A Framework for Enhancing Process Understanding using Multivariate Tools on Commercial Batch Process Data

A thesis submitted by

Matthew Molloy

for the award of Engineering Doctorate in Biopharmaceutical Process Development

Biopharmaceutical and Bioprocessing Technology Centre

School of Chemical Engineering & Advanced Materials

Newcastle University

May 2017

Abstract

A lot of effort is made by pharmaceutical companies on the research and development of new pharmaceutical products and processes using the latest in quality by design tools, and process analytical technologies. Older pharmaceutical processes that were developed without the use of these tools are, however, somewhat neglected. Significant quantities of process data are routinely collected and stored but the information contained within this data is not extracted.

Extensive literature on multivariate statistical process monitoring and control exists for exploring both batch and continuous process data. However, these methodologies rely on data from processes that are relatively well understood or controlled. Many industrial processes show batch to batch variability, which may be tolerated as it is not detrimental to the quality of the product, and the impact of this variability is not fully understood.

The thesis presents a framework for exploring historical batch process data, to extract insights on where process control can be improved. The challenges presented with commercial process data are discussed. Multivariate tools such as multi-way principal component analysis are used to investigate variability in process data. The framework presented discusses the pre-processing steps necessary with batch process data, followed by outlier detection, and finally multivariate modelling of the data to identify where the process could benefit from improved understanding and control.

This framework is demonstrated through the application to commercial process data from the active pharmaceutical drug substance manufacturing process of spironolactone at Piramal Healthcare, Morpeth, UK. In this case study, the process exhibits variability in drying times which traditional univariate data analysis has not been able to solve. The results demonstrated some of the challenges the use of the available data from commercial processes. Although the results from the multivariate data analysis did not show a significant statistical difference between the batches with long and short drying times, small differences were observed between these two groups. Further analysis of the crystallization process using infrared spectroscopic techniques which identified a potential root cause to the extended drying time.

Dedication

For my wife and my family.

Declaration

This thesis is submitted to the degree of Doctor of Engineering at Newcastle University. The research in this thesis was performed in the Biopharmaceutical and Bioprocess Technology Centre in the School of Chemical Engineering, under the supervision of Professor Elaine Martin and Prof. Jarka Glassey, and is my own work unless stated otherwise in the text. I certify that none of the material in this thesis has been submitted previously by me for a degree or any other qualification at this or any other university.

Acknowledgements

This EngD project was supported by the Engineering and Physical Sciences Research Council (EPSRC) and Piramal Healthcare, Morpeth.

I would like to thank all of those from Newcastle University involved with the supervision of this project including Dr. Ming Tham, Prof. Gary Montague, and Prof. Elaine Martin for their support and guidance throughout the various stages of this project. Thanks goes especially to Prof. Jarka Glassey for her help and support through the final stages of this process.

I would also like to thank my primary supervisor from Piramal Healthcare, Mike Devenport for the opportunity to work on this project and the guidance and support received throughout. I would also like to express my gratitude to all those at Piramal Healthcare who have guided and supported me during this process especially Tim Pearson and James Howells.

Thanks are also due to Bruker, especially Ali Gahkani, for the advice and loan of the IR spectroscopy instruments.

Many people have touched my life during my research period. I would like to offer my heartfelt thanks to the friends I have made, especially my fellow research students with whom I have shared many memorable experiences.

I would like to thank my friends and family for their continued support throughout the course of this EngD, especially my parents, for their encouragement. Finally, and most importantly I would like give thanks to my wife, Agnieszka, for her patience, support and encouragement that was much needed over the years.

Contents

List of Figures

List of Tables

Nomenclature

Θ	Scalar value from MI
σ_e	Variance of reconstructed process data error
σ_y	Variance of original process data (before compression)
$\sigma_{\hat{y}}$	Variance of reconstructed process data
$\bar{x}$	Sample mean
l	Data time lag
r_1	Sample autocorrelation
S	Diagonal matrix of the eigenvalues
T^2	Hotelling's T^2 statistic
x_i	Sample from process variable
$\mathbf{E}$	Residuals
$\mathbf{P}$	Matrix of principal component loadings
$\mathbf{P}$	Path scores for dynamic time warping
$\mathbf{p}$	Vector of principal component loadings
$\mathbf{T}$	Matrix of principal component scores
$\mathbf{t}$	Vector of principal component scores
$\mathbf{W}$	Warping path matrix
$\mathbf{X}_1$	Replication of trajectory 1 into a square matrix
$\mathbf{X}_2$	Replication of trajectory 2 into a square matrix
$\mathbf{X}$	Input variable matrix to PLS

X Matrix of input variables to PCA

Y Response variable matrix for PLS

AEEE Aldona ethyl enol ether

API Active pharmaceutical ingredient (drug substance)

ARMAX Auto-regressive moving-average with exogenous input model

ATR Attenuated total reflectance

ATR-FTIR Attenuated total reflectance Fourier transform mid-infrared spectroscopy

CMO Contract manufacturing organisation

COW Correlation optimised warping

D Matrix of absolute differences

DPCA Dynamic Principal Component Analysis

DPCA Dynamic principal component analysis

DTW Dynamic time warping

EMA European medicines agency

EWMA Exponentially Weighted Moving Average

FBRM Focused beam reflectance measurement

FDA Food and drug administration regulatory body

FIR Finite Impulse Response

FTIR Fourier transform infrared spectroscopy

I Batch mode

ICH Q10 International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use guideline on pharmaceutical quality systems

IIR Infinite Impulse Response

IR Infrared spectroscopy

J Variable mode

K Time mode

MIR Mid-infrared spectroscopy

MPCA Multi-way principal component analysis

MSPC Multivariate statistical process control

NIPALS Non-linear iterative partial least squares

NIR Near infrared spectroscopy

PAGA Peak alignment by genetic algorithm

PCA Principal component analysis

PTW Parametric time warping

Spironolactone 17-hydroxy-7α-acetylthio-3-oxo-17α-pregn-4-ene-21-carboxylic acid γ-lactone acetate

SSE Sum of squared errors

STW Semi-parametric time warping

SVD Singular value decomposition

B Between imputation variance component

e_i Reconstruction error of sample i

h Sampling interval

M Number of parameter estimates

m Number of archived values in data historian

N Number of values in data before compression

T Total imputation variance

W Within imputation variance component

$\boldsymbol{X}$ Matrix of Data

$\boldsymbol{y}$ Original process data (before compression)

y_i Value of original data at sample i

$\hat{\boldsymbol{y}}$ Reconstructed process data

$\hat{y}_i$ Value of reconstructed data at sample i

ALSIA Alternating Least Squares Iterative Algorithm

CF Compression Factor

DPCA Dynamic Principal Component Analysis

MAR Missing at random

MCAR Missing completely at random

MNAR Missing not at random

NI Non-ignorable

PCADA Principal Component Analysis Data Augmentation

PCAIA Principal Component Analysis Iterative Algorithm

PCA Principal Component Analysis

PDM Percentage Difference of between the Mean

PLS Projection to Latent Structures

RVC Ratio between reconstructed and original data standard deviation

SVD Singular Value Decomposition

Chapter 1. Introduction

1.1 Motivation and objectives

Pharmaceutical manufacturers spend a lot of time and money in the research and development phase of both drug substances and drug products using the latest in Process Analytical Technology and applying strategies such as Quality by Design, exploring design spaces and process robustness. Older pharmaceutical products already out at the manufacturing sites, or contract manufacturing organisations (CMOs) are, however, somewhat neglected. Older manufacturing processes exist where a lot of data is collected from the batches being manufactured which may contain useful information about a process.

The international conference on harmonisation of technical requirements for registration of pharmaceuticals for human use guideline on pharmaceutical quality systems (ICH Q10 (ICH, 2008)) suggests that pharmaceutical products in the commercial manufacturing stage of the pharmaceutical product life cycle should include activities to facilitate continual improvement. This is not only meant in regards to the pharmaceutical product quality but also to the process performance and controls. Additionally, ICH Q10 also recommends that improvement opportunities are identified and evaluated and the body of knowledge on the product and process are continually expanded.

Traditional approaches such as interrogating the process data in a univariate manner are not always suitable due to complex interactions and the dynamic nature of the processes involved. Other methods of changing one factor at a time, or even using factorial experimental designs, are often difficult to implement due to the large number of factors that could be explored. An alternative approach for increasing the understanding of processes is through the use of multivariate statistical projection techniques such as Principal Component Analysis (PCA). The use of multivariate methods can help in identifying some of the factors that should be further investigated reducing both the cost and the time taken for process improvement work carried out.

Furthermore, as recommended in ICH Q10, changes to commercial pharmaceutical manufacturing process should be controlled through a change control procedure.The use of a change control procedure is a regulatory requirement for any changes made to a registered pharmaceutical process where the change is made to a parameter, setting, or other information detailed in the regulatory filing. The change control procedure should document the change in detail and the risk assessment performed for the change carried out by suitably knowledgeable persons in areas including manufacturing, quality, and regulatory affairs. Following the change, the change control procedure should also detail a review of the change to confirm that the objectives set out for the change have been met, and there were no deleterious impacts on product quality as a result of the change. The ability to use the existing data of batches already manufactured to inform which changes to make, and therefore only raising change controls for large scale development trials on changes that are likely to significantly improve the process, will have a significant benefit to commercial pharmaceutical manufacturing organisations.

Statistical process control is a useful technology to identify unusual events in process data and can be applied both on-line as a batch evolves, or off-line after a batch has been manufactured (Grigg, 1998; Škulj et al., 2013; AlGhazzawi and Lennox, 2008; Masding and Lennox, 2010). Traditional statistical process control methodologies are typically univariate in nature and they only consider one process measurement at a time, ignoring the interactions between the process variables which may also contain important information. Multivariate statistical process control methodologies consider all of the process variables simultaneously, and can handle collinearity in the data, and are therefore more suited to the analysis of process data (Kourti, 2005; Kona et al., 2013; Bersimis et al., 2007; Martin et al., 1996b).

One such method is Principal Component Analysis (PCA), which through use of the covariance structure within a dataset with correlated variables, is able to obtain a smaller number of uncorrelated (orthogonal) principal components that describe the variability within the original data (Wold et al., 1987; Jolliffe, 2002). Traditional PCA is not ideally suited to the analysis of batch process data as the batch element is also a variable in the process and therefore batch process datasets are typically three-dimensional. The multi-way approach (Nomikos and MacGregor, 1994) can be applied to handle this additional dimension in the data through tools such as multi-way PCA (MPCA) .

Batch processes are often dynamic in their behaviours with the resulting data often being autocorrelated. Traditional PCA cannot deal with these dynamic behaviours, however, by

combining ideas from other methodologies such as auto-regressive moving-average with exogenous input (ARMAX) models the dynamic extension of PCA can be applied (DPCA) (Chen and Liu, 2002).

Monitoring of batch trajectories through the use of multivariate projection techniques has been achieved and a range of industrial applications demonstrating their applicability and effectiveness have been extensively reported in the literature (Gabrielsson et al., 2002; Simoglou et al., 2005; García-Muñoz et al., 2003; De Beer et al., 2009; Burggraeve et al., 2011; Sarraguça et al., 2010; Martin et al., 1996a). A number of these applications and tools, however, have been developed and demonstrated using simulated process data. Commercial manufacturing process data presents additional challenges with the availability and quality of the data that must be processed before multivariate tools can be applied.

Some of the failures that these multivariate tools can identify, can also be easily identified at the time by existing process control systems, such as equipment failures and processing errors. These tools are, however, useful for providing an early warning of batch failure where a corrective action can be applied, of for identifying more complex failures not picked up by simpler control and monitoring systems.

The primary focus of this thesis is not the identification of batch failures using the traditional multivariate statistical process control tools, but the use of these tools to identify subtle differences in historic process data that can inform on how to improve the control and operation of a process. Additionally, genuine data from commercial processes presents some challenges, be this through uneven batch data alignment, number of batches, and the quality of the data available (Wan et al., 2014; Laurí and Lennox, 2014; Camacho et al., 2015; Kassidas et al., 1998; García-Muñoz et al., 2011).

These challenges are demonstrated through the application of multivariate statistical tools to the commercial batch manufacturing process of spironolactone drug substance at Piramal Healthcare, Morpeth, UK. The spironolactone manufacturing process at Morpeth is important for the business however it currently exhibits significant variability in process yield and cycle time. Any improvements to understand or control the variability in yield and/or cycle time will have significant benefits for the company allowing for better control, planning, and use of their assets and workforce.

1.1.1 Objectives

The objectives of this thesis are as follows:

- Present a framework in which to apply multivariate tools and appropriate pre-processing techniques to industrial batch process data to extract understanding from the existing process data
- Test the framework on the Spironolactone Active Pharmaceutical Ingredient process data to extract information on the variable cycle times of the drying process.

1.2 Contributions of the thesis

The thesis presents a framework for exploring historical batch process data, to extract insights on where process control can be improved. More specifically the key contributions include:

- A framework for the application of multivariate statistical tools, specifically dynamic multi-way principal component analysis, to commercial manufacturing process data in the area of process control. The proposed methodology extends the existing work through presentation of a framework including the pre-processing of commercial process data, and more specifically details on how filtering and batch selection play an important part of the pre-treatment process.
- Details on some of the challenges associated with applying multivariate statistical process control tools to commercial process data. A good knowledge of the process and its data is required in order to, amongst other things, effectively align process events, and differentiate between unusual batch behaviours that are of interest or caused by known events that are not of as much importance. This enables the variation associated with a process to be reduced down to variation that describes changes in an attribute of interest.
- Demonstration of the proposed framework and challenges on commercial process data of the spironolactone drug substance manufacturing process at Piramal Healthcare, Morpeth, UK, to identify potential causes in variable drying times.
- Evaluation of infrared spectroscopic tools for obtaining understanding of a complex crystallization system at lab scale. The current challenges with instrument design are discussed. The benefit of using attenuated total reflectance Fourier transform

mid-infrared (ATR-FTIR) spectroscopy coupled with principal component analysis to the monitoring of the complex spironolactone crystallization is also presented.

- Identification of possible root causes of the variability in drying times of spironolactone drug substance through the proposed framework followed by spectroscopic analysis of the simplified crystallization process.

1.3 Layout of the thesis

The layout of the remainder of the thesis will be in the following structure.

Chapter 1 provided an introduction to the thesis including the motivations and objectives for the research into the use of multivariate statistical tools for identification of where process control can be improved. The main contributions of the thesis were also identified.

Chapter 2 presents the background to the spironolactone process as a case study on which the framework is presented. The background includes details of the process chemistry including the impurities and challenges with polymorphism. The manufacturing process is also described in addition to the control of the process and the data that is collected from the process.

Chapter 3 introduces the pre-processing techniques and considerations used within the framework including compressed data, missing data, alignment of batch data, centring and scaling, and filtering. Principal component analysis is introduced in chapter 4 along with some extensions to PCA including multi-way (MPCA), dynamic (DPCA) and some examples of industrial case studies of PCA available in the literature.

Chapter 5 describes the application of the framework to the Spironolactone case study beginning with the objectives and an overview of the modelling approach. Next the model development on the dryer data is presented, followed by the model development on the reactor data.

Chapter 6 describes the experimental work carried out using various infrared spectroscopic techniques on the crystallization of spironolactone. Focused beam reflectance measurement (FBRM), a tool for monitoring particle size distributions on-line, is discussed as are the transflectance near-infrared (NIR) and attenuated total reflection Fourier transform mid-infrared (ATR-FTIR) spectroscopic techniques. The analysis of the ATR-FTIR data using principal component analysis as a dimensionality reduction method are presented for this case study.

Finally, chapter 7 summarises the key findings in the thesis and identifies some areas for future work.

Chapter 2. Spironolactone background

Spironolactone is an important synthetic steroid lactone drug substance used primarily for the treatment of congestive heart failure, cirrhosis of the liver (both with and without hepatic ascitis), primary aldosteronism, and essential hypertension (Soliman et al., 1997; Marini et al., 2001; Chen et al., 2006). Spironolactone is a potassium sparing diuretic, promoting the excretion of water whilst supporting the retention of potassium. Spironolactone (Aldactone drug product) was granted marketing authorisation by the Food and Drug Administration (FDA) on 21st January 1960 with G.D. Searle and currently 6 others hold marketing authorisations (MA) with the FDA (Food and Drug Administration) for commercial production.

The manufacturing facility in Morpeth, founded by G.D. Searle in 1969, has changed ownership a number of times, and is currently owned by Piramal Healthcare. Throughout this period, spironolactone has been one of the main drug substances produced at the site. Although spironolactone has been off patent for a number of years, it still remains commercially important as it took over 40 years for another therapeutic compound (eplerenone) to be found that could compete with spironolactone, however this did not make it to market until 2002. Additionally, the number of patents issued for applications of spironolactone and similar molecules has increased over recent years with over 50 patents granted in 2012 (WIPO, 2012). Spironolactone therefore remains an important product for Piramal and any improvement in the process will have significant impact on the business.

This section introduces spironolactone and the associated manufacturing process as performed at Piramal Healthcare, Morpeth, starting with the process chemistry, followed by the manufacturing process, process control, and the associated process data.

2.1 Spironolactone process chemistry

2.1.1 Overview of synthesis

There are several manufacturing routes in the literature (Weier, 1978; Cella et al., 1961; Dryden and Markos, 1977; Somberg and Ranade, 2009) for the synthesis of spironolactone. A change in synthesis route would require an expensive and re-registration for spironolactone API and all of the associated drug products in each market in which it is registered. Although Europe has the EMA (European Medicines Agency) , each country would require a re-registration, and the requirements and time scales can vary significantly resulting in a very costly and complicated supply chain for spironolactone. A change in route would therefore only be considered for significant cost, environmental, legislative, or safety reasons. The synthesis route used at Morpeth can be split into three distinct parts, each producing a relatively stable intermediates or products.

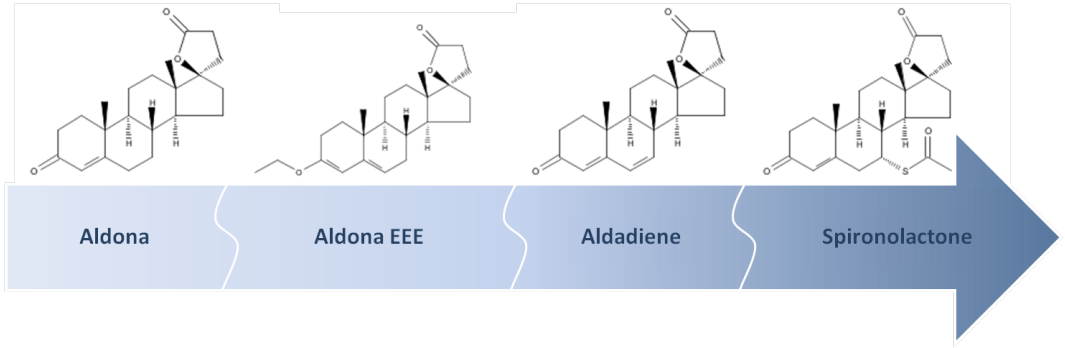


Figure 2.1: Overview of the synthesis route from aldona to spironolactone at Piramal Healthcare

The first of these is the conversion of the starting material, 17β-hydroxy-3-oxo-pregn-4-ene-21-carboxylic acid γ-lactone (aldona), to the first intermediate, 17β-hydroxy-3-oxo-pregn-3,5-diene ethyl enol ether-21-carboxylic acid γ-lactone (aldona ethyl enol ether or AEEE), using triethyl-orthoformate.

Following the isolation of AEEE, it next undergoes a mono-halogenation to yield 6-bromo AEEE, followed by a dehydrohalogenation and hydrolysis of the enol ether function to yield the second isolated intermediate, 17β-hydroxy-3-oxo-pregn-4,6-diene-21-carboxylic acid γ-lactone (aldadiene).

The final step to yield the spironolactone drug substance is the thiolacetylation of aldadiene using thiolacetic acid to give 17-hydroxy-7α-acetylthio-3-oxo-17α-pregn-4-ene-21-carboxylic acid γ-lactone acetate (spironolactone).

Spironolactone exhibits optical isomerism from the 7-acetylthio function of the molecule. This is seen during the synthesis of spironolactone from aldadiene where both the products 17-hydroxy-7α-acetylthio-3-oxo-17α-pregn-4-ene-21-carboxylic acid γ-lactone acetate (7α-spironolactone) and 17-hydroxy-7β-acetylthio-3-oxo-17α-pregn-4-ene-21-carboxylic acid γ-lactone acetate (7β-spironolactone) are formed (Somberg and Ranade, 2009). The thiolacetylation of aldadiene is reversible so the reaction system is able to convert the undesired 7β-spironolactone to the desired and efficacious 7α-spironolactone given appropriate conditions.

2.1.2 Spironolactone impurities

Both the chemistry involved in the synthesis of spironolactone and the raw materials used leave room for undesirable side reactions to take place throughout the process. This invariably leads to impurities in the final product (Chen et al., 2006). Thiolacetic acid for example, degrades in both the presence of oxygen and water to several different products. Furthermore, due to the presence of methanol as a reaction solvent, degradation of thiolacetic acid occurs during the thiolacetylation reaction to form hydrogen sulphide, which is then able to form a number of impurities with spironolactone. Although a number of measures have been implemented to remove oxygen from the manufacturing equipment (through the use of nitrogen purges and nitrogen blankets for example), small quantities of oxygen will enter the process train, be this through vacuum transfers from bulk drums to reagent header tanks where air may be drawn in, or raw material sampling for example. The reaction conditions also require moderate heat input which may again be favourable to the formation of undesirable products.

There are currently seven known impurities of spironolactone that are seen and have specification limits for the manufacturing process at Morpeth. The first of these is the starting material, aldona. This impurity may arise from either the aldona not being completely converted to AEEE , or through degradation of AEEE back to aldona, which is not removed completely during the filtration and centrifugation steps. The specification limits are not more than 0.1% and typically only 0.03% is seen in the analytical results for product.

Another impurity that can be found in the drug substance is the intermediate product aldadiene. This is again due to either the aldadiene not being completely converted to spironolactone, or through degradation of the spironolactone back to aldadiene, which is not

removed completely during the filtration step. The specification limits are not more than 0.5% and typically only 0.1% is seen in the analytical results for product.

The Δ20-spironolactone impurity arises from an impurity found in the aldona start material. This impurity is not removed during any of the isolations and is carried through the process to produce the impurity Δ20-spironolactone. In the batches where the quantities are above the limit of detection (0.02%) the typical quantities found are approximately 0.1%.

As discussed in section 2.1.1, spironolactone is optically isomeric and the thiolacetylation of aldadiene to yield spironolactone is reversible. The 7β isomer is the kinetically favoured and thermodynamically more stable isomer produced in parallel with the desired 7α isomer during the thiolacetylation. The 7α isomer is less soluble and therefore by crystallizing out the desired 7α isomer during the reaction, the 7β isomer is forced back to aldadiene through Le Chatellier's principle and a higher yield of 7α-spironolactone is obtained, however small quantities of 7β-spironolactone may be found in the finished product.

Another known impurity is 4-bromo spironolactone. This is formed through a side reaction during the synthesis of the aldadiene intermediate. It is typically a result of over bromination of the AEEE resulting in a failure to scavenge all of the bromine. The subsequent purification and isolation steps are unable to remove this impurity so it is carried through to produce 4-bromo spironolactone. Quantities observed in the product are typically around the limit of detection of 0.02%.

The two final known impurities of spironolactone are known as bis-spironolactone and per-spironolactone. These are both a result of degradation and side reactions involving thiolacetic acid. Thiolacetic acid degrades in the presence of oxygen to form the bis-thiolacetic acid. Thiolacetic acid also degrades in the presence of oxygen and hydrogen sulphide (formed through a reaction between thiolacetic acid and methanol) to form per-thiolacetic acid. These degradates compete with thiolacetic acid resulting in the formation of bis- and per-spironolactone impurities.

2.1.3 Polymorphism of spironolactone

In addition to the impurities associated with the spironolactone process chemistry, spironolactone also exhibits polymorphic behaviour. Polymorphism is the ability of a material to have more than one crystalline form. There are two types of polymorphism, monotropic

polymorphic materials have one polymorphic form that is more thermodynamically stable than another. Enaniotropic polymorphic materials however may have multiple thermodynamically stable polymorphic forms. Polymorphism is a complex phenomenon as McCrone (1965) details *"the number of [polymorphic] forms known for a given compound is proportional to the time and money spend in research to that compound"* (McCrone, 1965; Myerson, 1993). This is due to several factors that may have an influence on the polymorph forms including temperature, pressure, a change in the stirring of a reactor, impurity content and growth rate (Myerson, 1993).

The polymorphic form of both pharmaceutical drug substances and pharmaceutical grade excipients used in drug product formulations is important as different polymorphic forms have different physical properties. As the physical properties, such as solubility and compressibility, of a polymorph change the bioavailability of a drug and the manufacturability or formulation of a drug product respectively (Buckton, 2013; Myerson, 1993).

There have been several polymorphic forms of spironolactone reported in the literature (Marini et al., 2001, Agafonov et al. (1989), El-Dalsh et al. (1983), Espeau et al. (2007), Nicolaï et al. (2007), Agafonov et al. (1991) , and Salole and Al-Sarraj (1985)). The spironolactone polymorphic forms manufactured at Morpeth are predominantly spironolactone type II with some traces of type I. Nicolaï et al. (2007) claim the formation of the crystal structure of spironolactone indicates that it is first the solvated forms of spironolactone that are formed, which upon desolvation undergoes minimal change in the three dimensional lattice structure to form the type I polymorph.

The type I polymorph is less thermodynamically stable than the type II polymorph and therefore, when heat is applied (during the drying process for example), the type I polymorph is converted to the type II polymorph (Espeau et al., 2007 and Nicolaï et al. (2007)). The work that these conclusions are based on uses the ethanol solvated form of spironolactone, whereas the reaction system operated at Morpeth uses methanol and acetone as the solvents at this stage of the manufacture. Further work using a similar method to that carried out by Espeau et al. (2007) and Nicolaï et al. (2007) with the solvents in the concentrations used at Morpeth would be required before a solid understanding of the formation of the crystal form can be understood.

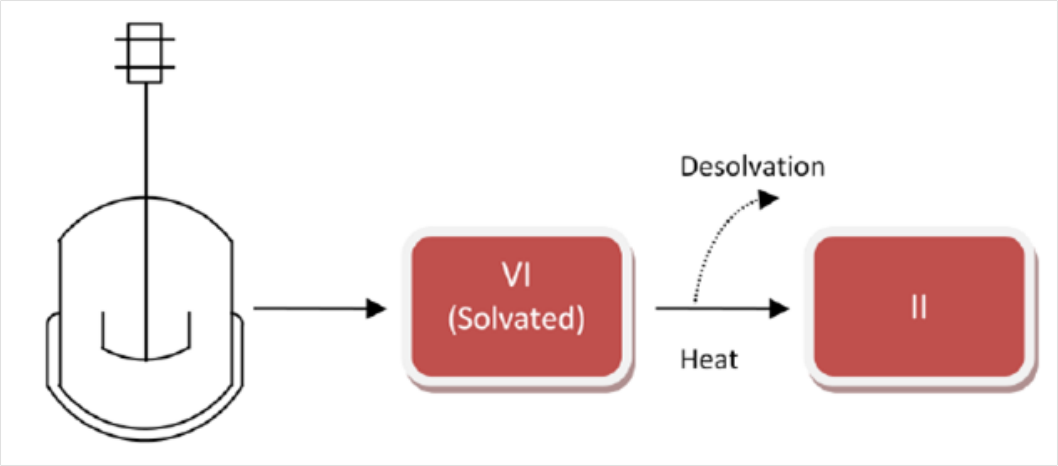


Figure 2.2: Spironolactone polymorph form II production through desolvation of the solvated form VI

Liebenberg et al. (2003) and Agafonov et al. (1991) are both in agreement that the type II spironolactone polymorph is more thermodynamically stable than the type I polymorph, and that the solvated form (VI)is created first before desolvation to the de-solvated polymorphic forms (figure 2.2). Agafonov et al. (1991) show however that the type I polymorph of spironolactone can be formed from the rapid cooling of spironolactone in acetone (solvent boiling down to 0 °C in a few hours) and the type II polymorph can be formed from natural evaporation of the solvent over several weeks. This suggests that the spironolactone in the acetone will crystallize to form the type I polymorph (figure 2.3).

Agafonov et al. (1991) state that both the type I and type II polymorphic forms are monotropic as no transformation between the two was seen upon heating. This understanding is challenged in Liebenberg et al. (2003) where gradual conversion from type I to the more stable type II was seen in the temperature range of 25 °C to 75 °C and rapid changes above 100 °C suggesting enantiotropic behaviour of the type I polymorph. This is then added to in Espeau et al. (2007) where it is claimed that the transition from type I to type II is not an enantiotropic related solid - solid transition, but a melting - recrystallization process occurring at about 100 °C to 120 °C.

In summary, the crystal properties of spironolactone are complex and, as a result, not fully understood. There is evidence of type I - type II transition; however it is unclear if this is an enantiotropic transition, a melt - recrystallization transition, or both. What is clear is that both the type I spironolactone polymorph and the type II spironolactone polymorph (the more thermodynamically stable form) will be produced in the reactor at Morpeth. Type I is formed

from the spironolactone in acetone, whereas type II is formed from the desolvation of methanol solvated form (VI).

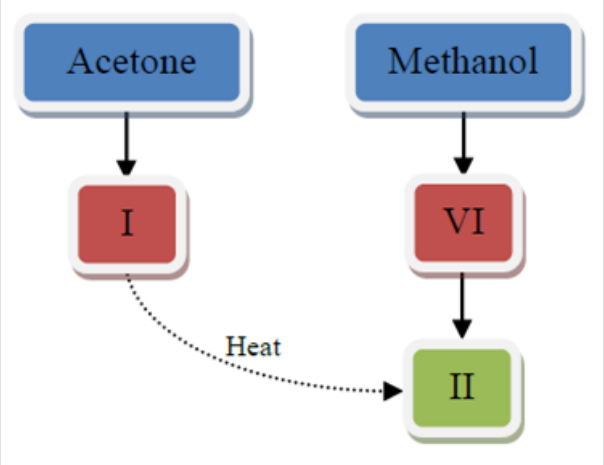


Figure 2.3: Spironolactone polymorphic form conversion routes from methanol and acetone crystallizations

2.2 Spironolactone manufacturing process

The process of manufacturing spironolactone from the start material aldona can be split into three parts. The first produces the intermediate aldona ethyl enol ether (AEEE) and is performed at Piramal Healthcare, Morpeth (figure 2.4). The process starts with washing the reactor with acid in preparation. Methanol is then charged to the reactor which is distilled. Subsequently toluene is charged to the reactor and the methanol toluene mixture is distilled. The reactor is finally rinsed with tetrahydrofuran to ensure that it is clean, dry, and free of traces of alkaline material from previous batches that could interfere with the chemistry of the next batch. When the reactor is confirmed as clean and dry through a visual inspection, the aldona is charged and dissolved in solvent and other reagents. A catalyst is then charged to the reactor to allow the reaction to take place. When the reaction is complete, the product (AEEE) crystallises out of solution and some of the solvents are distilled off. The AEEE is then isolated through crystallization, centrifuged and washed to remove the mother liquors. The AEEE is then dried under vacuum before being weighed into drums and subject to quality analysis consisting of material identification, assay, and impurity identification and quantification.

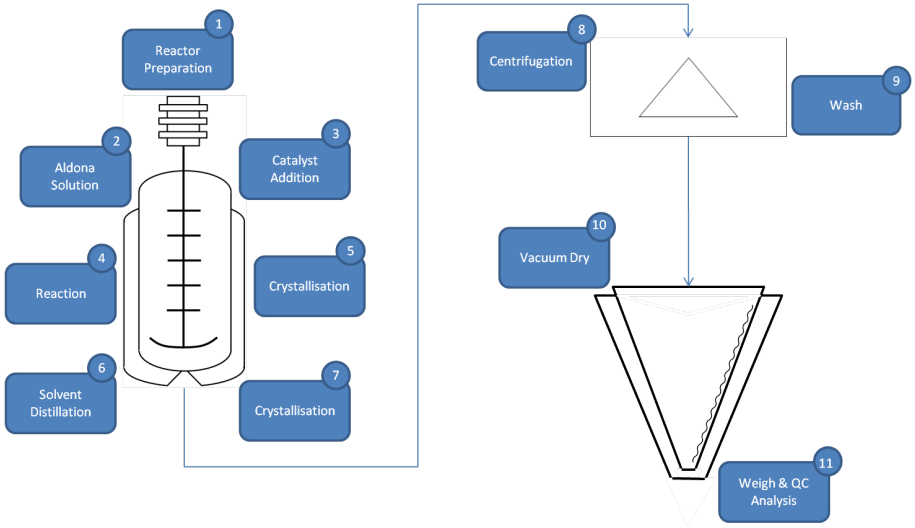


Figure 2.4: Aldona to aldona ethyl enol ether manufacturing process

The second manufacturing step (figure 2.5) is the conversion of AEEE to aldadiene. Again this process starts with reactor preparation ensuring that the reactor has been cleaned and dried. Solvents and reagents are then metered into the reactor before the AEEE is weighed and charged into the reactor to dissolve in the solvents. This is then brominated using hydrogen bromide. As this is an exothermic reaction, the additions are performed by the operator in small quantities and the cooling is applied manually. The primary control for the reactor temperature is through the rate of addition. After an in process test to confirm that the reaction is complete, a bromine scavenging step is performed to remove excess bromine and prevent the formation of the di-bromo impurity. Some of the solvents are then removed through distillation before the intermediate is acidified to hydrolyse the molecule. Water is then charged to the reactor to cause the intermediate product (aldadiene) to crystallise out of solution. The aldadiene is then washed in a DCS (Distributed Control System) controlled centrifuge before being dried under manual control to remove the residual traces of solvents. The aldadiene remains in situ in the dryer with no quality analysis or yield quantification performed other than loss on drying to reduce the manual handling of the intermediate. Piramal Healthcare, Morpeth report this part of the process to be robust with consistent yields and quality based on monitoring prior to the removal of the testing on the intermediate.

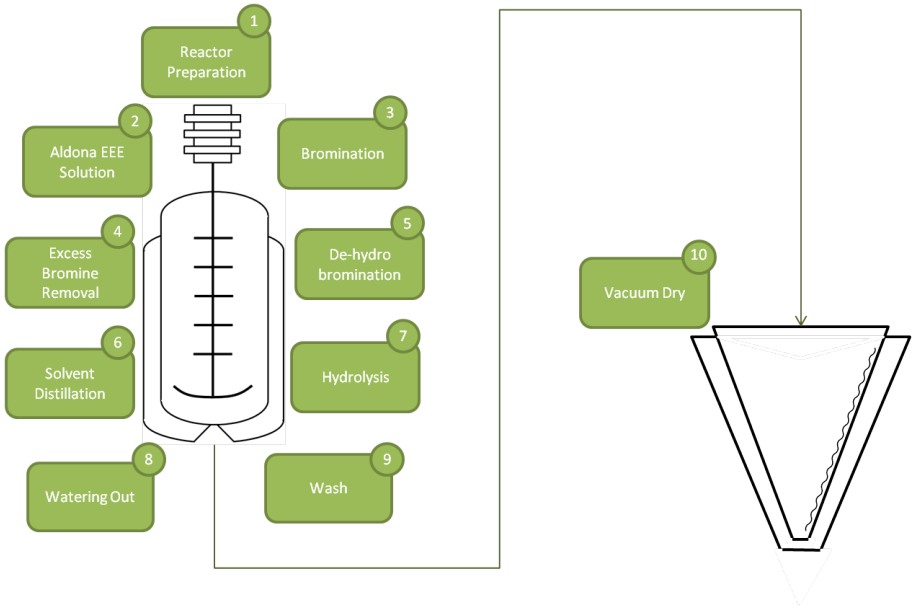


Figure 2.5: Aldona ethyl enol ether to Aldadiene manufacturing process

The final stage in the spironolactone manufacturing process is the conversion of aldadiene to spironolactone (figure 2.6). This process starts with the dissolution of the Aldadiene intermediate within the dryer using methanol and acetone as solvents. The solution is then transferred using a double diaphragm pump to a 500 gallon stainless steel reactor in which activated carbon is added as a colour treatment. The batch is then recirculated through a series of 1μm bag filters before being transferred to a 500 gallon glass lined reactor via a 0.45μm and 0.2μm cartridge filters. Thiolacetic acid is then charged to form spironolactone which is then isolated through the addition of more methanol and a reduction in temperature. The spironolactone slurry is then pumped to a Rosenmund pressure filter and washed with chilled methanol to remove the mother liquors. The spironolactone cake is then dropped into a conical screw agitated dryer with a hot water jacket to drive off the residual solvents. The dry product is then pneumatically conveyed to a jet mill where it is micronized to the desired particle size before being weighed into drums and sampled for QC release testing.

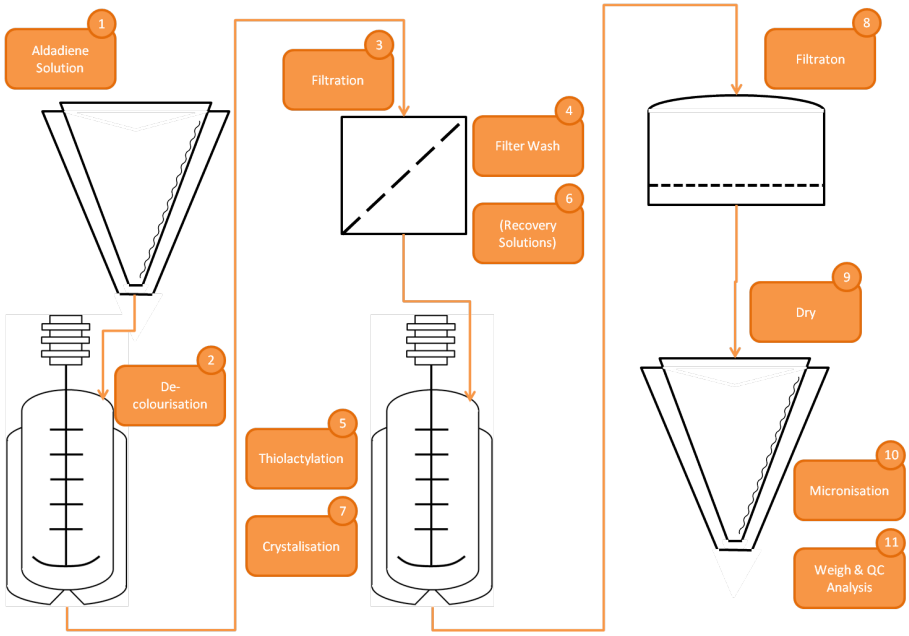


Figure 2.6: Aldadiene to spironolactone manufacturing process

As the manufactured AEEE is isolated, sampled for QC analysis, and a known quantity charged at the start of the Aldadiene process, this part of the process is independent from the Spironolactone manufacturing process and is therefore out of scope. There is no yield of aldadiene quantified however therefore it is difficult to separate this process from the spironolactone process. A simplified process schematic (figure 2.7) shows how the AEEE process is completely separate from the remainder of the process.

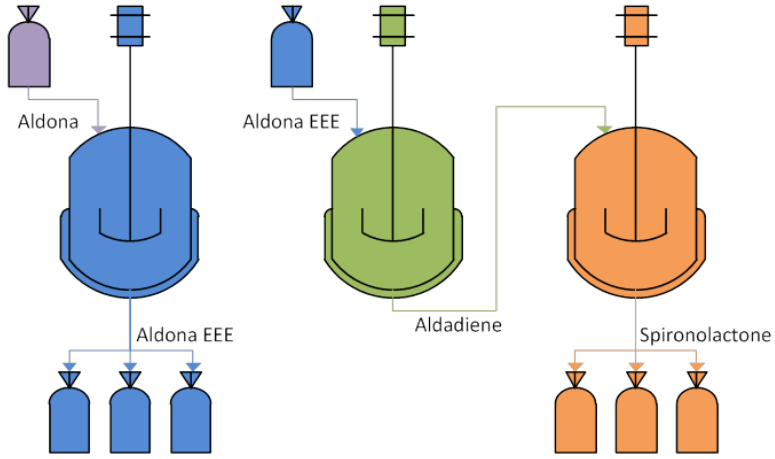


Figure 2.7: Simplified spironolactone manufacturing process schematic

Previous work had been carried out at Piramal Healthcare to attempt to map the losses of the process from AEEE through to Spironolactone. The losses from a batch throughout various stages of the process were compared to the theoretical losses to try and identify the greatest

improvement opportunity. The results from this work showed the yield for the aldadiene manufacturing process to be above 97%. The drying and micronization steps gave a yield of 93% and the thiolacetylation, isolation, and filtration gave a yield of 73%. It is clear, therefore, that the largest losses of product compared to the theoretical losses are in the aldadiene to spironolactone stage. Additionally, the drying process of the aldadiene to spironolactone stage is currently the bottleneck of the process. The aldadiene to spironolactone process is the manufacturing process to be considered during testing of the framework presented in this thesis.

It should also be noted that there may be significant errors in the data obtained for the yields quoted above arising from sampling and analytical errors. The filter bed will settle differently between each re-slurry operation and temperature changes in the filter may also lead to variable losses of product into the mother liquors over time. Without extensive sampling, which was not feasible due to time, cost and material handling issues, this will not be observed. Analytical inaccuracies may also be present, as the analytical methods have a limit of detection and also rely on sample preparation which can lead to less than 100% recovery of the compound of interest. Nonetheless, the difference in the recovered yields between the three parts of the process quantified show that the thiolacetylation and isolation process are significantly different to the remainder of the process; and looking for improvements in this part of the process is most likely to have the greatest impact on the overall process yield and cycle time.

2.2.1 Spironolactone process data

Data is generated throughout the batch manufacture in each of the major equipment items. For reactor R101, this includes the reactor weight, temperature, pressure, header tank (T104) weight, and temperature controller output. The filter F101 stores pressure data, however this is of little use as the range of the transmitter is very small therefore the pressure data is out of range for the duration of use of the pressure filter. The dryer D101 collects temperature data from both the contents probe and the jacket return, in addition to the pressure and the temperature controller output. The data recorded from the plant instrumentation is passed to a data historian through a compression algorithm. The algorithm is a modification of the boxcar compression method discussed in more detail in section 2.2.1.

Collection of process data

The frequency of data collection is controlled by the DCS. This scans each of the measurement instruments every 30 seconds and passes this data to the data historian for compression and storage. Most of the process variables have relatively slow dynamics, allowing the 30 seconds measurement frequency to be sufficient to observe the behaviours in the process. There are a few places where increased measurement frequency would increase the available resolution in the process data, such as during the nucleation event in the reactor. The temperature rise due to the latent heat of crystallization and subsequent temperature fall caused by the rate of cooling exceeding the rate of heating occurs within two to three minutes and therefore only a small number of data points capture this behaviour. With an increased measurement frequency the maximum temperature rise by the latent heat of crystallization could be more accurately measured.

The data recorded from the plant instrumentation is passed to a data historian through a compression algorithm. The algorithm is a modification of the boxcar compression method which also incorporates a gradient limit to further increase the compression achieved.

The compression algorithm starts from the last recorded data point and sets boxcar limits up on this point (i.e. current value plus or minus a small deviation). When the new data is passed to the algorithm, it calculates a back-slope (gradient) limit. The data passed from the plant is monitored until both the boxcar and the back-slope limits have been exceeded, or fifteen minutes has passed between recorded data points. The previous data point is then stored permanently in the data historian. This process is repeated for every measurement that is received from the plant.

This data is compressed using a modification to the boxcar compression method which also incorporates a gradient limit to further increase the compression achieved, as outlined in figure 2.8. The compression algorithm starts from the last recorded data point and sets boxcar limits on this point (i.e. current value plus or minus a small deviation). When the next data point is passed to the compression algorithm, it calculates a gradient (backslope) and again calculates deviation limits based on this backslope. Subsequent data is monitored by the algorithm until a data point falls outside both the boxcar and the backslope limits, or 15 minutes has passed since any new data was permanently stored. When these conditions have been met, the data point passed to the historian immediately before both the limits were exceeded is permanently stored and new boxcar and backslope limits are set.

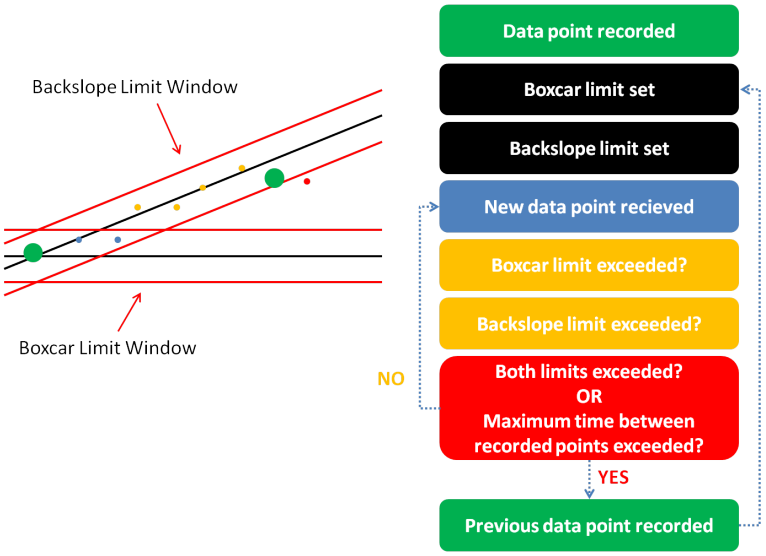


Figure 2.8: Process data compression algorithm

The limits used in the boxcar-backslope compression algorithm are tighter than the instrument tolerance permissible for instrument calibration. Although this results in some error in the interpolated data, the quantity of error is within the calibration tolerance on the measurement variable and therefore of small significance.

Accuracy of process data

There are a number of sources of error in the process data that can impact the accuracy and precision of the data. The first of these is the accuracy and precision of the measurement instruments as installed on the process. The instruments installed are calibrated to within 1% of the measured value (table 2.1) however the calibrations are typically much tighter than this. This indicates that the accuracy of the instruments could vary significantly (e.g. up to ± 1.6 °C for temperature and ± 30 kg for weights). The precision of instruments are typically at least one order of magnitude better than the accuracy of the instrument as is the case with the process measurements from this process. Relatively small changes (within the first decimal place) that are observed on the process variables are representative of the changes in process conditions local to the measurement device.

Another factor that can influence the accuracy of the process measurements is the physical location of the measurement device. The measurement indicated from devices such as temperature probes can only be the value experienced by the measurement device. The use of thermowells to site temperature probes can cause a lag in the measurement of the true temperature as heat has to be transferred through the thermowell material. The overall mass

and surface area of a thermowell may also smooth out some local variations in temperature. Additionally the point of measurement may not be representative of the conditions throughout the entire vessel. For example, if a reaction is fast and exothermic, the temperature at the addition point of a reagent will be greater than the temperature at the probe location.

Finally, the data that is obtained from the plant is compressed before it is stored. The action of data compression and recovery will add another error to that already present in the data. The deviation limits for both the boxcar and the back-slope used within the compression algorithm are shown in table 2.1. The tolerances used for the data historian deviation limits are much lower than the instrument calibration tolerances. This means that the error introduced through compression of the data is much less than, and therefore insignificant compared to, the error already in the measurement.

The error introduced to the data through the compression algorithm is significantly less than the instrument tolerances on the process plant and therefore the values obtained through interpolation of the stored data will give sufficient understanding of how the process data behaviour during the manufacture of the batches.

2.2.2 Data challenges and proposed data processing framework

Challenges with process data

There are several challenges associated with process data. The first of these is the alignment of the data. Not only does the start time of each batch need to be aligned with the start time of the other batches, some batches, or steps in the manufacture may take longer. This can lead to misalignment of process steps between batches in an analysis even if the two batches were aligned at the start. Techniques for overcoming data alignment challenges are discussed in more detail in section 3.3.

Another challenge in the analysis of process data is handling missing data. Missing data can present problems in the analysis of batch data as techniques requiring matrix manipulation (PCA for example) cannot perform the mathematics if any value in the matrix is missing (Lennox et al., 2001). As the dynamics of the process data are relatively slow, when small quantities of missing data are present this can be interpolated with little detriment to the reliability of the data. If the missing data however occurs in an area of process data with faster dynamics (e.g. the heat of crystallization) in which the data changes direction relatively

Table 2.1: Instrument calibration tolerances and historian deviation limits

Description	Calibration Tolerance	Scan Rate	Historian Tolerance
R101 Contents Temperature	-40 – 160 °C ±1%	30 seconds	0.1 °C
R101 Weight	0 – 3000 kg ±1%	30 seconds	0.1 kg
R101 Pressure (Blanket)	0 – 300 mmWG ±1%	30 seconds	1 mmWG
R101 Pressure (Full Range)	0 – 2.5 barA ±1%	30 seconds	0.01 barA
R101 Jacket Temperature Controller	0 – 100%	30 seconds	0.1 %
R102 Contents Temperature	-40 – 160 °C ±1%	30 seconds	0.1 °C
R102 Weight	0 – 3000 kg ±1%	30 seconds	0.1 kg
R102 Pressure (Blanket)	0 – 300 mmWG ±1%	30 seconds	1 mmWG
R102 Pressure (Full Range)	0 – 2.5 barA ±1%	30 seconds	0.01 barA
R102 Jacket Temperature Controller	0 – 100 %	30 seconds	0.1 %
T104 Weight	0 – 600 kg ±1%	30 seconds	0.1 kg
D101 Pressure (Full Range)	0 – 2.5 barA ±1%	30 seconds	0.01 barA
D101 Pressure (Vacuum)	0 – 1000 mbar ±1%	30 seconds	1 mbar
D101 Jacket Temperature Controller	0 – 100%	30 seconds	0.1 %
D101 Jacket Temperature	0 – 150 °C ±1%	30 seconds	0.1 °C
D101 Contents Temperature	-20 – 130 °C ±1%	30 seconds	0.1 °C
X101 Venturi Line Pressure	0 – 20 barG ±1%	30 seconds	0.1 barG
X101 Ring Line Pressure	0 – 20 barG ±1%	30 seconds	0.1 barG
X101 Feed Rate from Dosing Unit	0 – 200 kg hr^{-1} ±1%	30 seconds	0.1 kg hr^{-1}

quickly, interpolation can lead to unreliable results. When this is the case the batch has been removed from the analysis. Similarly, if the data missing spans a large time period, interpolation of the data may not give results that accurately reflect the process conditions, and therefore any batches with large expanses of missing data have been removed from the analysis.

Data processing framework

To address some of these challenges a framework is presented (figure 2.9) in which the process data is first pre-processed, followed by multivariate outlier detection, before modelling is

applied to interrogate the process data and extract information that can be used to enhance the understanding of the process. The components of the framework are discussed in the subsequent chapters, with chapter 3 detailing the pre-processing methods, and chapter 4 detailing the principal component analysis methods. Subsequently, the framework is tested on the spironolactone process data in chapter 5.

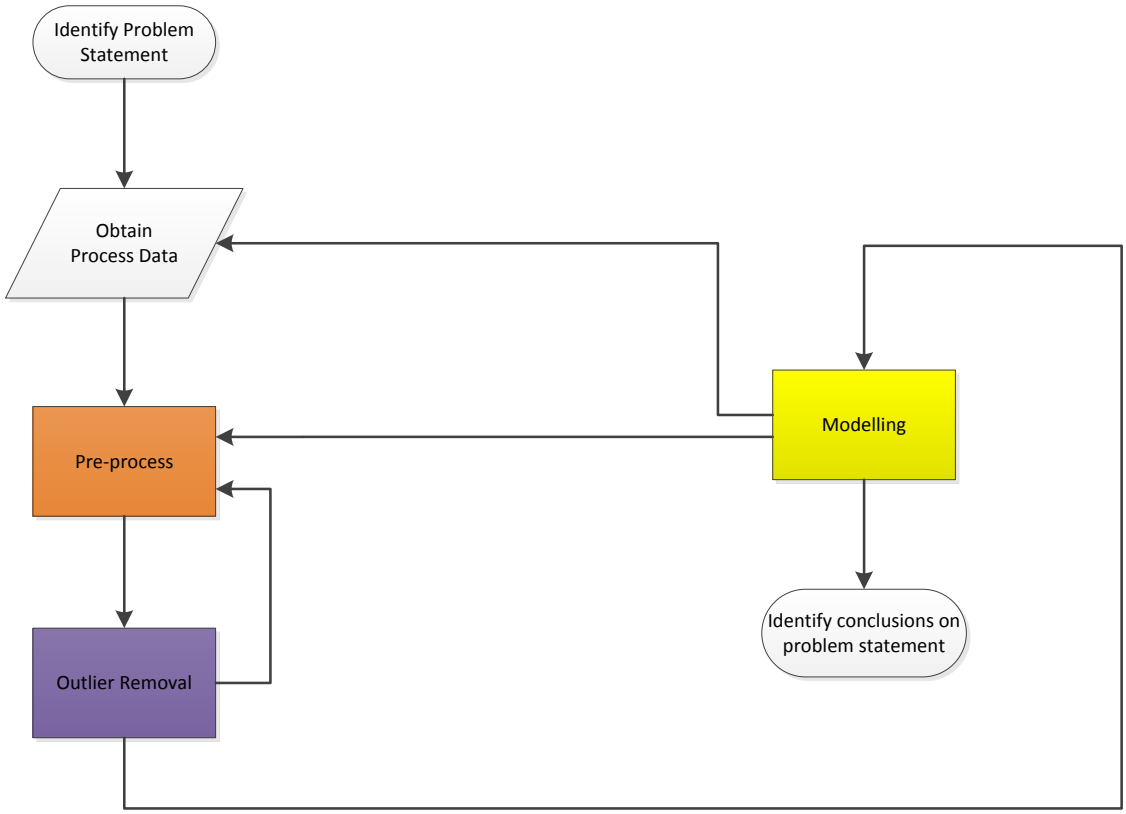


Figure 2.9: Overview of framework to extract multivariate information from batch process data (orange - pre-processing, blue - outlier detection, yellow - multivariate modelling, grey support the framework)

2.3 Summary

This chapter has introduced the spironolactone the spironolactone chemistry detailing the synthesis route employed at Piramal Healthcare, Morpeth, and potential impurities in the process. Subsequently, a summary of the polymorphic behaviour of spironolactone was presented. Although a lot of research has been performed on spironolactone, the solid state characteristics are still not fully understood in the literature, however, it is clear that

polymorphism should be considered with the solvent systems used in the spironolactone isolation process. Following the process chemistry, an overview of the manufacturing process was presented, including details on the process data available.

Chapter 3. Multivariate pre-processing methods

Pharmaceutical manufacturers are expected to consistently produce high quality products. In order to achieve this and ensure profitability, it is important that manufacturers are able to understand their processes and the effects of process parameters on the quality of products. There are a number of tools and methodologies that are available to help achieve this understanding, some of which are discussed in this chapter.

Prior to the application of a multivariate tool to interrogate the data, the data may first need to be pre-processed. Pre-processing is important to be able to get the most out of the data and remove artefacts in data that could encourage misleading results to be shown. On the other hand, care must also be taken when pre-processing data as this data manipulation can change the structure in the model and significantly alter the results obtained leading to misinterpretation of the original data.

This chapter provides an overview of the multivariate pre-processing methodologies including dealing with compressed data, dealing with missing data, data alignment, centring and scaling, and finally filtering are discussed.

3.1 Compressed data

Industrial processes can easily produce large quantities of data. Regulatory guidance in the pharmaceutical industry states that original data (including process data trends) that forms part of the batch record, and backups of the data, should be kept for a minimum of 1 year post expiry of the product or 5 years post release (whichever is longer) (European Commission, 2012), however, some data may be required to be kept for 30 years (Medicines and Healthcare Products Regulatory Agency, 2015). This results in the requirement for significant quantities of data to be securely stored for long periods of time.

There are two main reasons for the compression of process data. The first is to reduce the costs of long term data storage. Recent advances in storage media, however, have reduced the

benefit obtained by compression as the cost of data storage has significantly decreased in recent years. The second reason for compression of process data is to reduce the data transmission costs through telecommunications links. This is more applicable to the oil and gas industry where transmission of data via satellite links from off-shore platforms to on-shore headquarters is required, however, if pharmaceutical companies want to perform remote monitoring of processes through telecommunication links this may be of concern (Thornhill et al., 2004; Imtiaz et al., 2007).

3.1.1 Data compression algorithms

Data compression can fall into three classes; piecewise linear functional approximation, data transform, and vector quantization. Each of these methods is briefly discussed here, however, a more detailed comparison can be found in Watson et al. (1998).

Piecewise linear methods

Piecewise linear methods, also known as direct methods, use a set of rules on the data to determine which data points to keep and which can be ignored. Examples of these methods are the boxcar, backward slope, boxcar-backward slope, and swinging door algorithms (Watson et al., 1998). These are briefly introduced below before table 3.1 summarises their advantages and disadvantages with regards to process applicability.

1. Boxcar compression algorithm: The boxcar is a deviation limit applied to the value stored in the data historian. Boxcar compression works by testing all subsequent data points with these limits. When a new data point is found to fall outside of these limits, the data point that came immediately before (the last data point to fall within the limits) is recorded to the data historian (Watson et al., 1998). Figure 3.1 shows an example of boxcar compression.

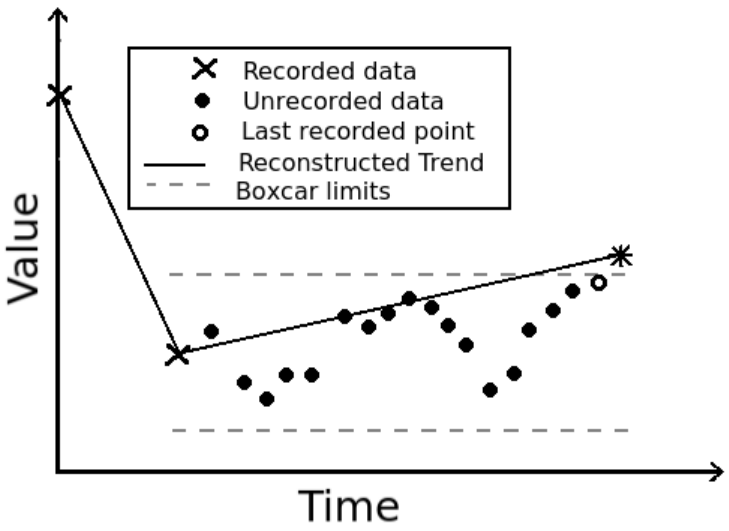


Figure 3.1: Boxcar compression. (Redrawn from Watson et al. (1998))

2. Backward slope compression algorithm: The backward slope is similar to the boxcar, however, the the deviation limit is set based on the gradient of the two previous recorded points in the data historian. All subsequent data points are tested against these limits until a data point falls outside. The last data point to fall within the limits is then recorded to the historian, and a new gradient limit is set based on the gradient between this data point and the value of the data point stored in the historian before it (Watson et al., 1998). An example of the backward slope compression is shown in figure 3.2.

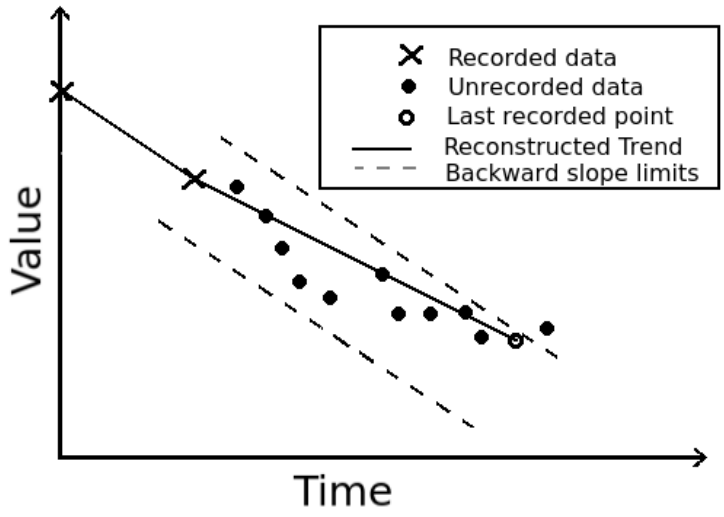


Figure 3.2: Backward slope compression. (Redrawn from Watson et al. (1998))

3. Boxcar backward slope compression algorithm: These two methods can be combined by changing the rule to store the next data point to be when both limits have been exceeded

(Watson et al., 1998). An example of the boxcar backward slope compression is shown in figure 3.3.

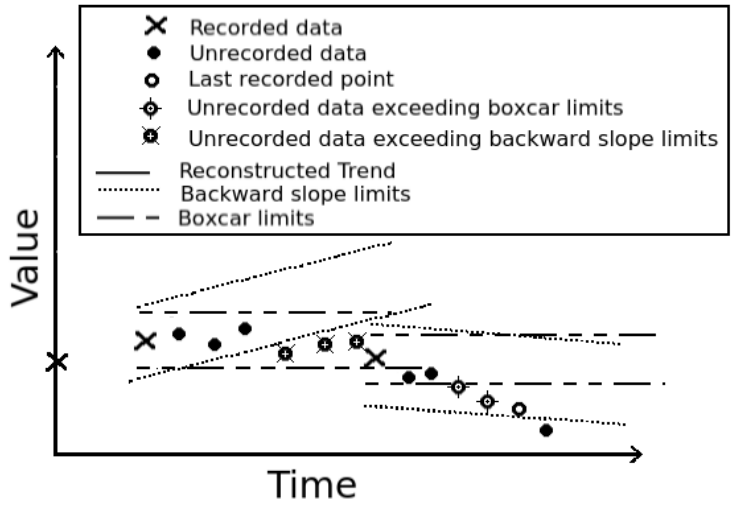


Figure 3.3: Boxcar backward slope compression.(Adapted from Watson et al. (1998))

4. Swinging door compression algorithm: The swinging door compression algorithm applies a gradient limit to the data to determine which data points need to be recorded. The gradient, however, is calculated from the last recorded value and the new value passed to the algorithm. This gradient is compared to the gradients obtained from a deviation limit around the last recorded value and the current values (Sivalingam and Hovd, 2011). The swinging door algorithm looks to approximate the original data into linear segments with each segment length maximised within tolerable error. This is achieved by placing two imaginary points (pivot points) a fixed distance above and below the last recorded value. Doors are then drawn between these points and the last recorded value, as closed doors (i.e. the lines between the pivot points and the new data intersect each other to the right of the original data point). When more data is available the doors are swung open to allow for all of the values to lie within the doors, with the top door swinging upwards, and the bottom door swinging downwards. The gradient of the doors is computed and for each subsequent value that becomes available the gradients between the new value and the pivot points are calculated. The doors are, however, only redrawn if the gradient for the top door increases with the new point, or the gradient for the lower door decreases with the new point (i.e. to allow all of the values to lie between the swinging doors). This process is repeated until the gradients of the swinging doors become divergent (i.e. the gradient of the new value to the last recorded value is less than the gradient of the upper swinging door, or more than the

gradient of the lower swinging door, resulting in the lines not intersecting to the right of the original data point) and the previous value is recorded to the data historian (Bristol, 1990) (frame 5 in figure 3.4).

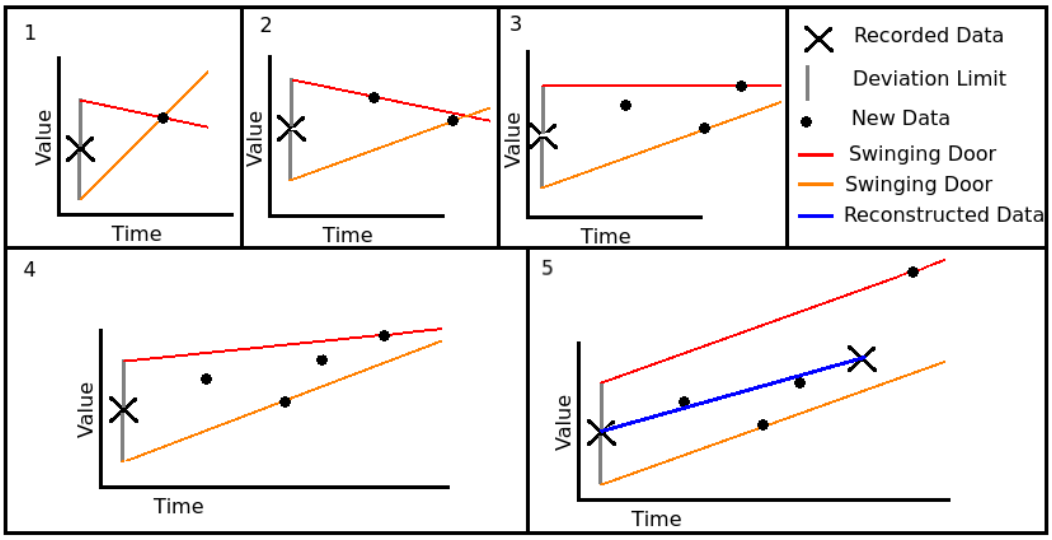


Figure 3.4: Swinging door compression algorithm

5. Piecewise linear on-line trending (PLOT) compression algorithm: The PLOT compression algorithm fits linear trends to segments of the process data based on minimising the error of the fit (RMSE). This method, however, goes further than the swinging door method as it decides if a value that falls outside of the limits is a genuine change in the process trend, or an outlier, by considering subsequent values before a decision to commit a value to the data historian is made. The algorithm uses the assumption that the sample interval is a constant, and each new value passed to the algorithm can therefore simply calculate a gradient without the requirement for additional temporal data to be included in the algorithm.

   The PLOT algorithm works by fitting a least squares straight line to all of the data since the last recorded value. From this data an estimate of the data variance is obtained which is then used to construct interval limits $(1 - \alpha)$ around the prediction of the next value. If the next value falls within the $1 - \alpha$ interval, the process is repeated including this new value in the fitting of the least squares straight line, and the estimation of the data variance. If, however, the value falls outside of the $1 - \alpha$ prediction interval, it is either indicative of a change in trend and therefore the last point to fall within the prediction interval is recorded to the data historian, or it is indicative of a potential outlier.

   The potential outlier is tested with the next value to be passed to the algorithm. If this

new value falls within the same prediction interval, and not the extrapolation from the potential outlier and the value immediately previous to it, then the value was considered as an outlier and omitted from least squares fitting and variance estimation. If this new value falls outside of the original prediction interval, but within the extrapolation interval, then this is considered as a new trend and the last value to fall within the prediction interval is recorded to the data historian. If the new value falls either within or outside of both the prediction interval and the extrapolation interval, the test is inconclusive and is therefore repeated with the next value to be passed to the algorithm (Mah et al., 1995).

Data transform methods

Data transform methods use a transformation of the data that has an inverse allowing for perfect reconstruction of the original data. There are many different transform methods available, including Laplace, Fourier, and wavelet transforms (Watson et al., 1998).

These transforms work by translating the original process data into a smaller number of uncorrelated transform coefficients (Barsanti and Athanason, 2013). The transform of the data can be restored to a perfect copy of the original data through the inverse transform (lossless compression). However, often the high frequency portion of the data is associated with noise, whereas the information of interest falls within the lower frequency portions of the data. By setting the high frequency portion of the transformed data to zero, known as thresholding, and storing the remaining transform coefficients in a data historian the original signal can be compressed (Watson et al., 1998; Nesic et al., 1996).

Applying the inverse transform to these compressed coefficients will yield a representation of the process data, with some losses of the original signal (lossy compression). This method however, lends itself to off-line processing of the data as the trend is required before its transform can be performed (Thornhill et al., 2004).

Each of the different transform compression methods work in the same way, first applying a transform, followed by thresholding, and finally applying the reverse transform to reconstruct the data. The difference between the methods is the function used to obtain a transform of the original data. When using the Fourier or cosine transforms, only the frequency information is retained in the transformed data, and the temporal information relating to this is lost. These transform methods can provide information on which frequencies were present in the input

signal, however, not where they occurred. The discrete wavelet transform, on the other hand, provides both the frequency information in the transform, and the time locations where these frequencies occurred. This is achieved by using multi resolution analysis where high temporal resolution is obtained for high frequencies at the expense of resolution of these frequencies, followed by reduced temporal resolution of the lower frequencies allowing for increased resolution of these frequencies. That is, the temporal location of high frequencies are known but the distribution of the frequencies is not well defined, whereas the distribution of the low frequencies is well defined, however, the temporal location of these frequencies in the entire signal is not known (Watson et al., 1998). This is represented in figure 3.5 showing the resolution of a discrete Fourier transform compared to that of a discrete wavelet transform. The resolution is indicated by the red boxes on the frequency time plot. As the Fourier transform only contains the frequency information, the bars are narrow on the frequency axis (high frequency resolution) but cover the entire time axis (no temporal resolution). This contrasts with the wavelet representation where for the high frequency region the bars are tall and thin representing high temporal resolution, but low resolution of the high frequencies, whereas the lower frequencies have short wide bars indicating high resolution of the frequencies, however, lower temporal resolution.

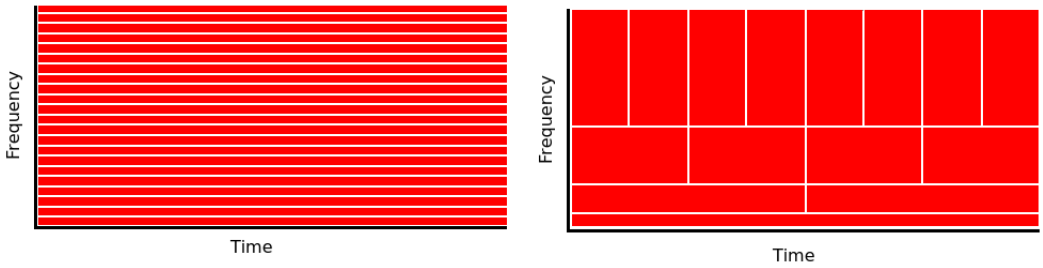


Figure 3.5: Resolution analysis representation of Discrete Fourier Transform (left) and Discrete Wavelet Transform (right). (Adapted from (Watson et al., 1998))

Vector quantization

Vector quantization is another compression technique that works akin to rounding numbers to a desired level of accuracy. This rounding process is not limited to number of decimal places however. Considering the scalar example, the possible values for the data is divided into segments called cells. The cell size is determined by the density of the data in the training data set. Where there is a high density of data, the cell sizes are small to minimize the error, whereas in low data density regions the cell sizes are increased allowing for reduced accuracy

where data is infrequent. This can be extended to vectors, where it is not the scalar that is segmented, but blocks of data. This vector quantization is achieved by taking a subset of the data and matching this with the data stored in the codebook. The index of the codebook that is matched is then recorded, instead of recording the quantized values and therefore compressing the data (Watson et al., 1998).

Table 3.1 summarises the advantages and disadvantages of the each of the compression methods detailed.

Table 3.1: Comparison of data compression methods

Compression Method	Advantages	Disadvantages
Boxcar	Simple Easy to implement Good for steady state processes Can reconstruct from paper copy of compressed data	Poor compression of ramps and steps Low compression ratio Poor performance for noisy data
Backward Slope	Simple Easy to implement Can reconstruct from paper copy of compressed data Improved compression of ramps and steps	Poor compression for noisy data Low compression ratio
Boxcar - Backward Slope	Simple Easy to implement Suitable for steady state, steps and ramps Better than individual boxcar or backward slope	Low compression ratio Poor compression for noisy data
Swinging Door	Simple Improved compression over boxcar backward slope Recording limit matched to noise	Low compression ratio compared to transform methods Poor performance with noisy data

Continued on next page

Table 3.1 – *Comparison of data compression methods (Continued from previous page)*

Compression Method	Advantages	Disadvantages
Piecewise Linear On-line Trending	Relatively simple Can reconstruct from paper copy of compressed data Suitable for steps, ramps and steady states Improved performance with noisy data and outliers Compression adapts to process variance Improved performance over swinging door compression	More complex to implement than swinging door compression Poorer compression compared to transform methods Slower computationally than swinging door
Transforms	High compression achievable Can achieve low reconstruction error Good for sudden changes (steps, ramps, spikes) and steady state data	Complex Off-line compression Need to select appropriate wavelet to achieve good compression and reconstruction accuracy PC required to reconstruct data
Vector Quantization	Compress infinite data once codebook is defined Suitable for steady state, ramps and steps	Codebook design is time consuming New codebook required for each variable of each process (not transferable) PC (and codebook) required to reconstruct data

A number of different techniques have been presented for compressing process data. If it is necessary to compress process data for long term storage or data transmission, the data transform method is recommended due to the low reconstruction error that can be achieved

with this method, and its ability to handle sudden changes in the data. Where simpler compression algorithms are required, the piecewise linear on-line trending method is preferred over the swinging door and boxcar methods due to its superior performance with noisy data. The boxcar and swinging door methods are not recommended for new systems, however, may be found on legacy systems, especially in the pharmaceutical area where such systems are required to be validated. If, however, compression of the process data can be avoided this is the preference for reasons to be discussed in the following section.

3.1.2 Challenges posed by compressed data

Reconstruction following data compression does not result in a perfect reconstruction of the original data. This can result in problems when trying to use the reconstructed data for analysis of the process. The first of these is a problem when using piecewise linear compression techniques. In this case, if the length of a feature present in the data is smaller than the length of the linear segments as a result of having a high compression factor (CF) , or wide tolerances in the algorithm. This may result in entire features in the data being lost during compression, or poor reconstruction of features where the full magnitude or time scale of the feature is lost.

Compression factor, in its simplest form, is a ratio of the number of factors in the data before compression (N) to the number of values stored in the data historian (m). There are other definitions of compression factor that take into account the need to store time stamps with each stored value to enable reconstruction for example (Thornhill et al., 2004).

$$CF = \frac{N}{m} \tag{3.1}$$

A second problem when using compressed process data is when the data is highly noisy, or entirely noise. The compression algorithm will sample this noise and can provide a reconstruction that appears to be showing a trend when the original signal was solely consisting of noise.

A third feature of compression techniques is the loss in accuracy of the process variable mean. In order to perform steady state assessments on process data, such as mass balancing in looking for leaks, or obtaining production rates, the mean of the data is required. Therefore, if

the mean of the reconstructed data does not accurately reflect the mean of the original data, some error will be present in the analysis. A method of quantifying the impact on the process mean is calculating the Percentage Difference of between the Mean (PDM). This is the differences in the means of the original data and the reconstructed data scaled by the variance of the original data (equation 3.2). This however required the original data to be available which is often not the case when dealing with historical process data.

$$PDM = 100 \times \frac{mean(\mathbf{y}) - mean(\hat{\mathbf{y}})}{\sigma_y} \tag{3.2}$$

Similar to the error in the mean of the data from reconstructed data, is the error in the process variance. Typically the variance in a process is correlated with variance in profit. Furthermore, statistical tools that use process measurement variance, such as principal component analysis, may be affected if the variance of the reconstructed data differs from that of the original data. In this case, two ratios are useful to determine the impact on the variance following compression and reconstruction. The first of these (equation 3.3) is the ratio between the standard deviation of the reconstructed data and the standard deviation of the original data (RVC) .

$$RVC = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} \tag{3.3}$$

The second of these (equation 3.5) is the ratio between the standard deviations of the reconstruction error (equation 3.4) and the original data.

$$e_i = y_i - \hat{y}_i \tag{3.4}$$

$$RVE = \frac{\sigma_e^2}{\sigma_y^2} \tag{3.5}$$

When the sum of RVC and RVE does not equal 1, the error in the reconstruction is correlated with the reconstruction. This means that information may have been lost in the compression and reconstruction process.

Finally, it has been shown that compression and reconstruction of process data can cause non-linearity in the reconstructed process data that was not present in the original data (Thornhill et al., 2004).

Thornhill et al. (2004) concluded that if the compression factor is sufficiently low (a compression factor of 3 or less) that the use of reconstructed data can be used with caution as the performance measures were found to be similar to that of the non-compressed data.

3.1.3 Dealing with compressed data

As part of the pre-processing steps of industrial data for analysis, determining if the data has been compressed and how compressed the data is. It is also important to understand what kind of compression algorithm was used for the compression as this will help inform whether the data may be suitable for use in analysis techniques.

The simplest method is the use of the ratio of the number of compressed data points to the corresponding number of original data points and calculate the compression factor from this (Thornhill et al., 2004).

This information is not always available however, therefore an alternative method at estimating the CF is required. Thornhill et al. (2004) proposed a method using the second differential for data that was compressed using piecewise linear methods. If the data was compressed using a piecewise polynomial technique, the fourth derivative would be required.

Compression detection method

Thornhill et al. (2004) present a method for detecting the level of compression of a reconstructed dataset in the absence of the original data where a piecewise linear compression method has been used. This method uses the fact that the second derivative of a straight line is zero. Therefore, with data reconstructed from piecewise linear compression, the second derivatives with a value of zero indicate linear segments (i.e. interpolated data), and non-zero second derivatives indicate where these linear segments are joined (i.e. an actual data point stored in the data historian). Therefore by counting the zero and non-zero second derivatives from reconstructed data, one can approximate the compression factor using the following equations (equation 3.6 and equation 3.7). In cases where the piecewise compression did not use a linear function, but a polynomial selection of the appropriate derivatives can be used to

the same effect. Thornhill et al. (2004) give the example of the cubic spline function requiring the fourth derivatives to estimate the compression factor.

$$\Delta(\Delta\hat{y}_i) = \frac{(\hat{y}_{i+1} - \hat{y}_i)/h - (\hat{y}_i - \hat{y}_{i-1})/h}{h} = \frac{\hat{y}_{i+1} - 2\hat{y}_i + \hat{y}_{i-1}}{h^2} \tag{3.6}$$

$$CF_{estimated} = \frac{N}{m} \tag{3.7}$$

The above is based on knowing the original number of samples in the data before compression was performed. Ideally this would be the same as the number of data points in the reconstructed data, however, this may not always be the case. If fewer samples are reconstructed than in the original data and the number of samples reconstructed is taken as an indication of the number of samples in the original data, this can lead to underestimation of the true compression factor of the reconstructed data. Thornhill et al. (2004) describe a method of identifying when inappropriate reconstruction intervals have been used, by looking at the pattern of the zero valued second derivatives (in the case of piecewise linear compression).

Additionally, a method for assessing the arithmetic accuracy of both the compression and the reconstruction process is detailed in (Thornhill et al., 2004). This is important as arithmetic errors can lead to inaccuracies in the estimation of the compression factor.

3.2 Missing data

When dealing with industrial data, missing data can occur for a number of reasons. Communication errors in the information technology infrastructure, power failures, measurement values exceeding instrument ranges, maintenance activities, instrument breakdowns, differing sampling or measuring frequencies, etcetera. This missing data can cause problems when analysing the data as many methods cannot deal with missing values.

3.2.1 Types of missing data

There are a number of types of missing data each requiring different considerations on how it can be dealt with. From the statistics literature, there are three main classifications. Missing completely at random (MCAR) data is when the missing data does not depend on the values of any other missing or observed data point, and the missing data and observed data are not statistically different. This type of missing data can occur when data is sampled at different rates or irregularly, a sensor fails, etc. In this situation, the mechanism behind the missing data can safely be ignored (Imtiaz and Shah, 2008; Walczak and Massart, 2001).

In missing at random (MAR) data, the probability that a value will be missing is dependant only on the observed values and not other missing data (Imtiaz and Shah, 2008; Walczak and Massart, 2001). An example of such a mechanism could be in a drying process where the drying end point is determined by the final moisture content determined by a loss on drying test that is performed off line. To avoid interrupting the drying process and repeatedly removing samples from the dryer, a temperature on the dryer could be monitored until it indicates that the material is close to being dry at which point samples can be taken for the off-line analysis. The values of the missing data during the time the process was not being sampled, can easily be modelled from the monitored process data to estimate their values. The mechanism behind the missing data can be ignored if there is enough information in the data to be able to model the missing values (Walczak and Massart, 2001).

Missing not at random (MNAR) , also known at non-ignorable mechanism (NI) is when the missing data depends on both the observed data and the missing data. In this case, the mechanism behind the missing data cannot be ignored and must be included in the analysis of the data. Examples of such missing data are if a measurement value goes beyond the range of the sensor, be this an on-line sensor such as temperature or pressure measurement, or below the limit of detection in an analytical measurement (Walczak and Massart, 2001; Imtiaz and Shah, 2008) resulting in the measurement not representing the true state of the process. In the case, however, where a measurement hits a constraint where the measured value is representative of the state of the process, for example, a valve position being fully open or closed, this is not missing data and can be ignored.

3.2.2 Dealing with missing data

Several methods are now discussed in how to deal with missing once it has been identified. Each method is discussed in turn, and a summary of this is provided in table 3.2.

Deleting data

The simplest method for dealing with data set with missing values is to completely disregard the record with the missing data. When dealing with industrial batch process data, the number of batches available is often limited, and reducing the number of batches due to small quantities of missing data can lead to significantly reducing the number of batches available for analysis. Often a high proportion of the available batches will have missing data of one form or another. As other techniques are available for dealing with missing data, deleting batches is not recommended unless the missing data is to such an extent that the other techniques would not be able to reconstruct the data with sufficient confidence (Baraldi and Enders, 2010).

Zero order hold

A technique often applied in data processing software is the zero order hold. This is where the reporting frequency of the data is higher than the measurement frequency and rather than interpolating data between known values, the value of the last known data is held until the next value becomes available. This results in a stepped appearance to the data, losing some of temporal correlation between the data, however, it can be suitable for the reconstruction missing data resulting from the compression of steady state data using the box-car method (Singhal and Seborg, 2003; Imtiaz and Shah, 2008).

Interpolation (first order hold)

Interpolation between known data points is often employed in data reconstruction techniques from compressed data (such as piecewise linear methods). Additionally it can be used where data is missing for other reasons, however it should be used cautiously as it is easy to lose information and features from the data if the interpolation is for longer than the frequency of the features in the data. Furthermore, using interpolation destroys the spatial correlation between the variables (Singhal and Seborg, 2003; Imtiaz and Shah, 2008).

Multiple imputation (MI)

Single imputation is a three step process whereby a distribution from which the data can be modelled is selected (Modelling), next the model parameters are estimated (Estimation), and finally the missing values are imputed from the estimated model parameters (Imputation).

Multiple imputation is similar to single imputation, however, rather than a single estimate of the model parameters being made, multiple (M) estimates of the model parameters are made, and the missing data is imputed from these estimates resulting in M complete data sets. The benefit of this method is that MI improves the estimation of variance in the data set over that of single imputation methods by taking into account the error introduced to the data set through the model parameter estimation.

Following MI, the scalar value of any datum (Θ) can be calculate as the mean of the corresponding scalar values from the M complete data sets (equation 3.8).

$$\Theta = \sum_{i=1}^{M} \frac{\Theta_i}{M} \tag{3.8}$$

For each of the scalars (Θ) the variance can be split into two parts. The first (W) is the contribution to the variance from the imputed scalar(s) and can be defined as shown in equation 3.9.

$$W_i = \frac{1}{n(1-n)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.9}$$

The second component of the variance comes from the noise introduced from the random model parameter selection, and is known as the between imputation variance (B) and is defined in equation 3.10. The total variance (T) is however not a simple summation and a correction factor is applied (equation 3.11). This correction factor allows estimation of the contribution of the between imputation variance to reduce as the number of imputations increases (Rubin, 1977; Walczak and Massart, 2001; Imtiaz and Shah, 2008).

$$B = \sum_{i=1}^{M} \frac{(\hat{\Theta}_i - \bar{\Theta})^2}{M-1} \tag{3.10}$$

$$T = \bar{W} + \left(\frac{1+1}{M}\right) B \tag{3.11}$$

Maximum likelihood (ML)

Maximum likelihood is a method for estimating the values of the missing data through an iterative approach. The method consists of two steps, an E-step (estimation step) where typically the missing values are estimated, and an M-step (maximization step). However, with multivariate data it may not be appropriate to directly estimate the missing values due to the presence of quadratic statistics from the sums of squares and sums of products across the variables. Rather the conditional multivariate distribution of the missing values should be computed with appropriate model parameters to obtain the means, mean squares, and mean products, before the missing data can be estimated (Dempster et al., 1976).

A complete data set ($\boldsymbol{X}$) can be described by the density function shown in equation 3.12 where Θ represents the density functions parameters (i.e. mean and variance for Normal distribution).

$$p(\boldsymbol{X}|\Theta) \tag{3.12}$$

Similarly the missing data can be described by the density function shown in equation 3.13 , where Φ represents the density function parameters.

$$p(M|\boldsymbol{X},\Phi) \tag{3.13}$$

These can be combined to give the density of function of both $\boldsymbol{X}$ and M (equation 3.14).

$$p(\boldsymbol{X},M|\Theta,\Phi) = p(\boldsymbol{X}|\Theta)p(M|\boldsymbol{X},\Phi) \tag{3.14}$$

In the case of MCAR and MAR data, equation 3.13 can be rewritten as equation 3.15 enabling the initial estimation of the missing data, however, for NI missing data, the missingness mechanism needs to remain.

$$p(M|\boldsymbol{X}_{observed},\boldsymbol{X}_{missing},\Phi) = p(M|\boldsymbol{X}_{observed},\Phi) \tag{3.15}$$

The next step is the maximization step. This uses a likelihood function on Θ, i.e. any function that is proportional to the density distribution of the observed values of $\boldsymbol{X}$, which is maximised to obtain the maximum likelihood estimate for the model parameters. Typically for the normal distribution an exponential likelihood function is used as this simplifies the maths when the expression is integrated (Dempster et al., 1976; Walczak and Massart, 2001; Imtiaz and Shah, 2008).

Expectation maximization (EM)

Expectation maximization is an iterative class of algorithms, closely related with the ML methods, used to estimate missing values. The method uses four steps as detailed below:

1. Initial guess of missing data. If multiple maxima are present in the log-likelihood function, the converged solution depends on the initial conditions, and therefore some thought should be given to how the initial estimates of the missing data are obtained. Typically this is the mean values for MCAR and MAR data, however for time series or longitudinal data, interpolation or extrapolation is more often used as this represents a better initial guess of the missing data.
2. Estimation of the parameters. This is done using the ML estimation of the parameters with the current estimate of the missing data present in the data set $\boldsymbol{X}$.
3. Re-estimation of the missing elements with the new parameters.
4. Re-estimation of the parameters. The re-estimation processes are repeated until convergence has been achieved.

Walczak and Massart (2001) states that the convergence of the EM algorithm is linear and proportional to the amount of missing data. Additionally, slow convergence has been reported for some data sets. Methods are reported in the literature for speeding up the convergence, however this is beyond the scope of this thesis (Walczak and Massart, 2001).

Data augmentation (DA)

Data augmentation is a group of iterative optimisation methods that estimate the missing data and then optimize the missing value, so that there is no change to the statistical properties of the data set. This is achieved through two steps. The first step is the imputation step whereby

using the current estimate of the model parameters, the value of the missing data is taken from the conditional distribution of the missing value. Subsequently, the new estimation of the parameters is calculated based on the current estimate of the missing value. This process is repeated until the algorithm converges.

Data augmentation is similar to expectation maximization with the imputation and posterior steps of data augmentation being similar to the E-step and M-step of expectation maximization respectively. Data augmentation can be easily applied with relatively simple steps to many missing data problems and has good convergence (Walczak and Massart, 2001; Imtiaz and Shah, 2008).

Principal component methods

There are a number of algorithms around that can be applied to multivariate data sets for principal component methods listed below (Nelson et al., 1996; Imtiaz et al., 2007; Imtiaz and Shah, 2008; Walczak and Massart, 2001).

1. Single Component Projection

   Single component projection is based on the NIPALS methods, however it is non-iterative and is applied to each dimension separately. As this looks at each component independently, any error introduced from the estimation of the missing data in one component is propagated through to subsequent components and can lead to large errors in the reconstructed data matrix $\boldsymbol{X}$ if there is more than 1 high variance dimension, that is more than the first principal component is of interest (Nelson et al., 1996).

   As with most reconstruction techniques, there will be some errors associated with the method. In this case some of the error occurs through variance being incorrectly assigned to the current scores from both subsequent and previous components due to subsequent components explaining large quantities of the variance in the data, and missing data causing a violation of the orthogonality assumption (Nelson et al., 1996).

2. Projection to the Model Plane

   A method to reduce the error from the single component projection method is to calculate the scores on all the components at the same time. This is known as projection to the model plane. A known error remaining in the reconstructed data from this method is due to the attribution of some of the variance to the incorrect scores. This error is

larger when missing data causes the variables to become almost collinear. By using ridge regression or principal component regression to calculate the scores, can reduce this error (Nelson et al., 1996).

3. Conditional Mean Replacement

   Conditional mean replacement builds on the EM algorithm discussed earlier by using EM to estimate the multivariate mean and covariance matrix from the data set with missing data. As with the traditional EM approach this method relies on the specification of an appropriate distribution to fit the data, and this is typically the Multivariate Normal distribution. A problem that may be seen with this method is that in cases where a large quantity of missing data exists, the co-variance can become ill-conditioned. This can be overcome by biased regression methods such as ridge regression, principal component regression, or projection to latent structures can be used to estimate the least squares and covariance matrix.

   Sources of error in the conditional mean replacement method can originate from a lack of information in the input data, numerical ill-conditioning, noise, or violations of the least squares assumptions (Nelson et al., 1996).

4. Iterative Imputation

   Principal Component Analysis Iterative Algorithm (PCAIA) is similar to the EM approach, however it uses PCA as the maximization step, using the PCA loadings rather than the conditional expected values. Similarly, PCAIA uses these parameters in the loadings to obtain new estimates of the missing data.

   There are six steps to the PCAIA method:

   (a) Reduce the data set to $< 30\%$ missing data. This is achieved by iteratively removing the rows of the data matrix, $\boldsymbol{X}$, with the most missing data until the quantity of missing data is $< 30\%$.

   (b) Fill the missing data with the unconditional mean values. Alternative initialization methods could also be used depending on the data set being studied.

   (c) The number of components required in the model is selected based on cross-validation. More details of cross-validation can be found in section 4.1.

   (d) Singular Value Decomposition (SVD) is then performed on the data set to calculate the loadings on the principal components.

(e) The loadings are then used to predict the noise free values of the missing data to be put back into $\boldsymbol{X}$.

(f) This process is repeated, from the estimation of the number of components to retain, until the sum of the squared errors between the observed values and the predicted values are below a predefined threshold.

The benefit of PCAIA over the previous methods is that it attempts to restore the correlation structure between the variables to enable construction of a multivariate model. This method is also applicable to time varying process data, where the value of a data point depends on the value of the previous point. To enable analysis of this type of data, the PCAIA method can be applied using dynamic PCA (DPCA) to account for this time dependency in the data (Imtiaz et al., 2007; Imtiaz and Shah, 2008; Walczak and Massart, 2001).

Care needs to be taken with PCAIA, however, as the random error is ignored during the prediction of the missing data and therefore the data matrix, $\boldsymbol{X}$ can become distorted and subsequently the covariance structure may get distorted. Another consequence of this is that the amount of missing data has a direct link to the order of the model (i.e. the number of components to retain) if a method other than cross validation is used to estimate the required order of the model (Imtiaz and Shah, 2008).

5. Principal Component Analysis Data Augmentation

   Principal Component Analysis Data Augmentation (PCADA) attempts to overcome some of the draw backs identified for the PCAIA method by considering the measurement error during the data imputation. Additionally, whereas PCAIA converges monotonically to a constant value, due to the random noise inclusion in PCADA the convergence monitor is not monotonic, but varies around a value when converged and therefore a bounded convergence limit rather than a constant value should be used to determine when the algorithm has converged on a solution. Although the predictions are improved with PCADA over PCAIA, the computational cost is significantly higher, therefore with a large data set it may be preferable to use PCAIA over PCADA (Imtiaz and Shah, 2008).

6. Alternating Least Squares

   Alternating Least Squares Iterative Algorithm (ALSIA) can be used with multi-way analysis, as would be required for batch process data. This method decomposes the data

matrix alternating between each mode to estimate the model parameters, followed by estimating the missing values of $\boldsymbol{X}$ iteratively until convergence is obtained. A summary of the ALSIA algorithm is detailed below (Walczak and Massart, 2001).

(a) Initialize B and C

(b) Estimate the missing elements of $\boldsymbol{X}$ using an appropriate method

(c) Calculate A from equation 3.16

$$[A] = svd(\boldsymbol{X}^{(I\times JK)}(C \otimes B), L) \tag{3.16}$$

(d) Calculate B from equation 3.17

$$[B] = svd(\boldsymbol{X}^{(J\times IK)}(C \otimes A), M) \tag{3.17}$$

(e) Calculate C from equation 3.18

$$[C] = svd(\boldsymbol{X}^{(K\times IJ)}(B \otimes A), N) \tag{3.18}$$

(f) Calculate Z from equation 3.19

$$Z = A'X(CB) \tag{3.19}$$

(g) Calculate the predicted values of $\boldsymbol{X}$ from equation 3.20

$$\boldsymbol{X}^{Predicted} = AZ(C' \otimes B') \tag{3.20}$$

(h) Update the missing values with the predicted values

(i) Repeat the process from step (c) until convergence is achieved

ALSIA has been shown to handle missing data quantities of up to 40% well, however, as with all iterative methods for the restoration of missing data, the variability in the data can be attenuated depending on the correlation of the variables and the amount of missing data (Walczak and Massart, 2001).

7. Trimmed Score Method

The trimmed score method is a technique in which the scores for the missing data are estimated based on the scores calculated from the observed measurements by imputing

the unconditional mean values (zero value) of the missing observations (Arteaga and Ferrer, 2002; Folch-Fortuny et al., 2015). Equation 3.21 shows the estimated scores:

$$\boldsymbol{\tau} = \boldsymbol{P}^{*T}\boldsymbol{z}^{*} \tag{3.21}$$

The method however is prone to large errors when influential variables are present on the missing data. Arteaga and Ferrer (2002) presents a method of detecting if this is the case, however, the technique is not recommended as there are more appropriate techniques to use that do not have these problems such as trimmed score regression.

8. Trimmed Score Regression
   An extension of the above trimmed score method is to apply it with regression. In the case of trimmed score regression, the unknown, or missing, scores are predicted by regressing the known scores as shown in equation 3.22 Arteaga and Ferrer (2002); Folch-Fortuny et al. (2015).

$$\boldsymbol{T}_{1:A} = \boldsymbol{T}^{*}_{1:A}\boldsymbol{B} + \boldsymbol{U} \tag{3.22}$$

   Where $\boldsymbol{B}$ is the least squares estimation matrix (equation 3.23):

$$\boldsymbol{B} = (\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{T}_{1:A} \tag{3.23}$$

   The predicted scores for the missing data can be defined by equation 3.24:

$$\boldsymbol{\tau}_{1:A} = \boldsymbol{\Theta}_{1:A}\boldsymbol{P}^{*T}_{1:A}\boldsymbol{P}^{*}_{1:A}(\boldsymbol{P}^{*T}_{1:A}\boldsymbol{P}^{*}\boldsymbol{\Theta}\boldsymbol{P}^{*}_{1:A})^{-1}\boldsymbol{P}^{*T}_{1:A}\boldsymbol{z}^{*} \tag{3.24}$$

   This method has been shown to produce very good reconstruction of the missing data for multivariate analysis with some of the lowest prediction errors of the techniques discuss here (Arteaga and Ferrer, 2002; Folch-Fortuny et al., 2015).

9. Known Data Regression The known data regression technique is very similar to the trimmed scores regression method, however it regresses the known data to find the scores of the missing data, rather than regressing the known scores, see equation 3.25 Arteaga and Ferrer (2002); Folch-Fortuny et al. (2015).

$$\boldsymbol{T}_{1:A} = \boldsymbol{X}^{*}_{1:A}\boldsymbol{B} + \boldsymbol{U} \tag{3.25}$$

   Where $\boldsymbol{B}$ is the least squares estimation matrix (equation 3.26):

$$\boldsymbol{B} = (\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{T}_{1:A} \tag{3.26}$$

The predicted scores for the missing data can be defined by equation 3.27:

$$\boldsymbol{\tau}_{1:A} = \boldsymbol{\Theta}_{1:A}\boldsymbol{P}^{*T}_{1:A}(\boldsymbol{P}^{*}\boldsymbol{\Theta}\boldsymbol{P}^{*T})^{-1}\boldsymbol{z}^{*} \tag{3.27}$$

Although studies found this technique marginally better in terms of prediction error than the trimmed scores regression technique, it is more computationally expensive than the trimmed scores regression Arteaga and Ferrer (2002); Folch-Fortuny et al. (2015). Furthermore, Folch-Fortuny et al. (2015) observed that the known data regression method is not suited to certain data sets, including batch process data and spectroscopic data, due to the high number of samples compared to observations that are typically associated with these data sets. Thus, the trimmed scores regression method is preferred for such data sets.

3.2.3 Summary of methods for dealing with missing data

A summary of the reconstruction methods for missing data is presented in table 3.2.

Table 3.2: Summary of missing data methods

Method	When to use	Advantages	Disadvantages
Delete Data	**Not recommended** unless the remaining data contains insufficient data to allow another technique.	Simple	Reduces the number of batches available in the analysis Loose information Only applicable for MCAR data
Zero order hold	Suitable for data compressed with a box-car compression algorithm	Simple Suitable for steady state data Spatial correlation between variables preserved	Results in stepped data Poor reconstruction of non-steady state data Temporal correlation between variables lost Correlation structure between variables is lost

Continued on next page

Table 3.2 – *Summary of missing data methods (Continued from previous page)*

Method	When to use	Advantages	Disadvantages
Linear Interpolation	Can be used with time series data however better methods exist Useful as initialization for some more complex algorithms	Temporal correlation between variables preserved	Spatial correlation between variables lost
Multiple Imputation	If the computational cost is acceptable the method is preferable over single imputation methods	Computationally more expensive than single imputation methods Variance is related to number of imputations	Improved estimation of model parameters Noise introduced to reduce the loss in variance for estimated missing values
Expectation Maximization	Suitable for data sets with small quantities of missing data	Convergence proportional to quantity of missing data Improved estimates of missing data over ML method	Convergence can be slow More computationally expensive than ML method
Data Augmentation	When the statistical properties of the data need to be retained	Retains the statistical properties of the data set Relatively simple to implement Good convergence reported	More computationally expensive than simpler methods

Continued on next page

Table 3.2 – *Summary of missing data methods (Continued from previous page)*

Method	When to use	Advantages	Disadvantages
NIPALS	Multivariate data sets	Relatively easy to implement	Not suitable for large amounts of missing data Reduces variance for missing data
Single Component Projection	Multivariate data sets when variance is explained in small number of principal components	Relatively easy to implement	Errors propagated through subsequent components leading to potential large errors
Projection to the Model Plane	Multivariate data sets	Reduced error over single component projection	Large errors possible when missing data causes collinearity in the variables
Conditional Mean Replacement	Multivariate data sets with small quantities of missing data	Convergence proportional to quantity of missing data	Relies on the underlying distribution to be known Large quantities of missing data can cause ill-conditioning of the covariance matrix
Iterative Imputation (PCAIA)	Multivariate data sets where the correlation between the variables is required	Attempts to restore correlation structure between variables	Random error ignored can lead to distorted covariance structures

Continued on next page

Table 3.2 – *Summary of missing data methods (Continued from previous page)*

Method	When to use	Advantages	Disadvantages
Principal Component Analysis Data Augmentation (PCADA)	Multivariate data sets with small quantities of missing data where covariance structure is important	Accounts for random error and therefore maintains covariance structure and variability in restored data	Convergence monitoring more difficult as may not be monotonic High computational cost
Alternating Least Squares (ALSIA)	Multi-way multivariate data sets with missing data up to 40%	Can handle relatively large quantities of missing data	Attenuation of variability in the restored data
Trimmed score method	Not recommended as trimmed scores regression has reduced error		Can introduce large errors in reconstructed data and reduce the orthogonality of the model variables
Trimmed score regression	Multivariate data sets including batch process data	Relatively fast and has very good reconstruction	
Known data regression	Multivariate data sets	Not suitable for data where the number of samples compared to observations is high (e.g. batch process or spectroscopic data)	Low prediction errors

Although there are cases where other techniques may be suitable, as discussed in table 3.2, the preferred missing data reconstruction technique is the trimmed scores regression as this has been shown to offer the best reconstruction and is comparatively computationally inexpensive to similar performing methods. Although NIPALS is not a particularly effective method, it is

commonly implemented in commercially available software (Folch-Fortuny et al., 2015), however, Folch-Fortuny et al. (2015) have made the Matlab code for the trimmed scores regression method available.

3.2.4 Extending missing data methodology to other applications

Other areas can also be formulated as missing data problems and the above methods employed, such as reconstruction of compressed data, and data alignment for data sets with variable batch lengths.

Data compression is a NI missing data mechanism, and typically has a large percentage of missing data resulting in difficulty in retaining the correlation structure of the data. Typically direct methods are used for compression in data historians, and interpolation is incorrectly used as a default method for reconstructing this data. A more appropriate method to use that may be able to preserve some of the correlation structure is PCAIA Imtiaz et al. (2007).

3.3 Data alignment

Due to the nature of batch processes, the total duration of a batch may vary from one batch to the next. This may be for many reasons. Consider a batch that is to be cooled from 60 °C to 40 °C using a water jacket from a tower cooled water supply with a fixed flow rate, for example. In the summer when the environmental temperature is higher than in the winter, the temperature of the cooling water may be higher. This may lead to the batch taking longer to cool down to the target temperature, and the batch will therefore take longer to complete.

Another cause of process delay may be due to operator interactions. Although increasing amounts of process automation are replacing tasks that process operators perform to improve safety, quality compliance, and efficiency, there will inevitably be parts of the process, especially in older manufacturing processes or on older manufacturing plants where the operators are required to provide input to the batch to allow batch progression. An example of this is in the spironolactone manufacturing process where during the crystallization, the control system prompts the operator to visually confirm that crystallization has occurred before the batch sequence will continue to the next phase. The time taken for the process operator to respond to this prompt may vary as the operator needs to go to the crystallization vessel to observe the crystallization before returning to the control room to accept the prompt. Often

process operators are not running only one process, especially in the case where there is a lot of process automation, resulting in further delays if the operator is not in the control room when the control system prompts for the input.

There are four main challenges that need to be overcome relating to batch process data alignment. The first is how the data is compressed and stored; the second is the frequency of measurement; the third is the unequal batch length inherent with batch processes; and the fourth is also related to data alignment but is aligning the batches that have differences in amplitude of measurement signals (e.g. subtle differences in vacuum pressure, or reflux temperature).

In order to compare variables of each batch over time, the time axis needs to be aligned. There are several methods to achieve this alignment (Wan et al., 2014), some of which are discussed further here. The process data at Piramal Healthcare consists of samples at 30 second intervals for each of the variables on the spironolactone plant. As the control and data collection system is relatively old and designed in a time when data storage was expensive, the raw data is compressed prior to being stored in the data historian. This results in a variable sampling time being observed when the raw data is recovered from the data historian. For example, figure 3.6 shows the differences in the time stamps recovered from the dryer contents temperature stored in the data historian between the 8^{th} and 9^{th} January 2012.

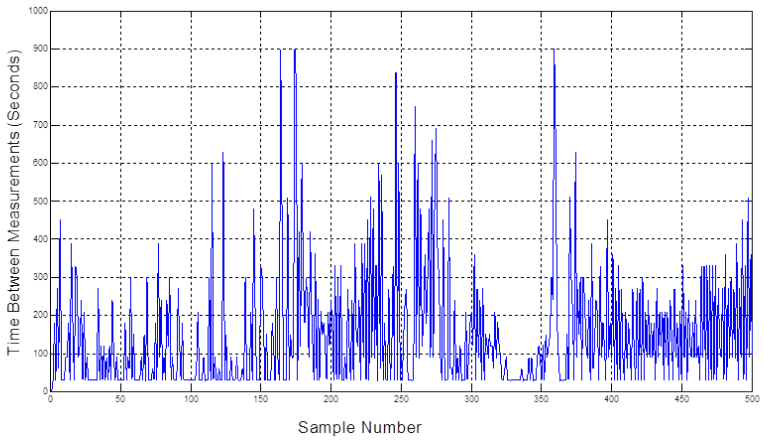


Figure 3.6: Sampling rate of compressed dryer temperature data following retrieval from data historian

The data recorded from the plant instrumentation is passed to a data historian through a compression algorithm. The algorithm is a boxcar backslope compression method and is detailed in section 3.1.

The algorithm starts from the last recorded data point and sets boxcar limits up on this point. When the new data is passed to the algorithm, it calculates a back-slope (gradient) limit. The data passed from the plant is monitored until both the boxcar and the back-slope limits have been exceeded, or fifteen minutes has passed between recorded data points. The data point prior to the data point that exceeds the limits is then stored permanently in the data historian. This process is repeated for every measurement that is received from the plant.

3.3.1 Cutting data

One method of data alignment is to simply trim the batches to the same (minimum) batch length (Marjanovic et al., 2006). This method is useful to obtain batches that are all of the same length, however care must be taken as if important information was contained in the data that has been cut this information will be lost. Additionally this method does not account for variations in the durations of different phases throughout the batch, therefore features of a batch (e.g. addition of a reagent) may remain misaligned. In certain cases it may be appropriate to trim data from either the start or during the progression of a batch. Again this may cause informative data to be lost, however it was also be a useful technique for aligning batch features.

3.3.2 Indicator variables

Another method for data alignment is the use of linear interpolation. This is where a new batch progression variable is created as a fixed length for each batch (or phase of a batch). The process variable samples from each batch are then interpolated to give batch variable trajectories of constant length for each batch (García-Muñoz et al., 2011; Fransson and Folestad, 2006; **?**). The benefit of this method is that it is simple and by performing this on each phase of a batch the impact of variability in process holds between phases can be reduced resulting in a data set with better alignment of batch features. Care must be taken with this method however as it cannot deal with variability that occurs during batch features, and the interpolation interval needs to be selected appropriately. If an interpolation interval is too long (i.e. reducing the overall number of samples to be analysed) batch features and events that occur rapidly may be omitted if the samples fall either side of the event. Similarly if the interpolation interval is too short (i.e. many more samples than originally present) an unimportant batch feature or process noise may be given more importance.

An adaptation of this method is the use of an indicator variable (García-Muñoz et al., 2003), that is a variable that changes monotonically and starts and ends at the same value for each batch. This indicator variable is then used as the batch progression indicator in exchange for the time which is included in the model as a process variable. A typical example of this in chemical engineering would be reaction conversion (from 0% to 100%). A monotonically changing variable may be difficult to obtain in batch processes due to the nature of operating in different modes, however this may be applied over a batch phase (e.g. heating the reactor contents to reflux, or cooling a batch from reflux to a batch recipe defined temperature).

3.3.3 Time warping

There are several techniques used for time warping the data including dynamic time warping (DTW), correlation optimised warping (COW), parametric time warping (PTW), peak alignment by genetic algorithm (PAGA), and semi-parametric time warping (STW) (**?**). These are at present used for alignment of chromatographic data however some of these techniques have been applied to process data (Fransson and Folestad, 2006). All of these techniques aim to align data vectors with a reference data vector.

Dynamic Time Warping (**?**)was first used for aligning the frequency data if spoken words for speech recognition. It has since found many different applications, from optical character recognition to motion tracking in games consoles, gene expression studies, and batch process monitoring. DTW non-linearly warps two trajectories so that the minimum distance between similar events is obtained. The algorithm starts by replicating the batch trajectory vectors into two square matrices ($\mathbf{X}_1$ and $\mathbf{X}_2$). The absolute difference is then taken to give a matrix of absolute differences $\mathbf{D}$.

$$\mathbf{D} = |\mathbf{X}_1 - \mathbf{X}_2{}^T| \tag{3.28}$$

The path scores ($\mathbf{P}$) are then calculated with the starting constraint that both trajectories start from time 0. This calculation of these path scores is biased to penalise for large differences from the diagonal resulting in minimising the warping distance. It is also constrained to only move down, right, or diagonally down and right by one unit. This is to prevent the warping patch from doubling back on itself.

$$\mathbf{P}_{(1,1)} = \mathbf{D}_{(1,1)} \tag{3.29}$$

$$\mathbf{P}_{(i,1)} = \mathbf{D}_{(i,1)} + \mathbf{P}_{(i-1,1)} \tag{3.30}$$

$$\mathbf{P}_{(1,j)} = \mathbf{D}_{(1,j)} + \mathbf{P}_{(1,j-1)} \tag{3.31}$$

$$\mathbf{P}_{(i,j)} = \mathbf{D}_{(i,j)} + min([\mathbf{D}_{(i-1,j)}, \mathbf{D}_{(i-1,j-1)}, \mathbf{D}_{(i,j-1)}]) \tag{3.32}$$

The warping path matrix (**W**) is then calculated from the path scores matrix (**P**). This is done by starting at $\mathbf{P}_{(I,J)}$ and finding coordinates of the minimum path back to $\mathbf{P}_{(1,1)}$. Again the algorithm is constrained to only moving left, up, or diagonally left and up by one unit to prevent the warping path doubling back on itself.

$$i = I \quad and \quad j = J \tag{3.33}$$

$$\mathbf{W}_{(1,[1\ 2])} = [i \quad j] \tag{3.34}$$

The new i and j are those which correspond to the location of the minimum of:

$$\mathbf{P}_{(i,j-1)}, \mathbf{P}_{(i-1,j-1)}, and \mathbf{P}_{(i-1,j)} \tag{3.35}$$

$$\mathbf{W}_{(2,[1\ 2])} = [i_{new} \quad j_{new}] \tag{3.36}$$

This process is repeated until $i = 1$ and $j = 1$.

As noted above, there are several constraints included in the algorithm. These include the start and end points of the algorithm being fixed and the step size is constrained to one unit and unable to move backwards. The data that is passed to the algorithm must therefore already have the start and the end times aligned, that is the first and last points in both vectors are aligned to the same features in the data. The data processing techniques that are applied to the data before DTW can have an effect on the warping as they change the features that the algorithm is aligning to.

For example, consider the two batch trajectories shown in figure 3.7. Batch 2 was shorter therefore the last data point was held until the two trajectories were the same length.

Batch 1 was transposed and then both vectors were replicated until a square matrix for each batch was obtained. The difference between the two was then calculated and absolute values taken to remove the negative signs from the data (figure 3.8). It can be seen that where the

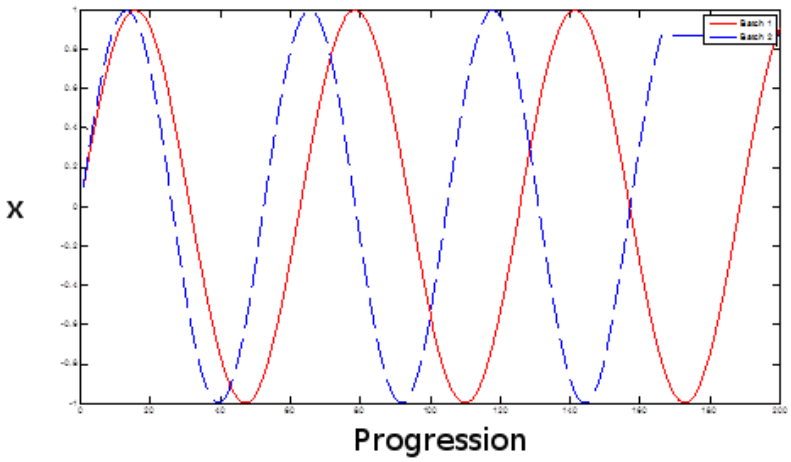


Figure 3.7: Example of simple batch trajectories to be aligned

values of both trajectories are the same, the difference is zero; this is shown in the differences matrix (dark blue). Conversely, the maximum difference that can be achieved here is 2, where one batch has a value of +1 and the other batch has a value of -1.

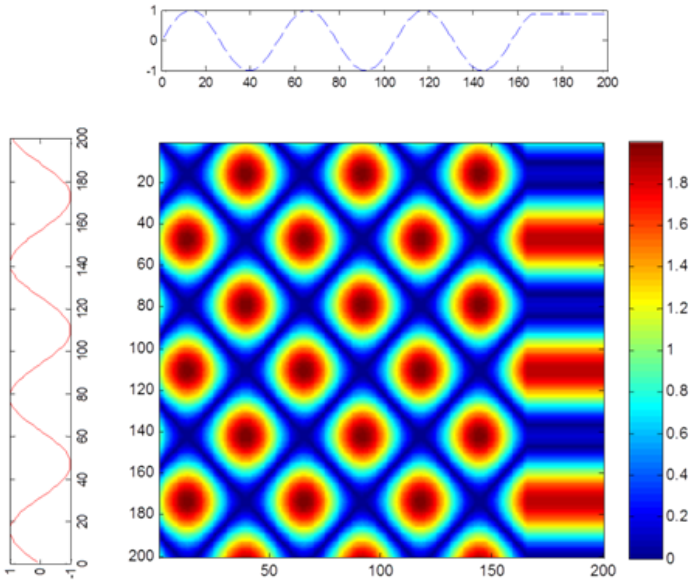


Figure 3.8: Absolute difference matrix for simple example batches 1 and 2

From this the warping path can be calculated (Figure 3.9). It can be seen how the algorithm has weighted the distance from the diagonal (no warping) to attempt to match the features that are relatively close together with each other.

The resultant warping vector can be extracted from this matrix and this can be used to re-scale the time vector of the original data, thus aligning the data (Figure 3.10). Modifications need to

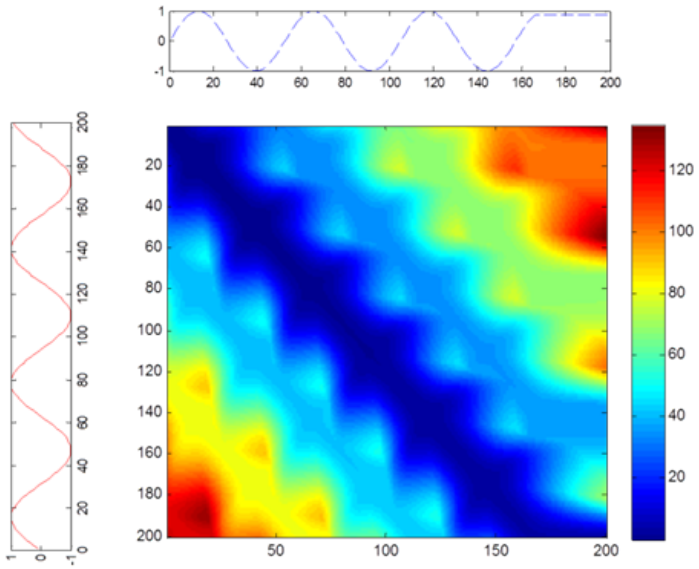


Figure 3.9: DTW warping path matrix for simple example batches 1 and 2

be made to the algorithm or data if the amplitude of the batch trajectory differs from batch to batch as this will lead to sub optimal alignment.

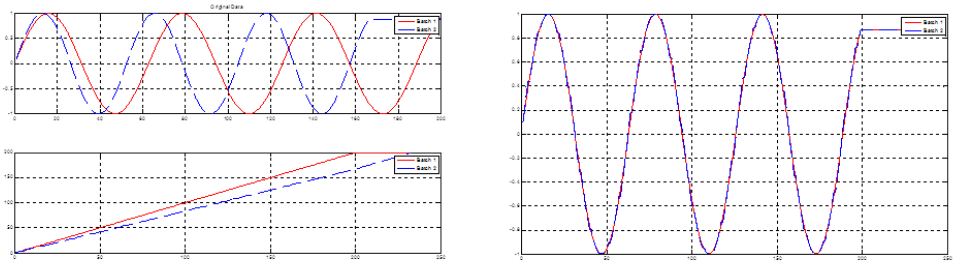


Figure 3.10: Original data (top left), warping vector (bottom left), and warped data (right) using DTW alignment algorithm

3.4 Centring and scaling

Centring of data is performed to remove common offsets within a data set. There are four benefits of centring data; firstly to reduce the rank of the model; secondly to increase the fit to the data; thirdly for specific removal of offsets; and fourthly to avoid numerical problems (Bro and Smilde, 2003). Centring should not be applied without scientific thought to the data however. There may be occasions where an offset in the data is of interest, for example a process in which the offset of a variable changes in different operations, the value of the offset may contain valuable information which the removal of the offset through centring, if

performed independently for each of the operations may remove this information from the model. As centring changes the structure of the model, the use of centring should be evaluated on a case by case basis.

Scaling does not alter the structure of the model and therefore has a less dramatic influence on the model. Nonetheless it may still be an important tool in pre-processing data as it can adjust scale differences and alter the weightings of different variables. It can also adjust for heteroscedastic data by scaling with the inverse of the standard deviation, and allow for different sizes of subsets of data (block scaling) (Bro and Smilde, 2003). Consider, for example, a dryer with pressure data measured in millibar (absolute) and temperature measured in centigrade. When the dryer is operating near ambient pressure, the pressure data with a value of approximately 1000 mbara will have a much greater influence on the model than the temperature measuring between 20 °C and 100 °C. Also, the drying regime at ambient pressure will give the pressure variable more influence than when the dryer is operating at reduced pressure (approximately 50 mbara). Scaling the data by the standard deviation will give all the variables at all time points equal importance in the model.

When handling multi-way data, the major non-linearities in the data can be removed by subtracting the average trajectory from each of the process variables, which is of benefit over traditional PCA where scaling by one mode causes the covariance structure to retain the non-linearities Nomikos and MacGregor (1994).

3.5 Filtering

Filtering can be used as a preprocessing tool to remove mean shifts in data through the use of exponentially weighted moving average (EWMA) filters (Miletic et al., 2004), or for removing some noise from a signal (i.e. increasing the signal to noise ratio) (Brown and Wentzell, 1999).

3.5.1 Infinite impulse response vs finite impulse response

Rani et al. (2011) compares Infinite Impulse Response (IIR) filters to Finite Impulse Response (FIR) filters for the removal of baseline noise (i.e. highpass filter) in electrocardiogram signals. Although both types of filter were able to remove the noise from the signals, the IIR was preferred over the FIR for the following reasons:

- IIR filters are less computationally expensive than FIR filters
- IIR filters are easier to implement
- The transition band on IIR filters is narrower than on FIR filters
- Although IIR filters inherently have phase distortion and delay this can be removed by applying the filter in two directions
- FIR filters of require higher orders and therefore have an increased phase delay

3.5.2 Types of IIR filters

There are many types of IIR filters however three of the most popular are Butterworth, Chebyshev, and Elliptic. The differences between these are discussed in turn below.

Butterworth

Butterworth filters are also known as maximally flat filters. This is because the characteristic of these filters is a flatness in the passband and stopband. A trade off from this, however, is that these filters have a wide transition band and therefore require relatively high orders to narrow this band (Orfanidis, 1996; Proakis and Manolakis, 2007; Rani et al., 2011). With the flat passband and stopband, this type of filter will introduce minimum distortion into the filtered data. Due to the width of the transition however, there is a risk that the filter will distort some of the signal around the frequencies being removed from the data, and that some of the noise frequencies will remain in the data. This type of filter may be suitable for multivariate analysis, however, the filter design should be carefully considered to minimise the risk of distorting the signal through inappropriate selection of cut off frequencies.

Chebyshev type I

By allowing for some oscillations in the passband, Chebyshev Type I filters have a narrower transition band than Butterworth filters, however the maximally flat stopband still causes the transition band to be relatively wide Orfanidis (1996); Proakis and Manolakis (2007); Rani et al. (2011). Depending on the amplitude of these oscillations, significant distortion of the original signal and introduce correlated errors into the data. This is undesirable in multivariate analysis and therefore Chebyshev Type I filters should be avoided.

Chebyshev type II

The Chebyshev Type II filter allows for oscillations in the stop band, whereas the passband is maximally flat. This again allows for a narrower transition band than the Butterworth filter (Orfanidis, 1996; Proakis and Manolakis, 2007; Rani et al., 2011). The flat passband in this filter is more appropriate for multivariate data analysis, however depending on the amplitude of the oscillations in the stopband, there is a risk that with this filter some noise and correlated error will be introduced to the filtered signal. Similar considerations for the cut off frequency as with the Butterworth filter needs to be made for the Chebyshev Type II filter to prevent distortions to the signal around the cut off frequencies. This type of filter may be considered for multivariate analysis if the Butterworth filter transition window width is too difficult to allow for successful noise attenuation.

Elliptic

Elliptic filters have the smallest transition window by allowing for oscillations in both the passband and the stopband (Orfanidis, 1996; Proakis and Manolakis, 2007; Rani et al., 2011). As with both the Chebyshev filters, these oscillations in the passband and stopband may cause distortion of the filtered signal and the introduction of correlated errors which would be undesirable for multivariate analysis. Elliptic filters are therefore not recommended for filtering signals for use in multivariate analysis.

3.5.3 Response type

There are four main response types for filters. The first of these is a lowpass filter which allows the low frequency components of a signal to pass through the filter whilst attenuating the high frequency portion of the signal. The second is the highpass filter. This filter allows high frequency portions of the signal to pass, whilst attenuating the low frequency portion of the signal. The third type is the band pass filter, which allows all frequency components within a band to pass through the filter whilst attenuating frequencies above and below the bandpass cut off frequencies. Finally, the fourth type is the bandstop filter. This filter attenuates the frequency components within the cut off limits, whilst allowing the frequencies above and below the cut off frequencies to pass through the filter. Although an ideal filter would have a clean and immediate cut off at the cut off frequencies, such filters do not exist and filters that

get close to this characteristic typically exhibit oscillations in the passband and/or stopband (also known as ringing). Filter design is therefore a trade off between the transition across from pass to stop, against the stability in the passband and/or stopband (Orfanidis, 1996; Proakis and Manolakis, 2007).

The type of filter response required is application dependent and requires prior knowledge of the process data and which frequencies of the data signal constitute noise and which frequencies should be retained in the signal.

3.5.4 Filter delay

Two types of delay can occur when applying a filter to a signal; constant filter delay where the delay is independent of the frequency, and frequency dependent delay where the delay changes as the frequency changes. This can easily be removed by applying the designed filter in both directions, i.e. forward then backwards across the signal. The result is zero-phase distortion of the filtered signal with a constant delay of zero. This process does however double the order of the designed filter as it is applied across the data twice, and the magnitude of the filter transfer function is squared (Orfanidis, 1996; Proakis and Manolakis, 2007; Rani et al., 2011).

When applying filters to the process data, it is recommended that the filter delay is removed from the filtered signal by applying the filter in both directions to prevent delay distortions from the filter being carried forward into the multivariate analysis.

3.5.5 Cautions on using filters in multivariate analysis

Brown and Wentzell (1999) discuss how applying smoothing filters, specifically polynomial least squares filters (i.e. Savitzky-Golay filters) to multivariate data can impact the subsequent multivariate analysis. In this paper it is recommended that these filters are not applied as a pre-processing tool in multivariate analysis as the side effects of applying these filters to remove noise is distortion of the original signal and the introduction of correlated errors to the filtered signal. Furthermore, the benefit of applying such filters is typically marginal in terms of improving the model predictive performance. Brown and Wentzell (1999) do, however, concede that filtering data might have its place where systems have a large noise component.

Although the arguments presented in Brown and Wentzell (1999) are focused on the lowpass

Savitzky-Golay filters specifically, the arguments presented could be brought into other types of lowpass filters, as no filter is perfect and always has some effect on the signal being processed which will carry into the multivariate analysis. Therefore if filters are to be applied to process data, there must be good justification present for the application of the filter.

Table 3.3 summarises the advantages and disadvantages of the each of the data filtering methods detailed.

Table 3.3: Comparison of data filtering methods

Filtering Method	Advantages	Disadvantages	Suitable for Multivariate Analysis?
FIR Filters		Computationally expensive Difficult to implement Wide transition band High orders required leading to increased phase delay	No
Butterworth (IIR)	Maximally flat Minimum distortion in filtered data	Wide transition band Risk of distorting data around cut-off frequencies	Yes
Chebyshev Type I (IIR)	Narrower transition band than Butterworth	Ripples in passband May introduce significant distortion and correlated errors to the data	No
Chebyshev Type II (IIR)	Narrower transition band than Butterworth	Ripples in stopband May introduce distortion and correlated errors to the data Careful selection of cut-off frequency required	Yes
Elliptic (IIR)	Smallest transition window	Ripples in both the passband and stopband May introduce distortion and correlated errors to the data	No

3.6 Summary

This chapter detailed the pre-processing methods that can be applied within the framework (figure 3.11). The importance of identifying compressed data and determining if the data is suitable for multivariate modelling was first discussed, if the original raw data is no longer available. This highlighted the importance of understanding the data, and how it has been treated prior to obtaining it for analysis, and presented some useful tools to help identify the compression and reconstruction methods used.

Subsequently, methods for dealing with missing data were presented. No one technique is appropriate for all data, therefore, it is again important to understand the history of the data, the level and type of missing data, and what analysis needs to be performed on the data before a technique for dealing with the missing data can be applied. It may be of use to investigate more than one method for dealing with the missing data, and compare results of the reconstruction, and potentially subsequent modelling across the methods selected.

Alignment of process data was next discussed, giving details of a number of tools in varying complexity that can be used to align the data. Again, these methods need to be selected on a case by case basis, and more than one method may be required in the same analysis to achieve alignment of the data.

Different methods for centring and scaling batch process data were subsequently presented and a discussion on the benefits and risks of such procedures. It should also be noted that case should be taken, especially with commercial software packages as these often automatically apply autoscaling to data as part of the PCA operation as a default setting.

Finally, different methods of filtering process data were presented and the risks and benefits of such a practice were discussed. Again if filtering is to be applied, the type of filter to be selected should be done on a case by case basis as each filter has different benefits and consequences, and filtering can easily distort data. If filtering is performed as part of the pre-processing procedure, it would be prudent to perform the analysis with both the filtered and unfiltered data and compare results to test that the filter has not distorted the data.

A summary of these techniques, and a recommended order of operations is presented in the framework shown in figure 3.11. The outlier removal and modelling procedures will be

elaborated on in chapter 4, before this framework is tested on commercial batch process data in chapter 5.

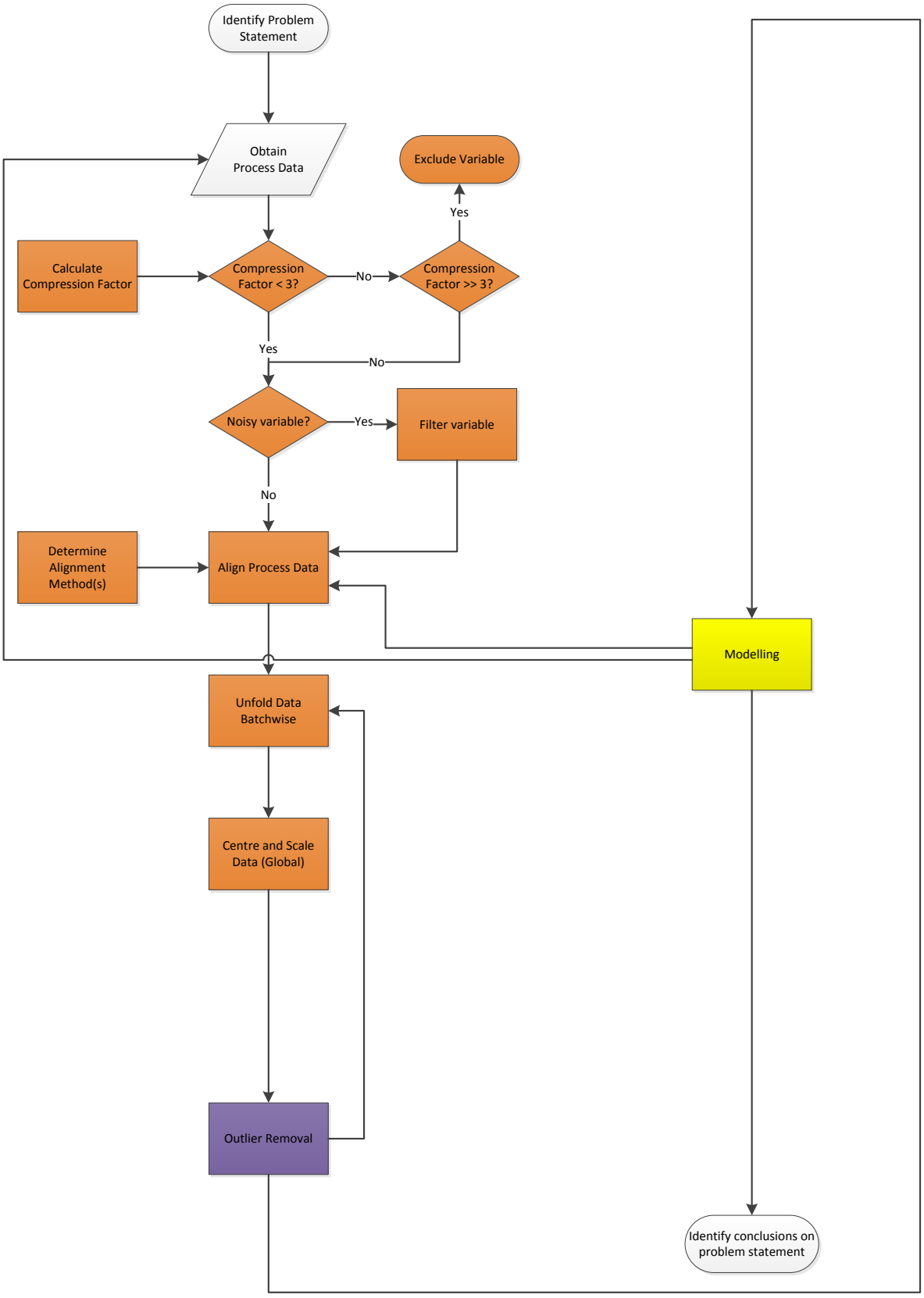


Figure 3.11: Overview of framework to extract multivariate information from batch process data highlighting (yellow) the pre-processing methods discussed in chapter 3

Chapter 4. Principal Component Analysis

The pre-processing techniques that contribute to the framework were discussed in chapter 3. This chapter will build on these and detail the next steps in the framework, namely outlier detection and the multivariate modelling. Principal component analysis is key in both of these steps, and is presented alongside some of the associated statistics. Subsequently, extensions to principal component analysis suited to batch process data are discussed including multiway and dynamic principal component analysis. Finally, some relevant applications of principal component analysis found in the literature are presented.

4.1 Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical projection technique in which the original data is orthogonally and linearly projected onto a space where the variance in the data is maximised. A result of this orthogonal projection is that it has the potential to reduce the number of variables needed to describe a system sufficiently, as the correlation between the variables is removed (Wold et al., 1987). PCA was first used in the field of biological science (Pearson, 1901). Since then it has found a very broad field of applications such as agriculture, biology, chemistry, climatology, demography, ecology, economics, food research, genetics, geology, meteorology, oceanography, psychology, quality control and more (Jolliffe, 2002). Since the advent of relatively powerful computers, PCA has become ever more prominent in a wide research area including chemical engineering. The main draw to PCA is its ability to reduce the dimensionality of highly dimensional data sets. PCA has other uses than just dimensionality reduction as identified in Wold et al. (1987) including problem simplification, data reduction, modelling, outlier detection, variable selection, classification, and prediction.

PCA takes a multivariate data set ($\mathbf{X}$) and finds a series of new variables, principal components, that are orthogonal to each other to describe the variation in the original data set.

Each principal component is a linear combination of the original variables. The first principal component is associated with the plane of most variation in the original multivariate data set. The second principal component, being orthogonal to the first principal component, accounts for the next largest plane of variation in the remaining data set. And so on. The later principal components, that account for small amounts of variation in the original multivariate data set can often be attributed to noise, and therefore excluded reducing the dimensionality of the data.

The application of principal component analysis has been widely reported in the literature in numerous areas. Among these is the application to process data, from both batch and continuous processes. There are several papers on the application of PCA to simulated processes, however the application to process data collected from a real process plant is of more interest as there are different challenges associated with processing of raw process data. The following gives a sample of applications of multivariate techniques to process data published in the literature, and the challenges that they identify.

Pöllänen et al. (2006b) describe the application of PCA based multivariate statistical process control (MSPC) charts to monitor the onset of crystallization using on-line spectroscopic data. The model developed was able to predict the polymorphic form prior to nucleation. Masding and Lennox (2010) go further with this in combining mechanistic process models with process data and multivariate statistical process monitoring techniques.

Burggraeve et al. (2011) describe a methodology for applying Projection to Latent Structures (PLS) and PCA techniques to build a monitoring system for the control of a fluidised bed granulation process, using batch temperature data and PAT tools. Here a PLS model is constructed on normal batches. A PCA model was then built on the scores from the PLS model, which is then used to test if new batches are statistically different from those that were used in the construction of the original models. Burggraeve et al. (2011) highlight the challenge of finding appropriate instrument positions to both obtain representative process measurements and prevent instrument fouling. De Beer et al. (2009) also comment on this challenge of obtaining representative samples throughout a batch, in the application of building a monitoring system for a lyophilization process using PAT and MSPC techniques.

Further work by Sarraguça et al. (2010) notes that it is also important to monitor all of the critical process aspects. This is achieved on the lyophilization process through the use of multiple spectroscopic technologies. A model is subsequently constructed from a set of

batches in 'normal operating conditions' which can then be applied to new batches to detect deviations from the normal batch trajectories. This is further commented on by Camacho et al. (2015) for the monitoring of fed-batch fermentation processes where it is difficult to measure primary quality variables.

Camacho et al. (2008) outline the application of multi-phase analysis, an extension to PCA to account for dynamic model structures, to the simulated data for the cultivation of yeast (*Saccharomyces cerevisiae*) and waste water treatment data collected from a laboratory sequential batch reactor. The study successfully developed a model that was capable of predicting the final phosphorus content of a batch from the first 85 sampling times (the anaerobic processing stage).

García-Muñoz et al. (2003) detail a study of an industrial drying process to which both PCA and PLS techniques were applied. García-Muñoz et al. (2003) also highlight the issue with process data alignment. The data given in the paper is misaligned due to different starting conditions of batches, and some of the operating conditions are different for each batch. This results in the data becoming misaligned within processing stages and an overall difference in batch length. García-Muñoz et al. (2003) did not have an appropriate variable (monotonically changing and starting and ending at the same value for each batch) to use as an indicator variable for re-interpolation. It was, however, possible to split the batch trajectory into three phases of which the first two could be aligned using various indicator variables within their respective phase. The third phase did not have an appropriate indicator variable available therefore this phase was re-sampled linearly to a constant length. Using this warped dataset, and incorporating warped time as a new variable, García-Muñoz et al. (2003) were able to show that the initial quality of the charged material had little impact on the final product quality, whereas the operating conditions, holding times, temperatures and pressures, were responsible for the quality of the final product.

4.1.1 Mathematical description of PCA

Taking $\mathbf{X}$ as the original data matrix the principal components of a system are the eigenvectors of the covariance matrix of $\mathbf{X}$. Thus, for each principal component (j) the scores ($\mathbf{T}$) can be described as a linear relationship to $\mathbf{X}$ by the loadings ($\mathbf{P}$).

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{P}_j \tag{4.1}$$

The loadings and scores for each principal component can be collected into two matrices **T** and **P** respectively. The loadings of the principal components are the eigenvectors of the covariance matrix of **X**. These describe the linear combination of the original variables required to map the data onto the principal component hyperspace. The loadings are therefore useful for interpreting the patterns observed in the scores plots and identifying variables that have significant contributions to the observations made from the scores plots.

The scores can be calculated from equation 4.2. These describe the relationship between the observations in the data matrix on the new principal component hyperspace. They are useful for observing patterns in the data and splitting the data into groups. For example one group of observations (or batches in multi-way PCA) that exhibit one behaviour, and another group of batches that exhibit a different behaviour.

$$\mathbf{t} = \mathbf{X} \cdot \mathbf{p} \tag{4.2}$$

There are a number of methods for calculating the principal components including Non-linear Iterative Partial Least Squares (NIPALS) and Singular Value Decomposition (SVD) Jolliffe (2002). The difference between the two methods is the NIPALS algorithm is an iterative process calculating the principal components sequentially, whereas the SVD algorithm calculates all of the principal components simultaneously.

The simplicity of the NIPALS algorithm and the ability to handle moderate amounts of randomly distributed missing data lends itself to integration in computer programming. The NIPALS algorithm is first introduced in section 3.2.2 and more details on the algorithm are detailed here. The procedure for calculating the principal components is as follows (Wold et al., 1987).

For each dimension of a scaled matrix **X**:

1. Set the scores vector (**t**) to the column in **X** with the largest variance.
2. Calculate the loading vector $\mathbf{p}^T = \mathbf{t}^T\mathbf{X} / \mathbf{t}^T\mathbf{t}$. The elements in **p** can be interpreted as the gradient in the linear regressions between **t** and the corresponding column in **X**.
3. Repeat the process until convergence between all the elements in two consecutive score vectors (within a pre-defined tolerance).
4. Calculate the residuals (**E**) from $\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T$.

5. Use **E** as **X** in the calculation of the next dimension. Continue until all of the required dimensions have been calculated.

4.1.2 Selection of number of principal components

The first principal component is the hyperplane to which the data with the largest variance has been projected. Subsequent principal components capture diminishing quantities of the variance within the data set until 100% of the variance has been explained. In most data sets, especially those originating from industrial scale process engineering applications, not all of the variance in the data is of interest, as some of that variance (often significant quantities) is related to noise in the data. It is therefore not appropriate to use the most principal components available to model the data set as the information of interest will be contained in the first few principal components. Including more principal components than required to a PCA model can even be detrimental to the sensitivity of the model. A methodology is required to determine which principal components should be retained in the model, and which principal components add little value to the model and can therefore be excluded.

There are a number of methodologies that have been proposed to select the number of principal components that should be included such as methods that consider the cumulative percentage of variance explained, minimum eigenvalues for subsequent principal components, and many cross-validation based methods (Valle et al., 1999).

Jolliffe (2002) suggests the cumulative variance captured by the model should be between 0.7 and 0.9. However, if the level of noise in the data is less than 10% this would result in some information being excluded from the model. This is unlikely to be a problem with a data set from an industrial process, however the problem that the noise in the data may contribute to more that 30% of the variance leading to noise being included in the model reducing its sensitivity. For example, consider the PCA model described in table 4.1 using this method five or more principal components would be retained.

Another method is to include only those principal components with eigenvalues greater than 1, known as the Kaiser-Guttman rule (Jackson, 1993). An eigenvalue of less than 1 indicates that there is less information in that principal component than in one original variable. Using this metric may however cause a loss of information if a principal component with an eigenvalue slightly smaller than 1 is excluded from the model when it contains non-noise information.

Table 4.1: Variance captured in PCA model for dryer process dataset

PC Number	Eigenvalue of Cov(x)	% Variance captured this PC	% Variance captured total
1	1660	27.73	27.73
2	969	16.14	43.87
3	937	15.60	59.48
4	482	8.03	67.50
5	458	7.64	75.14
6	366	6.09	81.23

A scree plot is another method to determine the number of principal components required (Jackson, 1993). It is another method relying on the eigenvalues of the covariance matrix of the original data. In a scree plot, the eigenvalues are plotted against the number of principal components in the model as a decreasing curve. The 'elbow' in the scree plot, i.e. the point at which the curve levels off is taken as the number of principal components required. There may however be more than one 'elbow' in the plot leaving some ambiguity as to how many principal components to retain. Again this method suffers the same potential problem as using the value of the eigenvalue, as principal components that contain useful information may be excluded as the describe significantly less variability than the principal components computed before them. The example scree plot in figure 4.1 indicates that three principal components should be retained for this model as this is the first point that does not fall on the level part of the scree plot moving from the right.

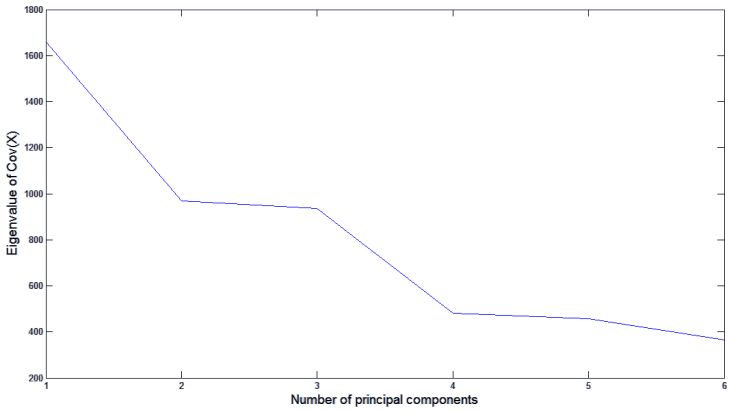


Figure 4.1: Scree plot for example PCA model on NIR methanol-acetone-water dataset used to find the number of principal components to retain

Cross-validation is a method in which samples (rows) in the data matrix ($\mathbf{X}$) are selectively

removed once and a PCA model is constructed with the remaining samples using 1 principal component (Giancarlo and Chiara, 2002). The removed sample data is then predicted using the model and the predicted Sum of Squared Errors (SSE) is calculated. This is repeated for the remaining samples (rows) in $\mathbf{X}$ that have not been removed until every sample has been removed once. The predicted sum of squared errors is then summed to give the total predicted SSE for 1 principal component. This is then repeated for increasing numbers of principal components. The optimal number of principal components is the number of principal components that give the lowest predicted Sum of Squared Errors. The example shown in figure 4.2 shows that six principal component should be retained in the model.

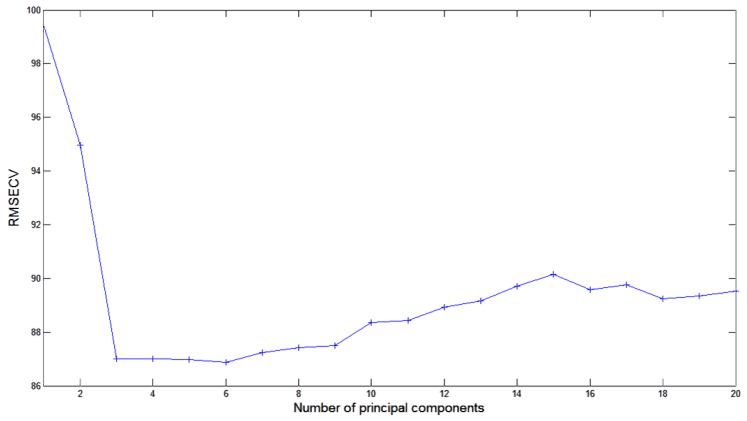


Figure 4.2: Cross-validation plot for example PCA model on NIR methanol-acetone-water dataset used to find the number of principal components to retain

In summary, there is no one method of selecting the number of principal components that is better than the others. The most appropriate method will change between data sets, and therefore the most appropriate methods need to be selected on a case by case basis.

4.1.3 Model quality indicators

Hotelling's T^2 statistic

The Hotelling's T^2 statistic is a measure of variation within the model. That is when applied to the scores, it provides a statistic for each sample describing how similar that sample is to the other samples in the model. Taking $\mathbf{t}_i$ to be the ith row of the scores matrix $\mathbf{T}_k$ with k scores vectors, and S is the diagonal matrix of the eigenvalues the Hotelling's T^2 can be described as follows (Simoglou et al., 2005; Kourti, 2005).

$$T_i^2 = t_i S^{-1} t_i^T \tag{4.3}$$

Q residuals

The Q-statistic is a lack of model fit statistic. It is a measure of how well each sample fits the model and considers variation not captured by the model. Taking $\mathbf{e}_i$ to be the ith row of the residuals matrix $\mathbf{E}$, $\mathbf{P}_k$ to be the first k columns of the loadings in $\mathbf{P}$ and the identity matrix $\mathbf{I}$, the Q statistic on the ith sample can be expressed as:

$$Q_i = \mathbf{e}_i \mathbf{e}_i^T = \mathbf{x}_i (\mathbf{I} - \mathbf{P}_k \mathbf{P}_k^T) \mathbf{x}_i^T \tag{4.4}$$

4.1.4 Contribution plots

Following the identification of a sample, or group of samples, of interest using the PCA scores, Hotelling's T^2 statistic, or the Q-residual statistic for example, it is desirable to know which original variables are responsible for those scores and make the sample(s) of interest.

Taking a data matrix $\mathbf{X}$ of n $\times$ m process variables and samples respectively. The scores for the k^{th} observation and the i^{th} principal component can be expressed as:

$$t_{k,i} = \sum_{j=1}^{m} x_{k,j} p_{j,i} \tag{4.5}$$

Where $p_{i,j}$ is the loading for the j^{th} variable on the i^{th} principal component. This can then be decomposed into the m contributions for each individual process variable (Conlin et al., 2000; Flores-Cerrillo and MacGregor, 2004). If more than one observation is of interest, the contributions can be summed for all of the observations of interest. For example, for the k^{th} observation to the K^{th} observation the score for the i^{th} principal component from the j^{th} variable can be given by:

$$c_{k:K,i} = \sum_{k=k}^{K} x_{kt,j} p_{j,i} \tag{4.6}$$

4.1.5 A visual example of PCA

Consider the case of Near Infra-Red (NIR) spectroscopic data for a three component system of methanol-acetone-water generated as part of a vibrational spectroscopic methods training module as part of doctoral training programme. Several spectra were obtained at different concentrations as shown in table 4.2. The NIR spectra obtained in transflectance mode for these samples are shown in figure 4.3.There are several things to note on the spectra. The first is the absorbance saturation around wavenumbers 5500 to 5400 cm^{-1} (point 1). This is due to the path length being too long for this system and therefore not allowing enough light of these wavenumbers to be passed back to the instrument to be detected resulting in a flat topped peak that is noisy and off the scale of the instrument.

Another feature of the spectra is shown in points 2 and 3. These indicate the spectra that are flatter either because the spectra were collected before the NIR probe entered the sample, or there was a bubble of air in the probe window (light path). The result is that the absorbance across the wavelengths is decreased due to the reduction in the amount of sample that the light has to pass through.

Table 4.2: Compositions of methanol-acetone-water samples for NIR spectroscopy

Sample	% Methanol	% Acetone	% Water
1	5.0	0.0	95.0
2	0.0	5.0	95.0
3	0.0	0.0	100.0
4	2.5	2.5	95.0
5	2.5	0.0	97.5
6	0.0	2.5	97.5

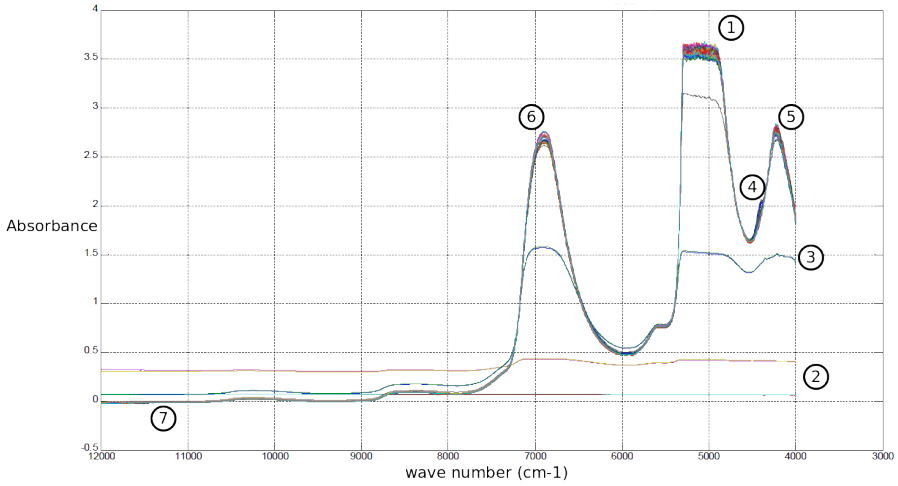


Figure 4.3: NIR spectra of methanol-acetone-water samples at various concentrations. Point 1 indicates detector saturation. Points 2 and 3 indicate effect of air on the spectra. Points 4 - 6 indicate where there are relatively large differences between samples in the collected spectra relating to the different component concentrations. Point 7 indicates a part of the spectra that changes little with changes in concentration.

As PCA is a dimensionality reduction technique based on the variability in the data it can be used to reduce the spectra into a small number of components required to determine the quantity of each of the three components in the samples. It does this by preferentially selecting variables (wavenumbers in this instance) that show large differences between samples (i.e. have high variance in the data set) such as those indicated by points 4, 5, and 6. Conversely PCA does not select variables that do not change between the samples by more than the noise in the data. An example of this is shown at point 7 where there is no absorbance by any of the three components and therefore the spectra does not change as the samples change.

If PCA stopped here however, there would still be a high dimensionality in the resulting data as there are a large number of wavenumbers that absorb light for some of the components but not all and therefore have high variance in the data set. The dimensionality reduction power comes from the use of the covariances. For example, consider the spectral data at point 6 where there is a high variance in the data (i.e. the spectra are able to discriminate between the quantities of the components in the sample). As the concentration of one of the sample components increases the absorbance at wavenumber 6900 cm^{-1} increases. So does however, the absorbance at wavenumber 6901 cm^{-1}, and 6902 cm^{-1} and so on. These wavenumbers have a high covariance in the dataset and indicate the same property of the system. PCA is able to combine these into one new variable (principal component) that summarises one characteristic

of the variance of the system. Conversely, as this is a three component system with components that absorb at different wavenumbers, where the concentration of one component (say methanol) increases but another (say acetone) is held constant, the wavenumbers at which methanol absorbs the light will indicate increasing absorbance, whereas the wavenumbers at which acetone absorbs will remain constant. These wavenumbers will have low covariance and therefore they will end up in separate components in the PCA model.

Applying PCA to the spectral data set for the three component mixtures will be able to reduce the data set from 2074 wavenumbers (variables) to just 3 principal components. From these three principal components, however, it will still be possible to determine the concentration of each of the components in the samples. Additional principal components will hold little information about the system (perhaps some temperature information, or other sources of systematic error) and will mostly comprise of noise.

Figure 4.4 shows the scores on principal components 1 through 4 inclusive for the methanol-acetone-water samples. In principal component 1, 2, and 3, the scores appear significantly different for the six samples where air is present and the absorbance is reduced as a result. These spectra would also be pulled out in the Hotelling's T^2 statistic (not shown). No such identification can be made from the scores on principal component 4 as this component contains little information form the system and is composed of mostly noise.

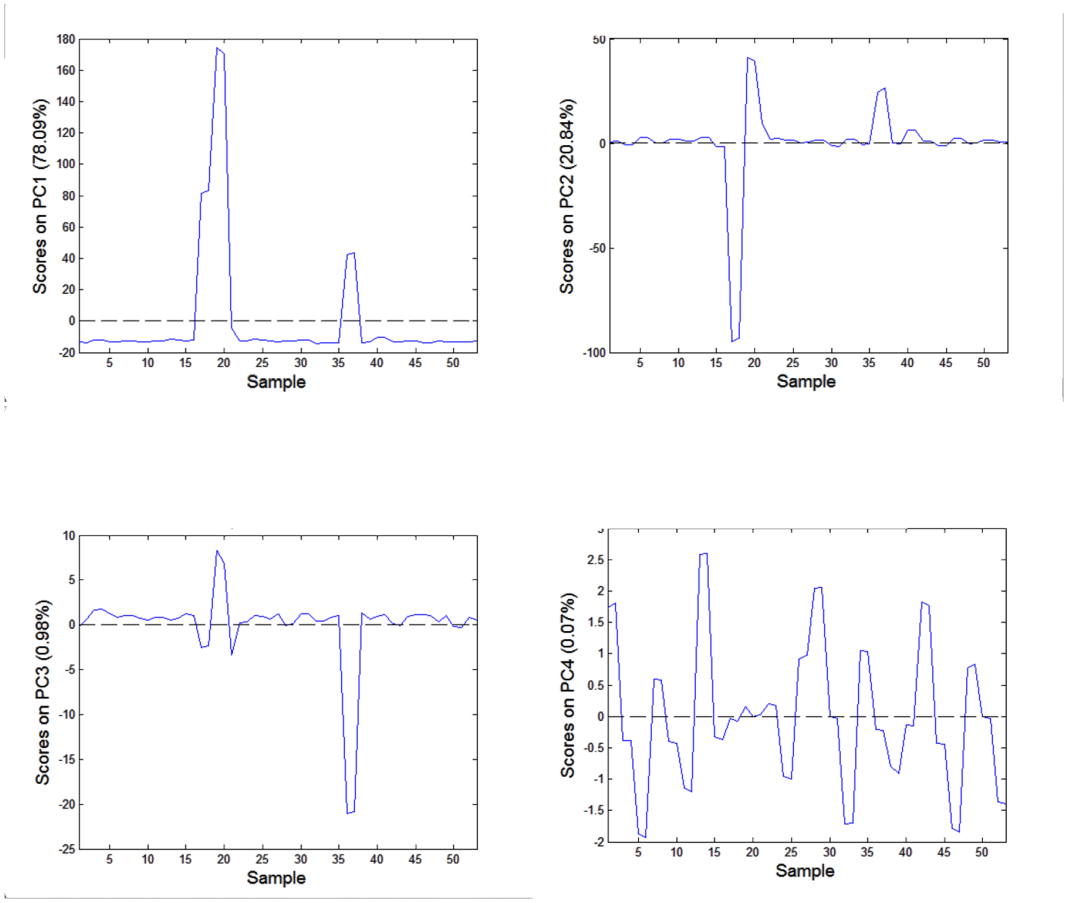


Figure 4.4: PCA scores on all methanol-acetone-water spectra on principal component 1 (top left, 78.09% variance captured), principal component 2 (top right, 20.84% variance captured), principal component 3 (bottom left, 0.98% variance captured), and principal component 4 (bottom right, 0.07% variance captured).

The presence of air in the transflectance NIR samples causes a baseline reduction in the absorbance. This is because there is a smaller quantity of sample material for the NIR light to pass through, thus, the samples for which an air bubble was trapped in the NIR probe window can be classed as outliers and removed. Following removal of the spectra containing air, a PCA model was built on the remaining data. Figure 4.5 shows the loadings on principal components 1 through 4 inclusive against wave number. Principal components 2 and 3 have captured the absorbance peaks for methanol and acetone, although confounded by the peaks attributed to water (see figure 4.6 for the pure methanol and pure acetone NIR spectra). The main variables in principal component 4 are the wave numbers that were saturated on the instrument and therefore the information contained in this principal component is mainly noise.

Figure 4.7 shows the scores on both principal component 1 and 3. This enables the data to be

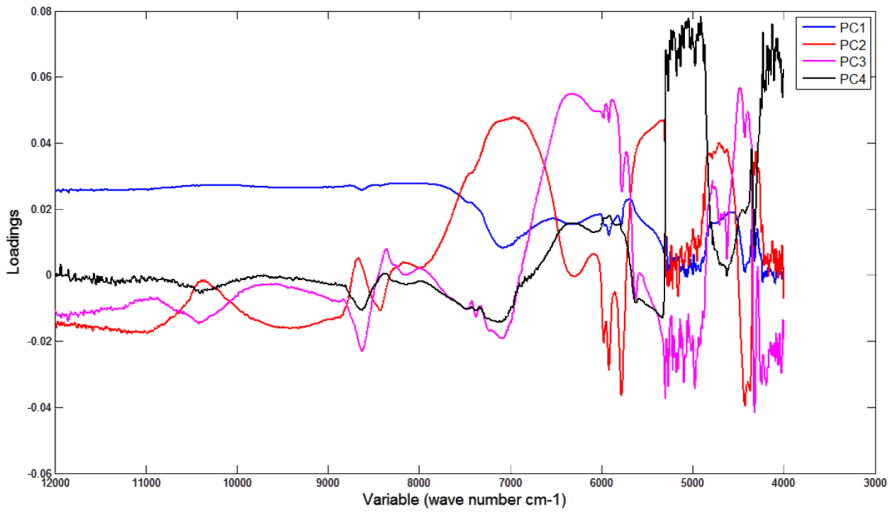


Figure 4.5: PCA loadings on principal component 1 (62.44% variance captured) (blue), principal component 2 (17.68% variance captured) (red), principal component 3 (10.62% variance captured) (magenta), and principal component 4 (5.56% variance captured) (black) plotted against wave number.

grouped into those samples with low concentrations of acetone (0%), those with high concentration of acetone(5%), and the others (2.5% acetone). A similar analysis can be performed for methanol and for water (not shown). PCA has been able to summarise the spectra for the samples collected and pull out the wavenumbers (variables) with the largest variance and condense the data set into three principal components that are able to describe the samples sufficiently to be able to determine the composition of each sample.

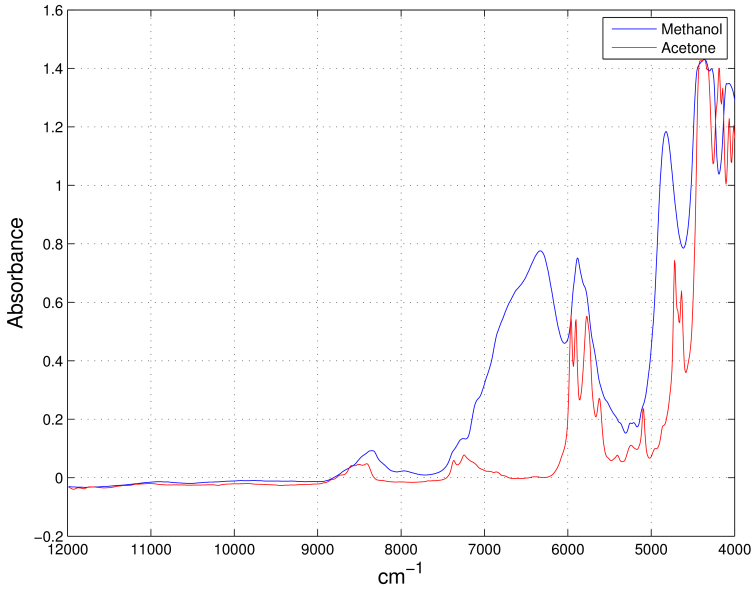


Figure 4.6: NIR absorbance spectra of samples of pure methanol (blue) and acetone (red)

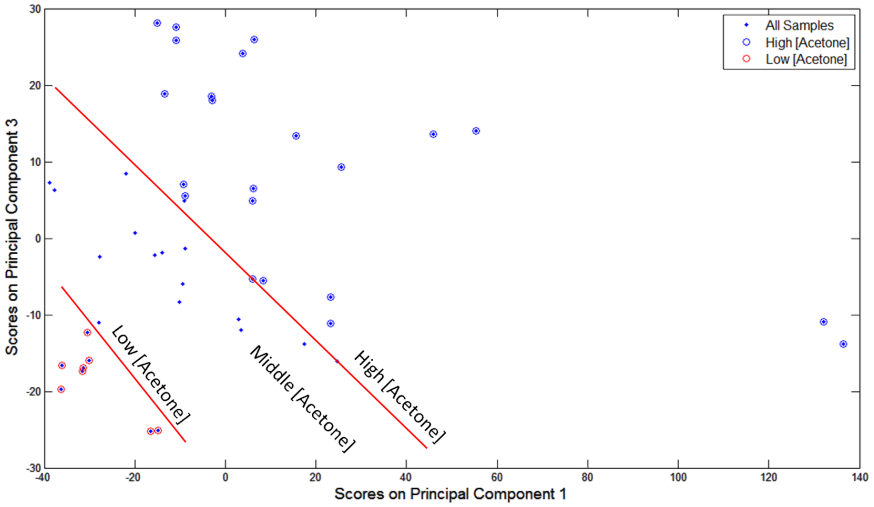


Figure 4.7: Scores on Principal Component 1 vs Scores on Principal Component 3 for methanol-acetone-water system grouped by acetone concentration. Blue points show all data. Blue circles indicate samples with high acetone concentration. Red circles indicate samples with low acetone concentration.

4.2 Multi-way principal component analysis

Multi-way PCA (MPCA) is an extension to PCA allowing for multi-dimensional datasets (Nomikos and MacGregor, 1994). Traditional PCA takes a two-dimensional data matrix $\mathbf{X}$ comprising of J variables by K samples. With batch process data however a third dimension is introduced, the batch. The resulting data matrix $\mathbf{X}$ for batch process data therefore consists of I batches, J variables, and K time points (or samples) as shown in figure 4.8.

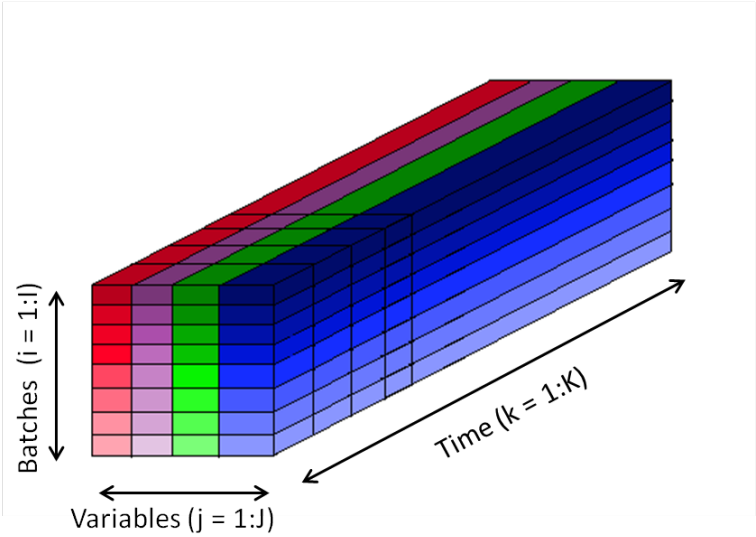


Figure 4.8: Three dimensional data matrix for batch process data

4.2.1 Unfolding

There are six possible ways to unfold a three-dimensional array to two-dimensions. Considering the three-dimensional data array $\mathbf{X}(I \times J \times K)$ the data can be unfolded as follows:

1. $\mathbf{X}(I \times JK)$
2. $\mathbf{X}(KI \times J)$
3. $\mathbf{X}(K \times IJ)$
4. $\mathbf{X}(I \times KJ)$
5. $\mathbf{X}(IK \times J)$
6. $\mathbf{X}(JI \times K)$

Matrices 1 and 4, 2 and 5, and 3 and 6 are equivalent and therefore only three methods will be discussed further. Depending on the way the three-dimensional data matrix is unfolded, different variability is analysed in the PCA model.

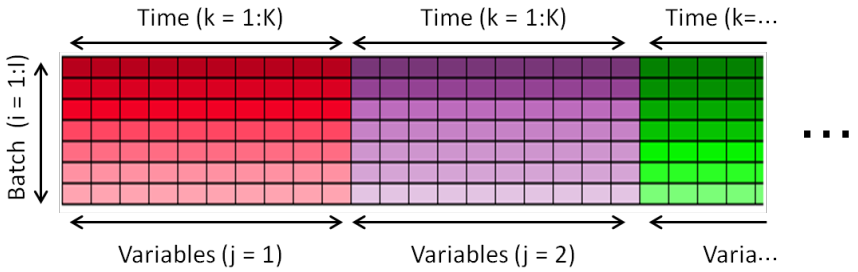


Figure 4.9: Unfolding **X** (IxJK)

Figure 4.9 shows the I $\times$ JK unfolding method (time mode, Angulo and Godo (2007)) . The resulting matrix consists of columns containing the process variables for each time point over all batches. The application of PCA to the data unfolded this way is used to analyse the batch-to-batch variation at each time point for each process variable. This lends itself to identification of batches that are different at a specific time point.

In order to unfold the data array in the time mode, it is important that the batch data is aligned and of the same length so the same time point is compared in each batch.

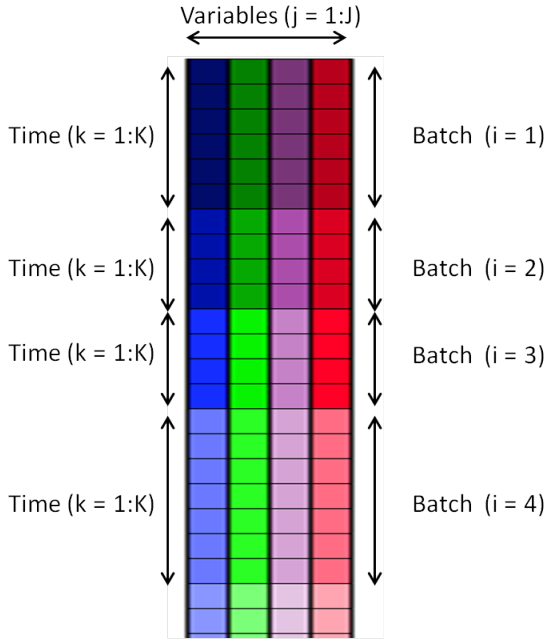


Figure 4.10: Unfolding **X** (IKxJ)

Figure 4.10 shows the IK $\times$ J unfolding method (batch mode) (Wold et al., 1987). Each row in the unfolded array consists of the data for an individual batch at an individual time point for all variables. The application of PCA to the data unfolded this way is used to analyse the variability between batches over the entire batch trajectory.

Unlike the I $\times$ JK unfolding method, the IK $\times$ J unfolding method does not require the batches

to be of the same length, and is therefore useful for on-line process monitoring. A large variation in the data however comes from the time dependant variation inherent with batch processes. This results in this time dependant structure being resolved in the first few principal components, thus pushing the non-time dependant variability into the later PCs and therefore requiring a greater number of PCs to be retained.

4.2.2 Outlier Detection

An important use of the batchwise unfolding method is outlier detection (Nomikos, 1996; Dahl et al., 1999; Petersen et al., 2008; Burggraeve et al., 2011; Ben Yahia et al., 2016). In this method, each of the scores represents a batch, and therefore any outlying batches will have either high Hotelling's T^2 statistic or Q-statistic. A high Hotelling's T^2 statistic indicates that the data for the batch is dissimilar to the other batches included in the model, whereas a high Q-statistic indicates that the batch does not fit the model based on variation remaining in the residuals.

The contribution plots for the respective statistic can be used to identify the variable and time in the batch that caused it to be flagged as an outlier. The batch data should then be interrogated to identify a cause for the different behaviour and considered for removal of the dataset. After each batch is removed from the dataset, the model should be reconstructed without this data and tested for further outliers. Similarly, only one outlier should be removed at a time, as a batch may be flagged as a false positive outlier due to the influence of the true outlier on the model.

By using these techniques following the pre-processing steps, previously discussed in chapter 3, the outlier removal procedure in the framework can be expanded as shown in figure 4.11.

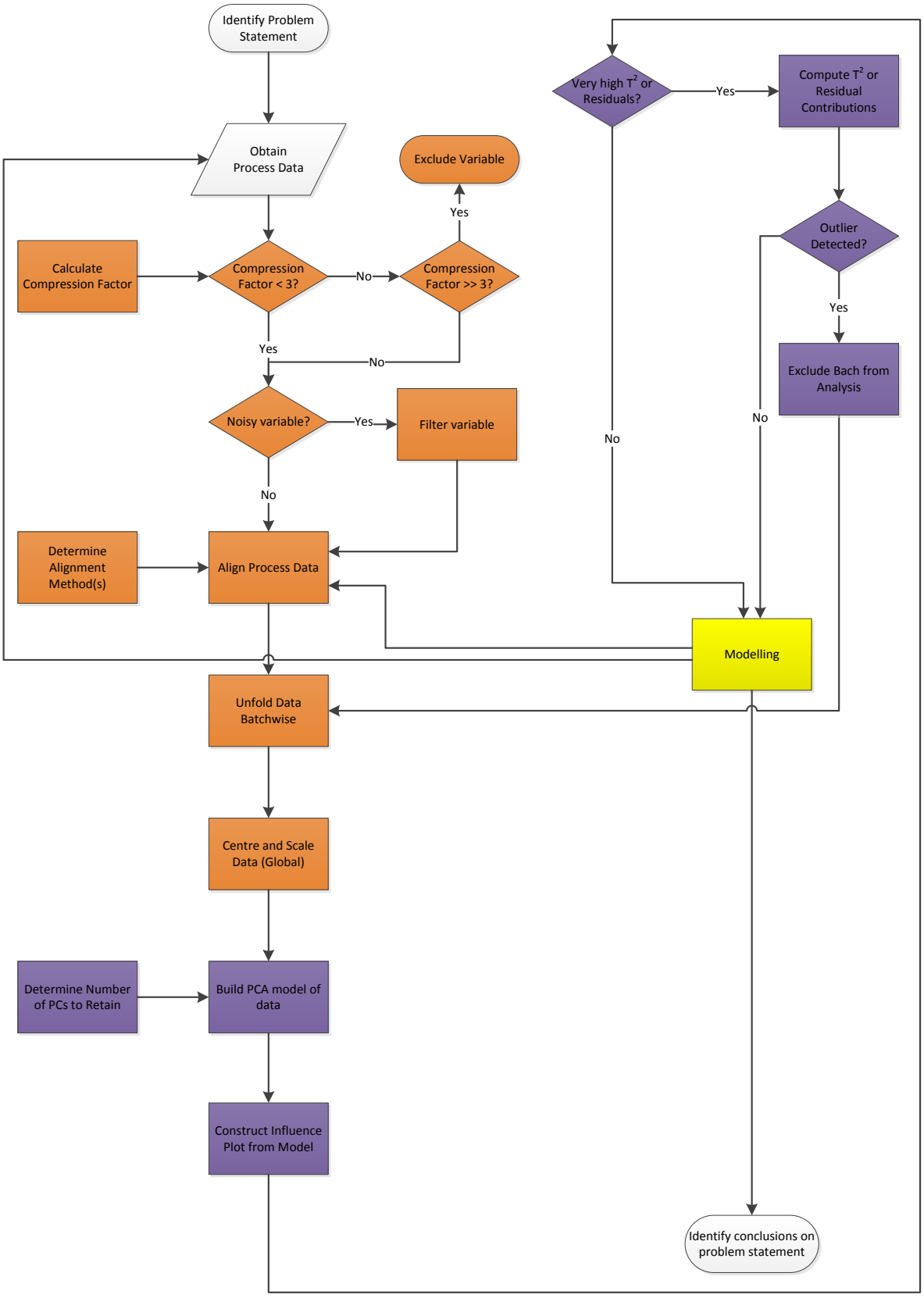


Figure 4.11: Overview of framework to extract multivariate information from batch process data expanding (purple) the outlier detection methods discussed in chapter 4

4.3 Dynamic principal component analysis

An extension to multi-way principal component analysis that deals with dynamic process data is dynamic principal component analysis. Batch processes are inherently dynamic and thus DPCA may be highly relevant for batch process data. Progressive time points in a large number of process variables are correlated with their previous values (auto-correlated). For example, considering a temperature variable during a heating process of a batch. The value of the temperature at time t will be related to the temperature value at time $t-1$ as the temperature dynamics are sufficiently slow to cause the temperature to slowly increase. Even measurements on process variables with much faster dynamics such as pressure or reaction conversion will still have some element of autocorrelation within the data given sampling time that is sufficient to capture these dynamics.

Static PCA fits a linear static model through the data provided to the PCA algorithm. Applying static PCA to data where dynamic information is present results in a linear static approximation of the dynamic information. Although the static PCA methods have been shown to detect and isolate disturbances from a dynamic process (Tennessee Eastman Simulation) due to the violation of the statistical assumptions required for static PCA, namely the assumption of time independence, the resultant principal component scores will be auto-correlated, and possibly cross-correlated. The results of such an approach could be misleading and lead to false alarms being detected in process monitoring especially for small disturbances (Ku et al., 1995; **?**). The current values of a dynamic process will depend on at least the previous value if not more than one previous values. This information should therefore be included in the PCA model as it is constructed to obtain valid results for a dynamic process.

An idea borrowed from ARMAX (auto-regressive moving average exogenous inputs) time series models is to incorporate the previous observations (lags) in each observation vector into the data matrix (Chen and Liu, 2002; Flores-Cerrillo and MacGregor, 2004). Merging the ARMAX model with PCA is known as Dynamic principal component analysis (DPCA) and allows for the time-dependant relations in measurements to be extracted into the principal components. Where the time dependant relationships are seen through autocorrelation within a variable, by augmenting the data matrix with lagged variables this autocorrelation becomes transposed and observable within the variables in the data matrix.

To apply DPCA, the correct number of lags to be included in the augmented data matrix X needs to be determined. Typically this is one or two for linear systems, however, for non-linear

systems more lags may need to be incorporated to obtain a better linear approximation of the non-linearity. To determine the number of lags required for each variable to augment the data matrix with the autocorrelation function can be used. Autocorrelation is the correlation of a data series with its own past values. The sample autocorrelation function can be described as:

$$r_1 = \frac{\sum_{k=1}^{m-l}(x_i - \bar{x})(x_{i+l} - \bar{x})}{\sum_{k=1}^{m}(x_i - \bar{x})^2} \tag{4.7}$$

where l is the appropriate time lag, x_i is a random variable and $\bar{x}$ is the sample mean. The value of the autocorrelation will be a value between zero and 1, where zero indicates no correlation (i.e. completely random data) and 1 indicates 100% correlation.

The sample autocorrelation cannot be solely relied upon to give the number of lags as it will only give the number of lags for a moving average system model. If the system exhibits auto-regressive behaviour the partial-autocorrelation function will also be required to accurately identify the number of lags required for the system. The partial autocorrelation is the correlation between x_i and x_{i+l} following the removal of their correlation on the variables between them (i.e. $x_i, \quad \ldots \quad, x_{i+l-1}$).

Ku et al. (1995) present an algorithm for determining the number of lags required for a dynamic system:

1. Set $l = 0$
2. Form $\mathbf{X} = [\mathbf{X_{(k)}} \quad \mathbf{X_{(k-1)}} \quad \ldots \quad \mathbf{X_{(k-l)}}]$
3. Perform PCA to calculate the PC Scores
4. Set $j = n \times (l+1)$ and $r_{(l)} = 0$
5. If j^{th} component represent linear relation go to 6, else, go to 7.
6. Set $j = j - 1$ and $r_{(l)} = r_{(l)} + 1$. Go to 5
7. Calculate the number of new relations (r_{new}):
   $r_{new} = r(l) - \sum_{i=0}^{l-1}(l - i + 1) r_{new}(i)$
8. If $r(new) \leq 0$ go to 9, else, go to 10.
9. Set $l = l + 1$ and go to 2.
10. Stop.

4.4 Partial Least Squares

Partial Least Squares or Projection on Latent Structures (PLS) regression is a technique combining features of PCA and multiple linear regression. The aim it to take a multivariate data set, and by reducing the dimensionality to latent variables which can be linearly regressed to predict one or more dependant variables. Whereas PCA decomposes in input data matrix **X** to obtain principal components that best explain the variance in **X**, PLS decomposes the input data matrix **X** with the response matrix **Y** to obtain latent variables that best predict **Y** from **X** (Kourti, 2005; Abdi and Williams, 2010).

As the aim of the PLS algorithm is to provide a good description of **X** and **Y** and the correlation between them, the output projection in PLS differ from those obtained with PCA. There are some examples in the literature on the use of PLS in modelling the quality attributes of pharmaceutical products however the predictive ability of PLS was found to be low for non-linear systems (Chen and Liu, 2002). Other examples in the literature show how PLS can be useful in determining the most important factors influencing a product quality attribute (Gabrielsson et al., 2002).

4.5 Application of PCA to process data

There are a small number of publications relating to the application of PCA and PLS to industrial applications, and as Gabrielsson et al. (2002) and García-Muñoz et al. (2003) indicate, even fewer on applications to pharmaceutical processes. In 1995, Lindberg and Lundstedt (1995) discuss the application of PCA to understanding the impact of particle size on dissolution of a poorly soluble drug product. Following a comprehensive analysis a relationship between PSD and dissolution was found in addition to effects from impurities, process parameters, and surface area (BET) analysis.

Patel and Podczeck (1996) present PCA analysis looking at the impact of microcrystalline cellulose variability on a capsule filling process, however only the scores are presented in this analysis. Similarly, Jover et al. (1996) fail to present the loadings from their PCA analysis on a tablet compression process from extruded pellets. Another application of PCA to a pellet to tablet process was presented in Pinto et al. (1997), however in this case the scores were not

presented. In 1998, Rotthäuser et al. (1998) discuss how PCA has been applied to the formulation development of an effervescent tablet.

Lennox et al. (2001) present an industrial case study on the application of PCA to a fed-batch fermentation process, in which details on pre-treatment including low pass filtering to noise reduction, and detection of outliers were performed. Also included is a discussion of missing data and how this can impact PCA, followed by details of an on-line monitoring scheme using PCA compared with PLS and artificial neural networks, to the fermentation process. Marjanovic et al. (2004) present another case study on a fermentation process investigating fault detection and the differences between the unfolding methods applied to two penicillin fermentation case studies and some simulated data. Another fermentation case study is presented by Ben Yahia et al. (2016) for the fed-batch Chinese Hamster Ovary (CHO) cell fermentation for protein production. Here PCA is used to identify outliers using the Hotelling's T^2 statistic during the initial stages of the fermentation as a pre-processing technique for an alternative modelling tool.

García-Muñoz et al. (2003) later describe how multi-block PLS (MacGregor et al., 1994) can be applied to a drying process where the input blocks are wet chemistry data from the wet cake feeding the dryer, and the dryer process data, and the output block is the data measured from the end of drying. This application includes a detailed discussion on how alignment can be achieved in batch process data, and shows that segmented indicator variables were used in this application. Burggraeve et al. (2011) present a fluid bed granulation case study where cross validation was applied to identify the number of principal components to retain. The case study presents the analysis as an online monitoring tool used for outlier and abnormal event detection using scores and Hotelling's T^2 statistic. The PLS model also presented was able to predict the quality attributes of the batches. Finally, Sun et al. (2017) present how multi-block PLS was applied to a paracetamol tablet dissolution problem to pull out the critical process parameters and material attributes that significantly influence the tablet dissolution in the batch wet granulation and compression processes. Furthermore, Wong et al. (2008) describe a method for incorporating spectroscopic data and process data in a multi-block approach to improve the process monitoring.

In addition to the pharmaceutical applications there are also a handful of other industrial case studies from batch processes in the literature. Initially most of these were around polymerization processes, for example, Kosanovich et al. (1994) discuss autoscaling the data as a pre-processing technique before splitting the data to similar operating phases and applying

PCA to the reactor data. Kosanovich et al. (1994) also show that the model built on one reactor can be applied on another reactor operating the same process so long as the input data to the models are appropriately scaled. Kourti et al. (1995) discuss the application of MPCA and MPLS to a batch polymerization process using historical process data for outlier detection. Another technique presented in this case study is to construct the PLS model using only the 'good' or 'normal' operating data and apply this model to the 'bad' or 'abnormal' operating data to identify the cause of the 'bad' or 'abnormal' batches, using squared prediction error analysis. Neogi and Schlags (1998) present the application of MPCA and MPLS to an emulsion batch polymerization process and realigning the data using reaction extent as an indicator variable. Similarly to Kourti et al. (1995), Neogi and Schlags (1998) also apply fault identification by constructing the model on 'normal' data and then applying this to the 'abnormal' data, and using MPLS to predict the quality of the product.

Martin et al. (1996a) discuss the case study of a simulated batch polymerization processes using cross validation to identify the number of principal components to retain, and presents an alternative statistic to Hotelling's T^2, based on a likelihood confidence region. This metric does not cover space in the multivariate hyperplane that does not have data in its confidence region. Dahl et al. (1999) also present on a batch polymerization case study using cross validation to identify the number of principal components to retain in the MPCA model, which is used for outlier detection. Dahl et al. (1999) also comment on how segmenting the batch data into similar operating regions can be useful to prevent erroneous large loadings in the analysis.

Gallagher et al. (1998) present research on a slightly different batch process, metal etching. Again this work discusses the importance of centring and scaling the data, however, it also goes into detail on the usefulness of the Hotelling's T^2 statistic and Q residuals, the impact of keeping redundant variables in the model, and dealing with shifting process means and covariance structures.

In Marjanovic et al. (2006) an interesting and very true observation is made regarding the difficulty to get accurate models with the data available from industrial processes. In this case study of a batch reactor data from a speciality chemical process the historical process data was not detailed enough for the analysis, therefore, data from new batches was collected ensuring all the information required was being captured at an appropriate frequency for the objectives of the modelling. In this case, MPLS was applied to estimate the end point of the batch, and discarding the initial data as an unusual alignment technique was chosen.

A process for manufacturing punctured seamless steel pipe was presented as a case study for MPCA by Xiao et al. (2016). In this example, the number of principal components selected was achieved through using a threshold of 0.9 on the cumulative variance captured by the model. Again the batch process data was broken up into multiple stages, and monitored using both squared prediction error and Hotelling's T^2 contribution plots.

The challenges associated with batch data in which there are multiple operating phases, as is often the case with pharmaceutical batch processes, is discussed in Wang et al. (2015) and Wang et al. (2016). Up-down multi-model dynamic principal component analysis is introduced in Wang et al. (2015) that first segments the data from each batch into a number of phases. Each phase segment is then grouped through use of a local group standardization clustering algorithm, and finally dynamic PCA models are constructed on the variable wise unfolded data for each group. A very similar approach is discussed in Wang et al. (2016), named local collection standardization multi-model dynamic principal component analysis. Here more details on the use of a variable similarity threshold value used in the clustering part of the method is given. This similarity threshold value determines the number of clusters generated and therefore impacts the accuracy and complexity of the model, with a low values resulting in coarse clustering, and poor accuracy, however, a reduced complexity of the model. The method in both papers is demonstrated on the batch data from a ladle furnace steel making process (Wang et al., 2015, 2016). This method, is perhaps at present too complex and unproven for adoption for industry, particularly in the pharmaceutical area. To increase the applicability of this approach to industry, more studies on the reliability of such an approach as a head to head with other techniques should be performed on commercial batch process data from the pharmaceutical industry. Furthermore, additional guidance would need to be developed on the selection of the clustering threshold parameter and the impact of pre-processing with these approaches.

Lv et al. (2016) presents another approach for multiple phase batch process systems, multiple-phase online sorting PCA. In this method, the data is first unfolded variable-wise and then the phases are determined through k-means clustering. The benefit of this approach is that it is automated and does not rely on the prior knowledge of the process data. Following phase clustering, a PCA model is constructed from the variable-wise unfolded clusters (Lv et al., 2016). Lv et al. (2016) describes how the increasing number of phase models improves the accuracy of the modelling, and also presents a method for automatically selecting the number of phase models based on the rate of convergence of the cluster sizes. Little information is

given, however, on quantifying the trade-off in accuracy that this method gives versus the computational effort involved.

Another approach which has potential to lend itself to adoption in industry is detailed in Westad et al. (2015). This method focuses on the automation of the synchronization of batch process data with reduced input for the modeller. The approach relies on unfolding the batch process data variable wise and applying PCA to pull out the features of the batch data in the resultant scores. This method, however, relies on consistent signatures between the batches which may not always be the case, especially when processes have a lot of manual intervention. The method is demonstrated through two case studies. The first is the application to process data from a batch chemical reaction, and the second is spectroscopic data collected during a fluidized bed drying process. The modelling approach is fairly standard, with a model first being built on the normal operating condition batches, and then applied to new batches to identify batches that fall outside the normal operating conditions through statistical metrics such as Hotelling's T^2 and residuals (Westad et al., 2015).

Souihi et al. (2015) present a method, orthogonal projections to latent structures, as an improved method over traditional PCA and PLS methods. The method is an extension to PLS that includes a filtering step to capture the structured variation uncorrelated to the response matrix. The impact of this when applies to time-varying batch process data is to cause a rotation in the latent variables to place more of the time variance information into the first latent variable, thus enhancing the interpretability of the model outputs over that of traditional PLS. Souihi et al. (2015) state that this method performs the same as traditional PLS in terms of prediction accuracy, however, the benefit comes from the improvement of the interpretation of the latent variables. This is demonstrated on a dataset obtained from a hydrogenation reaction comprising of both process data and spectroscopic measurements (Souihi et al., 2015).

Borchert et al. (2015) presents a case study on an upstream and downstream bio-process for the production of a potential vaccine for malaria. The approach taken is to build separate models for the upstream and downstream processes. The models are constructed on the auto-scaled data from normal operating condition batches and tested on the new batches as they are produced using statistics such as Hotelling's T^2 and residuals to determine if the new batches fit the data or show different behaviours. The study goes further to present how batch-wise unfolding is useful for offline monitoring of how similar or different batches are

and if there are any trends or clusters with the batches, whereas variable-wise unfolding is useful for online monitoring of a batch (Borchert et al., 2015).

Another case study in which the normal operating condition batches are first collected is presented in Sarraguca et al. (2015). In this case a cocrystallization study is performed and the spectroscopic data collected from the experiments is processed using PCA, following unfolding in the variable-wise direction and mean centring. The modelling was able to detect gross changes in solvent level, and the absence of input materials in the system. Additionally, interpretation of the loadings showed a potential form change obtained during one of the abnormal operating condition batches (Sarraguca et al., 2015). Only two principal components were retained in the model with a total of 97.4% variance captured. This would potentially be considered high for process data, however, is normal for spectroscopic data.

4.6 Summary

This chapter presented an overview of principal component analysis methods for batch data processing. Firstly, a description of PCA in general was presented followed by a number of methods for selecting the number of principal components to retain in a model. Although the cross validation methods are the most preferred in the literature, due to the ease in which they can be automated for on-line PCA applications, alternative methods are also presented as a method for manually sense checking the cross validation results. Furthermore, despite significant research effort in the area, there remains no consensus on which of the methods should be used. It would be prudent, therefore, if it is unclear how many principal components to retain in a model, that the modelling procedure be applied with different numbers of principal components.

Some model quality statistics including Hotelling's T^2 and Q residuals were also discussed in chapter 4, alongside the complementary contribution plots for interrogating the model to relate statistics of interest back to the original data. These tools have a place in multiple areas of the framework. The first is to identify potential outliers in te data, and subsequently interrogate the loadings to determine why the batch is an outlier. The importance of performing the outlier detection as an iterative procedure is discussed, and examples where these methods have been used in the literature for outlier detection are detailed.

The second area in the framework where these quality statistics are of use is in the modelling

section. More specifically to identify the difference between groups of batches. Again, recent examples of such procedures detailed in the literature are discussed.

Finally, multi-way and dynamic extensions to PCA that may be applicable to batch process data were next presented before relevant applications of principal component methods to process data found in the literature were summarised. Pulling all of these tools together results in the framework presented in in figure 4.12.

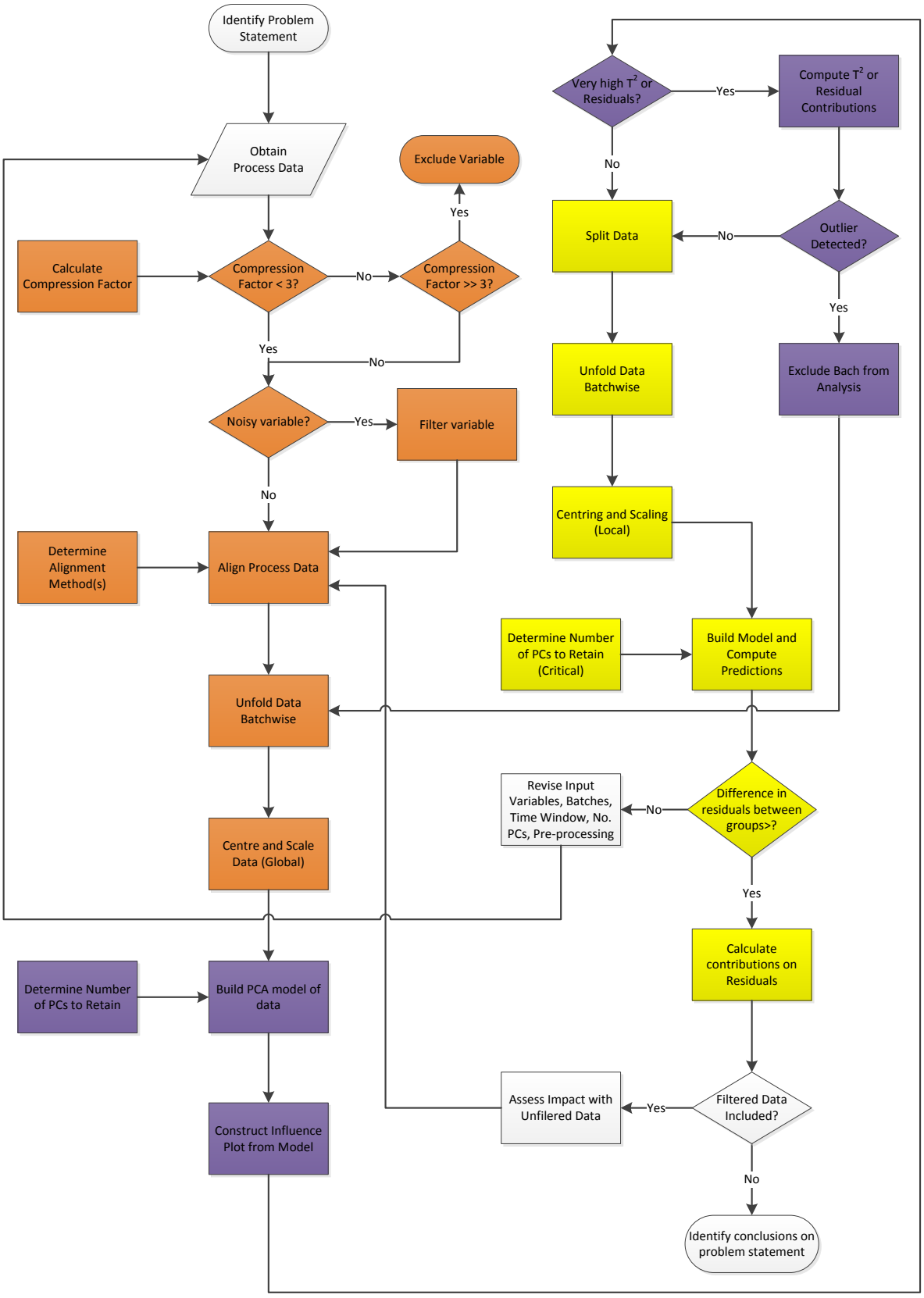


Figure 4.12: Overview of framework to extract multivariate information from batch process data highlighting (yellow) the PCA methods discussed in chapter 4

Chapter 5. Multivariate modelling of spironolactone process data

In this chapter the results of the modelling of the dryer and the reactor used in the manufacture of spironolactone (see chapter 2) are discussed. This chapter starts with a summary of the objectives of the modelling and an overview of the modelling approach employed. Then follows a brief overview of the drying process, pre-processing of the dryer data, finally the results of the modelling and the impact of filtering data has on the results of such models.

Following from the results of the modelling of the dryer data, the same methodology is applied to the reactor data, highlighting some different challenges with process data, namely alignment and outlier removal, and the impact of alignment on the results of the modelling.

5.1 Modelling objectives

Following isolation of the spironolactone drug substance through crystallization and filtration, the spironolactone is dried. This drying process is recipe controlled and described in section 5.3.1. At the end of the recipe driven process the dryer contents temperature must be greater than 80 °C. If this is not the case, the operator manually returns the dryer to vacuum drying until the contents temperature achieves the 80 °C end point. Under normal operating conditions the dryer is the bottle neck in the process with a cycle time of approximately 21 hours, however, this is convenient as it fits in with operator shift patterns allowing 1 batch per day to be manufactured. Over recent years however, the majority of batches have started to require additional drying following the end of the recipe process in order to reach the drying end point. This additional drying can be significant with some batches requiring 10 hours or more of additional drying. This has a significant impact on the ability to schedule and manufacture batches. The change in the drying time of the batch gradually became more prevalent over time and no single change to manufacturing process, equipment, or raw materials can be attributed to the cause of the change.

The objectives of this modelling are to use the available process data from both the reactor and

dryer within the proposed framework of this thesis to understand what is causing the variable drying time.

5.2 Modelling approach

An overview of the modelling approach is shown below. It comprises briefly of obtaining the process data, and pre-processing it, followed by outlier detection using batchwise MPCA, and finally interrogation of the differences between an 'ideal' set of batches, and a group of batches with a 'non-ideal' attribute (in this case drying time).

The pre-processing comprises of importing the data, determining how that data has been compressed and if it is acceptable for multivariate analysis, determining if filters need to be applied to any of the process variables, aligning the data, centring and scaling the data, and finally unfolding the data.

Multiple plots are used to aid in the analysis of the data. Influence plots are used to help with the detection of outliers accompanied by contribution plots on the residuals or Hotelling's T^2. Again the residuals and associated contribution plots are employed in interrogating the differences between the groups of data with 'ideal' and 'non-ideal' attributes.

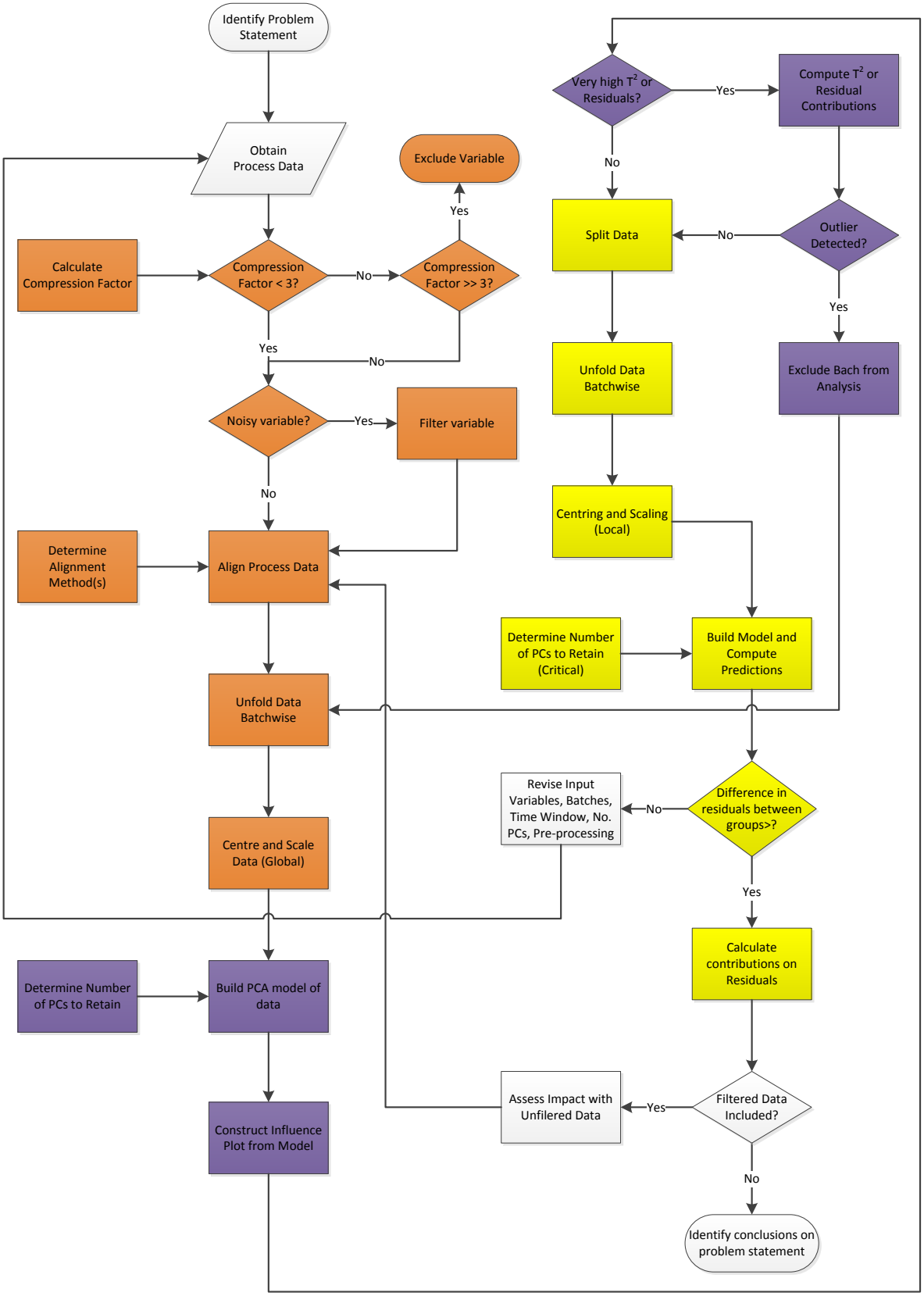


Figure 5.1: Proposed Framework

5.3 Multivariate modelling of spironolactone drying data

This section provides a high level summary of the drying process for spironolactone. A more detailed description of how the drying process is operated is discussed in chapter 2.

5.3.1 Summary of drying process

Spironolactone is dried in a conical screw agitated dryer with heat applied through a jacket on the dryer. Following isolation of the API in the reactor the batch is filtered in two parts. The first part, approximately half of the batch, is loaded onto a pressure filter and filtered. Following this process it is discharged from the filter directly into the dryer below. Whilst the second half of the batch is being filtered, the drying process is started on the first half of the batch. This involves reducing the pressure in the dryer with a vacuum pump and applying 25 °C to the jacket. The batch is agitated throughout this phase which lasts for 3 hours. After the three hours drying has elapsed, the jacket recirculation valves are closed and the dryer waits for the remainder of the batch to complete the filtering operation with no agitation. After the second half of the batch has completed the filtration, it is discharged from the filter into the dryer on top of the first half of the batch. The whole batch is then dried under vacuum for 2.5 hours with a jacket temperature of 25 °C following which the jacket temperature is ramped up to 90 °C over 2 hours. The batch is dried still under vacuum at 90 °C for 2 hours after which the pressure is restored to the dryer and the batch is 'de-odoured' with a jacket temperature of 90 °C and a small purge of nitrogen passing over the top of the batch to remove any odorous compounds from the batch. The de-odour takes 11 hours following which the batch is subject to a final 1 hour vacuum dry with the jacket at 90 °C. Following the final 1 hour vacuum dry, drying should be complete and the sequence calls for the pressure to be restored and cooling is applied to the jacket. However, if the batch temperature has not reached 80 °C drying is not deemed complete and the operator must manually return the dryer to vacuum dry until the target temperature has been achieved.

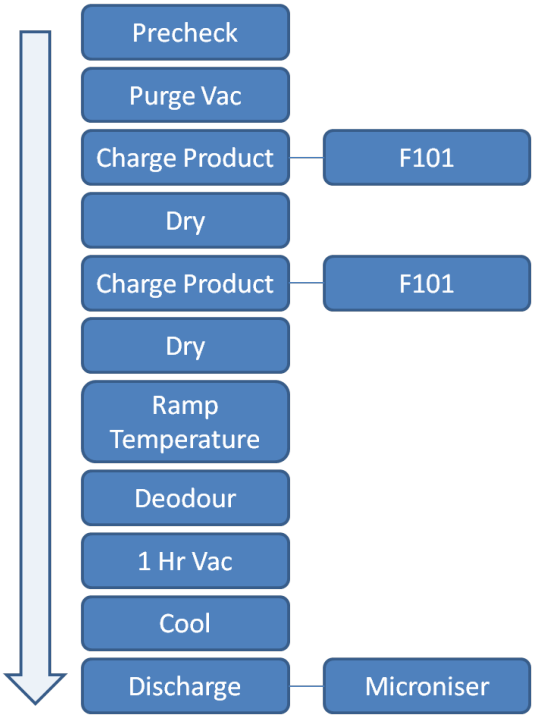


Figure 5.2: Overview of spironolactone drying sequence

5.3.2 Control of dryer

Aspen PROVOX is a DCS (Distributed Control System) used for the control of the dryer operation. It controls the sequencing of the drying process through a recipe that is loaded for each batch, it also sends signals to pneumatically actuated control valves around the dryer to effect control over the dryer jacket services, jacket temperature, pressure. PROVOX also has interlocks that include those required for charging and discharging of the dryer. Although there is a lot of automatic control implemented on the dryer through PROVOX and PID (proportional + integral + derivative) controllers, some manual intervention is also required during the drying process and PROVOX can prompt for these interventions. More details regarding the control of the dryer are discussed in section A.1.3.

5.3.3 Compressed data

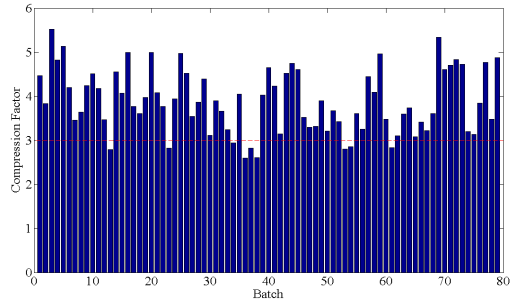
A set of 79 spironolactone (NMP) batches were collected and the data from the dryer selected. The data had been compressed when stored in the data historian, and subsequently reconstructed within the historian through linear interpolation back to the original sampling rate of 30 seconds (see chapter 2 for more details on the compression of the process data).

Because the original (uncompressed) data is not available, the compression factor of the data needs to be estimated. The method presented by Thornhill et al. (2004) where the second derivative of the reconstructed process data is obtained and values near zero (i.e. within the arithmetic accuracy of the data processing) and non-zero values to indicate linear interpolated and non-interpolated data respectively.

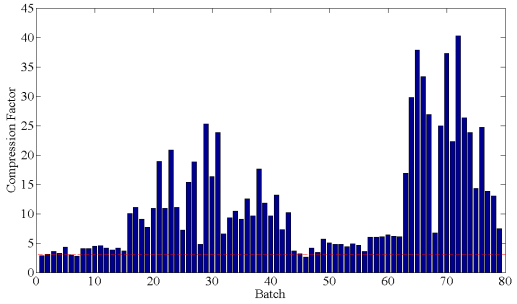
To check that the data has been reconstructed at an appropriate sampling frequency the pattern of second differentials can be observed. In this data there were not always two consecutive non-zero second differences where the linear segments were joined. This indicates that the sampling frequency is appropriate, whereas if the second differences gave a pattern in which there was always a pair of non-zero second differences where the segments were joined, the data may have been reconstructed at a lower sampling frequency than the original data was collected at and therefore the compression factor needs to be modified to account for this duplication in non-zero second differences.

The compression factor for each variable and each batch was calculated and are shown in the following plots. The estimated compression factors are summarised in table 5.2. Figure 5.3a shows the estimated compression factor of the contents temperature for each batch. The compression factor is around the limit of 3 proposed by Thornhill et al. (2004), however, some batches exceed this limit with compression factors approaching 5.5. Although this is above the limit it does not exceed it by a large margin and therefore the data may be suitable for use in the subsequent multivariate analysis, however the results should be used with caution.

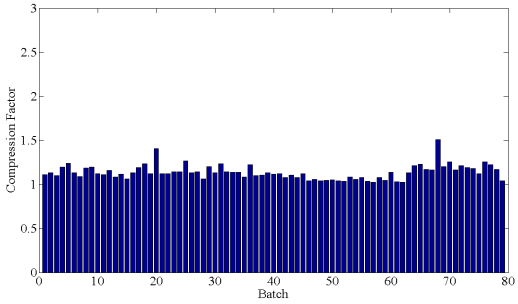
The compression factor for the dryer jacket temperature is shown in figure 5.3b. This indicates that the dryer jacket temperature data is too compressed to be suitable for use in multivariate analysis. Although some batches have relatively low compression factors (between 3 and 5), there are a few sustained periods where the compression factor becomes high (between 10 and 40). The compression factor would be expected to be high for this variable as it is a controlled variable and should, assuming good control of the variable, be held at a constant value or on a first order ramp. If controlled variables do not deviate from the set point (i.e. are well controlled) only a small number of data points will need to be recorded to capture the behaviour of the variable. Thornhill et al. (2004) also note that there are certain types of data where higher compression factors may be acceptable where the intended use of the variable is to record constant values, such as set-points, targets, and limits for example. The jacket temperature data may therefore be included model if required.



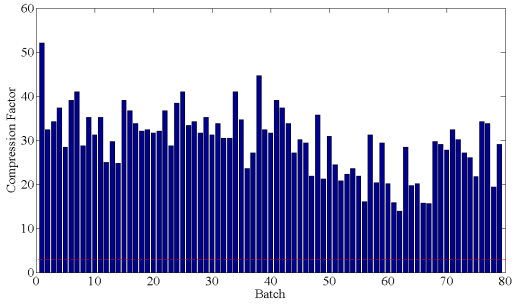
(a) Dryer contents temperature



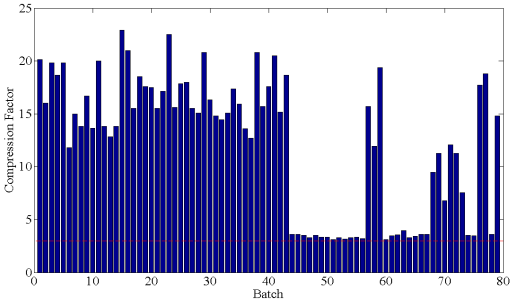
(b) Dryer jacket temperature



(c) Dryer controller output



(d) Dryer full pressure



(e) Dryer vacuum pressure

Figure 5.3: Estimated compression factor for the spironolactone dryer process data for each batch

From figure 5.3b there appears to be a time dependent structure in the compression factor of the data. This can be seen as the batches were plotted in the order they were manufactured, and a compression factor appears to rise and fall in discrete blocks. Poorly tuned PID controllers can lead to oscillations in the data, which would make the compression factor low as the data historian records more data to capture this oscillatory behaviour. The periods where the compression factor is high, could indicate where the PID controller parameters have been adjusted to obtain better control. This was confirmed by looking at the data from batches with high and low compression factors.

The compression factor for the dryer temperature controller output is shown in figure 5.3c. This indicates that this variable is suitable for multivariate data analysis with a compression factor of approximately 1.

The compression factor for the dryer full pressure data is shown in figure 5.3d. This indicates that this variable is not suitable for multivariate data analysis with a compression factor significantly above 3. Any information contained in this variable should however be also contained in the vacuum pressure variable and therefore the exclusion of this variable is of no consequence to the modelling.

The compression factor for the dryer vacuum pressure is shown in figure 5.3e. This indicates that for approximately half of the batches this variable is not suitable for multivariate data analysis with a compression factor of approximately 15 to 20, whereas the other half of the batches the data is suitable with compression factors of approximately 3.

There is a clear binary response on the level of compression for the batches. This is caused by the pressure back-pulses used to clear the dryer filter socks being inadvertently turned off. Without these pressure pulses, the pressure in the dryer is more stable and therefore fewer data points are needed to be recorded to capture this resulting in higher compression factors. The pressure data may be important in the model, and although the inclusion of the first half of the data may impact the relationships obtained from the multivariate analysis due to this over compression, excluding this data may also prove detrimental to the modelling ability. This requirement for the inclusion of this data in the models will therefore be evaluated during the model building process.

Table 5.2: Summary of the compression factors for the dryer data

Dryer variable	Compression factor	Outliers	Include variable?
Contents temperature	3 - 5	None	Yes
Jacket temperature	3 - 5 & 10 - 40	Batches 1 - 15 & 44 - 62 due to poorly tuned PID controller	No, compression factor too high for a large number of batches. Relevant information should be captured in contents temperature and controller output data.
Controller output	1	None	Yes
Full pressure	20 - 50	None	No, compression factor too high. relevant information should be captured in vacuum pressure data
Vacuum pressure	3 & 15 - 20	Yes, most batches prior to batch 45 did not have pressure purges activated to clear the filters	Yes, although the compression factor is above for half of the batches, this is a controlled variable and therefore higher compression factors can be accepted

5.3.4 Missing data and filtering

There was no missing data present in the data set obtained, thus only filtering had to be considered in this case.

Most of the process variables from the spironolactone drying process have some level of high frequency noise associated with the signal. This noise is low amplitude and, as Brown and Wentzell (1999) discuss in their paper on the hazards of filtering signals for use in multivariate analysis, removal of this noise will yield minimal improvement on the predictive power of the model, however may introduce distortion and correlated error into the filtered data. No filtering will therefore be performed to remove any measurement system noise from the data for the dryer contents temperature, jacket temperature, and pressure data.

There are, however, a couple of notable features in the data which may be considered for filtering. The most obvious of these is found in the temperature controller output signal. Due to a poorly tuned controller for the jacket temperature, the controller exhibits oscillatory behaviour (figure 5.4). This oscillations peak at 50% as a maximum limit was placed on the controller output to prevent temperature overshoots during the start of the heating phases which could cause the dryer operation to fail.

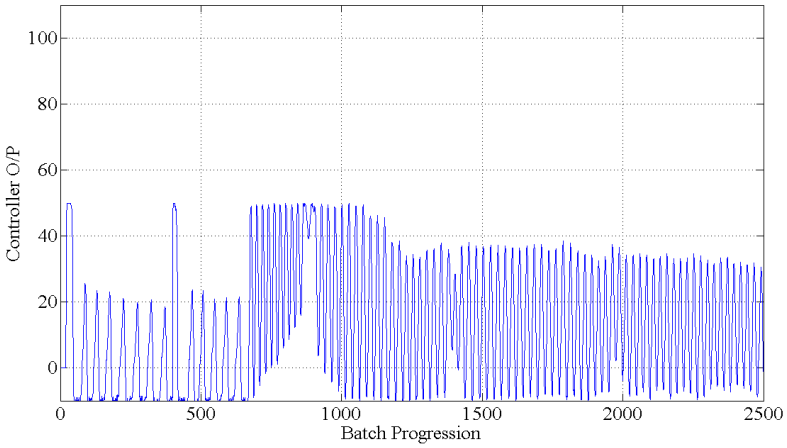


Figure 5.4: Temperature controller output with oscillations

This oscillatory behaviour in the controller signal is significant and as a result, it is a major contributor to the variance in the dryer process data and will therefore have significant impact on the PCA model. Constructing PCA models from the data with the large oscillations will lead to a model in which the oscillations are a significant contributor to the variance captured. This is undesirable as these oscillations are mostly noise induced from the poorly tuned

controller, masking the more interesting attribute of the mean controller position indicating how much steam is required by the dryer to maintain the temperature set point. Filtering this signal may prove beneficial to the modelling effort by removing these oscillations and revealing the local mean controller value. The impact of applying such a filter should, however, be assessed by running the analysis both with and without the filtering applied. In order to identify an appropriate filter to apply to the controller output signal, the single sided magnitude spectrum of the signal can be computed (figure 5.5).

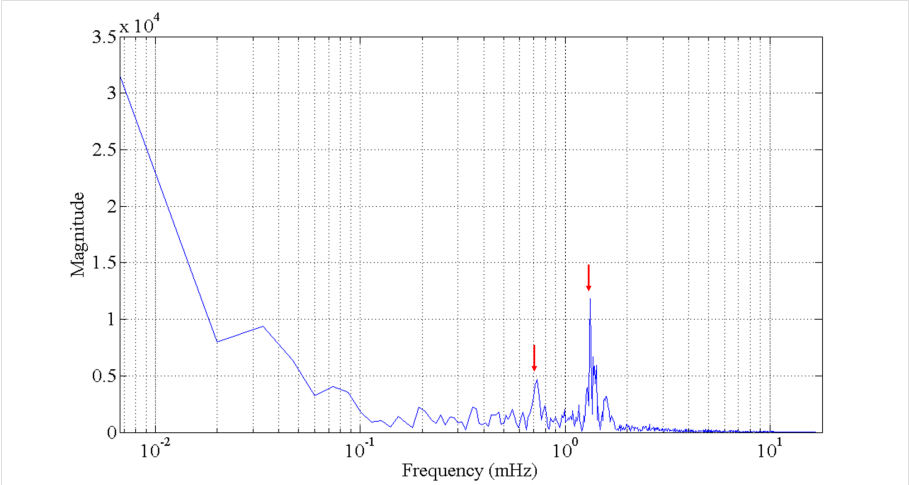


Figure 5.5: Single sided magnitude spectrum of the controller output signal (arrows indicating 0.7 mHz and 1.5 mHz)

From figure 5.5 the slow moving (i.e. low frequency) information from the controller output signal can be seen on the left side of the plot. There are also two significant peaks towards the higher frequency end of the spectrum where the magnitude increases. These are found at approximately 0.7 mHz and 1.5 mHz. These relate to the oscillations in the low temperature drying phases and the high temperature drying phase respectively, and can be confirmed by measuring the distance between oscillations in this period and converting it to a frequency. A band stop filter should therefore be designed around these two frequencies to eliminate the oscillatory behaviour from the signal and reveal any underlying structure in the signal.

A second order Butterworth bandstop filter was built on the controller output signal. The filter was applied in both the forward and the backward direction to in order to reduce the distortion and delay in the filtered signal. Figure 5.6 shows the a comparison between the original signal and the filtered signal, where the oscillations in the data have been significantly reduced revealing the structure in the controller output variable that was previously masked by the oscillations. There are however some artefacts visible in the filtered signal as a result of the

filtering. The impact of applying the filter to the controller signal is discussed in further in section 5.3.9.

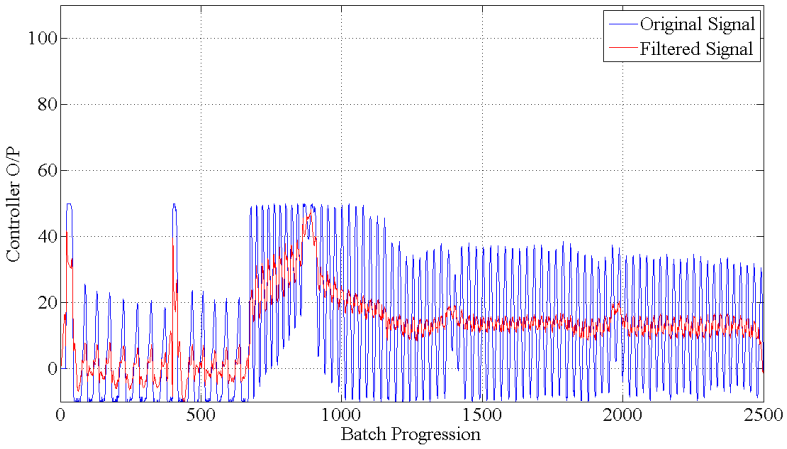


Figure 5.6: Original and filtered temperature controller output signal

5.3.5 Data alignment

In order to compare variables of each batch over time, the batch progression axis needs to be aligned (be this time or another variable). Due to differing starting conditions, environmental conditions and operator interaction throughout a process, the time taken for each operation, and even phase, may change significantly from batch to batch. An example of this behaviour on the drying data is shown in figure 5.7. Figure 5.7 is an illustration of how the requirement for operator prompts and waiting on other unit operations may cause extensions in the time profile of a batch.

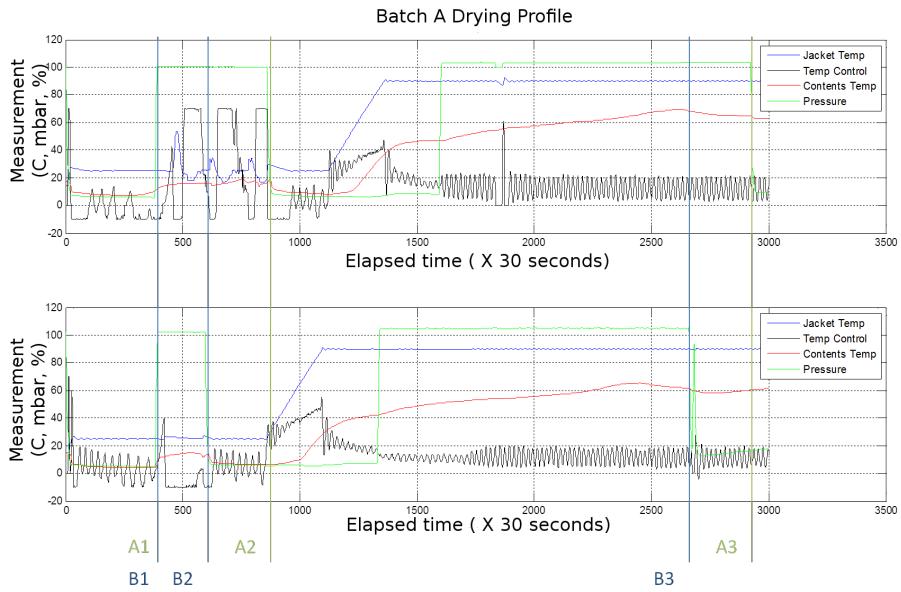


Figure 5.7: Misalignment in drying profiles of two batches due to process wait. A1 and B1 indicate the start of the second charge operation to the dryer for batch A and B respectively. A2 and B2 indicate the start of the second dry operation for batch A and B respectively. A3 and B3 indicate the end of the deodour operation for batch A and B respectively.

Figure 5.7 shows the drying profile for two batches of spironolactone. Both batches start at the same time on the time axis and they also both finish the first dry operation in the same amount of time (A1 and B1). The time taken to get the second part of the batch filtered and transferred to the dryer differ with batch A is longer (A2) than batch B (B2). The misalignment arising from this additional waiting time is carried throughout the batch trend and can be seen at the end of the sequence drying operation (A3 and B3).

There are several methods of overcoming this problem including data cutting, linear interpolation between two known features in the data, and time warping the data (García-Muñoz et al., 2011; Wan et al., 2014). These methods are discussed in detail in chapter 3.

Data cutting is the simplest method, however, it has severe limitations. Data cutting carries the risk that information can be lost if the data removed contained useful information. The dryer data is however an example of where such a technique may be employed to achieve alignment of the process data. As the drying process is recipe controlled, and each of the phases is controlled by a fixed time, the data should be relatively easy to align. There are three regions on the drying process data where the cutting technique has been used.

The drying sequence starts with a pre-check and a series of vacuum - nitrogen purges, followed

by a charge to the dryer of the first portion of the batch from the filter. The data available during these three operations will not contain any relevant information relating to the product or the dryer. The contents temperature probe is not in contact with the product for most of these operations, the jacket temperature is not being controlled and therefore the contents temperature and the controller output signals will not contain any relevant information, and other than to verify that the dryer nitrogen purge has been successful and the performance of the vacuum pump, the pressure measurements will contain no information regarding the spironolactone product. The start of drying was therefore taken from the point at which the temperature controller output starts to increase following the vacuum purges, indicating that the product has been charged, the agitator is moving and the drying sequence has started.

The second region to be cut from the drying data is the data collected between the end of the first dry phase and the the start of the second dry phase. Following the first dry phase of fixed duration, the control system stops the recirculation pump on the jacket water, stops the dryer agitator, and waits for the operator to confirm that the second portion of the batch has been charged to the dryer. The data obtained during this phase cannot be trusted to be representative of the conditions in the dryer. As the agitator is stopped, the contents temperature thermocouple is only measuring the temperature at 1 location in the dryer assuming that the probe is in contact with the batch whilst waiting for the second portion to be dropped into the dryer. The jacket on the dryer is in a no flow condition, therefore the temperature on the jacket outlet is not representative of the temperature in the jacket. Also, as a result of the no flow condition in the dryer jacket, the controller behaviour is unpredictable and either fully opens (within the constraints programmed in the DCS) or fully closes based on the signal received from the unreliable jacket temperature thermocouple. Similarly, the pressure monitoring ambient pressure in the system.

By removing the data between the start of the second dry and the end of the first by phases, this spurious data is excluded from the models, and the drying data becomes aligned. By removing this data, however, the elapsed time variable should be considered for inclusion in models as this may hold relevant information relating to different batch lengths that may be important.

The final section to be cut from the data is towards the end of the batch. One of the objectives of the modelling of the dryer data was to identify causes for extended drying (i.e. changes in contents temperature at the end of the deodour phase). Therefore, if information on the causes of extended drying is present in the drying data it should be present before the end of the

drying operation. Furthermore, following the deodour and final 1 hour vacuum drying phases, the sequence automatically applies cooling to the dryer jacket irrespective of the contents temperature. Therefore the information in what happens to the batch following the deodour phase is highly dependent on if the operator is around to switch the drying from automatic to manual. All data following the end of the deodour operation can therefore be cut from the model allowing for the remaining batch drying trajectory to be aligned and each batch to be of constant length.

5.3.6 Unfolding

Batchwise unfolding was employed for both the outlier detection and the exploratory modelling. For the outlier detection, the entire data set was unfolded in one block, whereas, for the exploratory dataset, a model calibration set was created by batchwise unfolding of the batches with fast drying times, and a second dataset was created by batchwise unfolding of those batches with long drying times.

5.3.7 Centring and scaling

Two methods of centring and scaling were applied to the data depending on the use in the modelling process. The first, used for outlier detection, takes all of the dryer data. For each variable the mean trajectory was calculated and subtracted from each batch. This removed the non-stationary information from the data caused by the different operating phases and helps give equal weight to each time point in the model. Additionally, dividing each batch by the standard deviation along the trajectory scales the data so that each variable has an equal weight.

The second method is similar to the first, however a only a subset of the batches was used to calculate the mean and this mean was subtracted from all of the batches. This again removes the non-stationary information from the data caused by the different operating phases giving equal weight to each time point (sample) in the model. However, by centring based on a subset of the data, the differences between the means of the two groups will be pulled out of the model. This approach is commonly used in batch monitoring where a new batch is scaled by the means of the batches that the model was build on (i.e. calibration data set). Again, all the batches were scaled by dividing by the standard deviation of the subset of batches used to calculate the mean.

5.3.8 Outlier detection

Following autoscaling and unfolding the dryer data batchwise, principal component analysis was applied to the data in order to identify any outliers in the data that should be removed prior to a more focused analysis of the data.

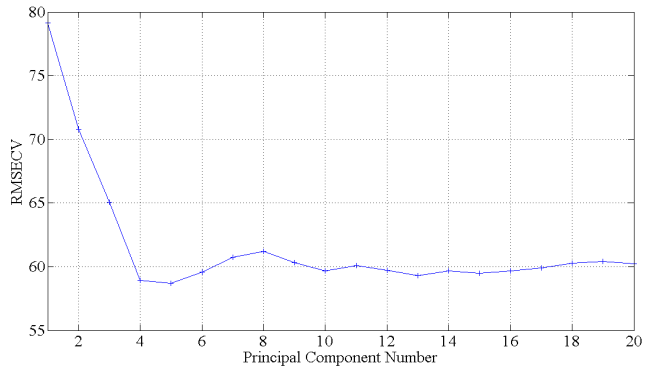


Figure 5.8: Root mean squared error for cross validation for dryer data

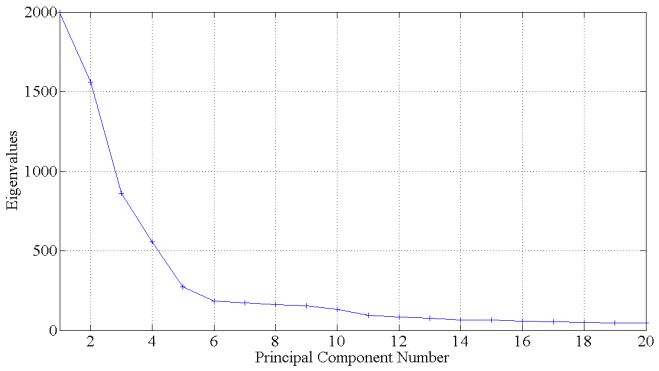


Figure 5.9: Scree plot for dryer data

The number of principal components to retain in this model was selected by leave one out cross-validation (figure 5.8) with five PCs retained. The scree plot in figure 5.9 shows the eigenvalues are still decreasing rapidly after four PCs and therefore six PCs should be retained. As this initial model is being used for outlier detection, it is not critical to capture all of the variability in the data, just enough to identify any batches that significantly deviate from a 'typical' batch. Furthermore, the inclusion of more PCs than necessary risks the inclusion of noise in the model, which may make it more difficult to identify outliers. The model with four PCs explained a total of 66.33% of the variance, with 26.62% explained in the first principal component, a further 20.82% variance explained in the second principal component, 11.49% explained in the third principal component, and 7.40% explained in the fourth principal component.

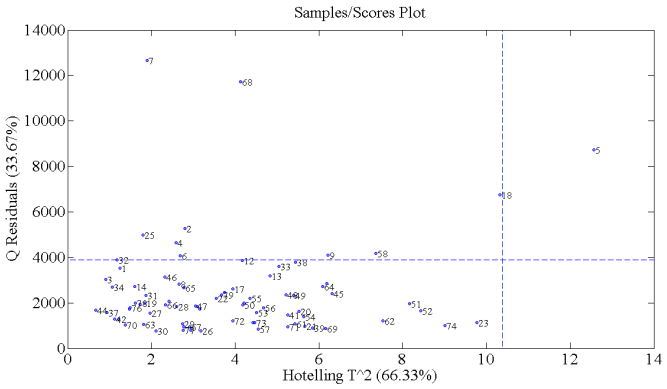
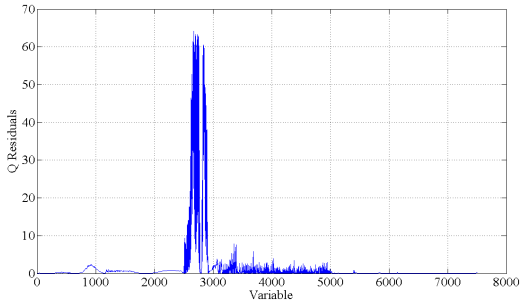


Figure 5.10: Influence plot of auto scaled dryer data unfolded in the batch direction with 95% confidence limits shown

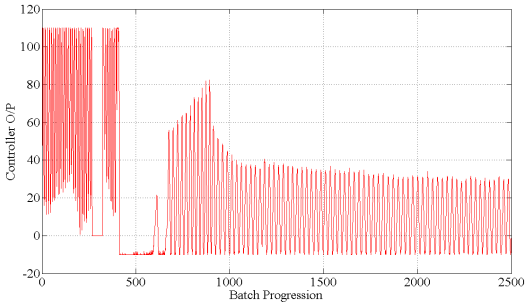
The influence plot in figure 5.10 identifies two batches that have high residuals (batch 7 and 68). These batches may be potential outliers and the contribution plots on the residuals should be interrogated to determine if these are true outliers.

The contributions on the residuals for batches 7 (figure 5.11a) and 68 (figure 5.12a) both identified the controller data to be abnormal for these batches, and there is also some abnormality in the pressure data during the deodour for batch 68. Batch 7 was found to be calling for excessive steam during the first dry phase, and subsequently no steam during the beginning of the second dry phase (figure 5.11b). At the start of the first dry phase, vacuum is pulled on the dryer. This is done in two stages, first using a liquid ring vacuum pump to get the initial partial vacuum, followed by a switch over to the high vacuum pump. The high vacuum pump however failed causing the DCS to shut the dryer down into a failed state with cooling applied to the dryer. This was immediately recovered by the operator. On further investigation into the DCS sequences, it was found that the fail phase switched the valves on the dryer jacket from recirculation to once through to allow for cold water to bring the temperature of the jacket down quickly. The recovery sequence however does not switch the valves back to recirculation when the steam controller is returned to use and as a result, the jacket is operated in a single pass mode. The temperature controller is not set up to operate in this mode, and as a result large oscillations in the controller output occur as the controller attempts to achieve the set point on the cold water entering the jacket which is likely to contain temperature disturbances as it is a common source of water for multiple plants.

Due to these unusual oscillations in the controller, oscillations were also induced in the jacket temperature which were large enough to cause the jacket temperature to exceed 30 °C which caused the dryer to fail to a safe mode and apply cooling to the dryer jacket until an operator



(a) Contribution on the Q residuals



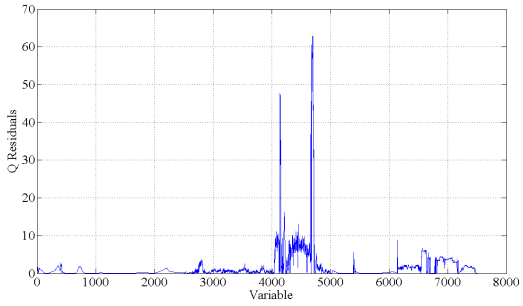
(b) Dryer temperature controller output

Figure 5.11: Dryer steam failure identified in (a) Q residuals for batch 7 drying data and confirmed in (b) by the dryer temperature controller output data for batch 7 showing batch failure behaviour in high steam demand in the first 400 time points, followed by long periods no steam demand up to approximately time point 600

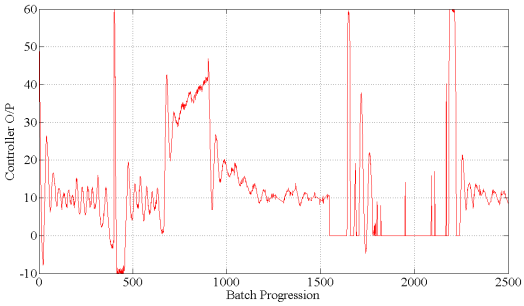
recovers the operation. The lack of steam at the start of the second dry phase was due to the steam controller fully opening the steam valve, with a no flow condition in the dryer jacket, during the pause whilst the second portion of the batch was charged to the dryer. This over heated the heat exchanger and jacket water around this area so that when the second dry phase was initiated and the flow returned to the jacket, the temperature rose rapidly above the set point. The controller responded by closing the steam valve, however, because during heating the jacket is a closed loop, it took a long time to remove this excess heat from the system. These behaviours are not normal behaviours expected within a batch and therefore the data should be removed from the analysis.

In the case of batch 68, the dryer moved onto a failed state multiple times throughout the deodour phase (figure 5.12b). As a result the operators tried several times to continue the drying including switching back to vacuum drying and then back to the deodour phase. The information is not available in the data to determine the root cause of these failures.

The influence plot (figure 5.10) also suggested that batches 5 and 18 did not fit the model well with relatively high residuals for both batches. Investigation using the contributions plots and original data found that these batches also had failures during manufacture where the temperature controller output moved to the safe condition of 0%. Batch 5 has a similar root cause for these failures as batch 7, where an unreliable high vacuum pump failed causing the dryer to move to a failed safe state. The recovery sequences do not however set the jacket valves to the correct flow path resulting in the controller being unable to heat the jacket water



(a) Contribution on the Q residuals



(b) Dryer temperature controller output

Figure 5.12: Dryer steam failure identified in (a) Q residuals for batch 68 drying data and confirmed in (b) by the dryer temperature controller output data for batch 68 showing batch failure behaviour manifested as periods of no steam demand during the deodour phase (time point 1500 onwards)

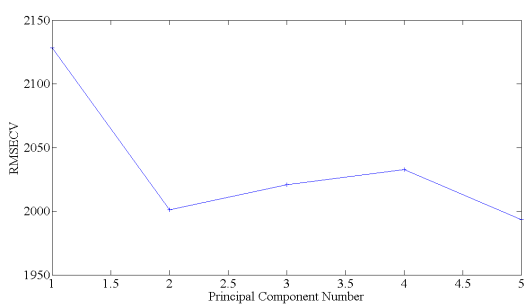
sufficiently. Batch 18 showed similar behaviour to batch 68, however only one failure event occurred during the deodour phase.

These batches were removed from the dataset as outliers and the process repeated with the remaining batches until the remaining batches did not have very high residuals or Hotelling's T^2 values. Although confidence limits are shown on the influence plot, these should be used as a guide only and not a hard rule as to whether a batch is an outlier or not (Bro and Smilde, 2014).

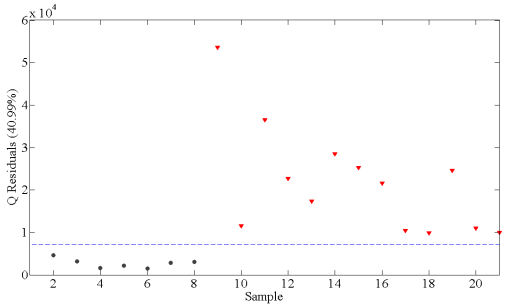
5.3.9 Principal Component Analysis

Following the removal of the outliers from the original data set of 79 batches, two subsets were pulled out of the data. The first was the desirable batches (i.e. those with an extended drying time of less than 3.5 hours), and the second was the undesirable batches (i.e. those with extended drying times of greater than 7.5 hours). These subsets consisted of 8 batches and 13 batches in the fast and slow drying groups respectively.

All of the batches were centred around the mean of the batches with short extended drying times, and scaled to unit variance. By removing the means of the fast drying batches from all of the data, the non-stationary behaviour will be removed from the model as this is not of interest, however, the information on how the slower drying batches deviate from the trajectories of the faster drying batches will be retained in the model. The data was then unfolded batchwise to allow the PCA model to be built.



(a) Root mean squared error for cross validation



(b) Q Residuals

Figure 5.13: (a) Root mean squared error for cross validation to select the number of PCs to retain (b) Q Residuals for the dryer data from the fast (black dots) and slow (red triangles) drying batches with the 95% confidence limit indicated

As the residuals in the model will be used in this analysis, it is important to choose the correct number of principal components to retain in the model. Leave one out cross validation was used to select 2 principal components to retain (figure 5.13a).

A principal component model was built on the data with fast drying times. This explained 59.01% of the variance in the data with 35.73% accounted for in the first principal component, and the other 23.29% in the second principal component. The principal component model was then applied to the slow drying batch data. The residuals plot shown in figure 5.13b shows that the slow drying batches are different to the faster drying batches and all of the slow drying batches falling above the 95% confidence interval limit. Examination of the contribution plots for the residuals would be prudent for these batches to identify which variables are contributing to the differences between the two groups of data.

The contribution plot on the residuals is shown in figure 5.14. This indicates 3 regions of interest that contribute significantly to the high residuals of the batches that have long extended drying times. There are two regions with extremely high contributions. The first at approximately variable 2900, and the second at around variable 5000. The first occurs in the pressure data when the dryer is pulling vacuum at the start of the second dry. For each batch there is a rapid decrease in pressure in the dryer (over 1 - 2 time points) which if not perfectly aligned will cause high variance between the batches. This is a very short event, in the contributions and is believed to be an artefact from the alignment process and therefore is not considered to have a genuine significance in determining the drying time of the batches.

The second region with extremely high contributions around variable 5000 corresponds to the temperature controller output signal at the beginning of the first drying stage. This is the same

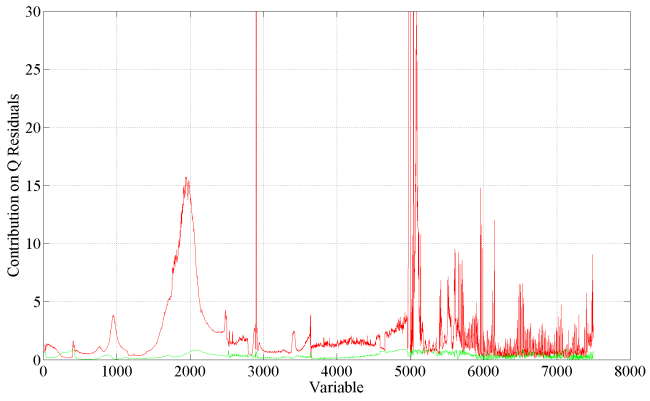


Figure 5.14: Contribution plot on the Q Residuals for the dryer data from the fast (green) and slow (red) drying batches

fault that was identified during the outlier detection, where the recovery from a failed vacuum pump caused a temporary error in the valve configuration on the dryer jacket leading to a delay in getting the heat into the jacket at the start of the drying process. This fault can have a genuine influence on the drying time of the batches, and is related to the manufacturing equipment and sequences. These can be relatively easily remedied through more investment in more reliable high vacuum process, and/or an update to the fail and recovery sequences.

Finally, a third region between variables 1500 and 2500 show significant contributions to the residuals for the slow drying batches. These are not as high as the previously discussed two instances, however they may still be significant and are comparable to the magnitude of the contributions in some of the controller fault at variable 5000. This corresponds to the contents temperature variable during the deodour, starting when the pressure is restored to just above ambient with the nitrogen sweep across the head space of the dryer however the contributions decent fairly rapidly from time point 2000 towards the end of the batch. This indicates that it is not a straight forward temperature effect such as hotter batches towards the end of the drying phase drying faster than batches that are cooler, but there is something more subtle occurring in the dryer. Further investigation reveals that this peak in the contributions coincides approximately with the point where an endotherm is observed in the dryer. This endotherm occurs at different times and temperatures for each batch, and is more pronounced in some batches than others. Spironolactone is known to have many polymorphic forms, therefore this endotherm may be related to a solid state transition occurring in the dryer as a result of thermal and mechanical energy inputs to the process. This drying endotherm has not been observed in the lab scale process, possibly due to different drying techniques, therefore future work could

be to either monitor online, or obtain samples from the dryer throughout the deodour phase to look for a solid state transformation. This may be achieved through techniques such as Raman spectroscopy (Chakravarty et al., 2009), or x-ray powder diffraction (Espeau et al., 2007; Nicolaï et al., 2007; Liebenberg, 2005).

Impact of not filtering the controller output signal

As the inclusion of filtered variables in multivariate analysis may lead to the inclusion of distorted data, for only small gains in the predictive ability of the analysis, the impact of the inclusion of the filtered variables was also assessed.

The same analysis was performed using the original (unfiltered) signal. Two principal components were retained as selected by cross validation and a principal component model was built on the data with fast drying times. This explained 55.31% of the variance in the data with 37.84% accounted for in the first principal component, and the other 17.48% in the second principal component. The principal component model was then applied to the slow drying batch data. The residuals plot shown in figure 5.15 shows that the slow drying batches are different to the faster drying batches and all of the slow drying batches falling above the 95% confidence interval limit. Examination of the contribution plots for the residuals would be prudent for these batches to identify which variables are contributing to the differences between the two groups of data. Filtering the controller output variable does however slightly improve the separation between the slow and fast drying batches, however both methods are still capable of distinguishing between these batches.

The contribution plot on the residuals is shown in figure 5.16. This indicates the same 3

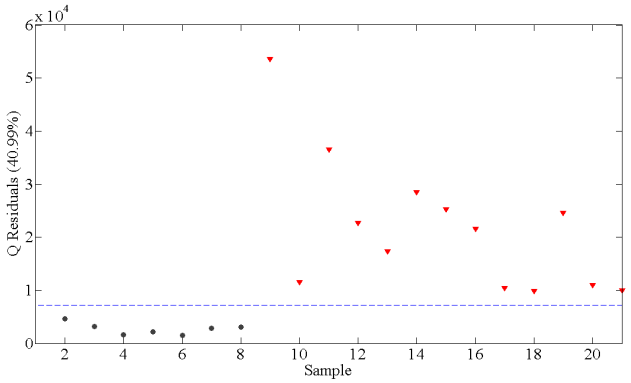


Figure 5.15: Q Residuals for the dryer data from the fast (black dots) and slow (red triangles) drying batches with the 95% confidence limit indicated

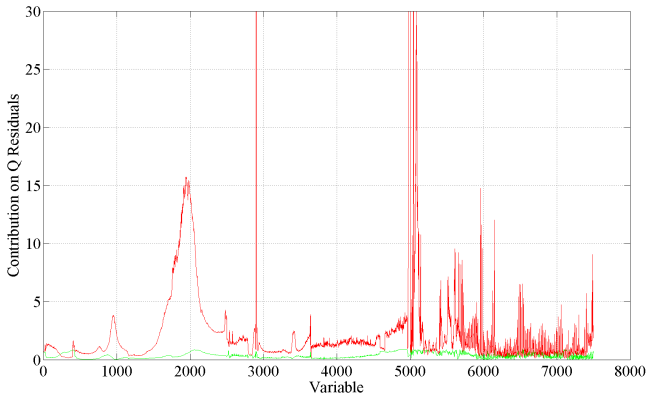


Figure 5.16: Contribution plot on the Q Residuals for the dryer data from the fast (black) and slow (red) drying batches

regions of interest as figure 5.14. The difference between the two is that some of the noise from the controller has been captured in the contributions for the controller, however the same structures are still visible beneath the noise. It is therefore recommended that the controller data does not need to be filtered for this analysis, and the original controller data should be used as the benefit in separation is marginal. However, it was useful to use the filtered variable to get confirmation of the structures within the controller data.

5.3.10 Conclusions

The proposed modelling framework has successfully been applied to the dryer process data and identified some of the challenges with pre-processing such data. The analysis was able to identify a number of different equipment issues that may have an impact on the economic operation of the process, and also identified the cause for the endotherm in the dryer should be further investigated as this is a significant contributor to the variability in drying times between batches.

5.4 Multivariate modelling of spironolactone reactor data

Following from the application of the modelling methodology to the dryer data (see section 5.3), in which some variability in the product was noted that caused variation in the endotherm observed in the dryer, the same modelling approach was applied to the reaction and initial crystallization data from the reactor to determine if the reactive crystallization is having an impact on the drying time. This section provides a high level summary of the reaction and crystallization process for spironolactone. A more detailed description of how the reactor is operated is discussed in chapter 2.

5.4.1 Summary of reaction and crystallization process

The thiolacetylation and isolation reactor is controlled by PROVOX through the sequence illustrated in figure 5.17. Each of the steps has the required settings for the batch contained within the batch recipe. It starts by checking that the reactor is ready and that a valid batch number has been set. The reactor is then evacuated approximately 100 mbarA and purged with nitrogen three times (vacuum-purge). The appropriate valves are then opened to allow the operator to transfer the colour treated aldadiene in methanol and acetone to the reactor via the 0.45μm and 0.2μm filters. After the transfer is complete, the reactor contents are heated to reflux for 6 minutes with the timer starting when the reactor contents exceed 60 °C. After the reflux, the steam supply to the reactor jacket is isolated and thiolacetic acid is charged to the reactor, controlled by weight change in the thiolacetic acid header tank, T104. The reactor is then returned to reflux at 62 °C for 20 minutes. The reactor jacket is then filled with cold water to start the batch cooling. As soon as the jacket is showing full of cold water by activation of a level switch in the jacket, the water is blown out of the jacket using compressed air and the batch remains slowly cooling until the operator confirms that crystallization has been observed in the reactor. Heat is then applied to the reactor to return the contents to reflux at a minimum of 64 °C for 100 minutes to ensure that the reaction goes to completion. After the 100 minute reflux, the steam is isolated and the condensate in the jacket blown clear with plant air. A quantity of methanol is charged to the reactor R101 controlled by weight change on the receiving reactor. There is then an option to add recovery material (spironolactone that has been recovered from the first wash of the process train with acetone). More methanol is then added to the reactor again controlled by the weight change in R101. The reactor is then cooled to 40 °C using cold water in the reactor jacket after which the jacket service is changed to

chilled glycol to bring the reactor contents down to -10 °C. The batch is held at -10 °C for a minimum of two hours before it is filtered. The batch is filtered in two parts in the Rosenmund pressure filter F101, the charge controlled by the weight change in reactor R101. Whilst the first load is being filtered the remaining spironolactone slurry in R101 remains held at -10 °C. More details on the control of the reactor are presented in section A.1.1.

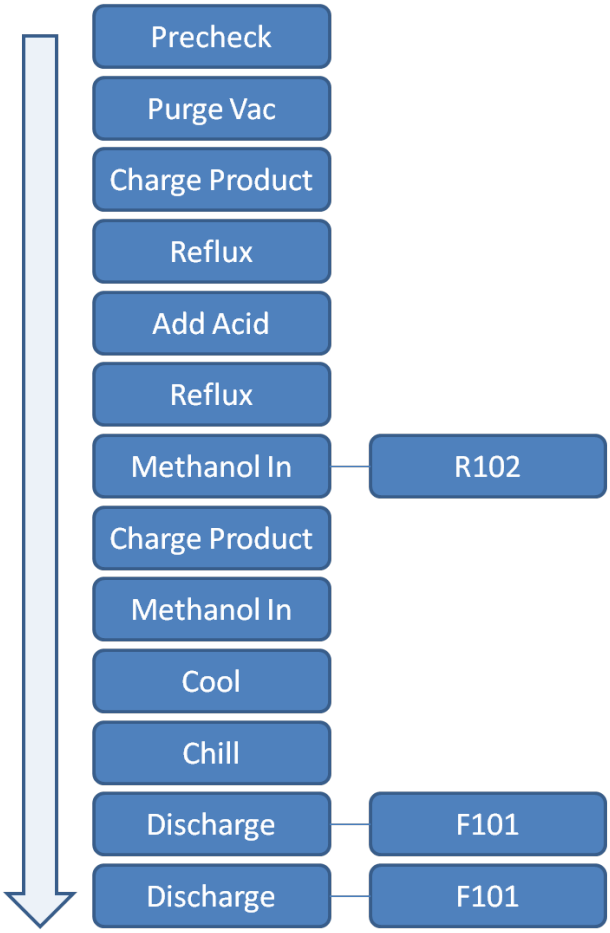


Figure 5.17: Overview of reactor R101 control strategy

For this analysis, only the 6 minute reflux, through to part way through the second reflux will be considered in the model as this is the parts of the reactor sequence where the spironolactone is made and initially isolated. This data includes the addition of the thiolacetic acid, the 20 minute reflux, primary nucleation, confirmation of crystallization, and final 100 minute reflux.

5.4.2 Control of reactor

Aspen PROVOX is a DCS (Distributed Control System) used for the control of the dryer operation. It controls the sequencing of the reactor through a recipe that is loaded for each

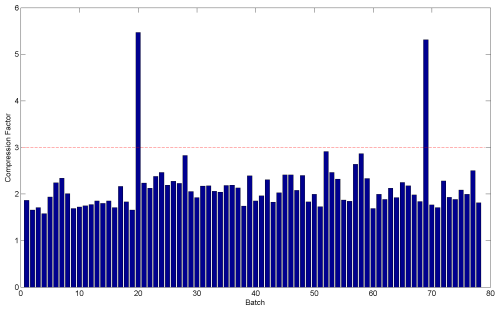
batch, it also sends signals to pneumatically actuated control valves around the actor to effect control over the reactor jacket services, jacket temperature, pressure. PROVOX also has interlocks that include those required for charging and discharging of the reactor. Although there is a lot of automatic control implemented on the reactor through PROVOX and PID (proportional + integral + derivative) controllers, some manual intervention is also required during the crystallization process and PROVOX can prompt for these interventions. More details regarding the control of the dryer are discussed in section A.1.1.

5.4.3 Compressed data

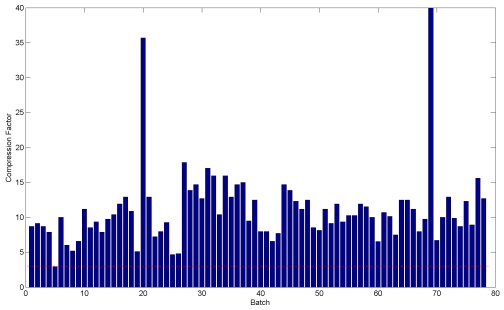
The reactor process data from the same set of 79 spironolactone (NMP) batches analysed in section 5.3 was analysed to estimate the compression factor and thus determine if the data was suitable to multivariate analysis. The estimated compression factor for each variable and each batch are shown in the following plots and summarised in table 5.7. Figure 5.18a shows the estimated compression factor of the contents temperature for each batch. The compression factor is generally below the limit of 3 proposed by Thornhill et al. (2004) at a value of approximately 2, however, there are two batches that exceed this limit with compression factors approaching 5.5. From looking at the data for these batches, they are clearly outliers and will therefore be removed from the analysis. The contents temperature data is therefore suitable for multivariate analysis from a compressed data perspective.

Figure 5.18b shows the estimated compression factor of the reactor weight for each batch. The compression factor is above the limit of 3 with most values in the region of 5 to 17. Again the two outlier batches have very high compression factors as the variable is relatively constant for these two batches. The reactor weight data is potentially too compressed to be used in multivariate analysis without introducing significant errors. Therefore the reactor weight variable should be considered for exclusion from the analysis.

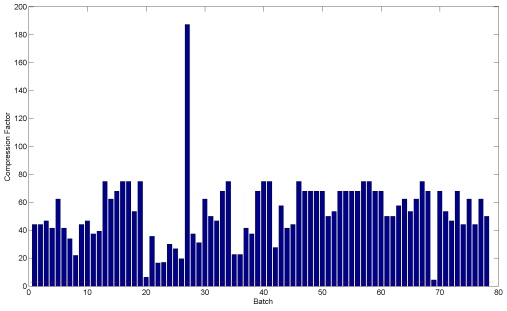
Figure 5.18c shows the estimated compression factor of the reactor temperature controller output for each batch. The compression factor is significantly above the limit of 3 with most values in the region of 20 to 80. During the sequences, heating and cooling is generally applied fully (i.e. full reflux requiring fully open steam, or rapid cooling requiring full flow on coolant), therefore the controller signal is relatively flat for most of the duration of the batch and only indicates if heating or cooling is being applied. The exception to this is when the batch is chilling at -10 °C, the reactor attempts to control this temperature to the set point



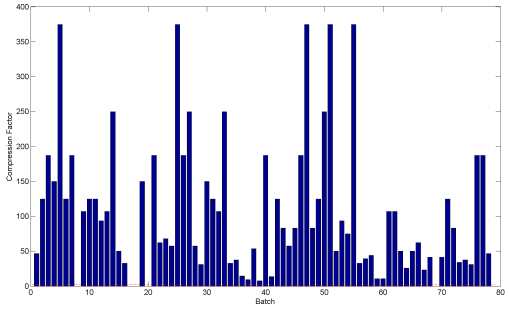
(a) Reactor contents temperature



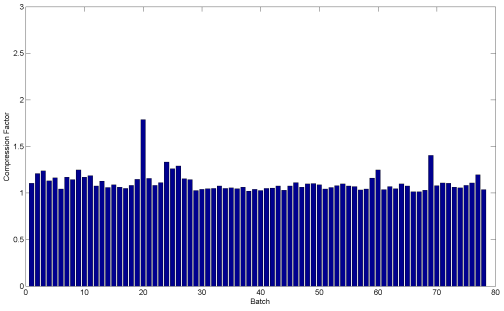
(b) Reactor weight



(c) Reactor controller output



(d) Reactor full pressure



(e) Reactor blanket pressure

Figure 5.18: Estimated compression factor for the spironolactone reactor process data for each batch

therefore there is some movement of the controller signal during this period. As the controller signal is effectively representing a set point, a high compression factor is acceptable and therefore this variable can be included in the model if required.

Figure 5.18d shows the estimated compression factor of the reactor full pressure for each batch. The compression factor is significantly above the limit of 3 with most batches showing compression factors in the range of 40 to 370. The sequence for the spironolactone reaction and crystallization is all performed at ambient pressure, however the pressure transmitter for this variable is ranged to cover vacuum up to 0 to 2.5 bara. The signal does not have the appropriate span to be sensitive enough to detect the small changes in pressure in the reactor during the operations and therefore the measurement is relatively flat throughout. There is however another pressure signal collected during processing which is more appropriately ranged where the pressure information is located. The full pressure variable should therefore be considered for exclusion from the analysis.

Figure 5.18e shows the estimated compression factor of the reactor full pressure for each batch. The compression factor is significantly below the limit of 3 with most batches showing compression factors of approximately 1. This variable is therefore suitable for multivariate analysis if required.

Table 5.7: Summary of the compression factors for the dryer data

Reactor variable	Compression factor	Outliers	Include variable?
Contents temperature	2	20 and 69 have compression factors of approximately 5	Yes, the compression factor is sufficiently low
Weight	5 - 17	20 and 69 have compression factors of 35 and 40 respectively	Consider for exclusion
Controller output	20 - 80	20 and 69 have compression factors of 180 and 4 respectively	Yes, variable is effectively measuring a set point and therefore can be included if required
Full pressure	40 - 350	Highly variable	No, compression factor too high and blanket pressure data may capture relevant information
Blanket pressure	1	None	Yes

5.4.4 Missing data and filtering

Similarly to the dryer data in analysed in section 5.3, there was no missing data present in the data set obtained, thus only filtering had to be considered in this case.

Most of the process variables from the spironolactone reactor process have some level of high frequency noise associated with the signal. However, as discussed in section 5.3 the risk of introducing distortion and correlated error into the data to remove low amplitude noise outweighs the benefits. Therefore, no filtering will be performed to remove any measurement system noise from the reactor data.

5.4.5 Data alignment

Similarly to the dryer data analysed in section 5.3, the reactor data required alignment of the time axis.

The first section of data to be cut is all of the data from the start of the vacuum nitrogen purges to the start of the 6 minute reflux. This data was removed as the focus of this analysis was on the reaction and initial crystallization of spironolactone only. The only information in this data that has been cut were related to the reactor preparation, and the transfer of the aldadiene and solvents to the reactor. The reactor preparation is not relevant for the analysis and can therefore be excluded from the model. The only relevant information in the charges to the reactor are the relative quantities of aldadiene, methanol, and acetone. These charges are however made in the aldadiene drier and another reactor manually through totalising flow meters. As the aldadiene intermediate is not quantified, there are a series of transfers, filtrations, solvent charges, and refluxes prior to the transfer into the spironolactone reactor (R101) the quantities of aldadiene, methanol, and acetone cannot be directly obtained from the reactor R101 weight during the charges. This information may be however indirectly obtained through the temperature achieved during reflux, with differing solvent and solute concentrations impacting the temperature at which the batch refluxes. The data obtained prior to the start of the 6 minute reflux may therefore be discarded from the analysis.

The second section of data to be discarded is the data post the 100 minute reflux. This focus of this analysis was on the reactive crystallization and therefore the cool down and subsequent hold following the initial crystallization and reaction completion is out of scope and will therefore be excluded from this analysis.

The resulting data still had some misalignments of the features, because, although the reactor is sequence controlled by time for the most part there are still some phases that are controlled by target, weights, temperatures, and an operator prompt to confirm that crystallization has occurred. Figures 5.19 through 5.21 show the partially aligned reactor process data for the reactor contents temperature, reactor weight and blanket pressure.

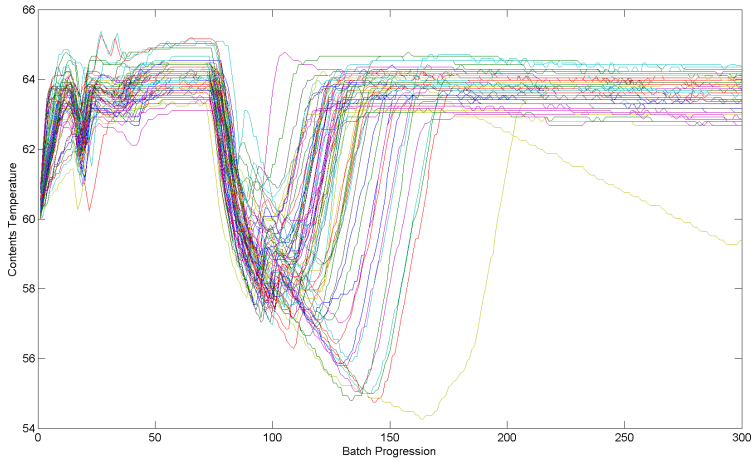


Figure 5.19: Pseudo aligned data for reactor contents temperature

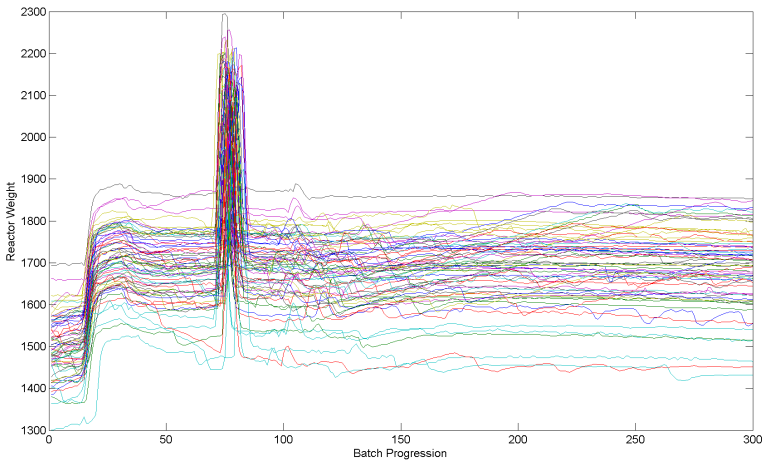


Figure 5.20: Pseudo aligned data for reactor weight

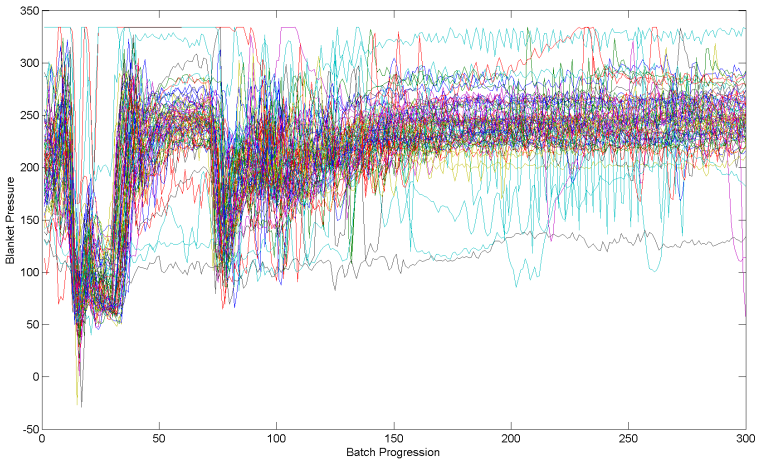


Figure 5.21: Pseudo aligned data for reactor blanket pressure

Two approaches to the alignment of the data were taken. The first was to use the above data as is, and the second to align the start and end of thiolacetic acid addition, the 20 minute reflux, the exotherm observed due to heat of crystallization, and the 100 minute reflux. Figures 5.22 to 5.23 show the aligned data for the reactor contents temperature, reactor weight, and blanket pressure. Due to the distortion of the batch time from this alignment, the aligned time variable was also included in the models.

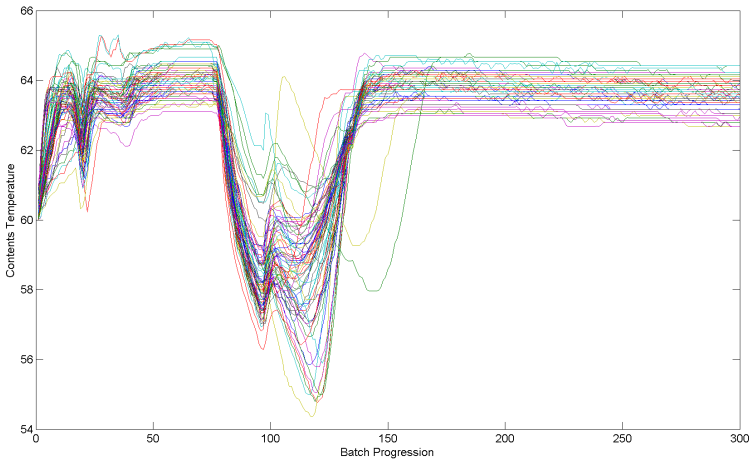


Figure 5.22: Aligned data for reactor contents temperature

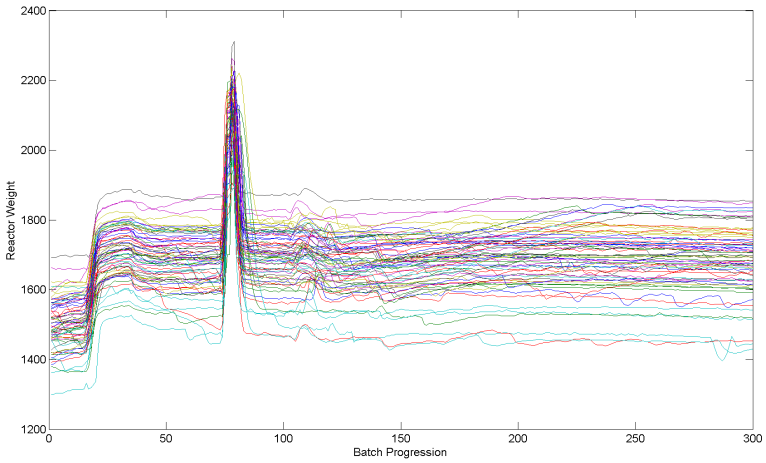


Figure 5.23: Aligned data for reactor weight

5.4.6 Pre-processing

Batchwise unfolding was employed for both the outlier detection and the exploratory modelling of the reactor data using the same approach as described in section 5.3.

Centring and scaling was again applied using the same two methods described in section 5.3, initially using all of the available data for outlier detection, and subsequently centring and scaling from a subset of the data to identify differences between groups of batches.

5.4.7 Outlier detection

Following autoscaling the dryer data and unfolding the data batchwise principal component analysis was applied to the data in order to identify any outliers in the data that should be removed prior to a more focused analysis of the data.

The number of principal components to retain in this model was 5 principal components, selected by leave one out cross-validation (figure 5.24). The model a total of 79.18% of the variance is explained with 38.82% explained in the first principal component, a further 19.70% variance explained in the second principal component, 9.53% explained in the third principal component, 6.55% explained in the fourth principal component, and 4.58% in the fifth.

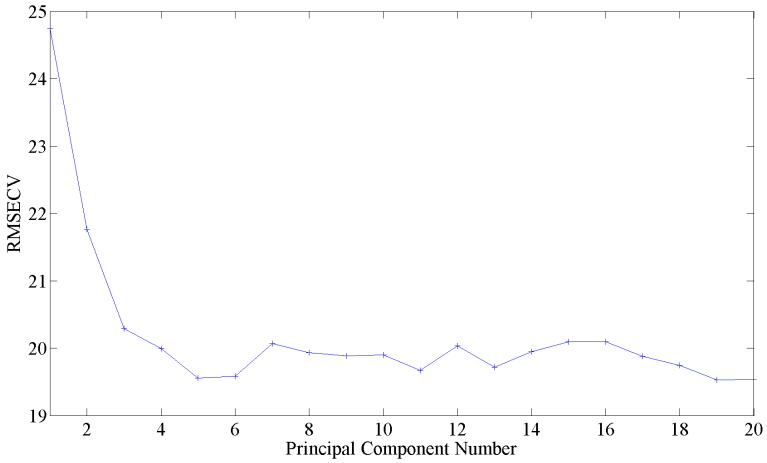


Figure 5.24: Root mean squared error for cross validation for dryer data

The influence plot in figure 5.25 identifies one batch that has high residuals (batch 39) and two that have high Hotelling's T^2 values. These batches may be potential outliers and the contribution plots should be interrogated to determine if these are true outliers.

The contributions on the residuals for batch 39 (figure 5.26a) identified the blanket pressure data to be abnormal for this batch in a number of different regions including the crystallization and 100 minute reflux (figure 5.26c), and there is also some abnormality in the temperature data during the same period (figure 5.26c). During the 100 minute deodour phase, a nitrogen sweep is operational which applies a blanket pressure across the head space of the reactor in order to remove methanethiol from the head space of the reactor as it is evolved during the reaction. The vent line from the reactor to the scrubber is open with a conservation valve in the line between the reactor and the scrubber to maintain some pressure in the reactor. The pressure is controlled through a pressure control valve on the nitrogen inlet to the reactor which is in turn controlled by a solenoid valve based on a recipe signal received via the DCS. Further details on the control of the pressure are available in section A.4 in appendix A.

Due to the correlation of the pressure rises with thermal events in the reactor, it is clear that there was a nitrogen supply fault to the reactor. The conservation valve maintains the reactor pressure at approximately 100 mmWG, however when the reactor contents are heated, through energy inputs through the reactor jacket, or energy releases during crystallization, this temperature increase increases the vapour pressure in the reactor. The conservation vent subsequently opens to slowly vent this increased pressure.

The root cause for the lack of nitrogen during this phase cannot be determined from the data

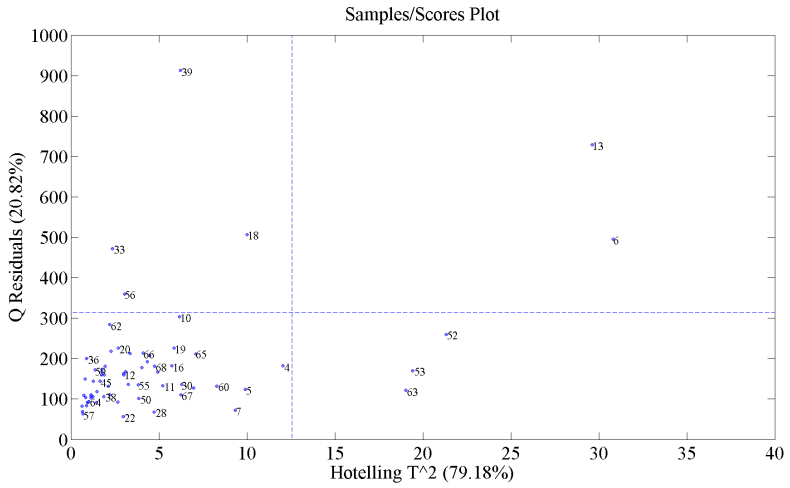
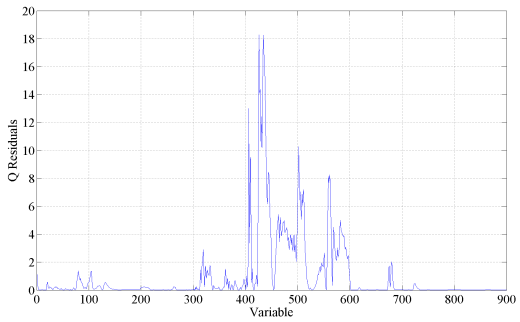
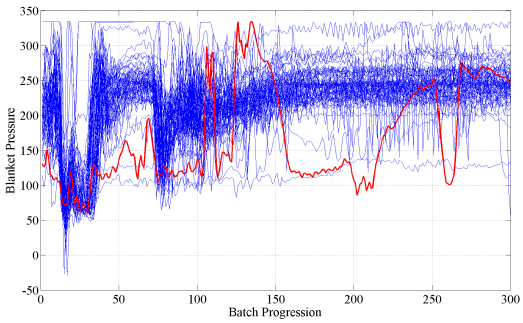


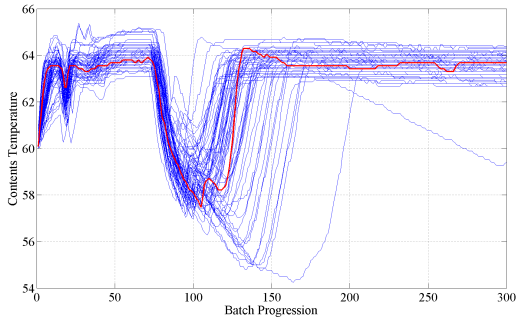
Figure 5.25: Influence plot of autoscaled reactor data unfolded in the batch direction with 95% confidence limits shown



(a) Q residuals for batch 39 reactor data



(b) Reactor blanket pressure data for batch 39



(c) Reactor contents temperature data for batch 39

Figure 5.26: Identification of batch 39 as an outlier (a) Contribution on the Q residuals for batch 39 reactor data (b) Reactor blanket pressure data for batch 39 showing atypical low pressure behaviour with pressure rises at approximately samples 100-110, 125-160, 210-250, and 260-300 (c) Reactor contents temperature data for batch 39 showing the locations of thermal events at approximately samples 100-110 (nucleation), 125 (return of reactor to reflux), 200-225 (Reflux temperature slowly increasing), and 260-300 (step change in reflux temperature following short drop in temperature)

available, however this is an atypical event, and therefore the batch should be excluded from the analysis.

In the case of batch 6, two abnormalities were highlighted in the batch. The first was a delayed nucleation event observed in the reactor. This was not a fault with the reactor, but a property of the batch, coupled with the random homogeneous nucleation event expected from an unseeded crystallization (Perry, 1997). For this alone, batch 6 cannot be removed from the analysis as this may be important in the model. The second fault however was a loss of heating to the reactor during the 100 minute reflux. This caused the temperature to drop significantly throughout the duration of this 'reflux' operation and is clearly an atypical event that. This batch may therefore be excluded from the analysis.

In the case of batch 13, there was a significant delay post the nucleation event before the operator answered the prompt to confirm that nucleation had been observed and that the sequence may proceed to the 100 minute reflux. As this batch was such an extreme delay (approximately 30 minutes) this batch may be removed form the analysis.

These batches were removed from the dataset as outliers and the process repeated with the remaining batches until the remaining batches did not have very high residuals or Hotelling's T^2 values. Although confidence limits are shown on the influence plot, these should be used as a guide only and not a hard rule as to whether a batch is an outlier or not (Bro and Smilde, 2014).

5.4.8 Principal Component Analysis

Following the removal of the outliers from the original data set of 79 batches, two subsets were generated. The first was the desirable batches (i.e. those with an extended drying time of less than 3.5 hours), and the second was the undesirable batches (i.e. those with extended drying times of greater than 7.5 hours). These subsets consisted of 9 batches and 11 batches in the fast and slow drying groups respectively.

All of the batches were centred around the mean of the batches with short extended drying times, and scaled to unit variance. By removing the means of the fast drying batches from all of the data, the non-stationary behaviour will be removed from the model as this is not of interest, however, the information on how the slower drying batches deviate from the

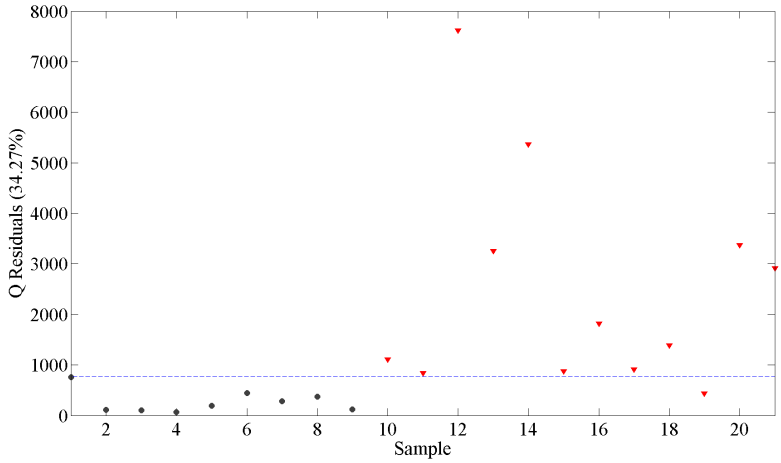


Figure 5.27: Q Residuals for the dryer data from the fast (black dots) and slow (red triangles) reactor batch data with the 95% confidence limit indicated

trajectories of the faster drying batches will be retained in the model. The data was then unfolded batchwise to allow the PCA model to be built.

As the residuals in the model will be used in this analysis, it is important to choose the correct number of principal components to retain in the model. Leave one out cross validation was used to select 2 principal components to retain.

A principal component model was built on the data with fast drying times. This explained 65.73% of the variance in the data with 46.37% accounted for in the first principal component, and the other 19.36% in the second principal component. The principal component model was then applied to the slow drying batch data. The residuals plot shown in figure 5.27 shows that the slow drying batches are different to the faster drying batches and all of the slow drying batches exhibiting higher residuals and mostly falling above the 95% confidence interval limit. Examination of the contribution plots for the residuals would be prudent for these batches to identify which variables are contributing to the differences between the two groups of data.

The contribution plot on the residuals is shown in figure 5.28. A number of regions of interest that contribute significantly to the high residuals of the batches that have long extended drying times can be seen. The largest, and therefore most significant from the data presented to the models, is around variable 140. This corresponds to the reactor contents temperature after the nucleation event. The variability in this region is generally cause by a delay in the operator confirming the crystallization has occurred and subsequently allowing the sequence to continue with the heat up and 100 minute reflux phase. Another region that is shown in the temperature

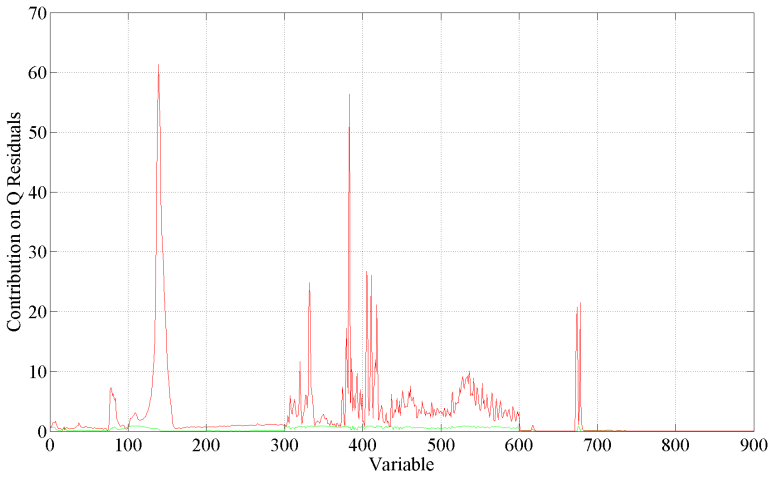


Figure 5.28: Contribution plot on the Q Residuals for the reactor data from the fast (green) and slow (red) drying batches

portion of the contributions plot is around sample 80. This corresponds to the start of the cool down prior to nucleation, and indicates that there is a differnece in the temperature data around this variable for fast and slow drying batches. Further investigation into the data presented to the model shows that the slow drying batches all entered the cool down phase prior to the fast drying batches. This is due to some misalignment of the phases within the data.

The second most influential region is found in the blanket pressure data at approximately sample 390 (figure 5.27). This again corresponds to the location of the nucleation event, and is likely due to the pressure rise in the reactor as a secondary effect of the temperature rise caused by the crystallization. The difference in the pressure and temperature could be either or both of two effects. One, a difference in temperature (and therefore pressure) rise from the crystallization event. Or two, a difference in the location of the crystallization event. In order to investigate these two events, the nucleation event needs to be aligned. The results of the analysis with the aligned features is discussed in section 5.4.9.

Another large difference between the batches with slow and fast drying is observed at approximately sample 680. This corresponds with the reactor jacket filling with and then immediately draining of cold water to initiate the cooling of the reactor to effect nucleation. The data shows that this event is also not aligned due to the variability in the duration of the thiolacetic acid charge. This variation may only be flagged as significant due to the misalignment of the data, and therefore the data should be re-analysed with the features aligned (section 5.4.9).

5.4.9 Principal component analysis with improved feature alignment in reactor data

The analysis on the reactor data indicated that the misalignment of the features in the reactor data were pulling out some variables as different in the model. There are a number of sources of misalignment in this data, the first of which is the time it takes to complete the addition of thiolacetic acid. Although this is a DCS controlled operation, where two valves in series are opened to allow the transfer of thiolacetic acid to the reactor up to a predefined loss of weight in the acid header tank. The valves are then cycled to add a series of smaller slugs of acid until the target acid charge has been achieved. All the valves are operated pneumatically, and there is some level of dynamics in the valves, and components associated with the control system that causes some variability in how fast the valves open and close. This variability, in addition to the dynamic accuracy on the header tank load cells used to control the charge may cause the reactor to require a different number of pulses of acid to achieve the targeted charge, and thus adding variability in the duration of the charge phase.

In order to identify if these were genuine differences in the correlation structure of the measured variables, or just due to the misalignment, the reactor data for this set of models was aligned using the second approach detailed in section 5.4.5.

The same modelling approach with batchwise unfolding and autoscaling for outlier detection, followed by unfolding and centring and scaling based in the data with fast drying times was applied. The difference between these models and the reactor models previously discussed was the inclusion of the augmented time variable as an indicator of batch progression.

A plot of the contributions on the Q residuals is shown in figure 5.29. The most obvious difference between the residuals plots for the misaligned and aligned data is the reduction in the relative contributions for variable 140 which is significantly reduced for the aligned data. This variable is still somewhat significant in the Q residuals indicating that there is remains a genuine difference in this region. This is caused by the delay from the point of nucleation for the operator to acknowledge nucleation and allow the reactor to stop cooling and return to reflux.

Another difference in the temperature region of the data is the disappearance of the peak at sample 80. This is because the data here is the temperature from the reduction in temperature through active cooling at the start of the nucleation phase. By aligning this data, it shows that the data is no longer different between the fast and slow drying batches.

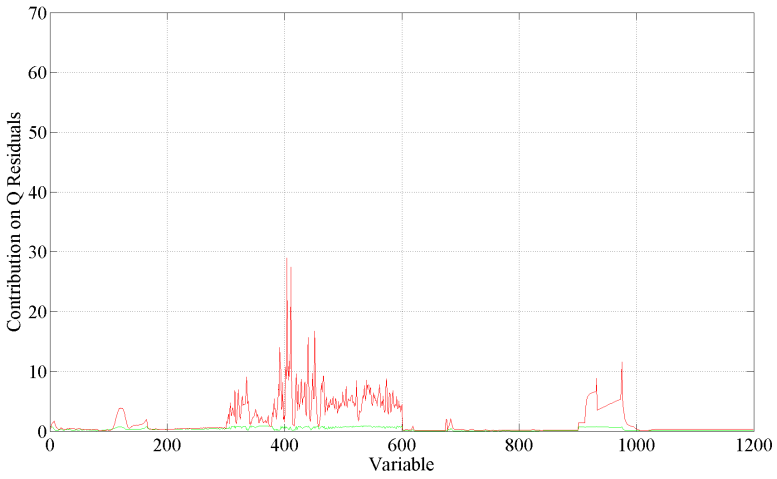


Figure 5.29: Contribution plot on the Q Residuals for the reactor data from the fast (green) and slow (red) drying batches for the data with feature alignment applied

The final difference in the temperature region is the relative increase in the residuals for the slow drying batches around sample 110. This corresponds with the heat rise from the nucleation event in the reactor. Now that this event is aligned between all of the batches it confirms that there is a difference in temperature difference in the reactor arising from the enthalpy release. This may be indicative of the degree of supersaturation of spironolactone at the nucleation event. This may be due to a poorer yield from the upstream processes which are unaccounted for resulting in a higher relative solvent charge. It may also be due to a difference in the ratio of solvents, as spironolactone is more soluble in acetone than methanol. It is not possible to determine if this is due to an increased yield of aldadiene and therefore spironolactone as this data is not collected. Neither is the information on the solvent charges collected, however the most significant contribution to the residuals shown in figure 5.29 remains to be from the blanket pressure data (samples 301 - 600). These are consistently high throughout the batch with the exception of samples in the region 350 - 390. This is the region where cooling is applied to the reactor.

The highest contributions in the pressure region come from approximately sample 400 which corresponds approximately to when heat is being applied to the reactor following nucleation to return the reactor to reflux. For the remaining samples in the pressure region have a relatively constant, and high, contribution. These are all the periods where the reactor is at reflux. The pressure data for these batches shows that there is a slight increase in blanket pressure for the batches with slow drying batches over those batches that dried quickly. This could indicate that there was a consistent fault on all of these batches with the pressure control either applying

slightly higher nitrogen flow to the reactor, or the vent valve maintaining the pressure slightly higher. However, a more likely explanation is that the ratio of the two solvents (methanol and acetone) was different between the two sets of batches.

The boiling point of acetone is lower than that of methanol (56.2 °C and 64.6 °C respectively (Amer et al., 1956)). changing ratios of these two solvents will therefore influence the vapour pressure of the system. The observation shown in the contribution to the residuals for the pressure data that may relate to solvent ratio differences is also supported by a slight increase in the baseline of the contributions on the residuals across the temperature samples indicating different boiling points of the systems.

By looking at the contributions on the residuals from the weight data, it is obvious that the high contribution around sample 680 has significantly decreased. This is due to the alignment of the addition of water to the jacket to initiate the cool down for nucleation. There is still a small amount of variability in this area, however this is relatively small, and is shown in both the fast and the slow drying batches indicating that the information in the model does not fully capture the variability in the data at this point. The reason for the variability in the data around this region is likely to be due to variation in the quantity of water filled to the reactor jacket. During this operation, the DCS fill the jacket with water until a high level switch in the jacket is activated. This prompts the DCS to isolate the cooling water and then drain the reactor jacket. There will be inherent variability in when this level switch is activated, and in the response times for the relevant valves. In addition to this, the weight data of the reactor collected here will be subject to the dynamic accuracy if the load cells. The PCA model is therefore unable to capture this variability with the number of retained principal components.

Finally, the contributions from samples 901 through to 1200 relate to the aligned batch progression index (i.e. time) of the data. There is a region between sample 913 and sample 980 where the addition of thiolacetic acid and subsequent return to reflux for 20 minutes is performed. This is capturing the information that for the slow drying batches, it took longer to add the thiolacetic acid and therefore longer for the batch to return to the reflux temperature as it had cooled further, than for the batches that dried quickly.

5.4.10 Conclusions

The proposed modelling framework (section 5.2)has successfully been applied to the reactor process data and identified some of the challenges with pre-processing such data. The analysis

was able to identify a number of batches that were labelled as potential outliers from the model. Some of these were attributed to process equipment malfunctions or extreme delays due to operator availability, and were therefore excluded from the analysis. However not all of the batches flagged as potential outliers were excluded from the analysis as some were abnormalities within the batch, such as a retarded nucleation event not attributable to an equipment malfunction and therefore potentially interesting to leave in the data set.

A comparison between two methods of alignment showed how important process feature alignment is in order to extract understanding from the models.

The analysis on the reactor data also identified several potential causes for differences in the drying time of spironolactone batches including:

- A difference in blanket pressure of the batches throughout the reflux phases of the batch. This is believed to be related to the composition of the solvents in the reactor.
- A delay in the time taken for the addition of the thiolacetic acid and subsequent additional heating time required to reach reflux for the slow drying batches. This may have an impact on the extent of the reaction prior to cooling for crystallization.
- A difference in the temperature rise in the system between batches due to the enthalpy of crystallization. This may also be related to both of the above points as a different solvent ratio may alter the solubility of spironolactone, and therefore the degree of supersaturation during crystallization. Also, if the reaction is allowed to proceed for longer prior to crystallization, the degree of supersaturation (at the same temperature during cooling) may increase. Both of these could cause an increase in the temperature rise observed as more nucleation would occur at this point.

5.5 Chapter summary

In this chapter, the framework described in earlier was tested on the historical spironolactone batch process data. Initially analysis of the dryer data was presented, describing how the data is first pre-processed and outliers are detected, and as a result identifying some parts of the processing equipment and control system that could be improved to obtain better reliability and efficiency of the drying process. Subsequently, multivariate modelling was applied to the data identifying changes in a product characteristic leading to variability in the endotherm

observed towards the end of the drying process. This variability may have been a result of the isolation process and therefore the reactor data was subsequently studied.

The data from the reactor was analysed within the same framework, and demonstrated the importance of alignment of the process data. The results of the modelling indicated that there was variability in the data that may be attributed to difference in the solvent ratios in the reactor, and variability in the time of addition of a key reagent. Differences in crystallization were also observed, which may be attributed to the variability in solvent ratios and reagent addition.

Overall the framework was successfully tested on the spironolactone process data leading to a hypothesis that variability in the reactive crystallization was resulting in differences in how the isolated spironolactone performs in the dryer. This hypothesis is tested in chapter 6.

Chapter 6. Spironolactone crystallization study

The PCA modelling of the dryer data in section 5.3 showed that the variation in the endothermic behaviour towards the end of the drying cycle was significant with regards to the variability in drying time for the product. The subsequent PCA modelling of the reactor data showed that there may be differences in the solvent ratios during the crystallization process that has an impact on the drying time of the product.

Spironolactone is known to exhibit polymorphism and can form solvates (Marini et al., 2001; Agafonov et al., 1989; El-Dalsh et al., 1983; Espeau et al., 2007; Nicolaï et al., 2007; Agafonov et al., 1991; Salole and Al-Sarraj, 1985) as discussed in chapter 2.1.3. This impact of solvent ratio on drying time alongside the variability in the endotherm observed in the deodour phase in the dryer could indicate that there is a variable degree of solvate formed batch to batch. The endotherm in the dryer may be indicative of a solid state transformation such as a desolvation or form conversion.

6.1 Objectives

A set of experiments were devised to investigate the impact of solvent ratio on the spironolactone crystallization. The objectives of this study were to:

- Determine if solvent ratio has an effect on the quantity of spironolactone that can be crystallized
- Determine if the solvent ratio impacts the particle size of the isolated spironolactone
- Determine if the solvent ratio has an effect on the form of the spironolactone isolated in the crystallization
- Determine if the solvent ratio and the degree of supersaturation has a significant impact on the particle size of the isolated spironolactone and the crystal form that is isolated.

In order to achieve these objectives a simplified isolation procedure was used. Rather than performing a large number of complex experiments converting aldadiene into spironolactone and simultaneously isolating the spironolactone, a recrystallization was performed to remove the uncertainty around the supersaturation which was instead controlled by solvent ratio and cooling rate.

6.2 Experimental design

6.2.1 Preparation of the RC1 reactor

Several experiments were performed back to back in the RC1 as the experiment only involved heating and cooling of spironolactone in methanol and acetone. The three factors that were being changed in the experiment were the cooling rate, the acetone to methanol ratio, and the concentration of spironolactone. Two of these factors, solvent ratio and spironolactone concentration, were changed through additions to the RC1. Due to the small probe length of one of the instruments however the RC1 was required to be operated at full working volume for every experiment. This meant that the RC1 needed to be completely discharged and set up again in order to change the solvent ratio or the spironolactone concentration. For each set of experiments, a concentration of spironolactone and a solvent ratio was selected and the cooling rate was changed. The spironolactone was completely re-dissolved by heating the RC1 to reflux and holding at reflux for a period of time between each experiment in the set.

Spironolactone API was obtained from Piramal healthcare from a single location in a single batch. Portions of this sample were used for each experiment on the RC1 reactor. The RC1 was set by charging a known quantity of spironolactone followed by the measured quantities of methanol and acetone from the plant at the desired ratio. The RC1 was instrumented with a glass thermocouple inside the reactor to monitor the contents temperature. There was also a thermocouple monitoring the temperature of the RC1 jacket heat transfer oil. A Mettler Toledo Focused Beam Reflectance Measurement (FBRM) probe was inserted into the reactor positioned slightly above the axial flow impeller in the RC1. Additionally a Near Infrared (NIR) transflectance probe was mounted into the reactor, however due to the short length of the probe it was mounted approximately one inch below the surface of the liquid. An Attenuated Total Reflectance (ATR) Mid Infrared (MIR) probe was also mounted in the reactor toward the bottom of the vessel. A glass baffle was also installed in the RC1 to aid mixing. Figure 6.1 shows the placement of the instruments in the RC1 reactor.

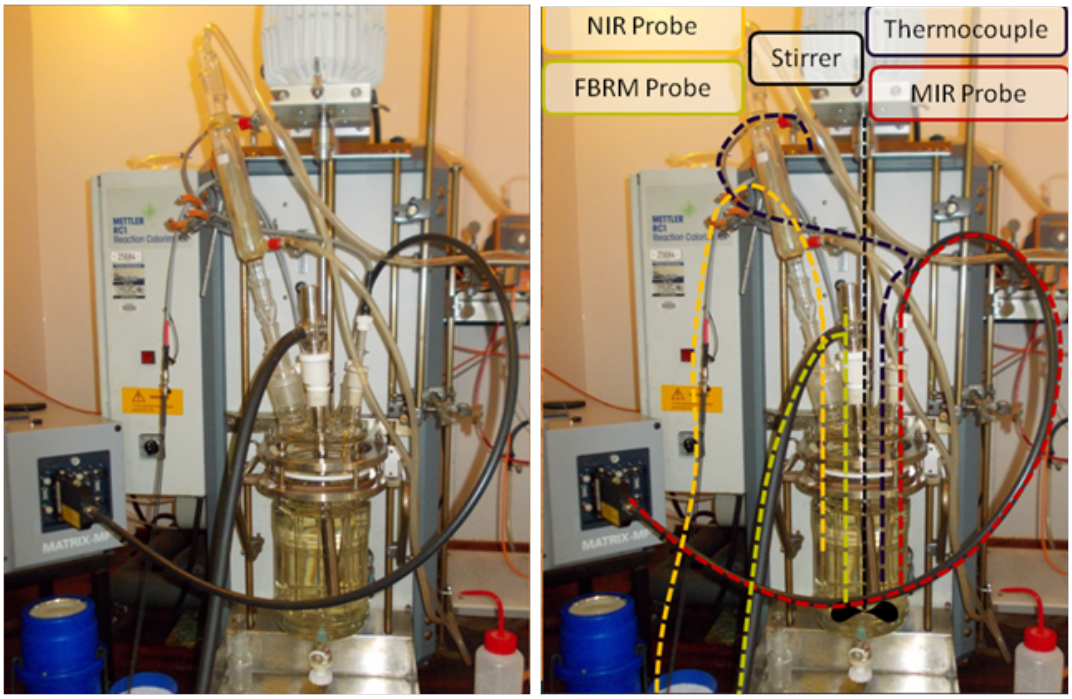


Figure 6.1: RC1 reactor with FBRM, ATR-FTIR, and NIR instruments

6.2.2 Typical RC1 crystallization procedure

After the RC1 was set up, it was programmed to perform a series of heating and cooling cycles. A typical cycle consists of firstly to ramp the stirrer speed up to 600 rpm to provide the flow required for the FBRM to measure the chord lengths (approximation of particle size). The next step is the dissolution of the spironolactone through heating the reactor contents. The heating must be controlled through the RC1 jacket temperature and not the contents temperature because the boiling point of the solvent will change as the ratio between methanol and acetone changes. The jacket temperature is therefore ramped to 70 °C (above the boiling points of both methanol and acetone) and held for 2 hours to ensure that the spironolactone is completely dissolved. The RC1 is fitted with a condenser to allow the solvent vapours to reflux back into the RC1. The final step is the controlled cool for crystallization. This is controlled using the RC1 contents temperature and is programmed with a final temperature of 25 °C and the cooling rate required for the experiment. Figure 6.2 shows a summary of the protocol followed for each experiment with the variable parameters for each experiment tabulated in table 6.1.

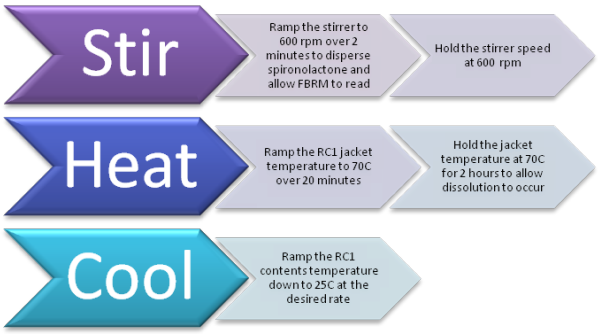


Figure 6.2: RC1 reactor protocol summary

Table 6.1: Spironolactone crystallization NIR study experimental parameters

experiment Number	Cooling Rate $^{\circ}$ C min^{-1}	Solvent Ratio Methanol:Acetone	Spironolactone g
M1_a	0.700	18:1	85.9
M1_b	0.700	18:1	85.9
M1_c	0.013	18:1	85.9
M1_d	0.700	18:1	85.9
M2_a	0.700	15:1	85.9
M2_b	0.700	15:1	85.9
M2_c	0.350	15:1	85.9
M2_d	0.700	15:1	85.9
M3_a	0.710	12:1	85.9
M3_b	0.120	12:1	85.9
M3_c	0.360	12:1	85.9
M3_d	0.070	12:1	85.9
M3_e	0.530	12:1	85.9

6.2.3 Deviations from the design

Ideally the experiment would be designed to use the same initial spironolactone concentration in the solvent system , however, spironolactone is only slightly soluble in methanol, whereas aldadiene (the precursor to spironolactone) is more soluble. Therefore, the spironolactone will not fully dissolve in the solvent system. A reduced concentration of spironolactone was therefore used to ensure complete dissolution at the start of each experiment. This is unlikely to impact the results of the experiment as the objectives of the study are how the solvent ratios affect the the crystallization of spironolactone, and during the full reaction-crystallization, not all of the spironolactone is immediately available as the reaction does not immediately go to completion. Any conclusions from this study could be verified using the full concentration of the spironolactone with the full reaction and crystallization performed.

6.3 Instrumentation

6.3.1 Focused Beam Reflectance Measurement (FBRM)

Focused Beam Reflectance Measurement (FBRM) is an instrument to measure particle and droplet size. The instrument works by focusing a laser on a sapphire window at the end of the probe which is in contact with the suspension. Pneumatics operate the optics which cause the focused laser to rotate at a high speed, scanning the suspension in contact with the sapphire window. When the laser hits a particle it is reflected back into the probe and recorded. The probe measures the chord length of particles; that is the distance that the laser was continuously reflected back into the instrument. The chord length does not give particle size as it is measuring a curved path across the surface of a particle, and depending how a particle is orientated the laser may not see the maximum diameter of the particle (figures 6.3-6.4). However, the chord length distribution gives an approximation of the particle distribution and shape and can be used for comparison between experiments.

FBRM was selected to give information on the particle size distribution throughout the crystallization experiments.

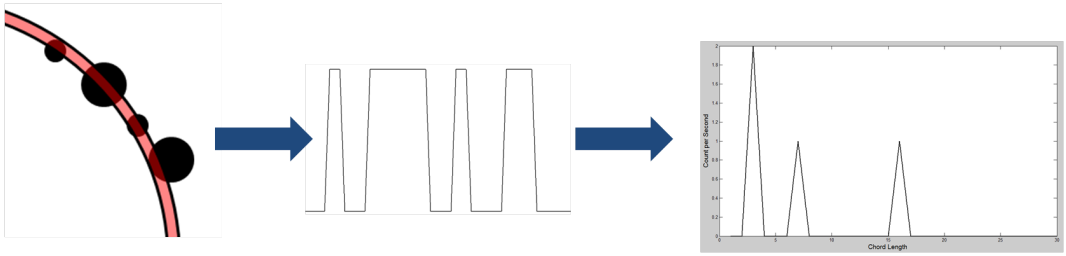


Figure 6.3: FBRM measurement principle

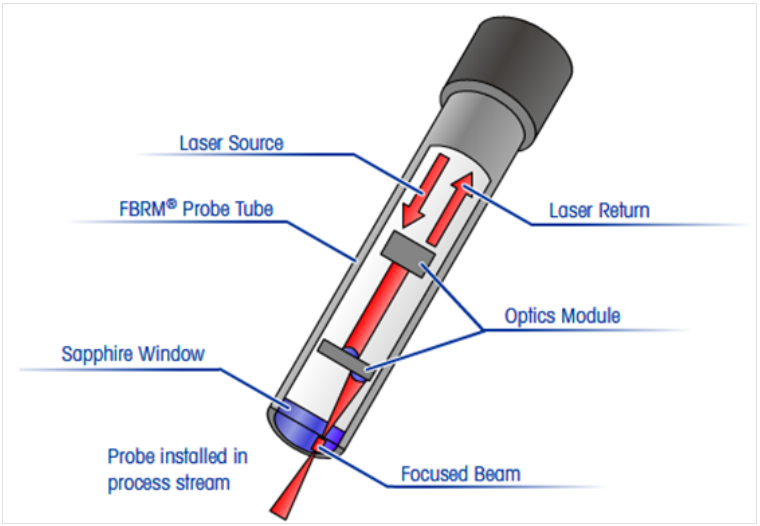


Figure 6.4: Diagram of FBRM probe operation

6.3.2 Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy ATR-FTIR)

Mid Infrared Spectroscopy (MIR or IR) is a vibrational spectroscopy technology based on the mid infra-red part of the electromagnetic spectrum (wavenumbers 4000 cm^{-1} to 400 cm^{-1} (Fujiwara et al., 2002; Fevotte, 2002; Pöllänen et al., 2006a; Sun, 2009)). Vibrational spectroscopy works on the basis that molecular bonds vibrate and as a result are able to absorb infrared frequencies corresponding to the frequency at which the bond vibrates. Bonds may exhibit six different modes of movement including symmetric stretching, antisymmetric stretching, bending, wagging, twisting, and rocking (figure 6.5). The frequency of vibration depends on the vibrational mode in addition to the mass of the atoms involved in the bond with smaller, lighter atoms having a higher frequency than larger, heavier atoms. The mode of vibration that is available also depends on the atomic structure and the available degrees of freedom for vibration in that mode.

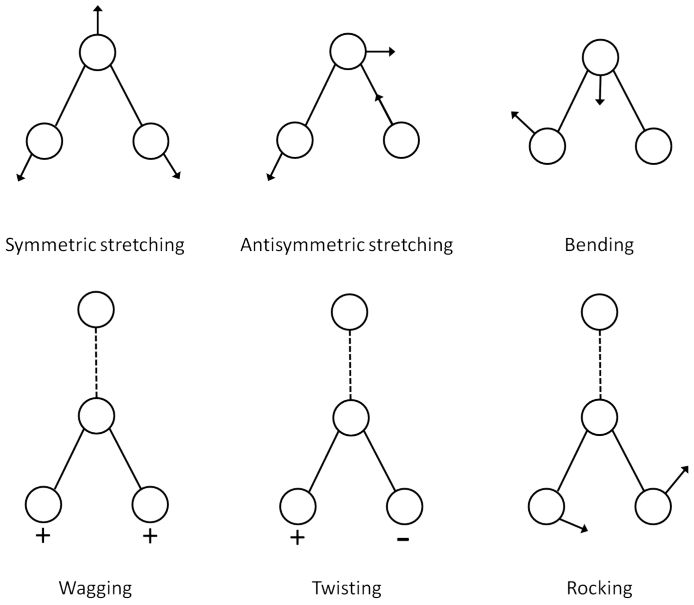


Figure 6.5: Vibrational modes (Sun, 2009).

Fourier Transform IR (FTIR) spectroscopy is a technique that uses the interference properties of light to generate a spectrum over a large range of wavenumbers in a very short amount of time. An interferometer is a device that creates a phase shift in the infrared light to cause constructive and destructive interference at different wave numbers. This is achieved by taking an infrared source and passing the beam of light through a beam splitter after which a portion of the light is directed towards a fixed position mirror and the other portion is directed towards a moving mirror. The split beams are then reflected off their respective mirrors and meet at the beam splitter again where they are combined and travel towards both the detector (through a sample of interest) and the source (figure 6.6). Because part of the beam that is combined is reflected off a moving mirror the path length the light has to travel can be changed and cause constructive or destructive interference according to the path difference and the wavelength of the light. When the path difference of the light is equal to (an integer multiple of) its wavelength the result is complete constructive interference and therefore the observed intensity is increased. When the path difference of the light is (an integer multiple of) half its wavelength complete destructive interference is observed (zero intensity). Any other path differences result in either partial constructive or partial destructive interference of the light (figure 6.7). By moving the mirror at a constant speed and measuring the intensity of light at the different wavelengths (wave numbers) whilst the mirror is moving, a very quick scan of the entire spectrum can be obtained. The resulting plot of path length (mirror displacement) against intensity is known as an interefereogram. The interefereogram can then be easily

transformed to a spectrum through a Fourier transform which takes the time (space) domain intensity and transforms it into the frequency domain intensity.

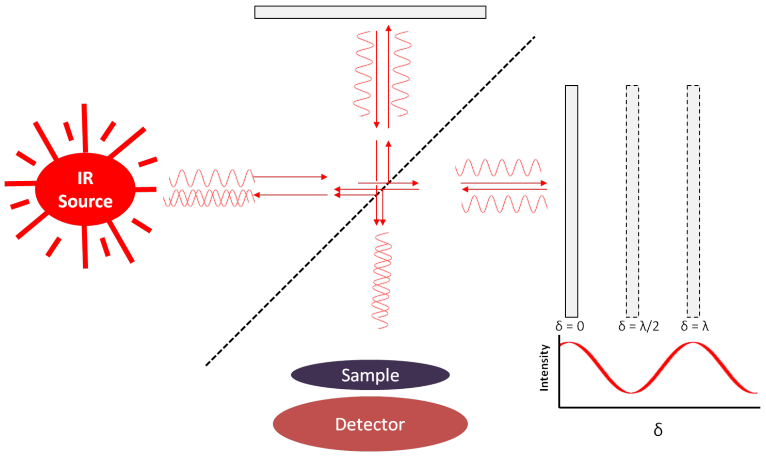


Figure 6.6: Principles of interferometry

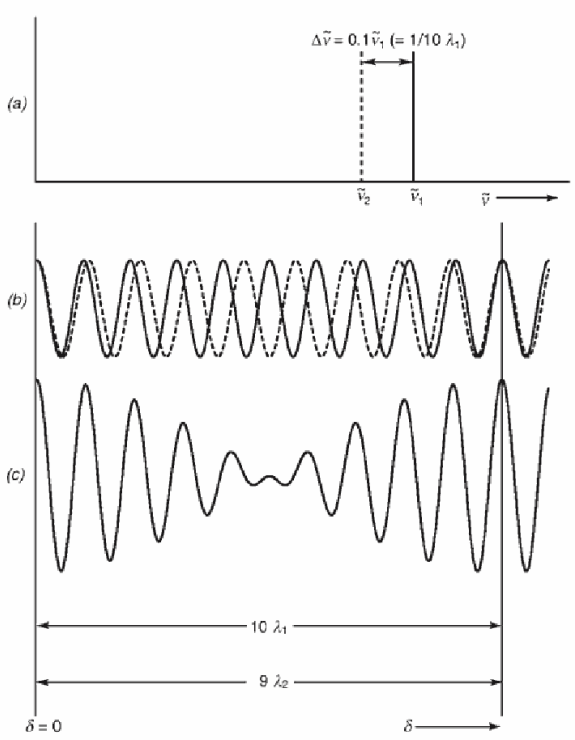


Figure 6.7: Example of how multiple wavenumbers produce information in the combined interefereogram (De Griffiths and Haseth, 2007).

Attenuated Total Reflectance (ATR) is a technology whereby MIR waves are subject to total internal reflection in a diamond crystal at the end of the probe. A phenomenon known as an evanescent wave is then produced that is able to penetrate into the sample. The evanescent wave is an electrical field that decays exponentially as the distance from the total internal reflection increases. The penetration depth of the evanescent wave is function of both the

material it is passing through and also the wavelength of the light that created it. The spectra is scaled to remove the effects of the relationship between wavelength and penetration depth so that only the interaction between the evanescent wave and the sample into which it penetrates remains.

ATR-FTIR was selected to monitor the concentrations of the solvents and spironolactone in solution throughout the experiments.

6.3.3 Transflectance Near Infrared Spectroscopy (NIR)

Near Infrared Spectroscopy (NIR) is a vibrational spectroscopy technology based on the near infrared part of the electromagnetic spectrum (wavenumbers 12000 cm^{-1} to 4000 cm^{-1} (Sun, 2009)).

There are three modes in which NIR spectroscopy can work, reflectance, transmission and transflectance. Reflectance spectroscopy works by detecting the NIR wave reflected off the sample however can only give information on the surface of a material. Transmission spectroscopy works by detecting the NIR wave transmitted through the sample and therefore allows for samples that are less uniform as the light passes through the full depth of the sample. Transflectance is a combination and works by detecting the NIR wave reflected off the sample and transmitted through the sample. In this work the transflectance mode was used due to the suspended solids in the sample during crystallization.

An NIR instrument was included to supplement the ATR-FTIR instrument in giving concentration information throughout the experiments. In addition to this, the NIR spectra may hold some information on the particle size which would be complementary to the FBRM data, and NIR may also be able to give some information on any polymorphic changes in the systems throughout the crystallization or between experiments.

6.4 Pre-processing of spectroscopic data

NIR spectra can be significantly influenced by light scatter during collection causing non-linearities in the spectra. This may be a result of temperature, or suspended particles in the system. Methods to remove such non-linearities are crucial in the pre-processing of the spectroscopic data before chemometric modelling is performed (Rinnan et al., 2009). There are a number of techniques widely known in the literature that can be employed to reduce or

eliminate the effects of scatter on spectral data, the most popular of which will be briefly discussed below.

6.4.1 Baseline correction

Baseline correction is a method of de-trending in which random offsets from spectra can be removed. The baseline can be removed from the spectra by fitting a low order polynomial through the spectra and subtracting the resulting function's curve from the spectra. The low order polynomial can be fitted through reference points on the spectra where little spectral information is present, or alternatively applied to the entire spectra if it is not possible to identify a low information region (Siesler et al., 2002).

6.4.2 Multiplicative scatter correction

Multiplicative Scatter Correction (MSC) is used to remove multiplicative and additive scatter effects from spectra. It works by using a reference spectra, which may also be the average spectra of a calibration set, which is used to calculate the scalar intercept and slope terms for each spectra to be corrected. This is achieved by fitting a least squares model between the reference spectra and the unseen spectra. The result of MSC is to remove background offsets and slopes from spectra, whilst maintaining the spectral features. One of the main challenges in MSC is to identify a suitable reference spectrum. This may be a spectrum collected under controlled conditions during a calibration procedure, or a composite mean spectrum from a calibration data set (Rinnan et al., 2009). There are a number of variations to MSC which are beyond the scope of this review. For more information the reader is directed to Rinnan et al. (2009).

6.4.3 Standard normal variate

Standard Normal variate (SNV) is a scaling technique that is equivalent to autoscaling the rows of the spectral data matrix. The corrected spectrum is the original spectrum with the mean subtracted and divided by the standard deviation. The result is similar to the MSC technique, removing the background offsets and slopes from the spectra whilst maintaining the spectral features. SNV employs means and standard deviations rather than least squares fitting, and can

therefore be more sensitive to noise in the spectra, however it is a simpler technique to employ as it does not require a reference spectrum (Rinnan et al., 2009).

6.4.4 Spectral derivatives

Taking derivatives of spectra can be employed to two main effects. Firstly, taking the first derivative of a spectrum will remove the baseline of the spectra, whilst employing second derivatives also removes linear trends. Secondly, applying derivatives to spectral data can help resolve overlapping peaks, at the expense of making the spectra more difficult to interpret. One of the most popular methods is Savitzky-Golay derivation. This methods applies a symmetrical window over the point to find the derivative and fits a polynomial to this window. From this the derivatives are calculated analytically. The result of this window means that the method also has smoothing properties on the spectra and can therefore reduce noise,with the larger the window, the greater the smoothing effect (Rinnan et al., 2009).

6.5 Application of process analytical technology to crystallizing systems

The application of techniques such as FBRM, NIR, and ATR-FTIR to the study of crystallization systems is well documented in the literature (Fevotte, 2002; Fujiwara et al., 2002; Yu et al., 2004; Pöllänen et al., 2006a; Kadam et al., 2010; Sarraguca et al., 2015).

Kadam et al. (2010) presents a comparative study of NIR and ATR-FTIR for the study of cooling crystallization. This results indicated that the ATR-FTIR performed better as the instrument was not as susceptible to fouling. Additionally, the statistical models were found to be more accurate for the ATR-FTIR instrument when extensive model calibration was performed, compared to the NIR instrument.

Yu et al. (2004) presents a good summary of process analytical techniques and their uses in crystallization studies. Two methods are presented as suitable for monitoring supersaturation, NIR and ATR-FTIR. Additionally, Yu et al. (2004) discusses how ATR-FTIR has been used to monitor polymorphic form, however this is as an offline technique on the solid phases, as the primary role in a slurry is to monitor the solution phase. Several techniques are discussed in relation to monitoring particle size including FBRM, and to some extend NIR. It is also noted that for monitoring particle shape, techniques that involve particle imaging are required such as Lasentec's PVM instrument. The application of PAT to monitor polymorphic form is

discussed and includes techniques such as NIR, Raman spectroscopy, x-ray powder diffraction and solid state NMR. Finally, some chemometric methods for analysing the data from PAT are presented before some case studies are presented looking at design of process control using FBRM and ATR-FTIR, monitoring of crystal shape with PVM, and monitoring of polymorphic form using Raman, NIR, and x-ray diffraction.

Fevotte (2002) discusses the application of ATR-FTIR for monitoring crystallization progress and highlights the ability of this technique to distinguish between polymorphic forms and solvates. Fujiwara et al. (2002) present how ATR-FTIR can be used to monitor the progress of crystallization processes and measure the metastable zone width under different crystallization conditions. This work was presented on a paracetamol crystallization system. Pöllänen et al. (2006a) present a study on the crystallization of sulphathiazole utilizing ATR-FTIR to monitor the concentration of the solute during the crystallization.

Another application of ATR-FTIR and NIR on a crystallization process is presented in Sarraguca et al. (2015) for the monitoring of a cocrystallization process where the ATR-FTIR was applied as an online instrument, whereas the NIR was employed offline on the isolated crystals, alongside other complementary techniques including x-ray powder diffraction and differential scanning calorimetry.

Since the crystallization study presented in this thesis was performed, Jiang et al. (2015) has published work using ATR-FTIR in addition to Raman spectroscopy to monitor the solid state transitions of spironolactone, however this was of the form I to form II transition via the 1-propanol and 2-propanol solvates. These tools were applied online in addition to standard offline analysis such as microscopy and powder x-ray diffraction.

6.6 Results

6.6.1 Temperature

The first group of experiments (M1_a to M1_d) were performed with a methanol to acetone ratio of 18:1, which is the typical solvent ratio used on the plant. Four crystallizations were performed with the first, second and fourth experiment at a cooling rate of 0.7 °C min^{-1} and the third experiment at a cooling rate of 0.013 °C min^{-1}.

By looking at the temperature data collected during the crystallization experiments the nucleation temperature can easily be identified. Figure 6.8 shows the rate of change of reactor

contents temperature plotted against the reactor contents temperature for the M1 series of experiments. The first two crystallizations with cooling rates of 0.7 °C min^{-1} show an exotherm due to the crystallization at approximately 52.6 °C, whereas in the fourth crystallization, with the same cooling rate, the crystallization occurred later at approximately 51.9 °C. This may be due to the random nature of homogeneous crystallization resulting in some variability in the crystallization. Unlike heterogeneous nucleation where seed crystals are added to the system to initiate nucleation, homogeneous nucleation can be influenced by a number of factors. Similar behaviour has been widely reported in crystallization studies in the literature (Kulkarni et al., 2013), where the induction time for crystallization showed variability for the same sample with the same degree of supersaturation. This was attributed to the stochastic nature of nucleation. In the spironolactone experiments where the concentrations remained unchanged, good control over temperature and agitation rate was achieves, and between experiments the system was refluxed for 2 hours to ensure complete dissolution had occurred, the most likely cause of the variability seen is also due to the stochastic nature of homogeneous nucleation.

Figure 6.8 also shows the temperature of the crystallization event for the M1_c experiment with the slower cooling rate occurred at a higher temperature (approximately 54.8 °C). This is due to the slower cooling rate allowing the kinetics of crystallization to cause the nucleation event at a smaller degree of supersaturation, compared to the systems that were cooled more rapidly. The temperature change at this event is also smaller than the equivalent event for the same system when cooled faster. This is likely due to two causes. The first is that the degree of supersaturation is less at the higher nucleation temperature and therefore less nucleation will occur at this location and consequently a smaller quantity of energy will be released. Secondly, due to the slower cooling rate, the control system on the reactor is able to apply relatively more cooling compared to the systems that are cooling faster arresting the observed exotherm in the reactor contents temperature.

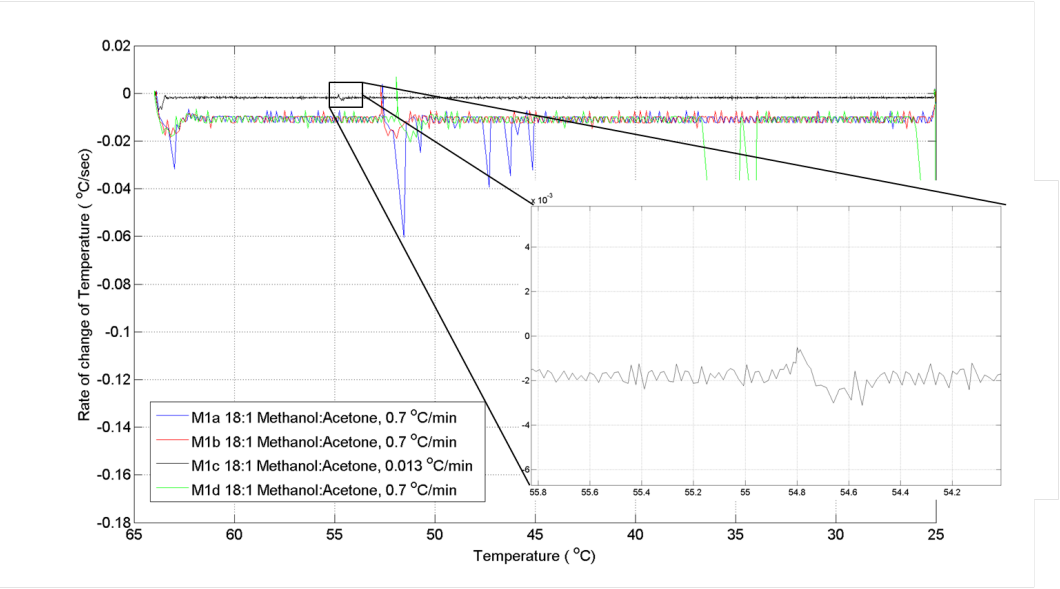


Figure 6.8: Rate of change of temperature vs. reactor contents temperature during crystallization experiments for system M1. Inset, blown up portion of the M1_c experiment to show the location of crystallization.

The second group of experiments (M2_a to M2_d) were performed with a methanol to acetone ratio of 15:1 to assess how the increase in acetone concentration can affect the crystallization of spironolactone. Four crystallizations were performed with the first, second and fourth experiment at a cooling rate of 0.7 °C min^{-1} and the third experiment at a cooling rate of 0.35 °C min^{-1}.

Similar to the M1 group of experiments, for the same cooling rate in the M2 group of crystallizations the nucleation occurs at approximately the same temperature with a small amount of variability (between 51.3 °C and 51.7 °C). The slower cooling rate for experiment M2_c again shows nucleation at a higher temperature and with a smaller impact on the cooling rate.

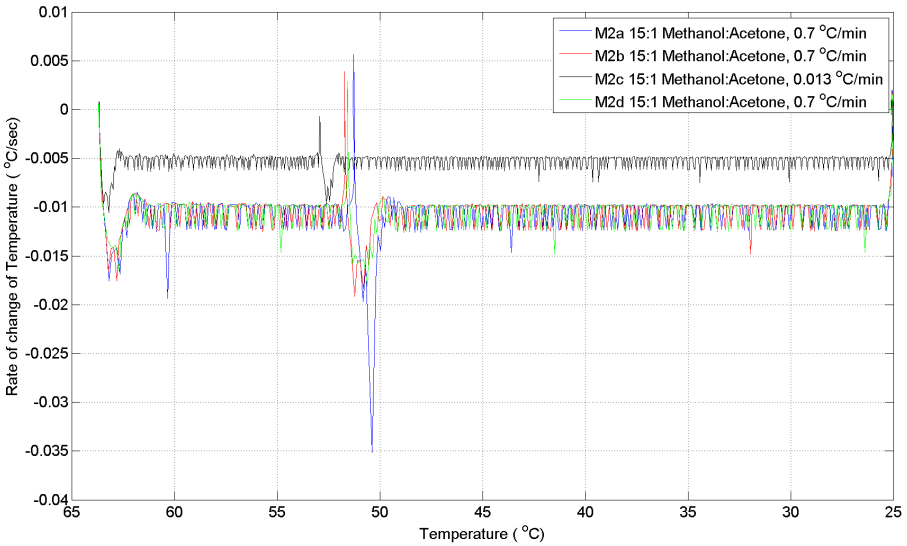


Figure 6.9: Rate of change of temperature vs. reactor contents temperature during crystallization experiments for system M2.

The third group of experiments (M3_a to M3_e) were performed with a methanol to acetone ratio of 12:1 to further assess how the increase in acetone concentration can affect the crystallization of spironolactone. Five crystallizations were performed, each with a different cooling rate including 0.71 °C min^{-1}, 0.12 °C min^{-1}, 0.36 °C min^{-1}, 0.07 °C min^{-1}, and 0.53 °C min^{-1}, for experiments M3_a to M3_e respectively. The order of the experiments was designed to minimise any effects of time that may be present in the system by varying the cooling rates in a pseudo-random order.

The temperatures at which the exotherm from the heat of crystallization was observed at are 50.8 °C, 53.5 °C, 51.9 °C, 53.6 °C, and 51.3 °C for experiments M3_a to M3_e respectively. The temperature at the onset of crystallization correlates strongly with the rate at which the solution was cooled with a correlation coefficient (R^2) of 0.9708, with the solutions that were cooled slowly crystallizing at a higher temperature than the solutions that were cooled quickly (figure 6.10).

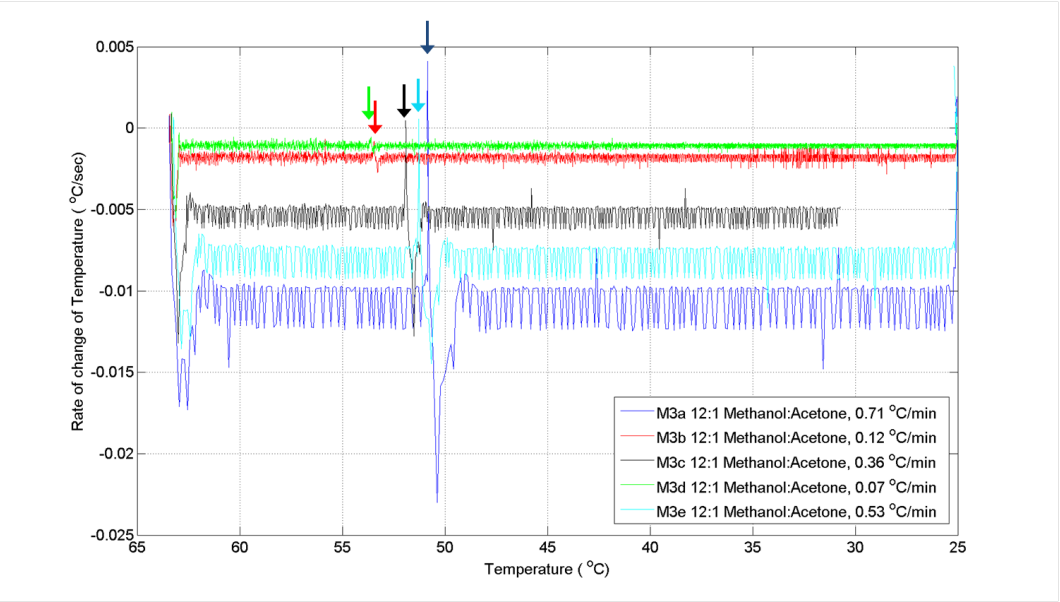


Figure 6.10: Rate of change of temperature vs. reactor contents temperature during crystallization experiments for system M3.

Martini et al. (2001) show results for the crystallization of lipid blends cooled at two different rates to a crystallization temperature where the temperature was held until crystallization was observed. They observed that the systems that were cooled slower had a shorter induction time to crystallization and concluded that this was because the system that was cooled slowly had more time to arrange the molecules in an structured order to enable the creation of stable nuclei. The systems that were cooled rapidly did not have as much time for the molecules to rearrange and therefore this rearrangement was conducted at the crystallization temperature and resulted in longer induction times. This same behaviour is occurring in the spironolactone system, however as no constant temperature is maintained for crystallization a lag time is not able to be calculated, and the result is that nucleation occurs at higher temperatures for experiments with slower cooling rates and therefore at a lower degree of supersaturation.

There was also a strong correlation between the concentration of acetone and the temperature of the onset of crystallization with the more acetone added, the cooler the solution when crystallization occurs. This is because spironolactone is a lot more soluble in acetone than in methanol, therefore as the amount of acetone increases, colder the solution must be to obtain the same degree of super saturation (figure 6.11 and figure 6.12).

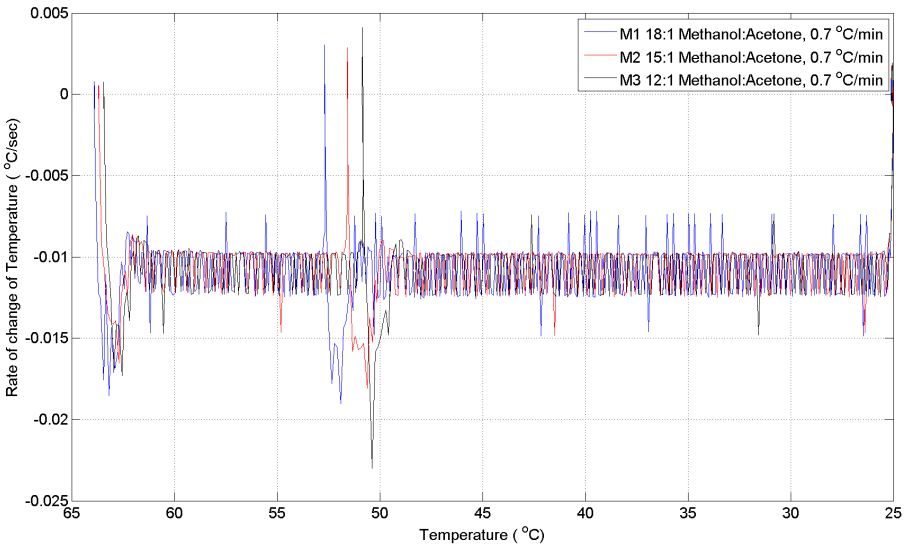


Figure 6.11: Rate of change of temperature vs. reactor contents temperature during crystallization experiments for system systems with varying quantities of acetone for the same cooling rate.

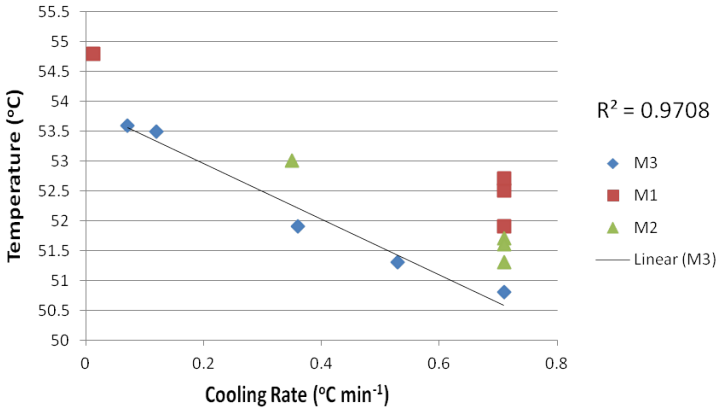


Figure 6.12: Temperature at the nucleation exotherm

6.6.2 Focused Beam Reflectance Measurement (FBRM)

Impact of cooling rate on final chord length distribution

The chord length distribution (CLD) at the end of the crystallization experiment (Figure 6.13) showed some differences between the experiments. Experiments M1_a and M1_b had very similar CLDs which is expected as both of these systems were crystallized under very similar conditions. Experiment M1_d shows a higher count of chord lengths between 1μm and 40μm and slightly fewer chord lengths measuring above 40μm.

Although M1_d was expected to give similar results to M1_a and M1_b, as noted earlier the nucleation event appeared to occur later at a slightly lower temperature. As the systems were otherwise the same this would indicate that there was a higher degree of supersaturation resulting in more homogeneous nucleation and therefore a suspension of smaller crystals. Experiment M1_c, on the other hand, shows a reduced count of chord lengths between 1μm and 50μm and higher count of chord lengths above 50μm. This system was cooled slower and as a consequence the nucleation event occurred at a higher temperature. The degree of supersaturation will therefore have been lower reducing the rate of nucleation events and thereby allowing crystal growth to be more dominant than in the other systems. This results in a suspension of crystals with a slightly larger size.

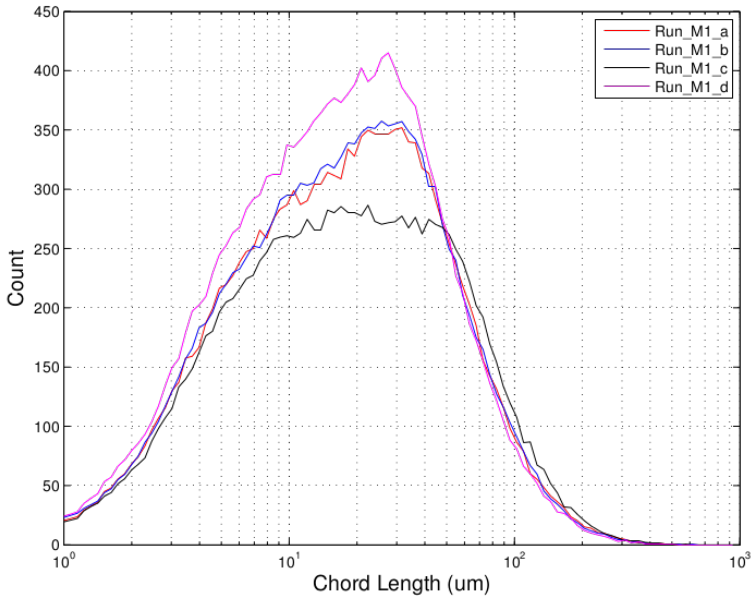


Figure 6.13: FBRM chord length distribution at end of experiments for group M1

Similar results can be seen for the second group of experiments (Figure 6.14). Experiment

M2_a, M2_b and M2_d all were carried out with the same cooling rate and saw nucleation occur at similar temperatures. The CLDs at the end of crystallization are also very similar as expected, showing that with consistent crystallization conditions, similar CLDs can be obtained. Experiment M2_c was cooled at a slower rate and again as a result saw the nucleation event occur at a higher temperature than with the solutions that were cooled faster. This affect of this was to yield a CLD with a reduced count of chord lengths between 1μm and 50μm and slightly increased count of chord lengths above 50μm.

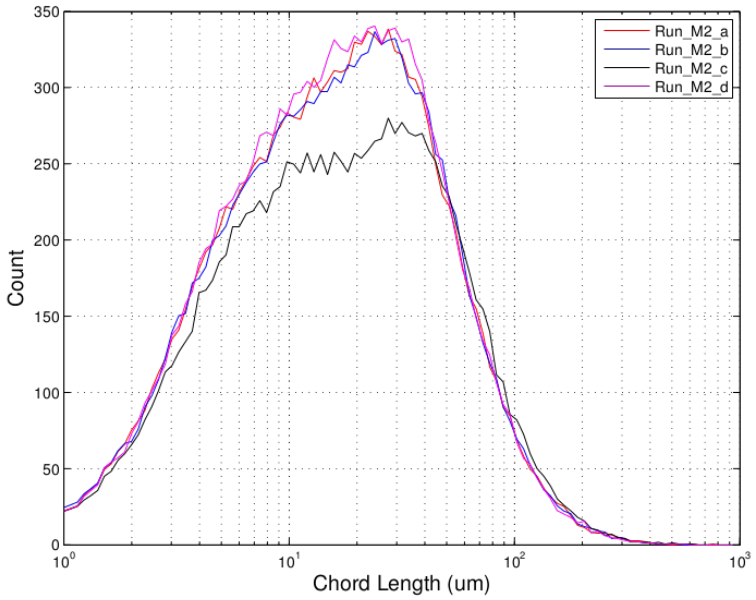


Figure 6.14: FBRM chord length distribution at end of experiments for group M2

The third group of experiments shows similar results to the first two groups of experiments (Figure 6.15), however another behaviour can be seen in the third group that was not noticed in the other groups due to the increased range of cooling rates used in the third group. Experiments M3_a and M3_e have very similar CLDs with M3_a having slightly more counts at chord length of 9μm to 20μm than M3_e.

Experiment M3_c shows that as the cooling rate decreased the count also decreased, however the distribution remained roughly the same shape and no increased count was noted at the larger chord lengths. Reducing the cooling rate further again reduced the chord count, however the shape of the distribution appears to change; M3_a, M3_c, and M3_e have the peak count between 20μm and 40μm with a high tail between 9μm and 20μm, M3_b and M3_d however have the peak count between 10μm and 20μm with a relatively high tail between 20μm and 60μm where the count for all of the experiments decreases relatively quickly. There

was also an increase in the count of chord lengths greater than 60μm for M3_b and M3_d compared to the other M3 experiments.

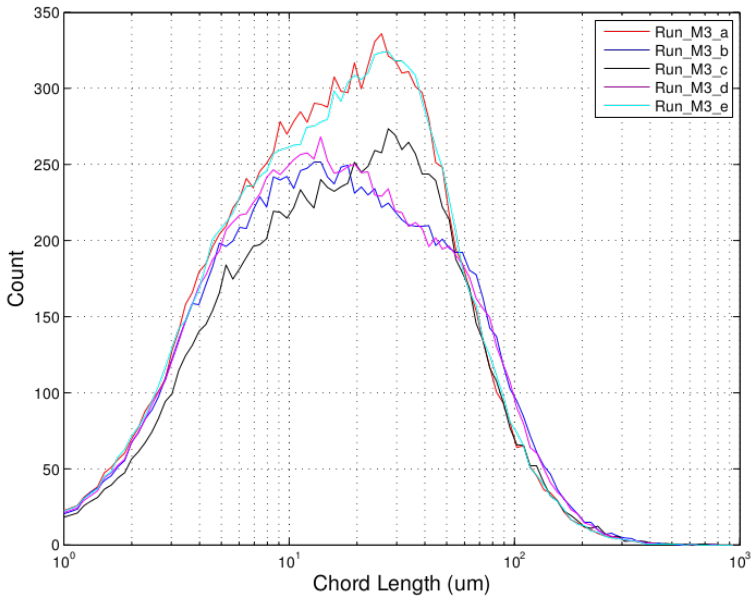


Figure 6.15: FBRM chord length distribution at end of experiments for group M3

Impact of acetone ratio on final chord length distribution

Figure 6.16 shows the affect of the acetone concentration on the final CLD of spironolactone. As the concentration of acetone increases, the count of chord lengths greater than 6μm decreases. Experiments M1_a and M1_b have a peak chord length of approximately 30μm and higher counts for the larger chord lengths than experiments M2_a, M2_b and M3_a. Experiments M2_a, M2_b and M3_a have similar CLDs which show fewer counts on the larger chord lengths than M1_a and M1_b. As spironolactone is soluble in acetone, when the concentration of acetone is higher, the degree of supersaturation of spironolactone at the same temperature will be lower than in the system with less acetone. This will retard the rate of nucleation and growth resulting in a higher proportion of smaller crystals. The total quantity of spironolactone that can come out of solution may also be reduced as a result of the increased quantity of acetone.

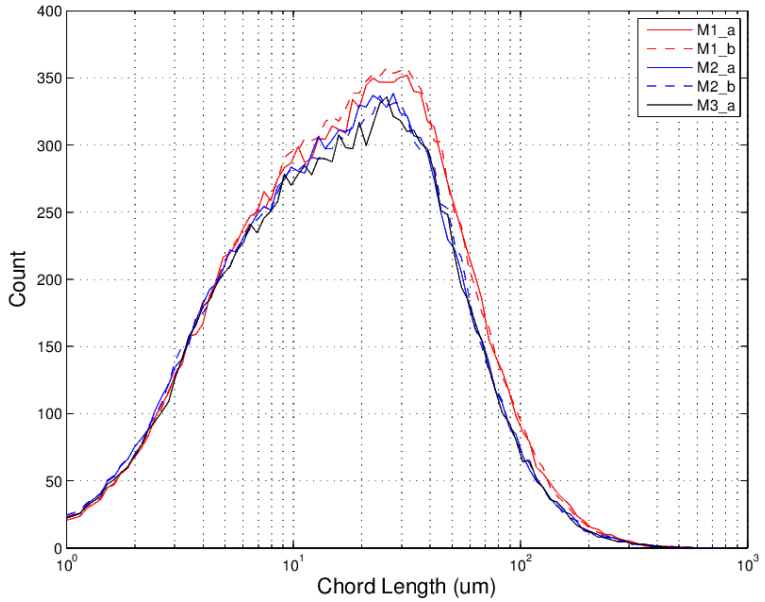


Figure 6.16: Effect of acetone concentration on Spironolactone chord length distribution

Identification of the nucleation event with FBRM

The nucleation event was not seen on the FBRM probe at the same time the crystallization exotherm was picked up on the RC1 data. For example the exotherm for experiment M1_a (Figure 6.17) was seen to start at 14:49:26 whereas the FBRM chord length count did not increase significantly until almost 3 minutes later at 14:52:51.

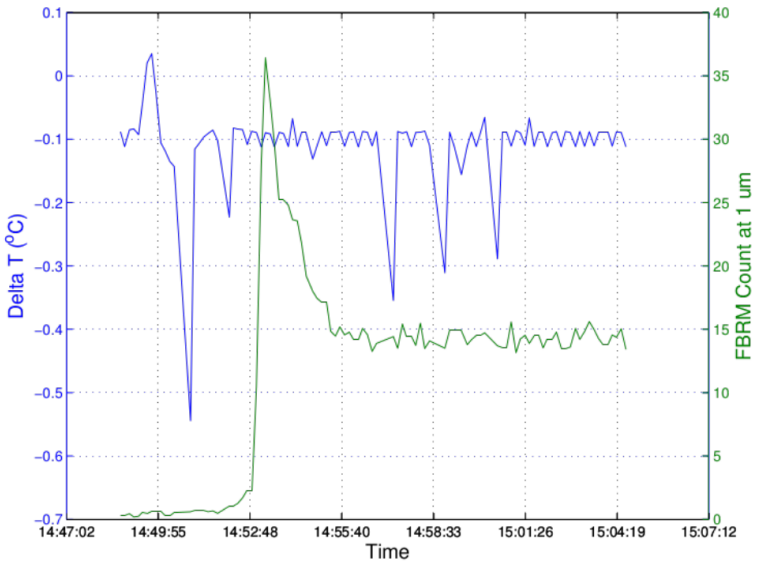


Figure 6.17: FBRM identification of nucleation in experiment M1_a

Experiment M2_a (Figure 6.18) also shows very similar behaviour with the exotherm from the head of crystallization being seen at 13:00:11, a little over 3 minutes before the FBRM chord length count began to increase at 13:03:21. And experiment M3_a (Figure 6.19) shows similar

results with the time delay being 3 minutes and 15 seconds between the exotherm and visible particles on the FBRM. This indicates the smallest chord lengths that are detected by the FBRM are around 1μm.

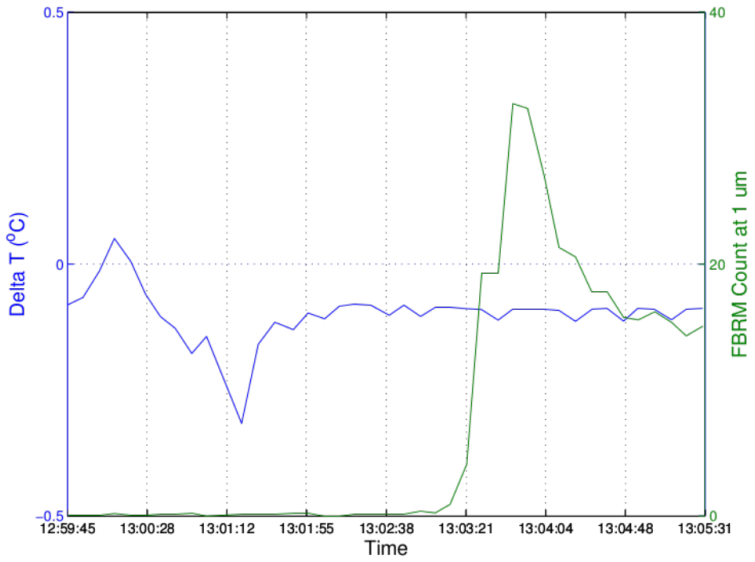


Figure 6.18: FBRM identification of nucleation in experiment M2_a

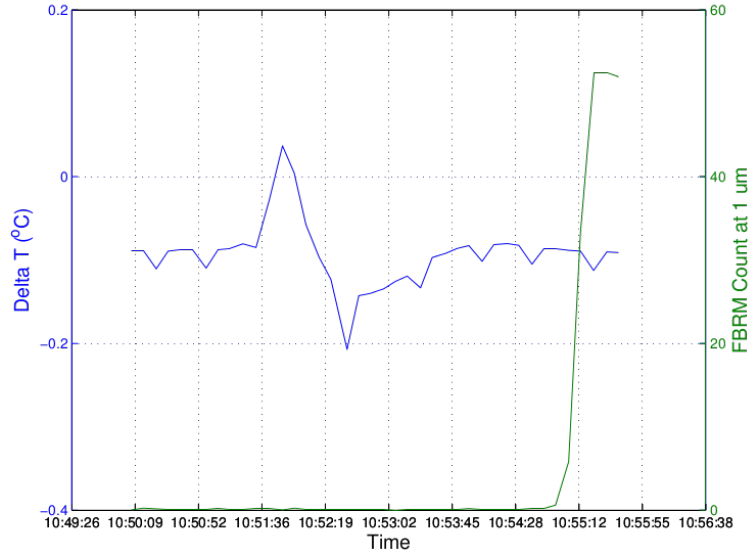


Figure 6.19: FBRM identification of nucleation in experiment M3_a

6.6.3 Transflectance Near Infrared Spectroscopy (NIR)

Near Infrared (NIR) spectra were collected in transflectance mode throughout each experiment from M1_a to M3_e inclusive. In addition to the crystallization experiments, NIR spectra were obtained for pure samples of methanol and acetone (figure 6.20). No pure spectra were

obtained for spironolactone as the spectra of the pure form would have saturated the instrument due to the crystalline nature of pure spironolactone. Both the spectra have absorptions at similar wavenumbers, however there are sufficient differences in each spectra, such as the shapes and locations of the peaks to be able to identify both species in a mixture.

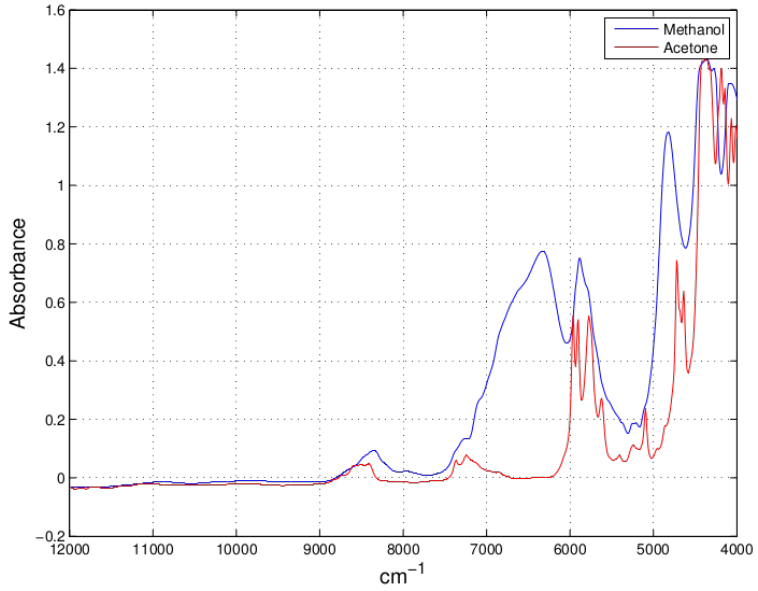


Figure 6.20: NIR spectra of pure methanol and pure acetone

The blue series on Figure 6.20 shows the spectra for methanol. A relatively weak peak can be seen at 8800 cm^{-1} to 8000 cm^{-1} which relates to an aliphatic CH_3 group. Decreasing in wavenumber, another weak peak can be seen at 7500 cm^{-1} to 7200 cm^{-1} which is also from the CH_3 group. Adjacent to this is a relatively strong broad peak from 7200 cm^{-1} to 6100 cm^{-1} which is the R-OH stretch. The peak at 6000 cm^{-1} to 5500 cm^{-1} is another relatively strong peak relating to the aliphatic R-CH_3 group. One of the strongest peaks on the methanol spectra is 5100 cm^{-1} to 4550 cm^{-1} which is a combination band of the R-OH and CH_3 groups. The most intense peaks are from 4500 cm^{-1} to 4000 cm^{-1} which relate to other combination bands.

The red series on Figure 6.20 shows the spectra of acetone. A group of relatively weak peaks can be seen around 8700 cm^{-1} to 8300 cm^{-1}. The first of these at 8586 cm^{-1} is the aliphatic C=O group and also present in this group is the second overtone of the aliphatic CH_3 group. Again the CH_3 peaks are also visible at 7400 cm^{-1} to 7100 cm^{-1}. There are three medium intensity peaks between 6000 cm^{-1} and 5700 cm^{-1}. The first at 5963 cm^{-1} is the C=OCH_3 group, and the second at 5901 is thought to be a CH or possibly the R-CH_3. The third of these at 5770 cm^{-1} may also be the R-CH_3 group. The final relatively strong peak that could be

identified was at 4720 cm^{-1} and is a combination band for the CH - C=O groups. Again the strongest bands are found in the first combination CH region at 4500 cm^{-1} to 4000 cm^{-1} (Workman and Weyer (2008); Socrates (2004)).

The NIR transflectance spectra was initially plotted in Figure 6.21 to identify any information that can be seen from the raw spectra. One hundred spectra were taken from around the region where nucleation was seen in the temperature data and plotted by colour over time. The first spectra, where the whole system is in solution, have a baseline of around 0 absorbance units in the larger wavenumbers (12000 cm^{-1} to 8000 cm^{-1}). The baseline increases over time as the system is cooled and the spironolactone crystallizes. This is due to the increased scatter caused by the crystals in suspension which affects the entire spectrum in addition to the effects of temperature on the NIR spectra.

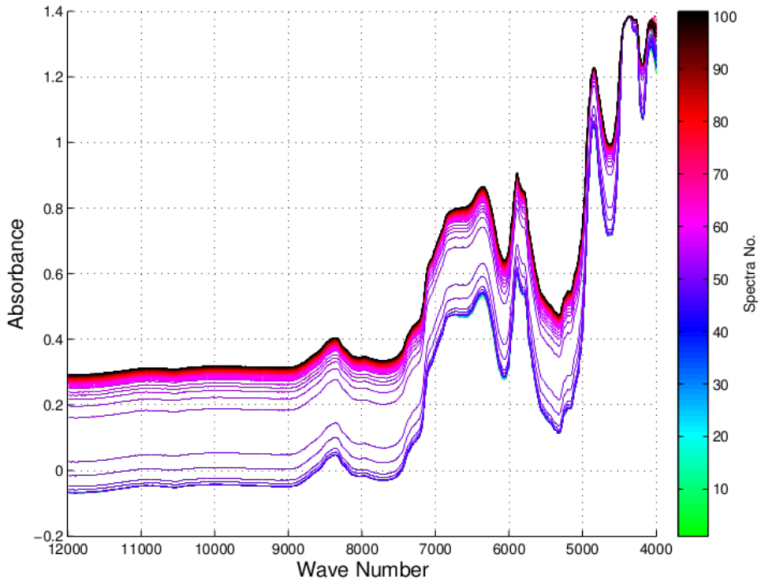


Figure 6.21: Raw NIR spectra collected during experiment M3_a

The raw NIR spectra is capable of detecting nucleation prior to the temperature increase was seen in the RC1 due to the energy release of crystallization; in this instance the NIR spectra in Figure 6.22 show a significant baseline shift at 10:51:23 whereas the temperature change was not seen until the next sample at 10:52:32. This may be due to the time delay in the system resulting from the distance of the nucleation events from the temperature probe, the heat transfer characteristics of the solution, and the dynamics of the temperature probe itself.

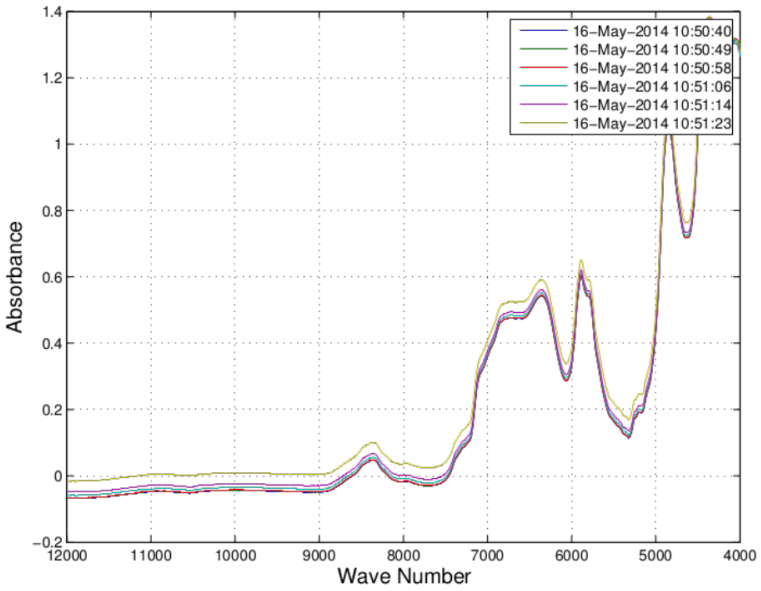


Figure 6.22: Raw NIR spectra of experiment M3_a around nucleation

All of the experiments exhibit the same behaviour. Another example of which is shown in experiment M1_a (Figure 6.23). This shows that the NIR spectra exhibits a shift starting at the sample 14:49:01, whereas the temperature increase is seen on the next sample at 14:49:18 (Figure 6.24).

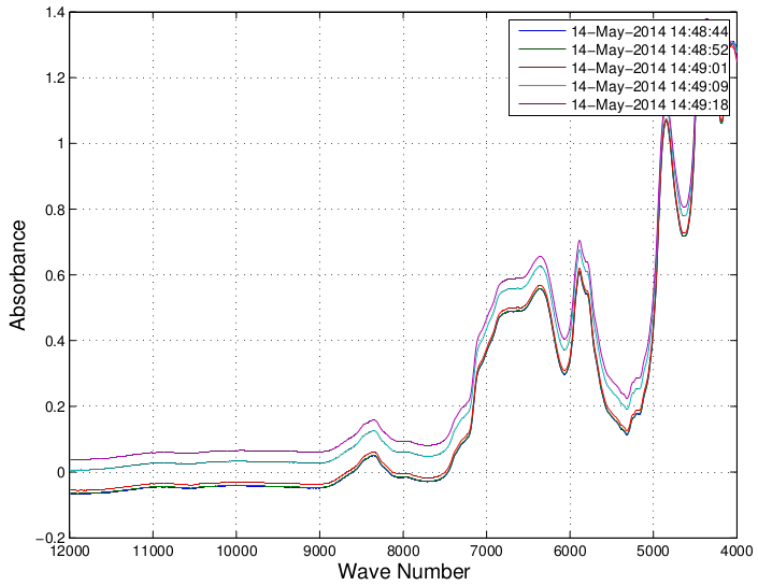


Figure 6.23: Raw NIR spectra of experiment M1_a around nucleation

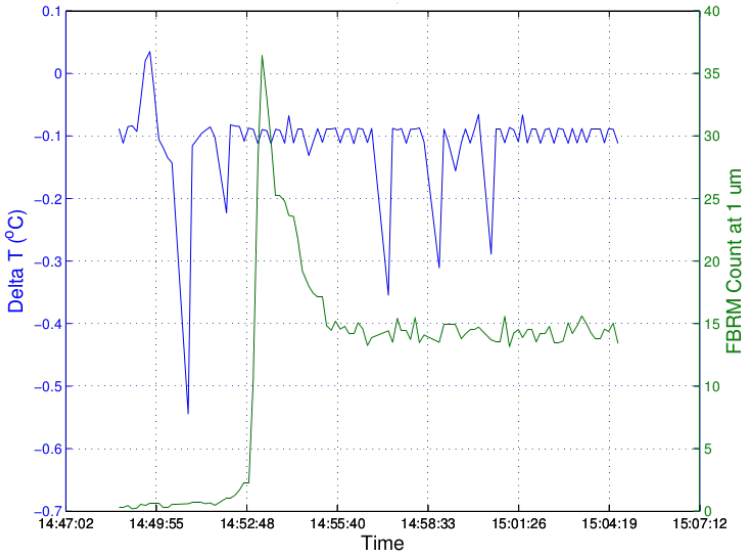


Figure 6.24: Differential temperature of experiment M1_a around nucleation

As the probe is a transflectance probe, there is a window (a gap in the probe in which the sample enters) between the optics of the probe and a mirror. NIR light enters the probe through one of the fibres, passes through the gap in the probe where the sample is, hits the mirror, passes back through the sample and into the probe where the NIR light is then carried through another fibre back to the spectrometer. It is in this window that solid material may collect. Figure 6.25 shows the construction of a transflectance NIR probe. Due to the relatively short length of the NIR probe, the tip of the probe with the window was only just submerged in the top of the reactor contents. The reactor contents were agitated from the bottom with a upward thrust propeller type stirrer as required for the FBRM probe, resulting in the NIR probe tip being submerged in relatively low velocity fluid compared with the thermometer, MIR probe and FBRM probe. Although care was taken to ensure that the window of the probe was rotated and slightly angled towards the direction of flow, the low velocity fluid found towards the top of the reactor may not have had enough energy to clear the solid material as it accumulated in the probe tip. This may have resulted in variable quantities of solid spironolactone accumulating in the sample window throughout the duration of the experiment. The window was completely cleared after every experiment by dissolution of the drug substance during the reflux prior to the start of each experiment.

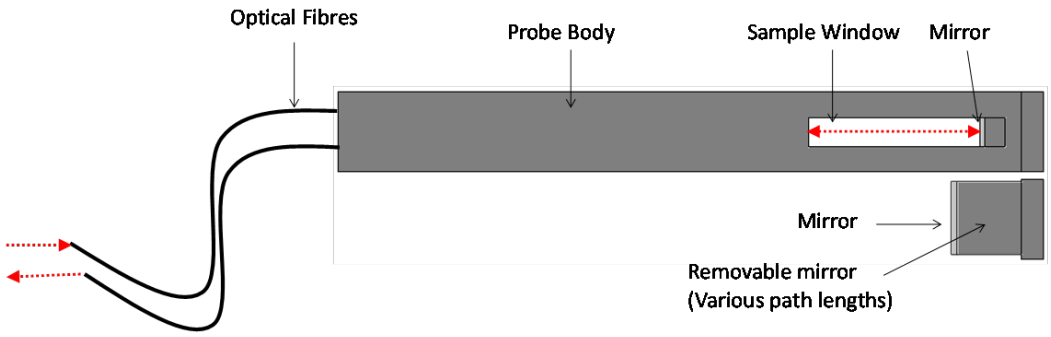


Figure 6.25: Transflectance NIR probe

As a result of this, the NIR spectra post nucleation showed a lot of scatter effects from the crystals that were collected in the probe window. There are several pre-processing techniques that can be applied to try and remove this scatter including taking derivatives, standard normal variate (SNV), multiplicative scatter correction (MSC), baseline correction, and combinations thereof. Unfortunately, these techniques were not able to remove the scatter in the spectra when comparing systems that were operated under identical conditions.

A principal component model was constructed on the NIR spectra obtained from the M1 group of experiments with as much of the scatter due to solids in the probe window removed as possible using a combination of SNV and second order derivatives. Three principal components were retained in the model, explaining 98.43% of the variation in the data, the second principal component captured a further 1.36% of the variance, and the third principal component captured a further 0.17%. The scores on principal components 1 to 3 are plotted in figures 6.26 to 6.28 respectively. Each of the experiments plotted was carried out under identical conditions. The scores up to sample 650 show a lot of noise. This is because these samples were collected when the systems were under reflux, therefore bubbles from the boiling solvents passing through the probe window cause significant variations in the absorbance of the NIR.

Following the reflux, cooling was applied to the systems, and the noise in the scores is significantly reduced and trends appear in the scores plots. Experiment M1_d appears to have a slight delay in the application of cooling compared to the other two runs. Other than this the experiments were replicates of the same condition and would therefore be expected to show the same trends and magnitudes in the scores plots. This is not the case in any of the three principal components , indicating that the principal component is still modelling undesired information, likely from the solids stuck in the window of the probe, in each of the three

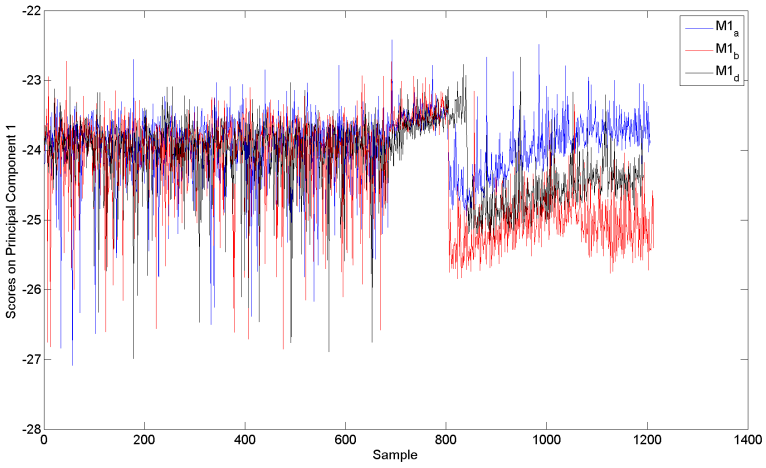


Figure 6.26: Scores plot on principal component 1 for NIR data from experiments M1

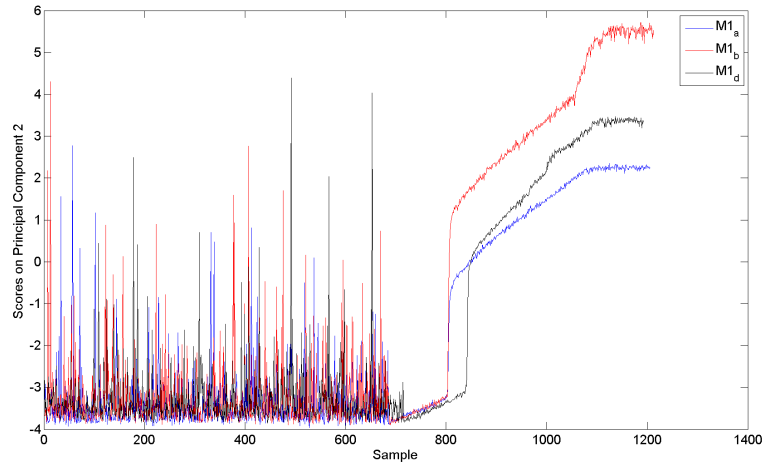


Figure 6.27: Scores plot on principal component 2 for NIR data from experiments M1

principal components, and therefore there is low confidence from further analysis of this data past the nucleation event.

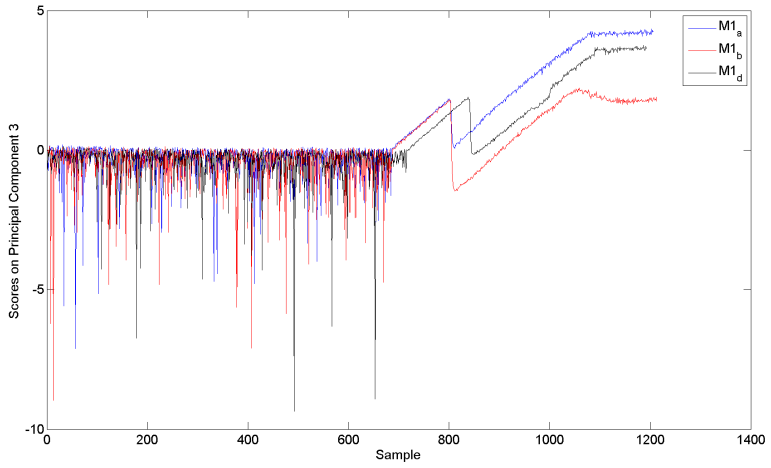


Figure 6.28: Scores plot on principal component 3 for NIR data from experiments M1

The loadings in the first principal component correspond to variation in the spectra due to methanol. This can be seen by comparing the main peaks in the loadings plot on principal component 1 (figure 6.29) against the spectrum for methanol (figure 6.30).

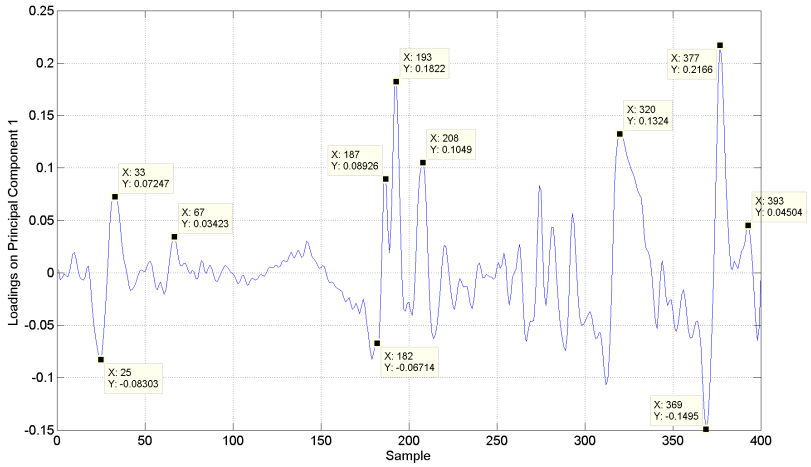


Figure 6.29: Loadings plot on principal component 1 for NIR spectral data

Similarly, the second principal component captures the information relating to the acetone in the system, however there is also some information from methanol retained in this principal component as shown by the loadings plots and NIR spectrum from acetone in figures 6.31 and 6.32 respectively.

Finally, principal component 3 captures again more information relating to methanol and acetone, however the largest loading corresponds to sample 316, which is in neither the methanol nor the acetone spectra, and therefore must relate to spironolactone (figure 6.33).

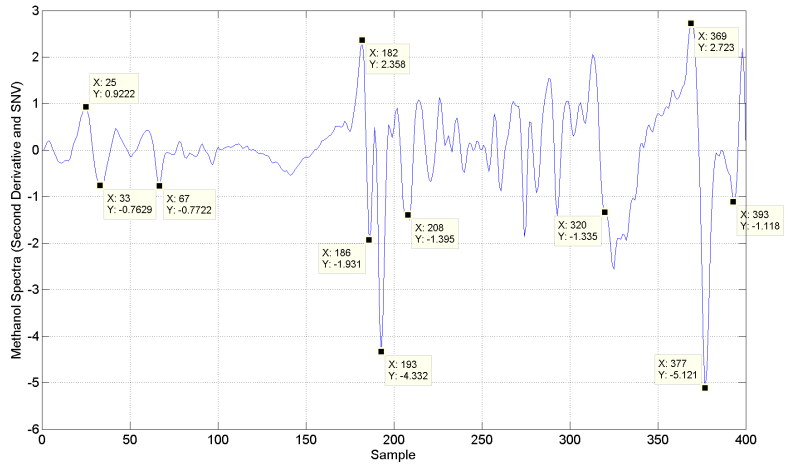


Figure 6.30: NIR spectra (second derivative and SNV) of methanol

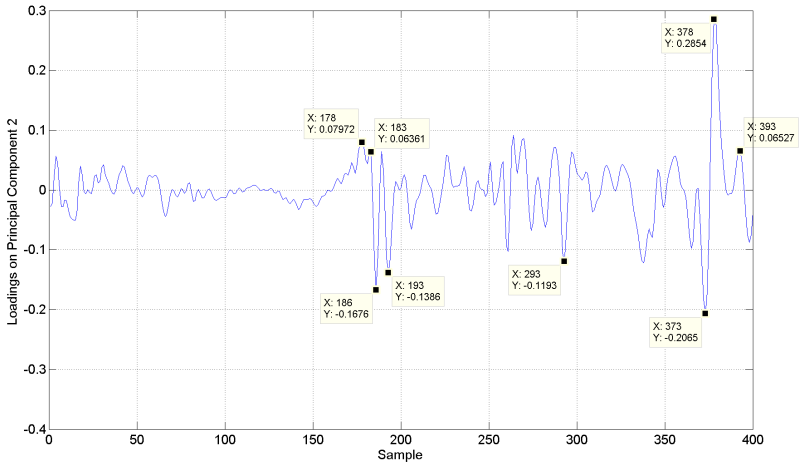


Figure 6.31: Loadings plot on principal component 2 for NIR spectral data

If this work was to be repeated, an alternative design of NIR probe would be recommended. Either a probe that is capable of reaching more turbulent zones in the reactor, another design of window on a transflectance probe that is less prone to collecting solid material, or another mode of NIR spectroscopy that removes the requirement for a window where solid material can gather such as a diffuse reflectance probe.

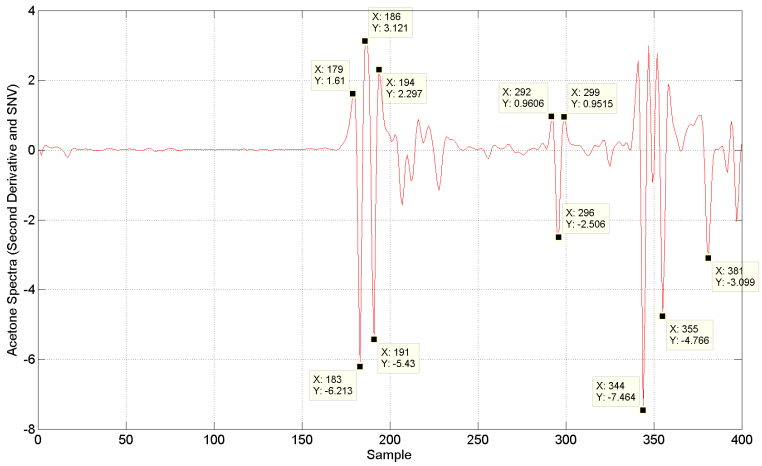


Figure 6.32: NIR spectra (second derivative and SNV) of acetone

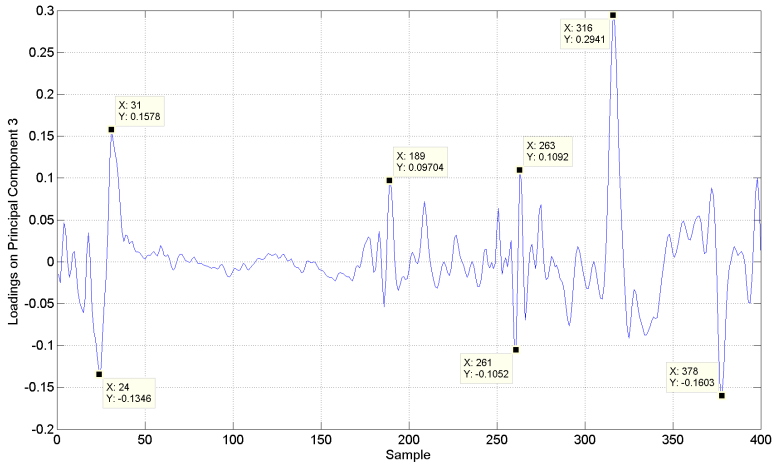


Figure 6.33: Loadings plot on principal component 3 for NIR spectral data

6.6.4 Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy (ATR-FTIR)

Attenuated Total Reflection Fourier Transform Mid Infrared (ATR-FTIR) spectroscopy was performed for three crystallization experiments. These included experiments M2_b, M2_c, M2_d, and M3_a. The MIR instrument was not available for the other experiments as liquid nitrogen, required for cooling the IR sensor, was not available. In addition to the crystallization experiments, MIR spectra were obtained for pure samples of methanol, acetone, and spironolactone (solid) at room temperature (figure 6.34, figure 6.35, and figure 6.36 respectively).

The peaks in the mid-infrared region are much more defined and narrower for specific bonds when compared to near-infrared absorption spectra. The strongest peak on the ATR-FTIR

spectra of pure methanol (figure 6.34) seen at 1022 cm^{-1} relates to the C - O stretch. A weaker double peak is also seen at 1420 cm^{-1} and 1454 cm^{-1} which are thought to be the C - H_3 bend and the C - O - H bend respectively. The two peaks at 2831 cm^{-1} and 2951 cm^{-1} are the C - H stretch and the broad peak at 3267 cm^{-1} is the O - H stretch (Coates, 2006).

Figure 6.34: Raw ATR-FTIR spectra of methanol

There is one strong peak and two medium intensity peaks on the ATR-FTIR spectra of acetone (figure 6.35). The strongest peak at 1709 cm^{-1} is the C = O stretch. One of the medium intensity peaks (1358 cm^{-1}) is a doublet with a weak peak at 1423 cm^{-1} and corresponds to the O = C - CH_3 group in acetone. The other medium peak at 1219 cm^{-1} is from the C - C (= O) - C bend (Coates, 2006).

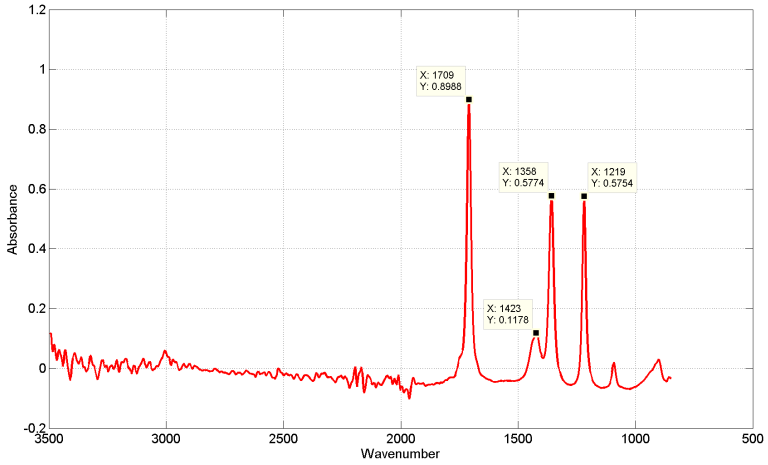


Figure 6.35: Raw ATR-FTIR spectra of acetone

The ATR-FTIR spectra of spironolactone (figure 6.36) shows two strong peaks relatively close together at 1678 cm^{-1} and 1771 cm^{-1} which correspond to the lactone and thiol groups

respectively. There is a weak peak next to the thiol peak (1616 cm^{-1}) which is related to a di-ketone, enol structure in the molecule. The weak peaks towards the higher wavenumbers between 2947 cm^{-1} and 2881 cm^{-1} relate to the CH_2 and CH_3 stretches (Coates, 2006).

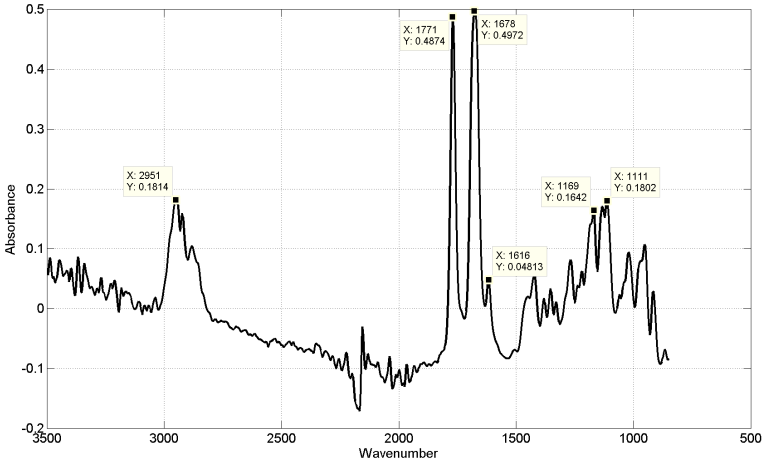


Figure 6.36: Raw ATR-FTIR spectra of spironolactone

Wavenumbers in the range 1763 cm^{-1} to 991 cm^{-1} were selected for the following analysis as these contained the relevant peaks for the three components. There was however an overlap between acetone and spironolactone at wavenumbers 1705 cm^{-1} to 1678 cm^{-1}, therefore, to increase the ability to interpret the model and pull out the differences between acetone and spironolactone, these overlapping wavenumbers were excluded. This did not exclude the entire peak, and the non-overlapping regions were retained.

A principal component analysis was constructed from the three pure component ATR-FTIR spectra after baseline correction. Three principal components were retained in the model, capturing 60.77% of the variation in the first component, a further 27.93% in the second principal component, and the remaining 11.3% in third component. Although typically models capturing 100% of the variation are not desired, in this case three components are required to pull out the information on the three components of the system.

As shown from both the scores and loading plots (figure 6.37 and figure 6.38), the first principal component captures the spectral information on mostly the methanol, however a small quantity of information relating to the acetone is also contained in principal component 1. Similarly, principal component 2 contains the spectral information mostly relating to the acetone, however some information relating to the methanol is also captured here. Finally, principal component three captures the spectral information relating to the spironolactone in the system.

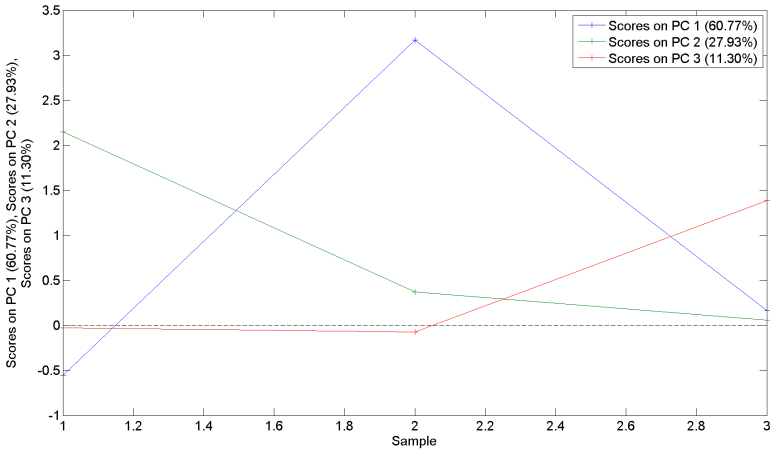


Figure 6.37: Sample scores plot on PCA model of ATR-FTIR pure component spectra of methanol, acetone, and spironolactone for principal components 1, 2, and 3 (blue, green, and red respectively)

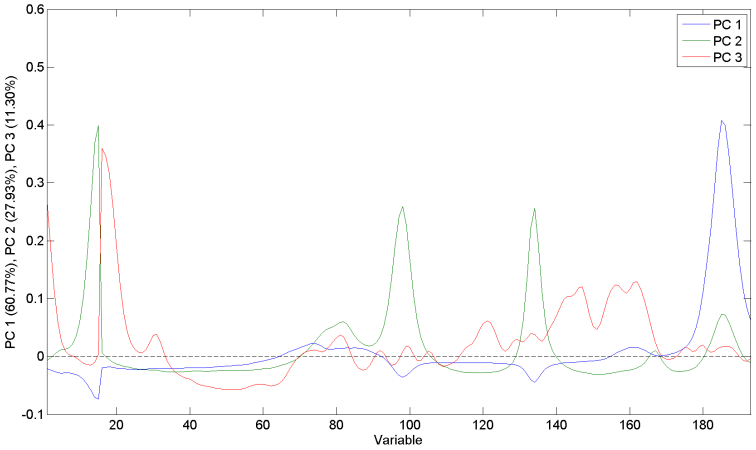


Figure 6.38: Variable loadings plot on PCA model of ATR-FTIR pure component spectra of methanol, acetone, and spironolactone for principal components 1, 2, and 3 (blue, green, and red respectively)

The ATR-FTIR spectra at 0.1 °C intervals between 63 °C and 25.1 °C were obtained and the same baseline procedure was applied. These spectra were then passed to the PCA model developed on the three pure component spectra. The scores plots on each principal component are show in figures 6.39 to 6.43.

The scores on principal component 1 show a linear increasing trend as the temperature is reduced. This indicates that there may be some temperature information contained within the first principal component. Figure 6.39 shows the scores plot for experiment M2_b along with the linear regression line constructed from the first 30 data points. This shows that there is an obvious change in gradient at approximately 53 °C following which the gradient on the scores increases. This is due to the changing methanol concentration visible to the ATR-FTIR probe during the crystallization of the spironolactone. This trend can be observed for each of the experiments. Finally, there is a slight offset shown in figure 6.40 with the scores on principal component 1 for experiment M3_a being slightly lower than for the experiments M2_b through M2_d. This is because principal component 1 also contains a small amount of information relating to the acetone concentration in the system, and experiment M3_a has an increased concentration of acetone.

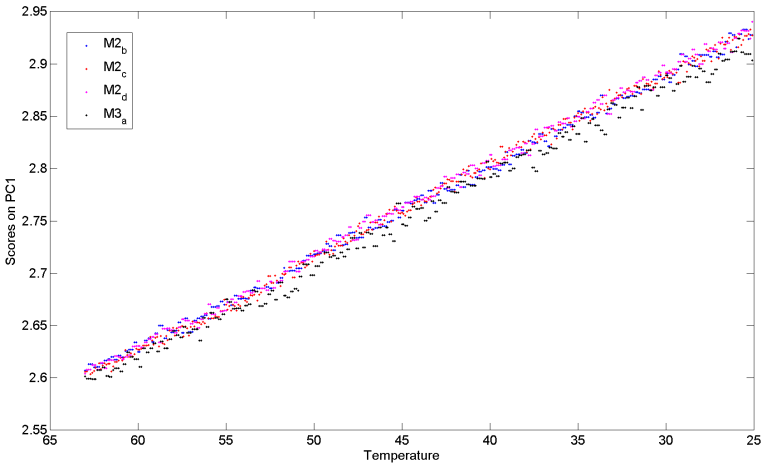


Figure 6.39: Temperature scores plot for PCA model on ATR-FTIR crystallization spectra on principal component 1

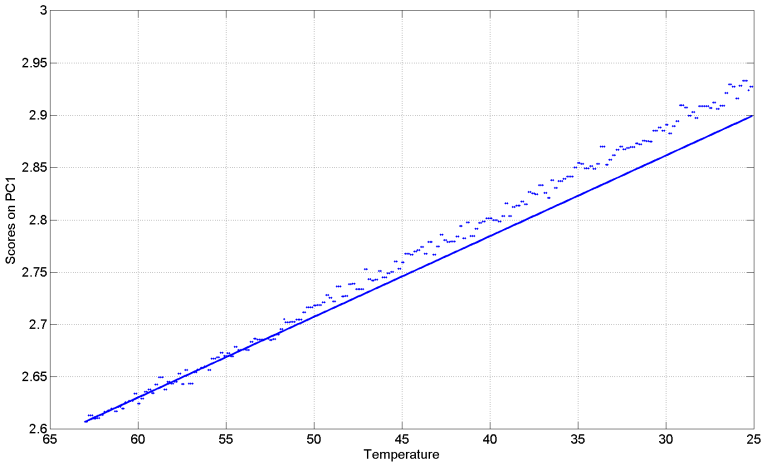


Figure 6.40: Temperature scores plot for PCA model on ATR-FTIR crystallization spectra from experiment M2_b with linear regression through first 30 samples on principal component 1

The scores on principal component 2 (figure 6.41) show a clear difference between experiment M2_b through M2_d, compared to experiment M3_a. This is due to the increased acetone concentration in experiment M3_a being described by principal component 2. There is also a non-linearity observed in the M2 group of experiments which is not as obvious in the M3 experiment. This may indicate that the acetone concentration is changing throughout the duration of the experiment, and does so differently between the high and low acetone concentration systems. Again there is an increasing trend as the temperature is reduced. This indicates that there may be some temperature information contained within the second principal component also.

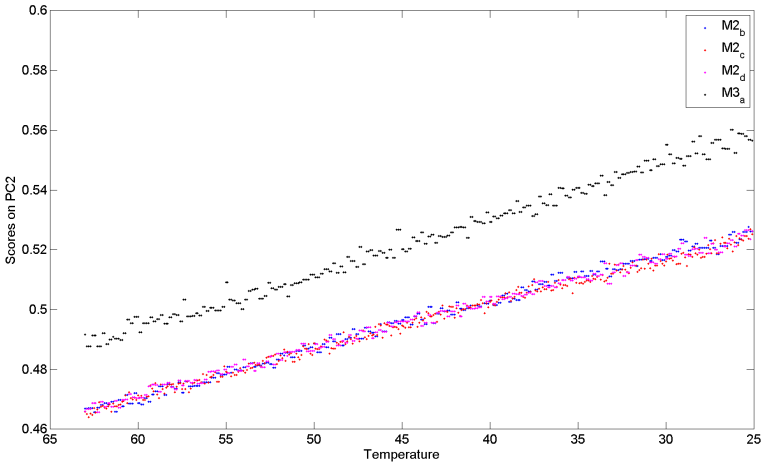


Figure 6.41: Temperature scores plot for PCA model on ATR-FTIR crystallization spectra on principal component 2

Similarly, plotting the scores on principal component 1 against the scores on principal component 2 shows that there is some relationship between the two (figure 6.42). Again the M3 experiment has higher scores in PC2 due to the increased concentration of acetone as expected. The trends between the two groups of scores however are not parallel between the M2 and M3 system. This again indicates that the concentration of solvents is changing at a different rate between the two systems. As methanol concentration is increasing due to removal of spironolactone from the solvent system (i.e. scores on principal component 1 are increasing) , the concentration of acetone is also increasing as would be expected (i.e. scores on principal component 2 are also increasing). However, the non-linearity in the relationship between principal component 1 and principal component 2, and the divergence between the trends for the M2 system and M3 system indicate that the rate or increase of one of the solvents is faster than the rate of increase of the other solvent. Therefore, one of the solvents must also be coming out of solution during the crystallization.

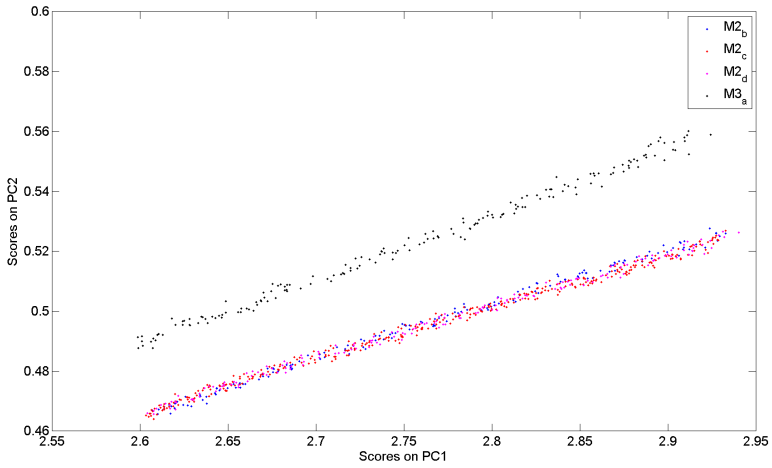


Figure 6.42: Scores plot for PCA model on ATR-FTIR crystallization spectra on principal component 1 against principal component 2

The scores plot or principal component 3 (figure 6.43) again shows some temperature sensitivity at the start with increasing scores between 63 °C to approximately 53 °C. At 53 °C there is a rapid decline in the scores on principal component 3 indicating crystallization of spironolactone which occurs at approximately the same temperature for all of the experiments in group M2, and a slightly lower temperature for experiment M3_a. This is expected as the spironolactone is soluble in acetone, therefore an increase in the amount of acetone in the system, increases the relative acetone to spironolactone concentration resulting in a reduction of supersaturation for the same temperatures and therefore a delay to the crystallization. This

increase in concentration of acetone can also be seen as a decrease in the scores of principal component 3 throughout the experiment. This is because additional acetone was added to the system without adjusting for solvent volumes resulting in a reduction of spironolactone concentration in the system. Another observation from the scores on principal component 3 is the scores reduce by approximately the same amount for each experiment indicating that the concentration of spironolactone has decreased by the same amount each time. Finally, the decreasing concentration of spironolactone plateaus at approximately 40 °C indicating that the crystallization is complete by this point and no further spironolactone is coming out of solution.

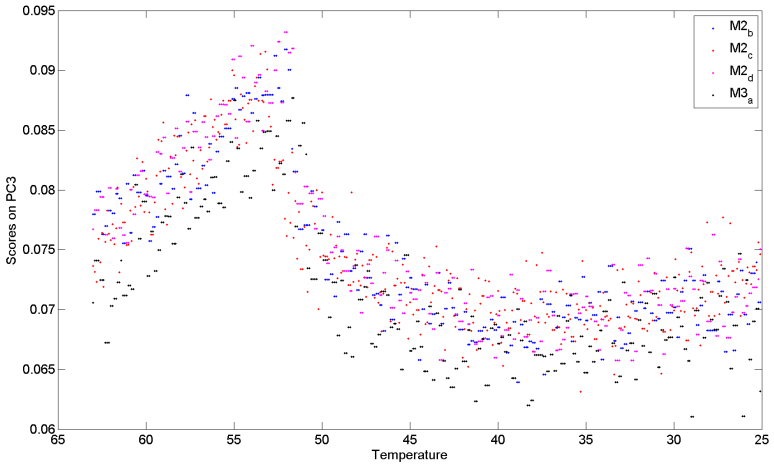


Figure 6.43: Temperature scores plot for PCA model on ATR-FTIR crystallization spectra on principal component 3

The spectra in the model were of pure components only and did not have any temperature data as each of the spectra were captured at room temperature. As a result, no all of the information from the spectra collected during the experiments will be captured in the principal components. Looking at the Q residuals plot (figure 6.44), the Q residuals appear to increase as the temperature in the system falls. This is contrary to the expectation that as the temperature of the system reduces towards the temperature of the pure spectra the model should become more accurate. This indicates that there is an effect due to the interactions between the three components in the system that changes the absorbance behaviour.

The contributions plot on the Q residuals (figure 6.45) shows a number of regions that have high contributions on the residuals. This plot is coloured by time, with the contributions from the first spectra shown in blue (the warmest samples), red contributions are from those spectra towards the end (cooling samples), and finally the black spectra are those obtained at the end

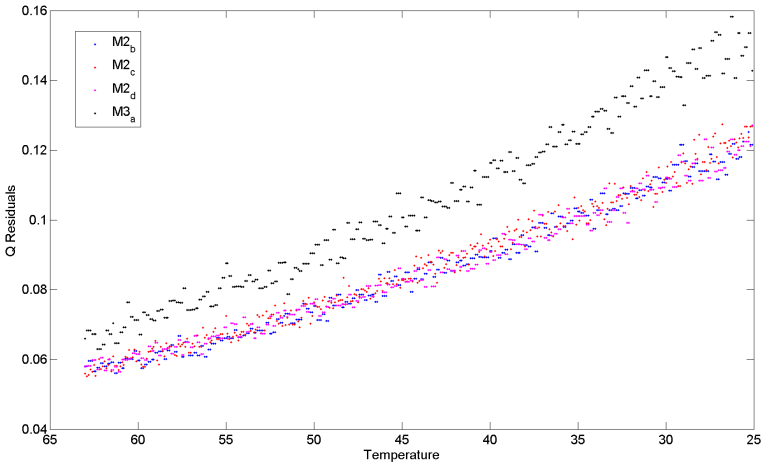


Figure 6.44: Temperature Q residuals plot for PCA model on ATR-FTIR crystallization spectra

of the experiment (coldest spectra). Overall there is a typically increasing trend in the contributions at a number of locations across the wavenumbers included in the model. There are however two regions were the contributions decreases (samples 2 - 14, and samples 70 - 80).

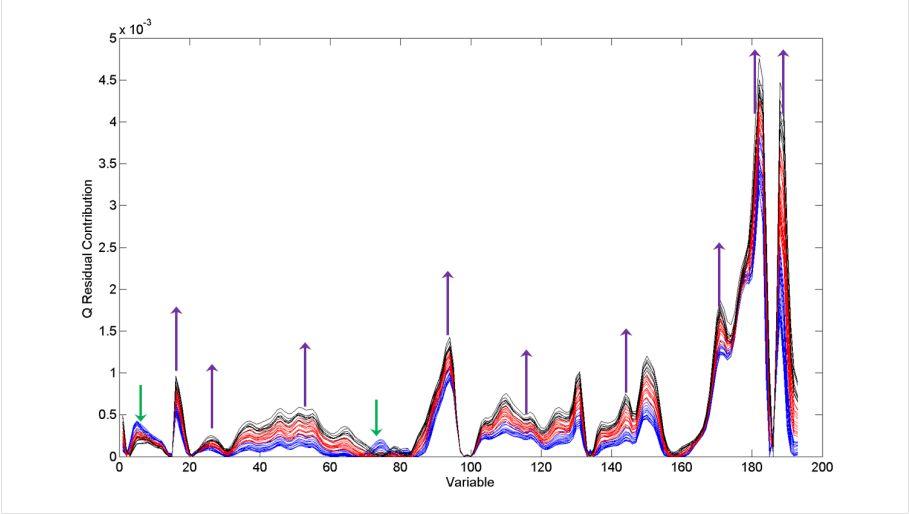


Figure 6.45: Sample Q residual contribution plot for PCA model on ATR-FTIR crystallization spectra for experiment M2_b

From analysing the spectra over time at these regions (not shown) it can be seen that the first peak relates to spironolactone and decreases over the duration of the experiment indicating that this material is being removed from the system. This is expected as the spironolactone crystallizes, the ATR-FTIR probe is no longer able to detect the spironolactone as the penetration depth of the probe is very small therefore a crystal face would have to be very close to the probes diamond to be detected. The model was constructed with samples of liquid

solvent, and crystalline spironolactone. The spironolactone ATR-FTIR spectrum may look slightly different when dissolved due to interactions with the solvents interfering with the vibrational characteristics of the molecule.

The second region that showed decreasing residual contributions (samples 70 - 80) is an area of the spectra where all three components are found to absorb the MIR light. Throughout the experiment, the absorbance at these wavenumbers increases. This could indicate that as the temperature is approaching that of the calibration data set (pure spectra at room temperature) the modelling is predicting more accurately in this region.

The most prominent peaks on the contribution plot are a double peak near the methanol peak. The main methanol peak is captured in the model and therefore has low Q residuals, however throughout the crystallization, the methanol peak becomes wider (figure 6.46). This broadening may be indicative of an increase in hydrogen bonding (Coates, 2006). This increase in hydrogen bonding may be indicative of the formation of a methanol solvate.

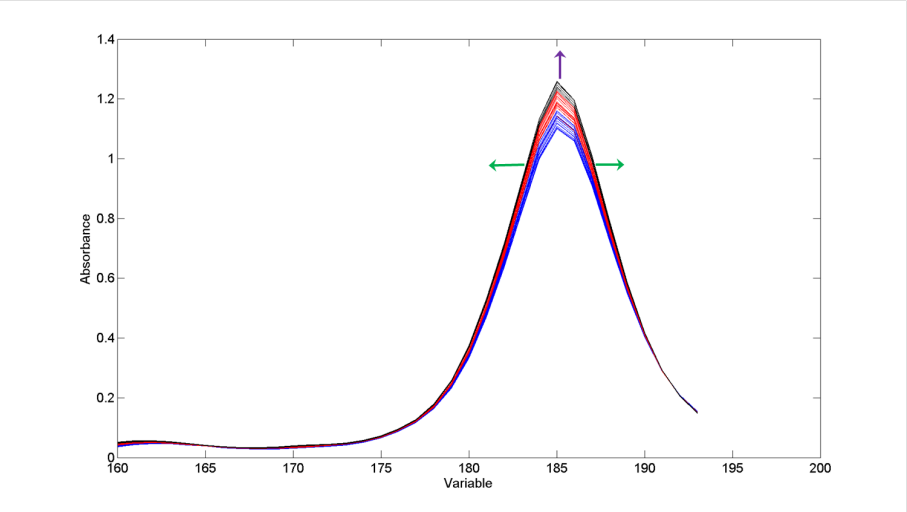


Figure 6.46: Baselined ATR-FTIR spectra M2_b of methanol peak wave numbers showing peak broadening throughout crystallization (blue = start, red = middle, black = end of experiment)

The smaller peak near the methanol peak at approximately sample 170 corresponds to a small increase in the absorbance throughout the experiment. This may be due to an interaction between spironolactone and methanol, however the absorbances in this region are weak for the pure spectra, therefore it is difficult to interpret this any further.

Another peak at approximately sample 90, corresponds to an increasing absorbance throughout the experiment at a region where the weak absorbances of the three components overlap.

Similarly the remainder of the peaks identified in the contribution on the Q residuals relate to overlapping weaker peaks from combinations of the solvents and drug substance.

6.7 Conclusions

From the temperature data collected during the crystallization experiments, the point of nucleation could be determined through the temperature rise caused by the heat of crystallization exotherm. There was some small variability in the temperature and time taken to achieve nucleation for repeats of the same system and same cooling rate due to the random nature of homogeneous nucleation.

Another observation from the temperature data was that there is a linear relationship between cooling rate and the temperature (or time) of nucleation. This was true for all of the systems with varying solvent concentrations, however, as the acetone concentration increased, the temperature that nucleation occurred reduced (the intercept of the linear trend decreased), and the gradient of the linear trend decreased.

The final observation from the temperature data was that as the cooling rate increased, the temperature rise due to the heat of crystallization exotherm also increased. This was due to the faster cooling causing a lower temperature of crystallization leading to an increase in supersaturation as crystallization is not only temperature dependent, but also time dependent. This increase in supersaturation caused a greater degree of crystallization to occur at nucleation releasing more energy to the system.

The FBRM data analysis showed that for systems with the same cooling rate, similar chord length distributions are obtained at the end of crystallization. Any differences in the CLD can therefore be attributed to process changes and not just random variation.

As the cooling rate increased (i.e. supersaturation increased) there was a small increase on the small chord lengths, and a corresponding decrease on the larger chord lengths indicating that the generation of more fines or less growth had occurred in these systems.

Conversely, as the concentration of acetone was decreased, the large particle count increased, and the overall counts increased indicating that more spironolactone was coming out of solution and crystal growth was a larger influence.

The FBRM data was able to detect nucleation, however this was approximately 3 minutes after

the nucleation was observed on the temperature data. The data was collected from the same computer and therefore the time stamps on the data were aligned.

The NIR data was able to detect the onset of crystallization through a baseline shift observed in the spectra. This was detected prior to the temperature increase observed in the date by approximately 1 minute (1 temperature sample). This may be due to a lag in the temperature measurement system and sampling frequency of the temperature data.

A principal component model constructed from the NIR spectra was able to pull out differences relating to the three components in the system (methanol, acetone, and spironolactone). The NIR probe window was fouled post nucleation with solid spironolactone, and therefore the data post nucleation was unreliable as pre-processing was unable to completely remove this information.

The loadings on the ATR-FTIR principal components show principal component 1 is mostly spectral information from the methanol, with some information on the acetone spectrum. Principal component 2 is mostly acetone spectral information with some information relating to the methanol spectrum. Principal component 3 is mostly spectral information relating to the spironolactone.

From observations in the ATR-FTIR scores the onset of crystallization can be observed, however there is a lot of noise in the scores for detecting this and also a temperature effect is observed in each of the three retained principal components. There is also evidence of loss of a solvent from the system from the scores plots on principal component 2, and the plot of the scores on principal component 1 against principal component 2.

There is some information not captured by the ATR-FTIR model which is shown in the prediction residuals for each of the experiments. Observation on the contributions to these residuals show there there is a broadening of the methanol peak in the spectra throughout crystallization indicating there there may be some increase in hydrogen bonding relating to the methanol, and perhaps due to the formation of a spironolactone methanol solvate.

The PAT tools evaluated in this study proved to be complementary overall providing all of the information required to met the objectives outlined at the start of this chapter. The NIR instrument, however, had severe limitations with the transflectance probe design resulting in poor quality data being collected throughout the experiments. PCA is a powerful tool for modelling spectroscopic data, however it is very sensitive to the pre-processing that is applied to the data.

Chapter 7. Conclusions and future work

7.1 Summary of thesis

The thesis presented a framework for exploring historical batch process data, to extract insights on where process control can be improved in established batch processes. The challenges presented with commercial process data were discussed. Multivariate tools such as dynamic multi-way principal component analysis were used to investigate variability in process data. The method detects batches with unusual events, such as equipment failures which, although important and of interest, were not the main focus. Following the identification and subsequent removal of these batches the true uncontrolled variation within the process was analysed to identify where the process could benefit from improved understanding and control.

This framework was demonstrated through application to commercial process data from the active pharmaceutical drug substance manufacturing process of spironolactone at Piramal Healthcare, Morpeth, UK. In this case study, the process exhibited variability in drying times which traditional univariate data analysis was not been able to attribute a root cause to. The results demonstrated some of the challenges associated with the use of the available data from commercial processes. Although the results from the multivariate data analysis did not show a significant statistical difference between the batches with long and short drying times, small differences were observed between these two groups. Further analysis of the crystallization process was carried out using infrared spectroscopic techniques which identified a potential root cause to the extended drying time.

Chapter 2 introduced the spironolactone drug substance manufacturing process as a case study on which the framework was tested. The background included details of the process chemistry including the impurities and challenges with polymorphism. The manufacturing process was also described in addition to the control of the process and the data that was collected from the process.

In chapters 3 and 4 an overview of some multivariate statistical tools was provided.

Pre-processing techniques including centring and scaling were discussed in terms of the challenges presented by industrial process data, namely that of the spironolactone manufacturing process. These challenges, such as alignment, data quality, quantity of data available, and noise, need to be handled on a problem by problem basis. More details on the challenges associated with the spironolactone process data were detailed in chapter 5. Principal component analysis and various extensions to handle such batch data were discussed.

Chapter 5 presented the multivariate model development on the spironolactone case study. First the model development on the dryer data was discussed and pre-processing including compression checks, missing data, filtering, alignment, centring and scaling, unfolding, and outlier detection and removal were performed. Following this Principal component analysis was performed by constructing the PCA model on the 'good' batches and applying this to the 'bad' batches to identify any differences between the two. This process highlighted a number of challenges with processing industrial process data and identified a number of equipment issues that should be resolved in the future. Additionally, this procedure identified the endotherm as important for the total cycle time. Subsequent to this, the reactor process data was investigated and the same procedure was applied. A comparison between two different alignment techniques showed how important process feature alignment is in order to extract understanding from the models. A number of potential causes or symptoms of the variability in drying time were identified in the reactor process data as a result of the modelling, some relating to variability in charges, some relating to operator variability, and others a combination of the two. This identified that the crystallization needed to be looked at in more detail, however the appropriate level of instrumentation was not available on the commercial reactor system, therefore, a lab scale investigation was initiated (chapter 6).

In chapter 6 a laboratory scale study of the crystallization of spironolactone was detailed. The study built on the observation on the dryer data analysis, which indicated that a cause of the variability in process drying time may have been a result of changes in crystal or polymorphic form (endothermic behaviour towards end of drying). The study investigated the impact of changing cooling rate and solvent ratio on the particle size of spironolactone using focused beam reflectance measurement. The FBRM was able to detect crystallization, although a delay was present compared to the heat of crystallization observed in the reactor temperature measurement. Other observations made with the FBRM instrument were consistent with crystallization theory in that the systems with faster cooling rates and therefore higher degrees of supersaturation presented smaller crystals than compared with the same systems cooled at a

slower rate. Transflectance NIR spectroscopy and ATR-FTIR spectroscopy were used to obtain information on the spironolactone and the solvents. The NIR spectroscopy data was found to be of limited use as the probe was too short to allow positioning in the reactor to keep the measurement window clear during the crystallization, thus hindering analysis of any information the instrument would have been able to collect. The ATR-FTIR instrument data was more successful and was able to detect crystallization before the crystallization was observed by the FBRM instrument. Additionally, the ATR-FTIR instrument indicated that there may be formation of a methanol solvate during the crystallization. The cause of the variability in the spironolactone drying time may therefore be a variation in the quantities of the solvents charged to the batch in the reactor, or the quantity of methanol solvate formed during crystallization which is converted back to the desired form through the endotherm in the drying cycle.

In summary, a framework (figure 7.1) was presented in which industrial batch process data is pre-processed as necessary (such as alignment, filtering, centring and scaling), unusual batches are identified and removed using multivariate statistical process monitoring tools, and a final multivariate model is built to identify regions of variability that relate to variations in product quality or process attributes. This framework was demonstrated on industrial data obtained from the spironolactone drug substance manufacturing process. An observation was made as a result of the PCA modelling of the process data enabling further investigation to focus on the crystallization. From this work, poor control of the solvents into the reactor was found to be a possible root cause for the variability in drying times. Furthermore, this case study demonstrated the importance on having appropriate instrumentation available to enable the identification of control improvements.

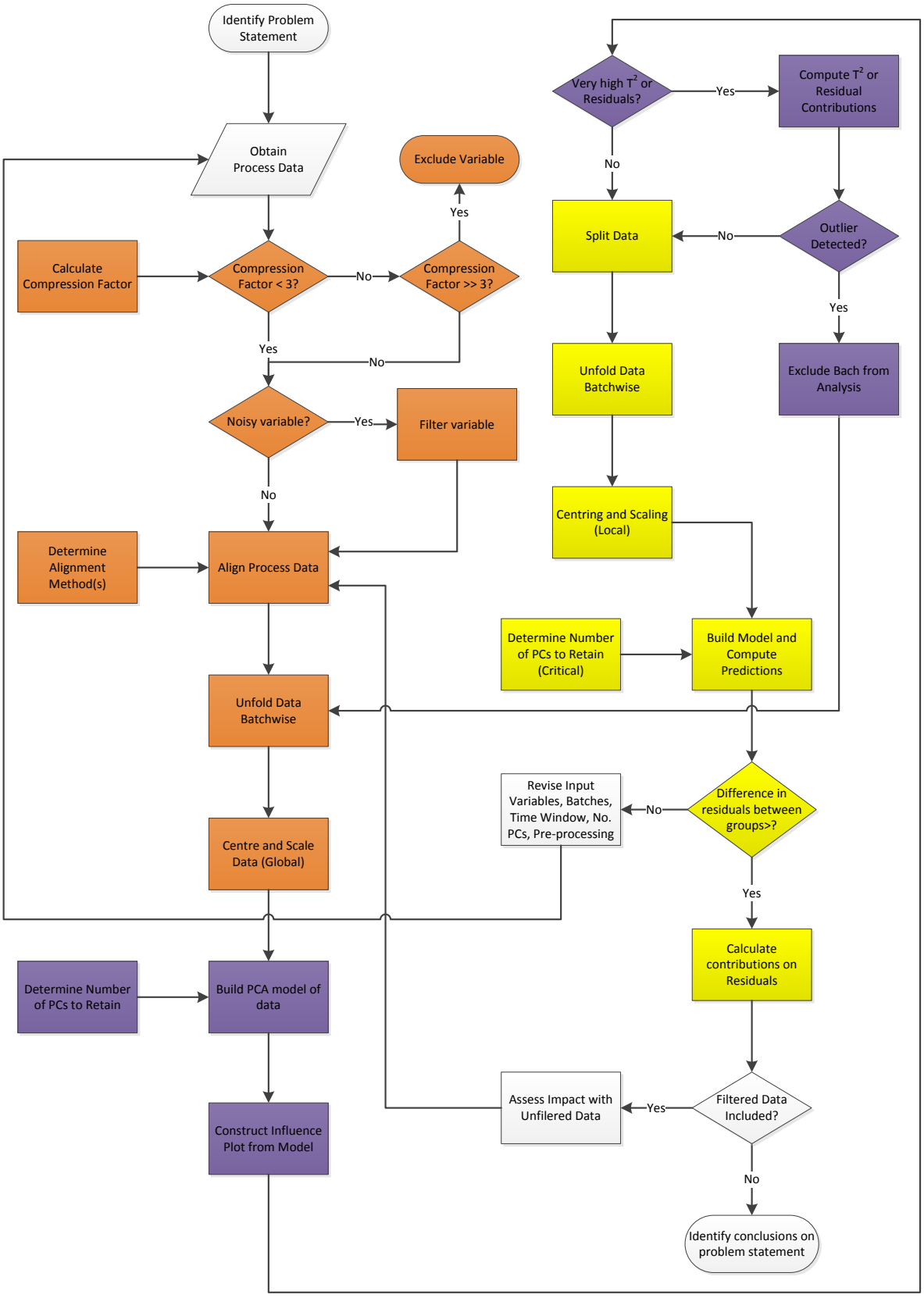


Figure 7.1: Overview of framework to extract multivariate information from batch process data (orange - pre-processing, blue - outlier detection, yellow - multivariate modelling, grey support the framework)

7.2 Future work

This thesis presented a framework for exploring historical batch process data, to extract insights on where process control can be improved. This was demonstrated by the application of multi-way principal component analysis (MPCA) and the dynamic extension of this (DPCA) to process data obtained from a commercial scale batch pharmaceutical manufacturing plant. Some of the challenges associated with such data have been demonstrated and the multivariate tool has been shown to identify atypical batches and problems with the control system of the manufacturing plant. Additionally, the multivariate data analysis has also identified an area to focus further studies on the spironolactone manufacturing process to address the challenges currently faced with uncontrolled variability in drying process cycle time.

In order to understand the crystallization and polymorphic behaviour of spironolactone further the following future work is recommended:

1. The work performed by Nicolaï et al. (2007) on spironolactone polymorphic forms should be continued and expanded to include the solvent systems used for the manufacture of spironolactone at Piramal Healthcare, namely acetone and methanol.
2. Furthermore, building on from the results discussed in chapter 6 samples of spironolactone should be isolated from the dryer both before and after the drying process and characterized by X-ray powder diffraction as a minimum to investigate the crystal forms found at the start and end of the drying process and validate the hypothesis from the NIR study. This could be further expanded to samples from multiple batches to investigate the batch to batch variability in crystal form produced.

For the crystallization monitoring using IR spectroscopic techniques discussed in chapter 6 the following future work is recommended:

1. One limitation of the study was the time available for the experimental work. As a result only a small number of experiments were able to be carried out. Further to this the ATR-FTIR instrument was only available for a small number of these runs due to the availability of liquid nitrogen in the instrument. Further repeats of the experiment would be preferred to obtain more statistical confidence in some the results obtained.
2. Another limitation of the study was the design of the NIR probe which was only able to reach the upper fill volumes of the reactor. The use of a more suitable NIR probe that is capable of integrating into a reactor at a depth allowing for the sample window to be

relatively free of solid fouling. This would be beneficial to retrieve any information that the NIR instrument is capable of obtaining, for example crystal shape / size characteristics. This would also need calibration work and/or verification through the use of a tool such as on-line video microscopy for example.

3. The initial set of experiments was set up using a factorial design approach that would allow for the examination of multiple factors with a limited number of experimental runs. This design could not be carried out, however, due to unforeseen circumstances including the inability to operate the reactor at any fill level other than full (due to the NIR probe penetration length). The number of runs and factors under investigation were therefore severely reduced in order to obtain some useful information whilst the instruments were all available. The randomised run order was also affected due to this redesign of the study, though the runs were performed in an order with limited repeats to try and account for any effects of run order. If the study were to be repeated, the use of a factorial design could be a useful way to investigate the solvent ratio, and cooling rate factors again in addition to the spironolactone concentration factor that could not be included in the study.

For the multivariate study of the spironolactone process data discussed in chapter 5 the following future work is recommended:

1. Due to the continuous changes made to the spironolactone process, especially the recent change from the OD to the NMP process, only a relatively small number of batches were available to include in the PCA models of the drying process. Further models could be constructed on a larger dataset when more batch data becomes available making the model more robust or more discriminatory. The selection of the batches to include in the training set for the dryer models was made to ensure that the training data set included batches across a relatively representative range of batches. An increased availability of batches may enable a more random batch selection process and improve the model.

2. The availability of more batches would allow for the PCA models to be constructed to investigate the variability in process yield. The number of batches available at present is insufficient as the dataset would need to be grouped into four subsets based on whether the batch was manufactured following a clean of the process train, and if recovery solutions were added to the batch. It may also be possible to use a multi-block approach

to combine four models, one on each of the sub groups, to investigate the variability of yields if the number of batches in one or more of the sub groups remains low.

3. The benefit of alternative approaches in terms of the ease of modelling and interpreting the models would be interesting to study for this process. Other multivariate techniques such as supervised distance preserving projections (SDPP) may be able to identify features of interest that have not been found in the variance identified by PCA (Zhu et al., 2013; Corona et al., 2014). Another approach that may improve the discriminatory power between the batches would be to pass the principal component scores into a classifier such as support vector machines SVM (Mahadevan and Shah, 2009), or partial least squares (PLS).

For the specific case study of the spironolactone process at Piramal Healthcare, Morpeth, UK, discussed in chapter 5, poor control of the solvent charges to the reactor were identified as a potential root cause to the drying time variability in addition to a potential methanol solvate created during crystallization. As a result, the following future work is recommended:

1. Implement improved standardization of the manufacture of batches, especially around the solvent charges at the start of the batch and throughout manufacture if recovery solutions (spironolactone in acetone) are to be added to the batch following cleaning. Improved standardization of the cooling during crystallization will also be beneficial in developing a better understanding on how variability in the process impacts the drying time and yields.

2. The lab scale investigation was a simplified study on only the crystallization. Commercial production occurs with the reaction and crystallization occurring simultaneously in order to drive the reaction to the desired isomeric form. A study on this simultaneous reaction and crystallization with an ATR-FTIR spectroscopic probe would identify if cooling rate was indeed an important factor.

3. Application of PAT tools to the reactor and dryer would provide valuable information on the process performance. Data from these instruments collected over several batches could then be combined with the currently available information to improve the power of the multivariate models in the proposed framework for identifying where the control of the process could be improved.

7.3 Business Impact

The application of the framework presented in the thesis can help businesses, operating batch processes, identify areas of their processes that could benefit from improved control. This information may be extracted from the data already present in the data historian, or additional monitoring may need to be set up, such as applying PAT in order to obtain measurements of the appropriate variability within the process data. This can increase the speed at which a business can identify problems in how a process is operated or controlled leading to variability in process or quality attributes, and reduce the number of experiments in the lab or plant trials required to diagnose problems. Furthermore, this framework can be easily adapted to be used with continuous processes by selecting discrete periods of time as pseudo batches.

Successful implementation of such a framework, as tested on the spironolactone process at Piramal Healthcare, within the wide pharmaceutical industry can help to reduce ongoing development time and costs, and identify areas where improved process control can be implemented. This can lead to improving the quality and yields of batch processes and reducing batch failures requiring costly rework or disposal, thus driving an improvement in the economy of such manufacturing processes and overall plant utilization. Furthermore, the utilization of the existing process data allows for confidence to be built to justify the complex regulatory hurdles that may be present to perform verification experiments on the commercial assets.

Appendix A. Detailed Description of Spironolactone Process and Control

A.1 Spironolactone process control

The aldadiene to spironolactone process is manufactured using a PROVOX DCS controlled plant. PROVOX was installed at Piramal Healthcare, Morpeth in 1995 as is supplied by Emerson Process Management Ltd. The DCS is linked up to AspenTech Process Information Management System (PIMS) for the purpose of batch data collection, and the production of electronic bath reports. The system has redundancy built in with a duplicate 'off-line' system to use as a 'hot-spare' backup. This section will cover more details of the process and how it is controlled at each stage discussing each processing unit in turn.

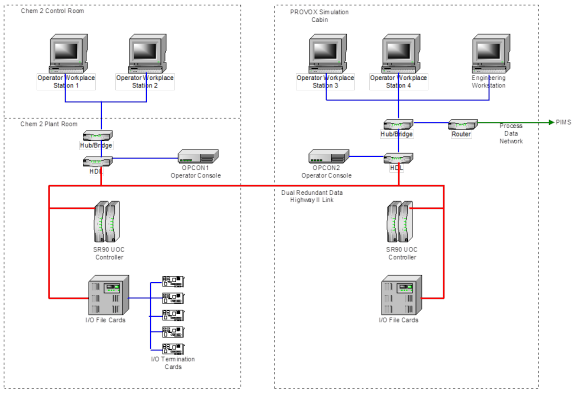


Figure A.1: PROVOX network schematic at Piramal Healthcare (Bell, 2008)

A.1.1 Reactor R101: thiolacetylation and isolation

The thiolacetylation and isolation reactor is controlled by PROVOX through the sequence illustrated in figure A.2. Each of the steps has the required settings for the batch contained within the batch recipe. It starts by checking that the reactor is ready and that a valid batch number has been set. The reactor is then evacuated approximately 100 mbarA and purged with nitrogen three times (vacuum-purge). The appropriate valves are then opened to allow the operator to transfer the colour treated aldadiene in methanol and acetone to the reactor via the

0.45μm and 0.2μm filters. After the transfer is complete, the reactor contents are heated to reflux for 6 minutes with the timer starting when the reactor contents exceed 60 °C. After the reflux, the steam supply to the reactor jacket is isolated and thiolacetic acid is charged to the reactor, controlled by weight change in the thiolacetic acid header tank, T104. The reactor is then returned to reflux at 62 °C for 20 minutes. The reactor jacket is then filled with cold water to start the batch cooling. As soon as the jacket is showing full of cold water by activation of a level switch in the jacket, the water is blown out of the jacket using compressed air and the batch remains slowly cooling until the operator confirms that crystallization has been observed in the reactor. Heat is then applied to the reactor to return the contents to reflux at a minimum of 64 °C for 100 minutes to ensure that the reaction goes to completion. After the 100 minute reflux, the steam is isolated and the condensate in the jacket blown clear with plant air. A quantity of methanol is charged to the reactor R101 controlled by weight change on the receiving reactor. There is then an option to add recovery material (spironolactone that has been recovered from the first wash of the process train with acetone). More methanol is then added to the reactor again controlled by the weight change in R101. The reactor is then cooled to 40 °C using cold water in the reactor jacket after which the jacket service is changed to chilled glycol to bring the reactor contents down to -10 °C. The batch is held at -10 °C for a minimum of two hours before it is filtered. The batch is filtered in two parts in the Rosenmund pressure filter F101, the charge controlled by the weight change in reactor R101. Whilst the first load is being filtered the remaining spironolactone slurry in R101 remains held at -10 °C.

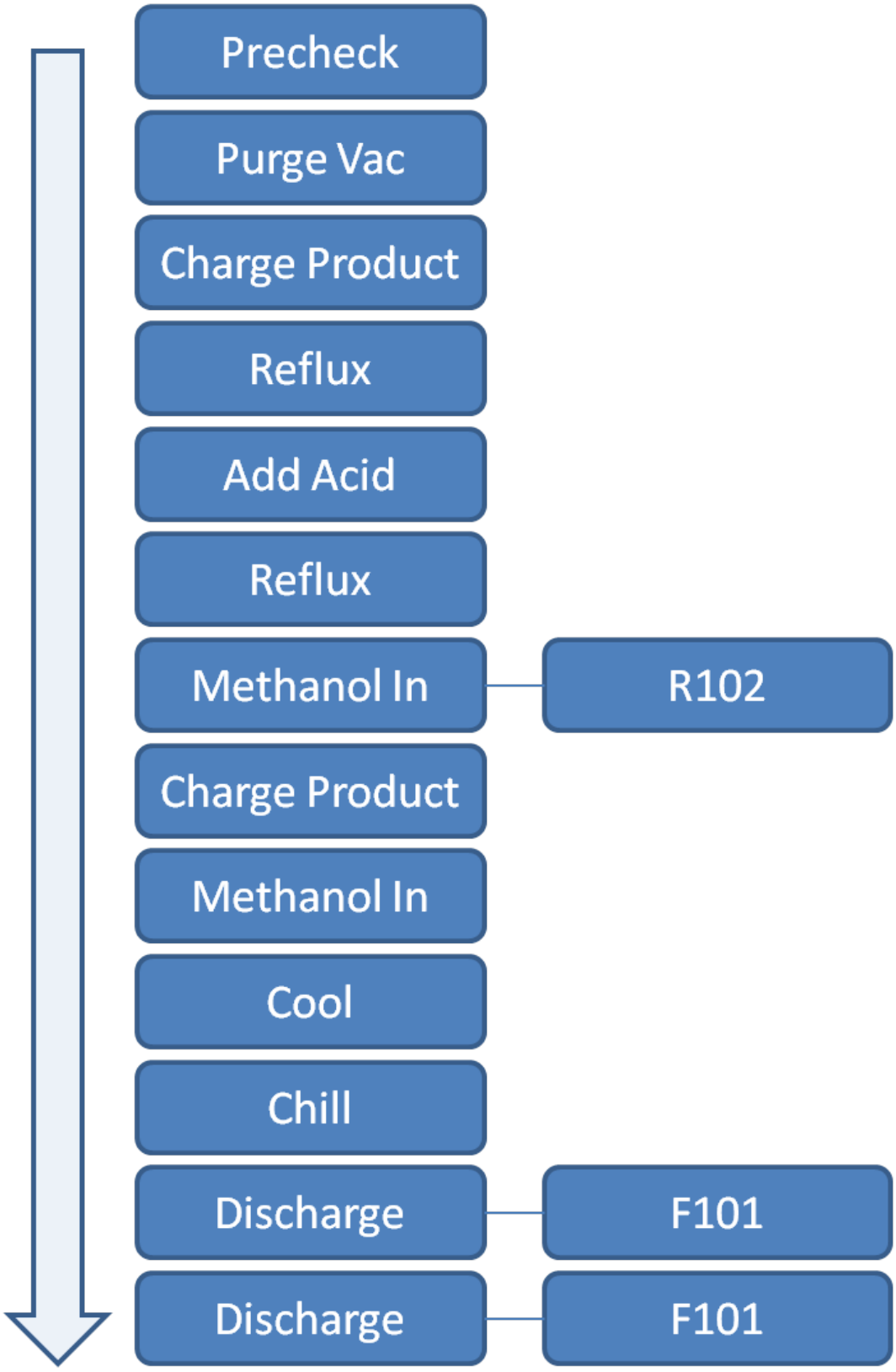


Figure A.2: Overview of reactor R101 control strategy

Reactor R101 temperature control

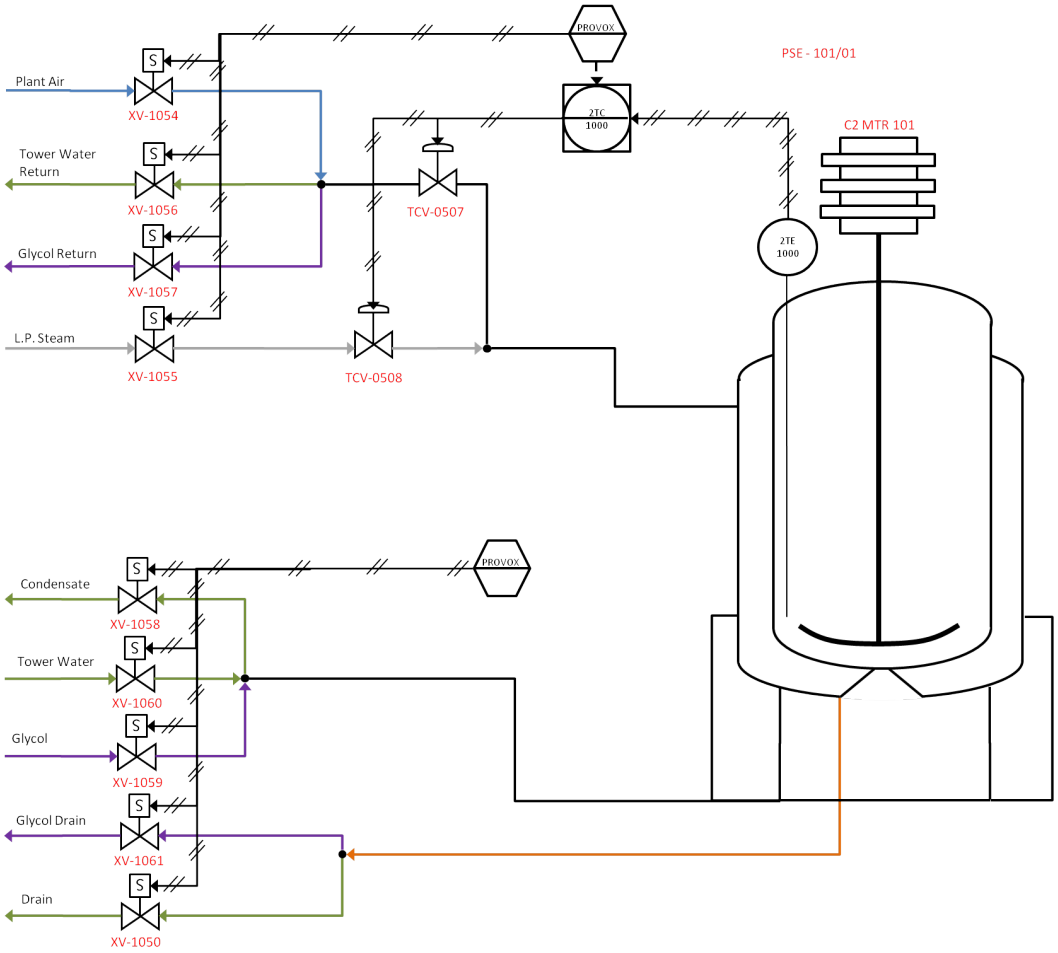


Figure A.3: Reactor R101 temperature control

The reactor contents temperature set-point and jacket service is specified in the PROVOX batch recipe, which dictates the solenoid control valves that are manipulated in order to deliver the appropriate heat transfer medium to the reactor jacket. The reactor jacket is blown clear between the use of both glycol and water to ensure that no water enters the glycol system which could freeze up the chillers, and conversely that no glycol enters the water system. It is not always necessary to blow the condensates clear as following steam to the jacket, the subsequent service called should only be either steam , or water.

The control valves (TCV-0507 and TCV-0508) are operated by the same controller (2TC1000) and have an opposing action. That is, when TCV-0507 is at 80%, TCV-0508 is at 20% and vice versa. This configuration allows for the same controller to be used for both heating and cooling operations and is possible due to the presence of automated isolation valves XV-1054 and XV-1057. The temperature is controlled through a closed loop feedback control by a

Proportional + Integral + Derivative (PID) controller embedded within the PROVOX software.

Reactor R101 pressure control

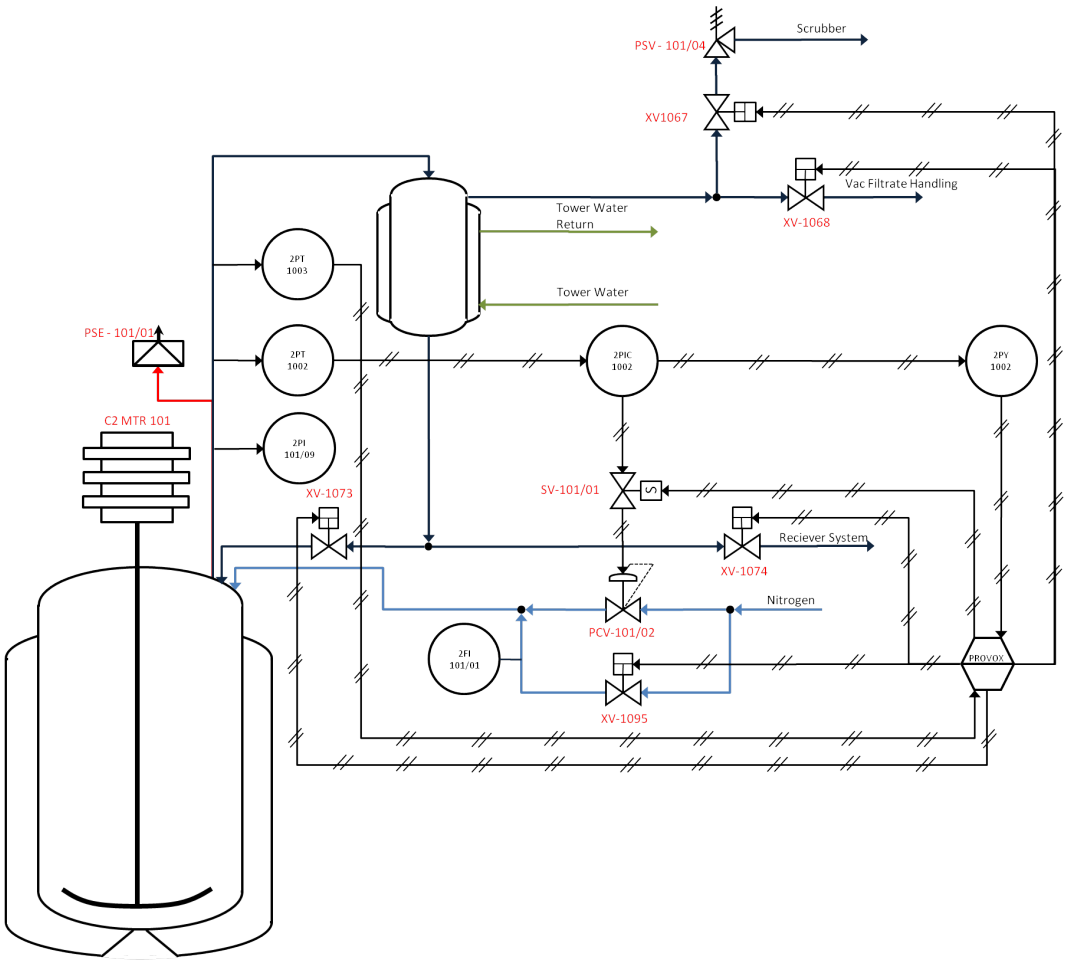


Figure A.4: Reactor R101 pressure control

Pressure in the reactor is controlled through the PROVOX DCS, determining which mode of operation is required and selecting the appropriate sequences and control as required. When vacuum is required, PROVOX ensures that all of the relevant valves are in the closed position through feedback on the limit switches on the valves. When the vacuum pump is available, PROVOX opens valve XV-1068 which remains open until the value for vacuum pressure, stored in the PROVOX sequence, has been seen in 2PT1003. Vacuum is then broken through modulation of control valve PCV101/02, again until 2PT1003 has a signal above that set in the PROVOX sequence. The vent valve XV-1067 is then opened allowing a nitrogen blanket to be maintained in the reactor without over pressuring.

When the solenoid valve SOV-1002 is energised, the control valve (PCV101/02) input is

vented resulting in the valve closing. The controller is a field mounted Proportional + Integral (PI) pneumatic controller and is operated in reverse acting mode. As pressure on the controller input increases, the controller output pressure decreases, closing the valve. This modulates the pressure in the reactor to ensure that there is a nitrogen blanket present when required.

During other operations in reactor R101, a 'nitrogen sweep' is required. This is where a larger flow of nitrogen is continuously blown into the head space of the reactor and vented to the scrubber system. This is achieved through PROVOX opening the bypass valve XV-1095 whilst the vent is open. The reactor pressure is controlled through a conservation vent (PSV101/04) on the vent line set to 1 barA.

PROVOX also controls the handling of the distillate and can either allow the reactor to operate in reflux with all of the distillate returning to the reactor, when XV-1073 is open (and XV-1074 is closed), or XV-1074 can be opened and XV-1073 closed to remove all the distillate to the tank receiver system.

Overpressure on reactor R101 is prevented through a bursting disc which vents the reactor contents to the roof of the chemical plant. The reactor R101 does not require under pressure protection as the vessel is rated to operate under vacuum.

Reactor R101 weight control

The aldadiene start material starts in the dryer where the intermediate product was dried. Solvent is charged to the dryer to dissolve the aldadiene through flow meter FM SOL/11 and the flow control valve FCV SOL/11. This is connected to the dryer through a flexible pipe (not shown in figure A.5). The flexible pipe work arrangement is changed to connect reactor R506 to the solvent meters and more solvent is charged to the batch through the same flow metering system (SOL/11). The batch is then filtered through the bag filters C2 FLT0101/02 and transferred to reactor R101. Reactor R506 is then washed twice with more solvent, again metered into the reactor, which is also routed through the filters to ensure that the entire batch has been transferred.

To start the reaction to produce spironolactone, thiolacetic acid (TAA) is charged from the header tank (T104) into the reactor R101. The quantity of TAA is controlled by weight change on T104. The weight is measured by WE1005 which is sent to PROVOX. When the weight change comes to within 25 kg of the desired mass of TAA, valve XV1079A is closed. As

XV1079A has restricted closing, the flow is only reduced and not completely stopped. When the weight change comes to within 5 kg XV1079B is also closed. PROVOX then pulses the flow of TAA into R101 by opening and closing valve XV1079B a predefined number of times. If the target weight is not achieved during these pulses the operator is asked if the pulses should be repeated in order to achieve the correct charge of TAA.

After the reaction, methanol is charged to the reactor R101 to aid with the crystallisation. This is controlled by the weight of R101 being monitored by PROVOX via WE1004 and WY1004. PROVOX opens valve XV1093 and waits until the change in weight exceeds the set point. Originally this was done twice for two subsequent additions of methanol with a gap between them to allow for recovery solutions to be charged if required. From the 31st March 2011, only the second addition of the solvent was carried out using this method. The first methanol charge is performed manually via R506 and measured through the solvent meter FM SOL/11 the same way that the batch is charged. This was done in an attempt to reduce seasonal variation in methanol temperature affecting crystallization, as the methanol can be heated to 25 °C in reactor R506 before it is charged to R101.

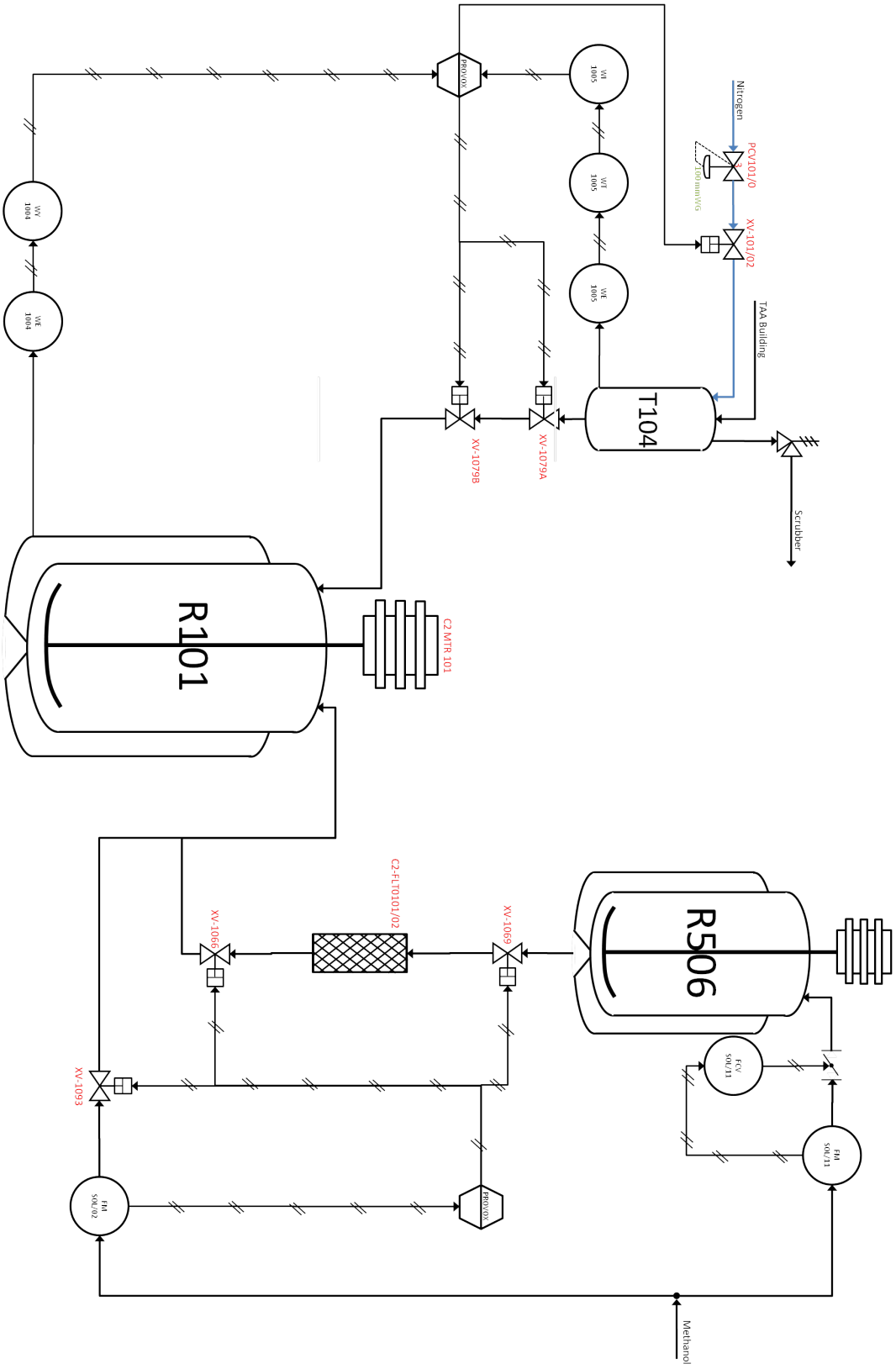


Figure A.5: Reactor R101 weight control

Reactor R101 agitator and level control

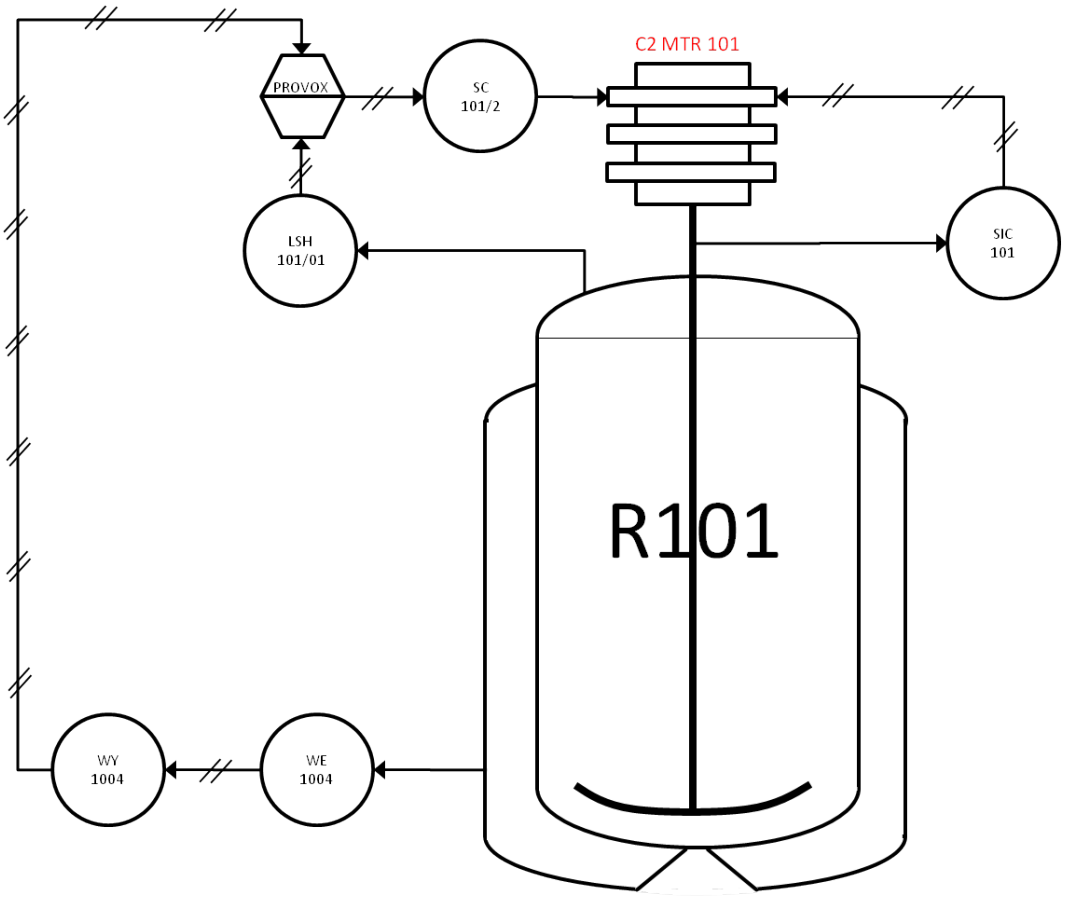


Figure A.6: Reactor R101 level control

The variable speed agitator is controlled locally on plant where the operator adjusts the power input until the desired speed is shown on SIC 101. PROVOX has overall control of the agitator however and turns power to the agitator on or off based on the weight of the vessel. If reactor R101 is more than 150 kg, solenoid SC 101/2 energises allowing the agitator to run and if the weight drops below 150 kg PROVOX stops the agitator by de-energising solenoid SC 101/2.

There is also a high level switch on reactor R101 (LSH 101/01) to prevent overfilling the reactor. This is linked to PROVOX which is monitored as part of the interlock logic to which will shut all inlet valves when the level switch is active.

A.1.2 Filter F101: isolation of spironolactone

The filter F101 is controlled by PROVOX through the sequence illustrated in figure A.7. The filter is first purged 3 times with nitrogen and vented to atmospheric pressure. The reactor

R101 contents are then transferred to the filter with a recycle back to the reactor. When the recycle stream is observed as clear by the operator, the return line to the reactor is closed and half of the batch is charged to the filter. This recycle is to ensure that a cake is built up on the filter plate to prevent losses of product through the filter plate. The filter is then pressurised and the filtrates are collected. When the liquid is no longer visible on the surface of the cake, cold methanol at -10 °C is charged to the filter in a displacement wash. The filter is again pressurised and the filtration liquors removed. When the liquid is not visible on the surface of the cake, more methanol at -10 °C is added and the cake re-slurried before the filter is pressurised and the filtration liquors again removed. Again when the liquors are not visible on the surface of the cake, methanol at -10 °C is added to the filter in a final displacement wash. This is again pressurised and the filtrates are removed. The filter cake is blown with pressurised nitrogen for at least 1 hour until no liquors can be seen by the operator flowing to the filtrate tanks. The filter cake is then discharged to the dryer before the process is repeated with the other half of the batch.

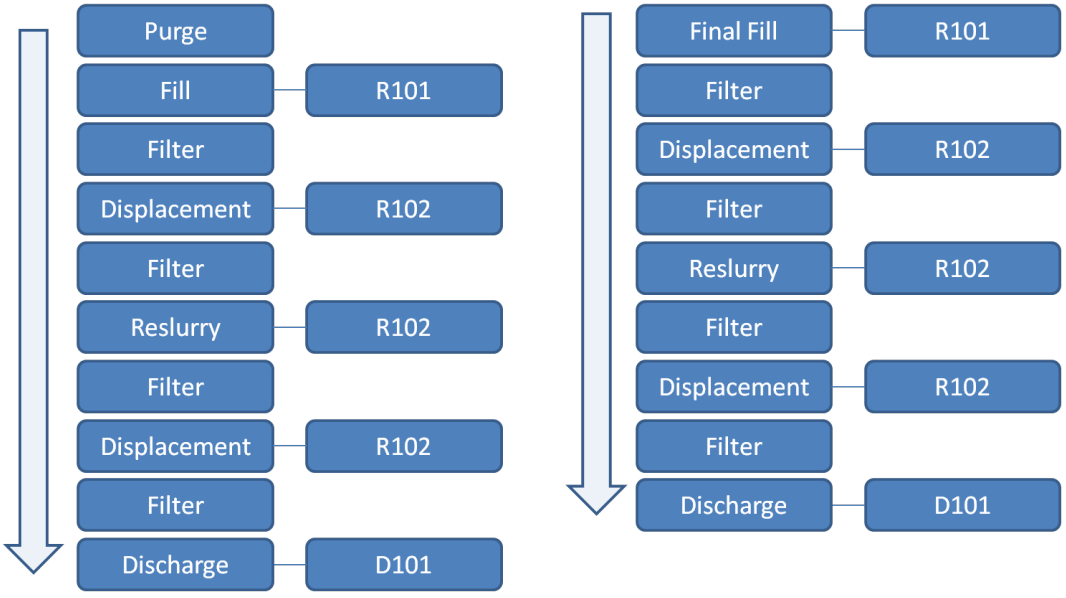


Figure A.7: Overview of filter F101 control strategy

Filter F101 pressure control

The filter is purged three times with nitrogen at the start of each batch. Before this is done, PROVOX opens the pressure equalisation valve XV1194 to equalise the pressure either side of the filter plate. There are two pressure switches on F101; PS F101P is set at 0.1 barG rising and therefore allows the operator to see if there is any pressure in the filter and prevent the discharge slide valve XV1061 from opening if this is the case. PS1110B is set at 1.2 barG

rising and is used as an indicator of successful pressurisation of the filter for purges and filtration. The purges are achieved through opening the vent (XV1193) until the pressure switch PS F101P is not active, meaning that the pressure is below 0.1 barG. The vent (XV1193) is then closed and PROVOX puts nitrogen into F101 by fully opening PCV 1130. PCV 1130 is fully opened by operating solenoid valve PY 1130 so that the signal to the control valve is to fully open. When the high pressure switch PS 1110B is active, meaning the pressure is in excess of 1.2 barG, the solenoid valve PY 1130 is changed back so PCV 1130 is closed at high pressures. The vent valve (XV 1193) is again opened to allow the pressure to fall until it falls below 1.2 barG and pressure switch PF F101P is not active. This process is repeated three times before PROVOX moves to the fill operation where the pressure equalisation valve XV1194 is closed.

The filter operation pressurises F101 to assist the filtration liquors move through the cake. To achieve this PROVOX closes the vent XV1194, and ensures the pressure equalisation valve XV1194 is closed, before setting PCV 1130 to fully open by manipulation of solenoid valve PY 1130. At this point the operator selects manual control from PROVOX. When the operator is satisfied that the filtration is ready to receive the next wash, they isolate the nitrogen to the filter to allow the pressure to drop below 1 barG before giving back control to PROVOX to allow the filter to vent. This is to ensure that the cake has de-watered and that the pressure in the filter is not too high, that venting it through the scrubber will cause the scrubber to overflow. PROVOX then vents the filter by opening the pressure equalisation valve XV1194 and then the vent valve XV1193 and putting the nitrogen back to modulating around atmospheric pressure. When pressure switch PS F101P is not active PROVOX closes XV1193 and XV1194 ready for the next wash or discharge. During washes ambient pressure is maintained through pressure transmitter PT1130 and controller PC1130. The controller is reverse acting; so on rising pressures the output reduces closing PCV1130. This can be overridden by PY1130 which upon receipt of a digital signal from PROVOX applies compressed air to at 1.4 barG to PY1130 to fully open PCV1130.

The pressure measurement that is available on the Process Information Management System (PIMS) comes from PT F101/1, a 0 – 300 mmWG pressure transmitter. When the filter is pressurised during the filtration, the pressure typically rises in excess of 2 barG (1500 mmWG) resulting in the pressure in PIMS being off scale for most of the filtration. An additional pressure transmitter (PT F101/2) was installed to the filter F101 with a range of 0 to 2 barG. This was connected to an electronic chart recorder PIR F101/2 in July 2012 to allow analysis

of the pressure changes in the filter. This measurement is not used for process control and, as the chart recorder is located inside an instrumentation panel, the measurement is not routinely available to the operators therefore they must use a field mounted pressure gauge, reading in barG, to monitor the pressure.

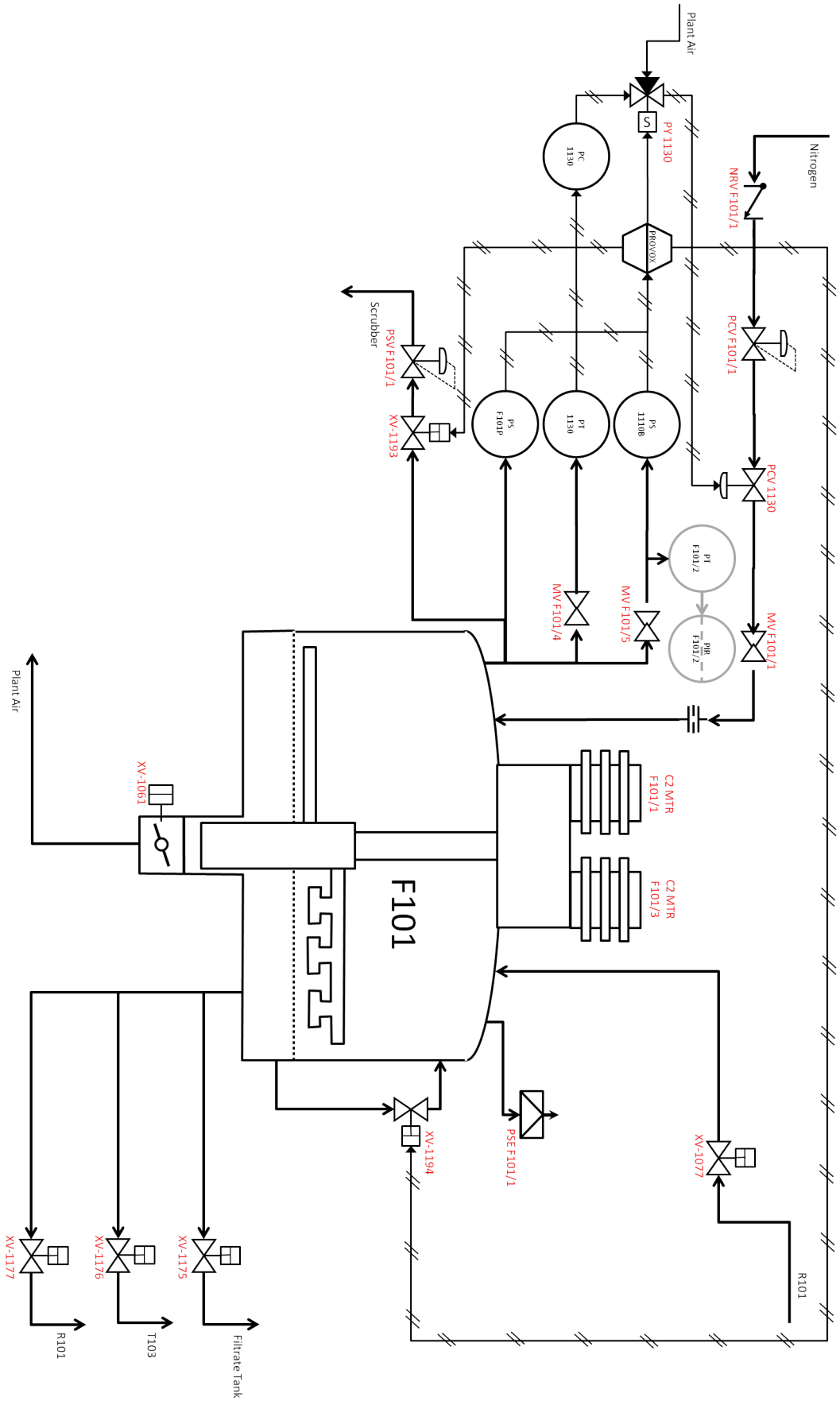


Figure A.8: Filter F101 pressure control

Filter F101 flow control

Spironolactone is filtered in two parts. To charge the each load the filter opens the reactor R101 discharge valve XV1051 and sets the path to the filter by opening valves XV1010 and XV1065 when both units are ready. PROVOX also closes the reactor recycle valve XV1064 and starts the air supply to the double diaphragm pump (between XV1051 and XV1010, not shown in figure A.9). The contents are recycled between the filter and reactor by opening XV1177 and XV1024, to allow a cake to build up on the filter plate. When the operator has confirmed that the filtrates are clear from LG F101/1, PROVOX stops the recycle to R101 by closing valves XV1177 and XV1024. It then opens the valve XV1175 to allow the filtrates to be removed to the filtrate tank. PROVOX continues to transfer material from R101 to F101 until the pre-calculated weight change of R101 (from WE1004) has been achieved. PROVOX will also stop the transfer from R101 to F101 is the high level switch in F101 (LS 1111) is active, or the transfer has taken too long.

The chilled methanol at -10 °C for all of the displacement washes and re-slurry washes comes from reactor R102. To carry out this solvent charge, PROVOX opens the R102 discharge valve (XV1052) and sets the path to F101 by opening valves XV1077 and XV1024. The weight of R102 (WE1002) is then monitored until the change in weight has exceeded the set point. The valves XV1052, XV1077 and XV1024 are then closed and PROVOX progresses the batch to the filter phase in the operation.

To discharge the filter load to the dryer, the operator must first ensure that there is no liquid present on the slide valve by opening a manual valve and draining any liquid that may be present. When both the filter F101 and dryer D101 are ready, PROVOX opens the slide valve XV1061. The discharge arm is set to rotate and slowly lowered until the bottom limit switch on the discharge arm is active.

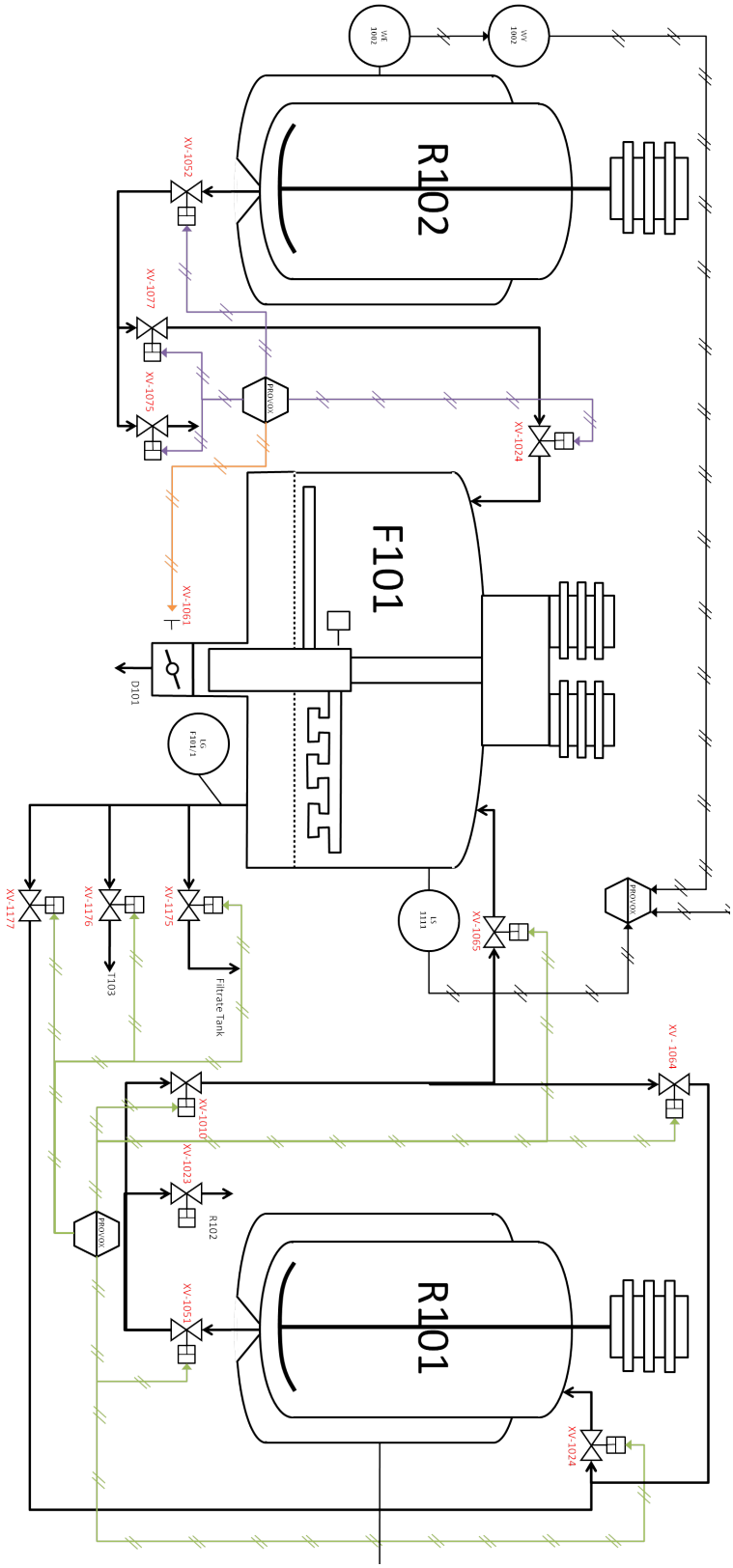


Figure A.9: Filter F101 flow control

A.1.3 Dryer D101: drying of spironolactone

The dryer is a conical, screw agitated, batch drier with contact heating from a water heated jacket. The reactor is controlled by a PROVOX sequence as shown in figure A.10. It starts by checking that the dryer is ready and that a valid batch number has been set. The dryer is then evacuated and purged with nitrogen three times. The dryer then communicates with the filter and when both units are ready the first filter load (half a batch) is dropped from the filter into the dryer below. This part batch is then dried under vacuum with a jacket temperature set to 25 °C for 3 hours. When the dry phase is complete, the dryer returns to atmospheric pressure and the jacket water circulation is turned off and the dryer waits for the filter to be ready to transfer the second load to the dryer. When the second filter load has been transferred to the dryer, it goes back into the dry phase under vacuum with a jacket temperature of 25 °C for 2 hours. The dryer then ramps the jacket temperature up to 90 °C over a period of 2 hours and holds this temperature for a further 2 hours. The dryer then enters the deodour phase, in which the pressure returns to atmospheric pressure and a nitrogen sweep of the dryer head space is carried out for 11 hours whilst the jacket temperature remains at 90 °C. The dryer then returns to vacuum for 1 hour with the jacket at 90 °C before the batch is cooled and transferred to the micronizer.

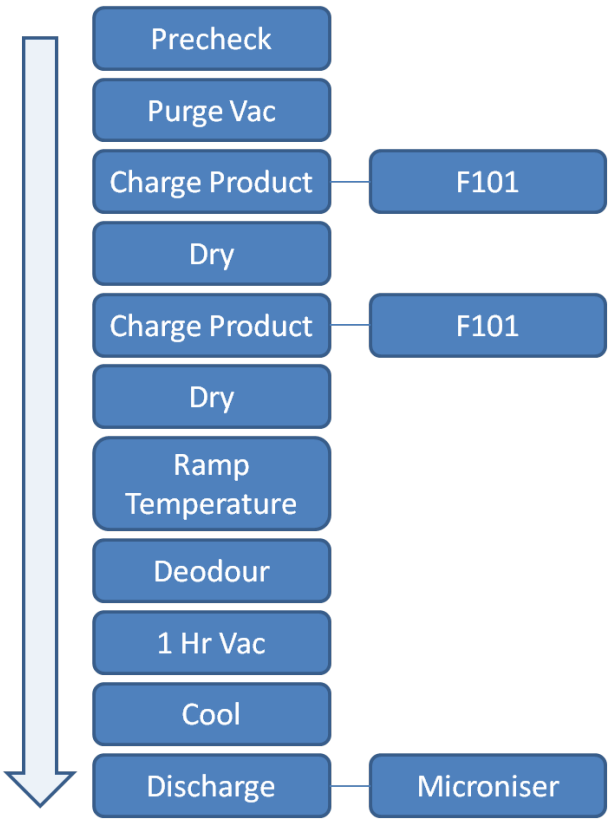


Figure A.10: Overview of control strategy for dryer D101

Dryer D101 temperature control

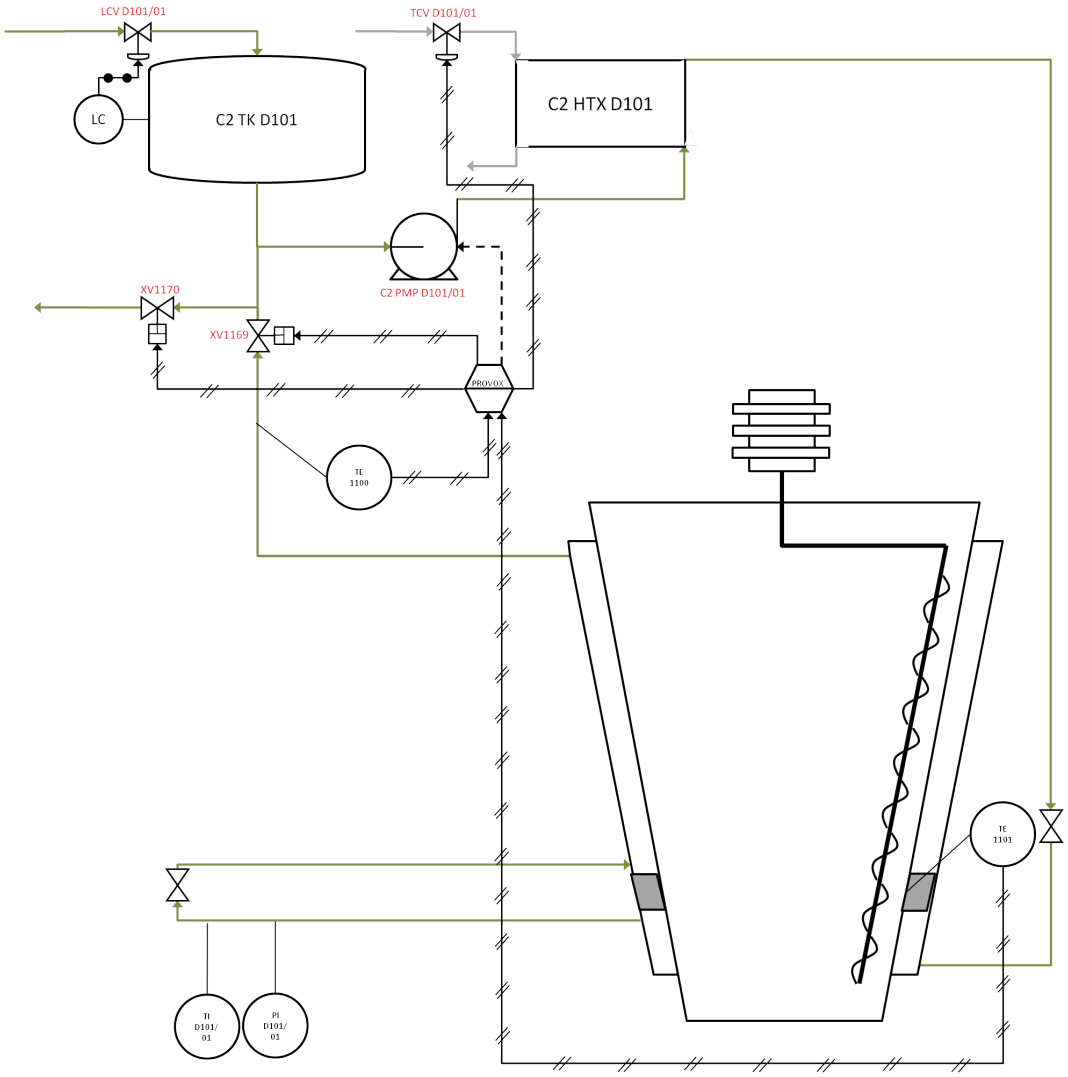


Figure A.11: Dryer D101 temperature control

The contents temperature in the dryer is only monitored and is not used for automatic control within the control sequence. The temperature is measured by a thermocouple (TE1101) in a thermo-pocket inserted near the bottom of the dryer. The contents temperature measurement is transmitted to PROVOX to allow the operators to view the measurement. It is this temperature measurement that the operators use to determine the end point for drying.

The dryer jacket temperature is controlled in two different ways, depending upon the phase's requirements in the sequence control. When the batch is in a phase requiring heat, hot water is circulated through the jacket at the set point temperature as required by the recipe. PROVOX closes the water outlet valve XV1170, opens the recirculation valve XV1169 and starts the pump C2 PMP D101/01. The temperature is controlled by a closed loop PID (Proportional + Integral + Derivative) controller embedded within the DCS. The temperature is measured

through a thermocouple mounted in the jacket water outlet line (TE1100). This is fed back to PROVOX which sends an output value to the steam control valve (TCV D101/01) on heat exchanger C2 HTX D101 where the circulating water is heated.

When the batch is in a location in the sequence that requires cooling, the valve configuration is changed so that XV1170 is open and XV1169 is closed. The pump remains on and cooled water is passed through the heat exchanger and dryer jacket, in a once through manner to cool the whole system.

The system is kept primed with water through the header tank C2 TK D101 which uses a mechanical device to close the valve LCV D101/01 when the tank is near full.

Dryer D101 pressure control

The dryer has three modes of pressure operation, high vacuum, low vacuum and atmospheric pressure. Low vacuum (typically 0.18 barA) is achieved through a liquid ring vacuum pump X103, which is shared with the other process units in the 'chemical 2' facility. High vacuum (typically 0.08 barA) is obtained through a high vacuum pump X106, to which dryer D101 has the exclusive use.

Spironolactone is initially dried under high vacuum conditions. PROVOX first brings the dryer down to 0.3 barA using the liquid ring pump. This is done by closing all the valves on the pressure system and starting the liquid ring vacuum pump (X103), if it is not in use by another unit. When X103 is available and running, valve XV1174 is opened until the pressure on PT1102 drops below 0.3 barA. The valve XV1174 is then closed and the high vacuum pump, X106, is selected. When X106 is running, the valve XV1180 is opened to reduce the pressure in D101 further. Vacuum is maintained by leaving X106 pulling vacuum on the dryer with XV1180 open. To return the dryer to atmospheric pressure PROVOX closes both XV1180 and XV1174 and stops the relevant vacuum pump. The vacuum breaker valve XV1171 is opened to bring the dryer back to 0.9 barA before the vent valve XV1173 can also be opened.

After drying, the dryer carries out a deodour phase where a nitrogen sweep over the top of the batch removes any mercaptans from the batch. This is achieved by closing all the pressure valves and opening both the sweep valve XV1182 and the vent valve XV1173. The pressure is controlled below 1.3 barA. This is achieved by modulating valve XV1182. If the pressure from

transmitter PT1102 is equal to or greater than 1.15 barA the sweep valve XV1182 is closed until the pressure drops below 1.1 barA.

To ensure that there is a flow of nitrogen through the dryer head space, PROVOX monitors the flow switch FIS1197. If FIS1197 indicates low flow and the sweep valve XV1182 is open PROVOX fails the operation and alerts the operator that there is no nitrogen sweep flow to the dryer.

To ensure that the filter on the dryer vapour outlet is clean, there is a back pulse system to inject pulses of nitrogen through the filter into the dryer. This knocks any accumulated product from the filter back into the batch. The system works by filling a reservoir with pressurised nitrogen. This nitrogen is released to the filter by opening valve KV1135 which is controlled by a timer.

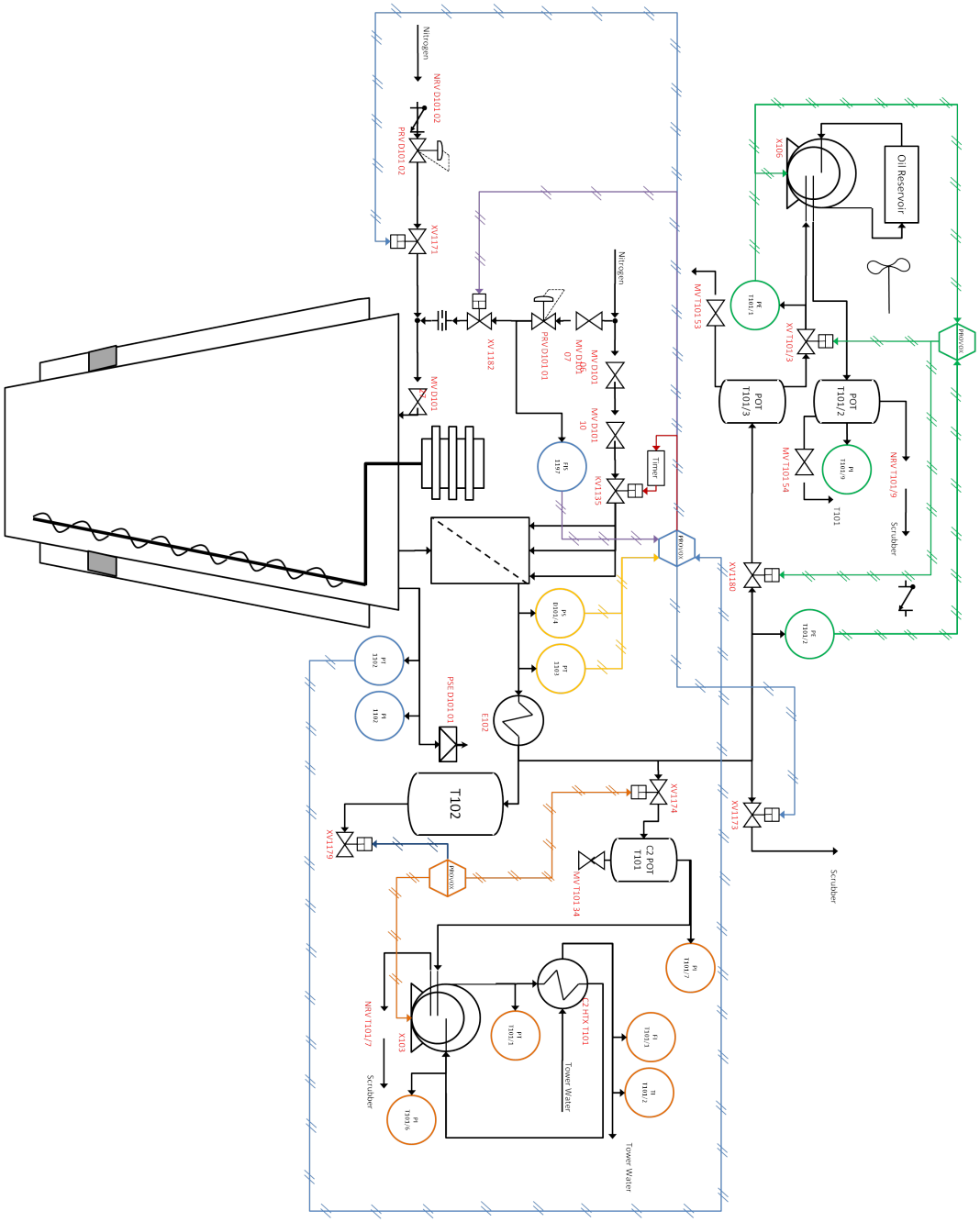


Figure A.12: Dryer D101 pressure control

References

Abdi, H. and Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Agafonov, V., Legendre, B., and Rodier, N. (1989). A new crystalline modification of spironolactone. *Acta Crystallographica Section C*, 45(10):1661–1663.

Agafonov, V., Legendre, B., Rodier, N., Wouessidjewe, D., and Cense, J.-M. (1991). Polymorphism of spironolactone. *Journal of Pharmaceutical Sciences*, 80(2):181–185.

AlGhazzawi, A. and Lennox, B. (2008). Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice*, 16(3):294–307.

Amer, H. H., Paxton, R. R., and Winkle, M. V. (1956). Methanol-Ethanol-Acetone. *Industrial & Engineering Chemistry*, 48(1):142–146.

Angulo, C. and Godo, L., editors (2007). *Artificial Intelligence Research and Development*. IOS Press, Amsterdam.

Arteaga, F. and Ferrer, A. (2002). Dealing with missing data in MSPC: Several methods, different interpretations, some examples. *Journal of Chemometrics*, 16(8):408–418.

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.

Barsanti, R. J. and Athanason, A. (2013). Signal compression using the discrete wavelet transform and the discrete cosine transform. *2013 Proceedings of IEEE Southeastcon*, pages 1–5.

Bell, A. (2008). IS System Overview - Provox.

Ben Yahia, B., Gourevitch, B., Malphettes, L., and Heinzle, E. (2016). Segmented linear modeling of CHO fed-batch culture and its application to large scale production. *Biotechnology and Bioengineering*.

Bersimis, S., Psarakis, S., and Panaretos, J. (2007). Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International*, 23:517–543.

Borchert, S. O., Voss, T., Schuetzmeier, F., Paul, J., Cornelissen, G., and Luttmann, R. (2015). Development and monitoring of an integrated bioprocess for production of a potential malaria vaccine with Pichia pastoris. *Journal of Process Control*, 35:113–126.

Bristol, E. H. (1990). Swinging door trending: adaptive trend recording. *Proceedings of the ISA National Conf.*, pages 749–753.

Bro, R. and Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33.

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9):2812.

Brown, C. D. and Wentzell, P. D. (1999). Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. *Journal of Chemometrics*, 13(2):133–152.

Buckton, G. (2013). *Aulton's Pharmaceutics*. Churchill Livingstone Elsevier, 4 edition.

Burggraeve, A., van den Kerkhof, T., Hellings, M., Remon, J. P., Vervaet, C., and Beer, T. D. (2011). Batch statistical process control of a fluid bed granulation process using in-line spatial filter velocimetry and product temperature measurements. *European Journal of Pharmaceutical Sciences*, 42(5):584–592.

Camacho, J., Lauri, D., Lennox, B., Escabias, M., and Valderrama, M. (2015). Evaluation of smoothing techniques in the run to run optimization of fed-batch processes with u-PLS. *Journal of Chemometrics*, 29(6):338–348.

Camacho, J., Picó, J., and Ferrer, A. (2008). Multi-phase analysis framework for handling batch process data. *Journal of Chemometrics*, 22(11-12):632–643.

Cella, J. A., Forest, L., and Tweit, R. C. (1961). United States Patent: Alkanoylthio-17-carboxyethyl-17-hydroxyandrosten-3-one lactones.

Chakravarty, P., Bhardwaj, S. P., King, L., and Suryanarayanan, R. (2009). Monitoring Phase Transformations in Intact Tablets of Trehalose by FT-Raman Spectroscopy. *AAPS PharmSciTech*, 10(4):1420–1426.

Chen, H., Wang, Y. F., Yang, Z. D., and Li, Y. C. (2006). Isolation and Identification of Novel Impurities in Spironolactone. *Journal of Pharmaceutical and Biomedical Analysis*, 40(5):1263–1267.

Chen, J. and Liu, K.-C. (2002). On-line Batch Process Monitoring using Dynamic PCA and Dynamic PLS Models. *Chemical Engineering Science*, 57:63–75.

Coates, J. (2006). Interpretation of Infrared Spectra, A Practical Approach. In *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, Chichester, UK.

Conlin, a. K., Martin, E. B., and Morris, a. J. (2000). Confidence limits for contribution plots. *Journal of Chemometrics*, 14(5-6):725–736.

Corona, F., Zhu, Z., de Souza Júnior, A., Mulas, M., Muru, E., Lorenzo, S., Barreto, G., and Baratti, R. (2014). Supervised Distance Preserving Projections: Applications in the quantitative analysis of diesel fuels and light cycle oils from {NIR} spectra. *Journal of Process Control*, Article in:http://dx.doi.org/10.1016/j.jprocont.2014.11.005.

Dahl, K. S., Piovoso, M. J., and Kosanovich, K. a. (1999). Translating third-order data analysis methods to chemical batch processes. *Chemometrics and Intelligent Laboratory Systems*, 46(2):161–180.

De Beer, T. R. M., Vercruysse, P., Burggraeve, A., Quinten, T., Ouyang, J., Zhang, X., Vervaet, C., Remon, J. P., and Baeyens, W. R. G. (2009). In-line and real-time process monitoring of a freeze drying process using Raman and NIR spectroscopy as complementary process analytical technology (PAT) tools. *Journal of Pharmaceutical Sciences*, 98(9):3430–3446.

De Griffiths, P. and Haseth, J. A. (2007). *Fourier Transform Infrared Spectroscopy*. John Wiley & Sons.

Dempster, A. P., Laird, N. M., and D.B. Rubin (1976). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):290–295.

Dryden, H. L. and Markos, C. S. (1977). United States Patent: Process for the preparation of 17-hydroxy-3-oxo-17-pregn-4-ene-21-carboxylic acid lactone.

El-Dalsh, S. S., El-Sayed, a. a., Badawi, a. a., Khattab, F. I., and Fouli, a. (1983). Studies on spironolactone polymorphic forms. *Drug Development and Industrial Pharmacy*, 9(5):877–894.

Espeau, P., Nicolaï, B., Céolin, R., Perrin, M.-A. A., Zaske, L., Giovannini, J., and Leveiller, F. (2007). Thermal behavior of orthorhombic polymorphs I and II of spironolactone. *Journal of Thermal Analysis and Calorimetry*, 90(2):341–342.

European Commission (2012). Volume 4: Good Manufacturing Practice Medicinal Products for Human and Veterinary Use. In *The Rules Governing Medicinal Products in the European Union*, chapter 4, pages 1–9. Brussels.

Fevotte, G. (2002). On-Line Monitoring of Batch Pharmaceutical Crystallization Using ATR-FTIR Spectroscopy. In *IFAC 15th Triennial Wold Congress*, Barcelona.

Flores-Cerrillo, J. and MacGregor, J. F. (2004). Multivariate monitoring of batch processes using batch-to-batch information. *AIChE Journal*, 50(6):1219–1228.

Folch-Fortuny, A., Arteaga, F., and Ferrer, A. (2015). PCA model building with missing data: New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 146:77–88.

Fransson, M. and Folestad, S. (2006). Real-time alignment of batch process data using COW for on-line process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 84(1-2 SPEC. ISS.):56–61.

Fujiwara, M., Chow, P. S., Ma, D. L., and Braatz, R. D. (2002). Paracetamol Crystallization Using Laser Backscattering and ATR-FTIR Spectroscopy: Metastability, Agglomeration, and Control. *Crystal Growth & Design2*, 2(5):363 – 370.

Gabrielsson, J., Lindberg, N.-O., and Lundstedt, T. (2002). Multivariate Methods in Pharmaceutical Applications. *Journal of Chemometrics*, 16(3):141–160.

Gallagher, N. B., Wise, B. M., Watts Buttler, S., White, D. D. J., and Barna, G. G. (1998). Development and Benchmarking of Multivariate Statistical Process Control Tools for a Semiconductor Etch Process: Improving Robustness Through Model Updating. Technical report, Eigenvector Research Inc.

García-Muñoz, S., Kourti, T., MacGregor, J. F., Mateos, A. G., and Murphy, G. (2003). Troubleshooting of an Industrial Batch Process Using Multivariate Methods. *Industrial Engineering and Chemistry Research*, 42(15):3592–3601.

García-Muñoz, S., Polizzi, M., Prpich, A., Strain, C., Lalonde, A., and Negron, V. (2011). Experiences in batch trajectory alignment for pharmaceutical process improvement through multivariate latent variable modelling. *Journal of Process Control*, 21(10):1370–1377.

Giancarlo, D. and Chiara, T. (2002). Cross-validation Methods in Principal Component Analysis: A Comparison. *Statistical Methods and Applications*, 11:71–82.

Grigg, N. P. (1998). Statistical process control in UK food production: an overview. *International Journal of Quality & Reliability Management*, 15(2):223–238.

ICH (2008). The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use: Pharmaceutical Quality System {Q10} ({ICH Q10}).

Imtiaz, S. A. and Shah, S. L. (2008). Treatment of missing values in process data analysis. *Canadian Journal of Chemical Engineering*, 86(5):838–858.

Imtiaz, S. a., Shoukat Choudhury, M. a. a., and Shah, S. L. (2007). Building multivariate models from compressed data. *Industrial and Engineering Chemistry Research*, 46(2):481–491.

Jackson, D. A. (1993). Stopping Rules in Principal Component Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, 74(8):2204–2214.

Jiang, C., Yan, J., Wang, Y., Zhang, J., Wang, G., Yang, J., and Hao, H. (2015). Isolation Strategies and Transformation Behaviors of Spironolactone Forms. *Industrial & Engineering Chemistry Research*, 54(44):11222–11229.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, second edition.

Jover, I., Podczeck, F., and Newton, M. (1996). Evaluation, by a statistically designed experiment, of an experimental grade of microcrystalline cellulose, Avicel 955, as a technology to aid the production of pellets with high drug loading. *Journal of Pharmaceutical Sciences*, 85(7):700–705.

Kadam, S. S., van der Windt, E., Daudey, P. J., and Kramer, H. J. M. (2010). A Comparative Study of ATR-FTIR and FT-NIR Spectroscopy for In-Situ Concentration Monitoring during Batch Cooling Crystallization Processes. *Crystal Growth & Design*, 10(6):2629–2640.

Kassidas, A., Taylor, P. a., and MacGregor, J. F. (1998). Off-line diagnosis of deterministic faults in continuous dynamic multivariable processes using speech recognition methods. *Journal of Process Control*, 8(5-6):381–393.

Kona, R., Qu, H., Mattes, R., Jancsik, B., Fahmy, R. M., and Hoag, S. W. (2013). Application of in-line near infrared spectroscopy and multivariate batch modeling for process monitoring in fluid bed granulation. *International Journal of Pharmaceutics*, 452(1-2):63–72.

Kosanovich, K. A., Piovoso, M. J., Dahl, K. S., MacGregor, J. F., and Nomikos, P. (1994). Multi-way PCA applied to an industrial batch process. In *American Control Conference, 1994*, volume 2, pages 1294–1298 vol.2.

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivaraite Statistical Process Control in Industry. *International Journal of Adaptive Control and Signal Processing*, 19:213–246.

Kourti, T., Nomikos, P., and MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control*, 5(4):277–284.

Ku, W., Storer, R. H., and Georgakis, C. (1995). Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 30:179–196.

Kulkarni, S. a., Kadam, S. S., Meekes, H., Stankiewicz, A. I., and Ter Horst, J. H. (2013). Crystal Nucleation Kinetics from Induction Times and Metastable Zone Widths. *Crystal Growth and Design*, 13(6):2435–2440.

Laurí, D. and Lennox, B. (2014). Expanded PLS algorithm: Modelling of batch processes. *Chemometrics and Intelligent Laboratory Systems*, 132:111–123.

Lennox, B., Montague, G. A., Hiden, H. G., Kornfeld, G., and Goulding, P. R. (2001). Process monitoring of an industrial fed-batch fermentation. *Biotechnology and Bioengineering*, 74(2):125–135.

Liebenberg, W. (2005). Crystal polymorphism and its occurance among active pharmaceutical ingredients in south africa. pages 1 – 27.

Liebenberg, W., Tonder, E. C. v., Dekker, T. G., and Villiers, M. M. d. (2003). Variable temperature X-ray powder diffractometry of spironolactone polymorphs. *Pharmazie*, 58(6):435–437.

Lindberg, N.-O. and Lundstedt, T. (1995). Multivariate data analysis of variables influencing the dissolution rate of prednimustine : a case of disconformity with the Noyes-Whitney equation. *European Journal of Pharmaceutics and Biopharmaceutics*, 41:101 – 113.

Lv, Z., Yan, X., and Jiang, Q. (2016). Batch process monitoring based on multiple-phase online sorting principal component analysis. *ISA Transactions*, 64:342–352.

MacGregor, J. F., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE Journal1*, 40(5):826 – 838.

Mah, R. S. H., Tamhane, A. C., Tung, S. H., and Patel, A. N. (1995). Process trending with piecewise linear smoothing. *Computers and Chemical Engineering*, 19(2):129–137.

Mahadevan, S. and Shah, S. L. (2009). Fault Detection and Diagnosis in Process Data using One-class Support Vector Machines. *Journal of Process Control*, 19:1627–1639.

Marini, A., Berbenni, V., Bruni, G., Maggioni, A., Orlandi, A., and Villa, M. (2001). Thermodynamics of a complex melting process : the case of spironolactone. *Thermochimica Acta*, 374(2):171–184.

Marjanovic, O., Lennox, B., Sandoz, D., and Lovett, D. (2004). Statistical Process Monitoring of Industrial Batch Processes. In *World Batch Forum European Conference*, pages 1 – 8.

Marjanovic, O., Lennox, B., Sandoz, D., Smith, K., and Crofts, M. (2006). Real-Time Monitoring of an Industrial Batch Process. *Computers and Chemical Engineering2*, 30:1476 – 1481.

Martin, E. B., Morris, A. J., Papazoglou, M. C., and Kiparissides, C. (1996a). Batch process monitoring for consistent production. *Computers & Chemical Engineering*, 20, Supple:S599 – S604.

Martin, E. B., Morris, A. J., and Zhang, J. (1996b). Process Performance Monitoring using Multivariate Statistical Process Control. *IEEE Proc. Control Theory Appl.*, 143(2):132–144.

Martini, S., Herrera, M. L., and Hartel, R. W. (2001). Effect of CoolingRate on Nucleation Behaviour of Milk Fat - Sunflower Oil Blends. *Journal of Agricultural and Food Chemistry*, 49:3223–3229.

Masding, P. and Lennox, B. (2010). Use of dynamic modelling and plant historian data for improved control design. *Control Engineering Practice*, 18(1):77–83.

McCrone, W. C. (1965). *Physics and Chemistry of the Organic Solid State*, volume 2. Interscience Publishers.

Medicines and Healthcare Products Regulatory Agency (2015). MHRA GMP Data Integrity Definitions and Guidance for Industry March 2015. Technical report, Medicines and Healthcare Products Regulatory Agency.

Miletic, I., Quinn, S., Dudzic, M., Vaculik, V., and Champagne, M. (2004). An industrial perspective on implementing on-line applications of multivariate statistics. *Journal of Process Control*, 14(8):821–836.

Myerson, A. S. (1993). *Handbook of Industrial Crystallization*. Butterworth-Heinemann Series in Chemical Engineering. Butterworth-Heinemann.

Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, 35(1):45–65.

Neogi, D. and Schlags, C. E. (1998). Multivariate Statistical Analysis of an Emulsion Batch Process. *Industrial & Engineering Chemistry Research*, 37(10):3971–3979.

Nesic, Z., Davies, M., and Dumont, G. (1996). Paper machine data compression using wavelets. *Proceeding of the 1996 IEEE International Conference on Control Applications IEEE International Conference on Control Applications held together with IEEE International Symposium on Intelligent Contro*, pages 161–166.

Nicolaï, B., Espeau, P., Céolin, R., Perrin, M.-A., Zaske, L., Giovannini, J., and Leveiller, F. (2007). Polymorph formation from solvate desolvation. *Journal of Thermal Analysis and Calorimetry*, 90(2):337–339.

Nomikos, P. (1996). Detection and Diagnosis of Abnormal Batch Operations Based on Multi-way Principal Component Analysis. *ISA Transactions1*, 35:259 – 266.

Nomikos, P. and MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8):1361–1375.

Orfanidis, S. (1996). *Introduction to Signal Processing*. Prentice Hall.

Patel, R. and Podczeck, F. (1996). Investigation of the effect of type and source of microcrystalline cellulose on capsule filling. *International Journal of Pharmaceutics*, 128(1):123–127.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Perry, R. H. (1997). *Perry's Chemical Engineers' Handbook*. McGraw-Hill, London, sixth edition.

Petersen, N., Stocks, S., and Gernaey, K. V. (2008). Multivariate models for prediction of rheological characteristics of filamentous fermentation broth from the size distribution. *Biotechnology and bioengineering*, 100(1):61–71.

Pinto, J. F., Podczeck, F., and Newton, J. M. (1997). Investigations of tablets, prepared from pellets produced by extrusion and spheronisation .1. The application of canonical analysis to correlate the properties of the tablets to the factors studied in combination with principal component analysis to selec. *International Journal of Pharmaceutics*, 147(1):79 – 93.

Pöllänen, K., Häkkinen, A., Reinikainen, S.-P., Louhi-Kultanen, M., and Nyström, L. (2006a). A Study on Batch Cooling Crystallization of Sulphathiazole Process Monitoring Using ATR-FTIR and Priduct Characterization by Automated Image Analysis. *Chemical Engineering Research and Design*, 84(A1):47 – 59.

Pöllänen, K., Häkkinen, A., Reinikainen, S.-P., Rantanen, J., and Minkkinen, P. (2006b). Dynamic PCA-based MSPC charts for nucleation prediction in batch cooling crystallization processes. *Chemometrics and Intelligent Laboratory Systems*, 84(1–2):126–133.

Proakis, J. G. and Manolakis, D. G. (2007). *Digital Signal Processing. Principles, Algorithms, and Applications*. Pearson Prentice Hall, fourth edi edition.

Rani, S., Kaur, A., and Ubhi, J. S. (2011). Comparative study of FIR and IIR filters for the removal of Baseline noises from ECG signal. *International Journal of Computer Science and Information Technologies*, 2(3):1105–1108.

Rinnan, A., van den Berg, F., and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222.

Rotthäuser, B., Kraus, G., and Schmidt, P. C. (1998). Optimization of an effervescent tablet formulation using a central composite design optimization of an effervescent tablet formulation containing spray dried l-leucine and polyethylene glycol 6000 as lubricants using a central composite design. *European Journal of Pharmaceutics and Biopharmaceutics*, 46(1):85–94.

Rubin, D. B. (1977). Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72(359):538–543.

Salole, E. G. and Al-Sarraj, F. A. (1985). Effect of Solvent-Deposition on Spironolactone Crystal Form. *Drug Development and Industrial Pharmacy*, 11(12):2061 – 2070.

Sarraguça, M. C., Beer, T. D., Vervaet, C., Remon, J.-P., and Lopes, J. A. (2010). A batch modelling approach to monitor a freeze-drying process using in-line Raman spectroscopy. *Talanta*, 83(1):130–138.

Sarraguca, M. C., Ribeiro, P. R., Dos Santos, A. O., and Lopes, J. A. (2015). Batch Statistical Process Monitoring Approach to a Cocrystallization Process. *Journal of Pharmaceutical Sciences*, 104(12):4099–4108.

Siesler, H. W., Ozaki, Y., Kawata, S., and HeiseH.M., editors (2002). *Near-Infrared Spectroscopy: Principles, Instruments, Applications*. Wiley-VCH, first edit edition.

Simoglou, A., Georgieva, P., Martin, E. B., Morris, A. J., and de Azevedo, S. (2005). On-line Monitoring of a Sugar Crystallization Process. *Computers and Chemical Engineering*, 29:1411–1422.

Singhal, A. and Seborg, D. (2003). Data compression issues with pattern matching in historical data. *Proceedings of the 2003 American Control Conference, 2003.*, 5:3696–3701.

Sivalingam, S. and Hovd, M. (2011). Effect of data compression on controller performance monitoring. *2011 19th Mediterranean Conference on Control & Automation (MED)*, pages 594–599.

Škulj, G., Vrabič, R., Butala, P., and Sluga, A. (2013). Statistical Process Control as a Service: An Industrial Case Study. *Procedia CIRP*, 7:401–406.

Socrates, G. (2004). *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. Wiley, third edit edition.

Soliman, O. A. E., Kimura, F., Hirayama, K., El-Sabbagh, H. M., El-Gawad, A. E.-G. H. A., and Hashim, F. M. (1997). Amorphous Spironolactone-Hydroxypropylated Cyclodextrin Complexeswith Superior Dissolution and Oral Bioavailability. *International Journal of Pharmaceutics*, 149(1):73–83.

Somberg, J. C. and Ranade, V. V. (2009). United States Patent: Synthesis and Separation of Optically Active Isomers and Cyclopropyl Derivatives of Spironolactone and Their Biological Action.

Souihi, N., Lindegren, A., Eriksson, L., and Trygg, J. (2015). OPLS in batch monitoring - Opens up new opportunities. *Analytica Chimica Acta*, 857:28–38.

Sun, D.-W. (2009). *Infrared Spectroscopy for Food Quality Analysis and Control*. Academic Press.

Sun, F., Xu, B., Zhang, Y., Dai, S., Shi, X., and Qiao, Y. (2017). Latent variable modeling to analyze the effects of process parameters on the dissolution of paracetamol tablet. *Bioengineered*, 8(1):61–70.

Thornhill, N. F., Shoukat Choudhury, M. A. A., and Shah, S. L. (2004). The impact of compression on data-driven process analyses. *Journal of Process Control*, 14(4):389–398.

Valle, S., Li, W., and Qin, S. J. (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Industrial and Engineering Chemistry Research*, 38(11):4389–4401.

Walczak, B. and Massart, D. L. (2001). Dealing with missing data: Part II. *Chemometrics and Intelligent Laboratory Systems*, 58(1):29–42.

Wan, J., Marjanovic, O., and Lennox, B. (2014). Uneven batch data alignment with application to the control of batch end-product quality. *ISA Transactions*, 53(2):584–590.

Wang, Y., Sun, F., and Jia, M. (2016). Online monitoring method for multiple operating batch processes based on local collection standardization and multi-model dynamic PCA. *The Canadian Journal of Chemical Engineering*, 94(10):1965–1976.

Wang, Y. J., Jia, M. X., and Mao, Z. Z. (2015). A fast monitoring method for multiple operating batch processes with incomplete modeling data types. *Journal of Industrial and Engineering Chemistry*, 21:328–337.

Watson, M. J., Liakopoulos, A., Brzakovic, D., and Georgakis, C. (1998). A Practical Assessment of Process Data Compression Techniques. *Industrial & Engineering Chemistry Research*, 37(1):267–274.

Weier, R. M. (1978). United States Patent: 7-(alkoxylcarbonyl)-6-alkyl/halo-17-hydroxyl-3-oxo-17-pregn-4-ene-21-carboxylic acid lactones and corresponding 21-carboxylic acids, their salts, and esters.

Westad, F., Gidskehaug, L., Swarbrick, B., and Flåten, G. R. (2015). Assumption free modeling and monitoring of batch processes. *Chemometrics and Intelligent Laboratory Systems*, 149:66–72.

WIPO (2012). WIPO - Search International and National Patent Collections.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52.

Wong, C. W. L., Escott, R., Martin, E. B., and Morris, J. (2008). The Integration of Spectroscopic and Process Data for Enhanced Process Performance Monitoring. *The Canadian Journal of Chemical Engineering*, 86:905 – 923.

Workman, J. and Weyer, L. (2008). *Practical Guide to Interpretive Near-Infrared Spectroscopy*. CRC Press, Boca Raton, Florida.

Xiao, D., Jiang, J., Mao, Y., and Lui, X. (2016). Process Monitoring and Fault Diagnosis for Piercing Production of Seamless Tube. *Mathematical Problems in Engineering*, 2016:1 – 13.

Yu, L. X., Lionberger, R. A., Raw, A. S., D'Costa, R., Wu, H., and Hussain, A. S. (2004). Applications of process analytical technology to crystallization processes. *Advanced Drug Delivery Reviews*, 56(3):349–369.

Zhu, Z., Similä, T., and Corona, F. (2013). Supervised Distance Preserving Projections. *Neural Process. Lett.*, 38(3):445–463.