FINGERPRINTING

OF

COMPLEX BIOPROCESS DATA

by

Nor Fadhillah Mohamed Azmin

A Thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Chemical Engineering and Advanced Materials

Newcastle University

January 2013

Abstract

The focus of the research is on the analysis of complex bioprocess datasets with the ultimate goal of forming a link between the data and its underlying biological patterns. The challenges associated with investigating complex bioprocess data include the high dimensionality of the underlying measurements, the limited number of “observations”, and the complexity of selecting meaningful features to characterise the data. Contained within these data is a wealth of information that can contribute to inferring process outcomes and providing insight into improving productivity and process efficiency. To address these challenges, there is a real need for techniques to analyse and extract knowledge from the data. This thesis investigates an integrated discrete wavelet transform (DWT) and multiway principal components analysis (MPCA) approach to extract meaningful information from different types of bioprocess data.

The integrated methodology is demonstrated by application to two types of bioprocess data: a near infrared (NIR) dataset collected from an industrial monoclonal antibodies (MAb) process, and an electrospray ionisation mass spectrometry (ESI-MS) dataset generated during the development of recombinant mammalian cell lines. The objective of the thesis was to develop a methodology that enabled the extraction of information from these two data sets. For the industrial NIR dataset, the genealogy or parent-child relationship of batch process from monoclonal antibodies (MAb) manufacturing was investigated whilst for the ESI-MS dataset goal was to identify characteristics that would enable the differentiation between high and low cell producers.

The main challenges of the NIR and ESI-MS data sets lay in the complexity of the spectra. The NIR spectra usually have broad overlapping peaks and baseline shifts. Furthermore, as the NIR spectra used in this thesis were collected from batch process, there is an extra dimension in the data that of batch. On the one hand, the extra dimension provides extra information but on the other, it presents a further challenge as the data now is three-dimensional and requires additional pre-processing, including data matrix unfolding and batch alignment. Similar to the NIR spectra, the ESI-MS dataset also faces the problem of baseline shifts along with other complexities including high noise to signal ratio, shifts in the mass-to-charge (m/z) ratio, and differences in signal intensities. These challenges lead to difficulties in extracting relevant information about the feature of interest. The proposed methodology was proven effective in extracting meaningful information from both data sets.

In summary, the proposed method which utilised the integration of discrete wavelet transform and multiway principal component analysis was able to differentiate the distinguished characteristics of the spectra in the datasets thereby providing understanding of the relationships between spectral data and the underlying behaviour of the process.

Acknowledgement

I would like to express my sincere thanks to my supervisor Professor Elaine Martin for her continuous support, encouragement, trust and above all her patience and Professor Gary Montague for his guidance in the early part of my PhD. I am also grateful to Dr. Ronan O'Kennedy for his great insight into bioprocess engineering and Dr. Christopher O'Malley who has helped considerably in terms of utilisation of data analysis tools. I would also like to thank my group members in SYSMA (System Modelling and Analysis).

I would like to acknowledge the funding from the International Islamic University Malaysia and Ministry of High Education Malaysia.

My greatest thanks is for my family, especially my husband, Amran, my daughter, Auni, and my grandmother Nor Akhmar, who have been so understanding and supportive during this journey.

Contents

Appendices

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Objective of the Thesis

The pharmaceutical industry develops and manufactures therapeutics that aid improvements in the health and quality of life of people. Therapeutics can be categorised into two major classes: small molecule and large molecule compounds. The first class, small molecule compounds, are synthesised from chemical reactions between different organic and/or inorganic compounds. The second class, large molecule therapeutics, also known as biopharmaceuticals, are derived from recombinant protein technology or new biotechnology. Small molecule drugs comprise a limited number of atoms and can generally be manufactured with high consistency (Rader, 2008) whereas the manufacture of biopharmaceuticals produces heterogeneous mixtures which translate into variability in process batches (Steinmeyer and McCormick, 2008). Alongside small molecules, biopharmaceuticals have an important role in the treatment of diseases.

Today's pharmaceutical industry has observed a shift in the direction of the market, moving towards innovation and the manufacture of biopharmaceutical therapeutics. As reported by Goodman (2009), constant or decreasing amounts in revenue portfolios of most pharmaceutical companies is due to the expiry of older blockbuster products patents that are mainly small molecules and the reassignment of resources towards research and development of biopharmaceutical therapeutics. The trend in patent filing also shows an increasing gap between patent filing for biopharmaceutical and small molecules drugs from 2007 to 2009. In 2009, 60% of the patents filed were for biopharmaceutical drug products (Philippidis, 2012). Key reasons for the increasing interest in biopharmaceutical therapeutics development lies in both their scientific and economic potential. On the scientific front, their high selectivity and specificity mean a lower probability of non-mechanism-based toxicity (Steinmeyer and McCormick, 2008; Nicolaides et al., 2006). On the economic front, biopharmaceutical therapeutics present high potential for robust sales even after patent expirary (Projan et al., 2004). This is due to biologic products being quite rare leading to minimum pricing competitions including biologic generics.

One of the fastest growing biopharmaceutical therapeutics in the current market is the monoclonal antibody whose global sales were reported at USD44.6 billion in 2011 and projected to rise to USD56 billion by 2016 (BCC Research, 2012). The first monoclonal antibody, murine monoclonal antibody, or muramonab, was introduced into clinical development in 1987 (Buss et al., 2012). Since then, antibody-based therapeutics have rapidly expanded their market share and become the leading blockbuster biopharmaceutical therapeutics during the last three decades (Ryu et al., 2012). Expressions of all commercial therapeutics monoclonal antibodies were derived from mammalian cells with the majority being from Chinese Hamster Ovary (CHO), mouse myeloma (NS0) and Sp2/0, a mouse myeloma often used in hybridoma fusion with CHO being the main choice (Kelley, 2009). While biopharmaceutical therapeutics promise favourable profit returns, their manufacture involves high production costs (Johnson-Leger et al., 2006) which is partly due to a longer development timeline i.e. time from product development to marketed products (Kozlowski and Swann, 2006), therefore it is imperative to speed up the development and the achievement of right-first-time production (Jungbauer and Gobel, 2011). Along with the high production costs, the manufacture of the monoclonal antibody presents unique challenges including; the development of cell lines that are capable of producing high yields to meet the market demands at a reasonable cost-of-goods (Nicolaides et al., 2006), the screening and selecting of highly productive and stable cell lines (Luo and Chen, 2007), and the manufacture of consistently high quality product (FDA, 2004). Improvements throughout process development and manufacture are essential in order to overcome these hurdles.

In an attempt to address these challenges, input is required from a number of areas including cell engineering, product development, data mining, and process modelling. A paper by Kelley (2009) suggested that the process development groups of the monoclonal antibody industry should switch their objectives from invention and innovation of new development technologies, to focusing on understanding the process fundamentals of their current platforms. This view is in line with the Pharmaceutical Quality for the 21st century initiative introduced by the U.S. Food and Drug Administration (FDA). The initiative includes:

(1) Process Analytical Technology (PAT) (FDA, 2004),
(2) the International Conference on Harmonization (ICH) guidance Q9: Quality Risk Management (FDA, 2006),
(3) the International Conference on Harmonization (ICH) guidance Q10: Pharmaceutical Quality System (FDA, 2009b), and
(4) the International Conference on Harmonization (ICH) guidance Q8(R2): Pharmaceutical Development (FDA, 2009a).

These initiatives describe the idea of the design space that connects product and process knowledge. FDA (2009a) defines a design space as “the multidimensional combination and interaction of input variables and process parameters that have been demonstrated to provide assurance of quality”. Furthermore, they highlight the utilisation of mathematical models as one of the current gaps and challenges in pharmaceutical process development and manufacturing. Building a bridge between product and process knowledge necessitates the implementation of analytical tools including process modeling and simulation which utilize data generated during development and production to help enhance process understanding. Rathore et al. (2010) emphasized the application of chemometric methods to ensure effective analysis of data from complex systems such as biopharmaceutical products. Data collected throughout process development which includes process and spectral data hold a wealth of information which provides insight which may improve the steps in process development (Bhushan et al., 2011; Rathore et al., 2011). Such information will also be invaluable to the progress of a bioprocess in industrial-scale operations, for example in process scale-up, process comparability and technical transfer between production sites.

A growing body of literature has investigated spectral data generated during bioprocess process development; for example, Damen et al. (2009), Pons et al. (2004), and Wan et al. (2001). The common objectives of these papers were the investigation of the spectral data for bioprocess monitoring and fault detection, and fingerprinting of the protein to serve as biomarkers for disease identification. While these objectives are important in their own right, there is a lack of contributions to process understanding to aid process development. This thesis focuses on the investigation of the underlying behaviour of the process through spectral data. It is hypothesised that this approach will contribute to knowledge enhancement in terms of process understanding.

The investigation of two different types of complex spectral data from the manufacture of monoclonal antibodies underpins this thesis: the first type is near infra-red spectra and the second is electrospray ionisation spectra. Spectral data of these types are typically complex, and of high dimensionality, and hence their analyses is a challenge. In this thesis the aim is to develop fingerprints of the bioprocess spectral data through the application of a combined discrete wavelet transform and the multivariate statistical technique of principal component analysis. Fingerprinting of the spectral data in the context of this thesis is of importance as it potentially provides a link between spectral data and the underlying pattern of the biological behaviour of the monoclonal antibody. This will not only add value to process understanding but will also establish a set of benchmarks for subsequent processes in product development. Information gained from the fingerprint can be used as a benchmark for selecting the appropriate batch to subcultured and as valuable information to rank cell line producer, in addition to information generated from quality parameters (e.g.: titre and product quality).

In the following sections, the objectives, contributions and contents of each chapter of the Thesis are described.

1.2 Contributions of the Thesis

The primary contribution of the thesis is in the field of the development of fingerprinting models for complex bioprocess data. More specifically the key contributions include:

(1) The development of a methodology for selecting the type of mother wavelet and wavelet decomposition level required for the analysis. Key decisions from this investigation form the basis of the subsequent model development.

(2) One of the challenges when analysing bioprocess spectral data is the complexity and multidimensionality of the dataset, with meaningful information masked by overlapping bands and/or high noise to signal ratio. The discrete wavelet transform technique, which was adopted as one of the stages in the model development, has an interesting characteristic, multiresolution analysis. In multiresolution analysis, the spectral data is decomposed into different levels of resolution; these levels are known as wavelet sub-bands. The wavelet sub-bands were used in two contrasting ways in this thesis. In the first case study, the concept of feature extraction was used to select meaningful wavelet coefficients from the wavelet sub-bands to represent the spectra in the subsequent analysis. Meanwhile, in the second case study, all wavelet coefficients in the wavelet sub-bands were retained and each wavelet sub-band was analysed individually.

(3) An application of the wavelet denoising algorithm to near infra-red spectra was investigated. It was demonstrated that denoising is not necessary for near infra-red spectra because of the low noise to signal ratio in the near infra-red spectra. It was also shown that denoising flattened the peaks of the spectra.

(4) A demonstration that distinct characteristics in the behaviour of the batches can be explained by the underlying genealogy. One of the case studies presented was a batch process in which batches from the process were organised into their genealogy or parent-child relationships. Strong similarities and differences between batches within and between families were identified.

(5) The development of a pre-processing technique for near infra-red batch data is proposed. It involves unfolding methods and the alignment of the three-dimensional near infra-red spectral batch data. The commonly applied unfolding approach for batch data has focused on process data whereas in this thesis the focus is on spectral data.

(6) The development of an integrated approach comprising the discrete wavelet transform technique and principal component analysis for application as a fingerprinting framework. The methodology is investigated through its application to near infra-red spectral data and electrospray ionisation data from the industrial and laboratory scale manufacture of monoclonal antibodies. It is hypothesised that the results obtained provide a meaningful link between the spectral data and their genealogy in the former and biological behaviour in the latter, hence leading to enhanced process understanding.

(7) The development of contribution plots of principal component scores of the wavelet sub-bands enables the characterisation of cell lines and replicates. Analyses of the contribution plots resulted in the establishment of the

characteristics that differentiated between high and low producer cell lines which potentially aid to better process development.

1.3 Thesis Outline

The following provides a summary of each of the chapters in the thesis.

Chapter 1 provides an introduction to the motivation and the objectives for undertaking research into the fingerprinting of complex spectral data generated from bioprocess. The contributions of the thesis are also identified and briefly discussed. Following this, an outline of the thesis is given. The thesis is organised into six chapters with Chapter 2 and Chapter 3 providing an overview of the techniques forming the basis of the fingerprinting framework, wavelet theory and principal component analysis. Investigations into the application of the techniques to fingerprint near infra-red and electrospray ionisation spectral data sets are discussed in Chapter 4 and Chapter 5 respectively. Following this, a summary of the key results and proposal for future work are presented in Chapter 6.

An introduction to wavelet transform theory is provided in Chapter 2. Initially the rationale for adopting the wavelet technique in this thesis is provided. Then a brief review of the Fourier transform and the short-time Fourier transform is given, prior to introducing the wavelet transform. The concept of the wavelet is placed within the general framework of time-frequency analysis. Following this, multiresolution analysis, which is one of the pivotal theoretical bases of the wavelet transform, is explained. Finally, an investigation of the application of wavelet denoising algorithm on near infra-red spectra is presented.

Chapter 3 provides an introduction to the multivariate statistical technique of principal component analysis (PCA), and its extension to batch processes multiway PCA. An overview of the algorithm is given, along with a discussion on the associated metrics necessary for the development of a fingerprinting framework. Particular attention is given to PCA since it is this technique which underpins the thesis.

The aim of the research, presented in Chapter 4, is to investigate the impact of genealogy on batch behaviour by developing a fingerprinting model from near infra-red spectral dataset generated from the industrial manufacture of monoclonal antibodies. Batch processes form a significant part of monoclonal antibodies production and generally exhibit batch-to-batch variation. There is thus a need to understand the connection between batch genealogy and their behaviour. Initially, a description of the process and the genealogy of the batches which represent the family structure or parent-child relationship of the batches is given. A brief background to near infra-red spectroscopy is presented followed by a brief discussion on a number of pre-processing techniques. Forming the basis of the fingerprinting framework is the integration of the discrete wavelet transform and principal component analysis. The motivation for this integrated approach and a number of applications reported in the literature are discussed prior to describing the application of the framework to the near infra-red batch spectral data set. The development strategy of the fingerprint representation, which includes a route for selecting the type of mother wavelet and wavelet decomposition levels required for the analysis, wavenumber selection, multiresolution analysis of the wavelet transform and feature extraction are also discussed.

Chapter 5 describes a case study to investigate the fingerprinting of electrospray ionisation spectra data from a different monoclonal antibodies process. One of the major challenges in the development process of monoclonal antibodies is to screen and select highly productive and stable cell line producers. There are two main objectives in this chapter. The first is to describe and establish the characteristics that differentiate between high and low cell line producers. Secondly it is to investigate the transferability of the integrated wavelet-principal component analysis framework to analyse a complex dataset of a different structure. A brief introduction to electrospray ionisation spectra is presented prior to introducing the process. A specific data pre-processing technique with regard to the electrospray ionisation spectral data is also described. More specifically in this chapter, a different perspective to the application of contribution plots of principal component scores is proposed. The standard use of contribution plots has been in process monitoring and fault diagnosis whilst this case study utilises the contribution plots to characterise cell line producers.

Finally, in Chapter 6, a summary of the key results are given with proposals for future work.

Chapter 2 Review of Discrete Wavelet Transform

2.1 Introduction

Over the past two decades, the development and application of the wavelet transform has attracted interest from both mathematicians and engineers. In the early stage of the wavelet transform development, much of the literature focused on its mathematical aspect. As the application of wavelet transform becomes widespread more literature relating to both the theory and application are published. More specifically the wavelet technique has been successfully employed as one of the signal processing tools in various fields of analytical chemistry including infrared spectrometry and mass spectrometry (Jetter et al., 2000). The focus of this chapter is to provide an introduction to the fundamental theory of the wavelet transform. Furthermore this chapter will define the basic terms and provide the theoretical background to the application of the wavelet transform in Chapter 4 and 5.

In this chapter, firstly the rationale for implementing the wavelet transform is given. Then the history of wavelet theory, which is inherently linked to the Fourier Transform and Windowed or Short-Time Fourier Transform, is provided. Therefore prior to introducing the wavelet transform, the Fourier transform and short-time Fourier transform are defined. Following this, an overview of the wavelet transform along with examples of wavelet families is described. Finally the concept of multiresolution, which is the key theoretical basis of the wavelet transform, is then discussed.

Numerous books have been published on the subject of wavelets including the book by Mallat (1998) which provides an introduction to the theory of wavelets. Other early technical references include Strang and Nguyen (1997), Meyer and Ryan (1993), Daubechies (Daubechies, 1992), and Meyer and Salinger (Meyer and Salinger, 1992).

Hubbard (Hubbard, 1996) provides an excellent introduction to wavelets from a less mathematical perspective. More recent books on wavelets focus on its application in specific fields, including Chau et al. (2004) in chemometrics, Gopalakrishnan (2010) on dynamic problems on structures including metallic, composite, and nano-composite, Sarkar et al. (2002) on engineering electromagnetics, and Tang et al. (2000) on pattern recognition. The rationale for the application of wavelet transform in this thesis is discussed in the next section, Section 2.2.

2.2 Rationale

The rationale behind implementing the discrete wavelet transform is based on the underlying advantages of the methodology. One advantage is its time-frequency localisation property which differs from the traditional Fourier techniques. This property enables the analysis of a signal in time and frequency domain simultaneously. As a result the wavelet transform can ‘zoom-in’ and ‘zoom-out’ of any part of the signal to be analysed. Furthermore, the application of wavelet has been proven to be more efficient and faster as compared to the Fourier transform in capturing significant information of a dataset (Lio, 2003).

Previous studies have shown that application of the wavelet is an efficient and practical way to extract information including identification of breakdown points trends, and self-similarity in signals from dataset by utilising this property (Li et al., 2011; Borah et al., 2007; Carreno and Vuskovic, 2007). In the study undertaken by Li et al. (2011), it has been proven that due to this property, the discrete wavelets transform is the appropriate technique to detect cutting tool failure in automatic machining processing. A system to recognize the cutting tool wear states was developed utilising the wavelet time-frequency information. Carreno and Vuskovic (2007) applied the discrete wavelet transform to extract important features from electromyographic signals (EMG) which consist of four grasping motions. The extracted features were then categorised for the purpose of controlling a prosthetic device. It was shown that the application of the discrete wavelet transform has significantly reduced the dimensionality of the extracted feature vectors whilst retaining the important information in the signals. Moreover, in the study of tea sorting process through image textures of tea granules, Borah et al. (2007) demonstrate that the first and second wavelet resolution are more sensitive to the image texture which enables the accurate classification of the tea images.

In view of the application of the wavelet transform in chemometrics, its application has been demonstrated in various chemometrics fields including signal denoising and compression, classification and multivariate calibration over the past two decades. For example, in data denoising and compression (Barclay et al., 1997), and classification and multivariate calibration (Pinto et al., 2011; Jouan-Rimbaud et al., 1997). Data denoising and compression work together by removing noise from the data by applying thresholding and simultaneously compressing the data. This results in a smoother and denoised ('cleaner') data, and reduction in dimensionality of the dataset in study reported by Barclay et al. (1997). Meanwhile, Pinto et al. (2011) and Jouan-Rimbaud et al. (1997) proved that classification and multivariate calibration models of diverse chemical datasets are improved through the application of the wavelet transform. Evidently, the application of the wavelet transform as a signal processing has proven to be advantageous hence the decision to utilise it in this thesis. The following section provides a historical review of the wavelet transform.

2.3 Historical Review of Wavelet

The work of Jean Morlet who developed wavelets as an oil prospecting tool in 1980 is considered the 'formal' root of the wavelet theory; as there were at least 15 distinct roots identified by Meyer (Hubbard, 1996). Morlet then collaborated with Alex Grossman to mathematically validate the empirical results of using wavelets to represent a signal by proving that the average value of the square of the signal is unaffected. This is a crucial condition as it means that a signal can be transformed into wavelets and then reconstructed to its original form.

Then in 1986, a work by Stephane Mallat justifies that many studies which appeared under different names including wavelets, the pyramid algorithms in image processing, the sub-band coding of signal processing, the quadrature mirror filters of digital speech processing were fundamentally the same. Furthermore, he demonstrated the use of a new function, the scaling function, to speed up the computation of wavelet coefficients. Also he describes a systematic way to construct new orthogonal wavelet basis functions which used truncated versions of infinite wavelets. The work of Mallat instigated Ingrid Daubechies to construct a set of orthogonal wavelet basis functions with compact support that can prevent errors due to the truncation. This new wavelet basis functions which is orthogonal and has compact support has become the foundation of wavelet application (Daubechies, 1992). The next section gives a definition of the Fourier transform and the short time Fourier transform which are part of the building blocks of the wavelet transform.

2.4 The Fourier Transform and Short Time Fourier Transform

A signal is defined as a mathematical function that transmits information which arises from either natural phenomenon or synthesized designs (Alterovitz, 2007) and is usually evolved from the field of time series analysis in modern signal processing. It is analysed in either a time-domain or a frequency domain by the application of various mathematical tools, including the Fourier transform and short time Fourier transform which are discussed in Section 2.4.1 and Section 2.4.2 respectively.

2.4.1 *The Fourier Transform*

The Fourier transform can be used to transform a signal from the time-domain to the frequency-domain. Transformation of a signal from one domain to the other is of importance as it leads to better understanding and additional insight of the signal under analysis.

The relationship between the time domain, *f*(t), and the frequency domain, *F*(ω), is established by the Fourier transform and inverse Fourier transform. The Fourier transform decomposes a signal into oscillatory functions and is define as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt, \qquad 2.1$$

and its inverse is given by

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}d\omega, \qquad 2.2$$

where $j =$ represents imaginary

$\omega =$ frequency in radians per seconds

$t =$ time in unit seconds

Both the Fourier transform and its inverse functions can be written in terms of sine and cosine functions by replacing the exponential term with Euler's equation:

$$e^{jk\theta} = \cos(k\theta) + j\sin(k\theta), \qquad 2.3$$

The revised FT and IFT are as follows:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)[\cos \omega t - j \sin \omega t] \, dt \, , \qquad 2.4$$

and,

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)[\cos \omega t + j \; sin\omega t] \, d\omega. \qquad 2.5$$

Equation 2.4 and 2.5 demonstrates that the transform uses sines and cosines functions as bases for the transformation between the time and frequency domains. The function *f*(t) in Equation 2.5 exhibits the time information but masks the information about frequency. Meanwhile the Fourier transform *F*(ω) in Equation 2.4 expresses information in terms of frequency of which the time information is contained in the phases of the displacement of the sines and cosines for each frequency. The phase displacement occurs when the sines and cosines are either combined amplifying the signals; or subtracted from each other thereby cancelling the signal. Consequently, a signal can only be studied either in the time or frequency domain through the transformation by the Fourier transform. This means that the application of the Fourier transform is appropriate for the analysis of stationary signals, a signal whose frequency does not change over time. The following example shows a non-stationary signal with two different frequencies (80Hz and 160Hz) which are buried in noise. Figure 2.1 shows that the frequency components of this signal are difficult to identify when the signal is analysed in the time-domain. However, performing the Fourier transform to the signal enables the identification of the frequency components (Figure 2.2) even though they were buried in the noise.

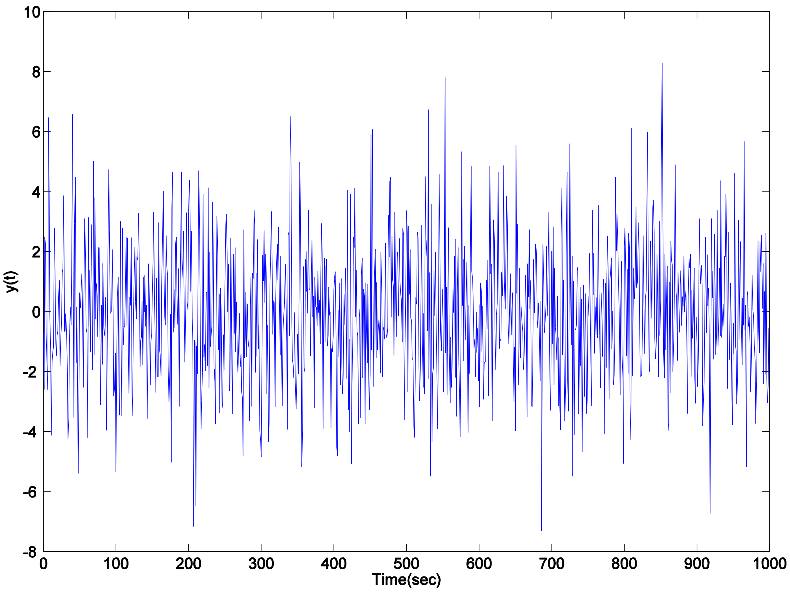


Figure 2.1 Example signal with two different frequencies buried in noise

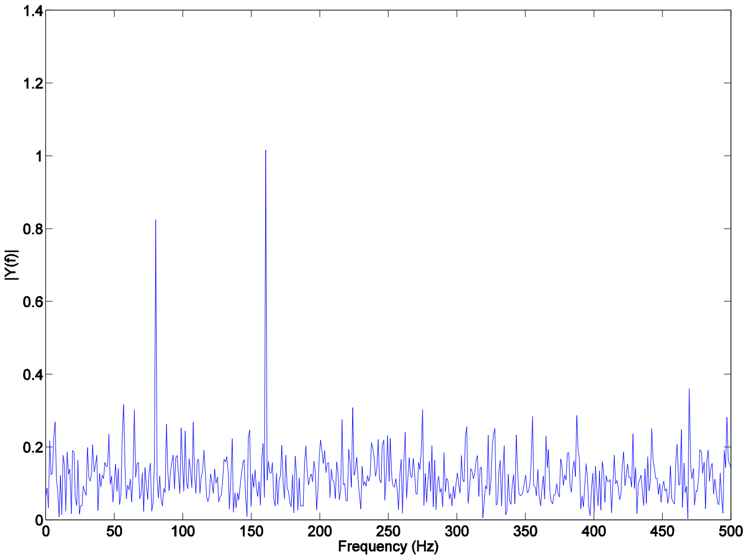


Figure 2.2 Fourier transform of the signal enables the extraction of the frequency components

2.4.2 *The Short Time Fourier Transform*

The inability of the Fourier transform to display both time and frequency simultaneously has become a limitation to its ability to analyse non-stationary signals. To address this limitation, a short time Fourier transform was introduced by Gabor in 1946 (Hubbard, 1996). In the short time Fourier transform, the sinusoidal wave in the Fourier transform is replaced by the product of a sinusoid and a time localisation window; i.e. replacing *f*(t) in Equation 2.1 with *f*(t)*g*(t - τ). Hence the short-time Fourier transform is defined as

$$F(\tau,\omega) = \int_{-\infty}^{\infty} f(t)g(t-\tau)e^{-j\omega t}dt. \qquad 2.6$$

The function *g*(t - τ) in Equation 2.6 is the time localisation window which allows the short-time Fourier transform to split the signal into blocks of equal length. The window, *g*(t) which is fixed in size then passes between analysing individual block i.e. the short-time Fourier transform of each block is calculated as shown in Figure 2.3. Hence the calculated short time Fourier transform is a description of the evolution of the signal frequency over time. With the window function, the short-time Fourier transform seems an ideal technique to examine a signal in both the time and frequency domain.

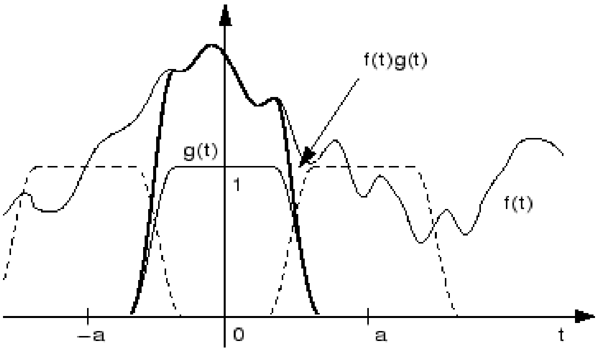


Figure 2.3 The short Fourier transform

However, since the same window size is used for all frequencies, the choice of window size is critical for the analysis of non-stationary signal. Figure 2.4 shows two sinusoids signal with two different frequencies, 100 Hz and 125 Hz. As shown in Figure 2.5 the short time Fourier transform of the signal, when the window size is too narrow, the transformed signal lacks resolution to display the frequency content. In contrast to Figure 2.6 of which the window size is appropriate, the two sinusoids with different frequencies can be identified. If the window size in increase and becomes too wide, the changes in the frequency content over time will become blurred.

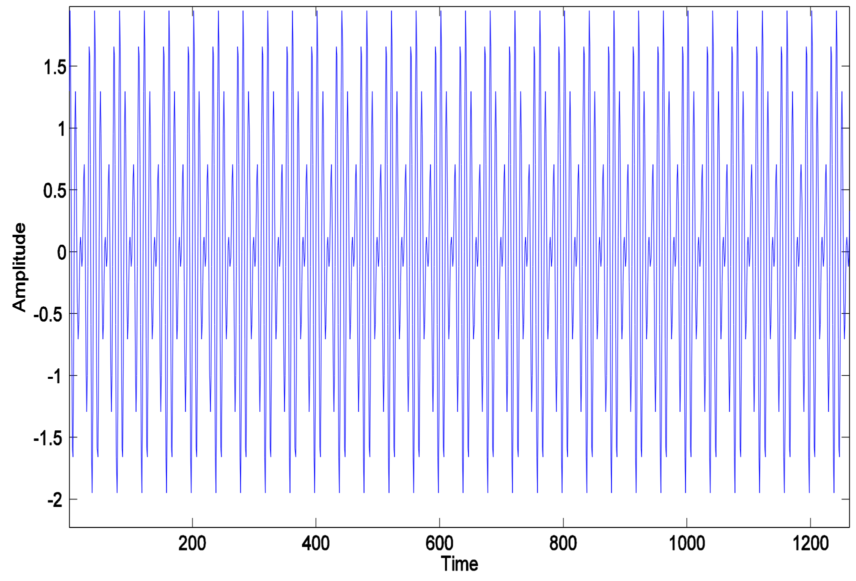


Figure 2.4 Example signal which comprises two different frequencies sinusoids.

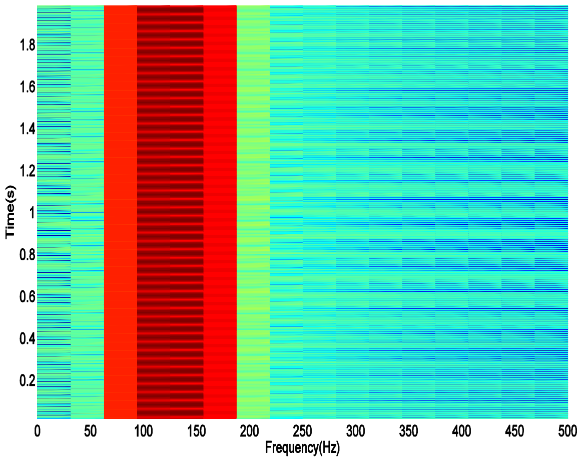


Figure 2.5 The short time Fourier transform of the example signal in Figure 2.4 using a 32-sample window

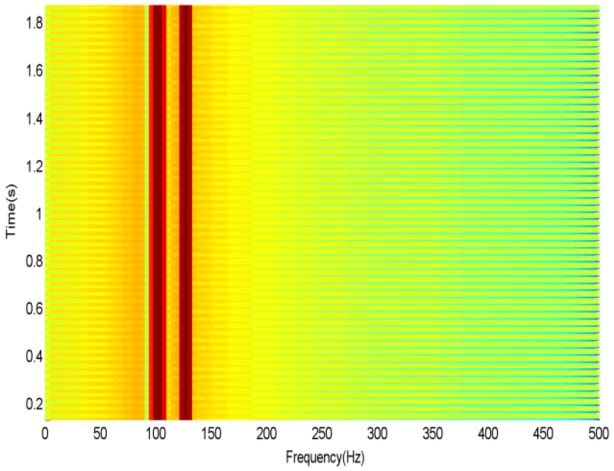


Figure 2.6 The short time Fourier transform of the example signal in Figure 2.4 using a 256-sample window

This example shows that the fixed resolution property of the short-time Fourier transform in which, the same window size is used for all frequencies limit its flexibility as there is a trade-off between time and frequency resolution. A narrow window which gives better time resolution is desirable for the analysis of the low frequency components of a signal. On the other hand, a wider window is preferable for the analysis of the high frequency component as it means better frequency resolution. Hence, provided that an appropriate window size is selected and the frequency of the signal is stationary within that window, the short time Fourier transform is applicable for the analysis of a signal in both the time and frequency domain. However, the analysis of many signals requires a signal processing tools with varying size of windows thereby instigated the work of Jean Morlet on the wavelet transform. The following section described the wavelet transform in the time-frequency domain framework with varying analysing window.

2.5 The Wavelet Transform

Wavelets means ‘little waves’ which are used to analyse signals or data. The wavelet transform is an extension to Fourier analysis. In the Fourier transform, the signal is decomposed into sine and cosine waves of various frequencies whilst in wavelet analysis, the signal is decomposed into a set of wavelets which later can be used to reconstruct the original signal without loss of information.

The main characteristic of the wavelet transform is its ability to process signals at different resolutions and hence it has been termed a ‘mathematical microscope’, due to its ability to assess both local and global features of signals by adjusting its “focus” (Hubbard, 1996). This gives the wavelet transform an advantage over the Fourier transform and the short time Fourier transform in the signal processing field.

The wavelet transform represents a signal as a set of basis functions i.e. wavelets which are generated by dilation and translation of a single function called the ‘mother wavelets’ $\psi(t)$,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|\alpha|}}\psi\left(\frac{t-b}{a}\right) \qquad 2.7$$

where a is the dilation parameter and b is the translation parameter. The dilation parameter a stretches or compresses the mother wavelet to generate wavelets that are used to captures different frequency components. The translation parameter b is the shifting parameter that determines the position of the wavelet to capture the time information of the signal under analysis. As the goal in the wavelet transform is to analyse a signal by measuring the similarity between the signal of interest and the analysing wavelet $\psi_{a,b}$, a and b are varied to achieve that objective. When $|a|$>>1 the mother wavelet is stretched (Figure 2.7(b)) and used to analyse the high frequency components whilst if $|a|$<< 1, the mother wavelet is compressed (Figure 2.7(c)) and used to study the low frequency components. The translated wavelet is shown in Figure 2.7(d).

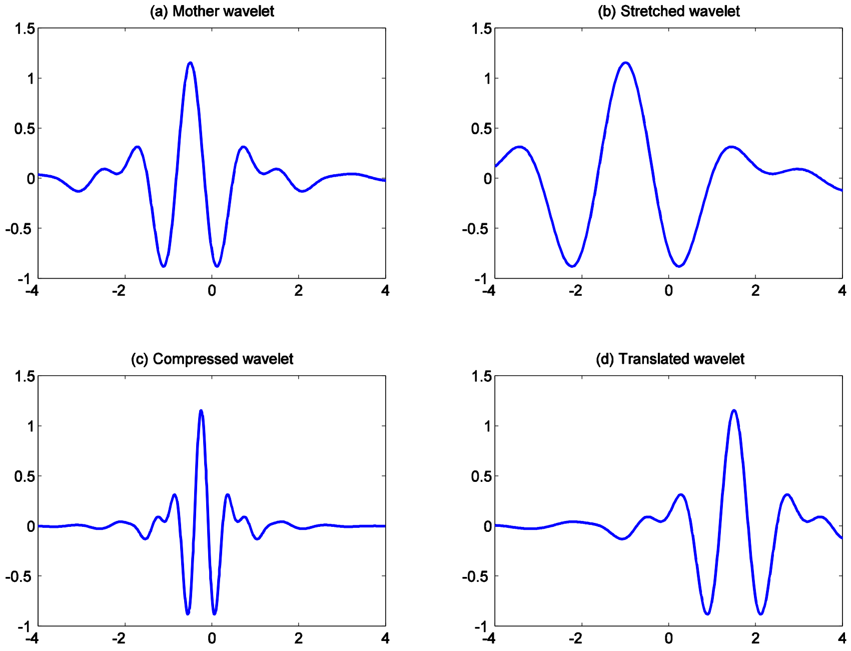


Figure 2.7 Example of a mother wavelet with its stretched, compressed and translated wavelets.

The mother wavelet selected to be the wavelet basis function for analysis requires to possess certain properties including number of vanishing moments, good time localisation and good frequency localisation. The number of vanishing moments is a property that enables the wavelet to suppress a polynomial where it determines the degree of accuracy of the wavelet in representing a signal (Strang and Nguyen, 1997).

Details of the good time and good frequency localisation are discussed in Section 2.5.2. There exist a number of wavelet transform, the two main forms are continuous and discrete. The next section describes the continuous and discrete wavelet transform.

2.5.1 *The Continuous and Discrete Wavelet Transform*

The type of input signal, and the translation and dilation needed to analyse the input signal determines which wavelet transform is applicable. The continuous wavelet transform (CWT) is defined as

$$CWT(a,b) = \langle f, \Psi_{\mathrm{a,b}} \rangle = |a|^{-1/2} \int_{-\infty}^{+\infty} f(t)\Psi * \left(\frac{t-b}{a}\right) dt\,; a, b \in R, a \neq 0 \qquad 2.8$$

which means it is the sum of the signal, $f(t)$, multiplied by the scaled and shifted versions of the wavelet function, Ψ, over all time. For the continuous wavelet transform, the dilation parameter a, and translation parameter b, change continuously which results in a smooth transformation of a signal and generation of a lot of wavelet coefficients.

In contrast, the dilation and translation parameters are discretised for the discrete wavelet transform and the wavelets are calculated on dyadic scales and position i.e. the scaling and translation of the wavelet is based on the power of two. This is a more efficient approach of calculating the wavelet as the wavelet are calculated at alternating scales and positions, which improves calculation time and data storage. Hence the discrete wavelet transform is selected as the wavelet form of choice in this thesis. The discrete wavelet transform is defined as

$$DWT(m,n) = \langle f, \Psi_{m,n} \rangle = a_0^{-m/2} \int_{-\infty}^{\infty} f(t)\Psi(a_0^{-m}t - nb_0)dt. \quad 2.9$$

where m and $n =$ the dyadic scale and position respectively,

a_0 and $b_0 =$ the discretised scaling and translation parameter respectively.

The discrete wavelet transform approach is implemented using complementary low and high pass filters. This concept was introduced by Mallat and Meyer and is termed multiresolution wavelet analysis (Vetterli and Herley, 1992). The following section discusses the concept of multiresolution analysis of the discrete wavelet transform.

2.5.2 *Multiresolution Analysis*

The key principal of multiresolution wavelet analysis is the decomposition and reconstruction of a signal using the wavelet transform. As a signal is analysed using the discrete wavelet transform, it is actually being decomposed into wavelet sub-bands which consist of wavelet coefficients. Figure 2.8 shows the fundamental of the decomposition process where the signal is passed through a low pass and high pass filter. The low pass and high pass filters referred to above are denoted as LP and HP are the key elements of the decomposition process. The function of the low pass filter is to remove the high-frequency component in the signal while retaining the low-frequency components which are termed as approximation coefficients. Meanwhile, the highpass filter is used to capture the ‘bump’ or the high frequency components while simultaneously removing the low-frequency components. The signal components that pass through the high pass filter are termed detail coefficients.

For the successive decomposition, only the low frequency component is passed through another set a low pass and high pass filter whilst the high frequency component is not analysed. As shown in Figure 2.8 at the first level of decomposition, the original signal was decomposed into an approximation and a detail sub-band, namely A1 and D1 respectively. Then the approximation sub-band A1 was decomposed into A2 and D2. An example of the decomposition process to a sample signal from the data of Chapter 5 is shown in Figure 2.9.

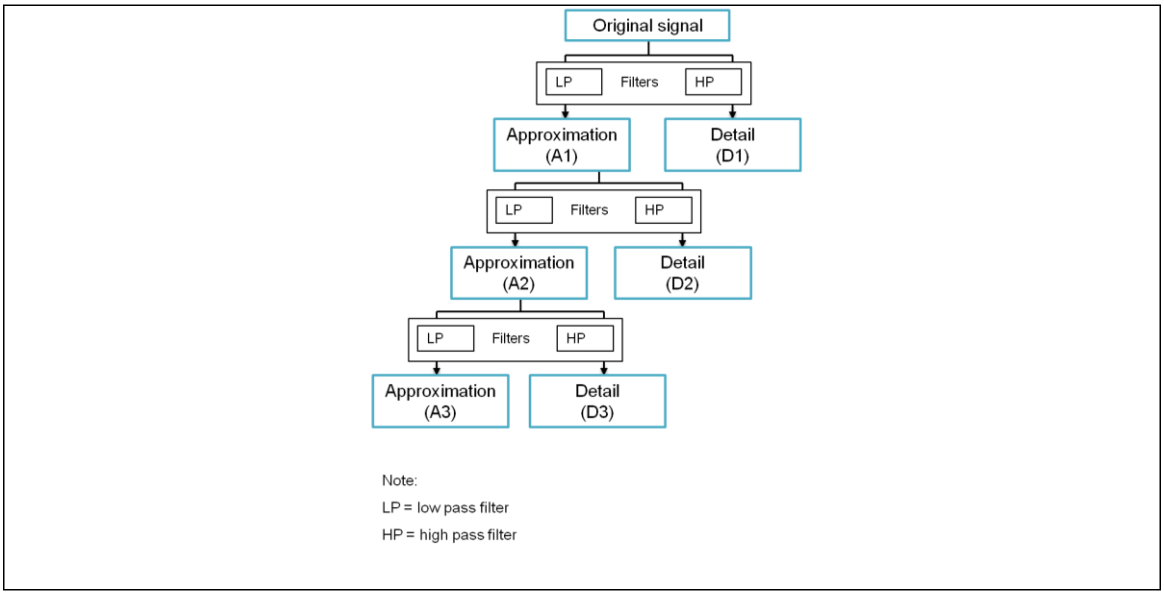


Figure 2.8 A wavelet decomposition process of a signal

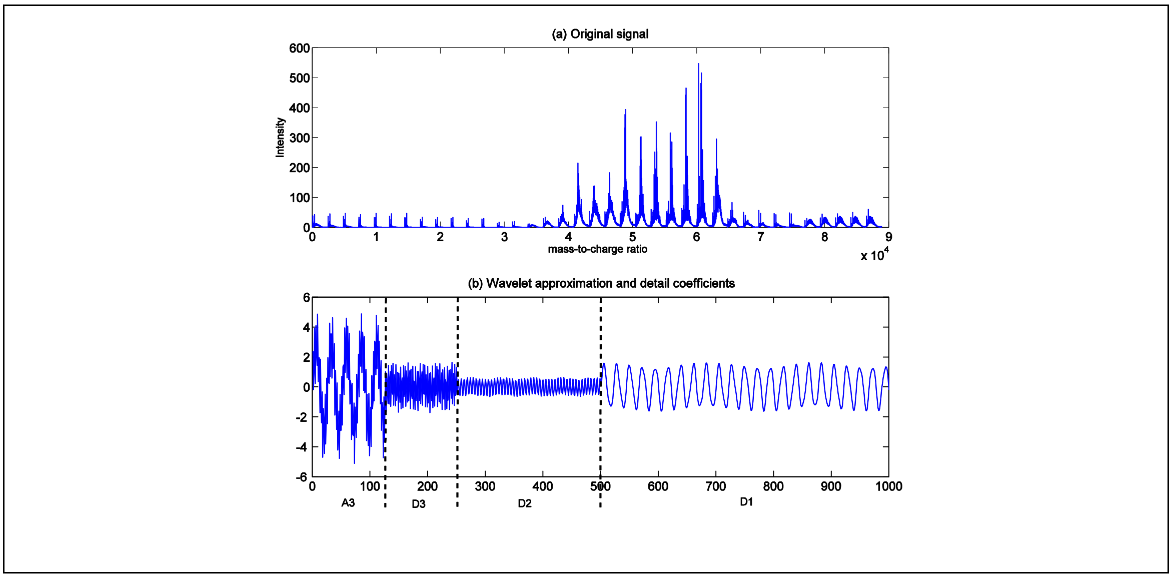


Figure 2.9 A multilevel decomposition performed to an ESI spectroscopy from the data of Chapter 5

The maximum number of decomposition depends on the length of the data to be analysed. For a signal with N data samples, the maximum scale is 2^n where n is the scale. At each successive decomposition (n-1) the length of the signal is halved due to downsampling. Figure 2.9 shows the length of the signal is halved as the decomposition level increase from 1 to 3.

Each of the low and high pass filters is coupled with downsampling where downsampling is the process of reducing the sampling rate of a signal. This approach is needed in the discrete wavelet transform computation as it avoids doubling the length of the original signal when the original signal is passed through the lowpass and highpass filters. A simple explanation of this is the length of output signal coming out of the filter is equal to the length of the input signal. Therefore, when the signal is passed through two parallel filters, its length doubles. By applying the downsampling procedure after each filter, the signal's length coming out of each filter is halved. Later when the two halves signals are combined, they will produce a signal of the same length as the input signal.

Figure 2.10 shows that the time-frequency plane is divided into tiles by the wavelet transform. The varying widths and heights of the tiles represent the compromise between the time and frequency resolutions. For example, at scale n=2, a signal component being analysed will result in a 'good time resolution' as the wavelet is wide and any peaks in the spectrum get smoothed out. A 'good frequency resolution' of the analysed signal can be observed at wavelet scale n=4 at which the frequency of the wavelet is narrower and the peaks are sharper and have larger amplitude. In essence, the different window sizes have enabled the wavelet transform to analyse nonstationary signals.

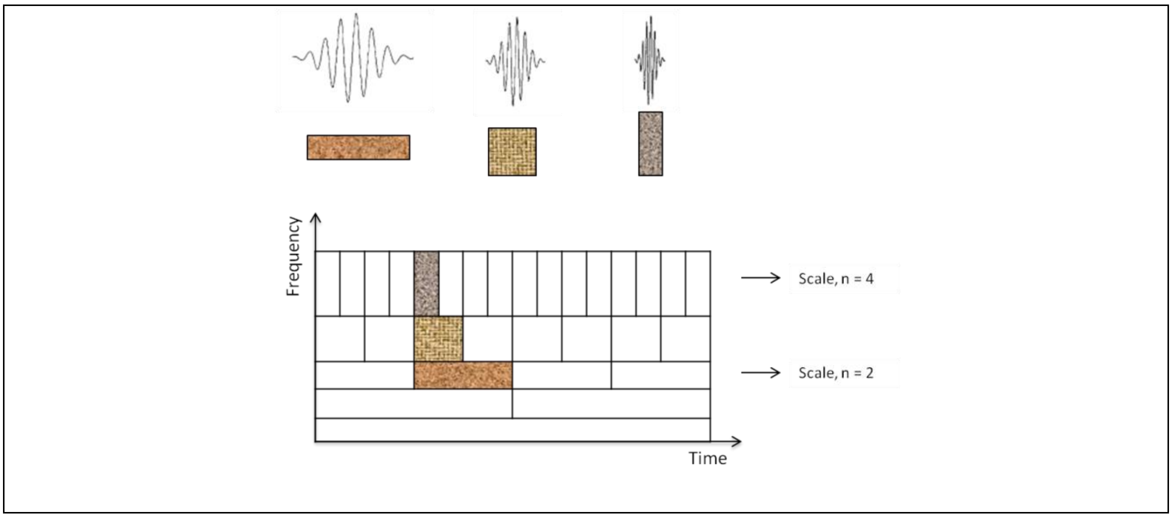


Figure 2.10 Tiling of the time-frequency plane defines the time-frequency boxes of a wavelet basis

On the other hand reconstruction is a process where the wavelet coefficients are reconstructed and summed into the original signal with minimum loss of information. The decomposed sub-bands will be used to reconstruct the original signal in the wavelet reconstruction process. A signal that has been decomposed into wavelet sub-bands can be reconstructed utilising the wavelet reconstruction algorithm of the discrete wavelet transform.

Mathematically, wavelet reconstruction is essentially the inverse discrete wavelet transform. The approximation and detail coefficients from the decomposition stage are upsampled in the reconstruction stage. The purpose of the upsampling operation is to retrieve full-length vectors of the original signal. Contrary to downsampling, upsampling is an approach where zeros are inserted between samples to lengthen a signal component. Once upsampled, the approximation and details coefficients are passed through lowpass and highpass filters respectively. The multiresolution property of the discrete wavelet transform allows different ways to reconstruct a signal. Figure 2.11 shows an example of three-level wavelet reconstruction for which there are three ways to reconstruct the original signal:

Original signal = A1 + D1

= A2 + D2 + D1

= A3 + D3 + D2 + D1

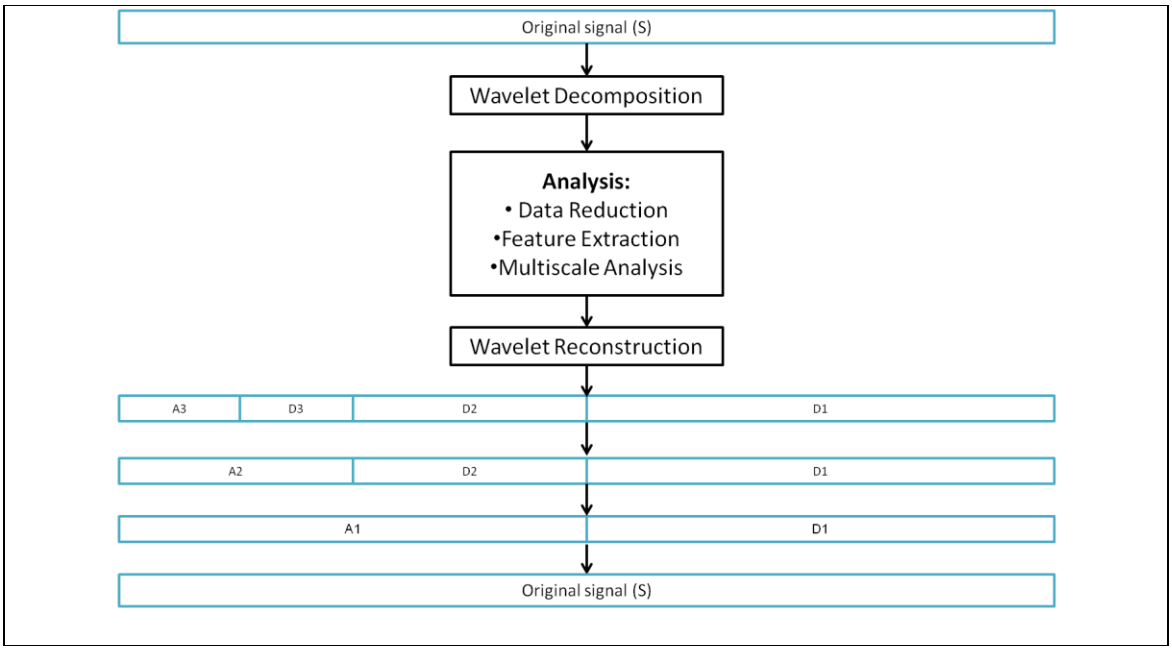


Figure 2.11 Several options to reconstruct a 5 level wavelet reconstruction

However, prior to reconstruction, these coefficients can be analysed (Mallet et al., 1997) for a number of purposes including wavelet denoising, data reduction, feature extraction, and multiscale analysis. Application of the discrete wavelet transform as data reduction and feature extraction technique are described in Chapter 4 whereas multiscale analysis is discussed in Chapter 5. An example of wavelet denoising is discussed in Section 2.6.

2.6 Wavelet Denoising

The objective of wavelet denoising is to remove noise from the data through the application of a thresholding algorithm. This has been shown in studies on NIR spectra, (Donald et al. 2005). Applying a denoising algorithm to the transformed spectra results in the removal of small-amplitude components in the transformed domain (Barclay et al., 1997). The corrected coefficients are considered ‘noise-free’ and the reconstructed signal from these 'noise-free' coefficients is termed the denoised signal.

In the application of the denoising technique, certain aspects require to be considered including selecting between global and level dependent thresholding and choosing between soft and hard thresholding. A global thresholding considers a constant threshold value for all wavelet decomposition levels whilst a level-dependent thresholding is where the threshold value is local to a particular decomposition level. The denoising threshold value δ is defined as

$$\delta = \sigma\sqrt{2\log L} \qquad 2.10$$

where $\sigma =$ standard deviation of the wavelet coefficients

$L =$ number of data points

In hard thresholding, the detail coefficients with an absolute value lower than the set threshold, δ, are set to zero. For soft thresholding the nonzero coefficients are shrunk towards zero by subtracting the threshold δ from the values larger than δ. Both global and level-dependent thresholding are investigated on a sample near infra-red spectrum with the soft thresholding option because in MATLAB hard thresholding is a default setting for data compression. Mother wavelet sym8 with five levels of decompositions is selected as the wavelet basis. The procedure for de-noising is as follows:

1. Wavelet decomposition: decompose the NIR spectrum using sym8 with a decomposition of level 5.
2. Threshold the detail coefficients using the global and level-dependent thresholding.
3. Reconstruct the signal the denoised signal

Figure 2.12 illustrates the original spectra and the denoised spectra for the application of wavelet denoising utilising the global thresholding. It is observed that the magnitude of the denoised spectra was smaller than that of the original spectra. This is due to the removal of the detail coefficients. The denoised spectrum of a wavelet denoising performed with the level-dependent thresholding appears to mapped on the original spectrum. However, a closer inspection reveals that there is a slight phase shift and slight decrease in the magnitude of the peaks as shown in Figure 2.13(b) and Figure 2.13(c). The results from this investigation are used in the development of the process representation in Chapter 4.

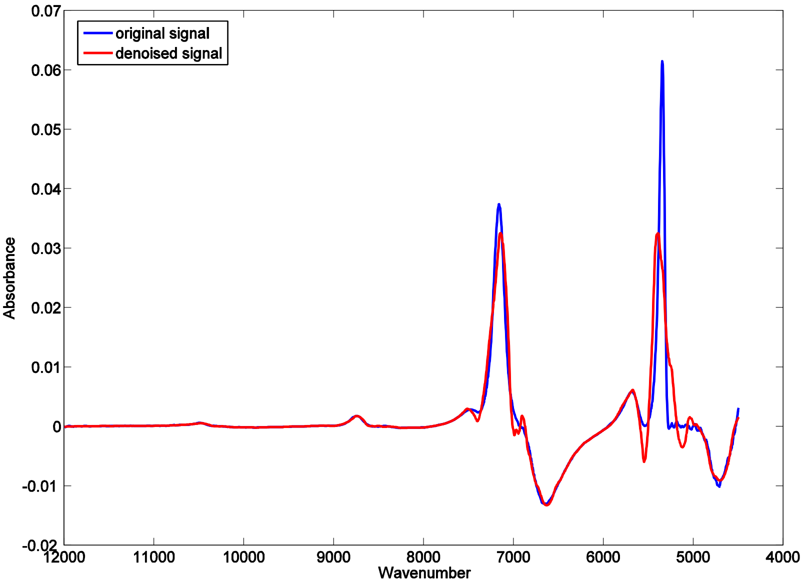
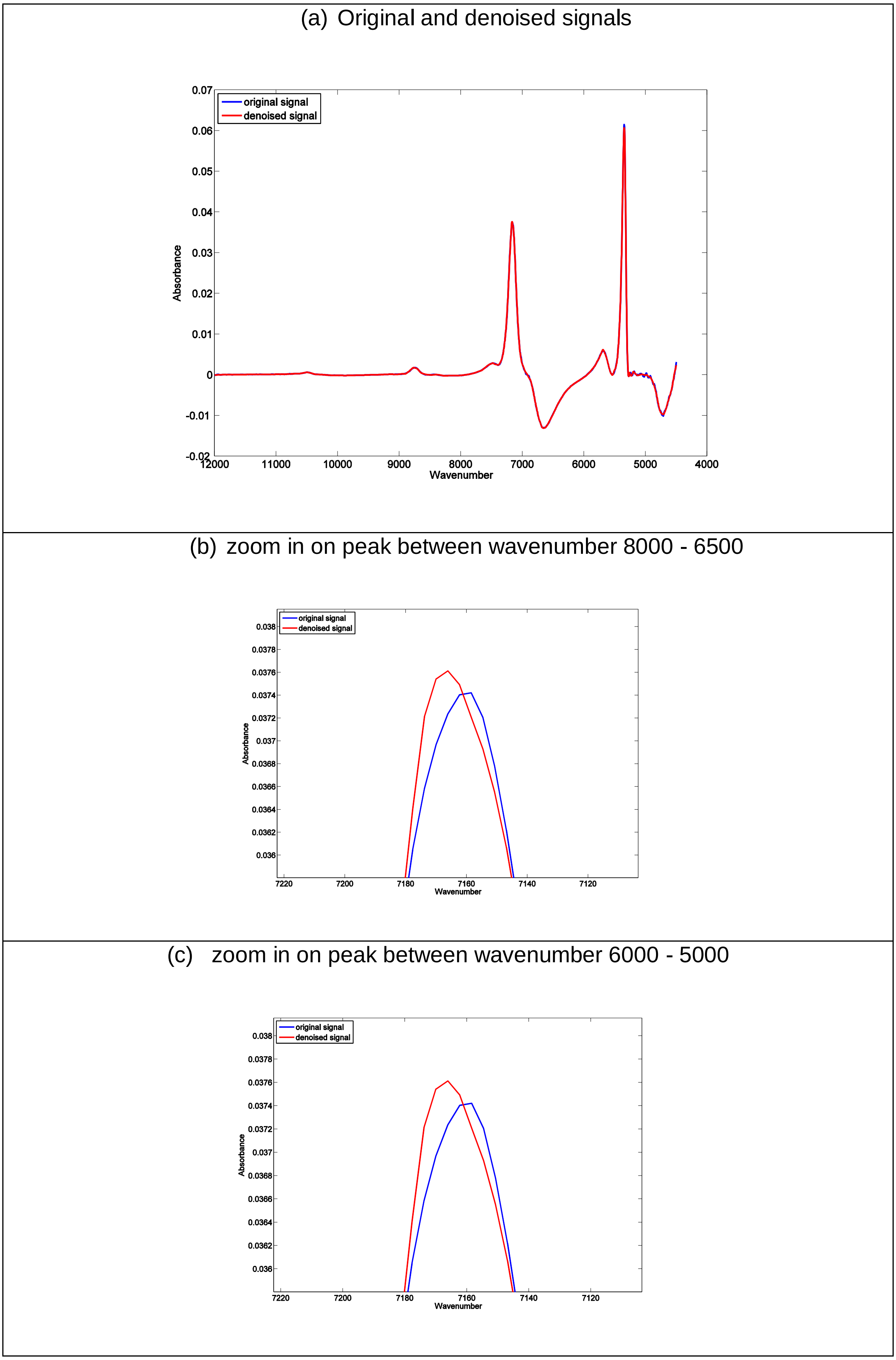


Figure 2.12 A wavelet denoising is perfomed with global thresholding on the NIR spectroscopy from the data of Chapter 4


(a) Original and denoised signals

(b) zoom in on peak between wavenumber 8000 - 6500

(c) zoom in on peak between wavenumber 6000 - 5000

Figure 2.13 Example of wavelet denoising with level dependent thresholding

2.7 Summary

Wavelet theory is heavily supported by mathematical theories. Its applications nevertheless extend to various scientific and engineering fields. This chapter has only provided a brief overview of the theory of the wavelets which are related to the studies of this thesis.

Firstly, the rationale of the implementation of the wavelet transform was discussed. Brief descriptions of the Fourier transform and short time Fourier transform were first given before discussing the wavelet transform in more depth. Then the key theoretical basis of the wavelet transform, multiresolution, was discussed. The multiresolution property of the wavelet transform allows for features of the signals (in the form of wavelet coefficients) at different resolution to be extracted. This is a powerful technique in extracting hidden features in signals masked by noise and overlapping spectra. As shown in the case studies presented in Chapter 4 and 5, features extracted using the multiresolution technique result in fingerprinting of the complex bioprocess data.

Finally, the investigation of wavelet denoising, one property of the discrete wavelet transform is performed. This investigation is a part of the foundation to the application of discrete wavelet transform in Chapter 4 and 5.

Chapter 3 Overview of Multivariate Data Analysis Technique

3.1 Introduction

As protein therapeutics increasingly become major player in the pharmaceutical industry (Goodman, 2009), biopharmaceutical companies are on the search for more efficient manufacturing strategies and an increase in process understanding to cope with the increasing demands. Hence consistency in the manufacturing of high quality product and reduction in product development time can be achieved.

From product development to manufacturing stages of the protein therapeutics, involves the measurement and recording of an enormous amount of data which represent an opportunity for data mining that can contribute to process understanding. However, this data is typically complex as it is multidimensional and comprises multiple variables hence necessitates the utilisation of multivariate data analysis.

Multivariate data analysis is recognized as one of the promising techniques for enhancing process understanding by the FDA's Process Analytical (PAT) initiative (FDA, 2004). It is the extension of bivariate data analysis to higher-order datasets (Acar and Yener, 2009). The objective of applying multivariate data analysis is to enable the analyses of data comprising multiple variables to capture hidden structures and underlying correlations between variables. The goal of this chapter is to present an introduction to the multivariate data analysis technique employed in this thesis. Also, it provides the theoretical background to the application of multivariate data analysis in the two case studies discussed in this thesis.

The next section discusses the definition of principal component analysis and its metrics as it is the specific form of multivariate data analysis employed in this thesis.

Following this, two techniques of data preprocessing are presented. The two techniques, mean centring and standardisation, are the techniques used in both case studies presented in this thesis. Finally the concept of multiway techniques which includes detailed discussion of batch data unfolding and two primary techniques of multiway principal component analysis are provided.

3.2 Principal Component Analysis

> *"Principal component analysis is the backbone of latent variable methods" (Geladi and Grahn, 1996).*

One form of multivariate analysis is principal component analysis (PCA). It is a technique in which the original variables are transformed into a new set of latent variables called principal components. The principal components are mutually orthogonal and are linear combinations of the original variables. The first principal component (PC1) defines the direction of greatest variability, with subsequent principal components explaining a decreased amount of variability within the data set. Consequently lower order principal components can be excluded as they characterise the noise in the process. By retaining a limited number of principal components the dimensionality of the problem is then reduced (Jolliffe, 2002).

The history of PCA goes back to 1901 when it was proposed by Karl Pearson but the algorithm was developed by Harold Hotelling in 1933 (Jolliffe, 2002). The application of PCA became more widespread with the advance in computer power. PCA has been widely applied and covers wide variety of areas such as chemistry and biology. Some of the major objectives of the application of PCA are:

- identifying principal components or new meaningful variables from the original variables to determine the underlying trend of the problem;
- dimensional reduction of the problem;

- exclusion of some of the original variables that only contribute limited information in the context of the problem.

The methodology of PCA is discussed in the next section and the statistical metrics are then introduced.

3.2.1 *Methodology of Principal Component Analysis*

Principal components are derived by projecting the samples in a data set onto new space. From data generated during a process, consider a data matrix $\boldsymbol{X}$ with n rows (samples) and p columns (variables). The samples are typically time points or batches for particular measurements of a variable such as wavenumber.

The derivation of the principal components of a data matrix $\boldsymbol{X}$ involves its decomposition into a sum of the outer product of vectors $\boldsymbol{t_i}$ and $\boldsymbol{p}_i$:

$$\boldsymbol{X} = \quad \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \cdots + \boldsymbol{t_r p_r^T} + \cdots + \boldsymbol{t_R p_R^T} \qquad 2.1$$

where vector $\boldsymbol{t}_\mathrm{i}$ is the scores vector, $\boldsymbol{p}_\mathrm{i}$ is the loadings vector and $\boldsymbol{R}$ represents the maximum number of principal components, i.e. min (n, p). The scores vector $\boldsymbol{t}_\mathrm{i}$ is the projection of the samples onto the principal components. It describes the relationships between samples and is a weighted linear combination of the original variables with the weight defined by the loadings, $\boldsymbol{p}_\mathrm{i}$. The scores vector is defined by

$$\boldsymbol{t_{id}} = \quad \boldsymbol{x_{ij} p_{jd}} \qquad 2.2$$

where $\boldsymbol{x}_{ij}$ is the element of the jth variable measured for the ith sample and $\boldsymbol{p}_{jd}$ is the vector of loadings for variable j in dimension d. The superscript T of the matrix $\boldsymbol{p}$ indicates the matrix needs to be transposed.

Hence for a particular sample the score is

$$\boldsymbol{t_{11}} = \boldsymbol{x_{11}p_{11}} + \boldsymbol{x_{12}p_{21}} + \cdots + \boldsymbol{x_{1p}p_{p1}} \qquad 2.3$$

The loading vector, $\boldsymbol{p}_i$ describes the relationships between variables and is defined as the eigenvector of the covariance matrix of $\boldsymbol{X}$:

$$\boldsymbol{cov}(\boldsymbol{X}) = \frac{1}{n-1}(\boldsymbol{X}^T\boldsymbol{X}) \qquad 2.4$$

For each $\boldsymbol{p}_i$, Equation 2.4 can be written as:

$$\boldsymbol{cov}(\boldsymbol{X})\boldsymbol{p_i} = \boldsymbol{\lambda_i p_i} \qquad 2.5$$

The eigenvalue λ_i for eigenvector $\boldsymbol{p}_i$ is a measure of the variance explained by each principal component with the largest eigenvalue corresponding to the first principal component thereby capturing the main source of variability in the data with subsequent principal components explaining less amount if variability.

Typically, the first few principal components account for most of the variability and have information in the data and hence Equation 2.1 can be simplified to

$$\boldsymbol{X} = \boldsymbol{t_1 p_1^T} + \boldsymbol{t_2 p_2^T} + \cdots + \boldsymbol{t_r p_r^T} + \boldsymbol{E} \qquad 2.6$$

where $\boldsymbol{E}$ is a residual matrix in which the variance not explained by the retained principal components is captured. The superscript $\boldsymbol{T}$ of the matrix $\boldsymbol{p}$ refers to matrix $\boldsymbol{p}$ being transposed. The method of selecting $\boldsymbol{r}$ is discussed in section 3.2.3. It is noted that there exists other methods of calculating principal components including singular value decomposition (SVD) and non-linear iterative partial least squares (NIPALS). In the cases where the columns of $\boldsymbol{X}$ are scaled, the covariance matrices in Equations 2.4 and 2.5 are replaced with correlation matrices.

3.2.2 *Scores and Loadings Plots*

As stated in the Section 3.2.1, scores convey knowledge on relationships between samples whilst loadings provide information on relationships between variables. There are several ways to represent scores and loadings matrices graphically. In this section, two types of graphical representations will be discussed.

The first graphical representation is bivariate scores plot and univariate loadings plot as these are the plots utilised in the analysis of the problems presented in this thesis. Figure 3.1 shows an example of bivariate scores plot of a sample data taken from Chapter 4 where the scores of principal component one is plotted against the scores of principal component two. The points plotted represent the position of the samples in the scores plane. The underlying trend that can be extracted from Figure 3.1 is the process batch was clustered according to their groups with Group 1 and Group 2 were positioned near to the origin. Bivariate scores plot of the remaining principal components can be plotted in the same manner, for example PC3 against PC4.

A univariate loadings plots is a bar chart of a particular loadings against variables. Figure 3.2 demonstrates an example of a univariate loadings plot from an analysis on a near infrared dataset where PC1 is plotted against the wavenumber of the near infrared spectra. The loadings plot helps identify the most important wavenumbers in terms of individual principal components. According to Figure 3.2, large magnitude of loadings comes from the wavenumbers 7200 to 6500, 5500 to 5000, and 5800 to 4500. Other than bar chart, the univariate loadings plot can be plotted as lines.

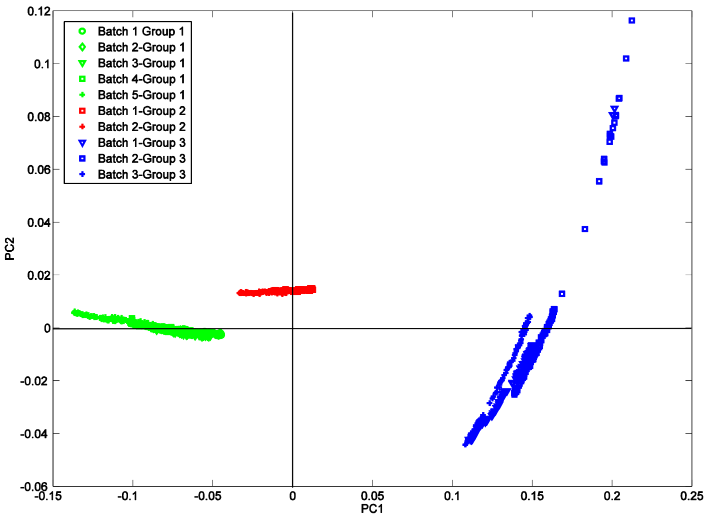


Figure 3.1 An example of bivariate scores plot (PC1 vs. PC2) generated from the NIR data from Chapter 4

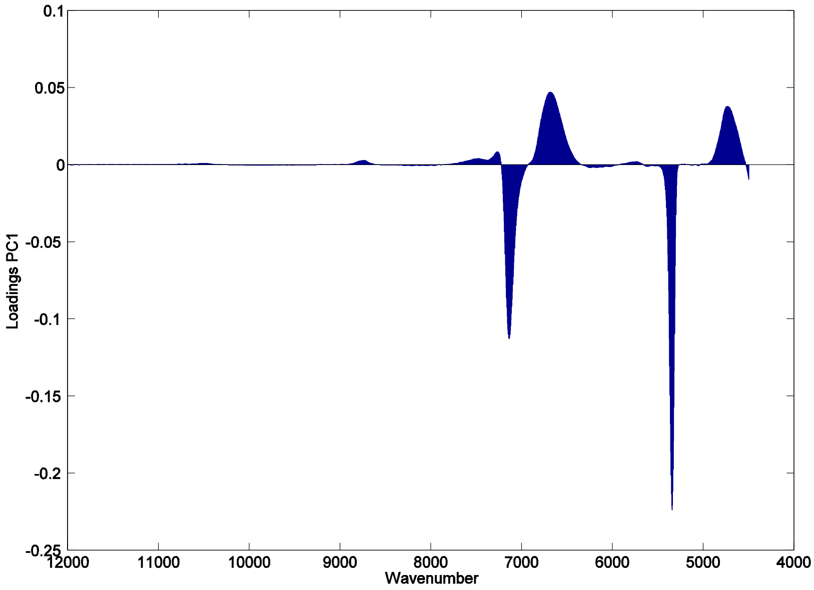


Figure 3.2 An example of univariate loadings plot generated from the NIR data from Chapter 4

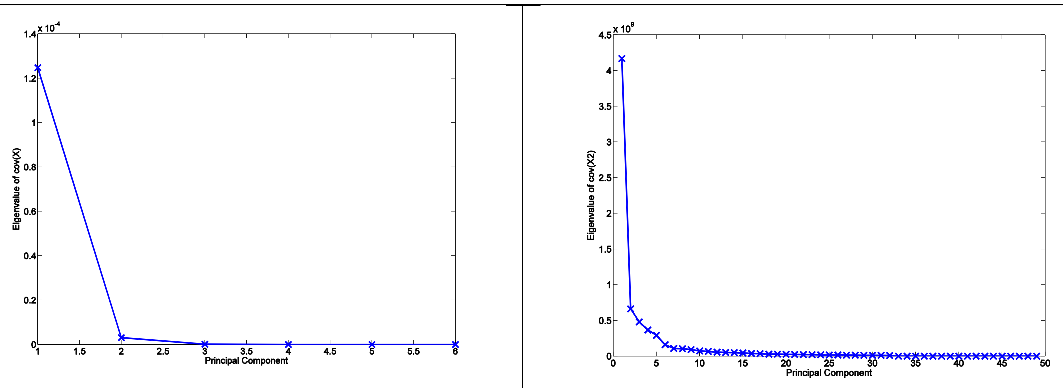
3.2.3 *Selection of the Number of Principal Components to be retained*

In constructing the PCA representation, one crucial step is to determine the number of principal components to retain to capture the main sources of variability in the data. A number of techniques have been proposed in the literature for selecting the number of principal components required to retain. One technique is to include sufficient principal components so that the cumulative variability explained is between 80% and 90% or more;

$$Percentage\ of\ variance\ explained = \quad 80\% < \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{R} \lambda_i} < 90\% \qquad 2.7$$

where λ is the eigenvalues of the dataset, r is the number of principal components being retained, and R is the maximum number of principal components.

Another technique is to consider the explicit values of the principal components, those whose eigenvalues are less than one should be excluded as they explain less variability than an individual variable. Alternatively, a plot of the eigenvalue against the number of principal components can be used to identify where an “elbow” in the curve occurs and define the number of principal components to retain as shown in Figure 3.3. In the near infrared data, the “elbow” in the curve is identified at the second principal component whilst in the electrospray ionisation mass spectrometry data it is detected at the sixth principal component.



(a)

(b)

Figure 3.3 Examples of number of principal components retained: (a) near infrared data of Chapter 4; (b) electrospray ionisation mass spectrometry data of Chapter 5

3.2.4 *Contribution Plots*

The contribution plots can be calculated for the scores provide insight as to which variables are responsible for the non-conforming behavior. The principal components scores can be described as a weighted sum of the process variables:

$$t_{ir} = \sum_{j=1}^{p} x_{ij} p_{jr} \qquad 2.8$$

where $x_{ij} =$ value to sample i and variable j

$p_{jr} =$ loading for variable j for principal component r.

The score $\boldsymbol{t}_{ir}$ for each sample i and principal component r can then be decomposed into j=*1,2,…,* p variables. The contribution of each process variable to the individual scores of the PCA model is given by:

$$Contribution_{ij} = \frac{t_{ir}}{\sigma_i^2} p_{ij} \left(x_j - \bar{x}_j \right) \qquad 2.9$$

where $t_{ir} =$ score of sample i for principal component r

$p_{ij} =$ loading for the sample i and variable j

$\sigma_i =$ standard deviation of sample i

$\bar{x}_j =$ mean of variable j

$x_j =$ corresponding variable j

The variables with large absolute contribution are expected to contribute to the cause of the sample not exhibiting similar performance to the other (Westerhuis et al., 2000). Other than the contribution plots to the scores, the contribution plots for metrics including Hotelling's T^2 and square prediction error (SPE) can also be calculated. However, it is not discussed as these two metrics are not utilised in this thesis.

3.3 Data Prepocessing

The data matrix $\boldsymbol{X}$ may require to be scaled prior to applying principal component analysis. The rational for scaling the data is that if the variables have significant difference and standard deviation then the main source of variability will be attributed to the variable with the large standard deviation. Consequently the principal component will not reflect the maximum scores of variability inherent to the process under investigation.

There are a number of techniques by which to scale a dataset. These can be categorized into a number of types (Brereton, 2009) where two main types are: (1) row (samples) scaling, and (2) column (variable) scaling. Each type of scaling provides a number of techniques to choose from. Two techniques, mean centring and standardisation, which are used in this thesis are discussed below. The application of mean centring in the context of this thesis is discussed in Section 4.10 and Section 5.6.2.

3.3.1 *Mean Centring*

Mean centring is defined as removing a constant offset across the columns or rows of a data matrix (Gurden et al., 2001). Column centring is when the column average, $\bar{x}_j$ is subtracted from each element in a column:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \qquad 2.10$$

then the mean centred element x_{ij}^* is given by

$$x_{ij=}^* \quad x_{ij} - \bar{x}_j \qquad 2.11$$

where n is the number of samples.

The second type, centring across the second mode or row centring is when the row average is subtracted from every element in the row. It can be mathematically expressed as

$$x_{ij}^* = \quad x_{ij} - \bar{x}_i \qquad 2.12$$

The mean of the ith row of data matrix $\boldsymbol{X}$ is given by

$$\bar{x}_i = \frac{\sum_{j=1}^{p} x_{ij}}{p} \qquad 2.13$$

where p is the number of variables.

Centring across either a column or row is known as single centring whilst double centring is where both form of single centring are applied (Bro and Smilde, 2003). The order is irrelevant as the application of single centring to one direction does not disturb the centring in the other direction.

3.3.2 *Standardisation*

Standardisation involves first mean centring and then scaling the data to unit standard deviation. The original sample i for variable j, that x_{ij} is transformed into:

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad 2.14$$

where $\bar{x}_j =$ mean of variable j

$s_j =$ standard deviation of variable j

Standardisation is applied so that all variables have a similar influence when developing the process representation.

3.4 Multiway Techniques

3.4.1 *Introduction to Batch Data Unfolding*

Batch process data typically comprises measurement of J process variables (j=1,2,…,J) recorded at regular time intervals (k=1,2,…,K) throughout the batch run. Similar data is collected for a number of batch runs i =1,2,…, I. This information can be organized into a three-dimensional data array, $\underline{\boldsymbol{X}}$, Figure 3.4.

Multiway principal component analysis is an extension of principal component analysis for three-dimensional data. It is performed by initially unfolding the three-dimensional data array to a two-dimensional matrix and then by the application of principal component analysis to the resulting two-dimensional matrix.

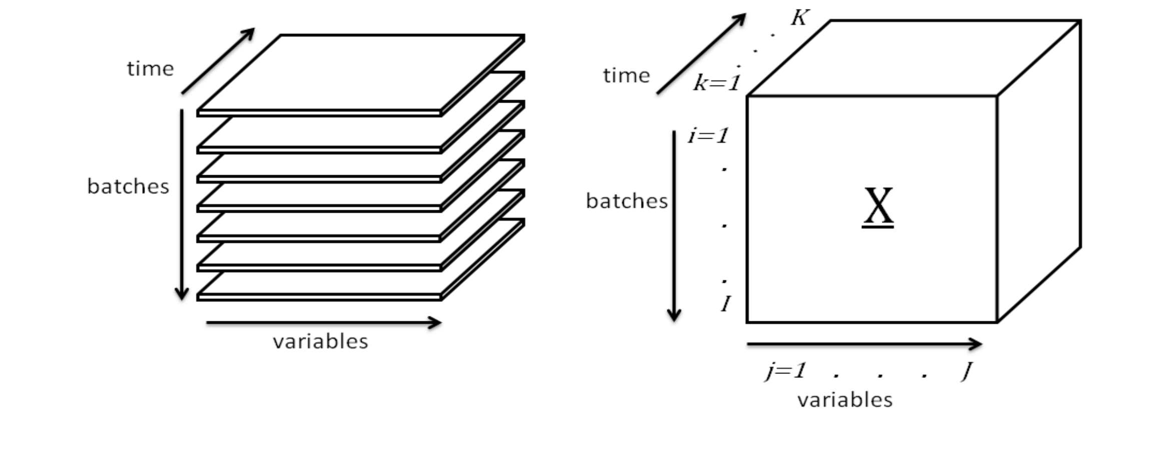


Figure 3.4 A three-dimensional data array, $\underline{\boldsymbol{X}}$, of batch process

There are three possible methods to unfold the data matrix, $\underline{\boldsymbol{X}}$. In each case the direction of one axis is preserved and the direction of the other two axes are transposed, resulting in 3 two-dimensional matrices: $\boldsymbol{A}$ (I x KJ), $\boldsymbol{B}$ (I x JK) and $\boldsymbol{C}$ (J x IK) illustrated in Figure 3.5.

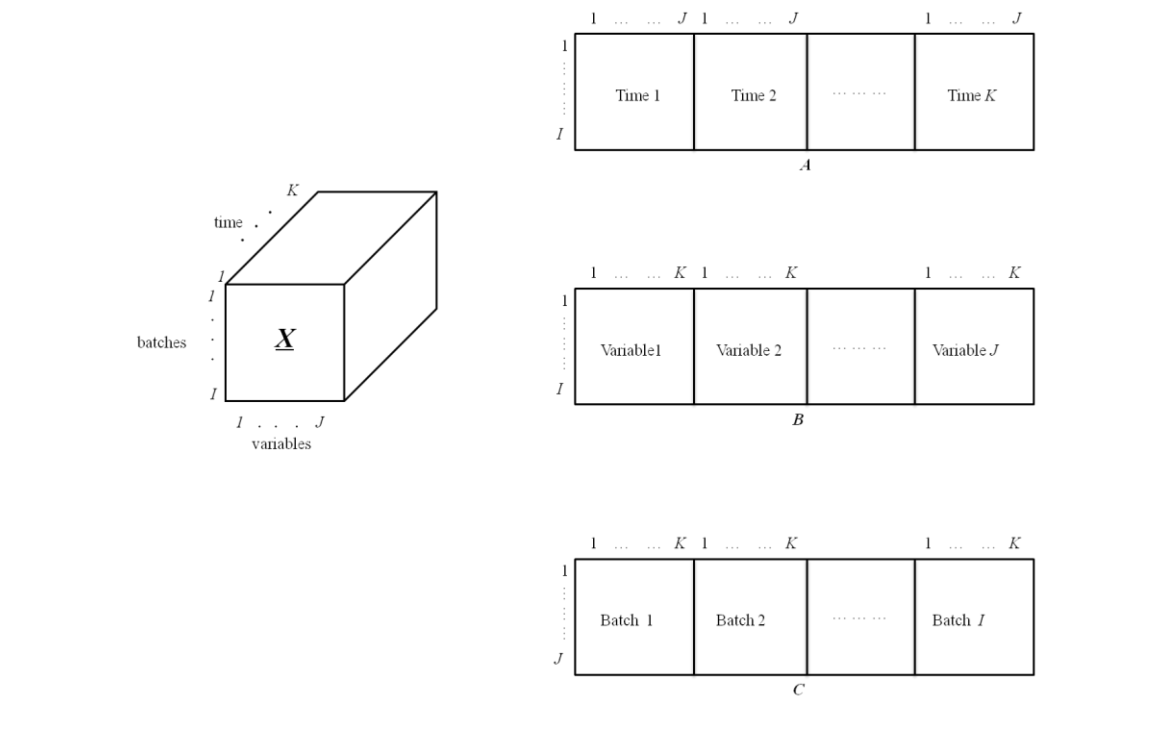


Figure 3.5 Unfolding methods of a three-dimensional data matrix

For matrix $\boldsymbol{A}$ and $\boldsymbol{C}$, measurements of variables logged at the same time were kept together for all batch runs. For matrix $\boldsymbol{B}$, measurements of an individual variable during the duration of the batch were kept together for all batch runs. These different unfolded matrices plus PCA on them corresponds to the analysis a different type of variability.

MPCA allows the description of the multiple batches operation by compressing the information contained in the data trajectories into low-dimensional spaces. It develops a data-based model by utilising a number of orthogonal latent variables. Once the data set is unfolded, the model is constructed in a manner that its maximizes covariance between the input and output data spaces (Wold et al., 1987). Consequently, the unfolding approach of choice depends on the objective of the data analysis. The subsequent section provides description the unfolding of matrix A.

3.4.2 *Multiway PCA - The Nomikos and MacGregor Approach*

The approach introduced by Nomikos and MacGregor (1994) is to unfold the three-dimensional data matrix to a two-dimensional data matrix by preserving the direction of the batches. As discussed in Section 3.4.1, when the direction of the batches is preserved, the results are two-dimensional data matrices $\boldsymbol{A}$ (Figure 3.5). Data matrix $\boldsymbol{A}$ is the format proposed by Nomikos and Macgregor (1994).

The two-dimensional data matrix $\boldsymbol{A}$ (I x KJ) shown in Figure 3.5 is obtained by slicing the three-dimensional matrix $\underline{\boldsymbol{X}}$ vertically and organising each vertical slice, side by side. The vertical slices (I x J) representing the values of all variables for all the batches at time interval k. Adopting this approach the information contained in the data matrix $\underline{\boldsymbol{X}}$ can be summarised with respect to both variables and their time variation, allowing the analysis of variability between the batches and subsequently making this approach an effective one to analyze historical data (Wold et al., 1998). Also, the goals and benefits of MPCA are similar to PCA. It has been shown that MPCA is statistically and algorithmically consistent with PCA (Nomikos and MacGregor, 1994).

The objective of MPCA is to decompose the matrix $\underline{\boldsymbol{X}}$ into a summation of the product of the score vector ($\boldsymbol{t}_r$) and the loading matrices ($\boldsymbol{P}_r$), plus a residual $\underline{\boldsymbol{E}}$:

$$\boldsymbol{X}_{IxKJ} = \sum_{r=1}^{R} \boldsymbol{t}_r \otimes \mathbf{P}_\mathrm{r} + \underline{\mathbf{E}} \qquad 2.15$$

or

$$\boldsymbol{X}_{IxKJ} = \sum_{r=1}^{R} \boldsymbol{t}_r \mathbf{p}_\mathrm{r}^\mathrm{T} + \underline{\mathbf{E}} \qquad 2.16$$

where R is the number of principal components.

This decomposition conforms to the principles of PCA as it separates the data into two parts (Jolliffe, 2002), the systematic variation ($\sum_{r=1}^{R} \boldsymbol{t}_r \otimes \boldsymbol{P}_r$), i.e. it expresses one fraction ($\boldsymbol{t}_r$) related to the batches I and a second fraction ($\boldsymbol{P}_r$) related to the variables J and their time variation K. The second part is the residual $\underline{\boldsymbol{E}}$, which typically describes the noise associated with the data.

The unfolded data matrix $\boldsymbol{A}$ is usually mean centred prior to performing PCA. Mean centring the matrix $\boldsymbol{A}$ results in the removal of the mean trajectory of all spectra thereby removing the main non-linear component in the data. As a result, applying PCA to the mean-corrected data is a study of the systematic variation in the resulting trajectories of the spectral data for all batch about the mean trajectory. The following section provides description another option to unfold the three-dimensional matrix $\underline{\boldsymbol{X}}$.

3.4.3 *Multiway PCA - The Wold, Kettaneh, Friden and Holmberg Approach*

In the original paper of Wold et al. (1998), they proposed an alternative approach to MPCA to that of Nomikos and MacGregor (1994). Their approach is based on unfolding the three-dimensional data matrix by preserving the direction of the variables. The resulting two-dimensional matrix consists of (I x K) rows and J columns, assuming equal batch length, Figure 3.6. Each row then comprises the data for all variables for an individual batch at time point k. The unfolding of this method is defined by:

$$\boldsymbol{X}_{IKxJ} = \sum_{r=1}^{R} \boldsymbol{t}_r \mathbf{p}_{\mathrm{r}}^{\mathrm{T}} + \underline{\mathbf{E}} \tag{2.17}$$

where $t_r =$ scores vector of size (1 xIK)

$p_r =$ loadings vectors of length (J x 1)

$\underline{\boldsymbol{E}} =$ residual matrix

$R =$ number of principal components

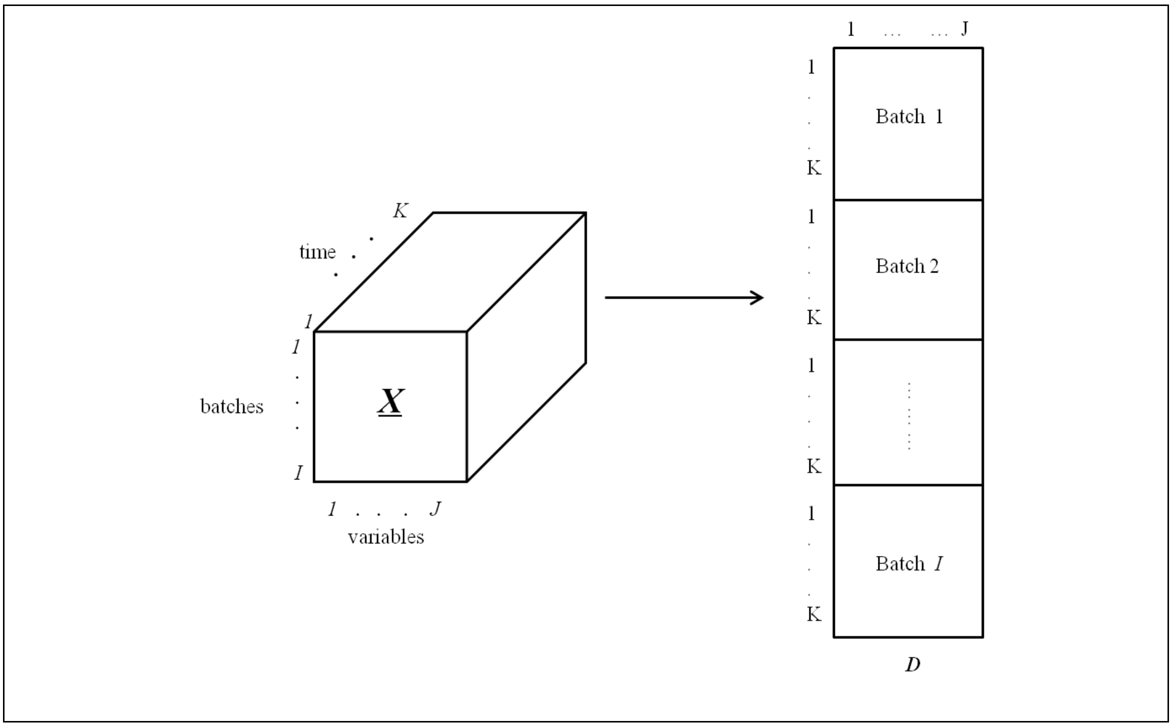


Figure 3.6 Unfolding method of Wold, Kettaneh, Friden and Holmberg (1998)

There are three differences between the Wold et al. (1998) approach and the Nomikos and MacGregor (1994) approach. The key difference is in terms of the length of batches in the three-dimensional matrix $\underline{\boldsymbol{X}}$. Wold et al. (1998) argued that although the MPCA approach proposed by Nomikos and MacGregor is powerful and effective in batch analysis specifically in batch monitoring, it has a downside in that it assumes the data for the complete batch is available.

The second difference concerns the rows of the unfolded matrices $\boldsymbol{D}$ and $\boldsymbol{A}$. The rows of matrix $\boldsymbol{D}$ consist of the data for all variables for an individual batch at time point k whilst the rows of matrix $\boldsymbol{A}$ comprises the data for all variables and time points for a batch. In other words, matrix $\boldsymbol{D}$ regards individual sample as a unit whilst matrix $\boldsymbol{A}$ regards the whole batch as a unit.

The third difference relates to the mean centring and scaling of the unfolded matrix. Prior to applying PCA to matrix $\boldsymbol{D}$, it is scaled to zero mean and unit variance. Mean centring of matrix $\boldsymbol{D}$ is done by subtracting the mean of each variable over all batches and all times from the trajectory of each variable in each batch. The mean-centred matrix $\boldsymbol{D}$ captures covariance among the variables (Westerhuis et al., 1999).

One limitation in the Wold et al. (1998) approach is the number of principal components required to explain the structured variation in the data. Westerhuis et al. (1999) claimed that the Wold et al. (1998) approach demands nearly as many latent variables as original variables to describe the same variation in the data as described by the Nomikos and MacGregor (1994) approach. Consequently, this does not contribute to the dimension reduction of the problem which is one of the objectives in applying PCA.

3.5 Summary

In this chapter, an overview of multivariate data analysis technique specifically PCA is provided. Statistical metrics utilised in this thesis are described with their application to samples of spectral data from subsequent chapters. MPCA, the extension of PCA is also discussed as the problems presented in Chapter 4 and 5 involve three-dimensional data matrix which require to be unfolded prior to analysis. Of the two unfolding methods discussed (Section 3.4.2 and Section 3.4.3), the Nomikos and MacGregor (1995) approach was utilised in the data analysis of both Chapter 4 and 5. Furthermore the next two chapters will described the application of combination of MPCA and the discrete wavelet transform as an approach to fingerprint bioprocess data.

Chapter 4 Integrated Modelling for NIR Industrial Process Data

4.1 Introduction

Biopharmaceuticals represents a rapidly evolving segment of the pharmaceutical industry. The manufacture of biopharmaceutical products is currently a business trend in the pharmaceutical industry with global revenues generated from this business reported at USD120 billion per annum and projected to increase to USD150 billion by 2015 (Butler and Meneses-Acosta, 2012). The economic success of the biopharmaceutical industry is largely a consequence of the manufacture of monoclonal antibodies produced from mammalian cell culture bioprocesses (Walsh, 2010) with the majority of commercial monoclonal antibodies derived from the expression of Chinese Hamster Ovary (CHO) cell lines (Kelley, 2009).

Monoclonal antibodies (MAb) are primarily produced through recombinant mammalian cell culture batch processes which are subjected to batch-to-batch variability (Ferreira et al., 2005). Typically monoclonal antibodies production involves upstream cell culture processes and downstream purification processes. It starts with a vial from the cell bank and ends with the final product (Rathore et al., 2011). Shukla and Thommes (2010) provide a general description of the monoclonal antibodies production process which is shown schematically in Figure 4.1. The cells in the vial are expanded through a series of seed batches in increasing volumes moving from shake flasks. The cell culture is then transferred to the production bioreactor where the cells continue to grow and the monoclonal antibody is expressed into the medium. Following this, the cell culture broth is harvested through centrifugation and filtration steps to remove cells and cell debris. The next stage involves product capture through Protein A affinity chromatography and a further one or two polishing chromatography steps to remove impurities. Finally, the ultralfiltration/ diafiltration step is performed to formulate the bulk drug product.

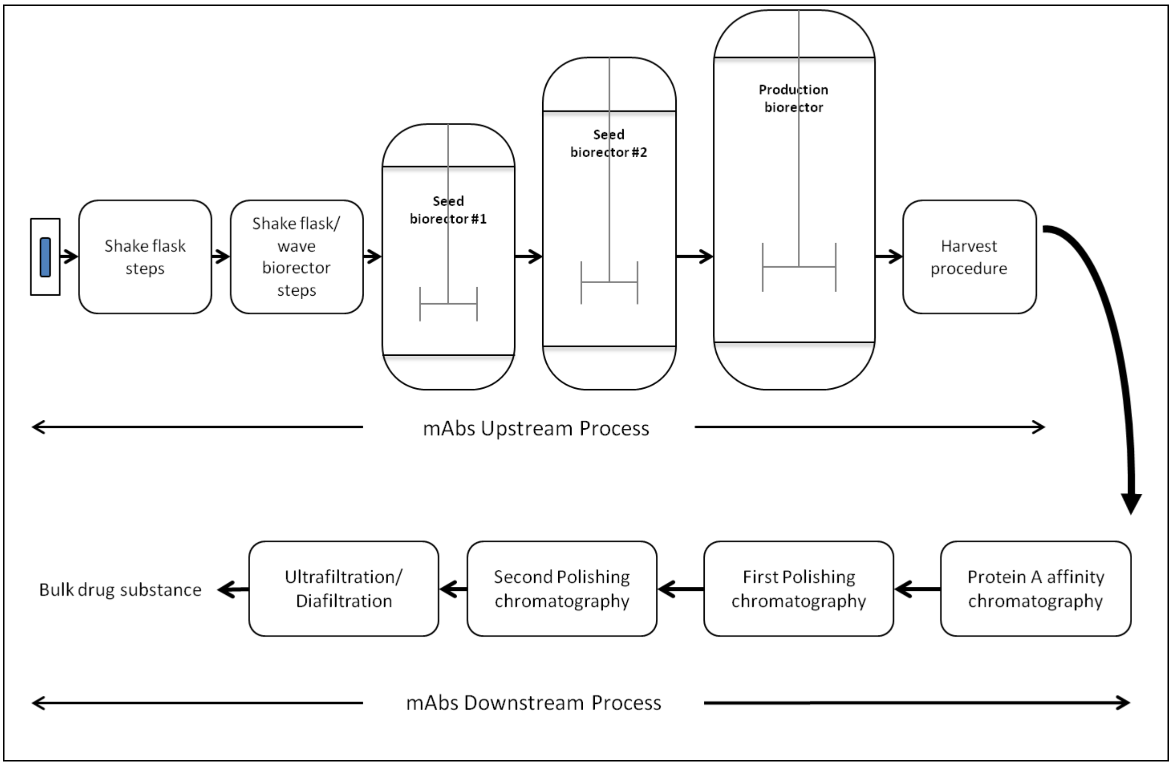

Shake flask steps
Shake flask/ wave biorector steps
Seed biorector #1
Seed biorector #2
Production biorector
Harvest procedure
mAbs Upstream Process
Bulk drug substance
Ultrafiltration/ Diafiltration
Second Polishing chromatography
First Polishing chromatography
Protein A affinity chromatography
mAbs Downstream Process

Figure 4.1 Upstream and downstream process of monoclonal antibody (Shukla and Thommes, 2010)

The relationship between process batches make up the genealogy or family tree with the genealogy of a production batch being complex. Inconsistency in process scale-up from the shake flasks to production scale and batch-to-batch variability are among the key challenges in industrial bioprocess (Schmidt, 2005) where producing consistent and good quality product is the central goal of process development (Steinmeyer and McCormick, 2008). To increase product quality, Jenzsch (2006) proposed maintaining good reproducibility of the batch process. This can be achieved through understanding of the variability of batch behaviour and hence knowledge of the genealogy (parent-child or seed-production relationship) is key requirement. Process understanding is a key component of the Process Analytical Technology (PAT) initiative introduced by the Food and Drug Administration (FDA) (FDA, 2004).

The case study described in this chapter is based on a monoclonal antibody (MAb) manufacturing process from a mammalian cell from which on-line traditional measurements, on-line spectral measurements, and off-line sample analysis are recorded. This data is a source of information from which enhanced process understanding can be attained.

The primary objective of this Chapter is to investigate whether process understanding may be enhanced by utilising information from the process genealogy. Such knowledge may potentially be invaluable in terms of industrial-scale operations, for example in process scale-up, process comparability, and technical transfer between production sites. The approach adopted is to develop and establish a fingerprint of the process NIR spectroscopy data to capture differences between batches. The methodology adopted is a combination of the wavelet transform and multiway PCA. It is shown that the proposed procedure is successfully able to describe distinctive characteristics of batches relative to the underlying genealogy.

Traditional chemometrics approaches applied to mammalian cell culture unit operations include principal component analysis (PCA) (Gunther et al., 2007) , partial least squares (Teixeira et al., 2009; Riley et al., 1999), and, multiway principal component analysis (MPCA) and multiway partial least squares (MPLS) (Ferreira et al., 2007; Cunha et al., 2002); with the primary focus being monitoring and fault detection. For example, Gunther et al. (2007) developed a PCA model to detect abnormal process conditions resulting from different fault types. Both Teixeira et al. (2009) and Riley et al. (1999) constructed PLS calibration models for process monitoring purposes, with the former’s objective being to correlate fluorescence maps with viable cells and recombinant protein concentrations and the latter’s objective being to predict analyte concentrations in NIR spectra. Studies by Ferreira et al. (2007) and Cunha et al. (2002) involved developing both MPCA and MPLS models. Ferreira et al. (2007) constructed an MPCA model to diagnose process faults and an MPLS model to predict final product concentration whilst Cunha et al. (2002) developed an MPCA model to assess the quality of seed batches and an MPLS to infer the final productivity.

The work described is novel, as so far neither the study of batch behaviour relative to the process genealogy nor the application of the integrated wavelet transform-multiway PCA framework to such data has been conducted.

4.2 Process Description

The process of interest is an industrial pilot-plant involving the manufacture of monoclonal antibodies and considers both the seed and production stages of the cell culture process. Three types of measurement data were collected from the process; on-line traditional measurements including alkali addition rate and dissolved oxygen rate, on-line spectral measurements i.e. in-situ near infrared (NIR), and off-line sample analysis such as viable and total cell count and media components.

Figure 4.2 shows a schematic of the cell culture process used in this research for the production of a monoclonal antibody. Cell culture from the shake flask was transferred into seed bioreactors which contained growth media that provide nutrients for the multiplication of cells. During the seed stage, the cell lines undergo inoculum expansion to accumulate sufficient cell concentration for inoculation of the production stage. The cell culture is then transferred to a production bioreactor in which the medium is designed for the cells to continue to grow and to express the desired MAb. The cells in each seed and production stage are subcultured on a 3-4 day cycle to allow synthesis and product secretion. Subculture is a procedure that removes the medium in the bioreactor and transfers the cells from a previous culture into a fresh growth medium. It enables further propagation of the cell line.

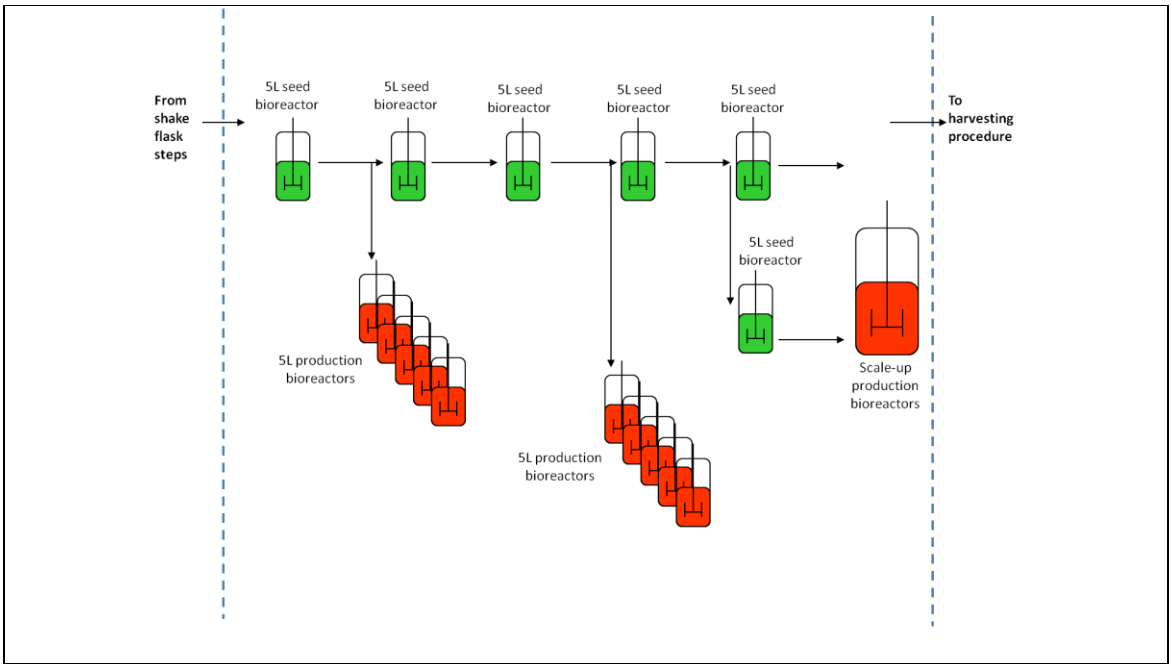


Figure 4.2 Schematic of monoclonal antibody cell culture process

There were seven seed and three production bioreactors selected for this study, where each operating bioreactor represents one process batch. Other batches were not selected because their relationship to other batch cannot be identified. All bioreactors had a working volume of 5 liter.

From a discussion with the engineer in charge of the MAb manufacturing process from CHO cell line, a specific protocol was developed to organize the batches into the genealogy shown in Figure 4.3. Firstly, batches cultured in the same bioreactor were identified and grouped together. Secondly, the batch with the lowest passage number and the earlier culture date was placed at the top of the genealogy. Passage number refers to the number of times the cells have been subcultured into a new vessel with the lowest number representing the earliest subculture. The rest of the batches in that group are then arranged according to the sequence of passage number and culture date. These procedures were repeated for another group of batches. As a result, the batches were organized into three groups which are named Family 1, Family 2 and Family 3. Family 1, the largest family in the genealogy comprises of four seed batches and one production batch. The first member of Family 2 and Family 3 were subcultured from Seed Batch 3 and Seed Batch 2 of Family 1 respectively. This results in a ‘cousin’ relationship between Family 1 and Family 2, and Family 1 and Family 3. Production batches of all families were subcultured at an identical sequence, resulting in the same passage number. It is hypothesised that the genealogy of the process batches may potentially affect batch behaviour.

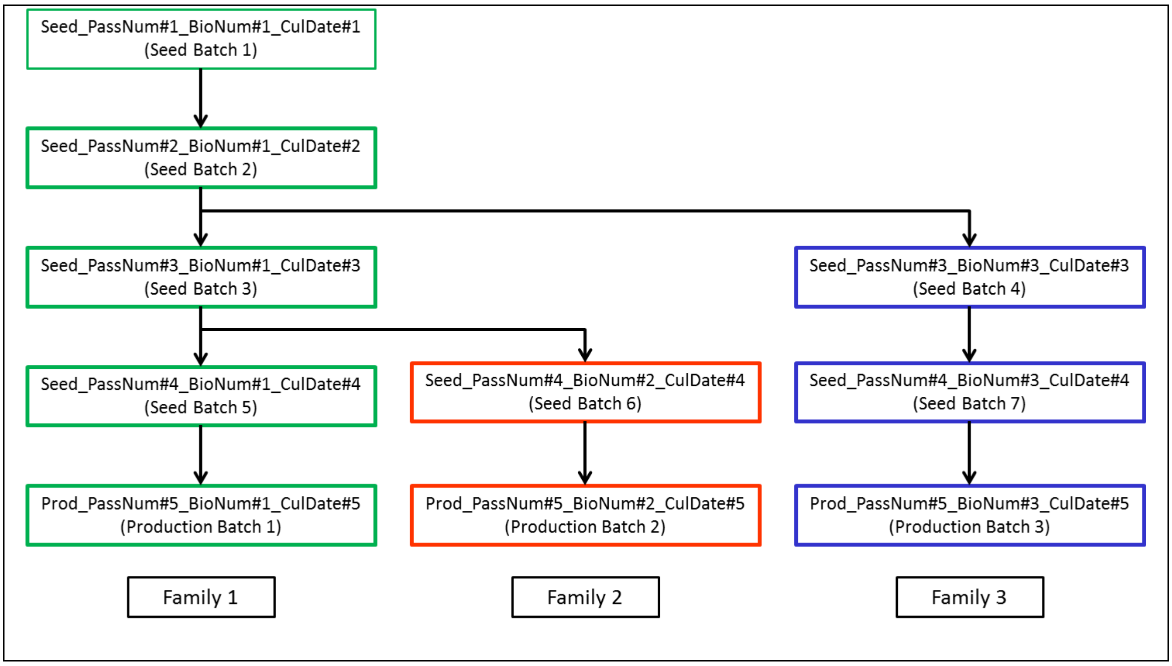

Seed_PassNum#1_BioNum#1_CulDate#1 (Seed Batch 1)
Seed_PassNum#2_BioNum#1_CulDate#2 (Seed Batch 2)
Seed_PassNum#3_BioNum#1_CulDate#3 (Seed Batch 3)
Seed_PassNum#3_BioNum#3_CulDate#3 (Seed Batch 4)
Seed_PassNum#4_BioNum#1_CulDate#4 (Seed Batch 5)
Seed_PassNum#4_BioNum#2_CulDate#4 (Seed Batch 6)
Seed_PassNum#4_BioNum#3_CulDate#4 (Seed Batch 7)
Prod_PassNum#5_BioNum#1_CulDate#5 (Production Batch 1)
Prod_PassNum#5_BioNum#2_CulDate#5 (Production Batch 2)
Prod_PassNum#5_BioNum#3_CulDate#5 (Production Batch 3)
Family 1
Family 2
Family 3

Figure 4.3 Genealogy of the process batches

A Bruker Matrix-F FT-NIR process spectrometer (Bruker Optics Ltd., Coventry, England, U.K.) with a transflectance probe was used to record the NIR spectra. The spectra lay between 800 and 2500 nm. The transflectance probe (precision sensing devices model 625) was operated at 1.0mm peripheral pathlength. The Matrix-F FT-NIR process spectrometer has an internal multiplexer for up to six fibre optics channels which allows for the simultaneous monitoring of multiple bioreactors run.

The NIR spectral measurements were recorded every 2.5 minutes in the seed and production bioreactors during the 3 and 4 day cycle. Over the duration of the batch, approximately 90 and 230 NIR spectral measurements were collected for seed and production batches respectively. Based on visual observation, Figure 4.4, changes in the NIR spectral measurements for a particular batch were slow due to the behaviour of the process. Cell culture process usually follows a characteristic growth pattern comprises of four phases: lag, exponential, stationary, and death as shown in Figure 4.5. When the cell is introduced into the growth medium, it grows slowly in the lag phase and then steadily increases in the exponential phase for a certain period. After that the subculture enters the stationary phase where the rate of growth of the cell slows down due to declining concentrations of nutrients and/or accumulating concentration of toxic substance. The subculture enters a death phase as the rate of growth declines. As a consequence of this behaviour, the NIR spectral measurements logged were averaged over every hour. The batch information is summarized in Table 4.1. Each batch was coded with a batch ID for subsequent analysis. The batch length between the seed batches did not differs much with three batches comprising 89 spectra, three with 92 spectra and one with 90 spectra whilst for the production batches, two batches had 236 spectra and one batch comprising 260 spectra.

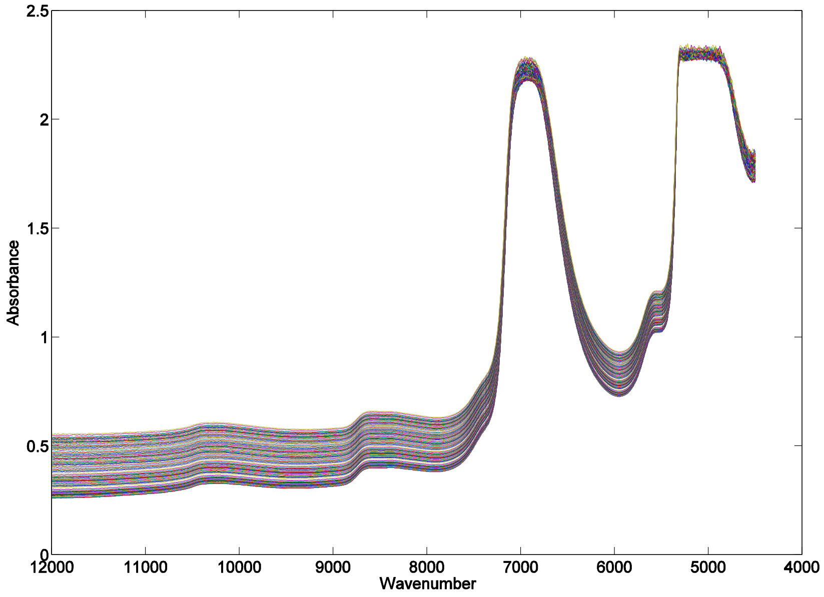


Figure 4.4 Evolution of raw spectra collected throughout Seed Batch 1

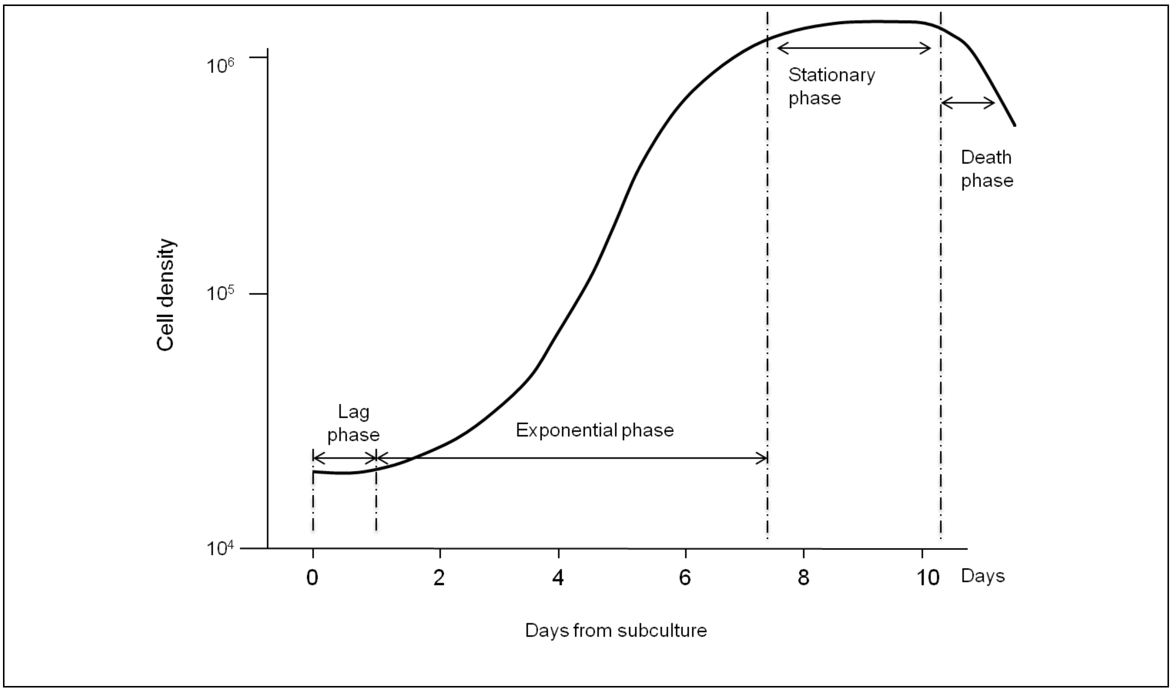


Figure 4.5 Growth culture showing the lag, exponential, stationary and death phases (Reference: Davis, 2011)

Table 4-1 Batch information for the MAb manufacturing process

Batch Name	Batch ID	Batch Length (number of spectra)
Seed Batch 1	S1	97
Seed Batch 2	S2	89
Seed Batch 3	S3	89
Seed Batch 4	S4	89
Seed Batch 5	S5	92
Seed Batch 6	S6	92
Seed Batch 7	S7	92
Production Batch 1	P1	236
Production Batch 2	P2	260
Production Batch 3	P3	236

4.3 Data Preprocessing

Prior to developing a fingerprint from the NIR spectra, the spectra need to be preprocessed. There are three stages of preprocessing: (1) NIR spectroscopy preprocessing which includes the removal of baseline shifts and first derivative smoothing (Section 4.4.1), (2) NIR batch unfolding (Section 4.5), and (3) alignment of NIR batch data (Section 4.6).

4.4 Near Infrared Spectroscopy: Literature Review on Its Application to Cell Culture and Multivariate Data Analysis

NIR spectroscopy is a methodology that uses the NIR region of the electromagnetic spectrum and lies between 700nm to 2500 nm, which is between the red band of visible light and the mid infrared (mid-IR) region. The absorption of electromagnetic radiation in the NIR region is caused by the combinations and overtones of the fundamental vibrations of molecules seen in the mid IR bands. Vibrations of –CH, –OH, –SH, and –NH bonds and their combination and overtones are observed in the NIR region (Roggo et al., 2007). The assignment of these hydrogen bonds to NIR bands is illustrated in Figure 4.6. As these bonds are essentially observed in all biological molecules, NIR spectroscopy becomes a theoretical means to measure the majority of the fundamental components in bioprocesses (Scarff et al., 2006).

Three measurement approaches are possible with NIR, off-line, at-line and on-line (Cervera et al., 2009). In off-line analysis, a sample is collected and analysed later, usually at a different location. Meanwhile, a sample is collected and analysed immediately in an at-line measurement approach. On the other hand, an on-line measurement eliminates the need for manual sample handling as the sample is analysed directly. There are two categories of on-line measurement, in-situ and ex-situ. A sampling device, for example a fiber-optic probe is positioned inside the bioreactor in in-situ whilst in ex-situ is where the sampling device is placed on a glass window inserted into the bioreactor.

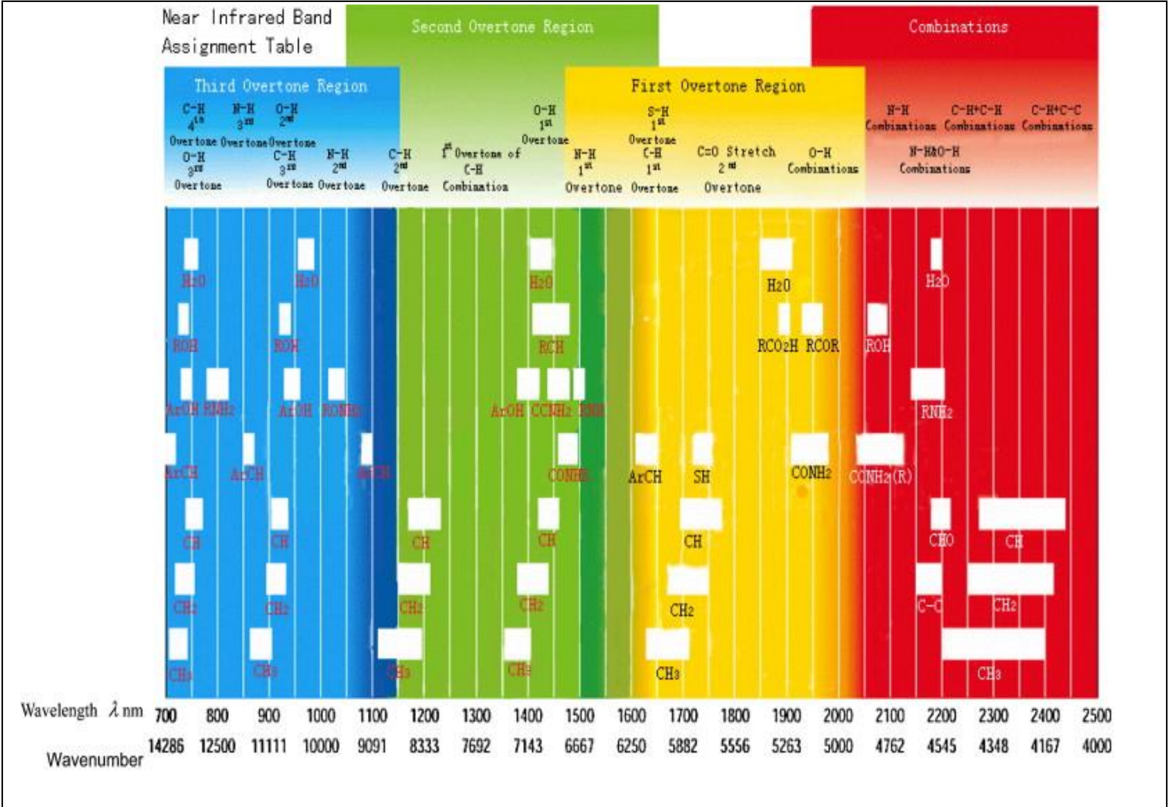


Figure 4.6 Near Infrared band assignment table (from Bruker Optics, Germany)

In recent years, the application of NIR spectroscopy in the area of cell culture system monitoring and control has expanded significantly (Scarff et al., 2006). This development was discussed in a review carried out by Cervera et al. (2009) where they showed that research in the field has advanced from simple systems with anaerobic conditions and/or low agitation to more complex systems with vigorous agitation and aeration. Configuration of probes has also progressed from at-line and ex situ to more challenging in situ implementation.

In the past decades, a number of studies (Arnold et al., 2003; Vaidyanathan et al., 2001b; Hagman and Sivertsson, 1998) have demonstrated that NIR spectroscopy is a reliable and robust tool for bioprocess monitoring. Arnold et al. (2003) and Hagman and

Sivertsson (1998) used near infrared spectroscopy as a technique to monitor the bioprocessing of mammalian cell cultures whilst Vaidyanathan et al. (2001b) monitored fungal bioprocesses. The analytes monitored in these three studies included glucose, lactate, and ammonia. Other than analytes, Hagman and Sivertsson (1998) also monitored the biomass and viability of the mammalian cells, and Vaidyanathan et al. (2001b) monitored biomass of the fungal. The calibration model developed in these studies demonstrated that NIR spectroscopy is a useful tool for bioprocess monitoring as it shows good predictive ability in terms of analyte concentrations.

Furthermore for bioprocesses monitoring and control diagnosis, NIR spectroscopy measurements offer many attractive features including real time measurements of a number of bioprocess variables simultaneously (Rodrigues et al., 2008; Arnold et al., 2003). Rodrigues et al. (2008) developed an NIR calibration model for the real time monitoring of the concentration of active pharmaceutical ingredient content, viscosity, nitrogen source and carbon source for an industrial fermentation process of an API. The feasibility of their proposed method was demonstrated through the satisfactory accuracy of the simultaneous monitoring of multi-parameter on on multiple fermentation bioreactors. Arnold et al. (2002) demonstrated that the NIR spectroscopy could be used to attain good predictive models for multiple-analytes through the application of an in-situ NIR to fed-batch industrial E. coli process.

Despite the advantages offered by NIR spectroscopy in bioprocess monitoring, its application is not yet routine due to two main reasons. The first challenge is related to the complexity of bioprocess datasets which is a consequence of the nature of microbial growth and product formation in batch cultivations (Clementschitsch and Bayer, 2006), and an extensive amount of data attained from a large number of process variables logged at high frequency (Gunther et al., 2007). The second challenge relates to the fact that raw NIR spectra have broad bands and baseline shifts leading to difficulties in interpreting the spectra (Roggo et al., 2007). Broad bands are due to the dispersion of the spectra, which occurs when radiation in the near infrared range excites overtone and combination vibrations in the sample material (Mark et al., 2010). This translates to absorbance at a specific wavenumber comprising of more than one chemical substance and, the chemical substance can also be absorbed at different wavenumber.

The combination of the aforementioned challenges necessitates the application of multivariate statistical techniques to extract real-time information from the NIR spectra. Challenges associated with complex bioprocess NIR datasets include multi-collinearity (McShane and Cote, 1998) and inherent correlation between different chemical substances (Petersen et al., 2010). These issues can be addressed through the application of multivariate statistical techniques (Zou et al., 2010). Furthermore, the combination of NIR spectroscopy and its analysis using multivariate statistical techniques has been proven to be effective in various fields of study including pharmaceutical (Luypaert et al., 2007), biodiesel (Balabin and Smirnov, 2011), agriculture (Sato, 1994), and wastewater treatment (Pons et al., 2004).

Three types of analysis performed on NIR using multivariate statistical techniques performance are mathematical pretreatment, classification and calibration modelling (Roggo et al., 2007). The cases of broad bands and poor baseline resolution discussed previously can be handled using mathematical pretreatment methods such as Savitzky-Golay derivatives, multiplicative scatter correction, standard normal variate and orthogonal signal correction (Cervera et al., 2009; Scarff et al., 2006). This matter is discussed further in Section 4.4.1. Classification methods categorize samples according to their spectra for the purpose of extraction of underlying trend in the dataset whereas calibration methods link the spectra absorbance value to quantifiable properties such as cell viability and concentration of analytes (for example, glucose and lactate).

Some of the commonly used classification and calibration methods are principal component analysis (PCA) (Tatavarti et al., 2005; Vaidyanathan et al., 2001a) and partial least square (PLS) (Petersen et al., 2010; Rhiel et al., 2002). Tatawarti et al. (2005) applied PCA to the NIR spectral data of a veterinary pharmaceutical drug to determine content uniformity, tablet crushing strength, and dissolution rate in the dosage samples. Differences in chemical composition and physical attributes of the drug samples were captured by principal component 1 and principal component 2 respectively. The application of PCA by Vaidyanathan et al. (2001a) was performed on NIR spectral data from culture samples of antibiotic production to monitor variations in the bioprocess through changes in the NIR spectral data. Scores and loadings of the PCA were able to capture process related changes including variations in medium composition and reactor configuration. Both Petersen et al. (2010) and Rhiel et al. (2002) built PLS calibration models to predict the concentration of analytes such as glucose and ammonia in NIR spectral datasets from different types of cell cultures. The results of Petersen et al. (2010) demonstrated that the PLS algorithm can satisfactorily predict glucose but not ammonia from on-line NIR spectroscopy of filamentous fermentation media whereas Rhiel et al.'s (2002) study showed that the PLS algorithm can selectively extract analyte specific information from the NIR spectroscopy of animal cell culture.

The main objective of this case study is to classify the spectra to allow the fingerprinting of the process batches and to link the fingerprint to the process genealogy, and ultimately relate knowledge acquired from this relationship to the off-line quality process measurements through the implementation of the combination of discrete wavelet transform and multivariate statistical analysis.

4.4.1 *Preprocessing of NIR Spectroscopy*

The raw NIR spectra are subjected to broad bands, baseline shifts and light scattering. These aspects can be addressed by pre-treating the raw NIR spectra with the Savitzky-Golay smoothing algorithm. Previous research has shown that the application of multiplicative scatter correction (Roychoudhury et al., 2007) and standard normal variate (Rodrigues et al., 2008) can be used to reduce light scattering effects whilst first derivatives (Ferreira et al., 2005) and second derivatives (Arnold et al., 2003) can remove baseline shifts. Roychoudhury et al. (2007) applied multiplicative scatter correction whereas Rodrigues et al. (2008) applied the standard normal variate as a pre-processing technique to handle the effects of light scattering in the NIR spectra collected from cell culture during the manufacture of monoclonal antibodies and antibiotic respectively. To remove baseline shifts, Ferreira et al. (2005) identified that the most appropriate result for their NIR spectral data generated from an industrial fermentation was first derivatives, whilst Arnold et al. (2003) performed second derivatives on the NIR spectra from the bioprocessing of industrial *E. coli.*

Second derivatives not only remove baseline shifts but also deconvolute the overlapping peaks seen in the raw spectra, Figure 4.8 shows the deconvolution of the overlapping peaks. However, second derivatives amplify the noise in the NIR spectra because they calculate the rate of change of the signal hence affecting the noise-to-signal ratio (Candolfi et al., 1999).

Figure 4.4 provides an example of the raw NIR spectra collected over the duration of a batch. It is evident from the illustration of the first derivatives and second derivatives of the NIR spectra, Figure 4.7 and Figure 4.8 respectively, that second derivatives amplify the noise in the raw spectra. Consequently for the purpose of this case study, the raw NIR spectra were corrected using first derivatives. This resolves baseline shifts while minimizing noise resulting from the derivatives. A window size of 15 and a second order polynomial were employed by the person in charge of handling the initial preprocessing of the raw spectral data. This pre-treatment replaces the raw NIR spectra with a set of data which shows clear absorbance peaks. As seen in Figure 4.7, two absorbance peaks were identified at the following wavenumber ranges: 8000 - 7000 and 6000 - 5000. All spectra manipulations were performed using PLS Toolbox, MATLAB 8.1.

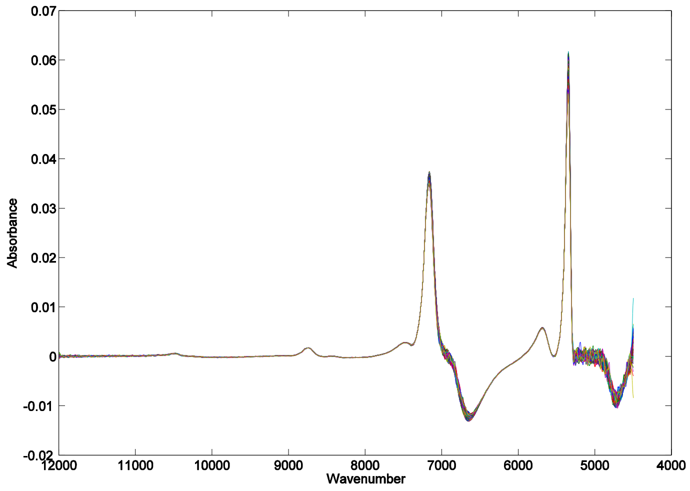


Figure 4.7 First derivative NIR spectra of Seed Batch 1

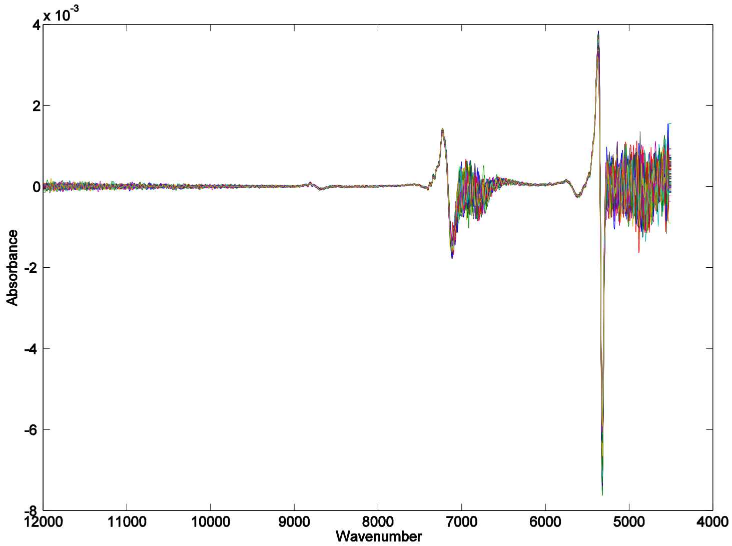


Figure 4.8 Second derivative NIR spectra of Seed Batch 1

4.5 NIR Batch Unfolding

In Section 3.4.1, an overview of batch analysis was provided. The concept is now transferred to NIR spectra logged throughout the duration of a particular batch. For batch NIR spectra the three dimensions are wavenumber, spectra recorded through the duration of the batch and batch number. Therefore, for a data array $\underline{\boldsymbol{Y}}$ of batches comprising NIR spectral measurements, the convention adopted is batches (I) x wavenumber (J) x number of spectra (K), Figure 4.9.

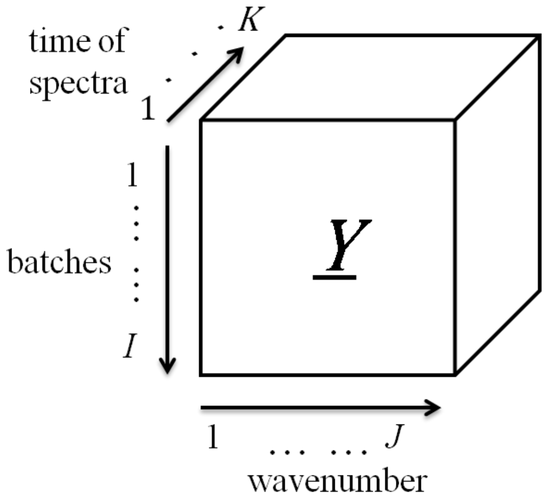


Figure 4.9 Batches of NIR spectroscopy data with respect to number of spectra

The three-dimensional data array $\underline{\boldsymbol{Y}}$ requires to be unfolded prior to subsequent analysis. In this study, the behaviour of the batches within a particular wavenumber is of interest hence variability in batch-to-batch behaviour for a specific wavenumber region is considered. Thus the three-dimensional data array $\underline{\boldsymbol{Y}}$ requires to be unfolded to allow these investigations to be performed. The unfolded matrix, $\boldsymbol{E}$ (K x IJ) is shown in Figure 4.10. This way of unfolding considers all batches as one object hence enables the possibility of comparing between spectra from different batch runs throughout the whole NIR spectrum or for selected wavenumber regions. The individual spectra information is captured by the principal component scores thereby enabling the study of similarities and differences between batches within a wavenumber region.

Loadings of the matrix $\underline{\boldsymbol{Y}}$ is the weight of the wavenumbers in terms of defining the individual principal component.

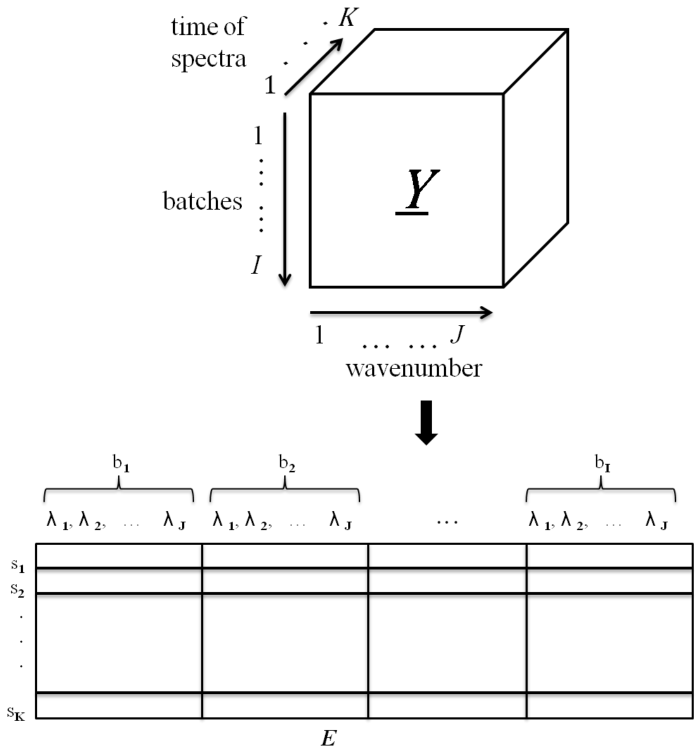


Figure 4.10 Three-dimensional NIR data array $\underline{\boldsymbol{Y}}$ and its unfolded matrix $\boldsymbol{E}$

4.6 Alignment of NIR Batch Data

By adopting the unfolding approach proposed in the previous section the number of spectra in each batch needs to be equal. As observed from Table 4-1 and illustrated in Figure 4.11, the batches from the manufacture of MAb were from seed and production and hence were of unequal duration. This is a further challenge in terms of the analysis performed in this study.

A number of methods have been proposed to resolve the issue of unequal batch length, cut to minimum length, multivariate dynamic time warping (Ramaker et al., 2003; Kassidas et al., 1998; Gollmer and Posten, 1996) and the use of an indicator variable (Nomikos and MacGregor, 1994). The batch processes in the aforementioned studies comprise process variables measurements whereas in this study the batch processes under investigation comprise NIR spectra. Of the three techniques, multivariate dynamic time warping and the use of an indicator variable are not applicable due to loose biological structure and no surrogate variable available. There are no references in the literature proposing a method to equalize the length of batches comprising NIR spectral measurements. A technique to address this problem is proposed in Section 4.6.2.

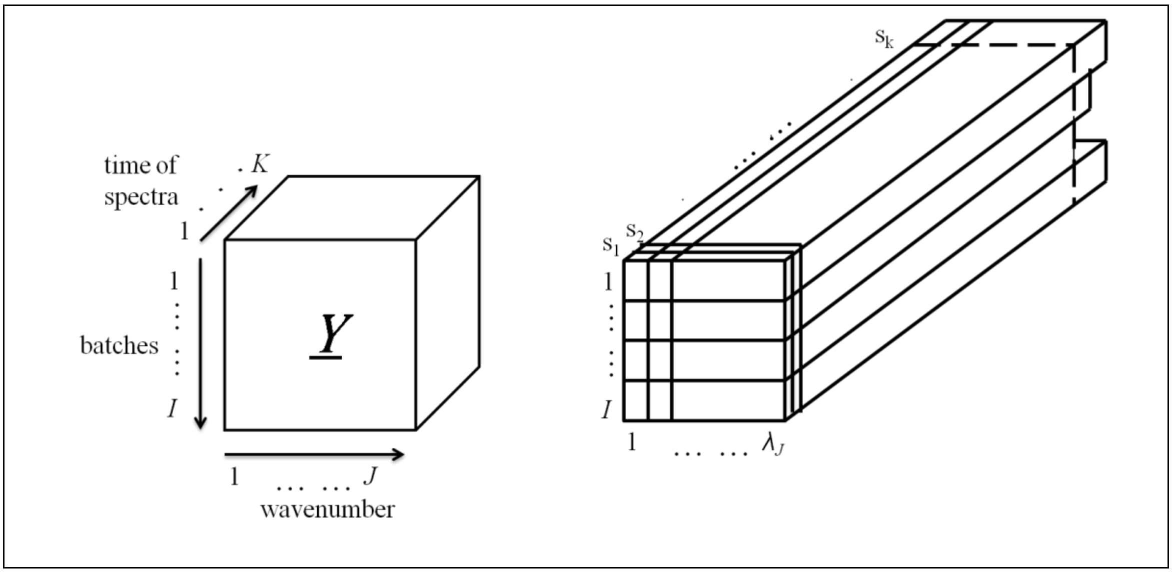


Figure 4.11 Illustration of unequal batch length in MAb manufacturing

4.6.1 *Cutting to Minimum Length*

With respect to batch NIR spectral measurements, cutting to minimum length means reducing the number of spectra in all batches to the selected minimum number of spectra, s_k. This technique may impact on the resulting fingerprinting as important information may be contained towards the end of batches runs. This technique is appropriate when the difference in batch length is small as successfully demonstrated by Gunther et al. (2007) in their fault detection and diagnosis analysis of data from an industrial pilot plant cell culture. In this case study, batch durations differed by less than 10%, therefore they selected the shortest batch duration and used this as a reference for other batches.

4.6.2 *Re-sampling Spectra Count*

The aim of the proposed approach is to include spectra throughout the duration of a process. The rationale is, if a batch is cut to a certain length to comply with a reference batch, the information towards the end of the process will be removed and this may be of interest. The procedure for this approach is to increase the step count or sampling rate of the long batches so that the number of spectra in that batch is relatively closer to the number of spectra in the reference batch. Details of this procedure are explained in the following example in which Seed Batch 3 (S3) is used as a reference batch and Production Batch 1 (P1) as a batch that needs re-sampling. As stated in Table 4-1, S3 and P1 have 89 and 236 NIR spectra respectively. Therefore, P1 needs to be re-sampled so that the number of NIR spectra is as close as possible to 89. To accomplish this, every other NIR spectrum was selected starting from the first one. This reduced the NIR spectra to 118. Figure 4.12 describes monoclonal antibodies cell culture growth phases relative to the NIR spectra collected throughout the duration of S3 and P1. It can be seen that re-sampling of the spectra takes account of the spectra throughout the duration of the batch whilst reducing the batch length. Once re-sampled, P1 is cut to a minimum length of 89 NIR spectra prior to developing the model using the matrix unfolding approach discussed in Section 4.5. Even though the re-sampled batch data still needs to be reduced, this only resulted in a few NIR spectra being removed in contrast to cutting to minimum length from original batch. If P1 is simply cut to synchronize with the length of S3, NIR spectra from the middle of the growth phase to the end of the batch run will be discarded.

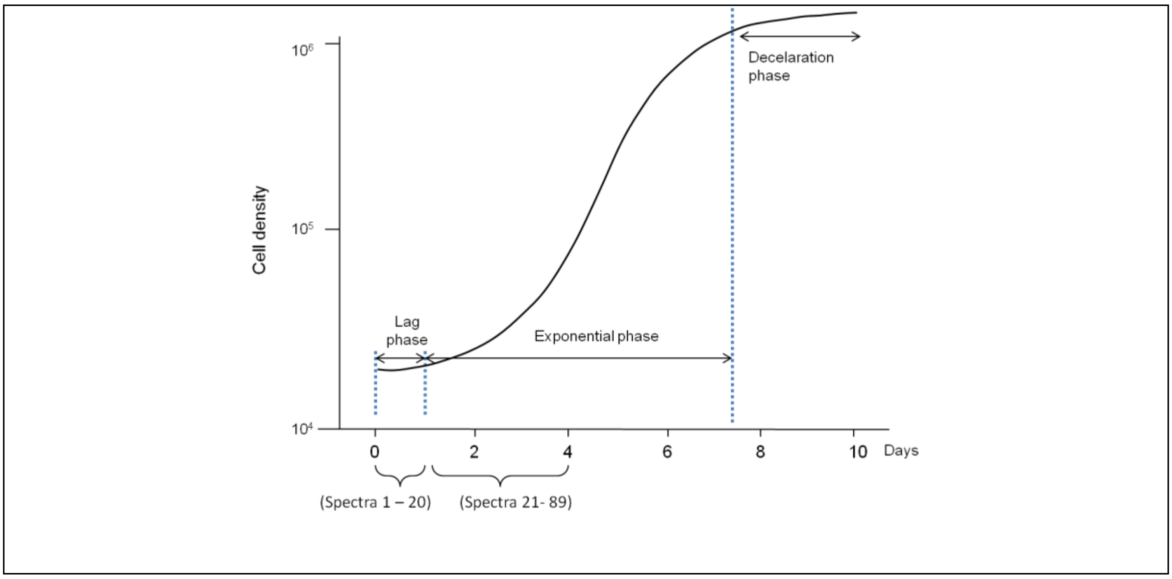


Figure 4.12 Cell culture growth phase relative to NIR spectra

4.7 Development of Process Representation

The focus of the analysis is to develop an integrated discrete wavelet transform-multiway PCA model to extract underlying behaviour and distinctive characteristics of the process batches, seed and production, i.e. to fingerprint the process. The ultimate goal is to link the behaviour and characteristics of the seed and production batches to the genealogy.

Prior to the development of the integrated discrete wavelet transform-multiway PCA model, an initial study was conducted. In the initial study, a range of approaches were investigated to determine the most appropriate approach to accomplish the objectives of this study. Results from this study formed the basis of the proposed approach. Details of the initial study are discussed in each development stage of the process representation.

Seven seed batches were included in the development of the fingerprint. The representation was then applied to the production batches. Differences between the batches were revealed via scores plots and were further investigated through the loading of the wavelet features to the scores. Having preprocessed the raw NIR spectra using first derivative, the next step was to develop a fingerprinting representation. Fingerprinting in the context of this thesis is defined as the underlying pattern of the data represented by the PCA metrics. A schematic of the development of the process representation is illustrated in Figure 4.13.

The development of the process representation comprises three key stages: investigation and implementation of the discrete wavelet transform, application of feature extraction, and projection of the selected features from the seed and production batches onto the PCA space. The first stage, the discrete wavelet transform, involves determining the appropriate wavelet family and decomposition level to be applied to the NIR spectra of the batches to transform them into different frequency components. Details of the discrete wavelet transform implementation to the NIR spectra is discussed in Section 4.9 with the selection of wavenumber intervals outlined in Section 4.9.2. The second stage describes the feature extraction step and is discussed in Section 4.10. Finally the third stage focus on the superposition of the production batch onto the developed representation and is discussed Section 4.11. Each stage starts with an initial study with the aim of selecting the optimal approach for that stage. Results from this study are discussed along with the subsequent development stages.

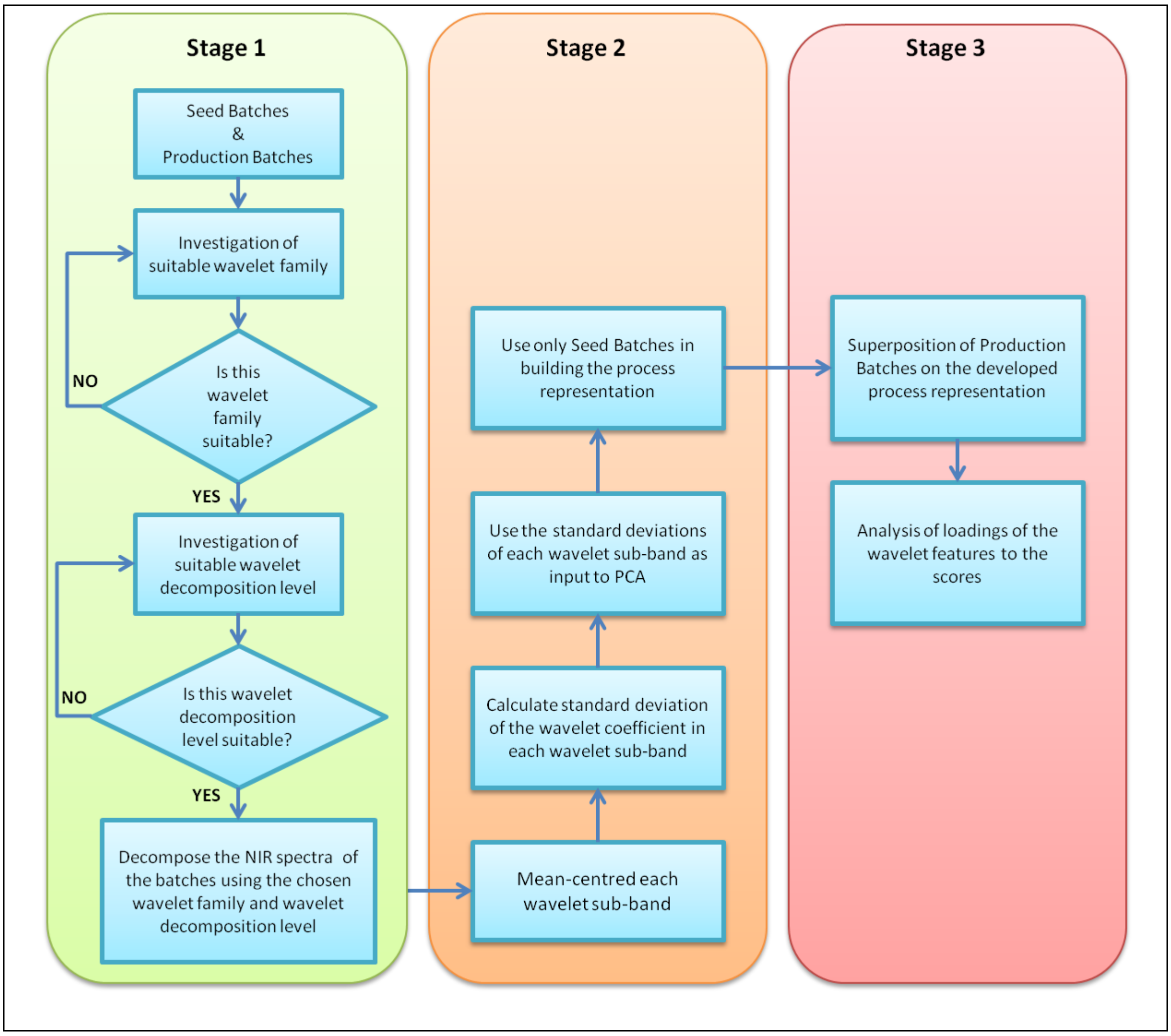

Stage 1
Seed Batches & Production Batches
Investigation of suitable wavelet family
Is this wavelet family suitable?
NO
YES
Investigation of suitable wavelet decomposition level
Is this wavelet decomposition level suitable?
NO
YES
Decompose the NIR spectra of the batches using the chosen wavelet family and wavelet decomposition level
Stage 2
Mean-centred each wavelet sub-band
Calculate standard deviation of the wavelet coefficient in each wavelet sub-band
Use the standard deviations of each wavelet sub-band as input to PCA
Use only Seed Batches in building the process representation
Stage 3
Superposition of Production Batches on the developed process representation
Analysis of loadings of the wavelet features to the scores

Figure 4.13 A schematic diagram of the development of process representation of the NIR data

4.8 Motivation for Integrated Modelling

The implementation of the combination of the discrete wavelet transform and MPCA was driven by results from the initial study. In the initial study, MPCA was applied to the first derivative NIR spectra (all wavenumbers) and it is observed that batches (seed and production) from the same family were clustered as shown in Figure 4.14. It is important to note that the goal of developing the process representation is to investigate whether process understanding may be improved by drawing on information from the process genealogy. Thus, this finding is agreeable as batches from the same family were expected to display similarities in their behaviour which is reflected in the bivariate scores plot. Figure 4.15 shows closer observation of the seed and production batches for Family 1. It is observed that the top batch in the genealogy is located at the right end whilst the bottom batch in the genealogy is located at the left end. It is observed that batch position in the cluster is sequential according to their position in the genealogy. The parent (S1) on the right end, the child (P1) on the left end, and the other batches following through. On the other hand, this finding is unsatisfactory as the correlation between families in the genealogy is not evident, for example S6 of Family 2 and S4 of Family 3 were subcultured from S3 and S2 respectively.

PC1 and PC2 explain approximately 99% of the variance whereas PC3 and PC4 explain of the order of 0.005%. The bivariate scores plot of PC3 and PC4 is shown in Figure 4.16 where the small magnitude of the variance explained translate into the batches overlapping and non-evident clustering. This preliminary result suggests that meaningful information is being masked in the NIR spectra hence the integration approach was proposed.

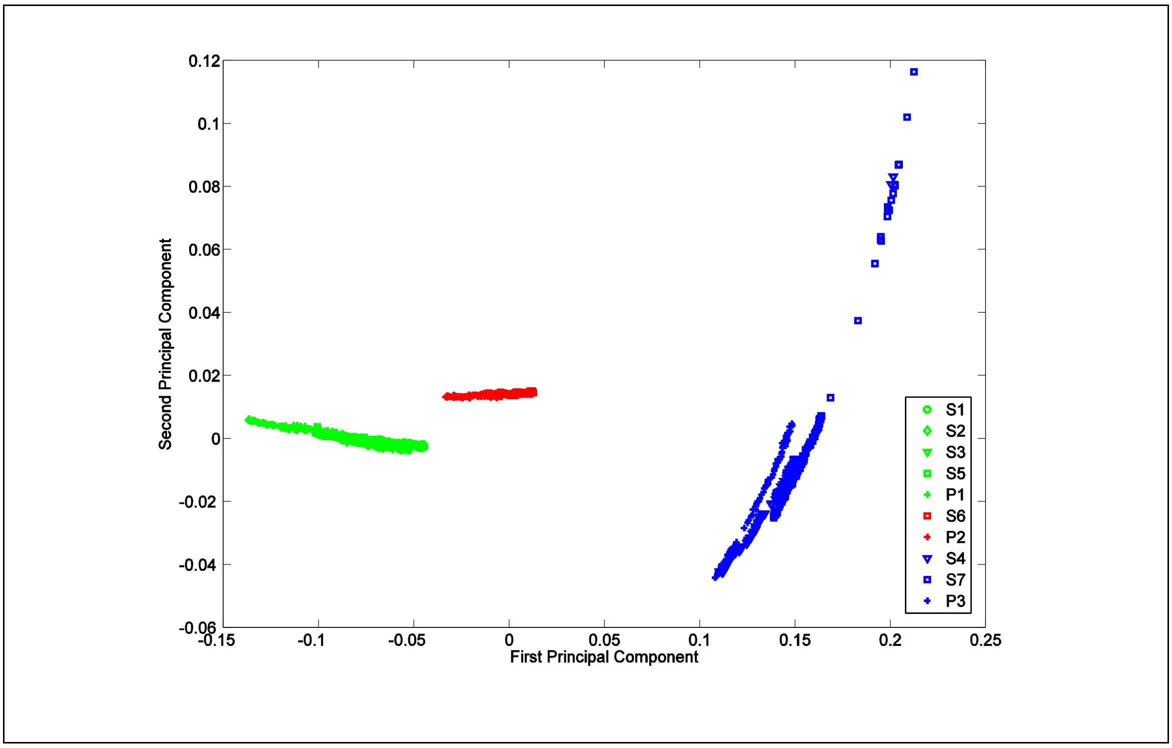


Figure 4.14 Bivariate scores plot of PC1 vs. PC2 for all wavenumbers

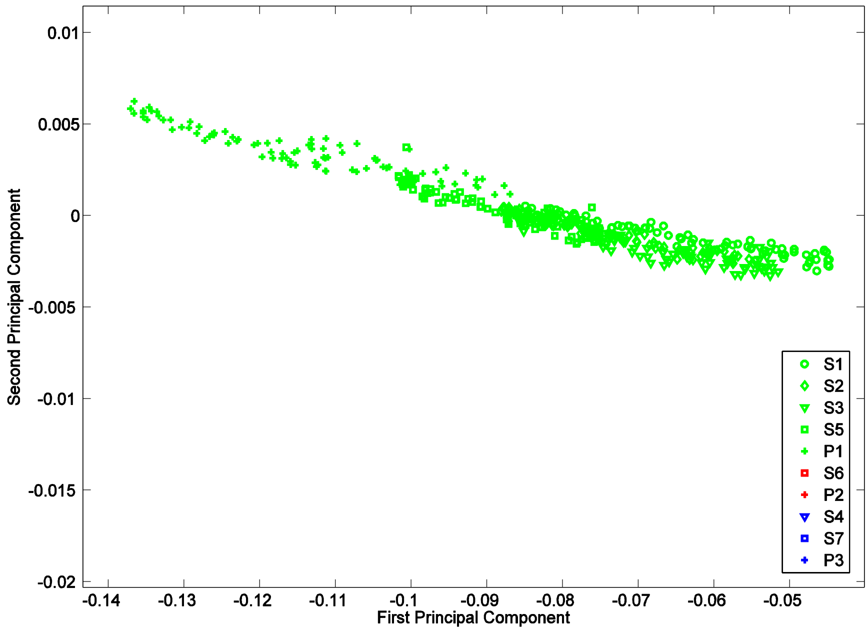


Figure 4.15 Zooming in on Family 1 on Figure 4.14

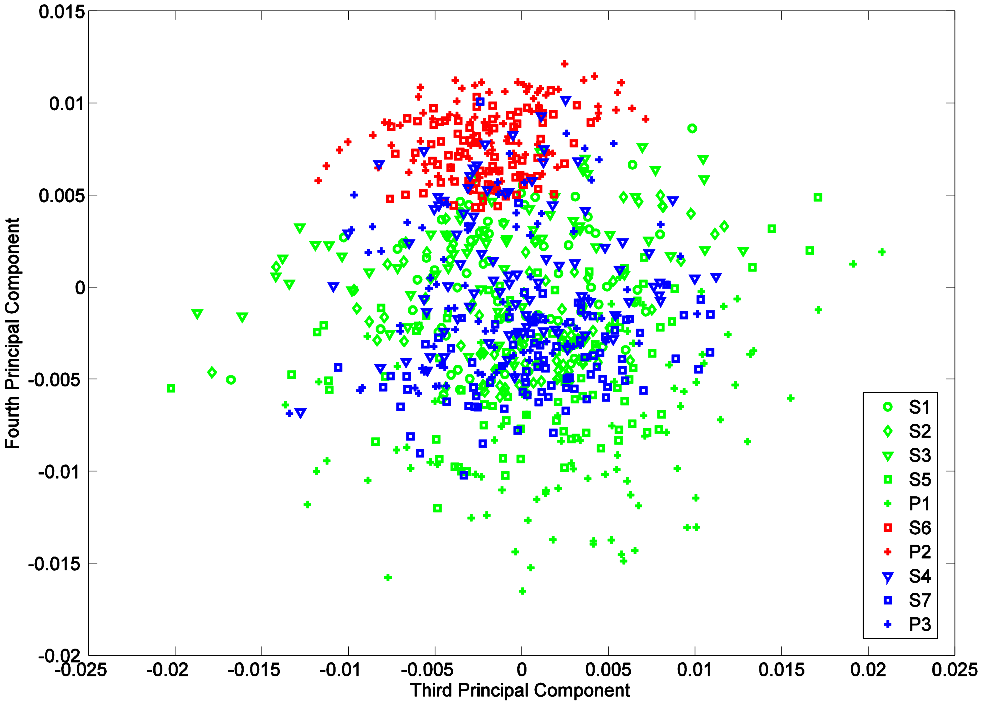


Figure 4.16 Bivariate scores plot of PC3 vs. PC4 for all wavenumbers

It has been demonstrated in a significant number of papers that multivariate statistical techniques can be very useful in extracting valuable information and demonstrating correlation structures in a dataset thereby enhancing process knowledge (Kirdar et al., 2007). Through scores and loading plots, Kirdar et al. (2007) successfully extracted process knowledge from cell culture process data comprising small scale (2 litres) and large scale (2000 litres) batches to assess scale up and comparability of the process. In terms of PCA applications, which focused on investigating NIR spectra generated from cell cultures to help identify bioprocess variations, Vaidyanathan et al. (2001a) concluded that PCA can identify variations in a bioprocess relative to changes in spectral information and can assess the structure of the data in terms of differences within and between process.

Furthermore, PCA has been proven to be an effective method for identifying abnormal process conditions in both continuous and batch industrial processes (Kourti, 2005). A process representation based on MPCA has been demonstrated to successfully distinguish between high and low productivity of industrial seed fermentation batches (Cunha et al., 2002). Their model was developed to investigate the benefits of including seed quality information into data-based models for final productivity estimation. Cunha et al. (2002) took two different approaches to handling different batch lengths in their production-scale seed data. Firstly, the on-line data from the first 24-hours of the cultivation were discarded, followed by cutting to the length of the shortest batch. In the second approach, they subsampled the on-line data at hourly intervals in reverse order from the end of the cultivation. Gunther et al. (2007) successfully developed a process representation based on PCA to detect and diagnose abnormal process conditions in an industrial fed-batch cell culture process. Three fault types, irregular thermal heating, elevated dissolved oxygen values, and large variations in agitation were detected by their model.

In another development, Luo and Chen (2007), and Teixeira et al. (2009) investigated a combined PCA and PLS approach. Despite a wide range of reactor operating conditions, the combined PCA and PLS model of Teixeira et al. (2009) accurately estimated the concentration of viable cells and the concentration of recombinant protein in mammalian cell cultures. Studies by Roggo et al. (2004) used PCA to compare pharmaceutical products produced at different manufacturing sites. Their PCA score plots demonstrated that spectra recorded at different manufacturing sites are statistically different due to differences in physical aspects including particle size and aspects of surface or density, and moisture content of the tablets.

The application of PCA however, faces two limitations, poor discriminatory power and large computational load, as identified by Feng et al. (2000) in a study pertaining to human face recognition. Hence, the combination of the wavelet transform and PCA approach was proposed. A study conducted by Shao et al. (1999) used the wavelet coefficients as inputs into a non-linear principal component analysis algorithm. The focus of the study was to implement non-linear PCA for process monitoring and fault detection on industrial process data. The application of the wavelet transform for identification purposes has also been successfully demonstrated for the detection of infrared spectra of benzenes (Bos and Vrielink, 1994). A study by Tian et al. (2005) proposed integrating wavelet coefficients with principal component analysis (PCA) for better extraction of defect information for pulsed eddy current non-destructive testing. The focus was on classifying and quantifying the defect signals extracted from the pulsed eddy current signals.

Other applications of the combination of the wavelet transform and PCA approach include de Bianchi et al. (2006) and Borah et al. (2007) where the wavelet transformation was used to consider extracted image features, in the area of facial expressions and tea granules respectively. The latter used a combination of the wavelet transform and principal component analysis to classify the tea granules. Even though both adopted the basic principles of integrating the wavelet transform with PCA their approaches differed. de Bianchi et al. (2006) in their research applied the discrete wavelet transform to images and used the generated wavelet coefficients to create matrices that contained key features of the original data. Subsequently, principal component analysis was applied to the matrices to find the projections of the original data. Meanwhile Borah et al. (2007) used the fast wavelet transform (Daubechies family) to decompose the images of the tea granules into sub-band images. Statistical features including mean, variance, entropy and energy, of the sub-band images were calculated. In this research, principal component analysis was used as a visualization method to differentiate between the statistical features.

Apart from quantification and classification, the integration of the wavelet transform and principal component analysis may also result in a reduction in the size of the dataset as suggested by Trygg et al. (2001). It was shown in this study that the NIR dataset collected from the on-line monitoring of wood chips was reduced 70 times from its original size.

Based on the literature discussed above, two hypotheses are proposed. The first is that by integrating the wavelet transform and principal component analysis, important information can be extracted and analysed from a large set of data. Secondly, the versatility offered by the integration of the wavelet transform and principal component analysis draws on the abilities of the wavelet transform to zoom in and zoom out of any part of the signal and also reduce data size. Furthermore the ability of principal component analysis to reduce the number of original variables into a smaller number of uncorrelated variables is utilized. The next logical step in this research is to apply the concept to analyse the NIR spectra from the seed and production batches.

4.9 Discrete Wavelet Transform

Within this section, two investigations are reported. The first relates to the type of wavelet family to be used in transforming the NIR spectra and the second concerns the selection of the decomposition level.

4.9.1 *Selection of Wavelet Family and Decomposition Level*

The two-dimensional NIR data matrix, $\boldsymbol{E}$, shown in Figure 4.10 was first analysed using the discrete wavelet transform decomposition. Prior to the analysis, the first task was to determine the appropriate mother wavelet and the number of decomposition levels to be used for the analysis. There is no defined approach for selecting either factor when applying the discrete wavelet transform.

However, Mallat (1998) emphasized selecting a mother wavelet that produces a maximum number of wavelet coefficients that are close to zero and therefore can efficiently approximate signals with a few non-zero coefficients. Properties of a mother wavelet including the number of vanishing moments and the support size of the wavelet determine the number of 'close to zero' wavelet coefficients of a signal. The number of vanishing moments is a criterion of wavelets that enables a wavelet to suppress a polynomial. The need to suppress a signal in the application of the wavelet transform is so that the remainder of the signal may be highlighted. As explained by Strang & Nguyen (1997) the number of vanishing moments in a wavelet determines the degree of accuracy of the wavelet in representing a signal. Also, numerous vanishing moments helps eliminate background effects in a signal (Chen et al., 2004). As for the support size, a minimal support of a mother wavelet means fewer large magnitude wavelet coefficients (Chau et al., 2004).

The Daubechies mother wavelets have a minimum size support for a given number of vanishing moments which effectively suppresses low degree polynomials present in NIR spectra (Esteban-Diez et al., 2004). Therefore in determining the appropriate mother wavelet for the subsequent analysis, two members of the Daubechies family were investigated: db3 and db5, where 'db' represents the wavelet name and the number next to the wavelet name represents the number of vanishing moments for the subclass of wavelet. Daubechies family members have been applied previously to NIR spectra, for example db4 and db6 (Cai et al., 2008), db4 (Esteban-Diez et al., 2004) , and db4, db6, db8,db10 and db12 (Bos and Vrielink, 1994). Comparing the results from wavelet decomposition performed with db3 level 3 (Figure 4.17) and db5 level 3 (Figure 4.18), it is observed that the high amplitude of wavelet coefficients in the sub-band D3 was evident in db5. Based on this db5 was deemed to be the more suitable of the two. Wavelet decomposition performed with db3 level 5 and db3 level 10 can be referred to in Appendix A.

The next stage was to determine the level of decomposition of the discrete wavelet transform. The maximum level of decomposition to apply depends on the total points in a spectrum, since each decomposition level involves a down-sampling by 2. Previous literature including Bruce et al. (2002) and Jouan-Rimbaud et al. (1997) applied different decomposition levels. Bruce et al. (2002) selected seven levels of decomposition with a range of mother wavelet including Haar, Daubechies, Biorthogonal, Coiflet and Symlet to extract features from agricultural hyperspectral data. The study showed that the extracted wavelet coefficients resulted in a significant increase in classification accuracy in contrast to techniques such as best spectral band selection and traditional PCA. Meanwhile, Jouan-Rimbaud et al. (1997) applied four levels of decomposition with two Daubechies family (db8 and db9) on NIR spectra datasets. Elimination of irrelevant signal component and noise from the original signal prior to multivariate calibration resulted in an improved PLS model.

A decomposition of level 5 was selected following the analysis of each batch using decomposition levels 3, 5 and 10. An example of the analysis on a single spectrum from Seed Batch 1 is shown in Figure 4.18, Figure 4.19, and Figure 4.20 respectively. Through visual comparison between Figure 4.18 and Figure 4.19, level 3 was not chosen because it only extracted a limited amount of the high frequency components from the original spectrum. Furthermore it was observed in Figure 4.20 that from A6 to A10 the wavelet approximation coefficients show little resemblance to the original spectra. Hence the decision not to use level ten since the aim of performing the wavelet decomposition is to efficiently represent the signal.

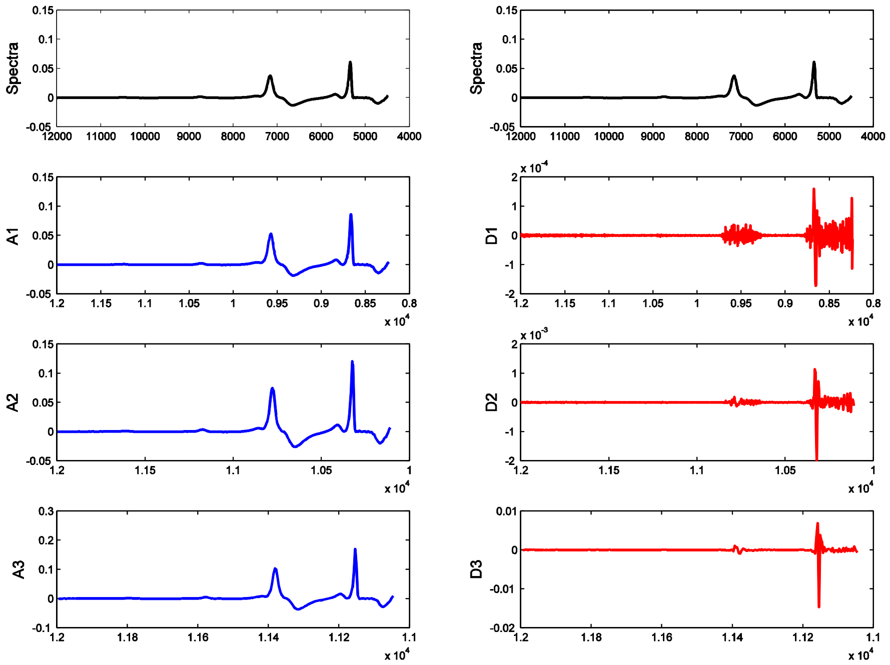


Figure 4.17 Wavelet decomposition using db3 with 3 levels of decomposition

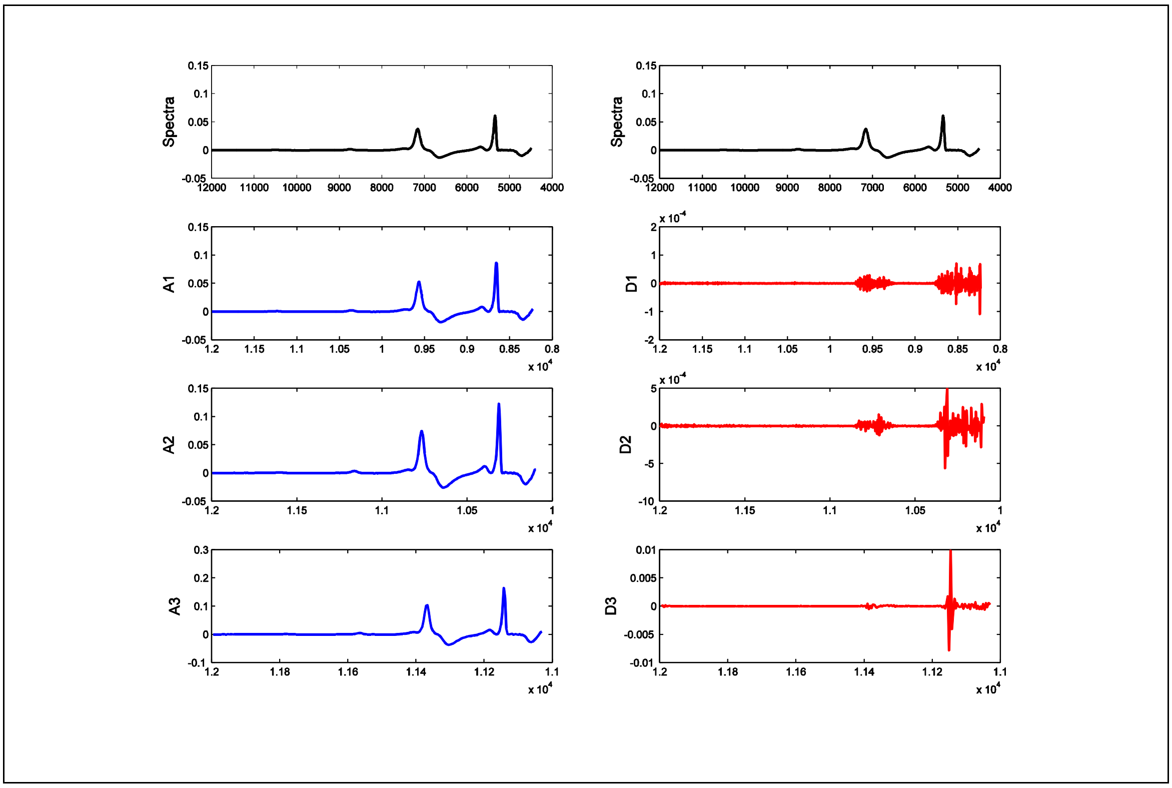


Figure 4.18 Wavelet decomposition using db5 with 3 levels of decomposition

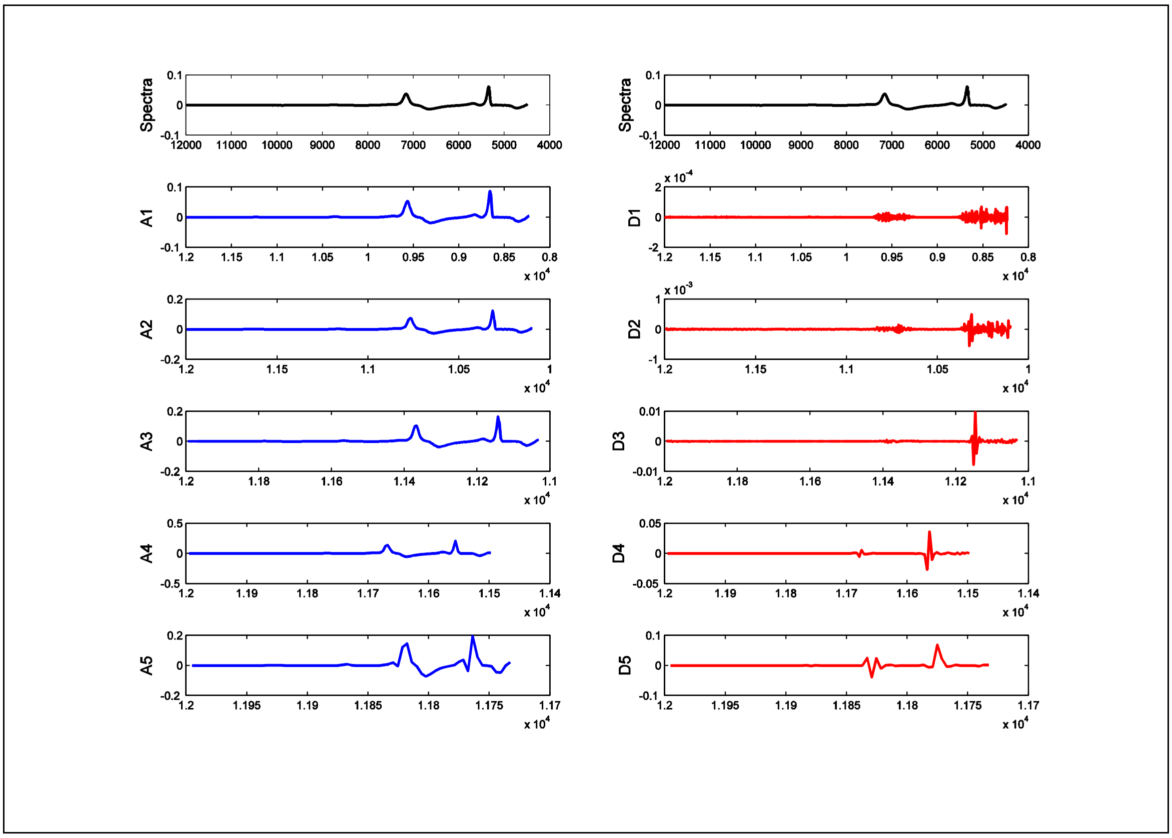


Figure 4.19 Wavelet decomposition using db5 with 5 levels of decomposition

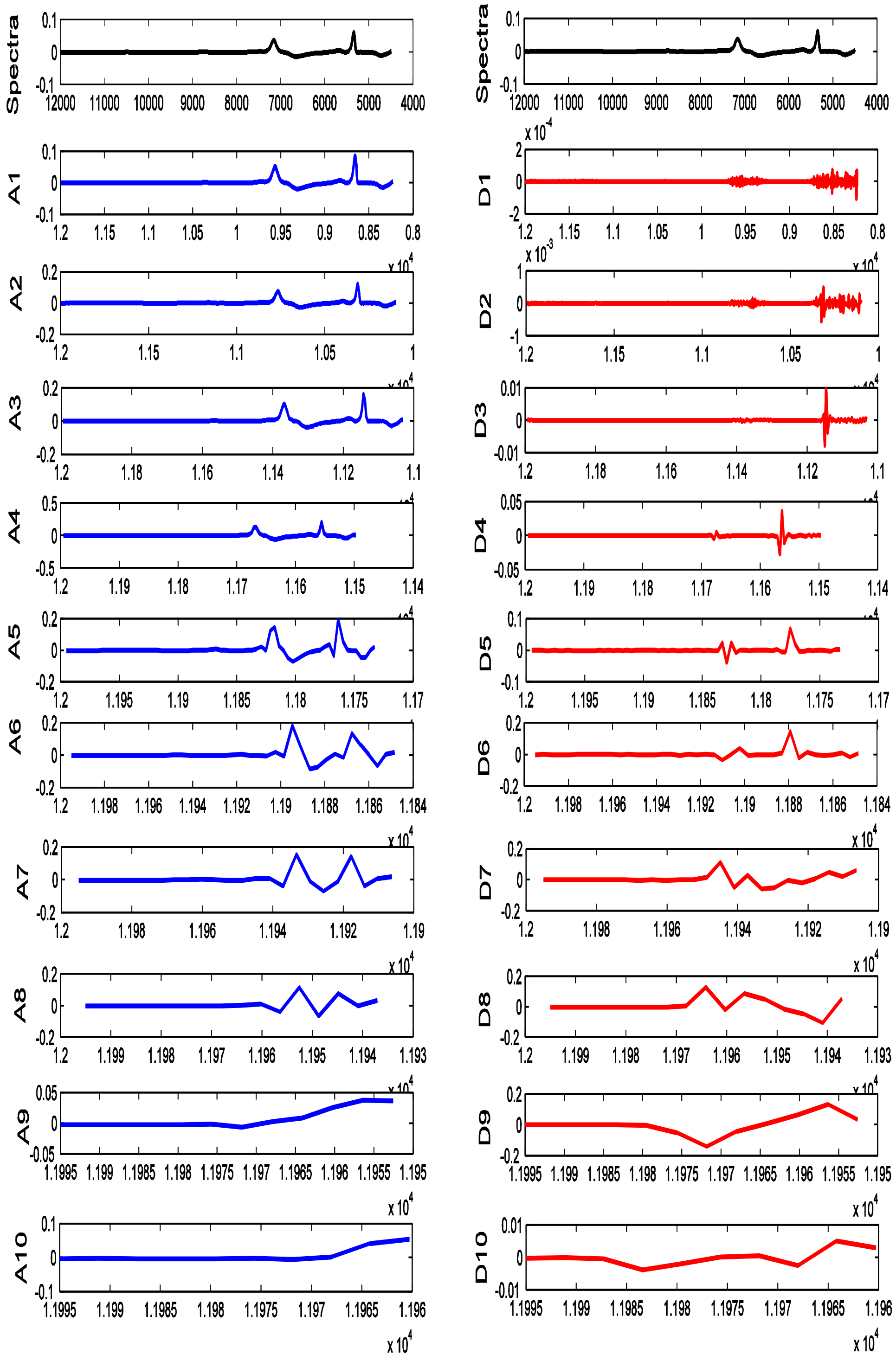


Figure 4.20 Wavelet decomposition using db5 with 10 levels of decomposition

4.9.2 *Wavenumber Interval Selection in Spectral Data Analysis*

Following the series of investigations in terms of the application of the discrete wavelet transform to the NIR spectra, the next step was to select appropriate wavenumber regions. Only selected regions of the spectrum were used to build the fingerprint representation because information of interest lay in a relatively small number of spectral regions. Essentially, the remainder of the spectral regions that are deemed uninformative were eliminated because they may lead to a degradation in the result. Namkung et al. (2008) built a PLS model on a selected spectral range for the analysis of etchant solutions and proved that the prediction selectivity was significantly degraded in a condition where whole spectral range was used for the model.

Glucose plays a vital role in the success of MAb manufacture as it is the primary carbon and energy source for mammalian cells. Arnold et al. (2003) suggested that monitoring the on-line NIR spectra measurements of glucose levels within animal cell culture systems may avoid premature cell death, thus leading to higher productivity. This indicates that the glucose spectral regions contain a wealth of information that may help determine the behaviour of the batches relative to their genealogy. The spectral regions used in this study were selected based on glucose (CH) overtones regions. These regions were first CH overtone (1600 – 1800 nm), second CH overtone (1100 – 1250 nm), third CH overtone (850 – 950 nm), and fourth CH overtone (700 – 800 nm). The spectral regions representing other important analytes including ammonia and lactate were not investigated because they are by-products of the process and this study aims to focus on the main analyte that contribute to growth in the cell culture.

The selection of wavenumber intervals in NIR spectral measurements applied to cell culture systems has been addressed in several small scale studies and the wavenumber intervals adopted vary according to the objectives and analytes of interest. The mammalian cell lines used in this study were used in previous studies. One key difference was that the previous models built using glucose spectral region were calibration models (Roychoudhury et al., 2007; Arnold et al., 2003; Hagman and Sivertsson, 1998). Although the models built by these authors were based on glucose wavenumber regions, the selected wavenumber regions varied. Roychoudhury et al. (2007) selected the first overtone of C-H combinations band (1333-1640 nm) whereas Arnold et al. (2003) opted to use C-H first overtone band (1650 – 1750 nm) and C-H combination band (2260 – 2290 nm) to build their calibration model for glucose. Hagman and Sivertsson (1998) did not specify the spectra regions used in their calibration model despite highlighting the need to use the spectral region which shows strong intensity of the analyte band and low intensity of the interfering compounds. For other types of cell culture systems, different wavenumber regions were chosen to model glucose. The choice of wavenumber regions, however, ultimately depends on the nature of application (Vaidyanathan et al., 1999; Ge et al., 1994).

4.10 Integrated Wavelet Decomposition-Multiway PCA Model

As outlined in Section 4.5, a batch process comprising NIR spectra is unfolded to a two-dimensional data matrix $\boldsymbol{E}$ (KxIJ) illustrated in Figure 4.10. The rows of matrix $\boldsymbol{E}$, k=1,2,...,K are number of spectra logged throughout i=1,2,…, I batch runs and the columns j=1,2,...,J are wavenumber of the NIR spectra. Information contained in the NIR spectra is the absorbance intensity of the animal cell culture cultivation in the bioreactor.

After the selection of appropriate wavenumber regions, a process representation was built. The full NIR spectra consisted of 1945 data points but by partitioning the spectra according to the CH bands as illustrated in Figure 4.21, the resulting spectra comprised 200 (CH1), 150 (CH2), 100 (CH3) and 100 (CH4) wavenumber in the four NIR overtone regions respectively. Each CH band was modelled separately as well as in combination, i.e. 550 wavenumbers.

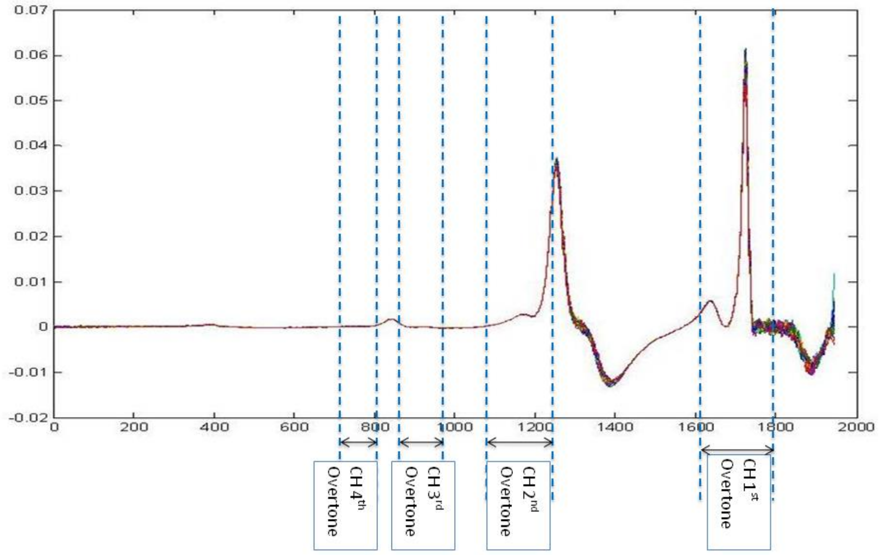


Figure 4.21 NIR spectra partitioned into CH bands to the four overtone regions.

Once the NIR spectra were partitioned, the discrete wavelet transform (db5) was applied to the partitioned columns of the unfolded matrix $\boldsymbol{E}$. This transformed the NIR spectra into wavelet coefficients thereby providing a compact representation that shows the energy distribution of the NIR spectra in time and frequency. The five levels of wavelet decomposition partitioned the NIR spectra into ten sub-bands; cD1, cD2, cD3, cD4, cD5, cA1, cA2, cA3, cA4 and cA5, where the first five are detail sub-bands and the latter five are approximation sub-bands. The letter ‘c’ in the abbreviation of the sub-bands stands for wavelet coefficient, ‘D’ for detail sub-band, ‘A’ for approximation sub-band, and the number represents the level of wavelet decomposition.

Each sub-band contains wavelet coefficients of the corresponding wavelet decomposition level. In wavelet terms, this procedure is known as multiresolution wavelet decomposition. The procedure of multiresolution wavelet decomposition for the first CH overtone region is shown in Figure 4.22. It summarises the number of wavelet coefficients in each sub-band following a discrete wavelet decomposition. These wavelet coefficients are often called features and using they have been found to be useful for discrimination (Bos and Vrielink, 1994). In their study they showed that the wavelet coefficients extracted features from the infra-red spectra that enabled the identification of benzenes.

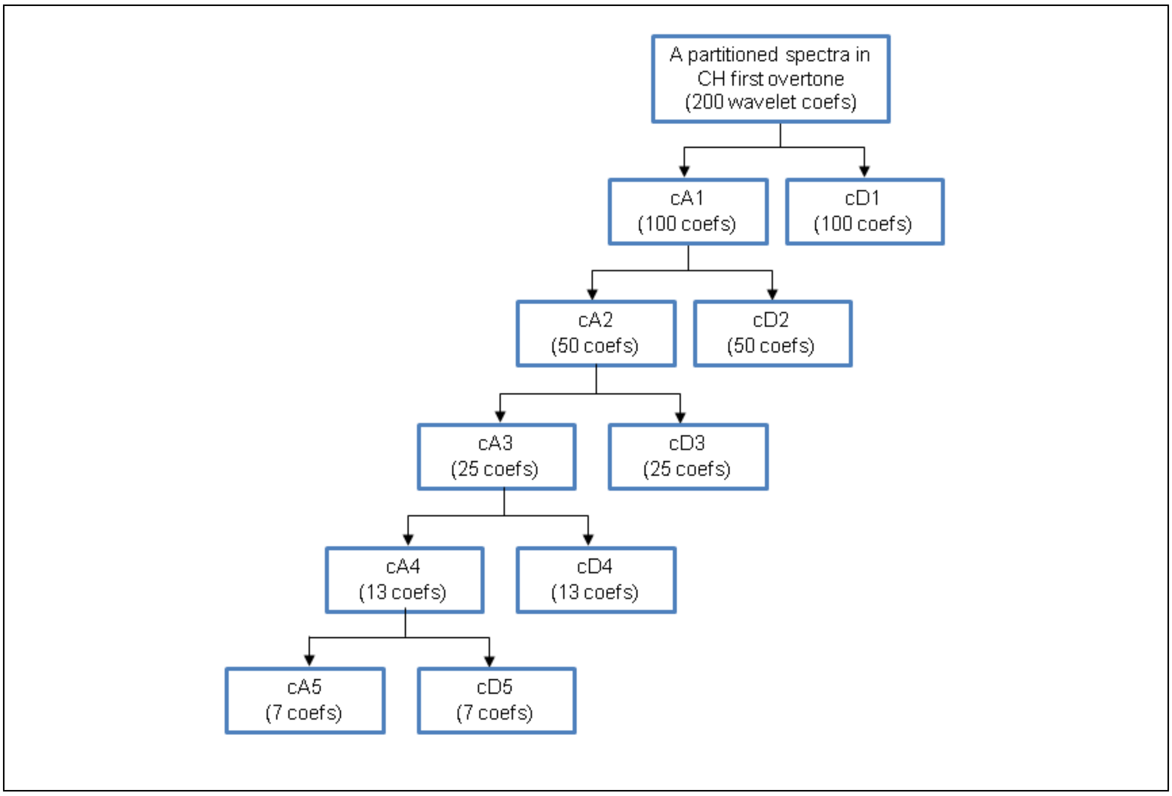


Figure 4.22 Wavelet decomposition of a NIR spectra in first CH overtone region in terms of its coefficients.

Based on the discrete wavelet transform theory discussed in Section 3.2, the original signal can be reconstructed from these sub-bands. However, reconstruction of the original NIR spectra directly after its decomposition defeats the purpose of transforming it in the first place. As pointed out by Daubechies (1992), the application of wavelet decomposition allows further manipulation of the NIR spectral data which has been transformed into wavelet coefficients. Therefore to extract the underlying information in the NIR spectra, manipulation of the wavelet coefficients is required to identify characteristics of the signal that were not apparent from the original NIR spectra. In this study, manipulation of the wavelet coefficient takes the form of calculating its statistical feature (standard deviation).

On completion of the wavelet decomposition, the wavelet coefficients were mean centred without scaling to unit variance. Scaling is omitted because the NIR spectra absorbance intensities are measured in the same units (Gurden et al., 2002). The process of mean centring the wavelet coefficients was to calculate the row-wise average and subtract it from each wavelet coefficient, thereby relocating the mean at the origin. After mean centring, the standard deviation of the mean centred wavelet coefficients was calculated. Details of these procedures are described in Figure 4.23 using sub-band cD_1 of the NIR spectra in the first CH overtone region.

It has been suggested that accuracy of the classification is often improved by representing signals by their important features (Cvetkovic et al., 2008; Jahankhani et al., 2006). Also, to prevent overfitting, it is best to keep the features used for classification to less than one-third of the number of points in the original dataset (Walczak et al., 1996). Therefore, prior to applying PCA, the mean centred wavelet coefficients were further manipulated by calculating their standard deviation. Other statistical features including mean, maximum and minimum of the wavelet coefficients were not considered as this study focuses on the measurement of dispersion between the wavelet coefficients.

Calculation of the standard deviation for the 6 sub-bands for each NIR spectra resulted in a total of six features per spectra. Each NIR spectra has now been reduced by approximately 320-fold, i.e. from 1945 points to 6 features. Figure 4.24 shows the six features per spectrum extracted from the seed batches. The x-axis represents the number of spectra from each batch where the batches were arranged according to the genealogy discussed previously. The y-axis represents the magnitude of the standard deviation in each sub-band. The top graph shows the features extracted from the sub-band cA5, followed by graphs of features extracted from sub-band cD5, cD4, cD3, cD2 and cD1. It is observed that the magnitude of the features in the sub-band cA5, cD5, cD4 and cD3 of batch S6 were lower that the features from other batches. The impact of this difference is shown in the subsequent PCA analysis (Section 4.11.1).

The new matrix contains the standard deviations of the mean centred wavelet coefficients and this will be used as an input for unsupervised principal component analysis. The loadings were then used to project the resulting standard deviation for the production batches onto the principal component space.

$$\boldsymbol{T}_P = \quad features_P * \boldsymbol{P}_S \qquad \text{Equation 4.1}$$

where

$\boldsymbol{T}_P =$ Matrix of the projected scores of the production batches

$features_P =$ standard deviation extracted from the production batches

$\boldsymbol{P}_P =$ Loadings matrix obtained from the PCA of seed batches

The results of the analyses are discussed in the following section. It is important to note that in this study the standard deviation was employed to represent the original spectra in the unsupervised principal component analysis.

Original matrix of cD$_1$ sub-band:

$$\begin{array}{c|cccc|cccc|c|cccc|}
 & \multicolumn{4}{c}{\overbrace{\hspace{5cm}}^{b_1}} & \multicolumn{4}{c}{\overbrace{\hspace{5cm}}^{b_2}} & & \multicolumn{4}{c}{\overbrace{\hspace{5cm}}^{b_I}} \\
 & \lambda_1 & \lambda_2 & & \lambda_J & \lambda_1 & \lambda_2 & & \lambda_J & & \lambda_1 & \lambda_2 & & \lambda_J \\
s_1 & s_{11}b_1 & s_{12}b_1 & \dots & s_{1J}b_1 & s_{11}b_2 & s_{12}b_2 & \dots & s_{1J}b_2 & \cdots & s_{11}b_I & s_{12}b_I & \dots & s_{1J}b_I \\
s_2 & s_{21}b_1 & s_{22}b_1 & \dots & s_{2J}b_1 & s_{21}b_2 & s_{22}b_2 & \dots & s_{2J}b_2 & & s_{21}b_I & s_{22}b_I & \dots & s_{2J}b_I \\
\cdot & \cdot & & & & \cdot & & & & & \cdot & & & \\
s_K & s_{K1}b_1 & s_{K2}b_1 & \dots & s_{KJ}b_1 & s_{K1}b_2 & s_{K2}b_2 & \dots & s_{KJ}b_2 & & s_{K1}b_I & s_{K2}b_I & \dots & s_{KJ}b_I
\end{array}$$

Mean of the sub-band is calculated as follows:

$$\bar{s}_K = \frac{s_{K1}b_1 + \cdots + s_{KJ}b_1 + s_{K1}b_2 + \cdots + s_{KJ}b_2 + \cdots + s_{K1}b_I + \cdots + s_{KJ}b_I}{number\ of\ spectra\ in\ row\ K} \qquad \text{Equation 4.2}$$

resulting in a column vector :

$$\begin{bmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \cdot \\ \bar{s}_K \end{bmatrix}$$

Thus matrix of the mean centred sub-band cD$_1$ is:

$$\left[\begin{array}{cccc|c|cccc}
\multicolumn{4}{c}{\overbrace{\hspace{7cm}}^{b_1}} & & \multicolumn{4}{c}{\overbrace{\hspace{7cm}}^{b_I}} \\
s_{11}b_1 - \bar{s}_1 = m_{11}b_1 & m_{12}b_1 & \dots & m_{1J}b_1 & \cdot\ \cdot & s_{11}b_I - \bar{s}_1 = m_{11}b_I & m_{12}b_I & \dots & m_{1J}b_I \\
s_{21}b_1 - \bar{s}_2 = m_{21}b_1 & m_{22}b_1 & \dots & m_{2J}b_1 & \cdot & s_{21}b_I - \bar{s}_2 = m_{21}b_I & m_{22}b_I & \dots & m_{2J}b_I \\
\cdot & & & & & \cdot & & & \\
s_{K1}b_1 - \bar{s}_K = m_{K1}b_1 & m_{K2}b_1 & \dots & m_{KJ}b_1 & & s_{K1}b_I - \bar{s}_K = m_{K1}b_I & m_{K2}b_I & \dots & m_{KJ}b_I
\end{array}\right]$$

Then, standard deviation, $sd_{s_K}b_I$, of the mean centred sub-band cD$_1$ is calculated as follows:

$$sd_{s_K} = \sqrt{\frac{(m_{K1}b_I - \bar{m}_K)^2 + (m_{K2}b_I - \bar{m}_K)^2 \dots + \left(m_{KJ}b_I - \bar{m}_K\right)^2}{(number\ of\ wavelet\ coefficients\ in\ the\ subband - 1)}} \qquad \text{Equation 4.3}$$

where $\bar{m}_K =$ mean of the mean centred matrix; calculated using a formula similar to Equation 4.2.

Figure 4.23 Derivation of the calculation of the mean centred sub-band and its standard deviation

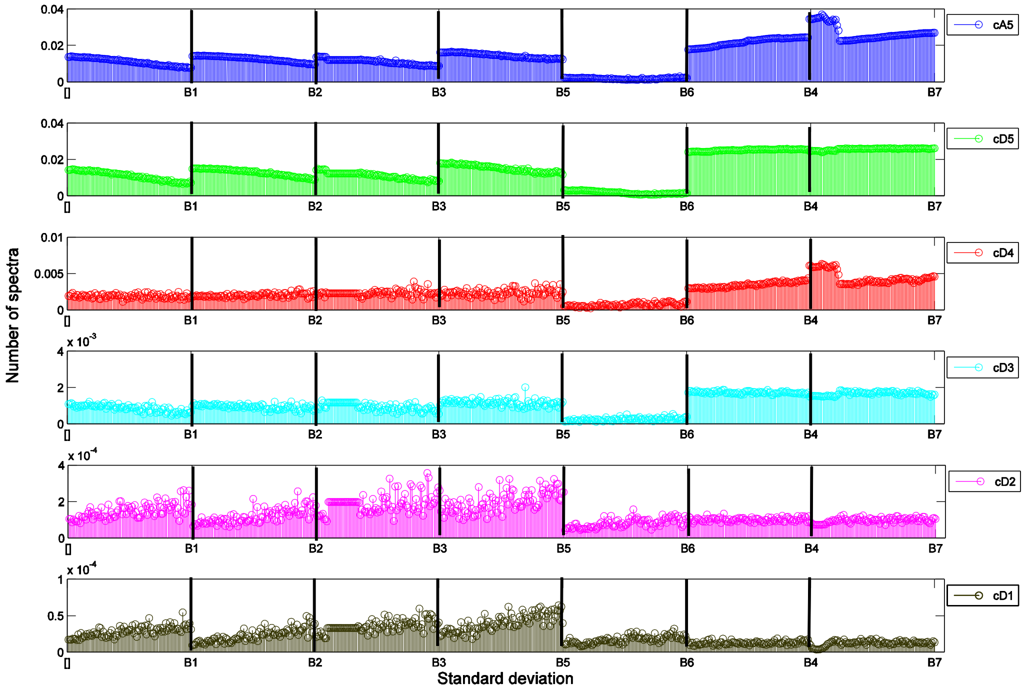


Figure 4.24 Features extracted from the seed batches in each sub-band

4.11 Results and Discussion of the Integrated Model

PCA was performed on the extracted features of the NIR spectra and it is found that two principal components capture more than 90% of the variance in the datasets, as shown in Table 4-2. Figure 4.25 to Figure 4.36 show the bivariate score plots for the four CH overtone regions for PC1 and PC2.

The bivariate scores plots provide a visual representation and a summary of the relationship between the extracted features of the spectral data enabling interpretation of the spectral fingerprints in the context of the genealogy of the batches. Wavelet features loading plots were then analysed to provide further in depth analysis.

The labeling convention used throughout the bivariate scores plots was based on colours being used to represent each family with different symbols defining the position of the batches in the genealogy. As shown in Figure 4.3, S3 and S4 were direct successors of S2, and were inoculated on the same date, thus they are represented by the same symbol. All production batches were represented with the symbol '+' despite being in different families. The reason is that all three production batches were cultured on the same date and there is only one production batch in each family.

It is important to note that square prediction error (SPE) and Hotelling T^2 were not analysed in this study because these two metrics are more applicable for process monitoring and fault detection.

Table 4-2 Percentage of variance explained by the individual principal components

Principal component		1	2	3	4
First CH Overtone	Individual %variance captured	97.49	2.38	0.12	0.01
	Cumulative %variance captured	97.49	99.87	99.99	100
Second CH Overtone	Individual %variance captured	92.50	7.48	0.02	-
	Cumulative %variance captured	92.50	99.98	100	-
Third CH Overtone	Individual %variance captured	99.57	0.23	0.16	0.04
	Cumulative %variance captured	99.57	99.80	99.96	100
Fourth CH Overtone	Individual %variance captured	81.47	12.23	5.24	1.00
	Cumulative %variance captured	81.47	93.70	98.95	99.95
ALL CH Overtone	Individual %variance captured	96.46	3.44	0.08	0.01
	Cumulative %variance captured	96.46	99.90	99.98	99.99

4.11.1 *Principal Component Scores Plots*

In the first CH overtone, the plane defined by the first two principal components (PC1 vs. PC2) explains 99.87% of the variance of the dataset. It is interesting to observe a separation between seed batches and production batches, with the exception of P2, along the PC2 axis as shown in Figure 4.25. The cluster of production batches (except for P2) are located to the top right of the separation line i.e. in the dotted circle. P2 however is interspersed between the clusters of Family 1 seed batches and Family 3 seed batches. The seed batches of Family 1 are located diagonally at the origin whereas the seed batches of Family 3 (blue) are spread diagonally to the right of PC2. Furthermore, it is observed that seed batch of Family 2 is placed to the left of Family 1 and spread vertically across PC1. Figure 4.26 zooms in on the features of S4 and S7 that deviated from the main cluster of seed batches which were located towards the bottom right of the figure. Further interrogation on the PC1-PC2 space reveals they are the extracted features at the end of S4 and the beginning of S7. Based on the genealogy, the cell culture in S4 was used to subculture S7 and this results in overlaps of the spectra. Meanwhile Figure 4.27 shows the area where S4 overlaps with seed batches of Family 1.

As described previously in Section 4.2, the spectra comprised 89 spectra. In context of the cell culture growth curve, approximately the first 20 spectra represent Day 1 (lag phase) whilst the remaining spectra represent Day 2 to Day 3 (spectra 21 to 70) and Day 4 (exponential phase i.e. spectra 71 to 89) of the subculture process (Figure 4.5).

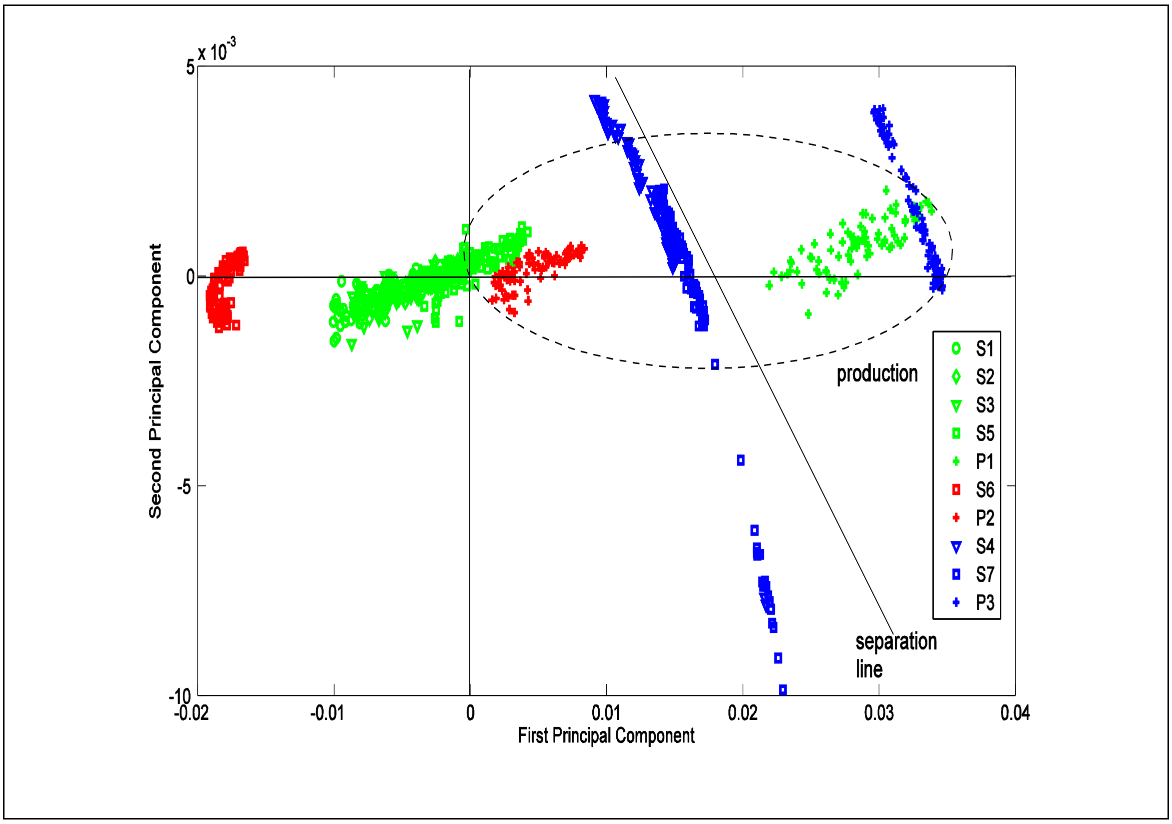


Figure 4.25 PC1 vs. PC2 for first CH overtone

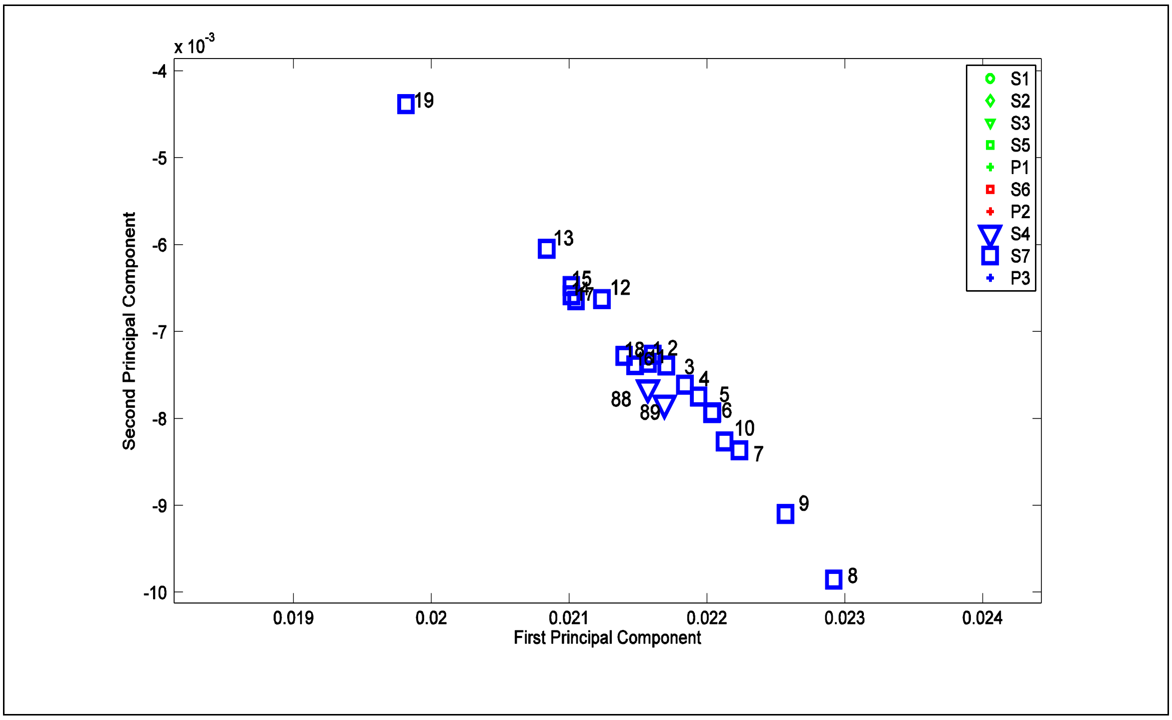


Figure 4.26 PC1 vs. PC2 for first CH overtone - features of S4 and S7 that deviate from the main cluster of seed batches

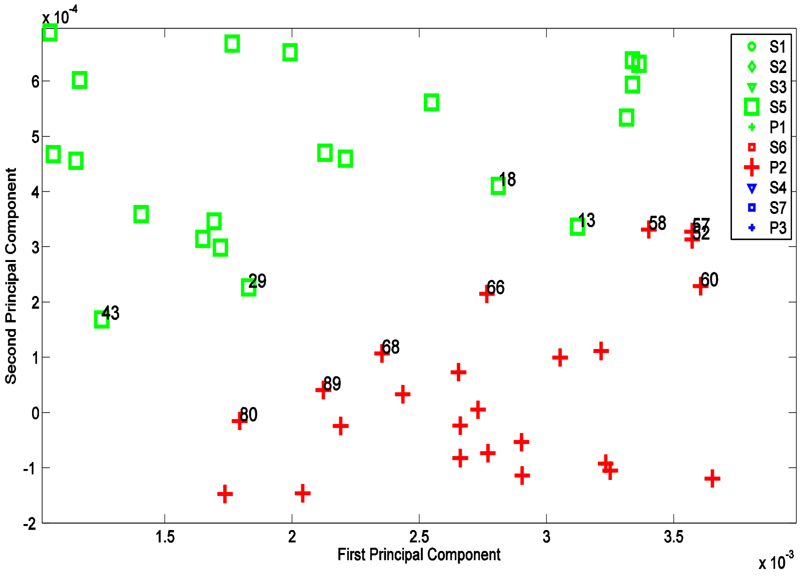


Figure 4.27 PC1 vs. PC2 for first CH overtone – features of P2 overlap with Family 1 seed batches

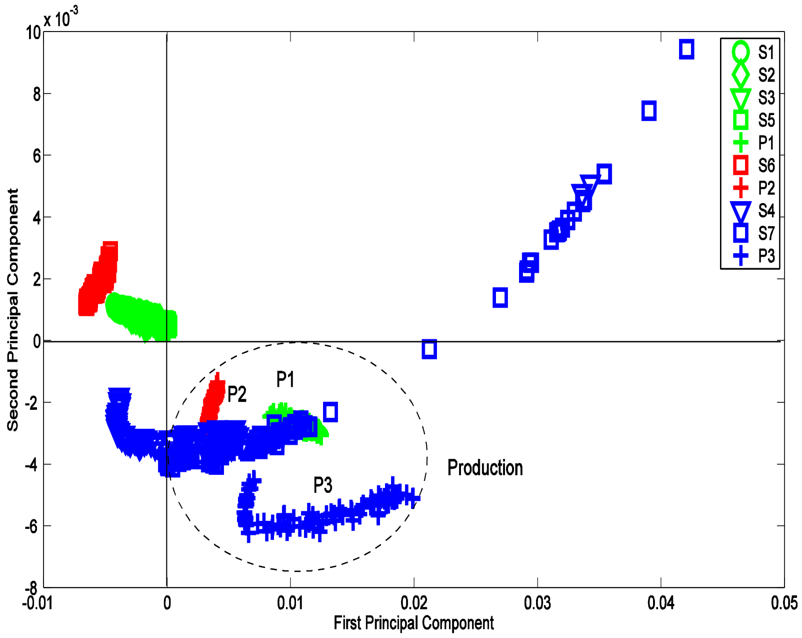


Figure 4.28 PC1 vs. PC2 for second CH overtone

It can be seen that for the second CH overtone the seed batches of Family 3 were scattered diagonally across PC1 and PC2 (Figure 4.28). PC1 and PC2 explained 99.98% of the variance in the features extracted from the second CH overtone. Some features from the cluster of Family 3 seed batches were detected at the top right, far from the majority of features for the same family. These are the same features identified in the first CH overtone analysis (Figure 4.29). Figure 4.30 and Figure 4.31 exhibit two areas where the overlaps occur: features from the end of S7 overlap with the middle of P1, and features from the middle of S4 merge with the start of P2, respectively. A cluster of Family 1 seed batches (S1, S2, S3 and S5) were located near the origin whilst at its top left is a S5 of Family 1 and S6 of Family 2. Extracted features of the production batches from Family 1 (P1) and Family 2 (P2) form a tight cluster within itself whereas the features of the production batch of Family 3 (P3) were spread out near to the cluster of Family 3 seed batches. Further discussion regarding the relationship between the principal component scores and the genealogy is discussed in Section 4.11.3.

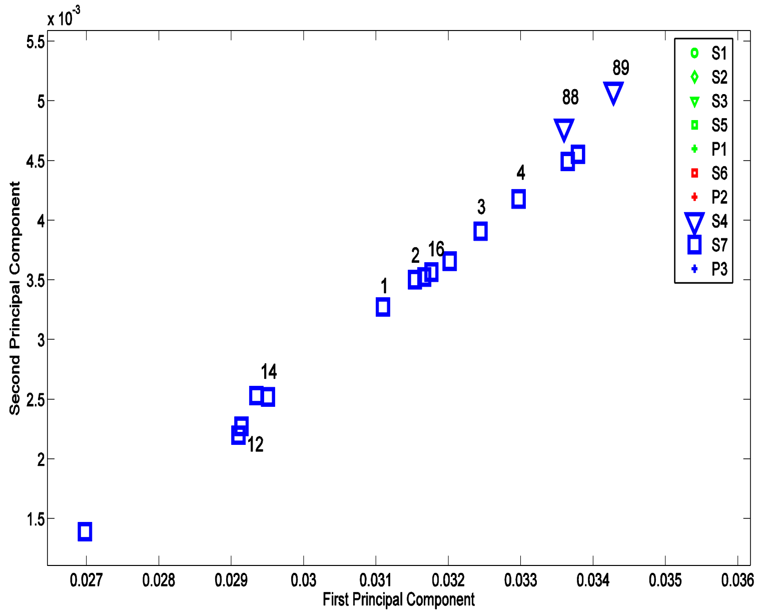


Figure 4.29 PC1 vs. PC2 for second CH overtone - features of S7 that deviate from the main cluster of seed batches

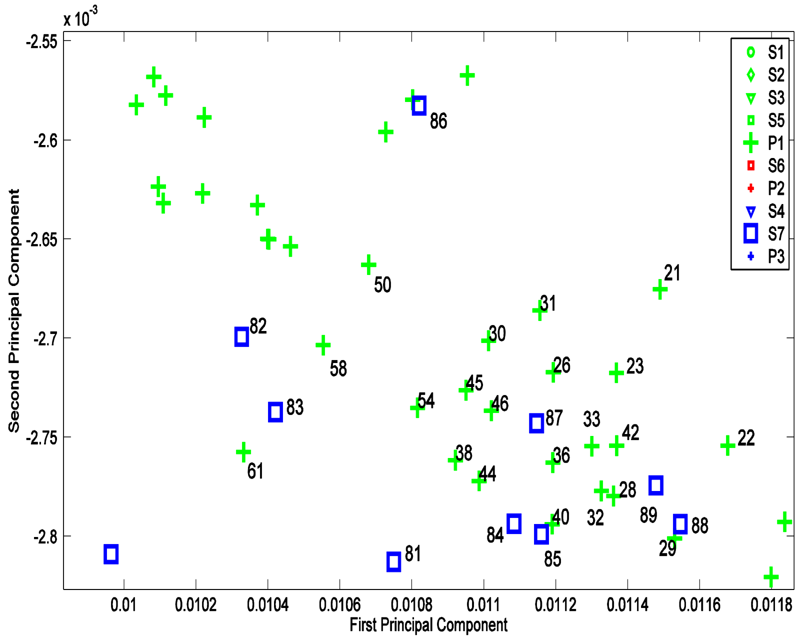


Figure 4.30 PC1 vs. PC2 for second CH overtone - overlap between features from end of S7 with middle P1

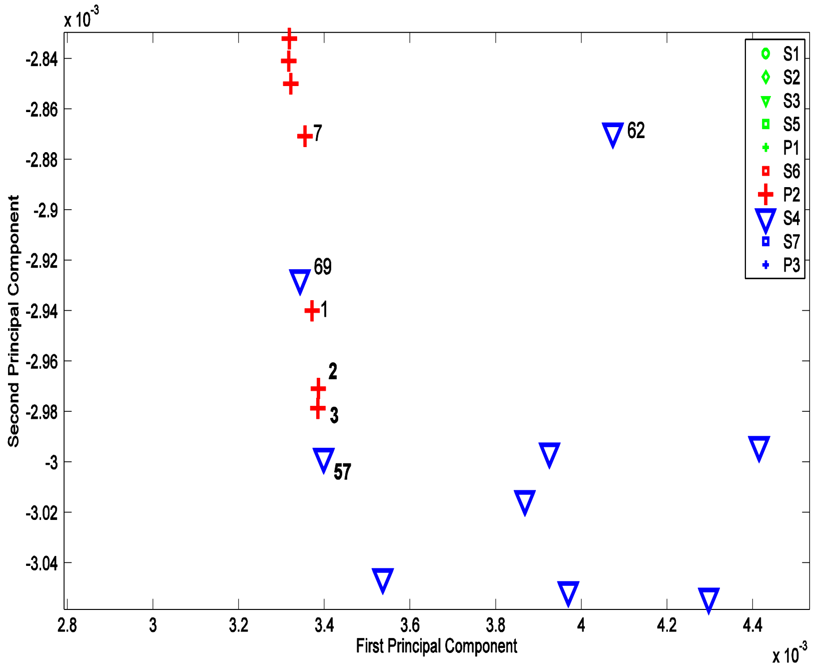


Figure 4.31 PC1 vs. PC2 for second CH overtone - overlap between features from middle of S4 with start of P2

For the third CH overtone, it is clearly seen in Figure 4.32 that the extracted features divide into two clusters: the seed batches to the left, and the production batches to the right of the separation line. There is however, a slight overlap between the Family 3 seed batches with the cluster of production batches along PC1. Figure 4.33 shows a closer view of the overlap. It can be seen that the features from the end of S4 and S7 are interspersed with features from the middle of P7. Also, features from the middle of S7 overlap with features at the start of P2.

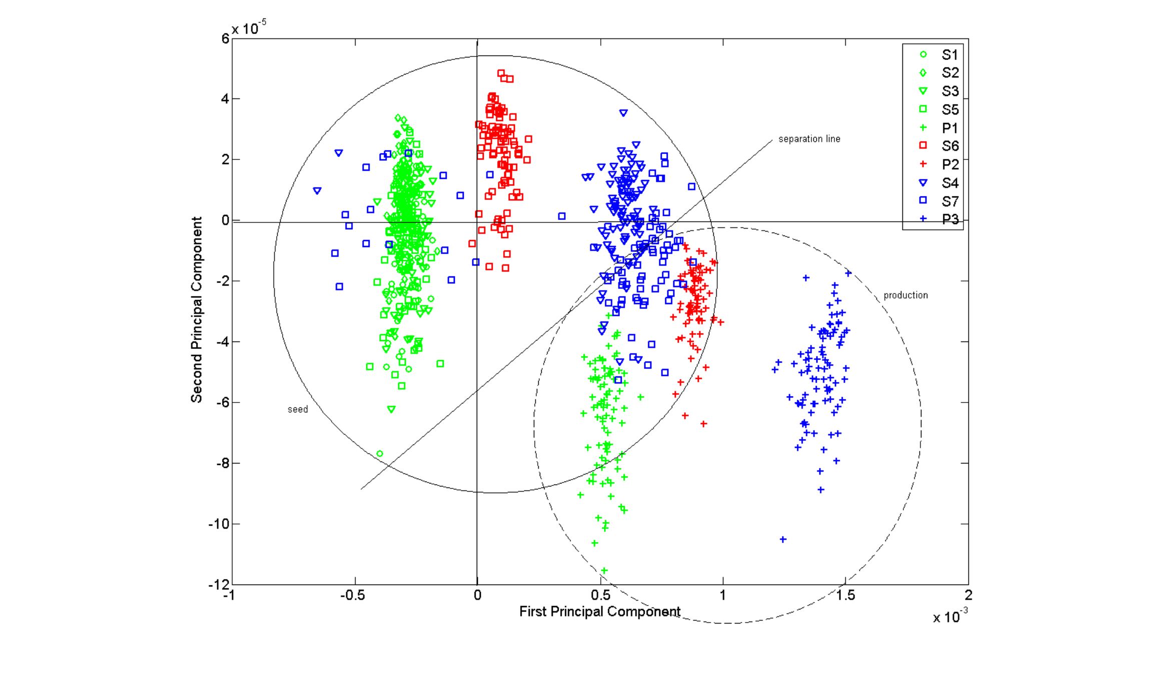


Figure 4.32 PC1 vs. PC2 for third CH overtone

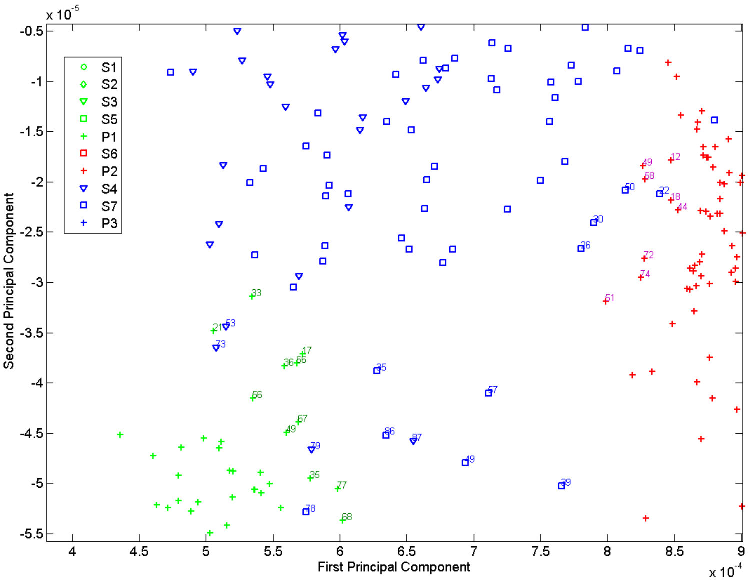


Figure 4.33 PC1 vs. PC2 for third CH overtone – overlap between seed and production clusters

Extracted features of the NIR spectra in the fourth CH overtone are more scattered compared to the other three CH overtones. Although the objectives differ to those in Roychoudhry et al. (2007), whose findings showed that the variance is greater at lower wavenumbers due to the energy throughput being directly proportional to the wavenumber. As shown in Table 4-2, the percentage of variance captured in the fourth CH overtone by the first four principal components is 94.85%, approximately 5% less than the other three CH overtones. The scores scatter plot of the NIR extracted features in fourth CH overtone (Figure 4.34) shows a slight resemblance to those in third CH overtone (Figure 4.32). The cluster of seed batches is located at the top left of the separation line whereas the cluster of production batches is placed at the bottom right. Also there is an overlap between clusters of seed batches and production batches. The overlap in the fourth CH overtone involves all except one of the seed batches (S1). Detailed discussion is presented in Section 4.11.3.

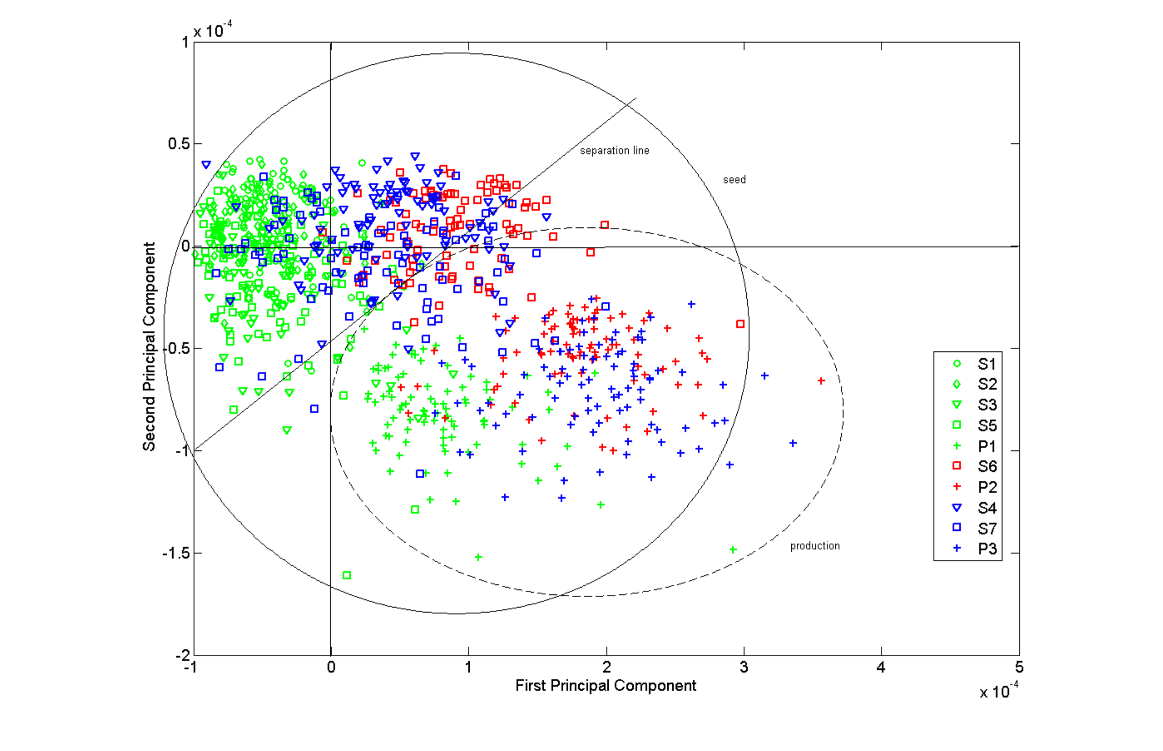


Figure 4.34 PC1 vs. PC2 for fourth CH overtone

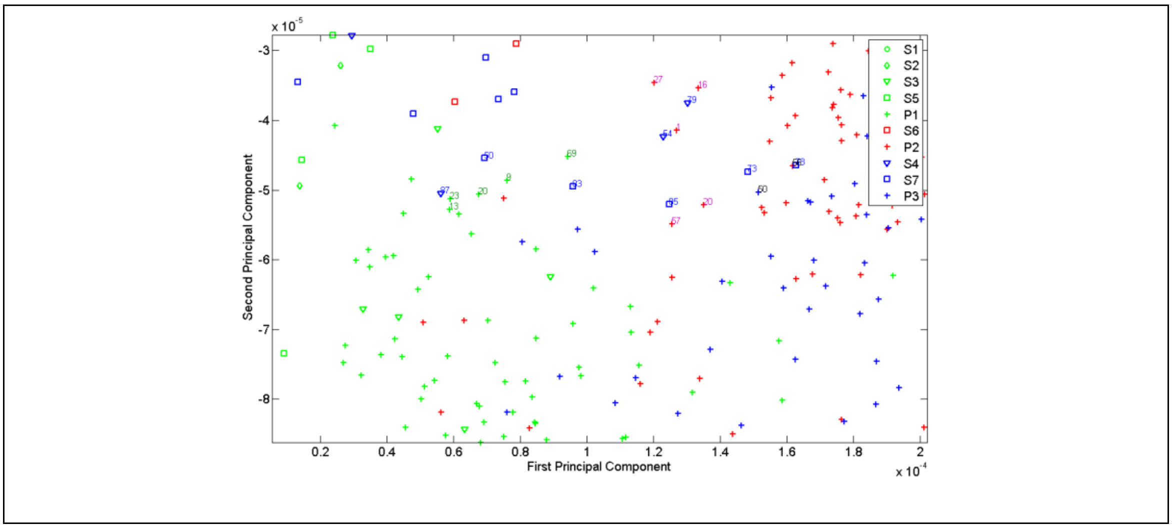


Figure 4.35 PC1 vs. PC2 for fourth CH overtone - overlap between seed and production clusters

Finally a model was established which combines the individual overtone CH regions. It is interesting to note that the resulting representation is heavily influenced by the first CH overtone region as shown in the bivariate scores plot (Figure 4.36) with it being very similar to the bivariate plot for the first CH overtone (Figure 4.25). This shows that the behaviour of the NIR spectra in the first CH overtone shapes the overall average behaviour. The percentage of variance explained by the first two principal components for the combined CH overtone model is 99.91%.

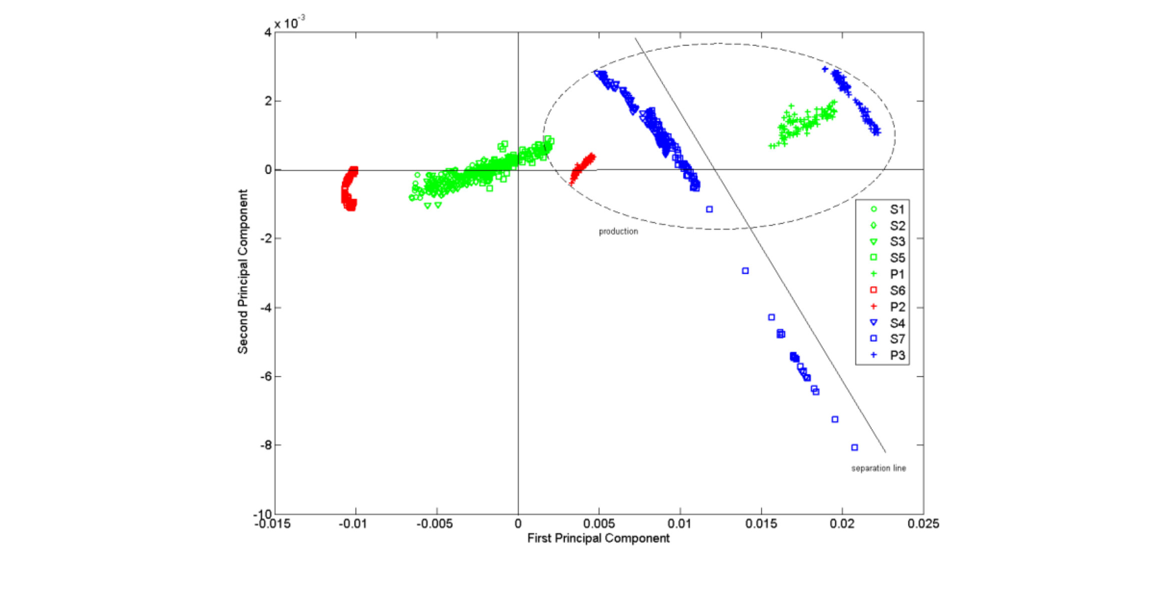


Figure 4.36 PC1 vs. PC2 for all CH overtones

The bivariate scores plots for PC3 and PC4 for the individual CH overtones and the combined CH overtones were included in Appendix A.

4.11.2 *Loadings Plots*

Six wavelet sub-bands were taken into account for feature extraction. Figure 4.37 to Figure 4.41 show the univariate loadings plot of PC1 to PC4 in terms of the wavelet features for the individual CH overtones and the combined CH overtones. All five figures show that features extracted from A5, i.e. approximation coefficients sub-band from decomposition at level 5, have a large impact in terms of defining the direction of greatest varibility in PC1. For all CH overtones, the features from A5 are the largest positive loadings for PC1 as demonstrated in Figure 4.37 to Figure 4.40. This association between the A5 features and PC1 is an advantage because it reflects the ability of the model to captures the underlying information.

The proportion of the sub-band A5 decreases whilst the proportion of the sub-band D1 increases as the principal component move from PC1 through to PC4. This is anticipated as it relates to the content of detail coefficients of level 1 which is mostly high frequency component or better known as noise.

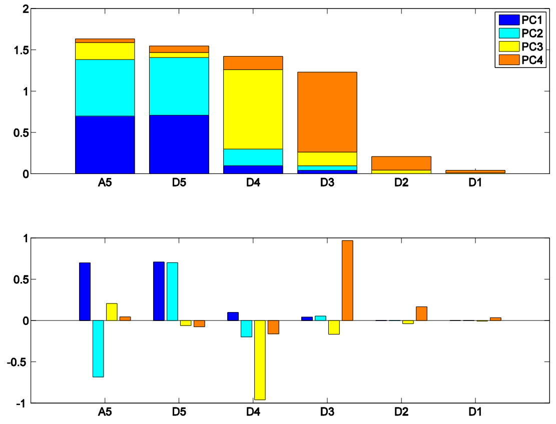


Figure 4.37 Univariate loadings plot of PC1 to PC4 in terms of wavelet features for first CH overtone

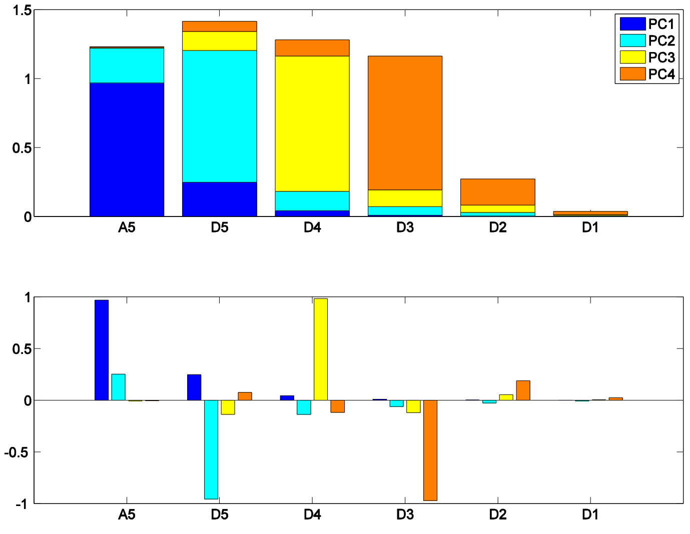


Figure 4.38 Univariate loadings plot of PC1 to PC4 in terms of wavelet features for second CH overtone

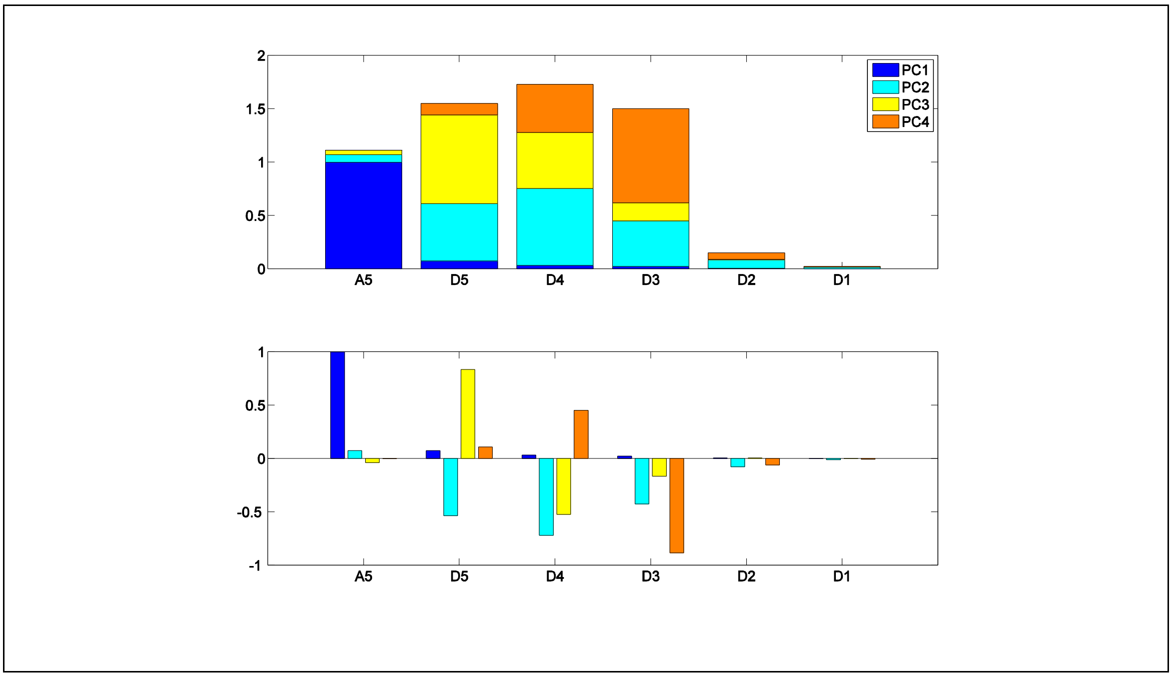


Figure 4.39 Univariate loadings plot of PC1 to PC4 in terms of wavelet features for third CH overtone

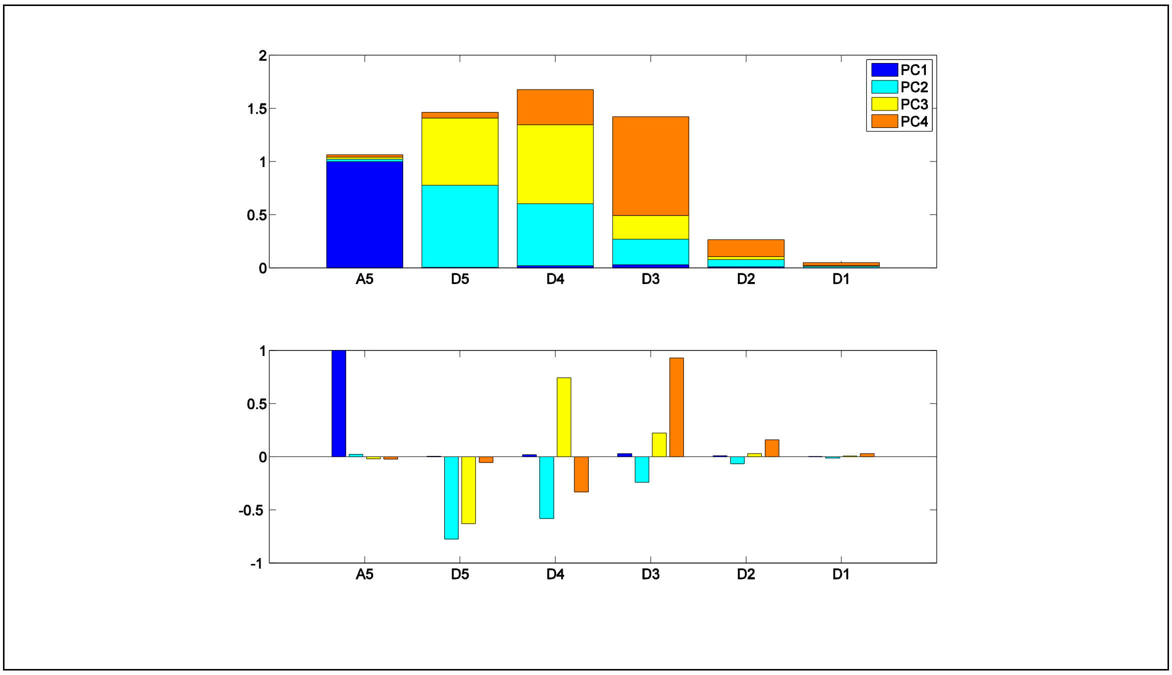


Figure 4.40 Univariate loadings plot of PC1 to PC4 in terms of wavelet features for fourth CH overtone

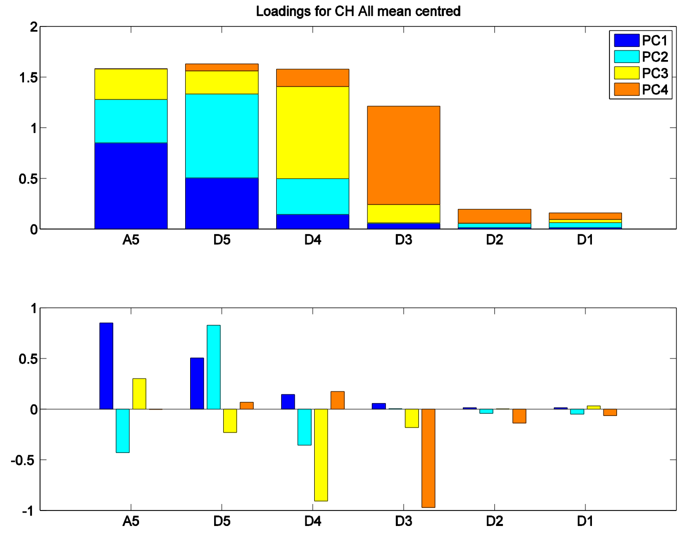


Figure 4.41 Univariate loadings plot of PC1 to PC4 in terms of wavelet features for all CH overtones

4.11.3 *Discussion*

Although complex to interpret, the scores and wavelet feature loadings plots are shown to demonstrate how underlying information can be extracted in terms of the relationship between the batch genealogy and process behaviour. Changes in the spectral information that correspond to process genealogy could be deciphered from these plots. These are discussed in detail in this section.

There are a number of factors which may cause the clustering behavior identified in terms of the spectral data for the seed batches within the same family and between different families, and the production batches of different families (Figure 4.25, Figure 4.28, and Figure 4.32). The factors vary from the apparatus used in the experiment such as the probes and bioreactors, to the process related measurements such as cell density, subculture ratio and passage number.

The probes and bioreactors have a strong influence in terms of the clustering of the batches. Optical differences between probes have been shown to cause the most variability in the NIR spectra and these differences are detectable in the scores plot (Roychoudhury et al., 2007). Based on the information from the engineer in charge of the manufacturing process, for the genealogy under study there exist optical differences between the probes used in different bioreactors. Minor differences in the alignment and calibration of each probe results in minor differences between the spectra collected. It is thus possible that the differences or similarities detected between the different families were a result of the optical differences or similarities of the probes. On the one hand, this variability is an advantage in that seed batches cultured in the same bioreactor with the same probe would be expected to form a cluster. On the other hand, the setting up of the probes is done manually. In a situation where the probes were not set up in a consistent manner by the operator, probe set up variability may occur. This can lead to reduced signal intensity and impact on the starting baseline and detector limit for saturation. Similar to probe set up variability, variability in bioreactor set up and operation may also contribute to differences between bioreactors with small differences in bioreactor set up and operation having a large impact on growth and substrate utilisation characteristics of a culture.

Aside from the apparatus used in the experiment, subculture ratio and passage number are other contributing factors that can impact the clustering behaviour of the batches. Subculture ratio which is also known as inoculum carry over ratio is a measure of how much an inoculum culture is diluted to achieve a specified starting viable cell concentration in a subsequent culture. Subculture ratio is of importance because it indicates the volume of carryover material from the inoculum. The volume being carried over from the previous to the next batch is measured based on the volume of alkali being carried over. It is a ratio of the final time of the batch run (in seconds) to subculture ratio. As some component in the inoculum is being carried over into the subsequent culture, this will impact on the startup and progression of the culture and spectral fingerprint. The volume of alkali carry over affects not only process behaviour but also the quality of the NIR spectra. In the alkali carried over, there is a substance named methyl red and the yellow colour of the methyl red base will change to red if the culture inside the bioreactor turns acidic. The red colour of methyl red may affect the quality of spectra collected by the spectrometer.

The graphs for subculture ratio for all three families are shown in Figure 4.42 (a), (b) and (c). During the MAb manufacturing process, the subculture ratio in S2 shows a significant drop from the subculture ratio of its predecessor, S1 as depicted in Figure 4.42 (b). S2 is then directly used to derive Family 3, as described in the genealogy of the batches shown in Figure 4.3. It is conjectured these two factors have caused the projection on principal components space of both seed batches in Family 3 to part with the projection of seed batches from Family 1 and 2.

Passage number of the mammalian cell culture is another factor to be considered. As a culture goes through successive passages in similar culture conditions, a culture can display adaptation characteristics such as faster growth and substrate utilisation, higher cell concentration and so forth.

On the other hand, the projections of Families 1 and 2 in the principal component space is strong evidence that support the hypothesis that batches originating from the same ancestor exhibit similar behaviour. The fact that Families 1 and 2 were projected near each other in the principal component space despite being inoculated in different bioreactors and different probes were used to obtain the spectra measurement should not be overlooked. Furthermore, through the plot of the subculture ratio of Families 1, 2 and 3 in Figure 4.42, it can be seen that the trend of the subculture ratio for Family 1 and 2 are similar compared to Family 3. This is probably another factor that contributes to the clustering pattern seen in Figure 4.25 to Figure 4.36.

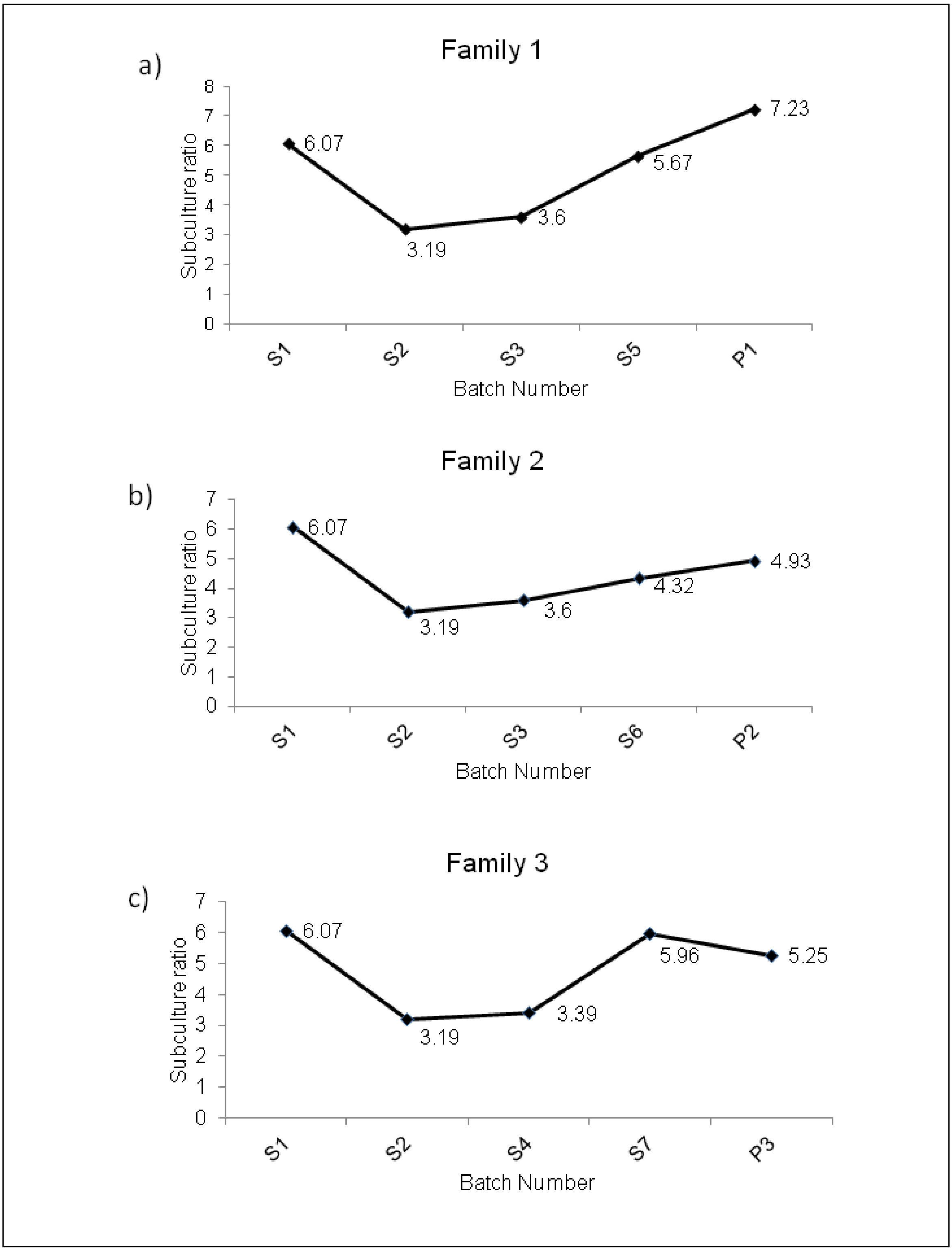


Figure 4.42 Subculture ratio of (a) Family 1. (b) Family 2, and (c) Family 3.

4.12 Summary

The capabilities of the integrated approach, which combines the discrete wavelet transform and PCA, were demonstrated by its application to NIR spectra from an industrial monoclonal antibody batch process. The NIR spectra were initially preprocessed using first derivatives prior to equalizing batch length using an approach that was based on the underlying biological process. The discrete wavelet decomposition with an appropriate mother wavelet and wavelet decomposition level was applied to the glucose overtone regions. The next stage was to apply multiway PCA to the decomposed NIR spectra from the seed batches. The production batches were then projected onto this representation and difference in behaviour were observed through loadings analysis.

This study demonstrates that the integrated approach was able to identify and differentiate using spectral features between different batches. The information extracted from the integrated model coupled with the information from the genealogy of the batches aid understanding of process behaviour. This development theoretically provides a powerful tool to enhance the understanding of process behaviour. It is possible to utilise the methodology to assist in scaling-up a process from pilot plant to full scale manufacturing plant. Results generated from this study can be utilised to select the suitable batches to be subcultured in the commercialization process based on the families in the genealogy that demonstrate stable and robust behaviour.

In the next chapter, the versatility of the integrated model is tested on a different type of dataset with a different objective. The dataset under investigation in Chapter 5 is in the form of electrospray ionisation mass spectra collected from another manufacture of monoclonal antibody cell lines where the main objective is the characterisation of the criteria that differentiate between high and low CHO cell lines.

Chapter 5 Enhanced Integrated Model Application on ESI Data

5.1 Introduction

The high cost of generating cell lines for recombinant monoclonal antibodies (MAb) production necessitates improvements in production techniques to ensure product quality, a reduction in development timelines, and an increase in cost efficiency. Enhancements in the initial steps of the development process would reduce the time from development to commercialization. One of the major challenges in the development process involves the rapid screening and selecting of highly productive and stable cell lines from the transfectant population (Li et al., 2010). Figure 5.1 shows diagrammatically the strategy for the selection of a recombinant MAb cell line.In paving the way towards commercialization, one key issue is to identify the criteria that differentiate between high and low producer cell lines.

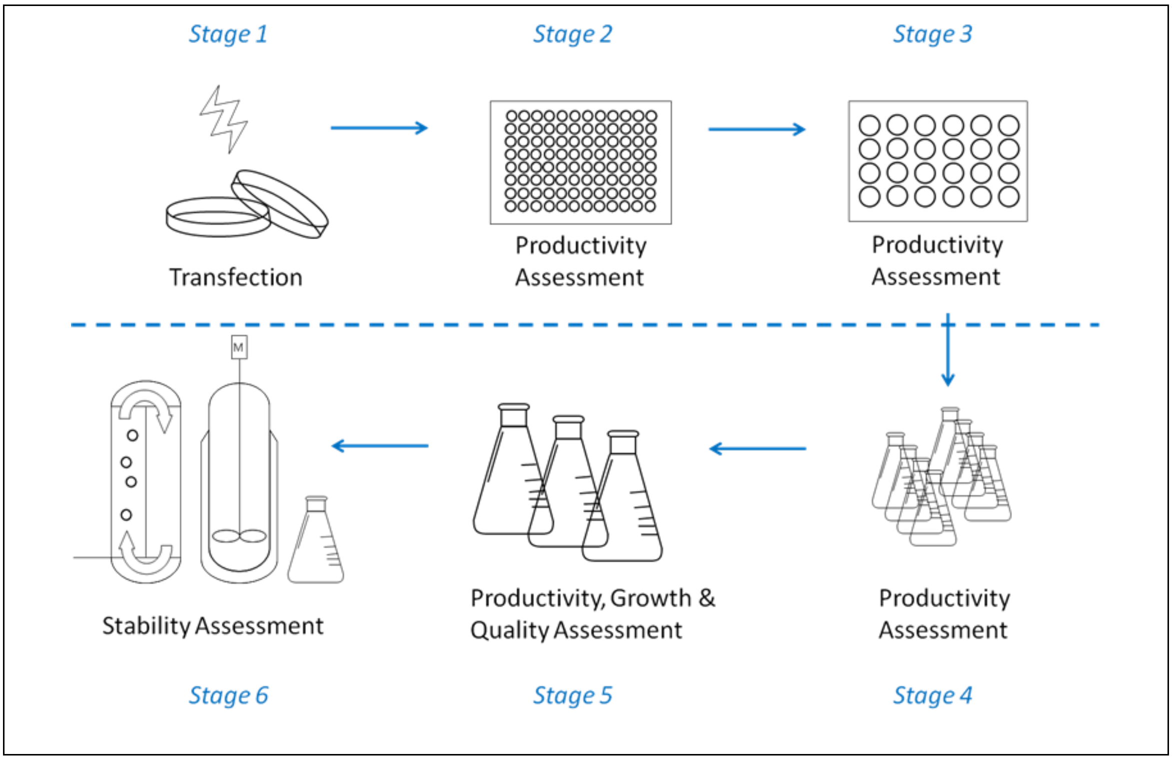


Figure 5.1 A strategy for recombinant MAb cell line selection

In Chapters 2 and 3, the background to the techniques used in the wavelet-PCA integrated approach were reviewed. Chapter 4 focused on the development of the wavelet and PCA integrated methodology and its application to investigate the impact of batch genealogy on process behaviour. Following the successful results from this case study, its more widespread applicability is investigated by applying it to analyse electrospray (i.e. electrospray ionisation, ESI) mass spectrometry data. The ESI data analysed in this Chapter was generated by Kent University.

The goal of this second study was to distinguish between and characterize the criteria that enable differentiation between high and low producer CHO cell lines. This approach exploits the concepts of data mining, signal processing and multivariate statistical analysis. The motivation for this study is discussed in the Section 5.2.

5.2 Motivation for Enhanced Model

Two research areas that may help enhance in the development of cell lines are: 1) the extraction of meaningful information from the generated data set which is generally large, multidimensional and complex, and 2) the identification of techniques to interpret the information inherent within the data. The extraction of information from a complex data set is related to the technical challenges summarised by Hilario et al. (2006). Firstly, fingerprinting of mass spectra involves the need to extract patterns from data that is contaminated by noise, and secondly, the need to manage 'high dimensional-small sample size' data sets. It was suggested by Arneberg et al. (2007) and van den Berg et al. (2006) for protein mass spectra that appropriate preprocessing methods can address these technical challenges. In addition separating biological variation in the mass spectrometry data from variability pertaining to the influence from measurement noise can improve the biological interpretability of the developed models (Archibald and Akin, 2000).

Compounding these technical challenges are issues related to the screening and selection of highly productive and reproducible cell lines from amongst the transfectant population in a limited time frame. Furthermore, the product quality and productivity of cell lines is highly dependent on cell culture conditions (Li et al., 2010).

Meanwhile, another ongoing issue is interpretability of models involving mass spectra classification. Hilario et al. (2006) stated that of the studies they reviewed many focused on the general performance of the models and failed to produce information on the model classifiers i.e. the mass-to-charge ratio of the ions, which would be of use to biomedical researchers. The definition of mass-to-charge ratio is explained in Section 5.4. It is emphasised that the interpretation and identification of biomarkers will be less complicated if the information on the direction and magnitude of the discriminatory mass-to-charge ratio is provided.

A considerable amount of literature has been published on mass spectrometry data fingerprinting and the objectives include investigation of preprocessing techniques for optimal feature extraction and fingerprinting of the mass spectrometry. Studies by Arneberg et al. (2007) and van den Berg et al. (2006) investigated the impact of different preprocessing techniques on the extraction of information from matrix-assisted laser desorption/ionisation mass spectrometry (MALDI-MS) and gas chromatography mass spectrometry (GC-MS) respectively. The impact of preprocessing techniques including smoothing, binning, centering, and scaling on the interpretation of the datasets was studied and it was shown that the outcome of the data analysis was significantly affected. For example, false biomarker candidates resulted from normalization of the spectral data without first addressing to the noise structure (Arneberg et al., 2007). Three important factors in determining a preprocessing technique that enable the extraction of information from a dataset are summarised by van den Berg et al. (2006), they are: (1) the biological question to be answered, (2) the properties of the data set, and (3) the chemometric tools selected.

In terms of mass spectrometry data fingerprinting, previous literature has focused on determining the fingerprint of the protein or metabolites, Zhang et al. (2006) and Danielsson et al. (2011). Zhang et al. (2006) showed that by fingerprinting the protein in MALDI-MS data the identification of three different types of mammalian cell was possible. Meanwhile Danielsson et al. (2011) successfully identified possible biomarkers of prostate and bladder cancer in human urine samples analysed with liquid chromatography–mass spectrometry (LC-MS).

Fingerprinting of mass spectrometry data largely depends on the application of chemometric tools including PCA and analysis of variance (ANOVA). Trim et al. (2008) investigated the application of PCA to aid the interpretation of MALDI-MS image data of brain regions. Scores plots from the supervised PCA showed differentiation between different brain regions whilst interpretation of the loadings plots enabled the identification of white and grey matter in the brain. Other applications of PCA include Mattoli et al. (2011) and de Souza et al. (2007). In these cases PCA was applied to electrospray ionisation mass spectrometry (ESI-MS) to fingerprint botanical dietary products and alcoholic beverages respectively. There was clear separation between different classes of samples in both studies. ANOVA was applied to capillary electrophoresis electrospray ionisation time-of-flight mass spectrometry (CE-ESI-TOF-MS) data generated from the analysis of human urine samples (Allard et al., 2008). Significant differences in intensities of the mass spectrometry due to the influence of three beverages (coffee, tea and water) were reflected in the first 10 principal components in that were presented to an ANOVA analysis.

Despite intensive ongoing research into the fingerprinting of mass spectra, no study exists which address classification in the context of Electrospray Ionisation-Mass Spectroscopy (ESI-MS) data using integrated discrete wavelet transform-PCA. The application of the discrete wavelet transform to extract features from other types of protein mass spectra has been limited. A study by Xia et al. (2007) applied the wavelet transform with PCA and artificial neural networks (ANN) to perform a pattern recognition task on a metabolimics dataset generated from gas chromatography-mass spectrum (GC-MS). Application of the wavelet transform allowed the decomposition of the spectra into wavelet sub-bands and the filtering of the noise component from the spectra through the wavelet denoising algorithm. The approach proposed was to remove all detail wavelet coefficients and thereby, reconstruct and classify the spectra using only the approximation wavelet coefficients. Contrary to Xia et al. (2007), Randolph and Yasui (2006) did not utilise using the wavelet denoising algorithm in the application of the multiscale wavelet transform to quantify MALDI-TOF mass spectrometry. Their methodology was based on a hypothesis that the identification and quantification of the signal content from the mass spectrometry was possible without prior estimation of the signal to noise ratio.

Furthermore, although a number of studies discussed previously adopted PCA as their chemometric tool, the studies only focused on the analyses of the principal component scores and loadings. There exists further a tool in PCA that can be of used in the fingerprinting of protein mass spectra, contribution analysis.

Contribution analysis was introduced in multivariate statistical process control to help identify which process variables were indicative of the changes in process signals for continuous and batch process (Conlin et al., 2000; Simoglou et al., 2000; Westerhuis et al., 2000). As described in Section 3.2.4, the contributions of individual variables to the principal component scores, squared prediction error (SPE or Q-statistic) and the Hotelling's T^2 (D-statistic) can be calculated. The basis of this technique is to compare the contributions of individual variables to the aforementioned statistics for a process which is recognized to not operate under normal conditions. The credibility of this approach has been successfully demonstrated by Simoglou et al. (2000), Nomikos (1996), Kourti et al. (1996) and Kourti et al. (1995). Even though the contribution analysis may not reveal the specific cause of the operational changes, it identifies the non-conforming variables and in conjunction with process understanding the source of the issue can be diagnosed.

This study introduces a new application of contribution analysis. The goal is to look at the contribution of the variables to help characterise the criteria that distinguish between high and low producing CHO cell lines. The integrated discrete wavelet decomposition-PCA approach results in the contribution analyses being performed on the wavelet coefficients.

The utilisation of mathematical models is highlighted as one of the current challenges in pharmaceutical process development and manufacturing in the Pharmaceutical Quality for the 21st Century initiative introduced by the U.S. Food and Drug Administration (FDA) (FDA, 2009a; FDA, 2009b; FDA, 2006; FDA, 2004). This is concurrent with a shift in focus in cell line process development, with the goal being to control product quality and the process as opposed to achieving high titre (Li et al., 2010). Assessment of product quality and process performance necessitates the implementation of analytical tools including process modeling and simulation which utilize data generated during development and production.

5.3 Process Description

The process forming the basis of this study is the manufacture of recombinant MAb cell lines. The selection strategy is as shown in Figure 5.1. Firstly, the host cell line is first transfected with an expression vector. The transfected cell lines are then diluted and cultivated in 96 well plates and screened for highly productive cell lines. The selected cell lines progress to the next stage with non- or low cell line producers being discarded. The selected cell lines are then cultivated in a set of 24 well plates followed by another screening. The next two stages involve the selected cell lines being cultivated in shake flasks to mimic a batch process and a fed-batch process. Each stage was followed by screening. After the screening in the fed-batch stage, the selected cell lines are cultured in small-scale bioreactors for cloning purposes, stability assessment and bioreactor studies. Cell pellets from the small scale bioreactor stage were collected and analysed using the liquid chromatography ESI-MS platform.

Based on a discussion with Kent University who were responsible in generating the ESI-MS data, cell lysis in different buffer systems was performed on the cell pellets for the liquid chromatography ESI-MS analysis. The cell pellets were of varying sizes and were from a range of different cell lines. This was followed by enzymatic digestion of the supernatant obtained from the cell lysis step. Next a series of high-performance liquid chromatography (HPLC) methods were examined to define the liquid chromatography method and the conditions that would result in appropriate mass spectra. Section 5.4 describes the process of electrospray ionisation mass spectrometry.

5.4 Electrospray Ionisation Mass Spectrometry

Mass spectrometry is a microanalytical technique that can be used to determine the elemental composition of an analyte through the measurement of the mass of gas-phase ions produced from molecules of an analyte. One of the unique features of mass spectrometry is its ability to produce and detect fragments of molecule that correspond to discrete groups of atoms of different elements that reveal structural features.

The tools used in the study of mass spectrometry are mass spectrometers and the generated data. The mass spectrometer measures the mass-to-charge ratio (m/z) of an ion instead of the mass of the ion. The mass-to-charge ratio, a dimensionless number, is the mass of the ion on the atomic scale divided by the number of charges that the ion possesses. A principle requirement of mass spectrometry is that the ions are in the gas phase before they can be separated according to their individual m/z values and detected to obtain the mass spectra. There are many different ionisation techniques available for producing gas-phase ions in mass spectrometry, electrospray ionisation is one example. Griffiths et al. (2001) reported that electrospray ionisation is the optimum method of ionisation for the widest range of biological macromolecules in their review of the mass analysis of ions.

ESI-MS is known for its ability to provide a simple, rapid, and sensitive analytical tool for the mass analysis of macromolecules. As was shown by Wan et al. (2001) in the determination of the glycoform amounts in a recombinant antibody produced in Chinese hamster ovary (CHO) cells, the application of ESI-MS enabled a fast analysis of the spectra and hence provided important product quality early in cell culture production.

Electrospray ionisation (ESI) is a soft ionisation technique that has been used to investigate noncovalent bonding of biological macromolecules such as protein and peptides. A noncovalent bond is defined as any relatively weak chemical bond that does not involve an intimate sharing of electrons (Lodish, 2007). Noncovalent bonding is the basis of many dynamic biological processes as it enables large molecules to bind specifically but transiently to one another.

The operating principle of ESI is based on electrical energy. This enables the transfer of ions from solution to the gas-phase prior to being subjected to mass spectrometric analysis. Figure 5.2 shows a schematic diagram of a typical ESI source with the zoom-in circle showing the mechanism of ion formation in ESI. A solution of the analyte is pumped into a small capillary/electrospray needle either from a syringe pump or as effluent flow from liquid chromatography. As shown in the zoom-in circle, the analyte solution then passes through the electrospray needle whose tip is of high potential difference. This generates a split of charge which forces the spraying of charged droplets from the needle. The electrosprayed droplets now possess an excess of positive and negative charges and have the same polarity as the charge on the needle. This causes the droplets to push away from the needle and move towards the source sampling cone on the counter electrode.

As the droplets traverse the space between the needle tip and the cone, they become gradually smaller due to the evaporation of the solvent in a drying gas at atmospheric pressure. This leads to an increase in surface charge density and a decrease in the droplet radius. The droplet will reach its ‘Rayleigh limit’, a point where its surface tension can no longer sustain the repulsion forces between charges resulting in a ‘Coulombic explosion’ hence ripping the droplet apart. The electric field strength within the charged droplet will eventually reach a critical point at which it is kinetically and energetically possible for the ions at the surface of the droplets to be ejected into the gas-phase as analyte molecules. The ions are sampled via a skimmer cone and then transported into the mass spectrometer. When the ionized proteins hit the detector of the mass spectrometer, information is compiled into a histogram. Through time, a series of consecutive histogram are registered for a number of samples. These histograms are known as the mass spectrum and form the basis of the raw data. Figure 5.3 shows a sample of a raw ESI-MS.

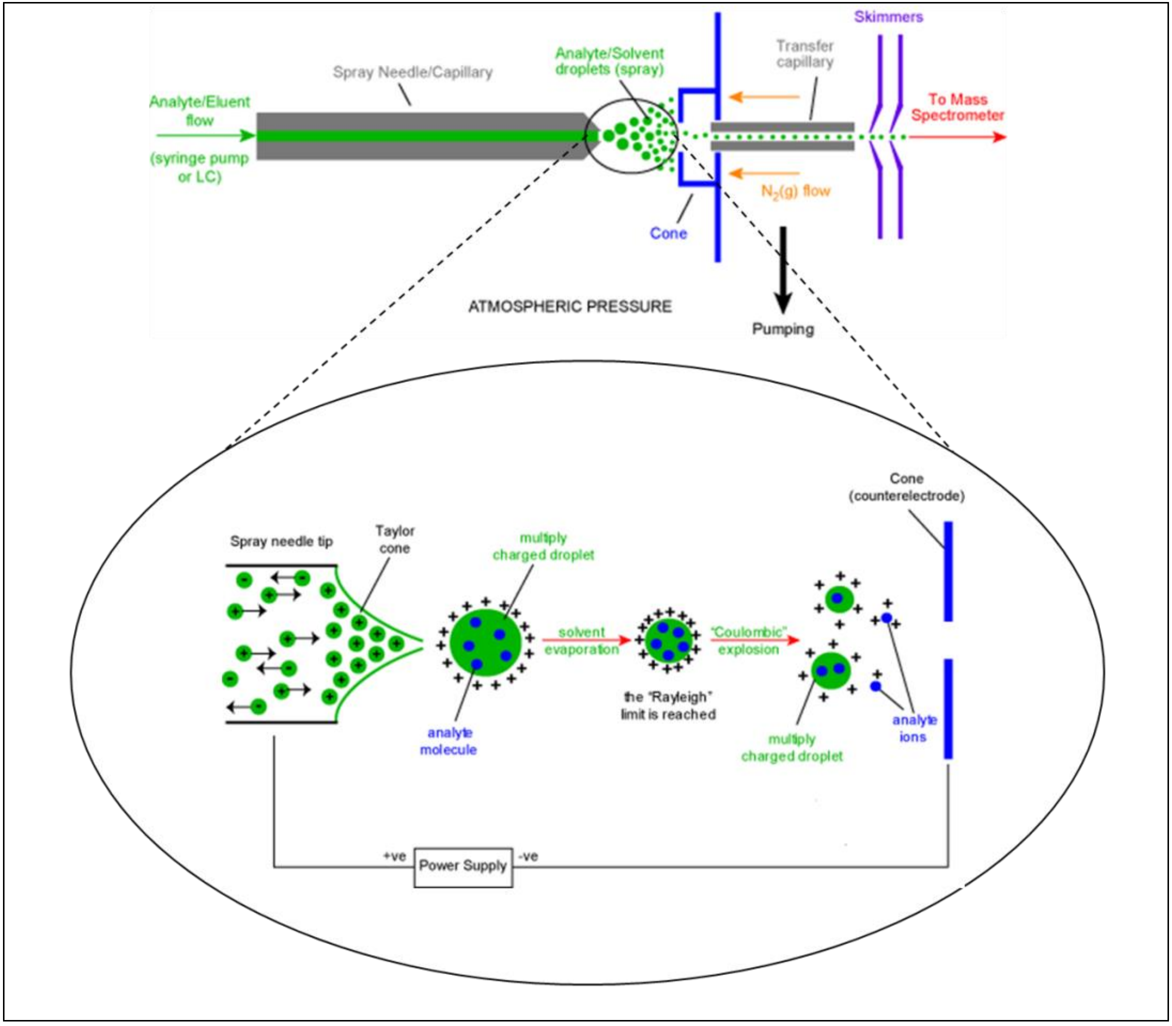


Figure 5.2 A schematic diagram of an ESI mechanism (Reference: School of Chemistry, University of Bristol)

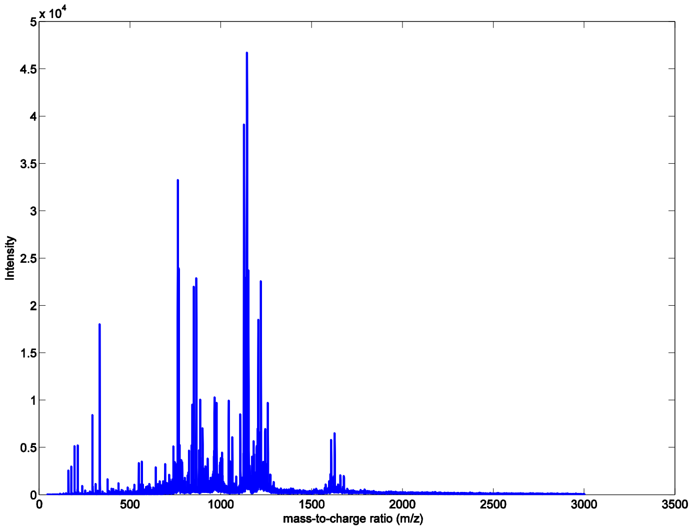


Figure 5.3 A raw ESI-MS spectrum from one of the CHO cell lines analysed in this Chapter

5.5 Data Preprocessing

A significant amount of data is produced by the liquid chromatography electrospray mass spectrometry (LC-ESI-MS) platform. In this study the order of the data point for each raw ESI-MS spectrum is more than 10^5. This necessitates preprocessing of the ESI-MS data prior to subsequent data analysis. Primarily, the goal of the preprocessing stage is to achieve an operational size data set and comparable spectra, enabling the application of multivariate analysis. The left diagram in Figure 5.4 shows a three-dimensional dataset established for each cell line replicate, that is, intensity of the ions (I), mass-to-charge ratio of ions as detected in the ESI-MS (m/z), and retention time (t_R). The three-dimensional dataset is concatenated to form a three-dimensional matrix $\underline{\boldsymbol{X}}$ whose convention is intensity of cell line (I of C_L) x mass-to-charge ratio (J) x retention time (K), where cell line, C =1 … L is shown in the right hand diagram of Figure 5.4. Retention time is defined as the elapsed time between the time of injection of a fluid and the time of elution of the maximum peak of that fluid i.e. the time from the column inlet to the detector. The files produced by the LC-ESI-MS platform were converted from the proprietary Bruker file format to a universal standard (mzML), using CompassExport.

The ESI-MS generated are of unequal length, therefore need to be aligned and binned to allow for comparison of multiple datasets from different samples or between replicates from the same sample. In aligning the spectra, the longest spectrum is selected as a reference thus the other spectra have to be upsampled by zero-padding to obtain equal length. Once aligned, the intensity of the ESI-MS were binned according to retention time and mass-to-charge ratio.

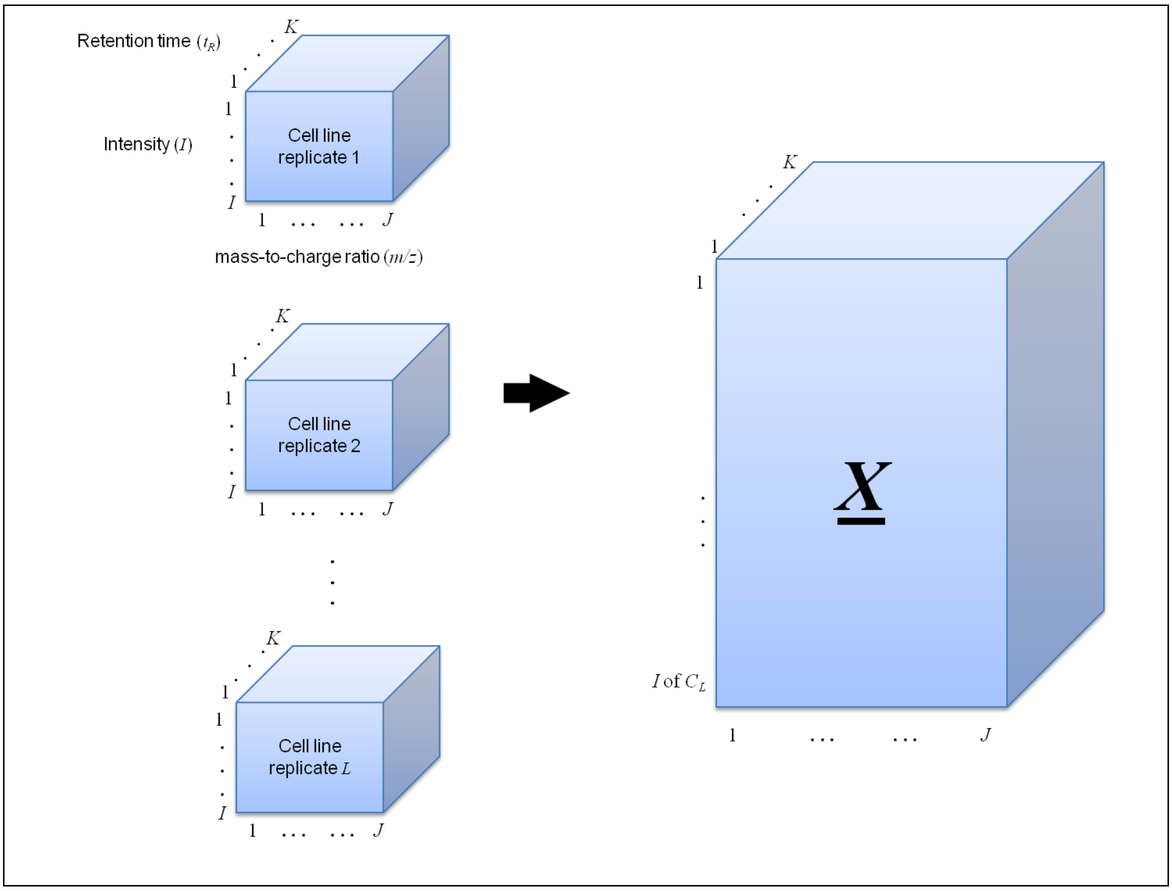


Figure 5.4 Three-dimensional data matrix of the ESI-MS data

5.5.1 *Binning*

Binning is one of the most widely applied preprocessing techniques to mass spectrometry data and is a standard approach in the analysis of electrospray ionisation-mass spectrometry (ESI-MS) data. The goal of the binning procedure is to extract information while performing data reduction prior to further data processing and analysis. The binning procedure was performed by dividing the retention time (elution time from the liquid chromatography system) and *m/z* range into equally spaced intervals. This is so that the intensity values are comparable across cell line replicates, *m/z* values, and retention time dimensions.

As discussed previously, the ESI-MS analysed in this study were collected by from Stage 6: small-scale bioreactor stage (Figure 5.1). The ESI-MS data generated is massive and highly complex. Using the whole spectral profiles may lead to problems including a large number of linearly correlated *m/z* ratios describing one profile and a shift in the *m/z* ratio between corresponding molecules in different spots i.e. alignment problem. Hence the binning approach will allow a compromise between full spectral profiling and peak integration (Arneberg et al., 2007). Peak integration is where adjacent *m/z* ratios are summarised throughout the spectrum and this requires the determination of an appropriate bin size that will ensure a good description of the full spectra with minimum information loss.

Determining the bin size is the critical stage in the binning procedure. If the bin size is too large, the binned data may have poor resolution of the peaks (Krishnan et al., 2012) and also there may also be the risk of different events registering in the same bin (Nielsen et al., 2010). Smith et al. (2006) suggested the use of overlapping adjacent bins instead of separate bins to avoid the risk of splitting a group of signals. Therefore, the bin size needs to be one that is comparable with the spectra peaks width in time and *m/z* ratio whilst retaining a good description of the features in the profiles (Fonville et al., 2011; Nielsen et al., 2010).

For the purpose of this study, the person in charge of the binning procedure considered the mass range of 100-2500 *m/z* with a bin size of 1 *m/z*. The time range selected was 0 – 2220 seconds with a bin size of 60 s. This range was selected following visual inspection of the ESI spectra showed that approximately 95% of the peaks were located in that range as shown in Figure 5.5.

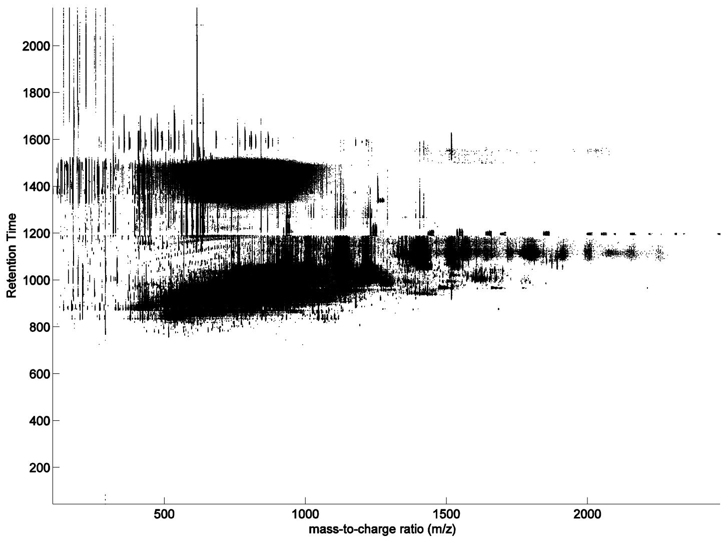


Figure 5.5 Peak intensities of the ESI-MS

Following the splitting of the retention time and *m/z* range into equally spaced bins, for each spectrum the raw intensities within the same *m/z* bin were summed to give a pair of intensity and mass-to-charge ratio value i.e. (*I*, *m/z*). This pair was then used to represent the relative intensities in that bin. On completion of the binning procedure, the three-dimensional data matrix $\underline{\boldsymbol{X}}$ was unfolded to a two dimensional data matrix $\boldsymbol{X}$ (C_L x (I x J x K)) as shown in Figure 5.6. The rows of $\boldsymbol{X}$ represents cell lines and their replicates and the columns represent the intensity of the ESI spectra after binning on retention time and *m/z* ratio. This method of unfolding allows the variability in the data to be observed as it accentuates the similarities and dissimilarities between cell lines and also cell replicates. The unfolded data matrix $\boldsymbol{X}$ consists of 50 cell replicates from 19 cell lines. Table 5-1 provides the cell line information used in this study including the three product quality parameters: Product Quality 1 (PQ1), Product Quality 2 (PQ2), and Product Quality 3 (PQ3). The cut-off value for PQ1 is 0.59, PQ2 is 0.53 and PQ3 is 0.34. A cell line is considered as high producing when all three of its product qualities are above the cut-off values. It is noted that the product quality information for PQ1, PQ2, and PQ3 has been normalised due to legality issue of restricted information.

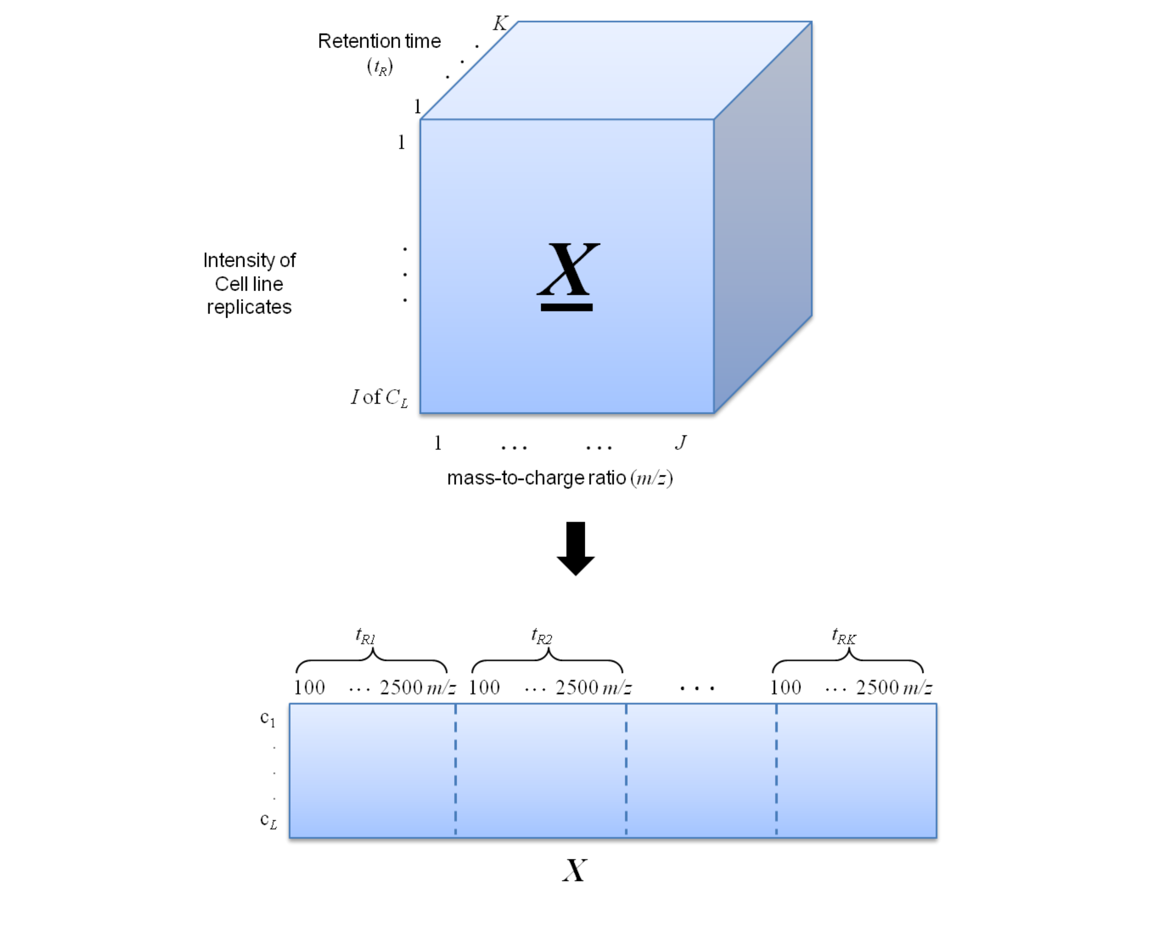


Figure 5.6 Unfolding of the three dimensional ESI-MS data set to two dimensional matrix after binning on retention time and mass-to-charge ratio

Table 5-1 Cell lines information and their corresponding product qualities

High/ Low Producing Cell	Cell Lines	Number of replicates	Product Quality 1 (PQ1)	Product Quality 2 (PQ2)	Product Quality 3 (PQ3)
High Producing Cell	CL10	2	0.71	0.83	0.40
	CL11	3	0.76	0.54	0.61
	CL12	5	1.00	0.67	0.69
	CL13	3	0.68	0.85	0.38
	CL16	3	0.66	0.77	0.40
Low Producing Cell	CL19	2	0.70	1.00	0.34
	CL2	2	0.38	0.71	0.24
	CL6	3	0.13	0.68	0.08
	CL8	3	0.09	0.68	0.05
	CL15	2	0.30	0.68	0.29
	CL17	3	0.35	0.58	0.26
	CL18	3	0.37	0.65	0.25
	CL1	2	0.44	0.11	0.81
	CL3	2	0.37	0.00	1.00
	CL4	2	0.52	0.47	0.47
	CL5	2	0.55	0.49	0.47
	CL9	2	0.54	0.20	0.79
	CL7	3	0.00	0.42	0.00
	CL14	3	0.35	0.45	0.32

5.6 Model Development

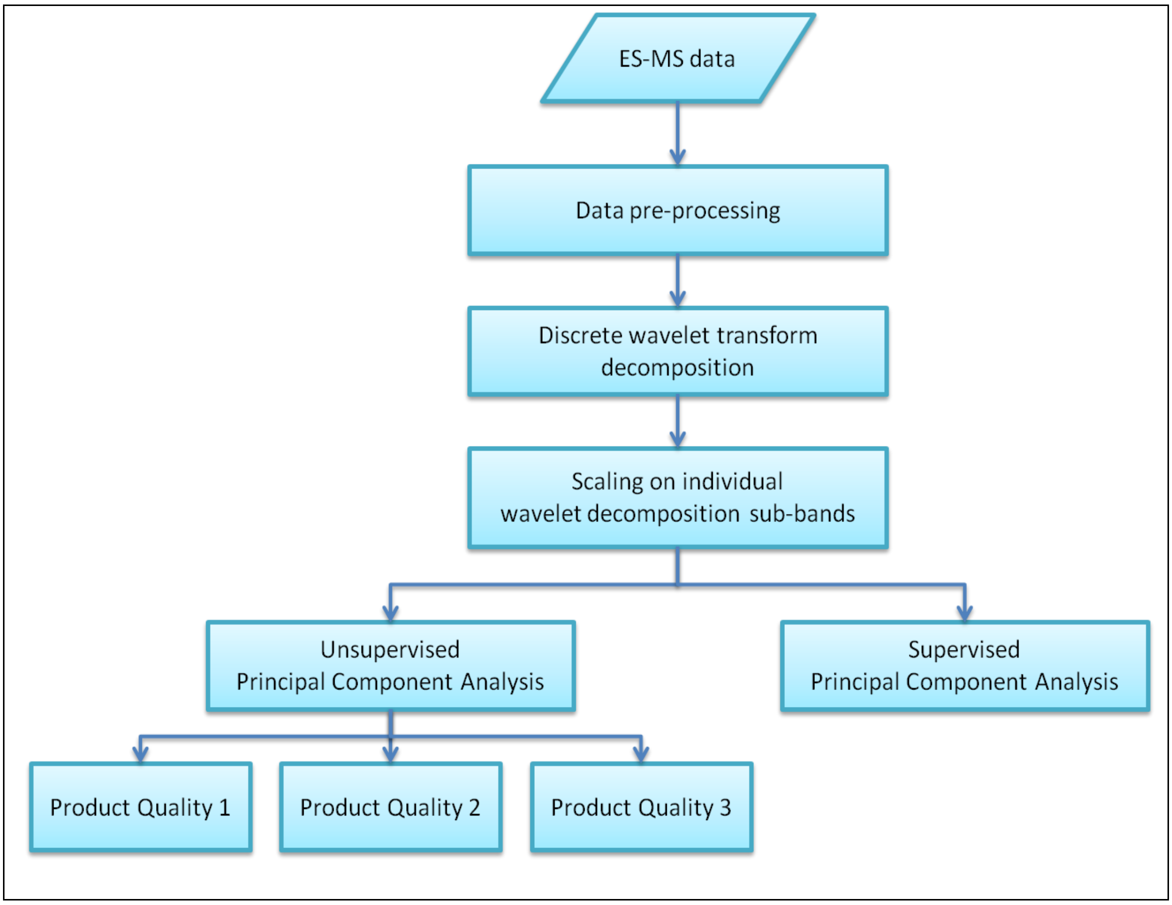

ES-MS data
Data pre-processing
Discrete wavelet transform decomposition
Scaling on individual wavelet decomposition sub-bands
Unsupervised Principal Component Analysis
Supervised Principal Component Analysis
Product Quality 1
Product Quality 2
Product Quality 3

Figure 5.7 Schematic of model development

Following the successful results from the NIR batch data fingerprinting approach (Section 4.10), the concept of integrating the discrete wavelet transform with MPCA as a method to analyse complex data was again implemented in this chapter. The goal of this study is to apply the concept to extract information from the ESI-MS dataset that would enable the characterisation of high and low cell line producers.

Figure 5.7 describes a schematic of the model development which consists of five stages. At the first stage, the ESI-MS dataset was pre-processed as has been discussed in Section 5.5. Once pre-processed, the discrete wavelet transform decomposition is applied to the ESI-MS dataset. Details of this stage are explained in Section 5.6.1. Following this, Section 5.6.2 describes the data scaling scheme performed to the wavelet sub-bands generated from the application of the discrete wavelet transform decomposition. Finally, unsupervised and supervised PCA are performed on the ESI-MS dataset and is discussed in Section 5.6.3. Results generated from both applications are each discussed in Section 5.7 and Section 5.8 respectively.

5.6.1 *Application of Wavelet to ESI-MS Data*

This section provides a detailed description of the application of the discrete wavelet transform decomposition to the ESI-MS dataset and how the method of analysis has been developed. As per the previous case study, the ESI-MS data was analysed using wavelets from the Daubechies family 5 (db5) with five levels of decomposition. This combination is selected as the analysing wavelet and decomposition levels because it has been proven to efficiently extract information in the previous case study. This transformation procedure which is also known as multiresolution wavelet decomposition decomposed each spectrum into five sub-bands of approximation spectra components (A_1 to A5) and five sub-bands of detail spectra components (D1 to D5). It is shown in Figure 5.8 that for 5 levels of decomposition, the multiresolution wavelet decomposition of the ESI-MS data has reduced the binned spectra to the finest approximation level (A5) by 95%.

Previously in Chapter 4, standard deviations of the wavelet coefficients from the sub-bands A5, D5, D4, D3, D2 and D1 were calculated and used to represent each NIR spectrum. In this study all wavelet coefficients in the sub-bands A5, D5, D4, D3, D2 and D1 were selected for the analysis of the ESI-MS data with each sub-band individually.

Prior to making the decision to use all wavelet coefficients in the sub-bands, an initial investigation is performed to determine the combination of discrete wavelet transform and PCA that is most suitable to analysis the ESI-MS data. The investigation involved the application of an unsupervised PCA to all wavelet coefficients and to the standard deviations of the wavelet coefficients. Both PCA applications are performed on combined sub-bands, which is similar to the application of PCA in Chapter 4.

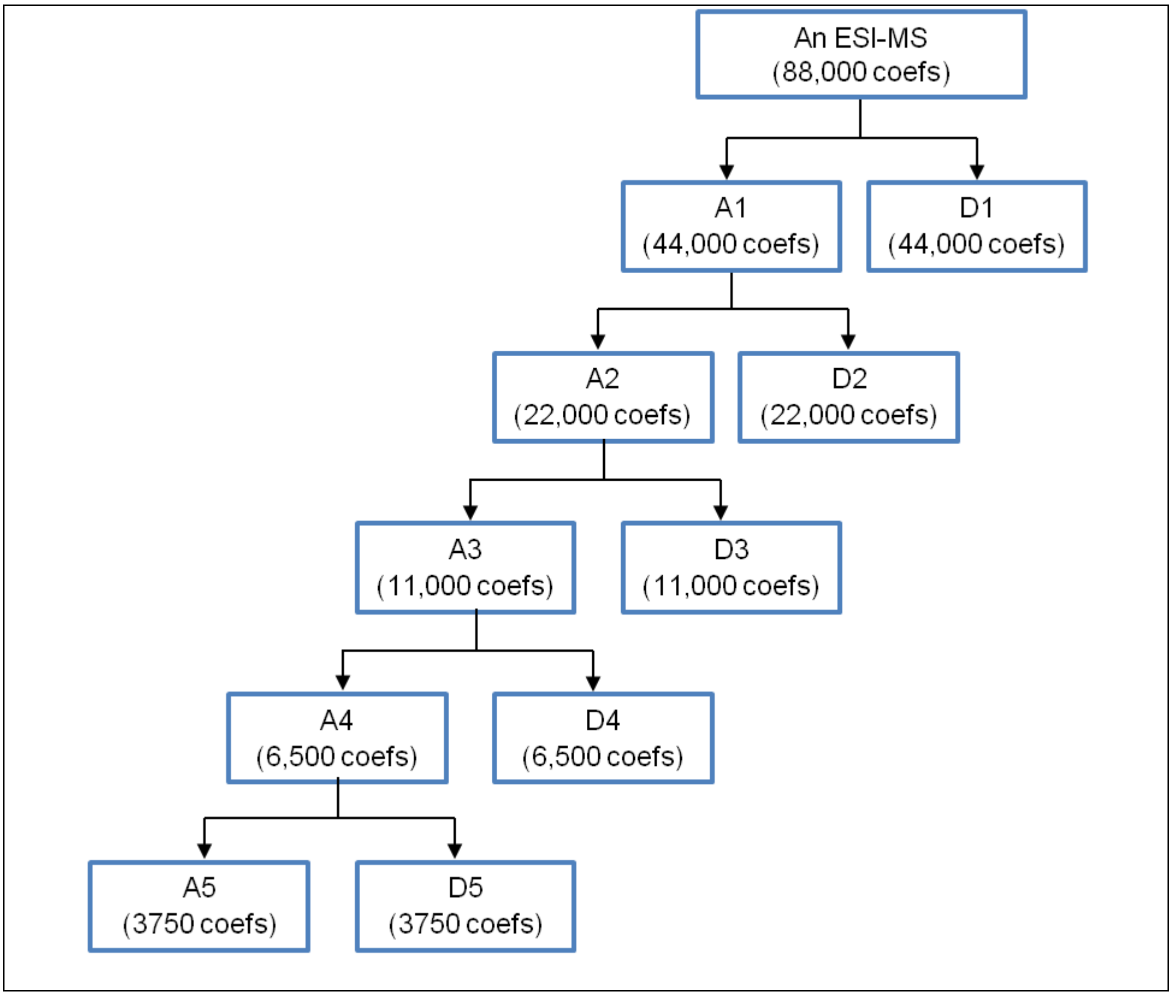


Figure 5.8 Multiresolution wavelet decomposition of an ESI-MS

A plot of the percentage variance explained for the former application is shown in Figure 5.9. For the latter application 100% of the variance of the standard deviation of the wavelet coefficients is explained by principal component 1 (PC1) when PCA was performed on the standard deviation of the wavelet coefficients. Meanwhile, ten principal components were required to explain approximately 50% of the variance in the wavelet coefficients. Based on the result of this investigation, it is hypothesised that potential information in the data is best to be extracted from all wavelet coefficients than from the standard deviation of the wavelet coefficients.

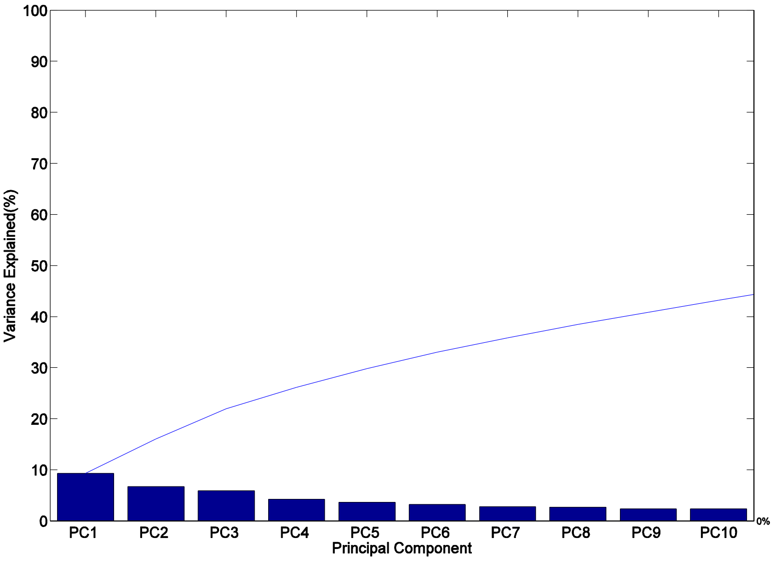


Figure 5.9 Percentage of variance explained for the application of PCA to all wavelet coefficients

5.6.2 *Data Scaling*

The ESI-MS was subjected to not only biological variation but also experimental and instrumental variation which affects the level of spectral response intensity (Fonville et al., 2011). Varying amounts of sample, degradation in the sample or variations in the instrument detector sensitivity are some examples of these variations. Scaling of the data removes the unwanted variations between measurements whilst retaining the meaningful biological variation (Katajamaa and Oresic, 2007). Therefore the sub-bands A5, D5, D4, D3, D2 and D1 acquired from wavelet decomposition were subjected to data scaling prior to subsequent analysis. Since the data has been decomposed into sub-bands previously, each sub-band is scaled individually. The data scaling of choice is standardisation, where the data is transformed to have zero mean and unit variance.

$$x_{standardised} = \frac{x - mean(x)}{std(x)} \tag{5.1}$$

Calculation of x_{stand} requires the mean and standard deviation of the sub-band. Figure 5.10(a) shows the sub-band A5 extracted from the two-dimensional data matrix following the application of the discrete wavelet transform. The columns of sub-band A5 are the wavelet coefficients relative to the intensity binned on retention time and mass-to-charge ratio whereas the rows are the cell lines and replicates. The statistical measurements can be calculated in two ways. They can either be calculated for an individual spectrum component in each sub-band, Figure 5.10(b), i.e. approach 1 or as illustrated in Figure 5.10(c), approach 2, they can be determined for a particular relative intensity binned on retention time and *m/z* ratio across all spectra for every in the sub-band. Approach 1 will combine the effect of each wavelet coefficient whilst approach 2 will combine the effect of each sample.

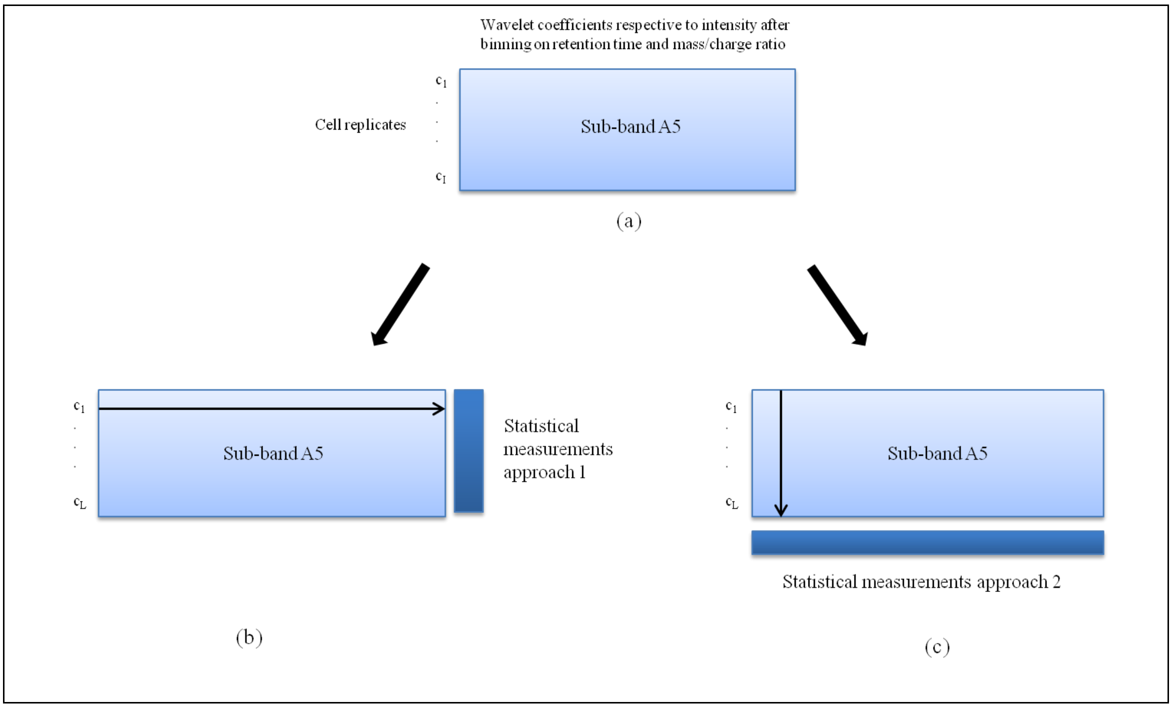

Wavelet coefficients respective to intensity after binning on retention time and mass/charge ratio
Cell replicates
c_1
c_I
Sub-band A5
(a)
c_1
c_L
Sub-band A5
Statistical measurements approach 1
(b)
c_1
c_L
Sub-band A5
Statistical measurements approach 2
(c)

Figure 5.10 (a) Sub-band A5, (b) statistical measurement of approach 1, and (c) statistical measurement of approach

Equations 5.2 and 5.3 describe the calculations using approach 1 whilst equations 5.5 and 5.6 describe the calculation using approach 2. In approach 1, the mean is given by

$$\overline{x_{M1}} = \frac{\sum_{i=1}^{N}(x_1 + x_2 + \cdots + x_N)}{N} \tag{5.2}$$

where

$\overline{x_{M1}} =$ mean of the spectrum component , approach 1

$x_i =$ wavelet coefficient i (i=1 … N) of the spectrum component in the sub-band

$N =$ number of wavelet coefficients of the spectrum component in the sub-band

and the standard deviation is given by

$$s_{M1} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x_{M1}})^2}{N-1}} \tag{5.3}$$

where

$s_{M1} =$ standard deviation of the spectrum component

$x_i =$ wavelet coefficient i (i=1 … N) of the spectrum component in the sub-band

$\overline{x_{M1}} =$ mean of the spectrum component, approach 1

$N =$ number of wavelet coefficients of the spectrum component in the sub-band.

In approach 2, the mean is given by

$$\overline{x_{M2}} = \frac{\sum\left(x_{1_{spectra\,1}} + x_{1_{spectra2}} + \cdots + x_{1_{spectraN}}\right)}{N} \tag{5.4}$$

where

$\overline{x_{M2}} =$ mean of the spectrum component, approach 2

$x_i =$ wavelet coefficient i (i=1 … N) of each spectra in the sub-band that correspond to a particular intensity

$N =$ number of cell lines replicates

and the standard deviation is given by

$$s_{M2} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_{i_{spectra\,i}} - \overline{x_{M2}}\right)^2}{N-1}} \tag{5.5}$$

where

$s_{M2} =$ standard deviation in the sub-band set using approach 2

$x =$ wavelet coefficient i (i=1 … N) of each spectra in the sub-band that correspond to a particular intensity

$\overline{x_{M2}} =$ mean of the spectrum component , approach 2

$N =$ number of cell lines replicates

Examples of the above techniques are shown in Figure 5.11 to Figure 5.14. Figure 5.11 shows a sample of the binned spectrum whilst Figure 5.12 shows the raw data for sub-band A5. Figure 5.13 and Figure 5.14 illustrate the plots of the standardised sub-band A5 using approaches 1 and 2 respectively.

A significant difference is observed between standardisation of the sub-band for the two approaches. The mean trajectory was removed from the spectrum for approach 2, Figure 5.14. However, in approach 1, the mean trajectory remains in the data as the trace in Figure 5.13 shows the same pattern as the raw data. Following this investigation, it was decided that standardisation using statistical measurement of approach 2 will be employed to scale the sub-bands.

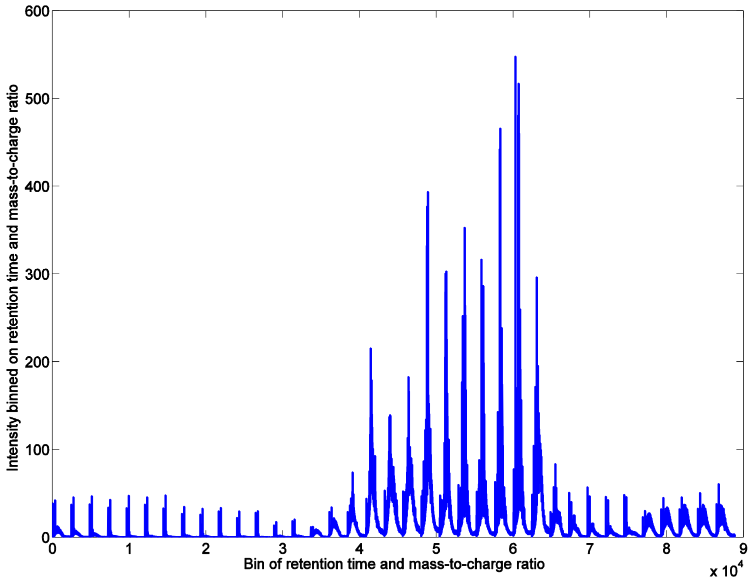


Figure 5.11 A sample of the binned spectrum

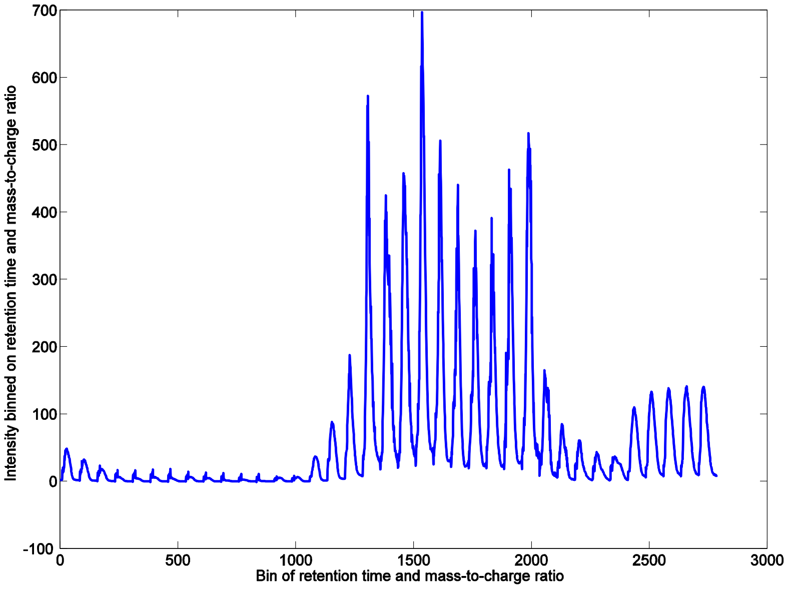


Figure 5.12 Sample spectrum of the sub-band A5

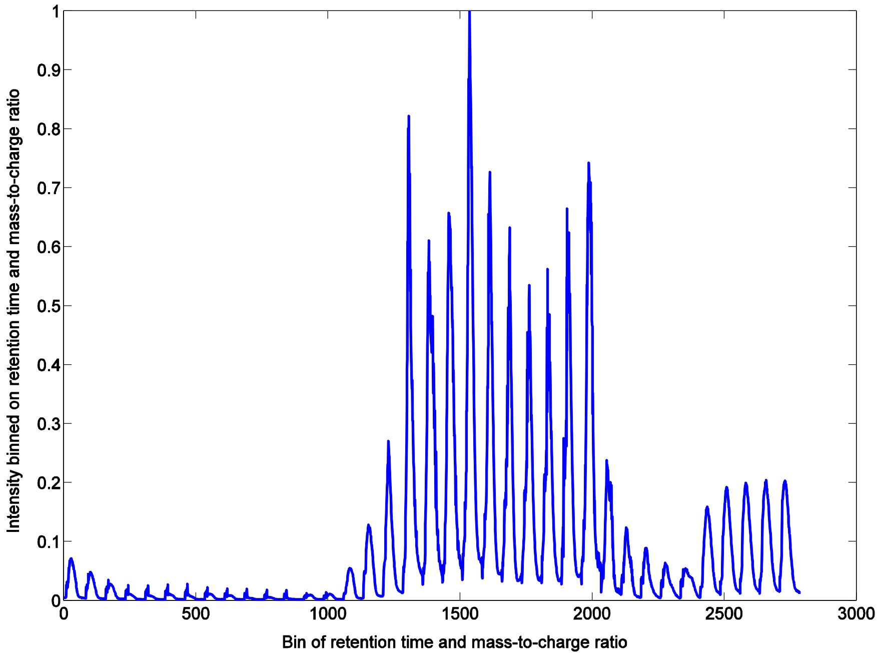


Figure 5.13 Standardising using statistical measurement of approach 1

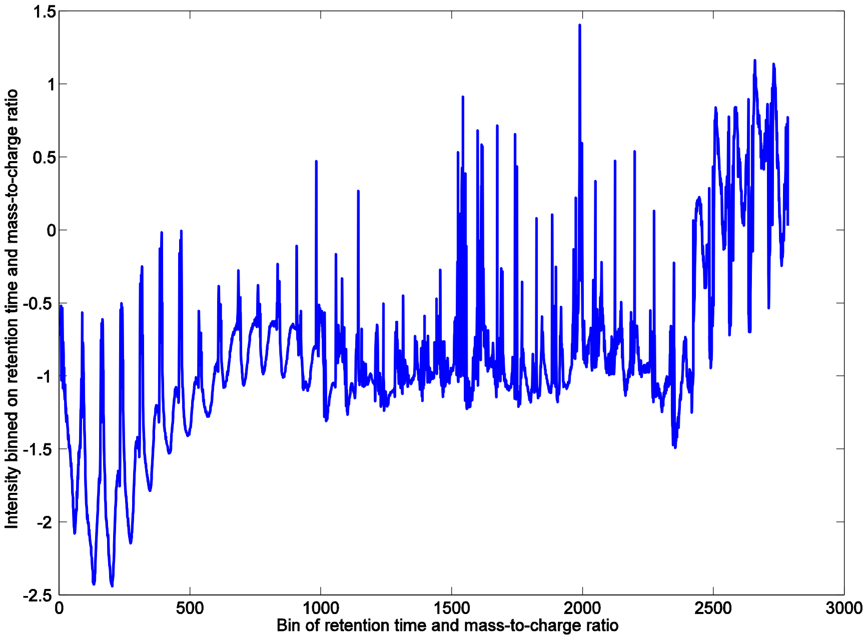


Figure 5.14 Standardising using statistical measurement of approach 2

5.6.3 *Application of Principal Component Analysis*

Once the sub-bands of the ES-MS data had been standardised, they were subjected to the application of principal component analysis. The multiresolution property of the discrete wavelet transform decomposition enables the application of PCA to the individual sub-bands realising the closer examination of the data at different resolutions. The PCA approach comprised two stages: an unsupervised analysis and a supervised analysis. In the unsupervised approach prior knowledge about the data was not used whereas in the supervised analysis the data were categorized into user-defined groups prior to analysis.

The unsupervised approach was adopted for the purpose of identifying hidden patterns (Lhoest et al., 2001), extracting and explaining key features of the data (Borah et al., 2007), and to identify those wavelet sub-bands making the largest contribution to the model. The supervised methodology was performed by superimposing the test set on the model built using the training set to project its behaviour.

5.7 Results and Discussion: Unsupervised Representation

After the application of the discrete wavelet transform decomposition to the data matrix $\boldsymbol{X}$, followed by standardisation of each sub-band, PCA was applied to each sub-band. As shown in the scree plots of the sub-bands (Figure 5.15) the number of principal components (PC) retained differs for each sub-band. Table 5-2 summarizes the number of principal component retained in each sub-band.

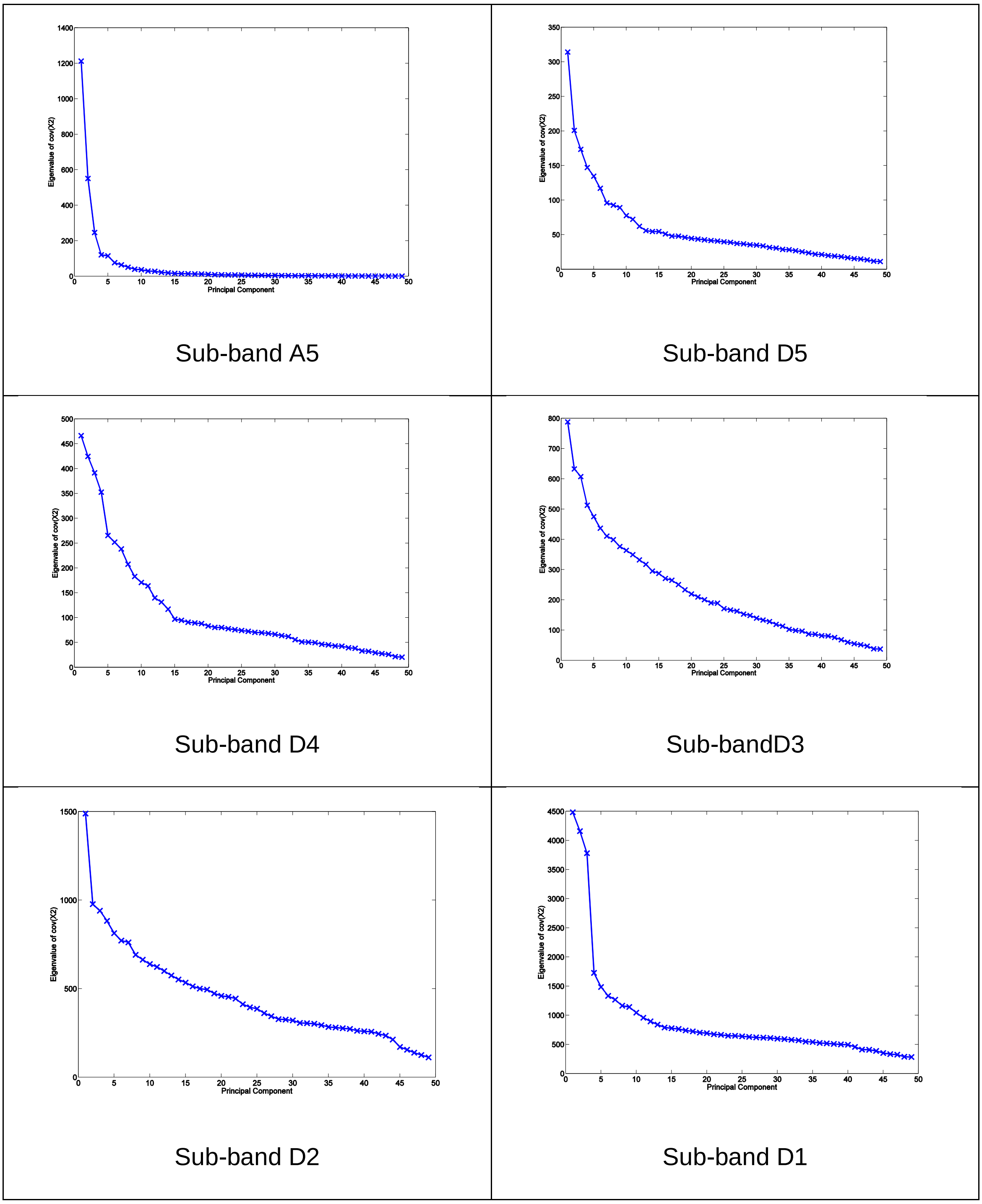
Table 5-2 Number of principal component retained in each sub-band and combined sub-bands

Sub-band	Number of principal component retained	Variance explained (%)
A5	6	83.8
D5	6	39.06
D4	14	63.02
D3	14	56.66
D2	6	17.39
D1	6	41.05
Combined	14	51.80

The bivariate scores plots for the combined sub-bands and for each sub-band are shown in Figure 5.16 to Figure 5.18 (combined sub-bands), Figure 5.19 (sub-band A5), Figure 5.20 (sub-band D5), Figure 5.21 (sub-band D4), Figure 5.22 (sub-band D3), Figure 5.23 (sub-band D2), and Figure 5.24 (sub-band D1). For the combined sub-bands the first fourteen PC were retained based on the highest number of PC retained in two of the sub-bands.

As can be seen from the bivariate scores plot of the combined sub-bands in Figure 5.16 (a), the cell replicates from the high and low cell producers were tightly clustered and located at the origin of the PC1-PC2 space. Interestingly, the bivariate scores plots of the wavelet detail sub-bands D5, D4, D3, D2, and D1 also show relatively similar patterns. Possible outliers in the bivariate scores plot of the combined sub-bands and each sub-band are identified and presented in Table 5-3. The number following the dash after the cell line ID represents the cell line replicate.

The same cell line replicates, CL6-3, CL10-2, and CL11-2, are identified as outliers in the bivariate scores plot of PC1 vs. PC2 in the combined sub-band and the sub-bands A5 and D1. Based on all three product quality parameters, cell lines CL10 and CL11 were categorised as high producer whilst CL6 was a low producer.



Sub-band A5

Sub-band D5

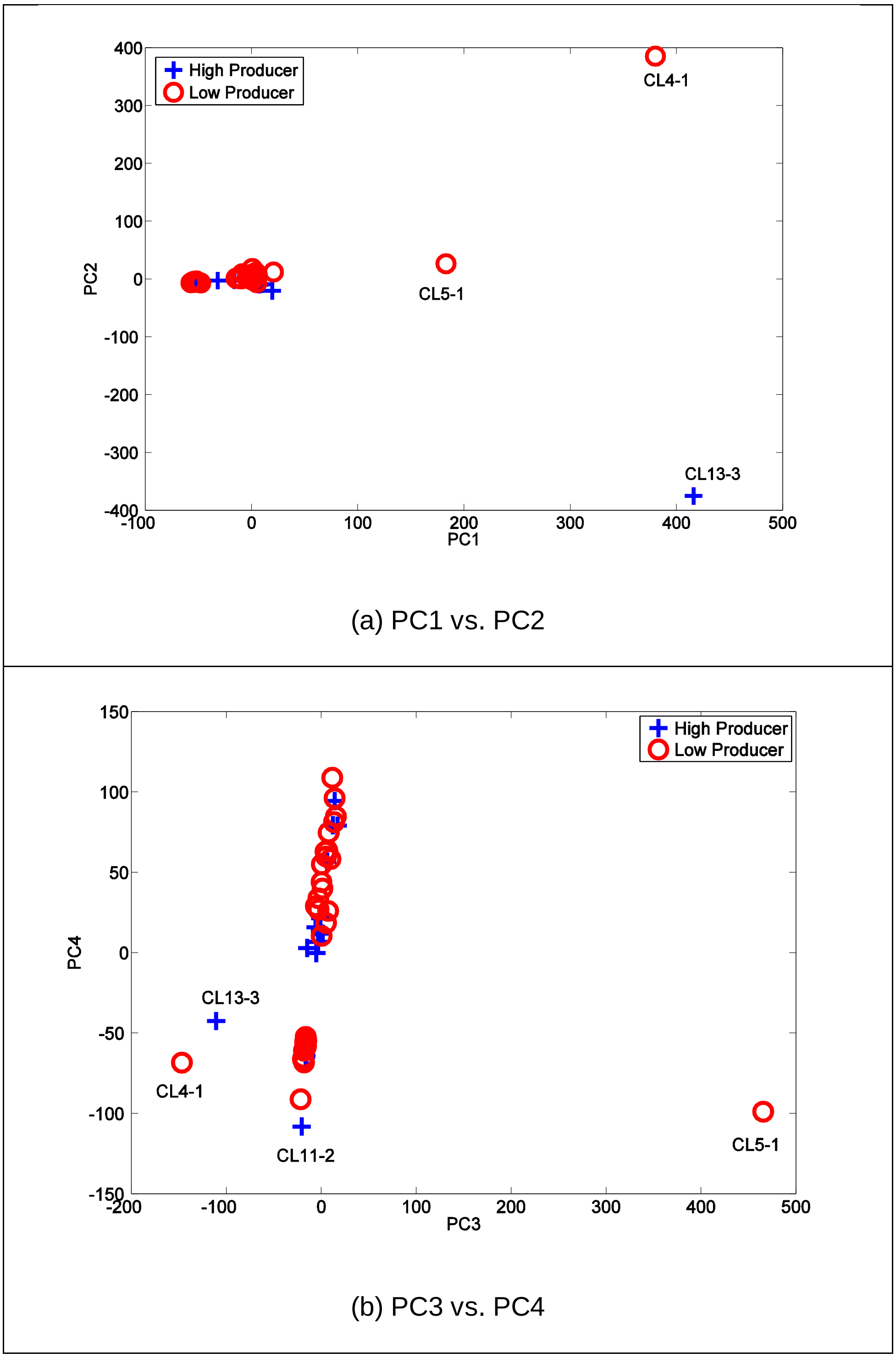
Sub-band D4

Sub-bandD3

Sub-band D2

Sub-band D1

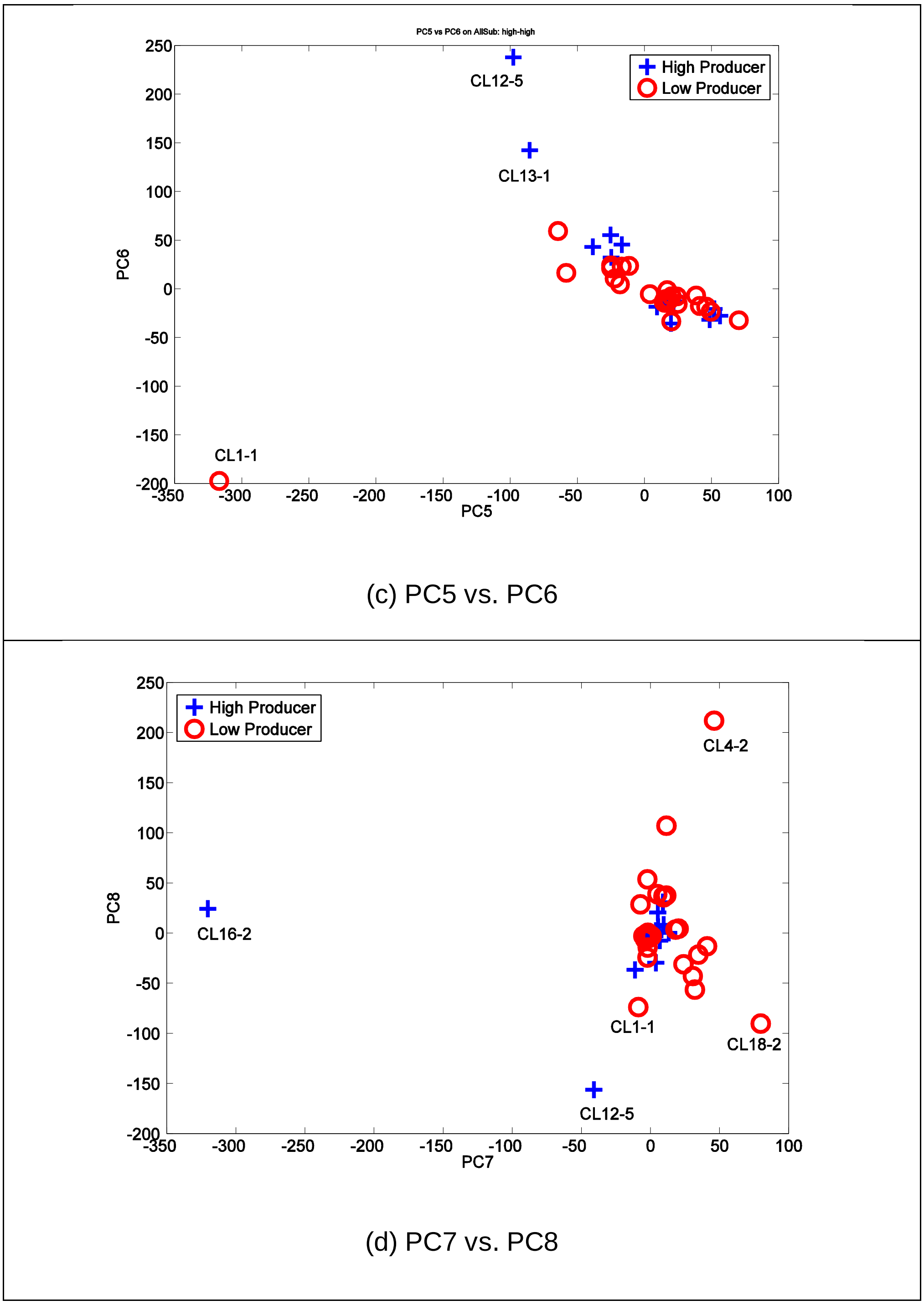
Figure 5.15 Scree plots for wavelet sub-bands A5, D5, D4, D3, D2 and D1



(a) PC1 vs. PC2

(b) PC3 vs. PC4

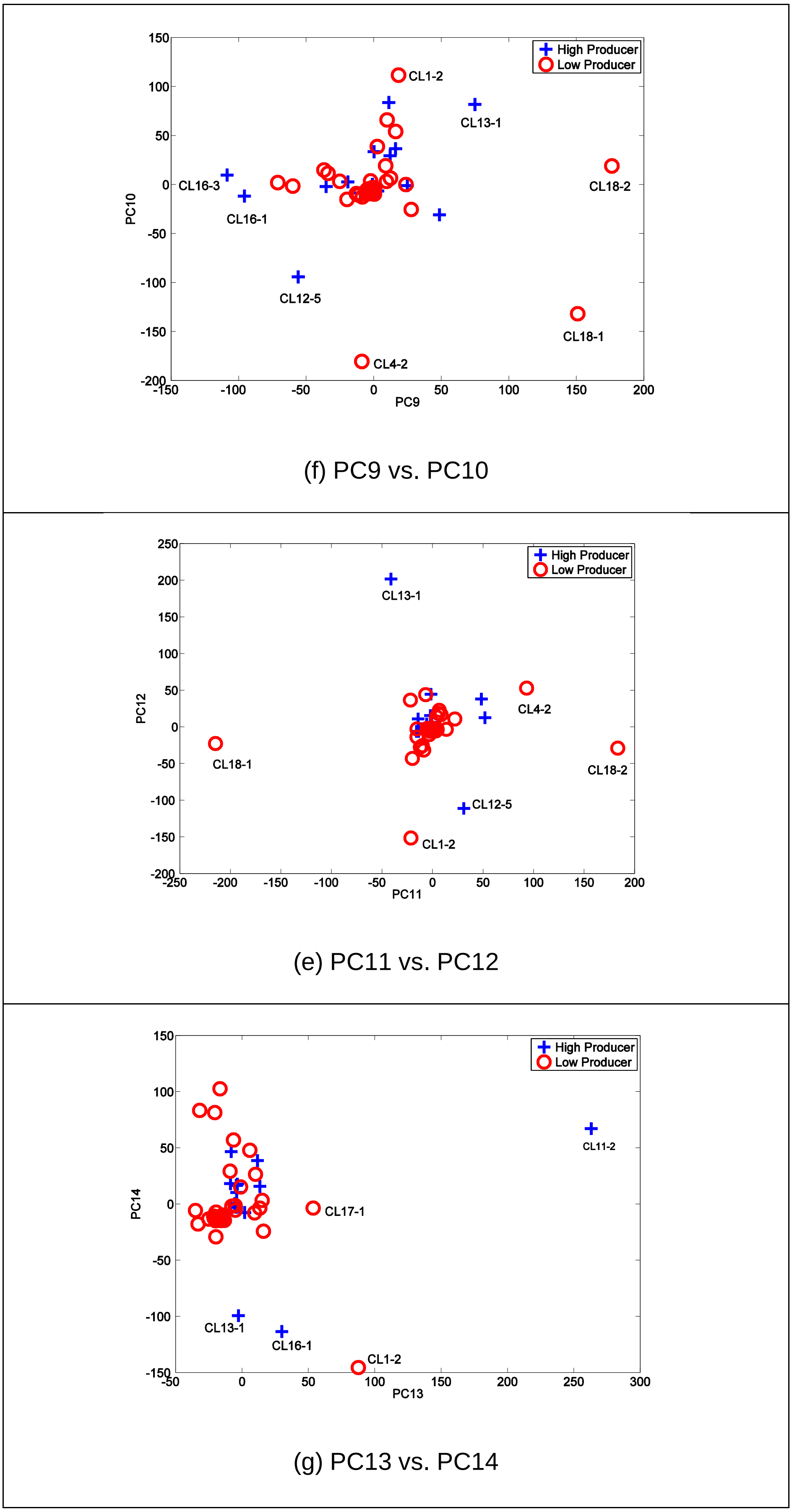
Figure 5.16 Bivariate scores plot of the combined sub-bands: PC1 vs. PC2 and PC3 vs. PC4



(c) PC5 vs. PC6

(d) PC7 vs. PC8

Figure 5.17 Bivariate scores plot of the combined sub-bands: PC5 vs. PC6 and PC7 vs. PC8

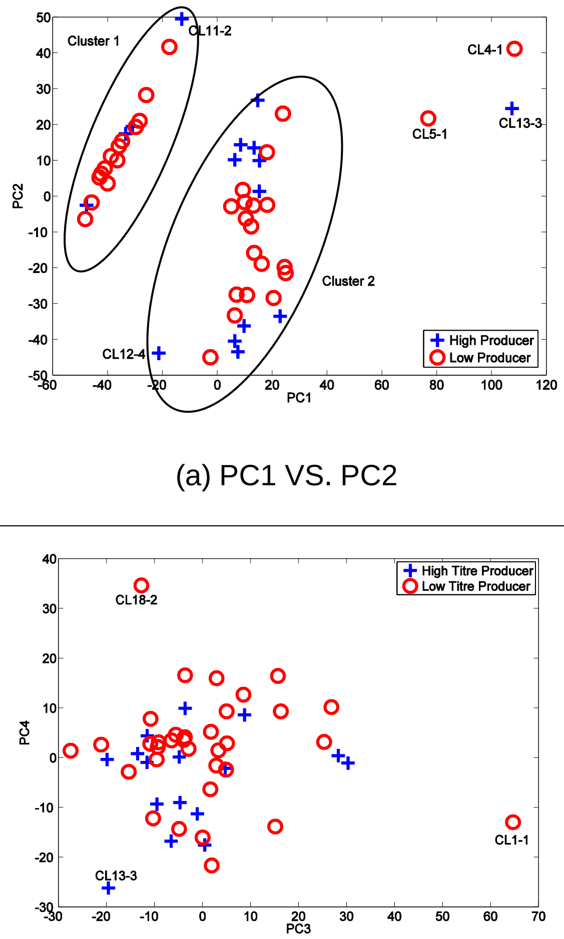


(f) PC9 vs. PC10

(e) PC11 vs. PC12

(g) PC13 vs. PC14

Figure 5.18 Bivariate scores plot of the combined sub-bands: PC9 vs. PC10, PC11 vs. PC12, and PC13 vs. PC14

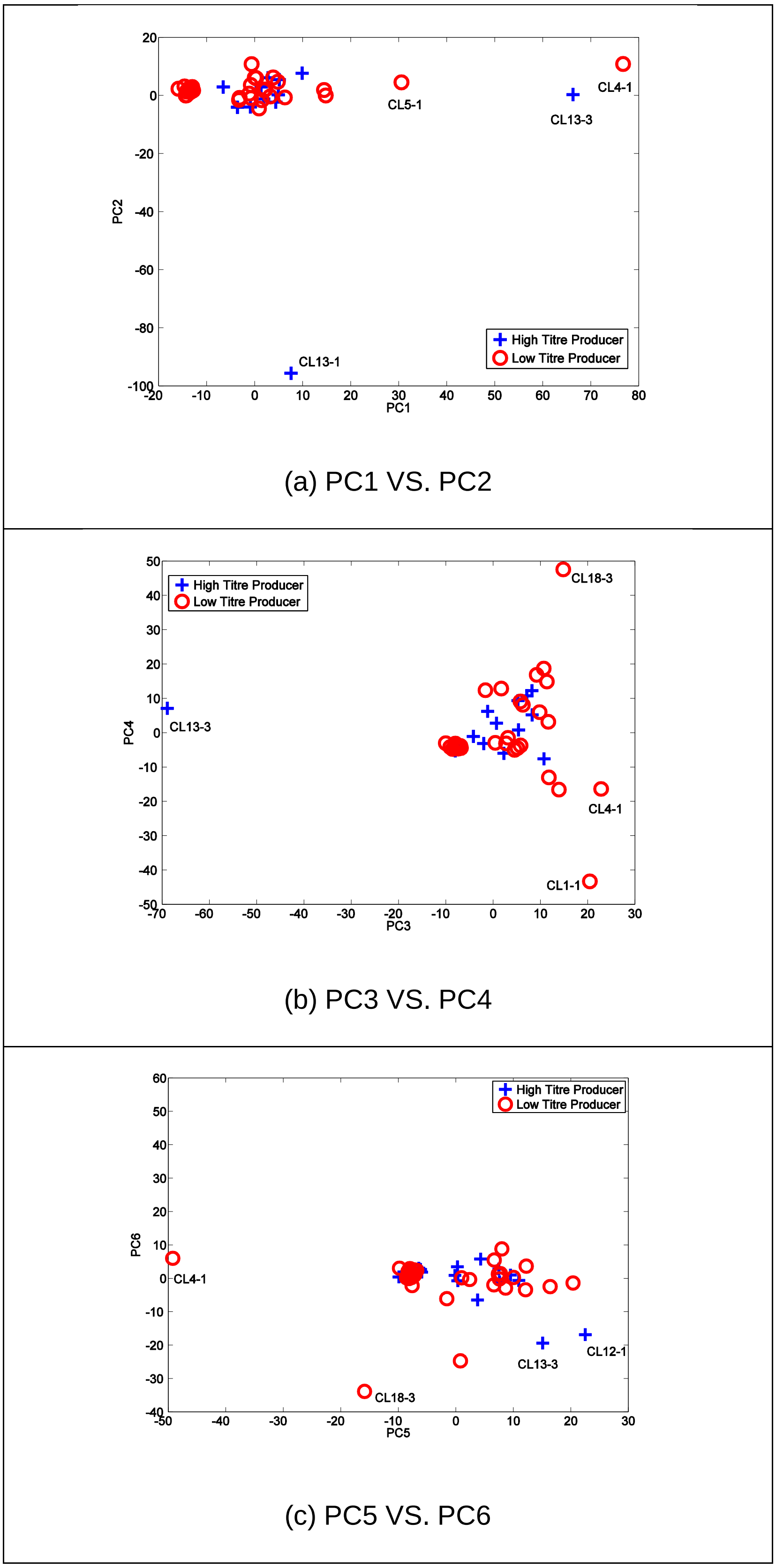


(a) PC1 VS. PC2

(b) PC3 VS. PC4

(c) PC5 VS. PC6

Figure 5.19 Bivariate scores plot of the sub-band A5

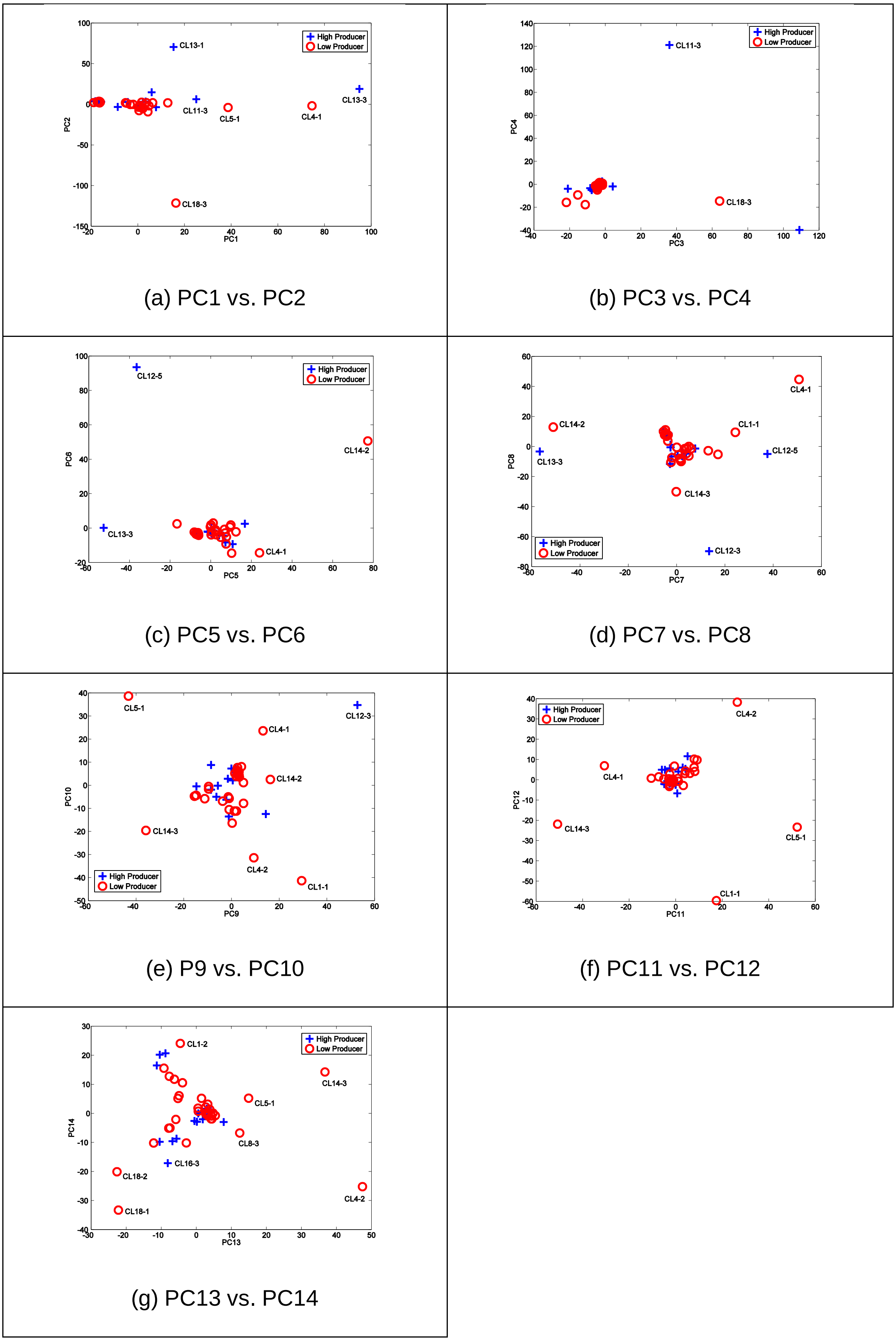


(a) PC1 VS. PC2

(b) PC3 VS. PC4

(c) PC5 VS. PC6

Figure 5.20 Bivariate scores plots of the sub-band D5



(a) PC1 vs. PC2

(b) PC3 vs. PC4

(c) PC5 vs. PC6

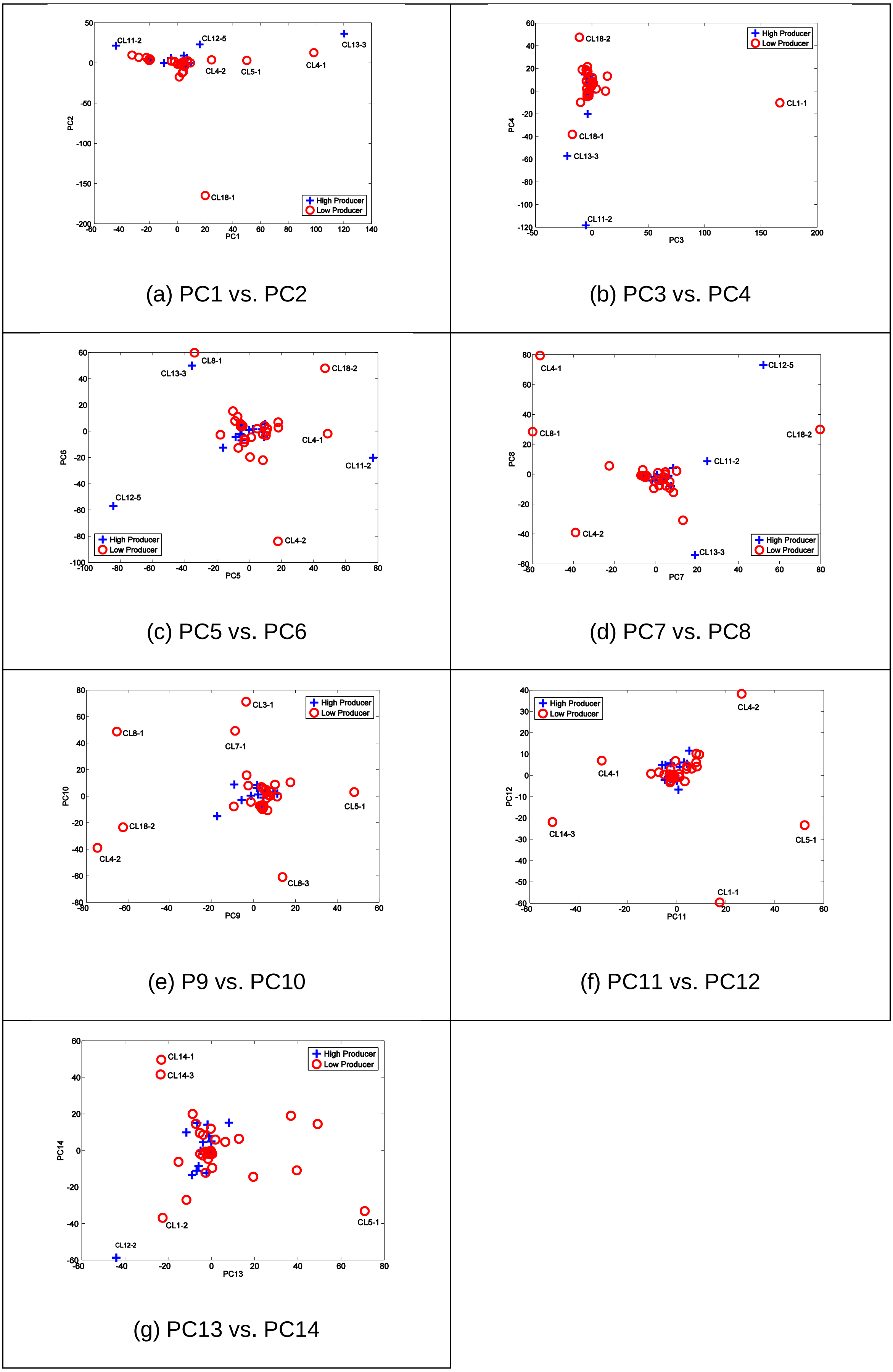
(d) PC7 vs. PC8

(e) P9 vs. PC10

(f) PC11 vs. PC12

(g) PC13 vs. PC14

Figure 5.21 Bivariate scores plots of the sub-band D4



(a) PC1 vs. PC2

(b) PC3 vs. PC4

(c) PC5 vs. PC6

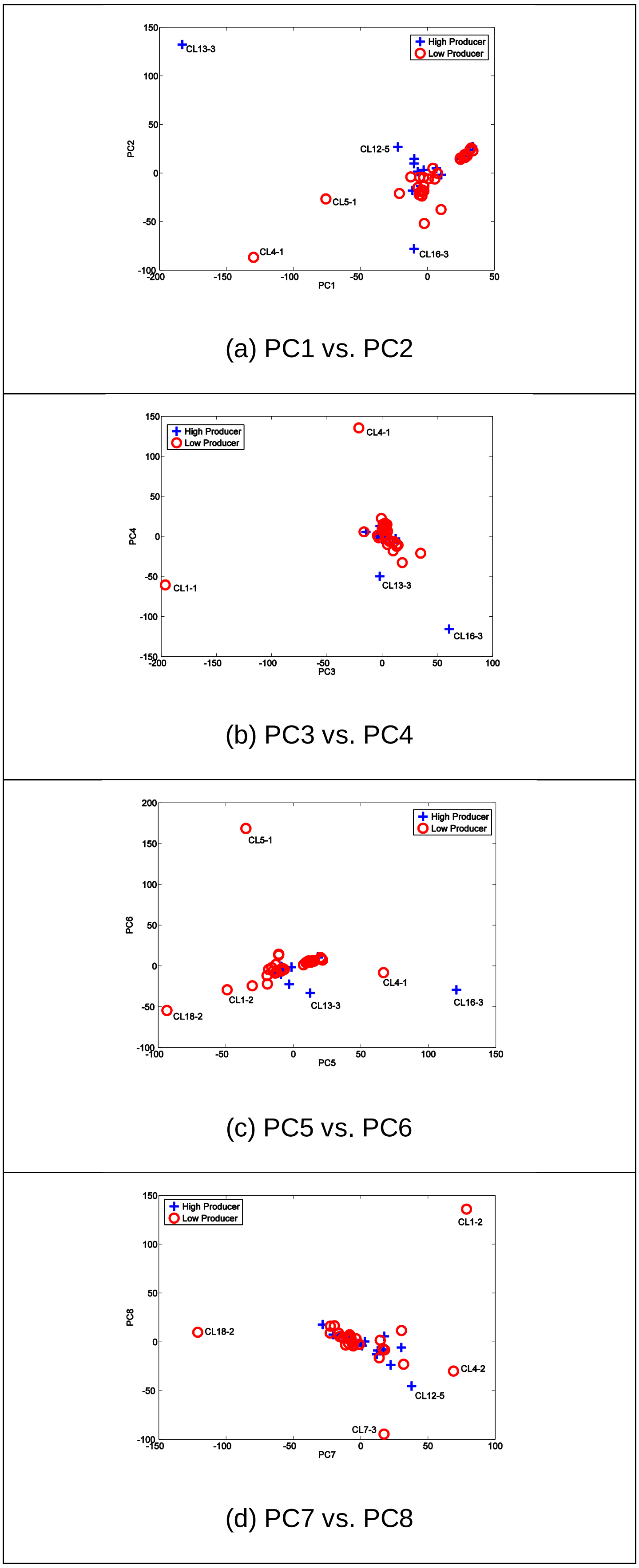
(d) PC7 vs. PC8

(e) P9 vs. PC10

(f) PC11 vs. PC12

(g) PC13 vs. PC14

Figure 5.22 Bivariate scores plot of the sub-band D3



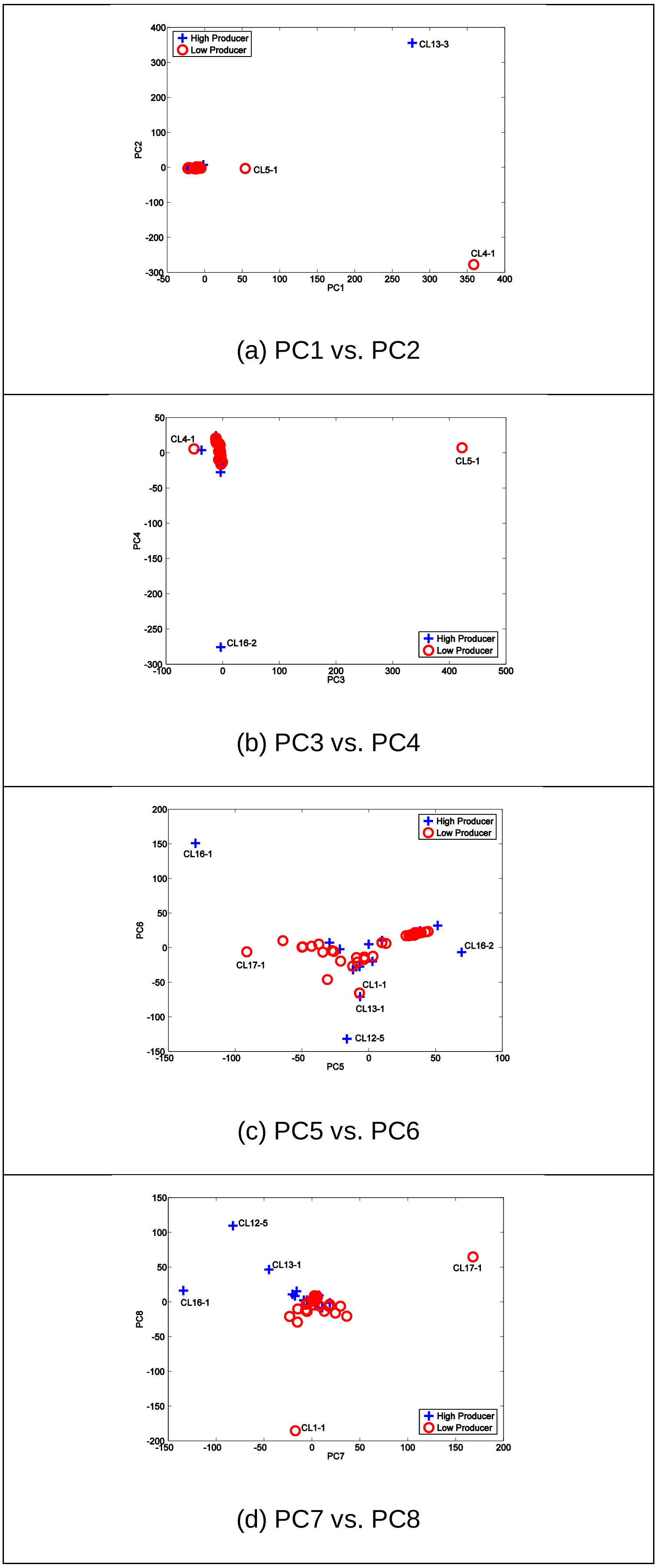
(a) PC1 vs. PC2

(b) PC3 vs. PC4

(c) PC5 vs. PC6

(d) PC7 vs. PC8

Figure 5.23 Bivariate scores plot of the sub-band D2



(a) PC1 vs. PC2

(b) PC3 vs. PC4

(c) PC5 vs. PC6

(d) PC7 vs. PC8

Figure 5.24 Bivariate scores plots of the sub-band D1

Table 5-3 Outliers in the retained principal components for the combined sub-bands and each sub-band

Bivariate scores plot	Outlier						
	Combined sub-bands	A5	D5	D4	D3	D2	D1
PC1 vs. PC2	CL6-3 CL10-2 CL11-2	CL6-3 CL1-2 CL11-2	CL6-2 CL6-3 CL10-2	CL6-2 CL6-3 CL8-2 CL10-2	CL6-3 CL7-3 CL10-2	CL6-3 CL11-2	CL6-3 CL10-2 CL11-2
PC3 vs. PC4	CL11-2	CL8-1 CL8-3	CL6-3 CL8-2 CL8-3	CL3-1 CL6-2 CL8-2	CL2-2 CL8-1 CL8-3	CL8-3 CL7-3 CL10-2	CL7-2 CL11-2
PC5 vs. PC6	CL5-1 CL6-2 CL8-3	None	CL8-2 CL10-2 CL11-2	CL5-1 CL6-3 CL13-2	CL2-2 CL4-2 CL5-1 CL6-3 CL8-1 CL10-2 CL11-1	CL7-3 CL8-1 CL11-2	CL5-1 CL7-1 CL7-2
PC7 vs. PC8	CL5-1 CL7-2 CL11-1	Not retained	Not retained	CL5-1 CL6-3 CL10-2 CL13-2	CL2-1 CL5-1 CL6-3 CL8-1 CL10-2 CL11-1	CL8-1 CL9-1 CL12-5	CL5-1 CL6-3 CL7-1 CL8-3
PC9 vs. PC10	CL7-3 CL8-1 CL11-1	Not retained	Not retained	CL4-2 CL8-3 CL11-2 CL13-3	CL4-2 CL5-1 CL8-1 CL9-2 CL11-1	Not retained	Not retained

Bivariate scores plot	Outlier						
	Combined sub-bands	A5	D5	D4	D3	D2	D1
					CL11-2 CL12-3		
PC11 vs. PC12	CL6-2 CL7-3 CL8-1 CL9-1	Not retained	Not retained	CL8-3 CL10-2 CL11-1 CL11-2 CL13-3	CL5-1 CL9-2 CL10-2 CL11-2 CL12-3	Not retained	Not retained
PC13 vs. PC14	CL2-2 CL6-2 CL7-1 CL9-1 CL10-1	Not retained	Not retained	CL7-3 CL11-1 CL13-3	CL4-1 CL11-2 32 CL13-3	Not retained	Not retained

Although they share the same outliers, it is evident that the bivariate scores plot of the combined sub-bands is a mirror of the bivariate scores plot of the sub-band D1. This indicates that the combined sub-bands were heavily influenced by the wavelet coefficients in sub-band D1. Furthermore, CL6-3 and CL11-2 are observed to be outliers in most of the bivariate scores plot. The characteristic of these cell line replicates will be investigated further in Section 5.7.1. Another PCA is performed with CL11-2 removed to examine its effect on the dataset. Interestingly, the bivariate scores plots of PC1 vs. PC2 generated for the combined sub-bands and each sub-band (Appendix B: Figure B.8 to Figure B.9) were similar to those shown in Figure B.1 and Figure B.2. The only difference is two of the outliers identified have changed, CL6-2 replaces CL6-3 and CL11-3 replaces CL11-2.

On the other hand, the bivariate scores plot of the wavelet approximation sub-band A5 reveals that the cell replicates were separated into two clusters along PC2 (Figure B.2). One group was clustered at the top left while the cell replicates in another group were located to the right and along PC1. Both clusters contain a mixture of high and low producers. No clear separation between high and low cell producers was observed in any of the bivariate scores plots.

Although no significant separation was observed, these results were nevertheless important. As was previously discussed in Chapter 3, the wavelet approximation sub-band contains the information in the spectra whilst the wavelet detail sub-band contains the noise in the spectra. The strong similarity between the bivariate scores plot of the combined sub-bands to those of the wavelet detail sub-bands indicate that the combined sub-bands were heavily masked by noise as discussed previously.

Percentage of variance explained in the retained principal components is presented in Table 5-2. For the combined sub-bands, the first fourteen principal components capture 51.8% of the variance in the dataset. 83.38% of the variance in the information extracted by the sub-band A5 is captured by the first six principal components. For the rest of the sub-bands, the variance captured is approximately less than 60%.

5.7.1 *Contribution Plots*

Score plots explain average behaviour of the cell replicates but a lack of clustering in the bivariate scores plots instigated another investigation of the ESI-MS dataset. By interrogating the underlying model, contribution plots may reveal the wavelet coefficients making the largest contribution to the model and causing the differences between cell replicates.

Based on the previous results, a different number of principal components were retained. Contribution analysis on the retained principal components from six wavelet sub-bands (A5, D5, D4, D3, D2 and D1) was performed to determine the characteristics of the high and low producers' cell lines. Figure 5.25 shows the breakdown of the contribution matrix. The rows represent data from 19 cell lines with varying number of replicates (Table 5-1), whilst the columns correspond to the contribution of the wavelet sub-bands to the retained principal component scores.

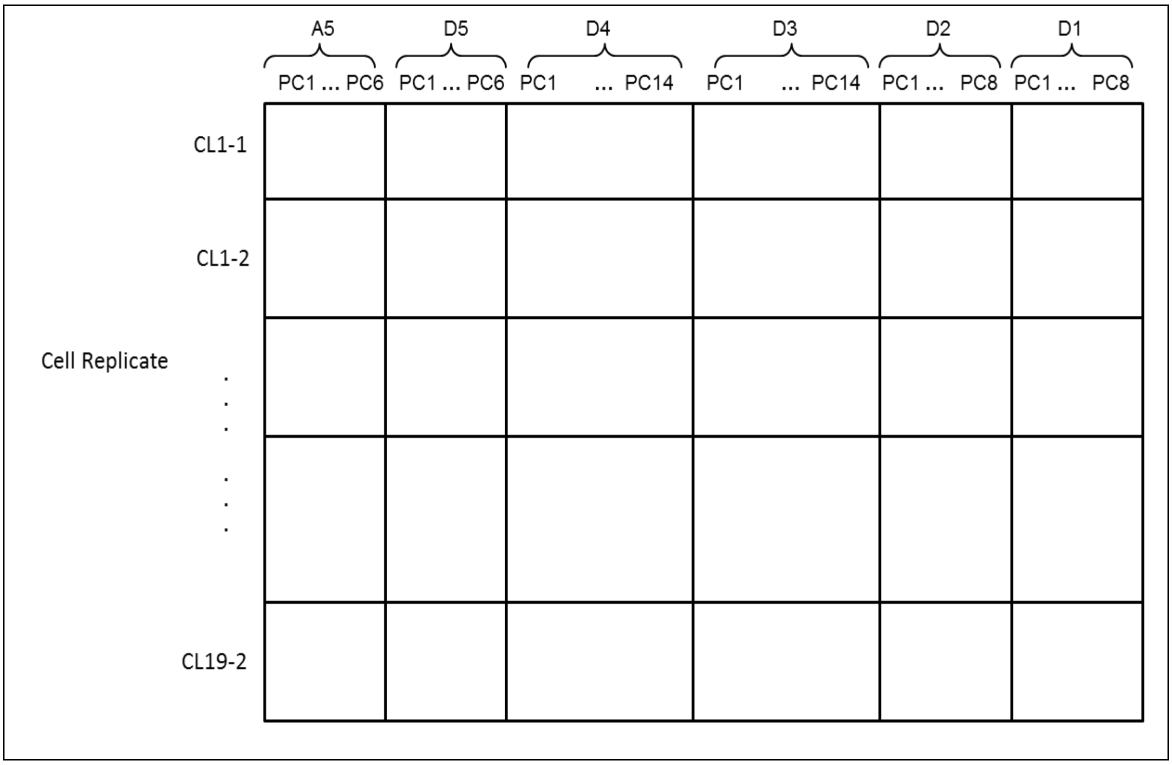


Figure 5.25 Matrix of the contribution of the wavelet sub-bands to the retained principal components

Figure 5.27 to Figure 5.35 and Figure B.1 to Figure B.10 provide the contribution plots of the wavelet sub-bands for the principal component scores for all cell lines and their replicates listed in Table 5-1. Figure 5.26 describes the legends for these figures. To establish the significance of the contribution analysis, an unsupervised PCA was also performed on the data matrix $\boldsymbol{X}$ without the application of the discrete wavelet transform. Based on the minimum and maximum number of principal component retained in the sub-bands, the percentage variance explained was calculated for the first six and fourteen principal components. It was found that 86.09% of the variance was captured by the first six principal components whilst 94.55% by the first fourteen principal components. Since the difference between percentages of variance is small, it is decided to retain six principal components for the application of PCA without discrete wavelet transform. The contribution plots on the retained principal components from the unsupervised PCA without discrete wavelet transform were plotted (Figure 5.36 and Appendix B: Figure B.11 to Figure B.18) and compared with the contribution plots of the wavelet sub-bands (Figure 5.27 to Figure 5.35 and Appendix B: Figure B.1 to Figure B.10).

These contribution plots are quite revealing in several ways. Firstly, it is interesting to discover that the cell replicates from the same cell line demonstrate diverse behaviour. This can be seen in both the integrated discrete wavelet decomposition-PCA model and in the stand alone PCA model. For example for the replicates of cell line CL12 and cell line CL14 which are shown in Figure 5.28 and Figure 5.35 respectively, CL12-5 exhibits different characteristics to the other four replicates. For the cell line CL14 all three replicates show distinctive characteristics with CL14-2 being significantly different.

Secondly, it can be seen that the significantly different cell replicates usually have a large contribution from wavelet sub-band D1. Figure 5.29 and Figure 5.30 respectively shows that CL13-3 and CL16-3 were significantly different to the other replicates from the same cell line. Further interrogation found that the two largest contributions to these two replicates were for the wavelet sub-bands D1 and D2 to PC1, PC2, PC3, PC4, PC5, and PC6. Also, it can be seen in Figure 5.32 and Figure 5.33 respectively that CL17-1 and CL18-1 differed from the other replicates from the same cell lines. It was interesting to observe that the largest contribution to the first replicate also came from the wavelet detail sub-bands D1 and D2. Similar behaviour was evident for CL14-2 as shown in Figure 5.35.

Another observation that can be drawn was that significant differences were identified between the integrated discrete wavelet decomposition-PCA model and the stand-alone PCA model. As can be seen from Figure B.10 (Appendix B), the highest contribution for cell replicate CL11-3 came from PC1. However, Figure 5.27 reveals that the contribution of the wavelet sub-band D4 to PC4 is the highest contributor. Similar behaviour was observed for CL4-2 (Figure 5.34). Further interrogation reveals that the highest contribution to this cell replicates was from the wavelet sub-band D3 to PC6.

Previously in the scores plots CL11-2 was identified as one of the significant outliers. From the contribution plot of cell line CL11 (Figure 5.27), it was observed that CL11-2 is significantly different from CL-1 and CL-3. The largest contribution to CL11-2 is the sub-band D4. When CL11-2 was removed and a second PCA was performed, the position of CL11-2 was replaced by CL11-3. Further interrogation of the contribution plot of CL11-3 found that the largest contribution to CL11-3 came from the sub-band D4. It is noted that the sub-band D3 and D4 were the two sub-bands that required the most principal components i.e. PC1 to PC14 to capture approximately 50% and 60% of the variance in the dataset. Meanwhile, the other significant outlier, CL6-3, was observed to have the largest contribution from the sub-band D3 (Figure 5.31).

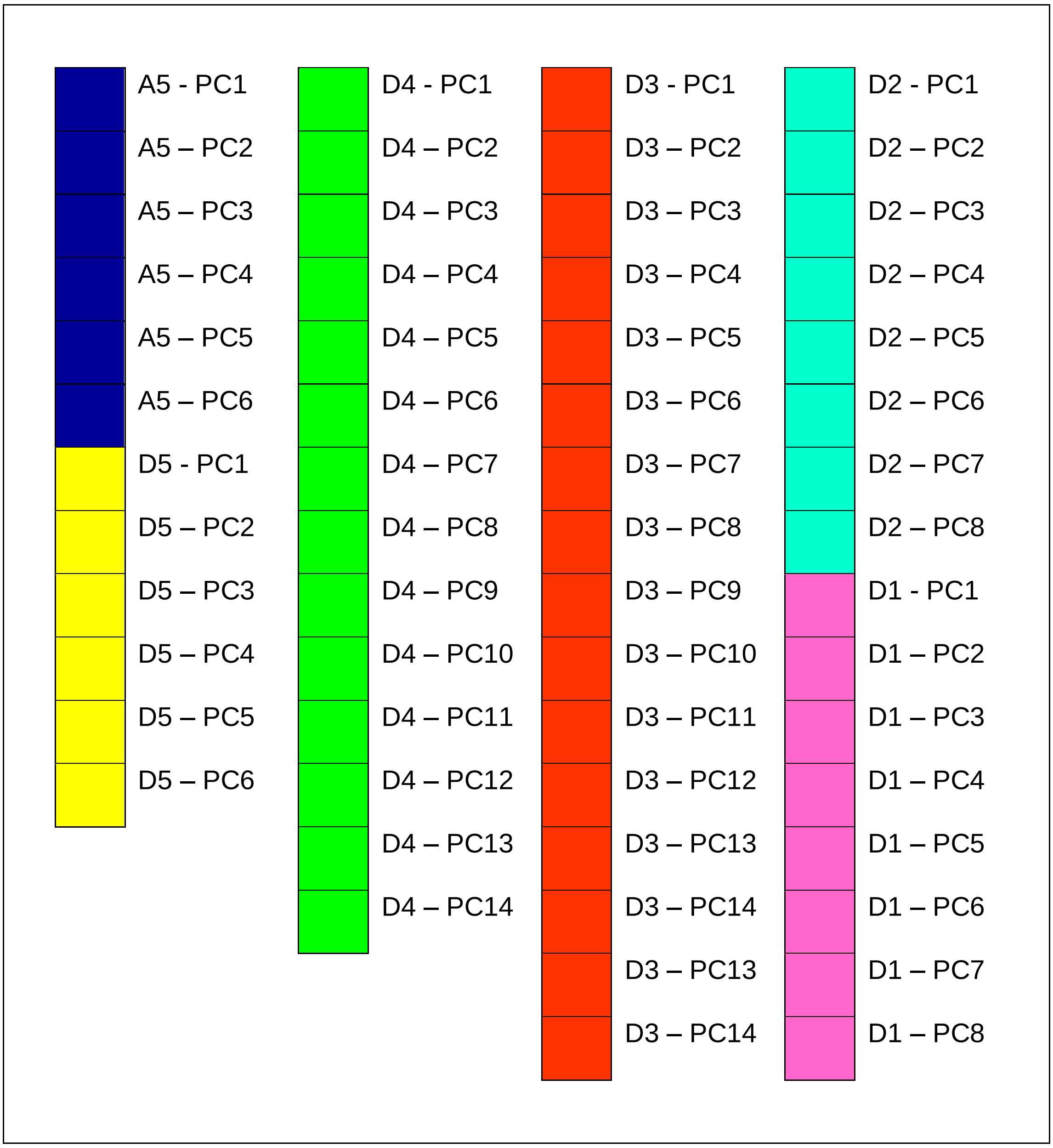


Figure 5.26 Legends for contribution plots figures (Figure 5.25 to Figure 5.34 and Figure B.4 to Figure B.10)

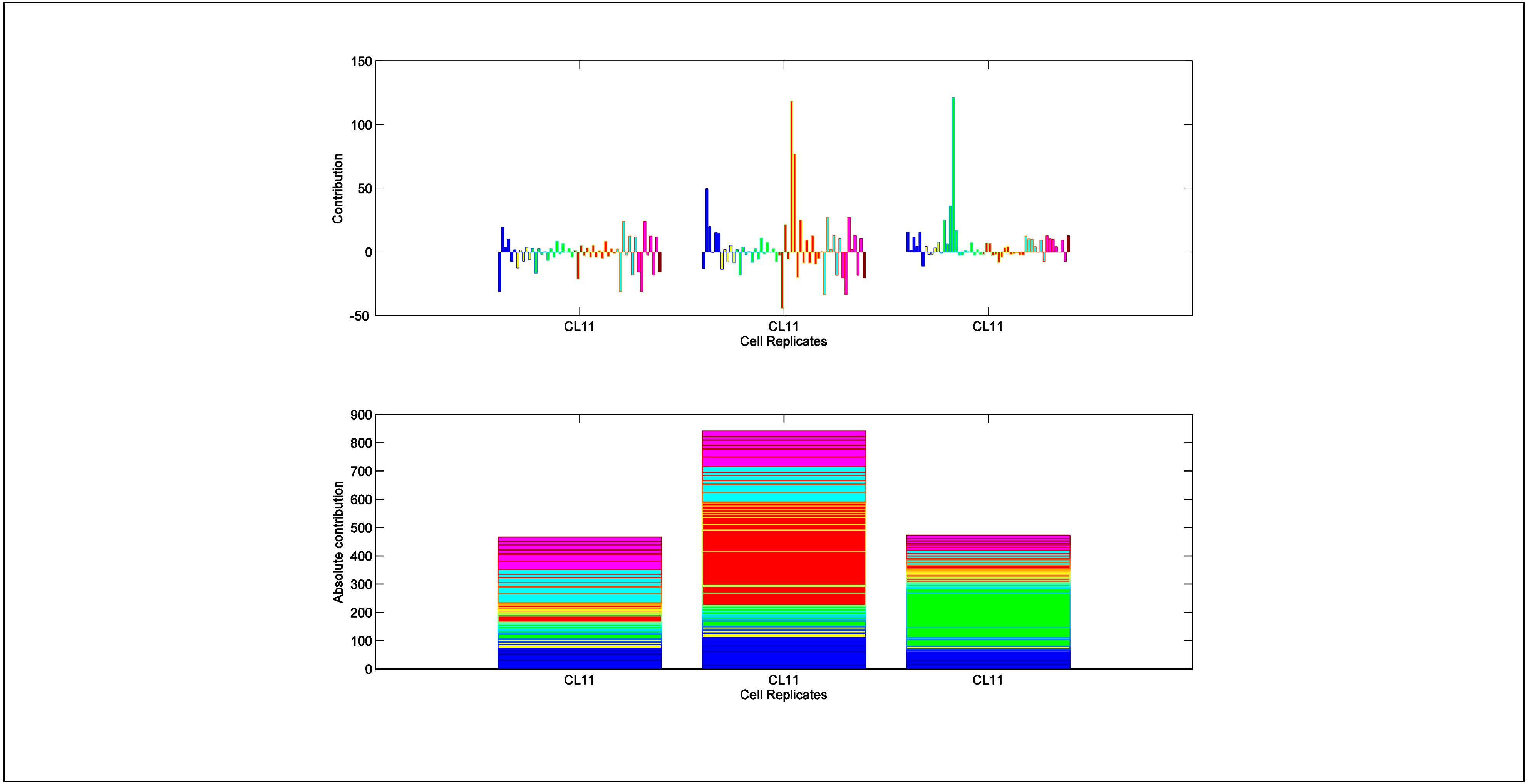


Figure 5.27 Unsupervised PCA: Contribution plots of the wavelet sub-bands on PC1 to PC6 for cell line CL11 (high producing cell line)

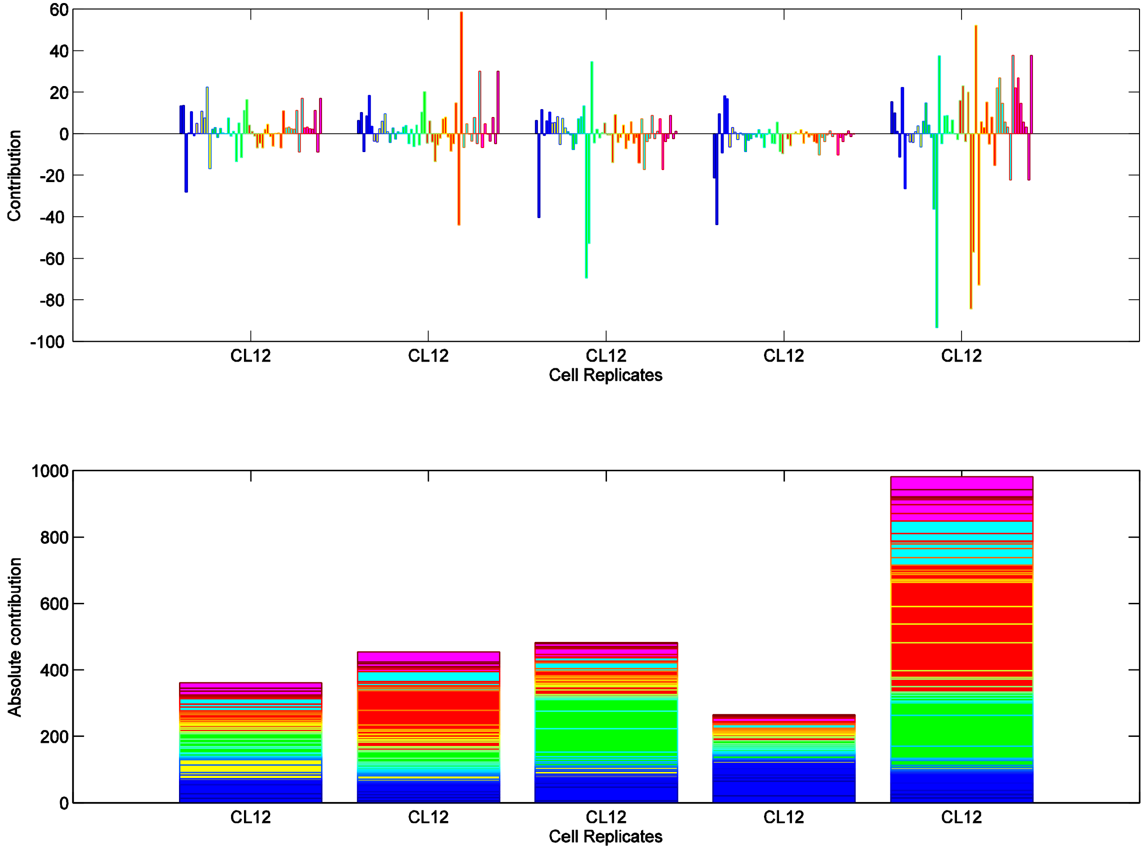


Figure 5.28 Unsupervised PCA: Contribution plots of the wavelet sub-bands on PC1 to PC6 for cell line CL12 (high producing cell line)

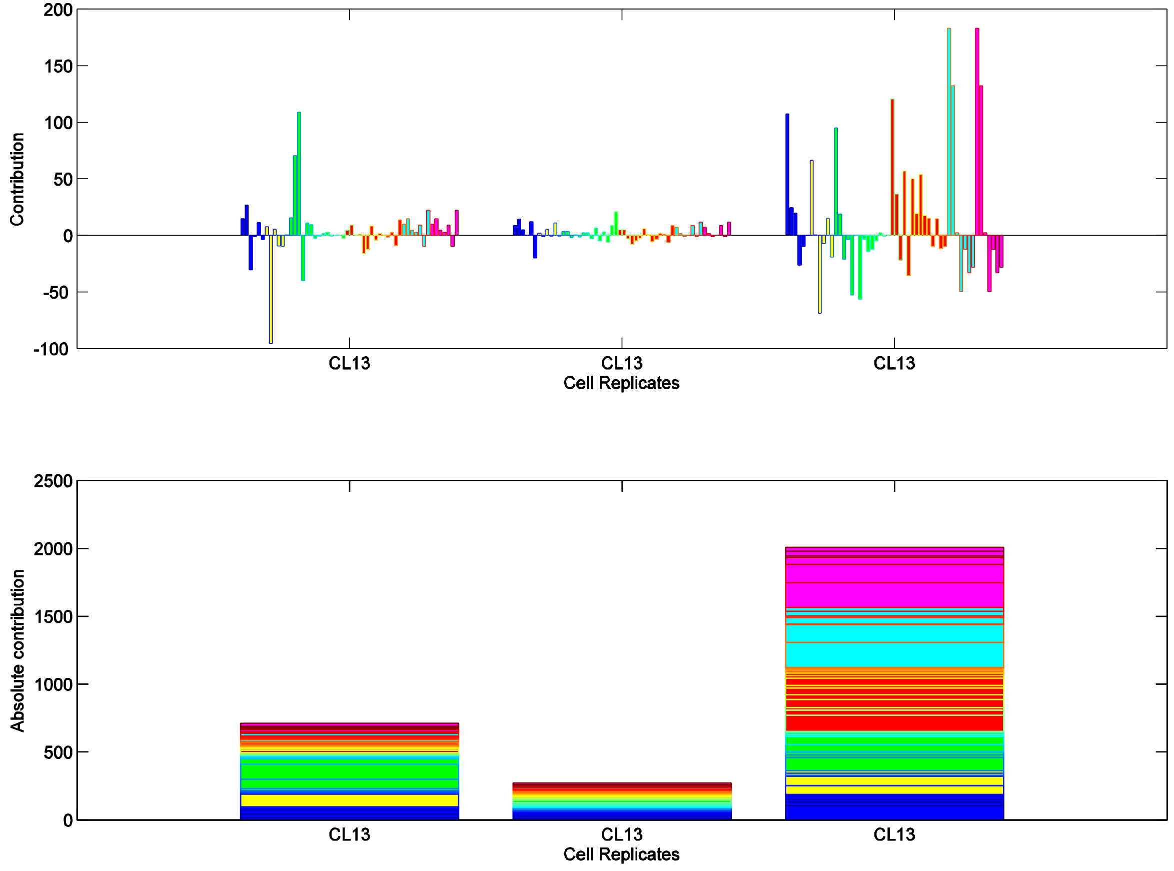


Figure 5.29 Unsupervised PCA: Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL13 (high producing cell line)

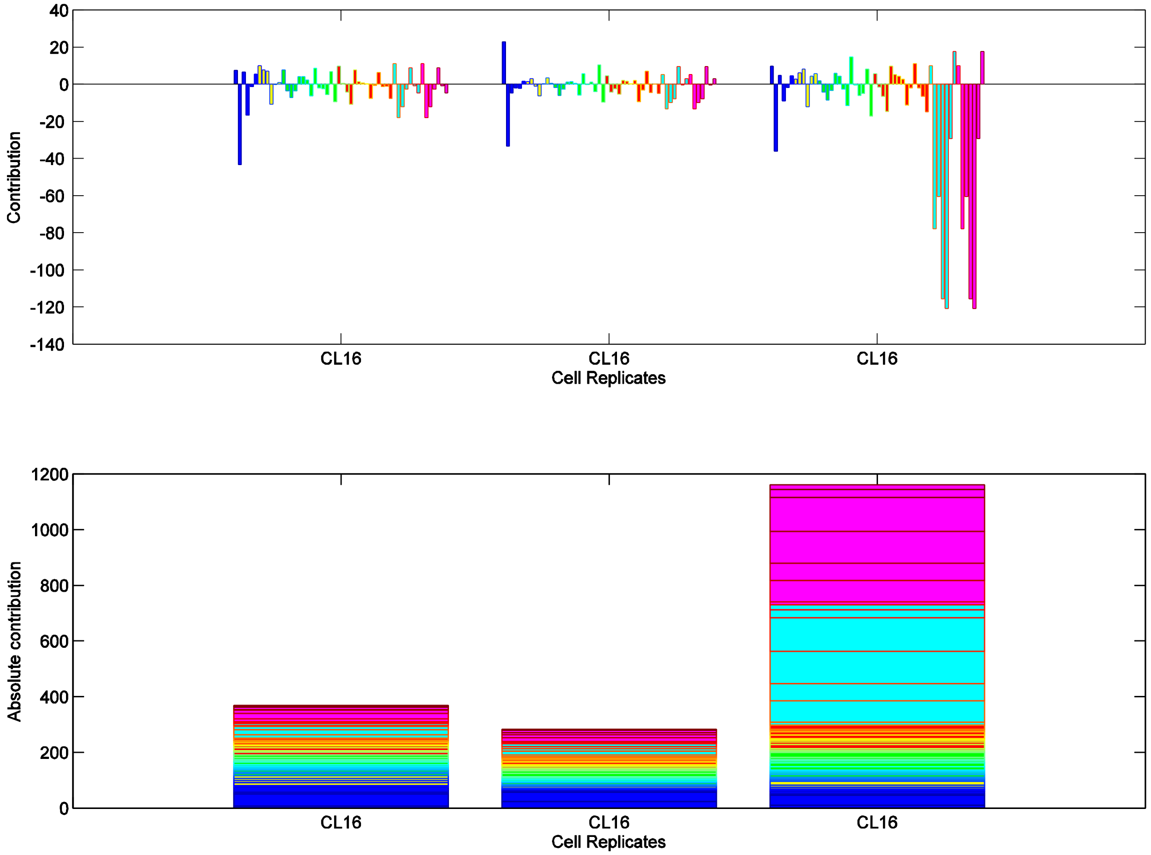


Figure 5.30 Unsupervised PCA: Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL16 (high producing cell line)

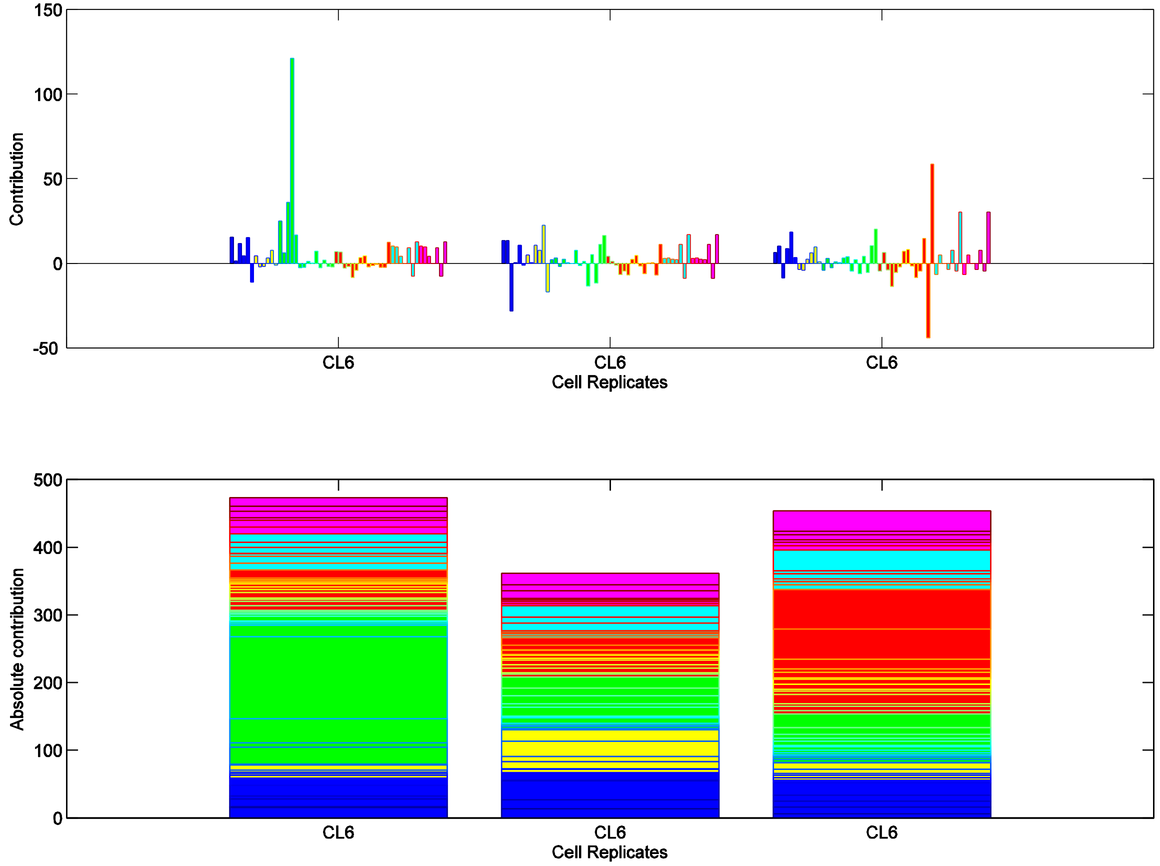


Figure 5.31 Unsupervised PCA: Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL6 (low producing cell line)

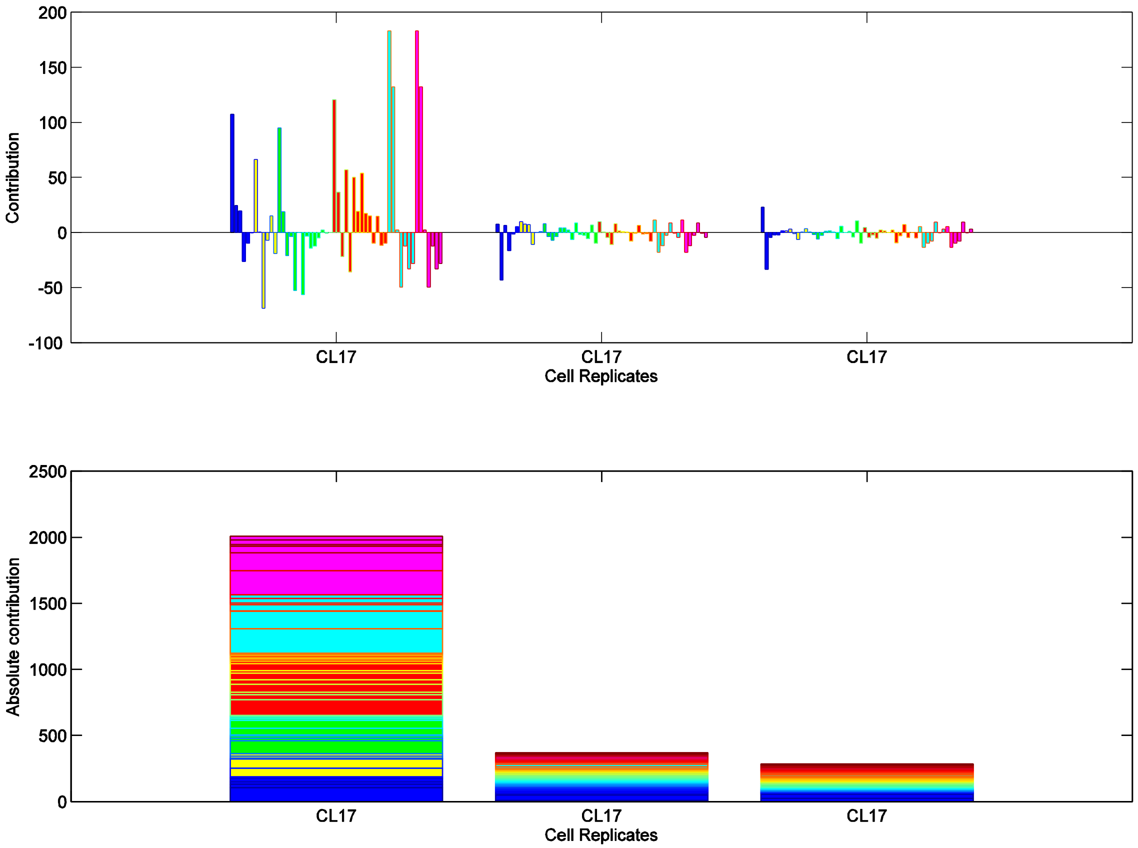


Figure 5.32 Unsupervised PCA: Contribution plots of the variables on PC1 to PC6 for cell line CL17 (low producing cell line)

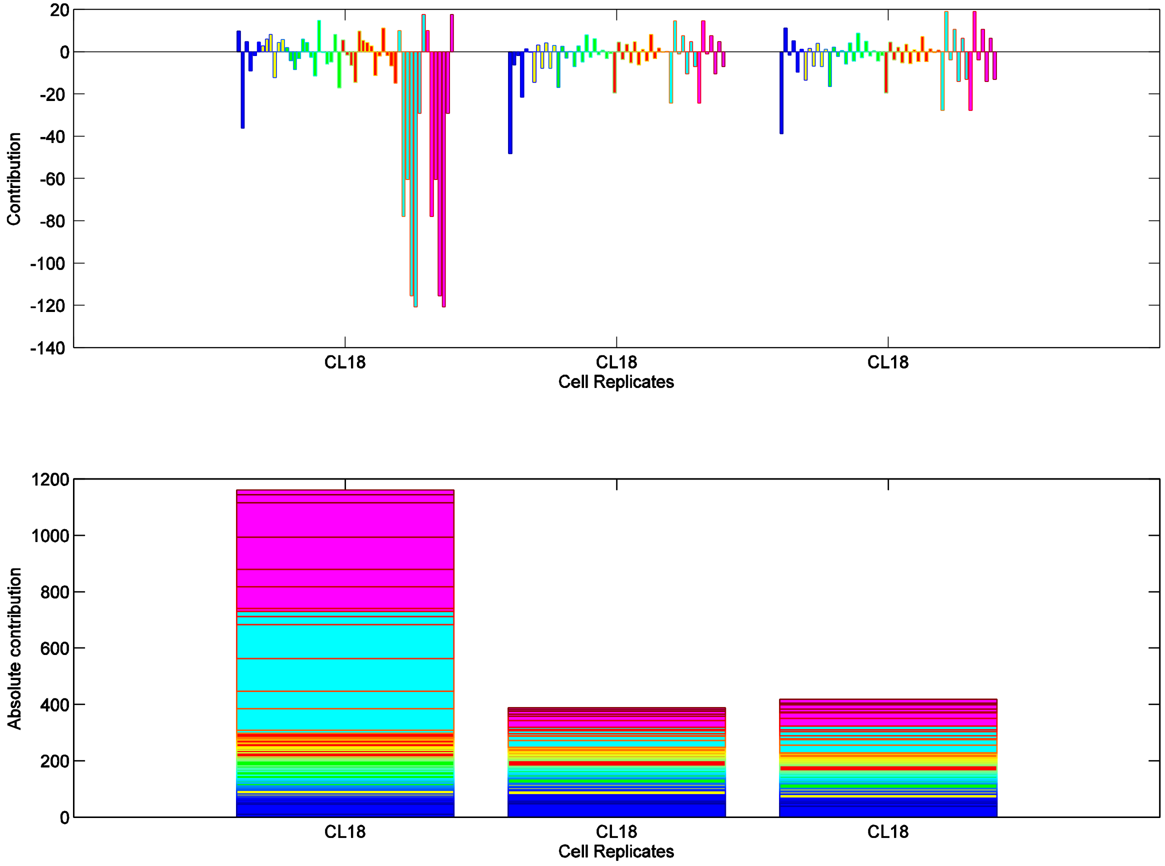


Figure 5.33 Unsupervised PCA: Contribution plots of the variables on PC1 to PC6 for cell line CL18 (low producing cell line)

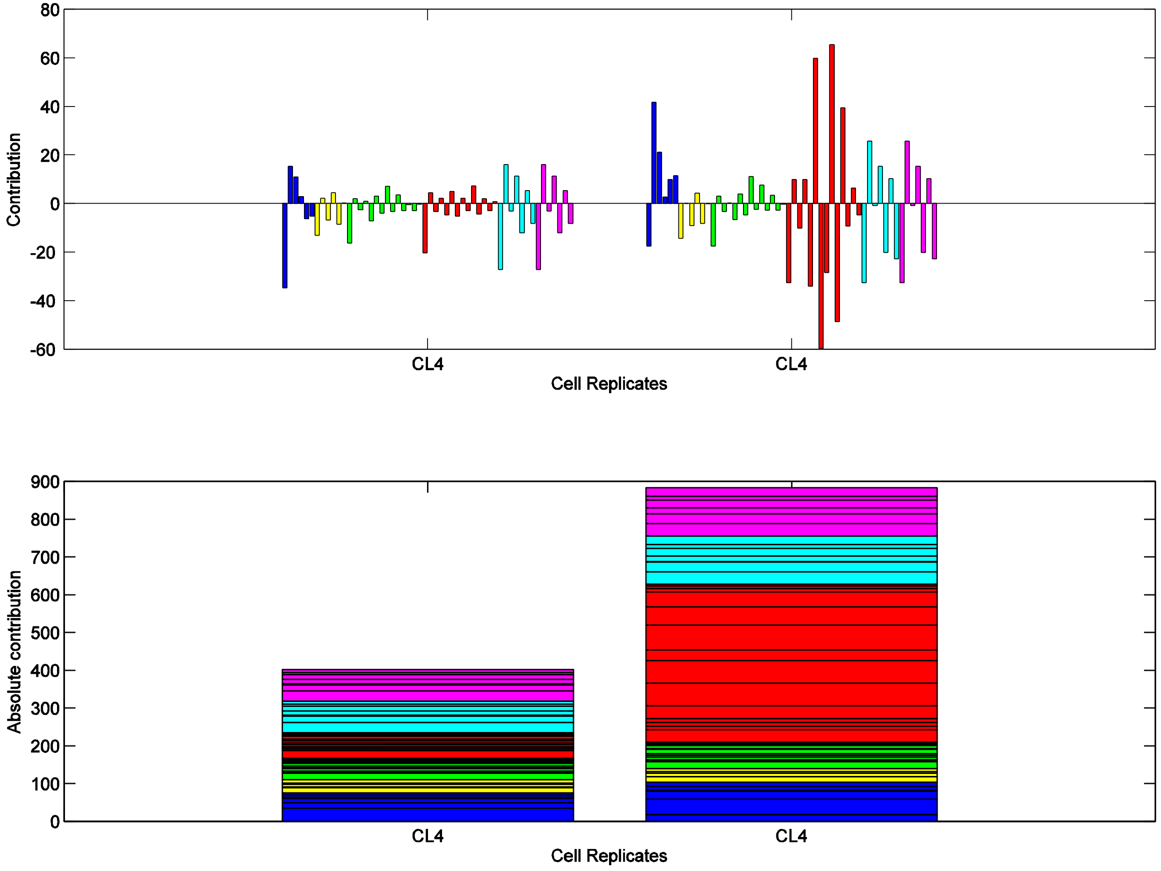


Figure 5.34 Unsupervised PCA: Contribution plots of the variables on P1 to PC6 for cell line CL4 (low producing cell line)

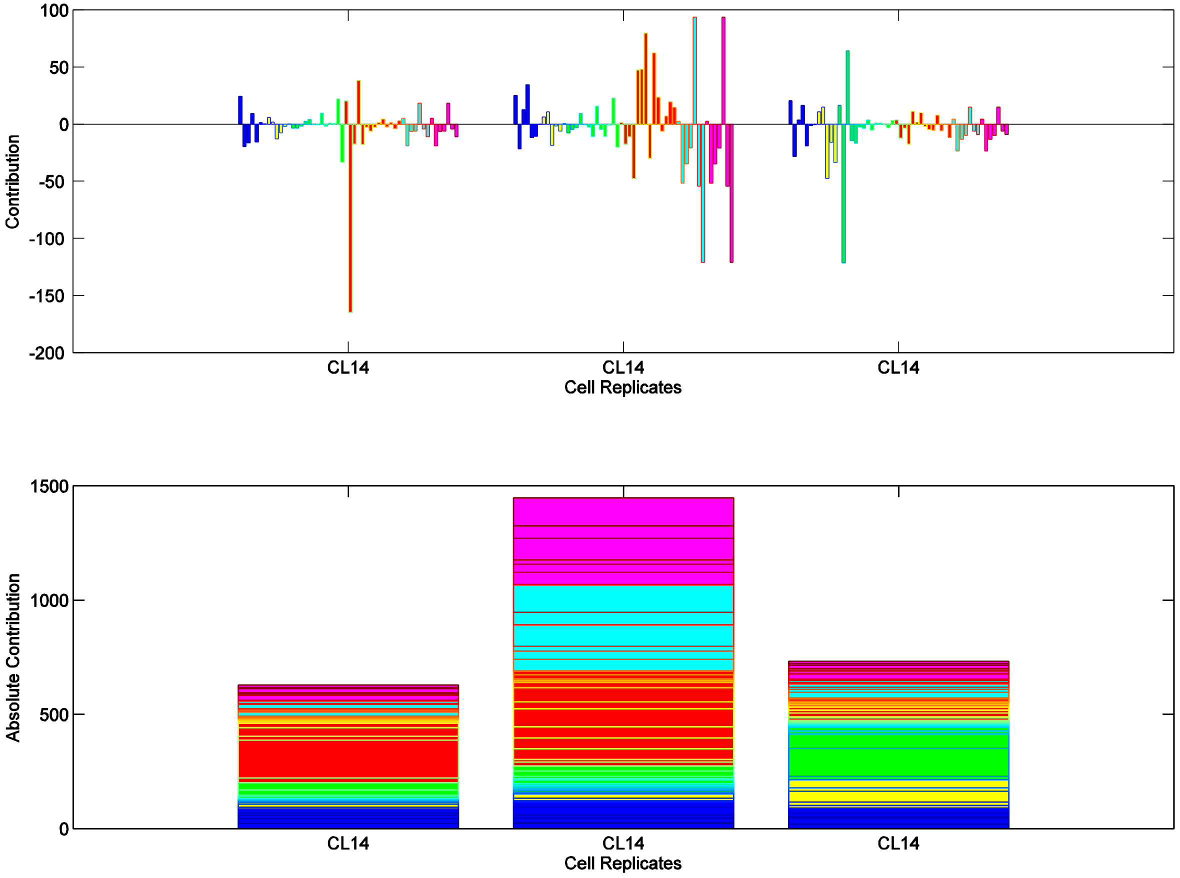


Figure 5.35 Unsupervised PCA: Contribution plots of the variables on P1 to PC6 for cell line CL14 (low producing cell line)

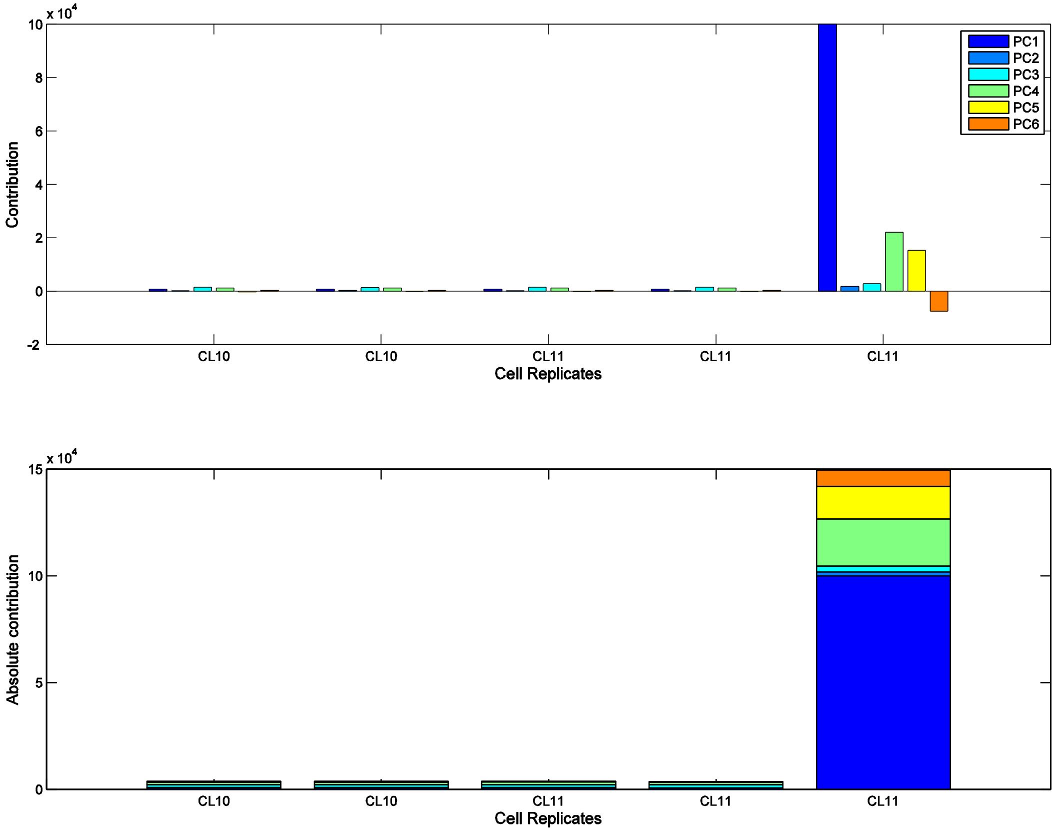


Figure 5.36 Contribution plots of the ESI spectra on P1 to PC6 (without discrete wavelet decomposition) for cell lines CL10 and CL11 (high producing cell line)

5.8 Results and Discussion: Supervised Model

A supervised model was developed utilising the measurements of three types of product quality parameters: PQ1, PQ2 and PQ3. These parameters were measured off-line. Based on a discussion with the engineer in charge of the ESI-MS data, each product quality has a cut-off value which is used to classify a cell line as high or low producer. A cell line is classified as high in a particular product quality group if its parameter measurement is equal to or above the respective cut-off value.

The cell lines were divided into a training set and a test set based on their measurements of the product quality parameters. The training set consists of cell lines that take low values for all the product quality parameters whereas the high cell lines were grouped into the test set. Table 5-4 shows a complete list of the training and test sets.

Figure 5.37 to Figure 5.43 shows bivariate scores plots for the first two principal components for each wavelet sub-band and the combined sub-bands. The most striking observation to emerge from the bivariate scores plots was that the outlying cell replicates CL13 and CL16 detected in the PC1-PC2 space of the whole spectra (Figure 5.43) were also identified in the PC1-PC2 space of the wavelet detail sub-bands D1 (Figure 5.42) and D2 (Figure 5.41). Interestingly, these outlying cell replicates were positioned almost at the same spot in all three bivariate scores plots. This observation was identical to the one found earlier in the unsupervised model. Similarity among these three bivariate scores plots gives a strong indication that the whole spectra were highly masked by noise. As with the unsupervised PCA, no evident clustering were observed in the bivariate scores plots.

Table 5-4 Cell lines information for the training and test sets

Test data set		Training data set	
Cell lines	Number of replicates	Cell lines	Number of replicates
CL10	2	CL19	2
CL11	3	CL2	2
CL12	5	CL6	3
CL13	3	CL8	3
CL16	3	CL15	2
		CL17	3
		CL18	3
		CL1	2
		CL3	2
		CL4	2
		CL5	2
		CL9	2
		CL7	3
		CL14	3
Total cell replicates	16	Total cell replicates	34

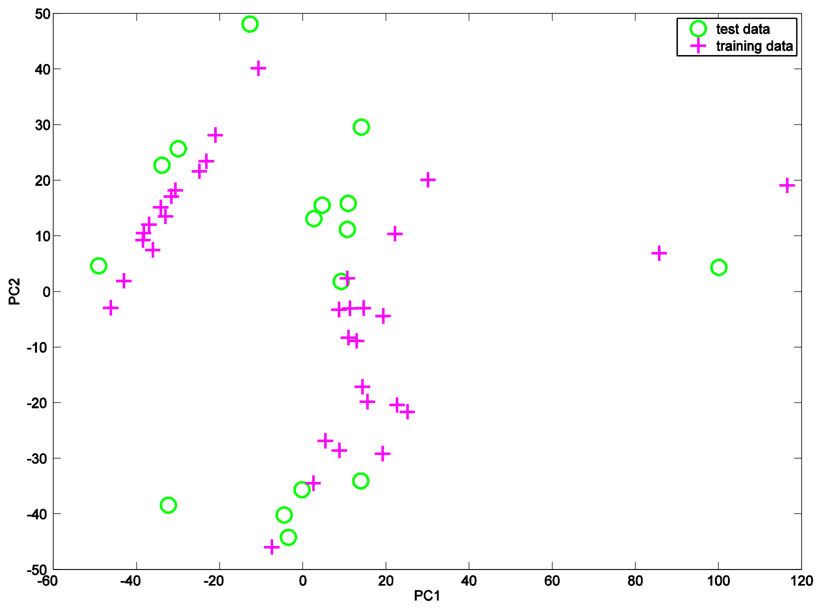


Figure 5.37 PC1 vs. PC2 for predicted wavelet sub-band A5

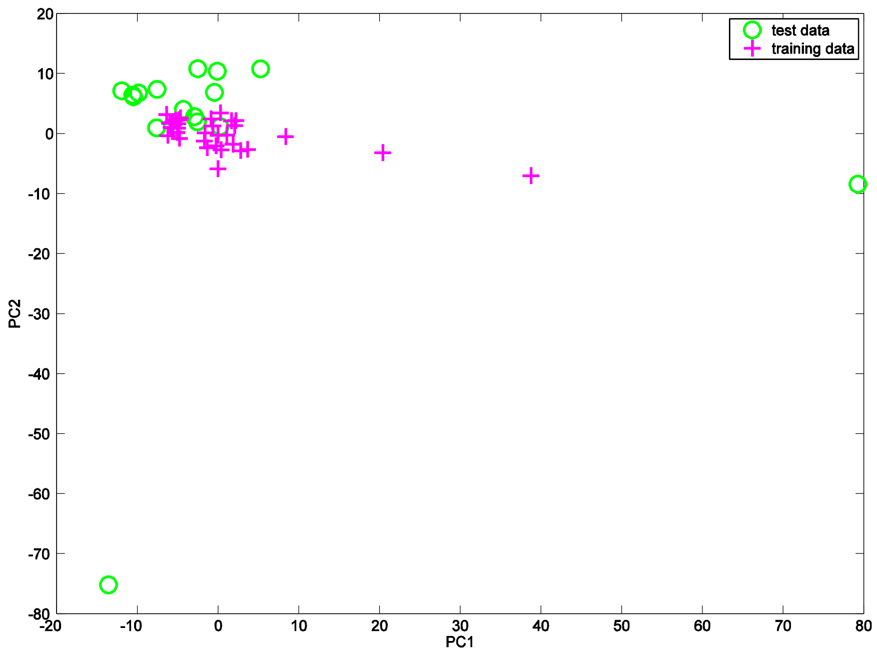


Figure 5.38 PC1 vs. PC2 wavelet sub-band D5

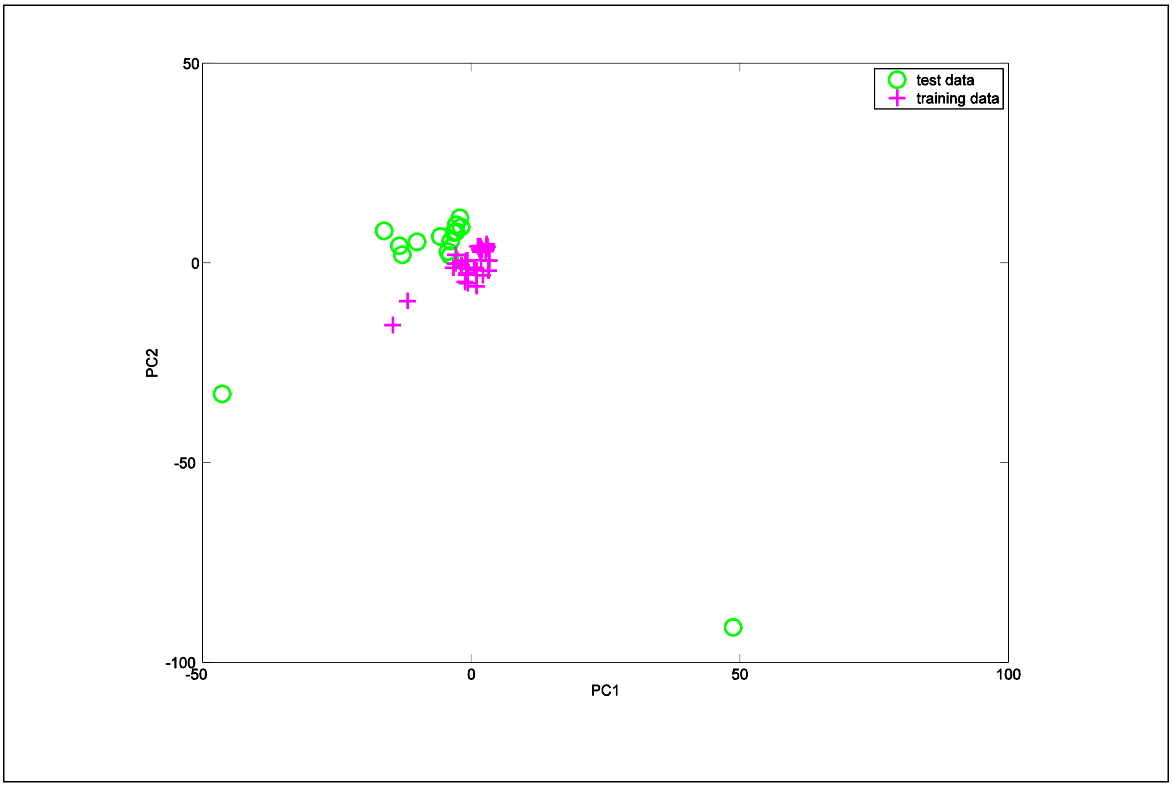


Figure 5.39 PC1 vs. PC2 for predicted wavelet sub-band D4

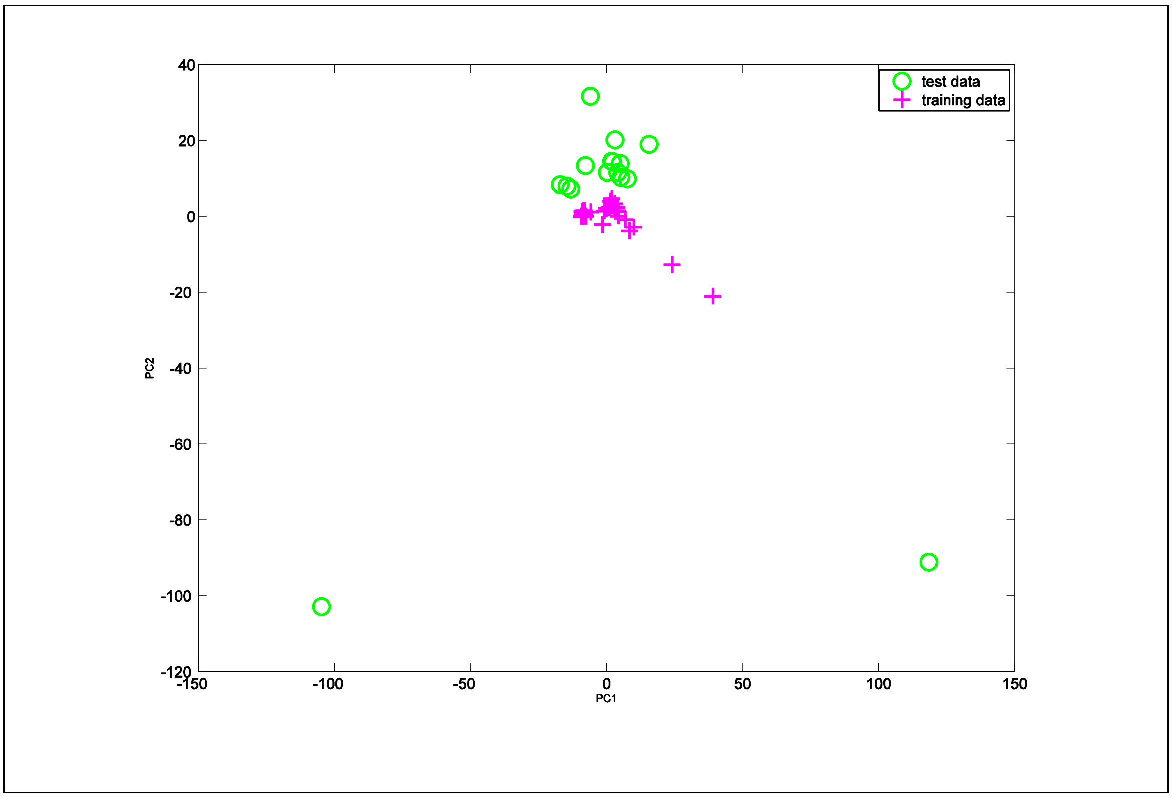


Figure 5.40 PC1 vs. PC2 for predicted wavelet sub-band D3

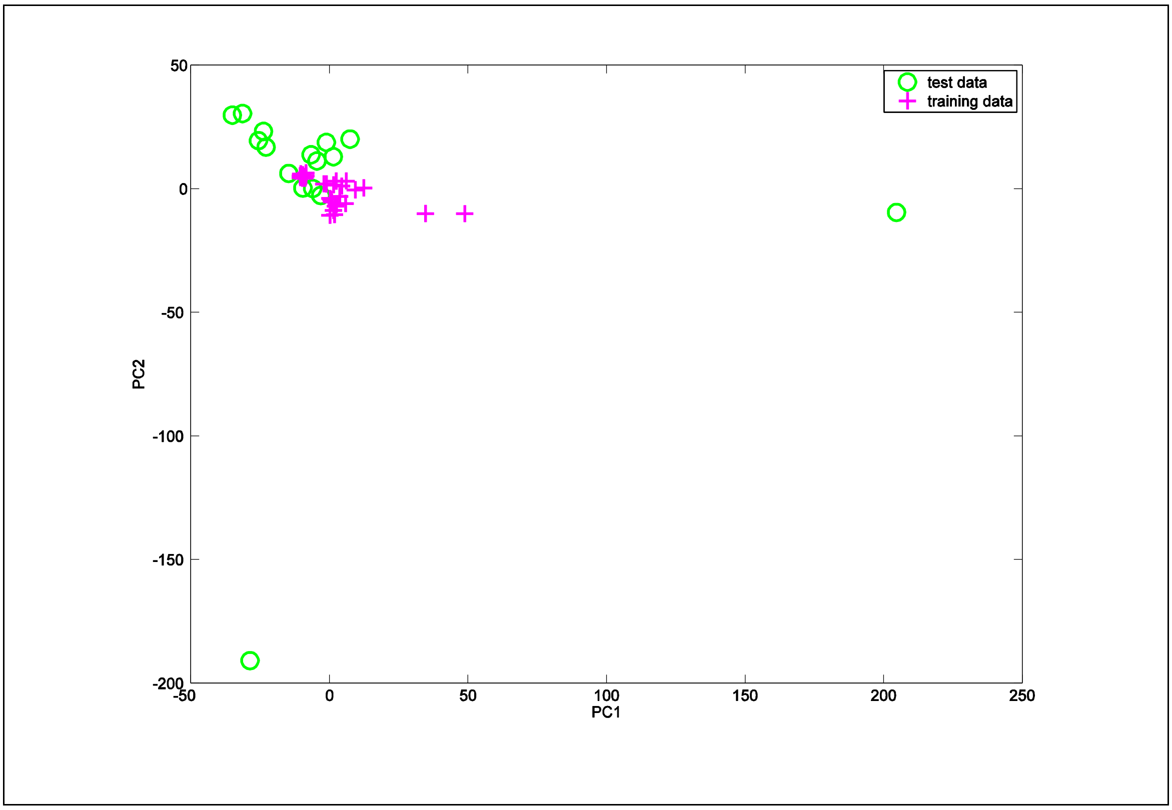


Figure 5.41 PC1 vs. PC2 for predicted wavelet sub-band D2

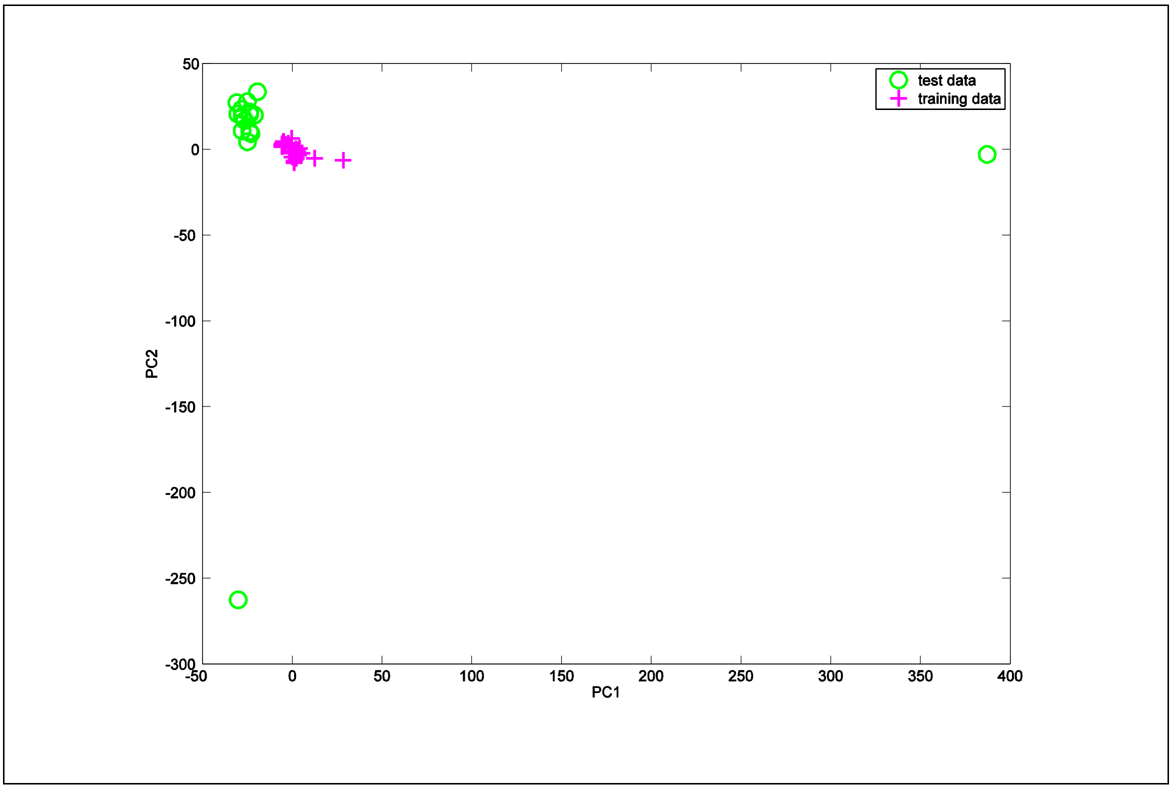


Figure 5.42 PC1 vs. PC2 for predicted wavelet sub-band D1

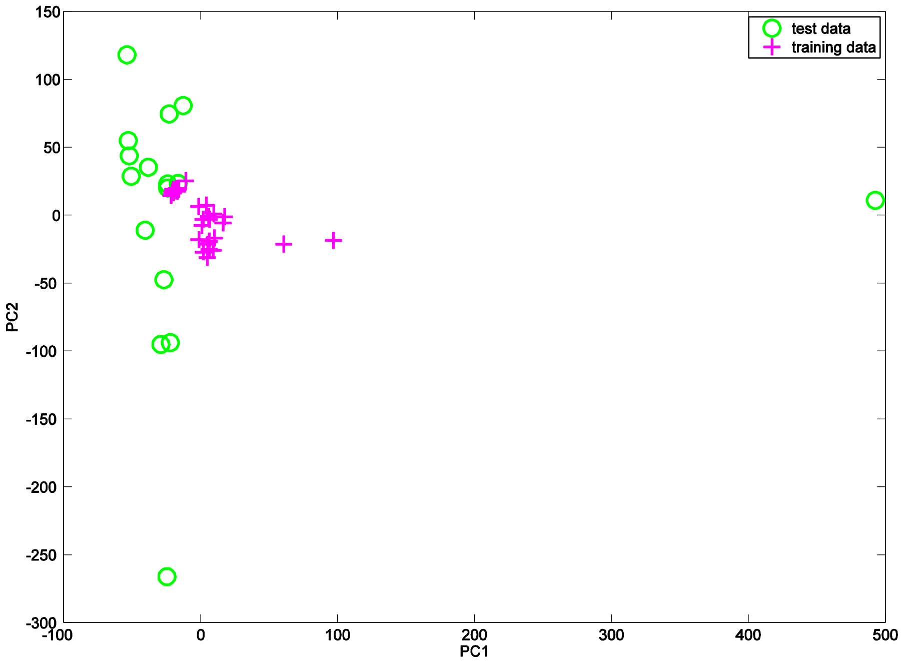


Figure 5.43 PC1 vs. PC2 for predicted for predicted combined sub-bands

5.9 Summary

The concept of developing a model to establish the criteria of the CHO cell lines producers has been introduced and the research has been observed to make a significant contribution to the analyses of a complex biological data set. Screening and selecting highly stable cell lines is a major challenge in the process development of the CHO cell lines producers. Hence, characterisation of the CHO cell lines producers is of high importance as the subsequent process stages condition on the previous screening. Despite its importance, methodologies for extraction of information from the complex biological data are limited. The development of the integrated discrete wavelet transform-PCA approach can help characterise the CHO cell lines. Also, it gives another perspective on the fingerprinting of the CHO cell lines producers.

More specifically in this chapter, a successful application of the integrated discrete wavelet decomposition-PCA platform in teasing out hidden information in the spectra has been proposed. This information was incorporated with the information on the quality parameters to help characterise the cell line producers. The benefits of the integrated discrete wavelet decomposition-PCA platform were highlighted. The challenges presented by the ESI-MS dataset were also discussed.

Although the ultimate objective of this study which is to characterise the criteria of the high and low CHO cell line producers is not fully justified, the study reveals a number of significant observation. Firstly, the bivariate scores plots show a strong relationship between the cell replicates from the same cell line. This was observed in the bivariate scores plot of the PCA on the dataset when a cell line replicate (CL11-2) was removed. It was demonstrated that another cell replicate from the same cell line i.e. CL11-3 was located at the position of CL11-2. Secondly, the contribution plots of the wavelet sub-bands reveal that most cell line replicates have different characteristic despite coming from the same cell line. This is surprising because the cell replicates that come from the same cell line are highly anticipated to have the same characteristic.

In conclusion, the differences in the characteristic of the cell line replicates may potentially be the key to answer questions including why cell culture from the same cell line behave differently. Therefore, fingerprinting of the CHO cell line producers from the perspective of this study requires further interrogation including the analysis of square prediction error and Hotelling’s T^2.

Chapter 6 Conclusions and Future Work

6.1 Conclusions

The aim of the thesis was to investigate strategies for the analysis of spectral measurements in two bioprocess applications. The outcome of the thesis was a number of contributions being made to the field of fingerprinting of complex bioprocess spectral data.

- The use of spectroscopy as a potential source of information to add value to the understanding of a process offers major opportunities. However, extracting meaningful information from spectral data can be hampered by the large number of variables, overlapping spectra, and the presence of a high noise to signal ratio in parts of the data.

- The successful application of fingerprinting to extract information and hence process understanding requires an approach that can extract latent information from complex data structure.

- A new approach based on the application of the combination of the discrete wavelet transform and the multivariate statistical technique of principal component analysis was proposed. This methodology offers the opportunity to extract features at different scales from the data. The flexibility of the proposed framework was demonstrated by its application to two contrasting sets of spectral data.

The application focus of the thesis was on complex bioprocess data, specifically spectral data from the manufacture of monoclonal antibodies from Chinese Hamster Ovary (CHO) cell lines. The proposed methodology was shown to be effective for the analyses of both near infra-red and electrospray ionisation spectra data sets. The raw near infra-red spectra are subjected to broad overlaps and baseline shifts whilst the electrospray ionisation mass spectra are subjected to high noise to signal ratio and shifts in mass-to-charge (m/z) ratio. The differences in structures of these data sets required the application of different pre-processing techniques, prior to the implementation of the proposed methodology.

A generic flow diagram of the framework developed in this thesis is shown in Figure 6.1. The framework comprises three main steps. Firstly, the spectral data sets require to be pre-processed with techniques that are suitable for their structure. Secondly, the discrete wavelet decomposition is performed on the pre-processed spectra. The wavelet coefficients resulting from the discrete wavelet decomposition of the spectra were then used as inputs to principal component analysis.

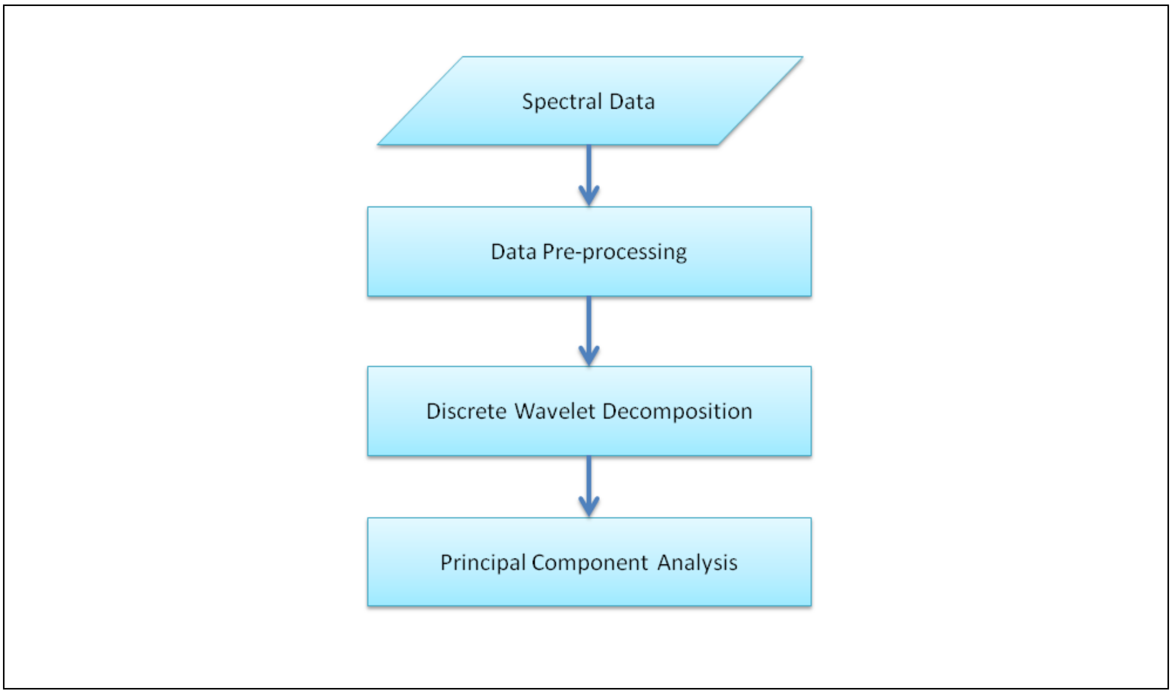


Figure 6.1 Generic flow diagram of the framework developed

Prior to developing a fingerprinting representation of the data sets, an investigation into the application of the wavelet denoising algorithm was undertaken. The wavelet denoising algorithm was applied to on the near-infra red data set to examine its effect on reducing the noise in the data. Based on the investigation, it was concluded that the implementation of the wavelet denoising algorithm can be omitted because it changes the peaks of the spectra. Flattened peaks were observed in the denoised spectra.

The development of a fingerprinting representation and its application to near infra-red spectral data generated from the industrial manufacture of monoclonal antibodies was described in Chapter 4. Data preprocessing techniques were reviewed prior to this. Of particular interest was the re-sampling of spectra due to the use of seed and production batches in the analysis. It was thus essential to the near infra-red spectral batch data. The approach to the re-sampling of the spectra aimed to include spectra throughout the duration of the growth curve.

A novel approach was also required for the near infra-red spectral batch data in terms of unfolding it from a three dimensional structure to a two dimensional matrix. The selection of informative spectral regions was also discussed and utilised in the development of the process representation. It has been shown in previous research that wavelength selection can improve the results by decreasing the influence of unwanted information contained in the spectra. More specifically, the wavelength regions selected were for glucose overtones as it is the primary carbon and energy source for mammalian cells, and hence observation of its consumption contributes to process understanding.

Furthermore, a route for selecting the mother wavelet and levels of decomposition were explained. The key advantage of the proposed approach is that it utilises the multiresolution property of the discrete wavelet transform in the feature extraction. Information wavelet coefficients were selected at different resolutions of the spectra and used as features to represent the spectra for subsequent analysis. It was demonstrated that the combination of the discrete wavelet transform and the multivariate statistical technique of principal component analysis was able to extract relevant features from the spectral data and establish a relationship between batch genealogy and spectral data behaviour. The genealogy enabled an explanation to be provided in terms of the distinguishing characteristics observed between the batches from same and/or different families.

In Chapter 5, the wider applicability of the framework developed in Chapter 4 was demonstrated through its application to a different form of spectral data. Electrospray ionisation spectral data from the laboratory scale development of monoclonal antibodies formed the basis of this chapter. The main challenge was to characterize the criteria that differentiated between highly productive and low producer cell lines. The wavelet coefficients of the proposed model were utilised differently in this chapter. The strategy adopted was to consider all the wavelet coefficients in each wavelet sub-band from the discrete wavelet transform decomposition. The wavelet sub-bands were analysed individually using contribution plots of the principal component scores. This approach differs in two ways from the approach undertaken in Chapter 4 where firstly, only certain wavelet coefficients from the wavelet sub-bands were selected to represent the spectra, and secondly, the wavelet sub-bands were combined prior to the subsequent analyses. The patterns contained within the contribution plots were used to establish the criteria for high and low cell line producers. In contrast to the standard use of contribution plots in multivariate statistical technique, this chapter introduced a different perspective on the use of contribution plots of principal component scores.

6.2 Recommendations for Future Work

This thesis has demonstrated two successful applications that address some of the emerging challenges in the bioprocess manufacturing industries. The bioprocess manufacturing industries present unique challenges including high production costs and the manufacture of consistently high quality product. Addressing these challenges requires input from multifaceted areas such as cell engineering, bioinformatics, process development and process modeling. The initial research and development into an integrated discrete wavelet transform and multivariate statistical technique representation has been undertaken in this thesis. Opportunities for further improvements and exploitation of the techniques are summarized below:

1) The data matrix unfolding in both case studies discussed in Chapter 4 and 5 adopted the Nomikos and MacGregor (1994) approach for the multiway principal component analysis. Further research requires to be performed using the approach proposed by Wold et al. (1987). As the Wold et al. (1987) approach does not require equivalent batch length; data from batch process of longer duration will have more features (wavelet coefficients) in the data representation which may potentially affect the interpretation of the behaviour of the projected batches.

2) The discrete wavelet transform with mother wavelet of Daubechies 5 (db5) was selected as the most suitable of the mother wavelets investigated in this thesis. Further research requires to be performed into the investigation of the wavelet packet as a platform to extract and select meaningful wavelet coefficients from the complex spectra. More specifically, the electrospray ionisation mass spectral data set for which the noise to signal ratio is high, the implementation of the wavelet packet is anticipated to extract more interesting information hidden within the noise.

3) The case studies carried out in this thesis used measurements collected from spectral data of different types, therefore a key part of any future work should involve incorporating process data or quality parameters into the model through the application of other types of multivariate statistical analysis such as Partial Least Squares (PLS).

4) Both of the data sets investigated in this thesis involved a number of preprocessing techniques. Further research requires to be performed investigating the range of preprocessing tools and the sensitivity of the analysis to these and to test the robustness of the data representation developed in this thesis.

Appendices

Appendix A: Appendices for Chapter 4- Integrated Modelling for NIR Industrial Process Data

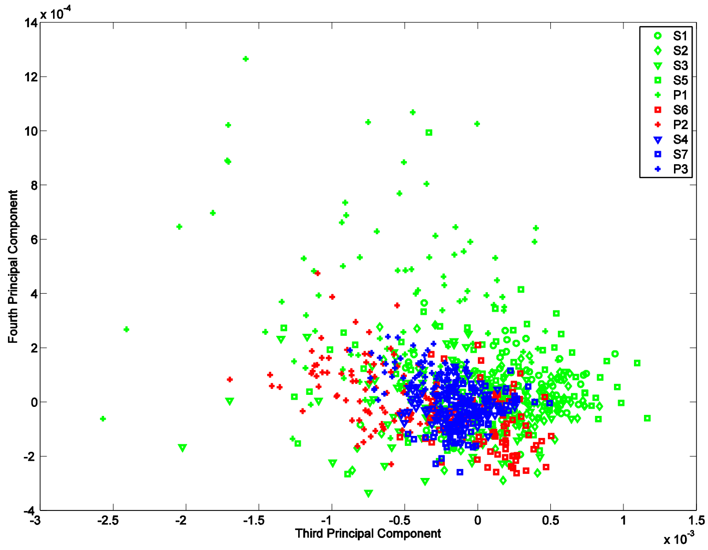


Figure A.1 PC3 vs. PC4 for first CH overtone

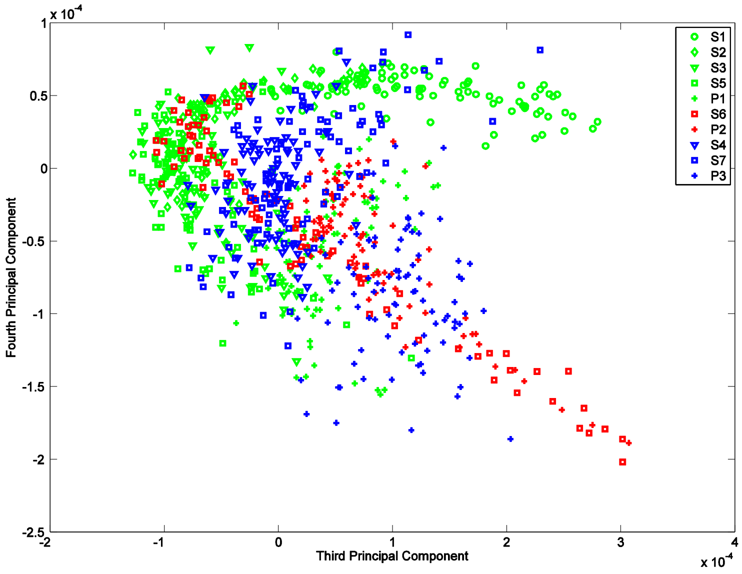


Figure A.2 PC3 vs. PC4 for second CH overtone

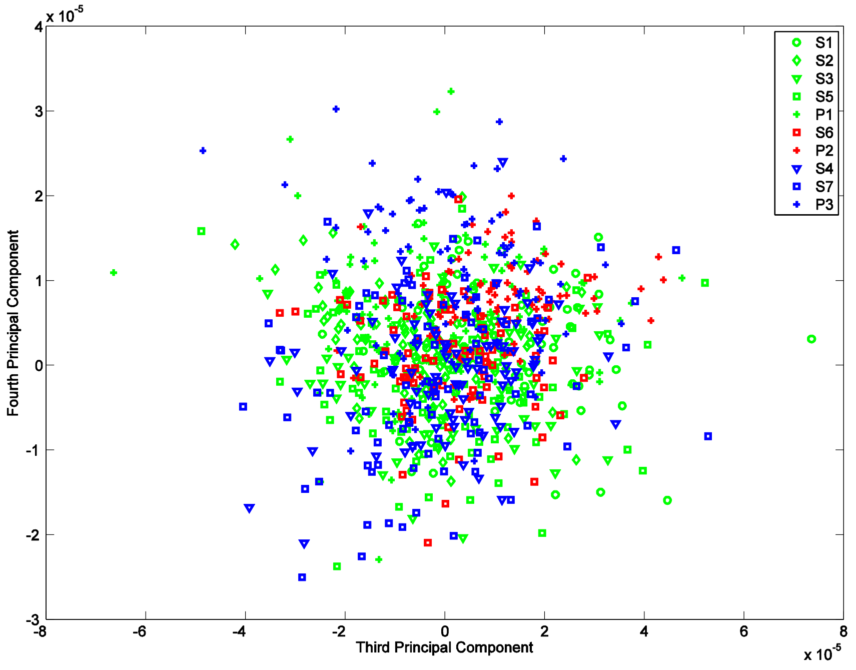


Figure A.3 PC3 vs. PC4 for third CH overtone

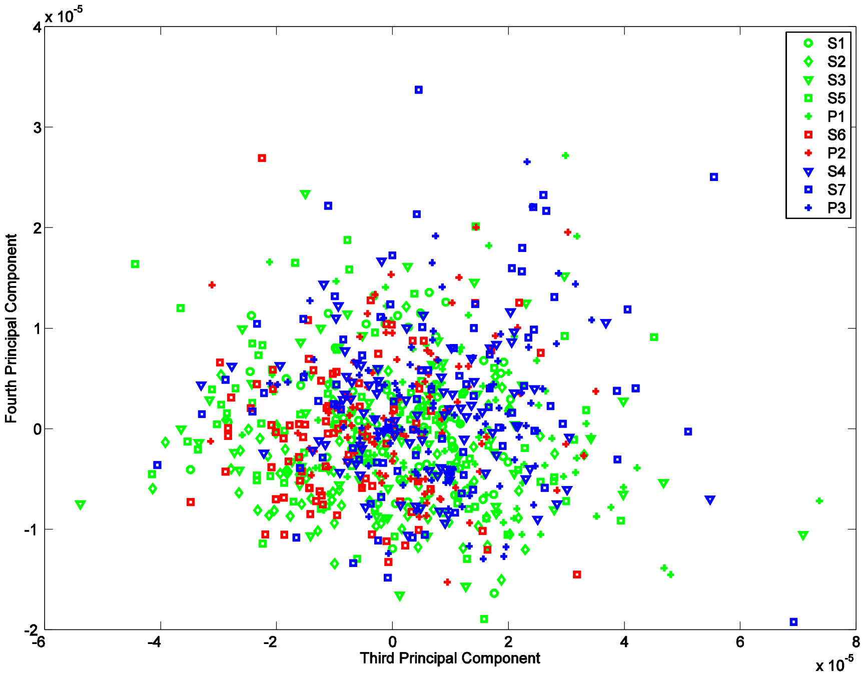


Figure A.4 PC3 vs. PC4 for fourth CH overtone

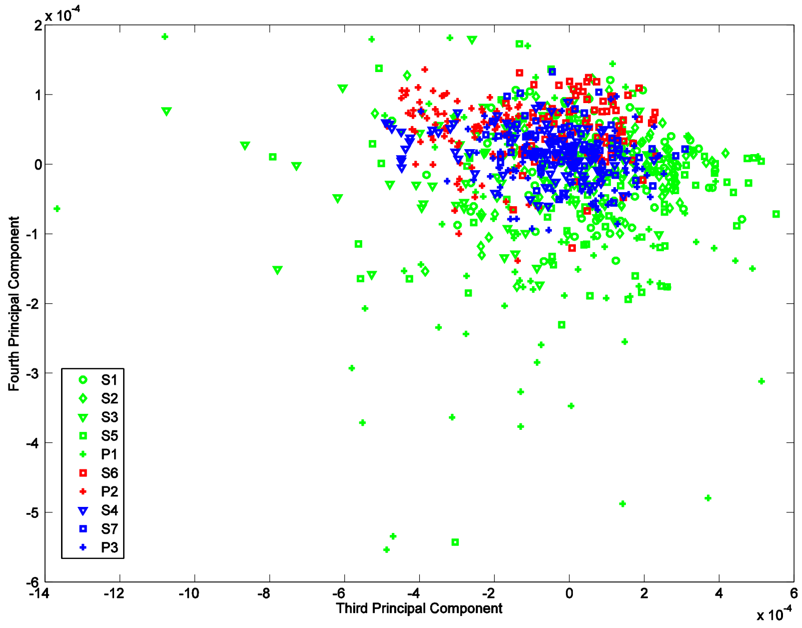


Figure A.5 PC3 vs. PC4 for all CH overtones

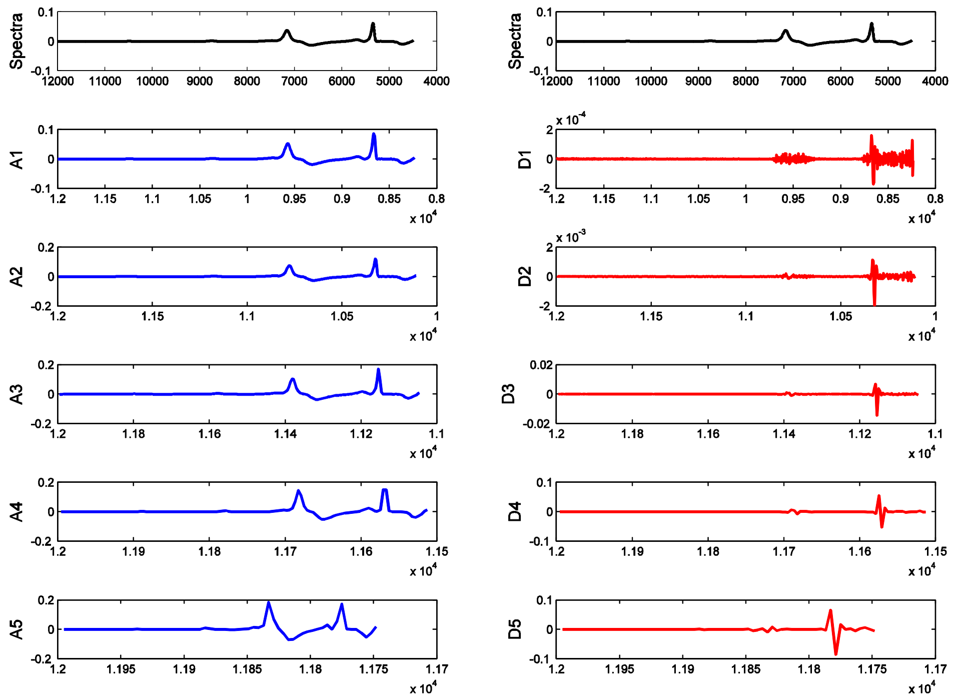


Figure A.6 Wavelet decomposition using db3 with 5 levels of decomposition

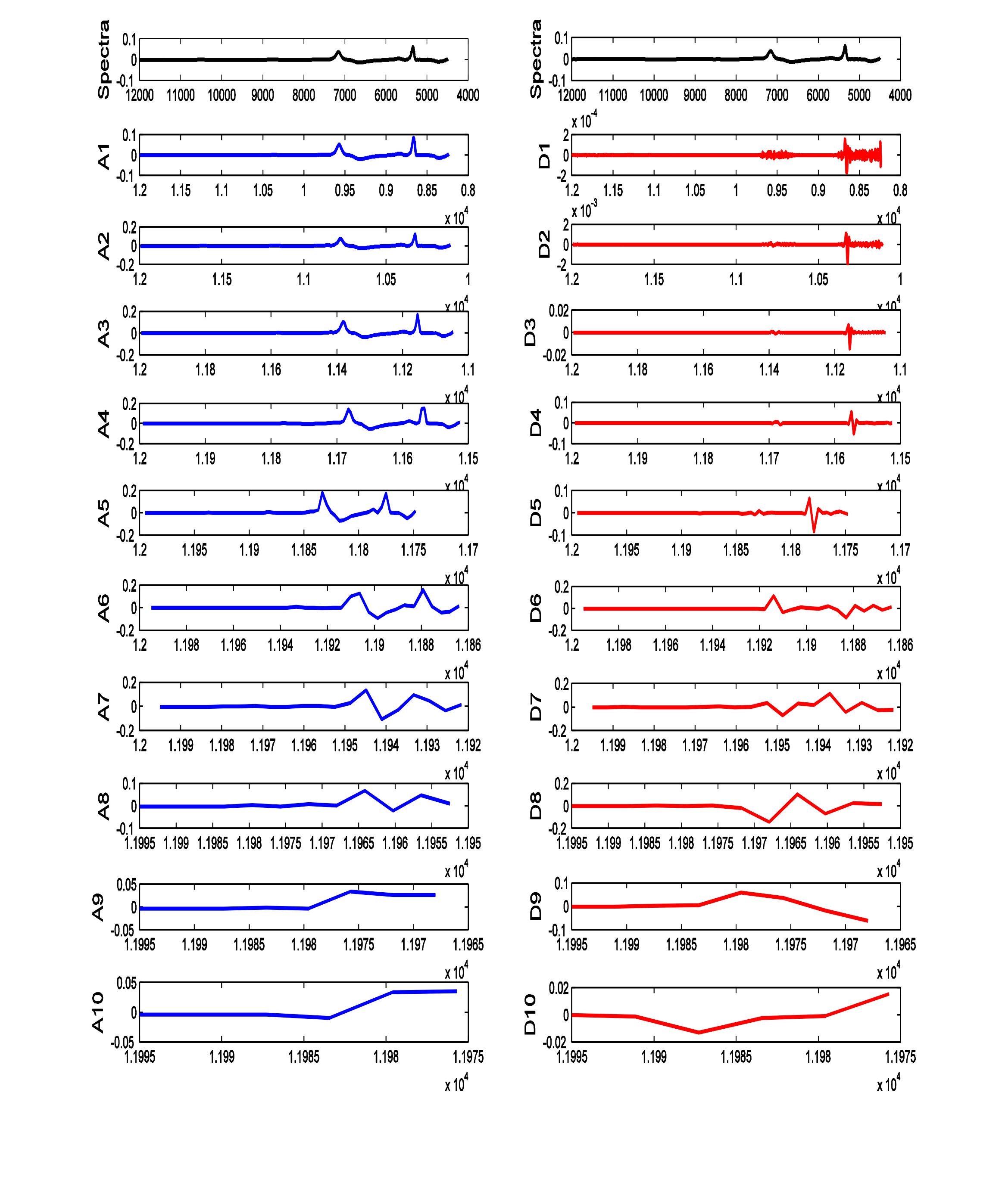
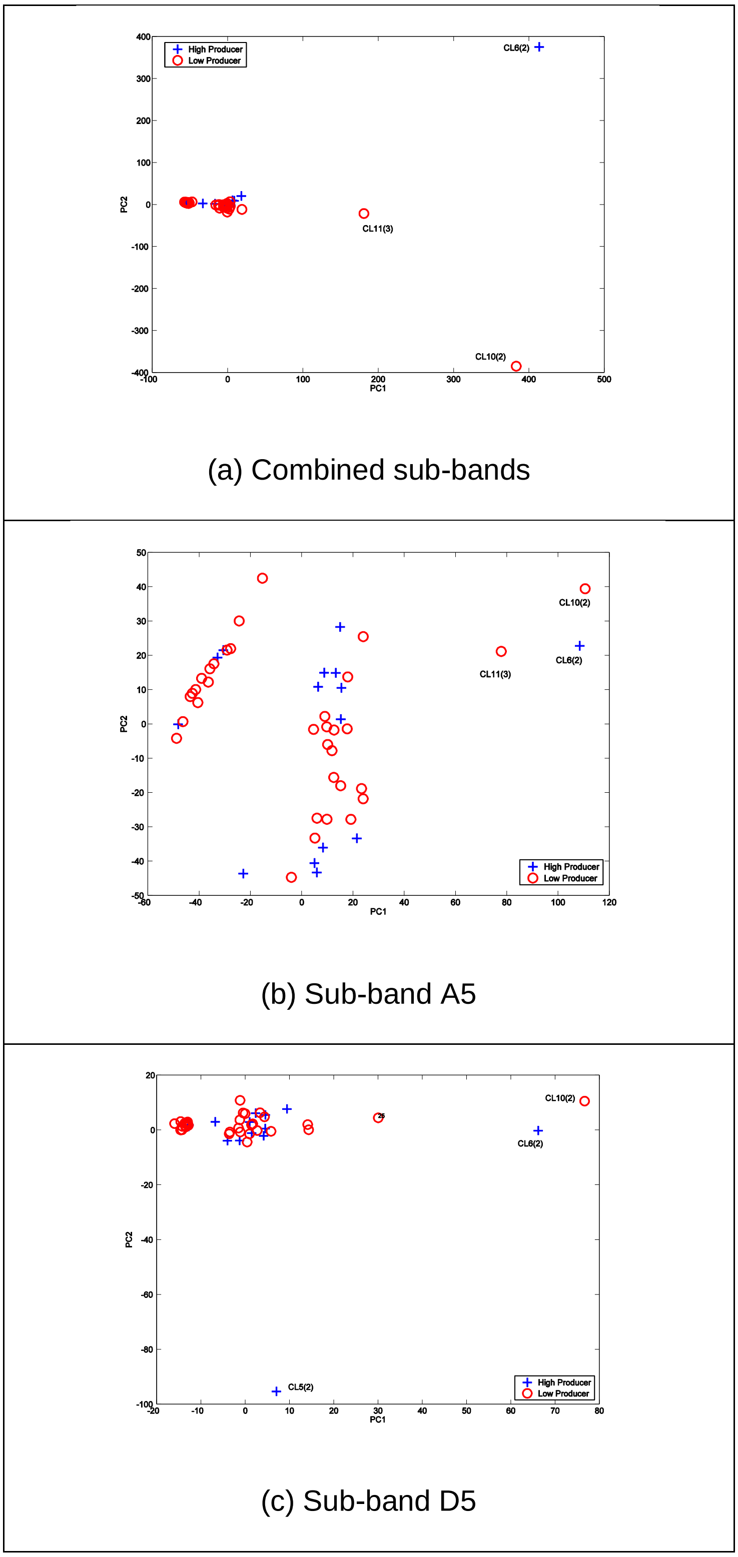


Figure A.7 Wavelet decomposition using db3 with 10 levels of decomposition

Appendix B: Appendices for Chapter 5- Enhanced Integrated Model Application on ESI Data

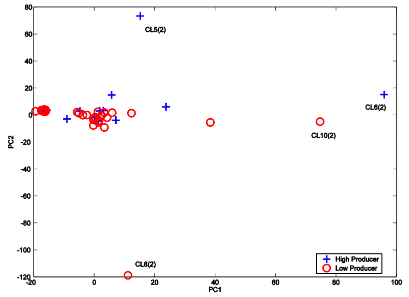


(a) Combined sub-bands

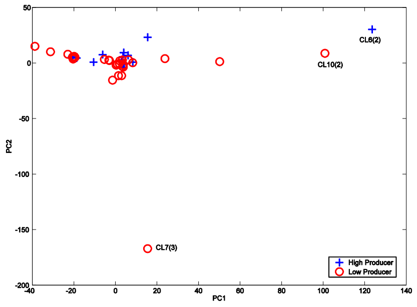
(b) Sub-band A5

(c) Sub-band D5

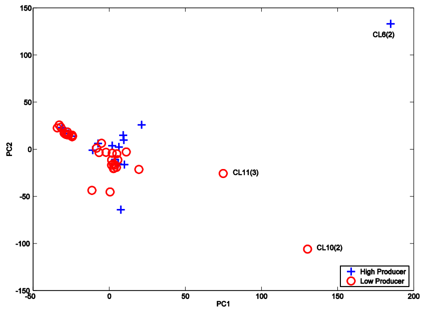
Figure B.8 Bivariate scores plot when CL11-2 is removed from the dataset for the combined sub-bands and sub-bands A5, D5 and D4



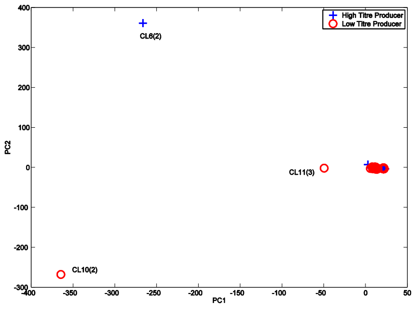
(d) Sub-band D4



(e) Sub-band D3



(g) Sub-band D2



(g) Sub-band D1

Figure B.9 Bivariate scores plot when CL11-2 is removed from the dataset for the sub-bands D3, D2 and D1

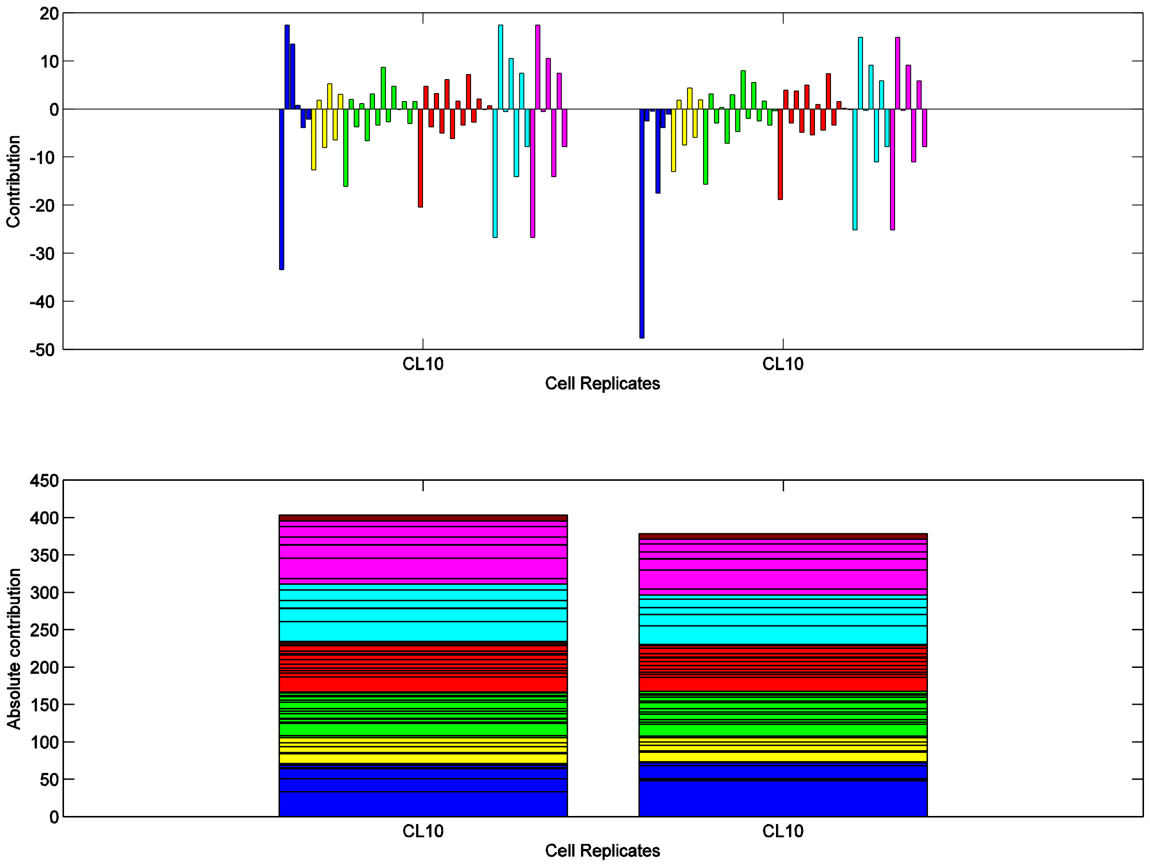


Figure B.1 Contribution plots of the wavelet sub-bands on PC1 to PC6 for cell line CL10 (high producing cell line)

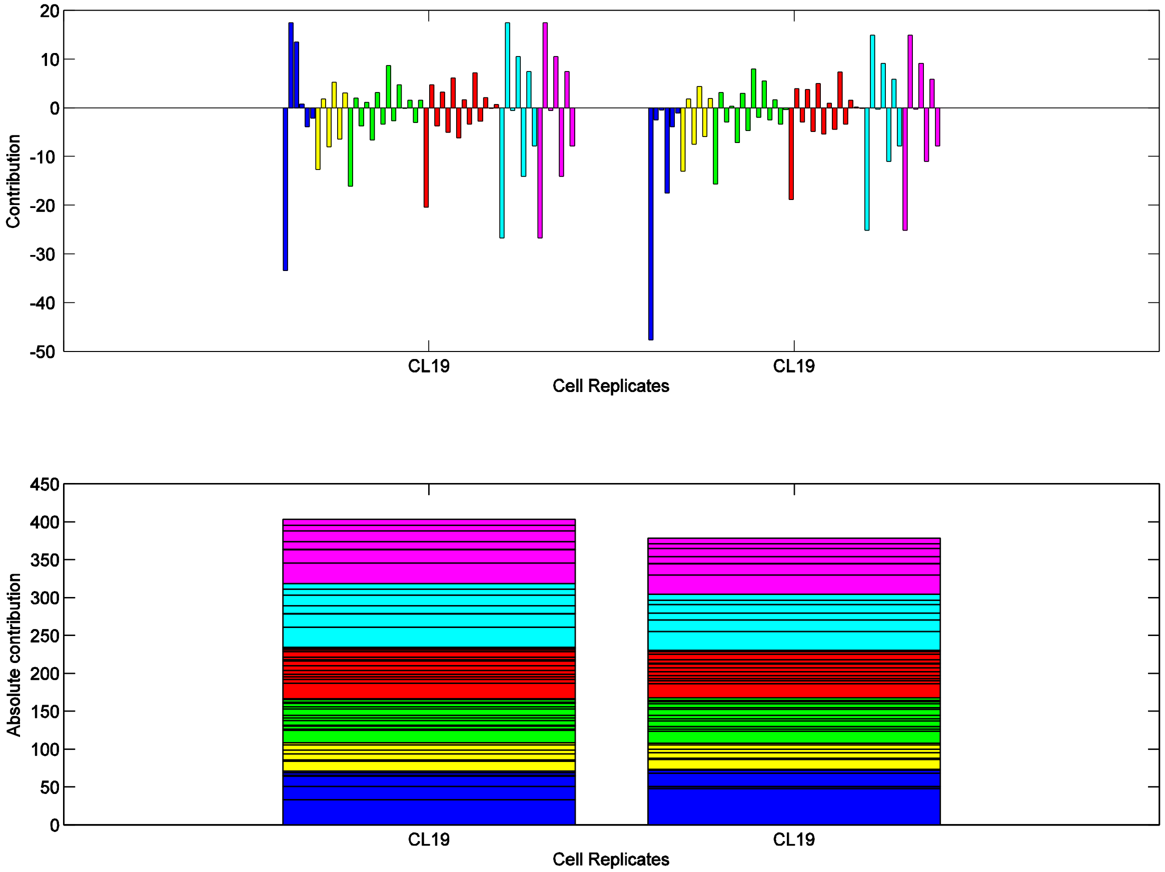


Figure B.2 Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL19 (low producing cell line)

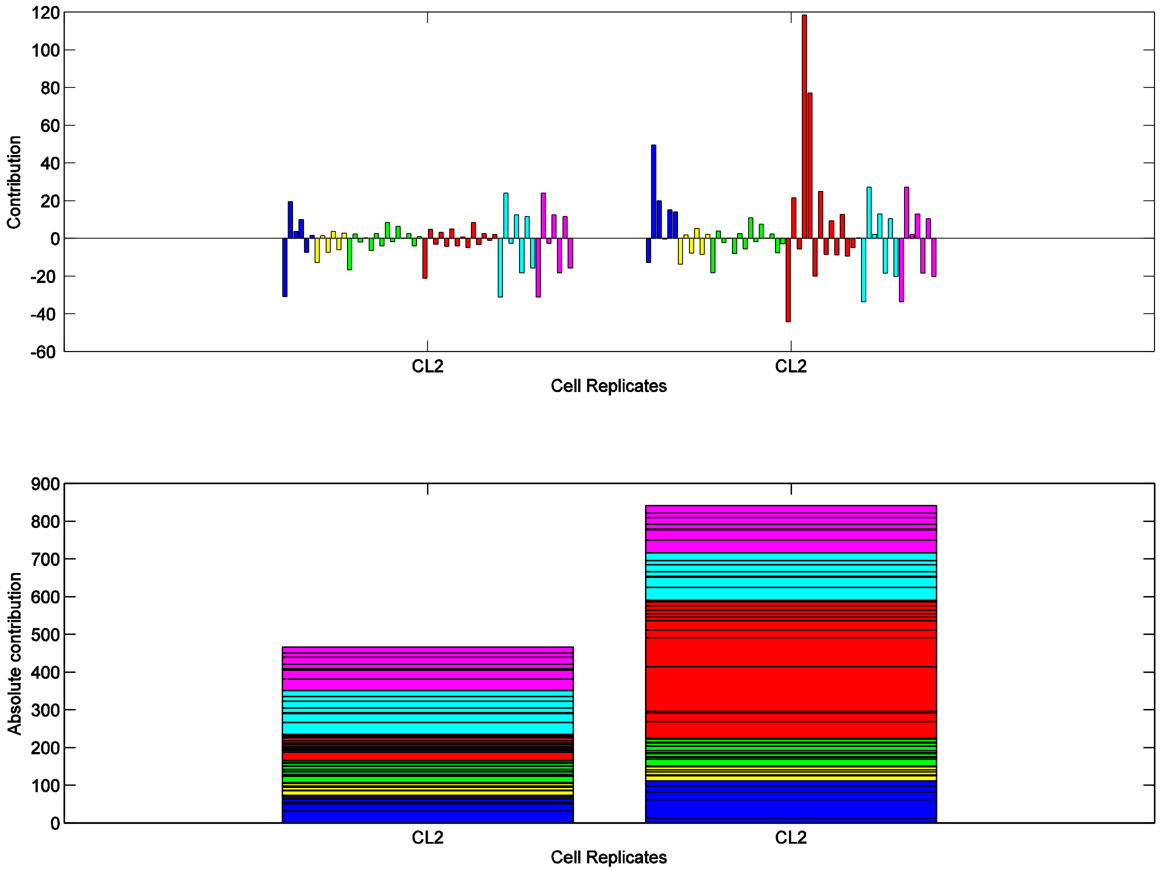


Figure B.3 Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL2 (low producing cell line)

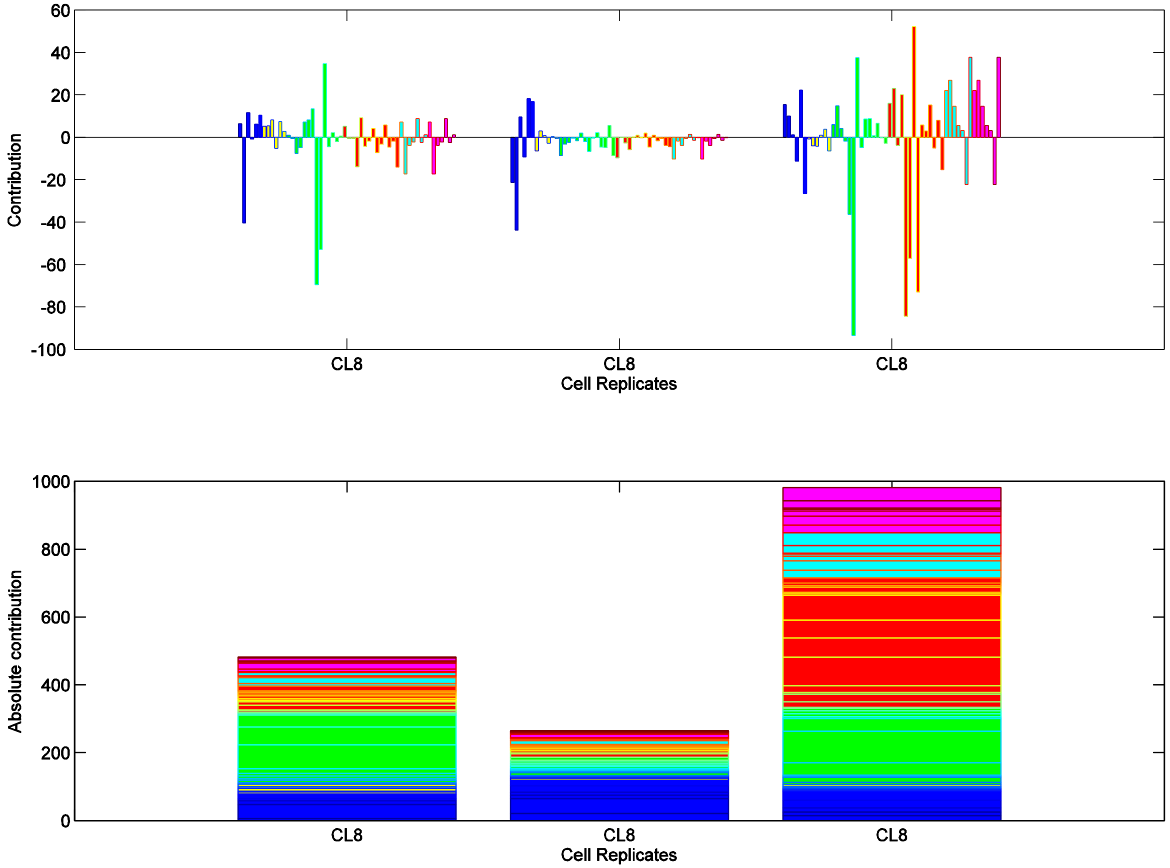


Figure B.4 Contribution plots of the wavelet sub-bands on P1 to PC6 for cell line CL4 (low producing cell line)

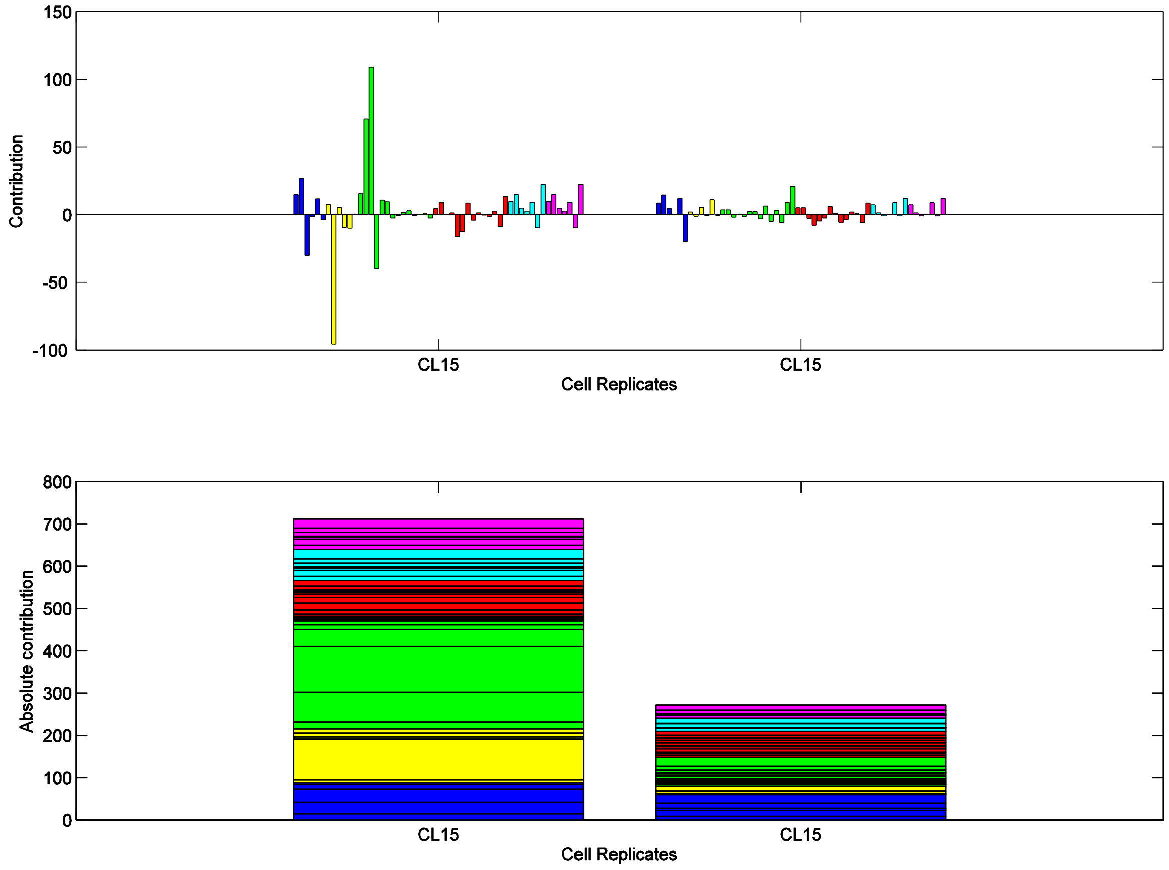


Figure B.5 Contribution plots of the variables on PC1 to PC6 for cell line CL15 (low producing cell line)

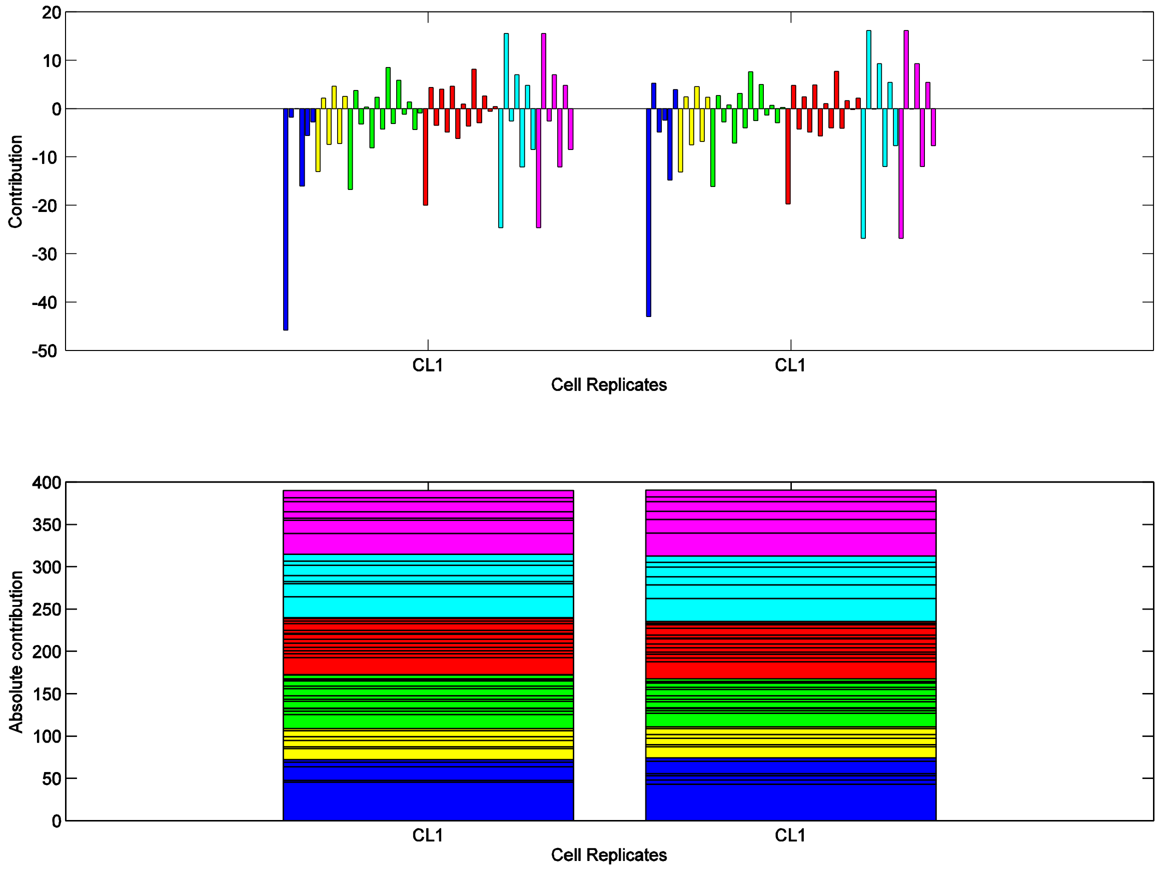


Figure B.6 Contribution plots of the variables on P1 to PC6 for cell line CL1 (low producing cell line)

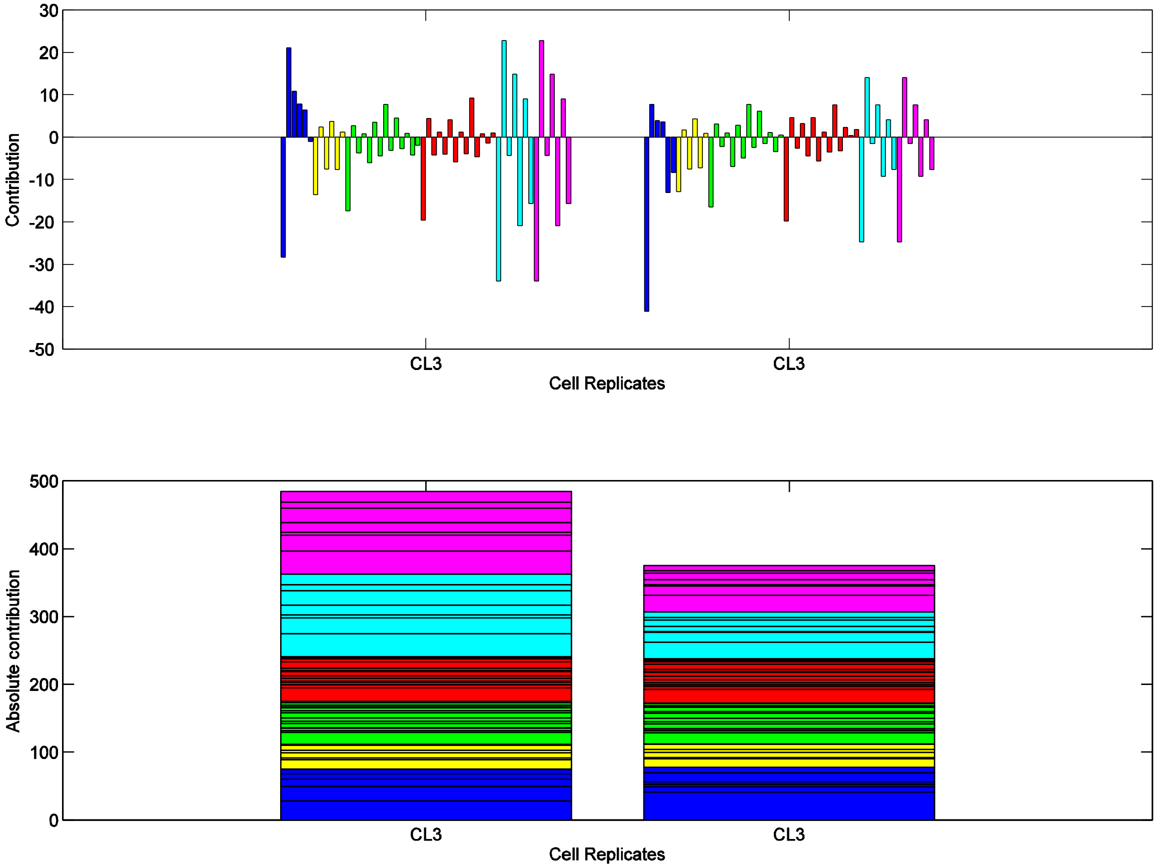


Figure B.7 Contribution plots of the variables on P1 to PC6 for cell line CL3 (low producing cell line)

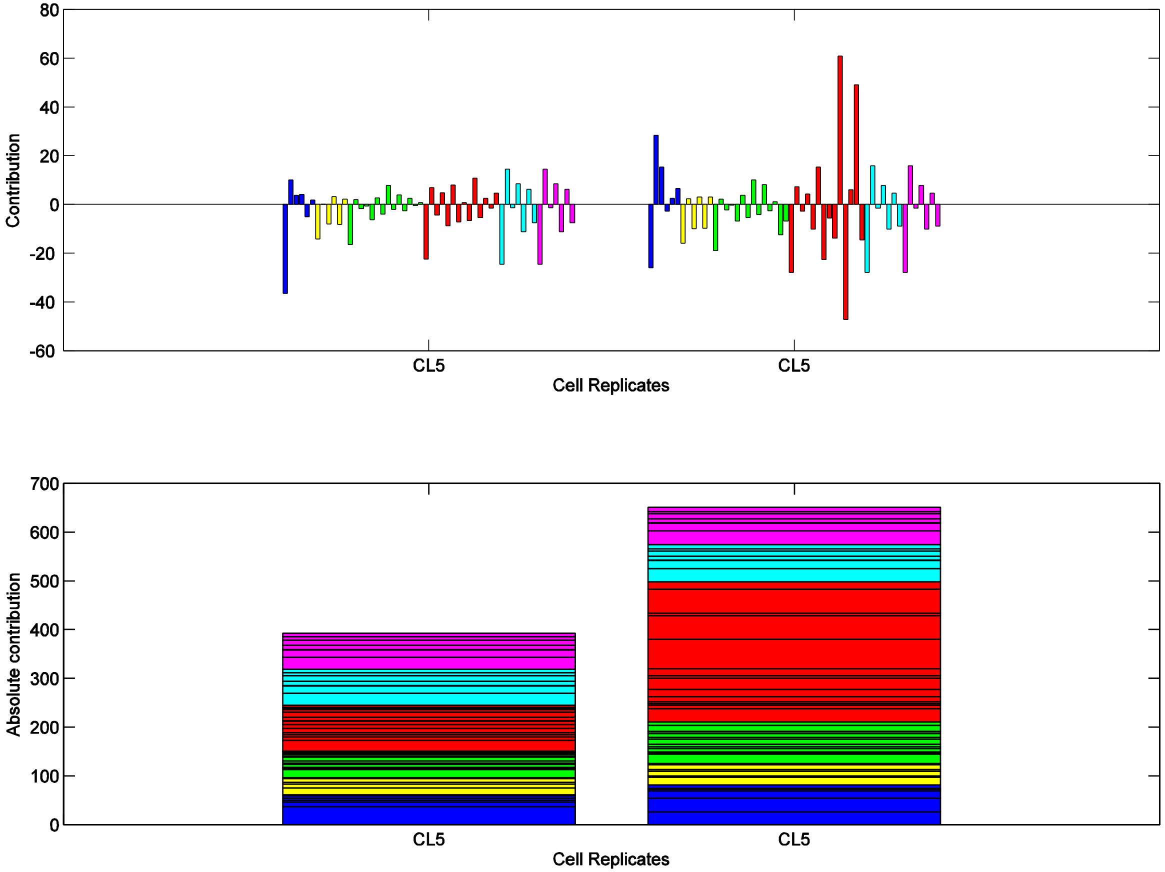


Figure B.8 Contribution plots of the variables on P1 to PC6 for cell line CL5 (low producing cell line)

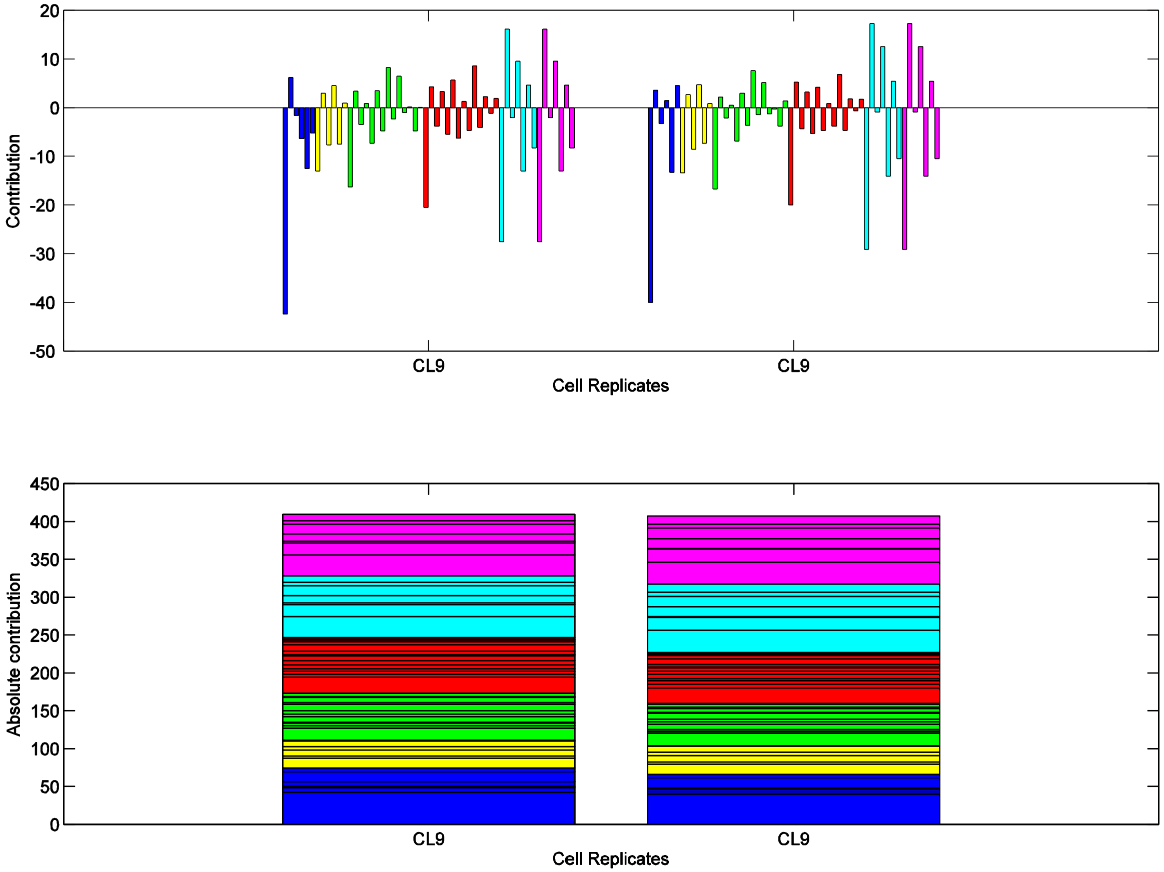


Figure B.9 Contribution plots of the variables on P1 to PC6 for cell line CL9 (low producing cell line)

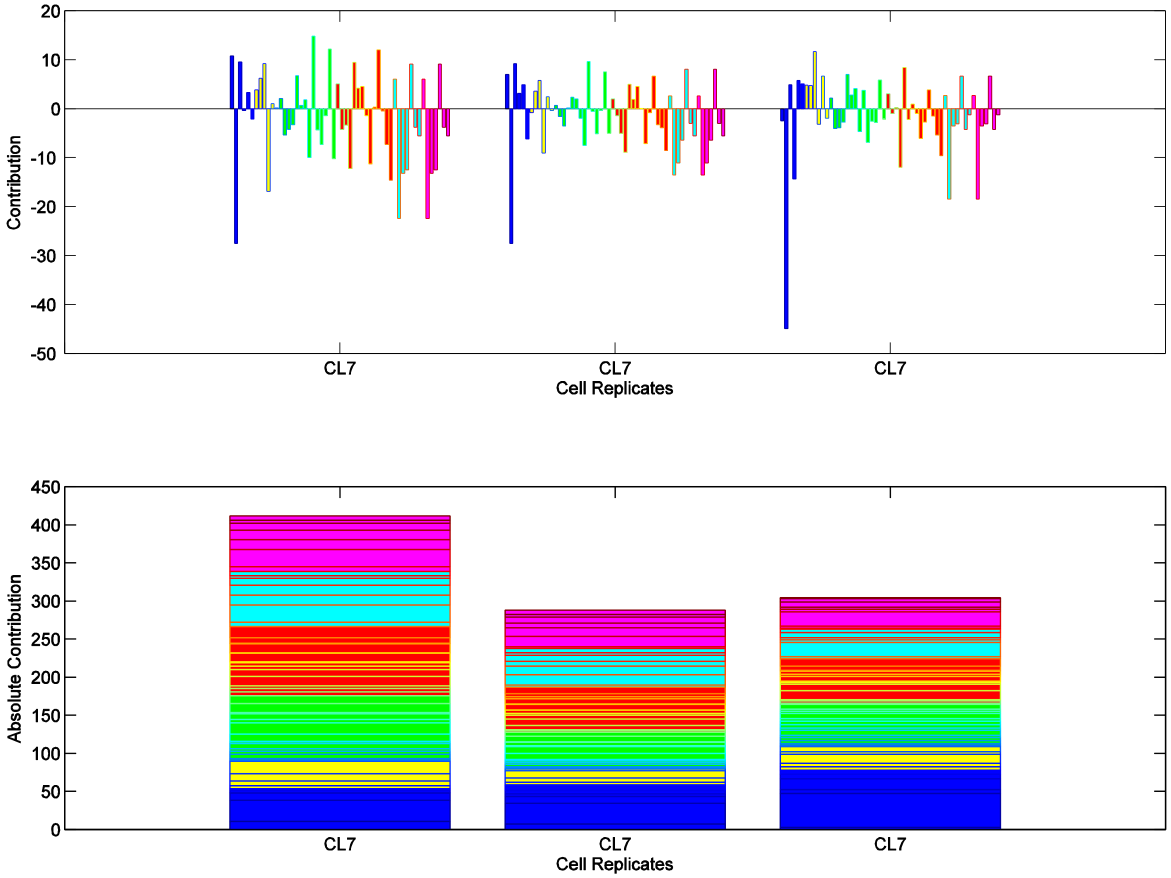


Figure B.10 Contribution plots of the variables on P1 to PC6 for cell line CL7 (low producing cell line)

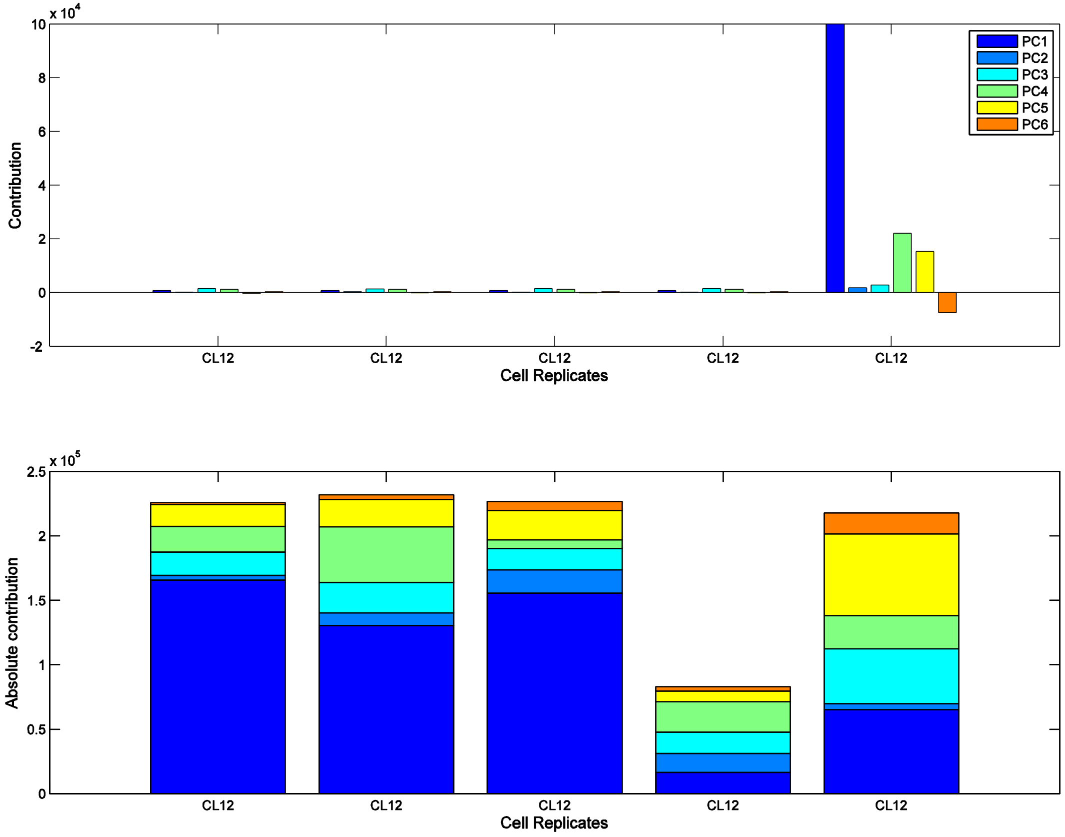


Figure B.11 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL12 and CL13 (high producing cell line)

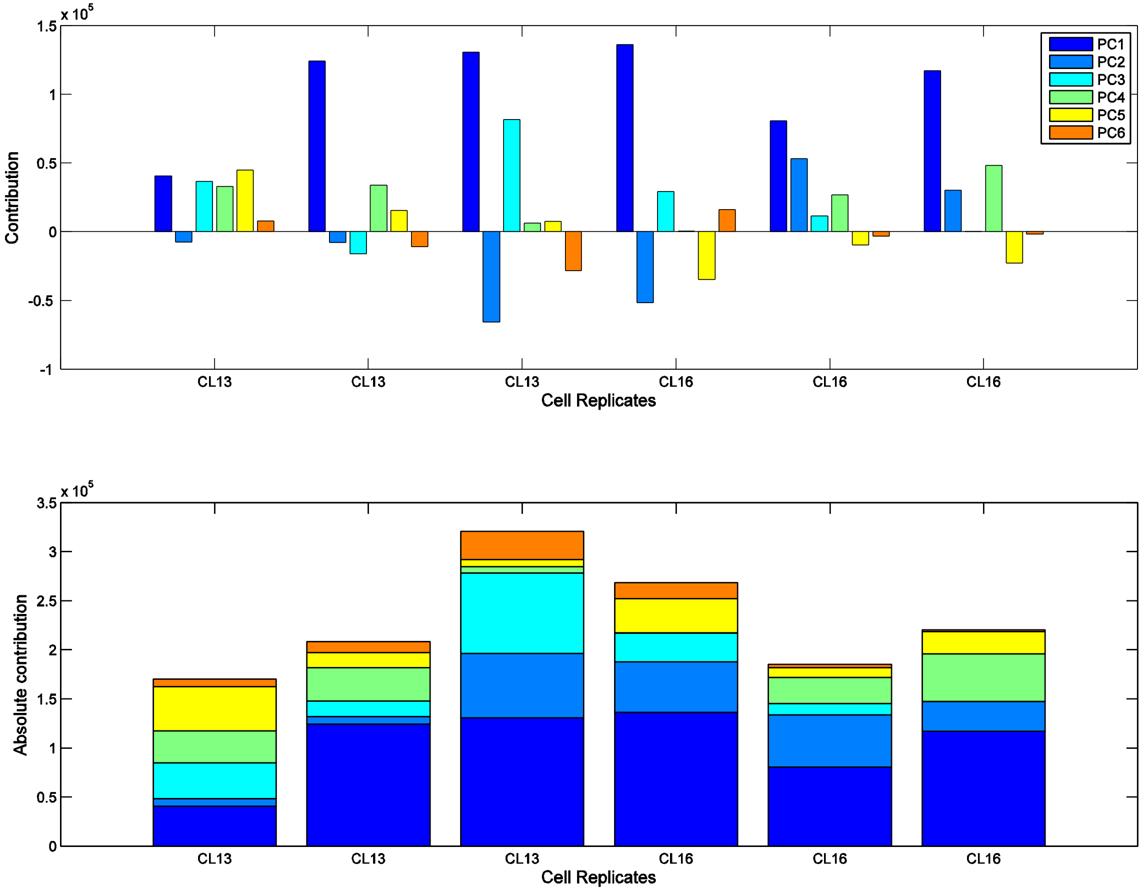


Figure B.12 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL13 and CL16 (high producing cell line)

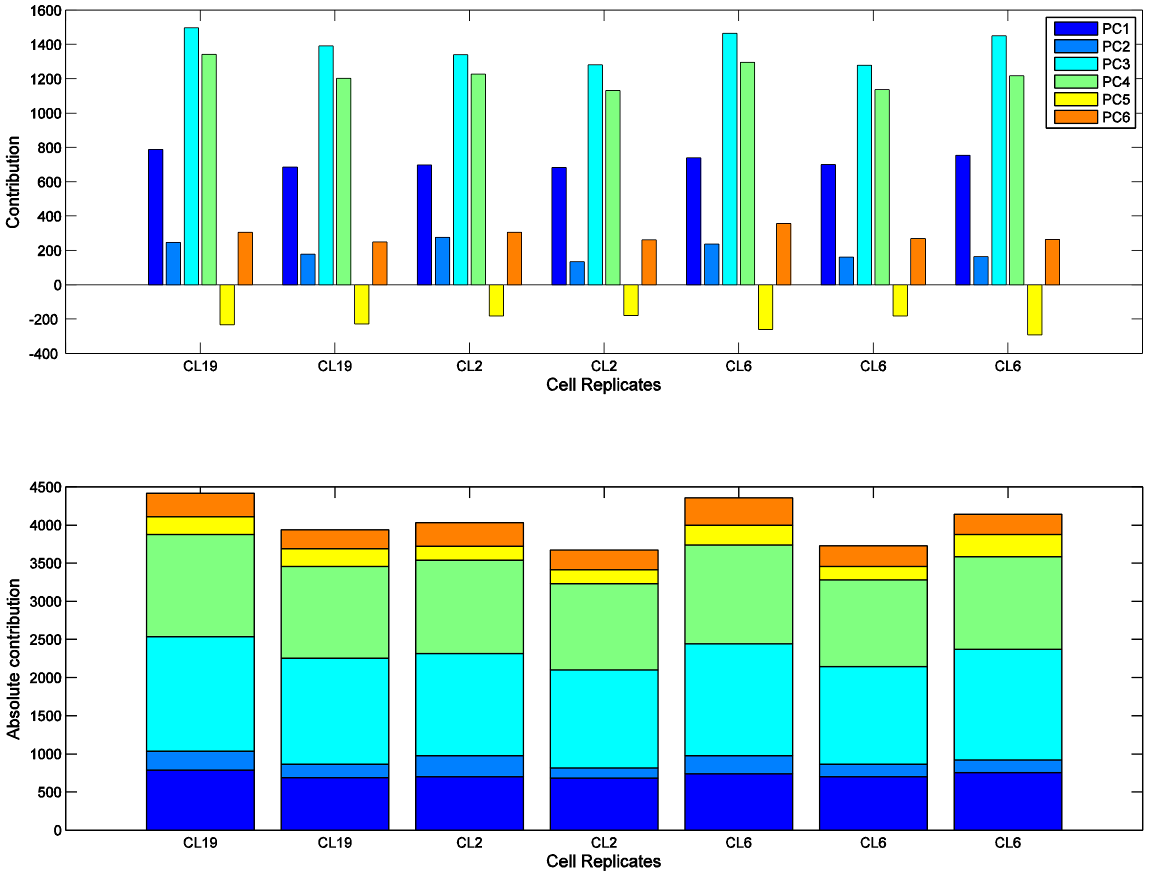


Figure B.13 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL19, CL2 and CL6 (low producing cell line)

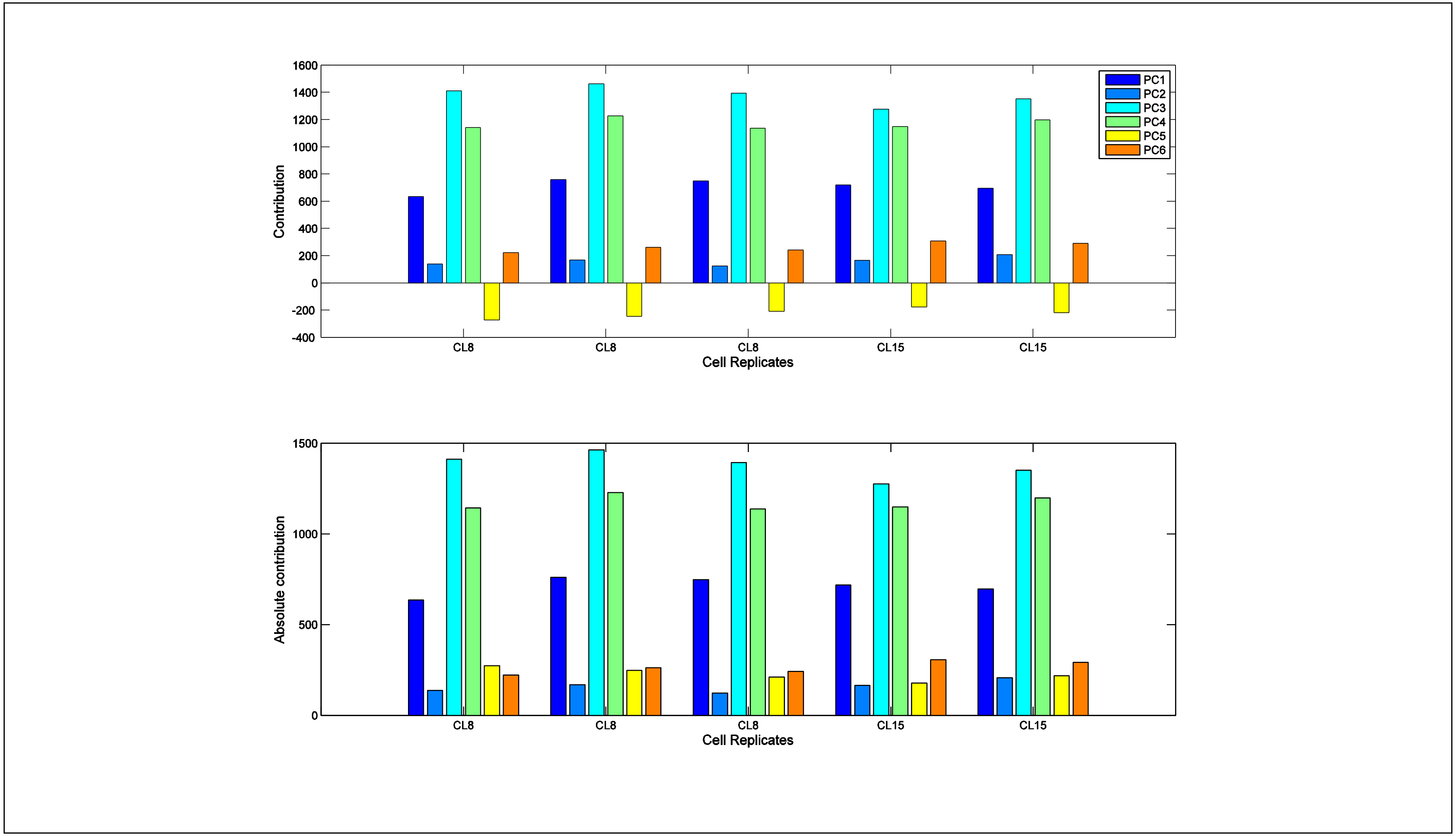


Figure B.14 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL8 and CL15 (low producing cell line)

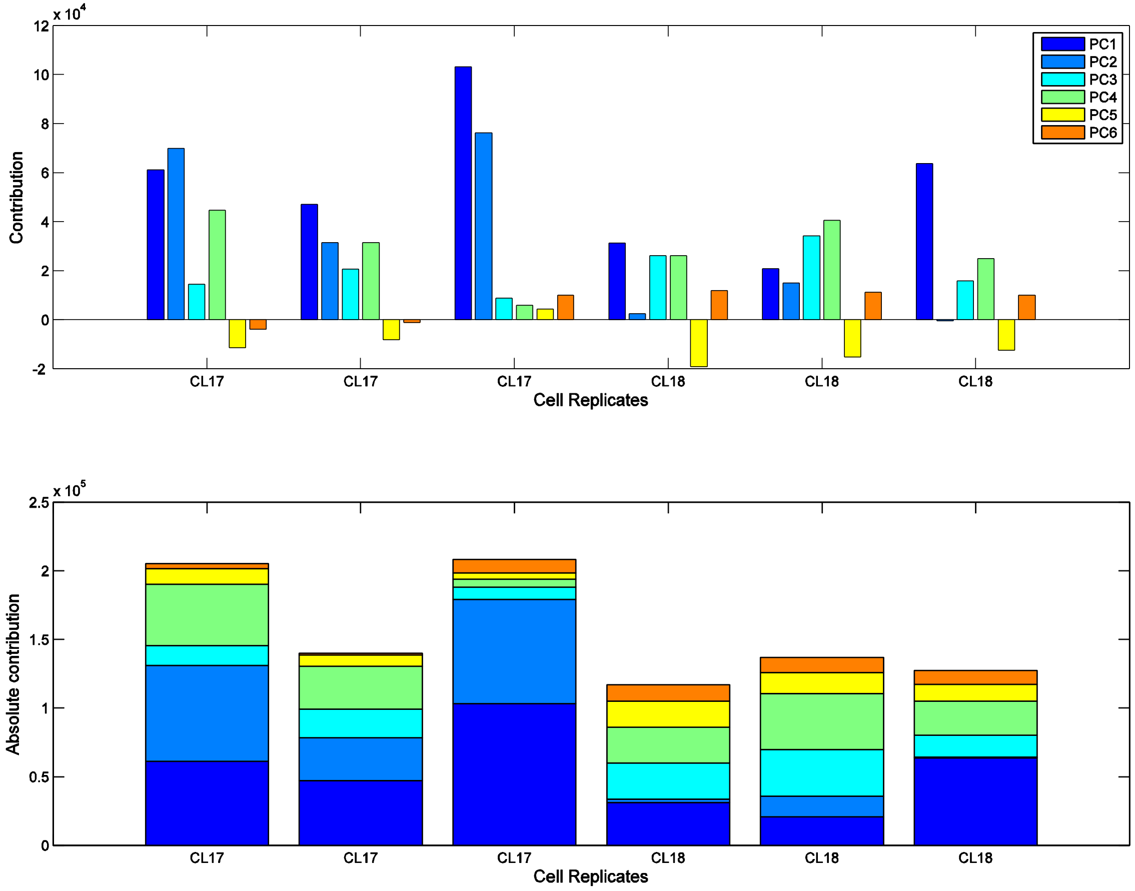


Figure B.15 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL17 and CL18 (low producing cell line)

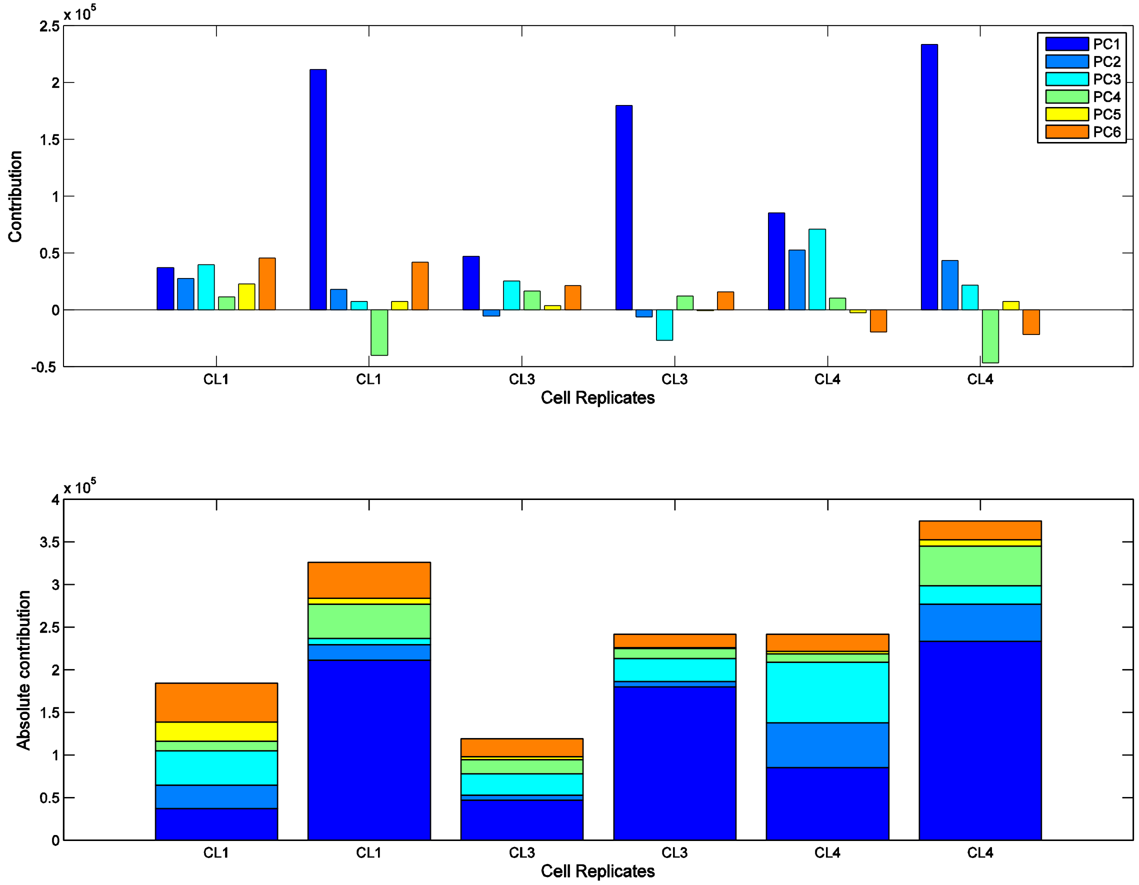


Figure B.16 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL1, CL3 and CL4 (low producing cell line)

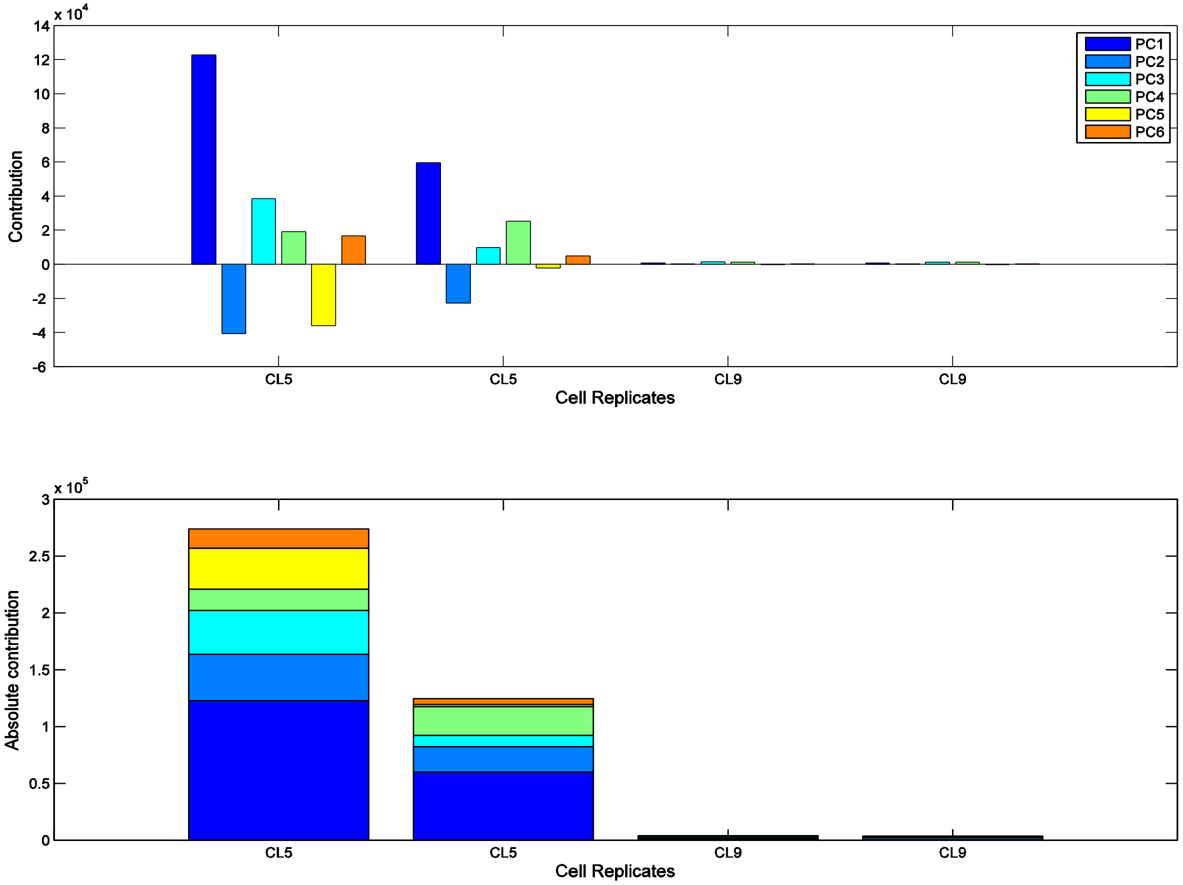


Figure B.17 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL5 and CL9 (low producing cell line)

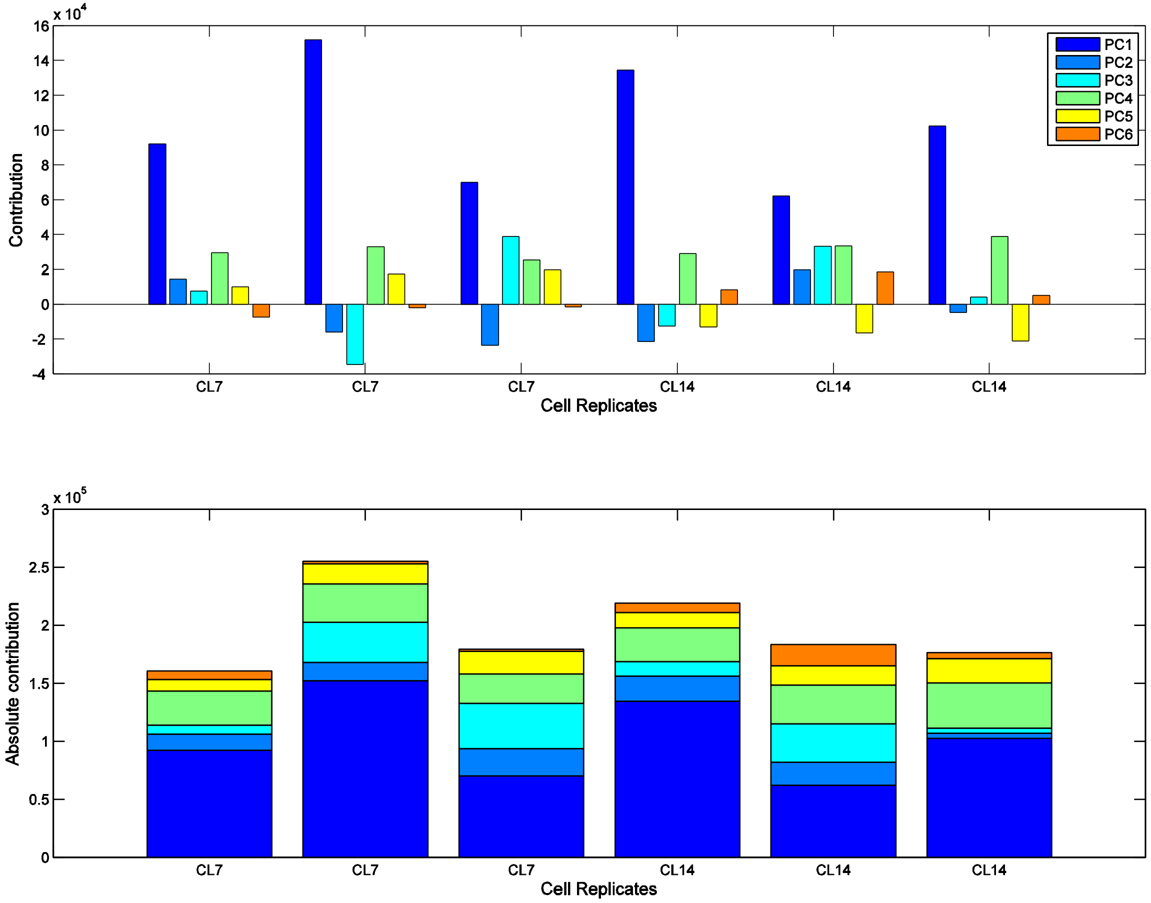


Figure B.18 Contribution plots of the ESI spectra on P1 to PC6 without discrete wavelet decomposition for cell lines CL7 and CL14 (low producing cell line)

References

Acar, E. and Yener, B. (2009) 'Unsupervised multiway data analysis: A literature survey', *IEEE Transactions on Knowledge and Data Engineering*, 21, (1), pp. 6-20.

Allard, E., Backstrom, D., Danielsson, R., Sjobberg, J. R. and Bergquist, J. (2008) 'Comparing capillary electrophoresis: Mass spectrometry fingerprints of urine samples obtained after intake of coffee, tea, or water', *Analytical Chemistry,* 80, (23), pp. 8946-8955.

Alterovitz, G. and Ramoni, M. F. (eds) (2007) *Systems Bioinformatics: An Engineering Case-Based Approach*. Boston, Massachusetts: Artech House.

Archibald, D. D. and Akin, D. E. (2000) 'Use of spectral window preprocessing for selecting near-infrared reflectance wavelengths for determination of the degree of enzymatic retting of intact flax stems', *Vibrational Spectroscopy,* 23, (2), pp. 169-180.

Arneberg, R., Rajalahti, T., Flikka, K., Berven, F. S., Kroksveen, A. C., Berle, M., Myhr, K. M., Vedeler, C. A., Ulvik, R. J. and Kvalheim, O. M. (2007) 'Pretreatment of mass spectral profiles: Application to proteomic data', *Analytical Chemistry,* 79, (18), pp. 7014-7026.

Arnold, S. A., Crowley, J., Woods, N., Harvey, L. M. and McNeill, B. (2003) 'In-situ near infrared spectroscopy to monitor key analytes in mammalian cell cultivation', *Biotechnology and Bioengineering,* 84, (1), pp. 13-19.

Arnold, S. A., Gaensakoo, R., Harvey, L. M. and McNeil, B. (2002) 'Use of at-line and in-situ near-infrared spectroscopy to monitor biomass in an industrial fed-batch Escherichia coli process', *Biotechnology and Bioengineering,* 80, (4), pp. 405-413.

Balabin, R. M. and Smirnov, S. V. (2011) 'Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data', *Analytica Chimica Acta,* 692, (1-2), pp. 63-72.

Barclay, V. J., Bonner, R. F. and Hamilton, I. P. (1997) 'Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression', *Analytical Chemistry,* 69, (1), pp. 78-90.

BCC Research (2012) 'Antibody drugs: Technologies and global markets'. Available at: http://www.bccresearch.com (Accessed: 14 December 2012).

Bhushan, N., Hadpe, S. and Rathore, A. S. (2011) 'Chemometrics applications in biotech processes: Assessing process comparability', *Biotechnology Progress,* 28, (1), pp. 121-128.

Borah, S., Hines, E. L. and Bhuyan, M. (2007) 'Wavelet transform based image texture analysis for size estimation applied to the sorting of tea granules', *Journal of Food Engineering,* 79, (2), pp. 629-639.

Bos, M. and Vrielink, J. A. M. (1994) 'The wavelet transform for preprocessing IR-spectra in the identification of monosubstituted and disubstituted benzenes', *Chemometrics and Intelligent Laboratory Systems,* 23, (1), pp. 115-122.

Brereton, R. (2009) *Chemometrics for pattern recognition.* Chichester: John Wiley & Sons.

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

Eugene N. B. (2000) *Biomedical Signal Processing and Signal Modeling*. New York: John Wiley & Sons.

Bruce, L. M., Koger, C. H. and Li, J. (2002) 'Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction', *IEEE Transactions on Geoscience and Remote Sensing,* 40, (10), pp. 2331-2338.

Buss, N., Henderson, S. J., McFarlane, M., Shenton, J. M. and de Haan, L. (2012) 'Monoclonal antibody therapeutics: history and future', *Current Opinion in Pharmacology,* 12, (5), pp. 615-622.

Butler, M. and Meneses-Acosta, A. (2012) 'Recent advances in technology supporting biopharmaceutical production from mammalian cells', *Applied Microbiology and Biotechnology,* 96, (4), pp. 885-894.

Cai, C. B., Han, Q. J., Tang, L. J., Nie, J. F., Ouyang, L. Q. and Yu, R. Q. (2008) 'Treating NIR data with orthogonal discrete wavelet transform: Predicting concentrations of a multi-component system through a small-scale calibration set', *Talanta,* 77, (2), pp. 822-826.

Candolfi, A., De Maesschalck, R., Jouan-Rimbaud, D., Hailey, P. A. and Massart, D. L. (1999) 'The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra', *Journal of Pharmaceutical and Biomedical Analysis,* 21, (1), pp. 115-132.

Carreno, I. R. and Vuskovic, M. (2007) 'Wavelet transform moments for feature extraction from temporal signals', *Informatics in Control, Automation and Robotics II*, pp. 235-242.

Cervera, A. E., Petersen, N., Lantz, A. E., Larsen, A. and Gernaey, K. V. (2009) 'Application of near infra-red spectroscopy for monitoring and control of cell culture and fermentation', *Biotechnology Progress,* 25, (6), pp. 1561-1581.

Chau, F.-T., Liang, Y.-Z., Gao, J. and Shao, X.-G. (2004) *Chemometrics: From basics to wavelet transform.* New York: John Wiley & Sons.

Chen, D., Shao, X. G., Hu, B. and Su, Q. D. (2004) 'A Background and noise elimination method for quantitative calibration of near infrared spectra', *Analytica Chimica Acta,* 511, (1), pp. 37-45.

Clementschitsch, F. and Bayer, K. (2006) 'Improvement of bioprocess monitoring: Development of novel concepts', *Microbial Cell Factories,* 5, pp. 19-29

Coifman, R. R. and Wickerhauser, M. V. (1992) 'Entropy-based algorithms for best basis selection', *IEEE Transactions on Information Theory,* 38, (2), pp. 713-718.

Conlin, A. K., Martin, E. B. and Morris, A. J. (2000) 'Confidence limits for contribution plots', *Journal of Chemometrics,* 14, (5-6), pp. 725-736.

Cunha, C. C. F., Glassey, J., Montague, G. A., Albert, S. and Mohan, P. (2002) 'An assessment of seed quality and its influence on productivity estimation in an industrial antibiotic fermentation', *Biotechnology and Bioengineering,* 78, (6), pp. 658-669.

Cvetkovic, D., Ubeyli, E. D. and Cosic, I. (2008) 'Wavelet transform feature extraction from human PPG, ECG, and EEG signal responses to ELF PEMF exposures: A pilot study', *Digital Signal Processing,* 18, (5), pp. 861-874.

Damen, C. W. N., Chen, W., Chakraborty, A. B., van Oosterhout, M., Mazzeo, J. R., Gebler, J. C., Schellens, J. H. M., Rosing, H. and Beijnen, J. H. (2009) 'Electrospray ionization quadrupole ion-mobility time-of-flight mass spectrometry as a tool to distinguish the lot-to-lot heterogeneity in N-glycosylation profile of the therapeutic monoclonal antibody trastuzumab', *Journal of the American Society for Mass Spectrometry,* 20, (11), pp. 2021-2033.

Danielsson, R., Allard, E., Sjoberg, P. J. R. and Bergquist, J. (2011) 'Exploring liquid chromatography-mass spectrometry fingerprints of urine samples from patients with prostate or urinary bladder cancer', *Chemometrics and Intelligent Laboratory Systems,* 108, pp. 33-48.

Daubechies, I. (1992) *Ten Lectures on Wavelets.* Philadelphia: Society for Industrial and Applied Mathematics.

Davis, J. M. (2011) *Animal Cell Culture: Essential Methods*. New York: John Wiley & Sons.

de Souza, P. P., Augusti, D. V., Catharino, R. R., Siebald, H. G. L., Eberlin, M. N. and Augusti, R. (2007) 'Differentiation of rum and Brazilian artisan cachaca via electrospray ionization mass spectrometry fingerprinting', *Journal of Mass Spectrometry,* 42, (10), pp. 1294-1299.

Esteban-Diez, I., Gonzalez-Saiz, J. M. and Pizarro, C. (2004) 'OWAVEC: a combination of wavelet analysis and an orthogonalization algorithm as a pre-processing step in multivariate calibration', *Analytica Chimica Acta,* 515, (1), pp. 31-41.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases', *AI Magazine,* 17, (3), pp. 37-54.

Feng, G. C., Yuen, P. C. and Dai, D. Q. (2000) 'Human face recognition using PCA on wavelet sub-band', *Journal of Electronic Imaging,* 9, (2), pp. 226-233.

Ferreira, A. P., Alves, T. P. and Menezes, J. C. (2005) 'Monitoring complex media fermentations with near-infrared spectroscopy: Comparison of different variable selection methods', *Biotechnology and Bioengineering,* 91, (4), pp. 474-481.

Ferreira, A. P., Lopes, J. A. and Menezes, J. C. (2007) 'Study of the application of multiway multivariate techniques to model data from an industrial fermentation process', *Analytica Chimica Acta,* 595, (1-2), pp. 120-127.

Fonville, J. M., Carter, C., Cloarec, O., Nicholson, J. K., Lindon, J. C., Bunch, J. and Holmes, E. (2011) 'Robust dataprocessing and normalization strategy for MALDI mass spectrometric imaging', *Analytical Chemistry,* 84, (3), pp. 1310-1319.

Ge, Z. H., Cavinato, A. G. and Callis, J. B. (1994) 'Noninvasive spectroscopy for monitoring cell-density in a fermentation process', *Analytical Chemistry,* 66, (8), pp. 1354-1362.

Geladi, P. and Grahn, H. (1996) *Multivariate Image Analysis.* Chichester: John Wiley & Sons.

Gollmer, K. and Posten, C. (1996) 'Supervision of bioprocesses using a dynamic time warping algorithm', *Control Engineering Practice,* 4, (9), pp. 1287-1295.

Goodman, M. (2009) 'Market watch: Sales of biologics to show robust growth through to 2013', *Nature Reviews Drug Discovery*, 8, (11), pp. 837. Available at: http://www.nature.com (Accessed: 5 November 2012).

Gopalakrishnan, S. and Mitra, M. (2010) *Wavelet Methods for Dynamical Problems: With Application to Metallic, Composite, and Nano-Composite Structures.* Florida: CRC Press.

Griffiths, W. J., Jonsson, A. P., Liu, S. Y., Rai, D. K. and Wang, Y. Q. (2001) 'Electrospray and tandem mass spectrometry in biochemistry', *Biochemical Journal,* 355, pp. 545-561.

Gunther, J. C., Conner, J. S. and Seborg, D. E. (2007) 'Fault detection and diagnosis in an industrial fed-batch cell culture process', *Biotechnology Progress,* 23, (4), pp. 851-857.

Gurden, S. P., Westerhuis, J. A. and Smilde, A. K. (2002) 'Monitoring of batch processes using spectroscopy', *AICHE Journal,* 48, (10), pp. 2283-2297.

Gurden, S. P., Westerhuis, J. A., Bro, R. and Smilde, A. K. (2001) 'A comparison of multiway regression and scaling methods', *Chemometrics and Intelligent Laboratory Systems,* 59, (1-2), pp. 121-136.

Hilario, M., Kalousis, A., Pellegrini, C. and Muller, M. (2006) 'Processing and classification of protein mass spectra', *Mass Spectrometry Reviews,* 25, (3), pp. 409-449.

Hagman, A. and Sivertsson, P. (1998) 'The use of NIR spectroscopy in monitoring and controlling bioprocesses', *Process Control and Quality,* 11, (2), pp. 125-128.

Hubbard, B. B. (1996) *The World According to Wavelets : The Story of a Mathematical Technique in The Making.* Wellesley, Massachusetts: A.K. Peters.

Jahankhani, P., Kodogiannis, V. and Revett, K. (2006) 'EEG signal classification using wavelet feature extraction and neural networks', *IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing, Proceedings*, pp. 120-124.

Jenzsch, M., Gnoth, S., Kleinschmidt, M., Simutis, R. and Lubbert, A. (2006) 'Improving the batch-to-batch reproducibility in microbial cultures during recombinant protein production by guiding the process along a predefined total biomass profile', *Bioprocess and Biosystems Engineering,* 29, (5-6), pp. 315-321.

Jetter, K., Depczynski, U., Molt, K. and Niemoller, A. (2000) 'Principles and applications of wavelet transformation of chemometrics', *Analytica Chimica Acta,* 420, (2), pp. 169-180.

Johnson-Leger, C., Power, C. A., Shomade, G., Shaw, J. P. and El Proudfoot, A. (2006) 'Protein therapeutics - lessons learned and a view of the future', *Expert Opinion on Biological Therapy,* 6, (1), pp. 1-7.

Jolliffe, I. T. (2002) *Principal component analysis.* 2nd ed: New York: Springer.

Jouan-Rimbaud, D., Walczak, B., Poppi, R. J., deNoord, O. E. and Massart, D. L. (1997) 'Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration', *Analytical Chemistry,* 69, (21), pp. 4317-4323.

Jungbauer, A. and Gobel, U. (2011) 'Biopharmaceutical process development - shortcut to market: An interview with Rolf Werner from Boehringer Ingelheim', *Biotechnology Journal,* 7, (1), pp. 14-16.

Katajamaa, M. and Oresic, M. (2007) 'Data processing for mass spectrometry-based metabolomics', *Journal of Chromatography A,* 1158, (1-2), pp. 318-328.

Kassidas, A., MacGregor, J. F. and Taylor, P. A. (1998) 'Synchronization of batch trajectories using dynamic time warping', *AICHE Journal,* 44, (4), pp. 864-875.

Kelley, B. (2009) 'Industrialization of mAb production technology: The bioprocessing industry at a crossroads', *Mabs,* 1, (5), pp. 443-452.

Kirdar, A. O., Conner, J. S., Baclaski, J. and Rathore, A. S. (2007) 'Application of multivariate analysis toward biotech processes: Case study of a cell-culture unit operation', *Biotechnology Progress,* 23, (1), pp. 61-67.

Kirdar, A. O., Green, K. D. and Rathore, A. S. (2008) 'Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application', *Biotechnology Progress,* 24, (3), pp. 720-726.

Kourti, T. (2005) 'Application of latent variable methods to process control and multivariate statistical process control in industry', *International Journal of Adaptive Control and Signal Processing,* 19, (4), pp. 213-246.

Kourti, T., Lee, J. and MacGregor, J. F. (1996) 'Experiences with industrial applications of projection methods for multivariate statistical process control', *Computers & Chemical Engineering,* 20, pp. S745-S750.

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), pp. 277-284.

Kozlowski, S. and Swann, P. (2006) 'Current and future issues in the manufacturing and development of monoclonal antibodies', *Advanced Drug Delivery Reviews,* 58, (5-6), pp. 707-722.

Krishnan, S., Vogels, J. T. W. E., Coulier, L., Bas, R. C., Hendriks, M. W. B., Hankemeier, T. and Thissen, U. (2012) 'Instrument and process independent binning and baseline correction methods for liquid chromatography-high resolution-mass spectrometry deconvolution', *Analytica Chimica Acta,* 740, pp. 12-9.

Lhoest, J. B., Wagner, M. S., Tidwell, C. D. and Castner, D. G. (2001) 'Characterization of adsorbed protein films by time of flight secondary ion mass spectrometry', *Journal of Biomedical Materials Research,* 57, (3), pp. 432-440.

Li, F., Church, G., Janakiram, M., Gholston, H. and Runger, G. (2011) 'Fault detection for batch monitoring and discrete wavelet transforms', *Quality and Reliability Engineering International,* 27, (8), pp. 999-1008.

Li, F., Vijayasankaran, N., Shen, A., Kiss, R. and Amanullah, A. (2010) 'Cell culture processes for monoclonal antibody production', *Mabs,* 2, (5), pp. 466-479.

Lio, P. (2003) 'Wavelets in bioinformatics and computational biology: state of art and perspectives', *Bioinformatics,* 19, (1), pp. 2-9.

Lodish, H. F. and Berk, A. (2012) *Molecular Cell Biology.*7th ed. Basingtoke: W. H. Freeman

Luo, Y. and Chen, G. X. (2007) 'Combined approach of NMR and chemometrics for screening peptones used in the cell culture medium for the production of a recombinant therapeutic protein', *Biotechnology and Bioengineering,* 97, (6), pp. 1654-1659.

Luypaert, J., Massart, D. L. and Heyden, Y. V. (2007) 'Near-infrared spectroscopy applications in pharmaceutical analysis', *Talanta,* 72, (3), pp. 865-883.

Mallat, S. G. (1998) *A Wavelet Tour of Signal Processing.* San Diego: Academic Press.

Mark, J., Andre, M., Karner, M. and Huck, C. W. (2010) 'Prospects for multivariate classification of a pharmaceutical intermediate with near-infrared spectroscopy as a process analytical technology (PAT) production control supplement', *European Journal of Pharmaceutics and Biopharmaceutics,* 76, (2), pp. 320-327.

Martin-Moe, S., Lim, F. J., Wong, R. L., Sreedhara, A., Sundaram, J. and Sane, S. U. (2011) 'A new roadmap for biopharmaceutical drug product development: Integrating development, validation, and quality by design', *Journal of Pharmaceutical Sciences,* 100, (8), pp. 3031-3043.

Mattoli, L., Cangi, F., Ghiara, C., Burico, M., Maidecchi, A., Bianchi, E., Ragazzi, E., Bellotto, L., Seraglia, R. and Traldi, P. (2011) 'A metabolite fingerprinting for the characterization of commercial botanical dietary supplements', *Metabolomics,* 7, (3), pp. 437-445.

McShane, M. J. and Cote, G. L. (1998) 'Near-infrared spectroscopy for determination of glucose lactate, and ammonia in cell culture media', *Applied Spectroscopy,* 52, (8), pp. 1073-1078.

Meyer, Y. and Ryan, R. D. (1993) *Wavelets: Algorithms & Applications.* Philadelphia: Society for Industrial and Applied Mathematics.

Meyer, Y. and Salinger, D. H. (1992) *Wavelets and Operators.* Cambridge; New York: Cambridge University Press.

Namkung, H., Lee, Y. and Chung, H. (2008) 'Improving prediction selectivity for on-line near-infrared monitoring of components in etchant solution by spectral range optimization', *Anal Chim Acta,* 606, (1), pp. 50-6.

Nicolaides, N. C., Sass, P. M. and Grasso, L. (2006) 'Monoclonal antibodies: A morphing landscape for therapeutics', *Drug Development Research,* 67, (10), pp. 781-789.

Nielsen, N. J., Tomasi, G., Frandsen, R. J. N., Kristensen, M. B., Nielsen, J., Giese, H. and Christensen, J. H. (2010) 'A pre-processing strategy for liquid chromatography time-of-flight mass spectrometry metabolic fingerprinting data', *Metabolomics,* 6, (3), pp. 341-352.

Nomikos, P. (1996) 'Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis World Batch Forum, Toronto, May 1996', *ISA Transactions,* 35, (3), pp. 259-266.

Nomikos, P. and MacGregor, J. F. (1994) 'Monitoring batch processes using multiway principal component analysis', *AICHE Journal,* 40, (8), pp. 1361-1375.

Petersen, N., Odman, P., Padrell, A. E. C., Stocks, S., Lantz, A. E. and Gernaey, K. V. (2010) 'In situ near infrared spectroscopy for analyte-specific monitoring of glucose and ammonium in streptomyces coelicolor fermentations', *Biotechnology Progress,* 26, (1), pp. 263-271.

Philippidis, A. (2012) 'Large molecules continue to gain favor', *Genetic Engineering & Biotechnology News*, 32, (10). Available at: http://genengnews.com (Accessed: 17 October 2012).

Pinto, L. A., Galvao, R. K. H. and Araujo, M. C. U. (2011) 'Influence of wavelet transform settings on NIR and MIR spectrometric analyses of diesel, gasoline, corn and wheat', *Journal of the Brazilian Chemical Society,* 22, (1), pp. 179-186.

Pons, M. N., Le Bonte, S. and Potier, O. (2004) 'Spectral analysis and fingerprinting for biomedia characterisation', *Journal of Biotechnology,* 113, (1-3), pp. 211-230.

Projan, S. J., Gill, D., Lu, Z. J. and Herrmann, S. H. (2004) 'Small molecules for small minds? The case for biologic pharmaceuticals', *Expert Opinion on Biological Therapy,* 4, (8), pp. 1345-1350.

Rader, R. A. (2008) '(Re)defining biopharmaceutical', *Nature Biotechnology,* 26, (7), pp. 743-751.

Randolph, T. W. and Yasui, Y. (2006) 'Multiscale processing of mass spectrometry data', *Biometrics,* 62, pp. 589-597.

Ramaker, H. J., van Sprang, E. N. M., Westerhuis, J. A. and Smilde, A. K. (2003) 'Dynamic time warping of spectroscopic batch data', *Analytica Chimica Acta,* 498, (1-2), pp. 133-153.

Rathore, A. S., Bhushan, N. and Hadpe, S. (2011) 'Chemometrics applications in biotech processes: A review', *Biotechnology Progress,* 27, (2), pp. 307-315.

Rathore, A. S., Bhambure, R. and Ghare, V. (2010) 'Process analytical technology (PAT) for biopharmaceutical products', *Analytical and Bioanalytical Chemistry,* 398, (1), pp. 137-154.

Rhiel, M., Cohen, M. B., Murhammer, D. W. and Arnold, M. A. (2002) 'Nondestructive near-infrared spectroscopic measurement of multiple analytes in undiluted samples of serum-based cell culture media', *Biotechnology and Bioengineering,* 77, (1), pp. 73-82.

Riley, M. R., Okeson, C. D. and Frazier, B. L. (1999) 'Rapid calibration of near-infrared spectroscopic measurements of mammalian cell cultivations', *Biotechnology Progress,* 15, (6), pp. 1133-1141.

Rodrigues, L. O., Vieira, L., Cardoso, J. P. and Menezes, J. C. (2008) 'The use of NIR as a multi-parametric in situ monitoring technique in filamentous fermentation systems', *Talanta,* 75, (5), pp. 1356-1361.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A. and Jent, N. (2007) 'A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies', *Journal of Pharmaceutical and Biomedical Analysis,* 44, (3), pp. 683-700.

Roggo, Y., Roeseler, C. and Ulmschneider, A. (2004) 'Near infrared spectroscopy for qualitative comparison of pharmaceutical batches', *Journal of Pharmaceutical and Biomedical Analysis,* 36, (4), pp. 777-786.

Roychoudhury, P., O'Kennedy, R., McNeil, B. and Harvey, L. M. (2007) 'Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses', *Analytica Chimica Acta,* 590, (1), pp. 110-117.

Ryu, J. K., Kim, H. S. and Nam, D. H. (2012) 'Current status and perspectives of biopharmaceutical drugs', *Biotechnology and Bioprocess Engineering,* 17, (5), pp. 900-911.

Sarkar, T., Salazar-Palma, M. and Wicks, M. (2002) *Wavelet Applications in Engineering Electromagnetics.* Norwood, Massachusetts: Artech House Books.

Sato, T. (1994) 'Applications of principal component analysis on near-infrared spectroscopic data of vegetable-oils for their classification ', *Journal of the American Oil Chemists Society,* 71, (3), pp. 293-298.

Scarff, M., Arnold, S. A., Harvey, L. M. and McNeil, B. (2006) 'Near infrared spectroscopy for bioprocess monitoring and control: Current status and future trends', *Critical Reviews in Biotechnology,* 26, (1), pp. 17-39.

Schmidt, F. R. (2005) 'Optimization and scale up of industrial fermentation processes', *Applied Microbiology and Biotechnology,* 68, (4), pp. 425-435.

Shukla, A. A. and Thommes, J. (2010) 'Recent advances in large-scale production of monoclonal antibodies and related proteins', *Trends in Biotechnology,* 28, (5), pp. 253-261.

Simoglou, A., Martin, E. B. and Morris, A. J. (2000) 'Multivariate statistical process control of an industrial fluidised-bed reactor', *Control Engineering Practice,* 8, (8), pp. 893-909.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) 'XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification', *Analytical Chemistry,* 78, (3), pp. 779-787.

Steinmeyer, D. E. and McCormick, E. L. (2008) 'The art of antibody process development', *Drug Discovery Today,* 13, (13-14), pp. 613-618.

Strang, G. and Nguyen, T. (1997) '*Wavelet Filter Banks'.* Wellesley: Wellesley-Cambridge Press.

Tang, Y. Y., Liu, J. and Yang, L. H. (2000) *Wavelet Theory and Its Application to Pattern Recognition.* Singapore: World Scientific Publishing.

Tatavarti, A. S., Fahmy, R., Wu, H., Hussain, A. S., Marnane, W., Bensley, D., Hollenbeck, G. and Hoag, S. W. (2005) 'Assessment of NIR spectroscopy for nondestructive analysis of physical and chemical attributes of sulfamethazine bolus dosage forms', *AAPS PharmSciTech,* 6, (1), pp. E91-9.

Teixeira, A. P., Portugal, C. A. M., Carinhas, N., Dias, J. M. L., Crespo, J. P., Alves, P. M., Carrondo, M. J. T. and Oliveira, R. (2009) 'In Situ 2D fluorometry and chemometric monitoring of mammalian cell cultures', *Biotechnology and Bioengineering,* 102, (4), pp. 1098-1106.

Trim, P. J., Atkinson, S. J., Princivalle, A. P., Marshall, P. S., West, A. and Clench, M. R. (2008) 'Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis', *Rapid Communications in Mass Spectrometry,* 22, (10), pp. 1503-1509.

Trygg, J., Kettaneh-Wold, N. and Wallbacks, L. (2001) '2D wavelet analysis and compression of on-line industrial process data', *Journal of Chemometrics,* 15, (4), pp. 299-319.

U.S. Department of Health and Human Services Foods and Drug Administration (2004) *Guidance for Industry, PAT- A framework for innovative pharmaceutical developing, manufacturing and quality assurance, US* [Online]. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070305.pdf (Accessed: 12 December 2012)

U.S. Department of Health and Human Services Foods and Drug Administration (2006) *ICH Guideline for Industry: Q9 Quality Risk Management* [Online]. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM073511.pdf (Accessed: 12 December 2012)

U.S. Department of Health and Human Services Foods and Drug Administration (2009a) *ICH Guidance for Industry Q8(R2) Pharmaceutical Development* [Online]. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM073507.pdf (Accessed: 12 December 2012)

U.S. Department of Health and Human Services Foods and Drug Administration (2009b) *ICH Guidance for Industry Q10 Pharmaceutical Quality System* [Online]. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM073517.pdf (Accessed: 12 December 2012)

Vaidyanathan, S., Arnold, S. A., Matheson, L., Mohan, P., McNeil, B. and Harvey, L. M. (2001a) 'Assessment of near-infrared spectral information for rapid monitoring of bioprocess quality', *Biotechnology and Bioengineering,* 74, (5), pp. 376-388.

Vaidyanathan, S., Harvey, L. M. and McNeil, B. (2001b) 'Deconvolution of near-infrared spectral information for monitoring mycelial biomass and other key analytes in a submerged fungal bioprocess', *Analytica Chimica Acta,* 428, (1), pp. 41-59.

Vaidyanathan, S., Macaloney, G. and McNeil, B. (1999) 'Fundamental investigations on the near-infrared spectra of microbial biomass as applicable to bioprocess monitoring', *Analyst,* 124, (2), pp. 157-162.

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. and van der Werf, M. J. (2006) 'Centering, scaling, and transformations: improving the biological information content of metabolomics data', *Bmc Genomics,* 7, pp. 15.

Vetterli, M. and Herley, C. (1992) 'Wavelets and filter banks: Theory and design', *Ieee Transactions on Signal Processing,* 40, (9), pp. 2207-2232.

Walczak, B., vandenBogaert, B. and Massart, D. L. (1996) 'Application of wavelet packet transform in pattern recognition of near-IR data', *Analytical Chemistry,* 68, (10), pp. 1742-1747.

Walsh, G. (2010) 'Biopharmaceutical benchmarks 2010', *Nature Biotechnology,* 28, (9), pp. 917-924.

Wan, H. Z., Kaneshiro, S., Frenz, J. and Cacia, J. (2001) 'Rapid method for monitoring galactosylation levels during recombinant antibody production by electrospray mass spectrometry with selective-ion monitoring', *Journal of Chromatography A,* 913, (1-2), pp. 437-446.

Westerhuis, J. A., Gurden, S. P. and Smilde, A. K. (2000) 'Generalized contribution plots in multivariate statistical process monitoring', *Chemometrics and Intelligent Laboratory Systems,* 51, (1), pp. 95-114.

Westerhuis, J. A., Kourti, T. and MacGregor, J. F. (1999) 'Comparing alternative approaches for multivariate statistical analysis of batch process data', *Journal of Chemometrics,* 13, (3-4), pp. 397-413.

Wold, S., Esbensen, K. and Geladi, P. (1987) 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems,* 2, (1-3), pp. 37-52.

Wold, S., Kettaneh, N., Friden, H. and Holmberg, A. (1998) 'Modelling and diagnostics of batch processes and analogous kinetic experiments', *Chemometrics and Intelligent Laboratory Systems,* 44, (1-2), pp. 331-340.

Xia, J. M., Wu, X. J. and Yuan, Y. J. (2007) 'Integration of wavelet transform with PCA and ANN for metabolomics data-mining', *Metabolomics,* 3, pp. 531-537.

Zhang, X., Scalf, M., Berggren, T. W., Westphall, M. S. and Smith, L. M. (2006) 'Identification of mammalian cell lines using MALDI-TOF and LC-ESI-MS/MS mass spectrometry', *Journal of the American Society for Mass Spectrometry,* 17, (4), pp. 490-499.

Zou, X. B., Zhao, J. W., Povey, M. J. W., Holmes, M. and Mao, H. P. (2010) 'Variables selection methods in near-infrared spectroscopy', *Analytica Chimica Acta,* 667, (1-2), pp. 14-32.