THE UNIVERSITY OF HULL

Application of Chemometrics in Process Analysis

being a Thesis submitted for the Degree of

MASTER OF PHILOSOPHY

in the University of Hull

by

Richmond Jerry Ampiah-Bonney,

BSc (Hons), University of Science and Technology, Kumasi, Ghana,

MSc, University of Hull, Hull, UK,

September 2006

Abstract

The acid catalysed esterification of ethanol by acetic acid, a batch process, has been investigated on a laboratory scale at the high temperature range of 78 – 80°C. The data has been collected by Raman Spectroscopy and successfully de-noised using Principal Components Analysis. The first principal component (PC1) was found to describe the fluorescence and other sources of noise in the data and the reconstituted data due to the variation captured in the second principal component (PC2) contained the actual Raman spectra. Thus the reaction profile as well as the profiles of individual reaction components have been clearly mapped out. Validation of this de-noising technique has been done by calculating the kinetics of the reaction with the reconstituted data, which has been found to follow the theoretical first order reaction kinetics. The effect of variable selection procedures on model building has been investigated using data from a continuous industrial process, for which reaction profiling as was done for the batch system is not applicable. Two variable selection techniques, General Randomised PRESS-based Elimination (GRAPE) and the genetic algorithm (GA), improve the prediction ability of MLR models by a great deal, indicated by Root Mean Square Error of Cross-Validation (RMSECV) values of 1.0649 – 1.1277 and 1.0977 – 2.0064 respectively. Predicted concentrations are a good estimate for the actual concentrations.

Acknowledgements

Dear name, the rock on which I built!
My shield and hiding place
My never-failing treasury, filled
With boundless source of grace

I wish to acknowledge the immense help and patience I received from my supervisor Dr. A.D.Walmsley throughout the period of my study. I am also grateful to Professor Haswell for his useful contributions, to Professor Townshend and indeed to every member of the Analytical staff at the School of Chemistry for their assistance in all kinds of ways. I acknowledge

- Centre for Process Analytics and Control Technologies (CPACT), UK.
- Clairet Scientific, UK (for providing the Raman spectrometer and probe).
- BP Amoco Chemicals Limited, Hull, for permission to use Naphtha data.
- School of Chemistry and Scholarships Committee, The University of Hull for funding.
- Dr. Adrie Dane, CPACT, Hull, for permission to use the GRAPE algorithm.

My profound thanks go to my wife Olivia and my sons Kodwo and Joojo for their unflinching support and faith in me through thick and thin. To my parents Comfort Asiedu and George, I am forever indebted for establishing me on so sure a foundation in life.

I am thankful to The Reverend David Frudd, Miss Angela Gage, Miss Carrie-Ann Langley and Dr. Sannie Chong for their support, encouragement and prayers that saw me through *those dark days*. I am also grateful to members of Trinity Methodist Church for their selfless assistance and support throughout my stay in the United Kingdom.

But over and above all these, all thanks and credit are due to the Lord Jesus Christ by whose hand alone I have crossed these murky grounds and moved on to the safety of the rock. I render profound gratitude to the Almighty God whose

.....bountiful care, what tongue can recite?
It breaths in the air; it shines in the light.
It streams from the hills, it descends to the plain
And sweetly distils in the dew and the rain.

Kingston Upon Hull,
August 2000 AD.

TABLE OF CONTENTS

CONTENT	PAGE
Abstract	2
Acknowledgement	3
Table of Contents	4
Introduction	9
CHAPTER ONE: Literature Review	11
1.1 Process Analysis Methods	11
1.2 Batch Processes	11
1.3 Control of Processes	12
1.4 Chemometrics and Applications	15
1.4.1 Multivariate Calibration	17
1.4.2 Multiple Linear Regression (MLR)	18
1.4.3 Factor Analysis and Principal Components Analysis	19
1.4.3.1 Principal Components Analysis (PCA)	21
1.4.4 Partial Least Squares (PLS) Regression	23
1.5 On-line Measurement Methods	25
1.5.1 Near-Infrared Spectroscopy (NIR)	25
1.5.2 Raman Spectroscopy	26
1.6 Theory	28
1.6.1 Calibration Methods	28
1.6.1.1 Cross-Validation	29
1.6.1.2 Variable Selection	29
1.6.1.3 General Randomised PRESS-based Elimination (GRAPE)	29
1.6.2 Univariate Linear Regression	30
1.6.3 Multiple Linear Regression	31
1.6.4 Principal Components Analysis	32
1.6.5 Principal Components Regression	40
1.6.6 Partial Least Squares	40

CHAPTER TWO: Experimental 42
2.1 Experiment 1: The Esterification of ethanol and acetic acid 42
2.1.1 Kinetics 42
2.1.2 Reagents 44
2.1.3 Apparatus and Procedure 45
2.2 Experiment 2: The Distilled Feed Column Data 46
2.3 Software 49

CHAPTER THREE: Results and Discussion 51
3.1 Experiment 1: Raman Monitoring of the esterification reaction 51
3.1.1 Data Analysis 56
3.1.2 Kinetics 70
3.1.3 Conclusion 72
3.2 Use of Raman Spectroscopy for Process analysis – Experiment 2 73
3.2.1 PLS Results 86
3.2.2 GRAPE Results 87

CONCLUSION 94
Appendix 95
References 98

FIGURES

Figure 1 *The use of MSPC in batch reaction control.* 14
Figure 2 *The basis of factor analysis - Reduction of two variables to one factor which is a linear combination of both variables.* 20
Figure 3 *Basis of factor analysis – Reduction of three variables to a plane that defines the essential information in the three variables.* 21
Figure 4 *Scatter diagram of data from two variables, giving little information* 33
Figure 5 *Data plotted on a new axis that captures the greatest variation in the data. The various groupings are easily identified.* 34
Figure 6 *Data plotted on another axis perpendicular to the axis chosen in figure 5.* 35

Figure 7	*Data plotted on two new axes to reveal the underlying patterns*	**36**
Figure 8	*Experimental set up showing the configuration of the Raman probe and other accessories in the 1-litre vessel.*	**46**
Figure 9	*Schematic of the Naphtha processing system.*	**47**
Figure 10	*System for obtaining Raman spectra from the distillation column*	**48**
Figure 11	*Raman spectrum of pure acetic acid, showing the prominent peaks at 896, 622 and 447cm-1 respectively.*	**51**
Figure 12	*Raman spectrum of pure ethanol, showing the prominent peaks at 930, 1056 and 1101cm^{-1} respectively.*	**52**
Figure 13	*Raman spectrum of pure ethyl acetate, showing the prominent peaks at 895, 683 and 427cm^{-1} respectively.*	**52**
Figure 14	*A plot of the Raman spectra of real-time monitoring of the acid-catalysed esterification of ethanol and acetic acid over the entire reaction period*	**54**
Figure 15	*A three-dimensional plot of the Raman spectra of real-time monitoring of the acid-catalysed esterification of ethanol and acetic acid.*	**55**
Figure 16	*Regression of a product profile (683cm^{-1}) against a reactant profile (930cm^{-1}) from the initial data, showing the apparent different stages of the esterification process*	**57**
Figure 17	*Schematic diagram of expected profiles of products and reactants during the reaction*	**58**
Figure 18	*Profiles of ethyl acetate and ethanol during the esterification..*	**59**
Figure 19	*Plot of eigenvalue versus number of PC's for Raman data*	**60**
Figure 20	*Regenerated data based on variation in PC1*	**62**
Figure 21	*Sample number versus scores on PC 1.*	**63**
Figure 22	*Profile of wavenumber 100cm^{-1} from the mean-centred data.*	**64**
Figure 23	*PC1 results: Result of removing the regenerated Raman data obtained from the variation captured by PC1, from the mean-centred Raman data*	**65**
Figure 24	*Profiles of ethyl acetate and ethanol during the esterification, using the data remaining after removing PC1.*	**66**

Figure 25	*PC2 results: Regenerated Raman spectra based on variation captured in the second PC*	**67**
Figure 26	*Sample number versus Scores for PC2. This shows the progress of the esterification reaction proper.*	**68**
Figure 27	*Profiles of the peaks for ethanol, ethyl acetate and acetic acid during the esterification, using PC2.*	**69**
Figure 28	*Reaction plots for ethanol and ethyl acetate during scans 18 to 29*	**71**
Figure 29	*Kinetic plots for ethyl acetate and ethanol during scans 18 to 29*	**72**
Figure 30	*Raman data from naphtha processing plant*	**74**
Figure 31	*Three-dimensional colourmap of Raman data from the DF plant*	**75**
Figure 32	*Mean centred Raman data from naphtha processing plant*	**76**
Figure 33	*Plot of eigenvalue versus number of PC's for the Naphtha Raman data*	**77**
Figure 34	*Regenerated naphtha data based on the variation captured in PC1*	**79**
Figure 35	*Sample number versus scores on PC1*	**80**
Figure 36	*Sample number versus scores on PC1, showing groupings*	**81**
Figure 37	*Regenerated naphtha data based on the variation captured in PC2*	**83**
Figure 38	*Score plot for Naphtha data, reconstituted from the variation captured by PC2*	**84**
Figure 39	*Profiles of peaks in PC2 data*	**85**
Figure 40	*Box plots of the prediction results of GRAPE and variable selection genetic algorithms (VSGA) on the naphtha Raman data*	**90**
Figure 41a	*Measured (solid line) and predicted (plus signs) concentration of the first six components using GRAPE/MLR..*	**91**
Figure 41b	*Measured (solid line) and predicted (plus signs) concentration for components 7-12 using GRAPE/MLR*	**92**

Figure 41c *Measured (solid line) and predicted (plus signs) concentration for components 13-17 using GRAPE/MLR* **93**

TABLES

Table 1 *Results from PCA model* **61**

Table 2 *Results showing percent variance captured by PCA model of naphtha data* **78**

Table 3 *Results of PLS modelling of original and denoised Raman Naphtha data* **87**

Table 4 *Results of modelling without variable selection* **88**

Table 5 *Results of variable selection experiments* **88**

Introduction

The need for better process control of industrial batch processes is important, since batch processes are finding increasing applications in the chemical industry, pharmaceuticals, bio-technical and in the brewery industry among many others[1]. Chemical manufacture by batch production forms a major and highly profitable sector of the chemicals industry. With the present trend towards batch production of high value added products like pharmaceuticals and fine chemicals, this sector is rapidly expanding. The cost of development of a fine chemical process is high. A significant part of this cost is in the laboratory experiments and extensive pilot plant trials required for scale up to full production. Strict safety rules also require the manufacturers to carry out more detailed safety studies.

The advantage of batch processes is their ability to produce high-value products within short manufacturing times. Moreover, the procedures in batch processes are relatively very simple; basically the reactants are loaded into the reaction vessel, processed under controlled conditions and then the completed product is discharged[2].

The variation in products from batch to batch needs to be minimised as much as possible. This brings in the need for better process control. Good process control design is a creative, dynamic, and iterative process. It demands an understanding of the big picture, the minute details, and the skills to balance them. Research that aims to advance optimisation and control with industrial applications are very relevant. Such work will involve non-linear process control techniques, process modelling and simulation and on-line real-time computer application control in order to achieve process automation.

Generally, laboratory analytical instrumentation is very accurate and versatile as well as expensive and delicate. Thus to be applied industrially, the instrument must be ruggedised or protected by some other means from the harsh operating conditions. The use of optical analytical devices allows the extension of the analytical instrument to various measurement points because of the use of low-loss optical fibres. This helps reduce cost. Another advantage is that the expensive and delicate instrumentation can be located in the relatively benign environment of a control room, with only the optical fibres and relatively simpler sampling equipment at the measurement point.

Many methods are used in process control, ranging from basic *knob-twisting* methods for controlling reaction conditions and feed rates, to advanced programmable

controllers. Batch reactions are complex in nature due to their finite duration, inconsistency in homogeneity and multicomponent nature. Even with extensive automation, control of a batch process is very challenging. Industry is rapidly moving away from the situation of doing post-production investigation due to its wastefulness both of resources and time[3]. One solution has been the on-line sampling and analysis of various stages of the process at regular time intervals. Besides the expenditure involved in equipment acquisition and set-up, it offers information only for correcting the fault as and where it is found.

A more pro-active approach is to create a model that is capable of predicting chemical changes within the reaction vessel, with a view to controlling the process from a remote position. The initial stage of this work is the continuous monitoring of the process in order to acquire information on all stages of the process and also to be able to follow all chemical processes going on. This calls for the use of an in-situ probe that conveys data to a spectrometer. Such data, when analysed with chemometric techniques, offers information vital to the modelling step.

As a starting point, we have elected to perform real-time monitoring of a relatively simple and well-known batch industrial process – the reaction of ethanol and acetic acid to produce ethyl acetate and water. Our aim is to establish the use of a spectroscopic probe placed in-situ to register all changes occurring in the reaction mixture in real time and to present data that can be use to trace and explain all the chemical and physical changes occurring. This is a prelude to process control. The kinetic (theoretical) characteristics of the process are compared with the data obtained experimentally to ascertain authenticity.

CHAPTER 1: LITERATURE REVIEW

1.1 Process Analysis Methods

Process analysis comprises continuous and discontinuous procedures for establishing the nature and properties of a process. Often used in connection with chemical processes in the chemical industry, the term "process analysis" is part of industrial applied analysis. In recent times on-line and in-line techniques have become more technically and economically important, making use of automation. Process analysis has been characterised by the use of sophisticated and expensive instruments for multicomponent analysis. However, with the increasing availability of spectroscopic methods of analysis, these have been substituted by relatively less expensive and simpler instruments[4]. Examples are the use of optical emission spectroscopy, x-ray fluorescence spectroscopy in industry[5,6]. Alongside these spectroscopic methods, classical methods like titrimetry, gravimetry, colorimetry and electrochemical methods as well as separation methods have always been used for reference.

Statistical process control is achieved when certain process variables are maintained close to their expected values[7]. In order to ensure that a process proceeds within the set values of these variables, the process must be monitored. This reduces variability, increases yield, and decreases (hazardous) waste and cost. Many companies in the pharmaceutical industry are developing methods for real-time process analysis. Real-time process analysis drastically reduces costly hold-up time where the analysis time may be longer than the processing time or where a particular batch may be held up waiting in an intermediate processing stage waiting for chemical analysis[3].

1.2 Batch Processes

Many chemical, pharmaceutical, biochemical, and other manufacturing processes are batch in nature. Process product quality variables are measured after the end of each batch, making it difficult to monitor the progress of the batch process or to control the product quality[8]. There are huge archives of routinely collected data on temperature, pressure, flow rates and other such process variables collected by on-line process computers within the duration of each batch, and these have been a rich source of batch process data for building and testing various multivariate techniques for batch

process analysis[4]. Batch processes can be found in a wide variety of industries--everything from hydrocarbon and chemical processing to food, pharmaceutical, and consumer goods manufacturing.

1.3 Control of processes.

Until recently, all control on the plant was manual. Valves were moved by hand, heat was turned on and off by flipping a switch, and levels were determined by watching gauges. These were later on replaced by automatic control, initially electrical and pneumatic forms and later on in much more sophisticated electronic configurations. The addition of computers and centralised digital control systems, then microprocessor technology, has led to the distributed control systems (DCSs) so commonly used today.

Pressure and level are two of the most fundamental measurements in process control. People have been measuring both for a long time, but the technologies involved are by no means static. New developments keep surfacing that help engineers make these measurements more accurately and with greater ease and flexibility.

Worldwide, flow measurement is the largest segment of the industrial measurement market, and that segment is growing rapidly, about twice as fast as the overall market. Flow is the measurement of confined fluid streams, liquid, gas, or vapour, due to either to gravity or to pressure produced by pumps and compressors. Reasons for measuring flow can be divided into four categories: control, indication, monitoring or totalizing, and custody transfer.

Temperature was one of the first variables to be measured in the process field and has been determined in many different ways. Virtually any physical property that changes with temperature has, at one time or another, been used as a basis for this measurement. Among the many methods still used in industry today are thermocouples (T/Cs), resistance temperature devices (RTDs), thermistors, electronic temperature sensors, bimetallic devices, filled devices, infrared (IR) devices, and acoustic pyrometry. Programmable controllers of today provide versatile control capabilities that make them suitable for a variety of continuous, batch, and discrete applications.

Important as these measurements are, they provide very little information about the chemical composition of a process. For this type of determination, analytical

measurements are required. Statistical process control enables to keep an industrial process under selected and controlled conditions. The traditional analysers and statistical tools are useful generally in a univariate way. In a multivariate approach, as required for complex batch processes, they allow for monitoring but do not explain and correct eventual detected changes. Multivariate statistical process control (MSPC) tries to detect variation due to special causes. One can define two types of variation in a process: common cause, variation that can be expected if the process is running under normal operating conditions (NOC); and special causes, variation that moves the process out of NOC. In continuous processes, multivariate control charts can be developed to control the process, which are comparable with univariate charts. MSPC for batch processes is inherently more difficult[9]. Batch processes that are very common in the chemical industry exhibit large variations in their operation. For batch processes the target values for the different process measurements and output variables are not clear and constant like in continuous processes. They depend on time (or conversion factor) and can be described as ideal temporal trajectories. MSPC can then be used to monitor the deviation from this ideal trajectory for a running batch. For batch MSPC the data can be arranged in a three-way array, and analysed with the batches considered as objects.

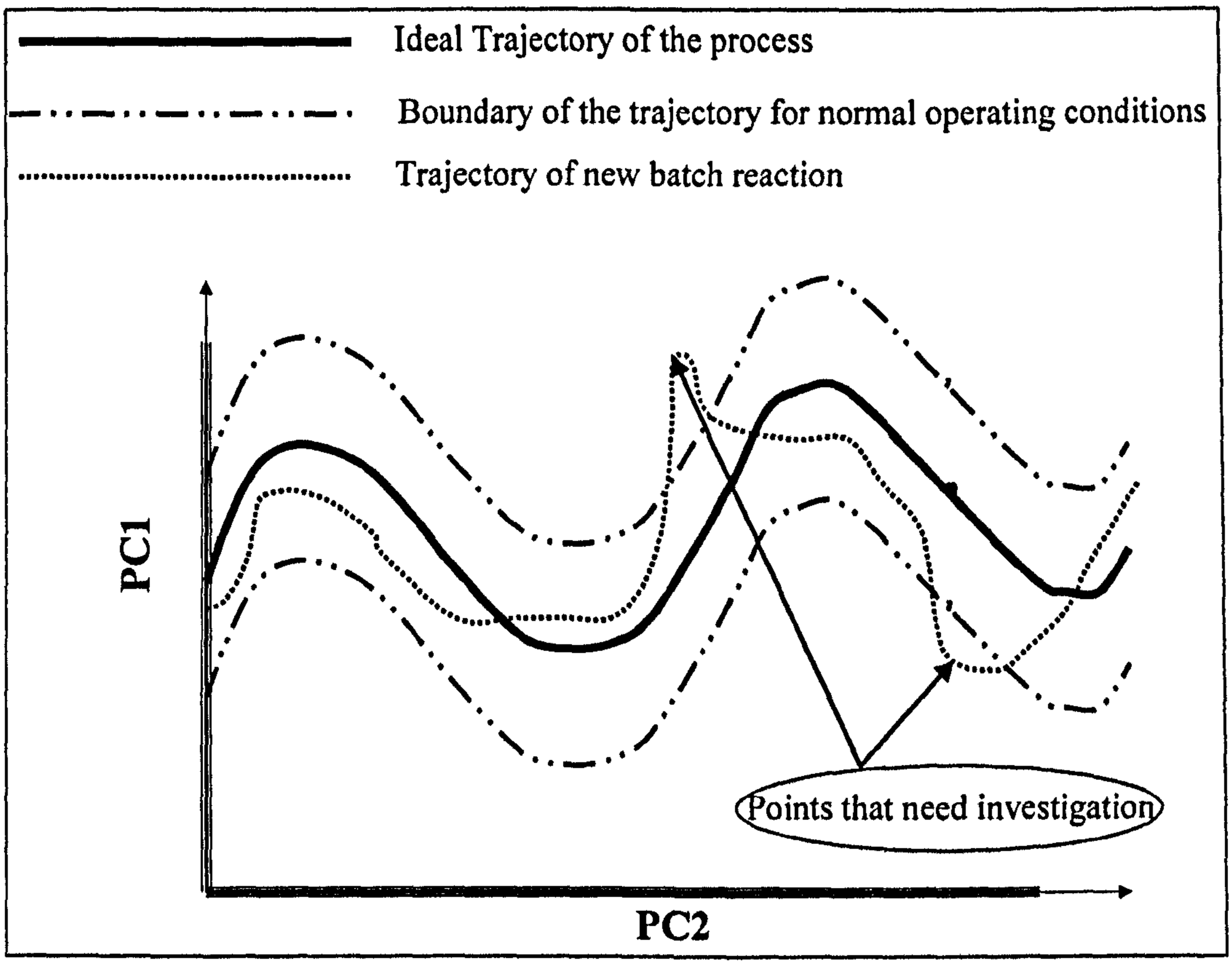

Figure 1: Illustrating the use of MSPC in batch reaction control. The normal operating conditions (NOC) are established within limits of the ideal trajectory, and the PC1-PC2 plot of a batch reaction is plotted. Points that fall outside the NOC require further investigation.

When a new batch starts, the available measurements are used to calculate the position of that batch in the PC1-PC2 plot. If during the run a batch falls outside the area of NOC, then action has to be undertaken.

The objectives of process control are:

- Stabilise the process
- Provide consistent operation from shift to shift
- Increase product quality
- Increase product yield
- Decrease variation in all process parameters
- Co-ordinate production scheduling
- Decrease energy use
- Compliance with environmental regulations.

There are a number of technical issues associated with chemical process control:

(1) most chemical processes are very non-linear,

(2) time lags and delays due to flow and heat conduction are prevalent,

(3) hysteresis is common,

(4) system drift is common,

(5) plants are often poorly instrumented, and

(6) first principles models are often unavailable or unattainable.

Fortunately, compared with some control problems, time scales are very long and computation time is not an issue. Many of these issues can be addressed by using control schemes which incorporate Non-linear Adaptive Computation (Neural Networks). First principle models are not needed if a network can adaptively capture the system performance as the plant operates. If the network is trained on-line, then plant drift can be tracked. The networks themselves are non-linear and usually have no trouble capturing the non-linearities in the processes. Non-linear accurate system models permit the system to be controlled into and out of unstable hysteresis regimes. Accurate non-linear models also allow accurate prediction farther into the future, thus permitting control in the presence of significant time lags and delays.[10] Modern DCSs are designed to gather, store, manipulate, and display process information in order to improve process control and achieve greater product consistency, quality, and output. They also can validate the process, aid in records management, help assure worker safety, and provide data needed to comply with environmental regulations.

1.4 Chemometrics and applications

Many chemical processes today are characterised by rich data that gives little direct information about the process. This data bank is seen as a *gold mine* of information provided that the relevant and important information is extracted effectively and quickly enough to be of use in quality and safety improvement, waste reduction and increased yield and consequently profits. Such a data extraction method should be able to overcome problems like undetected sensor failures, uncalibrated and misplaced sensors, lack of integrity of the data historian, and general human errors. With the increase in the effectiveness and application of spectroscopic techniques, large amounts of (complex) data can now be acquired in an impressively short acquisition time. One data analysis technique that has been applied successfully to

reduce large amounts of spectroscopic data into meaningful information is Chemometrics.

Chemometrics is the discipline concerned with the application of statistical and mathematical methods to chemical data[11,12]. A data collection task typically involves many measurements made on many samples. Such multivariate data has traditionally been analysed using one or two variables at a time. To determine the relationships among all samples and variables efficiently, we must process all of the data simultaneously. Chemometrics is the field of extracting information from multivariate chemical data using tools of statistics and mathematics, and can be typically used for one or more of three primary purposes:

- To explore patterns of association in data;
- To track properties of materials on a continuous basis; and
- To prepare and use multivariate classification models.

The algorithms in primary use in the field have demonstrated a significant capacity for analysing and modelling a wide assortment of data types for an even more diverse set of applications. A variety of powerful methods have been applied to the "supervised" analysis of multivariate data. In these methods, of which multiple linear regression (MLR), partial least squares regression (PLS) and principal components regression (PCR) are the most widely used, one seeks to relate the multivariate spectral inputs to the concentrations of target determinands, i.e. to generate a quantitative analysis, essentially via suitable types of multidimensional curve fitting or regression analysis [13,14]. Although non-linear versions of these techniques are increasingly available, [15,16] the usual implementations of these methods are linear in scope.

Patterns of association exist in many data sets, but the relationships between samples can be difficult to discover when the data matrix exceeds three or more features. Exploratory data analysis can reveal hidden patterns in complex data by reducing the information to a more comprehensible form. Such a chemometric analysis can expose possible outliers and indicate whether there are patterns or trends in the data. Exploratory algorithms such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) are designed to reduce large complex data sets into a series of optimised and interpretable views. These views emphasise the natural groupings in the data and show which variables most strongly influence those patterns.

The costs of making experiments are rapidly increasing, at the same time as the costs for making additional measurements on an ongoing experiment are decreasing due to the availability of electronic instrumentation such as spectrometers, chromatographs, etc. Hence, there is a tendency to make fewer and fewer experiments, but measure more and more data in each of them.

In many applications, it is expensive, time consuming or difficult to measure a property of interest directly. Such cases require the analyst to predict something of interest based on related properties that are easier to measure. One of the goals of chemometric analysis is to develop a calibration model, which correlates the information in the set of known measurements to the desired property. Chemometric algorithms for performing regression include PLS and PCR and are designed to avoid problems associated with noise and correlations in the data. Because the regression algorithms used are based in factor analysis, the entire group of known measurements is considered simultaneously, and information about correlations among the variables is automatically built into the calibration model. Chemometric regression lends itself handily to the on-line monitoring and process control industry, where fast and inexpensive systems are needed to test, predict and make decisions about product quality. Chemometrics methods are needed in order to extract useful, specific and selective information from these data. Standard chemometric methods have been found extremely useful in industry[17]. Methods such as multivariate calibration and other multivariate methods are being used increasingly in applications like the monitoring of beer production, the quality control of pharmaceutical formulations, cosmetic, and pulp and paper production as well as more recent applications in batch processes like biotechnical fermentation processes and wafer production in the semiconductor industry [18].

1.4.1 Multivariate Calibration

Calibration allows the user to relate instrumental measurements to the sample of interest. Multivariate calibration allows for the analysis of several measurements from several samples or specimens. This compares to univariate calibration, which involves the use of a single instrumental measurement to determine a single analyte. Either method may contribute to the two-step procedure where

1) Data is calibrated and

2) Predictions based on the calibration are made.

In calibration, indirect measurements are made from samples where the amount of the analyte has been pre-determined, usually by an independent assay or technique. These measurements, along with the pre-determined analyte levels, comprise a group known as the calibration set. This set is used to develop a model that relates the amount of sample to the measurements by the instrument. In some cases, the construction of the model is simple due to a certain relationship, such as Beer's Law in the application of UV and NIR spectroscopy. Other cases can be much more complex and, in these cases, construction of the model is the time-consuming step. Once the model is constructed, it can predict analyte levels based on measurements of new samples.

Another advantage of multivariate calibration is that it can be used to separate samples from interferences without the need of highly selective measurements for the analyte. In the case of HPLC, certain overlapping or anomalous peaks can be systematically separated or deleted from the data set based on certain linear combinations of measurements derived from one of several multivariate calibration techniques.

The multivariate calibration set contains multiple measurements from multiple sources of samples and pre-determined analyte amounts. The second stage is the prediction step for new sample levels, and this uses a model that provides the basis for the evaluation of a linear combination of the measurements. Calibration techniques (used in the calibration step) differ in determining coefficient values for the preceding equation (or a similar equation). Three of these methods to be discussed are multiple linear regression, principal components regression, and partial least squares.

1.4.2 Multiple linear regression (MLR)

Models constructed from spectroscopy are relatively simple due to linear combinations of the instrumental measurements, which makes the model correlation-based. Models for a broader range of conditions (i.e., measurements from several wavelengths) have been constructed in order to separate overlapping peaks elicited from the analyte plus other unknown components or conditions. These methods are based upon the following equation:

$$x_i = b_0 + b_1 * y_{i1} + b_2 * y_{i2} + \ldots + b_q * y_{iq} + e_i \qquad 1$$

where x_i is the analyte level of the ith specimen, y_{ij} is the jth instrumental measurement with the ith specimen, b represents the model parameters, and e_i is the error associated with y_i. From this equation, the analyte levels of new specimens can be predicted when the estimated b_j is substituted for a_j.

MLR does not require knowledge of the amount of interference or other samples in the calibration set. However the maximum amount of measurements (q) used in this method is restricted to approximately 2 to 10 in most cases. Therefore selecting an appropriate set of instrumental measurements is paramount. For example, in critical care environments, use of Visible-NIR spectroscopy non-invasively monitors oxygen saturation in arterial blood[19]. One wavelength in the visible spectrum monitors blood pulsatile volume and oxygen saturation; the other in the NIR measures pulsatile blood volume only. Since the interference in this case is the blood volume, the information from the two wavelengths can be used to filter out the interference and provide a measurement of the oxygen content.

1.4.3 Factor Analysis and Principal Components Analysis (PCA)

Factor analysis, of which the most frequently used variety is called PCA, is a mathematical technique performed on a set of variables to find its underlying dimensions or factors.

The main applications of factor analytic techniques are:

1. To reduce the number of variables and
2. To detect structure in the relationships between variables, that is to classify variables.

Therefore, factor analysis is applied as a data reduction or structure detection method. Factor analysis builds a model from data. The technique finds underlying factors, also called "latent variables" and provides models for these factors based on variables in the data. This technique can be very helpful in finding important underlying characteristics which might not themselves be observed, but which might be found as manifestations of variables which are observed. Factor analysis is also used for the combination of two variables into a single factor. The correlation between two variables is summarised in a scatter plot. A regression line can then be fitted that represents the "best" summary of the linear relationship between the variables. If a variable is defined that would approximate the regression line in such a plot, then that

variable would capture most of the "essence" of the two items. Subjects' single scores on that new factor, represented by the regression line, could then be used in future data analyses to represent that essence of the two items.

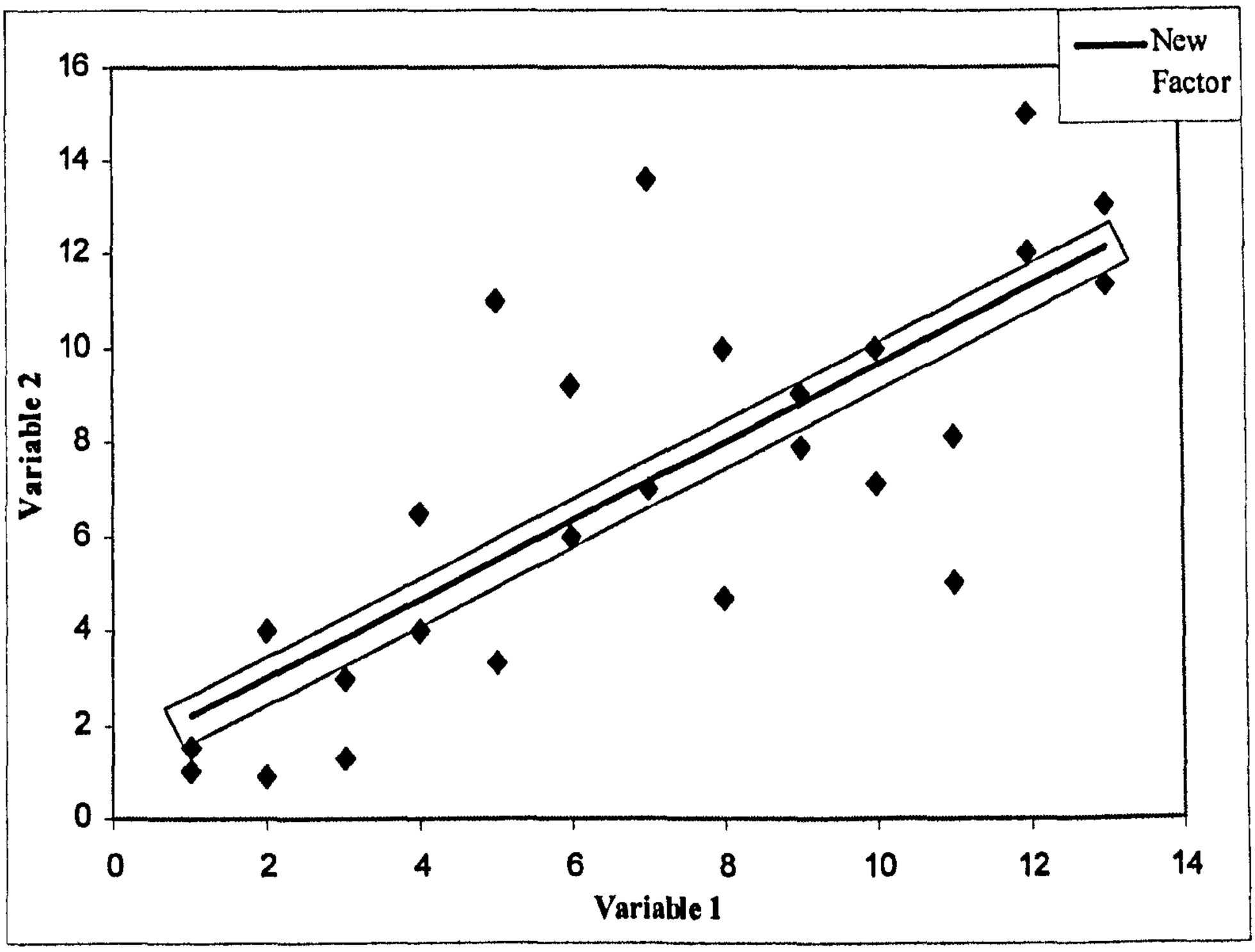

Figure 2: Illustrating the basis of factor analysis - Reduction of two variables to one factor which is a linear combination of both variables.

In a sense the two variables have been reduced to one factor which is actually a linear combination of the two variables. This phenomenon of combining two correlated variables into one factor illustrates the basic idea of factor analysis. If the two-variable example is extended to multiple variables, then the computations become more involved, but the basic principle of expressing two or more variables by a single factor remains the same. When there are more than two variables, they define a "space" just as two variables defined a plane. Thus, for three variables, a three-dimensional scatterplot can be plotted, and again a plane can be fitted through the data. With more than three variables it

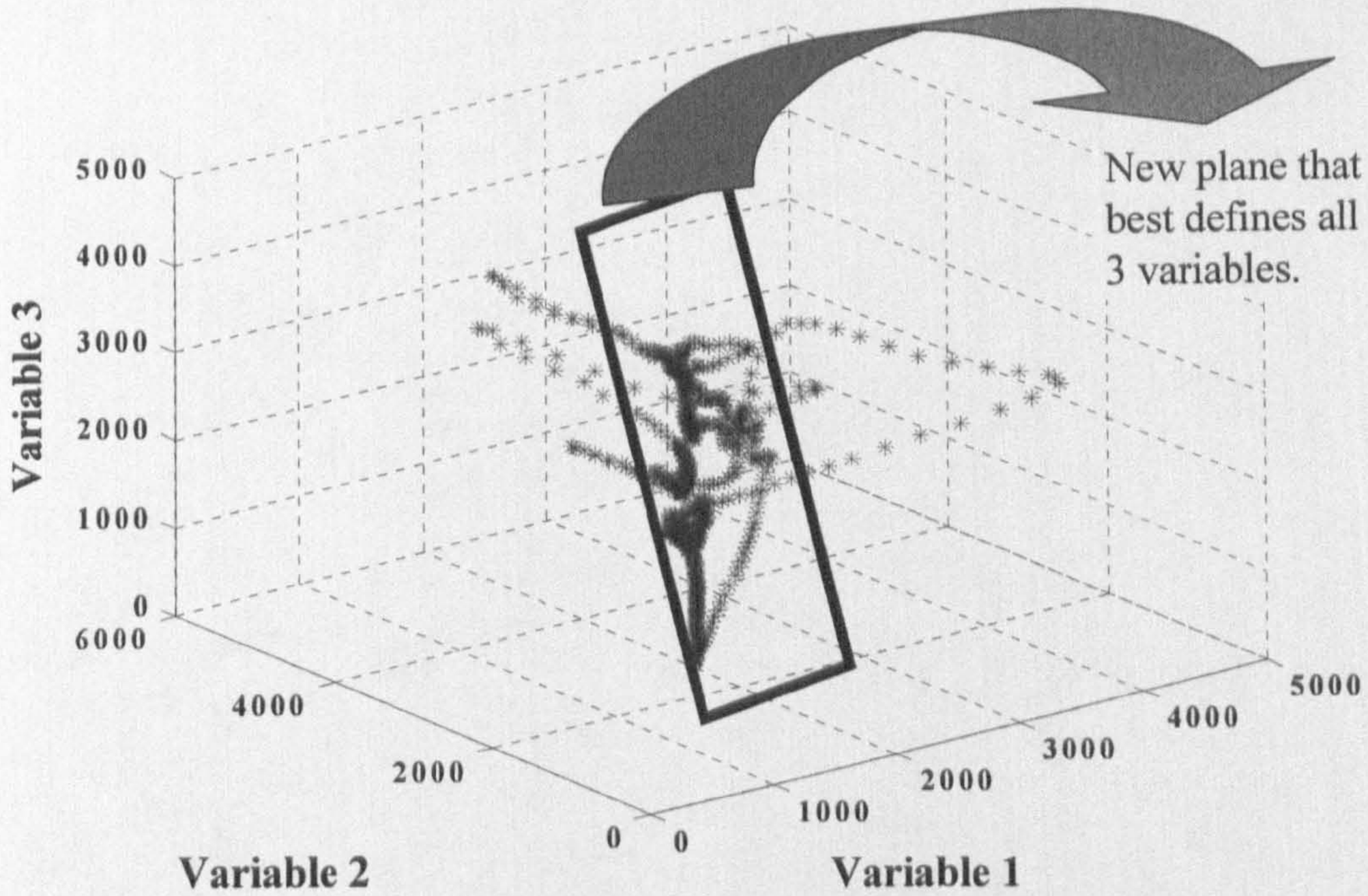

Figure 3: Basis of factor analysis – Reduction of three variables to a plane that defines the essential information in the three variables.

becomes impossible to illustrate the points in a scatterplot. However, the logic of rotating the axes so as to maximise the variance of the new factor remains the same.

1.4.3.1: Principal Components Analysis.

Modern laboratory instruments and measurement instruments in the chemical process industry are capable of acquiring an over-abundance of data, most of which is under-utilised or not utilised at all, as the information they give is not immediately useful. There is a great deal of correlation and redundancy in these measurements. Rather than being discarded, a better approach is to compress such data so that the essential information is retained and is more easily displayed than each of the original variables does. Since the essential information lies in how the variables change with respect to each other and not in any one single variable, the essential information must be extracted from the data, after employing some method of filtering or signal averaging to remove the large amount of noise that inevitably accompanies such large data.

PCA is a well-known technique of multivariate analysis that is useful in visualising and analysing large data sets. First proposed in 1901[20], it was not widely used until the arrival of modern computing technology. The main goal of PCA is to reduce the size of a data set which has a large number of intercorrelated variables and retain as much of the information present in the original data as possible. Suppose there are samples located in an environmental space or in species space for which all environmental variables or all species cannot be simultaneously envisioned. Then there would be the need for ordination methods. However, with more than three dimensions, these methods will not suffice. What PCA does is that it takes the cloud of data points and rotates it such that the maximum variability is visible, i.e. it identifies the most important gradients.

When it is used for modelling, PCA is further applied to determine the minimum dimensionality needed to reproduce the original information within experimental measurement error[21], i.e. PCA reduces the dimensionality of the problem in order to examine the important trends underlying the multivariate system.

Gurden *et al* [22] used PCA to successfully map out the process trajectory of an industrial reaction, along with other *statistical indicators* for the detection and diagnosis of process disturbances, and thus followed the operation of an industrial pilot plant. One major analytical tool both in industry and in research laboratories is HPLC, and like all chromatographic techniques, the polarity of the stationary phase as compared to the mobile phase determines the effectiveness of the method. Thus there are various polarity indicators, and Heberger conducted a comparison and evaluation of 8 of these and 30 stationary phases using PCA[23]. It was found that three principal components accounted for 99% of the total variance in the data, indicating that no single polarity variable is applicable alone. Further, a loadings versus scores plot showed significant groupings of the polarity indicators and stationary phases. An example of an environmental application of PCA is the work by Astorga-Espana *et. al.*[24], determining the levels of trace metals and some cations in a fish species native to the Canary Islands. PCA discriminated between the major cations and the trace metals and also between fish samples belonging to different seasons.

1.4.4 Partial Least Squares (PLS) regression

PLS is a multivariate calibration method that establishes a relationship between a set of independent variables **X** e.g. spectra, and dependent variables **Y** e.g. concentration given by

$$Y = X \times b + e \qquad 2$$

where b is the vector of PLS regression coefficients and e is the vector of errors that cannot be explained by the model. PLS is a full spectrum method, i.e. all the information contained in the spectra is available for the modelling[25]. In PLS, the original X variables are projected unto a reduced data space defined by new variables called PLS factors. The projection matrix is iteratively calculated from X and the Y variables such that the covariance between them is maximised among all factors.

PLS is a quantitative spectral decomposition technique that is closely related to PCR. However, in PLS, the decomposition is performed in a slightly different manner. Instead of first decomposing the spectral matrix into a set of eigenvectors and scores, and regressing them against the concentrations as a separate step, PLS actually uses the concentration information during the decomposition process. This causes spectra containing higher constituent concentrations to be weighted more heavily than those with low concentrations. Thus, the eigenvectors and scores calculated using PLS are quite different from those of PCR. The main idea of PLS is to get as much concentration information as possible into the first few loading vectors.

Actually, PLS simply takes advantage of the correlation relationship that already exists between the spectral data and the constituent concentrations. Since the spectral data can be decomposed into its most common variations, so can the concentration data. This generates two sets of vectors and two sets of corresponding scores; one set for the spectral data, and the other for the constituent concentrations. The two sets of scores are related to each other through some type of regression, and a calibration model is constructed.

As both the spectral and concentration data are decomposed simultaneously, and the scores are "exchanged" as each new factor is added to the model, PLS is a superior method.

One of the main advantages of PLS is that the resulting spectral vectors are related to the constituents of interest. This is an improvement upon PCR, where the vectors merely represent the most common spectral variations in the data, completely ignoring their relation to the constituents of interest until the final regression step.

There are two versions of the PLS algorithm; PLS-1 and PLS-2. The differences between these methods are subtle but have very important effects on the results. Like the PCR method, PLS-2 calibrates for all constituents simultaneously. The results of the spectral decomposition for both of these techniques give one set of scores and one set of eigenvectors for calibration. Therefore, the calculated vectors are not optimised for each individual constituent. This may sacrifice some accuracy in the predictions of the constituent concentrations, especially for complex sample mixtures. In PLS-1, a separate set of scores and loading vectors is calculated for each constituent of interest. In this case, the separate sets of eigenvectors and scores are specifically tuned for each constituent, and therefore, should give more accurate predictions than PCR or PLS-2.

The minor disadvantage in using the PLS-1 technique is about the speed of calculation. Since a separate set of eigenvectors and scores must be generated for every constituent of interest, the calculations take more time. For training sets with a large number of samples and constituents the increased time of calculation can be significant, but with modern computers, this is not a real problem. PLS-1 has the largest advantage when analysing systems that have widely varied constituent concentrations.

The advantages of PLS are:

- Single step decomposition and regression; eigenvectors are directly related to constituents of interest rather than largest common spectral variations.
- Calibrations are generally more robust provided that calibration set accurately reflects range of variability expected in unknown samples.
- Can be used for very complex mixtures since only knowledge of constituents of interest is required.
- Can sometimes be used to predict samples with constituents (contaminants) not present in the original calibration mixtures.

While all of these techniques have been successfully applied for spectral quantitative analysis, the arguments in the literature generally show that PLS has superior predictive ability. In most cases, PLS methods gives better results than PCR, and PLS-1 is more accurate than PLS-2. Unfortunately, there are no definite rules, and only good research practices can determine the best model for each individual system.

Disadvantages of PLS are:

- Calculations are slower that most Classical methods, especially PLS-1.
- Models are more abstract, thus more difficult to understand and interpret.
- Generally, a large number of samples are required for accurate calibration.
- Collecting calibration samples can be difficult; must avoid collinear constituent concentrations.

1.5 On-line Measurement methods

Because of their rapid response times and the ease with which they perform simultaneous multicomponent determinations, spectroscopic techniques are very useful in process analysis[26]. Among the spectroscopic techniques, vibrational spectroscopy is the most widely reported in process analysis[27].

1.5.1 Near Infrared spectroscopy (NIR)

NIR is more useful for quantitative analysis than for identification purposes. Some applications include the determination of water in a variety of samples, the quantitation of phenols, alcohols, organic acids and hydroperoxides, and the determination of esters, ketones and carboxylic acids. Thus it is a most suitable measurement technique for on-line real-time monitoring of the esterification process in this research. NIR has had many applications in process analysis, where it performs better than many alternative approaches. Recent advances in instrumentation and multivariate calibration have increased the utility and performance of NIR spectroscopy[28]. In comparison with infrared spectroscopy (IR), the NIR technology enables the direct in-line analysis of the reaction mixture by remote placing of the spectrometer using optical fibres[29]. NIR has been successfully applied in fermentation[30], polyesterification[31] and purity analysis[32] as well as for kinetic modeling[13]. In on-line applications, NIR has been used in tobacco processing, paper converting, textile manufacturing, cereal grains processing, oil in snack foods, protein in grains and weights and thickness of coatings[33]. Most NIR methods are temperature sensitive, particularly those involving water-based systems[34].

1.5.2 Raman spectroscopy

When electromagnetic radiation irradiates a molecule, the energy may be transmitted, absorbed or scattered. Of the scattered radiation, the strongest component is made up of Rayleigh scattering – an elastic collision between the incident photon and the molecule. Where the collision is inelastic, the energy of the molecule changes by an amount ΔE_m, characteristic of the molecule. This is called the Raman effect. The change in the energy of the molecule is equal to the difference in energy between the incident photon and the scattered photon[35].

Raman light sources are usually in the form of a laser, though the mercury arc was the initial source. The laser is almost ideal as a Raman spectroscopy light source because it has nearly complete linear polarisation, the (intense) beam can be focused on small sample volumes, and it is available in various wavelengths.

The use of Raman spectroscopy in industrial process control is relatively recent and rapidly increasing[36,37,38,39.] This is because Raman is more advantageous in many respects than many other conventional methods. Peaks in Raman spectra are abundant, well resolved and provide direct and clear chemical information since they correspond to fundamental transitions[40]. This is an advantage over the more commonly used NIR spectroscopy. Due to the selective nature of peaks in a Raman spectrum, it is also possible to use a few peak heights or peak areas to follow the progress of a reaction with time. Changes in laser intensity can be corrected by using an internal standard. Moreover, the light scattering nature of the Raman process allows simple and effective and stable fibre probe designs as compared to those needed for near- and mid-IR absorbance spectroscopy[41]. This allows Raman spectroscopy to be used for measuring several spectra simultaneously, and also allows remote monitoring of processes in hostile environments. Raman spectra yield more information about certain types of organic compounds than IR spectra, and studies yield useful information about olefinic functional groups and cycloparaffin derivatives that may not be disclosed by IR spectra. Because Raman spectra are less cluttered with peaks than IR spectra, peak overlap is less likely and quantitative measurements are simpler. Moreover, Raman instruments are not subject to attack by moisture, and small amounts of water in the sample cause no interference. Because the laser beams can be focused with precision, it is possible to quantify very small samples.

In their assessment of the application of Raman spectroscopy for on-line real-time multi-point analysis, Roberts *et al.*[42] confirmed the suitability of Raman spectroscopy for analytical applications and even for aqueous solutions because of the extremely weak response to water.

Recently, Raman spectroscopy has emerged as an important tool for the investigation of polymers. Raman spectra reveal information about the chemical nature, steric order, conformational order and orientation of the polymers[43].

In principle, since the intensity of the laser (source radiation) and the quantity of the scattering material present determine the intensity of a Raman band (assuming instrumental factors are invariant), Raman spectroscopy can be used quantitatively. However, reproducibility is almost impossible to control owing to the optical properties of the sample and its position with respect to the collection of optics of the input and focused laser beam. Consequently, most of the quantitative routines that have been used have been based on internal standards[44].

The drawback with this technique is that in addition to the Raman spectra, fluorescence spectra also arise, and this causes interferences. Raman and fluorescence spectroscopies are similar in their dependence on source intensity and signal loss. However, whereas Raman spectra always have multiple regions of uncorrelated spectral data, the same is not true for fluorescence spectra[12]. The presence of fluorescent reaction components and the occurrence of extra reflecting surfaces like bubbles enhance the occurrence of fluorescence spectra in Raman spectra. In fact, the fluorescence contribution can be so much as to render Raman data interpretation extremely difficult. Fluorescence can be minimised by a careful choice of the wavelength of the laser source for the Raman spectrometer, since the Raman scattering intensity decreases with laser wavelength whereas fluorescence gets less troublesome at longer laser wavelengths[7].

1.6: Theory

1.6.1 Calibration Methods

Multivariate calibration methods are employed here in order to develop a quantitative model to predict all the properties of interest simultaneously. Such a model will be described by

$$\varphi = g(X;\Theta), \qquad 9$$

where φ is the set of predicted variables ($y_1, y_2, y_3, \ldots, y_{Ny}$) from X, the set of predictor (independent) variables ($x_1, x_2, \ldots, x_{Nx}$)[45], g is a multivariate function and Θ is a vector or matrix of model parameters. Ideally, the difference between the actual y and the predicted φ should be minimal. For a linear model, $\varphi = \mathbf{XB}$, where **B** can be determined via latent variables as in principal components regression (PCR) and PLS[46]. Here, the predictors are compressed onto orthogonal factors or latent variables of which only those capturing the most variance are used to construct **B**. The number of factors to use can be determined by cross validation. The use of latent variables removes the problem of collinearity and reduces the influence of noise and experimental errors, thereby increasing the predictive ability of the calibration model built.

To cater for non-linearity, two non-linear models are provided. The polynomial PLS (n-PLS)[47] which is PLS with a polynomial inner relation, and then a multi-layer-feed-forward artificial neural network (MLF)[48]. For modelling non-linear relationships, a two layer MLF consisting of a layer with N_{hidden} non-linear units and a layer of N_y linear output units is most often applied. When bias terms are ignored (for notational convenience), the model is given by

$$\varphi = \Psi(\mathbf{W}_1\mathbf{X})\mathbf{W}_2, \qquad 10$$

where $\mathbf{W}_1$ (size N_x x N_{hidden}) and $\mathbf{W}_2$ (size N_{hidden} x N_y) are the model parameters or weights and $\Psi(x)$ is a non-linear transfer function, here the tangents sigmoid function

$$\Psi(x) = 1/(1 + e^{-2x}). \qquad 11$$

The weight values are obtained by learning or training in which the difference between φ and **y** for the calibration set is iteratively reduced. The optimal N_{hidden} can be obtained by using a test set or through cross validation.

1.6.1.1. Cross-validation

In order to build and validate a model, training and validation data sets are required. The training set is used to build the model and the validation set is used to test the performance of the model when presented with new data. A pre-requisite for this application is that there must be enough data to be split into training and validation sets. Where this is not the case, cross validation is applied. Here, the data (size $N_{samples}$) is split up into N_{splits} training sets (size N_{train}) and test sets (size $N_{test} = N_{samples} - N_{main}$). For each split a model is built using the training samples. Then the predictions for the respective test samples are used to calculate the Predicted Residual Error Sum of Squares (PRESS) thus:

$$PRESS = \sum_{i}^{N_{splits}} \sum_{j}^{N_{test}} (y_{ij} - \varphi_{ij})^2 \qquad 12$$

When it is used to assess the optimal number of latent variables in PLS and PCR, PRESS is calculated as a function of the number of latent variables. The number of latent variables that gives the minimum PRESS is used to build the actual model.

1.6.1.2. Variable Selection

Another alternative to PLS and PCR modelling is variable selection. The application of variable selection has the advantage of removing noise and uninformative variables prior to modelling to improve the robustness and predictive power of the model. Variable selection can also produce simpler models that are easier to interpret. Variable selection, as used in VS-MLR (variable-selection multiple linear regression) has been found to outperform both PLS and PCR. In variable selection, PRESS is used as a criterion, so that only variables that give the least PRESS are used to build the model.

1.6.1.3 General Randomised PRESS-based Elimination (GRAPE)[49]

In classical modelling techniques, the whole data set is used in building the model. The difficulties with this approach arise because of the inclusion of variables in the data that are a poor representation of the trend in the data. It is these variables that reduce the correlation between the real data and the predicted data from the model. There are other variables that accurately represent the trend in the data, and any model built with these variables highly correlates with the original data. Variable selection

techniques of model building seek to isolate only these variables for model building. GRAPE is an iterative variable selection process that consists of a random addition procedure and a PRESS-based elimination procedure that refines the set of selected variables by minimising PRESS. Another option in GRAPE is the use of PRESS-based addition. The steps in the recursive version of the GRAPE algorithm applied in this work are:

1. Create a trial solution S by randomly picking $N_s = N_{max}$ variables. N_{max} is a user-settable control variable.
2. Set $S_{best} = S$
3. Calculate $PRESS_S$
4. For each variable $i \in S$, calculate $PRESS_{S-i}$ (i.e. $PRESS$ of the trial solution S without the variable i.
- If $PRESS_{S-i} < PRESS_S$ then $S = S-i$
5. If $PRESS_S < PRESS_{Sbest}$, then $S_{best} = S$.
6. Add random variables $j \notin S$ to S until $N_S = N_{max}$.
7. Repeat steps 3-6 until a user-settable maximum number of iterations, $N_{iterations}$, is reached.

Thus in the first recursion, the number of variables to select from is equal to the total number of variables. In each subsequent recursion, only variables that were present at least once in S after step 4 and variables that were never picked in step 1 or 6 can be selected. This is repeated until no decrease in $PRESS$ is observed in a user-settable number of recursions counter $N_{recursions}$.

In order to make the results of the cross-validation independent of the number of samples, $PRESS$ is converted to the Root Mean Square Error of Cross-Validation (RMSECV) by:

$$RMSECV = \sqrt{PRESS/N_{samples}} \qquad 13$$

1.6.2: Univariate Linear Regression

This is commonly known as finding the line of best fit through a cloud of points. Linear regression is a method that fits a straight line through data. If the line is upward sloping it means that an independent variable has a positive effect on a dependent variable. If the line is downward sloping there is a negative effect. The

steeper the slope, the more effect the independent variable has on the dependent variable.

It is assumed that the relationship between a single X variable and one Y variable is linear, i.e.

$$Y=bX+a \qquad 14,$$

where b is the slope of the line and a is the intercept at the Y axis. In this text, the variance of a variable is a measure of the spread of a variable about its average value, and the covariance is a measure of the similarity of two variables. Variables having high covariance are strongly related to each other. To know the strength of this relationship, we also need to know the variance of the individual variables.

Univariate linear regression estimates the values of b and a by minimising the sum of squared vertical distances from points to the line. In other words, we choose a candidate slope, b and intercept, a. For each recorded (X, Y) pair, we square $Y - bX - a$ and add it to the total. The line having the smallest total is the best-fit line.

In practise, calculus gives us a formula for estimating b directly, and thence a, as follows: $\bar{b} = \frac{\text{Covariance(X, Y)}}{\text{Variance(X)}}$ **15,**

$\bar{b}$ standing for an estimate of b. a can be ignored if all the variables are centred before being used. This is done by calculating the average value of the variable and then subtracting this value from all sample values (mean centring). The value of a can be calculated after modelling using the estimated value of b and the subtracted averages.

When working with centred data, the linear regression equation for b in matrix form can be expressed as

$$\bar{b}=(X^TX)^{-1}X^TY \qquad 16$$

In this case if the variance of X is zero, then b cannot be estimated. This occurs when the X variable has the same value for all values of Y.

1.6.3: Multiple Linear Regression (MLR)

Correlation is a measure of relation between two variables. For example, a high correlation between purchases of say cheese and crackers indicates that these products are likely to be purchased together. Correlations may be either positive or negative. A positive correlation indicates that a high level of one variable will be accompanied by

a high value of the correlated variable. A negative correlation indicates that a high level of one variable will be accompanied by a low value of the correlated variable.

In the situation where there is more than one X variable, a linear regression can be formed with the assumption of a linear relationship. Then Y can be expressed as

$$Y = b_1x_1 + b_2x_2 + \ldots b_3x_3 + b_nx_n \qquad 17$$

As it happens, the matrix form of the linear regression equation, $b=(X^TX)^{-1}X^TY$ also works for multiple X variables. In this case, the resulting estimate of b is a vector containing the weights applied to the X variables.

In the case of multiple linear regression there are many situations when $(X^TX)^{-1}$ cannot be calculated. This situation arises whenever a (non-zero) weighted sum of the X variables gives a zero result, or one of the rows or columns of X contains all zeros. When such a weighted sum exists, the X variables involved are said to be collinear.

In practice, it is rare to be able to measure variables with absolute accuracy. So, even when some of the X variables are actually collinear, experimental values will not show this.

As the X variables become more and more collinear, the value of $(X^TX)^{-1}$ tends to zero. Small changes in collinearity alter this value radically. The effect of this on the model is particularly bad, because the model tends to amplify noise in the variables.

Also if the number of recorded samples is less than the number of X variables, then collinearity is guaranteed to occur. In this situation, the usual solution is to discard variables.

Therefore the disadvantages with MLR are

- It cannot handle collinearity.
- It is unstable with near collinearity.
- Relevant variables have to be discarded to avoid these problems.

1.6.4: Principal Components Analysis (PCA)

The instability of MLR when there are correlated X variables stresses the need to examine the structure within data sets. Finding such structure by hand can be extremely difficult, even in relatively simple cases.

Principal components analysis provides a method for finding structure in such data sets. Put simply, PCA rotates the data into a new set of axes, such that the first few axes reflect most of the variations within the data. By plotting the data on these axes,

we can spot major underlying structures automatically. Figures 4 to 7 illustrate this point.

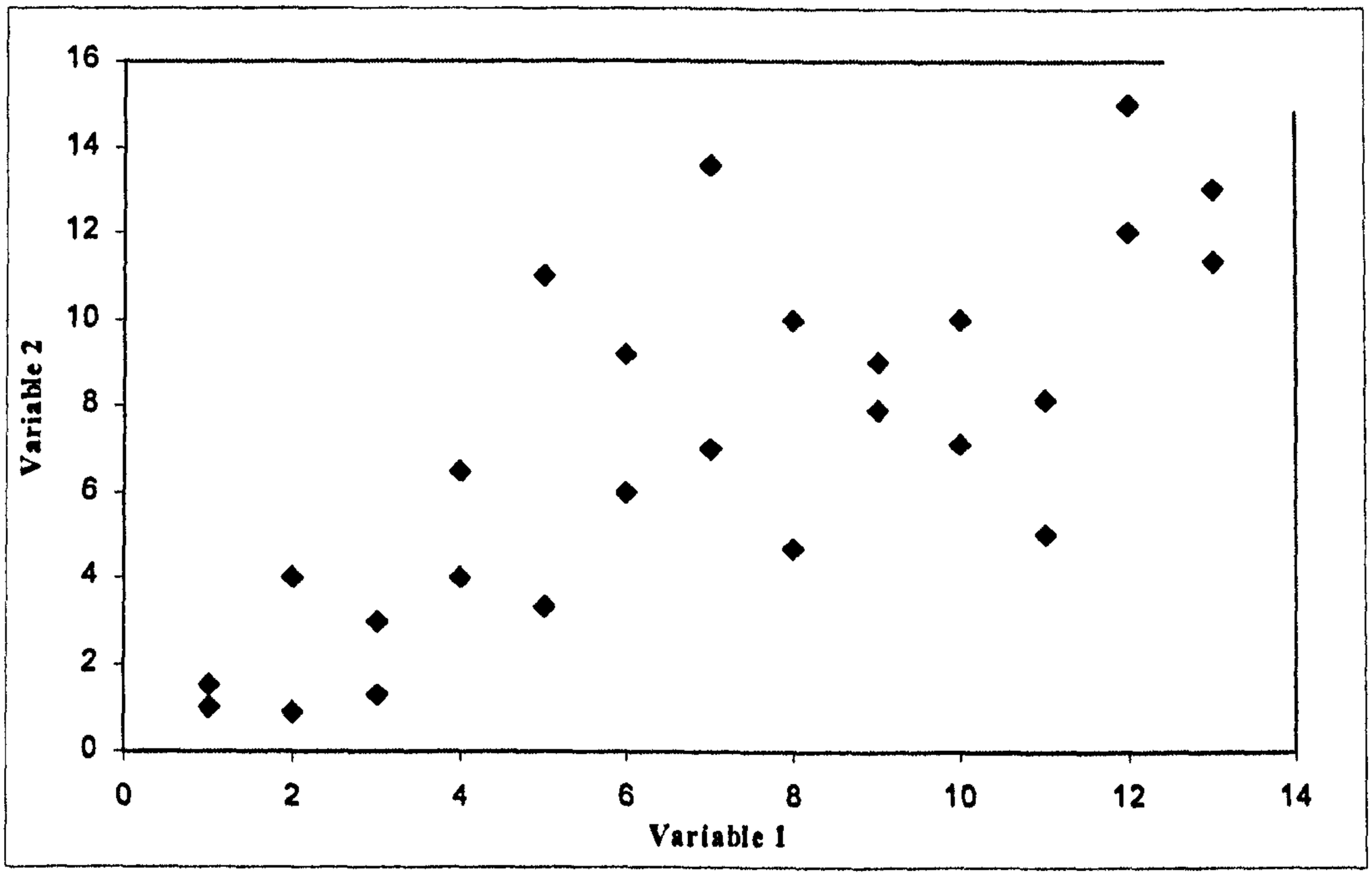

Figure 4: Scatter diagram of data from two variables, giving little information

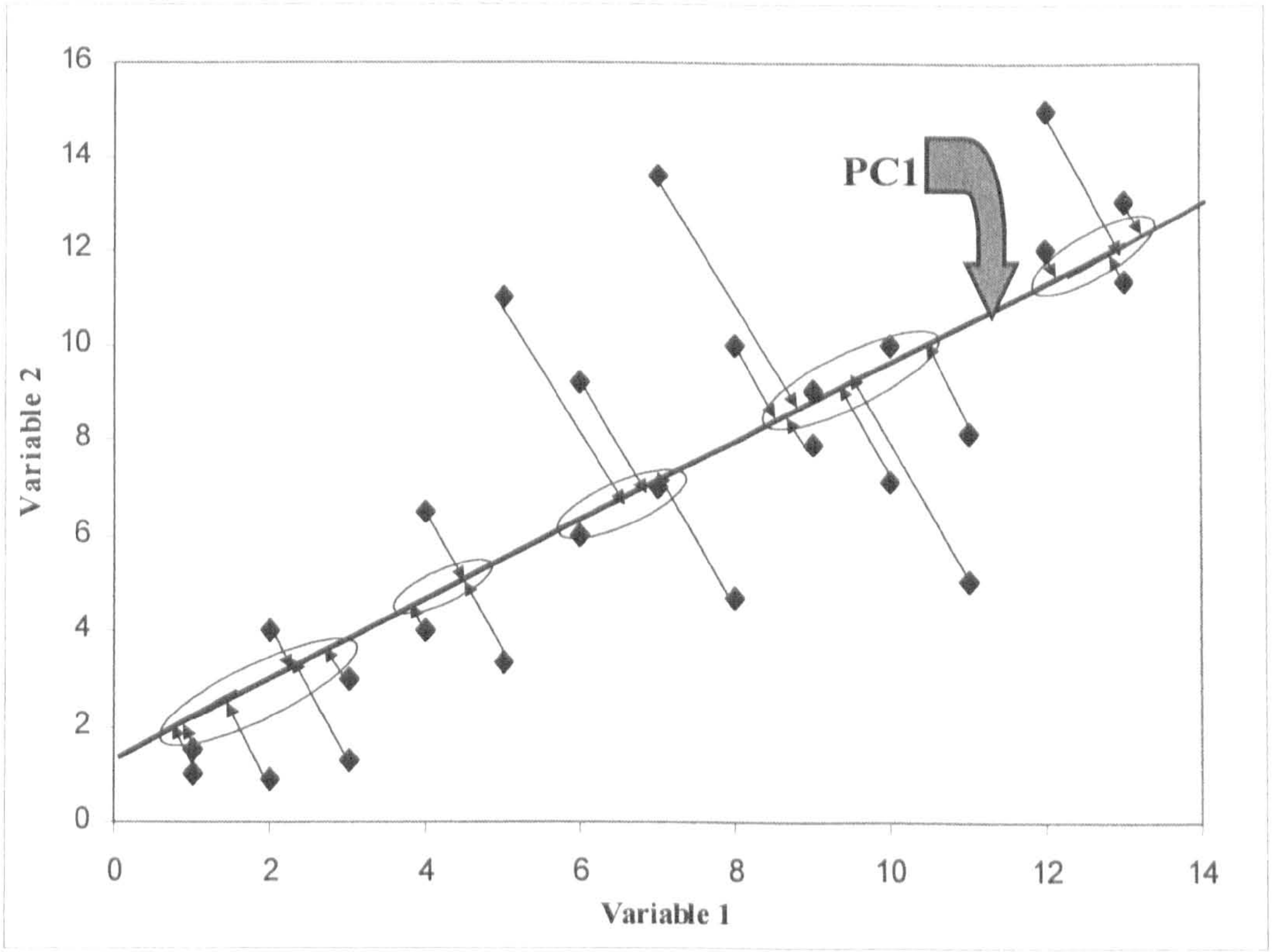

Figure 5: Data plotted on a new axis that captures the greatest variation in the data. The various groupings are easily identified. This axis represents the first principal component, PC1

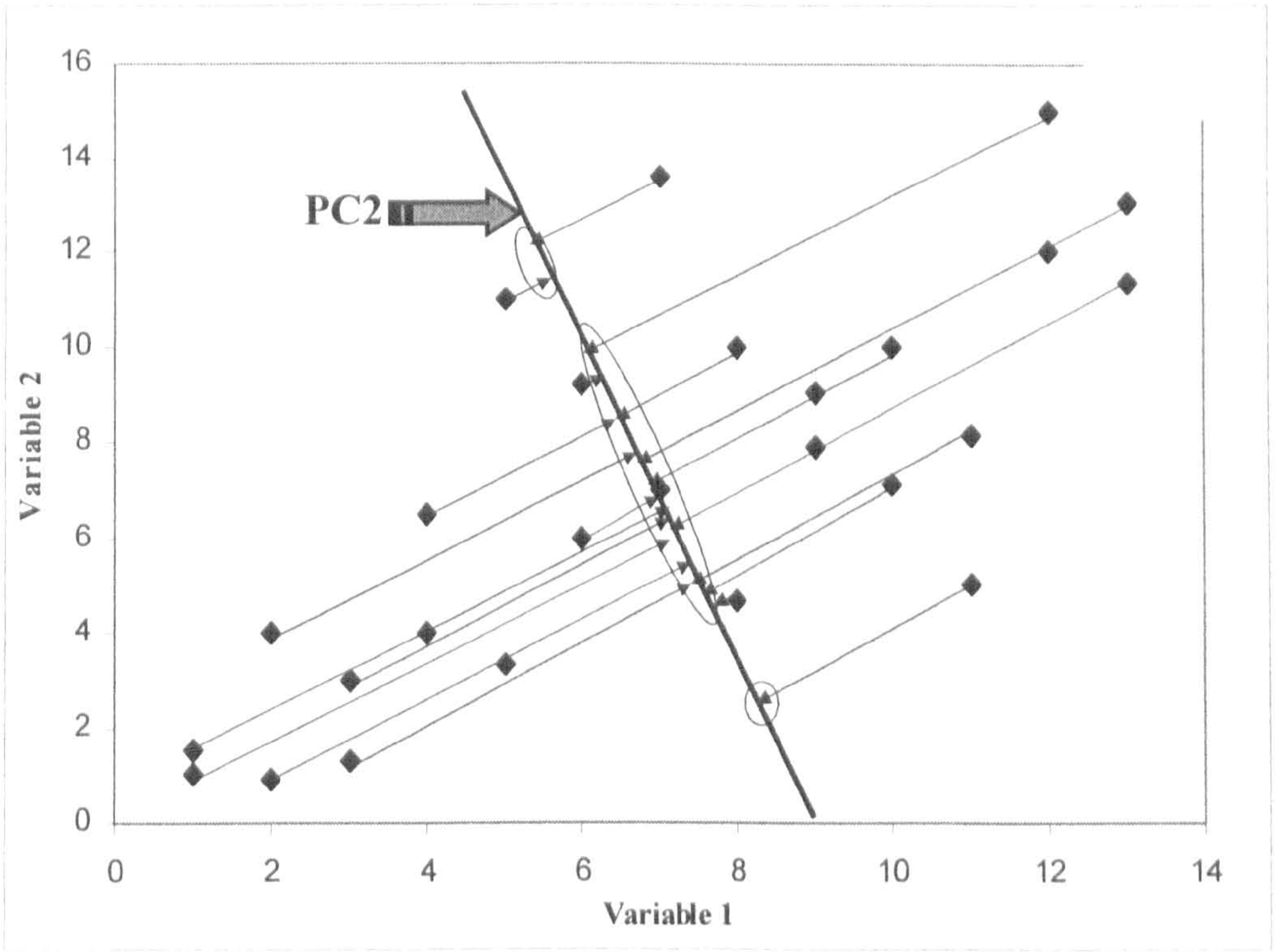

Figure 6: Data plotted on another axis perpendicular to the axis chosen in figure 5.

This is seen to capture less variation than the axis in figure 5, and the grouping here is different. This axis represents the second principal component, PC2

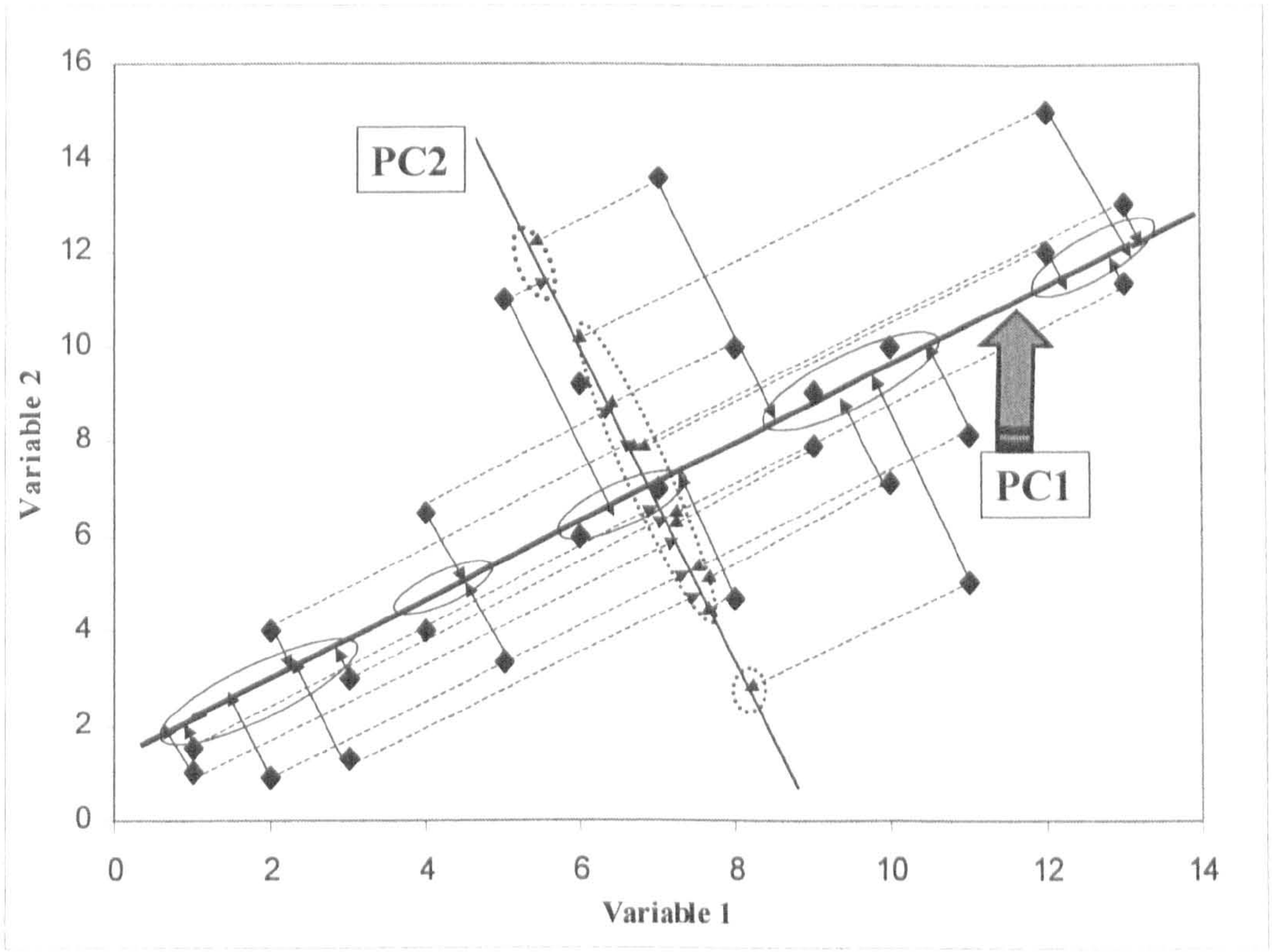

Figure 7: Data plotted on two new axes (PC1 & PC2) to reveal the underlying patterns. The principal component value is shown by the length of the arrow (continuous for the major axis and dotted-lines for the minor axis) linking the point to the particular axis.

The value of each point, when rotated to a given axis, is called the principal component value. PCA reduces the spectral data (**X**) into principal component scores (**T**) and loadings (**P**), according to the equation

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E} \qquad \textbf{18},$$

a linear transformation, where **E** is the X-residual matrix. The principal component scores are uncorrelated and are such ordered that the first few retain most of the variation present in all of the original variables. Thus only a few of the transformed variables are needed in further procedures[5]. The correlation or the covariance matrix of the variables is decomposed into eigenvectors, each with an associated eigenvalue. The matrix of eigenvectors, called the loadings, contains information on how the variables relate to each other while the scores give information on how the samples relate with each other. PCA is also used to detect outliers both in the samples and in the variables by observation of the residuals and by plotting the scores of the relevant

principal components against each other. Thus unwanted contributions can easily be identified and removed from noisy data [50]. In real samples, there are usually many different variations that make up a spectrum: the constituents in the sample mixture, inter-constituent interactions, instrument variations such as detector noise, changing environmental conditions that affect the baseline and absorbance, and differences in sample handling. Yet, even with all of these complex changes occurring, there should be some finite number of independent variations occurring in the spectral data. Hopefully, the largest variations in the calibration set would be the changes in the spectrum due to the different concentrations of the constituents of the mixtures. Ideally, a set of "variation spectra" that represented the changes in the absorbances at all the wavelengths in the spectra could be used instead of the raw spectral data for building the calibration model. There should be fewer common variations than the number of calibration spectra (in most cases), and thus, the number of calculations for the calibration equations will be reduced as well.

This "variation spectra" could be used to reconstruct the spectrum of a sample by multiplying each one by a different constant scaling factor and adding the results together until the new spectrum closely matches the unknown spectrum. Obviously, each spectrum in the calibration set would have a different set of scaling constants for each variation since the concentrations of the constituents are all different. Therefore, the fraction of each "spectrum" that must be added to reconstruct the unknown data should be related to the concentration of the constituents.

The "variation spectra" are often called eigenvectors (or spectral loadings, loading vectors, principal components or factors), for the methods used to calculate them. The scaling constants used to reconstruct the spectra are generally known as scores. This method of breaking down a set spectroscopic data into its most basic variations is called PCA.

Since the calculated eigenvectors came from the original calibration data, they relate to the concentrations of the constituents that make up the samples. The same loading vectors can be used to predict "unknown" samples; thus, the only difference between the spectra of samples with different constituent concentrations is the fraction of each loading vector added (scores).

The calculated scores are unique to each separate principal component and training spectrum, and can be used in place of absorbances in either of the classical model equations (CLS or ILS). Since the representation of the mixture spectrum is reduced

from many wavelengths to a few scores, it seems best to use the ILS expression of Beer's Law for calculating concentrations due to its ability to calculate concentrations among interfering species. It is important to note, however, that the calculations maintain the CLS averaging effect by using a large number of wavelengths in the spectrum (up to the entire spectrum) for calculating the eigenvectors. So, in effect, eigenvector models combine the best features of both the CLS and ILS methods together in the calculation. This is the main reason why eigenvector models are generally better than classical models in both accuracy and robustness.

PCA breaks apart the spectral data into the most common spectral variations (factors, eigenvectors, loadings) and the corresponding scaling coefficients (scores).

The trick in using these models comes in how the eigenvectors are calculated. These models base the concentration predictions on changes in the data, not absolute absorbance measurements (which are used in all the classical models). In order to calculate the PCA model, the spectral data must change in some way.

Multiple orthogonal factors: After the line on which the variance is maximal is established, there remains some variability around this line. In PCA, after the first factor has been extracted (that is, after the first line has been drawn through the data), another line is defined that maximises the remaining variability, and so on. In this manner, consecutive factors are extracted. Because each consecutive factor is defined to maximise the variability that is not captured by the preceding factor, consecutive factors are independent of each other, i.e. uncorrelated or orthogonal to each other. Basically, the extraction of principal components amounts to a variance maximising rotation of the original variable space. For example in a scatterplot, the regression line represents the original X-axis, rotated so that it approximates the regression line. This type of rotation is called variance maximising because the criterion for (goal of) the rotation is to maximise the variance of the "new" variable (factor), while minimising the variance around the new variable. One important point for consideration is how many factors to extract. As consecutive factors are extracted, they account for less and less variability. The decision of when to stop extracting factors depends on when there is only very little "random" variability left. The nature of this decision is arbitrary.

The defining characteristic that distinguishes between PCA and principal factors analysis is that in PCA it is assumed that all variability in an item should be used in the analysis, while in principal factors analysis only the variability that an item has in common with the other items is used. In most cases, these two methods usually yield

very similar results. However, PCA is often preferred as a method for data reduction, while principal factors analysis is often preferred when the goal of the analysis is to detect structure.

In PCA calibration, there can be problems with collinearity. If the concentrations of 2 important constituents in the calibration samples are always present in the same ratio (for example, 2:1 of A to B, such as if dilutions were made from a single stock sample), the model will only detect one variation, not two. As far as the model is concerned, all the absorbance peaks of constituent A increase or decrease when constituent B also increases or decreases, and vice versa. Thus, only one variation is detected: the changes in the spectrum of A+B. Therefore, it is very important when calibrating eigenvector models that the calibration data have concentrations of the individual constituents of interest present in evenly and randomly distributed ratios.

Before PCA is applied to a training set, the data is commonly mean centred. This means that the mean spectrum (average spectrum) is calculated from all of the calibration spectra and then subtracted from every calibration spectrum. Mean centring has the effect of enhancing the subtle differences between the spectra. This is very essential since eigenvector methods calculate the principal components based on changes in the absorbance data, and not the absolute absorbance. Thus anything that improves the ability of the calculation to detect the differences between the calibration spectra improves the model. Since the eigenvectors represent the changes in the spectral data that are common to all the calibration spectra, removing the mean simply removes the first most common variation before the data is even processed by the PCA algorithm.

PCA is effectively a process of elimination. By iteratively eliminating each independent variation from the calibration spectra in series, it is possible to create a set of eigenvectors (principal components) that represent the changes in the absorbances that are common to all. When the training data has been fully processed by the PCA algorithm, it is reduced to two main matrices: the eigenvectors (spectra) and the scores (the eigenvector weighting values for all the calibration spectra). By multiplying PC1 & PC2 (eigenvectors) by the set of representative scalar fractions (scores) and summing the results (along with the mean spectrum if the data was mean centred), the original calibration spectra can be recreated. The "spectral residual" is the difference between this reconstruction and the original.

1.6.5: Principal Components Regression (PCR)

PCA selects a new set of axes for the data. These are selected in decreasing order of variance within the data. They are also (of course) perpendicular to each other. Hence the principal components are uncorrelated. Some components may be constant, but these will be among the last selected.

The problem noted with MLR was that correlated variables cause instability. The solution is calculating principal components, throwing away the ones that only appear to contribute noise (or constants), and using MLR on the rest. This process gives the modelling method known as Principal Components Regression. Rather than forming a single model, as with MLR, models can now be formed using 1, 2, ... components, and the optimal number of components are decided. If the original variables contained collinearity, then some of these components will contribute only noise. So long as these are dropped, these models are guaranteed to be stable.

1.6.6: Partial Least Squares (PLS)

The intention, in using PCR, has been to extract the underlying effects in the X data, and to use these to predict the Y values. In this way, only independent effects are used, and low-variance noise effects are excluded. This improves the quality of the model significantly.

However, PCR still has a problem: if the relevant underlying effects are small in comparison with some irrelevant ones, then they may not appear among the first few principal components. This presents a component selection problem – it is not acceptable to just include the first n principal components, as these may serve to degrade the performance of the model. Instead, all components are extracted, and it is determined whether adding each one of these improves the model. This is a complex problem.

Partial Least Squares (PLS) regression solves the problem. The algorithm used examines both X and Y data and extracts components (now called factors), which are directly relevant to both sets of variables. These are extracted in decreasing order of relevance. So, to form a model now involves extracting the correct number of factors to model relevant underlying effects. [51]

For the two matrices, say a process variable data matrix **X** (n x m) and a matrix of corresponding product quality data **Y** (n x k), one would like to extract latent variables that not only explain the variation in the process data **X**, but that variation in **X** which is most predictive of the product quality data **Y**. PLS accomplishes this by working on the sample covariance matrix $(\mathbf{X}^T\mathbf{Y})(\mathbf{Y}^T\mathbf{X})$ such that the first latent variable

$$t_1=w_1{}^T x \qquad 19$$

is that linear combination of the x variables that maximises the covariance between it and the **Y** space. The first PLS loading vector w_1 is the first eigenvector of the sample covariance matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. Once the scores

$$t_1=\mathbf{X}w_1 \qquad 20$$

for the first component have been computed, the columns of **X** are regressed on t_1 to give a regression vector

$$p_1 = \mathbf{X}t_1/t_1{}^T t_1 \qquad 21$$

and the **X** matrix is deflated to give residuals

$$\mathbf{X}_2 = \mathbf{X}\text{-}t_1 p_1{}^T. \qquad 22$$

The second latent variable is then computed as $t_2=w_2{}^T x$ where w_2 is the first eigenvector of $\mathbf{X}_2{}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}_2$, and so on.

CHAPTER TWO: EXPERIMENTAL

2.1 Experiment 1: The esterification of ethanol and acetic acid

The reaction of interest, the sulphuric acid catalysed esterification of ethanol by acetic acid, is old and well established. The reaction is represented by the following scheme:

$$CH_3COOH + CH_3CH_2OH \xleftrightarrow[H^+]{} CH_3COOCH_2CH_3 + H_2O \qquad (23)$$

The reaction goes through a 5-step $A_{AC}2$ mechanism with known intermediates[52], shown by the following schematic:

$$RCOOH + H^+ \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} RC^+OOH_2 \qquad (24)$$

$$RC^+OOH_2 + R`OH \underset{k_{-2}}{\overset{k_2}{\longleftrightarrow}} RC(OH)_2OHR` \qquad (25)$$

$$RC(OH)_2OHR` \underset{k_{-3}}{\overset{k_3}{\longleftrightarrow}} RC(OH)(O^+H)_2OR` \qquad (26)$$

$$RC(OH)(O^+H)_2OR` \underset{k_{-4}}{\overset{k_4}{\longleftrightarrow}} RC^+(OH)OR` + H_2O \qquad (27)$$

$$RC^+(OH)OR` \underset{k_{-5}}{\overset{k_5}{\longleftrightarrow}} RCOOR` + H^+ \qquad (28)$$

where R = CH_3 and R` = CH_2CH_3. Step 27, which involves the production or consumption of a molecule of water, is the rate-determining step.

2.1.1 Kinetics

The necessity of knowledge of the kinetics of the reaction in order to adequately monitor the process cannot be overemphasised. First of all, the kinetic studies of an

experiment establish or confirm the mechanism involved. One obtains a precise ascertainment of the reaction parameters. To monitor a chemical process, knowledge of all the reacting species including reaction intermediates is very important in order to avoid underestimation, inconclusive generalisation and inexplicable changes. This knowledge comes from a kinetic study of the process. The study of the kinetics of the various elementary steps of a complex reaction reveals those steps that are crucial, for example the rate determining step. Kinetic studies give the order of the reaction with respect to each of the reacting species, and this is necessary in order to monitor the concentration of any constituent at any given time. For multi-step reactions, the kinetics show the lifetime and the conditions favouring the appearance or otherwise of each chemical species in the mixture. This is very important during monitoring, as the need for crucial timing cannot be overemphasized.

Generally, the reaction progress must show increasing product content and decreasing reactant content. For a second order reaction, the reciprocal of the amount of a reaction component must be directly proportional to the time. For this particular reaction therefore, the integrated form of its rate equation will be

$$\frac{1}{A_o - B_o} In \frac{B_o A}{A_o B} = kt, \tag{29}$$

where A_o and B_o are the initial concentrations of ethanol and acetic acid respectively, and A, B are the respective concentrations at time t, k being overall rate constant, i.e.

$$k = \frac{k_1 k_2 k_3 k_4 k_5}{(k-1)(k-2)(k-3)(k-4)(k-5)}. \tag{30}$$

It follows that a plot of $\frac{1}{A_o - B_o}$ versus t should be linear. The linearity of this plot is taken as the proof of the order of the reaction[53].

Although on the whole the reaction is second order, it is first order with respect to each of the reactants as well as each of the products. For first order kinetics, the rate of consumption of a reactant is proportional to the concentration of that reaction component, by definition, i.e.,

$$-d[A]/dt \propto [A], \quad (31)$$

and therefore

$$d[A]/dt = -k[A], \quad (32)$$

where k is the velocity constant, t is the time in seconds and $[A]$ is the concentration of the reaction component. The integrated form of Equation 32 is given as

$$In[A] = -kt + Constant \quad (33)$$

Therefore a plot of $log_{10}[A]$ against t should give a straight line with a slope of $-k/2.303$, as a confirmation that the reaction is of the first order[54].

2.1.2 Reagents

Glacial acetic acid (AR grade), ethanol, and concentrated sulphuric acid are the starting materials. Acetic acid and ethanol are the reactants while concentrated sulphuric acid catalyses the reaction.

Concentrated sulphuric acid:

This is a corrosive and irritant chemical. It is toxic when ingested. Contact with the skin and internal organs causes severe burns. In case of eye contact, immediate rinsing with copious amounts of water followed by seeking medical care is advised. Direct addition of water results in a violent reaction and is therefore not advised. The TLV exposure limit is $1mg/m^3$.

Acetic acid:

This is a flammable liquid that causes severe burns. Care should be taken not to breathe the fumes. In case of eye contact, immediate rinsing with copious amounts of water and seeking medical attention is advised. The TLV exposure limit is $25mg/m^3$.

Ethanol:

A highly flammable liquid, this substance must be kept away from any sources of ignition. Care should be taken to avoid skin contact. Precautionary measures should be taken against static discharge. The exposure limit is an 8hr TLV of $1900mg/m^3$.

In addition to the precautions mentioned above, the following precautions are taken during the experiment:

- The temperature of the heater, the flow of the tap feeding the condenser and the position of the reflux zone are constantly monitored throughout the procedure.
- The entire work is done in a fume cupboard that has been checked to be functioning properly, thus avoiding inhalation.
- All sources of ignition are kept away from the working area.
- All components are stored in a well-ventilated and separate place.
- All waste is stored for specialist disposal, nothing going down the drain.

2.1.3 Apparatus and Procedure

The reaction was performed in a 1-litre reaction vessel set in a water bath heated by a thermostated hotplate and fitted with a condenser and a glass stirrer driven by an electric powered motor. A model IMO/H0 Raman probe connected to an HL5-785-250 Kaiser optical system and a Hololab Series 5000 Raman Spectrometer was fitted at one of the probe-ports. The Raman probe uses a class 3B-focus laser. It was then connected to a computer with the *HOLOGRAM* software that controls the spectrometer and probe as well as receiving and storing data loaded on it (Figure 8).

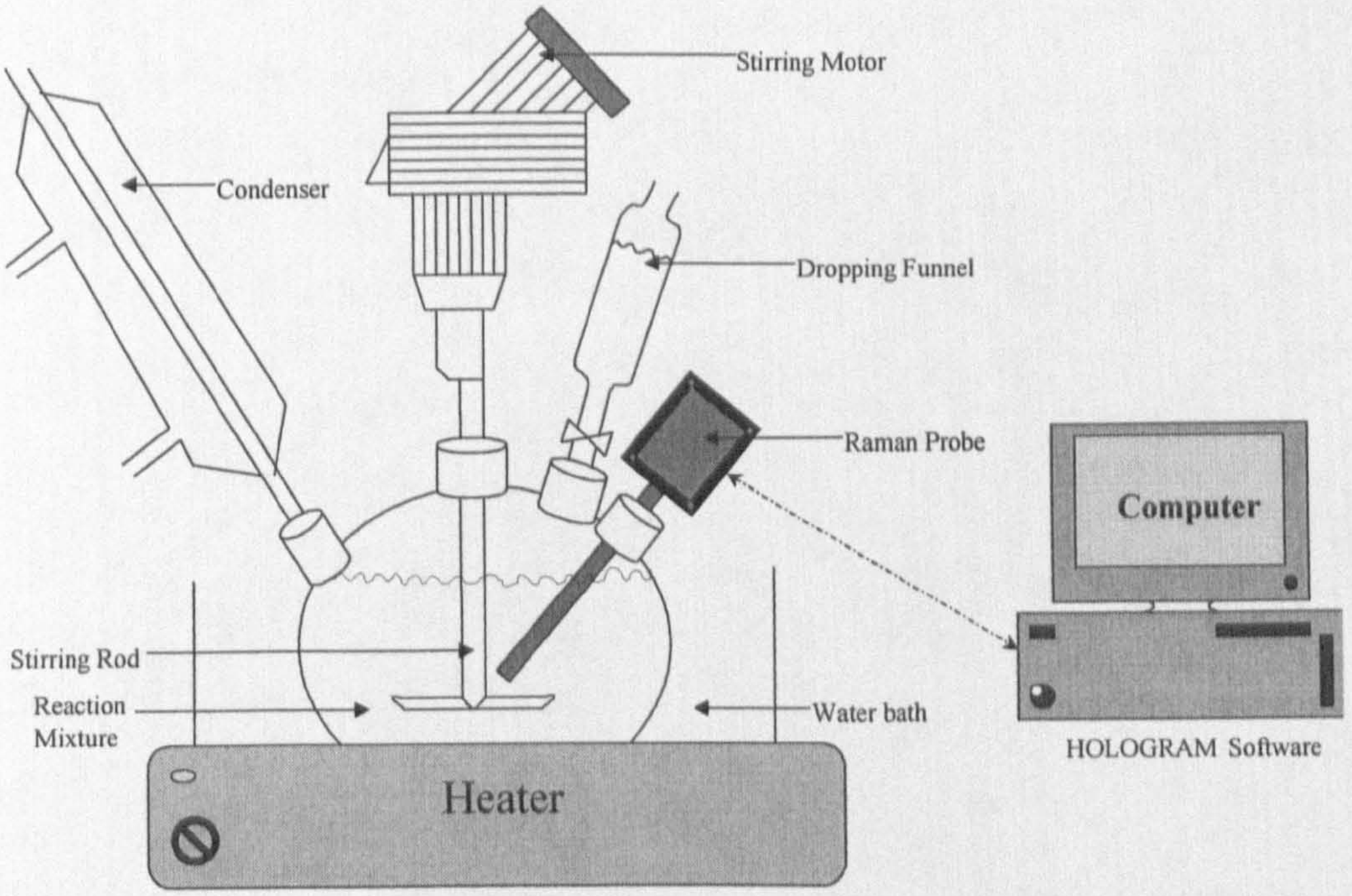

Figure 8: Experimental set up showing the configuration of the Raman probe and other accessories in the 1-litre vessel. The heater is controlled by a thermostat. The condenser, dropping funnel, Raman probe and the motor driving the glass stirrer are all held in place by clamps so as not to exert any pressure on the reaction vessel.

300ml of ethanol was poured into the reaction vessel, followed by 300ml glacial acetic acid. These volumes were used so that at least 60% of the length of the probe would be submerged in the liquid. That way, the probe head is deep enough inside the bulk of the reaction mixture to give a uniform and representative measurement. The reactants were slowly brought to 78°C while stirring. The reaction was initiated by the addition of 15ml concentrated H_2SO_4, the catalyst.

2.2 Experiment 2: The Distilled Feed Column Data[55]

This is real industrial data obtained from an industrial chemical processing plant. The Raman data is obtained from feed to a distillation column that separates raw materials into heads for onward processing into naphtha products and into tails that are sold as fuel (Figure 9).

The raw material is naphtha, which is made up of various levels of paraffin compounds like butanes, pentanes, dimethylbutanes, methylpentanes, hexane,

methylhexanes and heptane, aromatic compounds like benzene and toluene, and then naphthenes like cyclopentane, methylcyclopentane, cyclohexane and methylcyclohexane. The main reaction is oxidation of naphtha into various organic acids and acetone.

$$\textit{Naphtha} + \textit{air} \longrightarrow \textit{various organic acids} + \textit{acetone} \qquad (31)$$

The tailoring column (as shown in Figure 9) is basically a distillation column with trays stacked evenly upwards.

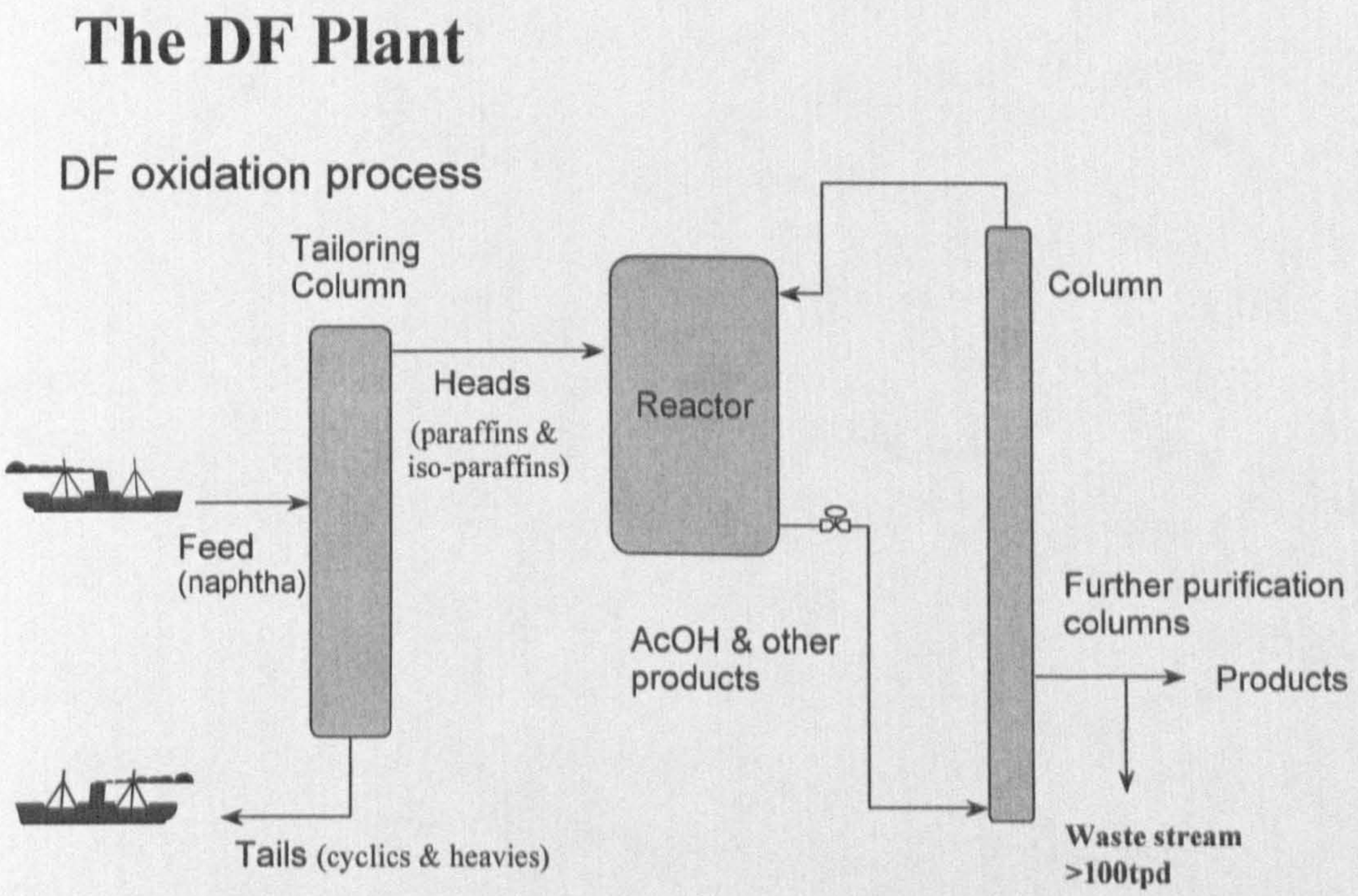

Fig. 9: Schematic of the Naphtha processing system on the DF Plant. Feed material is introduced to the column wherein heat from the base causes fractions up to C_6 to emerge as heads and on to the reactor for oxidation into various products.

The pre-heated feed enters the column around the middle tray and undergoes fractional distillation. The heavier materials (tails) descend to the lower trays while the higher trays end up rich in the more volatile components (heads). A re-boiler at the base of the column continually heats up the tails, so that lighter components keep rising up all the time. Only components with boiling points less than or equal to that

of cyclohexane pass out as heads. In effect, the naphtha tailoring distillation column minimises the amount of C_6's and above that gets to the reactor by removing cyclic and longer chain hydrocarbons. The heads vapours are condensed and then passed on to the reactor for the oxidation into acetic acid, other organic acids and acetone which are all very useful starting materials for the production of many chemical products. The oxidation reaction, after such tailoring, is more efficient and there is less material leftover to be recycled.

The Raman spectrometer is situated in the control room with a fibre-optic link to the probe head and laser inside a purged enclosure in an analyser hut, next to the distillation column (Figure 10). The design of the system is such that it measures the compositions of both the feed and heads. It takes about a minute to analyse one stream, then it switches to the other, and so on.

The spectrometer is connected to a computer network system for monitoring and control of the process, display and analysis of the acquired data. The whole system is designed to allow for safe and remote data capture and analysis, so that personnel are as far away from any potential hazards within the proximity of the plant area as possible.

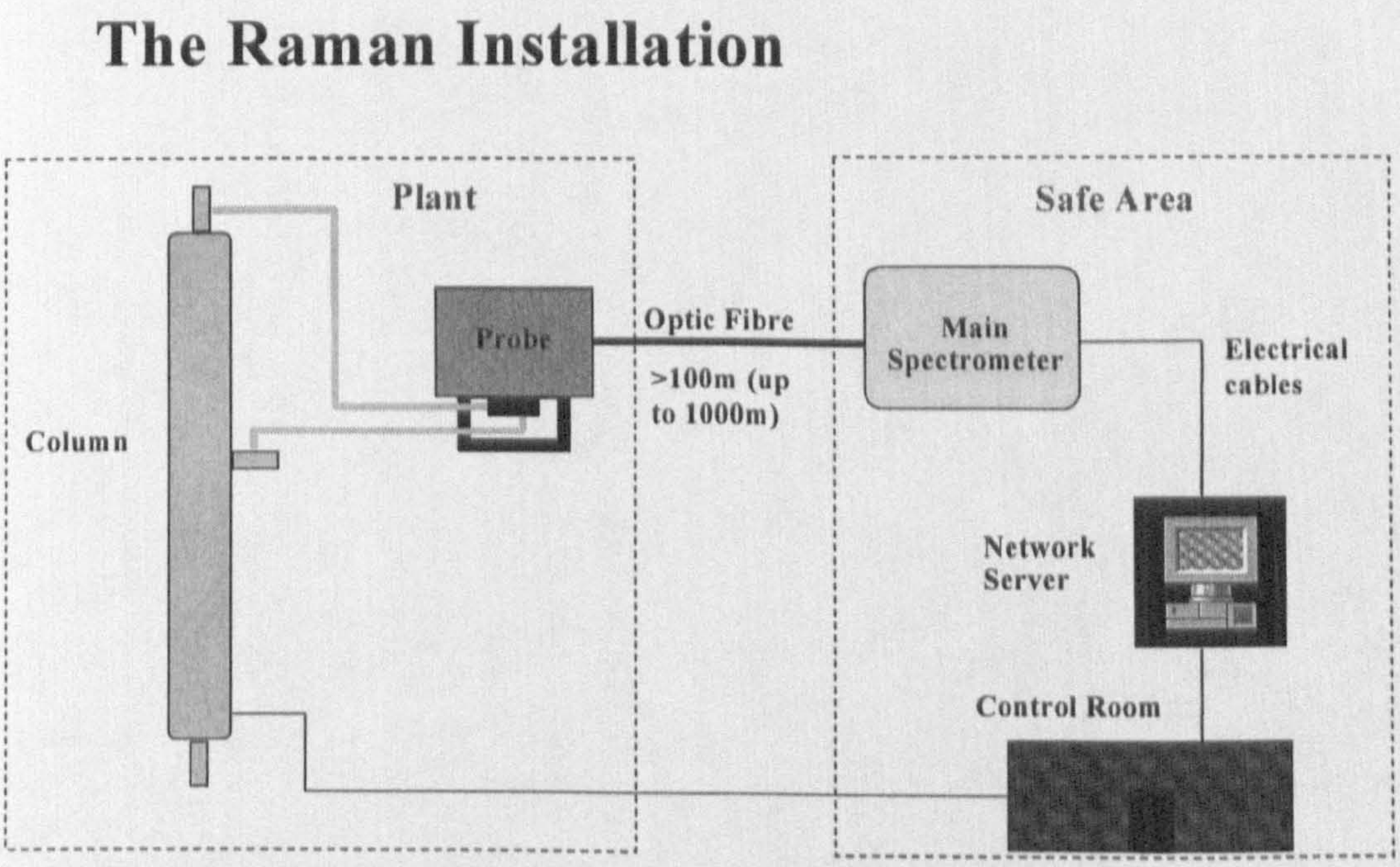

Figure 10: Set-up of the system for obtaining Raman spectra from the distillation column. Operating personnel control the system from the control room, a safe distance from the hazards in the proximity of the plant area.

The data set consists of 137 samples with 1925 variables of Raman spectra from the process. Some of the samples are spiked in order to make selected Raman peaks stand out since the fluorescence masks the Raman data in many places. Also, the spiking helps in calibrating the data.

The first 89 samples are distillation (separation) column feed data. Of these, the first 24 are not spiked, while samples 25 to 52 (except 49) are spiked with various combinations of some of the non-spiked samples. Samples 53 to 89 exhibit Laser Induced Fluorescence (LIF), the last twelve of which are spiked as earlier described.

The next forty samples are from the distillation column heads and are all spiked (except the last three) but are non-LIF. Then finally there are eight LIF non-spiked samples from the installation tank. The observation is that the feed data, being rich in the heavy components, show more fluorescence than the lighter heads. This information is shown in the Appendix.

The data is supplied with gas chromatography reference data size 137 by 17 to facilitate the building of calibration models. The data was obtained form BP Amoco Chemicals Limited, Saltend, Hull, UK.

2.3 Software

In this work, the data processing was done with the software MATLAB 5.2 (The MathWorks Inc., Natick, Massachusetts, USA) in a Windows operating environment. MATLAB has many powerful tools for manipulating, storing, and graphing n-dimensional data. MATLAB is an interactive, matrix-based system for scientific and engineering calculations with which complex numerical problems can be solved without actually writing a program. MATLAB is a high-performance language for technical computing. It integrates computation, visualisation and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include:

- Math and Computation
- Algorithm development
- Modelling, simulation and prototyping
- Data analysis, exploration and visualisation
- Scientific and engineering graphics

- Application development, including Graphical User Interface (GUI) building.

In this work, MATLAB has been used in the data analysis stage (Principal Component Analysis), the modeling stage (Partial Least Squares) and information visualisation.

The PCA algorithm was obtained from the PLS_Toolbox 2.0 (Barry M. Wise and Neal B. Gallagher, Eigenvector Research, Inc). All data used for the PCA was mean-centred by rows using an algorithm in the PLS_Toolbox. Smoothing of some of the spectral profiles was done using a 15-point Savitsky-Golay smoothing algorithm in the PLS_Toolbox.

Initial data collection and treatment was performed using Microsoft Excel, a spreadsheet software program included in Microsoft Office 97, patented and copyright owned by the Microsoft Corporation, USA.

CHAPTER THREE: RESULTS AND DISCUSSION

3.1. Experiment 1 - Raman monitoring of the esterification reaction

Initially, the Raman spectra of the Raman-sensitive components of the reaction, i.e. ethyl acetate, ethanol and acetic acid, were obtained separately. Figures 11, 12 and 13 show the spectra of these pure substances. These spectra give the naked appearance of the various reaction components individually, without the influence of the other components or the reaction conditions. More importantly, these spectra reveal, individually, the important peaks of each reaction component. Since these are pure compounds, the spectral intensity and therefore concentrations that they would show would be the maximum, and would thus be useful in the quantitative determination of the various reaction components when they are combined in the reaction mixture.

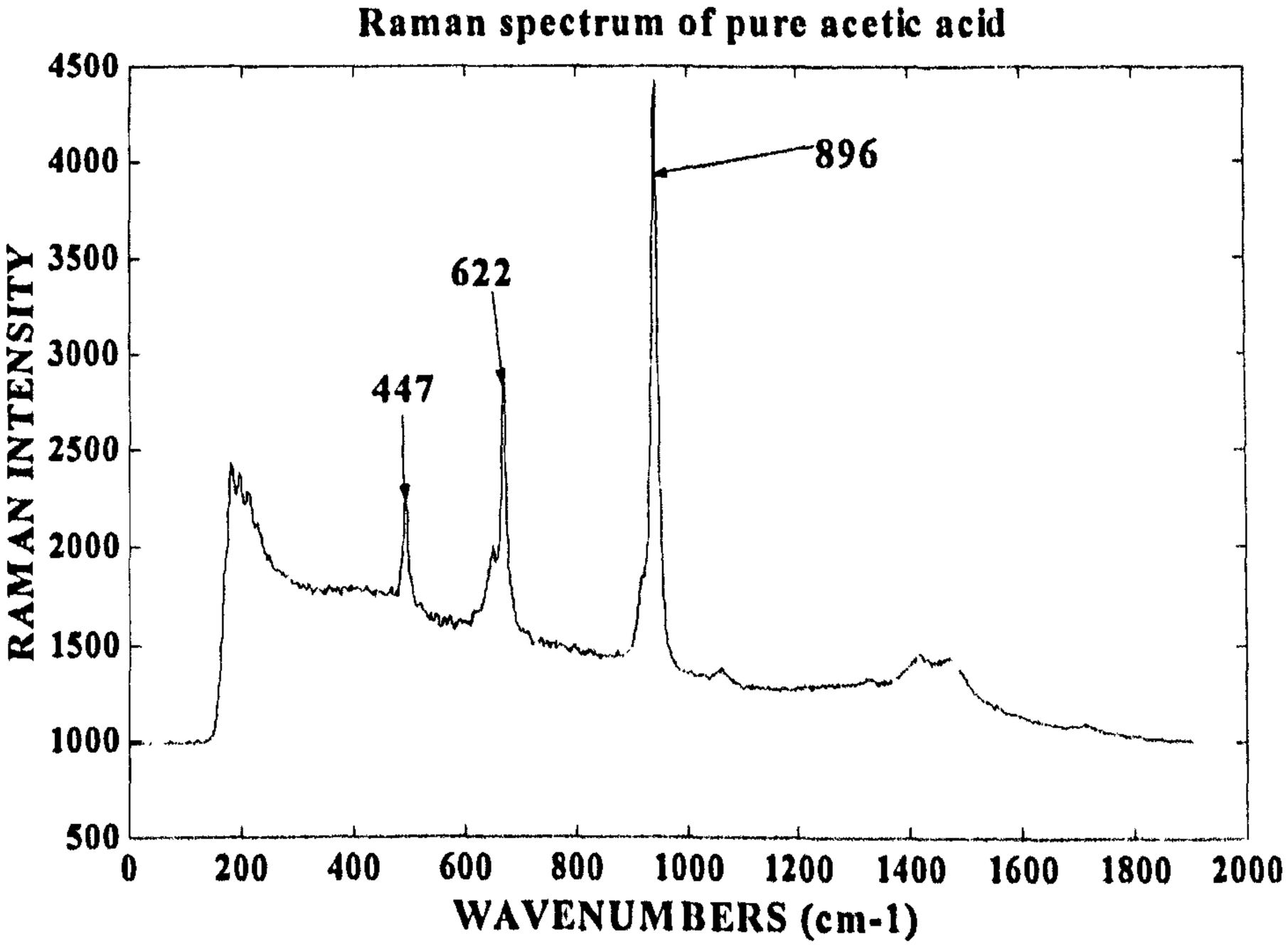

Figure 11: Raman spectrum of pure acetic acid, showing the prominent peaks at 896 due to C-O stretch, 622 due to OH in-plane-deformation, and 447cm-^{1}.

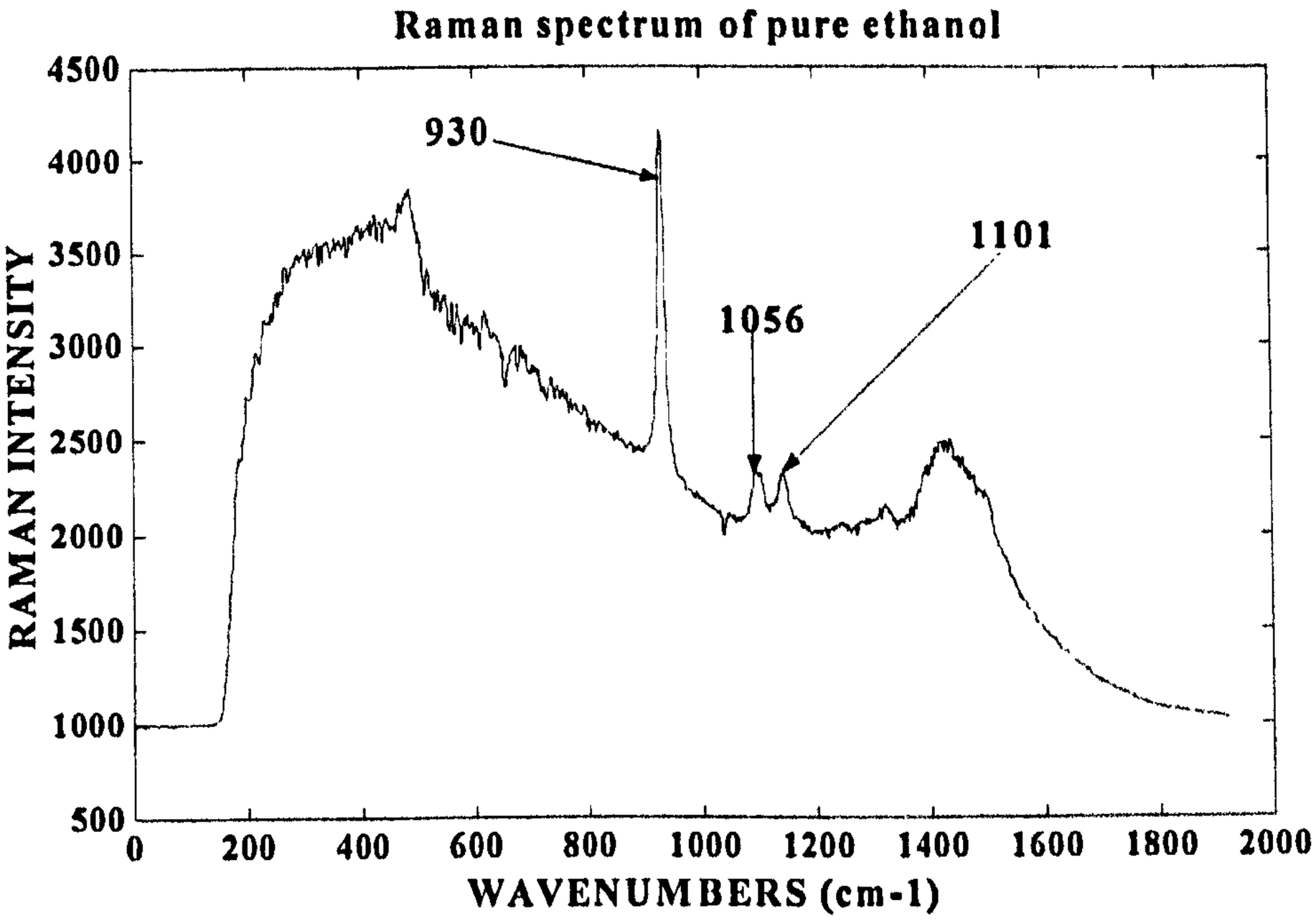

Figure 12: Raman spectrum of pure ethanol, showing the prominent peaks at 930, due to C-O stretching, 1056 and 1101cm^{-1} respectively.

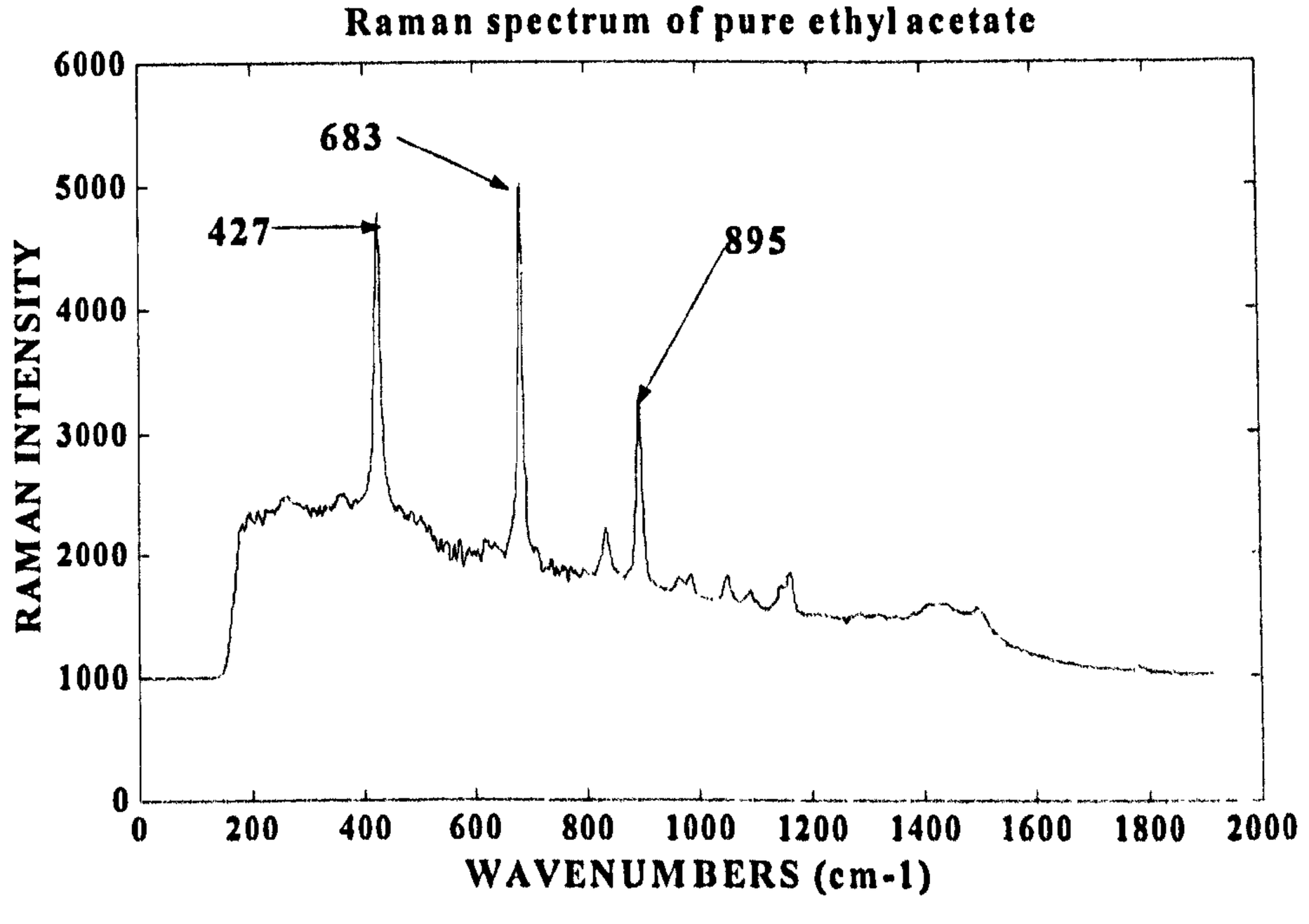

Figure 13: Raman spectrum of pure ethyl acetate, showing the prominent peaks at 895cm^{-1} due to weak O-ethyl bands, 683cm^{-1} and 427cm^{-1} due to C=O stretch.

Comparison of each of the pure spectra and reference to literature led to the assignment of the peaks in the reaction mixture to their respective compounds. Thus peaks at 427 and 683 cm^{-1} were identified with ethyl acetate (Figure 13). The $427cm^{-I}$ peak of ethyl acetate comes from the C=O stretch[56] in the compound, while a CH_3 rocking motion in the O-ethyl bonds produces the $895cm^{-I}$ peak[57].

The $930cm^{-1}$ peak identified as an ethanol peak (Figure 12) comes from the C-O stretch in CH_2OH, typical of primary alcohols. This peak falls in a band that involves C-C-O asymmetric stretching[58].

The $622cm^{-1}$ peak of acetic acid (Figure 11) is due to the in-plane OH deformation while the peak at $896cm^{-1}$ is due to a strong C-O stretch[59]. These functional group assignments to the peaks are a further confirmation that the peaks detected correspond to the respective compounds.

One of the advantages of Raman spectroscopy is that water does not interfere with the normal Raman signal. Therefore although water is a significant by-product of the reaction, it does not appear in the spectrum at all; neither does it affect the spectra of the other components.

Figure 14 is a plot of the Raman spectra of real-time monitoring of the esterification of ethanol and acetic acid. There is an obvious shift in baseline, all spectral intensities starting from around 1000 units, and a huge amount of noise especially in the wavelength region corresponding to the products. The noise is observed to be greater in intensity than the Raman data in many places. However, the detailed nature of the information captured is highly commendable, as shown in the 3-dimensional data plot (Figure 15). This ability to record a huge amount of data continuously is one of the high points of the Raman technique that makes it most suitable for continuous monitoring of chemical processes.

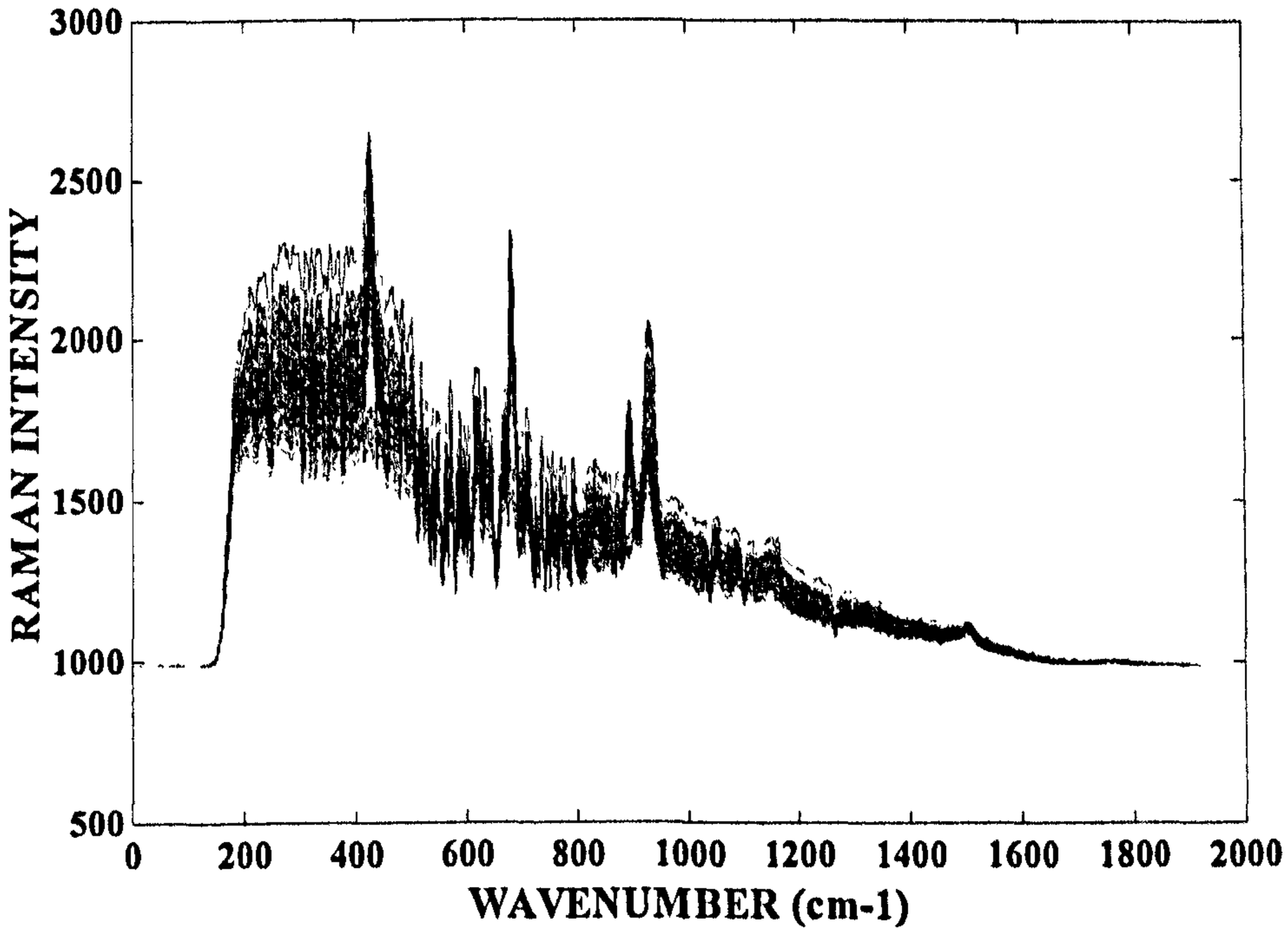

Fig. 14: A plot of the Raman spectra of real-time monitoring of the acid-catalysed esterification of ethanol and acetic acid over the entire reaction period. Each scan has 1918 data points, and the huge baseline shift and massive noise is clear.

Comparing Figure 14 with Figures 11, 12 and 13, the decrease in intensity of the Raman spectra is obvious. In Figures 11, 12 and 13, the peak heights range between 1000 and 5000 units approximately, whereas the range in Figure 14 is only 1000 to around 2500. This is explained by the fact that in Figure 14 all the components are mixed together and therefore their relative fractional compositions would be smaller in magnitude than that in the pure liquids.

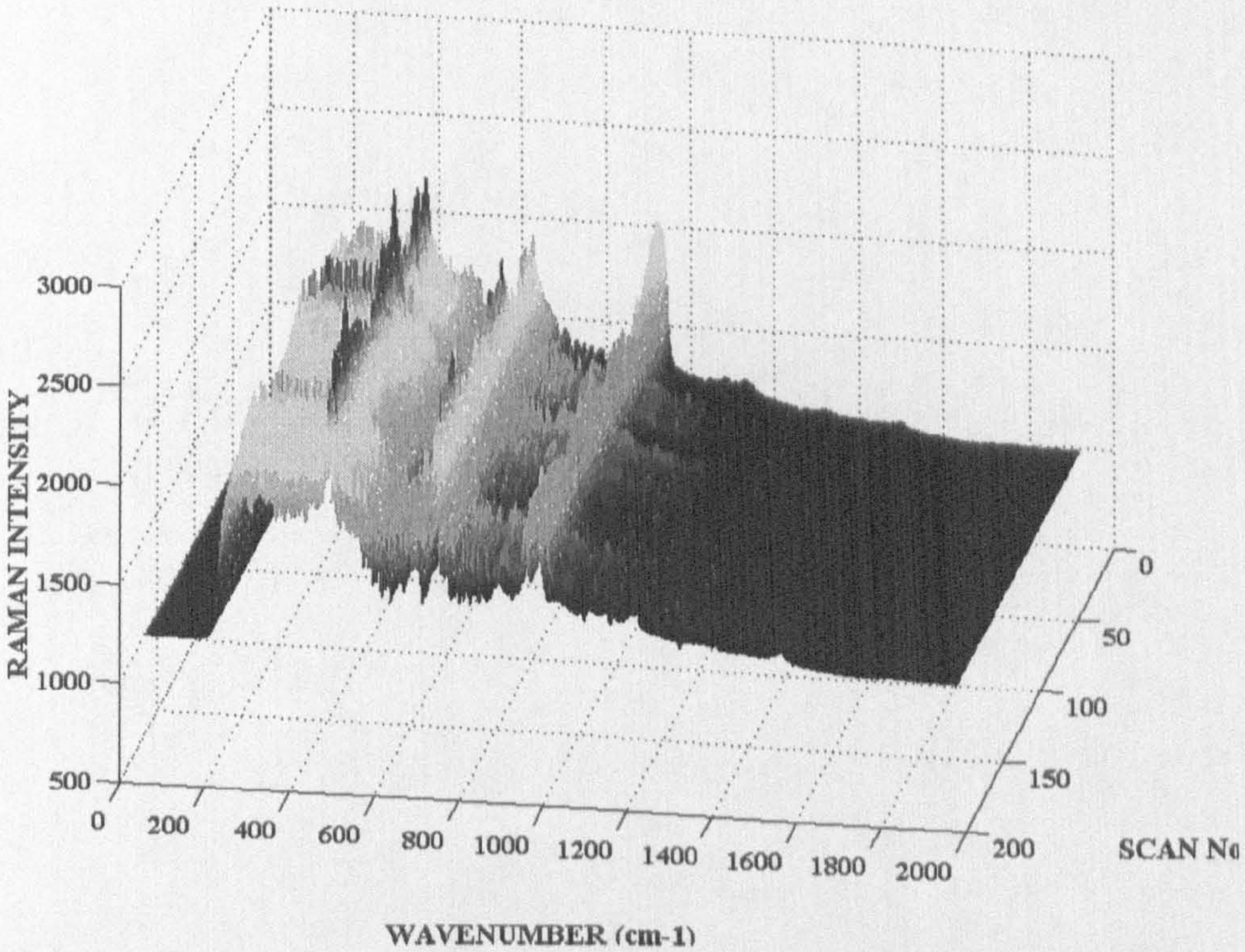

Figure 15: A three-dimensional plot of the Raman spectra of real-time monitoring of the acid-catalysed esterification of ethanol and acetic acid, showing all the 165 scans taken over the entire reaction period. Across the spectrum it can be observed the variation as well as peak appearance and disappearance as the reaction progresses.

Figure 15 also shows the progression of the wavelengths corresponding to the various reaction components along the time (Scan no.) axis. From a 'bird's eye view' sort of perspective the rise and fall of the peaks and the baseline are easy to observe. The diagram is a height-sensitive colour map, therefore the various colours show the intensity of the Raman spectrum at any particular point or area. The colours progress from blue-black to dark blue to light blue to yellow and then to red as Raman intensity increases. This makes the recognition of the highest and lowest intensities easy.

3.1.1 Data Analysis

Having identified the major wavenumbers relating to the various components in section 3.1, it is necessary to further establish their identity by seeing how they fit into the reaction process. Thus if a particular wavenumber represents ethanol for example, it is expected to progress like a reactant, and in an opposite way to another wavenumber that represents say ethyl acetate whose progress should be that of a product.

For a given reaction it is expected that the reactant-product relationship should have perfectly negative correlation as long as the process is in progress; in effect, as one increases, the other decreases. In order to establish this, a reactant profile has been regressed against a product profile using the original data, as shown in Figure 16. The diagram suggests four stages of the reaction: the initialising or settling down period **O** (stage one), the active reaction stage • (stage two), a transitional stage + (stage three), and the dormant stage ▲ (stage four).

Each symbol in Figure 16 represents a time span of 15 seconds. Thus during the first 3 minutes and 15 seconds of stage one, there was no particular pattern. We see a cluster of points rather than the expected linear arrangement. This suggests that though the process has been started, the chemical reaction has not started yet, hence the unsettled appearance. The system begins to look settled during the last 60 seconds of this stage, indicated by the straight-line formation of the 4 points. In stage two the points follow a straight-line formation with a negative gradient, just as is expected from a product-reactant plot. The 5 $\frac{1}{4}$ minutes of stage three again show no particular pattern while stage four, though it has an overall upward trend generally, also show no particular pattern from point to point.

The implication is that it is only during the 3-minute second stage, scans 18 to 29, that the reaction is in progress. Surely this is not acceptable, since it stands to reason that as soon as the catalyst was added to reactants at the desired temperature, the esterification should start. What we are observing in the non-linear sections is the result of the noise in the system overshadowing the real Raman data. Thus this noisy data must be separated from the data of interest, the pure Raman data. This is the objective of this research.

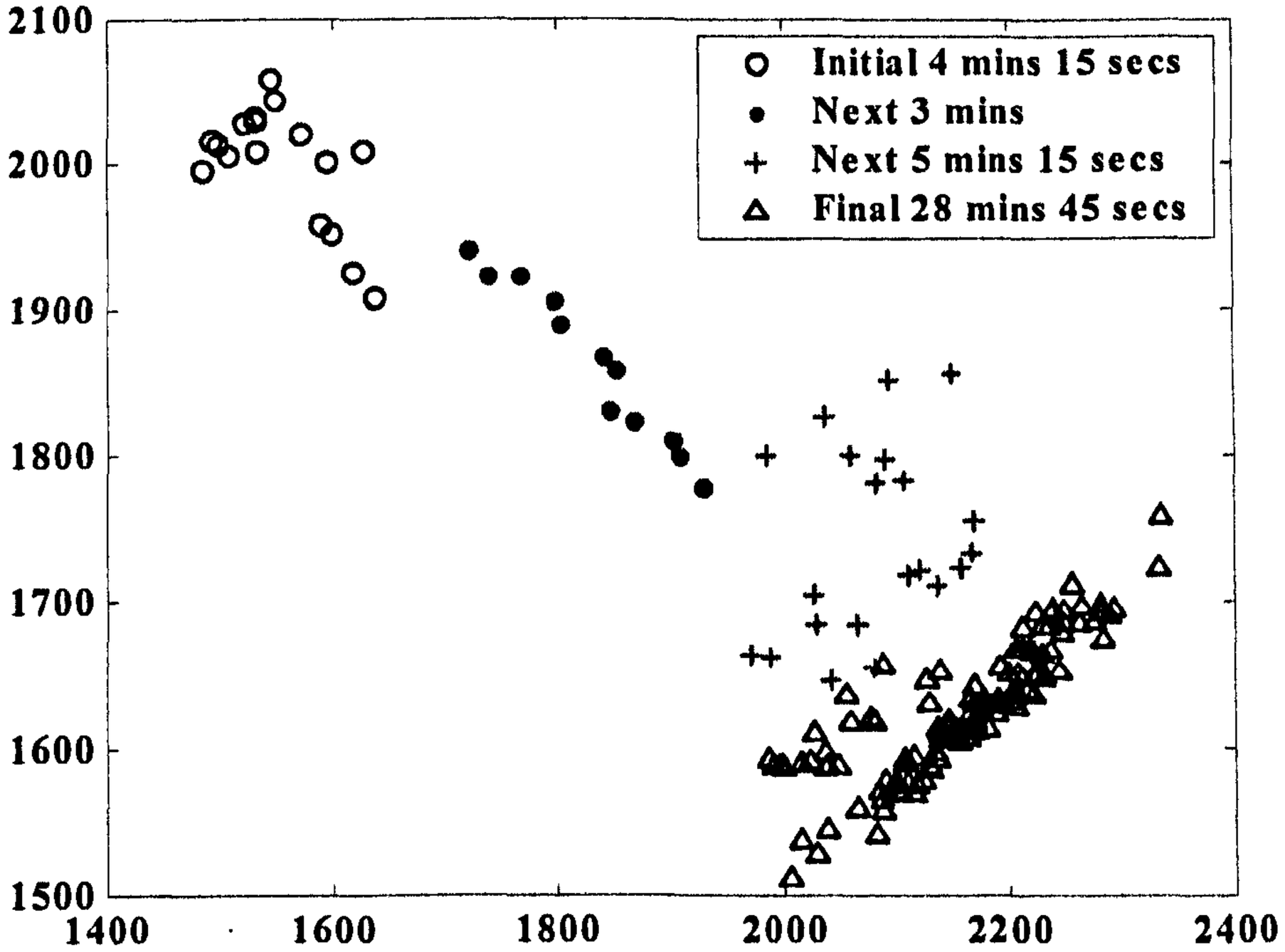

Figure 16: Regression of a product profile (683cm^{-1}) against a reactant profile (930cm^{-1}) from the initial data, showing the apparent different stages of the esterification process.

The product-reactant relationship during monitoring is expected to look like that shown in Figure 17. In effect, the product concentration begins from a minimum and keeps rising till the reaction is spent. Meanwhile, the reactant concentration having started from a maximum keeps decreasing till the reaction is spent. Thus devoid of any disturbance or noise effects, the profile of any product and reactant in the esterification reaction described in Section 2.1 is expected to be like Figure 17. To ascertain this, the profile of a product wavelength (683cm^{-1} for ethyl acetate) is traced together with that of a reactant wavenumber (930cm^{-1} for ethanol).

General profile of a reaction that tends towards equilibrium

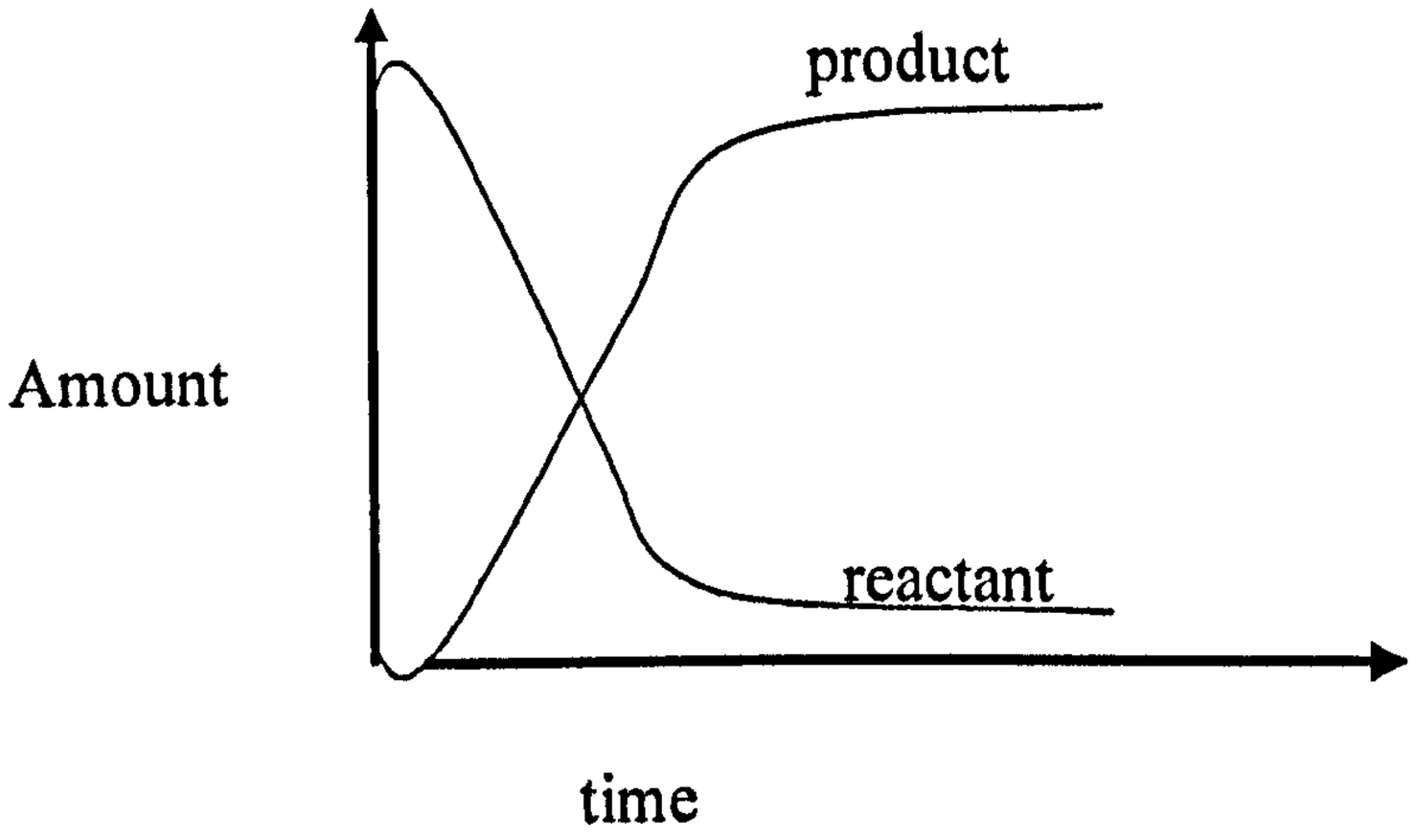

Fig.17: Schematic diagram of expected profiles of products and reactants during the reaction. The reactant starts from a high concentration level and then falls continuously till it settles at a constant value, while the product starts from a low concentration level and then rises continuously till it settles at a constant value.

A plot of the profiles of these peaks (930 & 683cm^{-1}) from the original (untreated) data during the monitoring period is presented in Figure 18. To allow for better viewing and presentation, the data was smoothed using a 15-point Savitsky-Golay algorithm.

The plot shows the expected product-reactant relationship only up to the region of scan 34, 8 $\frac{1}{2}$ minutes into the reaction. After this, there is an identical undulating pattern for both reactant and product, to the finish. Clearly this is not the result of a chemical reaction since reactant and product are always expected to behave opposite to each other and not in concord. This observation can be attributed to a dynamic effect such as a temperature fluctuation or the effect of stirring considering the similar 'wavelengths' of the two cycles, being around 50 scans (12.5 minutes) each. The undulating pattern, however, is therefore significant enough to completely overshadow the response of the reaction components. This is one effect that needs to be removed in order to see the full progress of the chemical reaction.

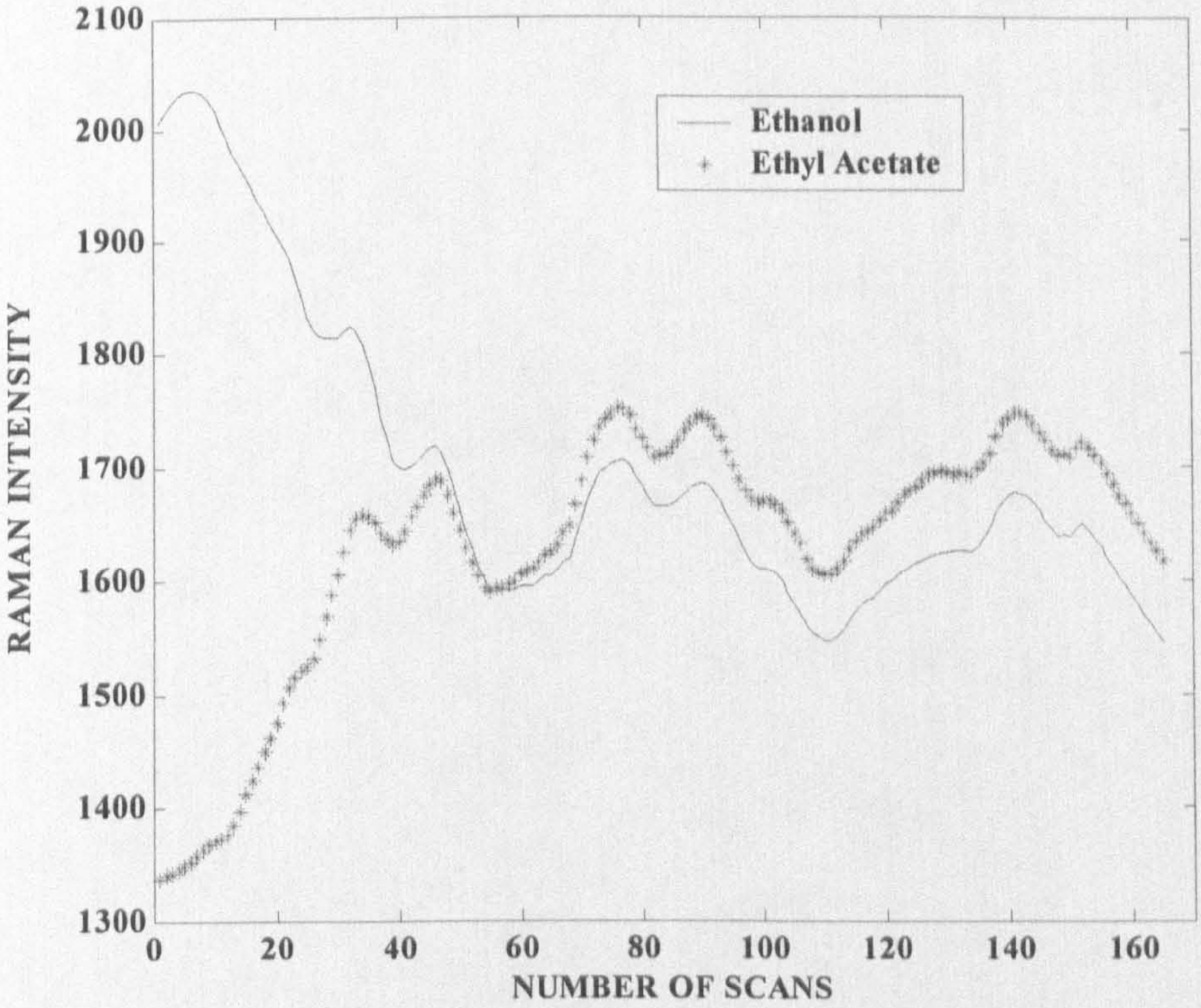

Figure 18: Profiles of ethyl acetate and ethanol during the esterification. Data smoothed using the 15-point Savitsky Golay smoothing algorithm.

To remove the effect of fluorescence and other non-chemical contributions, Principal Components Analysis (PCA) was performed on the mean-centred raw data. The data was first mean-centred in order to enhance the subtle differences between the spectra. PCA was selected as a technique because it has the ability of separating spectra that are due to independent variations occurring in the data. The assertion has been that the undulations in Figure 18 and the massive noise in Figure 15 are due to phenomena entirely different from and unrelated to the chemical reaction that is responsible for the Raman spectra, i.e. undulations caused by temperature fluctuations or the effect of stirring and noise due to fluorescence. Since these are independent of the esterification reaction, it follows from the theory of PCA that any variations due to them will be represented by different principal components from the principal component due to the variation in the chemical information. In this way, it should be possible to separate the fluorescence noise and undulations from the pure Raman spectra.

The first step in PC analysis is performance of eigen analysis to determine the optimum number of principal components required to build the PCA model. If the number of principal components used is too small, the model is inadequate and cannot represent the original data. Inclusion of too many principal components also leads to 'overfitting', since principal components that contain little information and mostly noise are also included in the model building exercise. In effect, any principal component that does not contribute significantly to the variation in the data should not be included in the PCA model building. One way of checking the contribution of the principal components is by the use of their corresponding eigen values. The smaller the eigenvalue, the less its contribution to the variation in the data.

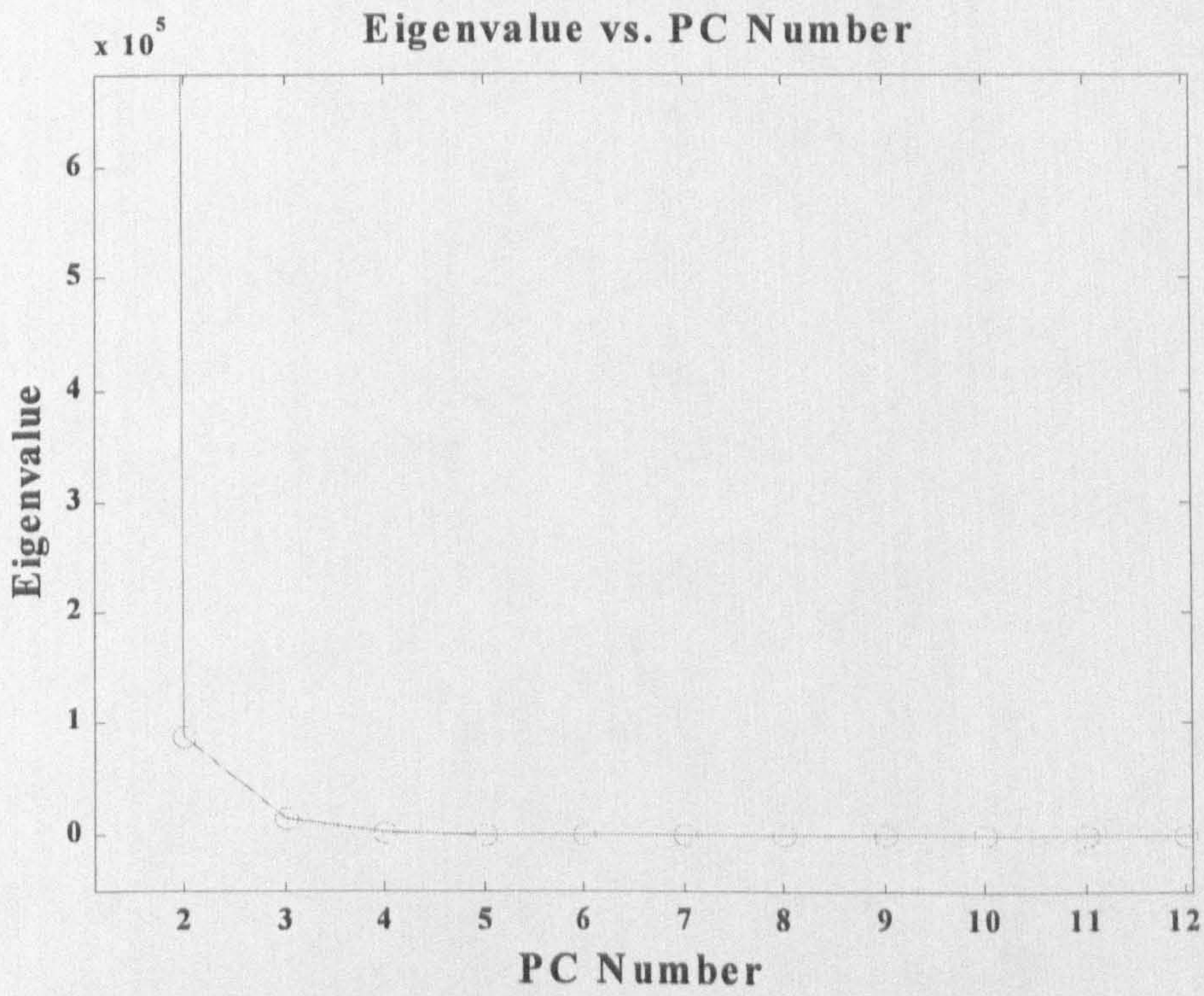

Figure 19: Plot of eigenvalue versus number of principal components for the Raman data, used to determine the optimum number of principal components to use in the PCA model building.

The optimum principal component number chosen is the one beyond which no significant change occurs in the eigenvalues of the subsequent principal components. Figure 19 shows a plot of eigenvalues against number of principal components (pc's) for the mean-centred esterification data. The huge difference in eigenvalue between

PC1 and PC2 indicates that PC1 contains the greatest fraction of the total variation and therefore needs to be retained. Similarly PC2 and PC3, though with smaller differences, can also be retained. The eigenvalue difference between PC3 and PC4, and then PC4 and PC5 are very small in comparison, but beyond PC5 there is no observable change in the eigenvalues. In order to avoid leaving out any relevant information in the PCA model building therefore, the first five principal components were used to construct the model. The data in PC6 and beyond only represent noise and contain no chemical information. They are therefore not useable for PCA. The result of the PCA modelling using 5 principal components is shown in Table 1. It is worth noting that before the PCA modelling, the data is mean centred.

PRINCIPAL COMPONENT NUMBER	EIGENVALUE OF COV(X)	% VARIANCE CAPTURED BY THIS PC	TOTAL % VARIANCE CAPTURED
1	$1.61e^{+007}$	99.34	99.34
2	$8.84e^{+004}$	0.55	99.88
3	$1.40e^{+004}$	0.09	99.97
4	$7.27e^{+002}$	0.00	99.97
5	$3.01e^{+002}$	0.00	99.98
6	$1.62e^{+002}$	0.00	99.98
7	$1.28e^{+002}$	0.00	99.98
8	$9.99e^{+001}$	0.00	99.98
9	$9.04e^{+001}$	0.00	99.98
10	$8.03e^{+001}$	0.00	99.98

Table 1: Results from PCA Model of esterification data.

The values of percentage variance captured confirm that the choice of 5 principal components was most appropriate for building the PCA model, since there is no increase in these values after 99.98% total variance at PC5. It is worth noting that although Table 1 ends with the 10th principal component, the value 99.98% remains the same through to PC25.

With the separation of the various groupings in the Raman data complete, the next stage is to reconstitute the Raman data using the values obtained from the PCA

model. Since PCA simply separates the data according to their contribution to the total variation, the nature of the original data is not affected. Therefore data reconstituted from the PCA should retain the nature and characteristics of the original data. The greatest contribution to variation in the data is PC1, and therefore data reconstituted from the first principal component is expected to bear the most important information in the data. The importance of the reconstituted data is expected to decrease with subsequent principal components.

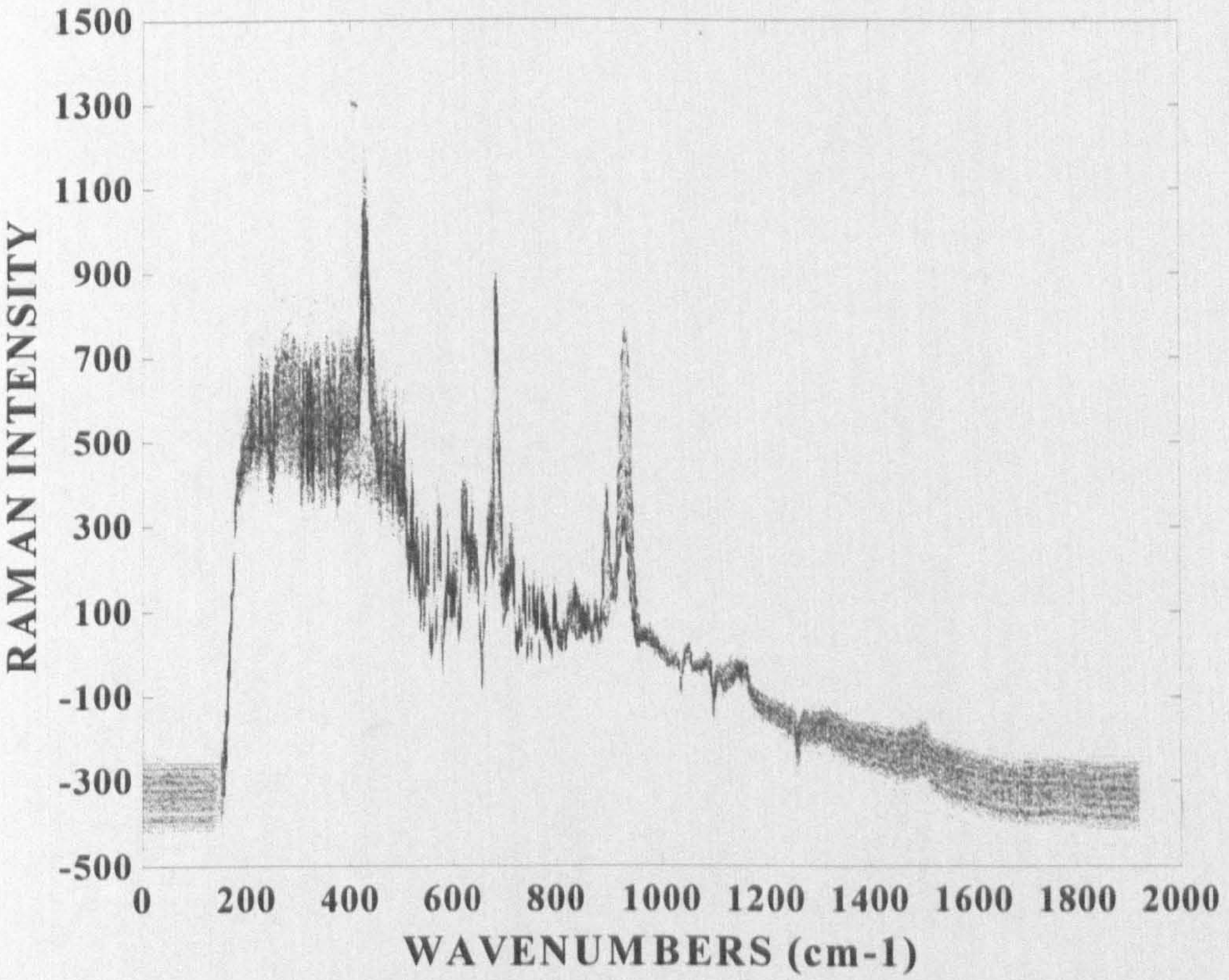

Figure 20: Regenerated data based on variation in PC1. This is very similar to the original data because PC1 contains 99.34% of the variation in the original data.

The scores and loadings of the first principal component (PC1) were mathematically recombined to regenerate the data based upon the variation in PC1. The resulting plot is shown in Figure 20.

The information in Figure 20 looks very much like the raw data, when compared with Figure 14. The huge baseline shift and the big noisy peaks are still present. This is to be expected since PC1 contains 99.34% of the variation in the data. The information in Figure 20 is therefore a very important part of the data. To determine the nature of

the data reconstituted from PC1, the scores of that principal component are observed. An observation of the scores on PC1 (shown in Figure 21) shows that the information here is definitely not that of the Raman spectra, considering the erratic nature of the score plot.

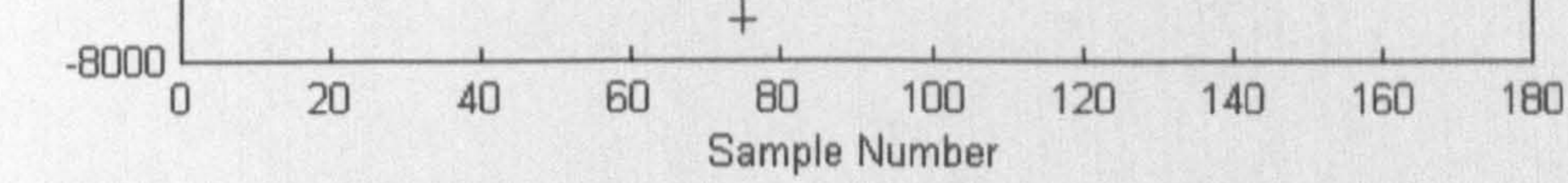

Figure 21: Sample number versus scores on PC 1. This diagram is obtained directly from the PCA algorithm in PLS_Toolbox. Here, each sample number represents a single scan

The systematic rise and fall in Figure 21 are a result of a noise effect that operates in cycles throughout the monitoring period. To ascertain the source of this noise, the profiles of wavenumbers 100cm^{-1} and 1850cm^{-1} (being baseline values) in the mean-centred data (Figure 14) were plotted for all scans (Figure 22), as they should represent the variation due to noise in the data.

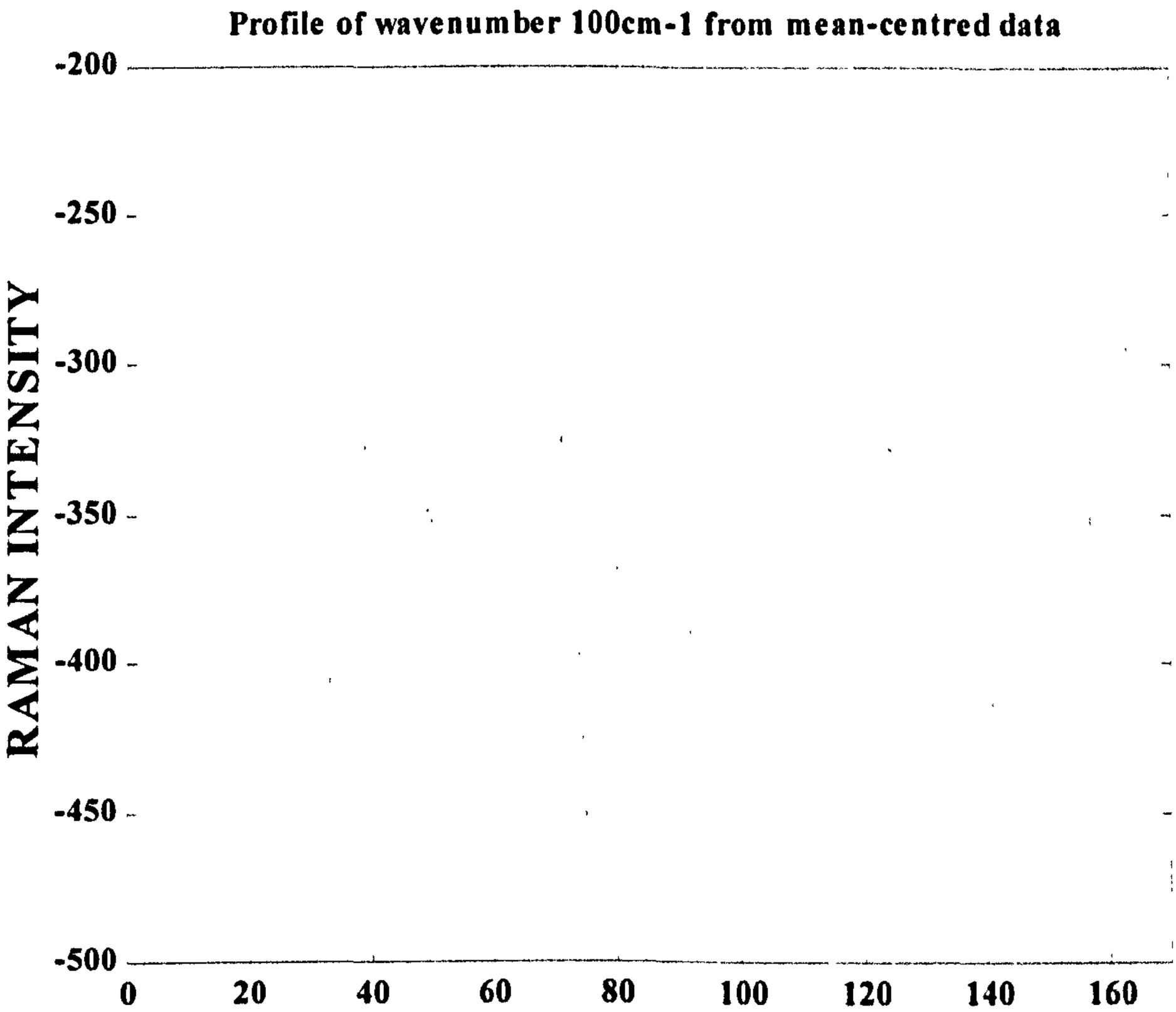

Figure 22: Profile of wavenumber 100cm^{-1} from the mean-centred data. This same profile is obtained for wavenumber 1850cm^{-1} at the opposite end of the spectrum. The similarity between this and the PC1 score plot gives an indication of the source of the PC1 data.

It is clear that Figure 22 is the exact replica of Figure 21, the PC1 score plot. Thus it is obvious that PC1 is only noise, specifically the baseline noise in the system. Since the fluorescence peaks form the major component, PC1 is largely the contribution from fluorescence, rather than the Raman data. This is a different situation from the usual case with PCA modelling of spectroscopic data, where the first principal component always contains the main chemical information. Therefore the main Raman data is represented by the rest of the principal components. This implies that if PC1 were removed from the data set, the leftover would be Raman data devoid of baseline shift and most of the noise. Figure 23 shows the nature of the data remaining after the contribution due to PC1 has been removed.

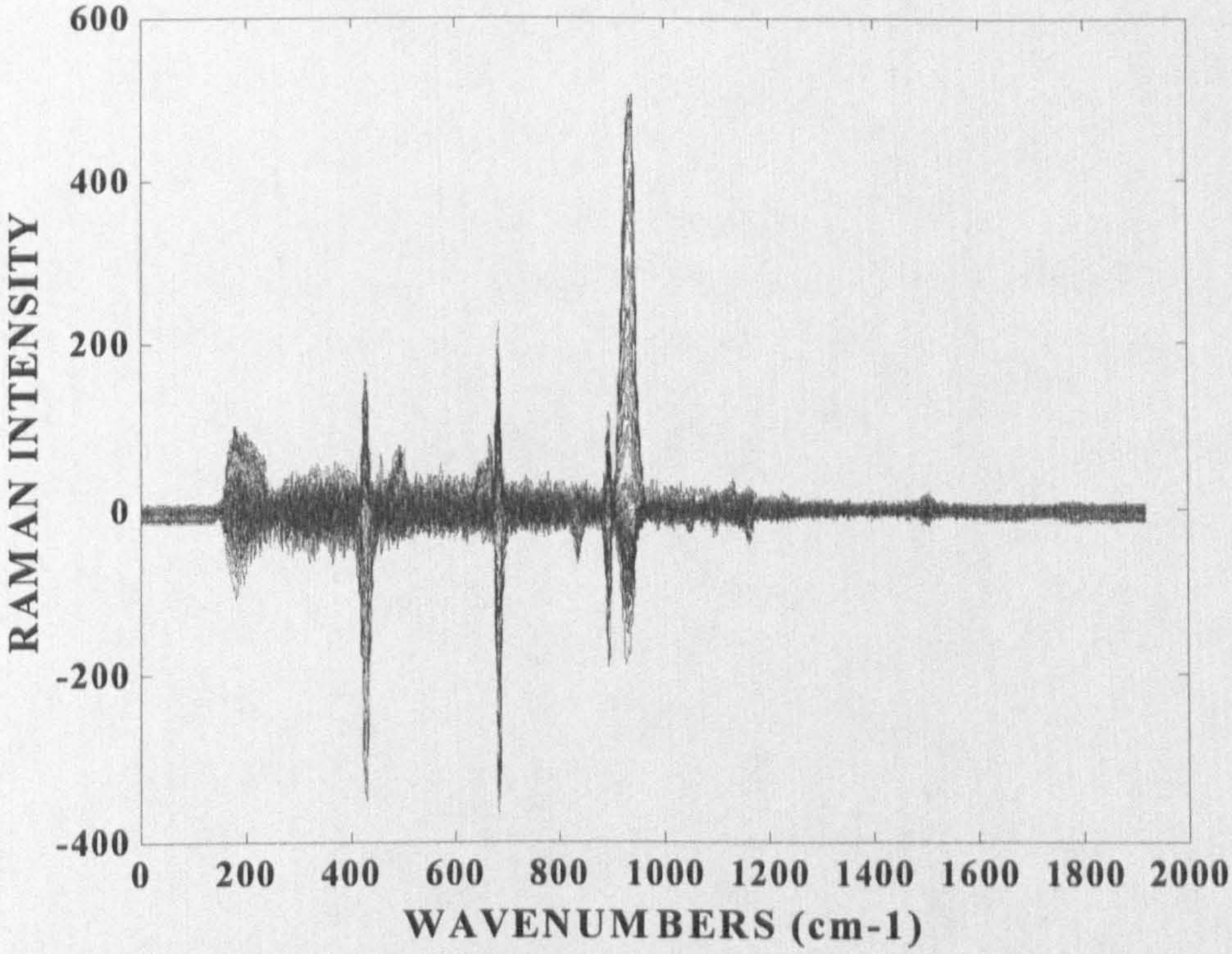

Figure 23: PC1 results: Result of removing the regenerated Raman data obtained from the variation captured by PC1, from the mean-centred Raman data

It can still be seen that most of the noise has been removed, and there appears to be a common well-defined baseline, which is about zero on the intensity axis because of the initial mean centering. The positions of the peaks are still maintained as in Figure14, confirming that this remaining data still contains the Raman information. However, the contribution of acetic acid is still not captured. It is most probably still overshadowed by the remaining noise still in the data.

With the expectation that the data has had the noise removed with the exclusion of the variation captured by PC1, the profiles of ethanol and ethyl acetate are mapped out once more, using the data shown in Figure 23. These profiles are shown in Figure 24. Compared with Figure 18, there appears to be little improvement in the quality of the data. The common regular undulation for both ethanol and ethyl acetate after 8.5 minutes is still present.

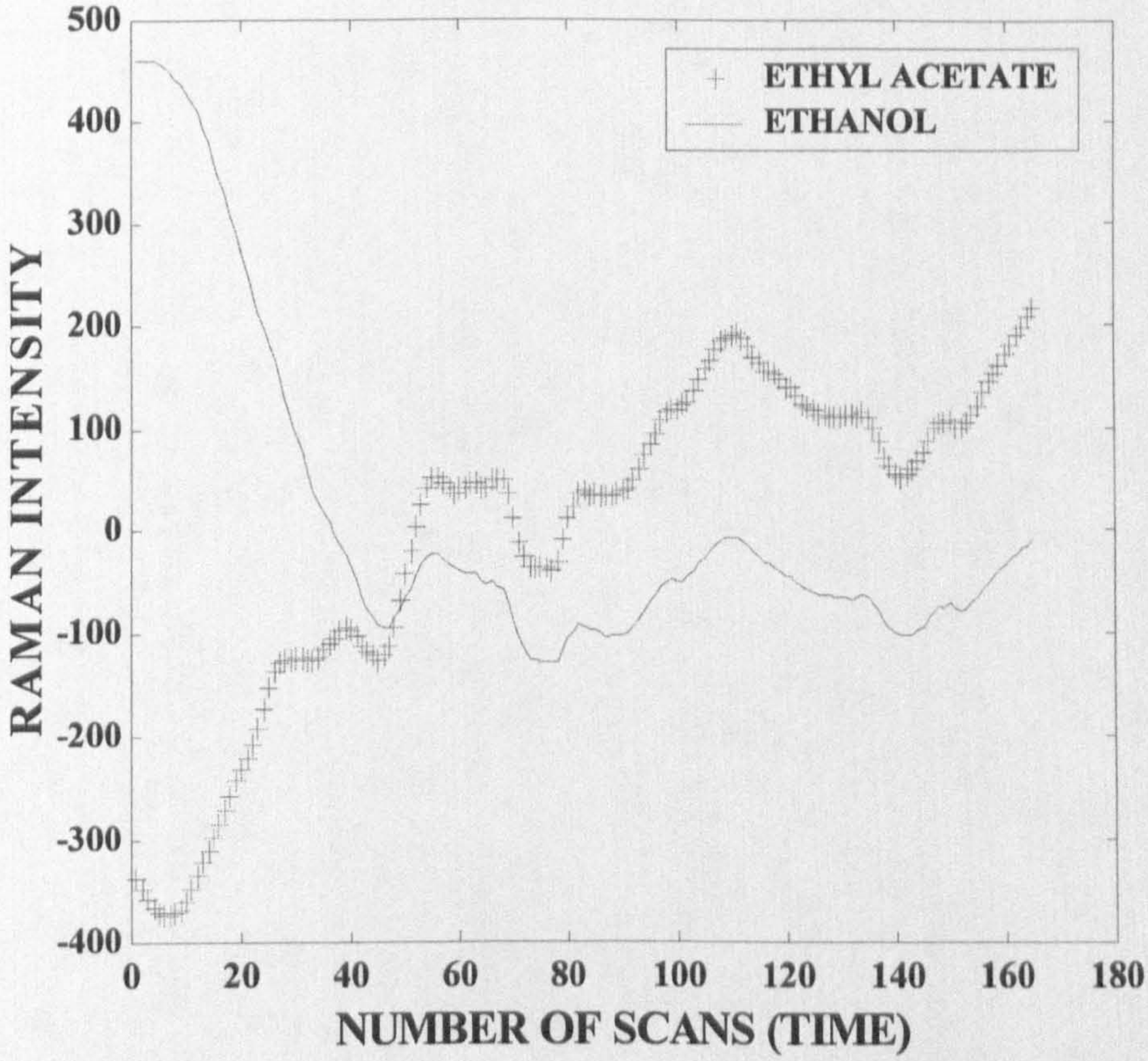

Figure 24: Profiles of ethyl acetate and ethanol during the esterification, using the data remaining after removing PC1. Data has been smoothed using the 15-point Savitsky Golay algorithm.

The noticeable difference is the magnitude of the Raman intensities, which drop from the 2050 – 1350 range down to the ~450 - 350 range. What we see is a removal of the fluorescence data that was very high in intensity, but no effect on the undulations due to the process dynamics, temperature fluctuations, etc., as the undulation is still present. It can therefore be inferred that the principal component(s) responsible for the variation due to the undulation is in one of PC's 2, 3, 4 or 5.

Since PC1 contains the fluorescence spectra, it follows that PC2 should contain the next greatest variation in the data, i.e. the pure Raman spectra. Thus the data regeneration stage was repeated, based on the variation in the second principal component, i.e. 0.55% of the total variation. The scores and loadings of the second principal component (PC2) were mathematically recombined to regenerate the data

based upon the variation in PC2. Figure 25 shows the regenerated Raman data using PC2 variation.

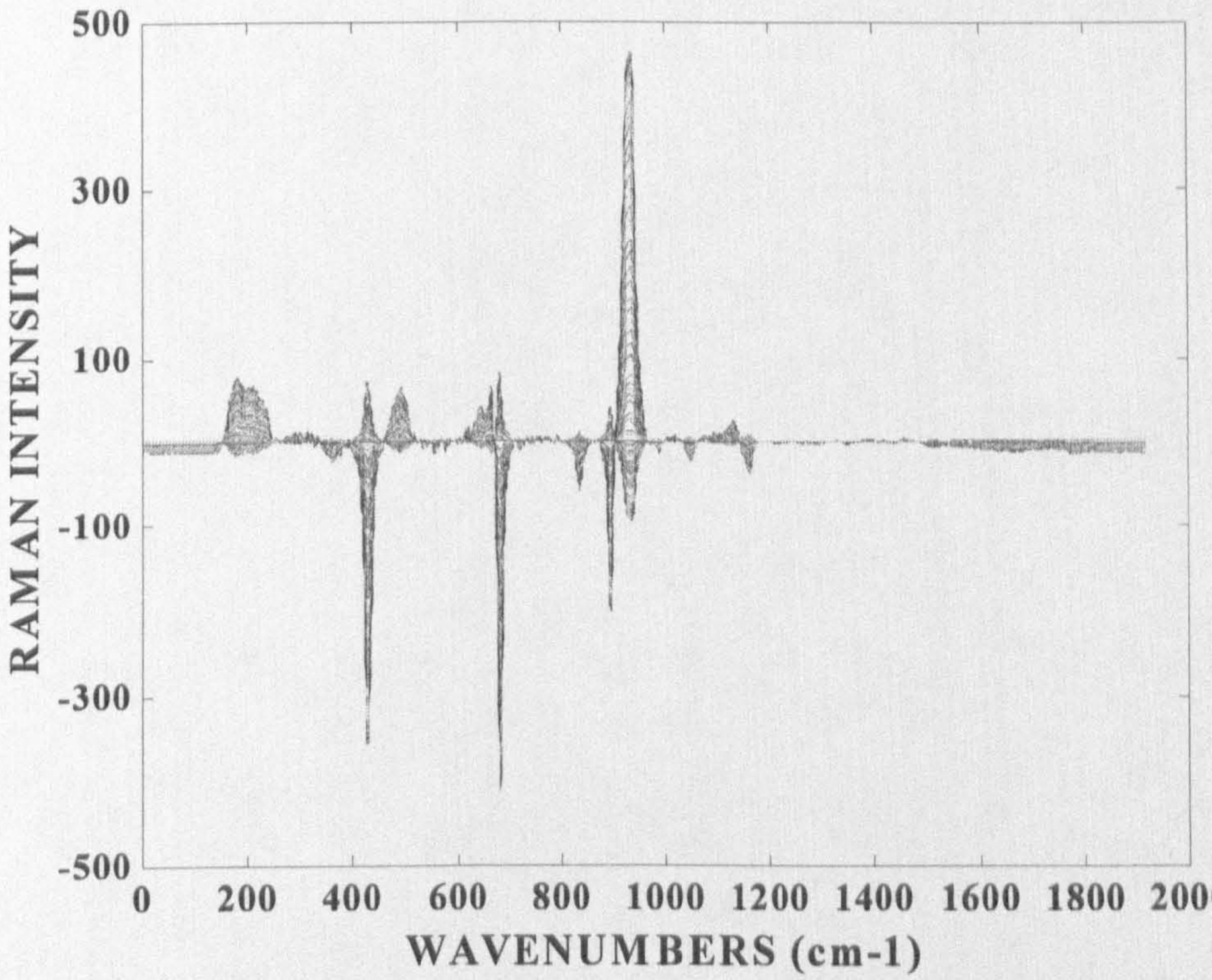

Figure 25: PC2 results: Regenerated Raman spectra based on variation captured in the second principal component.

From Figure 25, it is very obvious that all the noise has been removed, as the data now reflects the Raman spectra. There is a common well-defined baseline, which is about zero on the intensity axis because of the initial mean centering. In addition the acetic acid peaks at 180, 493, 670, and 943 cm^{-1} which were missing both in the original data and in PC1 are also captured, as are the peaks for ethanol and ethyl acetate. In comparison with Figure 23, this is a much more representative data set, with no noise at all and all the Raman data peaks standing out conspicuously. To determine the nature of the data reconstituted from variation captured by PC2, the scores of that principal component are observed by constructing a score plot. An examination of the scores for PC2 (Figure 26) shows the expected progress of the reaction from the beginning to the tapering end.

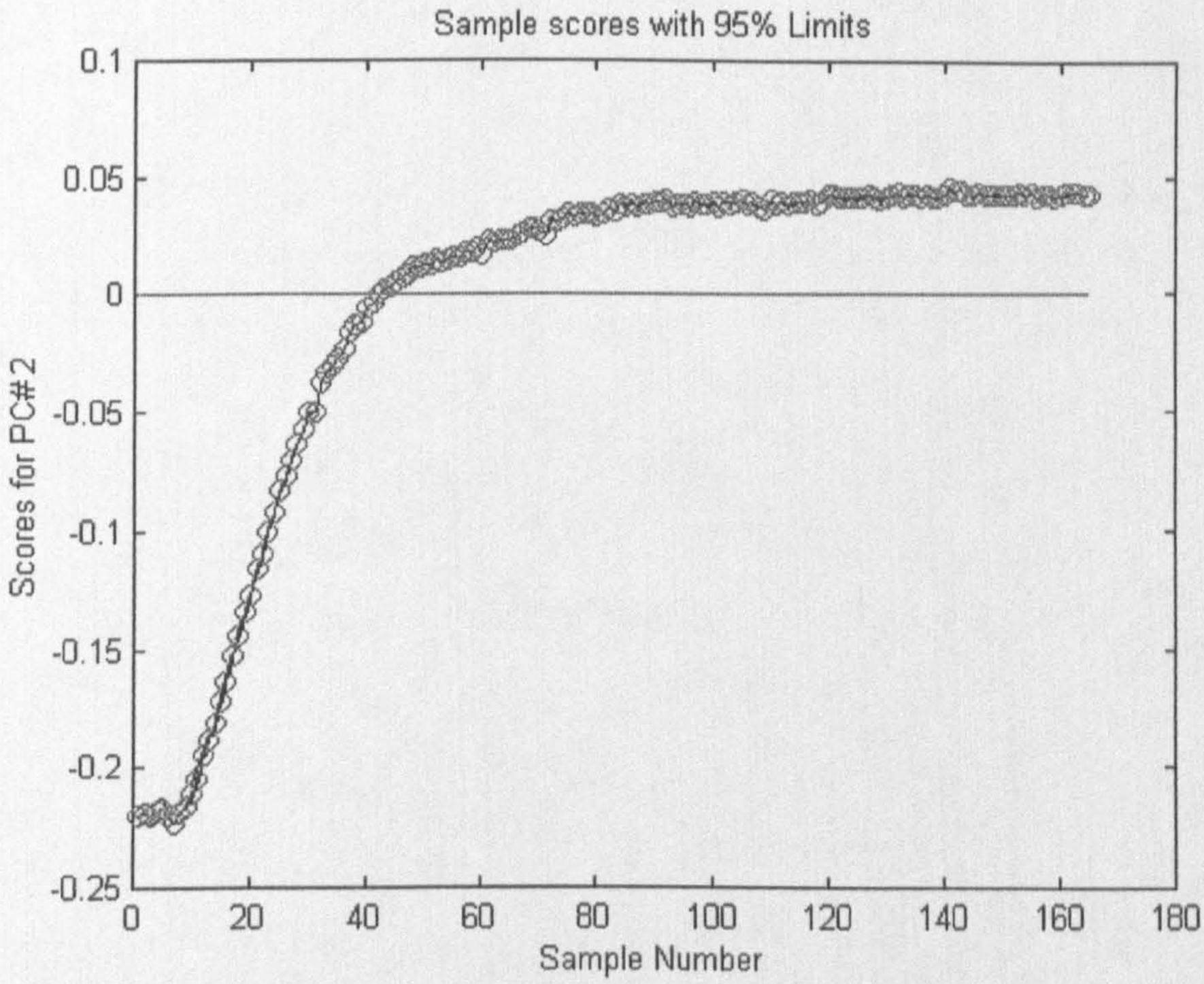

Figure 26: Sample number versus Scores for PC2. This shows the progress of the esterification reaction proper. This diagram is obtained directly from the PCA algorithm in PLS_Toolbox. Therefore, each sample number represents a single scan

The reaction progress can be seen to be 'static' until about scan 8, and then increase steadily at a fairly constant gradient until around scan 40, when the rate reduces until it reaches almost a constant value. When Figure 26 is compared with Figure 17, it is seen that the nature of the reaction follows the expected pattern. This confirms that PC2 contains largely pure Raman spectra coming from the esterification reaction.

Thus whilst PC1 captures the fluorescence data, PC2 captures only the Raman data. Again, the profiles of the peaks for ethanol, ethyl acetate and this time acetic acid are mapped out well in Figure 27.

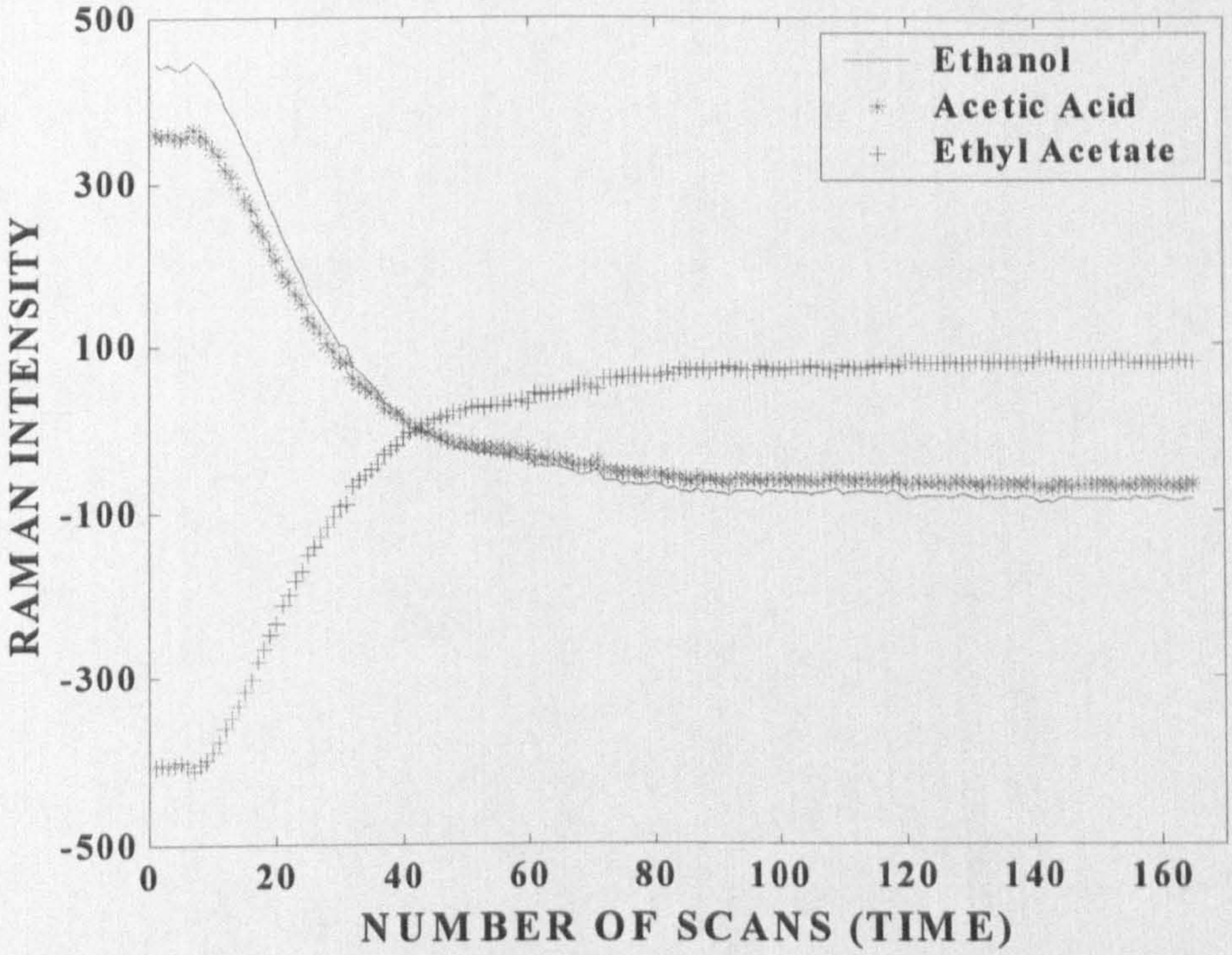

Figure 27: Profiles of the peaks for ethanol, ethyl acetate and acetic acid during the esterification, based on the variation captured in PC2.

Figure 27 is the desired reaction progress, showing the true account of the chemical process being monitored. All the noise and the undulations are removed, and the result is a perfect picture of a reversible reaction. From an initial starting plateau, ethanol and acetic acid concentrations decrease continuously while at the same time the ethyl acetate concentration increases. The gradients of both reactant and product profiles keep decreasing with time until it is almost zero, at which point the regeneration of both reactant and product is so minimal it is almost stopped. Comparison of Figure 27 with Figure 14 and Figure 18 demonstrates the progressive removal first of fluorescence noise and then noise due to the system dynamics, leaving behind a true reaction profile. In this way, principal components analysis has been used to filter out noise and purify spectroscopic data to give it more chemical meaning.

Thus we see the adverse effect of fluorescence spectra on Raman monitoring (Figure 14), and also the effectiveness of PC2 in selecting only the most relevant data (Figure 27). It is useful to know that whereas the data in Figures 18 and 24 were smoothed

(using Savitsky-Golay) for easier viewing, the data in Figure 27 has not undergone any such treatment. This shows how accurately PC2 has captured the Raman spectra.

3.1.2 Kinetics

Up to this stage, the reconstituted data due to variation captured in the second principal component has been observed to be of the nature of Raman spectra. Based on information from literature about the wavenumbers corresponding to the various reaction components, the reaction profiles have been plotted and have been seen to follow the trend expected of a typical reversible reaction. To theoretically validate the principal components results and so establish as a true representative of the process, either by comparison with concentration data (which is unavailable in this work) or by testing the reconstituted data against a well-known chemical theory. For the kinetic investigations the profile of ethyl acetate is selected as a product and that of ethanol as a reactant, using the PC2 data. In the esterification reaction, the amount of ethanol decreases while the amount of ethyl acetate increases with time. This trend is shown in Figure 28 for the period of scans 18 to 29, i.e. from 4.5 minutes to 7.25 minutes during the experiment. This is the period for which the reaction profile is straight in Figures 24, 26 and 27. Thus it has been shown that the reaction follows the expected reactant-product relationship.

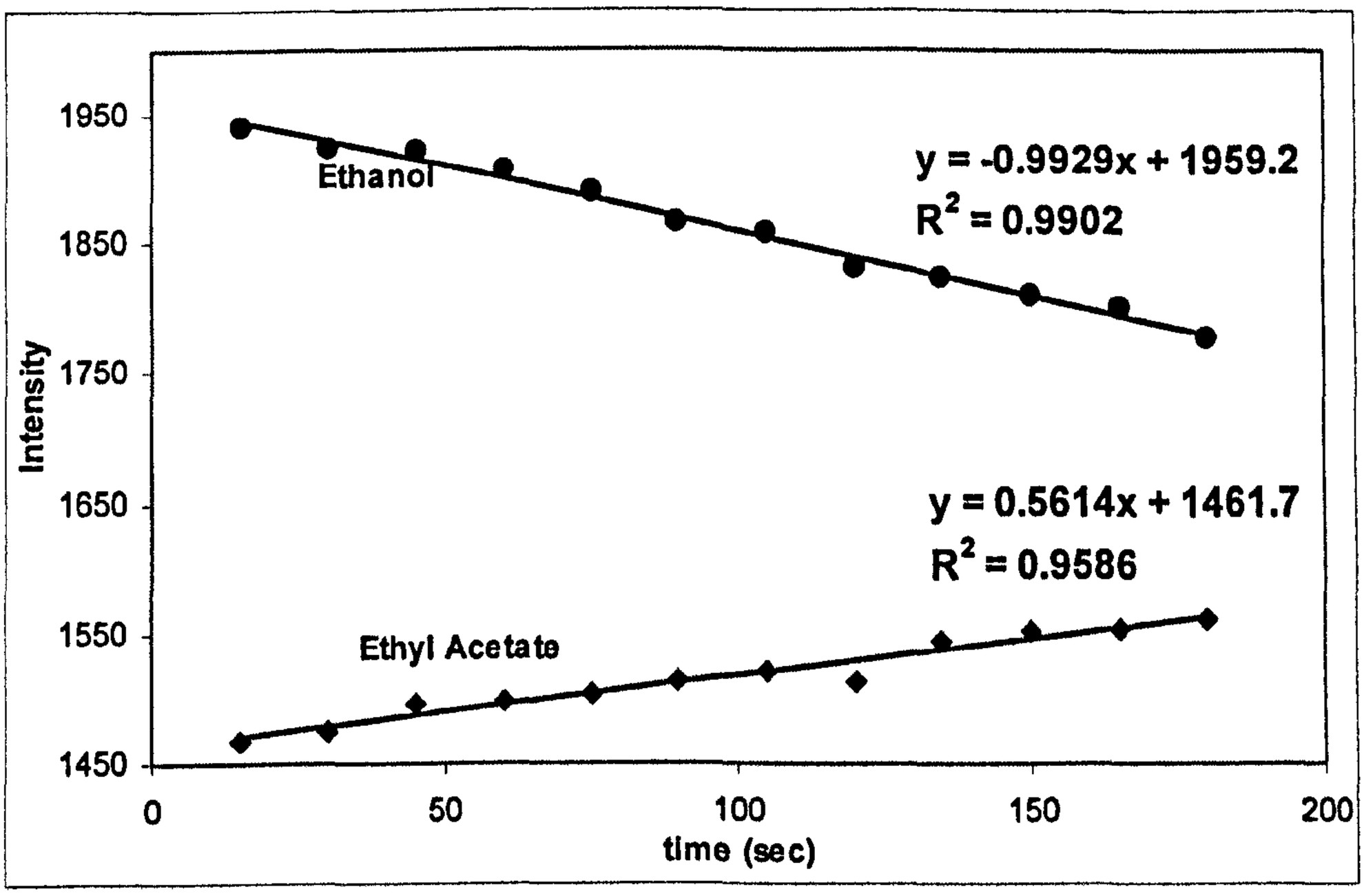

Figure 28: Reaction plots for ethanol and ethyl acetate during scans 18 to 29. This shows the amount of ethanol (reactant) decreasing while that of ethyl acetate (product) increases at the same time.

Hereafter the PC2 data is tested for compatibility with first order reaction kinetics, as discussed earlier on in the experimental section. The kinetic equation for a first order reaction is given as

$$In[A] = -kt + Constant \qquad (33),$$

where k is the velocity constant, t is the time in seconds and $[A]$ is the concentration or amount of the reaction component under analysis. Therefore a plot of $log_{10}[A]$ against t should give a straight line with a slope of $-k/2.303$, as a confirmation that the reaction is of the first order[60]. Thus the logarithms of the Raman intensities (representing the concentrations) of ethanol and ethyl acetate during scans 18 to 29, the region of highest reaction rate, are plotted against time as shown in Figure 29.

As expected, the kinetic plots are perfectly linear. The plot for ethyl acetate has the same gradient as the plot of ethanol, except in sign, i.e. whereas the ethanol gradient is –0.0002, that of ethyl acetate is +0.0002. This indicates that the rates of consumption of ethanol and production of ethyl acetate are the same.

Above all, Figure 29 agrees with equation 33, the relation for a first order reaction.

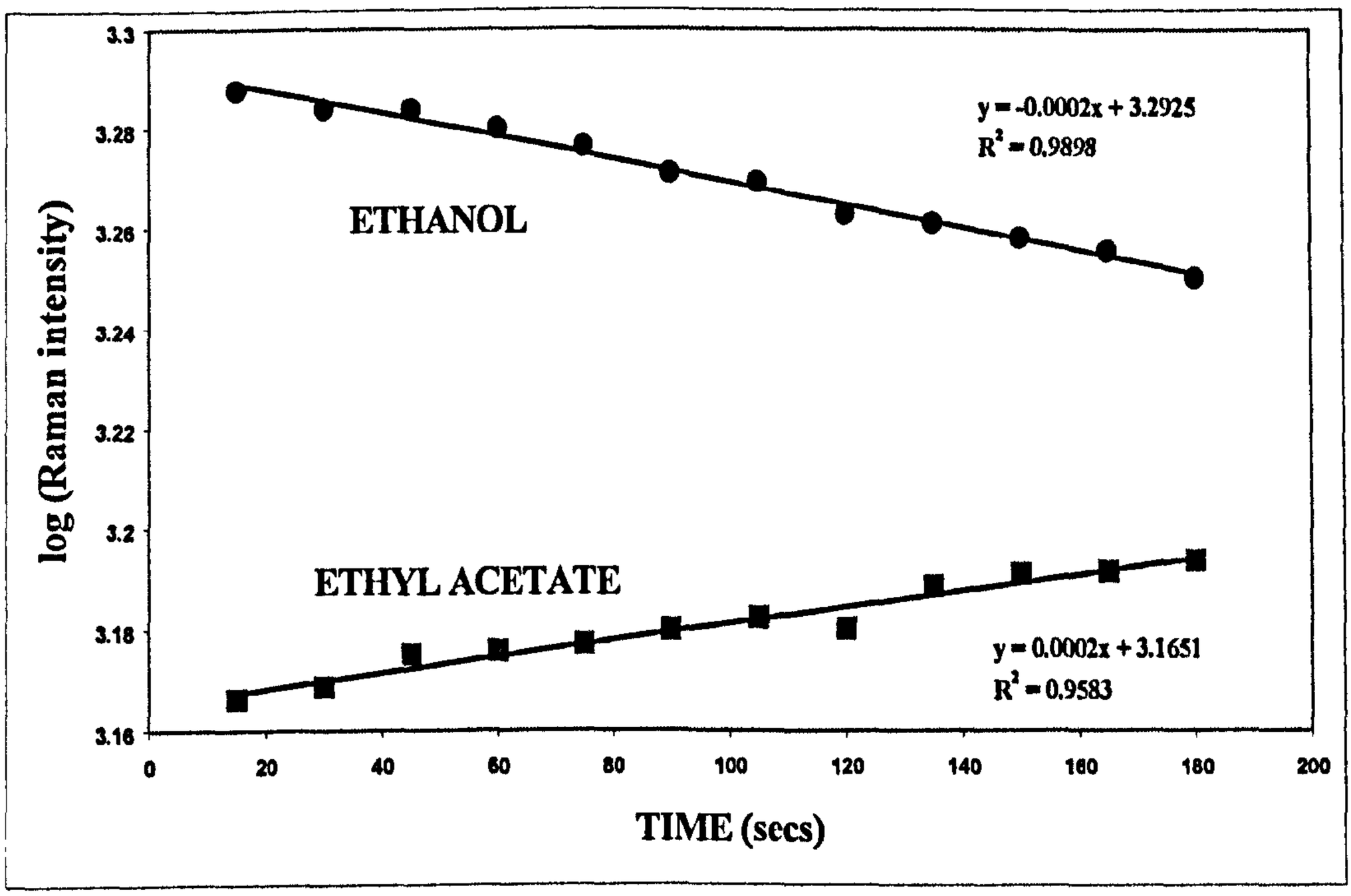

Figure 29: Kinetic plots for ethyl acetate and ethanol during scans 18 to 29. The plots of logarithms of the Raman intensities versus time give straight lines indicating compliance with first order reaction kinetics.

For ethanol, $-k/2.303$ is given by –0.0002, giving k a value of 4.6606×10^{-4}. For ethyl acetate, Equation 33 will take the form

$$In[A] = kt + Constant \qquad (34),$$

since ethyl acetate is a product and is released instead of being consumed. Therefore $k/2.303$ is given by 0.0002, giving k a value of 4.6606×10^{-4}.

3.1.3 Conclusion

Raman spectroscopy has been efficiently used to collect data on the esterification of ethanol and acetic acid. The data is very detailed, but contains massive noise as a result of fluorescence and dynamic effects such as a temperature fluctuation or the effect of stirring. This noise masks and/or distorts the Raman spectra, making it

difficult to observe the actual chemical process. Principal components analysis has been applied to the Raman data to remove noise. Fluorescence spectra was removed by PC1 whilst the second principal component contains all the pure Raman data, and clearly shows the progress of the reaction, totally devoid of noise and undulations due to the dynamics of the system which are contained in the subsequent principal components. This technique can be applied to other Raman data sets. However whether the Raman spectra will be described by PC1 or PC2 or whichever principal component depends on the degree of noise or fluorescence in the data. Thus with a less noisy data than that used in this work, PC1 may capture the pure Raman spectra, while in a noisier data set, the Raman data may be in PC3 for instance.

Principal Components Analysis has therefore been shown to be very useful in extracting useful information from raw data that is noise-ridden and difficult to explain. The fact that the data obtained from PC2 is actual Raman data has been confirmed by successful application of first order kinetic equations to the data.

3.2 Use of Raman Spectroscopy for process analysis – Experiment 2

In the first experiment, spectroscopic data from a chemical reaction was collected during the time the reaction was in progress. This data was analysed and then treated with principal components analysis to reveal the important chemical information after removing all noise from the data. The success of this exercise was proved by the fact that the regenerated data agreed with second order reaction kinetic laws.

It is worth noting that the batch process described in Chapter 2 was set up and performed under controlled conditions in the laboratory. Factors like volume and atmospheric temperature and pressure remained fairly constant throughout the experiment, and volumes involved were of the order of a few hundred millilitres. The reagents used were all of analytical grade and in a state of high purity. Desirable as these conditions are, they are different from what pertains on the industrial plant.

Following the success of monitoring the batch process with PC2 data, the next stage was to test the applicability of this technique to data from an industrial process that is characterised by large volumes of reagents in vessels that are subject to atmospheric conditions. Whereas the experimental work involved only four compounds (one of which is not Raman active), the real industrial data contained 17 components in various compositions that are sometimes of different orders. The difference in nature

between the two data sets emphasises the need to test the method that has worked perfectly within the confines of a laboratory under controlled conditions, on real life industrial data since that is where the method would ultimately be utilised.

The data from the industrial plant, described in detail in Chapter 2, is shown in the appendix. The appendix shows the various mixtures in the reaction vessel and how some of the samples have been spiked in order to enhance their responses.

Figure 30 shows the Raman data obtained from the processing of naphtha at the DF Plant, BP Chemicals Limited, Hull, UK, as described in Chapter 2.

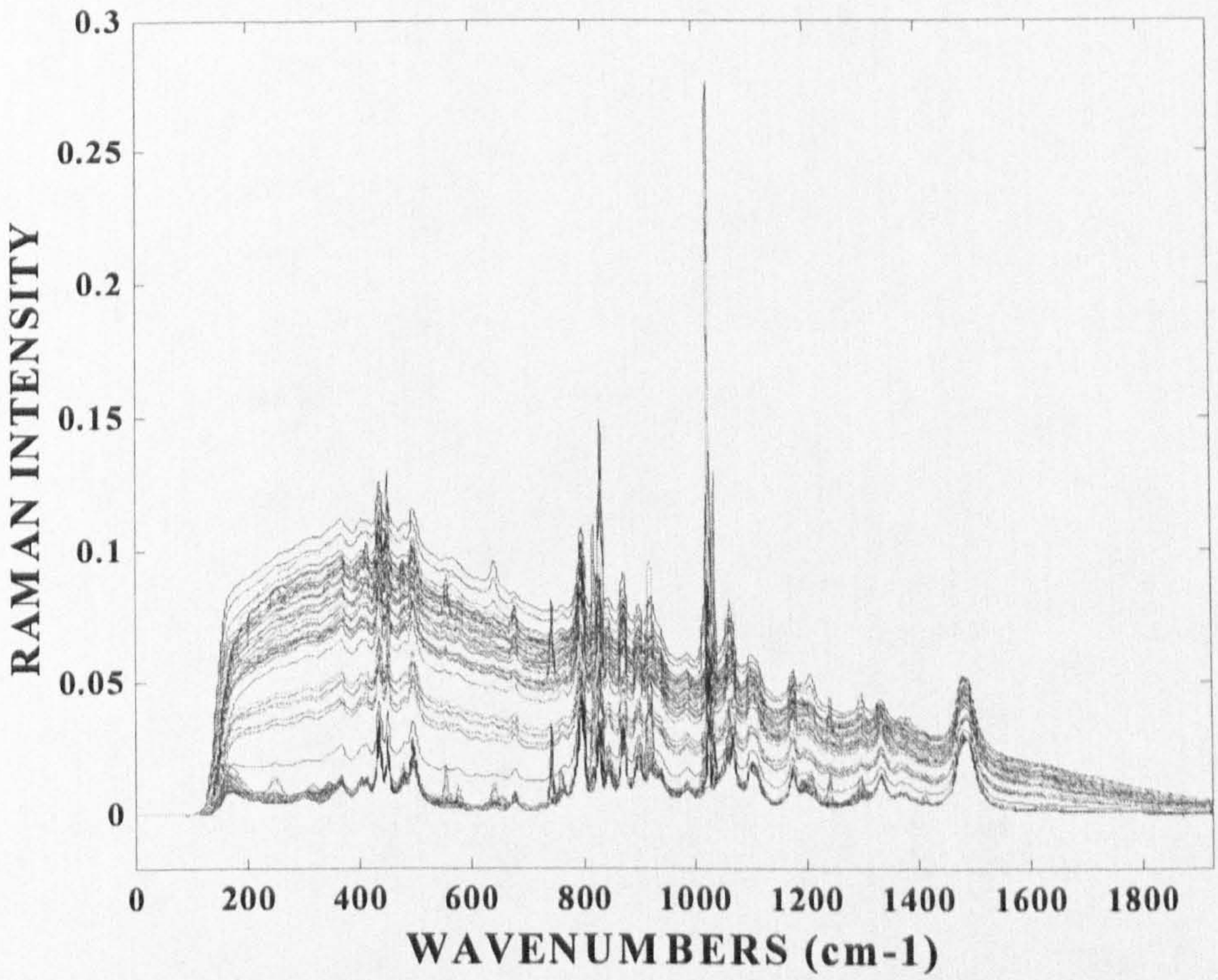

Fig. 30: Raman spectra of naphtha, made up of various levels of paraffin compounds, aromatic compounds, and then naphthenes, from the DF processing plant

The baseline is at zero intensity this time (as compared to Figure 14). The sharp peaks are registered where the samples have been spiked in order to enhance their responses. The peaks at 1037cm^{-1}, 1026 cm^{-1} and 833cm^{-1} stand out very conspicuously.

Figure 31 is a 'bird's eye view' sort of perspective of the DF naphtha data that shows the rise and fall of the peaks and the baseline very easily. The diagram is a height-sensitive colour map, therefore the various colours show the intensity of the Raman

spectrum at any particular point or area. The colours progress from blue-black to dark blue to light blue to yellow and then to red as Raman intensity increases. This makes the recognition of the highest and lowest intensities easy.

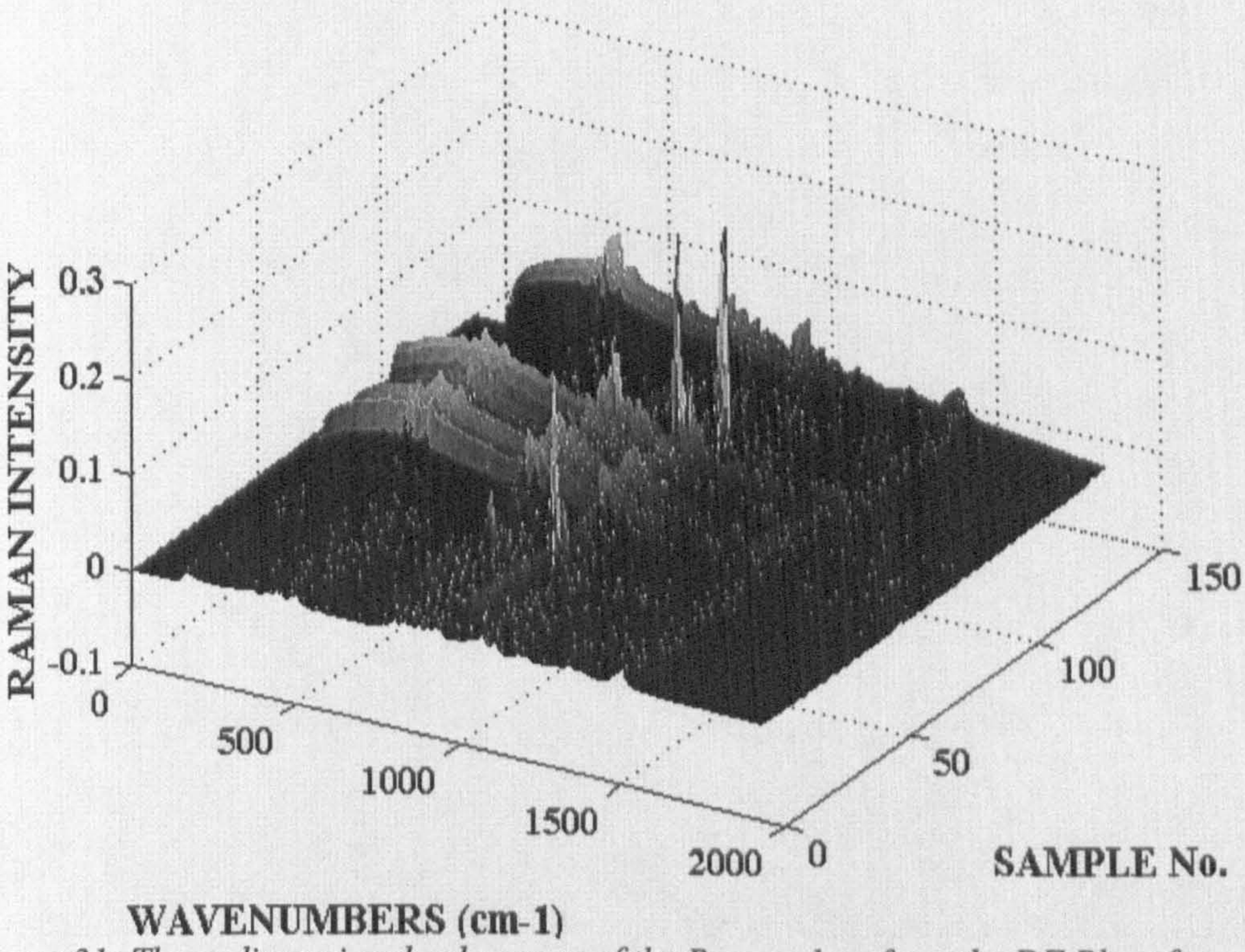

Figure 31: Three-dimensional colourmap of the Raman data from the DF Plant for naphtha processing. The highly spiked samples at wavenumbers 1037cm^{-1}, 1026 cm^{-1} and 833cm^{-1} stand out very conspicuously.

As a first step towards performing PCA, the data is mean-centred to enhance the subtle differences between the spectra. Figure 32 shows the mean-centred data that was used for analysis. The peaks in Figure 31 are all maintained in the same positions in Figure 32.

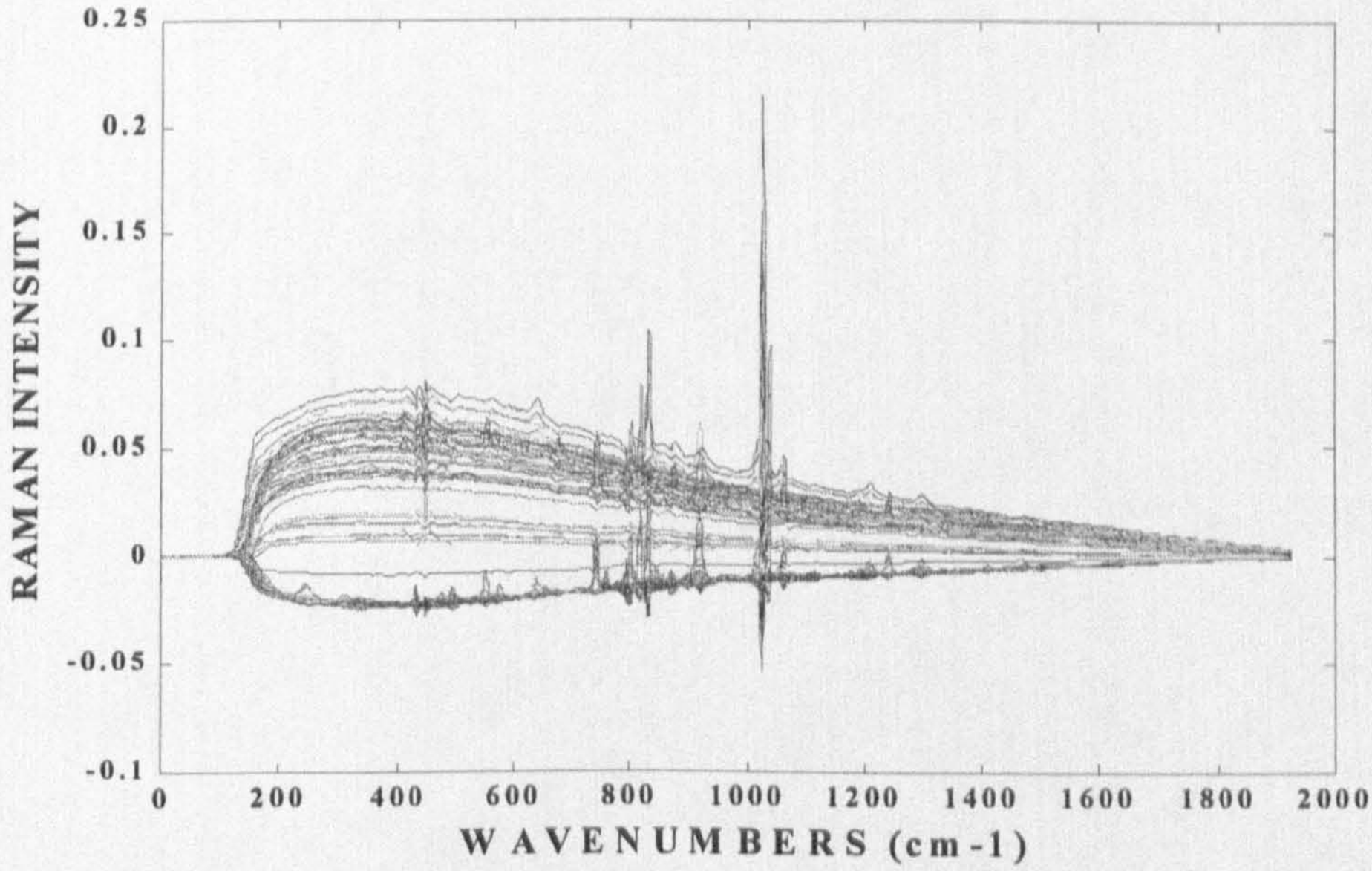

Fig. 32: Mean centred Raman data from naphtha processing plant

Eigen analysis is performed to determine the optimum number of principal components required to build the PCA model for the naphtha data. If the number of principal components used is too small, the model is inadequate and cannot represent the original data. Inclusion of too many principal components also leads to 'overfitting', since principal components that contain little information and mostly noise are also included in the model building exercise.

Figure 33 shows a plot of eigenvalues against number of principal components (pc's) for the mean-centred DF naphtha data. The huge difference in eigenvalue between PC1 and PC2 indicates that PC1 contains the greatest fraction of the total variation and therefore needs to be retained. Similarly PC2 and PC3, though with smaller differences, can also be retained. The eigenvalue difference between PC3 and PC4, and subsequent differences keep decreasing in comparison, but beyond PC9 there is no observable change in the eigenvalues. In order to avoid leaving out any relevant information in the PCA model building therefore, the first ten principal components were used to construct the model.

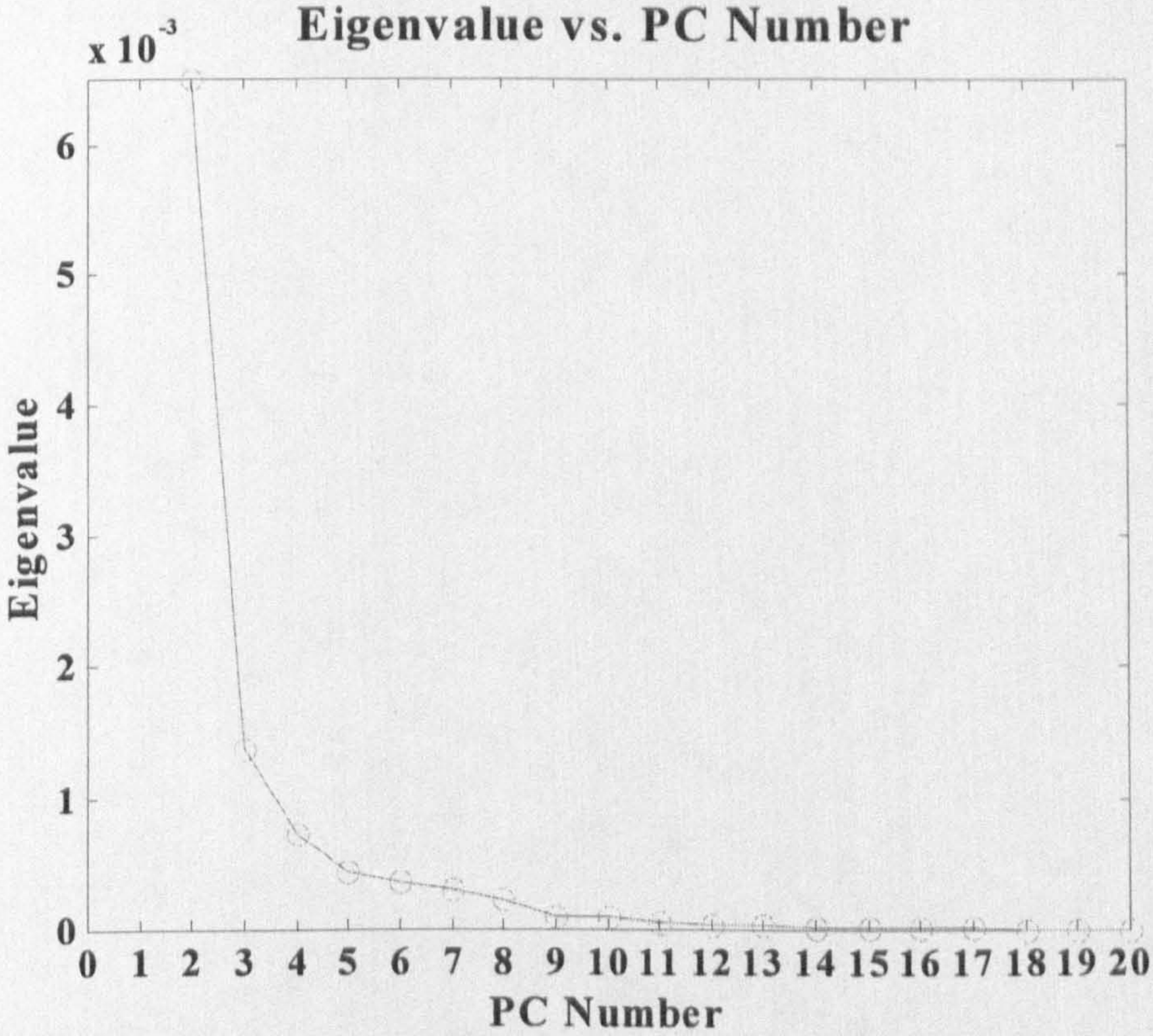

Figure 33: Plot of eigenvalue versus number of principal components for the Naphtha Raman data, used to determine the optimum number of principal components to use in the PCA model building

The result of the PCA modelling using 10 principal components is shown in Table 2. It is worth noting that before the PCA modelling, the data is mean centred. The eigenvalues of the first ten principal components supports the choice of ten principal components to build the PCA model, since after the 10^{th} principal component, there is little difference in the percentage variance captured by the subsequent principal components.

PRINCIPAL COMPONENT NUMBER	EIGENVALUE OF COV(X)	% VARIANCE CAPTURED BY THIS PC	% VARIANCE CAPTURED (TOTAL)
1	$7.47e^{-001}$	98.63	98.63
2	$6.51e^{-003}$	0.86	99.49
3	$1.38e^{-003}$	0.18	99.67
4	$7.25e^{-004}$	0.10	99.76
5	$4.37e^{-004}$	0.06	99.82
6	$3.70e^{-004}$	0.05	99.87
7	$3.09e^{-004}$	0.04	99.91
8	$2.25e^{-004}$	0.03	99.94
9	$9.52e^{-005}$	0.01	99.95
10	$9.28e^{-005}$	0.01	99.97
11	$5.67e^{-005}$	0.01	99.97
12	$4.27e^{-005}$	0.01	99.98
13	$3.25e^{-005}$	0.00	99.98
14	$2.52e^{-005}$	0.00	99.99
15	$2.37e^{-005}$	0.00	99.99
16	$1.66e^{-005}$	0.00	99.99
17	$1.13e^{-005}$	0.00	99.99
18	$8.92e^{-006}$	0.00	99.99
19	$7.42e^{-006}$	0.00	100.00
20	$5.35e^{-006}$	0.00	100.00

Table 2: Results showing percent variance captured by PCA model of naphtha data.

With the separation of the various groupings in the Raman data complete, the next stage is to reconstitute the Raman data using the values obtained from the PCA model. The scores and loadings of the first principal component (PC1) were mathematically recombined to regenerate the data based upon the variation in PC1. The resulting plot is shown in Figure 34.

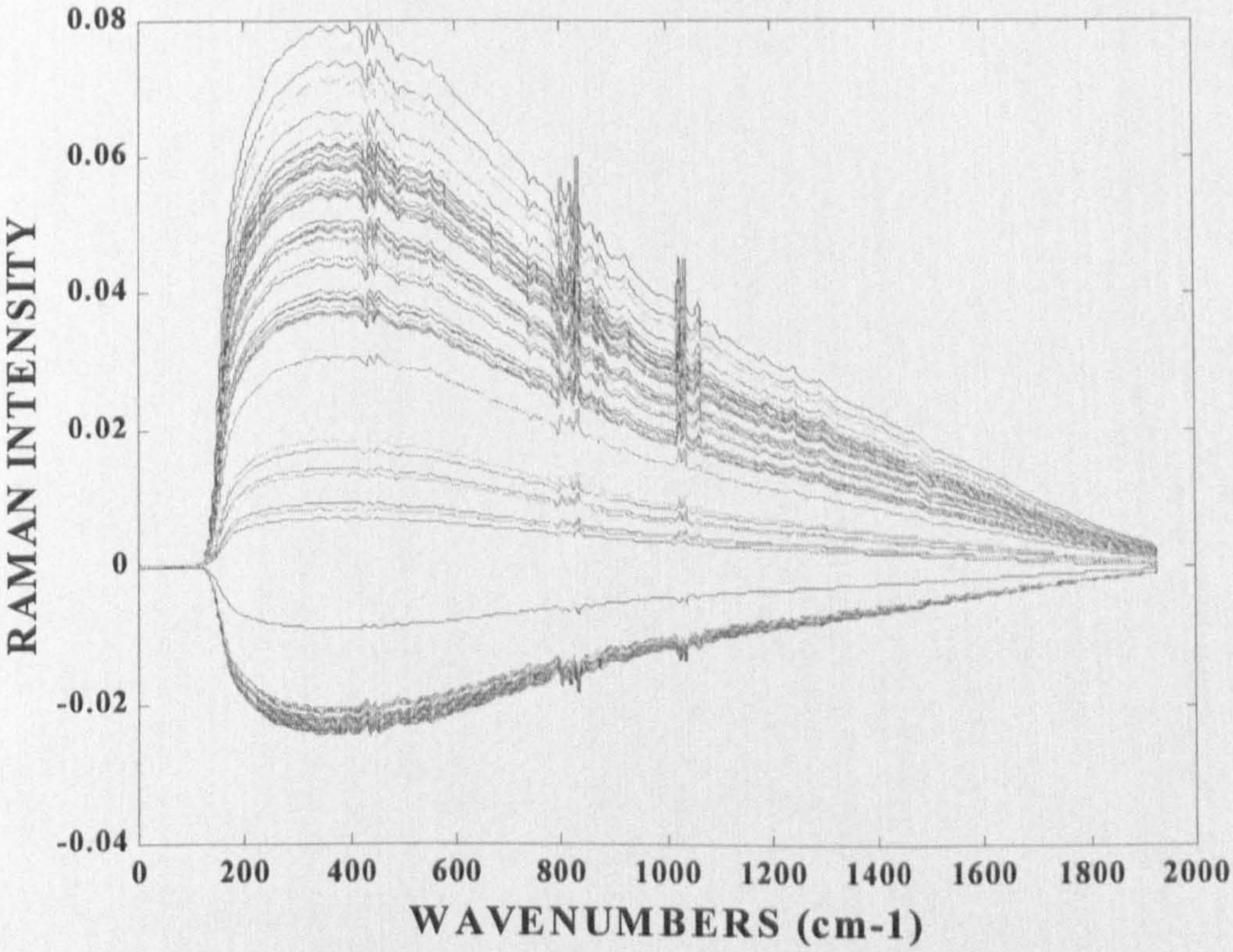

Figure 34: Regenerated naphtha data based on the variation captured in PC1

The information in Figure 34 looks very much like the mean-centred data in Figure 32. The intensities of both data sets are of the same order. This is to be expected since PC1 contains 98.63% of the variation in the data. However, the main spiked peaks are conspicuously absent. The information in Figure 34 is therefore a very important part of the data. To determine the nature of the data reconstituted from PC1, the scores of that principal component are observed. An observation of the scores on PC1 (shown in Figure 35) shows that the information here is definitely not that of the Raman spectra, considering the nature of the score plot.

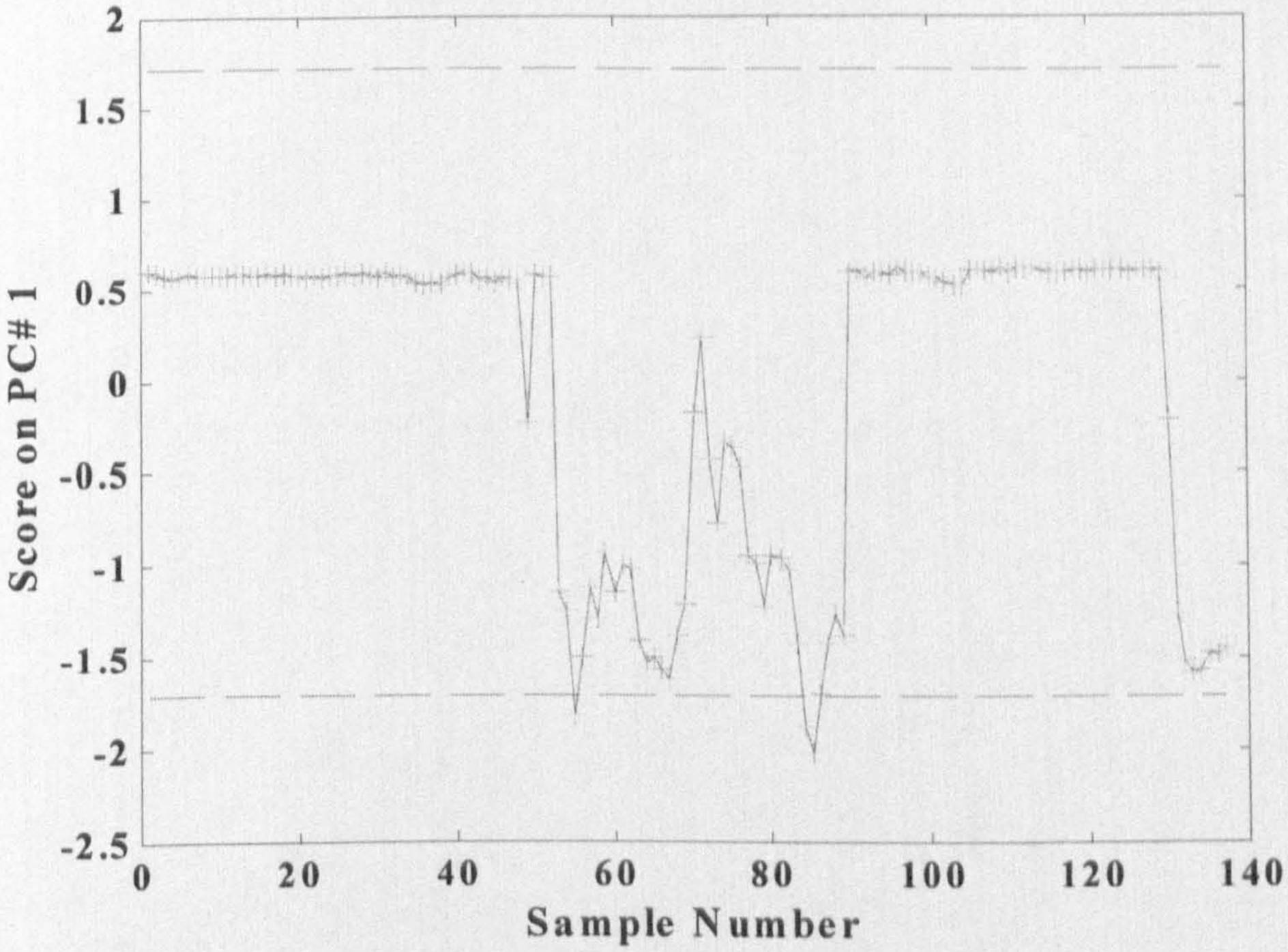

Figure 35: Sample number versus scores on PC 1. This diagram is obtained directly from the PCA algorithm in PLS_Toolbox. Here, each sample number represents a single scan.

An observation of Figure 35 shows a smooth progression broken by huge spikes at certain sample numbers. When this figure is analysed against the data in the appendix showing the nature of the 137 samples, it is seen that the smooth (straight) part of the graph represents those samples that do not show any laser-induced fluorescence, i.e. samples 1 to 48, 50 to 52 and then 90 to 129. These are all Raman spectra only. The protruding peaks are for samples that exhibit laser induced fluorescence (LIF). The first sample to '*jump*' out is sample number 49 from the tank blend, showing LIF. Then samples 53 to 77 which are not spiked and which exhibit LIF are in one group and distinctly separate from the purely Raman samples. Next come samples 78 to 89 which are spiked with various amounts of various components of the naphtha mixture and exhibit LIF. Then finally the installation tank blend samples, 130 to 137, which show LIF also stand out in one group well away from the non-LIF samples. The various groupings in the data according to exhibition of LIF is shown in Figure 36.

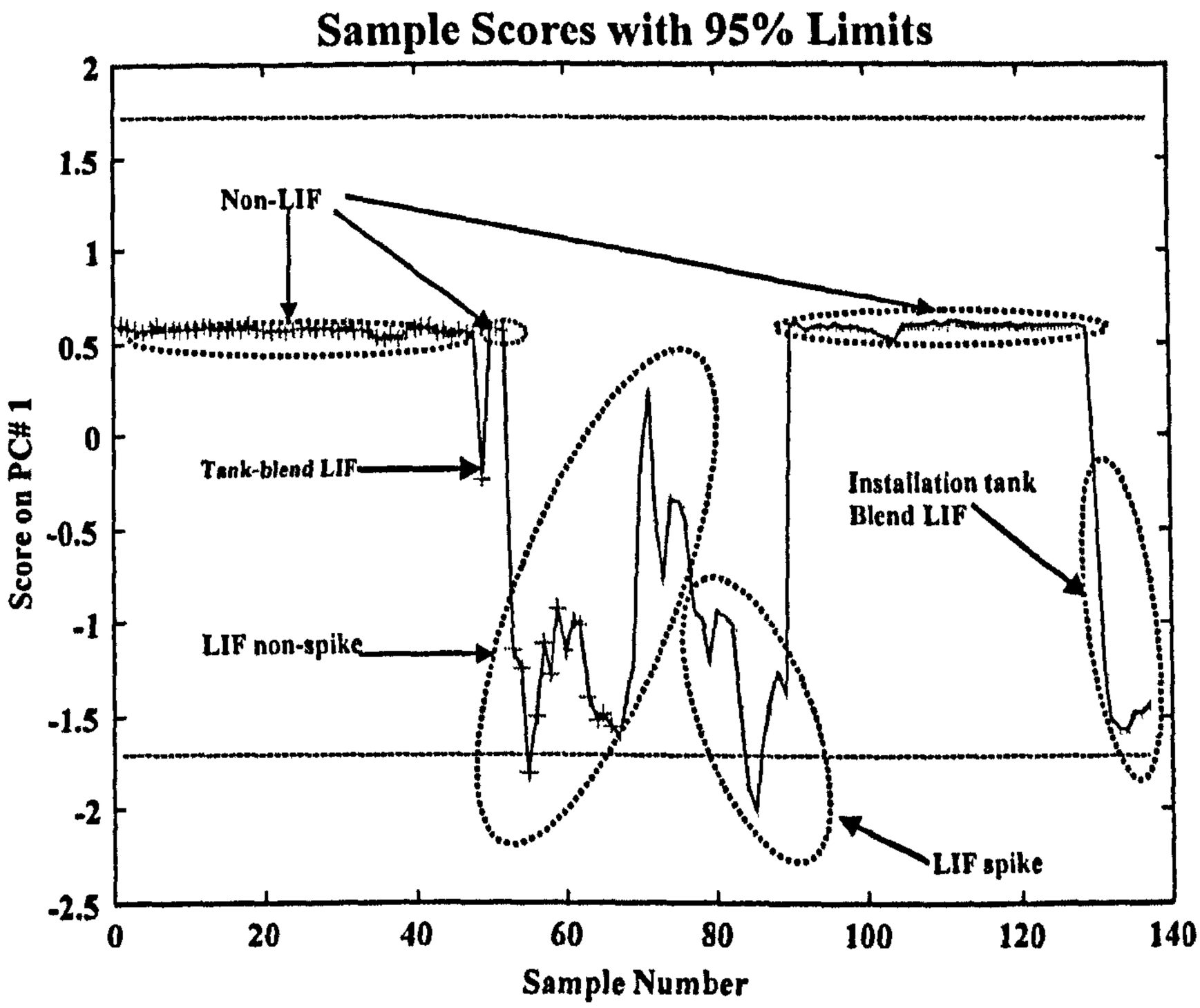

Figure 36: Sample number versus scores on PC1. This shows the various groupings in the data according to whether the sample shows purely Raman spectra or is affected by laser-induced fluorescence (LIF). This feature makes PCA a very useful and yet simple tool in chemical process control, since at a glance any deviation from normal behaviour can be detected. The groupings here agree entirely with the information on the data as given in the appendix.

This feature from the score plot of the first principal component makes PCA a very valuable and yet simple tool for chemical process control. Displayed on a control panel, this score plot easily shows when the reaction is giving good results and when the results are becoming undesirable, which samples give the undesirable results and consequently the location and time of the fault. Since the occurrence of fluorescence is detrimental to any analytical use of Raman spectroscopy, this method of separating the samples affected by fluorescence is very useful and valuable. The control analyst therefore has cause to determine whether the LIF samples come from one particular location in the plant or undergo one form of influence or another that is not experienced by the other samples that do not show LIF. In addition, like any control chart Figure 36 shows which samples are acceptable and which ones are definitely out of the acceptable range. So that samples 55 and 85 in Figure 36 fall out of the 95%

limits, although from the information in the appendix there is no clear difference between them and their nearest neighbours. This gives the opportunity to undertake close and specific analysis of these two samples to establish the cause of their deviation.

As a method of chemical process control, the results shown in Figure 36 is similar to that obtained in Figure 26 showing the progress of the esterification reaction. In Figure 26, the stages where the reaction is most active and where the reaction is slow or almost stagnant are clearly obvious. Therefore it has been shown that principal components analysis can be used to control both a simple chemical reaction performed in the laboratory under controlled conditions and with a small number of reaction components as well as to control an industrial scale chemical process in a chemical plant under varying conditions and constituted of many different reaction components. This is made possible by the ability of PCA to separate data into groups according to the effects influencing them.

The data regeneration process is repeated, this time using the variation captured by the second principal component, since PC2 has the next most important information in the data. The result (Figure 37) is a well-filtered Raman data set when compared to the original data (Figure 30).

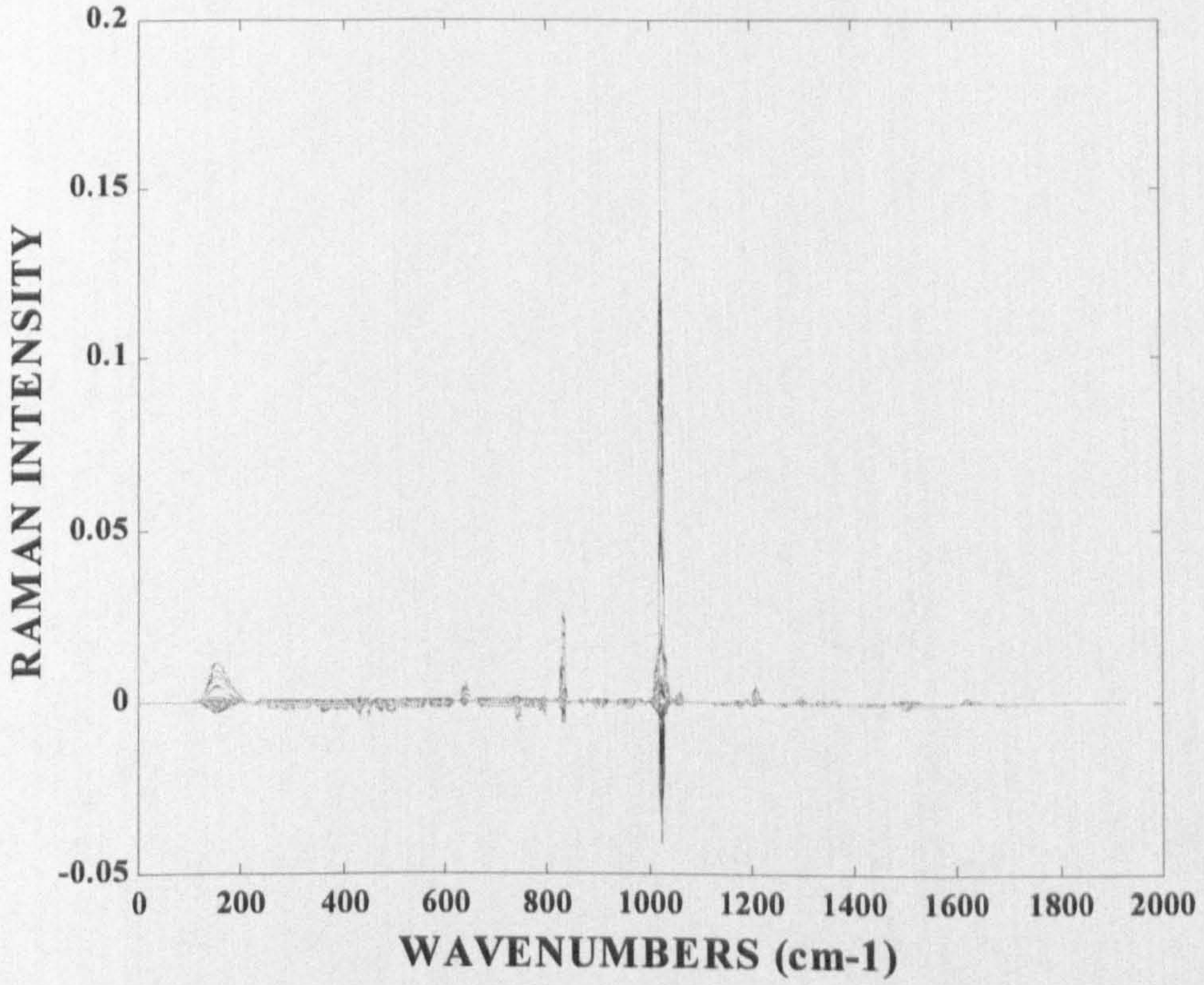

Figure 37: Regenerated naphtha data based on the variation captured in PC2. This is the expected nature of the Raman spectra.

To determine the nature of the data reconstituted from variation captured by PC2, the scores of that principal component are observed by constructing a score plot. An examination of the scores for PC2 (Figure 38) shows the expected progress of the naphtha monitoring process.

The score plot is the exact nature of how the Raman spectra are expected to look like. From the appendix and also from Section 2 of Chapter 2 (2.2), the samples are monitored for some time and then occasionally spiked with one or more of the components making up the naphtha mixture. This makes the Raman intensity of that sample rise sharply over and above the rest. In fact, these peaks are so high in intensity that they fall outside the limits of the model (shown in Figure 38 by dotted lines). Thus Figure 38 shows that PC2 is definitely a proper representation of the naphtha Raman data.

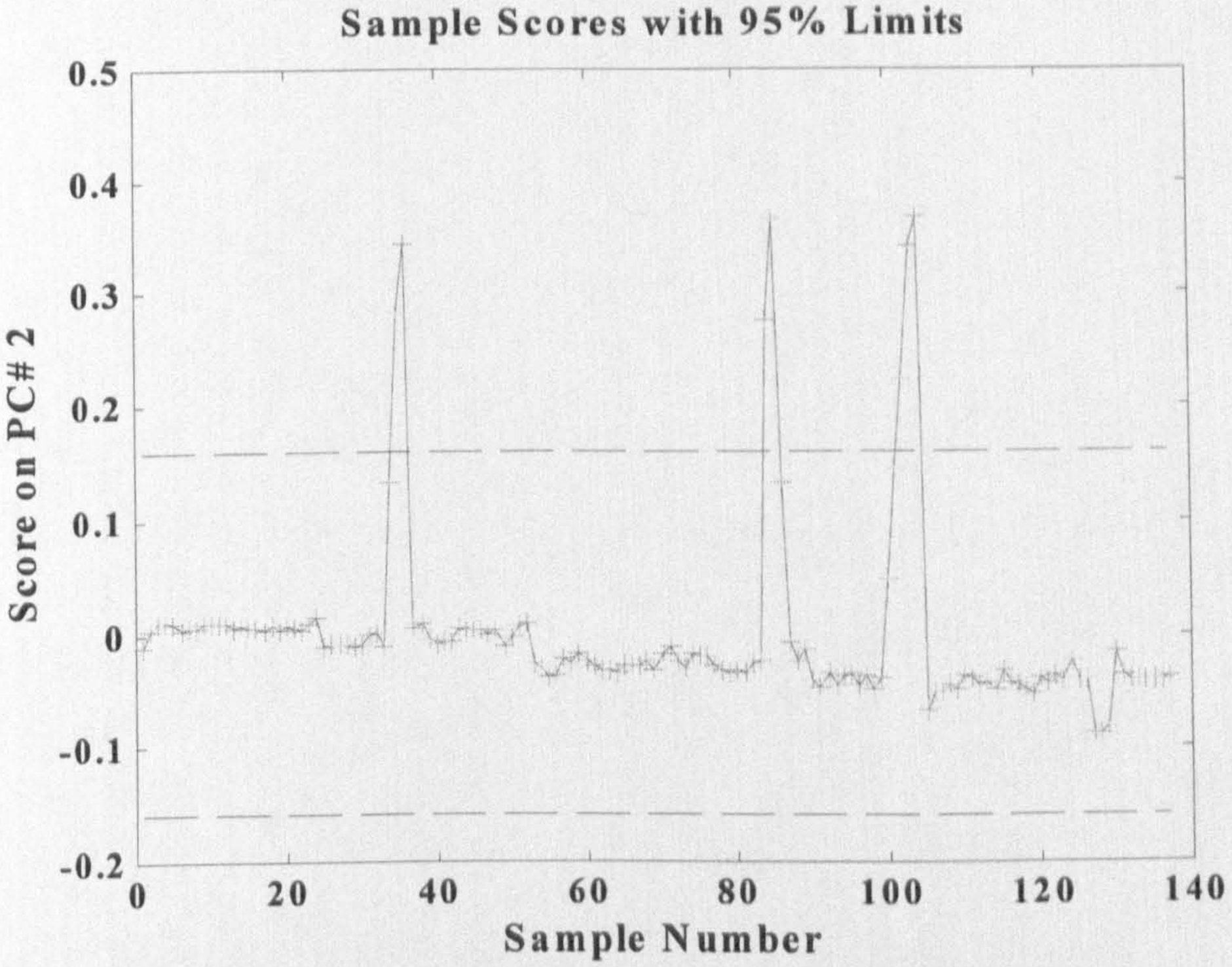

Figure 38: Score plot for naphtha data reconstituted from the variation captured by the second principal component. This profile looks very much like the actual chemical information in the Raman data, as the peaks where the samples were spiked stand out clearly out of the rest of the spectra that is relatively of very low intensity.

It is noted that in the monitoring of the naphtha data, the aim is to check the types of components that emerge as heads (or tails). Thus what is expected are profiles of the various samples showing the different components present each time, a reasonably different aim from that of the monitoring of the esterification reaction.

The peaks observed in Figure 38 occur at wavenumbers 1025, 833, 640 and 151 cm^{-1}. The profiles of the progression of these peaks within the reaction time are shown in Figure 39.

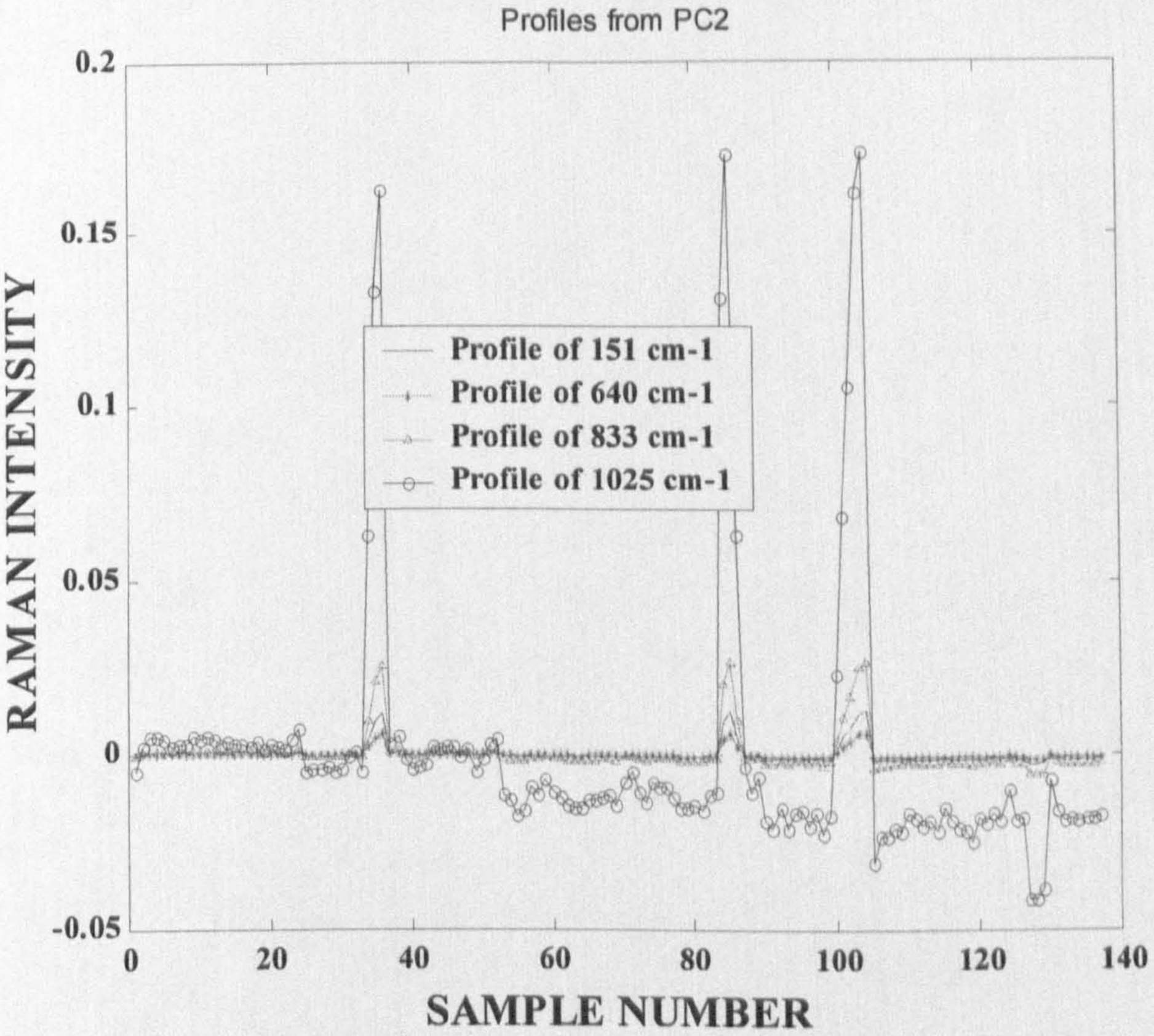

Figure 39: Profiles of peaks in PC2 data.

These are actually the change in concentration of one particular component of the continuous reaction mixture, and not a reaction profile such as that obtained in the batch reaction. This means that the system of observing the profiles of prominent wavenumbers (as used in the esterification) would be of little help if applied here.

Whereas the batch reaction has a starting concentration that continues to diminish or increase steadily as the reaction goes on and can therefore be mapped, the initial concentration in the continuous process keeps changing each time a new batch is detected. The trajectory of a reaction component is therefore not representative of the reaction profile.

This can be explained by the following:

- Unlike the esterification, which is a batch process, the Naphtha data comes from a continuous process. Thus instead of reactants starting from a high concentration level and gradually diminishing (and being regenerated), the reaction components are fed in afresh each time and in a continuous manner. Thus it is not possible to kinetically follow the profile of a reaction component.
- While monitoring the batch reaction, definite concentration values of the individual components are not required since the bulk volume does not change and quantities can be comfortably expressed as fractions of the initial concentrations. In the case of the continuous Naphtha process, actual concentration values are needed to correspond to the (Raman) spectroscopic data.

Thus the system of monitoring used in the esterification cannot be applied to the naphtha processing data. The most suitable method of analysis and evaluation therefore is to construct a model that would predict the concentrations of each reaction component such that the difference between the actual and predicted concentrations from spectral data would be minimal. Then subsequent concentrations of various components can easily be predicted from their Raman spectra. To do this would need a set of reference data made of a direct measurement of the same samples at the same time but using a different method that is tried and tested. This is the gas chromatograph data (137 samples by 17 components).

The data from the Raman spectra of the naphtha data was thus used to build calibration models for prediction of the concentration of each of 17 components making up the naphtha mix.

3.2.1 PLS Results

The mean-centred data was split into a training set and a validation set. These were then used to construct and validate a PLS model. The criteria for assessing the model were the correlation coefficient of the plot of predicted versus actual variables, and then the Root Mean Square Error of Validation (RMSEV) value. For good prediction of a set of real actual variables, the correlation coefficient must be as close to unity as possible. A perfectly correlated pair of data is one that has a correlation coefficient of value 1. Because the correlation coefficient is a ratio, its value lies between zero and one, one inclusive, i.e. $0 < C \leq 1$. The closer the correlation coefficient is to zero, the

poorer the correlation between the actual and predicted. The other criterion, the RMSEV, is a measure of the difference between the actual data and the predicted data. Therefore the smaller the value of the RMSEV, the better the prediction. The aim is therefore towards a smaller RMSEV and a correlation coefficient that is as close to unity as possible. The models were constructed using the original data and then the denoised data (i.e. data reconstituted from the variation captured by the second principal component). Table 3 shows the PLS results.

Data	Correlation coefficient from Predicted vrs Actual plot			RMSEV
	Model 1	Model 2	Model 3	
Original	0.9928	0.9955	0.9955	0.2138
Denoised	0.3207	0.2540	0.2180	2.3702

Table 3: Results of PLS modelling of original and denoised Raman naphtha data. This shows poorer prediction for the denoised data than for the original data.

It is clear from Table 3 that removing the noise in the data reconstituted by PCA makes the prediction by PLS rather worse in all cases. In all three models, the correlation coefficients are far closer to zero and the RMSEV is very large for the denoised data. The error is very high and the correlation between the original and predicted is practically non-existent. Therefore a better method is required.

3.2.2 GRAPE Results

The generalised randomised press-based elimination (GRAPE) algorithm randomly selects various variables to build the model and rejects those that do not contribute to reducing the PRESS. The final set of variables for which there is no decrease in PRESS is used to build the new data. Thus this variable selection method removes all the noisy data and retains the most important for model building.

To check the effect of the variable selection method GRAPE, a comparison is made between data modelled by MLR and then PLS first without variable selection (i.e. using the whole data set) and then with variable selection.

The result of MLR and PLS modelling without variable selection is shown in Table 4.

Method	*RMSECV*	N_{lv} / N_{hidden}
MLR	9.4624	
PLS	2.3592	49

Table 4: Results of modelling without variable selection. This shows a better result by PLS than by MLR modelling, though on the whole the RMSECV values are too high.

Here, the RMSECV values are undesirably high, especially for the MLR model. The values indicate a wide difference between the predicted and actual data since they are far from zero. Therefore variable selection experiments are carried out with $N_{max} = 50$ and $N_{max} = 75$. GRAPE and a genetic algorithm are applied to the data to create models, as shown in Table 5.

	GRAPE			GA		
	RMSECV	$\sigma(RMSECV)$	$N_{selected}$	*RMSECV*	$\sigma(RMSECV)$	$N_{selected}$
MLR (50)	1.1277	0.0790	45	1.0977	0.0649	44
MLR (75)	1.0649	0.0669	45	2.0064	1.2749	39

Table 5: Results of variable selection experiments.

Comparing Tables 3 & 4, there is a vast improvement in the RMSECV values with the use of GRAPE and GA. However, GRAPE is better than GA, having lower levels of error. When both GA and GRAPE are compared with the PLS result (Table 3), we see that normal PLS is more effective than both of these methods, though that result was obtained using 60 latent variables to build the PLS model. The advantage of the GA approach here is that it requires fewer number of selected variables.

A graphical presentation of these results is shown in the form of box-and-whisker plots in Figure 40. A box and whisker plot are produced for each column of the data. The box has lines at the lower quartile, median, and upper quartile values. The

whiskers are lines extending from each end of the box to show the extent of the rest of the data. The data with values beyond the ends of the whiskers are outliers.

Thus for the RMSECV plot, the most effective is GRAPE using N_{max}=75, because its corresponding boxplot has very short whiskers and a box with the smallest range. The use of the GA using N_{max}=75 had the widest box range and a very prominent outlier, and is therefore the least reliable. This observation is mirrored and enhanced in the $N_{selected}$ plot that shows the number of variables used for building the model. The plot for GRAPE with N_{max}=50 has the widest box range and prominent whiskers. This gives a wide range for $N_{selected}$. GRAPE with N_{max}=75 has a smaller box range with less prominent whiskers. The GA with N_{max}=50 has the smallest box range while the GA with N_{max}=75 has a wide box range, second in size only to GRAPE with N_{max}=50, as well as very long and prominent whiskers. In general, however, the method of choice is the GA with N_{max}=50 since it gives a small RMSECV as well as the smallest box range in the boxplot.

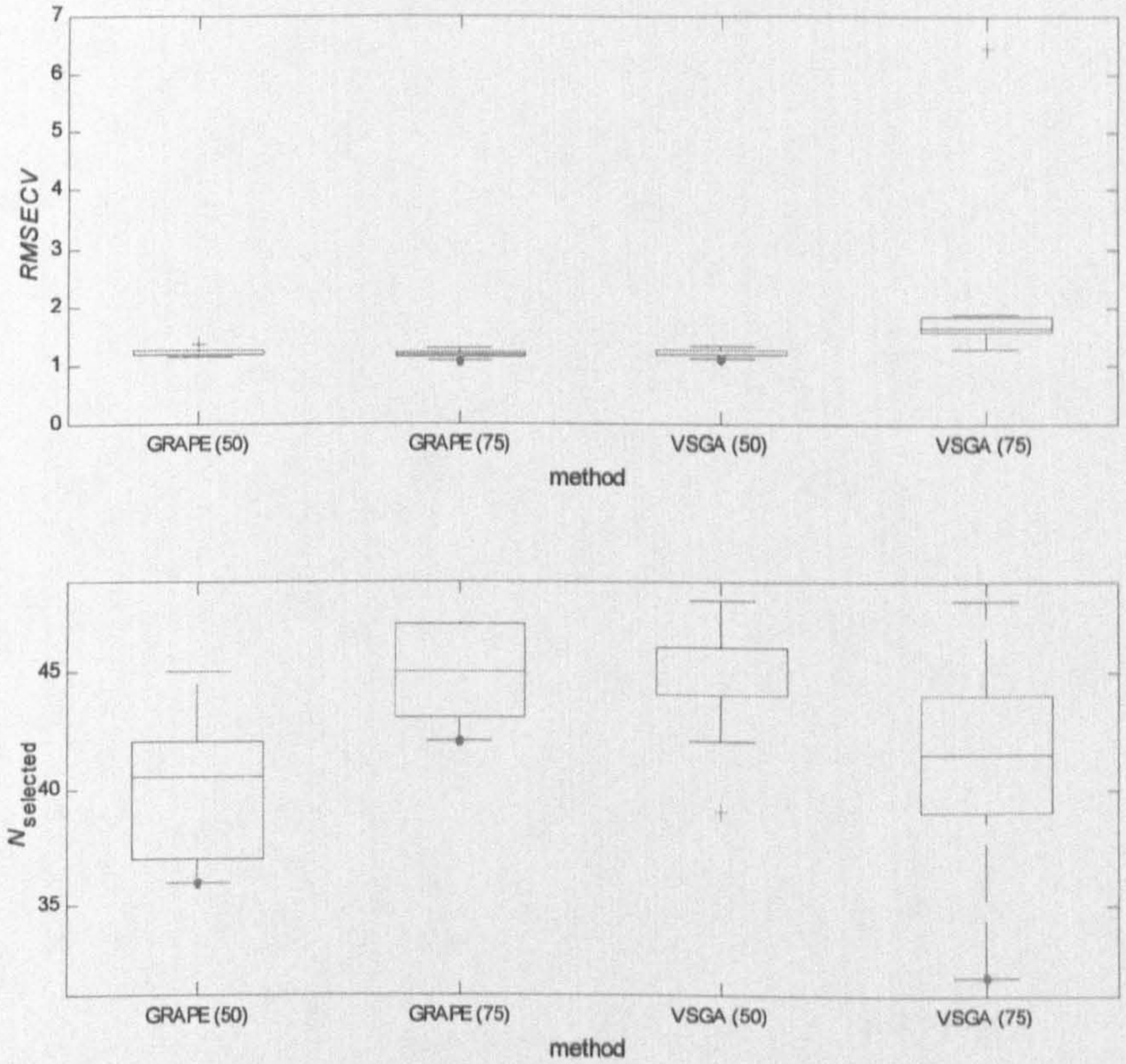

Fig. 40: Box plots of the prediction results of GRAPE and variable selection genetic algorithms (VSGA) on the naphtha Raman data.

Figure 41 shows plots of the original Raman data compared with the predictions from the application of GRAPE. The plots show the peaks representing the spiked samples where samples were spiked. In all the graphs, good prediction in all the 17 components of the naphtha mix is observed. The components are made up of paraffin compounds, aromatic compounds and naphthenes that cannot be named because of restrictions arising from matters of commercial and industrial confidentiality.

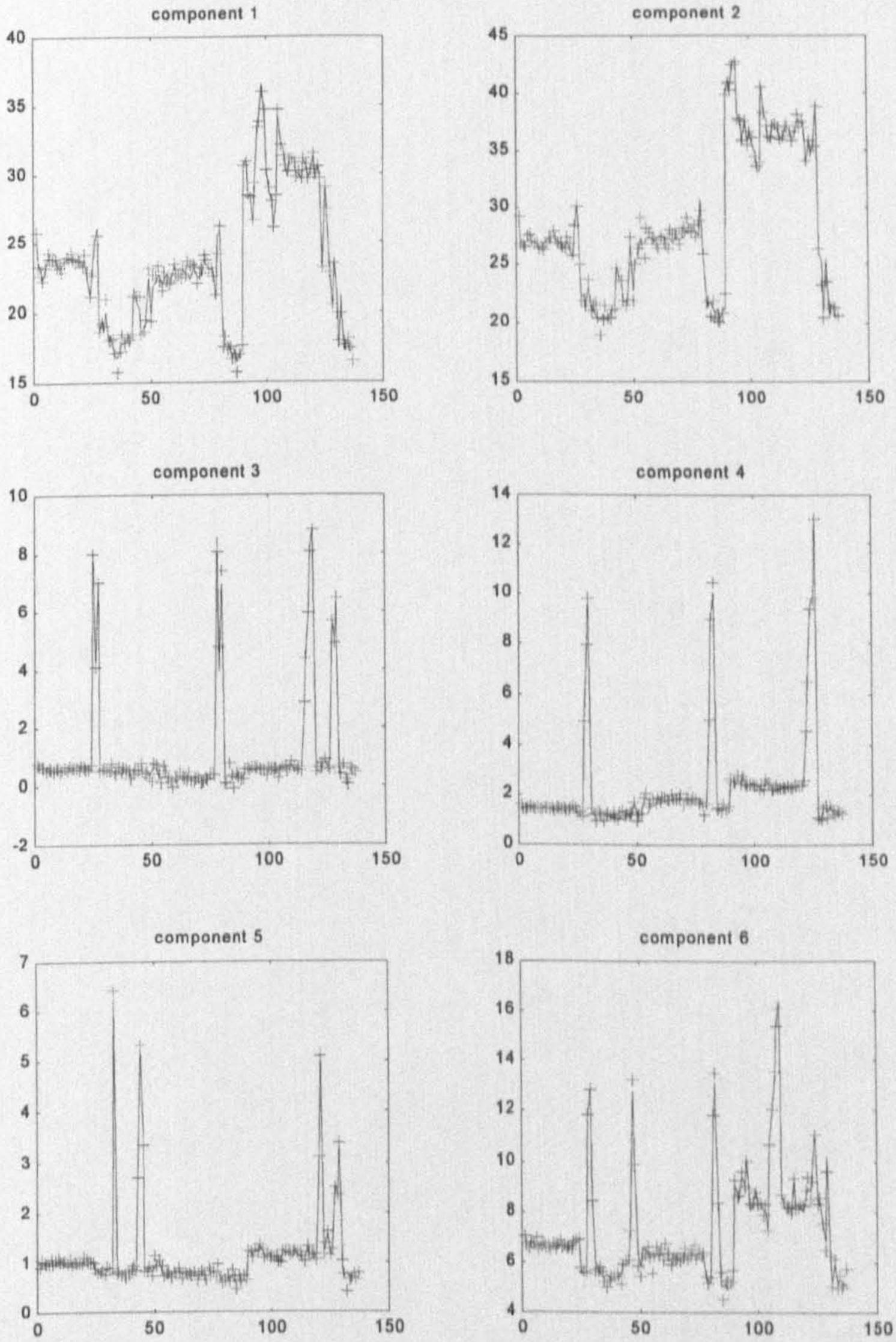

Fig. 41a: Measured (solid line) and predicted concentration (plus signs) of the first six components using GRAPE/MLR.

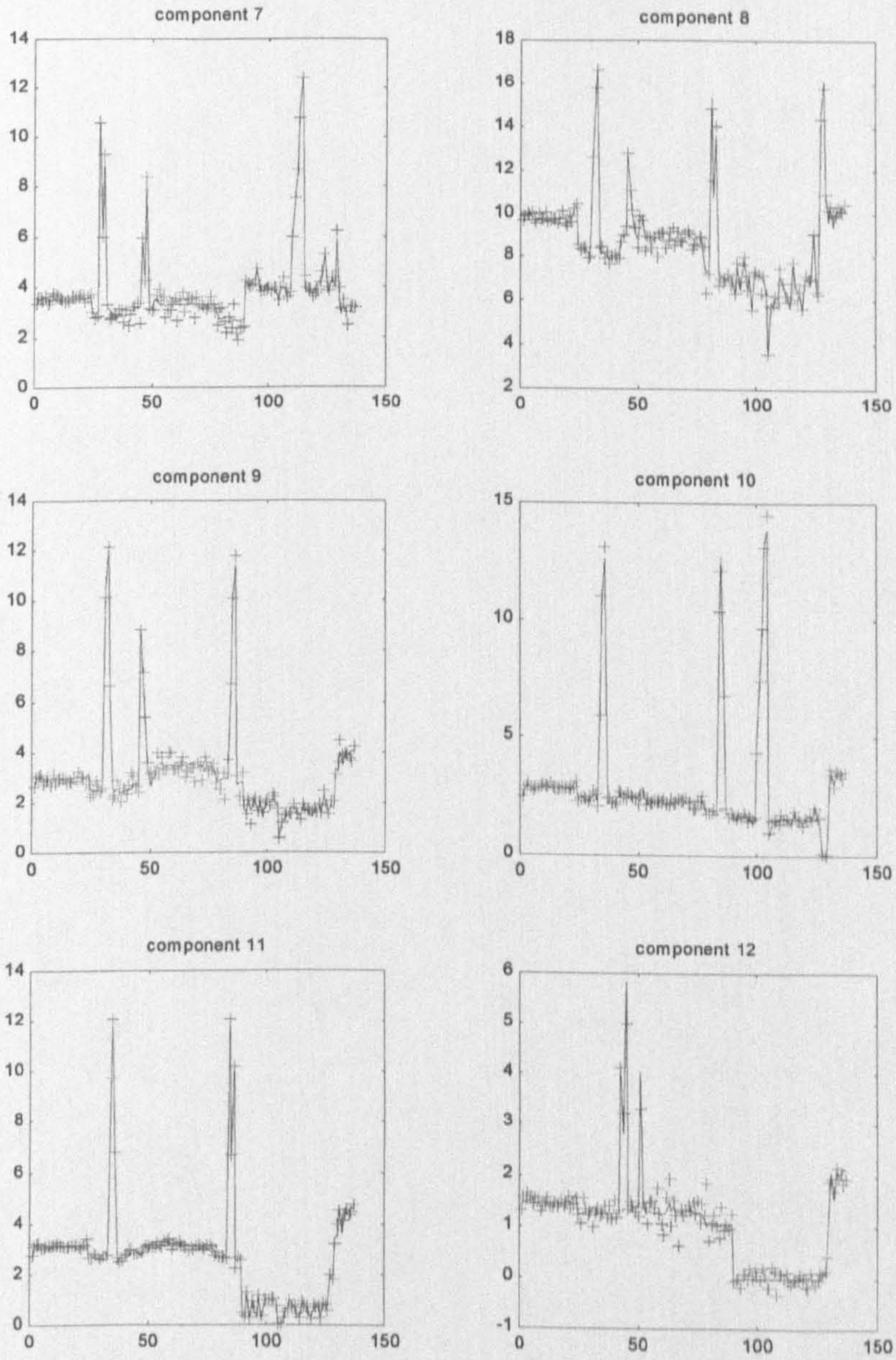

Fig. 41b: Measured (solid line) and predicted concentration (plus signs) for components 7-12 using GRAPE/MLR

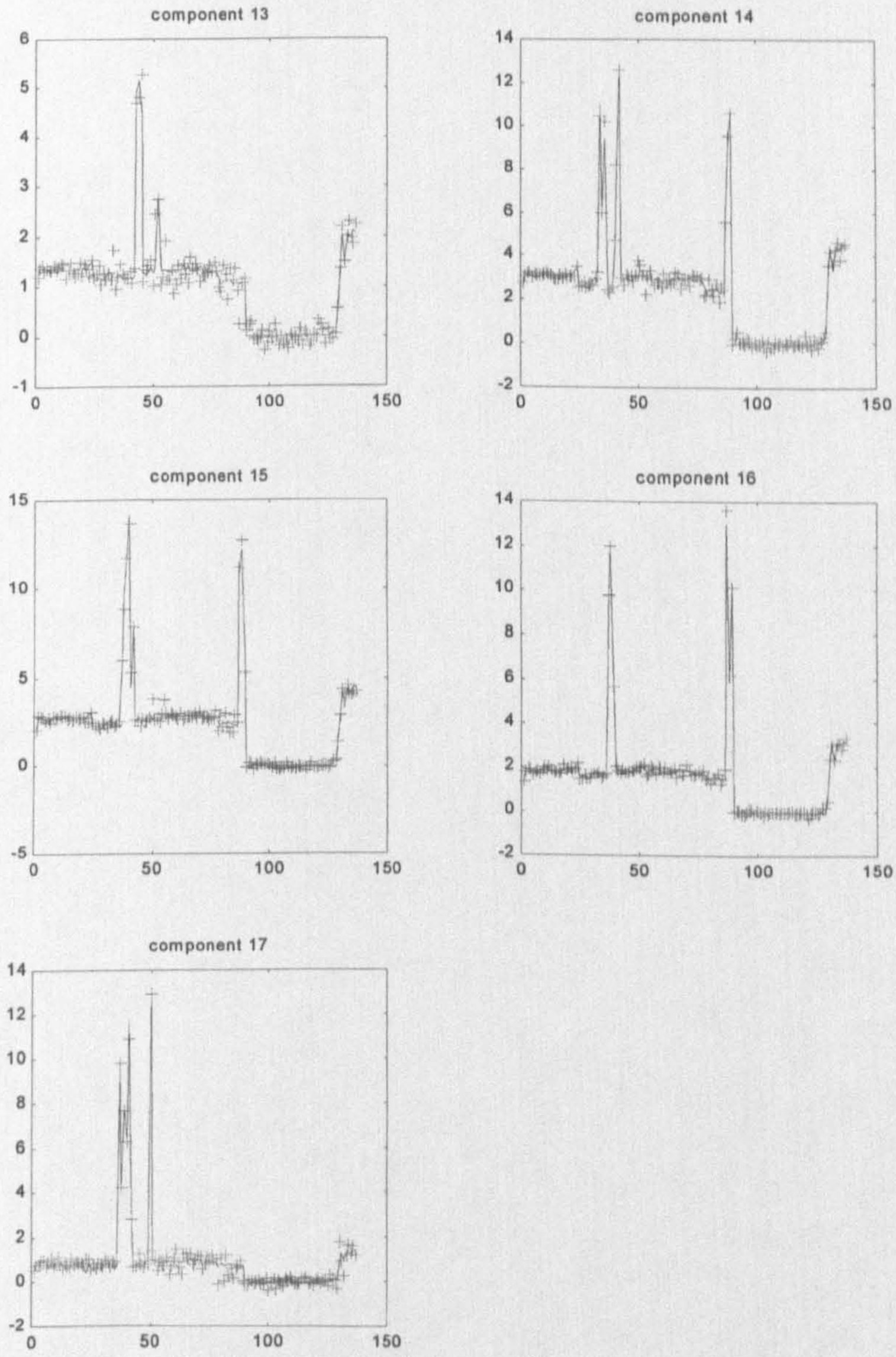

Fig. 41c: Measured (solid line) and predicted concentration (plus signs) for components 13-17 using GRAPE/MLR

CONCLUSION

For a complex continuous industrial process, the usefulness of the PCA technique for chemical process control has been successfully demonstrated by plotting the scores of the first principal component against the sample numbers. Monitoring the progress of each chemical component in the process with PCA is however not applicable, due to bulk changes in volume and concentration. This is compensated for by the use of variable selection algorithms in creating models that give excellent predictions. Prediction from PLS modelling gives better results than with MLR. However, when the variable selection methods are applied, the results are far better and the predictions more accurate than that acquired from the PLS model. Thus future runs in the continuous process can be assessed against the MLR model using variable selection.

APPENDIX

1.

Sample Number	Sample Type	Spike component
1	Non LIF	-
2	Non LIF	-
3	Non LIF	-
4	Non LIF	-
5	Non LIF	-
6	Non LIF	-
7	Non LIF	-
8	Non LIF	-
9	Non LIF	-
10	Non LIF	-
11	Non LIF	-
12	Non LIF	-
13	Non LIF	-
14	Non LIF	-
15	Non LIF	-
16	Non LIF	-
17	Non LIF	-
18	Non LIF	-
19	Non LIF	-
20	Non LIF	-
21	Non LIF	-
22	Non LIF	-
23	Non LIF	-
24	Non LIF	-
25	Non LIF Spike	1, 2, 3
26	Non LIF Spike	1, 2, 3
27	Non LIF Spike	1, 2, 3
28	Non LIF Spike	4, 6, 7
29	Non LIF Spike	4, 6, 7
30	Non LIF Spike	4, 6, 7
31	Non LIF Spike	5, 8, 9
32	Non LIF Spike	5, 8, 9
33	Non LIF Spike	5, 8, 9
34	Non LIF Spike	10, 11, 14
35	Non LIF Spike	10, 11, 14
36	Non LIF Spike	10, 11, 14
37	Non LIF Spike	15, 16, 17
38	Non LIF Spike	15, 16, 17
39	Non LIF Spike	15, 16, 17
40	Non LIF Spike	14, 15, 17
41	Non LIF Spike	14, 15, 17
42	Non LIF Spike	14, 15, 17
43	Non LIF Spike	5, 12 , 13
44	Non LIF Spike	5, 12 , 13
45	Non LIF Spike	5, 12 , 13

46	Non LIF Spike	6, 7, 8, 9
47	Non LIF Spike	6, 7, 8, 9
48	Non LIF Spike	6, 7, 8, 9
49	Tank Blend LIF	-
50	Non LIF Spike	17
51	Non LIF Spike	12
52	Non LIF Spike	13
53	LIF	-
54	LIF	-
55	LIF	-
56	LIF	-
57	LIF	-
58	LIF	-
59	LIF	-
60	LIF	-
61	LIF	-
62	LIF	-
63	LIF	-
64	LIF	-
65	LIF	-
66	LIF	-
67	LIF	-
68	LIF	-
69	LIF	-
70	LIF	-
71	LIF	-
72	LIF	-
73	LIF	-
74	LIF	-
75	LIF	-
76	LIF	-
77	LIF	-
78	LIF Spike	1, 2, 3
79	LIF Spike	1, 2, 3
80	LIF Spike	1, 2, 3
81	LIF Spike	4, 6, 8
82	LIF Spike	4, 6, 8
83	LIF Spike	4, 6, 8
84	LIF Spike	9, 10, 11
85	LIF Spike	9, 10, 11
86	LIF Spike	9, 10, 11
87	LIF Spike	14, 15, 16
88	LIF Spike	14, 15, 16
89	LIF Spike	14, 15, 16
90	Heads Spike Non LIF	2
91	Heads Spike Non LIF	2
92	Heads Spike Non LIF	2
93	Heads Spike Non LIF	2
94	Heads Spike Non LIF	2
95	Heads Spike Non LIF	1

96	Heads Spike Non LIF	1
97	Heads Spike Non LIF	1
98	Heads Spike Non LIF	1
99	Heads Spike Non LIF	1
100	Heads Spike Non LIF	10
101	Heads Spike Non LIF	10
102	Heads Spike Non LIF	10
103	Heads Spike Non LIF	10
104	Heads Spike Non LIF	10
105	Heads Spike Non LIF	6
106	Heads Spike Non LIF	6
107	Heads Spike Non LIF	6
108	Heads Spike Non LIF	6
109	Heads Spike Non LIF	6
110	Heads Spike Non LIF	7
111	Heads Spike Non LIF	7
112	Heads Spike Non LIF	7
113	Heads Spike Non LIF	7
114	Heads Spike Non LIF	7
115	Heads Spike Non LIF	3
116	Heads Spike Non LIF	3
117	Heads Spike Non LIF	3
118	Heads Spike Non LIF	3
119	Heads Spike Non LIF	3
120	Heads Spike Non LIF	5
121	Heads Spike Non LIF	5
122	Heads Spike Non LIF	4
123	Heads Spike Non LIF	4
124	Heads Spike Non LIF	4
125	Heads Spike Non LIF	4
126	Heads Spike Non LIF	4
127	Special Non LIF	-
128	Special Non LIF	-
129	Special Non LIF	-
130	Installation Tank Blend LIF	-
131	Installation Tank Blend LIF	-
132	Installation Tank Blend LIF	-
133	Installation Tank Blend LIF	-
134	Installation Tank Blend LIF	-
135	Installation Tank Blend LIF	-
136	Installation Tank Blend LIF	-
137	Installation Tank Blend LIF	-

Table 1: Data collected from Naphtha processing plant. (LIF = Laser Induced Fluorescence)

REFERENCES

1 Wold S., Kettaneh N., Friden H., and Holmberg A *Chemometrics and Intelligent Laboratory Systems*, **44**, (1998), 331-340

2 Dahl K.S., Piovoso M.J., Kosanovich K.A., *Chemometrics and Intelligent Laboratory Systems*, **46**, (1999), 161-180

3 Quinn A. C., Gemperline P. J., Baker B., Zhu M., and Walker D. S., *Chemometrics and Intelligent Laboratory Systems*, **45**, (1999), 199-214

4 Koch K.H., *Trends in Analytical Chemistry*, **12**, (1993), no.8, 333-339

5 McComb, M.E. and Gesser, H.D., *Talanta*, **49**, No.4, (1999) 869-879.

6 Mori, Y., *Bunseki Kagaku*, **49**, No.2, (2000), 131-132

7 Kourti T., MacGregor J.F., *Chemometrics and Intelligent Laboratory Systems*, **28**, (1995) 3-21.

8 Rännar S., MacGregor J.F., Wold S., *Chemometrics and Intelligent Laboratory Systems*, **41**, (1998) 73-81

9 Ad Louwerse & Age K. Smilde, *Multivariate Statistical Process Control of batch processes using three-way models*, PhD-project, University of Amsterdam, 1996

10 Roger Jones, Los Alamos National Laboratory and DuPont, March 18, 1994

11 Massart, D.L., Vandeginste B.G.M., Deming, S.N., Michotte, Y. and Kaufmann, L., *Chemometrics: A textbook* (Elsevier, Amsterdam, 1988).

12 Martens H., and. Næs T, *Multivariate Calibration*, John Wiley and Sons, New York, 1989.

13 Brereton, R.G.: *Multivariate Pattern Recognition in Chemometrics*. Elsevier, Amsterdam (1992).;

14 Brown, S.D., R.S. Bear and T.B. Blank: Chemometrics. *Analytical Chemistry*. 64 (1992) 22R-49R.

15 Wold, S., N. Kettaneh-Wold and B. Skagerberg: NONLINEAR PLS MODELLING. *Chemometrics and Intelligent Laboratory Systems*. 7 (1989) 53-65.

16 Taavitsainen, V.M. and P. Korhonen: NONLINEAR DATA ANALYSIS WITH LATENT VARIABLE, *Chemometrics and Intelligent Laboratory Systems*. 14 (1992)

17 Seasholtz, M.B., *Chemometrics and Intelligent Laboratory Systems*, **45**, (1999), 55-63

18 Wold S., Sjöström M., *Chemometrics and Intelligent Laboratory Systems*, **44** (1998), 3-14

19 Mendelson, Y., Cheung, P.W., Neuman, M.R., Flemin, D.G., and Cahn, S.D., *Advances in Experimental Medicine and Biology*, **159**, (1983), 93-102.

20 Kemsley E.K., *Chemometrics and Intelligent Laboratory Systems*, **33** (1996) 47-61

21 Andrews D.T., Chen L., Wentzell P.D., Hamilton D.C., *Chemometrics and Intelligent Laboratory Systems*, **34**, (1996), 231-244

22 Gurden S.P., Martin E.B., Morris A.J., *Chemometrics and Intelligent Laboratory Systems*, **44**, (1998) 319-330

23 Heberger, K., *Chemometrics and Intelligent Laboratory Systems*, **47**, 1999, 41-49

24 Astorga-Espana M.S., Pena-Mendez E.M., Garcia-Montelongo F.J., *Chemometrics and Intelligent Laboratory Systems*, **49**, (1999), 173-178.

25 Martens H., Næs T., *Multivariate Calibration*, John Wiley and Sons, New York, 1991.

26 Miller C.E., *Chemometrics and Intelligent Laboratory Systems*, **30**, (1995), 11-22

27 Blaser W., Bredeweg R., Harner R., LaPack M., Leugers A., Martin D., Pell R., Workman J., Wright L., *Analytical Chemistry*, **67**, (1995), 47R-70R.

28 Helminen J., Leppämäki M., Paatero E., Minnkinen P., *Chemometrics and Intelligent Laboratory Systems*, **44**, (1998) 341-352

29 Lang G., *Hydrocarbon Process*, **2**, (1994), 69-71

30 Classon R., *Chem. Eng.* (1993) 102-105

31 Heikka R., Immonen K., Minkkinen P., Paatero E., Salmi T., *Analytica Chimica Acta*, (1997), 1-8.

32 Blanco M., Coello J., Iturriaga H., Maspoch S., de la Pezuela C., *Analytica Chimica Acta*, **333**, (1996) 147-156

33 de Wit J.S., Baldwin E.K., *Process Control and Quality*, **4**, (1992) 21-30

34 Moessner R.C., *Process Control and Quality*, **2**, (1992), 237-247.

35 Colthup N.B., Daly L.H. Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, page 60.

36 Ward N., Edwards H., Johnson A., Fleming D., Coates P., *Appl. Spectrosc*, **50**, (1996), 812.

37 Cooper J.B., Wise K.L., Groves J., Welch W.T., *Analytical Chemistry*, 67, (1995), 4096

38 Cooper J.B., Wise K.L., Jensen B.J., *Analytical Chemistry*, 69, (1997), 1973

39 Svensson, O., Josefson, M. and Langkilde, F.W., *Chemometrics and Intelligent Laboratory Systems*, 49, (1999), 49-66

40 Colthup N.B., Daly L.H. Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, page 61.

41 Cooper J.B., *Chemometrics and Intelligent Laboratory Systems*, 46, (1999), 231-247

42 Roberts M.J., Garrison A.A., Kercel S.W., Muly E.C., *Process Control and Quality*, 1, (1991), 281-291.

43 Swierenga H., de Weijer A.P., van Wijk R.J., Buydens L.M.C., *Chemometrics and Intelligent Laboratory Systems*, 49, 1999, 1-17.

44 Townshend A., (Ed.-in-Chief), *Encyclopædia of Analytical Science*, 7, Harcourt Brace & Company, London, (1995), 4369-4429.

45 Massart, D.L. Vandeginste B.G.M., Buydens L.M.C., de Jong S., Lewi P.J., Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*, volume 20B of *Data Handling in Science and Technology*. Elsevier, Amsterdam, 1998.

46 P. Geladi, B.R. Kowalski, *Analytica Chimica Acta.*, 185 (1986) 1.

47 S. Wold, N. Kettaneh-Wold, B. Skagerberg. *Chemometrics & Intelligent Laboratory Systems*, 7, (1989) 53 65.

48 J.L. McClelland, D.E. Rummelhart, *Parallel Distributed Processing*, volume 1. MIT Press, London, 1988

49 Dr. Adrie Dane, CPACT, University of Hull, Hull, HU6 7RX, England.

50 Jackson J.E., *A User's Guide to Principal Components*, Wiley, NY, 1991.

51 Alun Jones, Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed SY23 3DA. auj@aber.ac.uk

52 Edinborough M., *Writing Organic Reaction Mechanisms*, Taylor and Francis, London, 1994, p140.

53 Atkins P.W., *Physical Chemistry*, Fourth Edition, Oxford University, Oxford, 1990, pp 786-787

54 Levitt B.P., (revised), *Findlays Practical Physical Chemistry*, 9th edition, Longman Group Limited, London, 1973, pages 335-336.

55 Data obtained from BP-Amoco Chemicals, Saltend, Hull, HU12 8DS, United Kingdom, with kind permission

56 Colthup N.B., Daly L.H, Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, page 390.

57 Colthup N.B., Daly L.H, Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, page 307.

58 Colthup N.B., Daly L.H, Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, page 333.

59 Colthup N.B., Daly L.H, Wiberley S.E., *Introduction to Infrared and Raman Spectroscopy*, 3rd Edition, 1990, Academic Press Ltd, London, pp 315-316.

60 Levitt B.P., (revised), *Findlays Practical Physical Chemistry*, 9th edition, Longman Group Limited, London, 1973, pages 335-336.