THE APPLICATION OF MULTIVARIATE STATISTICAL ANALYSIS AND BATCH PROCESS CONTROL IN INDUSTRIAL PROCESSES

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2010

Haisheng Lin

School of Electrical and Electronic Engineering

Contents

List of Tables

List of Figures

Abstract

To manufacture safe, effective and affordable medicines with greater efficiency, process analytical technology (PAT) has been introduced by the Food and Drug Agency to encourage the pharmaceutical industry to develop and design well-understood processes. PAT requires chemical imaging techniques to be used to collect process variables for real-time process analysis. Multivariate statistical analysis tools and process control tools are important for implementing PAT in the development and manufacture of pharmaceuticals as they enable information to be extracted from the PAT measurements.

Multivariate statistical analysis methods such as principal component analysis (PCA) and independent component analysis (ICA) are applied in this thesis to extract information regarding a pharmaceutical tablet. ICA was found to outperform PCA and was able to identify the presence of five different materials and their spatial distribution around the tablet.

Another important area for PAT is in improving the control of processes. In the pharmaceutical industry, many of the processes operate in a batch strategy, which introduces difficult control challenges. Near-infrared (NIR) spectroscopy is a non-destructive analytical technique that has been used extensively to extract chemical and physical information from a product sample based on the scattering effect of light. In this thesis, NIR measurements were incorporated as feedback information into several control strategies. Although these controllers performed reasonably well, they could only regulate the NIR spectrum at a number of wavenumbers, rather than over the full spectrum.

In an attempt to regulate the entire NIR spectrum, a novel control algorithm was developed. This controller was found to be superior to the only comparable controller and able to regulate the NIR similarly. The benefits of the proposed controller were demonstrated using a benchmark simulation of a batch reactor.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to offer my deepest gratitude to my supervisor Prof. Barry Lennox for his supervision, advice and guidance during my PhD study. Barry led me to the area of multivariate statistical analysis and batch process control. He gave me steadfast encouragement and support throughout my work. His insight into problems and his strict attitude to research impressed me a great deal. I attribute this thesis to his encouragement, effort and patience: without him this thesis would not have been completed.

Many thanks go in particular to my advisor Dr. Ognjen Marjanovic, who has given me so much support and help in this work. He has helped me enormously in many areas, from research study, to Matlab programming, to paper writing and so on. His selfless devotion to the work is greatly appreciated.

My special thanks go to Dr. Xiaozhi Gao and Dr. Hongwei Zhang, who helped me a great deal by introducing me to this great research opportunity. I would like to thank Dr. Zhengtao Ding, who gave me useful suggestions in many areas. I would like to thank all my friends in the control system centre. Thank you for all your support during my study and for making life so pleasant for me.

This research is support by the award of the Overseas Scholarship Scheme, the University of Manchester and Pfizer Incorporated. The financial support is much appreciated.

Last but not least, I would like to thank my dear parents who have given me unfailingly support and prayers. Their selfless contributions cannot be expressed in any words. I would like to thank my brother Dr. Enxin Lin who has helped me so much during my life. I gratefully thank my uncle Prof. Qingping Lin, who has inspired me to do research in the UK and has helped me a great deal in planning my career path.

Abbreviation

SOP	Standard Operating Procedures
FDA	Food and Drug Administration
PAT	Process Analytical Technology
ICA	Independent component analysis
PCA	Principal Component Analysis
MPC	Model Predictive Control
Wn-MPC	Wavenumber-Based Model Predictive Control
LMPC	Linear Model Predictive Control
Sc-MPC	PCA Score-Based MPC Control
NIR	Near-infrared Spectroscopy
FTIR	Fourier Transform Infrared Spectroscopy
IR	Infrared Spectroscopy
ASTM	American Society for Testing and Materials
PLS	Partial Least Squares
PC	Principal Component
IC	Independent Component
SIMCA	Soft Independent Modeling of Class Analogy
FCM	Fuzzy-C Mean
MCR	Multivariate Curve Resolution
ALS	Alternating Least Squares
MCR-ALS	Multivariate Curve Resolution-Alternating Least Squares
SVD	Singular Value Decomposition
API	Active Pharmaceutical Ingredient
DCLS	Direct Classical Least Squares
GLC	Globally Linearizing Control
GMC	Generic Model Control
DM	Dual-model
TC	Temperature Controller
CC	Cascade Control

ARX	Autoregressive with Exogenous Terms Model
RLS	Recursive Least Squares
SSQ	Sum of Square Errors
NIPALS	Non-linear Iterative Partial Least Squares
SIMPLS	Straightforward Implementation of a Statistically Inspired Modification of the Partial Least Squares
SCP	Single Component Projection
RMSEC	Root Mean Square Error of Calibration
TRSP	Reactor Temperature Set-Points
TR	Reactor Temperature
TJSP	Jacket Temperature Set-Points
TJ	Jacket Temperature
MVT	Manipulated Variable Trajectory

Notations

t_k	k^{th} score vector for PCA model / PLS model
p_k	k^{th} loading vector for PCA model / PLS model
e_k	k^{th} prediction error for PLS model
Z	Data matrix with m rows and n columns for PCA model
m	Number of rows in data matrix Z
n	Number of columns in data matrix Z
n_c	Number of principal components in PCA model
E	Information in Z that is not represented in the n_c principal components
Cov(.)	Covariance
λ_k	k^{th} eigenvalue of the covariance matrix
S	Vector of statistical independent latent variables for ICA model
A	Mixing matrix
W	Estimation of the pseudo inverse of A
$kurt(.)$	Kurtosis
$H(.)$	Entropy
$j(.)$	Negentropy
$\hat{y}(k)$	Future output at sampling time k
$y^{sp}(k)$	Set-point (reference trajectory) at sampling time k
$y_{\min}$	Minimum value for actual output
$y_{\max}$	Maximum value for actual output
$\hat{u}(k)$	Future input at sampling time k
$u(k)$	Actual input at sampling time k
$u_{\min}$	Minimum value for actual input
$u_{\max}$	Maximum value for actual input
M_A	Molecular weight of components of A, $kg/kmol$

M_B Molecular weight of components of B, $kg/kmol$

M_C Molecular weight of components of C, $kg/kmol$

M_D Molecular weight of components of D, $kg/kmol$

M_r The whole molecular weight of components A,B,C,D, $kg/kmol$

R_1 Reaction rate of reaction 1, $kmol/s$

R_2 Reaction rate of reaction 2, $kmol/s$

T_r Reaction temperature, $^\circ C$

T_j Jacket temperature, $^\circ C$

T_j^{sp} Jacket temperature set point, $^\circ C$

Q_r Heat released in reactor, Kw

Q_j Heat released in jacket, Kw

C_{pr} Mass heat capacity of reactor contents, $kJ/(kg\ ^\circ C)$

C_{pj} Mass heat capacity of jacket contents, $kJ/(kg\ ^\circ C)$

C_{pA} Mole heat capacity of component A, $kJ/(kmol\ ^\circ C)$

C_{pB} Mole heat capacity of component B, $kJ/(kmol\ ^\circ C)$

C_{pC} Mole heat capacity of component C, $kJ/(kmol\ ^\circ C)$

C_{pD} Mole heat capacity of component D, $kJ/(kmol\ ^\circ C)$

τ_j First order time constant, s

ρ_j Density of reactor contents, kg/m^3

V_j Volume of jacket, m^3

k_1 Rate constant for reaction 1, $kmol^{-1}s^{-1}$

k_2 Rate constant for reaction 2, $kmol^{-1}s^{-1}$

k_1^1 Rate constant 1 for reaction 1, $kmol^{-1}s^{-1}$

k_1^2 Rate constant 2 for reaction 1, $kmol^{-1}s^{-1}$

k_2^1	Rate constant 1 for reaction 2, $kmol^{-1}s^{-1}$
k_2^2	Rate constant 2 for reaction 2, $kmol^{-1}s^{-1}$
ΔH_1	Heat of reaction for reaction 1, $kJ/kmol$
ΔH_2	Heat of reaction for reaction 2, $kJ/kmol$
U	Heat transfer coefficient of reactor, $kW/(m^2\ ^\circ C)$
X^T	Data matrix
X_{on}^T	Online process variables
X_{off}^T	Offline measurements
u_c^T	Manipulated variables
$\hat{t}_{present}^T$	Projection of the X matrix onto the latent variable space
θ_i	Decision point
X_1^T	Known trajectory
X_2^T	Future unknown trajectory
P_1^T	Loading matrices for X_1^T
P_2^T	Loading matrices for X_2^T
$\hat{X}$	Predicted vector of the process variables
X_{nominal}	Predicted vector of the process variables
$\Delta X_{\min}$	Minimum value for $\hat{X} - X_{\text{nominal}}$
$\Delta X_{\max}$	Minimum value for $\hat{X} - X_{\text{nominal}}$
Q_1	Diagonal weighting matrices for the variables in y
Q_2	Diagonal weighting matrices for the variables in ΔX

Publications

- Lin, H., Lennox, B., Marjanovic, O. (2010) 'Application of End-Point Control and Trajectory Tracking to Batch Processes', *11th Computer Applications in Biotechnology*, Leuven, Belgium. Accepted.
- Lin, H., Marjanovic, O., Lennox, B., and Shamekh, A. (2009) 'Application of Near-infrared Spectroscopy in Batch Process Control'. *International Symposium on Advanced Control of Chemical Processes*, Istanbul, Turkey.
- Lin, H., Marjanovic, O., Lennox, B. (2008) 'Multivariate Statistical Analysis of Spectroscopic Data', *Proceedings of the UKACC International Conference on Control.* Manchester, UK.
- Shamekh, A., Lennox, B., and Lin, H. (2010) 'Comparison between Standard PLS, RLS, and Unbiased PLS', *Proceeding of Modelling, Identification, and Control,* Innsbruck, Austria.

1 Introduction

This chapter states the purpose of the work, lists the main contributions and provides an outline of the thesis.

This chapter is divided into the following sections:

1) Statement of the main subject matter of the thesis.
2) Contributions of this thesis.
3) Structure of this thesis.

1.1 Background Statement

The production of safe and effective medicines is of vital importance in the pharmaceutical industry, and in ensuring a healthy population. Furthermore, in recent years, the industry has been affected significantly by the growth in 'generic' pharmaceutical manufacturers, which has introduced a need in the pharmaceutical industry to improve the efficiency of production facilities. Finding methods to improve the efficiency of pharmaceutical manufacturing has therefore become an important issue. More recently, investment in the pharmaceutical industry has been focused on improving the efficiency by which products are manufactured.

Traditional pharmaceutical manufacturing uses batch processing with laboratory tests being conducted on collected finished samples to evaluate quality. It takes a considerable length of time for a tablet to be produced and tested, which can mean that substantial amounts of sub-standard material can be produced before problems are detected, which is clearly inefficient. Further inefficiencies are often claimed to be a result of the regulatory policies and practices in place in the industry. Current pharmaceutical manufacturing requires a plant to be operated under strict regulations described in the Standard Operating Procedures (SOPs). These regulations require extensive documentation of the manufacturing and supply chain processes. Under these regulations, products must

typically be manufactured under the same set of conditions each time, ignoring the impact of changes in raw material properties, for example.

From the public health view and from the perspective of improving efficiency and innovation, the best principles of quality management should be employed by the pharmaceutical industry. Efficient manufacturing can reduce operating costs, and accelerate the time it takes to market the product and make it available to the customer.

In order to encourage the pharmaceutical industry to introduce and develop innovative processing systems, the Food and Drug Administration (FDA) launched "*Guidance for Industry PAT—A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance*" [1], which describes a new regulatory framework (Process Analytical Technology, PAT). According to [1], PAT is: "*a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality*". The goal of PAT is: "*to understand and control the manufacturing process, which is consistent with our current drug quality system: quality cannot be tested into products; it should be built-in or should be by design*". In other words, PAT should be used to design well-understood processes that will ensure pre-defined quality in the manufacturing process. This will improve the efficiency of manufacturing.

Process analytical technology (PAT) is a revolutionary process quality control technique now being applied in the pharmaceutical industry. It is also applied in other industries such as the petrochemical and food industries [2]. Kourti [3] provides an overview of the process analytical technology initiative and of multivariate process analysis, monitoring and control. By using PAT, there is potential to reduce cycle times, decrease batch release times, reduce start-up times and improve project scheduling [4].

To gain the most from PAT, there are several areas in which it can be applied and many methods exist to aid in its application. These methods can be divided into four broad categories [1]:

1) Multivariate data acquisition and analysis tools
2) Process analytical chemistry tools
3) Process monitoring and control tools
4) Continuous improvement and knowledge management tools

Pharmaceutical processes are complex systems as the manufacturing of pharmaceuticals requires many different development procedures, each of which requires process design knowledge. Knowledge of the process is important for process quality control because it helps in understanding the relationships that exist in the process and provides useful information for process control purpose. Knowledge of the process may include biological, chemical and physical attributes of the materials being processed.

To effectively implement real-time monitoring and control of product quality during the manufacturing processes, knowledge and understanding of the process is critical. Unfortunately, knowledge of the process is often limited and a key area for PAT techniques is in the application of advanced sensors, such as chemical imaging, to gain a better understanding of the process. Chemical imaging techniques are non-destructive analytical techniques that are being used with increasing frequency in the pharmaceutical industry to collect important process data on-line. These methods are reviewed in Section 2.1.1. The data collected using chemical imaging techniques is typically of a large volume, which can make it difficult to interpret, and multivariate statistical analysis tools are often needed to extract critical information from the images for use in real-time control and quality assurance. A review of multivariate statistical analysis of chemical imaging data is given in Section 2.1.1. Finally, to improve product quality and increase processing efficiency, the application of process control techniques are beginning to be explored in the pharmaceutical industry. Section 2.2 provides a brief review of batch control techniques.

1.2 Contributions of this Thesis

The work described in this thesis focuses on multivariate statistical analysis of Raman images of a pharmaceutical tablet and quality control in batch processes. Major research contributions have been made in the following areas:

1) The extraction of chemical information from images of pharmaceutical tablets is an important issue. Various multivariate statistical analysis methods have been applied to it, but they all have disadvantages. Independent component analysis (ICA) is a powerful method for statistically separating mutually statistical independent components and in this work it is applied to identify the reference spectra of a pharmaceutical tablet's constituent compounds. Its performance is compared with the performance achieved using Principal Component Analysis (PCA). It is shown that ICA outperforms PCA. ICA was able to identify all the constituent compounds in the tablet, while PCA was only able to identify four, missing the most concentrated constituent.
2) Quality control is usually implemented by regulating several process variables. However, even if the process variables are well maintained, the quality of the final product cannot be guaranteed, due to the existence of disturbances such as changes in the raw material properties. To address this, two approaches for utilising NIR spectra in a feedback control system are proposed. Wavenumber-Based model predictive control (Wn-MPC) uses the measurement from a single peak as a controlled variable. In contrast, the second approach, PCA Score-Based MPC control (Sc-MPC) utilises PCA to extract the important information from the spectral measurement and control is applied to these measurements. The two methods were applied to the quality control and their performances compared with a traditional controller – cascade control. It is shown that Sc-MPC achieved satisfactory results with no user interaction. The performance of Wn-MPC, however, was heavily dependent on the selection of the wavenumber at which the NIR spectrum was controlled.
3) Quality control using Sc-MPC and Wn-MPC only regulates the NIR spectrum at specific wavenumbers. In contrast, the end-point controller introduced in Chapter 5 regulates the NIR measurements across all wavenumbers, and is able to achieve

improved control performances. Chapter 6 proposed a novel end-point controller which adjusts the trajectories of manipulated variables (MVTs) in the real process space, while taking into consideration process constraints. The technique is applied to track a set-point change in NIR spectrum on a simulated batch process. The results show that the proposed end-point control algorithm is better at tracking the set-point change than other applicable control systems.

1.3 Structure of the Thesis

This thesis is laid out in seven chapters and divided into three main parts. Chapter 2 presents a literature review on the application of multivariate statistical analysis methods to chemical imaging data and batch process control in industrial processes. Chapter 3 describes the application of multivariate statistical analysis methods to a Raman image of a pharmaceutical tablet. Chapters 4-6 focus on quality control of batch processes. Chapter 7 provides the conclusion of this thesis.

Following this introduction, Chapter 2 presents an overview of chemical imaging techniques, multivariate data analysis methods, and batch process control. NIR and Raman spectroscopy are the two chemical imaging techniques this chapter focuses on. An overview of the application of multivariate data analysis methods to chemical imaging data is presented, followed by a review of quality control in batch processes.

Chapter 3 describes the application of multivariate data analysis methods including PCA and ICA to identify the reference spectra of a pharmaceutical tablet's constituent compounds using Raman spectroscopic data. The spatial distribution of each of the constituent compounds from this tablet is estimated using the results from the multivariate analysis and this is presented in a clear graphical form. This provides a gross distribution of the constituent compounds in the tablet.

Chapter 4 proposes two quality control approaches by incorporating NIR spectra as feedback information. Both control approaches utilize MPC strategies. They are applied to regulate the NIR spectrum of the end product in a simulated batch reactor.

In Chapter 5, an end-point control method is used to control the product quality by regulating the entire end-point NIR spectrum. The method is applied to regulate the NIR spectrum and its performance is compared with that achieved using the control techniques introduced in Chapter 4.

Chapter 6 proposes a novel end-point controller which adjusts the trajectories of manipulated variables (MVTs) in the real process space, while taking into consideration process constraints. The performances of this controller are compared with the end-point controller proposed in previous studies by Flores-Cerrillo [5].

Finally, Chapter 7 provides conclusions of the thesis and gives suggestions for further work.

2 Literature Review on the Application of Multivariate Statistical Analysis Methods and Batch Process Control in Industrial Processes

This chapter provides an overview of the application of multivariate statistical analysis methods to extract important information from chemical imaging data. An overview of the application of batch process control techniques, using PAT, is also presented.

The chapter is divided into the following sections:
2.1) Overview of multivariate statistical analysis of chemical imaging data;
2.2) Introduction to batch process control;
2.3) Summary of this chapter.

2.1 Overview of Multivariate Statistical Analysis of Chemical Imaging Data

Knowledge of process is vital if it is to be effectively monitored and controlled in real-time. Chemical imaging techniques has been shown to be an excellent tool to collect process data online and provide a non-destructive way of gaining an understanding of a process. With the acquisition of chemical imaging data, multivariate statistical analysis methods are used to extract the critical information of the process. In this section, an overview of chemical imaging techniques is first provided, and then a review of the application of multivariate statistical analysis methods to this data is presented.

2.1.1 Chemical Imaging Techniques

To better understand what happens in their processes, many pharmaceutical companies have been exploring the use of various chemical imaging techniques, with Near-infrared (NIR) [6] and Raman spectroscopy [7] being the two most commonly applied techniques.

Chemical imaging techniques are non-destructive analytical methods which provide chemical and physical information regarding a sample of material. They have been shown to be capable of providing online quality measurements that can be used to improve process understanding and monitoring of the progression of batch processes. As such they have huge potential for application in process monitoring and control techniques. Chemical imaging has become a commonly used analytical technique in research and development, and is now being investigated in its application to quality assurance and quality control. It has been widely applied in solid samples [8], biology [9, 10], chemistry [11], medicine [12], agriculture [13], biotechnology [6, 14], and pharmacy [6].

Near infrared spectroscopy is a method which uses the near infrared region of the electromagnetic spectrum. The near infrared region is defined by the American Society for Testing and Materials (ASTM) as the spectral region with the wavelength range of 780-2526 nm. The wavelength covers ranges from the mid-infrared to the visible region. The NIR spectrum is generated by the overtones and combinations of fundamental vibrations of hydrogen bonds such as C–H, N–H, O–H and S–H.

The NIR region was identified in the 19^{th} century, and NIR spectroscopy was first applied in industry in the 1950s, and applications of NIR have gradually grown since then. The real breakthrough for using NIR spectra in process control and quality control occurred in the 1980s, and since then NIR spectroscopy has been applied in many industries. In the agricultural food industry, Slaughter et al. [15, 16] applied NIR to determine the soluble solids content of tomatoes. Balabin and Safieve [17] applied NIR spectroscopic data in the classification of gasoline by source and type in the petrochemical industry. In the environmental industry, Stallard et al. [18] applied NIR spectroscopy to the determination of motor oil contamination in sandy loam. In the pharmaceutical industry, NIR spectroscopy is widely recognized as a valuable tool for raw material testing, process monitoring and quality control [6, 19, 20], because it has the advantages that no sample preparation is required. Furthermore, the analysis is faster than other techniques, non-destructive, and can yield real-time measurements. The application of NIR spectroscopy in the pharmaceutical industry covers the following areas [6]:

- Identification and qualification of raw materials and intermediates [21-23].
- Analysis of intact dosage forms, including chemical, physical and related biopharmaceutical aspects [24-26].
- Process monitoring and process control [27-30].

Raman spectroscopy [31] is an alternative chemical imaging technique to NIR. It is a light scattering method where photons of light interact with a sample to produce scattered radiation of different wavelengths. Raman spectroscopy is based on the theory that when monochromatic light is incident on a sample, the light will interact with the sample and be reflected, absorbed or scattered. The scattered light then reflects information regarding the chemistry of the sample. Raman spectroscopy uses a laser to excite Raman spectra with a coherent beam of monochromatic light. The probe is then used to collect the Raman scatter. Following this, a spectrograph is used to separate the Raman scattered photons by wavelength, and a detector will record the intensity of the Raman signal at each wavelength.

Raman was discovered and named by Sir C. V. Raman in 1928 [32]. However it is only in the last two decades that Raman spectroscopy has developed into a mature technique for chemical analysis and has been applied in many areas. In cultural heritage, Sandalinas et al. [33, 34] applied Raman spectroscopy in the identification of pigments in Italian pottery from the sixteenth century. Welter et al. [33, 34] identified inorganic pigments in an ancient colored glass from Sir Lanka and Oman using Raman spectroscopy. In the biological industry, Raman spectroscopy has been applied to follow the distribution of the different compounds in plants in [35]. The application of Raman spectroscopy in food science was reviewed by Li-Chan [36]. Raman spectroscopy has also been widely applied in the pharmaceutical industry. In [37, 38], Šašić applied Raman spectroscopy to determine the spatial distribution of all components in a pharmaceutical tablet. Widjaja and Seah [39, 40] applied an analytical technique to detect and identify minor components of pharmaceutical drug tablets using Raman spectroscopy. The application of Raman spectroscopy in detecting minor component within pharmaceutical tablet can also

be found in [39, 40]. Hausman et al. [41] applied Raman spectroscopy as an on-line method to monitor the drug hydration state during drying.

Both NIR and Raman spectroscopy are widely used analytical chemical imaging techniques. However, Raman measurements contain more sharper bands and fewer overlapped bands [42], and hence it provides more selective spectral features for future analysis.

2.1.2 General Application of Multivariate Analysis of Chemical Imaging Data

Chemical imaging data can be used to extract important information relating to the progress of a process. It is therefore a useful tool for monitoring and controlling poorly understood systems. A major limitation with the technique, however, is that it typically results in very large data sets being produced. Research in this field has identified various methods that can be used to extract relevant information from this data. Two particularly powerful methods used for extracting this information are univariate and multivariate analysis methods.

Univariate analysis is used to analyze data by a single variable at a time and is the simplest method for revealing information from the data. However it requires the system to be clearly characterized. For example, in the analysis of the distribution of compounds in a pharmaceutical tablet [45], the application of univariate analysis requires that all the compounds that are present in the tablet, be known, and that the spectra of these compounds also be known in advance. These spectra are termed the reference spectra. It is also necessary to make sure that the chemical imaging spectra have unique and strong bands, which is not always possible.

Unfortunately, many chemical processes are complex, and the interaction of product and process variables exists. Furthermore, some constituents that are present in a sample may be unknown. In such cases, multivariate statistical analysis methods provide a more powerful method for extracting the relevant chemical information, such as concentration

and distribution of compounds, from the data. Multivariate statistical analysis is able to analyze several process variables simultaneously instead of singly, which often provides many benefits when analyzing large sets of data.

The most commonly used multivariate statistical analysis methods for extracting information from imaging data include principal component analysis (PCA), cluster analysis, multivariate curve resolution, partial least squares (PLS), and independent component analysis (ICA). The application of these methods to the analysis of chemical imaging data is briefly introduced below.

Principal Component Analysis (PCA)

PCA [43-45] is the algorithm that has probably been applied most frequently to extract information from chemical imaging data. It is a mathematical method which transforms a high-dimension dataset into a low-dimension dataset defined by what are referred to as principal components (PCs). The PCs are orthogonal to one another and are a linear combination of the original variables. They are identified such that they capture the main variation in the original data, with the first principal component explaining the largest variation within the data, and the second principal component having the second largest variation and so on. PCA has been applied in the analysis of chemical imaging data in various areas. For example, PCA was applied in the analysis of chemical imaging data to extract pure ingredient spectra from a pharmaceutical tablet [46]. Roggo et al. [47] applied PCA to infrared hyperspectral imaging of pharmaceutical solid forms to analyze a contamination on the surface of the tablet. Bacci et al. [48] applied PCA to analyze near-infrared spectra of alteration products in calcareous samples to discriminate the different reaction products. In [49], PCA was applied as a tool for the classification of materials using near-infrared spectroscopy. In this study, it was demonstrated that PCA was an effective and safe tool for classification purposes.

Cluster Analysis

Cluster analysis is used to assign a set of observations into several subsets (clusters) so that data in the same cluster contains observations with similar characteristics. These

clusters can be created using criteria such as distance and probability. Cluster analysis methods are typically classed as pattern recognition or data mining techniques. Cluster analysis techniques generally fall into two categories: unsupervised analysis and supervised analysis. A comprehensive review of cluster analysis techniques can be found in Jain et al. [50].

Supervised cluster analysis techniques require a set of samples that have already been assigned to a cluster for training. There are many supervised cluster analysis techniques. One widely used technique is soft independent modeling of class analogy (SIMCA) [51]. This method uses PCA or PLS to build a model to identify the clusters for the entire dataset. New observations are then classified into the identified clusters based on the best fit to the model. Smidt et al. [52] applied SIMCA to classify waste materials using Fourier transform infrared (FTIR) spectroscopy. In a further application, Candolfi et al [53] demonstrated the effectiveness of SIMCA in the identification of the excipients used in pharmaceutical process.

Unsupervised cluster analysis techniques do not require reference clusters to be known a-priori. There have been many unsupervised cluster analysis techniques that have been proposed. One commonly used unsupervised cluster analysis technique is fuzzy-C mean (FCM) variant [54]. FCM is a method of clustering that allows one data vector to belong to one or more clusters. It has been used to analyze near-infrared spectroscopic imaging data. For example, Wang et al [55] applied FCM in the analysis of Raman spectral imaging data, where it was used to extract chemical information from biological specimens. In [54], the FCM method was applied in the analysis of near-infrared spectroscopy imaging data. In their study, FCM was able to probe the spatial and spectral characteristics of the samples and could detect small sections of the image that contained novel and unanticipated spectral features.

Multivariate Curve Resolution

Multivariate curve resolution is a technique which resolves the components of a mixture by determining the response profiles (spectra, time profiles, PH profiles) and the

estimated concentrations. No prior information regarding the composition of the mixtures is required when using this technique. The most popular resolution method used in imaging analysis is multivariate curve resolution-alternating least squares (MCR-ALS) analysis [56]. MCR-ALS first estimates the number of compounds in a mixture and the spectral profiles and concentrations of each component using singular value decomposition (SVD) [57]. Following this, alternating least squares (ALS) [58] is applied to optimize the spectral profiles and the corresponding concentrations. MCR-ALS has been applied in analyzing NIR chemical imaging data to estimate the concentration of analytes in each pixel of a chemical image [59]. Azzouz and Tauler [60] applied MCR-ALS in the resolution and quantification of various analytes in different types of pharmaceutical and agricultural samples.

Partial Least Squares (PLS)

Partial least squares (PLS) [61, 62] is a method for finding a linear regression model by projecting the predicted variable matrix Y and the observed variable matrix X to a latent variable space. A PLS model is identified using a set of data containing measurements of X and Y. Once identified, this model can take new X measurements and use them to estimate the value of the output Y. PLS has been applied in the analysis of chemical imaging data. Burger and Geladi [63], for example, identified a PLS model based on reference materials of known composition and spectra from a hyperspectral image. The PLS model was then successfully used to predict unknown compositions from new images. Van den Broek et al. [64] applied PLS as a data decomposition technique for the classification of spectroscopic NIR images. In this study, PLS reduced the dimension of the data and increased the speed of online classification.

Independent Component Analysis (ICA)

Independent component analysis (ICA) [65, 66] is a method used to separate multivariate signals into sets of statistically independent non-Gaussian source signals. ICA has been widely used in blind signal separation [65-69]. It has also shown its capabilities in analyzing spectroscopic datasets. ICA has been applied in identifying the constituent compounds in commercial gasoline from infrared (IR) spectroscopy data [70, 71]. In a

further study, Kernel independent component analysis was applied to estimate the source spectra of the active components in samples of acetylspiramycin [72]. In this study, ICA was found to be capable of isolating the active components, excipients and other minor components. ICA was applied by Polder et al. [73] to estimate the important compounds in ripening tomatoes. Two primary independent components found in this study resembled the actual absorption spectra of the most important compounds.

All the multivariate statistical analysis methods described above have shown their capabilities in extracting important information from chemical imaging data in various industries. The application of multivariate statistical analysis to chemical imaging data in the pharmaceutical industry is still in its infancy, but applications are gradually increasing. The following section provides a literature review on the application of multivariate statistical analysis to chemical imaging data collected from pharmaceutical tablets.

2.1.3 Application of Multivariate Statistical Analysis to Chemical Imaging of Pharmaceutical Tablets

In the pharmaceutical industry, chemical imaging techniques have been applied in many areas, as described in Section 2.1.1. One important application area for development is in the imaging of solid dosage formulations [74-76]. In this area, Raman and NIR spectroscopy have been applied in the analysis of pharmaceutical tablets, with the aim being to identify the components and concentrations present in the tablets. The resulting chemical imaging data is complex and difficult to interpret because the particles in the tablets are of multiple types and mixed together. Extracting the quantitative information regarding the content of the active pharmaceutical ingredients (APIs) and the spatial distribution of the APIs in the tablet has become a key research area.

Applications of multivariate statistical analysis methods to extract chemical information from a pharmaceutical tablet have been presented by a number of authors. Šašić [77] has studied the application of Raman and NIR imaging techniques to determine the spatial distribution of five components in a pharmaceutical tablet. In this work, PCA was applied

to identify the reference spectra of the components and their corresponding concentrations in the tablet. Composite images (all components displayed in a single image) were produced to show the spatial distribution of all the individual compounds in a tablet. Similarly, Clark and Šašić [42] applied PCA to analyze the chemical imaging data of a pharmaceutical tablet. Zhang et al [78] applied four different multivariate data analysis methods to extract chemical information from the imaging data. PCA, cluster analysis, direct classical least squares (DCLS) and multivariate curve resolution (MCR) methods were applied to provide spatial distribution of the compounds of a pharmaceutical tablet. The relative merits and drawbacks of these methods were evaluated and compared.

PCA is a very popular multivariate statistical analysis method and has been applied in the analysis of chemical imaging data collected from pharmaceutical tablets. In these applications it has been shown to be effective in terms of its ability to identify the reference spectra of the constituent compounds from the imaging data [38], and has been applied to identify minor compounds in Raman imaging [39, 79]. Furthermore, it has been shown that through analysis of the PCA model it is possible to identify the spatial distribution and concentration of constituent compounds in the tablet [38].

Unfortunately, PCA is not always capable of identifying the constituent compounds in a tablet. This is particularly the case when the spectra of some of the compounds overlap. The phenomenon has been widely encountered in PCA applications and is referred to as “rotational ambiguity” [78].

Other multivariate data analysis methods also have their disadvantages when applied to chemical imaging data. PLS, for example, requires the reference spectra of the constituent compounds to be available [80], which is frequently unavailable. Cluster analysis has the problem of having to determine an appropriate number of clusters [81]. Selecting an unsuitable number of clusters can lead to poor results and the component spectra will not be identified. MCR-ALS has the drawback that there will be more than one solution, and

therefore cannot guarantee that the resolved component spectra will match the true component spectra [82].

Similarly to PCA, Independent Component Analysis (ICA) separates a multivariate signal, such as a spectroscopic measurement, into its constituent subcomponents, e.g. constituent compounds. However, ICA identifies mutually statistically independent components, rather than simply uncorrelated ones, as is the case with PCA [83]. In probabilistic theory, independence is a high-order statistic and it is a much stronger condition than orthogonality. Also, the condition of independence in the components allows ICA to overcome the rotational ambiguity problem encountered with PCA. It is important to note, however, that PCA and ICA are closely related. In fact, ICA can be considered as a "fine-tuning" form of PCA, since it rotates the principal components in order to remove high-order dependencies between the source signals [66, 83]. Also, PCA and ICA are related from a computational perspective, since they both rely on an eigenvalue decomposition for the identification of their corresponding models.

In [84], Shashilov et al. demonstrated ICA as a powerful tool for analyzing Raman spectra. In this study, ICA was able to extract the spectrum of β-sheet from Raman spectra while PCA and multivariate curve resolution could not separate the β-sheet constituent as an individual component. The applications of PCA and ICA in analyzing Raman spectra on paraffin-embedded skin biopsies has been compared in [85]. This work showed that PCA did not give good estimations of spectra and the associated concentration profiles. However, ICA was found to provide an improved estimate of the spectra of the chemical constituents. The estimated concentration profiles are not orthogonal and only have positive values, contrary to PCA. Chapter 3 will apply both PCA and ICA to the chemical imaging data of a common pharmaceutical tablet. Their capabilities in identifying the reference spectra of the constituent compounds and their corresponding concentrations are compared in this study.

2.2 Batch Process Control

An area which is likely to have a considerable impact on the PAT initiative is process control. Process control is used extensively in a variety of industries to ensure that the output of a process conforms to required specifications. Industrial processes are usually categorized into continuous or batch [86]. A continuous process is a process where feeds are continuously sent to the equipment, and product is continuously collected. The by-products are removed continuously. A batch process is a process in which a finite quantity of product is made over a period of time. Batch processes usually involve placing raw materials into a vessel, and then performing a series of unit operations (mixing, heating, reaction, distillation, etc). This is followed by the removal of the product, and cleaning of equipment. The process variables, such as heat, mass, temperature and concentration, vary over the duration of the batch, according to a "recipe".

Batch processes are gaining ever increasing importance in the manufacturing industries. They are particularly prevalent in the polymer, pharmaceutical and specialty chemicals industries where the focus is on the production of low-volume, high-value added products. Although prevalent in continuous operations, advanced process control is only now being explored in batch manufacturing. Examples now exist in specialty chemical production [87, 88], food [89], pharmaceuticals [90, 91], biochemicals [92], polymers [93], coatings [94] and many other processes.

The advantages of batch processing are [95, 96]:

1) It is flexible and able to manage many different product grades.
2) The process is repeated many times, and therefore the quality of production can be improved on a run-to-run basis.
3) Processing is typically very slow, and therefore improvements to operational conditions can be computed in real time.
4) It is economical for small volume manufacturing.
5) It is ideal if equipment needs regular cleaning or sterilization, as with many bio applications.

6) It can be scaled-up from laboratory to industrial application.

Unfortunately, batch processing introduces many difficulties not encountered in continuous production: the processing of raw materials into the end product takes place over a finite duration; there are rarely steady states conditions; process dynamics are typically time-varying and non-linear; quality measurement are often only available at the end of the batch [97].

Quality control of batch processes is usually implemented by regulating several process variables, such as temperature and pH, with the hope that keeping these variables fixed will ensure consistent end-point product quality. Unfortunately, variation in raw material properties means that this style of operation may not produce consistent product.

Advanced control methods have been shown to improve product consistency. Kravaris et al [98], for example, used globally linearizing control (GLC) for trajectory tracking. Clarke-Pringle and MacGregor [99] used a nonlinear adaptive controller, incorporating an extended Kalman filter to provide temperature control in a batch reactor. Wang et al [100] applied adaptive control, together with an extended Kalman filter, to a simulated batch styrene polymerization reactor to track a specific output in the presence of model uncertainty. Aziz et al [101] applied generic model control (GMC) coupled with a neural network, which estimated heat-release to track an optimal temperature profile.

In addition to regulating process variable measurements, there have been several studies in to optimizing process variable trajectories in batch processes. In these applications, the trajectories of process variables, such as temperature, were optimized at various decision points during the batch in an attempt to maintain final product quality. Crowley and Choi [102], for example, controlled the polymer weight chain length distribution by computing an optimal discrete sequence of reactor temperature set points using an optimization program based on sequential quadratic programming. Ruppen et al [103] successfully applied an on-line optimization strategy to the acetoacetylation of pyrrole with diketene in a laboratory-scale reactor.

However, it is shown in Chapter 4 of this thesis and in Lin et al. [104] that even in a simple chemical batch reactor, maintaining important process variables, such as reactor temperature at their set-points does not guarantee that the final product will meet its specifications. Changes in reaction rates, for example, caused by differing raw material properties can alter the reaction pathways and affect final product quality. Chapter 4 of this thesis proposes a novel control algorithm which is able to ensure product quality, even when the process is subject to large disturbances.

In addition to trajectory control, many research studies have investigated 'end-point' control of batch processes. End-point control attempts to optimize the operating conditions during the entire batch to ensure the quality of the product at the end point (end-point quality) meets target requirements. Yabuki et al. [105], for example, used simple empirical regression models to predict final product properties. If the prediction fell outside of a "no-control" region, a midcourse correction was made to bring the end product quality back to the target using online and offline measurements available up to the midpoint of the process. Flores-Cerrillo and MacGregor [5] proposed a method to regulate the complete quality variable trajectory by controlling the process in the reduced space of a latent variable model rather than in the real space of the manipulated variables. At various decision points during the batch, the quality of the end product was predicted using online and offline process variables measurements and corrective action was made, if necessary, to guarantee the quality of the end product. This end-point controller was successfully applied to regulate a simulated batch process. Pan and Lee [106] developed a data-based model to make online predictions of end-point product quality, and adjustments were then made to achieve desired product quality. The authors applied their technique to a methyl methacrylate (MMA) polymerization process. End product quality was predicted recursively based on reactor cooling rate, and a model predictive control approach was used to manipulate the reaction temperature.

2.3 Summary

This chapter has provided a literature review describing the application of multivariate statistical analysis methods to analyze chemical imaging data collected from pharmaceutical processes. Chemical imaging techniques, commonly used in the pharmaceutical industry, such as NIR and Raman spectroscopy were fist discussed. Then, the application of multivariate statistical analysis to data collected from these devices was reviewed. Finally, an introduction and a brief review of batch process control were provided.

3 Application of Multivariate Statistical Analysis to Raman Images of a Common Pharmaceutical Tablet

In this chapter, the application of multivariate statistical techniques to identify the composition of a pharmaceutical tablet is explored. To begin, two frequently applied multivariate techniques – independent component analysis (ICA) and principal component analysis (PCA) – are described. These two algorithms are then applied to two sets of simulated data to compare what information they can extract from spectral data samples. Following this, PCA and ICA are compared in terms of their capabilities to identify the composition of a pharmaceutical tablet from Raman spectroscopic data.

The chapter is divided into the following sections:

1) A description of PCA and ICA algorithms is provided.
2) The application of PCA and ICA to two sets of simulated datasets is presented.
3) PCA and ICA are applied to Raman image data collected from a common pharmaceutical tablet.
4) A summary of the chapter is provided.

3.1 Preliminary Methodology

As discussed in Chapter 2, PCA and ICA have been shown to have great potential in the analysis of high-dimensional datasets in the pharmaceutical industry, where production lines are becoming increasingly instrumented. The following sections describe PCA and ICA in detail. This is followed by the application of PCA and ICA to two sets of simulated datasets and then Raman spectroscopic data collected from a pharmaceutical tablet.

3.1.1 Principal Component Analysis

For a dataset with a large number of correlated variables, the power of principal component analysis (PCA) lies in its ability to condense the correlated information from hundreds or even thousands of dimensions into a small number of orthogonal principal components (PCs).

Given a data matrix X with m rows and n columns, with each sample being a row and each variable a column, PCA first identifies the direction in the data matrix that explains the greatest variance. Then the direction with the second largest variance is identified and so on. Each of the directions is chosen so that they are orthogonal to each other. And these new directions then become a new set of axes. The mathematical description of the PCA algorithm is shown as follows:

$$X = \sum_{k=1}^{n_c<n} t_k p_k^T + E \tag{3.1}$$

where t_k are referred to as score vectors and p_k are the loading vectors. E represents the information contained within the matrix X that is not represented in the first n_c principal components (a single principal component being the combined t_k and p_k pair).

t_k and p_k can be identified from Equation (3.1) using Singular Value Decomposition (SVD) [44], where the loading vectors p_k are defined as the eigenvectors of the covariance matrix: $\text{cov}(X) = \frac{X^T X}{m-1}$. Therefore, for each p_k

$$\text{cov}(X) p_k = \lambda_k p_k \tag{3.2}$$

Where λ_k is the eigenvalue of the covariance matrix $\text{cov}(X)$. The loading vectors p_k are orthonormal, while the score vectors t_k are orthogonal. For X and any p_k, t_k pair

$$X p_k = t_k \tag{3.3}$$

Thus, the score vectors t_k are the projection of X onto the loading vectors p_k.

The number of dimension is reduced in the data when n_c is less than n. n_c is an important parameter that must be determined when using PCA. It should be selected as small as possible without losing significant information in the dataset. There are several methods that have been proposed for selecting n_c. The simplest method is to analyze the variance captured by the n_c PCs [24, 107, 108]. The percentage of variance captured by n_c PCs is defined as:

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_{n_c}}{\lambda_1 + \lambda_2 + ... + \lambda_n} \times 100\% \tag{3.4}$$

Some studies suggest that n_c can be selected when the percentage of variance explained by adding additional principal components is smallest [45]. However, this can lead to over-fitting [45].

A more reliable method for selecting n_c is cross-validation [45]. A number of cross validation techniques have been proposed. A common approach involves dividing the dataset into P groups of equal size. One group of data is used to validate the model with the remaining $P-1$ groups used as training data. This is repeated P times, until every group is used once for validation. In the validation procedure, the prediction errors are calculated for each group of data and summed. This error is defined as $S(k)$ for k PCs, and the number of PCs is selected such that $S(n_c)$ is minimized.

In the following section, the algorithm for independent component analysis is detailed described.

3.1.2 Independent Component Analysis

Introduction

Independent component analysis (ICA) [66, 83, 109, 110] is a statistical method for finding latent signals that are statistically independent and nongaussian from a collection of data containing mutual signals. The ICA model can be expressed as

$$X = f(S) \tag{3.5}$$

were $X = (x_1,...,x_m)^T$ is an observed vector, $S = (s_1,...,s_n)$ is a vector with statistical independent latent variables, and f is a general unknown function. The model can be reduced to Equation (3.6) when the function is linear

$$X = AS \tag{3.6}$$

where A is an unknown $m \times n$ mixing matrix.

The task of the ICA algorithm is to identify the "mixing matrix" A, and the independent latent variables, S.

When using ICA, the linear model Equation (3.6) is identifiable under the following restrictions [111]:

- Independence: the source variables are assumed to be statistically independent.
- Nongaussianity: the independent components must have no more than one Gaussian distribution.
- The mixing matrix A is square and invertable.

Principles of ICA Estimation

The aim of the ICA algorithm is to find the independent components s_i and the mixing matrix A. There have been a number of ICA algorithms that have been proposed. The most common approach to solving the ICA problem is to firstly find a demixing matrix W, such that y_i in Equation (3.7) are estimates of s_i. Thus, W is an estimate of the (pseudo) inverse of A.

$$Y = W^T X = \sum_i w_i x_i \tag{3.7}$$

There are several algorithms that have been proposed to estimate independent components and mixing matrices. Two important criteria [66, 83] for the algorithms are: maximization of non-Gaussianity of the components and minimizing mutual information. These two criteria are now described.

Maximization of Non-Gaussianity of the Components

According to the central limit theory [112], the sum of independent non-Gaussian random variables is closer to being Gaussian than for the original random variables. Thus, a linear combination, $y_i = \sum_i z_i s_i$, of independent components s_i is maximally non-Gaussian if it is an independent variable. The task is then focused on finding w_i such that $y_i = \sum_i w_i^T x_i$ is as far from Gaussian as possible.

Non-Gaussianity can be measured by Kurtosis. The kurtosis of Y is defined as:

$$kurt(Y) = E(Y^4) - 3(E(Y^2))^2 \tag{3.8}$$

When the variables have positive Kurtosis, they are called subgaussian, and those with negative Kurtosis are called supergaussian.

Minimizing Mutual Information

Two variables are statistically independent when their mutual information is zero. Mutual information can be measured by entropy of the variables [109], where entropy of a discrete random variable Y is defined as:

$$H(Y) = -\sum_i p(Y = a_i) \log p(Y = a_i) \tag{3.9}$$

And the entropy of a random variable y with density $f(y)$ is defined as:

$$H(y) = -\int f(y) \log f(y) dy \tag{3.10}$$

When the variance of the variables is equal, the Gaussian variable has the largest entropy. Therefore, entropy can be used as a measure of non-gaussianity.

Negentropy [109] is another way to measure entropy. The negentropy of y is defined as:

$$J(y) = H(y_{gauss}) - H(y) \tag{3.11}$$

Where y_{gauss} is a Gaussian variable. Negentropy is zero for Gaussian variables. Thus, to maximize the independence between variables, the negentropy of the variables must be maximized.

Data Preprocessing for ICA

It is typically recommended that certain preprocessing methods be used to reduce the dimensionality of the data before applying ICA. It has been shown that poor results can be obtained when ICA is applied to a high dimensional data set without first reducing its dimensionality. This is particularly so when several of the original dimensions contain noise only [109]. It is therefore considered to be standard practice to apply PCA to reduce the dimensionality of the data, prior to using ICA.

Further to reducing the dimensionality of the data, it is also common to scale the vectors to zero mean and decorrelate them. The second-order dependencies between the observed vectors are removed by the decorrelation process. This procedure is also called whitening [109]. A zero-mean random vector $z = (z_1 ... z_n)^T$ is said to be white if its elements z_i are uncorrelated and have unit variance $E\{z_i z_j\} = \delta_{ij}$. Whitening transforms a random vector x into a whitened vector, z, by multiplying it by a linear matrix V:

$$z = Vx \tag{3.12}$$

Assuming the covariance matrix of the vector, x, is $C = E\{xx^T\}$, and the eigenvectors of the covariance matrix is $E_1 = (e_1, ... e_n)$, and the diagonal matrix of the eigenvalues of C is $D = diag(d_1 ... d_n)$, the linear matrix V is given by:

$$V = D^{-1/2} E_1^T \tag{3.13}$$

which indicates that z can be obtained by

$$z = D^{-1/2} E_1^T x \tag{3.14}$$

where $E_1 = (e_1 ... e_n)$ consists of a coordinate, and $E_1^T x$ is the projection of x in the coordinate. Multiplication by $D^{-1/2}$ and $E_1^T x$ tries to make the variance of the projections in $e_1 ... e_n$ of unit dimension. Each column of the matrix $E_1 = (e_1 ... e_n)$ denotes an eigenvector of n dimensions, and there are n such eigenvectors. The vector z is obtained by dividing each element of a column of E_1 by the square root of the corresponding eigenvalue ($z_{ij} = e_{ij} \lambda_j^{-1/2}$).

FastICA Algorithm

In the previous sections, the principles of ICA estimation and preprocessing techniques were introduced. In this section, the ICA algorithm itself is described. There are several different algorithms available for performing ICA. FastICA [109] is an efficient and popular algorithm for independent component analysis, and has become almost the standard algorithm for ICA applications. It is therefore the algorithm used throughout this work.

The procedures followed by FastICA [109], to find one independent component is as follows:

(1) Choose an initially random weight vector w.

(2) Let $w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w$.

(3) Let $w = w^+ / \|w^+\|$.

(4) Return to (2) until convergence is achieved.

The process above is repeated for all components that are identified. This can be achieved in parallel. However, $w_1^T x, ..., w_n^T x$ should be decorrelated before the next iteration.

3.2 Multivariate Data Analysis of Spectroscopic Data

In this section, PCA and ICA methods are applied to two sets of simulated data. The purpose of this exercise is to compare how well each of these techniques is able to identify the reference spectrum from a data set containing mixed spectra.

3.2.1 Generation of Data and Pre-processing

Generation of Data

Each simulated data set contained 500 measurements. Each measurement represented a spectral sample containing 200 spectral channels or wavenumbers. In each set of data four reference spectra were identified, with each spectrum containing a random number of 'peaks', with random location and width. Each measurement in the data sets was a linear combination of these reference spectra. Fig. 3.1 shows a plot of the four reference spectra (RS1, RS2, RS3 and RS4) for the first dataset, and Fig. 3.2 shows the reference spectra for dataset 2.

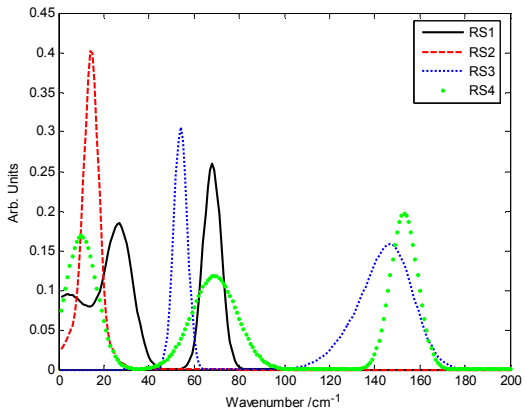

Figure 3.1: Reference Spectra of the Constituent Compounds of the Dataset 1

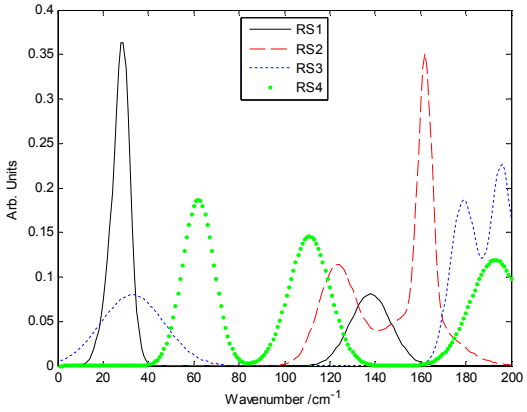

Figure 3.2: Reference Spectra of the Constituent Compounds of the Dataset 2

The concentrations of the 4 constituent compounds for each of the 500 samples were uniformly distributed random numbers ranging between 0.0 and 1.0. Individual

concentration values were made to conform to the "closure constraint", which states that all concentrations of all the constituent compounds sum up to 1 for every sample. The closure constraint was applied in the simulated datasets and was used to verify the estimated concentration.

Finally, the 'measured' spectra were created by multiplying reference spectra of the constituent compounds with the corresponding concentrations.

Pre-processing

Before PCA and ICA were applied to the 'measured' spectra, pre-processing of the data was undertaken. The main goal of the preprocessing was to remove the non-chemical information from the spectra. Autoscaling [113, 114] was used to convert the 'measured' spectra into normally distributed spectra with zero mean and unit variance.

Autoscaling [113, 114] adjusts a vector to a standard level by subtracting the mean of the vector and dividing the residual by the standard deviation of the vector. The autoscaling method converts the data to have zero mean and unit variance. The purpose of normalization is to remove systematic variation, which can mean some variables have higher amplitudes and variation, which can affect identification results. Without autoscaling, the variable with large absolute range variance will dominate the modeling process, and caused the inaccuracy modeling.

The auto-scaling formula is shown below:

$$\hat{D} = (D - \bar{D}) / \mathrm{var}(D) \tag{3.15}$$

Where D is the original data, $\bar{D}$ is the mean of the original data, $\hat{D}$ is the normalized data.

3.2.2 Results

In this example no measurement noise was applied to the data and hence both PCA and ICA models were developed with four principal/independent components according to Equation (3.1) and (3.6). Once the models were identified from the data, the loading vectors were analyzed, with the expectation that they would resemble the reference spectra. The concentrations of the constituent compounds in each sample were then estimated through analysis of the scores associated with each loading vector.

Results for Dataset 1

Plots of the loading vectors, corresponding to the PCA and ICA models, and the reference spectra most related to the constituent compounds present in dataset 1 are shown in Figure 3.3-Figure 3.6.

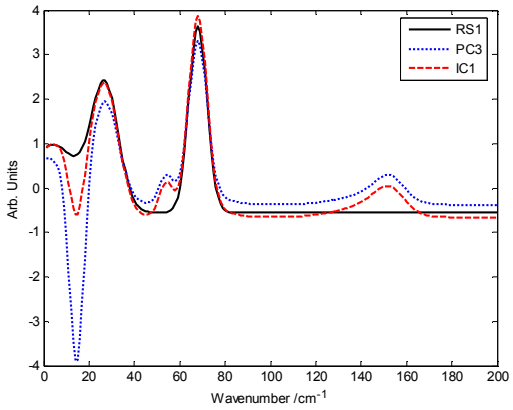

Figure 3.3: Plot of the 1st Reference Spectrum (RS1), 3rd PC and 1st IC

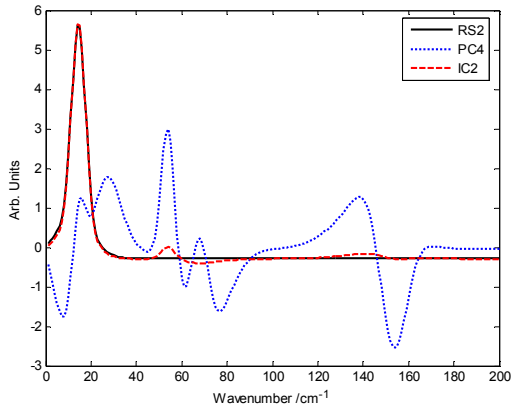

Figure 3.4: Plot of the 2nd Reference Spectrum (RS2), 4th PC and 2nd IC

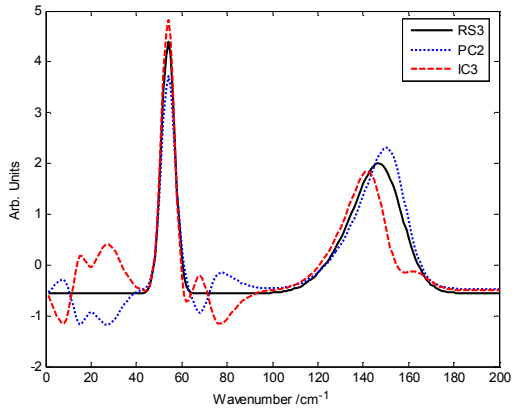

Figure 3.5: Plot of the 3rd Reference Spectrum (RS3), 2nd PC and 3rd IC

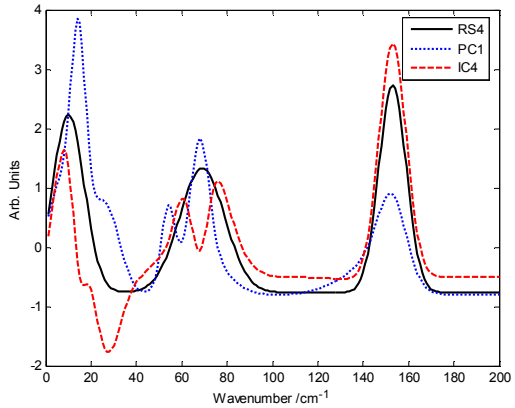

Figure 3.6: Plot of the 4th Reference Spectrum (RS4), 1st PC and 4th IC

Figure 3.3, 3.5, and 3.6, show that both loading vectors from PCA and ICA are able to approximate the 1st, 3rd and 4th reference spectra. However, it is shown in Figure 3.4 that,

while the 4th IC did manage to approximate 2nd reference spectrum, the PCA model failed in identifying this spectrum. Figures 3.3 to 3.6 also show that while the ICA model is able to estimate each of the reference spectra, there are significant differences between the actual and estimated spectra. The reason for this is that the reference spectra were not completely independent and there is some overlap between the spectra.

Subsequently, a comparison was made between PCA and ICA, in terms of their abilities to estimate the relative concentrations of the constituent compounds present in the analyzed samples. This comparison was conducted by applying correlation analysis between the estimated and the actual concentrations, the results of which are shown in Table 3.1. Correlation analysis was applied to provide a measure of similarity between the actual and estimated constituent compound. This table shows that ICA is more accurate when compared to PCA in terms of its ability to accurately estimate the relative concentration of each of the constituent compounds. In particular, the correlation between the true concentrations and those estimated using the ICA model is consistently greater than or equal to 0.81, with correlations below 0.2 for two of the spectra when using PCA.

Table 3.1: Cross-Correlation Coefficients between the Actual Concentrations of the Constituent Compounds and their Estimates Obtained using PCA and ICA Models for Dataset 1

	RS1	RS2	RS3	RS4
ICA	0.93	0.94	0.99	0.81
PCA	0.78	0.19	0.90	0.16

It is perhaps surprising that the PCA model clearly failed to accurately estimate the concentrations of the 4th constituent compound, even though Figure 3.6 demonstrates that it managed to capture the main features of this reference spectrum (RS4). This finding demonstrates that even if a particular reference spectrum is successfully inferred using a particular IC or PC it does not necessarily follow that the corresponding score values will accurately reflect concentration values of the corresponding constituent compound.

Results for Dataset 2

The ability of the PCA and ICA models to infer reference spectra of the constituent compounds present in dataset 2 is demonstrated in Figure 3.7 -Figure 3.10.

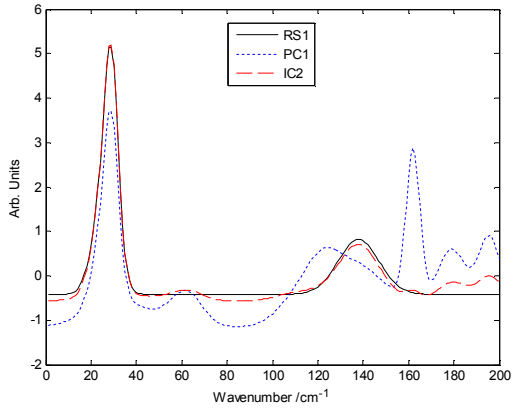

Figure 3.7: Plot of the 1st Reference Spectrum (RS1), 1st PC and 2nd IC

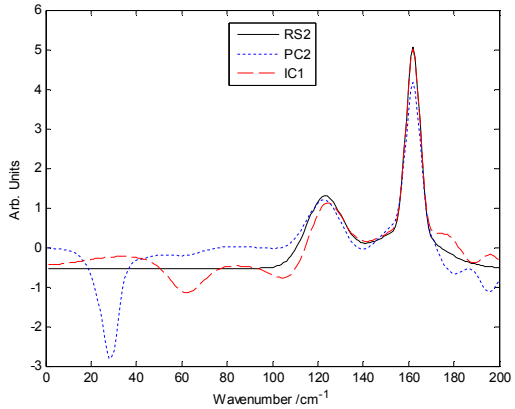

Figure 3.8: Plot of the 2nd Reference Spectrum (RS2), 2nd PC and 1st IC

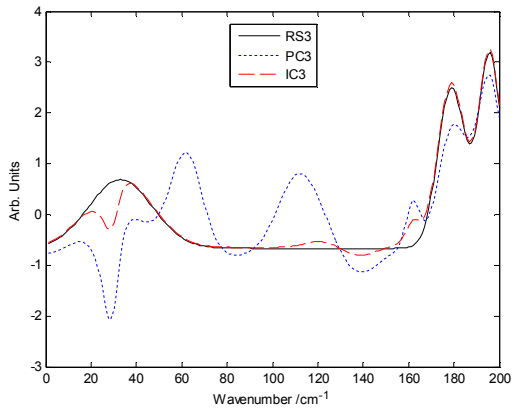

Figure 3.9: Plot of the 3rd Reference Spectrum (RS3), 3rd PC and 3rd IC

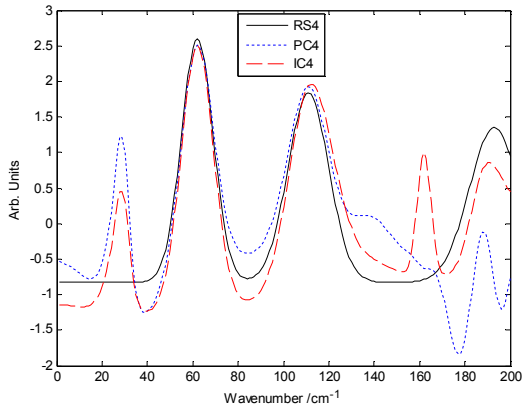

Figure 3.10: Plot of the 4th Reference Spectrum (RS4), 4th PC and 4th IC

Figure 3.7 – 3.10 demonstrate that both PCA and ICA managed to identify the reference spectra of all the constituent compounds present in dataset 2, to varying degrees of accuracy.

As with the previous data set it is evident that PCA is less able to accurately identify the independent spectra from the data than ICA. The results suggest that PCA is not as capable as ICA in separating out the independent spectra. For example, PC1, shown in Figure 3.7, identifies features related to three different reference spectra (RS1, RS2 and RS3). Also, PC3, shown in Figure 3.9, captures reference spectrum RS3 as well as the two peaks contained in the reference spectrum RS4.

As with the previous example, the ICA and PCA models were analyzed further and the concentrations of each compound in the samples estimated. The correlation coefficients between the estimated and actual concentrations are provided in Table 3.2.

Table 3.2: Cross-Correlation Coefficients between the Actual Concentrations of the Constituent Compounds and their Estimates Obtained using PCA and ICA Models for Dataset 2.

	RS1	RS2	RS3	RS4
ICA	0.97	0.98	0.96	0.94
PCA	0.67	0.96	0.54	0.78

Table 3.2 shows that the ICA model is capable of estimating the concentrations of all the constituent compounds with a high level of accuracy. In fact, the cross-correlation between the actual concentrations and those estimated using ICA is greater than or equal to 0.94. In contrast, the correlation coefficient for three out of four constituents is less than 0.9 when using PCA. Furthermore, PCA provided particularly poor estimates for the first and third constituent compounds, with the cross-correlations coefficients being equal to 0.67 and 0.54, respectively. This poor performance is believed to be because of the lack of selectiveness observed with the loadings vectors for PC1 and PC3, which contained features from several different reference spectra. The limited capabilities with PCA are expected to be generic, which suggests that for spectral applications, ICA is likely to be the better algorithm to apply.

3.3 Application of Multivariate Statistical Analysis to Raman Images of a Pharmaceutical Tablet

Using data collected from a Raman spectroscopic analysis of a real pharmaceutical tablet, this section compares the relative abilities of PCA and ICA to detect the constituent compounds within the tablet and to estimate their concentrations and spatial distribution.

3.3.1 Data Acquisition and Pre-processing

Data Acquisition

In this chapter, Raman data collected from a Renishaw Ramascope System 1000 using Wire V.1.3 software are analyzed. The spectra were obtained by exciting one tablet with a laser line at 782 nm and 100 mW. The composition of the tablet was known a-priori and was as follows: Avicel (microcrystalline cellulose) at 48% w/w; API (active pharmaceutical ingredient) at 24% w/w; DCP (di-calcium phosphate) at 24% w/w, Explotab (sodium starch glycolate) at 3% w/w; MgSt (Magnesium Stearate) at 1% w/w. Therefore, the main constituent compounds were Avicel, API and DCP, with Explotab and MgSt minor components. Data were collected through a 20× objective. Images with

an area of approximately $2 \times 2 \ mm^2$ were obtained with a spatial resolution of $25 \ \mu m$. Typical acquisition time was approximately 21 hours. Overall 6000 spectra, measured at 576 different spectral channels, i.e. wavenumbers, were used in the following statistical analysis.

Chemical imaging data obtained by Raman spectroscopy can be considered to be a three-dimensional cube with two spatial dimensions, known as a hypercube, and a third spectral wavelength dimension (See Figure 3.11) [115, 116]. The hypercube can be viewed as a series of spatially resolved spectra (called pixels) or, as a series of spectrally resolved images (called image planes or channels). Choosing one pixel, the intensity for this pixel plotted as a function of the wavelengths is the standard Raman spectrum, or a single-pixel spectrum, shown in the lower left of Figure 3.11. Choosing one wavelength, the intensities for all pixels is an image of that location, which is referred to as a single-channel image, which is shown in the lower right of Figure 3.11.

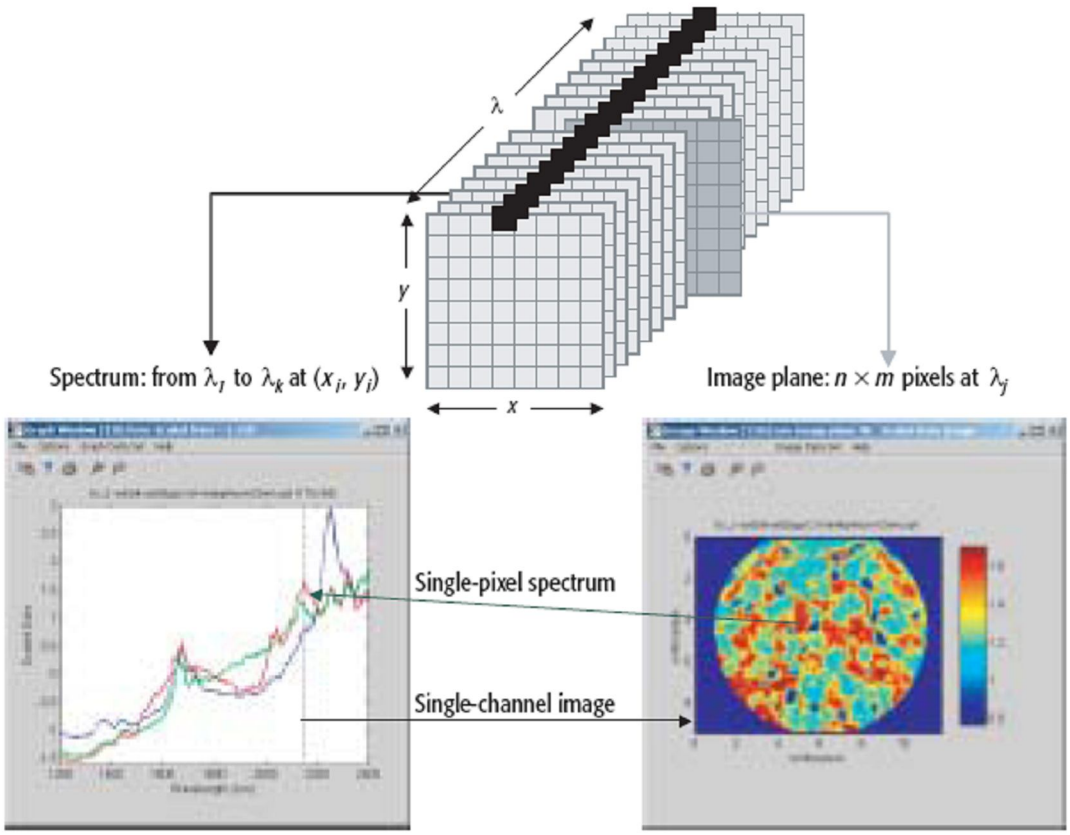

Figure 3.11: Schematic Representation of a Spectral Imaging Hypercube Showing the Relationship between Spatial and Spectral Dimensions [116].

The collected spectral data forms what is known as a three-way matrix. Unfortunately, most statistical methods are applicable to two-dimensional data sets only [117, 118]. Although three dimensional methodologies such as PARAFAC [119], are available, they are very complex and require significant computation, which restricts their use. Studies have shown that there are benefits in describing three-way curve resolution problems in a two-way form. The unfolding procedure [120-122], also called reorganization, is used to change the three-way array into a two-way array, as shown in Figure 3.12. Following this transformation, traditional curve resolution methods can be applied to the two-way array.

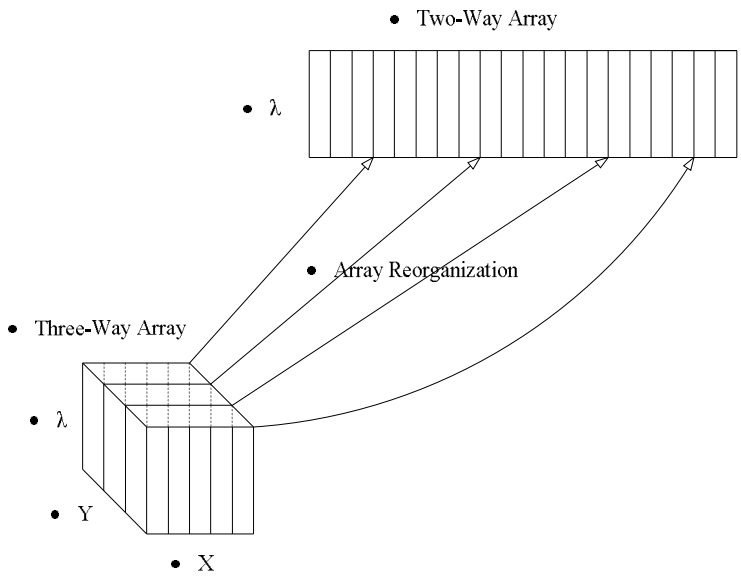

Figure 3.12: Diagrammatic Representation of Three-way Array Conversion into a Two-way Array by Array Reorganization (Also Called Array Unfolding)

Reference Spectra for the Pharmaceutical Tablet

Reference spectra were obtained by scanning the pure constituent compounds. Figure 3.13 and Figure 3.14 display the reference spectra of the five constituent compounds.

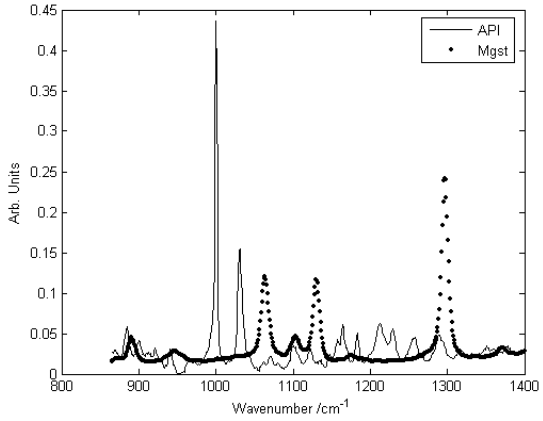

Figure 3.13: Plot of the Reference Spectra of API and MgSt

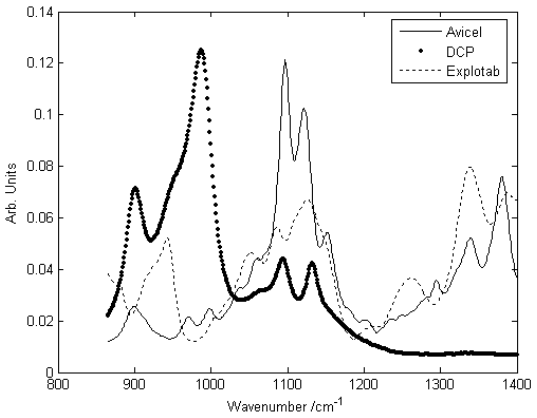

Figure 3.14: Plot of the Reference Spectra of Avicel, DCP and Explotab

Note that there is a substantial overlap between the reference spectra of Avicel and Explotab. There is also some overlap between Avicel, Explotab and DCP in the wavenumber range 1050-1200. These observations were confirmed by performing correlation analysis (cross-correlation coefficient between the Avicel and Explotab reference spectra was found to be equal to 0.6745). The correlation coefficients between the reference spectra are shown in Table 3.3. Such significant levels of correlation were expected to make the subsequent analysis of the data a challenging task for multivariate statistical methods [78].

Table 3.3: Correlation Coefficients between Reference Spectra

	API	Avicel	DCP	Explotab	Mgst
API	1.0000	-0.1065	0.0602	-0.1782	-0.0680
Avicel	-0.1065	1.0000	-0.2012	0.6745	0.2120
DCP	0.0602	-0.2012	1.0000	-0.3638	-0.1525
Explotab	-0.1782	0.6745	-0.3638	1.0000	0.0909
Mgst	-0.0680	0.2120	-0.1525	0.0909	1.0000

Pre-processing

As with the earlier simulated examples, the data was firstly autoscaled to zero mean and unit variance. This removes the influence of system variance on the image data. Following this, a Savitzky-Golay filter [114, 123-125] was used to smooth the image data.

The details of autoscaling and Savitzky-Golay Filter method are described in the following section.

Savitzky-Golay Filter

The Savitzky-Golay filter [114, 123-125] is a smoothing filter which performs a polynomial regression on a set of data within a window to determine the smoothed value for each point. To smooth a set of data using the Savitzky-Golay Filter algorithm, the order and frame size of the window should be chosen first. A high polynomial order tends to be good at keeping narrow and high features, while a low polynomial order is good at smoothing broad features. A small frame size tends to be good at keeping narrow and high features, while large frame size is good at smoothing broad features. After deciding upon the order and frame size, least squares is used to find the polynomial curve that best approximates the data in fixed frame size windows. The window contains several data points including the central point to be smoothed and points in the neighborhood. The data within the window are used to smooth the central point using the least squares algorithm, while the neighboring points themselves do not change. The neighboring points are adjusted as the window moves.

When the frame size of the window is $2N+1$ and the order of polynomial is m, the polynomial curve is of the form.

$$f_i = \sum_{k=0}^{k=m} b_{mk} i^k = b_{m0} + b_{m1} i + b_{m2} i^2 + ... + b_{mm} i^m \tag{3.16}$$

where i changes from $-N$ to N. f_i is substituted by the original data D_i to calculate the parameters $b_{m0}, b_{m1}, b_{m2}, ..., b_{mm}$, using least squares.

The derivatives of this equation are:

$$df_i / di = b_{m1} + 2b_{m2} i + 3b_{m3} i^2 + ... + m b_{mm} i^{m-1} \tag{3.17}$$

$$d^2 f_i / di^2 = 2b_{m2} + 3*2*b_{m3} i + ... + m*(m-1)*b_{mm} i^{m-2} \tag{3.18}$$

.....

$$d^m f_i / di^m = m!*b_{mm} \tag{3.19}$$

Since the data being smoothed is the central point, $i = 0$. The value of the m -th derivative for the central point is as follows.

$$f_0 = b_{m0} \tag{3.20}$$

$$df_0 / di = b_{m1} \tag{3.21}$$

$$d^2 f_0 / di^2 = 2b_{m2} \tag{3.22}$$

.....

$$d^m f_i / di^m = m!*b_{mm} \tag{3.23}$$

Using the Savitzky-Golay Filter, the n -th derivative of the spectra can be determined. In spectroscopy, the first derivative removes the baseline offset variance in the spectra, and the second derivative removes both the baseline offset difference and the baseline slope difference. As such these derivatives are quite useful when analyzing spectroscopic data.

3.3.2 Multivariate Data Analysis in Raman Images of a Common Pharmaceutical Tablet

In this section, the multivariate data analysis methods of PCA and ICA are applied to Raman images of a pharmaceutical tablet. The results obtained using PCA and ICA are then compared.

Application of PCA

The Raman images of pharmaceutical tablet are analyzed by PCA in this section. Figure 3.15-3.19 show a selection of the identified scores compared with the reference spectra that they most closely resemble. A number of observations can be made from the results.

1) Five of the eight scores were found to resemble the spectra of the constituent compounds.

2) Two of the scores were found to be similar to the reference spectrum for the API. The reason for this is that it was found that the locations of the peaks for this particular component were shifted in some samples by one wavenumber. Unfortunately with the analysis methods applied in this paper, this issue means that two independent reference spectra were identified. This is highlighted in Figure 3.15 and Figure 3.16, which show the reference spectra for the API and PC2 and PC6, which are all very similar in structure.
3) None of the components matched the spectra of Avicel, which was the component with the highest concentration in the tablet.
4) Out of the eight retained principal components, the 1st, 4th and 7th PCs were found to contain no useful information relating to any of the constituent compounds. Furthermore, these components did not appear to be composed of noise as there is clear structure to them. This poses a serious challenge in situations when the reference spectra of the constituent compounds are unknown as it is not clear which components contain useful information and which do not. The 4th and 7th components are shown in Figure 3.20 and Figure 3.21. These figures illustrate that there is no clear similarities between the reference spectra and either of these components. The first principal component was found to be almost identical to the mean measured spectrum, as shown in Figure 3.22. Therefore, the first principal component was observed to contain no consequential information regarding the reference spectra of the constituent compounds.
5) The scores beyond the eighth PC were found to contain no information relating to the reference spectra.
6) In this analysis, eight PCs were retained in the model. This number was chosen as the eighth PC was the last PC to contain clear information regarding the reference spectra. In situations where the reference spectra are unavailable it is not obvious how many components should be retained.

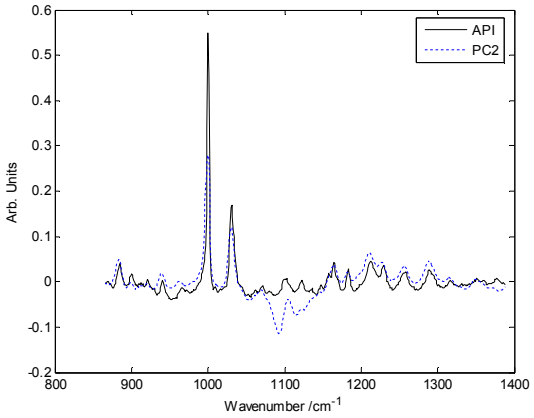

Figure 3.15: Reference Spectrum of API and 2nd PC

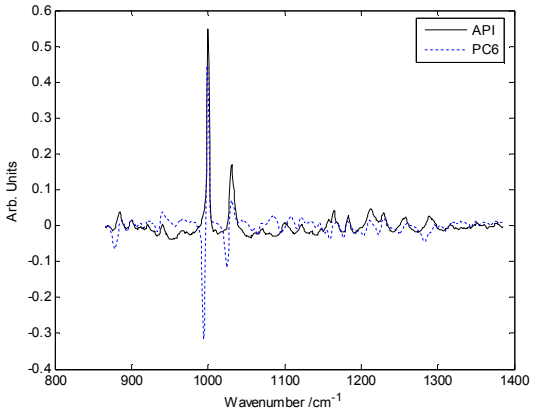

Figure 3.16: Plot of the 6th PC and Reference Spectrum of API

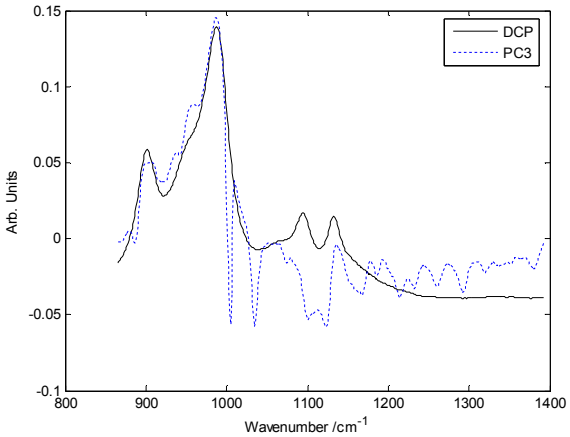

Figure 3.17: Plot of the 3rd PC and Reference Spectrum of DCP

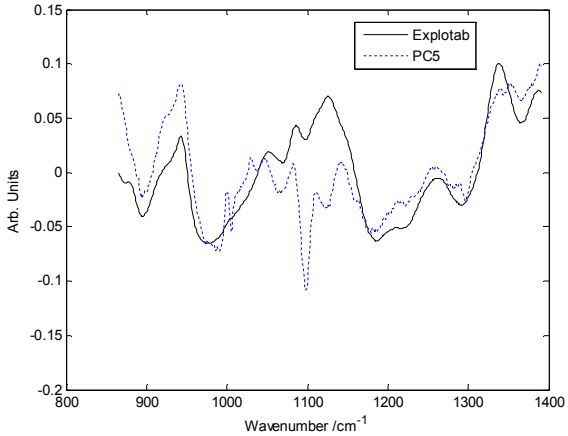

Figure 3.18: Plot of the 5th PC and Reference Spectrum of Explotab

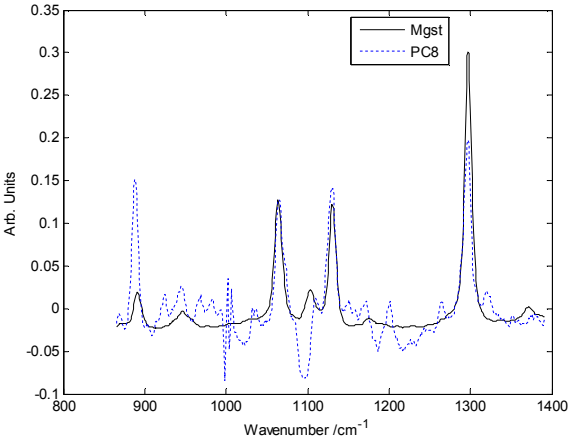

Figure 3.19: Plot of the 6th PC and Reference Spectrum of MgSt

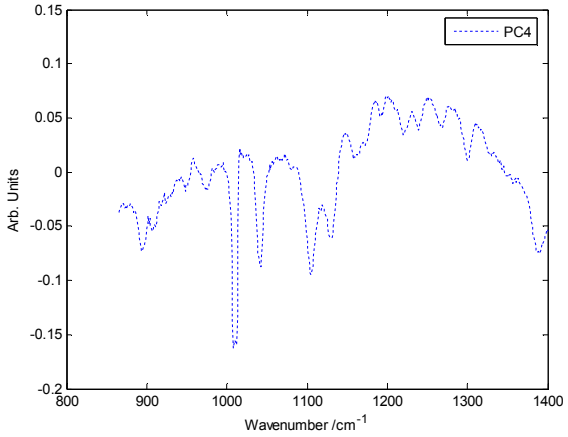

Figure 3.20: Plot of the 4th PC

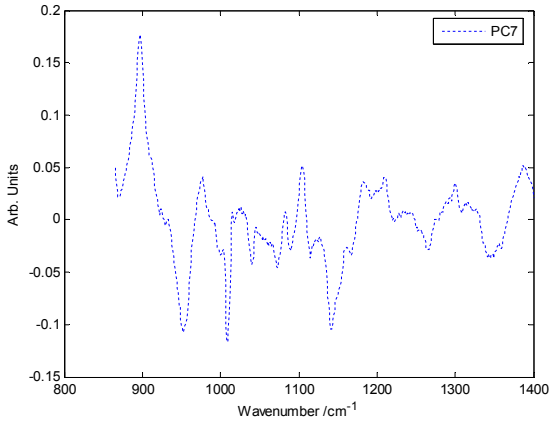

Figure 3.21: Plot of the 7th PC

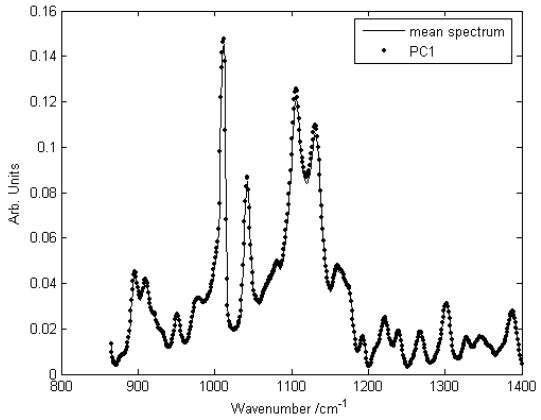

Figure 3.22: Plot of the 1st PC and the Mean of the Measured Spectra

When applying PCA, each PC is calculated by maximizing the amount of variance it can explain. The PCs will therefore not necessarily correspond to one specific chemical component, as argued by Zhang et al. [78] and Vrabie et al. [126]. In fact, principal components will typically be linear combinations of the reference spectra. This phenomenon is generally known as "rotational ambiguity" and is particularly evident when several reference spectra overlap significantly [78].

Application of ICA

When ICA was applied to the data collected from the tablet, the following observations were made:

1) ICA was able to identify the reference spectra of the constituent compounds far more readily than PCA.
2) ICA successfully identified the reference spectra of all of the five constituent compounds. However, in order to identify all of the reference spectra it was necessary to calculate seven independent components in the ICA model. As with PCA, without knowledge of the reference spectra it is not obvious how many components should be used in the model.
3) The reference spectrum for the API was captured by the 1st and the 5th IC as shown in Figure 3.23 and Figure 3.24 respectively. As with the analysis using PCA, the shift in this spectrum meant that it needed to be described by two components. The 6th IC identified the reference spectrum of Avicel, as shown in Figure 3.25, while the 3rd IC estimated the reference spectrum belonging to DCP, as shown in Figure 3.26. Finally, the reference spectra corresponding to Explotab and MgSt were identified by the 7th and the 4th IC as illustrated in Figure 3.27 and Figure 3.28, respectively.
4) There was one independent component (IC2) which was found to contain no useful information regarding the reference spectra of any of the constituent compounds. As with PCA, this poses a serious problem when attempting to identify reference spectra in situations when such reference spectra are unknown *a priori*. This IC is shown in Figure 3.29.

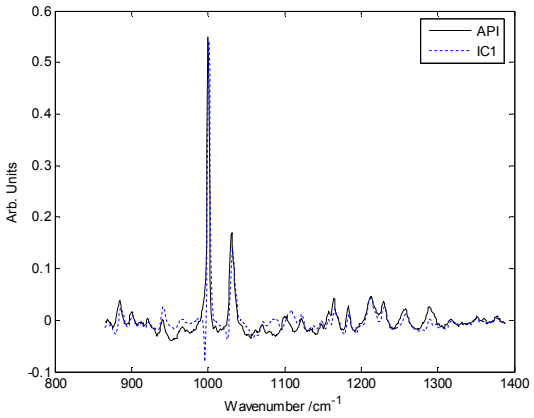

Figure 3.23: Plot of the 1st IC and the Reference Spectrum of the API

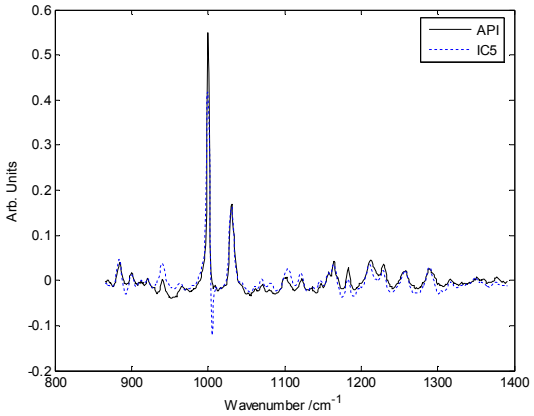

Figure 3.24: Plot of the 5th IC and the Reference Spectrum of the API

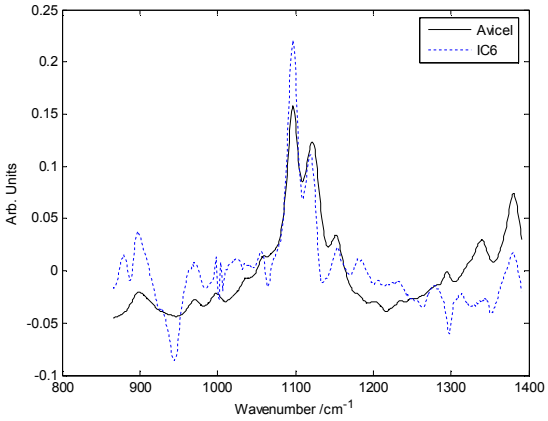

Figure 3.25: Plot of the 6th IC and the Reference Spectrum of the Avicel

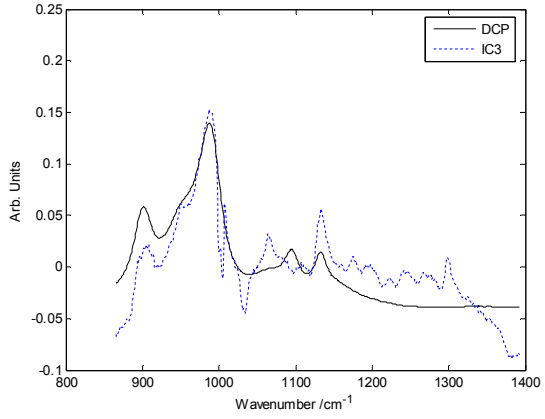

Figure 3.26: Plot of the 3rd IC and the Reference Spectrum of the DCP

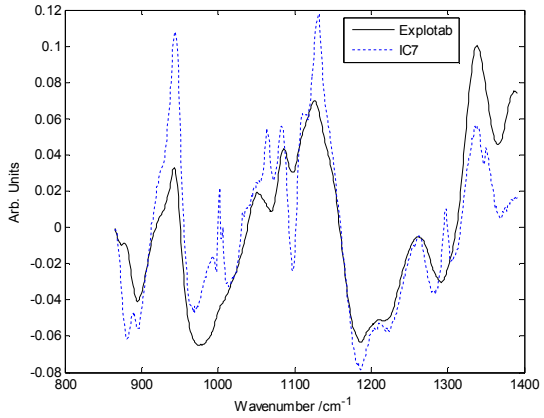

Figure 3.27: Plot of the 7th IC and the Reference Spectrum of the Explotab

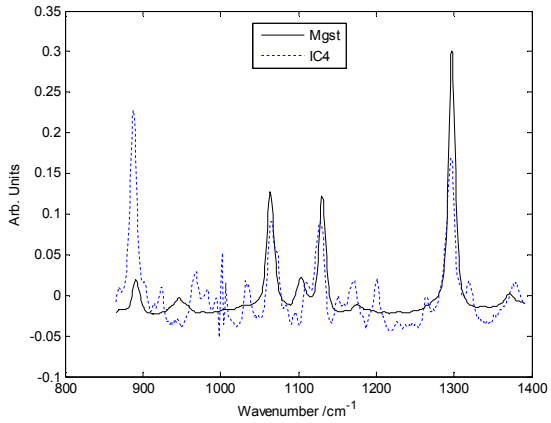

Figure 3.28: Plot of the 4th IC and the Reference Spectrum of the MgSt

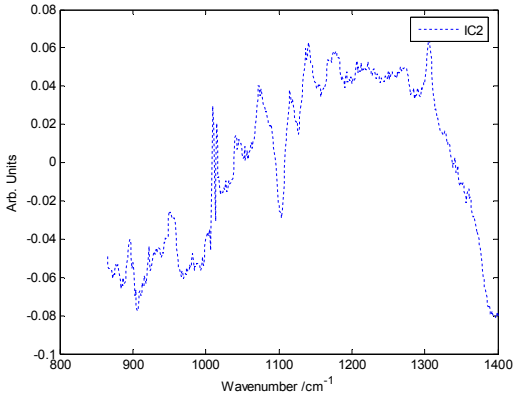

Figure 3.29: Plot of the 2nd IC

Careful analysis of the components identified using PCA and ICA suggest that ICA provides the clearer differentiation of the reference spectra. Furthermore, when ICA was

applied, all the reference spectra were identified and only one of the seven components contained information which did not appear to be related to the reference spectra. In contrast PCA failed to identify one of the reference spectra and three of the eight components contained information unrelated to the reference spectra.

For PCA and ICA to gain wider acceptance in the applied spectroscopy community it is important to address a number of practical issues that hamper their effectiveness. In particular, choosing the number of components retained in the model and also differentiating between those components that carry relevant information and those that depict mainly noise are non-trivial tasks. This is particularly true in those circumstances for which the number of constituent compounds and/or the shapes of the reference spectra of the constituent compounds are unknown. Also, the issue of dealing with inevitable peak shifts present in measured spectra remains to be comprehensively addressed.

3.3.3 Image Maps

The production of an image of a tablet describing the distribution of all the constituent compounds would be of great benefit to the pharmaceutical industry [42].

Having identified the reference spectra for the individual constituents, it is possible to construct a map showing the relative distribution of the individual compounds in the tablet. Given the improvement in using ICA rather than PCA, only the image maps produced using ICA are provided here.

Using the appropriate ICs from the developed ICA model, the relative concentrations of the five constituent compounds were determined for all 6000 pixels. These samples covered an area of $2 \times 2 \ mm^2$ on the tablet. The image of each component is obtained according to the concentration of corresponding constituent compounds. Because the reference spectrum for the API was captured by the 1st and the 5th IC, the concentration of the API is calculated as the addition of the weights of the 1st and the 5th IC.

In order to illustrate the level of spatial distribution of the constituent compounds, spatial distribution images corresponding to the score of each IC are deployed and shown in Figure 3.30. Note that pixels coloured white represent those segments of the tablet for which a particular score had a large value (i.e. high concentration of the associated compound) and pixels coloured black represent segments of the tablet with small score values (i.e. low concentration of the associated compound).

Figure 3.30 shows the score images of IC1 and IC3 to be very similar, which is unsurprising since they both relate to the same constituent compound (API) although their loadings are slightly shifted versions of each other. Hence, it is observed that a slight shift between the loadings of IC1 and IC3 does not seem to have a consequential impact on the resulting image maps. Score images of IC1 (API) and IC4 (MgSt) appear to be related even though there is an insignificant overlap between the corresponding reference spectra of API and MgSt, as shown in Figure 3.13. Hence, the similarity of these score images indicates that the spatial distributions of these two compounds may be similar.

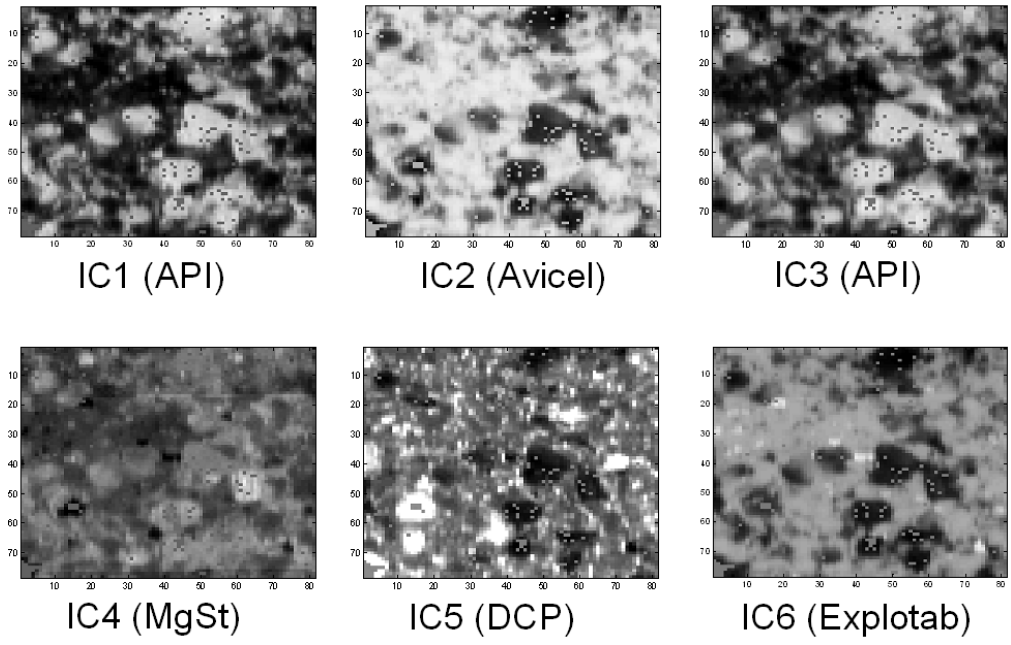

Figure 3.30: Score Images of the ICA Model

Finally, a composite image where the images for all compounds are shown to give a better understanding of the distribution of each compound is also identified. To make a composite image, each component is given a single colour. In a composite image, each

pixel is allocated to one constituent compound. Thus, the binarization threshold for each compound was selected such that the overall percentage of pixels allocated to that particular compound is close to the percentage concentration of that compound in a tablet. There are 6000 pixels in the tablet, and the composition of the tablet was as follows: Avicel at 48% w/w; API at 24% w/w; DCP at 24% w/w, Explotab at 3% w/w; MgSt at 1% w/w. Thus, the numbers of pixels allocated to Avicel, API, DCP, Explotab and MgSt are 2880, 1440, 1440, 180 and 60 respectively. If the thresholds for all compounds are selected simultaneously, some pixels will be allocated to two or more compounds, and some pixels will not be allocated to any compounds. To solve this problem, the low-weight compounds such as Explotab and MgSt pixels are firstly granted sufficient numbers of pixels, then high-weight compounds such as API, DCP and Avicel are granted sufficient numbers of pixels.

The basic threshold selection process is shown below (from low-weight compound to high-weight compound):

1) The threshold for MgSt is firstly selected to make sure that the number of MgSt pixels with values above this threshold is 60.
2) In the remaining pixels, 180 are allocated to Explotab by selecting an appropriate threshold.
3) In the remaining pixels, 1440 are allocated to DCP by selecting an appropriate threshold.
4) 1440 are allocated to API by selecting an appropriate threshold.
5) 2880 are allocated to Avicel by selecting an appropriate threshold.

After the threshold selection process, each pixel in the tablet is allocated to a unique compound.

A composite image map showing the gross spatial distribution of all five constituent compounds is shown in Figure 3.31. There is more than one pound in a pixel. In this image map, a pixel is assigned to a compound because this compound is the main

compound in this pixel. This is an illustrative map showing the distribution of the compounds, and it is not a map which shows the actual distribution of the compounds.

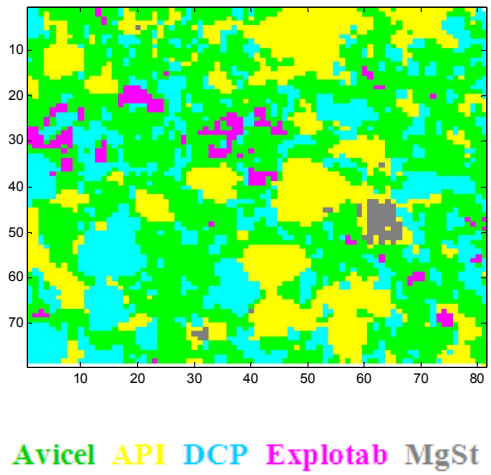

Figure 3.31: Image Map of all the Constituent Compounds

By observing Figure 3.31 it appears that Avicel and API are the most abundant compounds present in the tablet while Explotab and MgSt appear to be the least abundant constituents. This finding is expected due to the fact that the threshold concentration for each constituent compound was determined from knowledge of the absolute concentrations of each of the compounds within the tablet. Also, note that Figure 3.31 shows the most prevailing compound in each pixel rather than the only compound present in each pixel. Hence, it does not provide information regarding each compound's spatial distribution.

3.4 Summary

This chapter has briefly described the algorithms of PCA and ICA, and applied these techniques to two simulated sets of data. This study was used to illustrate how the algorithms work when applied to spectral data. Following this, the relative abilities of PCA and ICA to infer reference spectra of the five constituent compounds present in a common pharmaceutical tablet was compared. In this study, 6000 Raman observations

were made of a single tablet and ICA and PCA were applied to this data. The analysis showed the relative abilities of PCA and ICA to identify the constituent compounds from the data. ICA was able to identify the reference spectra of all five constituent compounds within this tablet, whereas PCA was only able to identify four, missing the most concentrated constituent. Finally, the spatial distribution of each of the five constituent compounds was estimated using the results of the ICA analysis and presented in a clear graphical form. This provided a gross distribution of the constituent compounds in the tablet.

4 Trajectory Tracking of Batch Processes

This chapter addresses the issue of quality control in the pharmaceutical industry by incorporating NIR measurements as feedback information in process control systems. Two approaches to this are proposed, each utilizing Model Predictive Control (MPC) technology, but in each case different control variables are used. The proposed approaches are applied to control quality in a simulated chemical batch reactor. The performances of the two controllers are compared with the performance achieved by a control system which does not feedback the NIR measurements.

This chapter is divided into the following sections:

1) Description of Model Predictive Control (MPC) algorithm;
2) Introduction to the model of the batch reactor used in this study;
3) Presentation of the proposed trajectory tracking methods;
4) Comparison of the performance of the two control methods;
5) Summary of the chapter.

4.1 Model Predictive Control (MPC)

Model predictive control (MPC) [95] is an advanced control technique widely applied in industrial process control. It has the following advantages in process control [127]:

1) It handles multivariable control, multi-input multi-output processes naturally.
2) It can take account of actuator limitations.
3) The control concepts are very intuitive and the tuning straight forward.
4) It allows operation closer to constraints compared with conventional control, which frequently leads to more profitable operation. Remarkably short pay-back periods have been reported.
5) Control update rates are relatively low in these applications, so that there is sufficient time for the online computations.
6) Process constraints can be treated readily in the optimization process.

Based on these properties, MPC is employed in this chapter in the two proposed methods for controlling the quality of batch processes. MPC is now described in further detail.

MPC relies on a model of the plant, which is typically obtained using system identification techniques. It utilizes this model to predict the future response of the plant, and then attempts to optimize plant behavior by adjusting a sequence of future manipulated variable changes. Only the first manipulated variable change is applied and the optimization procedure is repeated at the next control interval.

MPC has been applied in many different areas. Zhu et al. [128], for example, applied linear model predictive control (LMPC) to control large-scale gas pipeline networks. The approach was evaluated using a simulated industrial-scale oxygen pipeline network. Kiran and Jana [129] applied MPC to control the feed rate in a fed-batch yeast fermentation process. In [130], MPC techniques were applied to optimize oil yield. The process was identified using multivariable methods and a predictive controller was implemented on the real plant. Improved oil yield and extraction performance was achieved using the control technique. In [131], MPC was combined with a knowledge based control system to control the flow rate of each reflux in a petroleum plant. The combined control system achieved satisfactory performance when controlling the pressure in the plant. To obtain a comprehensive understanding of MPC algorithms and their applications, several textbooks are available [127, 132-136].

MPC operates by tracking one or more control variables to their set point while maintaining the manipulated variables within specific ranges. The basic concept for MPC is shown in Figure 4.1.

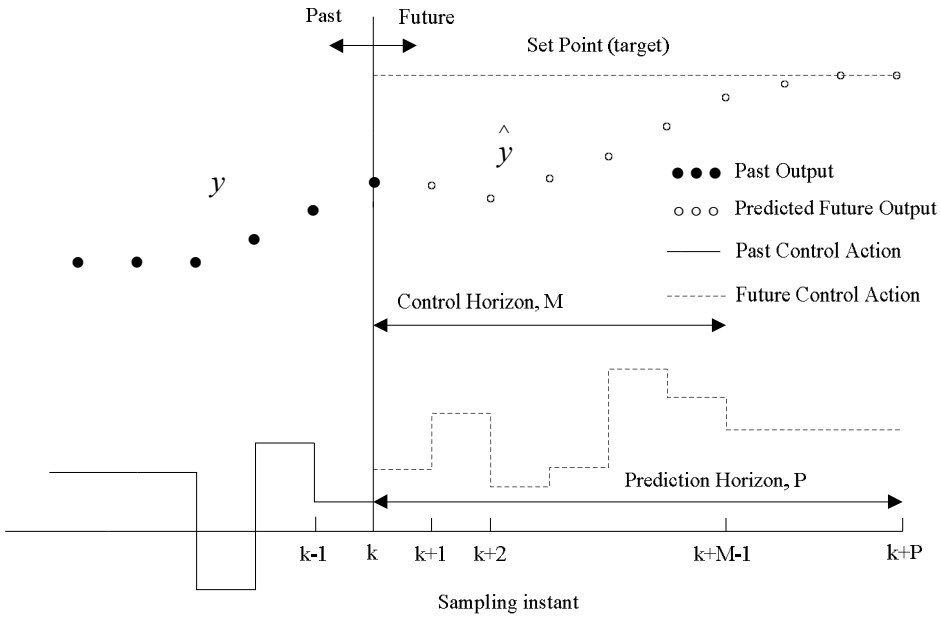

Figure 4.1 Basic Concept for Model Predictive Control [95]

The implementation of MPC consists of three steps [95]:

1) The future outputs $\hat{y}(k+j), j=1...P$ over the prediction horizon P are predicted at each sampling instant k. These predictions are based on the process model and future control signals $\hat{u}(k+j), j=0...M-1$.

2) A set of future control signals $\hat{u}(k+j), j=0...M-1$ for the control horizon M are calculated at each sampling instant k using the predicted future outputs and set point. The objective is to keep the future outputs as close to the set point as possible.

3) Only the first control signal $\hat{u}(k)$ is sent to the process. In the next sampling instant $k+1$, the same procedure is performed, and the control signal $\hat{u}(k+1)$ is obtained.

To identify the necessary control moves subject to process constraints, Δu is determined such that it minimizes the quadratic objective:

$$\min \quad \sum_{j=0}^{P-1} \left\| \hat{y}(k+j) - y^{sp}(k+j) \right\|^2 Q_1(j) + \sum_{j=0}^{M-1} \left\| \Delta u(k+j) \right\|^2 Q_2(j) \tag{4.1}$$

subject to the constraints:

$$\Delta u_{\min} \le \Delta u(k+j) \le \Delta u_{\max} \quad j = 0,1,2,...,M-1 \tag{4.2}$$

$$u_{\min} \le u(k+j) \le u_{\max} \quad j = 0,1,2,...,M-1 \tag{4.3}$$

$$y_{\min} \le \hat{y}(k+j) \le y_{\max} \quad j = 0,1,2,...,P-1 \tag{4.4}$$

where, $y^{sp}(k+j)$ is the reference trajectory, $Q_1(j)$ and $Q_2(j)$ are the j th elements in the weighting matrices Q_1 and Q_2. P and M are the prediction and control horizon respectively. The weighting matrices, prediction horizon and control horizon are the important factors that influence the performance of MPC algorithm. The weighting matrices are selected according to the relative importance of the variables, with the most important variable typically having the largest weight. In this work, the weight matrices Q_1 and Q_2 are selected as I and 0 separately. The prediction horizon P and control horizon M are selected as 100 and 10 separately.

4.2 Batch Reactor Model

The batch reactor process is shown schematically in Figure 4.2 [137]. Steam is used to heat up the reactor, while the coolant is used to cool down the reactor. The reactor temperature is affected by disturbance variables, including the reactant feed temperature and the rate at which the reaction in the vessel proceeds. The jacket temperature, T_j, is controlled using a proportional controller, which switches between providing cooling water or steam to the jacket depending on the jacket temperature set point, T_{jsp}. The reactor temperature, T_r, is controlled by regulating the jacket temperature set point, T_{jsp}, using an outer, or primary controller. In this chapter, for simplicity, proportional-integral-derivative controller is used.

Batch reactors have nonlinear characters and they operate over a wide temperature range. An optimal temperature profile is the precondition of manufacturing good quality product.

Optimal temperature profiles typically consist of heating up the reactor in the initial phase, maintaining the reactor temperature at a specific temperature for an appropriate amount of time and then cooling down the reactor. Temperature control of batch reactor can be a difficult task because the process is nonlinear and the operating state is not fixed. The reactor temperature may run away if heat generated exceeds the cooling capacity of the reactor, which can cause great risk to the plant personnel and equipment. Thus, the rate of change of reactor temperature should be carefully controlled. Cott and Macchietto [138] used generic model control (GMC) to track the reactor temperature in this particular batch process. Dual-model (DM) controller was used to achieve a minimum time control strategy in tracking reactor temperature by Shinskey [139]. The reactor temperature reaches the set point by maximum heating the reactor, and the error was reduced by maximum cooling the reactor. The PID controller was used to regulate the reactor temperature while it was in the set point.

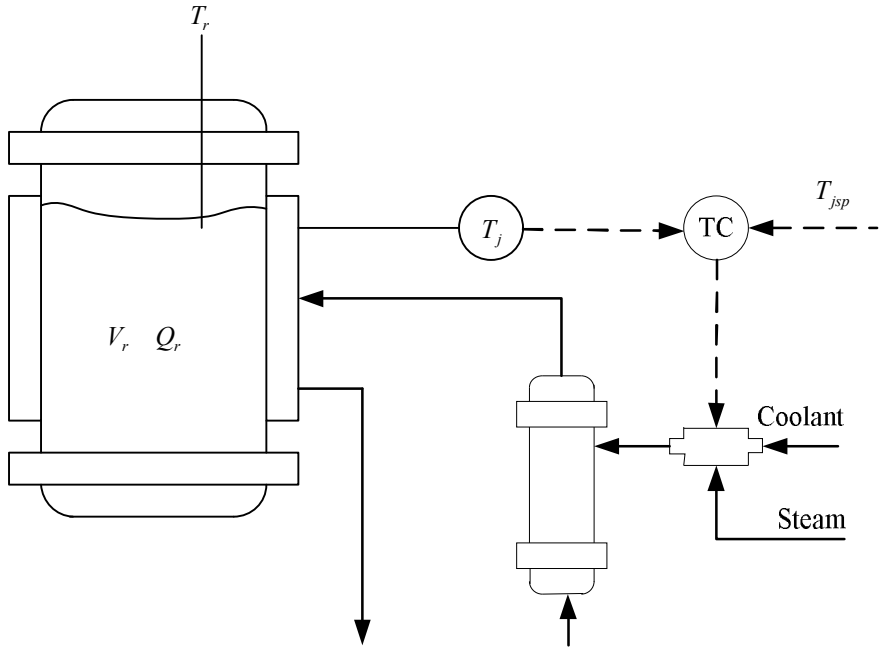

Figure 4.2 A Batch Reactor with a Reactor Temperature Control System (TC: Temperature Controller) [137]

The reaction taking place in the batch reactor, to be studied in this chapter is as follows [138].

$$A + B \rightarrow C\,;\ A + C \rightarrow D \tag{4.5}$$

where A, B are the raw material, C is the desired product and D is the unwanted byproduct. The operating objective is to maximize the production of C and minimize the production of D.

The model equations are shown below:

$$\frac{dM_A}{dt} = -R_1 - R_2 \tag{4.6}$$

$$\frac{dM_B}{dt} = -R_1 \tag{4.7}$$

$$\frac{dM_C}{dt} = R_1 - R_2 \tag{4.8}$$

$$\frac{dM_D}{dt} = R_2 \tag{4.9}$$

$$\frac{dT_r}{dt} = \frac{(Q_r + Q_j)}{M_r C_{pr}} \tag{4.10}$$

$$\frac{dT_j}{dt} = \frac{(T_j^{sp} - T_j)}{\tau_j} - \frac{Q_j}{V_j \rho_j C_{pj}} \tag{4.11}$$

$$R_1 = k_1 M_A M_B \tag{4.12}$$

$$R_2 = k_2 M_A M_C \tag{4.13}$$

$$k_1 = \exp(k_1^1 - \frac{k_1^2}{T_r + 273.15}) \tag{4.14}$$

$$k_2 = \exp(k_2^1 - \frac{k_2^2}{T_r + 273.15}) \tag{4.15}$$

$$Q_r = -\Delta H_1 R_1 - \Delta H_2 R_2 \tag{4.16}$$

$$M_r = M_A + M_B + M_C + M_D \tag{4.17}$$

$$C_{pr} = \frac{(C_{PA} M_A + C_{PB} M_B + C_{PC} M_C + C_{PD} M_D)}{M_r} \tag{4.18}$$

$$Q_j = UA(T_j - T_r) \tag{4.19}$$

where the constant parameter values of the model are as follows:

$C_{PA} = 18.0 kcal / kmol°C$

$C_{PB} = 40.0 kcal / kmol°C$

$C_{PC} = 52.0 kcal / kmol°C$

$C_{PD} = 80.0 kcal / kmol°C$

$\Delta H_1 = -10000.0 kcal / kmol$

$\Delta H_2 = -6000.0 kcal / kmol$

$C_p = 0.45 kcal / kg°C$

$C_{pj} = 0.45 kcal / kg°C$

$U = 9.76 kcal / \min m^2 °C$

$\rho_j = 1000.0 kg / m^3$

$k_1^1 = 20.8057 \ kmol^{-1} s^{-1}$

$k_1^2 = 10000 \ kmol^{-1} s^{-1}$

$k_2^1 = 38.9057 \ kmol^{-1} s^{-1}$

$k_2^2 = 17000 \ kmol^{-1} s^{-1}$

$V_j = 0.6921 m^3$

$A = 6.24 m^2$

$\Delta t = 0.2 \min$

$\tau_1 = 3.0 \min$

$W_r = 1560.0 kg$

The NIR spectra for A, B, C and D are acquired from [140].

4.3 Control Methodology

4.3.1 Cascade Control (CC)

The control systems that are proposed in this chapter are built above standard feedback control mechanism. [95] showed that the batch reactor studied here can be regulated using a cascade controller. Such a strategy is applied in many industrial batch control systems and was therefore considered a sensible low level control system to adapt in this work. Reactor temperature is typically controlled by manipulating the flow of a coolant or steam through a jacket surrounding the reactor. Feedback control takes action to correct any deviation in the controlled variables from their desired set-points. The disadvantage with feedback control is that corrective action is only taken after the controlled variables deviate from their set points. For processes that have large time constants, cascade control, is often applied. Cascade control consists of two control loops: the primary and secondary loops. The secondary control loop is able to identify the disturbances quality and take actions to reduce its influence one the controlled variable.

The structure of cascade control for the batch reactor is shown in Figure 4.3. The primary control loop, also known as the master loop, regulates the reactor temperature by adjusting the jacket temperature set-point. The secondary control loop, known as the slave loop, regulates the jacket temperature by manipulating the flow of either coolant or steam into it. Hence, the manipulating variable of the master control loop is the set-point for the slave control loop.

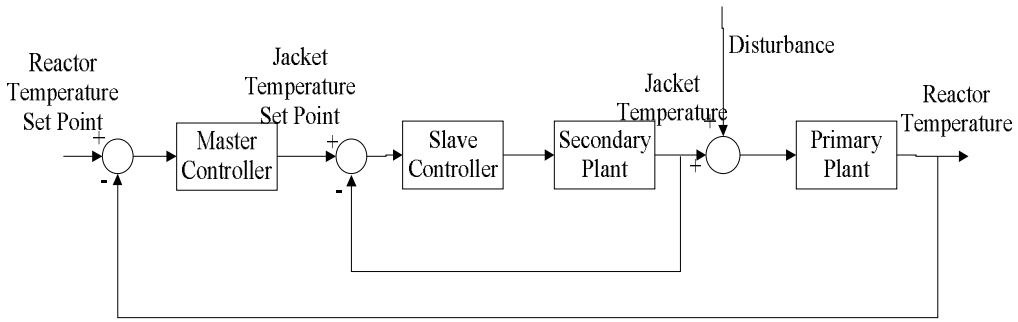

Figure 4.3 Control of Reactor Temperature Using CC System

In this work, a PI controller was used in the slave loop while a PID controller is employed in the master loop, as this was found to achieve good temperature control in the simulation. The proportional gain and integral gain used in PI controller were 10.7574 and 53.4882 separately. The proportional gain, integral gain and derivative gain used in PID controller were 13, 28.658 and 0.42884 separately.

Note that in the CC system described above, product quality is controlled implicitly, through the regulation of reactor temperature. Unfortunately, this type of control structure is unable to cope with disturbances that affect the underlying relationship between the reactor temperature and the product quality. As a result, the temperature profile required to provide a product of the correct quality may change. Without knowledge of an 'optimal' temperature profile required to meet product quality specifications, the CC is likely to result in unsatisfactory product quality. This is demonstrated later in this chapter.

4.3.2 Wavenumber-Based MPC Control (Wn-MPC)

As discussed in Chapter 2, spectroscopic instrumentation is being increasingly used in batch processes to provide measurements that are in some way related to product quality. By incorporating these measurements, as feedback information, into the control system, product quality can be regulated more explicitly, compared with the traditional CC scheme. One possible control system structure that incorporates NIR spectra as feedback information is shown in Figure 4.4.

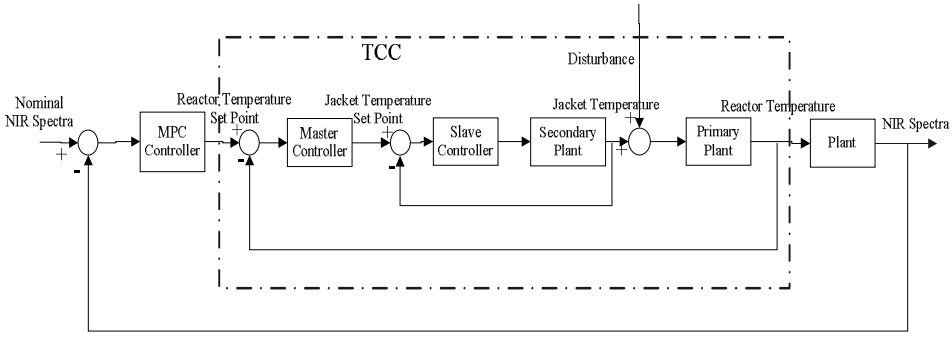

Figure 4.4 Basic Structure of Wn-MPC Control

This proposed control structure incorporates the CC system from Figure 4.3 and augments it with an additional outer control loop, namely the MPC control loop. The manipulated variable of the MPC controller is the reactor temperature set-point while the controlled variables are the intensities of the NIR spectra at a particular set of wavenumbers. Hence, within this control system structure, the CC system can be viewed as a slave controller with MPC as a master controller. This control system structure will be referred to in this thesis as Wn-MPC.

The reference profile for the wavenumber is obtained by collecting NIR spectra from a 'nominal' or 'golden' batch, This batch will have progressed with no major disturbances using the standard CC control scheme and would be considered to be an 'ideal' batch.

Since each wavenumber in the NIR spectra represents a candidate variable to be used as feedback information, there may be hundreds of potential controlled variables. Therefore, a serious practical problem that arises when attempting to implement Wn-MPC is the decision as to which wavenumbers should be regulated by the control system. The limited number of manipulated variables available in this system will prevent the entire NIR spectra being controlled explicitly. Currently, there are no clear guidelines as to which wavenumber should be selected for control purposes and therefore the sensible selection of appropriate wavenumbers is investigated in this chapter.

4.3.3 PCA Score-Based MPC Control (Sc-MPC)

In order to incorporate information from the entire spectral measurements into the feedback signal, a modified control system structure was used, as shown in Figure 4.5.

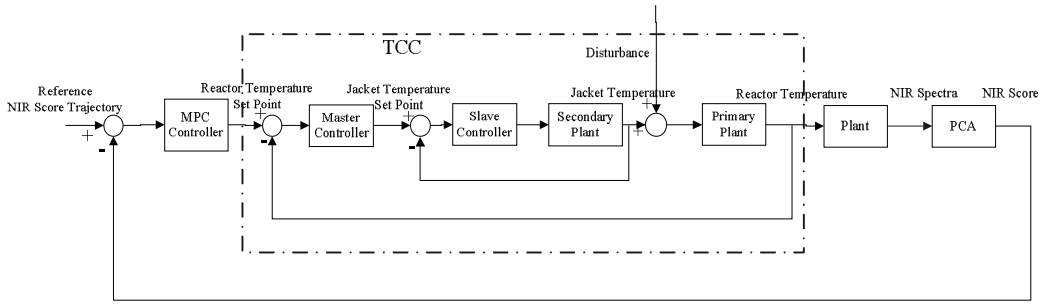

Figure 4.5 Basic Structure of the Sc-MPC Control

This control scheme differs from Wn-MPC in that it pre-processes the feedback measurement using PCA. The result of the PCA processing is a small set of variables, or scores, that contain information related to all of the measured wavenumbers. The first PC captures the most of the variances in the dataset. ICA identifies the independent variables in the dataset, and each IC only captures a small variance in the dataset. Thus, PCs are used here. This is in contrast to Wn-MPC where the feedback information relates to only a select number of wavenumbers.

In this chapter it is assumed that a PCA model is constructed using NIR spectra collected from a nominal batch. This nominal batch is run in the absence of any major disturbances using CC scheme. Hence, the resulting NIR spectra are assumed to represent a reference profile that is to be replicated by Sc-MPC.

4.4 Case Study

The three different control systems, described in Section 4.3, were evaluated by introducing a disturbance into the simulation and observing the response of the control loop system. In this study the disturbance was chosen to be a reduction in the value of the reaction rate constant k_1 by 8%. The objective of the three methods was to regulate the end-point NIR spectra to its nominal values, as shown in Figure 4.6.

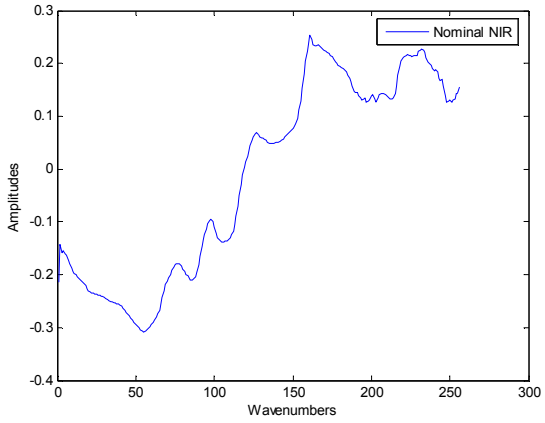

Figure 4.6 Nominal NIR Spectra of End Product

4.4.1 Prediction Model Identification

The main feature of MPC is that it forecasts process behavior and then determines future manipulated variables. The forecasting is based on the process model. Therefore, to develop the MPC scheme it is necessary to build a dynamic model relating the manipulated variables to the controlled variables. In this study an autoregressive with exogenous terms (ARX) model [141] was used. The ARX model is simple in structure and although only linear, has been shown in many studies to provide acceptable performance when used with ARX system [142, 143]. In all the studies in this thesis, recursive least squares (RLS) was used to identify the model. RLS has been shown to provide excellent identification capabilities, particularly for MPC application [144]. The training data for the model was obtained using the CC system structure, shown in Figure 4.3. To excite the process dynamics, the reference temperature trajectory was perturbed for three batches by adding a PRBS signal of amplitude 0.1 ^{0}C and switching time of 60 seconds. In this case study ARX based prediction models were developed with $n_y = 2$ and $n_u = 80$ using Akaike information criterion [145]. The data-driven identification method of RLS was used to develop dynamic models for both Wn-MPC and Sc-MPC controllers. In cross validation tests using an unseen data set from 10 further batches, $n_y = 2$ and $n_u = 80$ was found to provide the most accurate model.

4.4.2 Wavenumber Selection

In the case of Wn-MPC, candidate controlled variables were taken to be those wavenumbers that corresponded to a local peak in the measured NIR spectrum. In this case study the wavenumbers corresponding to local peaks in the NIR spectra and, therefore, representing the candidate controlled variables were 2, 77, 98, 127, 161 and 232, as illustrated in Figure 4.7.

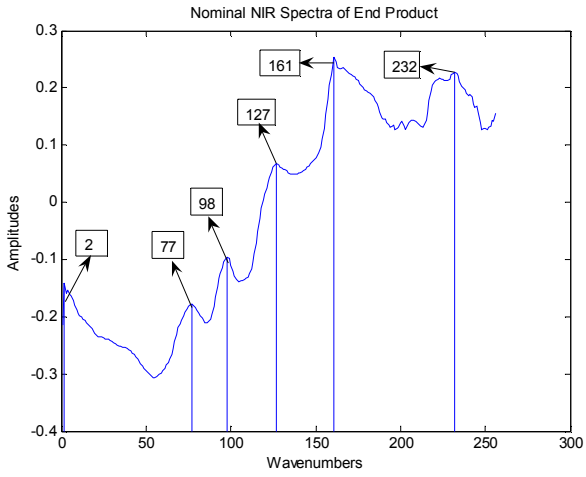

Figure 4.7 Selection of Spectral Peaks As Controlled Variables

For each of these wavenumbers, a prediction model was identified and the corresponding MPC controller was constructed and evaluated. The corresponding controllers are referred to with the chosen wavenumber written within brackets following the label Wn-MPC. For example, Wn-MPC(127) relates to the Wn-MPC controller that utilizes wavenumber 127 as the controlled variable.

4.4.3 PCA Model Development

A PCA model was developed using NIR spectra collected from a single nominal batch. The first PCA score captured 93.8% of the variation present in the NIR spectra and was used as a reference trajectory in the subsequent implementation of Sc-MPC controller. The loadings vector associated with the first PCA score was then used in real-time to compute a score value from the measured NIR spectra.

4.4.4 Results and Discussion

For each controller (CC, Wn-MPC and Sc-MPC) the process was perturbed using the disturbance described in Section 4.4. The resulting NIR spectra for each of the control systems, compared with the reference spectrum are plotted in Figure 4.8 and Figure 4.9.

Figure 4.8 shows the NIR spectra obtained when CC, Sc-MPC and Wn-MPC(77) were used to regulate the process. Sc-MPC can be seen to outperform both CC and Wn-MPC(77). In fact, the NIR spectrum obtained when using Sc-MPC controller was found to be very similar to the reference spectrum, as shown in Figure 4.8. On the other hand, both CC and Wn-MPC(77) clearly failed to reject the disturbance as evidenced by considerable deviation of their respective NIR spectra from the reference spectrum.

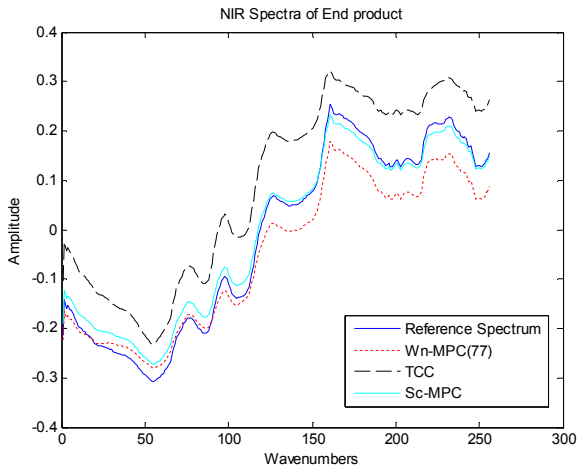

Figure 4.8 NIR Spectra of End Product Obtained When Using CC, Wn-MPC(77) and Sc-MPC

The reason for the difference in performance between CC and Sc-MPC lies in the fact that the CC control system does not consider the NIR spectra. Hence, its reference temperature profile is not adjusted to account for the presence of the disturbance, which has modified the underlying relationship between temperature and product quality. In constrast, Sc-MPC regulates the NIR spectra directly. The results also show that Wn-MPC(77) also provided poor control performance. This was because the spectral data contained in wavenumber 77 appeared not to be sufficient to characterize the entire NIR spectrum.

The performances obtained by controlling NIR trajectories at different wavenumbers (77 127 161) using Wn-MPC differed significantly, as shown in Figure 4.9. This figure shows that the controllers regulate the spectra around the specific wavenumber, for which they were designed, but that errors are introduced at other wavenumbers.

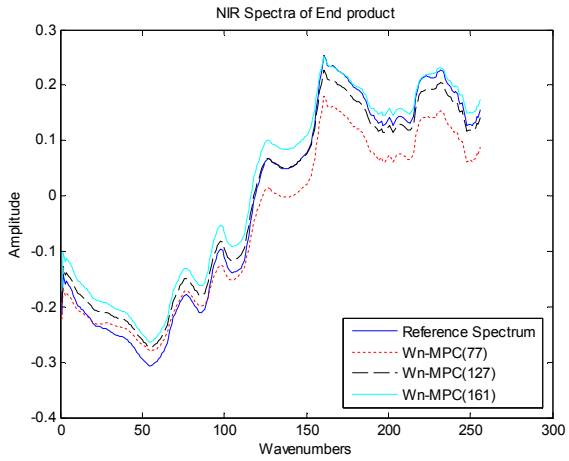

Figure 4.9 NIR Spectra of End Product Obtained When Using Wn-MPC(127), Wn-MPC(161) and Wn-MPC(77)

Wn-MPC controller was developed to regulate every wavenumber. The sum of the squared errors of the NIR spectra and its nominal values when using Wn-MPC for each wavenumber are calculated and displayed in Figure 4.10. This figure shows a large variation in performance achieved by Wn-MPC controllers that utilise different wavenumbers as their controlled variables. This suggests that the performance of Wn-MPC is very much reliant on the selection of the wavenumber which is to be controlled. As such, the performance of this controller is likely to be very unreliable.

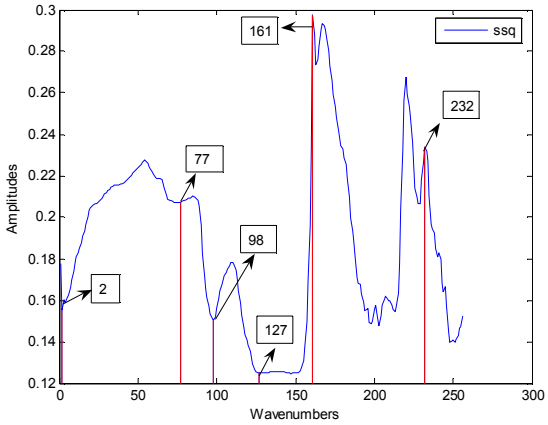

Figure 4.10 The Sum of Square Errors (SSQ) by Wn-MPC at Different Wavenumbers

Even if the Wn-MPC is used with an optimally selected wavenumber, which is wavenumber 127 in this particular case study, the resulting control performance was found to be very similar to the performance of the Sc-MPC controller. This is demonstrated in Figure 4.11 where the NIR spectra displayed were obtained when the process was being controlled using Sc-MPC and Wn-MPC(127).

The improvement in performance delivered by Wn-MPC(127) is not considerable and the trial-and-error procedure involved in selecting this wavenumber is likely to be prohibitively time-consuming and expensive. On the other hand, Sc-MPC delivered performance that was similar to that of the Wn-MPC(127) controller, but did not require any trial and error selection of control variable. Hence, Sc-MPC controller was found to require minimal user interaction when selecting an appropriate controlled variable while also delivering a highly satisfactory performance.

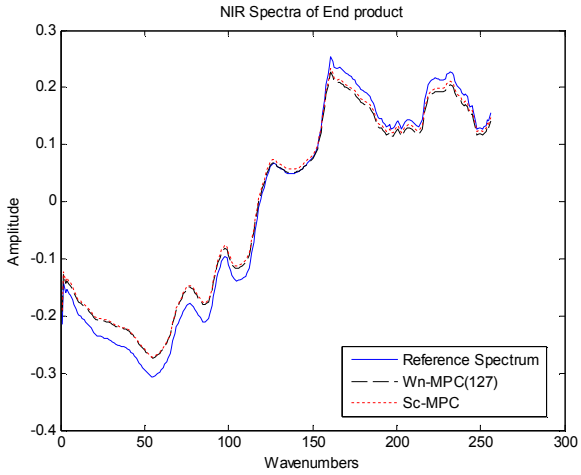

Figure 4.11 NIR Spectra of End Product Obtained When Using Wn-MPC(127) and Sc-MPC

4.5 Summary

This chapter investigated the ability of two different control strategies to regulate a simulated batch reactor using the NIR measurements in the feedback loop. One of the controllers, Wn-MPC used the spectral intensity at a specific wavenumber (spectral channel) corresponding to a peak in the NIR spectra as the single measurement feedback to the controller. The other controller used the multivariate statistical tool, Principal Component Analysis (PCA), to extract and condense the main features present in all of the wavenumbers into a single composite variable that was then controlled. This controller was referred to as Sc-MPC. The performance of these controllers was compared with a more traditional batch control strategy – cascade control. This controller was found to be inadequate when disturbances were introduced that altered the underlying relationship between reactor temperature and product quality.

By implementing the three controllers on a simulated batch reactor, it was observed that Sc-MPC achieved satisfactory control while also requiring no user interaction when deciding on the variable to be controlled. On the other hand, the performance achieved by Wn-MPC was found to be highly dependent on the choice of the wavenumber that was to be controlled. Due to the lack of rigorous guidelines when selecting appropriate

wavenumber and the resulting trial and error approach necessary to determine optimal wavenumber, it is questionable whether Wn-MPC can be used as a practical solution.

5 End-point Control of Batch Processes

This chapter studies the application of an end-point control method for incorporating NIR spectral measurements in a feedback control strategy for batch processes. The benefit of the end-point controller is that it can be used to regulate NIR measurements across all wavenumbers, rather than a composite, or specific selection of the wavenumbers, as investigated in Chapter 4. The end-point controller is applied to a simulated batch reactor, where its response to a reaction rate and set point change is investigated. The performance of this controller is also compared to that achieved by the PCA Score-Based MPC Controller (Sc-MPC).

This chapter is divided into the following sections:

1) Description of partial least squares and the end-point control algorithm;
2) Introduction to the proposed control methodology – PCA Score-based MPC control (Sc-MPC) and end-point control;
3) Application of PCA Score-based MPC controller (Sc-MPC) and end-point controller to a simulated batch process;
4) Summary of the chapter.

5.1 Preliminary Methodology

Quality control is very important in batch processing. In Chapter 4, NIR measurements were fedback to a control system that regulated quality in a simulated batch process. In that control system, the intensity of the NIR measurement at a single wavenumber was selected as the variable to be controlled. Although this control system was found to perform reasonably well, errors were introduced in the NIR measurement at non-controlled wavenumbers. This suggests that the control system cannot ensure that all product quality requirements will be met. To ensure the final product quality requirements are met, the entire NIR spectrum should be controlled. This chapter investigates an end-point control algorithm which is applied to control the entire NIR spectrum. The performance achieved by the end-point controller is compared with that of

Sc-MPC, which condensed the information in the NIR measurement using PCA and then regulated a single score. To begin, the end-point control algorithm and the partial least squares (PLS) algorithm used in model building are described.

5.1.1 Partial Least Squares (PLS)

Partial least squares (PLS) [61] is a method used to find a linear regression model by projecting the predicted variable matrix Y and the observed variable matrix X to a latent variable space. Identification then takes place in this score space. The model for PLS is shown in Equations (5.1) and (5.2).

$$X = \sum_{k=1}^{np<nx} t_k p_k^T + E = TP^T + E \tag{5.1}$$

$$Y = \sum_{k=1}^{np<nx} t_k q_k^T + F = TQ^T + F \tag{5.2}$$

where T is a matrix of the PLS scores which capture most of the variability in the data. P and Q are loading matrices for X and Y, respectively. E and F are the residual errors. np and nx are the number of latent variables and dimension of matrix X, respectively.

Several algorithms have been proposed for calculating the PLS model parameters. The most popular method is the Non-linear Iterative Partial Least Squares (NIPALS) algorithm [146]. In NIPALS, a pair of scores is obtained by decomposing X and Y. The predicted variable matrix X and observed variable matrix Y are deflated by the variance captured by the existing scores. The remaining data matrices are used to acquire the next pair of scores. Another method – Straightforward Implementation of a Statistically Inspired Modification of the Partial Least Squares (SIMPLS) [147] – is also a widely used algorithm for calculating PLS models. The advantage with this technique is that it is faster in the calculation and thus saves time. In this work, SIMPLS was applied.

A particular challenge when using PLS is identifying an appropriate number of scores or latent variables. Several methods have been proposed for helping with this, of which

cross validation [45] is the most commonly applied, and was the method applied in this study.

5.1.2 End-Point Control Algorithm

Flores-Cerrillo and MacGregor [5] proposed a method of end-point quality control. In their controller, the quality of the end-product was regulated by adjusting the trajectory of the manipulated variables in a reduced space of latent variable models. This controller was referred to as end-point control. The end-point control algorithm proposed by Flores-Cerrillo and MacGregor [5] is described in more detail in this section.

Model Building

The model used in end-point control is obtained using PLS and the dataset used to identify the model must contain data that is representative of the process being studied. In other words, the dataset should contain data collected while the process is being excited. The most appropriate method for exciting batch processes is an open research question. In this work, PRBS was applied, which is consistent with the work of Flores-Cerrillo and MacGregor [5]. The dataset used for identification should be arranged into a three dimensional array X, and a two dimensional matrix Y ($K \times M$), as shown in Figure 5.1.

X^T is a data matrix composed of online process variables X_{on}^T, offline measurements X_{off}^T and manipulated variables u_c^T. X_{on}^T, X_{off}^T, and u_c^T consist of l, r and n variables respectively. In Figure 5.1, $X_{on,j}^T = \left[x_{on,1}, \quad \cdots \quad x_{on,f}\right]_j$, $X_{off,s}^T = \left[x_{off,1}, \quad \cdots \quad x_{off,g}\right]_s$ and $u_{c,m}^T = \left[u_{c,1}, \quad \cdots \quad u_{c,w}\right]_m$ represents, respectively, f on-line measurements for the j th variable, g off-line analysis for the s th variable, and w manipulated variables for the m th variable.

As mentioned in Chapter 3, most statistical methods are suitable for two dimensional datasets only. It is therefore necessary to convert the three dimensional dataset into a two

dimensional dataset. In this work, the three dimensional X matrix is unfolded into a two dimensional ($K \times N$) matrix, as shown in Figure 5.1, where K is the number of batches, and $N = f \times l + g \times r + w \times n$.

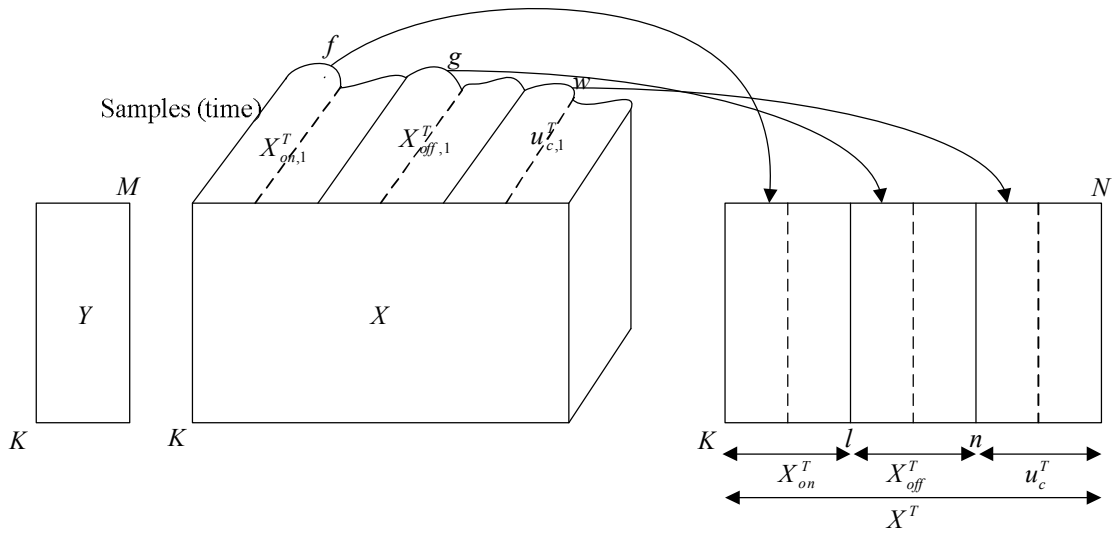

Figure 5.1 Unfolding of Database for Model Building [5]

Control

Following the development of the model, the control system can be formulated. Flores-Cerrillo and MacGregor [5] proposed using decision points during the batch. At each decision point, a forecast of the end-point quality was made. If this estimate indicated that quality requirements would not be met then appropriate control action was taken.

At decision point θ_i, the end-point quality is predicted using the PLS model:

$$\hat{t}^T_{present} = X^T W = [X^T_m \quad u^T_c]W = [X_{m,measured,\theta_i^T} \quad X^T_{m,future} \quad u_{c,implemented,\theta_i^T} \quad u^T_{c,future}]W \tag{5.3}$$

$$\hat{y}^T = \hat{t}^T_{present} Q^T \tag{5.4}$$

where $X_{m,measured,\theta_i^T}$ and $u_{c,implemented,\theta_i^T}$ are the measured variables and implemented manipulated variables in the period ($0 \le \theta \le \theta_i$). $X_{m,future}$ and $u_{c,future}$ are the future

variables and the future manipulated variables in the period $\theta_{i+1} \le \theta \le \theta_f$, i.e. in the remainder of the batch. $X_{m,fugure}$ is obtained using the PLS model and the missing data algorithms available in the literature [148]. Single Component Projection (SCP) [149] is the simplest method for missing data prediction and was the technique applied in this Chapter. W and Q are projection matrices obtained from the PLS model. $\hat{t}_{present}$ is the projection of the X matrix onto the latent variable space. $\hat{y}$ is the predicted output.

Following the prediction procedure, necessary changes in the scores, Δt, are identified which will ensure that the predicted end-point measurement will match the set-point, y_{sp}. Flores-Cerrillo and MacGregor [5] proposed identifying the value of Δt which minimized the following cost function.

$$\min \quad (\hat{y} - y_{sp})^T Q_1 (\hat{y} - y_{sp}) + \Delta t^T Q_2 \Delta t + \lambda T^2 \tag{5.5}$$

st $\hat{y}^T = (\Delta t + \hat{t}_{present})^T Q^T$

$$T^2 = \sum_{a=1}^{A} \frac{(\Delta t + \hat{t}_{present})_a^2}{s_a^2}$$

$$\Delta t_{\min} \le \Delta t = t^T - \hat{t}_{present} \le \Delta t_{\max}$$

where Q_1 and Q_2 are the diagonal weighting matrices for the variables in y and Δt respectively. T^2 is the Hotelling's statistics. s_a is the variance of the score t_a. λ is the weighting matrix which constrains the regions of the score space. $\Delta t_{\min}$ and $\Delta t_{\max}$ are the constraints which define the minimum and maximum values for Δt. This final constraint is included to limit the action of the control system.

To identify the value of Δt which minimizes Equation (5.5), the vector X is considered to be made up of a series of known trajectories, X_1, and future trajectories, X_2. Where $X_1^T = \left[X_{m,measured(0:\theta_i)}^T \quad u_{c,implemented(0:\theta_i)}^T \right]$ is the known trajectory during the time interval (0: θ_i), i.e. these are the measurements recorded earlier in the batch.

$X_2^T = \left[X_{m,future(\theta_i:\theta_f)}^T \quad u_{c,future(\theta_i:\theta_f)}^T \right]$ is the future unknown trajectory during the time interval $(\theta_i{:}\theta_f)$. The change in the future process measurements, $X_{m,future}$, and the manipulated variables, $u_{c,future}$, can be estimated by inverting the PLS model.

$$X_2^T = (t^T - X_1^T W_1)(P_2^T W_2)^{-1} P_2^T \tag{5.6}$$

where $t^T = \Delta t^T + \hat{t}_{present}^T$ is the future score. P_1^T and P_2^T are the loading matrices for X_1^T and X_2^T respectively. W_1 and W_2 are the weighting matrices for X_1^T and X_2^T respectively. The controller then applies the manipulated variable changes contained within X_2 until the next decision point is reached.

5.2 Control Methodology

In the studies undertaken by Flores-Cerrillo and MacGregor [5], the end-point controller was used to regulate a single end-point quality measurement. In this chapter, the end-point controller is applied to regulate the end-point NIR spectral measurements that are recorded in the batch simulation introduced in Section 4.4. The performance of the end-point controller is then compared with that achieved using the Score-Based MPC Controller (Sc-MPC). The structure of the end-point controller is described in more detail in the following section.

5.2.1 End-Point Control

The structure of end-point controller is shown in Figure 5.2, which was described in [150]. The key difference when compared to Sc-MPC is that the end-point controller augments the CC system with an end-point controller while Sc-MPC augments the CC system with a conventional MPC controller. The manipulated variable of the end-point controller is the reactor temperature set-point while the controlled variables are the intensities of the end-point NIR spectrum measured over all wavenumbers. The end-point control system adjusts the reactor temperature set point to minimize the difference between the estimated end-point NIR spectrum and the reference spectrum, which is

obtained from the reference batch. The controller can either act at every sampling instant and re-compute the trajectories of the manipulated variables, or as proposed by Flores-Cerrillo and MacGregor [5], the controller can act at a number of decision points during the batch.

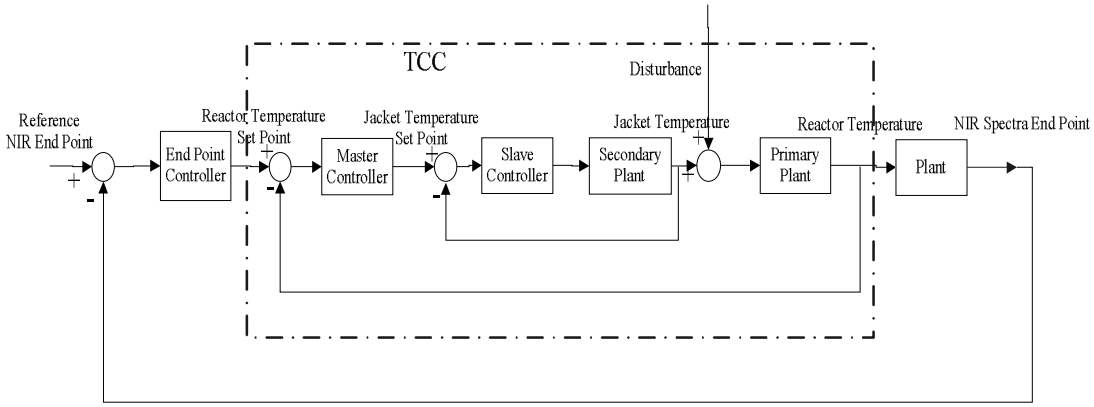

Figure 5.2 Basic Structure of End-Point Control ([150])

The model utilised by the end-point controller was identified using PLS. The data for this model was collected from a varying number of batches, each of which had a PRBS signal with amplitude 0.2 $^{\circ}$C with switching time of 60 seconds applied to the reactor temperature set-point. For application to real batch processes, an important factor will be the effect that the number of batches of data there are available for model identification. Section 5.3.1 describes the effect that increasing the number of batches of data from 3 to 90 had on the performance of the model and controller. This study found that there was limited benefit in using more than 10 batches of data to identify a model and a further 10 to validate it. Therefore in this section it is assumed that only 10 training batches are available. Following the development of the PLS model, the control methodology was applied in two stages. The first stage was to predict the values of the future outputs at each decision point using online and offline process measurements and MVTs available up to the decision point. To predict outputs $\hat{y}$, future measurements $X_{m,fugure}$ needed to be estimated. These were computed using the PLS model and the missing data projection method of Single Component Projection (SCP) [149]. The second stage of the controller

was to determine the necessary control actions to regulate the outputs. The control actions were first taken in the latent variable space and then model inversion was used to obtain appropriate MVTs. The two-stage procedure was taken at every decision point until the batch terminated.

5.3 Case Study

In this section, the Sc-MPC and end-point controller were compared in terms of their ability to regulate the end-point NIR spectrum in the presence of an unmeasured disturbance, and also in their ability to track a set point change.

5.3.1 Case Study 1: Disturbance Rejection

The disturbance considered here was a change in reaction rate k_1^1. Specifically, k_1^1 was decreased by 1.5% at the very start of a batch. This represents a typical disturbance which may occur on a process such as this.

In Sc-MPC, the model used in the MPC controller had an ARX structure, as described in Section 4.4.1, and was identified using the data collected from 10 batches. Similarly, for the end-point controller, 10 batches of data were used to build the PLS model. Cross validation revealed that the root mean square error of calibration (RMSEC) over the 10 validating batches did not decrease appreciatively after five latent variables were used, as shown in Figure 5.3. It shows that the error keeps reducing! Thus, five latent variables were selected when identifying the PLS model. The predicted and actual outputs, in this case the NIR spectra at wavenumber 127, for the testing batches are compared in Figure 5.4. It can be seen that the predicted outputs were different from the actual outputs. However, the prediction accuracy was assumed to be acceptable.

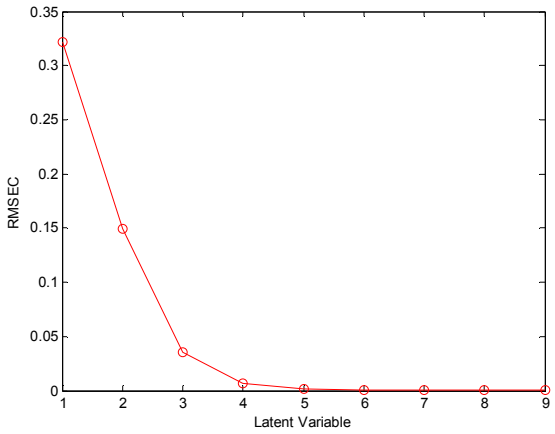

Figure 5.3 Root Mean Square Error of Calibration (RMSEC)

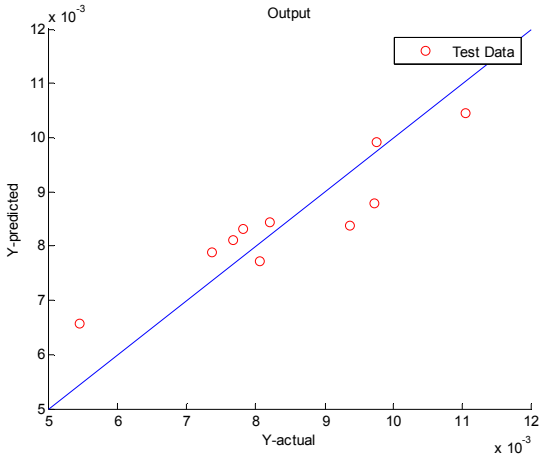

Figure 5.4 The Comparison of Predicted and actual Outputs in the test Data

The resulting NIR spectra at the end-point, obtained when using the end-point controller and Sc-MPC during the unmeasured disturbance are compared in Figure 5.5. The NIR spectrum obtained by Sc-MPC is similar to the nominal NIR spectra, which is desired. However, the NIR spectrum obtained using the end-point controller is significantly different from the nominal NIR spectra, illustrating that this controller is unable to reject the disturbance.

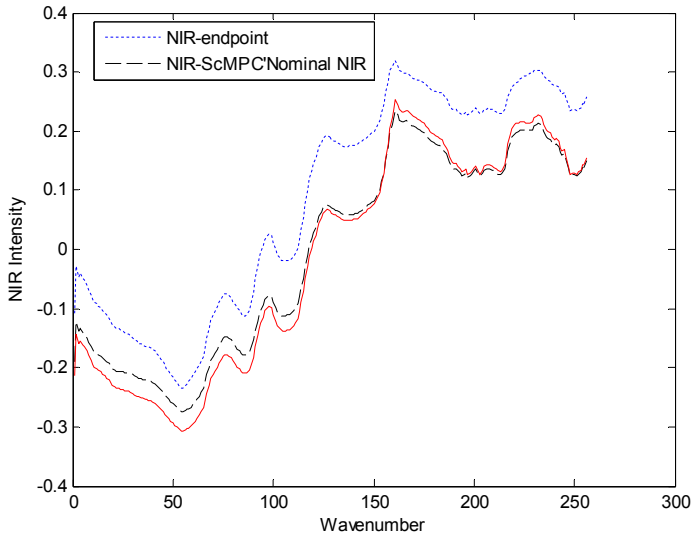

Figure 5.5 Comparison of NIR Spectra at Batch End-point

Figure 5.6 shows the reactor temperature for each of the control systems. This figure confirms that the end-point controller makes little change to the reactor temperature trajectory. However, Sc-MPC increases the temperature to reduce the impact of the disturbance. Sc-MPC adjusts the reactor temperature at every sampling number, while the end-point controller adjusts the reactor temperature at decision point 400.

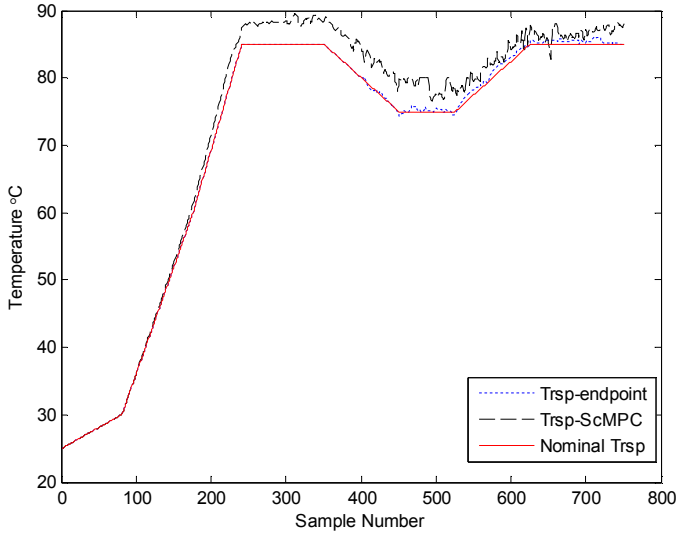

Figure 5.6 Comparison of Reactor Temperature Set-Point

A number of tests were conducted with varying sizes of reaction rate disturbance. The performances of Sc-MPC and the end-point controller for these disturbances are compared in Table 5.1. For different reaction rate changes, the sum of squared (SSQ) residuals of controlled and nominal NIR spectra for Sc-MPC and the end-point controller are provided in Table 5.1. From this table it can be seen that Sc-MPC was able to control the end-point NIR spectrum when the reaction rate changes were within a certain range between -1.5% and 1.5%. Figure 5.7 shows the performance of the control system when the reaction rate was changed by -2%. This figure shows that a significant difference exists between the controlled and nominal NIR spectra, indicating the poor performance of the controller.

In contrast, SSQ for the end-point controller was large for all reaction rate changes investigated. Thus, the end-point controller was unable to reject the disturbances. This result might be expected as there is no real feedback in the end-point controller to reflect the effect of the disturbance. The end-point controller acts more in a feedforward manner.

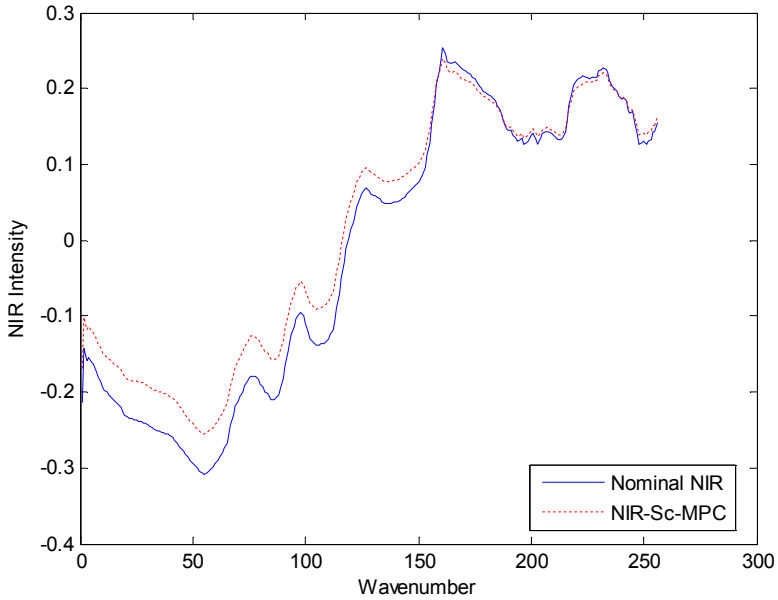

Figure 5.7 Comparison of NIR Spectra at Batch End-point

Table 5.1 Control Performances of Sc-MPC and the End-Point Controller

Percentage of Reaction Rate Change	SSQ for Sc-MPC	SSQ for the End-Point Controller
-3%	0.8355	8.8246
-2.5%	0.5140	6.4571
-2%	0.3261	4.3531
-1.5%	0.1135	2.5823
-1%	0.0478	1.2173
-0.5%	0.0102	0.3323
0.5%	0.0075	0.2936
1%	0.0257	1.2767
1.5%	0.0499	3.0106
2%	2.5040	5.5547

5.3.2 Case Study 2: Set Point Change

The performance of the end-point control method in tracking a set-point change is demonstrated in this test. The algorithms were first applied to track a set-point change in the NIR spectrum at the specific wavenumber of 127, and then to a set-point change across the entire NIR spectrum.

Set-Point Change for the NIR Spectrum at Specific Wavenumber 127

In this case study, the end-point control algorithm was applied to track a set-point change in the NIR spectrum at the end point of the batch, at the specific wavenumber of 127. The wavenumber of 127 was selected as this was the position where one of the peak

intensities was located. In this test the set-point for the NIR spectrum at wavenumber 127 was increased by 0.01, compared with its nominal value.

In this test, a single decision point was used, with that point occurring at sample number 400, with each batch lasting for 750 sampling points. The effect altering the decision point is discussed in the following section, where it is shown that the decision point of 400 samples provided the most accurate results.

When this control system was first implemented, the results were very poor. Investigation showed that this was because the control system did not consider the constraints on the process. For example, the reactor and jacket temperatures can only increase up to a maximum rate. Without considering this limit the performance of the control system was compromised. Based on dynamic analysis of the simulation, upper and lower limits for the rate of change of reactor temperature were set to $+12.5^{o}$ c/min and -3.75^{o} c/min respectively. For jacket temperature, the upper and lower limits were set to be $+25^{o}$ c/min and -37.5^{o} c/min respectively. When the rate of change of reactor or jacket temperature exceeds the limit, it is constrained to the limit. This is the hard constraints that are physically impossible to exceed

Figure 5.8 to 5.16 show the results when using the end-point controller. From these figures, the following observations can be made:

- The controlled and nominal NIR spectra of end product, were obtained and compared in Figure 5.8. For clarity, the NIR spectrum between wavenumbers 120 and 135 is magnified and is shown in the inset in the lower right corner of the figure. It can be seen that the controlled NIR spectrum of end product deviated from the nominal NIR spectrum and tracked the NIR set-point change well. The deviation of the controlled NIR spectrum at wavenumber 127 from its nominal value was 0.0116, which compares well with the set-point change, which was 0.01. The error was caused by plant-model mismatch.
- To achieve the required end-point quality, the controller adjusts the actual reactor temperature set point following the decision point. The actual and nominal reactor

temperature set-points are compared in Figure 5.9. The actual and nominal reactor temperature set-points between samples 380 and 750 are magnified and are shown in the inset. It can be see that the actual reactor temperature set-point deviates slightly from the nominal reactor temperature set-point. With the adjustment of reactor temperature set-point, the reactor temperature also changes accordingly. Figure 5.10 compares the actual and nominal reactor temperatures, with the period between samples 380 and 750 magnified in the inset. A small change between the actual and nominal reactor temperature can be observed.

- To achieve the reactor temperature set-point, the jacket temperature set-point has been adjusted. The actual and nominal jacket temperature set-points are compared in Figure 5.11. It can be seen that the jacket temperature set-point required a significant adjustment. With the adjustment of the jacket temperature set-point, the jacket temperature changes accordingly. A comparison of the actual and nominal jacket temperatures is shown in Figure 5.12.
- The limits imposed on the rate of change of reactor and jacket temperatures should also be checked. These changes are shown in Figure 5.13 and Figure 5.14 respectively, and it can be observed that they are all within constraints.

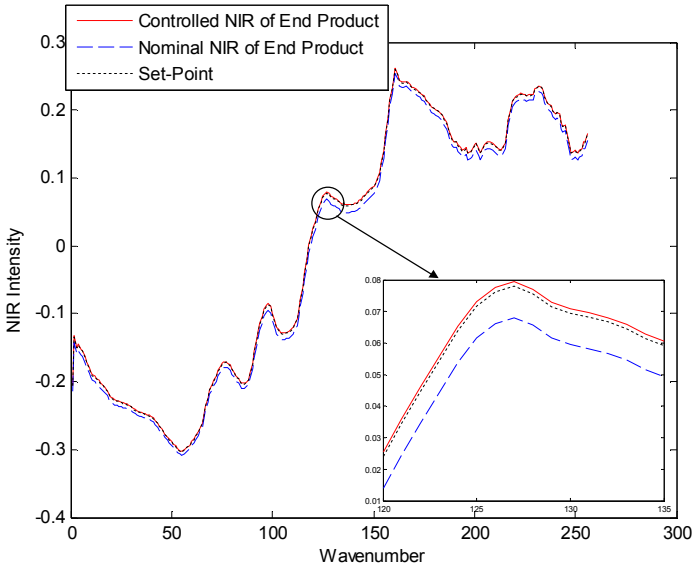

Figure 5.8 Comparison of Controlled and Nominal NIR Spectra of End Product, and the Set-Point

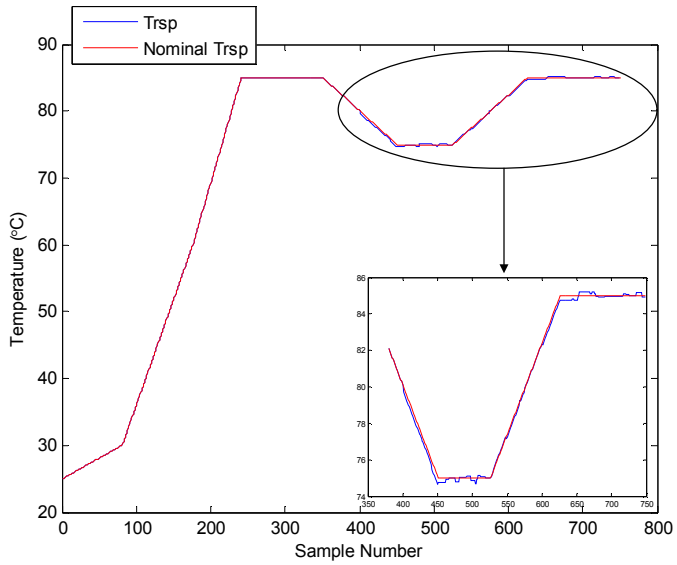

Figure 5.9 Comparison of Actual and Nominal Reactor Temperature Set-Points (Trsp)

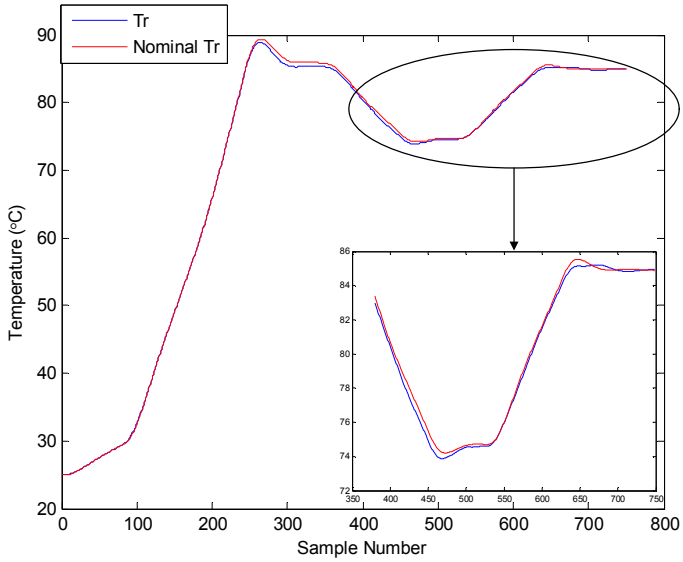

Figure 5.10 Comparison of Actual and Nominal Reactor Temperature (Tr)

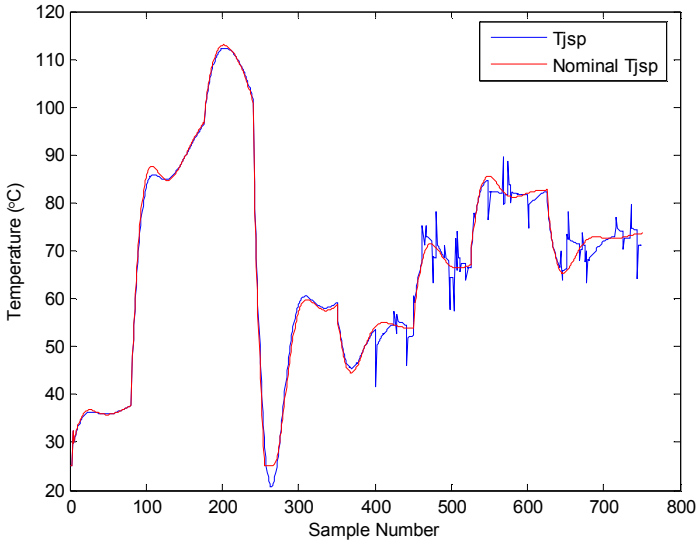

Figure 5.11 Comparison of Actual and Nominal Jacket Temperature Set-Points (Tjsp)

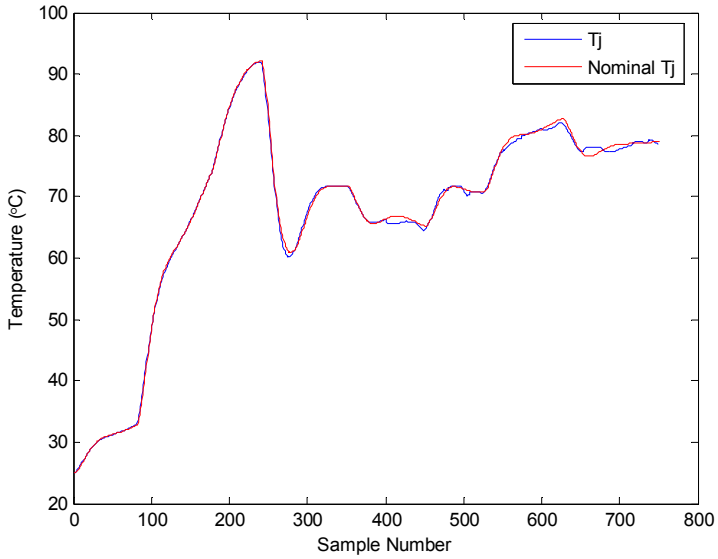

Figure 5.12 Comparison of Actual and Nominal Jacket Temperature (Tj)

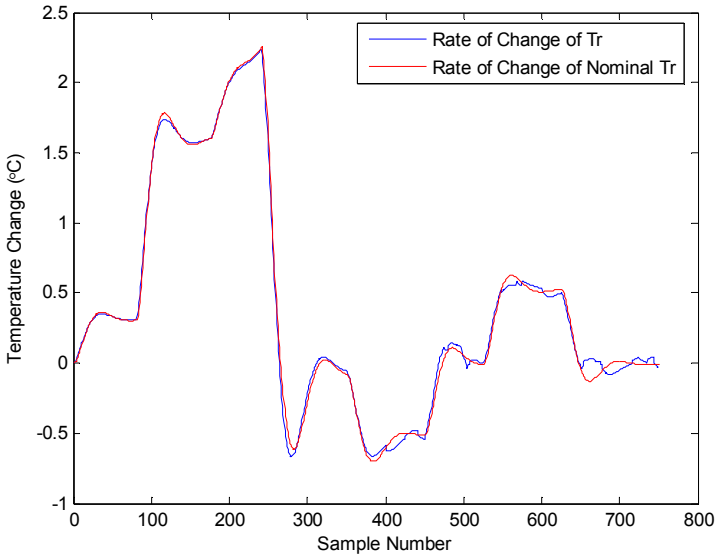

Figure 5.13 Comparison of Change Rate of Actual and Nominal Reactor Temperature (Tr)

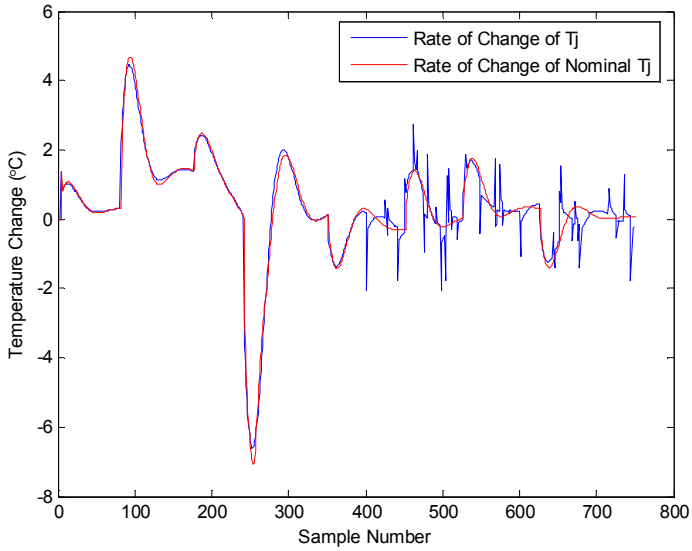

Figure 5.14 Comparison of Change Rate of Actual and Nominal Jacket Temperature (Tj)

In conclusion, the end-point controller was able to track the set point change in this case study. The process variables were adjusted to achieve the set-point change, and they were well within the constraints. However, the selection of decision points and the number of batches used to identify the model are important to consider.

The influence of the number of batches used in modeling was first tested. The set-point change chosen was also the same, namely a step change of 0.01 in the NIR intensity at a wavenumber of 127. The decision point was initially chosen to be at sample time 400, approximately half-way through the batch. The performances of the controller when different numbers of batches were used to identify the model are compared in Table 5.2. In this study a total of 100 batches of data were made available, and the numbers of batches k selected for model building were 3, 5, 10, 20, 50 and 90. For each test, k batches of data were randomly selected to identify the PLS model, and the controller was then applied. The difference between the controlled variable, actual NIR spectrum at wavenumber 127, and its nominal value, were then calculated. This procedure was repeated 20 times, and means and standard deviations of the control performance were determined and shown in Table 5.2. This table shows that the difference between the actual and nominal NIR values is relatively small. The standard deviations can be seen to reduce as the number of batches increases to 10, which means the control performance of the control system is more consistent. However, the performance of the controller does not improve considerably when more than 10 batches are used to identify the model. This suggests that for this study 10 batches is reasonable fro model identification.

Table 5.2. The Comparison of Control Performance with Different Number of Batches

Number of Batches (k)	Mean of difference between NIR Spectrum and its Nominal Value at Wavenumber 127	Standard Deviation of difference between NIR Spectrum and its Nominal Value at Wavenumber 127
3	0.0112	0.0017
5	0.0115	0.0024
10	0.0113	3.9244×10^{-4}
20	0.0114	6.3113×10^{-4}
50	0.0115	3.1866×10^{-4}
90	0.0116	3.2389×10^{-4}

The selection of different decision points is also very important and this too was tested. Various decision points were used in the end-point controller and the performance

compared, as shown in Table 5.3. This table shows that the performances of the controllers were similar, until the decision point was increased to 700. At this point the consistency and performance of the controller reduced. The reason for this is that selecting a decision point too close to the end of the batch will mean there is insufficient time for the process variables to adjust. Although the results when more than one decision point is used are encouraging, more work is required to determine whether they would provide consistently improved control. This is the subject of future research.

Table 5.3 The Comparison of Control Performance with Different Decision Points

Decision Points	Mean of difference between NIR Spectrum and its Nominal Value at Wavenumber 127	Standard Deviation of difference between NIR Spectrum and its Nominal Value at Wavenumber 127
200	0.0118	2.3560×10^{-5}
300	0.0115	2.7777×10^{-4}
400	0.0111	4.3483×10^{-4}
500	0.0111	5.9751×10^{-4}
600	0.0118	6.8148×10^{-4}
700	0.0124	0.0019
[300 500]	0.0114	2.9739×10^{-4}
[500 700]	0.0113	6.2975×10^{-4}
[200 400 600]	0.0116	9.3067×10^{-5}
[500 600 700]	0.0114	8.4277×10^{-4}
[200 300 400 500 600 700]	0.0116	2.1750×10^{-4}

The performance of the end-point controller when it was applied to track different set-point changes is shown in Table 5.4. The set-point changes were chosen as -0.015, -0.01, -0.0075, -0.005, -0.0025, 0.0025, 0.005, 0.0075, 0.01, 0.015, and 0.02. This table shows that control performance is, as might be expected, better for small changes in set-point. Further investigation found that the process variables, such as the rate of change of jacket

temperature, exceeded constraint limits for the larger set-point changes. For Flores-Cerrillo's end-point controller, if the process variables are within the constraints, the set-point change will be well tracked. However, if the process variables exceed the constraints, the values of process variables are constrained and the process variables required in the controller are not fully implemented, and thus the performance of the controller will deteriorate. This is investigated further in Chapter 6, where a novel end-point controller which takes into consideration process constraints is proposed, and its ability to track larger set-point changes is compared with the Flores-Cerrillo's end-point controller.

Table 5.4. The Comparison of Control Performance for Different Set-Point Changes

Set Point Change	Mean of difference between NIR Spectrum and its Nominal Value at Wavenumber 127	Standard Deviation of difference between NIR Spectrum and its Nominal Value at Wavenumber 127
-0.015	-0.0085	0.0054
-0.01	-0.0042	0.0043
-0.005	-0.0002	0.0032
0.005	0.0075	8.9101×10^{-4}
0.01	0.0112	3.9964×10^{-4}
0.015	0.0148	0.0016

With respect to the ability of Sc-MPC to track a set-point change, the situation becomes complicated. To track a change in the NIR end-point set-point, a new trajectory for the first score must be determined. If an example trajectory exists, as was the case in the first test, then this new trajectory can be applied. However, if a trajectory is unavailable then it may be difficult to compute. The trajectory of the first score is a linear combination of the trajectories of NIR spectra at different wavenumbers. Therefore, finding the trajectories for the NIR spectra at every wavenumber is required which may present difficulties.

Set-Point Change for the Entire NIR Spectrum

To guarantee product quality in a batch process, the entire NIR spectrum of end product should be controlled. This section expands the application of the end-point control methods to a more complicated situation, where the set-point of the entire end-point NIR spectrum changes.

The model used for the controllers was once again identified using the PLS algorithm. On this occasion, the output used in the training data was the entire NIR spectrum of the end product. In other words there were 256 output variables. The number of batches selected for model building, as before, was 10. The decision point was also chosen at sample time 400. The constraints imposed for the process variables were the same as before. In this section, the process variables and end-point conditions obtained with the application of Sc-MPC and the end-point control methods are compared.

Set-point tracking for the entire end-point NIR spectrum was implemented using the previous end-point control method in which the MVTs were manipulated in the reduced space of a latent variable model without consideration of constraints. The following observations were made using this method:

- Figure 5.15 shows the actual and nominal NIR spectra of the end product, as well as the new set-point for the NIR spectrum. The actual NIR spectrum is almost overlapped with the new desired NIR spectrum instead of the nominal NIR spectrum, which shows that the set-point change has been tracked successfully.
- The reactor temperature set-point was adjusted to track the set point change in the NIR spectrum. The comparison of actual and nominal reactor temperature set-points is shown in Figure 5.16. A small change can be observed between the actual and nominal reactor temperature set-points. As a consequence, the actual reactor temperature can also be seen to change from the nominal conditions, as shown in Figure 5.17. The adjustment of reactor temperature is reasonable.
- To achieve the reactor temperature set-point, the jacket temperature set-point was adjusted. The comparison of the actual jacket temperature set-point and its nominal value is shown in Figure 5.18.

- To identify the limitation of the process, the rates of change of reactor and jacket temperatures are shown in Figure 5.20 and 5.20. It can be seen in these two figures that the rates of change of reactor and jacket temperatures were all within the constraints.

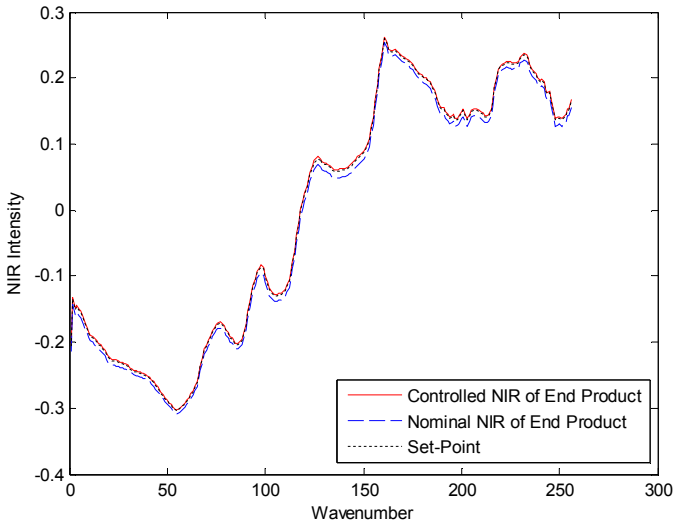

Figure 5.15 Comparions of Nominal NIR, Actual NIR and NIR Set Point

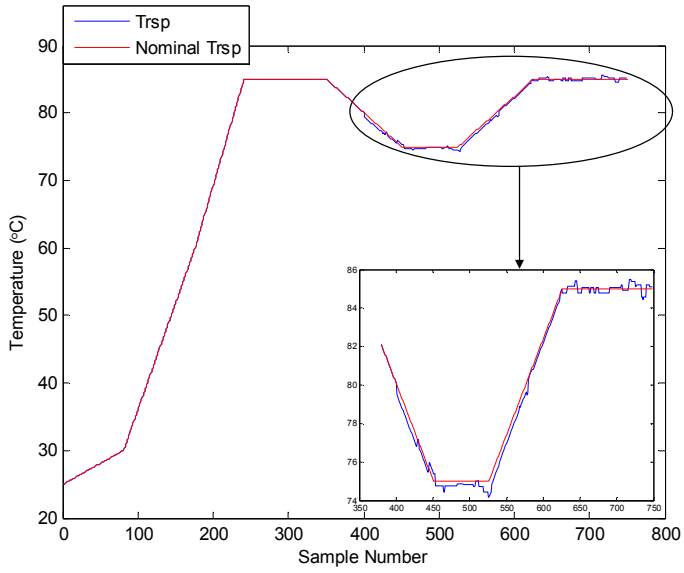

Figure 5.16 Comparison of Actual and Nominal Reactor Temperature Set Points (Trsp)

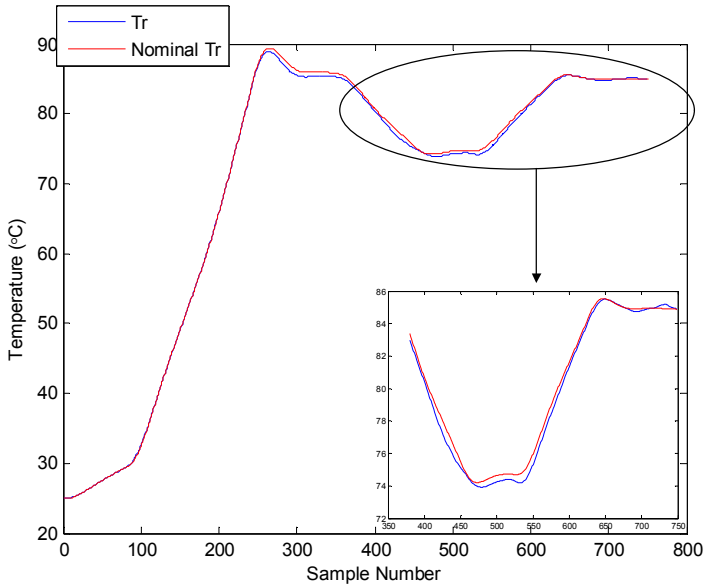

Figure 5.17 Comparison of Actual and Nominal Reactor Temperature (Tr)

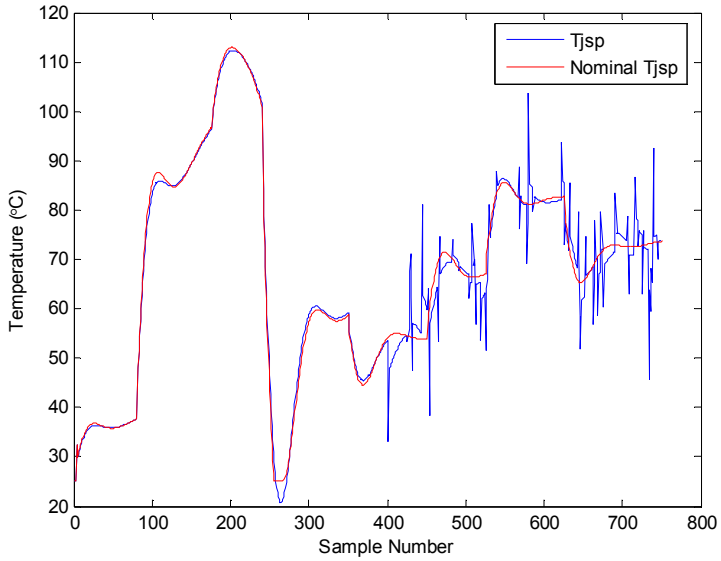

Figure 5.18 Comparison of Actual and Nominal Jacket Temperature Set Points(Tjsp)

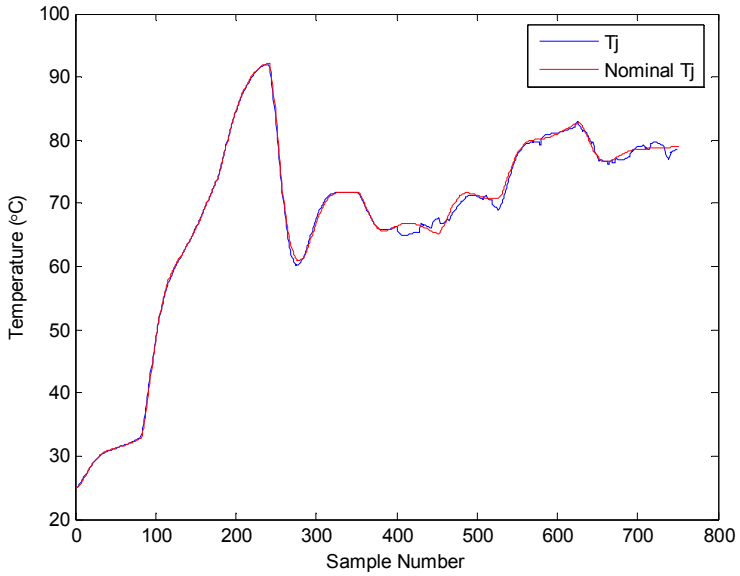

Figure 5.19 Comparison of Actual and Nominal Jacket Temperature (Tj)

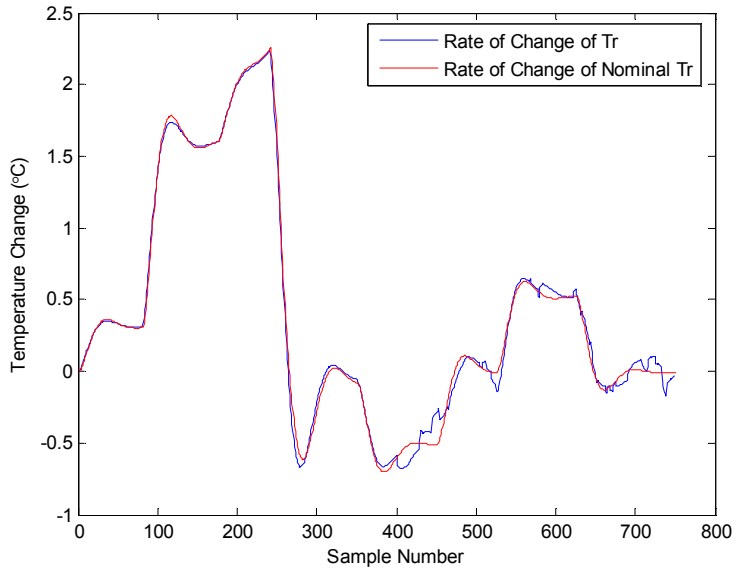

Figure 5.20 Comparison of Actual and Nominal Rates of Change for Reactor Temperature (Tr)

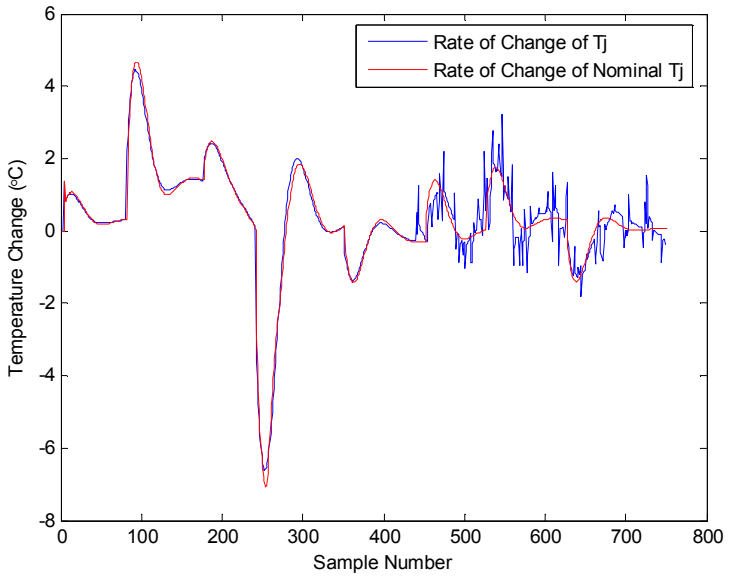

Figure 5.21 Comparison of Actual and Nominal Rates of Change for Jacket Temperature (Tj)

5.4 Summary

This chapter described the application of PCA Score-Based MPC (Sc-MPC) and end-point control to the quality control in batch processes. The first method regulated a single score obtained in the NIR measurement using PCA, while the second method regulated the full end-point NIR spectrum.

Detailed comparison of these methods showed that the end-point method could track set-points very well. However, the performance of end-point method depended on the model of the batch and it was unable to cope with unmeasured disturbances effectively. This limitation could potentially be addressed by selecting other process variables in the model and this is the subject of on-going research. In contrast it was not possible to employ the score based controller for set-point tracking but it was able to reject unmeasured disturbances very well.

6 Quality Control in Batch Processes Using a Proposed End-point Controller

This chapter proposes a novel end-point controller which adjusts the trajectories of manipulated variables (MVTs) in the real process space, while taking into consideration process constraints. To begin, the disadvantages of the Flores-Cerrillo's end-point controller, in which the MVTs are manipulated by controlling the process in the reduced space of a latent variable model, are presented. The two end-point control methods are then applied to track set-point changes, and their performances are compared.

This chapter is divided into the following sections:

1) Presentation of the disadvantages of the Flores-Cerrillo's end-point controller.
2) Proposal of a novel end-point controller.
3) Application of the two end-point controllers to the quality control of a batch process.
4) Summary of this chapter.

6.1 Disadvantages of the Flores-Cerrillo's End-Point Controller

The end point control algorithm proposed by Flores-Cerrillo and MacGregor [5] is an effective method for controlling product quality in a batch process, however the approach assumes that there are no hard constraints on the process variables. Unfortunately, to ensure the maximum control performance, it is necessary to consider the process constraints. To address the constraints problem, Flores-Cerrillo and MacGregor [151] proposed two alternative algorithms. In both of these algorithms, the new MVTs are computed such that the change in scores, t, is as small as possible, while the quality targets are met and the process constraints are not violated. However, t is a linear combination of different process and manipulated variables and has little physical meaning, as shown in Equation (5.3). The constraints on the changes to scores t, $\Delta t_{\min}$ and $\Delta t_{\max}$, are difficult to define. Furthermore, a small change in Δt might lead to a

significant change in the new calculated trajectory, X_2, and a correspondingly large change in the manipulated variables. An illustrative example is shown here.

Suppose the current values of the scores are t_0 and a small change in the scores Δt is applied. Where

$$t_0^T = [1.0894 \quad 1.8847 \quad -0.1564 \quad 0.4015 \quad -0.1930] \tag{6.1}$$

$$\Delta t^T = [0.1 \ -0.2 \ 0 \ 0 \ 0] \tag{6.2}$$

The actual values for the new scores t_1 are now:

$$t_1^T = \Delta t^T + t_0^T = [1.1894 \quad 1.6847 \quad -0.1564 \quad 0.4015 \quad -0.1930] \tag{6.3}$$

Figure 6.1 shows the values of the MVTs calculated by inserting the t_0 and t_1 vectors into Equation (5.6). This figure shows that although the t-vector has only changed slightly, it has had a significant effect on the calculated MVT. This simple example illustrates the danger of solving the control problem in the latent variable space. To address this problem, a method is proposed for solving the control problem, using a MPLS model, in the real-space.

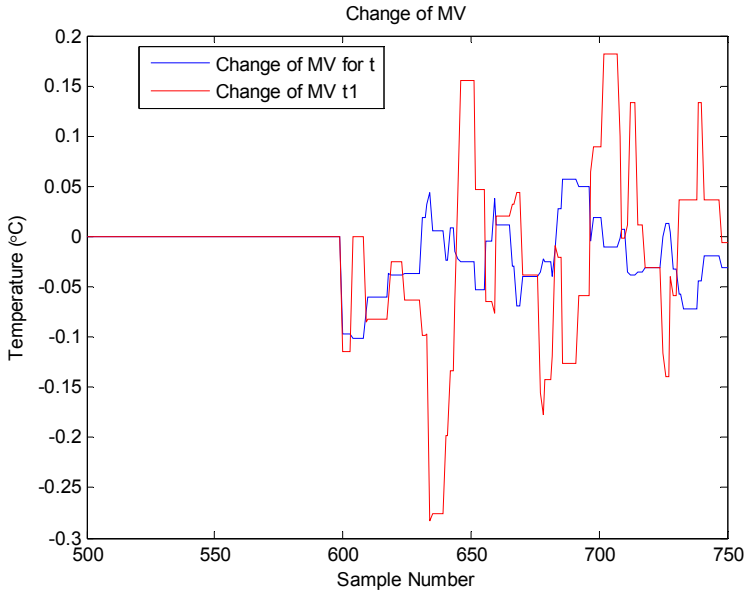

Figure 6.1 Change of Manipulated Variables for t and t_1

6.2 Proposed End-Point Controller

This section proposes a novel end-point control method which manipulates the MVTs in the real space of the manipulated variables, while taking into consideration process constraints.

The data requirements and model building procedure are the same as those described in Section 5.2.1 for Flores-Cerrillo's end-point controller. The difference is in how the control problem is formulated and solved. The control system, once again, makes use of decision points. At each decision point, the end-point outputs are predicted using the MPLS model. If the predicted outputs are sufficiently different from the desired outputs then control action is taken. The control action is computed to minimize the quadratic objectives:

$$\min\ (\hat{y}-y_{sp})^T Q_1 (\hat{y}-y_{sp}) + (\hat{X}-X_{\text{nominal}})^T Q_2 (\hat{X}-X_{\text{nominal}}) \tag{6.4}$$

st $\hat{y} = \hat{X} P Q^T$

$\Delta X_{\min} \le \hat{X} - X_{\text{nominal}} \le \Delta X_{\max}$

where P and Q are loading matrices in the PLS model, respectively. $\hat{y}$ and y_{sp} are the predicted and desired outputs. $\hat{X}$ and X_{nominal} are the predicted and nominal vectors of the process variables. Q_1 and Q_2 are the diagonal weighting matrices for the variables in y and ΔX respectively. $\Delta X_{\min}$ and $\Delta X_{\max}$ are the constraints which define the minimum and maximum values for $\hat{X} - X_{\text{nominal}}$ respectively.

The computation of the quadratic objectives requires all the variables in the vector $\hat{X} - X_{\text{nominal}}$ to be calculated. The vector $\hat{X} - X_{\text{nominal}}$ consists of known trajectories

$\hat{X}_1 - \hat{X}_{\text{nominal},1}$ and future trajectories $\hat{X}_2 - \hat{X}_{\text{nominal},2}$, $\hat{X} - X_{\text{nominal}} = \begin{bmatrix} \hat{X}_1 - X_{\text{nominal},1} \\ \hat{X}_2 - X_{\text{nominal},2} \end{bmatrix}$. Only future trajectories $\hat{X}_2 - \hat{X}_{\text{nominal},2}$ need to be calculated, as $\hat{X}_1$ is already known.

The minimization of quadratic objectives can therefore be restated as follows:

$$\min \ (\hat{y} - y_{sp})^T Q_1 (\hat{y} - y_{sp}) + (\hat{X} - X_{\text{nominal}})^T Q_2 (\hat{X} - X_{\text{nominal}})$$

$$= (\hat{y} - y_{sp})^T Q_1 (\hat{y} - y_{sp}) + \begin{bmatrix} \hat{X}_1 - \hat{X}_{\text{nominal},1} \\ \hat{X}_2 - \hat{X}_{\text{nominal},2} \end{bmatrix}^T \begin{bmatrix} Q_{21} & \\ & Q_{22} \end{bmatrix} \begin{bmatrix} \hat{X}_1 - \hat{X}_{\text{nominal},1} \\ \hat{X}_2 - \hat{X}_{\text{nominal},2} \end{bmatrix} \quad (6.5)$$

The output and input should satisfy the PLS model.

$$(\hat{y} - y_{sp})^{\mathrm{T}} = (\hat{X} - X_{\text{nominal}})^T \theta = \begin{bmatrix} (\hat{X}_1 - \hat{X}_{\text{nominal},1})^T & (\hat{X}_2 - \hat{X}_{\text{nominal},2})^T \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$= (\hat{X}_1 - \hat{X}_{\text{nominal},1})^T * \theta_1 + (\hat{X}_2 - \hat{X}_{\text{nominal},2})^T * \theta_2$$

$$\Delta \hat{X}_{\min,2} \le \hat{X}_2 - \hat{X}_{\text{nominal},2} \le \Delta \hat{X}_{\max,2}$$

where Q_{21} is the diagonal weighting matrix corresponding to ΔX_1

Q_{22} is the diagonal weighting matrix corresponding to ΔX_2

θ^T is the regression coefficient of the PLS model

θ_1^T is the regression coefficient relevant to ΔX_1

θ_2^T is the regression coefficient relevant to ΔX_2

$X_{\text{nominal},1}$ is the nominal value for X_1

$X_{\text{nominal},2}$ is the nominal value for X_2

$\Delta \hat{X}_{\min,2}$ and $\Delta \hat{X}_{\max,2}$ are the constraints which defines the minimum and maximum value for $\hat{X}_2 - \hat{X}_{\text{nominal},2}$ respectively.

Choosing $Q_1 = I$ and $Q_2 = 0$, gives the minimum variance controller, and the quadratic programming objective becomes:

$$\frac{1}{2}(\hat{X}_2 - \hat{X}_{\text{nominal},2})^T H(\hat{X}_2 - \hat{X}_{\text{nominal},2}) + f^T(\hat{X}_2 - \hat{X}_{\text{nominal},2}) \tag{6.6}$$

where

$$H = 2(\theta_2^T Q_1 \theta_2 + Q_{22}) \tag{6.7}$$

$$f^T = 2(\theta_1^T \hat{X}_1 - y_{sp})^T Q_1 \theta_2 \tag{6.8}$$

$\hat{X}_2 - \hat{X}_{\text{nominal},2}$ can be computed by minimizing the quadratic objective Equation (6.6) and thus $\hat{X}_2$ can be obtained. Following this, a series of calculated manipulated variables are applied to the system to adjust the output to be as close to the desired output as possible.

6.3 End Product Quality Control of Batch Processes

In this section, the capabilities of Flores-Cerrillo's end-point controller and the proposed end-point controller in regulating the product quality are compared. First, these algorithms are applied to track the set point change for the NIR spectrum at a specific wavenumber 127. Following this, they are applied to track a set point change over the entire NIR spectrum.

6.3.1 Case Study 1: Control of NIR Spectrum at a Specific Wavenumber

Comparison of Flores-Cerrillo's End-Point Controller and the Proposed End-Point Controller

The ability to track a set-point change in the NIR spectrum at wavenumber 127 was implemented using both Flores-Cerrillo's and the proposed end-point controllers. The decision point was selected at sampling 400. To achieve consistent results, 20 sets of data were collected. Each set of data contained the measurements collected from 10 batch runs. Based on the results reported in Chapter 5, for each set of data a PLS model was

identified using 10 training batches. The controller was then applied to regulate the nominal batch through a series of set-point changes to the spectral intensity at wavenumber 127.

The results of the control runs are presented in Table 6.1. For each set-point change (+0.015 to -0.015) the mean product quality (achieved spectral intensity at wavenumber 127) is provided, together with the standard deviation of the product quality over the 20 repetitions. The results in Table 6.1 show that for set-point changes of 0.005, 0.01 and 0.015, the mean and standard deviation of product quality for the two controllers was very similar. Further investigation found that the process variables for the two controllers in this case were all within the constraints and similar control performances were achieved. However, in tracking the set-point changes of -0.005, -0.01 and -0.015, the control performance of the proposed end-point controller was significantly better than Flores-Cerrillo's end-point controller. Further investigation found that the process variables for the proposed end-point controller were well within the constraints because these were adhered to by the controller. However, the process variables in Flores-Cerrillo's end-point controller did not respect the constraints, and thus the process variables required were not fully implemented, which introduced errors. Further details for the set-point change of -0.01 are now provided.

Table 6.1 Performance of two End-Point Controllers in Tracking Set-Point Change

Set-Point Change	Mean Product Quality (Flores-Cerrillo's End-Point Controller)	Standard Deviation of Product Quality (Flores-Cerrillo's End-Point Controller)	Mean Product Quality (Proposed End-Point Controller)	Standard Deviation of Product Quality (Proposed End-Point Controller)
0.015	0.0148	0.016	0.0149	0.019
0.01	0.0112	$3.9964*10^{-4}$	0.0110	$6.6515*10^{-4}$
0.005	0.0075	$8.9101*10^{-4}$	0.0065	$8.6834*10^{-4}$
-0.005	$-1.6862*10^{-4}$	0.0032	-0.0037	0.0011
-0.01	-0.0042	0.0043	-0.0084	0.0011
-0.015	-0.0085	0.0054	-0.0138	0.0018

For this set-point change, the following observations could be made regarding the two controllers:

- The controlled and nominal end-point NIR spectra compared in Figure 6.2. For clarity, the NIR spectrum between wavenumbers 120 and 135 is magnified and is shown in the inset in the lower right corner of the figure. It can be seen that the NIR spectrum obtained using the proposed end-point controller was closer to the set-point than the NIR spectrum obtained using Flores-Cerrillo's end-point controller. The set-point change was -0.01 for the NIR spectrum at wavenumber 127. After the application of the two controllers, the peak intensity of the NIR spectrum at wavenumber 127 had changed by -0.0084 for the proposed end-point controller, and changed by -0.0045 for Flores-Cerrillo's end-point controller. It shows that the proposed end-point controller was better at tracking the set-point change than Flores-Cerrillo's end-point controller.
- To achieve the required end-point quality, the controller adjusts the actual reactor temperature set point following the decision point. The actual and nominal reactor temperature set-points are compared in Figure 6.3. The actual and nominal reactor temperature set-points between samples 380 and 750 are magnified and are shown in the inset. It can be seen that the actual reactor temperature set-point used in Flores-Cerrillo's controller has a slightly bigger adjustment from the nominal value than that used in the proposed controller. Figure 6.4 compares the actual reactor temperatures obtained using the two controllers and their nominal values, with the period between samples 380 and 750 magnified in the inset. The actual reactor temperatures obtained using the two controllers both have small changes from the nominal value. It is evident that the temperature changes required by Flores-Cerrillo's are significantly larger than those required by the proposed controller.
- To achieve the reactor temperature set point, the jacket temperature set point has been adjusted. The actual jacket temperature set points used in the two controllers and their nominal values are compared in Figure 6.5. It can be seen that the jacket temperature set point required a significant adjustment for Flores-Cerrillo's controller, while it needed a smaller adjustment for the proposed controller. With the adjustment of jacket temperature set point, the jacket temperature changes

accordingly. A comparison of the actual and nominal jacket temperatures is shown in Figure 6.6. The adjustment of the actual jacket temperature used in Flores-Cerrillo's end-point controller was larger than that used in the proposed end-point controller.

- The limits imposed on the rates of change of reactor and jacket temperature should also be checked. These changes are shown in Figure 6.7 and Figure 6.8 respectively. The rate of change of the actual reactor and jacket temperature for Flores-Cerrillo's end-point controller both exceeded the constraints, while those for proposed end-point controller were within the constraints.
- Flores-Cerrillo's end-point controller does not directly consider the process constraints in the control algorithm, and it only constrains the process variables when the process variables exceed the limits after the control action. Thus, large process variables changes are required to achieve good control performance. However, these changes are not achievable in the process, and thus the set point of the NIR spectrum was not well tracked. However, the proposed end-point controller explicitly considers the process constraints in the control algorithm, and thus the process variables were well within the constraints. The desired process variable changes required to achieve good control performance were fully implemented and thus the set point of NIR spectrum was well tracked.

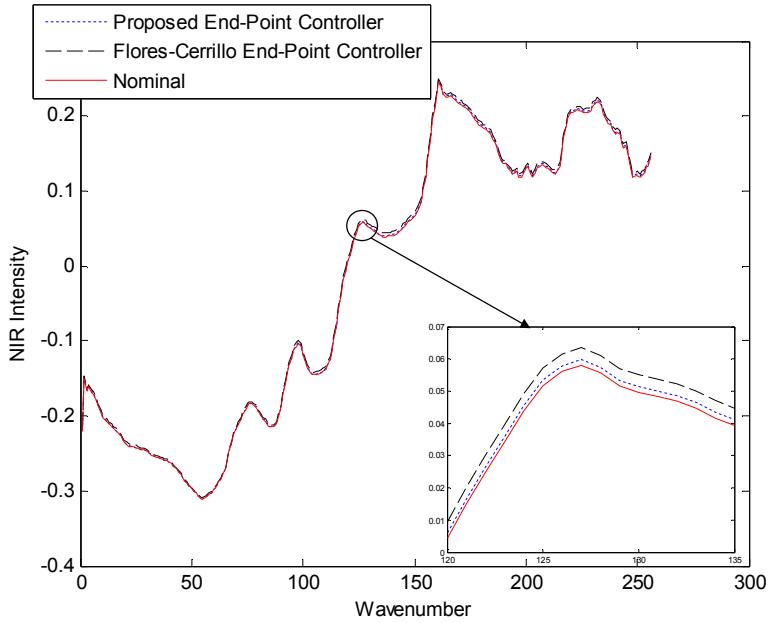

Figure 6.2 Comparison of NIR Spectra obtained using two Controllers and the Set-Point

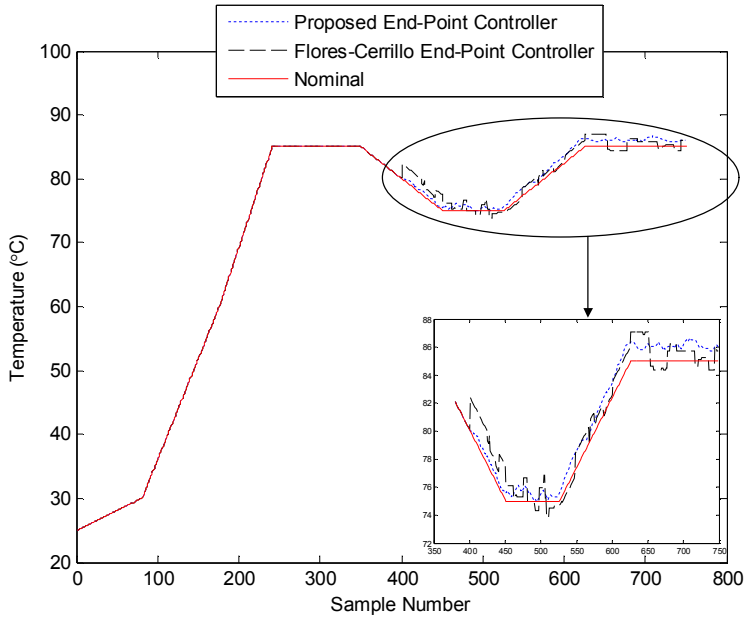

Figure 6.3 Comparison of Actual Reactor Temperature Set Points (Trsp) and their Nominal Values

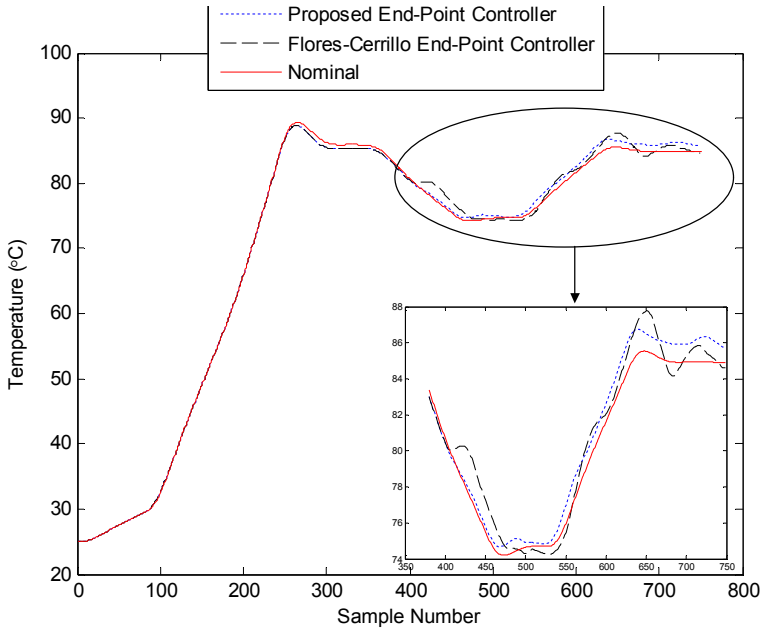

Figure 6.4 Comparison of Actual Reactor Temperatures (Tr) and their Nominal Values

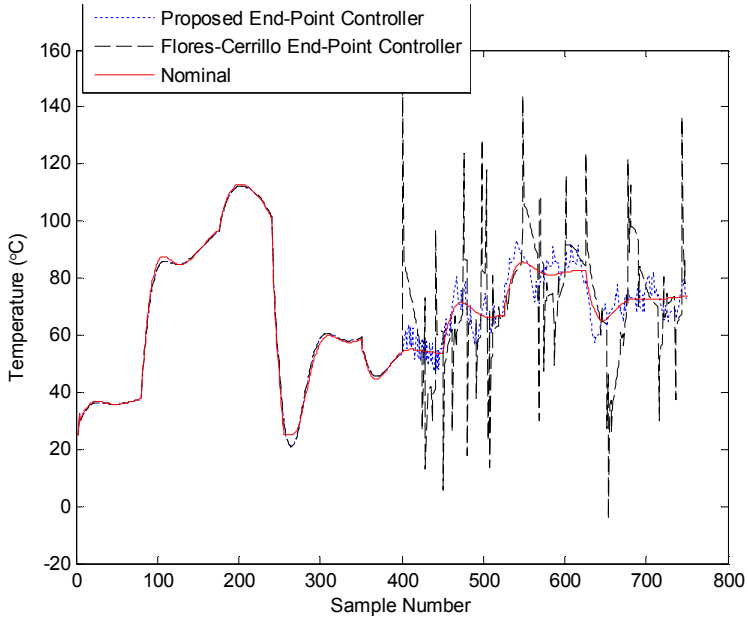

Figure 6.5 Compariosn of Actual Jacket Temperature Set Points (Tjsp) and their Nominal Values

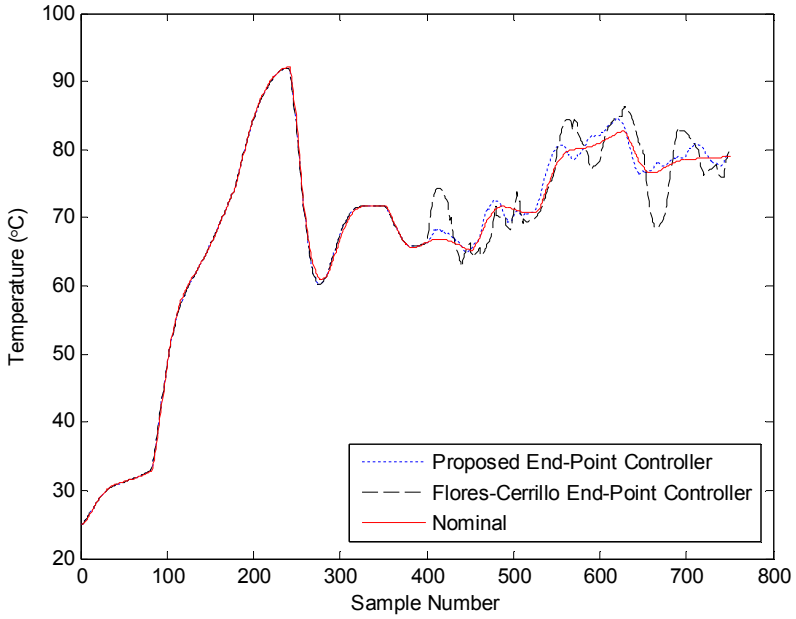

Figure 6.6 Comparison of Actual Jacket Temperatures (Tj) and their Nominal Values

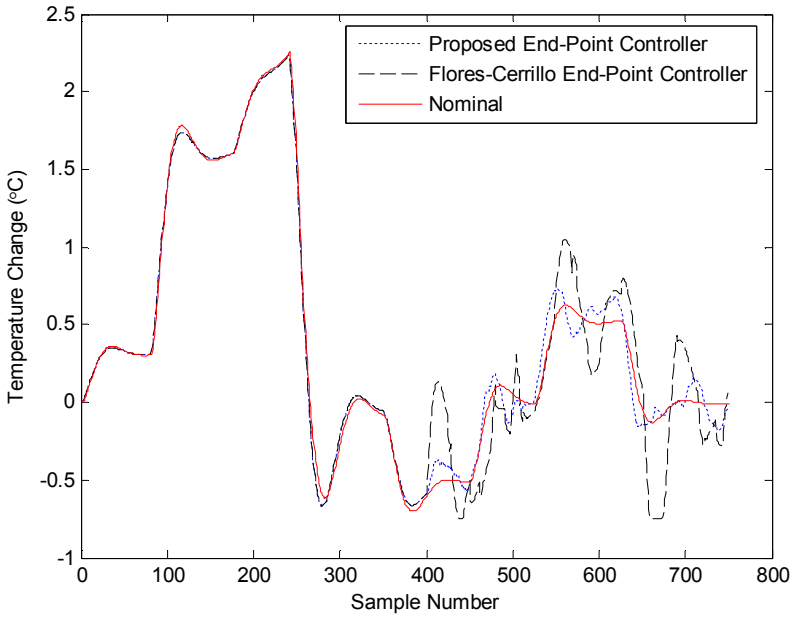

Figure 6.7 Comparison of Actual Rates of Change of Reactor Temperature (Tr) and the Nominal Values

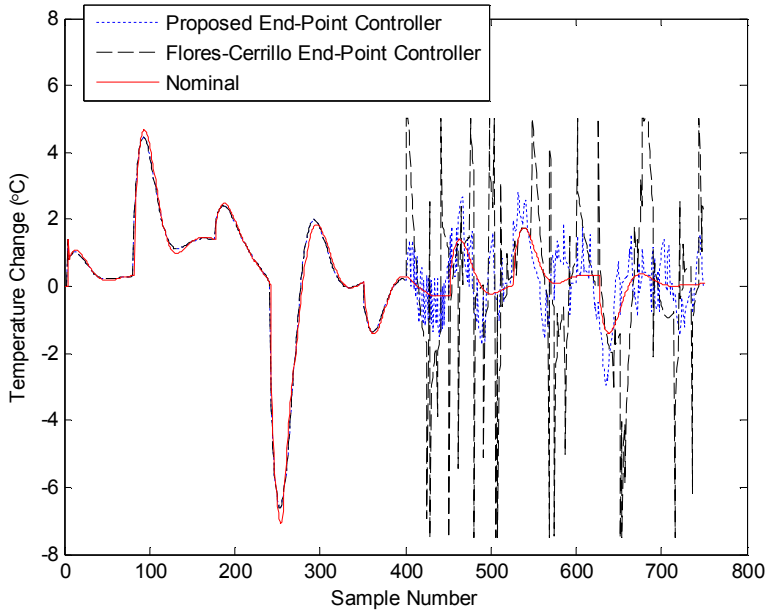

Figure 6.8 Comparison of Actual Rates of Change of Jacket Temperature (Tj) and the Nominal Values

6.3.2 Control of the Entire NIR Spectrum of End Product

The tracking of set-point changes over the entire end-point NIR spectrum using the proposed end-point controller and Flores-Cerrillo's end-point controller is compared in this section. The decision point was selected again to be at sample number 400. The following observations could be made using these methods:

- Figure 6.9 shows the actual NIR spectra obtained using the two controllers, and their nominal values. The actual NIR spectrum obtained using the proposed end-point controller was closer to the set points than those achieved using Flores-Cerrillo's end-point controller.
- The reactor temperature set point was adjusted to track the set-point change in the NIR spectrum. The comparison of actual reactor temperature set points used in the two controllers and their nominal values are compared in Figure 6.10. The proposed end-point controller required a reduced change in the reactor temperature set point than that required by Flores-Cerrillo's end-point controller. The actual reactor temperatures used in the two controllers are compared in Figure 6.11. A smaller

change can also be observed in the reactor temperature when the proposed end-point controller was used.

- To achieve the reactor temperature set point, the jacket temperature set point is also adjusted. The comparison of the actual jacket temperature set points used in the two controllers and their nominal values is made in Figure 6.12. The change of jacket temperature set point used in Flores-Cerrillos's end-point controller was sharper and more significant than that used in the proposed end-point controller. The jacket temperature was also tuned to adapt to the adjustment of jacket temperature set point. The actual jacket temperatures and their nominal values are compared in Figure 6.13. A smaller adjustment was required in the jacket temperature when using the proposed end-point controller.
- Finally, the limits imposed on the process were checked. The rates of change of reactor temperature used in the two controllers and their nominal values are compared in Figure 6.14. The rate of change of reactor temperature used in Flores-Cerrillo's end-point controller exceeded the constraints, while that used in the proposed end-point controller was well within the constraints. In Figure 6.15, the rates of change of jacket temperature used in the two controllers and their nominal values are compared. The rate of change of jacket temperature used in Flores-Cerrillo's end-point controller was sharper and more significant than that used in the proposed end-point controller. The reason is the same as that in cast study 1. The process variables required to achieve good performance were not fully implemented bu Flores-Cerrillo's controller, thus the set-point change was not well tracked.

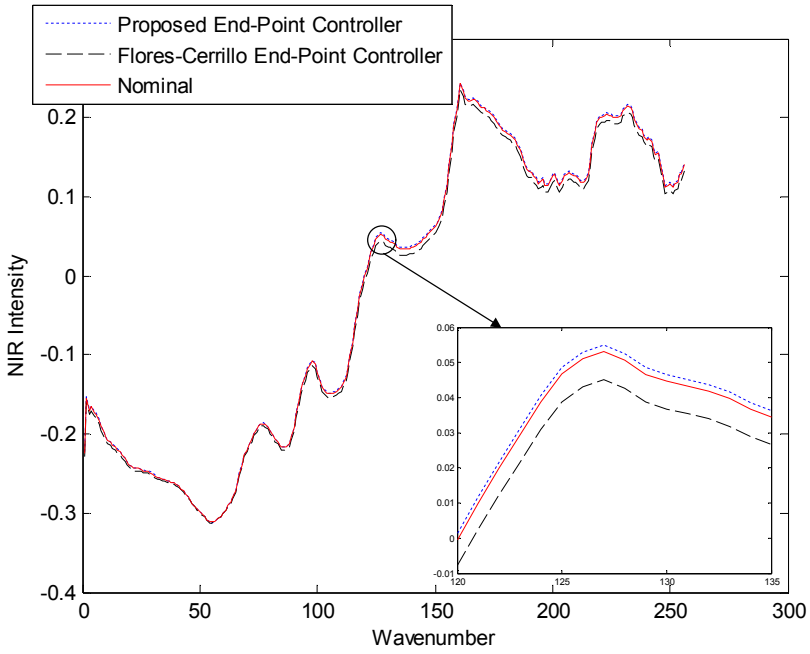

Figure 6.9 Comparison of NIR Spectra obtained using two Controllers and the new Set Point

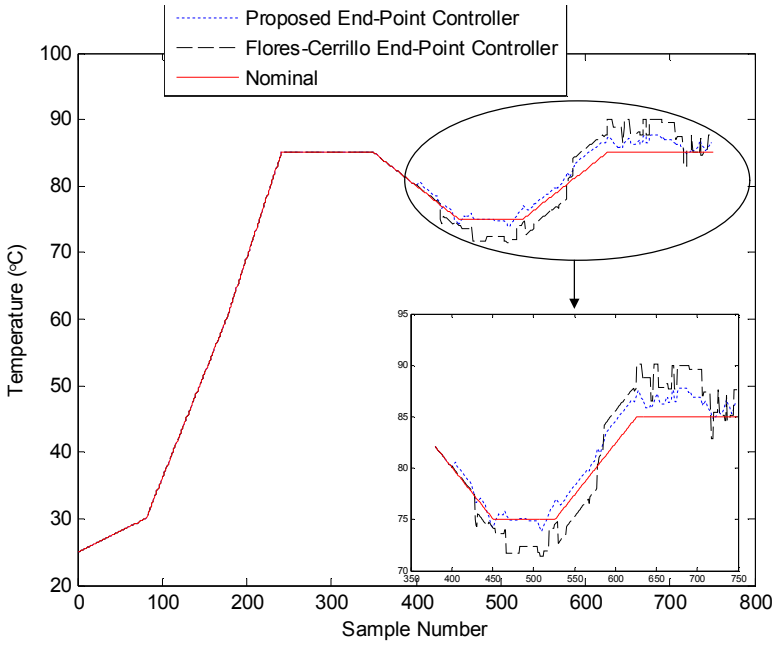

Figure 6.10 Comparison of Actual Reactor Temperature Set Points (Trsp) and their Nominal Values

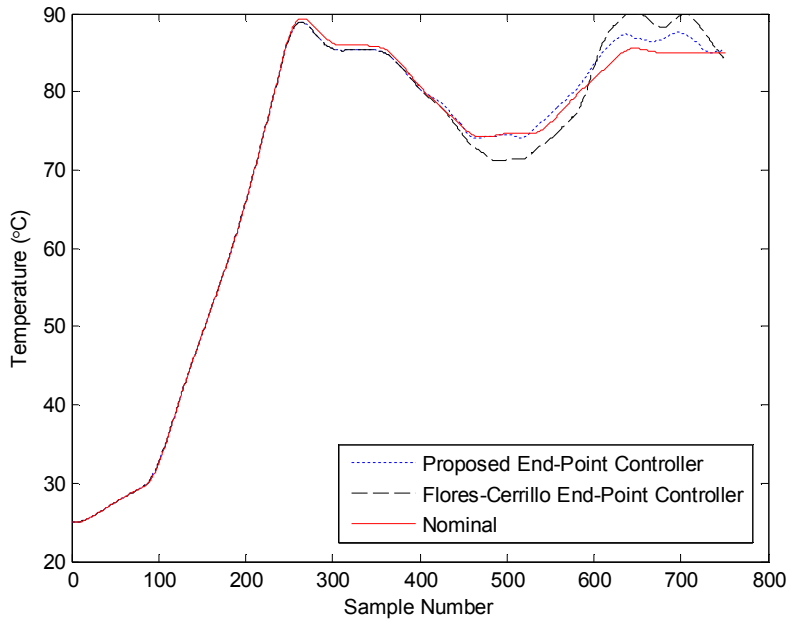

Figure 6.11 Comparison of Actual Reactor Temperatures (Tr) and their Nominal Values

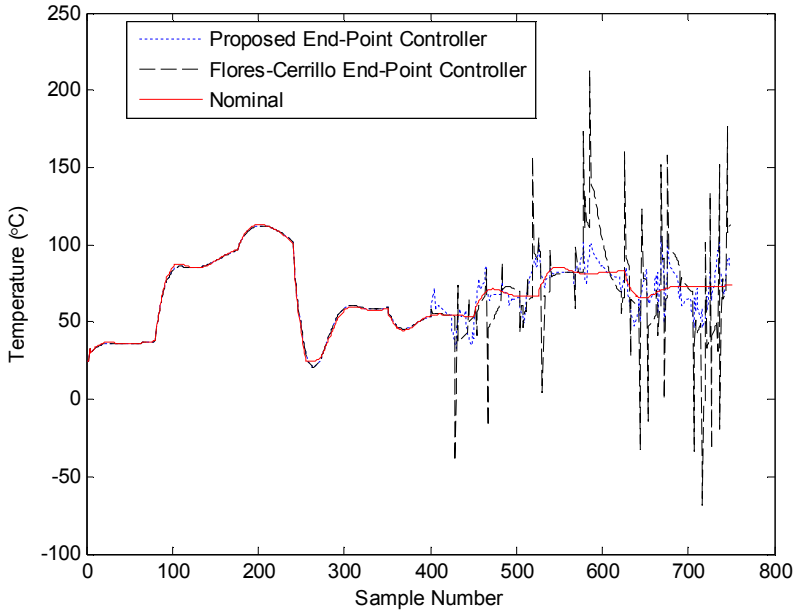

Figure 6.12 Comparison of Actual Jacket Temperature Set Points (Tjsp) and their Nominal Values

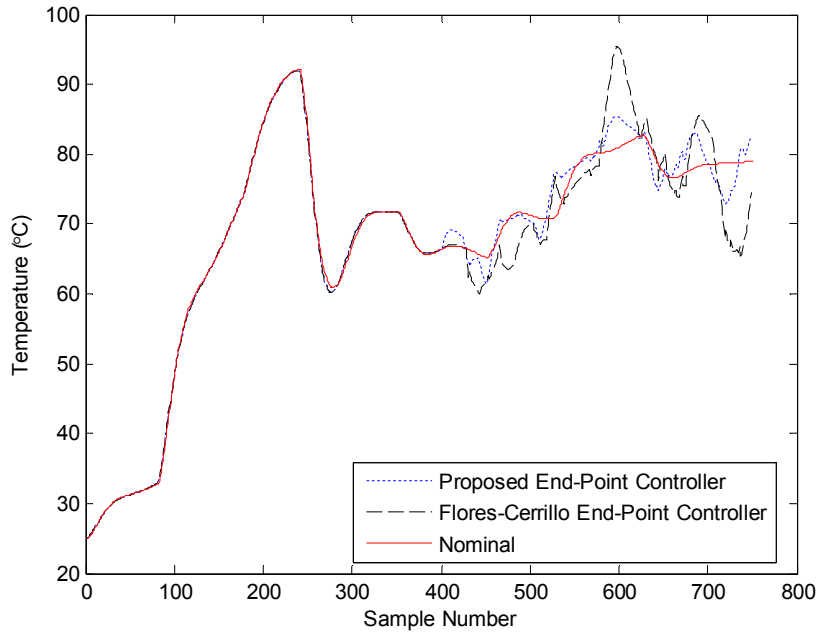

Figure 6.13 Comparison of Actual Jacket Temperatures (Tj) and their Nominal Values

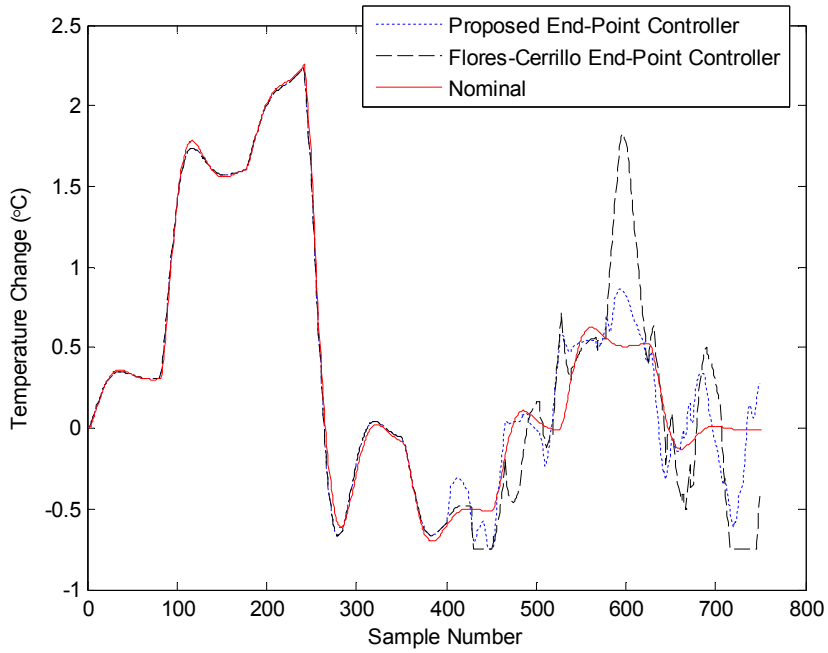

Figure 6.14 Comparison of Actual Rates of Change of Reactor Temperature (Tr) and the Nominal Values

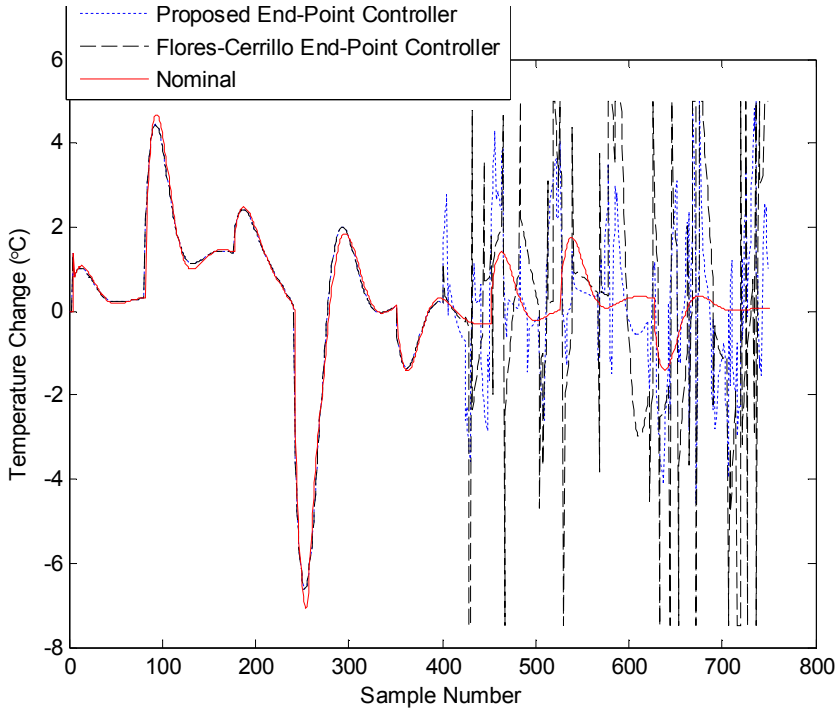

Figure 6.15 Comparison of Actual Rates of Change of Jacket Temperature (Tj) and the Nominal Values

The above results show that the proposed end-point controller is better at tracking the large set-point change than Flores-Cerrillo's end-point controller. With the application of the proposed end-point controller, the actual process variables are fully implemented and good set-point tracking is achieved. However, Flores-Cerrillo's end-point controller does not directly take into consideration process constraints in the control algorithm, and the actual process variables are not fully implemented in the process, and thus the performance of the controller is reduced.

6.4 Summary

This chapter has presented two end-point control methods to track set-point changes in the NIR spectrum. Flores-Cerrillo's end-point controller calculates the MVTs in the reduced space of a latent variable model. The proposed end-point controller calculates the MVTs by controlling in the real space of the process while taking into consideration process constraints. The capabilities of the two end-point control methods in tracking the

set-point change in the NIR spectrum of end product are compared. The results show that the proposed end-point control method is better at tracking the set-point changes.

7 Conclusions and Future Work

This chapter provides conclusions of the work presented in this thesis and suggestions for further work which should be explored to extend the results of the thesis.

The chapter is divided into the following sections:

1) Summary of the main research from the work reported in this thesis.
2) Suggestions for future work in this field.

7.1 Summary and Conclusions

This thesis has focused on the application of multivariate statistical analysis methods to Raman images of pharmaceutical tablets and various control strategies for application to batch processes. Chapter 2 provided a literature review of the multivariate statistical analysis methods and batch process control in industrial processes. Chapter 3 addressed multivariate statistical analysis methods and their applications in the pharmaceutical industry. Different strategies for controlling product quality in batch processes were discussed in Chapters 4-6.

7.1.1 Summary of Multivariate Statistical Analysis

To better understand the process, chemical imaging techniques including NIR spectroscopy and Raman spectroscopy are being used with greater frequency in the pharmaceutical industry. To extract useful information from the large quantities of data these instruments routinely collect, multivariate data analysis methods, PCA and ICA were applied to extract the chemical information from the measurements.

Chapter 2 presented a literature review of the chemical imaging techniques and multivariate statistical analysis methods. Specific attention was paid to the application of multivariate statistical analysis methods in the pharmaceutical industry.

In Chapter 3, information regarding the composition of a pharmaceutical tablet was extracted using both PCA and ICA. The ability of each of these techniques to extract the spectrum and concentration of specific compounds in the tablet was compared and their ability to provide a robust method for identifying the composition of pharmaceutical tablets assessed. A map showing a visual representation of the gross spatial distribution of the constituent compounds was provided.

7.1.2 Summary of Batch Process Quality Control

Quality control is a key issue in batch processes. Quality control can be achieved by controlling several process variables and making them track specified trajectories. An overview of quality control in batch processes was given in Chapter 2.

The general problem in batch processes is that fixed process variable trajectories do not guarantee consistent product quality. Change in reaction rates or inclusion of a new raw material, for example, can introduce new reaction pathways, and cause the quality of the final product to change significantly. To address this problem, two control approaches where NIR spectra were incorporated as feedback information are proposed in Chapter 4. One approach, Wn-MPC, selects a single wavenumber, corresponding to a specific spectral peak, as a controlled variable. The second approach uses PCA to extract the information from the spectral measurement and uses the compressed information from this analysis as a controlled variable. This method was referred to as PCA Score-Based MPC (Sc-MPC). The two control strategies were applied to control final product quality in a simulated chemical reaction process and their performance compared. Sc-MPC achieved a satisfactory performance with no user interaction. In contrast, the performance achieved using Wn-MPC was highly dependent on the choice of wavenumber that was to be controlled.

To control the entire NIR spectrum, an end-point control algorithm was used in Chapter 5. The end-point control algorithm and Sc-MPC were applied in two case studies: reaction rate change and set point change. The performances of the end-point control and Sc-MPC

in these two case studies was compared, and it was shown that the end-point controller was able to track set point changes well.

In Chapter 6, a novel end-point controller which adjusted the trajectories of the manipulated variables (MVTs) in the real process space, while taking into consideration process constraints was proposed. This proposed end-point control algorithm and the Flores-Cerrillo's end-point controller were applied to track set-point changes in the NIR spectrum, and their performance compared. The results showed that the proposed end-point control method was better at tracking set-point changes.

7.2 Recommendations for Future Work

Though there is much work left to do, only the most promising directions are suggested:

(1) In the application of PCA and ICA to analyze Raman images of pharmaceutical tablets, there are some issues that needed to be addressed. In the PCA and ICA models, choosing the number of components retained is not obvious, especially when the reference spectra of the constituent compounds are unknown. It is difficult to decide which component has relevant features and which mainly consists of noise. Another issue is the peak shift problem, which causes the PCA/ICA methods to identify two components. This issue also needs further research.

(2) In batch process control, the performance of Wavenumber-Based MPC control (Wn-MPC) is highly dependent on the selection of the wavenumber which is to be controlled. Choosing the wavenumber at which the NIR spectrum is controlled is a non-trivial task.

(3) In the end-point controller, the selection of decision points is very important. There are no general guidelines for the selection procedure. This is an issue that needs further research.

8 Reference

1. FDA, *Guidance for Industry PAT—A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance*. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070305.pdf, 2004.
2. Narinder Singh Sahni, T.I., Tormod Næs *The use of experimental design methodology and multivariate analysis to determine critical control points in a process*. Chemometrics and Intelligent Laboratory Systems, 56(2):105-121, 2001.
3. Kourti, T. *The Process Analytical Technology initiative and multivariate process analysis, monitoring and control.* Analytical and Bioanalytical Chemistry, 384(5):1618-2650, 2006.
4. Rockwell, *Rockwell Automation Support For Process Analytical Technology in MES*. Available from: http://literature.rockwellautomation.com/idc/groups/literature/documents/wp/life-wp004_-en-p.pdf
5. Flores-Cerrillo, J. and MacGregor, J.F. *Control of batch product quality by trajectory manipulation using latent variable models.* Journal of Process Control, 14(5):539-553, 2004.
6. Reich, G. *Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications.* Advanced Drug Delivery Reviews, 57(8):1109-1143, 2005.
7. McCreery, R.L., *Raman spectroscopy for chemical analysis* Wiley-Interscience, 2000.
8. Huang, S.X. and Chen, Y. *Ultrasensitive fluorescence detection of single protein molecules manipulated electrically on Au nanowire.* Nano Letters, 8(9):2829-2833, 2008.
9. Evans, C.L. and Xie, X.S. *Coherent Anti-Stokes Raman Scattering Microscopy: Chemical Imaging for Biology and Medicine.* Annual Review of Analytical Chemistry, 1:883-909, 2008.
10. Lee, S.C., Kim, K., Kim, J., Lee, S., Yi, J.H., Kim, S.W., Ha, K.S., and Cheong, C. *One micrometer resolution NMR microscopy.* Journal of Magnetic Resonance, 150(2):207-213, 2001.
11. Treado, P.J. *Chemical imaging: Visualizing material and process chemistry.* Abstracts of Papers American Chemical Society, 210(1-2):100, 1995.
12. Baianu, I.C., Costescu, D., You, T., Lozano, P.R., Hofmann, N.E., and Korban, S.S. *Near-infrared microspectroscopy, fluorescence microspectroscopy, infrared chemical imaging and high-resolution nuclear magnetic resonance analysis of soybean seeds, somatic embryos and single cells*. 2004, Amer Oil Chemists Soc: Champaign, USA. p. 241-273.
13. Priore, R.J., Olkhovyk, O., Drauch, A., Treado, P., Kim, M., and Chao, K. *Recent advances in chemical imaging technology for the detection of contaminants for food safety and security* In Sensing for Agriculture and Food Quality and Safety, Orlando, FL, USA, 2009.
14. Raghavachari, R., *Near-infrared Applications in Biotechnology* CRC Press, 2000.

15. Guthrie, J. and Wedding, B. *Robustness of NIR calibrations for soluble solids in intact melon and pineapple.* Journal of Near Infrared Spectroscopy, 6(1):259–265, 1998.
16. Slaughter, D.C. *Non-destructive determination of soluble solids in papayas using near infrared spectroscopy.* Journal of Near Infrared Spectroscopy, 7(4):223–228, 1999.
17. Balabin, R.M. and Safieva, R.Z. *Gasoline classification by source and type based on near infrared (NIR) spectroscopy data*. Fuel, 87(7):1096-1101, 2008.
18. Stallard, B.R., Garcia, M.J., and Kaushik, S. *Near-IR Reflectance Spectroscopy for the Determination of Motor Oil Contamination in Sandy Loam.* Applied Spectroscopy, 50(1):299-424, 1996.
19. Luypaert, J., Massart, D.L., and Heyden, Y.V. *Near-infrared spectroscopy applications in pharmaceutical analysis.* Talanta, 72(3):865-883, 2007.
20. Burns, D.A. and Ciurczak, E.W., *Handbook of Near-Infrared Analysis.* CRC, 2001.
21. Kramer, K. and Ebel, S. *Application of NIR reflectance spectroscopy for the identification of pharmaceutical excipients.* Analytica Chimica Acta, 420(2):155-161, 2000.
22. Wang, Z. and Xiang, B. *Application of artificial neural network to determination of active principle ingredient in pharmaceutical quality control based on near infrared spectroscopy.* Microchemical Journal, 89(1):52-57, 2008.
23. Thosar, S.S., Forbess, R.A., Kemper, M., and Shukla, A.J. *Determination of copolymer ratios of poly(lactide-co-glycolide) using near-infrared spectroscopy.* Journal of Pharmaceutical and Biomedical Analysis, 20(1-2):107-114, 1999.
24. Kourti, T., Lee, J., and Macgregor, J.F. *Experiences with industrial applications of projection methods for multivariate statistical process control.* Computers & Chemical Engineering, 20(Supplement 1):S745-S750, 1996.
25. Ito, M., Suzuki, T., Yada, S., Kusai, A., Nakagami, H., Yonemochi, E., and Terada, K. *Development of a method for the determination of caffeine anhydrate in various designed intact tables by near-infrared spectroscopy: A comparison between reflectance and transmittance technique.* Journal of Pharmaceutical and Biomedical Analysis, 47(4-5):819-827, 2008.
26. Schneider, R.C. and Kovar, K.-A. *Analysis of ecstasy tablets: comparison of reflectance and transmittance near infrared spectroscopy.* Forensic Science International, 134(2-3):187-195, 2003.
27. De Braekeleer, K., De Maesschalck, R., Hailey, P.A., Sharp, D.C.A., and Massart, D.L. *On-line application of the orthogonal projection approach (OPA) and the soft independent modelling of class analogy approach (SIMCA) for the detection of the end point of a polymorph conversion reaction by near infrared spectroscopy (NIR).* Chemometrics and Intelligent Laboratory Systems, 46(2):103-116, 1999.
28. Shannon C. Harris, D.S.W. *Quantitative real-time monitoring of dryer effluent using fiber optic near-infrared spectroscopy.* Journal of Pharmaceutical Sciences, 89(9):1180-1186, 2000.
29. Miwa, A., Yajima, T., and Itai, S. *Prediction of suitable amount of water addition for wet granulation.* International Journal of Pharmaceutics, 195(1-2):81-92, 2000.

30. Otsuka, M. *Chemoinformetrical evaluation of granule and tablet properties of pharmaceutical preparations by near-infrared spectroscopy*. Chemometrics and Intelligent Laboratory Systems, 82(1-2):109-114, 2006.
31. Ian R. Lewis, H.G.M.E., *Handbook of Raman Spectroscopy* CRC, 2001.
32. Krishnan, C.V.R.a.K.S. *A new type of secondary radiation.* Nature, 121(3048):501-502, 1928.
33. C. Sandalinas, S.R.-M., A. López-Gil, J. Miralles *Experimental confirmation by Raman spectroscopy of a PbSnSb triple oxide yellow pigment in sixteenth-century Italian pottery.* Journal of Raman Spectroscopy, 37(10):1146-1153, 2006.
34. N. Welter, U.S., W. Kiefer *Characterisation of inorganic pigments in ancient glass beads by means of Raman microspectroscopy, microprobe analysis and X-ray diffractometry.* Journal of Raman Spectroscopy, 38(1):113-121, 2006.
35. Bernhard Schrader, H.S., Malgorzata Baranska, George N. Andreev, Caroline Lehner, Juergen Sawatzki *Non-destructive Raman analyses – polyacetylenes in plants*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 61(7):1395-1401, 2005.
36. Li-Chan, E.C.Y. *The applications of Raman spectroscopy in food science.* Trends in Food Science & Technology, 7(11):361-370, 1996.
37. Widjaja, E. and Seah, R.K.H. *Application of Raman microscopy and band-target entropy minimization to identify minor components in model pharmaceutical tablets.* Journal of Pharmaceutical and Biomedical Analysis, 46(2):274-281, 2008.
38. Šašić, S. *An In-Depth Analysis of Raman and Near-Infrared Chemical Images of Common Pharmaceutical Tablets.* Applied Spectroscopy, 61(3):239-250, 2007.
39. Sasic, S. *Raman mapping of low-content API pharmaceutical formulations. I. Mapping of alprazolam in Alprazolam/Xanax tablets.* Pharmaceutical Research, 24(1):58-65, 2007.
40. Šašić, S. and Clark, D.A. *Defining a Strategy for Chemical Imaging of Industrial Pharmaceutical Samples on Raman Line-Mapping and Global Illumination Instruments.* Applied Spectroscopy, 60(5):494-502, 2006.
41. Hausman, D.S., Cambron, R.T., and Sakr, A. *Application of on-line Raman spectroscopy for characterizing relationships between drug hydration state and tablet physical stability.* International Journal of Pharmaceutics, 299(1-2):19-33, 2005.
42. Clark, D. and Sasic, S. *Chemical images: Technical approaches and issues.* Cytometry Part A, 69A(8):815-824, 2006.
43. Smith, L.I., *A tutorial on Principal Components Analysis*. Available from: http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf 2002.
44. D.P. Berrar, W.D., M. Granzow, eds. *Singular value decomposition and principal component analysis*. 2003, Kluwer: Norwell, MA. p. 91-109.
45. Johansson, E., Kettaneh-Wold, N., Wold, S., and Ericksson, L., *Multi- and Megavariate Data Analysis.* Umetrics Academy, 2001.
46. Clarke, F. *Extracting process-related information from pharmaceutical dosage forms using near infrared microscopy.* Vibrational Spectroscopy, 34(1):25-35, 2004.

47. Roggo, Y., Edmond, A., Chalus, P., and Ulmschneider, M. *Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms.* Analytica Chimica Acta, 535(1-2):79-87, 2005.
48. Bacci, M., Chiari, R., Porcinai, S., and Radicati, B. *Principal component analysis of near-infrared spectra of alteration products in calcareous samples: An application to works of art.* Chemometrics and Intelligent Laboratory Systems, 39(1):115-121, 1997.
49. Bossart, R., Keller, H., Kellerhals, H., and Oelichmann, J. *Principal components analysis as a tool for identity control using near-infrared spectroscopy*. Journal of Molecular Structure, 661-662:319-323, 2003.
50. Jain, A.K., Duin, R.P.W., and Mao, J. *Statistical Pattern Recognition: A Review.* IEEE Transactions on Pattern Analysis and Machines Intelligence, 22(1):4-37, 2000.
51. Wold, S. and Sjostrom, M. *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy.* Chemometrics: Theory and Application:243-282, 1977.
52. Smidt, E., Meissl, K., Schwanninger, M., and Lechner, P. *Classification of waste materials using Fourier transform infrared spectroscopy and soft independent modeling of class analogy*. Waste Management, 28(10):1699-1710, 2008.
53. Candolfi, A., De Maesschalck, R., Massart, D.L., Hailey, P.A., and Harrington, A.C.E. *Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA*. Journal of Pharmaceutical and Biomedical Analysis, 19(6):923-935, 1999.
54. Mansfield, J.R., Sowa, M.G., Scarth, G.B., Somorjai, R.L., and Mantsch, H.H. *Analysis of Spectroscopic Imaging Data by Fuzzy C-Means Clustering*. Anal. Chem. , 69(16):3370-3374, 1997.
55. Yu-Ping Wang, Y.W., and Paulette Spencer *Fuzzy Clustering of Raman Spectral Imaging Data with a Wavelet-Based Noise-Reduction Approach.* Applied Spectroscopy, 60(7):826-832, 2006.
56. Tauler, R. *Multivariate curve resolution applied to second order data.* Chemometrics and Intelligent Laboratory Systems, 30(1):133-146, 1995.
57. Massart, D.L., Vandeginste, B., Buydens, L., Jong, S.d., Lewi, P., and Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics.* Elsevier Science, 1997.
58. Karjalainen, E.J. *The spectrum reconstruction problem: Use of alternating regression for unexpected spectral components in two-dimensional spectroscopies.* Chemometrics and Intelligent Laboratory Systems, 7(1-2):31-38, 1989.
59. Amigo, J.M., Cruz, J., Bautista, M., Maspoch, S., Coello, J., and Blanco, M. *Study of pharmaceutical samples by NIR chemical-image and multivariate analysis.* TrAC Trends in Analytical Chemistry, 27(8):696-713, 2008.
60. Azzouz, T. and Tauler, R. *Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples*. Talanta, 74(5):1201-1210, 2008.
61. Martens, H. and Naes, T., *Multivariate calibration.* Wiley, 1989.
62. Wold, S., Sjöström, M., and Eriksson, L. *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 58(2):109-130, 2001.

63. Burger, J. and Geladi, P. *Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples*. Analyst, 131:1152-1160, 2006.
64. van den Broek, W.H.A.M., Derks, E.P.P.A., van de Ven, E.W., Wienke, D., Geladi, P., and Buydens, L.M.C. *Plastic identification by remote sensing spectroscopic NIR imaging using kernel partial least squares (KPLS).* Chemometrics and Intelligent Laboratory Systems, 35(2):187-197, 1996.
65. Ikeda, S. and Toyama, K. *Independent component analysis for noisy data - MEG data analysis*. Neural Networks, 13(10):1063-1074, 2000.
66. Hyvarinen, A. and Oja, E. *Independent component analysis: algorithms and applications.* Neural Networks, 13(4-5):411-430, 2000.
67. Martín, J.C.G., Spietz, P., Orphal, J., and Burrows, J.P. *Principal and independent components analysis of overlapping spectra in the context of multichannel time-resolved absorption spectroscopy.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 60(11):2673-2693, 2004.
68. Yamamoto, H., Hada, K., Yamaji, H., Katsuda, T., Ohno, H., and Fukuda, H. *Application of regularized alternating least squares and independent component analysis to HPLC-DAD data of Haematococcus pluvialis metabolites.* Biochemical Engineering Journal, 32(3):149-156, 2006.
69. Brandstein, M.S. and Silverman, H.F. *A robust method for speech signal time-delay estimation in reverberant rooms*. In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 1997.
70. Kardamakis, A.A., Mouchtaris, A., and Pasadakis, N. *Linear predictive spectral coding and independent component analysis in identifying gasoline constituents using infrared spectroscopy*. Chemometrics and Intelligent Laboratory Systems, 89(1):51-58, 2007.
71. Pasadakis, N. and Kardamakis, A.A. *Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis.* Analytica Chimica Acta, 578(2):250-255 2006.
72. Wang, G., Ding, Q., Sun, Y.a., He, L., and Sun, X. *Estimation of source infrared spectra profiles of acetylspiramycin active components from troches using kernel independent component analysis*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 70(3):571-576, 2008.
73. Polder, G., van der Heijden, G.W.A.M., and Young, I.T. *Tomato sorting using independent component analysis on spectral images*. Real-Time Imaging, 9(4):253-259, 2003.
74. Steven E. J. Bell, L.J.B., D. Thorburn Burns, Andrew C. Dennis and S. James Speers *Tracking the distribution of ecstasy tablets by Raman composition profiling: A large scale feasibility study*. Analyst, 128:1331-1335, 2003.
75. Slobodan Šašić, D.A.C., John C. Mitchell and Martin J. Snowden *A comparison of Raman chemical images produced by univariate and multivariate data processing—a simulation with an example from pharmaceutical practice.* Analyst, 129:1001-1007, 2004.
76. Slobodan Šašić, D.A.C., John C. Mitchell and Martin J. Snowden *Raman line mapping as a fast method for analyzing pharmaceutical bead formulations.* Analyst, 130:1530-1536, 2005.

77. SASIC, S. *An In-Depth Analysis of Raman and Near-Infrared Chemical Images of Common Pharmaceutical Tablets*. Applied Spectroscopy, 61(3):239-250, 2007.
78. Zhang, L., Henson, M.J., and Sekulic, S.S. *Multivariate data analysis for Raman imaging of a model pharmaceutical tablet.* Analytica Chimica Acta, 545(2):262-278, 2005.
79. Šašić, S. and Whitlock, M. *Raman Mapping of Low-Content Active-Ingredient Pharmaceutical Formulations. Part II: Statistically Optimized Sampling for Detection of Less Than 1% of an Active Pharmaceutical Ingredient.* Pharmaceutical Research, 62(8):833-935, 2008.
80. Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., and Jent, N. *A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies*. Journal of Pharmaceutical and Biomedical Analysis, 44(3):683-700, 2007.
81. Korol, A.B., *Microarray cluster analysis and applications*. Available from: http://www.science.co.il/enuka/Essays/Microarray-Review.pdf, 2003.
82. Shariati-Rad, M. and Hasani, M. *Application of multivariate curve resolution-alternating least squares (MCR-ALS) for secondary structure resolving of proteins.* Biochimie, 91(7):850-856, 2009.
83. De Lathauwer, L., De Moor, B., and Vandewalle, J. *An introduction to independent component analysis.* Journal of Chemometrics, 14(3):123-149, 2000.
84. Shashilov, V.A., Xu, M., Ermolenkov, V.V., and Lednev, I.K. *Latent variable analysis of Raman spectra for structural characterization of proteins.* Journal of Quantitative Spectroscopy & Radiative Transfer, 102(1):46-61, 2006.
85. Vrabie, V., Gobinet, C., Piot, O., Tfayli, A., Bernard, P., Huez, R., and Manfait, M. *Independent component analysis of Raman spectra: Application on paraffin-embedded skin biopsies*. Biomedical Signal Processing and Control, 2(1):40-50, 2007.
86. Rippin, D.W.T. *Batch process systems engineering : a retrospective and prospective review.* Computers & Chemical Engineering, 17(1):1-13, 1993.
87. Faggian, A., Facco, P., Doplicher, F., Bezzo, F., and Barolo, M. *Multivariate statistical real-time monitoring of an industrial fed-batch process for the production of specialty chemicals.* Chemical Engineering Research and Design, 87(3):325-334, 2009.
88. Lann, M.V.L., Cabassud, M., and Casamatta, G. *Modeling, optimization and control of batch chemical reactors in fine chemical production.* Annual Reviews in Control, 23:25-34, 1999.
89. Alonso, A.A., Perez-Martin, R.I., Shukla, N.V., and Deshpande, P.B. *On-line quality control of non-linear batch systems: Application to the thermal processing of canned foods.* Journal of Food Engineering, 19(3):275-289, 1993.
90. Fujiwara, M., Nagy, Z.K., Chew, J.W., and Braatz, R.D. *First-principles and direct design approaches for the control of pharmaceutical crystallization.* Journal of Process Control, 15(5):493-504, 2005.
91. Preuβ, K., Lann, M.-V.L., Cabassud, M., and Anne-Archard, G. *Implementation procedure of an advanced supervisory and control strategy in the pharmaceutical industry*. Control Engineering Practice, 11(12):1449-1458, 2003.

92. Jr, R.S. and Almeida, P.I.F. *Design of a fuzzy system for the control of a biochemical reactor in fed-batch culture.* Process Biochemistry, 37(5):461-469, 2001.
93. Kreft, T. and Reed, W.F. *Predictive control and verification of conversion kinetics and polymer molecular weight in semi-batch free radical homopolymer reactions*. European Polymer Journal, 45(8):2288-2303, 2009.
94. Engers, B. and Bauer, H.U. *Cost-effective PVD coatings in batch systems*. Surface and Coatings Technology, 116-119:705-710, 1999.
95. Seborg, D.E., Edgar, T.F., and Mellichamp, D.A., *Process dynamics and control.* Wiley, 2004.
96. Sharratt, P.N., *Handbook of Batch Process Design*. Springer, 1997.
97. Lennox, B., Montague, G.A., Hiden, H.G., Kornfeld, G., and Goulding, P.R. *Process monitoring of an industrial fed-batch fermentation.* Biotechnology and Bioengineering, 74(2):125-135, 2001.
98. Kravaris, C., Wright, R.A., and Carrier, J.F. *Nonlinear controllers for trajectory tracking in batch processes*. Computers & Chemical Engineering, 13(1-2):73-82, 1989.
99. Clarke-Pringle, T. and MacGregor, J.F. *Nonlinear adaptive temperature control of multi-product, semi-batch polymerization reactors.* Computers & Chemical Engineering, 21(12):1395-1409, 1997.
100. Wang, Z.L., Pla, F., and Corriou, J.P. *Nonlinear adaptive control of batch styrene polymerization.* Chemical Engineering Science, 50(13):2081-2091, 1995.
101. Aziz, N., Hussain, M.A., and Mujtaba, I.M. *Performance of different types of controllers in tracking optimal temperature profiles in batch reactors.* Computers & Chemical Engineering, 24(2-7):1069-1075, 2000.
102. Crowley, T.J. and Choi, K.Y. *Experimental studies on optimal molecular weight distribution control in a batch-free radical polymerization process*. Chemical Engineering Science, 53(15):2769-2790, 1998.
103. Ruppen, D., Bonvin, D., and Rippin, D.W.T. *Implementation of adaptive optimal operation for a semi-batch reaction system.* Computers & Chemical Engineering, 22(1-2):185-199, 1998.
104. Lin, H., Marjanovic, O., Lennox, B., and Shamekh, A. *Application of Near-infrared Spectroscopy in Batch Process Control*. In International Symposium on Advanced Control of Chemical Processes,, Istanbul, Turkey, 2009.
105. Yabuki, Y., Nagasawa, T., and MacGregor, J.F. *Industrial experiences with product quality control in semi-batch processes.* Computers & Chemical Engineering, 26(2):205-212, 2002.
106. Pan, Y. and Lee, J.H. *Recursive data-based prediction and control of product quality for a PMMA batch process*. Chemical Engineering Science, 58(14):3215-3221, 2003.
107. MacGregor, J.F. and Kourti, T. *Statistical process control of multivariate processes*. Control Engineering Practice, 3(3):403-414, 1995.
108. Kresta, J.V., Macgregor, J.F., and Marlin, T.E. *Multivariate statistical monitoring of process operating performance.* The Canadian Journal of Chemical Engineering, 69(1):35-47, 1991.

109. Hyvärinen, A., Karhunen, J., and Oja, E., *Independent Component Analysis* New York ; Chichester : Wiley, 2001., 2001.
110. Choi, S., Cichocki, A., Park, H.-M., and Lee, S.-Y. *Blind Source Separation and Independent Component Analysis: A review.* Neural Information Processing - Letters and Reviews, 6(1):1-57, 2005.
111. Comon, P. *Independent Component Analysis, a New Concept.* Signal Processing, 36(3):287-314, 1994.
112. Tijms, H., *Understanding Probability: Chance Rules in Everyday Life*, 2004.
113. Beebe, K.R., Pell, R.J., and Seasholtz, M.B., *Chemometrics : a practical guide.* New York ; Chichester : Wiley, 1998, 1998.
114. Bakeev, K.A., *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries* Blackwell Publishing Limited, 2005.
115. IV, F.W.K., Lee, E., Kidder, L.H., and Lewis, E.N. *Near infrared spectroscopy-the practical chemical imaging solution.* Spectroscopy Europe, 3(14):12-19, 2002.
116. Lewis, E.N., Schoppelrei, J., and Lee, E. *Near-infrared Chemical Imaging and the PAT Initiative.* Spectroscopy, 19(4):26-36, 2004.
117. Sanchez, F.C., Toft, J., Kvalheim, O.M., and Massart, D.L. *Eigenstructure Tracking Analysis for Assessment of Peak Purity in High-Performance Liquid-Chromatography with Diode-Array Detection.* Analytica Chimica Acta, 314(3):131-139, 1995.
118. Vanslyke, S.J. and Wentzell, P.D. *Real-Time Principal Component Analysis Using Parallel Kalman Filter Networks for Peak Purity Analysis.* Analytical Chemistry, 63(21):2512-2519, 1991.
119. Smilde, A., Bro, R., and Geladi, P., *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley & Sons, 2004.
120. Wold, S. *Multi-way principal components-and PLS-analysis.* Journal of Chemometrics, 1(1):41-56, 1987.
121. Geladi, P. and Grahn, H., *Multivariate Image Analysis*. Wiley 1997.
122. Andrew, J.J. and Hancewicz, T.M. *Rapid analysis of Raman image data using two-way multivariate curve resolution.* Applied Spectroscopy, 52(6):797-807, 1998.
123. Savitzky, A. and Golay, M.J.E. *Smoothing + Differentiation of Data by Simplified Least Squares Procedures.* Analytical Chemistry, 36(8):1627-&, 1964.
124. Press, W.H., *Numerical recipes in C : the art of scientific computing* Cambridge : Cambridge University Press, 1992.
125. Orfanidis, S.J., *Introduction to signal processing* Englewood Cliffs, N.J. : Prentice-Hall International, 1996.
126. Vrabie, V., Gobinet, C., Piot, O., Tfayli, A., Bernard, P., Huez, R., and Manfait, M. *Independent component analysis of Raman spectra: Application on paraffin-embedded skin biopsies.* Biomedical Signal Processing and Control, 2:40-50, 2007.
127. Maciejowski, J., *Predictive Control With Constraints.* Addison Wesley Longman, 2002.

128. Zhu, G.Y., Henson, M.A., and Megan, L. *Dynamic modeling and linear model predictive control of gas pipeline networks*. Journal of Process Control, 11(2):129-148, 2001.
129. Kiran, A.U.M. and Jana, A.K. *Control of continuous fed-batch fermentation process using neural network based model predictive controller*. Bioprocess and Biosystems Engineering, 32(6):801-808, 2009.
130. Bordons, C. and Nunez-Reyes, A. *Model based predictive control of an olive oil mill*. Journal of Food Engineering, 84(1):1-11, 2008.
131. Irizuki, Y., Suzuki, K., Murata, S., Uno, H., and Tani, T. *Design of combined control system incorporating knowledge based system and model predictive multivariable control system and its application to refining plant*. In Second International Conference on Knowledge-Based Intelligent Electronic Systems, Adelaide, Australia, 1998.
132. Prett, D.M. and Garcia, C.E., *Fundamental Process Control*. Butterworth-Heinemann, USA, 1988.
133. Bitmead, R., Gevers, M., and Wertz, V., *Adaptive Optimal Control: The Thinking Man's Gpc* Prentice Hall, USA, 1990.
134. Soeterboek, R., *Predictive Control: A Unified Approach* Prentice Hall Intemational, Cambridge, UK, 1992.
135. Camacho, E.F. and Bordons, C., *Model Predictive Control in the Process Industry* Sprhger-Verlag, Berlin, Germany, 1995.
136. Camacho, E.F. and Bordons, C., *Model Predictive Control*. Springer-Verlag, 1999.
137. Wonhui Cho, T.F.E., Jietae Lee *Iterative learning dual-mode control of exothermic batch reactors*. Control Engineering Practice, 16:1244-1249, 2008.
138. Cott, B.J. and Macchietto, S. *Temperature Control of Exothermic Batch Reactors Using Generic Model Control*. Industrial & Engineering Chemistry Research, 28(8):1177-1184, 1989.
139. Shinskey, F.G., *Process Control Systems*. USA: McGraw-Hill, 1996.
140. *Data sets for multi-variate data analysis*. Available from: http://www.models.life.ku.dk/Marzipan, 2004.
141. Ljung, L., *System identification : theory for the user*. Upper Saddle River, N.J. : Prentice Hall, 1999.
142. Huo, H., Zhu, X., and Tu, H. *Iterative learning control of SOFC based on ARX identification model*. Journal of Zhejiang University Science - A 8(12):1921-1927, 2007.
143. Huusoma, J.K., Poulsenb, N.K., Jørgensena, S.B., and Jørgensen, J.B. *ARX-Model based Model Predictive Control with Offset-Free Tracking* Computer Aided Chemical Engineering, 28:601-606, 2010.
144. Shamekh, A., Lennox, B., Sandoz, D., and Marjanovic, O. *PLS and Its Application within Model Predictive Controllers*. In Proceedings of the 17th IFAC World Congress, COEX, Korea, South, 2008.
145. Fukata, K., Washio, T., and Motoda;, H. *A Method to Search ARX Model Orders and Its Application to Sales Dynamics Analysis*. In Proceedings of the Sixth IEEE International Conference on Data Mining, Washington, DC, USA, 2006.
146. Wold, H., *Path models with latent variables: The NIPALS approach*. Academic Press, 1975.

147. de Jong, S. *SIMPLS: An alternative approach to partial least squares regression.* Chemometrics and Intelligent Laboratory Systems, 18(3):251-263, 1993.
148. Arteaga, F. and Ferrer, A. *Dealing with missing data in MSPC: several methods, different interpretations, some examples.* Journal of Chemometrics, 16(8-10):408-418, 2002.
149. Nelson, P.R.C., Taylor, P.A., and MacGregor, J.F. *Missing data methods in PCA and PLS: Score calculations with incomplete observations.* Chemometrics and Intelligent Laboratory Systems, 35(1):45-65, 1996.
150. Lin, H., Marjanovic, O., Lennox, B. *Application of End-Point Control and Trajectory Tracking to Batch Processes*. In 11th Computer Applications in Biotechnology, Leuven, Belgium, 2010.
151. Flores-Cerrillo, J. and MacGregor, J.F. *Iterative learning control for final batch product quality using partial least squares models*. Industrial & Engineering Chemistry Research, 44(24):9146-9155, 2005.