.4 and 4.5. Unless stated otherwise, the empirical models were developed using the entire temperature profile to predict the key compositions in both the distillate and the bottoms. Flowrate measurements are included in several cases to illustrate specific points but are not generally included in the models since they do no improve the predictions. This same result was obtained by Mejdell and Skogestad (1991a).

Data sets containing variation in the variables listed in Figures 4.4 and 4.5 were used to develop the inferential models. The data includes cases with only one variable changed from the base conditions and other cases with many variables changed concurrently. The ranges of variability for the inputs are presented in Tables 4.3 and 4.4. The ranges of variations for the BTX column are similar to those published
by Tolliver and McCune (1980) on the same system, while the ranges for the MAW column are similar those used by Latosinsky (1988) when identifying models from the pilot plant of the extractive distillation column. The individual simulations were performed without incorporating noise into the measurements. White noise, with specified standard deviation, was added to the data in the post processing stage. For the open-loop reference sets measurement noise was added to the temperature measurements (creating signal to noise ratios ranging from 35:1 to 5:1 for the MAW column and from 60:1 to 30:1 for the BTX column); however, the product compositions were simulated without noise. The effects of incorporating noise in the quality variables was investigated and it was determined that up to signal to noise ratios of 5:1, there was little or no effect on the models obtained; the same number of latent variables were identified and the final model coefficients were essentially unchanged. Before building models using these reference sets, standard pretreatment of the data was performed. Checks for outliers were performed. The system was
linearized through the use of the logarithm of the composition and the data was mean centered and scaled. For the steady-state results no scaling was performed as all the temperatures were measured in the same units.

When comparing the inferential models obtained it is often useful to compare the model coefficients. In order to make this comparison more comprehensible the coefficients are presented graphically, where the abscissa indicates the process variable (e.g. tray number) and the ordinate is the corresponding coefficient in the model. This representation makes it easy to recognize variables having a large effect in the inferential model. Figures 4.6 and 4.7 show the inferential models developed for prediction of the controlled variables under steady-state open-loop base case conditions for both columns. Also, the quality of the predictions were evaluated by plotting the predicted value vs. the true measured value with perfect prediction represented by a solid line (e.g. Figure 4.2). The predictions are compared on the basis of the change in the logarithm of the outlet composition, the variable of interest.
Table 4.3 Simulation conditions for the BTX distillation column.

<table>
<thead>
<tr>
<th></th>
<th>Base Case Conditions</th>
<th>Variation in Steady-state reference set²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs¹</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feed Flowrate</td>
<td>275</td>
<td>±13%</td>
</tr>
<tr>
<td>Feed Temperature</td>
<td>390.9</td>
<td>±10°C</td>
</tr>
<tr>
<td>Feed composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzene</td>
<td>10.0</td>
<td>±10%</td>
</tr>
<tr>
<td>Toluene</td>
<td>45.0</td>
<td>±10%</td>
</tr>
<tr>
<td>Xylene</td>
<td>45.0</td>
<td>±10%</td>
</tr>
<tr>
<td><strong>Manipulated variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distillate Flowrate</td>
<td>151.5</td>
<td>±6%</td>
</tr>
<tr>
<td>Reboiler Duty</td>
<td>11.32×10⁶</td>
<td>±8%</td>
</tr>
<tr>
<td>Pressure on all trays</td>
<td>1 atm</td>
<td>constant</td>
</tr>
<tr>
<td>Tray Efficiency</td>
<td>100%</td>
<td>constant</td>
</tr>
<tr>
<td><strong>Output produced</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distillate Flowrate</td>
<td>151.5</td>
<td></td>
</tr>
<tr>
<td>Distillate Composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzene</td>
<td>18.15</td>
<td></td>
</tr>
<tr>
<td>Toluene</td>
<td>80.84</td>
<td></td>
</tr>
<tr>
<td>Xylene</td>
<td>1.01</td>
<td>X_D 13.1 - 0.001%</td>
</tr>
<tr>
<td>Bottoms Flowrate</td>
<td>123.5</td>
<td></td>
</tr>
<tr>
<td>Bottoms Composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzene</td>
<td>0.0</td>
<td>X_B 13.1 - 0.001%</td>
</tr>
<tr>
<td>Toluene</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>Xylene</td>
<td>98.97</td>
<td></td>
</tr>
</tbody>
</table>

¹ These were varied individually and in combination
² The reference data set contains 76 points
Table 4. Simulation conditions for the MAW distillation column.

<table>
<thead>
<tr>
<th>Inputs 1</th>
<th>Base Case Conditions</th>
<th>Variation in Steady-state reference set²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feed Flowrate</td>
<td>110.0</td>
<td>±10%</td>
</tr>
<tr>
<td>Solvent Flowrate</td>
<td>44.5</td>
<td>±10%</td>
</tr>
<tr>
<td>Feed Temperature</td>
<td>315.5</td>
<td>±10°C</td>
</tr>
<tr>
<td>Solvent Temperature</td>
<td>333.15</td>
<td>±10°C</td>
</tr>
<tr>
<td>Steam Temperature</td>
<td>405.0</td>
<td>±2.5°C</td>
</tr>
<tr>
<td>Feed composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetone</td>
<td>66.0</td>
<td>68%-63%</td>
</tr>
<tr>
<td>Methanol</td>
<td>30.0</td>
<td>28%-33%</td>
</tr>
<tr>
<td>Water</td>
<td>4.0</td>
<td></td>
</tr>
</tbody>
</table>

| Output produced       |                      |                                          |
| Distillate            |                      |                                          |
| Methanol              | 90.29%               | 90.6 - 89.7                             |
| Acetone               | 1.15%                | 0.937 - 1.534                           |
| Bottoms               |                      |                                          |
| Methanol              | 7.61%                | 8.97 - 6.60                             |
| Acetone               | 16.49%               | 15.1 - 18.4                             |

¹ These were varied individually and in combination
² The reference data set contains 80 points, chosen randomly from a normal distribution with these standard deviations.
Figure 4.6 The inferential models for the BTX column developed from the reference set described in Table 4.2. The plots show the model coefficients used to relate tray temperature on tray $i$ to the controlled variable $X_D$ in (a) and $X_B$ in (b).
Figure 4.7 The inferential models for the MAW column developed from the reference set described in Table 4.3. The plots show the model coefficients used to relate tray temperature on tray i to the controlled variable $X_D$ in (a) and $X_B$ in (b).

4.6 Results of Steady-State Case Studies

The following results demonstrate the ability of the PLS algorithm to determine an inferential model with good predictive accuracy, when provided with a proper reference data set. Two advantages of using PLS to obtain this inferential model have already been shown when discussing the formulation of the algorithm. By using only the appropriate number of latent variables the PLS model avoids over-
fitting and obtains more accurate predictions than the standard multiple linear regression (MLR) formulation (Figure 4.1). The BTX column, which contains more tray temperatures in the model, is more susceptible to overfitting when MLR is used and shows a larger benefit to using PLS. The same conclusion could be reached by comparing the regression coefficients obtained using PLS and MLR, and the confidence regions for the coefficients obtained using MLR. For the MAW column the models obtained by the two methods are very similar and the confidence regions for the MLR coefficients are small. Whereas for the BTX column the MLR confidence regions are very large and no coefficients are outside these limits, indicating no significant model was found. The benefits of using PLS increase with increasing amounts of correlation among the data (either resulting from many measurements or from the nature of the data); for applications with low correlation, other multiple regression methods will also provide adequate results. However, even in these situations PLS provides the advantage that predictions calculated using the latent variable formulation of the model (equations 4.7 to 4.9) will be more robust to missing data (i.e. sensor failure), as illustrated in Figure 4.2 and 4.3.

Regardless of which estimation method will be used in the model building, MLR or PLS, the first concern should be that the reference data adequately represents the process operation. The following results clearly demonstrate how the types of input variation in the reference data set affect the inferential model. The importance of including variation from all important inputs is demonstrated using results from the MAW column, which had models developed from two reference data sets. One reference data set had variation in the manipulated variables only, while the other had variation in both the manipulated and disturbance variables. The two inferential
models developed from these reference data sets are compared in Figures 4.8 and 4.9, which show their abilities to predict distillate compositions for a test set. As shown in the figures, the predictions are much better for the cases with a richer reference set, i.e., for the data set which included changes in both manipulated and disturbance variables. This same calculation was also performed for the simpler BTX column and the results show that there is very little difference between the two reference sets. The reason for this apparent contradiction is that the disturbances and the manipulated variable changes in the BTX column change both the tray temperatures and the exit compositions in a similar way, therefore the model developed based on the manipulated variable changes provides reasonable prediction in all cases. Depending on the process, the difference between the two reference data set designs may be large or small, but since we can rarely know the extent apriori; the general principle applies that all variability should be included. In the work presented here the models are derived using 'good' reference sets.

Figure 4.8 Prediction of the methanol composition in the distillate for the MAW column using a model developed from a data set containing typical variation of both manipulated and disturbance variables.
Figure 4.9 Prediction of the methanol composition in the distillate for the MAW column using a model developed from a data set containing only variation in the manipulated variables. This shows the degradation of the prediction when an improper data set is used.

If a simulation is being used to generate the reference data set the importance of adding noise to the data is demonstrated in cases with the BTX distillation column. For this example the measurements for three flows were added to the base case data set. The model was initially fit using the reference data without noise. The resulting model coefficients for the inferential model, i.e. equation (4.4), are plotted in Figure 4.10 as solid squares. The large values for the flow rate coefficients indicate the importance of flow variables in the model. Subsequent models were developed with the same reference data except that noise was added to the flows and temperatures. As indicated in the Figure, the coefficients for the temperatures did not change significantly since the expected noise levels were low; however, the noise significantly affects the coefficients of the flows, resulting in reducing the influence of
flows on the predictions. Naturally, this would not be an issue when using real process data, as long as no additional filtering or smoothing is applied to the reference data which would not be applied in the inferential controller.

![Graph showing signal to noise ratio effects on model coefficients for distillate composition in the BTX column.](image)

**Figure 4.10** The effects of the signal to noise ratio on the model coefficients for distillate composition in the BTX column. The abscissa consists of the measured input variables in $X$, the tray temperatures, reboiler duty, distillate flow, and feed flow.

As discussed previously, another critical factor to consider in the collection of a reference set for building empirical models is the process and control system structure that exists at the time of data collection. To illustrate this, three separate sets of reference data were developed for the BTX column by varying input variables in a simulation with the three control structures shown in Figure 4.11: open-loop, single-loop feedback, and cascade. For the cases which involved control, the set point (of the highest level controller if in a cascade) was varied in place of the variation in the distillate flow rate, and all other variation was the same for all cases.
For this demonstration, an additional variable was added to the reference set, the distillate to feed ratio. The variable was added because it is a measure of the effect of the manipulated variable. The model coefficients for these cases are given in Figures 4.12 and 4.13, and they show significant differences. The model coefficient for the distillate to feed ratio is shown in the box insert in Figures 4.12 and 4.13 for the different cases. The coefficient for this variable changes sign between the model based on open-loop data and those based on feedback. This is to be expected since it is well known in the process identification literature that when fitting empirical models containing manipulated variables to data collected under feedback one identifies the reciprocal of the process controller model rather than the open-loop process model (e.g. Box and MacGregor, 1974; Ljung et al., 1974). With steady-state data such as used here the negative inverse of the feedback controller gain would result. Since the temperature profile in the vicinity of the control tray, 12, is closely correlated with this distillate flow, its coefficients are also forced towards negative values as feedback is introduced (Figures 4.12 and 4.13). Further discussion on the effects of feedback are given in MacGregor et al. (1991).

![Diagram](image)

Figure 4.11 Schematics of different control structures applied to the BTX column
Figure 4.12 Comparison of the model coefficients obtained from open-loop data versus that obtained from closed-loop data (feedback) for the BTX column.

Figure 4.13 Comparison of the model coefficients obtained from open-loop data versus that obtained from closed-loop data (cascade) for the BTX column.

The dramatically different coefficients in the three inferential models will obviously provide very different predictions. The model to use will be the one based
on data collected under conditions closest to those expected in the final application. If the model is to be used only for monitoring of the open-loop process, then the model developed using open-loop data should be employed. If it is intended to be used in an inferential feedback loop then the model developed under the cascade feedback would most closely resemble the structure of the final application. This is demonstrated in Figure 4.14, where inferential models developed for the three different control structures were used to predict future data, collected under cascade conditions. The best predictions were obtained when the same structure is used in the model development and the final control structure. The open-loop model predictions were very poor.

![Figure 4.14](image)

Figure 4.14 Comparison of the predictions obtained using three models developed under different control structures for the BTX column showing the better accuracy of models developed with feedback data. The data used for the comparison was collected under cascade control.

This situation gives rise to a dilemma. In order to develop an inferential model for feedback control one must collect data while operating under this model.
Obviously, if the latter situation existed one would not generally have need for the model. Therefore, by necessity developing an inferential model from plant data will be an iterative procedure. The first iteration will be to develop an approximate model for data collected under whatever feedback scheme is currently in existence (or open-loop in no feedback exists). This model would then be applied in an inferential feedback scheme, and its coefficients re-estimated after sufficient data has been collected under this mode of operation. This iterative procedure was used for the two distillation studies in this work, and in neither case were more than two iterations required to obtain a model of sufficient accuracy. An alternative on-line recursive identification approach was employed by MacGregor and Wong (1980).

Steady-state case studies can also be used to demonstrate the value of using all measurements in the model. A measure of the control performance was taken to be the offset, i.e. difference between the desired value of the composition and the value achieved via inferential control, without analyzer feedback. Two inferential strategies were considered; 1) control using a single tray temperature and 2) inferential control using all tray temperatures. Figure 4.15 shows the steady-state offset for these two inferential control strategies, along with no feedback control for comparison, as the most important disturbance, feed composition, varies. The PLS-developed inferential model using all tray temperatures performs better, resulting in a very small offset compared with the single-tray model.
Figure 4.15 Comparison of steady state offsets, caused by a change in the feed composition, under different inferential control schemes for the BTX column.

4.7 Scaling

The scale dependence of PLS models was introduced in Chapter 2, where it was shown that by scaling the X block variables, the order in which these variables are included in the model could be influenced. In this section two aspects of scaling are demonstrated using the distillation column simulations.

First, the steady-state simulation of the BTX distillation column can be used to demonstrate some interesting effects of variable scaling on the model coefficients obtained when using PLS. The general effects of variable scaling were discussed in Chapter 2; however it is also of interest to investigate the effects of specific scaling. Consider a situation where two variables are scaled such that their variance dominates the X block. Further suppose that one of the variables is highly correlated with the Y block variable of interest and the second has very little correlation to the same
variable. How would this scaling affect the inferential model? The results of this situation are shown in Figure 4.16 where tray temperatures 12 and 28 are scaled such that they dominate the X block variance. The model shown is for the prediction of $X_D$ and the results shown agree with what is expected from theory. The first PLS dimension includes both scaled temperatures; but as more PLS dimensions are calculated and the model is refined the tray temperature with little correlation to the Y block variable of interest losses all significance in the inferential model. The final inferential model is dominated by the variable with the highest variance and the highest correlation with Y. (For prediction of $X_B$ tray temperature 28 is retained and tray temperature 12 becomes insignificant).

![Figure 4.16 Model coefficients for the PLS model for the prediction of $X_D$ based on open-loop steady-state data from the BTX column, when the variances of tray temperatures 12 and 28 were increased 100 times. The plot shows that while both temperatures are included in the model after one PLS dimension (LV 1), only the temperature related to $X_D$ remains in the final model (LV 3).](image-url)
In the previous section the models were developed without scaling the temperature profile. This was done because the temperatures were measured in the same units and by not scaling the PLS-developed model emphasized those temperatures with both high predictive ability and high variance. This was a successful approach and provided accurate predictions of the key outlet compositions. The PLS-developed steady-state models for the MAW column are summarized in Figure 4.7. These models provided excellent steady-state predictions; however, examining the regression coefficients it is clear that both models rely heavily on tray temperatures at the bottom of the column which could cause difficulties if the model for $X_D$ (Figure 4.7a) were used in dynamic predictions of the distillate composition. The temperature variability at the bottom of the column is much greater than that near the distillate, thus the emphasis on the temperatures near the bottom of the tower. The high dependence of the PLS model for $X_D$ on the temperatures at the bottom of the MAW column can be modified by a different scaling approach. Several possible approaches exist. One common approach is to use unit scaling; however, due to the low signal to noise ratios at the top of the column it was felt that the scaling suggested by Martens and Naes (1989) would be more appropriate. This approach provides an alternative to simply scaling by using a weighting equal to the reciprocal of the standard deviation ($W_x = 1/S_x$), by using the weighting, $W_x = 1/(S_D + S_x)$, where $S_x$ is an estimate of the standard deviation of the noise. The estimate of the standard deviation of the noise can be obtained from previous information about the system or, as Martens and Naes suggest, by performing a PLS calculation on the data and obtaining an estimate of the noise from the unexplained variation in each variable. This approach was used to estimate the standard deviation of the noise and the results
agreed with the known levels of noise added to the temperature measurements. A third weighting was suggested by Mejdl and Skogestad (1991b), where \( W_x = \frac{1}{s_d \times s_x} \). In their simulation this weighting provides slightly better performance than \( W_2 = \frac{1}{s_d \times s_x} \). The only drawback to this scaling is that with signal to noise ratio of less than 1.0 the weighting is zero or negative. This does not provide problems in distillation column applications with high signal to noise ratios; however, if this method were used to develop inferential models for other systems, this weighting should be applied with caution. Using the Martens and Naes scaling a new model was developed for the prediction of \( X_b \), which provided more emphasis on the temperatures near the top of the column. These two models are compared in Figure 4.17. The Martens and Naes model and the model with no scaling have similar steady-state predictions (Figure 4.18) but as would be expected from the plots in Figure 4.17, when used for dynamic prediction (Figure 4.19) the Martens and Naes scaled model generally provides a slower but smoother response to disturbances entering the column. The new model for \( X_b \) showed little change from the unscaled version because the bottom temperature had both the highest variance and the most predictive power. All the models in the next section are developed using Martens and Naes scaling.
Figure 4.17 Model coefficients for the PLS model, for the prediction of $X_D$, based on open-loop steady-state data from the MAW column, comparing two methods of scaling.

Figure 4.18 Comparison of steady-state prediction of the distillate composition leaving the MAW column based on different scaling.
Figure 4.19 Comparison of the open-loop dynamic response to a change in the feed flowrate in the MAW column obtained using PLS models based on the two different scaling procedures. The Martens and Naes scaling generally provides a smoother response.

4.8 Results of Dynamic Case Studies

The steady-state cases demonstrated that improved inferential models could be developed by a combination of the PLS modelling approach and proper reference data design. For successful application for control, the inferential prediction should have a "good" dynamic response. By good dynamics, we generally mean 1) a fast response between the manipulated and predicted value, 2) a smooth, monotonic response to all inputs, and 3) a fast indication of the disturbances, perhaps before the product composition is upset.

The work in this section represents an initial study into using PLS-developed inferential models for prediction of the dynamic response of a process to different disturbances. Two types of models are developed, one based solely on steady-state data and the second based on dynamic data. A dynamic simulation of the MAW column is used to generate the dynamic data and to evaluate the models.
The development of steady-state models requires less data and engineering effort than dynamic models. Thus, it would be advantageous if the models developed in the previous section could be applied directly for control. The model for the prediction of $X_b$ in MAW, Figure 4.7b, relies almost entirely on reboiler temperature changes to predict the exit composition and thus the predictions and the actual open-loop dynamic response of the bottoms composition are very similar regardless of the disturbance, a typical response is shown in Figure 4.20. Conversely, for the prediction of $X_d$, the model depends on several temperatures throughout the column and the prediction of the open-loop dynamic response of the distillate composition will depend on the disturbance entering the column. Examples of open-loop predictions of the dynamic responses for the MAW distillate composition for various input step changes are shown in Figure 4.21 and are compared with the actual distillate composition. These results show that the dynamic prediction using an inferential model which contains all tray temperatures depends on the disturbance entering the system. The response to a change in the feed flowrate (Figure 4.21a) shows a very desirable response if the prediction were to be used in a cascade type of control scheme; however, the predicted responses to changes in feed composition (Figure 4.21b) and solvent temperature (Figure 4.21c) would cause difficulties in the same control scheme.
Figure 4.20 Comparison of the predicted open-loop dynamic response from the steady-state PLS model under Martens and Naes scaling (dashed line) and the actual simulated response (solid line) for the key component in the bottoms for the MAW column. The disturbance shown is a 10% increase in feed flowrate.

Alternately, it may be advantageous to consider fitting a dynamic inferential model. Many possibilities exist for the structure of the dynamic model. Certainly, a fundamental model could be used; however, it is often beneficial to retain the simplicity and generality of an empirical model. The goal is to choose a structure, which provides sufficient flexibility to model the dominant dynamic characteristics of the process. The structure used in this study expands the X matrix to include not only the present values of every input measurement, but also the input measurements from two past intervals. In this way dynamic information is incorporated into the reference set (the difference between two input measurements at different times gives an estimate of the dynamics at that time). Thus, the model has the form
Figure 4.21 Comparison of the predicted open-loop dynamic response from the steady-state PLS model under Martens and Naes scaling (dashed line) and the actual simulated response (solid line) for the key component in the distillate for the MAW column. The following disturbances are shown: a) a 10% increase in feed flowrate, b) a change in the feed composition from 30/66/4 to 34/62/4, and c) a 15 degree increase in the solvent temperature.
\[ \log(\hat{X}_{D,i}) = \alpha_0 + \alpha_1 T_{i,i} + \alpha_2 T_{i-1,i} + \ldots + \alpha_{i-1} T_{i-1,i} + \alpha_i T_{i-2,i} + \ldots + \alpha_{n} T_{n,i-2} \] (4.10)

Open-loop dynamic data was generated using a dynamic simulator and noise was added to the measurements as noted previously. The reference set consists of 15 dynamic responses to step changes in all the manipulated and disturbance variables, both individually and in combination. The model now has 45 X-block variables (15 temperatures at three time intervals) and the reference set contains 1000 points. Using dynamic data expands the data set tremendously, since for every steady-state point an entire dynamic profile must be incorporated. The benefits of using an efficient method such as PLS become even more apparent with data set of this size.

The inferential model for the prediction of \( X_D \) using the dynamic data emphasizes the temperatures at the top of the column. Comparison of the open-loop model predictions, calculated from the dynamic model, with the actual distillate compositions for a feed flowrate change is given in Figure 4.22. Clearly, the dynamic response follows the true response much better than with the steady-state model (Figure 4.21a); however, the advantages of the advance warning of the disturbance is lost. Depending on the application, a better dynamic model may be more useful, for example, in designing feedforward controllers or in models in a model-predictive controller.

Using the steady-state model or the full dynamic model represent two extremes in the predictions of the dynamic response. It may be beneficial to calculate a compromise of these methods. This can be accomplished by calculating an inferential model which predicts the exit composition at some time \( i \) in the future, i.e. a step ahead prediction \( \log(\hat{X}_{D,i+i}) \). Figure 4.23 show results from the PLS models
Figure 4.22 Comparison of the predicted open-loop dynamic response from the dynamic PLS model under Martens and Naes scaling (dashed line) and the actual simulated response (solid line) for the key component in the distillate for the MAW column. The disturbance shown is a 10% increase in feed flowrate.

developed from dynamic data, the changes shown in this figure are analogous to those shown in Figure 4.21. The steady-state models used in Figure 4.23 are based on the final point in the dynamic simulations and are not the same as the steady-state models used in Figure 4.21, which are based on 76 steady-state values. Figure 4.23a shows that as the step ahead parameter \(i\) is increased the emphasis on temperatures at the top of the distillation column decreases and the response obtained from the model resembles that obtained from the steady-state model.

The varied forms of the predicted dynamic responses, from smooth early prediction to non-minimum phase type responses, obtained using inferential models based on multiple temperatures indicate that the result is highly dependent on the type of disturbance entering the system and the dominant temperatures in the model. The varied forms of the responses for the prediction of \(X_d\) mean that it would be difficult to use these models in a model based controller for the MAW column.
Figure 4.23 Prediction from various inferential model for the distillate composition of the MAW distillation column. The plots shows that as the dynamic model predicts further into the future the dynamic response of the prediction tends to that of the steady-state model. The following disturbances are shown: a) a 10% increase in feed flowrate, b) a change in the feed composition from 30/66/4 to 34/62/4, and c) a 15 degree increase in the solvent temperature.

Conversely, the models for the prediction of $X_D$ provide a prediction which closely matches the true response and thus, could easily be used in a model based controller.
In a cascade type control scheme the prediction of $X_B$ would improve the control only if there was significant sampling time associated with the composition measurement but would presumably be more reliable and less costly than an on-line GC. (In all the plots presented in this section the actual measurements are plotted with the same sampling time as the temperature measurements.)

Although the inferential models would be difficult to use in a model based controller, could these predictions be used as the controlled variable in a simple proportional controller? The results of using the inferential model in the control of $X_B$ would depend on the model used but as the steady-state models are the easiest to obtain, and thus most likely to be implemented; these have been used to investigate the closed loop responses when using a proportional controller to adjust the solvent flow rate to the MAW tower. (Latosinsky (1988) determined that using the solvent flow as the manipulated variable provided better closed loop performance, than using reflux flow for this column.) The dynamic responses to feed flowrate and feed composition disturbances and a setpoint change in the distillate composition are given in Figure 4.24. The controlled variable in these simulations is the key distillate composition. These plots compare the responses obtained when the PLS model prediction is the controlled variable and when the actual measured composition is used as the controlled variable in the control scheme. The response of the inferential model based control is well behaved and represents reasonable dynamic performance. Although no general conclusions can be made regarding the use of a steady-state inferential model for controlling the dynamic plant and the dynamics of each system must be evaluated individually, these results are encouraging.
In the previous section one of the conclusions obtained from the examining the BTX column was that the open-loop inferential model will give poor results in the closed loop implementation. Yet, with the MAW column the open-loop model was used effectively in the control scheme just mentioned. A closed-loop steady-state model was also developed for this column and the regression coefficients are compared with the open-loop model in Figure 4.25. This plot shows that there is only a small difference in the two inferential models. The heaviest weight is still on the tray 2 temperature. This is also true in the dynamic PLS model. However, the general conclusion from the last section that the model should be developed from the same data as the final implementation is still a sound rule to follow, but there are times when the difference may be small.
Figure 4.24 Comparison of the closed-loop dynamic responses obtained for the MAW column using the open-loop steady state PLS model (dashed line) and using the actual controlled variable in the control scheme, to the following manipulations: a) a change in the feed flowrate, b) a change in the feed composition and c) a setpoint change to the distillate composition.
Figure 4.25 Comparison of the PLS model coefficients obtained using open and closed loop steady state data for the MAW column.

4.9 Conclusions

A procedure for developing inferential process models using PLS has been outlined. It has been shown to be a very powerful approach to building such models when there are large numbers of highly correlated measured variables. By retaining all the measurements without overfitting the data PLS is able to utilize all the information obtained from the process measurements. It has also been shown that PLS models will be extremely robust to missing data and recognized sensor failures, an important feature of any inferential control scheme.
The importance of obtaining a representative reference data set for the process to build these models has also been demonstrated. In particular it is important that the data contain a representative sample of the effects of all the manipulated variables and the process disturbances, and that the measurement noise levels be typical of those that will be encountered in the final control scheme. Furthermore, it was shown that the data must be collected while the process is operating under a feedback structure very similar to the one that is intended to be used. An iterative procedure was suggested to accomplish this. These points were illustrated through the development of PLS inferential models for two simulated distillation columns.

The preliminary investigation of using inferential models based on multiple sensors located throughout the plant showed that the form of the predicted response depended on the combination of the disturbance entering the system and the location of the dominant sensors in the model. Before addressing the question of how can the predicted response be made more uniform it is important to define the purpose of the model. If the model it to be used in model based controllers then the predictions should follow the true response, while for models used in cascade control schemes the objective function would be to obtain a fast indication of the effect of the disturbance on the controller variable. Further investigation of this area is necessary in order to evaluate the potential of multi-sensor inferential models. The results from the predictions of $X_g$ indicate that one way to restrict the model prediction to correlate with the actual response of the controlled variable is to restrict the location of the sensors to the vicinity of the controlled variable. A more difficult task will be to define a general strategy for choosing sensors which will provide early warning of disturbances but provide monotonic responses in the prediction. It has also been
shown that scaling can be used to influence the importance of various measurements in the final model and this ability may be used to advantage in this problem. Finally, the ability to use dynamic data, steady-state data and any combination by using the step ahead parameter (i) provides flexibility to the modelling procedure.

4.10 Notation

A - the optimal number of latent vectors
α - model coefficients
E - the residual matrix of X
F_i - flow of stream i
F - the residual matrix of Y
I - the identity matrix
k - the number of process/operational/input measurements
m - the number of quality/output measurements
n - the number of observations
p_a - the loading vector for the a^{th} latent vector of X calculated by PCA or PLS
P - the loading matrix for X calculated by PCA or PLS
p_i - the loading for the i^{th} X variable
q_a - loading vector for a^{th} latent vector of Y
Q - loading matrix for Y
S_D - estimate of the standard deviation of the measurement
S_E - estimate of the standard deviation of the noise
t_a - the latent variable for X associated with the a^{th} PLS dimension, the elements of this vector are the scores for each observation
\( T \) - the matrix of scores associated with \( X \)

\( T_i \) - the \( i^{th} \) temperature measurement

\( u_a \) - the latent variable for \( Y \) associated with the \( a^{th} \) PLS dimension, the elements of this vector are the scores for each observation

\( U \) - the matrix of scores associated with \( Y \)

\( w_a \) - the loading vector used for prediction, calculated from PLS, for the \( a^{th} \) latent vector

\( W \) - the matrix of prediction loadings for \( X \)

\( W_E \) - weights used in scaling

\( w_i \) - the prediction loading for the \( i^{th} \) \( X \) variable

\( X \) - (\( n \times k \)) matrix of the process measurements

\( x \) - (1 x \( k \)) vector of the process measurements from one observation

\( X_B \) - the composition of the light key in the bottoms product

\( X_D \) - the composition of the heavy key in the distillate product

\( Y \) - (\( n \times m \)) matrix of the product quality measurements

\( y \) - (1 x \( m \)) vector of the product quality measurements from one observation

**Superscript**

\( T \) - the transpose of the matrix

\( ^\wedge \) - the prediction of the variable

**Abbreviations**

BTX - Benzene, Toluene and Xylene distillation column

MAW - Methanol, Acetone and Water distillation column

MLR - Multiple Linear Regression

NIPALS - Nonlinear Iterative Partial Least Squares
PCA - Principal Component Analysis

PLS - Partial Least Squares or Projection to Latent Structures

PRESS - Prediction Error Sum of Squares

SSE - Sum of Squared Error

SVD - Singular Value Decomposition

4.11 References


Chapter 5. Dynamic Model Identification

The PLS method described in Chapter 2 is a linear method for modelling the relationship between a matrix of correlated inputs and a matrix of outputs. This method was successfully applied to steady-state applications in Chapters 3 and 4. In Chapter 4 the input matrix was augmented with past values of the measurements to capture the dynamic characteristics of the problem. In this chapter this approach is applied to the identification of dynamic process models to be used in model based controllers.

5.1 Introduction

Identification of dynamic models is an integral part of the application of process control. Many different methods have been used including: i) graphical techniques, e.g. process reaction curve method (Stephanopoulos, 1984); ii) non-parametric methods, e.g. frequency domain methods (Jenkins and Watts, 1969); and iii) parametric methods, both parsimonious and non-parsimonious (Ljung, 1987). A great deal of effort has been expended on this topic and the progress to 1971 has been reviewed by Astrom and Eykhoff (1971); since then hundreds of articles have been published on this topic, as well as many books, e.g. Box and Jenkins (1970), Astrom (1970) and Ljung (1987). The methods used in practice depend on the demands of the individual identification problem and all of the above methods are still being used. Improvements in the designs of perturbation sequences, the advent of cheap computing power and the implementation of model based and multivariate controllers has also influenced the choice of identification methods, particularly
increasing the popularity of parametric methods.

Identification procedures based on maximum likelihood estimation or general prediction error methods using parsimonious models (e.g., Box and Jenkins models) work extremely well on processes when the input variations are well designed, when the number of input variables is not too large, and when the process can be described by low order models. The successful application of these methods in the process industries is widespread. However, in many model predictive control schemes being used today (e.g., Dynamic Matrix Control (DMC) (Cutler and Ramaker, 1976)) the number of input and output variables is very large (applications involving 20 or more inputs have been reported) and non-parsimonious finite impulse or step response models are being used (involving typically 30 or more parameters relating each input to each output). The data sets used in these multivariate identifications involve many highly correlated input variables with designed perturbations in only a few at a time (typically one at a time). Impulse and step response models are used not only because they fit naturally into the final model predictive control algorithms, but because the multivariable control scheme is usually being applied around a large section of the process within which there exist many lower level controllers and recycle loops. In this situation low order transfer function models characteristic of most open-loop processes are often not adequate and the selection of appropriate parsimonious model structures necessary for Box and Jenkins identification is not easy.

Two popular non-parsimonious model structures are: i) the Finite Impulse Response (FIR) model which models the relationship between the present and past inputs to the output; and ii) the autoregressive exogeneous variable (ARX) model
which includes fewer past input values but adds past output values to the model. Both of these structures are highly over-parameterized and they are limited in the choice of noise models, which causes some difficulties in the parameter estimation but they have some advantages over the parsimonious Box and Jenkins (1970) models (discussed in section 5.2 on model structures).

The variables in these non-parsimonious models are the lagged inputs (and lagged outputs), and are highly correlated. The difficulty in applying ordinary least squares in this situation is that with limited and poorly designed data sets the parameter estimates obtained have large variances and depend strongly on the data set. The model over-parameterization and correlation among the variables makes this an ideal application for PLS which obtains more stable estimates of the model parameters (i.e. estimates with smaller variances). PLS is, however, only one of the methods which can be used for parameter estimation using these model; an alternative approach to stabilizing the parameter estimates is to use constrained least squares or regularization methods (RR). The regularization methods are extensions of the least squares method with improved mean square error and smoothness properties, and allow for the estimation of a much larger number of parameters in ill-conditioned multivariable situations.

Simulation studies of a multiple input single output (MISO) system are used to compare the results obtained for the two non-parsimonious model structures using PLS, RR and LS. The transfer functions and noise models for this process were chosen to represent typical problems faced in industry. To evaluate the estimation methods several identifications are performed using data with varying quantities and types of information. The estimated parameter results are compared to the true
model parameters. The results are analyzed using information regarding the limiting properties of the estimation methods to determine the suitability of the identification procedures. The results are also compared to those obtained using the parsimonious transfer function models and maximum likelihood estimation in two cases. The goal of these simulation studies is three fold. The first is to investigate the applicability of PLS regression for dynamic model identification and the second is to compare the results with the more common methods to determine if the more complex PLS calculations provide significant benefits.

One of the perceived advantages of applying non-parsimonious models instead of the iterative procedure necessary for the parsimonious Box and Jenkins identification is the simplicity of the methods. Once the data has been collected and the order of the FIR or ARX models chosen, the model parameters can be calculated directly. This black box approach makes these models attractive to the process engineers for model identification because little expertise in identification is needed. The third goal of the simulation studies is to show the limitations of the FIR and ARX model structures, especially for processes with autocorrelated disturbances.

Preliminary results from this investigation were presented at the IFAC Conference on Identification (MacGregor et al., 1991) and these are largely contained in Section 5.4. The application of PLS techniques to the estimation of non-parsimonious models for single-input single-output (SISO) systems has also been reported by Ricker (1988) and Wise and Ricker (1990). They report that PLS is able to handle the over-parameterized FIR identification better that SVD or least squares but Ricker (1988) indicates that for such simple systems the time series methods of Box and Jenkins (1970) are more efficient.
The RR and LS methods are only applicable in single output systems; however, in many multivariate applications several output variables are to be controlled. The second part of this chapter investigates whether the multivariate PLS2 algorithm provides better estimates for multiple input multiple output (MIMO) systems, than RR or PLS1, by considering all the outputs as a group.

5.2 Model Structures

In process control the effort of developing a process model from theory and verifying it with dynamic data is often not justified because a linear input output model is sufficient for controller design and tuning. Parametric process models can be identified in two forms: i) parsimonious, transfer function models or ii) non-parsimonious, impulse or step response models.

In the simulation studies the data were generated using transfer function models (Table 5.1, section 5.4 and equation 5.39, section 5.5); however, the results of the non-parsimonious estimation are compared on the basis of impulse response functions and step response functions. In order to facilitate comparison of results the transfer function models used in the simulation can be represented as the following finite impulse functions:

\[
y(t) = G_o(z^{-1})u(t) + D_o(t)
\]

for MISO processes and

\[
y(t) = G_o(z^{-1})u(t) + D_o(t)
\]

for MIMO processes, where \(z^{-1}\) is the backward shift operator. \(D_o = H(z^{-1})a_t\) is the unmeasured disturbance or the noise in the process and \(a_t\) is a white noise sequence. \(G_{oj}(z^{-1})\) is a polynomial expression in \(z^{-1}\), the impulse transfer function, for the true
deterministic relationship between input $i$ and output $j$. These are defined as the true models of the process and are used as a basis of comparison for the identification methods.

In model predictive control it is common to use non-parsimonious dynamic models in the impulse or step response form. A finite impulse response (FIR) model is of the form:

$$y_j(t) = \sum_{i=1}^{m} v_{ij}(z^{-1})u_i(t) + e_j(t) \quad (5.3)$$

where $m$ is the number of input variables and

$$v_{ij}(z^{-1}) = v_{io} + v_{i1}z^{-1} + \ldots + v_{ir}z^{-r} \quad (5.4)$$

is the finite impulse response model for the $i$-th input, the estimate of $G_{oi}$. The truncation point $r$ is taken to be greater than the settling time of the process (i.e. the time necessary for the process to reach steady-state following a disturbance). In practice the truncation point must be selected iteratively, adjusting this value so that all the trailing $v_{ir}$'s are small compared to their confidence intervals. The step weights can be calculated, for each input, as the cumulative sum of the corresponding impulse weights. The step weight of input $i$ at lag $k$, $S_{ik}$ is given as:

$$S_{ik} = \sum_{l=0}^{k} v_{il} \quad (5.5)$$

The number of parameters in FIR models is very large ($p=m(r+1)$) since the number of lags used is usually in the order of $r=30$ to 50. If the true underlying process is well approximated by a low order transfer function model, e.g. first order with only 2 parameters, then these impulse and step response models are over-
parameterized (i.e., non-parsimonious). This leads to ill conditioned parameter estimation situations, especially if the number of inputs \((m)\) is also large and the process identification experiments were not well designed.

Alternatively the autoregressive with exogeneous variable (ARX) model structures can be used.

\[ A(z^{-1}) Y_t = \sum_{i=1}^{m} B_i(z^{-1}) u_t + e_t \]  

(5.6)

In this form \(A(z^{-1}) = 1 + a_1 z^{-1} + \ldots + a_k z^{-k}\) is usually taken to be of fairly low order and the \(B_i(z^{-1}) = b_{i0} + b_{i1} z^{-1} + \ldots + b_{ik} z^{-k}\) are usually taken to be of sufficiently high order to include a few lags beyond the longest dead-time in the process. While the ARX model is still non-parsimonious fewer parameters are usually required than are used in FIR models.

One of the advantages of both the FIR and ARX model forms is that they are linear, and linear regression methods can be used in the estimation. Further, these model structures can represent any linear process, regardless of the complexity, as long as the truncation points are sufficiently large. The disadvantage is the large number of parameters to be estimated. The combination of a FIR model structure and LS estimation results in the assumption of a white noise model for the disturbance, limiting their applicability. When using ARX models there exists some flexibility in modelling correlated disturbances; however, the denominator of the process model and the noise model must be the same order which causes some difficulties, which are discussed in section 5.4.
In contrast the transfer function models (Box and Jenkins, 1970)

\[ y_t = \sum_{i=1}^n \frac{\omega_i(z^{-1})z^{-b_i}}{1 - \delta_i(z^{-1})} u_t + \frac{C(z^{-1})}{F(z^{-1})^p} a_t \]  \hspace{1cm} (5.7)

where: \( z^{-1} \) is the backwards shift operator;

\( b_i \) is the deadtime associated with input \( i \);

\( d \) is the order of differencing necessary for stationarity;

and \( a_t \) is a white noise sequence

allows independent and parsimonious parameterization of the process and the noise models; however, identification of the model in this form requires an iterative technique to determine the correct model orders (ie. order of the polynomials \( \omega_i, \delta_i, C \) and \( F \)) and a non-linear parameter estimation method. Determining the correct model order for the process and noise in complex or multivariate systems can be difficult, requiring significant experience to achieve an accurate identification.

Excellent methods for identifying transfer function models for low order systems are well established, the goal of this work is to investigate dynamic model identification of complex and multivariate systems using non-parsimonious models. Non-parsimonious models will be identified in both the FIR and ARX model forms to illustrate the differences in these models. The results of the non-parsimonious identification will be compared to parsimonious estimation of transfer function models obtained using the Identification toolbox for MATLAB to illustrate the differences between the two approaches.
5.3 Identification Methods

Three methods will be used to estimate process models for both the FIR and ARX model forms: i) linear least squares (LS); ii) Regularization methods (RR) and iii) PLS. Many excellent references are available for LS, including Draper and Smith (1981); however, it is still useful to include a general overview and the limiting properties of this method as these will be used in the interpretation of the results. The PLS method has been discussed in detail in Chapter 2 and will not be discussed here. In subsection 5.3.2 the RR method is introduced as one of the constrained least squares methods.

5.3.1 Properties of Least Squares Estimates

The identification methods (RR and PLS) used in this chapter are modifications of the Least Squares approach in that they use the same model structure and a modified least squares criterion. In order to understand the results obtained in the following sections it is important to establish the properties of Least Squares estimation. The theory presented here follows the development shown in Ljung (1987). The models presented in this section are for MISO processes where the true model can be represented by equation 5.1.

The model of the system used in the regression is

\[ y(t) = X^T(t)\theta + N(t) \]  \hspace{1cm} (5.8)

N(t) is an estimate of D_o(t); however, for the LS estimates in this section, N(t) is taken to be white noise. Thus

\[ \hat{y}(t) = X^T(t)\hat{\theta} \]  \hspace{1cm} (5.9)
where $y(t)$ is the prediction of the output, $y$, at time $t$, given the estimated regression vector $\hat{\theta}$. When estimating the model in the FIR format, $X$ will contain only the present and past input terms and the regression coefficients $\hat{\theta}$ are the impulse weights $\nu_i$, the estimates of the true impulse weights $G_{oi}$. However, when estimating an ARX model, $X$ will include the present and past inputs as well as the past output terms, and $\theta$ is a vector of the parameters of $B(z^{-1})$ and $A(z^{-1})$. The results of this estimation can be compared to that obtained using the FIR model by dividing $A(z^{-1})$ into $B(z^{-1})$ to obtain estimates of the true impulse weights $G_{oi}$.

Knowing the values of the output variable, $y(t)$, and the estimates of the parameters $\hat{\theta}$, the prediction error can be expressed as

$$e(t, \theta) = y(t) - X^T(t)\hat{\theta}$$

(5.10)

The object of least squares regression is to choose $\theta$ such that the second norm of the prediction error is minimized. Formally this can be expressed as the least squares criterion

$$V_n(\theta, Z^n) = (y - X\theta)^T(y - X\theta)$$

(5.11)

The minimum of this function is found as

$$\frac{\partial}{\partial \theta}[(y - X\theta)^T(y - X\theta)] = 0$$

(5.12)

This is a quadratic function in $\theta$ and thus the analytical solution to this minimization is the familiar normal equations for least squares

$$\hat{\theta}_{n}^{LS} = \arg\min_\theta V_n(\theta, Z^n) = [X^TX]^{-1}X^Ty$$

(5.13)
The result obtained will depend in the sequence of data, $Z^N (= y_N, y_{N-1}, \ldots; u_{1,N}, u_{1,N-1}, \ldots; u_{2,N}, u_{2,N-1}, \ldots)$, which is available; however, as $N$, the number of data points approaches infinity some limiting properties of the least squares estimate can be obtained.

Suppose that the data were generated by

$$y(t) = X(t)\theta_0 + D_o(t)$$

(5.14)

Substituting the value $y(t)$ in equation 5.18 a relationship between the estimated parameters and the true parameters can be developed.

$$\hat{\theta}_o^{LS} = (X^TX)^{-1}X^T(X\theta_0 + D_o)$$

(5.15)

$$= \theta_0 + (X^TX)^{-1}X^TD_o$$

(5.16)

The desired properties would be for

1. $\hat{\theta}_o^{LS}$ be close to $\theta_0$
2. $\hat{\theta}_o^{LS}$ to converge to $\theta_0$ as $N \to \infty$

If $D_o(t)$ is small then $\hat{\theta}_o^{LS}$ will be close to $\theta_0$ and the first condition will be satisfied.

In order to satisfy the second condition

$$(X^TX)^{-1}X^TD_o(t) \to 0 \quad as \quad N \to \infty$$

(5.17)

In order for this to be true two conditions must be satisfied

i) $X^TX$ must be non-singular

ii) $X^TD_o \to 0 \quad as \quad N \to \infty$. This will be true if either

a) $D_o(t)$ is a sequence of independent random variables with zero mean

(white noise).
b) the past input sequences in $X(t)$ are independent of the noise sequence $D_o(t)$.

Under either of these conditions $E(X(t)D_o(t)) = 0$.

Stating this another way is to say that the terms in $X$ and the noise sequence $D_o(t)$ must be independent. This is always the case if $D_o(t)$ is a white noise sequence. It is also true if the perturbations to the process $u(t)$ and the noise are independent, and the FIR model form (which contains no past output values) is used. However, if the noise is structured (i.e., related in time) and the ARX formulation (which contains past output values) is used then $X$ and $D_o(t)$ are not independent, thus $X^T D_o = 0$ as $N \to \infty$. There will be a bias in the estimate of $\theta^{LS}$. Again we emphasize that these limits apply only as $N$ approaches infinity; the number of samples needed to approach these results in practice will depend on the signal to noise ratio and the model being estimated.

The least squares criterion in equation 5.16 can also be expressed in terms of the power spectrum of the variables. This alternative frequency domain formulation, presented below, will allow easier interpretation of how bias is introduced into the estimates when different model structures are used. This in turn will allow easier interpretation of the results presented in the following sections.

Following the development presented by Ljung (1987, Section 8.5) the least squares criterion can be expressed as

$$
\lim_{N \to \infty} V(\theta, Z^N) = \frac{1}{4\pi} \int_{-\infty}^{\infty} \phi_e(\omega, \theta) d\omega
$$

(5.18)

where $\phi_e(\omega, \theta)$ is the spectrum of the prediction errors, $e(t, \theta)$. Recalling the true process model
and the regression model which can be written in terms of the inputs

\[ y(t) = G(z^{-1})u(t) + D_o(t) \]  \hspace{1cm} (5.19)

the prediction error can be expressed as

\[ e(t, \hat{\theta}) = H^{-1}(z^{-1}, \hat{\theta})[(G_o(z^{-1}) - G(z^{-1}, \hat{\theta}))u(t) + D_o(t)] \]  \hspace{1cm} (5.21)

and the spectrum of the prediction error can be written as

\[ \phi_e(\omega, \hat{\theta}) = \frac{|G_o(e^{j\omega}) - G(e^{j\omega}, \hat{\theta})|^2 \phi_u(\omega) + \phi_d(\omega)}{|H(e^{j\omega}, \hat{\theta})|^2} \]  \hspace{1cm} (5.22)

The least square criterion can then be written as

\[ \hat{\theta}_N^{LS} = \arg \min_{\theta} \frac{1}{4\pi} \int_{-\pi}^{\pi} \left[ |G_o(e^{j\omega}) - G(e^{j\omega}, \hat{\theta})|^2 \phi_u(\omega) + \phi_d(\omega) \right] \frac{1}{|H(e^{j\omega}, \hat{\theta})|^2} \omega d\omega \]  \hspace{1cm} (5.23)

This formulation allows the noise model to depend on the regression parameters, but can be simplified if a fixed noise model (i.e. one that does not depend on \( \theta \), \( H(e^{j\omega}, \hat{\theta}) \to H(e^{j\omega}) \)) is assumed. In this case equation 5.23 becomes

\[ \hat{\theta}_N^{LS} = \arg \min_{\theta} \frac{1}{4\pi} \int_{-\pi}^{\pi} \left[ |G_o(e^{j\omega}) - G(e^{j\omega}, \hat{\theta})|^2 \phi_u(\omega) \right] \frac{1}{|H(e^{j\omega})|^2} \omega d\omega \]  \hspace{1cm} (5.24)

If a FIR model structure is being identified and a fixed noise model (e.g. white noise) is being assumed, then equation 5.24 is the expression for the least squares criterion in the frequency domain. Inspection of this expression shows that; \( |G_o(e^{j\omega}) - G(e^{j\omega}, \hat{\theta})|^2 \) is the difference in amplitudes of the true and estimated process. This term is multiplied by the spectrum of the input signal divided by the spectrum of the noise model. The spectrum of the white noise sequence is 1.0, therefore only the spectrum of the input signal determines the relative weighting on the estimates
of amplitudes. Thus, the model will provide a better estimate in the frequency range which contains the most information, i.e., the low frequency range and the FIR model will tend to accurately identify low frequency components (e.g., the steady-state gain) at the expense of the high frequency components. However, when the ARX model form is used the noise model, \( N(t) = H(z^{-1}, \theta)c(t) = A(z^{-1}, \theta)^{-1}e(t) \), will depend upon the parameters \( \theta \), thus

\[
\frac{1}{|H(e^{j\omega}, \theta)|^2} = |A(e^{j\omega}, \theta)|^2 = |1 + a_1 e^{j\omega} + a_2 e^{2j\omega} + \ldots + a_p e^{pj\omega}|^2
\]

(5.25)

and equation 5.23 is appropriate. Since the AR term, \( |A(z^{-1})|^2 \), will in general have high gains at high frequencies the least squares fit of the ARX model will emphasizes the high frequency components of the model (for an example see Ljung, pg 227), thereby leading to poor estimates of the steady-state gain.

### 5.3.2 Constrained Least Squares Approach

Ordinary least squares regression leads to unbiased parameter estimates if the errors, \( e_n \), are independent of the elements of \( X \). However, if the disturbances \( D_i \) in equations 5.1 and 5.2 are autocorrelated then LS will not give unbiased estimates. Furthermore, if the regressor variables in \( X \) are not independent, but exhibit a high degree of correlation, the least squares parameter estimates will have very large variances. Since the elements of \( X \) for models (5.3) through (5.6) are simply lagged values of the inputs and outputs, this correlation among the parameters is a major problem in fitting these models. In these cases we use biased forms of estimation as alternatives to the least squares procedure. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in the stability of
the regression coefficients (i.e. regression coefficients will be similar for similar data sets). PLS is a biased estimator. By retaining that portion of the information in the data which is relevant to the prediction of the output variables the biased model will be better than the unbiased model in a mean squared error sense.

Another class of biased estimators considered here is usually referred to as regularization methods. They satisfy the following constrained least squares criterion

\[ V_N(\theta, Z^N) = [(y - X\theta)'(y - X\theta) + \gamma q(\theta)'q(\theta)] \]  

(5.26)

where \( \gamma \) is a scalar and \( q(\theta) \) is a vector valued function of the parameters \( \theta \). The most common choices for the parameter constraint terms \( q(\theta)'q(\theta) \) in (5.26) are:

\[ \theta'\theta = \theta' I \theta \]  

(5.27)

\[ \nabla \theta' \nabla \theta = \theta' K_1 \theta \]  

(5.28)

\[ \nabla^2 \theta' \nabla^2 \theta = \theta' K_2 \theta \]  

(5.29)

In equations 5.9 and 5.10 the matrices \( K_1 \) and \( K_2 \) are given by:

\[
K_1 = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 & 0 & 0 \\
-1 & 1 & 0 & \ldots & 0 & 0 & 0 \\
0 & -1 & 1 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & -1 & 1
\end{bmatrix}
\]

(5.30)

\[
K_2 = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 & 0 & 0 \\
-1 & 1 & 0 & \ldots & 0 & 0 & 0 \\
1 & -2 & 1 & \ldots & 0 & 0 & 0 \\
\ldots & \ldots & \ldots & \ddots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & 1 & -2 & 1
\end{bmatrix}
\]

Using equation 5.27 we obtain the well known Ridge Regression estimator (Hoerl and Kennard, 1970). In this estimator the length of the regression vector \( \theta \) is constrained. Using the constraint in equation 5.28 the size of changes in adjacent values of the parameter estimates are penalized. This form of regularization is particularly appropriate for the identification of step and impulse response models.
because it forces successive values to be similar, thereby smoothing the estimated impulse response plot. The last constraint form (equation 5.29) penalizes the rate of change in the parameter estimates forcing an even greater amount of smoothing. Equations 5.27 to 5.29 are all quadratic forms of the parameters \( \theta' \mathbf{H} \theta \) and the minimization of (5.26)

\[
\frac{\partial}{\partial \theta} [(y - X\theta)'(y - X\theta) + \gamma q(\theta)'q(\theta)] = 0
\]

(5.31)

is given by

\[
(X'X + \gamma \mathbf{H})\theta = X'y
\]

(5.32)

Due to the addition of the diagonally dominant \( \mathbf{H} \) matrix to the \( X'X \) matrix in equation 5.32, these regularization methods are much better conditioned problems. For many cases it can be shown that, with an appropriate choice of the ridge parameter \( \gamma \) the parameter estimates obtained from them have much smaller mean squared error than the least squares solution. Methods for selecting appropriate values of the ridge parameter \( \gamma \) are discussed in several sources (Hoerl and Kennard (1970), Myers (1989)). In general too low a value of \( \gamma \) yields estimates with low bias but large variances, while too large a value of \( \gamma \) yields overly smoothed and biased estimates.

When RR is applied in the FIR estimation the \( K \) is for all the inputs; for example, if 5 inputs with 35 parameters each are being identified then \( K \) will be a \((175 \times 175)\) matrix. Generally if models for several inputs are being identified, then there should be no constraint between the last parameter of one input and the first parameter of the next input. However, in this application this was not forced since the last lag of one input and the first lag of the next input should both be zero. RR of the forms presented in equations 5.28 and 5.29 were not used in the estimation
of the parameters in the ARX models because constraining the values of $A(z^1)$ in this way is not appropriate. Using RR of the form presented in equation 5.29 for the identification of FIR models was investigated but it was found that this form provided too strong constraint and could not accurately model the sudden change required at the first significant term in the finite impulse response. The RR results presented in sections 5.4 and 5.5 were obtained with the constraint equation 5.28.

5.4 Comparison of Identification Methods (MISO)

A multiple input, single output (MISO) system of the following form

$$Y_t = \sum_{i=1}^{5} \frac{\omega_i(z)}{1 - \delta_i(z)} u_{it} + D_t = \sum_{i=1}^{5} Z_i + D_t$$

(5.33)

was simulated to obtain data used to compare the estimation methods described in the previous section. The transfer functions were chosen to represent typical process dynamics and their equations, gains and settling times, the time necessary for the process to reach steady-state following a disturbance, are summarized in Table 5.1. For the base case identification the inputs to this system ($u_t$'s) were chosen to represent a good experimental design. Independent PRBS signals were applied to the first and third inputs and autoregressive signals representing normal process drifts were applied to the remaining inputs. The equations used, as well as the relative magnitudes for the signals, are given in Table 5.2. Three types of disturbance or noise models ($D_t$) were used in the simulations: a white noise sequence, an autoregressive noise sequence and an autoregressive non-stationary noise sequence. The data were generated as responses from individual processes ($Z_t$) which were then
summed with the noise sequence. This was done so that each individual process or
their combination could be studied and the true model for the system can be expressed
by equation 5.1.

The objective of this study was to investigate the ability of the various
identification methods to identify the impulse response functions for the MISO
process. The two non-parsimonious model structures of equations 5.3 and 5.6 were
used. In the FIR model (5.3) 35 impulse weights were estimated for each of the
inputs, this is sufficient to include all the response, since the settling time for each
input is less than 35 sampling intervals (see Table 5.1). This meant that 175
parameters were estimated in each identification. For the ARX model (5.6) the
present and 11 past inputs and 4 past outputs were employed resulting in a total of
64 parameters to be estimated. The results for each identification method and each
model structure were compared by computing the impulse and step weights from
the fitted models, and calculating the mean square error (MSE) - defined as the sum
of squares of the differences of the estimated step or impulse weights from the true
ones.

\[
MSE_{\text{impulse}} = \sum_{i=1}^{34} \sum_{t=0}^{34} (\hat{v}_{it} - v_{it})^2 \quad \text{and} \quad MSE_{\text{step}} = \sum_{i=1}^{5} \sum_{t=0}^{34} (\hat{S}_{it} - S_{it})^2 \quad (5.34)
\]

Note that each of these measures a different property of the response; \( MSE_{\text{impulse}} \)
emphasizes the accuracy of the initial response while the \( MSE_{\text{step}} \) concentrates on
the final gain of the process.

The PLS, LS and RR algorithms used in this analysis were programmed in
MATLAB - 386. When using PLS the data was autoscaled (mean centered and
scaled to unit variance) and the optimal number of components in PLS was deter-
mined by minimizing PRESS using cross-validation. The RR results for this section were obtained by T. Kourtis, and her contribution is acknowledged here. The regularization results are obtained by constraining the first differences between adjacent impulse weights (equation 5.28). When using RR the data remained unscaled and the magnitude of the diagonal terms in the $X^T X$ matrix were approximately 30 and for this scaling $\gamma = 10$ gave the results with the lowest prediction error.

Table 5.1 Summary of transfer functions used to simulate MISO system

<table>
<thead>
<tr>
<th>Input Variable Number</th>
<th>Transfer Function Equation</th>
<th>Steady State Gain</th>
<th>Settling Time (sampling periods)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Z_u = \frac{0.15z^{-4}}{1 - 0.85z^{-1}} u_t$</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>$Z_{2u} = \frac{0.045z^{-1}}{(1 - 0.85z^{-1})(1 - 0.7z^{-1})} u_2$</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>$Z_{3u} = \left[ \frac{-0.2z^{-1}}{1 - 0.8z^{-1}} + \frac{0.12z^{-5}}{(1 - 0.8z^{-1})(1 - 0.6z^{-1})} \right] u_3$</td>
<td>0.5</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>$Z_{4u} = 0.0 u_4$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>$Z_{5u} = \frac{0.3z^{-8}}{1 - 0.7z^{-1}} u_5$</td>
<td>1</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 5.2 Summary of Inputs, Disturbances and the Output signals

<table>
<thead>
<tr>
<th>Input 1</th>
<th>Equation</th>
<th>Variance of Input Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input 2</td>
<td>( u_2 = \frac{0.23}{1 - 0.85z^{-1}} a_2 )</td>
<td>0.181</td>
</tr>
<tr>
<td>Input 3</td>
<td>PRBS, switching interval 6 periods</td>
<td>1.0</td>
</tr>
<tr>
<td>Input 4</td>
<td>( u_4 = \frac{0.45}{(1 - 0.80z^{-1})} a_4 )</td>
<td>0.552</td>
</tr>
<tr>
<td>Input 5</td>
<td>( u_5 = \frac{0.75}{(1 - 0.75z^{-1})} a_5 )</td>
<td>1.4175</td>
</tr>
<tr>
<td>White noise</td>
<td>( D_t = a_t )</td>
<td>0.0104</td>
</tr>
<tr>
<td>Non-Stationary Noise</td>
<td>( D_t = \frac{1}{(1 - 0.99z^{-1})(1 - 0.4z^{-1})} a_t )</td>
<td>0.287</td>
</tr>
<tr>
<td>Autoregressive Noise</td>
<td>( D_t = \frac{1}{(1 - \alpha z^{-1})} a_t ), ( \alpha = 0.5, 0.7, 0.97 )</td>
<td>( \alpha_1 = 1.01 ), ( \alpha_2 = 1.08 ), ( \alpha_3 = 1.02 )</td>
</tr>
<tr>
<td>Output</td>
<td>( Y_t ) (without any noise)</td>
<td>2.2295</td>
</tr>
</tbody>
</table>
Figure 5.1 Base Case. Input sequences showing the first 300 sampling intervals.

a) Identification of models with white noise disturbances
Figure 5.1 Base Case. Output sequences showing the first 300 sampling intervals.
The base case is characterized by i) uncorrelated inputs with good excitation; ii) a white noise disturbance \((D_i=a_i)\); iii) PRBS in inputs 1 and 3 with switching intervals of 6 sampling periods, and autoregressive inputs in 2, 4 and 5; and iv) large amount of data, 850 sampling periods. The input and output sequences for the base case are presented in Figures 5.1 and 5.2, respectively.

The results of fitting the impulse response model (5.3) for 35 lags of each input are shown in Figures 5.3, 5.4 and 5.5 for the LS, RR and PLS methods, respectively. In these figures the impulse weights for the five transfer functions are plotted at the top, and the step weights at the bottom. The arrows show for each transfer function the impulse (or step) weight at lag=0. The true impulse and step responses are shown as dashed lines while the corresponding estimated responses are shown as solid lines joining each estimate (indicated by a *). The RR and PLS results were almost identical (see Table 5.3, where for each method and structure the MSE, and the variance of the residuals in the output, are given). For the base case the FIR model has the correct noise model and even though the process model is over-parameterized, all methods provide excellent estimates of the step responses. However, due to the over-parameterization, the LS estimates of the impulse response show a higher variability than those from RR and PLS.
Figure 5.3  Base Case. Impulse and Step Responses for FIR model estimated by LS. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.4  Base Case. Impulse and Step Responses for FIR model estimated by RR. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.5  Base Case. Impulse and Step Responses for FIR model estimated by PLS. True response - dashed line; estimated weights - *, joined by solid line.
Table 5.3 Base Case Results

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Model Structure</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
<th>Variance Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>FIR (35,0)</td>
<td>0.027</td>
<td>0.019</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.007</td>
<td>0.202</td>
<td>0.013</td>
</tr>
<tr>
<td>RR</td>
<td>FIR (35,0)</td>
<td>0.009</td>
<td>0.013</td>
<td>0.009</td>
</tr>
<tr>
<td>PLS</td>
<td>FIR (35,0)</td>
<td>0.024</td>
<td>0.029</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.016</td>
<td>0.394</td>
<td>0.016</td>
</tr>
</tbody>
</table>

The PLS estimates of the impulse and step responses obtained from fitting the ARX model (5.6) are shown in Figure 5.6 and are similar to those obtained from LS. The coefficients of an ARX model using the described RR method (i.e. constraining the first differences) were not calculated. The initial responses using either LS or PLS are well estimated, but poorer estimates of the steady state gain are obtained than when using an FIR model form. This is a typical result of using the popular ARX formulation to estimate impulse and step weights. The reason for this result can be explained by referring to the properties of least squares estimates. The properties below equation 5.17 do not hold for this ARX structure and hence even LS gives biased estimates. Furthermore, in the frequency domain equation 5.25 shows that when using the ARX form the past output terms force accurate prediction of the higher frequency components at the expense of the lower frequency components. Alternately this behavior can be explained by comparing the true model of the process and the model forced by the ARX formulation:

$$y_t = \frac{B(z^{-1})}{A(z^{-1})} u_t + \frac{1}{A(z^{-1})} a_t$$

(5.35)
The ARX model forces the disturbance model to be 4-th order autoregressive in nature rather than white noise, thereby introducing some bias into the estimates of the dynamic model.

When RR or LS is used the calculation is straightforward once the $\gamma$ is chosen for RR (see Kennard and Hoerl (1970) or Myers (1989)); however, when using PLS a further calculation must be performed to determine the appropriate number of latent variables to include in the model. As explained in Chapter 2 cross validation was the method used to make this determination. It was found that for situations such as the base case, where there was a large amount of very good data; the minimum in the PRESS is very broad and making the selection of the proper number of latent variables ambiguous (i.e. the cross validation criterion remains very close to 1.0 over a broad range of latent variables). This means that several different PLS models provide almost equivalent predictions of the output. Normally, when the prediction of $y$ is the ultimate goal; this would not cause a problem as any of the models, preferably the most parsimonious one, would be sufficient. In this application the object is to identify the true impulse weights, and since the problem is over-parameterized, several different combinations of impulse weights provide the same prediction of $y$. With each additional latent vector a greater proportion of the $X$ matrix information was used in the model; however, no clear trends were found as to which of the models in this broad range was the best; thus, the models shown in this work correspond to those indicated by cross-validation (i.e. stop when CSV $> 1.0$).
Figure 5.6 Base Case. Impulse and Step Responses corresponding to ARX model estimated by PLS. True response - dashed line; estimated weights - *, joined by solid line.
Another interesting feature of the PLS method when fitting the FIR model form was that a much higher dimension latent vector representation was necessary to adequately estimate the dead time when longer switching intervals for the PRBS inputs were used. This can be demonstrated by collecting a second data set with a PRBS switching interval of one and then refitting. Figure 5.7 compares the finite impulse responses for input 1 in the base case, (a), and input 1 in the data with a PRBS switching interval of 1, (b), for the second PLS dimension. With the fast switching interval an adequate estimate of the dead time is available after 2 PLS dimensions, while for the base case, 7 PLS dimensions are needed (Figure 5.5). More latent vectors were needed for a precise estimate of the dead time with slower switching because less information about the dead time is included in the data set. The data set with fast switching more observations are made in the range where the dead time affects the result; therefore, a precise estimate of the deadtime becomes more important to the prediction of y. Another major reason is that for a PRBS of 1 the elements of X are uncorrelated, while for an interval of 6 they are highly correlated. In this latter case PLS distributes the estimated parameters over the correlated inputs (MacGregor et al., 1991). By adding more PLS dimensions in the base case data set the correlation structure in the X space is explained and the dead time estimate becomes more accurate. In data sets with high signal to noise ratios the additional latent vectors needed to accurately estimate the dead time are easily calculated; however, in data sets with significant noise it may not be possible to calculate enough PLS dimensions without increasing the prediction error in y. Neither LS or RR suffers from this problem since they use all the data in X.
Figure 5.7 Comparison of the impulse weights obtained after 2 PLS dimensions for Input 1 for conditions as in the base case but with two different switching intervals in the PRBS signal: 1 sampling interval in (a) and 6 sampling intervals in (b).
The data collected for the base case contains large amounts of information and all the methods including least squares provide good estimates; however, this quality of information is rarely available except from simulations. In order to investigate the effects of less information another set of identifications using the FIR model structure were performed. A new data set was generated which contained the same transfer functions and noise as the base case but the number of observations was reduced to 200 and the level of excitation of the inputs was reduced to 1/3 of the that used in the base case (Table 5.4). The ability of all methods to obtain accurate responses is reduced with reduced information content (e.g. compare PLS results Figures 5.5 and 5.8); however, the LS estimates of the impulse weights are clearly the worst (Figure 5.9).

The low information content in this data set prevents accurate estimation using non-parsimonious models. More accurate estimates can be obtained by estimating fewer parameters. Thus using a parsimonious model to estimate the transfer function structure and parameters may be beneficial. Although additional effort (sometimes very difficult) is required to identify the model structures for each of the inputs, the resulting estimates are more accurate than using non-parsimonious models (see Table 5.4).
Figure 5.8 Reduced Data Set. Impulse and Step Responses for FIR model estimated by PLS. True resp. dashed line; estimated weights - *, joined by solid line.
Figure 5.9 Reduced Data Set. Impulse and Step Responses for FIR model estimated by LS. True response - dashed line; estimated weights - *, joined by solid line.
Table 5.4 Reduced Information Content

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Model Structure</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>FIR (35,0)</td>
<td>6.01</td>
<td>33.03</td>
</tr>
<tr>
<td>RR</td>
<td>FIR (35,0)</td>
<td>0.131</td>
<td>10.37</td>
</tr>
<tr>
<td>PLS</td>
<td>FIR (35,0)</td>
<td>0.191</td>
<td>7.57</td>
</tr>
<tr>
<td>parsimonious estimation</td>
<td>as in Table 5.1</td>
<td>0.034</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Another common situation when performing an identification study is for certain inputs to the process to be correlated; especially those inputs which are not being perturbed by the experimenter. In this study the autoregressive inputs represent inputs which are not controlled in the experiment but are varying due to variations in upstream processes and correlation among these variables was investigated. Another set of data were simulated with the same transfer functions, noise model and excitation as the base case but the five inputs were correlated. The correlation matrix for the five inputs were:

\[
\begin{pmatrix}
0 & u_1 & u_2 & u_3 & u_4 & u_5 \\
-0.12 & 1 & 0.16 & -0.16 & 0.27 \\
-0.27 & 0.93 & 0.036 \\
1 & -0.31 & 0.015 \\
1 & 0.001 \\
1 & 0.50
\end{pmatrix}
\]
The results of this case study are summarized in Table 5.5. The high correlation between \( u_2 \) and \( u_4 \) and the low signal to noise ratio for the second input, result in noisy estimates of the impulse weights (FIR model) for the second and fourth transfer functions when LS is used. However, the estimates of the gains and the settling times are comparable to the base case results. RR and PLS provide smooth impulse and step responses similar to those obtained for the uncorrelated base case. Again, the ARX models give poor gain estimates. (Table 5.5).

Table 5.5 Correlated Inputs with White Noise

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Model Structure</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
<th>Variance Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>FIR (35,0)</td>
<td>0.233</td>
<td>0.150</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.080</td>
<td>0.412</td>
<td>0.0124</td>
</tr>
<tr>
<td>RR</td>
<td>FIR (35,0)</td>
<td>0.015</td>
<td>0.084</td>
<td>0.009</td>
</tr>
<tr>
<td>PLS</td>
<td>FIR (35,0)</td>
<td>0.025</td>
<td>0.057</td>
<td>0.0116</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.022</td>
<td>0.509</td>
<td>0.0158</td>
</tr>
</tbody>
</table>

b) Models with Non-Stationary and Autocorrelated Noise.

The focus of the case studies will now change from processes where the disturbance is a white noise sequence to processes which include a non-stationary autocorrelated disturbance. The data used in this identification contains the same set of inputs and transfer functions as the original base case, but the noise added is the non-stationary noise described in Table 5.2. For this identification two approaches were used: i) fitting the raw data and ii) fitting the first differences of
the inputs and the first difference of the output. In both approaches all the methods provided a poorer estimate of the steady state gains due to the presence of the non-stationary noise.

i) Fitting the Raw Data. Identification of process models using FIR and either RR or PLS worked very well for situations with only white noise sequences for the disturbance, except for the low information case study. This is to be expected since the process models are linear and the assumed FIR noise model is white noise; however, when applying this model structure to processes which contain structured noise, especially non-stationary noise, the limitations of the white noise assumption become apparent. The noise structure assumed when FIR models are used can not account for the non-stationarity in the data; however, the non-stationarity is an important part of the response in the output and the estimation method will attempt to account for this variation through the model parameters. This is possible because the model is over-parameterized. The results of identifying the process models using raw data when non-stationary noise is present reflect this argument (see Table 5.6). The estimates obtained from RR and PLS for the FIR model are shown in Figure 5.10 and 5.11, respectively. The initial response of the models associated with inputs 1, 2, 3 and 5 are well identified by all the methods; however, the estimates of the steady state gain are poor. Also, the model for input 4 is grossly misestimated in all cases, which leads to the extremely high MSE for the step responses. The poor estimates of the gain can be explained in the following way: since the white noise structure assumed for the noise model does not correctly model the system, a bias in the estimates is obtained.
Figure 5.10 Non Stationary and Autocorrelated Noise. Raw Data. Impulse and Step Responses for FIR model estimated by RR. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.11 Non Stationary and Autocorrelated Noise. Raw Data. Impulse and Step Responses for FIR model estimated by PLS. True response - dashed line; estimated weights - *, joined by solid line.
The residuals with the FIR models using undifferenced data are very poor (see Table 5.6). FIR models force the estimator to explain the non-stationary variation of $Y_i$ with the process inputs. When utilizing the ARX models the residuals drop significantly, because these models allow for $Y_i$ variation that does not come from the inputs to be modeled by the noise term (see equation 5.35). However, the reduction in the variance of the residuals does not necessarily mean that the true dynamic model is well estimated. Although this model form does provide a noise model which can account for the non-stationarity in the output, the non-stationarity also affects the process model. Identification using the ARX model produced a better estimate of the process model for input 4, which accounts for the reduction in the MSE for the step response; however, for the remaining inputs the estimates of the steady state gains and the settling times were worse than those obtained from the FIR models. The reason for this bias in the steady-state gain can again be explained by referring to equation 5.24.

Table 5.6 Non-Stationary Noise - No Differencing of Data

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Model Structure</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
<th>Variance Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>FIR (35,0)</td>
<td>0.067</td>
<td>23.04</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.021</td>
<td>6.80</td>
<td>0.0026</td>
</tr>
<tr>
<td>RR</td>
<td>FIR (35,0)</td>
<td>0.059</td>
<td>22.95</td>
<td>0.228</td>
</tr>
<tr>
<td>PLS</td>
<td>FIR (35,0)</td>
<td>0.095</td>
<td>26.65</td>
<td>0.258</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.042</td>
<td>5.93</td>
<td>0.004</td>
</tr>
</tbody>
</table>
164

ii) Fitting the First Differences of the Inputs and Output. The non-stationarity in the disturbances accounts for the poor estimates of the process models. By differencing, both the input and the output data, the non-stationarity is removed and a model can be estimated using the transformed data. The results of this identification are summarized in Table 5.7, showing the improved estimates in all cases except the three dimensional PLS model (the reasons for this result are discussed in the following paragraph). The improvement in the estimates of the steady state gains for the FIR models is particularly evident (compare Figure 5.10 and 5.12 for RR and Figures 5.11 and 5.13 for PLS). By removing the non-stationarity from the data, the noise model assumed with the FIR model structure is a better estimate of the remaining structure of the noise in the data. The improvement for the ARX model is not as large because the ARX model was able to account for some of the non-stationarity in its structure when using undifferenced data.
Figure 5.12 Non Stationary and Autocorrelated Noise. Differenced Data. Impulse and Step Responses for FIR model estimated by RR. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.13  Non Stationary and Autocorrelated Noise. Differenced Data. Impulse and Step Responses for FIR model estimated by PLS (16 dimensions). True response - dashed line; estimated weights - *, joined by solid line.
Identifying a process model from differenced data using the PLS method and a FIR model shows another interesting point about the PLS method. Using this method and cross-validation to indicate the optimal number of latent vectors results in a model with 3 PLS dimensions and an extremely poor estimate of the process model (except for input 4 which is accurately estimated). By differencing the data to remove the non-stationary part of the noise some information about the steady state gains (i.e., the low frequency content) of the data is lost and the higher frequency information becomes more dominant, including the autoregressive noise in the output, thus the very poor estimate of the gain (Table 5.7). However, by calculating 16 PLS dimensions a very good estimate of the model is obtained! Consider the structure of the data in X the PRBS input which contained a series of steps has been converted by differencing to a few impulses, one each time the level in the PRBS changes. When PLS is applied to this data set it uses the dominant information in X to explain the dominant information in y; i.e., it uses the impulse type information in the X block to explain the higher frequency (initial responses) in the output. Before sufficient PLS dimensions are calculated to accurately estimate the steady state gain, cross-validation indicates a small increase in the PRESS. Insufficient number of PLS dimensions for accurate estimation of the steady-state gain are indicated due to the small amount of steady state information in y and the increased importance of the noise. If additional PLS dimensions are calculated, the increase in the PRESS is minor, and the small amount of information concerning the steady state gains is accessed and an accurate estimation of the process model is obtained (see Figure 5.13). By calculating the extra PLS dimensions a higher portion of the X block structure is explained and the model then becomes based on a wider frequency range.
of the inputs than the 3 dimensional PLS model which focussed primarily on the
high frequency range information. It is important to understand that calculating
extra PLS dimensions improves the estimate only if the information necessary for
the improvement is contained in the minor variations in the data. For example, in
the identification using undifferenced data, the poor estimate of the gain is due to
the inadequacy of the model structure and calculating extra PLS dimensions only
aggravates the problem.

Table 5.7 Non-Stationary Noise - Differenced Data

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Model Structure</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
<th>Variance Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>FIR (35,0)</td>
<td>0.0074</td>
<td>1.599</td>
<td>0.00246</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.0095</td>
<td>3.347</td>
<td>0.00267</td>
</tr>
<tr>
<td>RR</td>
<td>FIR (35,0)</td>
<td>0.0069</td>
<td>1.518</td>
<td>0.00261</td>
</tr>
<tr>
<td>PLS</td>
<td>FIR (35,0)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.0880</td>
<td>35.90</td>
<td>0.00667</td>
</tr>
<tr>
<td></td>
<td>FIR (35,0)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.0057</td>
<td>0.765</td>
<td>0.00247</td>
</tr>
<tr>
<td></td>
<td>ARX (12,4)</td>
<td>0.0122</td>
<td>3.894</td>
<td>0.00277</td>
</tr>
</tbody>
</table>

(a) 3 dimensional PLS model
(b) 16 dimensional PLS model

(c) Generalized Least Squares

The major difficulty in using the FIR or ARX formulations directly in the
identification is the inability to correctly model the noise, e.g. the white noise
assumption when using FIR and least squares estimation method. The ARX for-
mulation does provide a noise model, it is constrained to be the same as the
denominator of the process model, and this is rarely the correct form. Box and
Jenkins (1970) use the independently parameterized transfer function and disturbance models and a non-linear least squares method to overcome the problem of correctly modelling the structure of the process and the noise. The methods used in this work are linear therefore independently parameterized Box and Jenkins model cannot be used, but an alternative approach is to use generalized least squares (Goodwyn and Payne; 1977). In this method an independently parameterized autoregressive noise model which is independent of the denominator of the process model is used. The process model and noise models parameters are calculated separately and the algorithm iterates on these values till convergence.

The model form used in the GLS algorithm is the following

\[ A(z^{-1})y_t = B(z^{-1})u_t + H(z^{-1})e_t \]

and \( H(z^{-1}) = 1/R(z^{-1}) \). In the following calculations \( R(z^{-1}) \) is a second order polynomial. The present algorithm has been modified from the one presented by Goodwyn and Payne (1977) by replacing a LS estimation in step 3 by a PLS estimation.

1) Set \( \hat{R}(z^{-1}) = 1 \)
2) Form \( y_t^* = \hat{R}(z^{-1})y_t \) and \( u_t^* = \hat{R}(z^{-1})u_t \)
3) Use PLS to obtain estimates for the parameters in \( A \) and \( B \) (denoted by \( \hat{A} \) and \( \hat{B} \), respectively).
4) Calculate: \( e_t = \hat{A}(z^{-1})y_t - \hat{B}(z^{-1})u_t \)
5) Form the model \( R(z^{-1})e_t = a_t \) and obtain the least squares estimate of \( R \) (denoted by \( \hat{R} \)).
6) If converged - stop, otherwise go to 2).

This method is similar to the filtering approach suggested by Ljung (1987) except that the noise filter is calculated from the data rather than chosen by the user.
Using generalized least squares seems to answer the difficulties faced by the FIR and ARX models as it provides independent estimation of the process and the noise models. However, there are three potential difficulties: i) the method is iterative and for large problems the increase in computational time can be significant; ii) it does not directly address the problems encountered when using the ARX model (which needs moving average terms in the noise model to cancel the autoregressive terms forced on the noise model by this formulation); and iii) by estimating the noise and process models separately this method forces filtering of the data before calculating the final model parameters. Filtering works well if the noise and the model are in different frequency ranges; however, if the noise and the process have the same frequency range, as is often the case in chemical engineering, then identification is more difficult by also filtering some of the process information.

A SISO process simulation (input 1 in MISO case study) with different autoregressive noise models were used to investigate this algorithm. The input variation and the number of data points is the same as for the base case in the MISO identification but the noise level was increased in this simulation so that the signal to noise ratio was approximately 1. The results of this identification are summarized in Table 5.8. The estimates of the noise model for all these cases was very good; however, filtering the noise did not improve the estimate of the impulse or step responses. Figures 5.14 and 5.15 show the impulse responses, step responses and the original and estimated output values for the first 300 data points using FIR estimation with AR noise 2. The plots are the parameters after the first iteration (ie. no filtering, PLS model) and after the second iteration (ie. after convergence) of the GLS algorithm. The impulse response after filtering the noise provides a better
estimate of the initial response but is less smooth than for first iteration and the step response is nearly the same. The data sequences show how the filtering removes a large portion of the lower frequency variation. Further insight into this result can be obtained by examining equation 5.23. This equation shows that if an autoregressive noise model replaces a white noise model then the term $H(e^{iu})$ changes from a value of 1.0 to an expression (similar to the one for the ARX model shown in equation 5.25), which will produce a higher weighting on the high frequency components in the model. For AR3 the inverse of the noise model is a strong high pass filter and the $MSE_{step}$ experiences a dramatic increase (similar to PLS model with three dimensions for the differenced data and non-stationary noise).
Figure 5.14 Generalized Least Squares. Impulse and Step Responses and Predictions after the first iteration; for FIR model with AR2 estimated by PLS. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.15 Generalized Least Squares. Impulse and Step Responses and Predictions after the second iteration; for FIR model with AR2 estimated by PLS. True response - dashed line; estimated weights - *, joined by solid line.
Using the ARX model formulation combined with the GLS algorithm produced an interesting result. The impulse response between the PLS estimate (iteration 1) and the combined PLS and GLS estimate after convergence (iteration 3) experienced only a minor improvement and the final steady state gain error improved from 22% to 17%, but the values of the parameters estimated changed dramatically. Figure 5.16 shows the parameter estimates obtained from the first and final iterations of the GLS algorithm. The first 12 variables are the present and past input terms \((B(z^{-1}))\) and variables 13-16 are the past output variables \((A(z^{-1}))\). After the first iteration (the PLS estimates) this plot resembles the expected values for the ARX form, Figure 5.16a, (i.e. significant terms in \(A(z^{-1})\), variables 13-16). However, after the convergence of the GLS algorithm the past output terms in the model have very small values and the present and past inputs resemble FIR estimates, Figure 5.16b. By using GLS, the noise term filters the data to such an extent that the correlation between the present and past output terms is captured by the \(B(z^{-1})\) terms. This is yet another way of saying that the lower frequency information has been filtered from the data.

Application of the BJ model identification to this SISO problem is straightforward and excellent results are obtained for both the process and noise models.
Figure 5.16 Generalized Least Squares. Parameter Estimates for the ARX model with AR2 estimated by PLS; after the first iteration (a) and after convergence (b). The first 12 variables are the parameters in $B(z^{-1})$ and variables 13 - 16 are the parameters in $A(z^1)$. 
\[ y_t = \frac{0.16 \cdot z^{-1}}{1 - 0.847z^{-1}} u_t + \frac{1}{1 - 0.68z^{-1}} a_t \]  

(5.38)

The corresponding \( \text{MSE}_{\text{impulse}} \) and \( \text{MSE}_{\text{step}} \) for this model are 0.0003 and 0.065, respectively.

Table 5.8 Summary of generalized least squares results.

<table>
<thead>
<tr>
<th>Noise model</th>
<th>Following Iteration</th>
<th>PLS Dimension</th>
<th>( \text{MSE}_{\text{impulse}} )</th>
<th>( \text{MSE}_{\text{step}} )</th>
<th>Estimates of ( (1 - \phi_t z^{-1} - \phi_z z^{-1}) e_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR1 (=\frac{1}{1 - 0.5B} a_t )</td>
<td>1 (\cdot)</td>
<td>2 (\cdot)</td>
<td>0.0102</td>
<td>0.2295</td>
<td>0.505, -0.0267</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>0.0260</td>
<td>0.4005</td>
<td>0.501, -0.0167</td>
</tr>
<tr>
<td>AR2 (=\frac{1}{1 - 0.7B} a_t )</td>
<td>1 (\cdot)</td>
<td>3 (\cdot)</td>
<td>0.0084</td>
<td>0.1127</td>
<td>0.693, -0.0195</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>0.0156</td>
<td>0.1640</td>
<td>0.703, -0.0206</td>
</tr>
<tr>
<td>AR3 (=\frac{1}{1 - 0.97B} a_t )</td>
<td>1 (\cdot)</td>
<td>9 (\cdot)</td>
<td>0.0054</td>
<td>0.1359</td>
<td>0.967, 0.0342</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>0.0255</td>
<td>8.25</td>
<td>0.968, 0.0261</td>
</tr>
<tr>
<td>ARX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR2 (=\frac{1}{1 - 0.7B} a_t )</td>
<td>1 (\cdot)</td>
<td>3 (\cdot)</td>
<td>0.0115</td>
<td>0.7618</td>
<td>0.843, -0.128</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>0.0108</td>
<td>0.348</td>
<td>0.713, -0.018</td>
</tr>
<tr>
<td></td>
<td>3 (\cdot)</td>
<td>3</td>
<td>0.0102</td>
<td>0.340</td>
<td>0.705, -0.013</td>
</tr>
</tbody>
</table>

* indicates convergence

5.5 MIMO: Simulation and Identification

The results from the MISO case studies show that PLS and RR methods provide nearly equivalent results for systems with only one output; therefore the
additional computations involved with PLS do not seem justified. Now consider MIMO systems. If the outputs are independent then the problems can be treated as a series of MISO problems, and the conclusions from the previous section apply. Alternately, if the output variables are correlated then they should be considered together. Further, if the output variables are correlated then each output variable contains information about the others and if any of the outputs has a poor signal to noise ratio or outliers are present, an identification method that accounts for the relationship between these variables should be able to use the information in all the variables to obtain better estimates for the variable with poor signal to noise ratio or outliers. An example for this type of identification in steady-state was presented by Wold and Kettaneh-Wold (1989). The reported application contained 8 input variables and 11 quality variables. An experimental design was carried out to determine how the product could be improved. Unfortunately, because of the low information content of the individual quality variables no significant correlation between the inputs and the individual quality variables was found. Further investigation using PLS regression on the entire quality space determined a significant model in two latent variables was identified which could be used for prediction. The success of the PLS regression derived from the ability to model the correlation among the quality variables which contained the information necessary for the discrimination of a model.
The MIMO process simulated for this investigation is the following:

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    y_3
\end{bmatrix} = \begin{bmatrix}
    1.0 & 0.7 & 0.0 \\
    0.6 & 1.0 & 0.0 \\
    0.5 & 0.5 & 0.1
\end{bmatrix} \begin{bmatrix}
    u_1 \\
    u_2 \\
    u_3
\end{bmatrix} + \begin{bmatrix}
    c_1 \\
    c_2 \\
    c_3
\end{bmatrix} \begin{bmatrix}
    a_1 \\
    a_2 \\
    a_3
\end{bmatrix}
\]

\[y_2 = \frac{1z^{-1}}{1-0.8z^{-1}} \begin{bmatrix}
    0.6 & 1.0 & 0.0 \\
    0.5 & 0.5 & 0.1
\end{bmatrix} \begin{bmatrix}
    u_1 \\
    u_2 \\
    u_3
\end{bmatrix} + c_2 a_2 + c_3 a_3 \tag{5.39}\]

The identification has been simplified by considering a process where the dynamics and the deadtime for each process are the same and only the gains vary. Note that the third output is a linear combination of the first two outputs with a small additional contribution from each of the inputs, i.e.

\[y_3 = (y_1 + y_2) \frac{5}{1.05} + 0.05u_1 - 0.05u_2 + 0.10u_3 + N \tag{5.40}\]

Two cases were considered: i) the first considers a process where one of the variables has a poor signal to noise ratio and ii) the second process contains data corrupted by outliers. It was hypothesized that through the use of PLS2 and considering the outputs as a group, the estimates for the variable with the poor signal to noise ratio or outliers would be better that using RR or PLS1 on each output variable individually.

1) **Process with poor signal to noise ratio**

Data for this study were simulated using equation 5.39 with \( c_1 = 1.0, c_2 = 0.2 \) and \( c_3 = 0.1 \). A PRBS is used to perturb input 1 and autoregressive input sequences are used for inputs 2 and 3. The variances for the inputs, outputs and the noise sequences are given in Table 5.9. This identification is concerned with determining the models for the first output, \( y_1 \), which has a poor signal to noise ratio; however, output \( y_3 \) contains information regarding \( y_1 \) and has a better signal to noise ratio.
Comparison of the results obtained for $y_1$ are summarized in Table 5.10. Figures 5.17 to 5.19 show the results for each method for 400 data points. These plots show the true and estimated impulse and step weights for the transfer functions between each of the inputs and the first output. Both the plots and the MSE calculation show that there is an advantage to using PLS2 for the data set with fewer data points; however, for the case with 800 data points the results form all the methods are approximately the same.

Table 5.9 Variances Inputs, Outputs and White Noise for MIMO Simulation

<table>
<thead>
<tr>
<th></th>
<th>Variance of inputs</th>
<th>Variance of $a_t$</th>
<th>Variance of outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.46</td>
<td>21.7</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>1.71</td>
<td>15.4</td>
</tr>
<tr>
<td>3</td>
<td>0.9661</td>
<td>1.69</td>
<td>5.56</td>
</tr>
</tbody>
</table>
Figure 5.17 MIMO identification with poor signal to noise ratio. Impulse and Step Responses for FIR model estimated by PLS2 for the first output. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.18  MIMC identification with poor signal to noise ratio. Impulse and Step Responses for FIR model estimated by PLS1 for the first output. True response - dashed line; estimated weights - *, joined by solid line.
Figure 5.19 MIMO identification with poor signal to noise ratio. Impulse and Step Responses for FIR model estimated by RR for the first output. True response - dashed line; estimated weights - *, joined by solid line.
Table 5.10  Results from MIMO Identification for output with poor signal to noise ratio using FIR model structure.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Data set 1 (600 points)</th>
<th>Data set 2 (400 points)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{MSE}_{\text{impulse}}$</td>
<td>$\text{MSE}_{\text{Step}}$</td>
</tr>
<tr>
<td>PLS 2 (Multivariate)</td>
<td>0.5198</td>
<td>14.79</td>
</tr>
<tr>
<td>PLS 1</td>
<td>0.55</td>
<td>12.04</td>
</tr>
<tr>
<td>RR ($\gamma=10$)</td>
<td>0.70</td>
<td>14.37</td>
</tr>
</tbody>
</table>

**ii) The process with outliers**

This case study considers the same process as i) but the noise level on $y_1$ is lowered, $c_1=0.2$, and outliers, summarized in Table 5.11, are introduced. Again the results presented focus on the identification of the transfer functions associated with the first output. The outliers have a major effect on the results, Table 5.12, and again considering the entire data set using PLS2 shows an improvement in the MSE for both the impulse and step weights.

Another advantage of using PLS in this identification is that, at each PLS calculation, a plot of the data and the regression line between the scores for the $X$ matrix and the scores for the $Y$ matrix can be presented to the user. These plots show the scatter of the points around the line and indicate the strength of the correlation between latent vectors in $X$ and $Y$ at this PLS dimension. Outliers also become obvious when viewed on these plots. Once identified the outliers can be removed and the prediction greatly improved.
Table 5.10  Results from MIMO Identification for output with poor signal to noise ratio using FIR model structure.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Data set 1 (800 points)</th>
<th>Data set 2 (400 points)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{MSE}_\text{Impulse}$</td>
<td>$\text{MSE}_\text{Step}$</td>
</tr>
<tr>
<td>PLS 2 (Multivariate)</td>
<td>0.5198</td>
<td>14.79</td>
</tr>
<tr>
<td>PLS 1</td>
<td>0.55</td>
<td>12.04</td>
</tr>
<tr>
<td>RR ($\gamma=10$)</td>
<td>0.70</td>
<td>14.37</td>
</tr>
</tbody>
</table>

**ii) The process with outliers**

This case study considers the same process as i) but the noise level on $y_t$ is lowered, $c_1=0.2$, and outliers, summarized in Table 5.11, are introduced. Again the results presented focus on the identification of the transfer functions associated with the first output. The outliers have a major effect on the results, Table 5.12, and again considering the entire data set using PLS2 shows an improvement in the MSE for both the impulse and step weights.

Another advantage of using PLS in this identification is that, at each PLS calculation, a plot of the data and the regression line between the scores for the $X$ matrix and the scores for the $Y$ matrix can be presented to the user. These plots show the scatter of the points around the line and indicate the strength of the correlation between latent vectors in $X$ and $Y$ at this PLS dimension. Outliers also become obvious when viewed on these plots. Once identified the outliers can be removed and the prediction greatly improved.
Table 5.9 Outliers in the first output of the MIMO Simulation

<table>
<thead>
<tr>
<th>Points</th>
<th>Difference from base case</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-70</td>
<td>+4.0</td>
</tr>
<tr>
<td>200-220</td>
<td>-4.0</td>
</tr>
<tr>
<td>320</td>
<td>+8.0</td>
</tr>
<tr>
<td>370</td>
<td>+8.0</td>
</tr>
<tr>
<td>150</td>
<td>-9.0</td>
</tr>
</tbody>
</table>

Table 5.12 Results from MIMO Identification for output with outliers using FIR model structure.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>MSE Impulse</th>
<th>MSE Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS 2 (no outliers)</td>
<td>0.26</td>
<td>0.90</td>
</tr>
<tr>
<td>PLS 2</td>
<td>0.90</td>
<td>17.4</td>
</tr>
<tr>
<td>PLS 1</td>
<td>1.06</td>
<td>29.0</td>
</tr>
</tbody>
</table>

5.6 Conclusions

The comparison of PLS, RR and LS identification of non-parsimonious (FIR and ARX) models for MISO systems shows that the biased estimates obtained from PLS and RR provide smoother estimates of the impulse and step weights (i.e., with smaller variance from their true values). PLS and RR provide similar estimates but RR would be preferable for its simplicity. All methods suffer a reduction in accuracy with poorly designed data sets or significant noise. The advantage of using the FIR
and ARX models over the parsimonious Box and Jenkins models is that they can be used to model processes where low order models are not appropriate. These methods are also simpler to apply since there is no need to iteratively determine a parsimonious model order for each transfer function and are directly applicable in many model based controllers. There are, however, drawbacks to these models; since more parameters are being estimated, more information is necessary to obtain an estimate which is as accurate as that obtained using the parsimonious Box and Jenkins approach.

A further drawback to the application of any of these methods is the restrictive structure of the models that are linear in the parameters which prohibit an accurate characterization of the structured noise. Using a method such as PLS and an ARX model provides enough flexibility that accurate prediction of the output is possible; however, the model obtained does not reflect the true process parameters, but is rather, an empirical model which is only accurate as long as the process and noise structures do not change.

When using PLS in the identification the structure of the information content and the stopping criteria used must be considered. This work showed that for many of the case studies the minimum of the prediction error curve (as indicated by PRESS) is quite flat and that in some situations, i.e. to obtain a better estimate of the dead time and in order explain a higher proportion of the X block structure, it is beneficial to use a higher dimensional PLS model. Careful experimental design and pretreatment (e.g. differencing or scaling) is necessary because the PLS method extracts those parts of the model which dominate the information content in the data. If the data is noisy and the information relevant to the model is masked by other variation
in the data, then poor estimates can be obtained.

Use of multivariate PLS regression in the identification of MIMO processes did not show as large an improvement as was expected. Multivariate identification is beneficial when there are common parameters in the identification. Although correlation among the output variables was present the large number of latent variables meant that only a limited dimensional reduction (actually no dimension reduction in the output space) was achieved. The regression model was not adequately constrained to ignore the extraneous information contained in the noise and outliers.

5.7 Nomenclature

\( a_i \) - the \( i \)th white noise sequence (\( var = 1.0, \mu = 0 \))

\( D_o \) - true disturbance model

\( D_i \) - disturbance/noise at time \( t \)

\( e_i \) - modelling error at time \( t \)

\( G_{ai} \) - the true impulse weights between input \( i \) and output \( j \)

\( G_o \) - the true process model

\( H(z^{-1}) \) - estimate of the noise

\( K_i \) - matrices used in RR (see equation 5.30)

\( m \) - number of inputs

\( N(t) \) - estimate of the noise

\( p \) - number of parameters to estimate

\( r \) - truncation point for the FIR model

\( R(z^{-1}) \) - noise model used in GLS
u_t - the input to the process at time t
X - the matrix of inputs in the PLS regression
v_t(z) - polynomial of impulse weights for the FIR model
Y_t - the true value of the process output from the simulation
y_t - the modelled process output
Y - the matrix or vector of outputs in the PLS regression
Z^N - the sequence of input data
Z_t - the simulated output for each transfer function
z^{-1} - backward shift operator
\phi_u(\omega) - power spectrum of u
\theta - the estimated model parameters
\omega - the frequency
\omega_i(z) - numerator of the process transfer function i, a polynomial in z

Abbreviations

ARX - Autoregressive with Exogeneous Variable model (12,4), Eq. 4.
DMC - Dynamic Matrix Control
FIR - Finite Impulse Response function (35,0), Eq. 1.
LS - Ordinary Linear Least Squares
MIMO - Multiple Input Multiple Output
MISO - Multiple Input Single Output
MSE - Mean Squared Error
PLS - Partial Least Squares or Projection to Latent Structures
PRBS - Pseudo Random Binary Sequence
RR - Regularization Regression
SISO - Single Input Single Output

5.8 References


Chapter 6. Summary and Conclusions

The PLS method was developed as an algorithm, by H. Wold and improved by subsequent researchers, which could be used to gain information from multivariate data sets containing many correlated variables. The wide applicability of the method and its success, especially in the natural sciences, lead to increased usage despite the limited understanding of the properties of the method. The data collected from many chemical engineering processes exhibit many of the same characteristics as those in chemistry, biology, etc., where PLS was successfully applied (i.e. many correlated measurements of a relatively small number of independent phenomena). In the past the analysis of these data sets has been difficult due to the inter-related nature of the measurements; however, by defining new latent variables PLS is able to summarize the information content of these large data sets which facilitates both the interpretation of the data and building predictive models.

The objectives of this work were to investigate the applicability of using PLS in three areas of chemical engineering: 1) for performance monitoring, 2) for inferential model development and 3) in dynamic model identification. One further goal of this work was pursued in all aspects of this work; this was to gain understanding of the properties of the PLS method in the context of chemical engineering. The specific properties of interest were: the effects of scaling, the definition of the objective function for the PLS calculations and finally general understanding of the properties of PLS and how they impact on its applicability.
Understanding of the properties of the PLS method were obtained by applying the method to data obtained through simulation studies of chemical processes rather than from a theoretical development. During the course of this work other researchers were also investigating these aspects of the PLS method from a theoretical perspective and many of the mathematical and statistical properties of PLS have been illuminated in the past three years. Of particular relevance to this work are the insights provided by Hoskuldsson (1988) on the relationship of PLS to other methods and its objective function and to Phatak et al. (1991) on the effects of scaling when using PLS1.

The combination of the theoretical work done by these researchers and the work done in this thesis provide the basis for the following insights to the PLS method:

1) The PLS method is scale dependent. Of particular importance is the relative scaling of the variables in the individual X and Y blocks, with the scaling of the X block variables having the largest influence on the PLS model. Variable scaling will determine the order in which phenomena are extracted from the data sets and which variables will dominate the PLS model. If large numbers of PLS dimensions are used in the model the order in which different phenomena are incorporated in the model is less important, than when only a limited number of PLS dimensions can be used and important effects may be lost. No single scaling approach is appropriate for all applications. The scaling chosen for each application should reflect the user's knowledge of the system (e.g. are the measurements naturally scaled or are some variables critical to a successful model) and the objectives of the investigation.
2) No single mathematical expression for the objective function of PLS has been formulated, but the objective of PLS calculation at each dimension has been stated by Hoskuldsson (1988) and can be paraphrased as finding the linear combinations in X and in Y such that the covariance between these linear combinations is the maximum of all possible pairs of linear combinations. This means that the model obtained using PLS does not simply minimize the squared error of the Y variable predictions but provides a compromise between prediction of Y and explanation of the data structure in X. The balance of this compromise at any PLS dimension is affected by the scaling. Thus, PLS does not necessarily rely on those X variables which are most predictive of Y; rather, the method uses those variables in X which exhibit sufficient variation and predictive ability for Y.

3) PLS is an empirical method and the success of each PLS application is highly dependent on the structure and the information content of the X and Y blocks. Further, the PLS algorithm does not allow heuristic user input at intermediate stages of the calculation. However, the algorithm does allow extreme flexibility in the structure of the X and Y blocks which in turn allows the modelling of a wide variety of processes: linear, non-linear, stationary or time variant. Thus, successful application is highly dependent on the set up of the X and Y blocks, but once this is accomplished PLS is easily automated. Finally, by representing the original variables by a linear combination an averaging of errors is accomplished and by using the PLS formulation, as described in chapter 4, the predictions are robust to missing data.
Monitoring performance and detecting faults is an integral part of the successful operation of any process. Performance can be monitored by comparing the actual results to the predictions from a mechanistic model, or by using statistical process control (SPC) charts (e.g. Shewart charts, CUSUM charts or EWMA charts to compare the current state of the process against "normal operating conditions". The challenge currently facing process monitoring is the enormous amount of correlated data being collected from a multitude of sensors every few seconds, minutes or hours. This "data overload" and the lack of appropriate analytical tools has meant that very little is being done to utilize this wealth of information.

The first application of PLS to a chemical engineering problem considered in this work was performance monitoring of a multivariate processes. A multivariate SPC procedure has been proposed for handling large numbers of process and quality variables. Multivariate statistical procedures are used to reduce the dimensionality of these large and highly correlated data sets down to a few latent variables which contain most of the information about the process behaviour under normal operating conditions. By plotting the projections (rows) of new process observations over time on this low dimensional plane one is able to detect larger than normal process variations, and by also plotting the squared prediction errors, (ie. the perpendicular distances from the plane) one is also able to detect major changes in the behaviour of the process caused by new events.

By compressing all the information on the process down to low dimensional spaces, and using simple plots of the data in these spaces, together with meaningful control limits, the essential idea and philosophy of Shewart's (1931) SPC methods have been preserved and extended to handle the large number of variables collected
in most process industries today. The tools necessary to establish the multivariate charts (PCA and PLS) may be more complex than usually used in univariate SPC, but from the user’s point of view, the presentation of the data, and the interpretation of the results is almost as simple.

A natural application of PLS is to build empirical models so that quality measurements can be predicted from process measurements. This model building can be achieved using classical regression techniques for situations where truly independent process variables are chosen to predict the quality variables. The opportunity for PLS lies in applications with many correlated process variables. In the second application, a procedure for developing inferential process models using PLS has been outlined. It has been shown to be a very powerful approach to building such models when large numbers of highly correlated measured variables are available. By retaining all the measurements without overfitting the data PLS is able to utilize all the information obtained from the process measurements. It has also been shown that PLS models will be extremely robust to missing data and sensor failures, an important feature of any inferential control scheme.

The importance of obtaining a representative reference data set for the process to build these models has also been demonstrated. In particular, it is important that the data contain a representative sample of the effects of all the manipulated variables and the process disturbances, and that the measurement noise levels be typical of those that will be encountered in the final control scheme. Furthermore, it was shown that the data must be collected while the process is operating under a feedback structure very similar to the one that it is intended to be used. An iterative
procedure was suggested to accomplish this. These points were illustrated through the development of PLS inferential models for two simulated distillation columns in chapter 4.

A preliminary investigation into obtaining dynamic inferential model for use in inferential control schemes showed that the dynamic prediction obtained depended on the disturbance entering the process, the formulation of the model (e.g., using steady-state or dynamic data) and on the variables used in the model. The predictions varied greatly and no single formulation was found for which the predictions for all disturbances were acceptable. Further work in this area is necessary to formulate what is a desirable dynamic response and how to formulate the PLS problem so that this desirable dynamic response can be achieved.

In the final application the use of PLS for the identification of dynamic process models (for use in model predictive control) was investigated. In many model predictive control schemes being used today (e.g., Dynamic Matrix Control) the number of input and output variables is very large and non-parsimonious finite impulse or step response models are being used. The data sets used in these multivariate identifications involve many highly correlated input variables with designed perturbations in only a few at a time. Impulse and step response models are used not only because they fit naturally into the final model predictive control algorithms, but because the multivariable control scheme is usually being applied around a large section of the process within which there exist many lower level controllers and recycle loops. In this situation low order transfer function models characteristic of most open-loop processes are often not adequate and the selection of appropriate parsimonious model structures necessary for Box and Jenkins identification is not
easy.

The comparison of PLS, RR and LS identification of non-parsimonious (FIR and ARX) models for MISO systems shows that the biased estimates obtained from PLS and RR provide smoother estimates of the impulse and step weights (i.e. with smaller variance from their true values). PLS and RR provide similar estimates but RR would be preferable for its simplicity. All methods suffer a reduction in accuracy with poorly designed data sets or significant noise. The advantage of using the FIR and ARX models over the parsimonious Box and Jenkins models is that they can be used to model processes where low order models are not appropriate. These methods are also simpler to apply since there is no need to iteratively determine a parsimonious model order for each transfer function and are directly applicable in many model based controllers. There are, however, drawbacks to these models; since more parameters are being estimated, more information is necessary to obtain an estimate which is as accurate as that obtained using the parsimonious Box and Jenkins approach.

A further drawback to the application of any of these methods is the restrictive structure of the models that are linear in the parameters which prohibit an accurate characterization of the structured noise. Using a method such as PLS and an ARX model provides enough flexibility that accurate prediction of the output is possible; however, the model obtained does not reflect the true process parameters, but is rather, an empirical model which is only accurate as long as the process and noise structures do not change.

When using PLS in the identification the structure of the information content and the stopping criteria used must be considered. This work showed that for many
of the case studies the minimum of the prediction error curve (as indicated by PRESS) is quite flat and that in some situations, i.e. to obtain a better estimate of the dead time and in order to explain a higher proportion of the X block structure, it is beneficial to use a higher dimensional PLS model. Careful experimental design and pretreatment (e.g. differencing or scaling) is necessary because the PLS method extracts those parts of the model which dominate the information content in the data. If the data is noisy and the information relevant to the model is masked by other variation in the data, then poor estimates can be obtained.

Use of multivariate PLS regression in the identification of MIMO processes did not show as large an improvement as was expected. Multivariate identification is beneficial when there are common parameters in the identification. Although correlation among the output variables was present the large number of latent variables meant that only a limited dimensional reduction (actually no dimension reduction in the output space) was achieved. The regression model was not adequately constrained to ignore the extraneous information contained in the noise and outliers.

The applications of PLS to chemical engineering problems discussed in this work progress from using the method for obtaining insight into multivariate steady-state processes for performance monitoring, to developing dynamic process models for prediction and control. In the development of multivariate monitoring charts and inferential model the unique properties of PLS allow the benefits obtained from the retention of the important information, contained in numerous correlated measurements of underlying phenomena, without the associated problems of high dimensionality and over-parameterization. In dynamic model identification these
same properties are focussed on variables correlated in time. By using PLS, over-parameterized FIR and ARX models can be used to model multivariate processes for model predictive controllers such as DMC.
Appendix
This is a brief guide to the computer simulation programs used to generate the data presented in Chapters 3 and 4 of this thesis. The case studies presented in those chapters are based on computer simulations of three processes: a Fluidized Bed Reactor (FBR), a distillation column separating a Benzene-Toluene-Xylene (BTX) mixture and an extractive distillation of a Methanol-Acetone-Water (MAW) mixture. The BTX column was simulated using the SimSci package. The programs used for the simulation of the FBR and the MAW column were developed by other students at McMaster and the versions used in this work are archived with other programs and results generated in the course of this work. The following descriptions will allow future researchers to find and run these programs.

The Fluidized Bed Reactor (FBR) Simulation

The FBR simulation was used to generate data for the performance monitoring study of Chapter 3. This simulation package was developed by Jeff Kelly and the details of the program can be found in his thesis. The program used in this study was modified so that entire sets of steady-state conditions could be simulated in one run, this version of the program can be found in the [KRESTA.FBR] directory. The simulation is executed using the FBR_SS_SIMULATION.EXE program. This program can be run interactively or as a batch program using a command file (eg. NFB1.COM). The only inputs necessary are the name of the output file and the number of points to be generated. The program is currently configured to generate data points normally distributed around the base case. Alternately, the program can be used to generate any sequence of conditions by making the appropriate modifi-
cations to the SET_CONDITIONS.FOR subroutine and recompiling the program. An example of the output is provided in Figure A.1, each value is identified with a superscript and an explanation is given in the second part of the figure.

The Extractive Distillation Column Simulation

The extractive distillation of a Methanol-Acetone-Water (MAW) mixture was simulated using code developed by Andrew Chin. This code was used to simulate both steady-state and dynamic data, and is archived in the [KRESTA.EDC_PROG] directory. The steady state results were generated using the SSTATE.EXE program. This program was executed using batch programs, eg. [KRESTA.EDC]INPUT.COM, a partial listing of this batch program and an explanation of the inputs is given in Figure A.2. A corresponding partial listing of the output from this program is given in Figure A.3. The dynamic results were generated using the MAIN.EXE program. The initial conditions as well as any changes in the inputs to the column occurring during the simulation must be programmed directly into the MAIN.FOR program and the entire package recompiled using MAKE.MMS. The output for the dynamic simulation is in the same format as for the steady-state simulation.

Other Programs

Several other programs are also archived in my directories, many of which would be of little interest to other students. The notable exception is the SIMCA package used for calculating the PLS models. The executable code for this package is in the [KRESTA.SIMCA] directory and the source code is in the [KRESTA.SIMCA.SOURCE] directory. This is an early version of PLS programs and has
very limited plotting capabilities and for this reason the MATLAB version of PLS programmed by Bert Skagerberg was used in Chapter 5; however, this package allows different scaling of variables in both the X and Y blocks which is useful in many cases. The input format and directions for running this software can be found in the appropriate SIMCA manual.
<table>
<thead>
<tr>
<th>Concentration of butane</th>
<th>Concentration of hydrogen</th>
<th>Hydrogen to butane ratio</th>
<th>Volumetric flow in</th>
<th>Temperature at inlet</th>
<th>Temperature at outlet of reactor</th>
<th>Cooling oil temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1886</td>
<td>27.5761</td>
<td>8.6484</td>
<td>0.001330</td>
<td>520.89</td>
<td>529.77</td>
<td>535.67</td>
</tr>
<tr>
<td>4.3049</td>
<td>0.4031</td>
<td>0.5418</td>
<td>1.5045</td>
<td>24.0104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5561</td>
<td>0.2394</td>
<td>0.3217</td>
<td>0.5282</td>
<td>0.1293</td>
<td>0.9874</td>
<td></td>
</tr>
<tr>
<td>2.8466</td>
<td>28.2444</td>
<td>9.9223</td>
<td>0.001205</td>
<td>525.32</td>
<td>533.14</td>
<td>539.48</td>
</tr>
<tr>
<td>4.8010</td>
<td>0.3993</td>
<td>0.4251</td>
<td>1.1278</td>
<td>24.3378</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7934</td>
<td>0.2323</td>
<td>0.2473</td>
<td>0.6038</td>
<td>0.1383</td>
<td>0.7030</td>
<td></td>
</tr>
<tr>
<td>2.7126</td>
<td>28.1829</td>
<td>10.3824</td>
<td>0.001418</td>
<td>528.98</td>
<td>530.77</td>
<td>537.24</td>
</tr>
<tr>
<td>3.9641</td>
<td>0.3567</td>
<td>0.4451</td>
<td>1.2093</td>
<td>24.9002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6370</td>
<td>0.2373</td>
<td>0.7961</td>
<td>0.5542</td>
<td>0.1159</td>
<td>0.1691</td>
<td></td>
</tr>
<tr>
<td>2.9725</td>
<td>27.5623</td>
<td>9.2727</td>
<td>0.001528</td>
<td>534.88</td>
<td>532.01</td>
<td>540.24</td>
</tr>
<tr>
<td>4.7249</td>
<td>0.3981</td>
<td>0.4565</td>
<td>1.2498</td>
<td>23.7060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7428</td>
<td>0.2311</td>
<td>0.2650</td>
<td>0.5795</td>
<td>0.1399</td>
<td>0.9577</td>
<td></td>
</tr>
<tr>
<td>2.5208</td>
<td>28.2289</td>
<td>11.1986</td>
<td>0.001700</td>
<td>531.15</td>
<td>529.64</td>
<td>535.07</td>
</tr>
<tr>
<td>2.8485</td>
<td>0.2768</td>
<td>0.4188</td>
<td>1.3561</td>
<td>25.8494</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4459</td>
<td>0.2376</td>
<td>0.3596</td>
<td>0.4620</td>
<td>0.0843</td>
<td>0.9776</td>
<td></td>
</tr>
</tbody>
</table>

**Figure A.1** Partial printout from the FBR simulation and a description of the variables.
set def sys$sysdevice:[kresta.edc]
run/nodebug sstate

40.60612
106.9208
336.7764
317.9809
403.7463
64.70321  31.29679  4.000000
edct

RUN/NODEBUG SSTATE
40.15733
103.8275
340.2365
322.4985
409.3708
63.10598  32.89402  4.000000
edct

Flowrate of solvent
Flowrate of feed
Temperature of solvent
Temperature of feed
Reboiler steam temperature
Inlet feed composition: Methanol, Acetone, Water

Figure A.2  Input file for the MAW column simulation.
2.18096 336.77640 0.01720 317.98090 64.70 31.30 4.00
74.456 66.103 62.739 64.222 63.626 66.257 65.121 63.909
62.985 62.386 59.562 58.695 57.999 57.284 44.886
1.0000 303.0000 403.7463 71.1099 111.5520
0.98691 0.90194 0.01412 0.08395
2.91429 0.07649 0.17996 0.74355
1.16250 1.14728

Row 1  Solvent flowrate
Row 2  Temperature profile: (1-8)
Row 3  Temperature profile (9-15)
Row 4  Pressure
Row 5  Distillate flowrate
Row 6  Bottoms flowrate
Row 7  Reflux ratio

Figure A.3 Output file from the MAW column simulation.