A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Submitted for the Degree of Doctor of Philosophy from the University of London by

Martin E. Pate

September 1998.

The Advanced Centre for Biochemical Engineering, The Department of Biochemical Engineering, University College London, Torrington Place, London, WC1E 7JE.
Abstract

Liquid chromatography is becoming increasingly important for the final purification of biomolecules. Traditionally, chromatography has been modelled using mathematical techniques which require experimental determination of the physico-chemical data for the separation of interest. These methods are both time-consuming and very complex and have only been truly successful in the prediction of binary separations and where the species in a mixture do not interact significantly. This thesis investigates the use of Principal Component Analysis (PCA), a multivariate statistical technique for the modelling and scale-up of chromatographic separations using a range of stationary phase chemistries and explores the utility of the approach as an engineering tool for rapid process development.

The reversed-phase separation of a semi-purified erythromycin feed was used as the feedstock throughout the study. Separations were performed on columns with various geometries and stationary phases. An isocratic solvent comprising 45:55 (v/v) of acetonitrile/water was used to effect the separation into 4 major components and at least 10 minor components.

In a first set of chromatograms, experimental design techniques were used to investigate the effects of four process variables (load volume, load concentration, temperature and pH of buffer) on the chromatogram shapes from each of 4 columns. The choice of appropriate data pre-processing prior to PCA was investigated in order to achieve maximal analytical performance from the statistical method. The results showed that when the retention times of elution peaks changed due to variations in temperature and pH, it was necessary to align the main product peak in order to gain most benefit from the PCA. Correlations were derived which enabled accurate chromatogram predictions (>95%) to be made using data from columns with fivefold change of scale and a fivefold change in sample size (25-fold scaling factor overall).

A second set of chromatograms were generated in which only the amount of sample load was varied. Sample concentrations of 20mg/mL were separated by each column, the sample volume applied being in the range 1-10% of the bed volume of each
column which included realistic non-linear, overload conditions. The principal components derived reflected similar properties of the chromatograms regardless of scale and stationary phase. These similarities were correlated, enabling predictions to be made from small (5cm length × 4.6mm diameter) to large-scale (up to 60mm diameter). The overall scaling factor in this set of chromatograms was in the region of 5000-fold.

The use of data reduction techniques was investigated throughout the study so as to minimise the number of runs required at large-scale whilst maintaining highly accurate (>98% accuracy) predictions. Results showed that considerable reductions in the sizes of data sets could be made (>75% reduction) without significant loss in the quality of the data provided that attempts were not made to extrapolate too far outside the limits of the experimental conditions.

PCA appears to be a very promising technique for the rapid and reliable modelling and scale-up of performance requiring minimal experimental work and achieving greater accuracy than with traditional mathematical approaches and makes recommendations for future work involving examining the potential relationships between PCA models and the underlying physico-chemical events controlling chromatographic separations.
Acknowledgements

I would like to thank my supervisor Dr Nigel Titchener-Hooker, firstly for providing the opportunity to study for this PhD and secondly for his continual support, encouragement and knowledge throughout the past three years. I would also like to thank my advisor Mrs Nina Thornhill for her advice regarding the PCA aspects of the research.

I would like to acknowledge the immense contribution to the latter stages of this thesis of Professor Michael Turner. His expertise and zeal have helped to develop, I believe, an important and radical method for the modelling of complex chromatographic separations.

I would like to thank some of the technical staff for their help in getting the Prochrom system up and running. These include the members of the electronics and engineering workshops. Thanks also to George Habib for writing a LabView program to control the solenoid valves on the Prochrom system.

I would like to acknowledge my colleague Angela Scholtzova for her continual support in all of the practical aspects of my research, especially with the commissioning of the Prochrom system. Angela was responsible, amongst many other things, for the HPLC method development of the erythromycin separation which is used throughout the department.

A final acknowledgement must be given to Lisa and Lady (a Jack Russell) for putting up with me.
# Table of Contents

Abstract 2  
Acknowledgements 4  
Table of Contents 5  
List of Tables 12  
List of Figures 14  
Nomenclature 19  
Abbreviations 22  

1. Introduction 24  

1.1. Summary 25  
1.2. Introduction 25  
1.3. Process Integration 25  
1.4. Analytical and Preparative Chromatography 26  
1.4.1. Background to HPLC 27  
1.4.2. Chromatographic Parameters when scaling up 28  
1.4.3. Chromatographic Techniques. 29  
1.4.3.1. Ion exchange chromatography 29  
1.4.2.2. Hydrophobic interaction chromatography 30  
1.4.3.3. Affinity chromatography 30  
1.4.3.4. Size exclusion chromatography 30  
1.4.4. Chromatographic Selection and Integration 31  
1.4.5. High Performance Liquid Chromatography 32  
1.5. Monitoring and Control in Bioprocesses 33  
1.5.1. The benefits of computers in monitoring and control 34  
1.5.2. Electronic Batch Recording (EBR). 35  
1.6. Artificial Intelligence in bioprocessing 36  
1.7. Principal Component Analysis 36  
1.7.1. General application of PCA to Bioprocesses 38
2. Chromatographic modelling and scale-up - theory.

2.1. Introduction

2.2. The Mass-Balance Equation

2.2.1. The solutions of the mass balance equation

2.2.1.1. The ideal model.

2.2.1.2. The Equilibrium -Dispersive model

2.2.1.3. Lumped Kinetic Model.

2.2.1.4. The effect of Mass transfer on the mass balance equation

2.3. Adsorption Isotherms

2.3.1. Linear and Non-linear chromatography

2.3.2. Single component isotherms

2.3.2.A. The Langmuirian Isotherm

2.3.2.B The Bi-Langmuirian isotherm

2.3.2.C. The Fowler isotherm

2.3.2.D. Freundlich isotherm

2.3.2.E. S-shaped isotherms

2.3.3. Multicomponent or Competitive adsorption isotherms

2.3.3.A. The Competitive Langmuirian isotherm

2.3.3.B. The competitive bi-Langmuirian isotherm

2.3.3.C. The Ideal Adsorbed Solution

2.3.3.D. The Statistical Isotherm

2.3.3.E. The competitive Fowler Isotherm

2.3.3.F. The competitive Freundlich-Langmuirian Isotherm.

2.4. The Determination of Single Component Isotherms

2.4.1. Frontal Analysis.

2.4.2. Frontal Analysis by Characteristic Point

2.4.3. Elution By Characteristic Point
2.4.4. Retention Time Method. 63
2.5. Summary 64

3. Principal Component Analysis - Theory 65

3.1. Introduction 66
3.2. Mathematical description of PCA 66
3.3. Scaling of variables prior to PCA 68
3.4. Normalising the samples 68
3.5. Mean Centring 69
3.6. Calculating the Principal Components 69
  3.6.1. The NIPALS Algorithm 70
  3.6.2. Decomposition of the Variance-Covariance Matrix. 72
3.7. An example of PCA being applied to chromatography data. 72
  3.7.1. The scores and loadings for PCI 73
  3.7.2. Percentage variance associated with PCI. 76
  3.7.3. Reconstruction of Chromatogram data. 77
3.8. How many PCs are required for a model? 77
  3.8.1. The leave-one-sample-out-at-a-time procedure 77
  3.8.2. The Scree Graph and the Log-Eigenvalue Diagram 78
  3.8.3. Discussion about the number of PCs 78
3.9. Classification of samples 79
  3.9.1. SIMCA (Soft Independent Modelling of Class Analogies) 80
  3.9.2. The measurement of distance between samples. 80
  3.9.3. The SIMCA Model. 82
    3.9.3.1. The residual standard deviation (RSD) of a sample 83
    3.9.3.2. F-Test statistics for the comparison of \( s_k^2 \) and \( s_c^2 \) 83
    3.9.3.3. Upper and Lower limits for scores values of each PC 84
3.10. Other multivariate statistical techniques. 85
  3.10.1. Multiway Principal component analysis (MPCA) 85
  3.10.2. Partial Least Squares (PLS) 86
  3.10.3. Principal component regression (PCR) 87
3.11. Conclusions 88

4. Pre-processing of chromatographic data for Principal Component Analysis 89

4.1. Summary 90

4.2. Introduction 90

4.3. Materials and Methods 93

4.3.1. The Chromatographic Process 93

4.3.2. Generation of historical data sets 94

4.3.3. Flow Rate Transformation 96

4.3.4. Sample Load Adjustment 96

4.5. Results and Discussion 96

4.5.1. Analysis of the gross principal component model for dataset X1 after flow rate transformation 98

4.5.2. Analysis of the disjoint model for the cluster at constant vc 102

4.5.3. Analysis of the gross principal component model for dataset X2 104

4.5.4. Analysis of data set X1 with adjustment for sample load (after flow rate transformation) 105

5. Conclusions 107

5. The Use of Principal Component Analysis for the Modelling of High Performance Liquid Chromatography 109

5.1. Summary 110

5.2. Introduction 111

5.2.1. The Modelling of Liquid chromatography 112

5.2.2. Principal Component Analysis 113

5.3. Materials and methods 114

5.3.1. The Separation Method 114

5.3.2. The Erythromycin Samples 115

5.3.3. The Modular HPLC system 116
Table of Contents

5.3.4. Column Cleaning Procedure. 118
5.3.5. Generation of data sets at the 3 scales of operation. 119
5.3.6. Data pre-processing 121
5.3.6.1. Retention time transformation 121
5.3.6.2. Centring of chromatograms 122
5.4. Results and Discussion 122
5.4.1. Analysis of the full factorial data set for the 25 cm column (with retention time adjustment) 122
5.4.2. Selection of subsets of chromatograms - the framework for scale-up predictions 125
5.4.3. PCA of subsets of chromatograms (adjusted for retention time effects). 125
5.4.4. Analysis of chromatograms derived from columns of different lengths 129
5.4.4.1. Retention times 129
5.4.4.2. Principal component scores 129
5.4.5. Relationship between principal components and chromatographic profiles 131
5.4.6. Prediction of chromatographic profiles 132
5.4.6.1. Goodness of fit 134
5.5. Conclusions 136

6. The use of Principal Component Analysis for the modelling and scale-up of non-linear chromatography. 139

6.1. Summary 139
6.2. Introduction 139
6.3. Materials and methods 140
6.3.1. The test HPLC system 140
6.3.2. Columns 141
6.2.3.1. HPLC Column Packing 141
6.4. Results and Discussion 143
6.4.1. An overview of the chromatograms from the three columns. 143
6.4.2. Extraction of a PCA model for the CG71 methacrylate column 146
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4.2.1. <strong>Subtraction of the mean chromatogram</strong></td>
<td>146</td>
</tr>
<tr>
<td>6.4.2.2. <strong>Extraction of PC1</strong></td>
<td>146</td>
</tr>
<tr>
<td>6.4.2.3. <strong>Extraction of PC2 and PC3</strong></td>
<td>147</td>
</tr>
<tr>
<td>6.4.2.4. <strong>Correlations between the amount of sample chromatographed and the values in the scores matrices</strong></td>
<td>151</td>
</tr>
<tr>
<td>6.4.3. <strong>Reconstruction of CG71 chromatograms</strong></td>
<td>152</td>
</tr>
<tr>
<td>6.4.4. <strong>Extraction of PC models for the CG300 and the PLRP1000-S polystyrene columns</strong></td>
<td>156</td>
</tr>
<tr>
<td>6.4.5. <strong>Alternative models</strong></td>
<td>158</td>
</tr>
<tr>
<td>6.4.5.1. <strong>Model based on scaling of a mid-range chromatogram</strong></td>
<td>158</td>
</tr>
<tr>
<td>6.4.5.2. <strong>The use of differentials of the mean chromatogram</strong></td>
<td>158</td>
</tr>
<tr>
<td>6.4.6 <strong>Comparison of Principal Components extracted from different columns</strong></td>
<td>160</td>
</tr>
<tr>
<td>6.4.6.1. <strong>Loadings</strong></td>
<td>160</td>
</tr>
<tr>
<td>6.4.6.2. <strong>Scores</strong></td>
<td>161</td>
</tr>
<tr>
<td>6.4.7. <strong>Correlations between Principal Components from different chromatograms.</strong></td>
<td>161</td>
</tr>
<tr>
<td>6.5. <strong>Conclusions</strong></td>
<td>164</td>
</tr>
</tbody>
</table>

**Chapter 7. The use of PCA for the Modelling of Pilot-Scale Chromatograms.**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1. <strong>Summary</strong></td>
<td>167</td>
</tr>
<tr>
<td>7.2. <strong>Materials and Methods.</strong></td>
<td>167</td>
</tr>
<tr>
<td>7.2.1. <strong>Overview</strong></td>
<td>167</td>
</tr>
<tr>
<td>7.2.2. <strong>Preparation of Stationary Phase Slurry</strong></td>
<td>169</td>
</tr>
<tr>
<td>7.2.3. <strong>Column Packing</strong></td>
<td>169</td>
</tr>
<tr>
<td>7.2.4. <strong>Sample Injection</strong></td>
<td>170</td>
</tr>
<tr>
<td>7.2.5. <strong>Column Unpacking</strong></td>
<td>170</td>
</tr>
<tr>
<td>7.2.6. <strong>Problems Associated with Commissioning of the Prochrom HPLC System</strong></td>
<td>170</td>
</tr>
<tr>
<td>7.2.1. <strong>The Columns and their Separation Methods</strong></td>
<td>171</td>
</tr>
<tr>
<td>7.2.2. <strong>The Sample Loads</strong></td>
<td>172</td>
</tr>
<tr>
<td>7.3. <strong>Results and Discussion</strong></td>
<td>173</td>
</tr>
</tbody>
</table>
7.3.1. An overview of the chromatograms from the two semi-preparative columns 173

7.3.2. An overview of the chromatograms from the Pilot (Prochrom) Scale Column 174

7.3.1. Extraction of PCA models for the 2 semi-preparative columns 175

7.3.1.1. Loadings for the 2 semi-preparative columns 176

7.3.1.2. Scores for the 2 semi-preparative columns 177

7.3.4. Extraction of a PC model for Column 7.3 (12cm×60mm pilot scale) 178

7.3.4.1. Loadings Profiles for the 60mm Diameter Pilot Scale Column 179

7.3.4.2. Scores for the 60mm Diameter Pilot Scale Column 179

7.3.5. Prediction of all 4 chromatograms using the 3 PC model. 180

7.3.6. The Prediction of a Test Chromatogram within the range of Sample Loads for the Pilot-Scale column. 183

7.4. Conclusions 184

8. The use of Finite Difference Techniques 187

8.1. Introduction 187

8.2. The Principle of Finite Difference Methods 187

8.3. An Example involving Finite Difference Techniques 188

8.4. Discussion about the Finite Difference Approach 193

9. Overall Conclusions and Discussion 195

10. Future work 201

Appendix 1. Column Efficiencies 203

Appendix 2. Review of the Use of Artificial Intelligence In Bioprocessing 204

Appendix 3. References 214
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Differences between upstream and downstream processes</td>
</tr>
<tr>
<td>1.2.</td>
<td>Some of the operational differences between analytical and preparative HPLC.</td>
</tr>
<tr>
<td>2.1.</td>
<td>Summary of the main differences between linear and non-linear chromatography.</td>
</tr>
<tr>
<td>2.2.</td>
<td>Summary of the differences between ideal and non-ideal chromatography.</td>
</tr>
<tr>
<td>3.1.</td>
<td>A table giving 11 UV absorbance values which map out the shape of a set of example chromatograms. This shows the data in the raw form prior to subtraction of the mean chromatogram.</td>
</tr>
<tr>
<td>4.1.</td>
<td>A description of the levels for the factors (process variables) used to generate the datasets $X_1$ and $X_2$. ‘+1’ represents a high setting, ‘-1’ a low setting and ‘0’ a centre value for the factor.</td>
</tr>
<tr>
<td>4.2.</td>
<td>Summary of the experimental design used to generate the first dataset $X_1$. The table describes a 2-level full factorial design and a centre point.</td>
</tr>
<tr>
<td>4.3.</td>
<td>A summary of the design for generating the second dataset $X_2$.</td>
</tr>
<tr>
<td>5.1.</td>
<td>Table showing the values of each variable at their respective levels. The volume of sample loaded onto the column is proportional to the size of the column.</td>
</tr>
<tr>
<td>5.2.</td>
<td>Table showing the levels of the variables for the full factorial data set using the 25 cm column. The design includes one-at-a-time changes of each variable at four levels (+, -, +2 and -2).</td>
</tr>
<tr>
<td>5.3.</td>
<td>Table showing the range of both sample volumes and mass of material added to each of the three columns.</td>
</tr>
<tr>
<td>5.4.</td>
<td>The critical subset of chromatograms from which scale-up predictions were made.</td>
</tr>
<tr>
<td>5.5.</td>
<td>Table showing the clusters selected to comprise the three, five and seven cluster subsets.</td>
</tr>
<tr>
<td>5.6.</td>
<td>The goodness of fit from 3, 5, 7 and all 9 cluster models for chromatograms at 25 cm scale predicted using a 1PC model prediction from 5 cm data.</td>
</tr>
<tr>
<td>6.1.</td>
<td>Table of mass loaded onto each column type. Each mass loaded was proportional to the bed volume of each column. All samples had concentration of 20 mg/mL.</td>
</tr>
</tbody>
</table>
### 6.2. Mean goodness of fits for how a 3 PC model describes the column 3 (CG71) data.

### 6.3. The percentage variance associated for the first 3 PCs for each column.

### 7.1 Table of mass loaded onto each column type. All samples had concentration of 20 mg/mL.

### 7.2 The bed volumes for the 3 columns together with the range of sample volumes applied to each column.

### 7.3 The percentage variance associated for the first 3 PCs for both semi-preparative columns.

### 7.4 The percentage variance associated for the first 3 PCs for Column 7.3 (60mm diameter pilot scale).
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Legend</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The mean and maximum purification factors expected for four common types of chromatographic separation.</td>
<td>32</td>
</tr>
<tr>
<td>2.1</td>
<td>An example of a Langmuirian isotherm.</td>
<td>51</td>
</tr>
<tr>
<td>2.2</td>
<td>A concave-up Langmuirian-type isotherm.</td>
<td>53</td>
</tr>
<tr>
<td>2.3</td>
<td>A comparison between the effects of different-shaped isotherms on the shapes of single-component elution profiles. Linear, Langmuirian and anti-Langmuirian isotherms are compared</td>
<td>53</td>
</tr>
<tr>
<td>2.4</td>
<td>A typical example of a competitive isotherm for a binary system. Combinations of C₁ and C₂ allow the calculation of q₂.</td>
<td>57</td>
</tr>
<tr>
<td>2.5</td>
<td>A typical recording obtained for the frontal analysis method of isotherm determination.</td>
<td>62</td>
</tr>
<tr>
<td>2.6</td>
<td>A typical recording obtained for the frontal analysis by characteristic point method of isotherm determination.</td>
<td>63</td>
</tr>
<tr>
<td>3.1</td>
<td>A block diagram illustrating how the first two principal components are extracted using the NIPALS algorithm.</td>
<td>71</td>
</tr>
<tr>
<td>3.2</td>
<td>The chromatograms which are comprised of the data from Table 3.1 This figure shows the data in its raw form prior to mean centring.</td>
<td>73</td>
</tr>
<tr>
<td>3.3</td>
<td>The loadings profiles for the first 2 PCs of the example data set X. These are common to each sample. The amount of each of the PCs contained in each sample is given by the scores values.</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>The PC1-PC2 scores plot for the example matrix X. One trend of this plot is that PC1 scores reflects the sample load applied.</td>
<td>76</td>
</tr>
<tr>
<td>3.5</td>
<td>A scree plot for the first 5 PCs of the example data set. From this plot, an estimate can be made as to the number of principal components for a suitable model.</td>
<td>78</td>
</tr>
<tr>
<td>3.6</td>
<td>Two samples plotted in 2-dimensional variable space. The difference between the sample vectors define the Euclidean distance dₐᵢ.</td>
<td>81</td>
</tr>
<tr>
<td>3.7</td>
<td>The sample vectors are expressed as the centre of gravity and a residual vector. Class membership can be decided by comparing the residual distances to the centre of gravity of the samples used to model the class.</td>
<td>81</td>
</tr>
<tr>
<td>3.8</td>
<td>Two PCs have been fitted to the samples from Figure 3.7. Instead of comparison with the centre of gravity which assumes a spherical distribution, each sample is related to the lines defined by the PCs.</td>
<td>82</td>
</tr>
<tr>
<td>3.9</td>
<td>By using the spread of sample scores along each PC, upper and lower limits can be defined. The cylinder described using this technique is used to assess</td>
<td>85</td>
</tr>
</tbody>
</table>
acceptance or rejection of samples to or from a membership cluster.

3.10. A diagram of a 3-dimensional data matrix $X$ with dimensions $(I \times J \times K)$.  
There are $I$ samples, $K$ measured variables and $K$ time increments.

3.11. The algorithm for performing MPCA

4.1. An example of a typical chromatogram obtained from the test SEC system illustrating the presence of three peaks.

4.2. An overview of all 72 chromatograms in $X_1$. There are 3 distinct groupings corresponding to high, centre and low flow rates.

4.3 Scores plot for PC1 and PC2 for the gross model for the full factorial data set $X_1$ without pre-processing. There are 9 very compact clusterings.

4.4. The same 72 chromatograms as in Figure 4.2 after pre-processing where the data was transformed from time to volume in order to compensate for the flow rate variable.

4.5 Scores plot of the first two principal components for the gross (transformed) model for $X_1$. The plot shows the qualitative positions of clusters of similar chromatograms.

4.6. Scores plot of PC1 and PC3 for the gross (transformed) model for $X_1$. The plot is very similar to that for PC1 and PC2.

4.7. Disjoint model for cluster obtained at constant $v_c$ with the outliers removed. Indicated on the plot are the effects of increasing temperature or flow rate.

4.8. Magnified scores plot of cluster obtained at constant $v_c$ from Figure 4.4 to be compared with the disjoint model approach (Figure 4.7). It is clear that distinctions between flow rate and temperature are visible at this scale.

4.9 Scores plot for PC1 and PC2 in the gross model for the one at a time data set $X_2$ without adjustment for flow rate (Chandwani et al, 1997). This plot reveals the presence of six distinct and compact clusters of apparently very similar chromatograms.

4.10. Scores plot of the first two components for the gross (transformed) model for $X_2$. The plot shows the qualitative positions of clusters of similar operating conditions. There are 9 distinct clusters as predicted by the experimental design from Table 4.3.

4.11. All 72 chromatograms adjusted to constant area to reduce the influence of the mass loaded in the analysis. These chromatograms have previously been transformed from time to volume to account for the dominating effect of flow rate.

4.12. Scores plot for the first two principal components for data set $X_1$ following transformations to account for both flow rate and mass load effects.

5.1. A schematic of the Beckman analytical HPLC system comprising UV detector, autosampler and solvent delivery module all linked to an IBM personal computer.

5.1b All 108 Chromatograms for the 25 cm column prior to retention time
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1c</td>
<td>All 108 Chromatograms for the 25 cm column after retention time adjustment.</td>
</tr>
<tr>
<td>5.2a</td>
<td>Two chromatograms with different retention times before RT adjustment.</td>
</tr>
<tr>
<td>5.2b</td>
<td>The same two chromatograms after RT adjustment.</td>
</tr>
<tr>
<td>5.3</td>
<td>The loadings profiles for the first two principal components of the complete data set (with retention time adjustment).</td>
</tr>
<tr>
<td>5.4</td>
<td>Scores plot for the first two PCs in the data set with RT adjustment. This plot shows a definite trend of the chromatograms with respect to the mass of material loaded onto the column, the larger loads having high scores in PC1 and the smaller loads scoring low in PC1.</td>
</tr>
<tr>
<td>5.5</td>
<td>A comparison between the PC1 scores generated using just 9 groups from 36 (27 out of 108 chromatograms) (x-axis) and between the actual PC1 scores generated using all 108 chromatograms.</td>
</tr>
<tr>
<td>5.6</td>
<td>The PC1 loadings profiles for the 25 cm column obtained using all 36, 9 and 3 groups (108, 27 and 9 chromatograms).</td>
</tr>
<tr>
<td>5.7</td>
<td>A comparison between the PC1 scores generated using just 3 groups from 36 (9 out of 108 chromatograms) (x-axis) and between the actual PC1 scores generated using all 108 chromatograms.</td>
</tr>
<tr>
<td>5.8</td>
<td>Retention time correlations between 5 and 15 cm columns and 5 and 25 cm columns.</td>
</tr>
<tr>
<td>5.9</td>
<td>The PC1 loadings profiles of the nine cluster models for the 25 cm, 15 cm and 5 cm columns. To allow easier comparability, the loadings at the 5 cm and 15 cm scale have been stretched to align with the main peak of the 25 cm loadings.</td>
</tr>
<tr>
<td>5.10</td>
<td>The PC1 loadings profile for the model using only the low sample loads.</td>
</tr>
<tr>
<td>5.11a</td>
<td>PC1 scores correlations from 5 cm to both 15 and 25 cm column length using a nine cluster model. The 5 to 15 cm correlation is more linear than the 5 to 25 cm.</td>
</tr>
<tr>
<td>5.11b</td>
<td>PC1 scores correlations from 5 cm to 25 cm column using a 3 cluster model.</td>
</tr>
<tr>
<td>5.12</td>
<td>Chromatogram CC at 25 cm predicted from 5cm data using the three and nine cluster models.</td>
</tr>
<tr>
<td>5.12</td>
<td>A flow-sheet showing the main stages in the use of PCA for the prediction of chromatograms from 5 cm to 25 cm scale.</td>
</tr>
<tr>
<td>6.1a</td>
<td>All chromatograms for column 1 (5cm×4.6mm PLRP, 8 μm).</td>
</tr>
<tr>
<td>6.1b</td>
<td>All chromatograms for column 2 (15cm×4.6mm CG300, 35 μm).</td>
</tr>
<tr>
<td>6.1c</td>
<td>All chromatograms for column 3 (15cm×7.5mm CG71, 35 μm).</td>
</tr>
<tr>
<td>6.2a</td>
<td>The mean chromatogram of the CG71 column data (column 3).</td>
</tr>
<tr>
<td>6.2b</td>
<td>The original CG71 data (column 3) after the subtraction of the mean.</td>
</tr>
</tbody>
</table>
6.3a. The PC1 loadings profile for the CG71 column (column 3).
6.3b. The residuals after extraction of PC1 for the CG71 column (column 3).
6.4a. The PC2 loadings profile for the CG71 column (column 3).
6.4b. The residuals after extraction of PC2 for the CG71 column (column 3).
6.5a. The PC3 loadings profile for the CG71 column (column 3).
6.5b. The residuals after extraction of PC3 for the CG71 column (column 3).
6.6. Scores against mass loaded for the CG71 column (column 3). The relationship for PC1 is linear, PC2 is quadratic and PC3 is cubic.
6.7. A diagrammatic representation showing the reconstruction of a chromatogram using PCA. The contribution for each PC is the product of the scores and loadings. The sum of these products are added to the mean chromatogram to complete the prediction.
6.8. All chromatograms from the CG71 column (column 3) reconstructed using a 3PC model. The chromatograms are visually very similar to the original set.
6.9. The retention times of the shoulder of the main peak for column 3. The actual retention times are compared with those predicted from a 3PC model.
6.10. Figure showing actual with predicted chromatograms for the lowest sample load (with poorest fit) from column 3.
6.11a. The loadings profiles for PLRP1000 column (column 1).
6.11b. The loadings profiles for CG300 column (column 2).
6.12a. Scores against mass loaded for PLRP1000 column (column 1).
6.12b. Scores against mass loaded for CG300 column (column 2). Both of these plots are strikingly similar to the scores observed for the CG71 column (Fig. 6.5).
6.13. The reconstruction of the 0.46mg chromatogram compared with the actual chromatogram for column 2.
6.14. The set of chromatograms for column 2 each predicted from a simple scaling of a mid-range chromatogram.
6.15. The differentials of a mid-range chromatogram from the CG71 set of chromatograms. This technique of modelling the separation was also ineffective.
6.16. PC1 scores for columns 2 and 3 predicted from column 1 (PLRP1000).
6.17a. PC2 scores for columns 2 and 3 predicted from column 1 (PLRP1000).
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.17b</td>
<td>PC3 scores for columns 2 and 3 predicted from column 1 (PLRP1000).</td>
<td>163</td>
</tr>
<tr>
<td>7.1</td>
<td>A schematic for the preparative chromatography (Prochrom) system</td>
<td>168</td>
</tr>
<tr>
<td>7.2a</td>
<td>All chromatograms for column 7.1 (15cm×7.8mm CG300, 35μm).</td>
<td>174</td>
</tr>
<tr>
<td>7.2b</td>
<td>All chromatograms for column 7.2 (15cm×10mm CG300, 35μm).</td>
<td>174</td>
</tr>
<tr>
<td>7.3</td>
<td>All 4 chromatograms for column 7.3 (12cm×60mm CG300, 35μm).</td>
<td>175</td>
</tr>
<tr>
<td>7.4</td>
<td>The lowest sample load (95.3 mg) from column 7.3.</td>
<td>176</td>
</tr>
<tr>
<td>7.5</td>
<td>The loadings profiles for column 7.1 (15cm×7.8mm)</td>
<td>177</td>
</tr>
<tr>
<td>7.6</td>
<td>Scores against mass loaded for column 7.1 (15cm×7.8mm).</td>
<td>178</td>
</tr>
<tr>
<td>7.7</td>
<td>The loadings profiles for the first 3 PCs for Column 3 (60mm diameter pilot scale).</td>
<td>179</td>
</tr>
<tr>
<td>7.8</td>
<td>Scores against mass loaded for column 7.3 (12cm×60mm).</td>
<td>180</td>
</tr>
<tr>
<td>7.9</td>
<td>The 4 pilot-scale chromatograms predicted using a 3 PC model. All chromatograms had a goodness-of-fit of 100%.</td>
<td>181</td>
</tr>
<tr>
<td>7.10a</td>
<td>The prediction of the lowest sample load chromatogram from the pilot-scale with the exclusion of PC3. The prediction is shown with the actual chromatogram</td>
<td>182</td>
</tr>
<tr>
<td>7.10b</td>
<td>The prediction of the lowest sample load chromatogram from the pilot-scale with PC3 included. The prediction is shown with the actual chromatogram and has a goodness-of-fit of 100%.</td>
<td>182</td>
</tr>
<tr>
<td>7.11</td>
<td>The 600 mg test chromatogram shown together with the 4 original pilot-scale chromatograms.</td>
<td>183</td>
</tr>
<tr>
<td>7.12</td>
<td>The predicted chromatogram at 600mg based on a 3PC model compared with the actual chromatogram obtained with 600mg sample load.</td>
<td>184</td>
</tr>
<tr>
<td>8.1</td>
<td>A set of single-component elution profiles generated from a Langmuirian isotherm using the finite difference approach.</td>
<td>189</td>
</tr>
<tr>
<td>8.2</td>
<td>A set of single-component elution profiles generated from an anti-Langmuirian isotherm using the finite difference approach.</td>
<td>189</td>
</tr>
<tr>
<td>8.3</td>
<td>The loadings profiles of the first 3 PCs for the Langmuirian set of chromatograms generated using the finite difference method.</td>
<td>190</td>
</tr>
<tr>
<td>8.4</td>
<td>The loadings profiles of the first 3 PCs for the anti-Langmuirian set of chromatograms generated using the finite difference method.</td>
<td>190</td>
</tr>
<tr>
<td>8.5</td>
<td>The scores for the first three PCs with increasing sample load for the Langmuirian isotherm generated using the finite difference approach.</td>
<td>191</td>
</tr>
<tr>
<td>8.6</td>
<td>The scores for the first three PCs with increasing sample load for the Langmuirian isotherm generated using the finite difference approach.</td>
<td>192</td>
</tr>
</tbody>
</table>
Nomenclature

\( \alpha \) Level of Confidence in F-Test
\( \alpha \) Relative retention value
\( \chi \) Empirical interaction parameter in Fowler isotherm
\( \varepsilon \) Column voidage
\( \varepsilon_p \) Particle porosity
\( \lambda_i \) Eigenvalue associated with \( p_i \)
\( \lambda \) \( k \times k \) matrix of eigenvalues
\( \tau \) Tortuosity factor
\( \tau \) Time increment for finite difference technique
\( \theta \) Fractional surface coverage
\( A \) Constant from Knox or Van Deemter Equation
\( a \) Constant in Langmuirian and Freundlich isotherms
\( a_i \) Constant for \( i^{th} \) component in Competitive Langmuirian Isotherm
\( a_n \) Constant in bi-Langmuirian isotherm (where \( n = 1 \) or \( 2 \))
\( B \) Constant from Knox or Van Deemter Equation
\( b \) Constant in Langmuirian and Fowler isotherms
\( b_i \) Constants in Statistical Isotherm (where \( i = 1, 2 \ldots 5 \))
\( b_j \) Constant for \( j^{th} \) component in Competitive Langmuirian Isotherm
\( b_n \) Constant in bi-Langmuirian isotherm (where \( n = 1 \) or \( 2 \))
\( C \) Constant from Knox or Van Deemter Equation
\( C_i \) Local concentration of component \( i \) in the mobile phase, \( \text{mol/m}^3 \)
\( C_{i,j}' \) Concentration at time position \( n \) and space position \( j \) in finite differences, \( \text{mol/m}^3 \)
\( C_{ip} \) Concentration of \( i \) in the stagnant film, \( \text{mol/m}^3 \)
\( D_{a,i} \) Apparent axial dispersion coefficient of \( i \) in the mobile phase, \( \text{m}^2/\text{s} \)
\( d_i \) Average residual distance
\( D_{ip} \) Diffusion coefficient in the macropores, \( \text{m}^2/\text{s} \)
\( d_{ul} \) Euclidean Distance between 2 samples
\( D_{L,i} \) Axial dispersion coefficient of \( i \) in the mobile phase, \( \text{m}^2/\text{s} \)
Nomenclature

E Residual matrix after extracting k PCs
\( e_k \) Residual distance comparing \( x_k \) with \( x \)
F Phase ratio
\( G_{nj}^{\prime} \) Concentration at time position \( n \) and space position \( j \) in finite differences, mol/m³ where \( G=(C+Fq)/u \)
H Height Equivalent to a theoretical plate, m
\( h \) Space increment for finite difference technique, m
k Number of PCs extracted in a model
\( k_{av} \) Average capacity factor
\( k_n \) Capacity factor for peak \( n \)
\( k_a \) Rate constant for adsorption process, mol⁻¹m³s⁻¹
\( k_d \) Rate constant for desorption process, mol⁻¹m³s⁻¹
L Column length, m
\( m \) Number of chromatograms/objects/samples (rows) in data matrix \( X \)
N Number of theoretical plates per metre
n Constant in Freundlich isotherm
\( n \) Number of variables measured to represent the outcome of an experiment (number of columns in the data matrix \( X \))
p Loadings matrix
\( p_i \) Loadings matrix of PC
\( q_i \) Local concentration of component \( i \) in the stationary phase, mol/m³
\( q_{li} \) Saturation capacity of column, mol/m³
\( R_p \) Average radius of a particle, m
\( R_s \) Resolution factor
\( s_c \) Mean residual standard deviation
\( s_k \) Residual standard deviation of sample \( k \)
\( s_{ka} \) Spread of the scores along each PC
\( t \) Scores matrix
\( t_{\text{lower,}a} \) Lower limit for scores for the \( a \)th PC
\( t_{\text{min,}a} \) Minimum value of the object scores on the \( a \)th PC
\( t_{\text{upper,}a} \) Upper limit for scores for the \( a \)th PC
\( t_0 \) Column hold up time, s
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_i$</td>
<td>Scores matrix of PC$_i$</td>
</tr>
<tr>
<td>$t_n$</td>
<td>retention time of a chromatographic peak n, s or min</td>
</tr>
<tr>
<td>$t_{new}$</td>
<td>new time value for each of the 300 absorbance values in any chromatogram, s or min</td>
</tr>
<tr>
<td>$t_{xo}$</td>
<td>original time value for each of the 300 absorbance values, s or min</td>
</tr>
<tr>
<td>$u$</td>
<td>Linear flow velocity, m/s</td>
</tr>
<tr>
<td>$V_0$</td>
<td>Void volume, m$^3$</td>
</tr>
<tr>
<td>$V_s$</td>
<td>Volume of Adsorbent, m$^3$</td>
</tr>
<tr>
<td>$w_b$</td>
<td>width of a chromatographic peak profile at half the peak height, min</td>
</tr>
<tr>
<td>$w_n$</td>
<td>baseline width of a chromatographic peak profile n, min</td>
</tr>
<tr>
<td>$W$</td>
<td>Weights vector for use in PLS</td>
</tr>
<tr>
<td>$X$</td>
<td>Mean-centred data matrix for use in both PCA and PLS</td>
</tr>
<tr>
<td>$Y$</td>
<td>Mean-centred data matrix for use in PLS</td>
</tr>
<tr>
<td>$x_c$</td>
<td>Centre of gravity</td>
</tr>
<tr>
<td>$x_s$</td>
<td>Sample vector for comparison with $x_c$</td>
</tr>
<tr>
<td>$X_m$</td>
<td>Mean chromatogram</td>
</tr>
<tr>
<td>$X^T$</td>
<td>Transpose of Data matrix X</td>
</tr>
<tr>
<td>$z$</td>
<td>axial dimension, m</td>
</tr>
<tr>
<td>$Z$</td>
<td>Covariance matrix</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/D</td>
<td>Analog to Digital converter</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AU</td>
<td>Absorbance units</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine Serum Albumin</td>
</tr>
<tr>
<td>C</td>
<td>Abbreviation for sample concentration variable</td>
</tr>
<tr>
<td>CG300</td>
<td>Polystyrene matrix manufactured by Tosohaas</td>
</tr>
<tr>
<td>CG71</td>
<td>Methacrylate matrix manufactured by Tosohaas</td>
</tr>
<tr>
<td>cov(X)</td>
<td>Covariance of matrix X</td>
</tr>
<tr>
<td>EBR</td>
<td>Electronic Batch Recording</td>
</tr>
<tr>
<td>F</td>
<td>Abbreviation for flow rate variable</td>
</tr>
<tr>
<td>FACP</td>
<td>Frontal Analysis by Characteristic Point</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>GC</td>
<td>Gas Chromatography</td>
</tr>
<tr>
<td>HELP</td>
<td>Heuristic Evolving Latent Projections</td>
</tr>
<tr>
<td>HETP</td>
<td>Height Equivalent to a Theoretical Plate</td>
</tr>
<tr>
<td>HIC</td>
<td>Hydrophobic Interaction Chromatography</td>
</tr>
<tr>
<td>HPLC</td>
<td>High Performance Liquid Chromatography</td>
</tr>
<tr>
<td>IAS</td>
<td>Ideal Adsorbed Solution</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid Chromatography</td>
</tr>
<tr>
<td>LEV</td>
<td>Log-eigenvalue</td>
</tr>
<tr>
<td>LHS</td>
<td>Left hand side</td>
</tr>
<tr>
<td>MB</td>
<td>Megabyte</td>
</tr>
<tr>
<td>MPCA</td>
<td>Multiway Principal Component Analysis</td>
</tr>
<tr>
<td>NIPALS</td>
<td>Non-Iterative Partial Least Squares</td>
</tr>
<tr>
<td>P</td>
<td>Abbreviation for pH variable</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Principal Component Regression</td>
</tr>
<tr>
<td>PLRP 1000</td>
<td>Polystyrene matrix manufactured by Polymer Labs</td>
</tr>
</tbody>
</table>
PLS  Partial Least Squares
PRESS  Prediction Error for the Sum of Squares
RAM  Random Access Memory
rDNA  Recombinant DNA
RHS  Right hand side
RSD  Residual Standard Deviation

$RT_A_{centre}$  Mean Retention time for erythromycin A (centre point chromatograms), s
$RT_A_{test}$  Retention time for erythromycin A for test chromatogram, s
SEC  Size Exclusion Chromatography
SIMCA  Soft Independent Modelling of Class Analogy
SSE  Sum of squares of errors
SSE$_{\text{max}}$  Maximum sum of squares of errors
T  Abbreviation for temperature variable
u  Abbreviation for flow rate variable
UCL  University College London
UV  Ultra Violet
V  Abbreviation for sample volume variable
1. Introduction

1.1. Summary
1.2. Introduction
1.3. Process Integration
1.4. Analytical and Preparative Chromatography
  1.4.1. Background to HPLC
  1.4.2. Chromatographic Parameters when scaling up
  1.4.3. Chromatographic Techniques.
    1.4.3.1. Hydrophobic interaction chromatography
    1.4.2.2. Ion exchange chromatography
    1.4.2.3. Size exclusion chromatography
    1.4.2.4. Affinity chromatography
    1.4.2.5. High Performance liquid Chromatography
  1.4.3. The modelling of Liquid Chromatography
1.5. Monitoring and Control in Bioprocesses
  1.5.1. The benefits of computers in monitoring and control
  1.5.2. Electronic Batch Recording (EBR).
1.6. Artificial Intelligence in bioprocessing
1.7. Principal Component Analysis
  1.7.1. General application of PCA to Bioprocesses
  1.7.2. PCA Applied to Chromatography data
1.8. Thesis Aims
1. Introduction

1.1. Summary
This chapter examines the importance of developing new and improved techniques for the analysis of complex bioprocess data, specifically from the chromatographic stage in the purification of a biomolecule. Artificial intelligence methods which have been successful in some areas of bioprocess monitoring and control are surveyed. Although detailed in subsequent chapters, a brief introduction is given to chromatography theory and Principal Component Analysis (PCA), explaining why PCA has potential for the analysis of complex chromatographic data and forms the basis of this thesis.

1.2. Introduction
Because of the dynamic nature and non-linearities typical of bioprocesses, there is a need for novel techniques to model the unit operations comprising such sequences (Nugent and Olson, 1990). It is important that these techniques should provide the user with quality information since this helps address some of the problems encountered in areas such as process monitoring, control and in making scale-up predictions. The main purpose of the work presented in this thesis is to make use of the multivariate statistical technique of Principal Component Analysis to provide such information from one part of a bioprocess-the chromatographic stage. This step has been chosen for analysis since it is difficult to predict its behaviour adequately with existing mathematical techniques.

1.3. Process Integration
For the sake of simplicity bioprocesses may be divided into two sections - upstream which involves cell growth and product generation, and downstream where the aim is the recovery, isolation and purification of product. Downstream processing operations handle much smaller volumes than upstream e.g. the volume of material used for each chromatographic run may be in excess of 1000 times less than the original volume in the fermenter [Sofer and Nyström (1989)]. Going from upstream
to downstream reduces the number of components and mixtures become more homogeneous. Upstream stages tend to be much more automated and computer-controlled than the downstream operations. This division occurs because the upstream unit operations are concerned with product maximisation whilst the downstream operations are mainly concerned with product purification and hence quality control. Several other differences exist between upstream and downstream operations [Wheelright (1991)] and are summarised in Table 1.1.

The effective integration of the upstream and downstream sections is essential in bioprocesses. However due to the traditional segmentation between upstream and downstream operations in bioprocessing sub-optimal processes can occur. The integration of both sections prior to optimisation should lead to a more efficient bioprocess [Biedermann (1993)].

<table>
<thead>
<tr>
<th>Upstream</th>
<th>Downstream</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live organisms</td>
<td>No organisms</td>
</tr>
<tr>
<td>Large volumes</td>
<td>Progressively smaller volumes</td>
</tr>
<tr>
<td>Many components</td>
<td>Progressively fewer components</td>
</tr>
<tr>
<td>Heterogeneous mixture</td>
<td>Homogeneous solution</td>
</tr>
<tr>
<td>Capital intensive</td>
<td>Labour intensive</td>
</tr>
<tr>
<td>Automated</td>
<td>Manual</td>
</tr>
<tr>
<td>Few operations</td>
<td>Many operations</td>
</tr>
<tr>
<td>Lower cost materials</td>
<td>Higher cost materials</td>
</tr>
</tbody>
</table>

Table 1.1 Differences between upstream and downstream processes (Wheelright, 1991).

This thesis focuses on one of the unit operations of a bioprocess - the chromatographic stage. The following section provides an introduction to chromatography which is often the final purification stage of a bioprocess.

1.4. Analytical and Preparative Chromatography

Liquid chromatography may be defined as a means of separation of solutes by making use of their different adsorption and desorption characteristics when a number of
solutions are percolated through a packed column, which usually contains a solid adsorbent (Subramanian, 1991). One of the most common types of chromatography is reversed-phase high performance liquid chromatography (HPLC). The reversed-phase HPLC separation of erythromycin forms the basis of the work in this thesis. Having first been developed for analytical usage, HPLC is being increasingly used as an industrial purification technique (preparative and process scale). There are many differences between analytical and preparative HPLC. Some of these are outlined in Table 1.2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Analytical</th>
<th>Preparative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile phase</td>
<td>Wide range</td>
<td>More restricted</td>
</tr>
<tr>
<td>Flow rate</td>
<td>Often theoretical optimum</td>
<td>Often higher than theoretical optimum</td>
</tr>
<tr>
<td>Back pressure</td>
<td>Unimportant within system constraints</td>
<td>Important</td>
</tr>
<tr>
<td>Speed of run</td>
<td>Important (short)</td>
<td>Important (longer)</td>
</tr>
<tr>
<td>Sample load</td>
<td>Non-overload</td>
<td>Overload</td>
</tr>
<tr>
<td>Sample volume</td>
<td>Small</td>
<td>Large</td>
</tr>
<tr>
<td>Temperature control</td>
<td>Sometimes desirable</td>
<td>Less common</td>
</tr>
<tr>
<td>Sample stability</td>
<td>Important</td>
<td>Very important</td>
</tr>
<tr>
<td>Quantity of product</td>
<td>μg</td>
<td>g</td>
</tr>
<tr>
<td>Typical column dimensions</td>
<td>25cm×4.6mm</td>
<td>90cm×60mm</td>
</tr>
</tbody>
</table>

Table 1.2. Some of the operational differences between analytical and preparative HPLC. (Table from Subramanian, 1991).

1.4.1. Background to HPLC

Tswett (1903) is credited with discovering chromatography when he observed separation of chloroplast pigments from a series of coloured bands in a packed column. In the 1940s, Martin and Synge devised their theory of partition chromatography which led to Martin's Nobel prize for chemistry. This was essentially liquid-liquid chromatography and in 1952 Martin and James invented gas-liquid chromatography which provided better efficiencies.
In the 1960s, Giddings compared and contrasted all chromatography systems and pioneered the shift towards liquid chromatography (LC). In the late 1960s, from the enhanced understanding about optimum conditions needed for LC, HPLC was devised. HPLC was able to utilise new precision components including accurate pumping systems, columns with precisely machined internal surfaces and efficient packing and detectors capable of detecting very low concentrations of solutes.

1.4.2. Chromatographic Parameters when scaling up

In attempting to scale up a chromatographic separation, it is naturally desirable to make use of small-scale chromatographic data from which to predict production-scale behaviour. All production-scale separations start with analytical runs and hence when scaling-up, it is important to consider several important parameters which will aid the transition in scale (Subramanian, 1991).

Preparative columns are normally operated above optimum flow rate and overloaded to increase throughput. Both these reduce the column efficiency, expressed as a number of theoretical plates. The yield achieved should be maximised to give the required purity e.g. 99%+ etc. There are several parameters used to assess chromatographic performance which include:

Resolution, \( R_s \):

\[
R_s = \frac{2(t_2 - t_1)}{w_1 + w_2} = \frac{1}{4} (\alpha - 1)\sqrt{N} \left( \frac{k_{av}}{k_{av}^* + 1} \right)
\]  

(1)

Capacity factor, \( k_{av}^* \):

\[
k_{av}^* = \frac{t_1 - t_o}{t_o}
\]  

(2)

Relative retention value, \( \alpha \):

\[
\alpha = \frac{k_1^*}{k_2}
\]  

(3)

Column Efficiency, \( N \):

\[
N = 5.54 \left( \frac{t_1}{w_{1/2}} \right)^2
\]  

(4)
where $t_1$ and $t_2$ are the retention times of two different peaks with peak widths $w_1$ and $w_2$. $t_s$ is the retention time of a non-binding material and $w_{1/2}$ is the peak width at half peak height. As a guide, a well-packed analytical column may have a value of $N$ around 30,000 plates per metre (Subramanian, 1991). There may be a reduction in efficiency when scaling up mainly due to the fact that uniformity of packing is more difficult to achieve at larger scales for several reasons e.g. the wall effects reduce with increasing diameter and the weight of packing means that the bottom of the bed experiences more compression than the top.

The modelling of chromatographic behaviour is especially problematic. This is particularly the case when a mixture has many components and when the sample size is so large that the adsorption process is non-linear. The problems of modelling non-linear chromatography are detailed in Chapter 2.

1.4.3. Chromatographic Techniques.

The efficient large-scale purification of biomolecules has become very important over the past 10-15 years. Purification schemes often begin with homogenisation and precipitation stages which are well-suited to dealing with large quantities of material. Chromatography is normally the final stage of a bioprocess. Some of the most common chromatographic techniques are outlined in this section. These are *Ion exchange chromatography*, *Hydrophobic interaction chromatography*, *Affinity chromatography* and *Size exclusion chromatography*. High Performance Liquid Chromatography is quickly gaining prominence for the purification of high-purity specification biomolecules like antibiotics and hormones such as insulin. The following section will outline the advantages HPLC over these techniques and briefly explain how it works.

1.4.3.1. *Ion exchange chromatography*

Ion exchange chromatography involves the reversible exchange of ions with an ion exchange media. Many biological molecules have a net charge, the size and sign of which depend on the pH of the solution. It is this property which allows selective adsorption onto a matrix, either anion or cation. The strength and selectivity of the
binding are governed by the pH and ionic strength of the mobile phase and the iso-electric point of the protein. Bound substances are then eluted sequentially either by increasing the ionic strength or changing the pH. This technique is selective and can be used to isolate proteins in one step [Sofer and Nystrom (1989)].

1.4.3.2. Hydrophobic interaction chromatography

Hydrophobic interaction chromatography (HIC) is a technique whereby components are separated due to the different strengths of hydrophobic interactions with an uncharged adsorbent media containing hydrophobic groups. The main advantage of HIC is its ability to separate components of similar size and charge on the basis of their differences in hydrophobicity. Most proteins contain hydrophobic regions on their surface which allow interaction with a hydrophobically derivatized matrix in aqueous conditions. Phenyl, butyl and octyl residues are the common ligands used in HIC and the hydrophobic properties of phenyl matrices are suitable for purifying most proteins under mild conditions.

1.4.3.3. Affinity chromatography

This process relies on the selective and reversible formation of a complex between the solute to be recovered and a complementary binding substance (ligand) immobilised on an adsorbent matrix. Many types of ligand have applications especially those showing high specificity towards a particular type of molecule (e.g. monoclonal antibodies) and group-specific ligands exhibiting specificity towards classes of biomolecules (e.g. immobilised metal ions).

Because of the high selectivity of affinity media, large volumes of dilute product solution can be processed and the number of processing steps can be reduced. Although the capacity of the gel is not reduced by the binding of foreign proteins as in other chromatography methods, affinity chromatography is more susceptible to fouling and has a lower capacity.

1.4.3.4. Gel Filtration or Size exclusion chromatography
Size exclusion chromatography (SEC) also known as gel filtration or gel permeation is predominantly governed by the 'molecular sieve effect' where the 3-D network in the beads hinders diffusion of the solute into the interior liquid contained within the gel beads. This hindrance is related to the molecular dimensions of the solute and the density of the network. Totally excluded solute molecules, passing only through the void volume of the gel bed, are eluted first, the other molecules following in increasing order of intraparticle diffusivity. Fractionation on the basis of molecular size is achieved.

Column loading is usually applied at low ionic strength and neutral pH and the product is eluted using a salt gradient or by reducing pH.

1.4.4. Chromatographic selection and integration

A publication by Bonnerjea et al (1986) focused on the use of these four common chromatographic techniques. This was done by analysing data from 100 research papers in an attempt to see which chromatographic techniques were most widely used and how they were integrated into purification schemes. Gel Filtration was almost always used as the final step in a purification scheme. Ion-Exchange was seen to be very common, being used in 75% of all purification schemes. Affinity Chromatography was used in 60% of schemes and HIC was the least popular, only being used in 33% of schemes.

There were definite trends observed in the paper Bonnerjea et al (1986) which showed the order which the chromatographic methods were placed in a purification scheme. Ion-exchange was often the first step and Gel Filtration the last step with either (or both) of Affinity and HIC being intermediate steps. The main reason why Ion-Exchange is usually the first step is that the stationary phase media is generally less expensive than the other methods. It is also efficient at reducing protein loads and removing fouling substances. Gel Filtration, separating by molecular size is a popular final purification stage, despite its low capacity for loaded protein. This is because it deals well with the removal of self-aggregates of otherwise purified proteins (Bonnerjea et al, 1986).
Figure 1.1. shows the average and maximum purification factors which can be obtained using the four chromatographic techniques.

![Figure 1.1. The mean and maximum purification factors expected for four common types of chromatographic separation.](image)

**1.4.5. High Performance Liquid Chromatography**

High Performance Liquid Chromatography (HPLC) is gaining increasing popularity in the purification of high purity biomolecules. HPLC achieves high resolution separations of complex mixture and gives this much faster than in other forms of chromatography (Bidlingmeyer, 1992). It is particularly beneficial when the materials to be separated are very similar (e.g. isomers) and when very pure products are needed. HPLC utilises precision components including accurate pumping systems and columns (usually stainless steel) with precisely machined internal surfaces. It has only been in the last 10-15 years that these high precision components have been available for large-scale production purposes.

Reverse phase chromatography is the most common technique used in high performance liquid chromatography. In conventional chromatography the stationary phase is polar and mobile phase non-polar, but in reverse phase chromatography the

---

1 Purification Factor = \(\frac{\%\text{Purity}}{\%\text{Impurity}}\). A 99.9\% purity would have purification factor=999.
opposite is true and solutes attach more strongly to the mobile phase than the stationary phase. The components with a higher polarity elute more quickly from the column than the less polar ones which are retained longer by the column.

Another problem typical of bioprocesses are those of monitoring and control. The following section looks at why computer techniques are being used for analysis of bioprocess data in order to improve such monitoring and control.

1.5. Monitoring and Control in Bioprocesses

In bioprocesses, the regulatory control of variables like temperature, pressure, liquid level and nutrient feed flow rates can be handled using conventional control techniques. However since the biochemical processes involved in production are non-linear and dynamic, conventional control techniques cannot really be used as a secure method of regulating bioprocesses to meet process objectives [Wilson (1991)].

Over the past years, the use of fermentation for the production of a wide range of molecular weight compounds has increased sharply. One of the main goals is to devise methods which will increase yield. This has traditionally been done by selecting micro-organism strains with elevated performance or by optimising operating conditions within the bioreactor. The last few years have seen an upsurge of monitoring and control of biochemical growth conditions within culture media [Merbel et al (1996)].

Traditionally analyses have operated off-line which involves sampling and subsequent transfer of samples to a laboratory for analysis. If analytical equipment is located near to the process the analysis is described as being at-line. These methods are laborious and time-consuming and not attractive because of these inefficiencies.

In fermentation, non-evasive techniques for monitoring involving spectroscopic probes are attractive but there are very few applications of these in-line techniques. These are usually based on electrochemical detection or using biosensors and the limited use is due to problems associated with fouling of sensors by broth and the
inability to cope with sterilisation [Merbel et al (1996)]. With on-line methods, a representative sample is withdrawn from the process, pre-treated and introduced directly into an analytical system which is interfaced with the bioreactor. Most applications of on-line control in bioprocessing have involved Expert Systems, a form of Artificial Intelligence (see Appendix A2).

One of the reasons for the lack of application of conventional control in chromatography is that models describing separations are complex and based on the physical properties of the system. For process monitoring these models must be very accurate implying the complexities will increase and the computer power needed will be substantial. The fact that biochemical processes result in interactions between variables leads to the need for alternative methods for the rapid monitoring of chromatography. The advent of artificial intelligence (AI) (see Appendix 2) and increased use of multivariate statistics have enhanced overall productivity in bioprocesses although applications to chromatography have been limited.

1.5.1. The benefits of computers in monitoring and control

The advantages of using computers for the monitoring and control of chemical and biochemical processes are enormous. The linkage of computers and monitoring devices provides a means for the monitoring and storage of process information. With this flow of data, it is possible to specify operations within a sequence in an integrated manner in order to obtain improved yield and productivity. The stored information is accessible for control engineers to implement optimal control strategies, for troubleshooting purpose and in order to achieve compliance with good manufacturing practice. Cott and Macchietto (1989) list five benefits derived from using computers to aid process operation and integration which are:

1. Increased production obtained from reduced cycle times, higher yields and improved planning
2. Improved product quality, achieved through greater repeatability of processing tasks from batch to batch
3. Reduced operating costs by better utilisation of raw materials, energy, utilities, operating staff and processing units
4. Improved plant and product safety by closer monitoring and faster control response
5. Higher degree of flexibility allowing reduced response times to changes in product demands, raw materials etc.

N.B. *The previous list is useful but contains some repetition and possible inaccuracies. E.g. the improved planning mentioned in 1. Should, I believe been placed in 3.*

Although the above study dealt with chemical plants, some benefits will apply to bioprocesses. Three papers by Locher et al (1991, 1991 and 1992) and one by Sonnleitner et al (1991) offer similar guidelines but concentrate on the *sui generis* features of bioprocesses which are not typical of other processes. These publications highlight the importance of the need for a sound understanding of biochemical reactions or metabolic control mechanisms in the installation of effective bioprocess control. The papers also recognise the need for effective integration between upstream and downstream operations, proposing that ‘excellent automation will come from exploitation of currently available logistic and electronic tools and the development of sensors and intelligent analytical systems that are appropriate for use under aseptic conditions in a technical environment’.

1.5.2. *Electronic Batch Recording (EBR).*

The Food and Drug Administration (FDA) specify that pharmaceutical batch production records must be kept for a minimum of 7 years. Traditionally this has been paper-based but more recently Electronic batch recording (EBR) has become more universal (Dean and Symcox, 1995). The main benefits appear to be ease of handling and storage of data and hence improvements in production efficiency and quality control. Databases can also be used for reasons such as fault diagnosis, automatic resource allocation and job-scheduling (Endo and Nagamure, 1987). It is naturally of benefit if some of this archived information could be used to provide rapid regulatory information for future batches from historical data. This archived data could be analysed by statistical techniques to enhance the predictive abilities, e.g. In his thesis Chandwani (1995) attempted to use the multivariate statistical
technique of Principal component analysis (PCA) to analyse a mimicked archive of chromatograms. Test chromatograms were compared with the archive and efforts were made to establish whether or not they were in process specification.

The following sections looks at some of the advanced methods which are being used to aid data analysis techniques in bioprocessing including chromatography. Effectively these can be divided into Artificial Intelligence (AI) techniques and multivariate statistical techniques. Although it is not used for the work in this thesis, the next section on Artificial Intelligence is included for completeness.

1.6. Artificial Intelligence in bioprocessing

Artificial intelligence (AI) is that branch of computer science concerned with designing intelligent computer systems where intelligence is directly associated with human intelligence [Barr and Feigenbaum (1982)]. Winston (1984) defines AI as the study of ideas that enables computers to be intelligent. There have been recent advances producing numerous expert system shells, the most widely acknowledged being G2 (Gensym Corporation, Cambridge, MA, USA) which has been adopted in many industries and has the facility to incorporate all facets of AI namely expert systems, neural networks and fuzzy logic [Kama and Breen (1989), Joseph et al (1992), Willis et al (1992)] and also many algorithms and simulation tools. Since it is not used in this thesis, a review of the uses of Artificial Intelligence in bioprocessing is given in Appendix 2.

This appendix shows the utility of AI applications in some aspects of bioprocessing. Most applications have been in fermentation and the few downstream application such as for chromatography, have been met with limited success. This is mainly due to the complexities associated with such approaches. It will be shown in the next section that multivariate techniques like Principal Component Analysis have greater potential for bioprocessing applications. These techniques are more straightforward and have been met with greater success especially with chromatographic applications.

1.7. Principal Component Analysis
Perhaps the most feasible method for the analysis of the complex relationships between bioprocess variables is with the appropriate usage of advanced multivariate statistical techniques [Mellinger (1987)]. These techniques are powerful because they are able to detect subtle changes from the interacting variables such as those which influence the shapes of chromatography profiles.

Principal Component Analysis (PCA) is a standard multivariate statistical method, originally developed by Pearson (1901) and is used to extract meaningful information from correlated data. A thorough explanation of the theory behind PCA (and summaries of other multivariate techniques) is given in Chapter 3. Only a brief resume of PCA theory is given here. The section also details recent applications of PCA in bioprocessing with a particular focus on chromatography.

Since its initial application in psychology, the approach has been applied to problems which range from the interpretation of ecological data to taxonomy. Given many sets of similar data, the method eliminates redundant information and retains only the main features common to all the sets [Wold 1978]. A fundamental requirement of PCA is that the data must be in a matrix form. The main goal of PCA is to establish relationships between the m rows (experimental samples) and n columns (variables) of a matrix with dimensions m×n. These relationships are able capture the essential patterns within the data. The relationship between the m samples are known as the scores and the relationship between the n variables are called the loadings. It is particularly the scores plots which have made PCA a very strong method for analysis of complex industrial processes [Jokinen (1994)].

This approach has been used many times in a wide range of disciplines ranging from astronomical data interpretation to the detection of similarities between tea samples with respect to their constituents in order to help with quality control [Liu et al (1987)]. Heyer and Schloerb (1997) demonstrated PCA to be a valuable tool in astronomy. Large amounts of spectroscopic imaging observation data of the interstellar medium were analysed and the PCA models generated allowed the extraction of physical features to be made, as well as the filtering out of random noise.
Other astronomical applications include using PCA for galaxy classification (Murtagh and Heck, 1987).

There have been numerous applications of PCA in a process environment. e.g. in chemical engineering PCA is regularly used as a tool in sensor fault detection (see Dunia et al, 1996 and Wise and Gallagher, 1996).

1.7.1. General application of PCA in Bioprocesses

Nearly all modern industrial chemical and biochemical processes are heavily computerised and so produce large volumes of data. By studying and analysing this data, understanding of processes is enhanced which leads to higher quality products and increased profitability (Kosanovich and Piovoso, 1991). Today’s competitive marketplace demands the need to produce high quality products and any tool that can provide a measure of performance of a process is an integral and welcome addition to any manufacturing operation. Typically, this measurement is the output of a model defined by comparing true outputs with the outputs predicted from the process model.

Since 1990, there has been a marked increase in the use of PCA and other statistical techniques for the analysis of bioprocess data. The powers of PCA have been used for fermentation monitoring and control, for which there are several examples including Saner and Stephanopolous (1992), Keller et al (1994), Ignova et al (1995) and Kurtanjek (1995). Ignova et al used PCA combined with Neural Networks to correlate the performance of the main production fermenter in order to identify poorly performing fermenters at an early stage. Current fermentation data was compared with 20 historical fermentations. Linear PCA was found to be incapable of distinguishing between well and badly performing fermentations in the scores plots, however when PCA was combined with the neural network, the two types of fermentation were distinguishable in the scores plots.

1.7.2. PCA Applied to Chromatography data

Kärsnäs and Lindblom (1992) investigated the selectivities of different Hydrophobic interaction (HIC) media for protein separation. The use of PCA revealed that
different media could be divided into several groups via the resultant scores plots. For example, some media were selecting mainly according to protein hydrophobicity whereas with other media, the charge (or lack of charge) on the protein was the most important factor. PCA and other multivariate techniques were found to be valuable tools in understanding and optimising HIC.

An early use of statistical techniques in the analysis of chromatograms was demonstrated by Fernando Faigle et al (1991). The effect of column temperature on overlapping peaks of mixtures of toluene, isoctane and ethanol was investigated using Gas Chromatography. The statistical techniques employed were Partial Least Squares (PLS) and Principal Component Regression (PCR), both of which are similar to PCA and involve the use of scores and loadings plots. Although only a small number of samples were used (14) at each of three temperatures (105, 120 and 130 °C), this paper highlighted the potential of statistical techniques for the analysis of chromatographic data by revealing distinct clusters via scores plots. The study also used complete chromatography profiles with 41 detection readings. However, the only variable tested was temperature and so the potential of statistical techniques like PCA to analyse multivariate and interacting bioprocess data was not investigated.

A study by Malmquist and Danielsson (1994) appreciated the fact that different combinations of process variables lead to different shaped profiles which make data interpretation a complex issue. This paper also realised that the best way to represent ‘variables’ on which to perform PCA would be to use detection readings at regular time intervals and not resolution criteria such as retention time and peak width etc. The study appreciated the problems of base-line shift of peaks evident, especially with variations in flow rate which may cause significant information to be concealed. Proposed methods to ‘align’ successive chromatograms included a ‘time warping’ function as proposed by Reiner et al (1979), but this was only used for visual inspection and not for use with multivariate techniques. Other methods included using compression/expansion of the time scale [Andersson and Hämäläinen (1994)]. Another proposition was ‘normalisation to constant area’, but the method employed by Malmquist and Danielsson was a combined procedure. The first step involved
comparing the sample chromatogram with a target. The second step involved a retention time correction shift and the last stage involved fine tuning correction factors. This study was useful for the understanding of effective pre-processing prior to the implementation of PCA.

Liang et al (1994) also appreciated the problems of comparing chromatograms of differing base-line positions. PCA again was performed on complete chromatograms and was performed locally on selected regions of chromatograms in order to find chromatographic regions with similar concentration patterns.

Malmquist (1994) made use of PCA for the analysis of peptide mapping which is important in the quality control of \textit{rDNA} derived proteins. Once again the whole chromatogram was used as input data thus eliminating the need for retention and resolution data from each profile. Peptide mapping is a very demanding analytical technique and is vital for the assessment of amino-acid sequence integrity in proteins. Multivariate techniques like PCA are able to provide an unbiased evaluation and are capable of handling experimental variations. A set of reference chromatograms were used and PCA was performed on the dataset and test samples were classified using SIMCA (soft independent modelling of class analogy). PCA was able to identify relationships between test samples and historical data but no attempt was made to analyse the effect of varying the operating conditions.

Some of the most interesting and efficacious applications of PCA in the analysis of chromatogram data has been conducted by Kvalheim and his co-workers (predominantly with Liang). There have been many such publications throughout the 1990’s.

These publications focus on the problem of resolving co-eluting components into the pure constituents. Ensuring peak purity in liquid chromatography is a very demanding problem and impurities of less than 1% need to be detected.
Kvalheim and Liang (1992) introduce a new technique (and used in the subsequent papers) to help solve this problem. It is called Heuristic Evolving Latent Projections or HELP and involves the use of PCA for the analysis of multi-wavelength chromatograms. A single chromatographic analysis is performed at many different wavelengths e.g. Kvalheim and Liang (1992) used 32 wavelengths at 5 nm intervals between 210 and 365 nm for a separation of two isomers. The technique has the advantage of being able to determine the number of species present in a mixture from a single analysis, as well as being able to resolve the mixture into spectra and concentration profiles of the pure constituents.

The data presentation prior to PCA for the HELP method is as follows. Each chromatogram is scanned at m wavelengths and contain n time points. The m wavelengths are the columns of the matrix and the n time points are the rows. (This is analogous with the type of matrix used in this thesis where the chromatogram samples comprise the rows of the matrix and the time points comprise the columns.

The data matrix is then decomposed using PCA and the resulting scores and loadings plots reveal information about the numbers of species in the mixture. Local PCA is also performed on critical regions of the chromatograms e.g. the from and back edges of peaks to reveal more qualitative information.

Noise was a major consideration in the work described in these publications as the presence of peaks from more minor peaks species may be confused with noise. Noise effects such as those resulting from drifting baselines need to be accounted for. Such pre-treatments were focused on in Keller et al (1992). One technique for filtering noise was to use PCA on so-called zero-component regions of the chromatogram where there is a known absence of chemical species. An appreciation that some pre-processing of data was made in Liang et al (1994). Here baseline shifts of elution peaks (due to noise or otherwise) were accounted for prior to analysis.

The HELP methodology described in these papers by Kvalheim and Liang was generally validated using synthetic mixtures. Also the techniques detailed were for
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 1

analytical chromatography. Some of the principles detailed in the publications could be valuable in the resolving of peaks (and thus provide an assessment of peak purity) arising from a complex industrial separation such as the purification of erythromycin described in this thesis. The application however would require modification to allow for non-linear adsorption properties experienced when operation is under overload conditions.

A PhD thesis by Chandwani (1995) sought to examine how PCA could be used as a chromatographic pattern recognition tool. A size exclusion chromatographic (SEC) system was used to separate three proteins: myoglobin, ovalbumin and bovine serum albumin (monomer, dimer and trimer respectively). The mixture was made up synthetically to a fixed ratio of 17:2:1.

Experimental design techniques were use to investigate 4 process variables: temperature, load volume, load concentration and flow rate. The idea was to characterise the separation process using PCA. Clusterings in the resulting scores plots were used to identify those separations which were similar but not easily detectable to the human eye. Future runs could then be tested using statistical techniques like the F-test to see whether or not they were in process specification.

1.8. Thesis Aims

The aims of this thesis are to investigate the following:

- To attempt to use PCA to model the complex separation of crude erythromycin which would be similar to the type of separation which may be obtained following a fermentation.
- The importance of data pre-processing prior to PCA in order to maximise the ability of the PCA models.
- To examine ability PCA to help with predictions of realistic, non-linear separations which result when columns are operated under overload conditions.
• To compare PCA models which result when performing the same separation on columns with different stationary phases hence exhibiting different adsorption properties.
• To compare PCA models which result when performing the same separation on columns of different dimensions whilst maintaining the type of stationary phase.
• To use PCA on a set of chromatogram data generated using finite difference techniques.
• To outline the potential limits of using the PCA method for the modelling of chromatography.

1.9. Thesis Contents
The outline of the subsequent chapters in this thesis are as follows:

• Chapter 2. A theory chapter about the mathematical approach to the modelling of chromatography.
• Chapter 3. A theory chapter about Principal Component Analysis
• Chapter 4. A results chapter about the importance of pre-processing prior to PCA. These results were based on Chandwani's data from his PhD thesis (1995).
• Chapter 5. A results chapter about pre-processing of data from the erythromycin separation.
• Chapter 6. The use of PCA to model the erythromycin separation for three columns with different sizes and stationary phases. Realistic overload conditions were applied to each column.
• Chapter 7. The use of PCA to model the erythromycin separation for three columns of increasing size with the same stationary phase.
• Chapter 8. The use of PCA to model some artificial chromatogram data generated using finite difference techniques.
• Chapters 9, 10 and 11 provide an overall discussion, conclusions and future work respectively.
2. Chromatographic modelling - theory.

2.1. Introduction

2.2. The Mass-Balance Equation

2.2.1. The solutions of the mass balance equation.

2.2.1.1. The ideal model.

2.2.1.2. The Equilibrium - Dispersive model

2.2.1.3. Lumped Kinetic Model.

2.2.1.4. The effect of Mass transfer on the mass balance equation

2.3. Adsorption Isotherms

2.3.1. Linear and Non-linear chromatography

2.3.2. Single component isotherms

2.3.2.A. The Langmuirian Isotherm

2.3.2.B. The Bi-Langmuirian isotherm

2.3.2.C. The Fowler isotherm

2.3.2.D. Freundlich isotherm

2.3.2.E. S-shaped isotherms

2.3.3. Multicomponent or Competitive adsorption isotherms

2.3.3.A. The Competitive Langmuirian isotherm

2.3.3.B. The competitive bi-Langmuirian isotherm

2.3.3.C. The Ideal Adsorbed Solution

2.3.3.D. The Statistical Isotherm

2.3.3.E. The competitive Fowler Isotherm

2.3.3.F. The competitive Freundlich-Langmuirian Isotherm.

2.4. The Determination of Single Component Isotherms

2.4.1. Frontal Analysis

2.4.2. Frontal Analysis by Characteristic Point

2.4.3. Elution By Characteristic Point

2.4.4. Retention Time Method

2.5. Summary
2. Chromatographic modelling - theory.

2.1. Introduction

The differences between analytical and preparative chromatography were outlined in section 1.4. Accurate mathematical modelling of chromatographic processes is extremely complex and based around second order partial differential equations for each species. The complexities are particularly great when chromatography is used as a purification stage in a bioprocess (preparative scale) rather than when it is used as an analytical tool. This is mainly because preparative scale chromatography involves the separation of very complex mixtures which have been heavily loaded. This section focuses on these problems and explains why mathematical modelling is so complex and includes an examination of adsorption isotherms which plot the relationship between the concentration of sample at equilibrium between the mobile and stationary phases. These isotherms generally require pure components for their experimental determination. A brief insight into how adsorption isotherms are generated is given at the end of the chapter.

2.2. The Mass-Balance Equation

The mass-balance equation which describes chromatographic separations was first derived by Wicke in 1939 and further developed by Wilson (1940) and DeVault (1943), both independently of each other. The equation is given in (1) without proof which can be found in Guichon et al, 1994.

\[
\frac{\partial C_i}{\partial t} + F \frac{\partial q_i}{\partial t} + u \frac{\partial C_i}{\partial z} = D_{L,i} \frac{\partial^2 C_i}{\partial z^2}
\]  

(1)

\[C_i = \text{local concentration of component } i \text{ in the mobile phase, } \text{mol/m}^3\]

\[q_i = \text{local concentration of component } i \text{ in the stationary phase, } \text{mol/m}^3\]

\[z = \text{axial dimension, m}\]

\[D_{L,i} = \text{axial dispersion coefficient of } i \text{ in the mobile phase (this includes contribution to axial dispersion due to molecular diffusion and non-homogeneity of flow), m}^2/\text{s}\]

\[F = \text{phase ratio } = (1-\varepsilon)/\varepsilon = \text{volume stationary phase/volume mobile phase}\]

\[u = \text{Linear Velocity of mobile phase, m/s}\]

The mass balance equation is a second order partial differential equation. The first two terms on the left-hand side of the equation are accumulation terms in the mobile and stationary phases. The third term on the left hand side is a convective term. The term on the right hand side is a diffusion term. Each component present in the system
requires its own equation, including each component in the mobile phase. This means that a pure component eluted by pure mobile phase must have 2 mass balance equations, although it has been shown (Riedo and Kovaitz, 1982) that the mass balance for the mobile phase can be dropped if it is pure. There are several fundamental assumptions on which equation (1) are based. These include:

- The mobile phase is incompressible. This hold well up to a pressure of 500 bar.
- The partial molar volumes of sample concentrations are the same in both mobile and stationary phases.
- The column is radially homogeneous i.e. has constant quality of packing across the width of the column. To satisfy this, appropriate flow distribution is required for larger columns.
- The mobile phase has constant viscosity.
- The axial dispersion coefficients, $D_{L,i}$ are constant however, in multicomponent non-linear systems the $D_{L,i}$ values depend on the other components. The $D_{L,i}$ is a lumped parameter and includes: packing tortuosity; anastomosis (causes turbulent mixing); $B/u$ and $A_{u}^{1/3}$ of the Knox equation or $A$ and $B/u$ from the Van Deemter equation (Guiochon et al, 1994)
- The solvent is not adsorbed.
- No heat effects which would require a heat balance.

2.2.1. The solutions of the mass balance equation.

When solving equation (1), initial and boundary conditions must be pre-defined, otherwise there would be an infinite number of solutions. One of the commonly used boundary conditions is

$$C_i(z, t=0) = 0 \quad \text{for } 0 \leq z \leq L \text{ where } L \text{ is the column length} \quad (2)$$

For other commonly used initial and boundary conditions, consult Guiochon et al (1994). There have been several models proposed which rely on simplifications of equation (1). Some of these models are given as follows:
2.2.1.1. The ideal model.

Chromatographic separations can be either ideal or non-ideal. With ideal chromatography, the assumption is made that the columns have infinite efficiency (number of theoretical plates) therefore no axial dispersion or band broadening occurs as material passes through the column. Ideal chromatography also assumes there is a constant equilibrium between mobile and stationary phases and so the elution profiles are only a function of the equilibrium thermodynamics. Ideal chromatography is of very little practical interest.

By contrast, non-ideal chromatography is much more realistic, firstly because columns have a finite efficiency and so separated components will undergo band broadening as they travel through columns. There will also be band broadening due to the kinetics of the adsorption/desorption process. There are several more minor (but significant) factors which are present in non-ideal systems which serve to enhance the complexities of the models. These include eddy diffusion, molecular diffusion and the kinetics of the adsorption mechanism intermediates. A summary of the differences between ideal and non-ideal chromatography is given in Table 2.1.

<table>
<thead>
<tr>
<th>IDEAL CHROMATOGRAPHY</th>
<th>NON-IDEAL CHROMATOGRAPHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Columns with infinite efficiency</td>
<td>1. Columns with finite efficiency</td>
</tr>
<tr>
<td>2. No axial dispersion</td>
<td>2. Axial dispersion</td>
</tr>
<tr>
<td>3. Elution profiles only a function of equilibrium thermodynamics.</td>
<td>3. Additional contribution to band broadening from equilibrium thermodynamics and from adsorption/desorption.</td>
</tr>
<tr>
<td>4. Very little practical interest</td>
<td>4. Eddy diffusion and adsorption intermediates are significant.</td>
</tr>
<tr>
<td></td>
<td>5. Realistic</td>
</tr>
</tbody>
</table>

Table 2.1. Summary of the differences between ideal and non-ideal chromatography.

The ideal model assumes that there is no axial dispersion and that the column is infinitely efficient. Furthermore it is assumed that the mobile and stationary phases are constantly at equilibrium. Equation (1) becomes:

$$\frac{\partial C_i}{\partial t} + F \frac{\partial q_i}{\partial t} + u \frac{\partial C_i}{\partial z} = 0$$

(3)
This model ignores the influence of axial dispersion and mass transfer kinetics.

2.2.1.2. The Equilibrium -Dispersive model

\[
\frac{\partial C_i}{\partial t} + F \frac{\partial q_i}{\partial t} + u \frac{\partial C_i}{\partial z} = D_{ai} \frac{\partial^2 C_i}{\partial z^2}
\]  

(4)

The equilibrium dispersive model assumes that all contributions to non-equilibrium can be lumped into an 'apparent' axial dispersion term, \(D_{ai} \). The value of \(D_{ai} \) can be evaluated as follows

\[
D_{ai} = \frac{HL}{2t_o} = \frac{Hu}{2}
\]  

(5)

Where \(H \) is the HETP (height equivalent to a theoretical plate =1/N from section 1.4); \(L \) is the column length and \(u \) is the velocity of the mobile phase. \(t_o \) is the hold up time = \(L/u \). The model further assumes that \(D_{ai} \) are constant and independent of sample concentration. This model nearly always provides an excellent approximation especially with the high efficiencies of modern columns.

2.2.1.3. Lumped Kinetic Model.

In the lumped kinetic model, the accumulation term in the mass balance equation is replaced with a kinetic equation. An example of a lumped kinetic model is the Langmuir kinetic model which is as follows.

\[
\frac{\partial q_i}{\partial t} = k_s(q_{s,i} - q_i)C_i - k_d q_i
\]  

(6)

\(q_{s,i} \) is the saturation capacity of the column and has the same units as \(q_i \). \(k_s \) and \(k_d \) are the rate constants for adsorption and desorption. There are many other kinetic-based models which are well-documented in Guiochon et al, 1994. It is evident that when columns with very high efficiencies are used, the effect of the kinetic term is negligible in the overall mass-balance.

2.2.1.4. The effect of Mass transfer on the mass balance equation

The accumulation term \(\frac{\partial q_i}{\partial t} \) in the mass balance equation (1) can be modified to take account of intraparticular mass transfer controlled by diffusion into the macropores of
the stationary phase. In their review of liquid chromatography modelling, Bellot and
Condoret (1991) report that the term can be replaced with

$$\frac{3D_{i,p}\varepsilon_p}{R_p}\left(\frac{\partial C_{i,p}}{\partial r}\right)_r = R_p$$

The diffusion coefficient in the macropores ($D_{i,p}$) is usually less than the diffusion coefficient, $D_{i,m}$, of the component in the mobile phase. A tortuosity factor, $\tau$, is used to relate the two, i.e.

$$D_{i,p} = D_{i,m}/\tau$$

$\tau$ represents the random orientation of pores and variations in pore diameter. The mass balance equation (1) can be modified further to account for more complex mass transfer phenomena and for differences in particle shape and size. These are not reported here but can be located in Bellot and Condoret (1991).

### 2.3. Adsorption Isotherms

Once the mass balance for a single component has been performed (equation 1) and then modified to account for, e.g., equilibrium kinetics, the only requirement for the determination of the shape of an elution profile is the adsorption isotherm. Adsorption isotherms are used to provide the information regarding the concentrations $C_i$ and $q_i$ of component $i$ in the mobile and stationary phases. The calculation of each single component elution profile and predicting them accurately is only a step, albeit an important and necessary one, towards the prediction of the complete elution profiles of a mixture. The following sections document some of the most common single component isotherms. More realistic, competitive isotherm models are also discussed. Some of the problems and complexities associated with such models are also highlighted.

#### 2.3.1. Linear and Non-linear chromatography

A linear adsorption isotherm is the most simple adsorption mechanism in chromatography. The equilibrium isotherms are straight lines, indicating that the
sample load size in the mobile phase is proportional to that on the stationary phase. The assumption is made that each sample's isotherm does not depend on the presence or concentration of the other samples. It is further assumed that the properties of the resulting elution profiles (i.e. retention times, shape, peak maxima etc.) are not a function of the sample load applied. Unfortunately linear chromatography can only be approximated using very small sample loads at analytical scale and thus is of little practical benefit for production.

<table>
<thead>
<tr>
<th>LINEAR CHROMATOGRAPHY</th>
<th>NON-LINEAR CHROMATOGRAPHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Linear adsorption isotherms</td>
<td>1. Non-linear adsorption isotherms</td>
</tr>
<tr>
<td>2. Shape of isotherms not ( f(\text{other samples}) )</td>
<td>2. Shape of isotherms are ( f(\text{other samples}) )</td>
</tr>
<tr>
<td>3. ( R_T ), peak shapes, peak maxima not ( f(\text{sample loads}) )</td>
<td>3. ( R_T ), peak shapes, peak maxima are ( f(\text{sample loads}) )</td>
</tr>
<tr>
<td>4. Analytical scale loads only</td>
<td>4. All preparative applications</td>
</tr>
<tr>
<td>5. Little practical relevance.</td>
<td>5. Realistic and very complex models.</td>
</tr>
</tbody>
</table>

Table 2.2. Summary of the main differences between linear and non-linear chromatography.

Non-linear chromatography on the other hand is much more realistic. Here the adsorption isotherms are non-linear, and which also depend on the presence and concentration of other samples in the mixture. The retention time, shape and maxima of the peaks are a function of the sample size of each component in the mixture. Very complex mathematical models are a necessity with non-linear chromatography. All preparative applications are modelled using non-linear models, as are all overloaded columns at analytical scale. Table 2.2 summarises the differences between linear and non-linear chromatographic systems.

Although non-linear, non-ideal systems are the most realistic and complex to describe, it is also of interest to study non-linear, ideal systems as these allow the influence of thermodynamics to be assessed without the confusion of kinetics and mass-transfer. The next section surveys the most common single components adsorption isotherms.
2.3.2. Single component isotherms

A. The Langmuirian Isotherm

One of the most common adsorption isotherm is the Langmuirian isotherm. The concentration of component i adsorbed onto the stationary phase, \( q_i \), is related to the concentration, \( C_i \), in the mobile phase as follows

\[
q_i = \frac{aC_i}{1 + bC_i}
\]  

(9)

where \( a \) and \( b \) are constants. A typical Langmuirian isotherm is given in Figure 2.1. The shape is generally referred to as 'convex' or 'concave-down'. As \( C_i \) becomes large, equation \( q_i \) will tend towards \( q_s \) which is the column saturation capacity which is the value of the asymptote indicated by the dotted line.

\[
q_s = \frac{a}{b}
\]  

(10)

Figure 2.1. An example of a Langmuirian isotherm.

Langmuir isotherms are often written in their reduced form \( \theta = \frac{q_i}{q_s} \). \( \theta \) is often referred to as the 'fractional surface coverage'. The Langmuir isotherm can be linearised in several ways. Combining equations 9 and 10 gives

\[
\frac{C_i}{q_i} = \frac{1}{a} + \frac{C_i}{q_s}
\]  

(11)
Plotting a graph of $\frac{C_i}{q_i}$ vs. $C_i$ allows the experimental determination of constants $a$ and $b$. The gradient is $\frac{1}{q_i}$ and the intercept is $\frac{1}{a}$.

The Langmuirian isotherm was originally developed for gas-solid equilibria. There are several assumptions which hold for gas-solid systems which do not quite apply in liquid-solid systems. These include:

- ideal solutions
- monolayer coverage of the stationary phase
- adsorption layer is ideal
- no solute-solute interactions in the monolayer
- no solvent-solute interactions.

These assumptions become less valid with increasing concentration. Nevertheless, experience has shown that the Langmuirian isotherm equation (9) is an excellent approximation for single-component equilibria in liquid chromatography (Guiochon et al, 1994). Examples in the literature of successful use of Langmuirian isotherms include Jacobson et al,(1984) and Huang and Horvath (1987). Jacobson et al fitted accurate Langmuirian isotherms to a series of hydrophilic organic compounds comprising of p- and o-toluidine and o- and p-cresol with a Spherisorb ODS-2 stationary phase. Huang and Horvath used Langmuirian isotherms with similar success for the adsorption of many di- and tri-peptides also with a Spherisorb ODS-2 stationary phase.

Langmuirian-type isotherms with a negative $b$ are possible. These result in a 'concave-up' isotherm which approaches a vertical asymptote at $C = 1/b$ (Figure 2.2). These are often called 'anti-Langmuirian isotherms'. Figure 2.3 illustrates the differences between linear, Langmuirian and anti-Langmuirian isotherms. The first row shows the shapes of the respective isotherms. The second row shows the shapes of typical elution profiles using the assumption of an ideal model i.e. an infinitely efficient column with no band-broadening, with increasing load. The third shows the shapes of typical elution profiles with increasing load, using the more realistic non-ideal situation of band broadening from the assumption of finite column efficiency.
Figure 2.2. A concave-up Langmuirian-type isotherm.

Figure 2.3. A comparison between the effects of different-shaped isotherms on the shapes of single-component elution profiles. Linear, Langmuirian and anti-Langmuirian isotherms are compared. The first row shows typical shapes of these isotherms; the second row shows typical elution profiles resulting from an infinitely efficient column (no band broadening) and the third row shows typical elution profiles resulting from a column of finite efficiency (hence band broadening).
Several other isotherms have been proposed which are an improvement on Langmuirian isotherms in some situations. These include:

**B. The Bi-Langmuirian isotherm**

In many cases, the surface of the stationary phase is not homogeneous. For example a surface may be covered with two kinds of sites which behave independently. The resulting equilibrium isotherm is formed from the addition of the two independent contributions from the two types of site. Since in most cases the Langmuirian isotherm is appropriate to account for single-component adsorptions onto a homogeneous surface (Guiochon et al, 1994), the following isotherm results:

\[ q_i = \frac{a_i C_i}{1 + b_1 C_i} + \frac{a_i C_i}{1 + b_2 C_i} \]  

(12)

A guide as to whether an adsorption is bi-Langmuirian is that when \( \frac{C_i}{q_i} \) is plotted against \( C_i \) (equation 11) as for an assumption of a Langmuirian isotherm, a straight line does not result. A bi-Langmuirian isotherm requires the use of two isotherm curves per species. Some good examples of bi-Langmuirian isotherms can be found in Jacobson and Guiochon, (1992) and Jacobson and Guiochon, (1990). These publications describe the adsorption of the enantiomers of mandelic acid, of the N-benzoyl derivative of glycine adsorbed on bovine serum albumin (BSA), immobilised in silica. These studies showed that the adsorption data were too curved to fit a Langmuirian model. A set bi-Langmuirian isotherms were seen to be the best model, probably due to the existence of two types of binding sites-enantioselective and non-enantioselective.

**C. The Fowler isotherm**

First designed by R.H. Fowler in 1939, the Fowler isotherm corrects for first-order deviations from the Langmuirian isotherm (Guiochon et al, 1994). For liquid-solid equilibria, the isotherm is defined as

\[ bC_i e^{-\theta} = \frac{\theta}{1-\theta} \]  

(13)
Where $\theta$ is the fractional coverage and $\chi$ is an empirical interaction energy parameter. Unlike the Langmuirian isotherm, the Fowler isotherm cannot be inverted and solved for $q_i$ in its closed form. This makes it difficult to use in combination with the mass balance equation (1) for the prediction of elution profiles. A special case of the Fowler isotherm is when $\chi = 0$ when it is equivalent to the Langmuirian isotherm. As $\chi$ increases, this isotherm becomes more linear. In a few cases, the Fowler isotherm is able to fit experimental data better than a Langmuirian one. An example is given in Zhu et al (1991) who reported that both 2-phenylethanol and 3-phenylpropanol adsorbed onto 10$\mu$m ODS silica using 50/50 methanol/water mobile phase was better modelled using a Fowler isotherm than a Langmuirian one.

**D. Freundlich isotherm**

In 1885 C. Boedeher proposed the following isotherm for the adsorption of polar compounds onto polar adsorbants.

$$ q_i = aC_i^{1/n} \quad (14) $$

Where $a$ and $n$ are constants and $1/n < 1$.

It is known as the Freundlich after H. Freundlich who popularised it in the 1920s. The Freundlich isotherm appears to account well for the adsorption of certain proteins onto ion exchangers (Guiochon et al, 1994). There are several problems with the Freundlich isotherm because it is not thermodynamically consistent. The isotherm is suggestive of infinite retention times under analytical conditions. Guiochon et al (1994) recommends that stationary phases which follow a Freundlich isotherm under experimental conditions should be avoided in both analytical and preparative chromatography.

**E. S-shaped isotherms**

These are sometimes similar to anti-Langmuirian isotherms and are frequently observed in gas-solid chromatography. At low concentrations, the shape of the isotherm is concave-upwards, indicating that the amount adsorbed onto the stationary phase at equilibrium increases more rapidly than the concentration in the mobile phase. These isotherms are rare in liquid chromatography.
There are a vast number of other suggested equations for isotherms. They include the Toth isotherm, the Unilan isotherm and the Martire isotherm. These have more applications with gas-solid equilibria. Brief descriptions can be found in Guiochon et al (1994).

2.3.3. Multicomponent or Competitive adsorption isotherms

The previous sections described single component isotherms which are required to assist in the calculation of single component elution profiles. However, in all cases of practical importance in liquid chromatography, we are dealing with multi-component mixtures and not pure components. The complexities of multi-component systems derive from the competition between the different components for interaction with the stationary phase. The surface area of the available stationary phase is linked and the total number of molecules which can be accommodated is finite. As the concentration increases, the molecules have to 'compete' for adsorption sites. Furthermore, not only are the adsorption isotherms for each component non-linear, they are dependent on each other, especially with non-dilute mixtures (Guiochon et al, 1994). Additionally the amount of a component adsorbed at equilibrium almost always decreases with increasing concentration of another adsorbed component.

This section discusses a number of mathematical models which have been used to describe competitive isotherm data. These mechanisms are mainly extensions of the single-component isotherms described in section 2.3.2. Although theoretically the competitive isotherms can be used for any number of components, it is only with binary mixtures that any real success has been achieved. The main description will be of the competitive Langmuirian isotherm. Other isotherms such as the competitive bi-Langmuirian, the ideal adsorbed solution, the statistical isotherm, the competitive Fowler isotherm and the competitive Freundlich-Langmuir isotherm will also be discussed.

A. The Competitive Langmuirian isotherm

The single-component Langmuirian isotherm model (section 2.3.2 A) can be extended to multi-component systems (Guiochon et al, 1994). With several components in a
mixture, the amount of each component adsorbed at equilibrium is naturally less than if just a single component were present. For the $i^{th}$ component of a multi-component mixture, the competitive Langmuirian isotherm is defined as

$$q_i = \frac{a_i C_i}{1 + \sum_{j=1}^{n} b_j C_j}$$

(15)

where $n$ is the number of components in the mixture. $a_i$ and $b_i$ are coefficients for component $i$. The derivation of equation (15) is given in Guiochon et al (1994) who used a binary solution to illustrate this isotherm. The ratio $a_i/b_i$ is the column saturation capacity for component $i$, i.e. the maximum amount of component $i$ which can be bound.

Competitive Langmuirian isotherms are represented by surfaces in a suitable space. For a binary mixture there will be two 3-dimensional isotherms which contain simultaneous plots of $(q_1, C_1, C_2)$ and $(q_2, C_1, C_2)$. Figure 2.4 shows such a plot for $(q_2, C_1, C_2)$. The isotherm is in the form of a plane. From any combination of $C_1$ and $C_2$ (concentrations of components 1 and 2 in the mobile phase), the concentration of component 2 in the stationary phase, $q_2$ can be calculated.

Figure 2.4. A typical example of a competitive isotherm for a binary system. Combinations of $C_1$ and $C_2$ allow the calculation of $q_2$. 

57
One of the important properties of competitive Langmuirian isotherms is that they are unable to detect a situation when the elution order of the components changes. For example the elution order of *trans*-and *cis*-androsterone switches around when sample size increases (Huang and Guiochon, 1989). This, however is a rare situation.

Instead of using multi-dimensional surfaces like the one in Figure 2.4, competitive isotherms (for binary systems) can also be represented using *contour maps*. A series of horizontal intersections are made on the isotherm surfaces (see Golshan-Shirazi et al, 1991). C	extsubscript{1} and C	extsubscript{2} are plotted against each other on 2-dimensional graphs. A family of straight lines result which are used to represent how q	extsubscript{1} and q	extsubscript{2} are related to C	extsubscript{1} and C	extsubscript{2}.

The competitive Langmuirian isotherm model is the most attractive of all proposed competitive isotherm models. It is the most simple and is able to describe a wide variety of binary systems. The ease of calculation of parameters a	extsubscript{i} and b	extsubscript{i} enhances its attractiveness. Generally speaking, this isotherm tends to account for the adsorption behaviour of similar compounds in a binary mixture e.g. geometric isomers or close homologs [Guiochon et al, 1994]. On the other hand, it has been seen that the competitive Langmuirian isotherm is not a good model or binary mixtures when the difference between the column saturation capacities for the two components exceeds 5-10% [Guiochon et al, 1994]. With the case of *cis-* and *trans-* androsterone on silica, the pure components are described very well by single-component Langmuirian isotherms. However their saturation capacities differ by 30% and so the competitive Langmuirian isotherm accounts poorly for the competitive adsorption data (Huang and G. Guiochon, 1989). In cases like these, there is a need for other competitive isotherms.

**B. The competitive bi-Langmuirian isotherm**

By analogy with the competitive Langmuirian isotherm (equation 15), when a surface is covered by two kinds of site, the two components in a binary mixture can be accounted for using the following competitive bi-Langmuirian isotherm
which is the same as for the single-component case (equation 12) but for two components. An example of a successful application of a competitive bi-Langmuirian isotherm is reported in Charton et al (1993) who used one to model the adsorptive behaviour of a 1:1 mixture of N-benzoyl-D and L-alanine.

**C. The Ideal Adsorbed Solution**

When extending Langmuirian isotherms for two pure components to a competitive Langmuirian isotherm, the saturation capacity of the column is often not the same (Guiochon et al, 1994). For this reason, the ideal adsorbed solution (IAS) model was developed by Myers and Prausnits (1965). It was extended by Levan and Vermeulen (1981) who developed a competitive binary isotherm equation which accounted for differences in column saturation capacities for the two components when the single-component isotherms are either Langmuirian or Freundlich.

The single-component Langmuirian isotherm (9) can also be written as

$$ q_i = \frac{a_i C_i}{1 + b_i C_i} $$

where \( q_i \) is the column saturation capacity for component \( i \). Levan and Vermeulen extended this model to a binary system with the equation

$$ q_i = \frac{b_i q_{i,s} C_i}{1 + b_i C_i} + \frac{C_i}{\partial C_i} \ln(1 + b_i C_i + b_2 C_2) $$

where \( q_{i,s} \) is the weighted average monolayer capacity. The terms in this isotherm can be expanded by a Taylor series which converges very rapidly. The Levan-Vermeulen equation was used by Golsham-Shirazi et al (1991) to account for the experimental result obtained from the binary separation of *cis-* and *trans-* androsterone, two very close isomers which are very difficult to separate due to having very similar initial slopes to their isotherms.
D. The Statistical Isotherm

The usual form of isotherm equations are written as the ratio of linear polynomials. It is therefore a reasonable assumption to use higher polynomials to describe adsorption behaviour. Beyond the competitive Langmuirian isotherm (equation 11), the next approximation is the ratio of two second-order polynomials (Guiochon et al, 1994).

For binary mixtures, Ruthven and Goddard (1986) derived the following competitive isotherms

\[
\frac{q_1}{q_s} = \frac{b_1 C_1 + b_2 C_1 C_2 + 2b_3 C_1^2}{1 + b_1 C_1 + b_2 C_2 + b_3 C_1 C_2 + b_4 C_1^2 + b_5 C_2^2}
\]

(19)

and

\[
\frac{q_2}{q_s} = \frac{b_1 C_2 + b_2 C_1 C_2 + 2b_3 C_2^2}{1 + b_1 C_1 + b_2 C_2 + b_3 C_1 C_2 + b_4 C_1^2 + b_5 C_2^2}
\]

(20)

These equations have 7 parameters instead of 4 in the competitive Langmuirian isotherm. 6 of the parameters can be derived from single-component isotherms and only \(b_j\) needs to be determined from the mixture. Depending on the relative values of the coefficients, the isotherm may be convex, concave or have an inflection like in an S-shaped isotherm (Svoboda, 1990).

E. The competitive Fowler Isotherm

An extension of the single-component Fowler isotherm (equation 13), the competitive Fowler isotherm is usually defined as

\[
b_i e^{-z_i (\theta_1 + \theta_2)} C_i = \frac{\theta_i}{1 - \theta_1 - \theta_2}
\]

(21)

where the parameters are as for the single-component case (equation 13). Sometimes the competitive Fowler isotherm is written with the approximation of the \(b_i\) coefficients being equal, thus reducing the degrees of freedom. An example of a successful application of the competitive Fowler isotherm can be seen in Zhu et al (1991) in which the binary mixture of phenylethanol and 3-phenylpropanol were separated on C18 silica with water/methanol mobile phase. Here the fit to experimental data was better than either competitive Langmuirian or bi-Langmuirian isotherms.
F. The competitive Freundlich-Langmuirian Isotherm.

Several hybrids of the common models have been suggested. One of the most popular is an isotherm which combines both the Langmuirian and Freundlich isotherms. The Freundlich-Langmuirian isotherm addresses the problem of strong adsorption observed at low concentrations and rapid subsidence at increasingly large concentrations (Guiochon et al, 1994). The competitive Freundlich-Langmuirian isotherm has been used to fit experimental data obtained with various dyes. An example is given in Al-Duri et al (1992).

The next section gives a brief insight into how single-component adsorption isotherms are obtained in practice.

2.4. The Determination of Single Component Isotherms

This section summarises four common methods for the determination of single component isotherms which are used nowadays. These basically involve working backwards from a set of elution profiles (determined experimentally) to solve the mass balance equation (1). It is also possible to determine competitive isotherms for two components. This is however extremely difficult and much more time-consuming than for single components (Guichon et al, 1994) and will not be discussed in this section.

2.4.1. Frontal Analysis.

Isotherms have been traditionally measured by allowing stationary phases to reach equilibrium with a series of known sample concentrations. These static methods involve the measurement of weight change over a long period which is often several days (Guiochon et al). As an alternative, it is possible to use the gradient facility of a chromatography system to deliver step gradients of known sample concentration. This method is known as Frontal Analysis and is described in Golshan-Shirazi and Guiochon (1988). Figure 2.5 shows a typical recording obtained.
A typical recording obtained for the frontal analysis method of isotherm determination.

The area under the curve to at each step change gives the amount of sample adsorbed and is as follows

\[
q_{i+1} = q_i + \frac{(C_{i+1} - C_i)(V_{i+1} - V_o)}{V_a}
\]  

(22)

Where \( V_o \) is the void volume and \( V_a \) is the volume of adsorbent. Hence a series of \( C_i \) and \( q_i \) values can be obtained at each step increment and the isotherm constructed.

2.4.2. Frontal Analysis by Characteristic Point

Instead of a series of concentration steps like in Frontal Analysis, one large concentration step is pumped into the column. After a plateau has been formed for a considerable time, pure mobile phase is then pumped in. A typical concentration profile obtained using Frontal Analysis by Characteristic Point (FACP) is shown in Figure 2.6. A mass balance is written for each point on the rear of the elution profile. The amount adsorbed is calculated by integration of the area under the peak, starting at \( C=0 \) at the tail end.

So \( q_i \) is evaluated from \( C_i \) using the following equation

\[
q_i = \frac{1}{V_a} \int_0^{C_i} (V_1 - V_o) dc
\]

(23)
Thus each point on the rear of the profile gives 1 point on the isotherm. This method has the main disadvantage of being inaccurate with low-efficiency columns and should only be used with columns with several thousand theoretical plates (Guan et al, 1994).

![Figure 2.6. A typical recording obtained for the frontal analysis by characteristic point method of isotherm determination.](image)

**2.4.3. Elution By Characteristic Point**

This method is very similar to FACP and involves a single injection pulse of a large sample. The isotherm is derived from the rear of the elution profile using equation (23). Once again highly efficient columns are a necessity (Guiochon et al, 1994).

**2.4.4. Retention Time Method.**

If the isotherm is Langmuirian, the analytical solution of the ideal model (see equation 3) can be inverted to derive the isotherm parameters from one large and one very large sample injection (Golshan-Shirazi and Guiochon, 1989). The very small load gives the retention time under linear conditions which allows the calculation of the initial slope of the isotherm. The large sample load permits the determination of the loading factor and hence the saturation capacity, $q_s$. From these, the Langmuirian parameters $a$ and $b$ can be calculated. The main drawback of this rapid method is that it will fail if the isotherm is not Langmuirian.
Perhaps the main drawback of the four methods discussed for isotherm determination is the need for pure components. This is particularly problematic when crude mixtures (like the crude erythromycin in this study) contain potentially a large number of species. This apart, even if single component isotherm data are available, it is never known how applicable they will be to multicomponent separations. Sometimes single component isotherms have been adequate for use in binary separations (Felinger and Guiochon, 1996) but this is often not the case, especially when a mixture is complex and contains many species.

2.5. Summary

This chapter has addressed the problems associated with the mathematical approach to the modelling of chromatography. A second order partial differential mass balance is required for each type of species involved in the separation. In order to help solve the equations, knowledge of the all physico-chemical data is necessary. This data ranges from mass-transfer and diffusive effects to the knowledge of the isotherms for each component. Most of the widely-used isotherms, both for single-components and multi-component separations have been described in this chapter as well as some of the common methods used to generate them for single components.

Despite these phenomena being well understood, predictions had real success in the prediction of elution profiles of binary mixtures. This bottleneck presents a real challenge for chemists and chemical engineers in making acceptable predictions of realistic, multi-component (much larger than 2 components) non-linear chromatographic separations. The main objective of this thesis is to make use of the multivariate statistical technique of Principal component analysis (PCA) to help tackle this problem. The idea is to use PCA to model a complex separation in the absence of any mathematical models or physico-chemical data. The next chapter gives an explanation of the theory behind PCA describing how it works and how it is used in practice.
3. Principal Component Analysis - Theory

3.1. Introduction
3.2. Mathematical description of PCA
3.3. Scaling of variables prior to PCA
3.4. Normalising the samples
3.5. Mean Centring
3.6. Calculating the Principal Components
  3.6.1. The NIPALS Algorithm
  3.6.2. Decomposition of the Variance-Covariance Matrix.
3.7. An example of PCA being applied to chromatography data.
  3.7.1. The scores and loadings for PC1
  3.7.2. Percentage variance associated with PC1.
  3.7.3. Reconstruction of Chromatogram data.
3.8. How many PCs are required for a model?
  3.8.1. The leave-one-sample-out-at-a-time procedure
  3.8.2. The Scree Graph and the Log-Eigenvalue Diagram
  3.8.3. Discussion about the number of PCs
3.9. Classification of samples
  3.9.1. SIMCA (Soft Independent Modelling of Class Analogies)
  3.9.2. The measurement of distance between samples.
  3.9.3. The SIMCA Model.
    3.9.3.1. The residual standard deviation (RSD) of a sample
    3.9.3.2. F-Test statistics for the comparison of $s_k^2$ and $s_c^2$
    3.9.3.3. Upper and Lower limits for scores values of each PC
3.10. Other multivariate statistical techniques.
  3.10.1. Multiway Principal component analysis (MPCA)
  3.10.2. Partial Least Squares (PLS)
  3.10.3. Principal component regression (PCR)
3.11. Conclusion
Chapter 3. Principal Component Analysis - Theory

3.1. Introduction

Principal component analysis (PCA) is a multivariate statistical technique concerned with extracting the key information from large and often complex data sets, whilst maintaining the majority of the variance from the original data but reducing significantly the size of the data set. Since many of the variables in a data set are themselves correlated and are effectively ‘saying the same thing’ they are effectively redundant. In PCA these variables from the data set are replaced with a much smaller number of uncorrelated or orthogonal principal components which, when combined with linear weighting factors, are able to describe the original data set to a very high degree of accuracy. By reducing the data set and abstracting the main trends PCA often reveals relationships which were not previously obvious from the original data and thereby allows interpretations that would not ordinarily result from an analysis of the unprocessed data set. PCA was developed at the turn of the century by Karl Pearson (1901) and was refined in the 1930’s by Harold Hotelling. Both used it for psychological data interpretation. Some of the practical applications of Principal Component Analysis were discussed in Chapter 1. The aim of this section is to provide an insight into the theory behind PCA including how the principal components are extracted from an original data set. Two different PCA algorithms are discussed—the NIPALS (non-iterative partial least squares) and the decomposition of covariance method. Although not used in this thesis, sample classification SIMCA (Soft Independent Modelling of Class Analogies) is discussed only for the sake of completeness. A simple example on the use of PCA for the analysis of chromatogram data is given in section 3.7. The chapter concludes with a brief look at other common multivariate statistical techniques.

3.2. Mathematical description of PCA

A fundamental requirement of PCA is that the data must be in a matrix form. The main goal of PCA is to establish relationships between the m rows (experimental samples) and n columns (variables) of a matrix X (dimensions m×n). In a
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 3

In the chromatographic context, the \( m \) samples are the individual runs, each comprising \( n \) UV absorbance values. Thus a set of 10 chromatograms each having 300 UV absorbance values would be arranged into an original matrix with dimensions 10×300. Mathematically, PCA relies on an eigenvector decomposition (eigenanalysis methods can be found in most good mathematical text books e.g. Stroud, 1989) of the correlation or covariance matrix of the data set. The covariance matrix of \( X \) is defined as

\[
\text{cov}(X) = \frac{X^TX}{m-1}
\]  

(1)

This assumes that the columns of \( X \) have been 'mean centred', i.e. adjusted to have a mean of zero by subtraction of the original mean from each column (see section 3.5). If the columns have been 'autoscaled' i.e. adjusted to zero mean and unit variance by dividing by its standard deviation, equation (1) gives the correlation matrix of \( X \). (see section 3.3) (Wise and Gallagher, 1996).

PCA decomposes matrix \( X \) into the sum of the products of pairs of vectors \( t_i \) and \( p_i \) with a residual matrix \( E \).

\[
X = t_1p_1^T + t_2p_2^T + t_3p_3^T + t_4p_4^T + \ldots + t_kp_k^T + E
\]  

(2)

\( k \) is the number of principal component contained in the model which must be less than or equal to either of \( m \) (number of samples) or \( n \) (number of variables). The \( t_i \) vectors are known as the \textit{scores} and the \( p_i \) vectors the eigenvectors or \textit{loadings}. The scores represent relationships between the variables and the loadings those between the samples. For each \( p_i \),

\[
\text{cov}(X) p_i = \lambda_i p_i
\]  

(3)

where \( \lambda_i \) is the eigenvalue associated with eigenvector \( p_i \). The \( t_i \) scores vectors form an orthogonal set \( (t_i^T t_j = 0 \text{ for } i \neq j) \) and the \( p_i \) loadings vectors are orthonormal \( (p_i^T p_j = 0 \text{ for } i \neq j \text{ and } p_i^T p_i = 1 \text{ for } i = j) \). The \( \lambda_i \) are a measure of the amount of variance described by each PC. PC1 captures the greatest amount of variation, PC2 the next and so on.
Generally speaking, the original data can be adequately described by using far fewer PCs than original variables. Furthermore, the principal components are often far more robust indicators of process conditions than individual variables due to the signal averaging aspects of PCA (Wise and Gallagher, 1996).

### 3.3. Scaling of variables prior to PCA

The variables in the original data matrix serve to influence the PCA according to their variance (and hence their correlations with other variables). If some variables have much larger dominance than others, they will dominate and in some extreme cases, the first PC may contain contributions from only one or a few variables. For example a data set of chemical compounds might be characterised by several variables such as their solubility in water, their molecular weight and density etc. Because all of these parameters have different units, it is customary to standardise the variables prior to analysis to allow more realistic correlations to be drawn from the principal components. This is usually done by giving all variables a mean of zero (by subtracting the mean of each variable from each sample) and a variance of one. Unit variance for each variable is achieved by multiplying each variable by the inverse of its standard deviation. This is commonly referred to as ‘autoscaling’. Other scalings are sometimes used for particular purposes in PCA including taking logs of exponential data prior to adjustment of the mean and variance.

In a chromatographic context where the variables are UV absorbance values, the mean of zero is achieved by mean centring (section 3.5). There is no need to adjust the absorbance values to unit variance as this will greatly affect the peaks shapes and hence the quality of the resultant principal component analysis.

### 3.4. Normalising the samples

Normalisation of samples prior to PCA is sometimes performed in order to give the samples the same relative size. In chromatography peak size is related to the amount of sample injected and it may be desirable to account for this before analysis. The method, which essentially involves converting chromatograms to a common area was first recognised by Kvalheim (1985). This study, however used peak areas as the
original variables for the PCA and not, as is the case with this thesis, UV absorbance values which plot out the entire chromatograms. The merits of sample normalisation and other pre-processing techniques will be discussed in future results chapters.

3.5. Mean Centring

The original data set is commonly centred before PCA is applied. In such cases the mean chromatogram from the original data (or a reference chromatogram) is subtracted from every individual chromatogram. The principal components generated from the mean-centred data set are indicative of changes in UV absorbance and not the absolute values. Mean centring has the effect of enhancing the more subtle differences between chromatograms and since this improves the ability of the calculation to detect differences between chromatograms will result in a more accurate PCA model to the original data set. Mean centring is a logical step when considered in the context of how PCA calculates the loadings. Since the loadings represent the changes in the data which are common to every chromatogram, removing the mean simply removes the first and most common variation before the data has been subjected to the PCA algorithm.

3.6. Calculating the Principal Components

There are two different methods used to calculate the Principal Components from a data set; the NIPALS Algorithm and the Decomposition of Covariance. The following descriptions of these methods assume that the matrices involved have the following dimensions: $X$ is a matrix of $m$ rows $\times$ $n$ columns, $t$ is the scores matrix with dimensions $m \times k$ ($k$ is the number of principal components extracted), $p$ is the loadings matrix (dimensions $k \times n$) and $\lambda$ is an $k \times k$ matrix of eigenvalues. When used, the subscripts on the matrices indicate the matrix row.

NOTE: The eigenvalues matrix $\lambda$ is a diagonal square matrix. i.e. it only has values on its diagonal and is zero everywhere else. The eigenvalues should always descend along the diagonal since the first principal component (eigenvector) represents the largest variation, the second is the next largest and so on.
3.6.1. The NIPALS Algorithm

The NIPALS (non-iterative partial least squares) is the most common method for calculation of the Principal Components of a data set. The advantages of this method over Decomposition of Covariance method are three-fold: firstly, the covariance matrix $X^TX$ does not need to be determined; secondly, missing data can be handled; and thirdly PCs are extracted in order of decreasing importance [Wold et al (1987), MacGregor et al (1991), Kvalheim (1993)]. The flexibility of the NIPALS algorithm in coping with missing data makes it an ideal choice for analysis by PCA of data generated from industrial processes which may contain incomplete data. Only a description of the mechanism is presented since full details can be found in Geladi and Kowalski (1986).

The NIPALS algorithm gives more numerically accurate results when compared with the Decomposition of Covariance method, but is slower to calculate. The steps are as follows:

1. As a first approximation, set the PC$_1$ loading to the first sample: $p_1=x_1$
2. Calculate the eigenvalue: $\lambda_{ij} = (\sum t_i^2)^{1/2}$
3. Normalise the PC$_i$ loading: $p_i = p_i/\lambda_{ij}$
4. Compute the scores values: $t_i = X/p_i^T$
5. Check for convergence by comparing these scores to the scores from the previous pass for this eigenvector. If this is the first pass for the current (PC$_i$) loading, or scores are not the same continue with step 5. If the scores are the same, go to step 8.
6. Recompute the PC$_i$ loading: $p_i = X^T/t_i$
7. Go to step 2.
8. Calculate the residual matrix. Proceed with the next principal component loading (PC$_{i+1}$). Go to step 1.
9. Stop calculating at PC$_k$ when the residual matrix reaches a certain degree of sparsity.
The NIPALS algorithm can be summarised diagrammatically in Figure 3.1 for the first 2 PCs. A mean or reference matrix is initially subtracted from the original data and the remaining matrix $X$ is broken down into the product of two smaller matrices and a residual matrix ($E_i$). The smaller matrices have dimensions $(1 \times n)$ and $(m \times 1)$ respectively. The $(1 \times n)$ matrix, the loadings matrix for PC1, is a common component of all of the experimental samples. The $(m \times 1)$ matrix is the scores matrix for PC1 and represents the amount of the PC1 loadings matrix which is present in each experimental sample. Collectively they represent the first principal component (PC1).

Subsequent principal components are derived in a similar manner from the residual matrix ($E_i$) until a model is obtained in which the number of PCs is deemed sufficient to describe adequately the data set (see section 3.8). In this way PCA models are able to explain a large percentage of the variation in the original data matrix in terms of a small number of principal components.
3.6.2. Decomposition of the Variance-Covariance Matrix.

The Decomposition of the variance-covariance matrix method is faster than the NIPALS algorithm but tends to produce numerical errors in the higher principal components when performed on computers using only single precision. This is due to the data compression step performed at the beginning to calculate the variance-covariance matrix $Z$ (which has dimensions of $m \times m$). When using this algorithm, it is necessary to use double precision numbers, or only use it to calculate the first few Principal Components. The steps are as follows:

1. Compute the variance-covariance square matrix: $Z = X^TX$
2. Initialise scores to an arbitrary number: $t_i = 0.1$
3. Compute new scores: $t_i = Zt_i^T$
4. Calculate the eigenvalue: $\lambda_{ij} = (\Sigma t_i^2)^{1/2}$
5. Normalise the scores: $t_i = t_i / \lambda_{ij}$
6. Check for convergence by comparing these scores to the scores from the previous pass for this PC loading. If this is the first pass for the current PC loading, or the scores are not the same, go back to step 3. If the scores are the same continue with step 7.
7. If $i = p$ (number of PCs) go to step 9; otherwise compute the residual variance-covariance matrix for the next PC loading: $Z = Z - (t_i^T t_i)\lambda_{ij}$
8. Increase the PC loading counter, $i = i + 1$ and go back to step 3
9. Compute all of the PC loadings: $p = t^T X$

3.7. An example of PCA being applied to chromatography data.

The following section provides a very basic example to illustrate the use of PCA being applied to some chromatography data. The example shows just one elution peak obtained from 7 different sample load sizes. For simplicity each sample comprises just 11 UV values -one every 0.15 minutes (in practice each sample would contain several hundred to ensure good resolution and accurate results). The 11 UV values for each sample (together with the mean chromatogram) are given in Table 3.1 and the profiles themselves in Figure 3.2. The example then shows how the data matrices are presented in such an analysis.
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 3

Table 3.1. A table giving 11 UV absorbance values which map out the shape of a set of example chromatograms. This shows the data in the raw form prior to subtraction of the mean chromatogram. The mean chromatogram is also shown.

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>0.00</th>
<th>0.15</th>
<th>0.30</th>
<th>0.45</th>
<th>0.60</th>
<th>0.75</th>
<th>0.90</th>
<th>1.05</th>
<th>1.20</th>
<th>1.35</th>
<th>1.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample 1</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.010</td>
<td>0.019</td>
<td>0.015</td>
<td>0.009</td>
<td>0.005</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>sample 2</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.007</td>
<td>0.030</td>
<td>0.037</td>
<td>0.025</td>
<td>0.014</td>
<td>0.008</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>sample 3</td>
<td>0.004</td>
<td>0.005</td>
<td>0.008</td>
<td>0.024</td>
<td>0.074</td>
<td>0.064</td>
<td>0.040</td>
<td>0.023</td>
<td>0.013</td>
<td>0.008</td>
<td>0.005</td>
</tr>
<tr>
<td>sample 4</td>
<td>0.009</td>
<td>0.025</td>
<td>0.101</td>
<td>0.086</td>
<td>0.052</td>
<td>0.029</td>
<td>0.017</td>
<td>0.010</td>
<td>0.007</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>sample 5</td>
<td>0.010</td>
<td>0.016</td>
<td>0.098</td>
<td>0.146</td>
<td>0.098</td>
<td>0.057</td>
<td>0.032</td>
<td>0.019</td>
<td>0.011</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>sample 6</td>
<td>0.010</td>
<td>0.014</td>
<td>0.033</td>
<td>0.155</td>
<td>0.137</td>
<td>0.087</td>
<td>0.051</td>
<td>0.030</td>
<td>0.019</td>
<td>0.012</td>
<td>0.008</td>
</tr>
<tr>
<td>sample 7</td>
<td>0.012</td>
<td>0.017</td>
<td>0.039</td>
<td>0.182</td>
<td>0.160</td>
<td>0.101</td>
<td>0.060</td>
<td>0.035</td>
<td>0.021</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td>Mean</td>
<td>0.007</td>
<td>0.012</td>
<td>0.041</td>
<td>0.086</td>
<td>0.080</td>
<td>0.056</td>
<td>0.034</td>
<td>0.020</td>
<td>0.012</td>
<td>0.007</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Figure 3.2. The chromatograms which are comprised of the data from Table 3.1. This figure shows the data in its raw form prior to mean centring.

3.7.1. The scores and loadings for PCI

The following data matrix $X$ shows the mean centred data (to 3 decimal places), ready for the first principal component to be extracted.

$$X = \begin{bmatrix}
-0.006 & -0.010 & -0.039 & -0.084 & -0.070 & -0.038 & -0.019 & -0.011 & -0.007 & -0.005 & -0.003 \\
-0.005 & -0.009 & -0.037 & -0.079 & -0.050 & -0.019 & -0.009 & -0.006 & -0.004 & -0.002 & -0.002 \\
-0.003 & -0.006 & -0.033 & -0.062 & -0.006 & -0.008 & -0.006 & -0.003 & -0.001 & -0.001 & -0.000 \\
0.003 & 0.014 & 0.060 & 0.000 & -0.028 & -0.027 & -0.018 & -0.010 & -0.005 & -0.003 & -0.002 \\
0.003 & 0.004 & 0.057 & 0.060 & 0.018 & 0.001 & -0.002 & -0.001 & -0.001 & 0.000 & 0.000 \\
0.003 & 0.002 & -0.007 & 0.069 & 0.057 & 0.030 & 0.017 & 0.011 & 0.007 & 0.005 & 0.003 \\
0.005 & 0.005 & -0.002 & 0.096 & 0.080 & 0.045 & 0.025 & 0.015 & 0.009 & 0.005 & 0.003
\end{bmatrix}$$

The scores and loadings matrices $t_i$ and $p_i$ for the first principal component calculated using the NIPALS algorithm (section 3.5.1) are as follows.
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 3

\[ X = t_1p_1^T + E_1 \]  

(5)

Which is the same as equation (1) for just one principal component. The following matrices display the numerical values (to 3 decimal places) of the scores and loadings for the first PC.

Loadings matrix \( p_1 \)

\[
\begin{bmatrix}
-0.124 & 0.042 \\
-0.102 & 0.060 \\
-0.055 & 0.213 \\
-0.012 & 0.757 \\
0.067 & 0.531 \\
0.132 & 0.025 \\
\end{bmatrix}
\]  

(6)

Scores matrix \( t_1 \)

\[
\begin{bmatrix}
-0.124 & -0.102 & -0.055 & -0.012 & 0.067 & 0.132 \\
0.042 & 0.060 & 0.213 & 0.757 & 0.531 & 0.025 \\
\end{bmatrix}
\]

The following matrix shows the product of the scores and loadings matrices for PC1 (equation 6)

\[
\begin{bmatrix}
-0.052 & -0.074 & -0.264 & -0.939 & -0.656 & -0.320 & -0.169 & -0.066 & -0.064 & -0.026 \\
-0.004 & -0.006 & -0.022 & -0.077 & -0.054 & -0.026 & -0.014 & -0.008 & -0.005 & -0.006 \\
-0.002 & -0.003 & -0.012 & -0.042 & -0.029 & -0.014 & -0.007 & -0.004 & -0.003 & -0.003 & -0.001 \\
-0.001 & -0.001 & -0.003 & -0.009 & -0.006 & -0.003 & -0.002 & -0.001 & -0.001 & -0.001 & 0.000 \\
0.003 & 0.004 & 0.014 & 0.051 & 0.036 & 0.017 & 0.009 & 0.005 & 0.003 & 0.003 & 0.001 \\
0.004 & 0.006 & 0.020 & 0.070 & 0.049 & 0.024 & 0.013 & 0.007 & 0.005 & 0.006 & 0.002 \\
0.006 & 0.008 & 0.028 & 0.100 & 0.070 & 0.034 & 0.018 & 0.011 & 0.006 & 0.007 & 0.003 \\
\end{bmatrix}
\]

(7)

The residual matrix below, \( E_1 \), is calculated by subtracting the product of the previous matrices from the original matrix \( X \). The second PC is derived in the same way.

\[
E_1 = \begin{bmatrix}
-0.001 & -0.003 & -0.013 & 0.010 & -0.005 & -0.006 & -0.003 & -0.001 & -0.001 & -0.001 \\
-0.001 & -0.003 & -0.015 & -0.002 & 0.004 & 0.007 & 0.004 & 0.003 & 0.001 & 0.001 \\
-0.001 & -0.003 & -0.021 & -0.021 & 0.023 & 0.022 & 0.013 & 0.007 & 0.004 & 0.002 \\
0.003 & 0.014 & 0.063 & 0.009 & -0.002 & -0.024 & -0.016 & -0.009 & -0.005 & -0.002 \\
0.001 & 0.000 & 0.043 & 0.009 & -0.018 & -0.017 & -0.011 & -0.007 & -0.004 & -0.002 \\
-0.001 & -0.003 & -0.027 & -0.001 & 0.007 & 0.007 & 0.005 & 0.003 & 0.002 & 0.002 \\
-0.001 & -0.003 & -0.030 & -0.004 & 0.010 & 0.011 & 0.007 & 0.004 & 0.002 & 0.001 \\
\end{bmatrix}
\]

(8)

If \( t_1p_1^T \) (7) is added to \( E_1 \) (8), the result is numerically identical to the original matrix \( X \) (4) to 3 decimal places.
The loadings profiles for PC1 and PC2 are given in Figure 3.3. These are functions of time. The scores plot for the first two PCs is given in Figure 3.4. This plot shows the relationship between the 7 chromatographic samples.

The PC1 loading represents the pattern in the data set (Table 3.1) which is common to all samples and explains the largest amount of the original variance. PC2 loading represents the largest amount of variance remaining after the extraction of PC1.

The scores values represent how much the loading profiles for each PC need to be scaled in order to describe the original data set. Certain trends can be picked up in the scores plots e.g. one trend of the scores plot in Figure 3.4 is that, as the sample load increases, so does PC1. It appears that samples 1 and 2, and samples 6 and 7 are the most similar due to their respective close proximity.

![Figure 3.3. The loadings profiles for the first 2 PCs of the example data set X. These are common to each sample. The amount of each of the PCs contained in each sample is given by the scores values.](image-url)
Figure 3.4. The PC1-PC2 scores plot for the example matrix X. The plot shows the relationships between the samples. One trend of this plot is that PC1 scores reflects the sample load applied (the higher the load, the higher PC1 score).

### 3.7.2. Percentage variance associated with PC1.

To calculate the percentage variance explained by PC1, it is necessary to calculate the total variance in the original data set X. This is simply a summation of the variances of the 11 UV variables across the 7 samples.

\[
\text{Total variance in } X = 0.01211.
\]

The amount of variance explained by PC1 is the sum of the products of the scores matrix with its transpose, divided by \((m-1)\) where \(m\) is the number of samples.

\[
\text{Variance explained by PC1} = \frac{\sum t_1^T t_1}{m-1}
\]

From the values of PC1 scores (equation 6) the variance calculation is equal to 0.009919. Hence the percentage variance explained by PC1 is
100*(0.009919/0.01211) = 81.19 %. The percentage variances associated with the other PCs are calculated in exactly the same way.

### 3.7.3. Reconstruction of Chromatogram data.

The original chromatogram data may be reconstructed using the PC model generated in order to assess how well the model fits the data. This is done by a summation of the product of the scores and loadings for each sample and adding in the mean chromatogram $X_m$. For example the reconstruction of sample 1 using a 2PC model would be as follows.

\[
\text{Sample 1 (reconstructed)} = (t_1p_1^T) + (t_2p_2^T) + X_m
\]

The reconstruction would then be compared with the original sample and an assessment of the error made. This will be detailed in full in future results chapters.

### 3.8. How many PCs are required for a model?

Several methods have been proposed which set out to estimate the number of principal components needed for optimal PC models. The most common methods are outlined below.

#### 3.8.1. The leave-one-sample-out-at-a-time procedure

A much used technique which is well suited for small sample sets is the leave-one-sample-out-at-a-time procedure (Wold, 1978). In this method, one row is deleted from the original data matrix $X$ and the principal components calculated. The $m$ PC models are developed (same as the number of samples) and used to predict the left-out sample with 1, 2, 3 etc. principal components. The deviation between the actual values and predicted values is used to estimate an overall prediction error. The model providing the minimum prediction error is finally calculated with all of the samples included. This method uses the PRESS (prediction error for the sum of squares) algorithm which is used for calculating the prediction error and is documented in Wold (1978). The main drawback in using this approach is the computational time required and hence the reason why it tends only to be used with small data sets. Since chromatography data sets are potentially very large, this method is impractical for the work in this thesis.
3.8.2. The Scree Graph and the Log-Eigenvalue Diagram

Scree graphs and Log-eigenvalue diagrams are essentially the same and involve plotting the eigenvalues derived after the extraction of each successive principal component against the PC number. The number of components is decided upon when the slope of the curve becomes approximately horizontal. Both techniques are well documented in Jolliffe (1986). Figure 3.5 illustrates a scree graph for the first 5 PCs of the example data set (Table 3.1). From this plot, it can be estimated that a model containing 3 PCs would be suitable. Log-eigenvalue (LEV) diagrams, the same as scree graphs but using the log of the eigenvalues plotted against PC number. LEVs serve to eliminate noisy components and are popular in meteorology where PCA is used to correlate parameters such as wind speed, temperature and air pressure to help with weather forecasting.

![Figure 3.5. A scree plot for the first 5 PCs of the example data set (Table 3.1). From this plot, an estimate can be made as to the number of principal components for a suitable model](image)

3.8.3. Discussion about the number of PCs

There are other theoretical techniques which set out to optimise the number of PCs. These can be located in Jolliffe (1986) and include a technique called Partial Correlation where PCs are not retained if they are dominated by a single variable whose correlations with all the other variables are close to zero.
Despite these algorithms which set out optimise the number of PCs, perhaps as good a method as any is to use the number of PCs which explain a certain percentage of the total variation of the original data set. It may be decided to select 95% or 99% as the desired amount of variation and use the required number of PCs to satisfy this. One can also argue that some of the minor PCs are not useful (e.g. they may contain noise) and should be excluded. Since the methods mentioned previously were developed when computers were not as powerful as today, it made sense to minimise the number of PCs. However modern computers are able to generate models with many PCs very rapidly for very complex data sets, so it is conceivable that any number of PCs could be used in a model. In practice however few models use more than 4 PCs. Wold (1987) proposed that negligible useful information is carried from the 4th PC onwards.

For the work in this thesis, generally 3 PC models have been used which describe a very high amount of variance (typically >98%).

3.9. Classification of samples

Because of the ability of computers to handle vast amounts of data, classification methods have become increasingly important in diverse areas such as chemistry, biochemistry, geology and medicine (Kvalheim, 1993). The question often arises whether samples are similar in some respect; e.g. in structural, toxic or carcinogenic properties. Test samples can be compared with classes or clusters of validated models to assess similarity via goodness-of-fit criteria. A classic example of a successful empirical classification was the work conducted by Mendelev in the 1860s who grouped similar chemical elements together to form the Periodic Table. New elements were postulated and compared with established classes and successfully positioned within the most suitable groups. This early use of multivariate classification made a huge impact in the development of chemistry, although it was several decades later until a theoretical classification model was developed based on quantum physics.

The classification method SIMCA (Soft Independent Modelling of Class Analogies), developed by Wold and his co-workers (1976, 1978, 1981) is such a method where each class of samples is developed using its own PC model.
3.9.1. **SIMCA (Soft Independent Modelling of Class Analogies)**

A brief summary of SIMCA classification is presented in this section. The application of SIMCA in this thesis has not been necessary and so has only been included for completeness. A more rigorous description can be found in Chandwani (1995). The nature and complexity of chromatogram data means that clusters of diverse shapes and sizes are obtained. Attempts to account for these diversities would create confusion. Furthermore, the aims of this thesis are not to develop a pattern recognition tool which assigns samples to particular clusters, which was the aim of Chandwani's (1995) thesis—rather the overall goal of this thesis is to use of PCA as a prediction tool for the scale-up of chromatography.

The SIMCA classification method which is also known as Disjoint PC modelling was first used by Wold (1976). There are three basic steps initially employed which are:

- A PC model is developed for the entire data set. The resulting scores plots reveal clusters of similar samples.
- PC models are developed for each cluster (these are naturally independent of each other). Each, so-called disjoint model must be mean-centred individually before the extraction of the PCs.
- The disjoint models can then be refined (outlier removal etc.), so improving their predictive abilities.

The effectiveness of using SIMCA classification appears to be limited by the size of each disjoint model. It has been suggested that each disjoint model should have a minimum of around 5 samples and preferably more than 10. Moreover, Dröge and van't Klooster (1987) and Dröge et al (1987) suggested that as the ratio of variables to samples n/m increases, assigning samples to particular cluster gets more inaccurate.

3.9.2. **The measurement of distance between samples.**

The definition of distance is central to all classification. The Euclidean distance in variable space is often used for measuring similarity between samples. Figure 3.7 illustrates this in 2-dimensional space. The generalisation to M-dimensional space is straightforward and is given by the following formula:
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 3

\[ d_{kl}^2 = (x_k - x_l)^T (x_k - x_l) \] (11)

Figure 3.6. Two samples plotted in 2-dimensional variable space. The difference between the sample vectors define the Euclidean distance \(d_{kl}\).

The role of distance in multivariate classification derives from the assumption that proximity in multivariate space is indicative of similarity between samples.

If there is a group of samples and we wish to determine whether each sample belongs as a member of a particular group, one way achieve this is by comparing every sample, \(x_k\) to the centre of gravity of all samples, \(X^T\). The centre of gravity defines the model so that each sample can be described by \(x_c\) and a residual distance \(e_k\). This is illustrated in Figure 3.7.

Figure 3.7. The sample vectors are expressed as the centre of gravity and a residual vector. Class membership can be decided by comparing the residual distances to the centre of gravity of the samples used to model the class.

81
The class membership can be decided on by comparing the individual residual distance $e_k$ for each sample (total number $m$) with the average residual distance.

$$d_c^2 = \frac{\sum e_k^T e_k}{m} \tag{12}$$

The main drawback using this method of assigning cluster membership is that it is only applicable with spherical structures (radius $d_c$) so that cluster members must be within the sphere. It may be more beneficial for each point to be related to a line defined by the principal components. It is this approach which forms the basis of SIMCA.

### 3.9.3. The SIMCA Model.

When PCA is performed on the samples in Figure 3.7, the straight lines describing the first two principal components may look like the ones shown in Figure 3.8. Instead of relating each point to the centre of gravity as in the $d_c^2$ equation (equation 12), each point is related to lines defined by the PCs (Figure 3.8).

This allows the expression of a more complex data structure resulting from the correlation between variables. This correlation, expressed in terms of the PCs, provides a wider range of data structure to be determined. Another advantage of using the SIMCA approach over a purely distance-based approach is that the modelling into PCs efficiently filters noise. This provides a basis for the detection of outliers.

![Variable space](image)

**Figure 3.8.** Two PCs have been fitted to the samples from Figure 3.7. Instead of comparison with the centre of gravity which assumes a spherical distribution, each sample is related to the lines defined by the PCs.
3.9.3.1. The residual standard deviation (RSD) of a sample

Having decided on the number of principal components, the distance \( s_k \) of sample \( k \) to the model is given by:

\[
s_k^2 = \frac{e_k^T e_k}{n - a}
\]

Where \( n \) = number of variables and \( a \) = number of principal components. The division by \( n - a \) provides a distance measure which is independent of the number of variables and corrected for the loss of freedom due to the fitting of \( a \) PCs. In SIMCA terminology, \( s_k \) is the residual standard deviation (RSD) of sample \( k \).

The \( m \) RSDs in a class can be arranged into a vector \( s \) with dimensions \((m \times 1)\). The mean RSD of the class is then defined as:

\[
s_c^2 = \frac{s^T s}{m - a - 1}
\]

The division by \( m - a - 1 \) gives a scale which is independent of the number of samples \( m \) and corrected for the loss in degrees of freedom due to column centring and fitting of a PCs.

3.9.3.2. F-Test statistics for the comparison of \( s_k^2 \) and \( s_c^2 \)

The comparison of RSD for sample \( k \), \( s_k^2 \) with the mean RSD for the class, \( s_c^2 \) gives a direct measure of its similarity with the class model. The F-test (see any standard statistical text book) is a standard method for comparison of variances and the first application in PCA was by Wold (1976) for the comparison between \( s_k^2 \) and \( s_c^2 \). The degrees of freedom used to obtain the critical F-value are \((n-a)\) for \( s_k^2 \) and \((n-a)(m-a-1)\) for \( s_c^2 \).

The F-test can be used to calculate an upper limit for the RSD for samples that belong in the class:

\[
s_{\text{max}} = s_c^2 F_{\text{crit}}
\]

\( F_{\text{crit}} \) is usually determined for a given level of confidence 95% \((\alpha=0.05)\) or 99% \((\alpha=0.01)\). This means that the probability of a sample failing (false negatives) when it
should be a member will be 1 in 20 when $\alpha=0.05$ and 1 in 100 when $\alpha=0.01$. On the other hand, the risk of false positives will be greater for $\alpha=0.01$ than for $\alpha=0.05$.

The F-test method for classification has been criticised most notably by Dröge and van't Klooster (1987) and Dröge et al (1987). They concluded that this method gave model cylinders which were too narrow when the number of variables was much greater than the number of samples, thus endangering misclassification. Wold and Sjöström (1987) suggested improvements such as reducing the degrees of freedom when calculating $F_{crit}$.

### 3.9.3.3. Upper and Lower limits for scores values of each PC

By using the F-test to calculate the upper and lower limits for acceptable residual distances, the SIMCA model is closed around the PCs though the model is still open along each PC. Wold (1976) used the extremes of the scores for each PC to define the upper and lower limits for the scores.

$$t_{lower,a} = t_{min,a} - \frac{1}{2}s_{la}$$  \hfill (16)

$$t_{upper,a} = t_{min,a} + \frac{1}{2}s_{la}$$  \hfill (17)

Where $t_{lower,a}$ and $t_{upper,a}$ are the lower and upper values of the scores for the $a^{th}$ PC. $t_{min,a}$ is the minimum value of the object scores on the $a^{th}$ PC. $s_{la}$ is the spread of the scores along each PC and given in Wold (1976) as being:

$$s_{la}^2 = \frac{t_1^T t_s}{m}$$  \hfill (18)

An illustration of the upper and lower limits for a single-PC model is given in Figure 3.9. Essentially a cylinder for each PC with radius $s_i$ defines the boundary for membership. Wold (1976) provided no theoretical foundation for the choice of the upper and lower limits for the scores.

Currently available multivariate software packages are able to provide data using both the RSD F-test and upper/lower limit of scores criterion. The user can use both techniques when assessing whether or not to accept a sample.
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 3

Figure 3.9. By using the spread of sample scores along each PC, upper and lower limits can be defined. The cylinder described using this technique is used to assess acceptance or rejection of samples to or from a membership cluster.

3.10. Other multivariate statistical techniques.

In addition to PCA, there are other multivariate statistical techniques. These are very similar to PCA and include Multiway PCA (MPCA), Partial least squares (PLS) and Principal component regression (PCR). The following subsections give a very brief explanation of these techniques.

3.10.1. Multiway Principal component analysis (MPCA)

Multiway PCA (Nomikos and Macgregor, 1995 and Wise and Gallagher, 1996) is an extension of PCA inasmuch as the original data set $X$ is not 2-dimensional but multi-dimensional, commonly 3-D. This concept is illustrated in Figure 3.10. MPCA is most often used with batch processes where measurements describing the current state of each batch is monitored at regular time increments.

Figure 3.10. A diagram of a 3-dimensional data matrix $X$ with dimensions ($I\times J\times K$). There are $I$ samples, $K$ measured variables and $K$ time increments.
The PCA algorithms work in exactly the same way as for PCA. First of all though, the block matrix needs to be unfolded so that it is effectively a 2-dimensional structure. This is illustrated in Figure 3.11.

Wise and Gallagher (1996) applied MPCA to the monitoring of a nuclear waste storage tank. The waste constituents were being converted into hydrogen, ammonia and nitrogen dioxide gases. These gases were periodically released using a pump to agitate the waste slurry. The concentrations of each gas were measured using various analytical techniques as a function of time. The procedure was done 57 times (I=57) with 280 time increments each (K=280). The variables were the concentrations of the 3 gases (J=3). The original matrix $X$ with dimensions $57 \times 3 \times 280$ was broken down into matrix $X$ with dimensions $57 \times 840$ before applying PCA in the usual way.

![Figure 3.11. The algorithm for performing MPCA (adapted from Wise and Gallagher, 1996).](image)

3.10.2. Partial Least Squares (PLS)

Partial least squares uses the same techniques as in PCA to extract scores and loadings, but is applied to 2 data sets $X$ and $Y$. $X$ like in PCA is an $(m \times n)$ data set of process measurements. $Y$ is an $(m \times q)$ matrix of process variable settings. Similar data pre-treatments procedures (mean centring and adjustment to unit variance) to
PCA are usually performed (sections 3.3-3.5). In a chromatographic context, \( X \) is the data matrix with \( m \) samples each having \( n \) UV absorbance values, and \( Y \) is a matrix of \( q \) process variable settings for each sample. These \( q \) variables may comprise settings such as temperature, flow rate, mass load and pH. PLS is usually used as a process control tool whereby the reason(s) for failure, i.e. which process variable(s), can rapidly be determined (and thus corrected) with a high degree of accuracy. Difficulties often arise when attempting to predict many process variables from few process measurements. For this reason, \( n \gg k \) is required for models of high accuracy.

The goal of PLS is to make a future prediction of a \( y_i \) value (the process variable settings) from an \( x_i \) value (a chromatographic run). PLS is described in several publications. Some of the best descriptions can be found in Wold et al, 1984; Geladi and Kowalski, 1986; Kresta et al, 1993 and Wise and Gallagher, 1996. The NIPALS algorithm (section 3.6.1.) is used to develop PC models for \( X \) and \( Y \). These are developed so that the covariance between the 2 data sets is maximised. As well as extraction of scores and loadings, an additional set of vectors are calculated known as weights, \( W \). These weights are required to maintain orthogonal scores. A unique feature of PLS is that it can be used to form models relating more than one predicted variable (in \( Y \)) to many predictor variables (in \( X \)) (Wise and Gallagher, 1996).

### 3.10.3. Principal component regression (PCR)

In PCR, instead of regressing process variables (temperature, flow rate etc.) on to the measured variables (UV absorbance), the process variables are regressed on to the PC scores of the measured variables (which are orthogonal and therefore well-conditioned) (Wise and Gallagher, 1996).

The purpose of the regression model is to predict the properties of interest for new samples. Thus, the number of PCs should be determined which optimise the predictive ability of the PCR model. A cross-validation technique is performed which splits the data into training and test sets. A prediction of the residual error is on the test samples is determined as a function of the number of PCs in the regression model with the test data. The procedure is usually repeated several times, with each sample
3.11. Conclusion

This chapter has focused on the theory behind Principal Component Analysis and its application to chromatography data. The two most common methods of extracting the PCs were discussed—the Decomposition of the Variance-Covariance Matrix and the NIPALS algorithm which was the method used by the software employed throughout this study. The following chapter details the application of PCA to an experimental Size Exclusion Chromatography system. The raw data used in Chapter 4 was from Chandwani’s PhD thesis (1995) and an important pre-processing stage was required prior to analysis.
4. Pre-processing of chromatographic data for Principal Component Analysis

4.1. Summary

4.2. Introduction

4.3. Materials and Methods

4.3.1. The Chromatographic Process

4.3.2. Generation of historical data sets

4.3.3. Flow Rate Transformation

4.3.4. Sample Load Adjustment

4.5. Results and Discussion

4.5.1. Analysis of the gross principal component model for dataset X1 after flow rate transformation

4.5.2 Analysis of the disjoint model for the cluster at constant vc

4.5.3 Analysis of the gross principal component model for dataset X2

4.5.4. Analysis of data set X1 with adjustment for sample load (after flow rate transformation)

5. Conclusions
Chapter 4. Pre-processing of chromatographic data for Principal Component Analysis

4.1. Summary

This chapter examines the selection of the appropriate representation of chromatogram data prior to using principal component analysis (PCA), a multivariate statistical technique, for the diagnosis of chromatogram data sets. Historical data was used in which the effects of four process variables were investigated; flow rate, temperature, loading concentration and loading volume, for a size exclusion chromatography system used to separate three components (monomer, dimer, trimer). Major positional shifts in the elution peaks that result when running the separation at different flow rates caused the effects of other variables to be masked if the PCA was performed using elapsed time as the comparative basis. Two alternative methods of representing the data in chromatograms are proposed. In the first data were converted to a volumetric basis prior to performing the PCA, while in the second, having made this transformation the data were adjusted to account for the total material loaded during each separation. Two data sets were analysed to demonstrate the approaches. The results show that by appropriate selection of the basis prior to the analysis, significantly greater process insight can be gained from the PCA and demonstrates the importance of pre-processing prior to such analysis.

4.2. Introduction

The multivariate statistical technique of principal component analysis (PCA) offers potential for monitoring and control of bioprocesses. [Saner and Stephanopolous (1992)] used PCA for the monitoring and control of the fermentation stage, but until now there have been very few applications of PCA for downstream operations which include chromatography.
Instead of using mathematical models for predictions and diagnosis, historical data can be exploited for diagnosis of future batches. Under the Food and Drug Administration (FDA) guidelines, production records must be kept for a minimum of seven years, so this archived data is potentially very valuable since the records contain information regarding those batches which have gained approval and those which have not. Another advantage of using such statistical methods is that since actual data is being utilised, any interactions between the process variables will be an intrinsic property of the historical data.

Principal component analysis (PCA) is a multivariate statistical technique used to extract meaningful information from sets of correlated data. This approach has been used in applications ranging from ecological data interpretation to taxonomy following its initial application in psychology. PCA is described in many statistical text books e.g. Fry (1992). PCA eliminates redundant information in the dataset by retaining only the main features inherent within it [Wold et al, 1987]. The theory behind PCA is given in chapter 3.

A number of researchers have examined the application of PCA to the analysis of chromatography data. The main literature review is located in section 1.7.2 and only a brief resume is given here. An early use of such statistical techniques was for the analysis of chromatograms obtained by gas chromatography [Fernando Faigle et al, 1991]. The effect of column temperature on overlapping peaks of mixtures of toluene, isooctane and ethanol was investigated. The statistical techniques employed were partial least squares and principal component regression. These methods combine a regression modelling approach with a principal component analysis. Although the study only looked at the effects of three temperatures (378, 393 and 403 K) the paper highlighted the potential of using statistical techniques for chromatographic data analysis in general. A second study [Malmquist and Danielsson, 1994] noted that different combinations of process variables led to uniquely shaped profiles which made data interpretation complex. In this paper PCA was used to analyse chromatograms for peptide mapping as a quality control aid for recombinant DNA-
derived proteins. More recently a similar PCA technique for the diagnosis of a size exclusion chromatography system was reported [Chandwani et al, 1997]. In that work, clusters of similar chromatograms were observed in the PCA projections but the interpretation was complicated by the dominating effect of flow rate on the analysis.

Principal component analysis is able to detect subtle changes in the nature of chromatograms which are not visible to human operators. In this chapter the focus is made on the effect of data pre-processing to procure maximal efficacy from the analysis. In Chandwani et al (1997) the importance of the flow rate variable on the PCA was recognised but not accounted for prior to analysis. This led to difficulties in determining the subtle contributions of process variables such as temperature on the chromatogram shapes. The importance of effective pre-processing of chromatogram data prior to performing PCA has been noted before [Andersson and Hamalainen, 1994] where iteration techniques were employed which adjusted the retention times and shapes of test peaks in order to match target peaks. This pre-processing was performed on four test samples containing differing concentrations of the same constituents. The study led to a more effective use of PCA by accounting for phenomena such as retention time drift between samples. However the main benefit of pre-processing appeared to be that of reducing the number of principal components needed to describe the data. No investigation of the effects of other process variables such as temperature, flow rate or load volume-themselves resulting in retention time differences between samples was reported.

Different pre-processing techniques have been applied in order to gain improved diagnosis from the PCA with respect to isolation of product from contaminants [Sanchez et al, 1994]. The pre-processing used different centring techniques which included column centring, log column centring and log row centring. Although not directly related to the work reported in this chapter, Andersson and Hamalainen (1994) and Sanchez et al (1994) serve to display that intelligent use of pre-processing prior to PCA can improve the quality of information being derived from this technique.
This chapter examines the additional diagnostic information that may be obtained from pre-processing prior to PCA. The problem of distortion to the diagnosis caused by the effects of flow rate changes is addressed by analysing the chromatogram data on a volumetric basis rather than on the conventional time basis and we seek to examine whether accounting for the amount of material loaded onto the column could provide additional diagnostic insight using PCA, with respect to the minor process variables.

4.3. Materials and Methods

4.3.1. The chromatographic process

PCA was applied to data sets of chromatograms obtained from a test size exclusion chromatography system (SEC) which mimics a real industrial scale separation for the isolation of a therapeutic protein in its monomeric form, free from contaminating dimer and trimer protein aggregates [Chandwani et al (1997), Chandwani’s thesis (1995)]. For a more complete description of the materials and methods, consult Chandwani’s PhD thesis (1995). Only a brief summary is presented here.

The test SEC system comprised of a column packed with Sephacryl® S-200 SF resin (Pharmacia Biotech AB, Uppsala, Sweden), a volume of 23.4 ml and internal diameter 10 mm. The mobile phase was Tris base with the pH being set to 7.2-7.5 using hydrochloric acid [Chandwani et al, 1997]. The column was controlled using a Beckman System Gold® modular system (Beckman Instruments Inc., CA, USA) comprising an autosampler, dual pump and UV detector linked to a central IBM personal computer. The synthetic mixture separated by the column comprised myoglobin (horse heart), ovalbumin (grade VI) and bovine serum albumin (BSA) (Sigma Chemical Co., MO, USA) which represented the monomer, dimer and trimer respectively. The protein ratio in the synthetic mixture was held constant at 17:2:1 in order to reduce one factor leading to batch-to-batch variability and also to reduce the number of factors influencing the separation. The synthetic mixture was made up in the mobile phase.
The process variables studied were: total protein concentration (C) and volume (V) of the load; flow rate of the mobile phase (F) and the column temperature (T). A typical UV profile obtained from the separation is given in Figure 1 and illustrates the presence of three distinct peaks corresponding to the three proteins contained within the mixture. When the combination of high volume and high concentration was examined, the resulting chromatograms had flattened peaks. This was due to the UV detector reaching saturation point.

![Figure 4.1. An example of a typical chromatogram obtained from the test SEC system illustrating the presence of three peaks.](image)

### 4.3.2. Generation of historical data sets

Two data sets were generated using an experimental design approach. The first dataset, \( X_1 \), used a full factorial design approach and in the second dataset, \( X_2 \), just one of the four process variables was varied for any one separation. The chosen experimental design was comparable with that which may be evident in archived bioprocess production records (Parenteral Drug Association, 1992). A high level ' +1 ', a low level ' -1 ' and a centre point level were assigned for each process variable. The actual values for the four variables at each of these three levels are given in Table 4.1.
The design matrices for $X_1$ and $X_2$ are displayed in Tables 4.2 and 4.3 respectively. $X_1$ contained 72 chromatograms (16 from the full factorial design from Table 4.2 run four times each, plus eight centre runs). $X_2$ contained 47 chromatograms (eight from the experimental conditions from table 4.3 run five times each, plus seven centre runs).

<table>
<thead>
<tr>
<th>level</th>
<th>T°C</th>
<th>V%</th>
<th>C/mg.mL$^{-1}$</th>
<th>F/mL.min$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>18.0</td>
<td>1.0</td>
<td>1.00</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>23.0</td>
<td>4.0</td>
<td>2.00</td>
<td>0.15</td>
</tr>
<tr>
<td>+1</td>
<td>28.0</td>
<td>7.0</td>
<td>3.00</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 4.1. A description of the levels for the factors (process variables) used to generate the datasets $X_1$ and $X_2$. '+1' represents a high setting, '-1' a low setting and '0' a centre value for the factor. The volume of the load ($V$) is expressed as a percentage of the bed volume (23.4 mL).

<table>
<thead>
<tr>
<th>Trial</th>
<th>T</th>
<th>V</th>
<th>C</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>9</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>10</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>11</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>13</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>14</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>15</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>16</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.2. Summary of the experimental design used to generate the first dataset $X_1$. The table describes a 2-level full factorial design and a centre point. Each of the 16 combinations of factors was repeated 4 times and the centre point repeated 8 times, generating 72 chromatograms in total. The values for each level are given in Table 1. The chromatograms were arranged as a row in matrix $X_1$.

<table>
<thead>
<tr>
<th>Trial</th>
<th>T</th>
<th>V</th>
<th>C</th>
<th>F</th>
<th>repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>v</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>5</td>
</tr>
<tr>
<td>t</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4.3. A summary of the design for generating the second dataset $X_2$. The values for the levels for each factor are given in Table 2. The 47 chromatograms were arranged in rows in matrix $X_2$. 
The labelling convention used in this chapter is that upper case is used for high level setting of the variables and lower case for low level. Thus TvcF represents a chromatogram obtained under the conditions of high temperature, low volume, low concentration and high flow rate. Groups of chromatograms in this chapter are observed to have concentration and volume in common. An example is of such a group is cluster 'constant vC' which contain chromatograms obtained from settings of low volume and high concentration.

4.3.3. Flow Rate Transformation
In order to realise a more accurate representation of the data it was necessary to compensate for the dominating influence of flow rate. This was effected for all chromatograms by transforming the time variable to a volumetric basis. This transformation served to align the chromatograms. Without such a transformation, comparisons between chromatograms will tend to compare regions containing peaks in one chromatogram with zero data in another. Hence this pre-processing enabled better comparisons between peak shapes between chromatograms to be made.

4.3.4. Sample Load Adjustment
Since it is known that the area under a chromatogram is proportional to the mass of material loaded, it was decided to investigate whether adjustment of the areas of each chromatogram to a common value could result in improved diagnostics from the PCA. The reason for this adjustment was to assess whether the PCA could isolate very subtle changes in the chromatograms and also to ascertain whether too much pre-processing could reduce the utility of the PCA technique.

4.5. Results and Discussion
An overview of all 72 chromatograms in dataset X1 is given in Figure 4.2. It shows three distinct peak positions corresponding to high, centre and low flow. From earlier work [Chandwani et al, 1997] it was known that if the PCA were performed on this conventional time basis, the dominant effect of flow rate tended to obscure some of the more subtle process variations. To illustrate this, Figure 4.3 shows a plot of the first two PCs after performing PCA on dataset X1. A five principal component model.
was proposed using this technique where PC1 accounted for 73.32% of the variation, PC2 17.37%, PC3 7.57%, PC4 1.01% and PC5 0.46%. The amount of variation accounted for by the first two PCs is 90.69%. There are nine distinct clusterings, all of which are very compact.

![Graph showing high and low flow rates](image)

Figure 4.2. An overview of all 72 chromatograms in X1. There are 3 distinct groupings corresponding to high, centre and low flow rates. Due to these significant positional shifts, the effects due to the other process variables (e.g. concentration, temperature etc.) tend to be masked when performing the PCA.

From the factorial design (Table 4.2), it appears that there should be 17 clusters: 16 from the 16 different operating conditions and one for the set of centre runs. It was found with this form of data presentation that the temperature variable appeared to have a negligible effect on the shape of the chromatograms. This was surprising and it was concluded that the effect of temperature was hidden by the dominating effect of the flow rate.
Figure 4.3. Scores plot for PC1 and PC2 for the gross model for the full factorial data set X1 without pre-processing. There are 9 very compact clusterings. Due to the dominance of flow rate, the temperature variable appears to have only a negligible effect on the shape of the chromatograms.

4.5.1. Analysis of the gross principal component model for dataset X1 after flow rate transformation

Figure 4.2 shows the three peak positions of the 72 chromatograms corresponding to the three different flow rates employed in this study while Figure 4.4 shows the relative positions of these chromatograms after transformation for flow rate. Not only do the peaks shift so they are aligned, they also broaden/narrow in the process in accordance with the flow rate at which they have been generated. Hence the chromatograms obtained at the high flow rate are the narrowest on the time basis, and broaden after transformation, while those generated at the low flow rate are the broadest on the time basis and narrow on transformation.
Figure 4.4. The same 72 chromatograms as in Figure 4.2 after pre-processing where the data was transformed from time to volume in order to compensate for the flow rate variable. Notice how the peaks broaden/narrow in the process. Subsequent PCA was performed on data pre-processed in this way.

Following the flow rate transformation process, data set $X_1$ was centred so that the mean of the centre point profiles were subtracted from each of the other profiles. This was to ensure that the resulting scores clusterings were positioned relative to the centre point scores located around zero values in the scores plane. PCA was then performed on data set $X_1$. The gross (transformed) principal component (PC) model, for the 72 chromatograms in dataset $X_1$ resulted in a three PC model accounting for 90.63% of the total variation in $X_1$. [PC1 accounted for 86.94%, PC2 2.17% and PC3 0.52%]. Relationships amongst the 72 chromatograms are shown in the scores plot of the first two principal components (Figure 4.5). The plane spanned by these first two PCs accounts for 89.11% of the variation in $X_1$. This scores plot reveals the presence of five distinct clusters of chromatograms - four clusters each containing 16 chromatograms and a cluster of the eight centre point chromatograms. Each cluster contains chromatograms obtained under the same conditions of load volume and concentration. It is evident from Figure 4.5 that the cluster spread reduces going from
right to left on this scores plot. Cluster VC is the most spread and cluster vc the least. This trend reflects the actual size of the chromatograms - cluster VC contains the largest profiles and vc the smallest. When the PCA is performed on the entire dataset, the smaller profiles display least variation in size and shape, probably due to the fact that resolution of the peaks is not as good in the case of the separations performed with large load volumes. Another possible reason for this increased spreading with the larger volume separations is the fact that the separations were performed by the addition of several 'slugs' of material rather than a single injection and this tended to enhance the effects of diffusional band broadening in the column [Chandwani et al, 1997].

Figure 4.5 Scores plot of the first two principal components for the gross (transformed) model for XI. The plot shows the qualitative positions of clusters of similar chromatograms. e.g. Members of cluster constant VC had high concentration and high volume in common. The eight centre runs also form a distinct cluster.

If the data were not transformed prior to performing the PCA, the scores plot for the first two principal components would appear as shown in Figure 4.3. The clusters in Figure 4.3 are very compact. Using the conventional time basis to represent the data leads to certain misleading diagnostic information being drawn. For example cluster TVCF, tVCF (marked A) and cluster tVCf, TVCf (marked B) in Figure 4.3 appear to be very distant from each other within the scores plane. However when these same
chromatograms are displayed on a volumetric basis (Figure 4.5), these chromatograms appear in the same cluster. Since from Figure 4.5, it is evident that the temperature or flow rate effects are not as significant as those due to concentration and volume, it is correct to expect that groups A and B should be in close proximity. It is the ability of the PCA to detect the actual shapes and subtle changes in the transformed profiles that enhances the amount of diagnostic information obtained.

It is concluded that analysis on the pre-processed volumetric basis is preferable to analysis performed on the conventional time basis where the dominating flow rate effect tended to result in these subtle changes resulting from the other process variables being masked. Since it is known from a process viewpoint that an increase in flow rate will cause separated material to elute earlier, the transformation of the chromatograms to a volumetric basis to account for the time shifts is a beneficial method of data pre-processing.

Figure 4.6. Scores plot of PC1 and PC3 for the gross (transformed) model for X1. The plot is very similar to that for PC1 and PC2 (figure 4.5) and reveals that the higher component, PC3 reveals no more information regarding subtle changes in the chromatograms. Hence only a 2 principal component model is necessary to describe the gross (transformed) data set X1.

Often in PCA examination of scores plots for the higher components may reveal some subtleties concerning minor variables. However in this case no additional information was gained from PC3 or higher. Figure 4.6 show a scores plot of PC1 and PC3 for the
gross (transformed) model for dataset \( \mathbf{X}_1 \). On inspection this plot is very similar to that for PC1 and PC2 (Figure 4.5). In order to isolate the temperature and flow rate effects closer examination of the individual clusters was required. For the well-spread clusters (constant VC and constant Vc) some of these effects are visible from the gross model, but obviously with the tighter clusters (constant vC and constant vc) it is necessary to 'break up' these clusters by the development of disjoint models in order to reveal the subtleties. A disjoint principal component model contain data only from the cluster of interest in order to reveal the minor differences between chromatograms not apparent from the gross model (Wold et al, 1987). The tightest cluster (constant vc) was used as the example in this chapter.

4.5.2 Analysis of the disjoint model for the cluster at constant vc

It was apparent, having created a disjoint model for the cluster at constant vc that there were three qualitative outliers. When seeking to discover why these three outliers may have occurred, it was discovered that the apparent flow rates were outside the acceptable tolerance limit of ±2%. The pump was subsequently recalibrated. After removal of outliers a new disjoint model was developed for the remaining 13 chromatograms in cluster vc. A three principal component model accounting for 98.91% for the variation was proposed for the disjoint model of vc minus outliers (PC1 90.96% of the variation, PC2 5.18% and PC3 2.75%). The scores plot for the first two principal components is shown in Figure 4.7.

This plot gives a better indication of what the scores plot for cluster vc should look like if it not were for the outliers. The clusters indicate distinctive splitting of all the operating conditions and the arrangement of the clusters appears quite symmetrical. Additionally one can assess qualitatively the effect of how the positions on the scores plot alter with an increase (or decrease) of flow rate and temperature. The alternative method of investigating the subtle contributions of temperature and flow rate is simply to magnify cluster vc from Figure 4.5. Figure 4.8 is the scores plot from figure 4.5 but focused very closely in on cluster vc. Similar clusterings isolating the minor effects due to flow rate and temperature are observed, although rather more outliers are evident than with the disjoint model.
Figure 4.7. Disjoint model for cluster obtained at constant vc with the outliers removed. This plot gives an indication of what the disjoint model scores plot should look like if it not were for the outliers. Indicated on the plot are the effects of increasing temperature or flow rate.

This magnified plot has the advantage that the position of each chromatogram is relative to the gross model, whereas the disjoint model shows the positions relative to members of that cluster. The detection of outliers is made straightforward when the analysis is performed on the volumetric basis. Without transformation the outliers are impossible to detect.

Figure 4.8. Magnified scores plot of cluster obtained at constant vc from Figure 4.4 to be compared with the disjoint model approach (Figure 4.7). It is clear that distinctions between flow rate and temperature are visible at this scale. This magnified plot has the advantage over a disjoint model in that the position of each chromatogram is relative to the gross model. The arrows indicate which cluster the respective point should belong to.
4.5.3 Analysis of the gross principal component model for dataset X2

Performing PCA on dataset $X_2$ (one at a time variable changes from Table 4.3) without transforming from time to volume resulted in the scores plot for the first two principal components shown in Figure 4.9 [Chandwani et al., 1997]. This plot reveals the presence of six distinct clusters of apparently very similar chromatograms.

![Scores plot for PCI and PC2 in the gross model for the one at a time data set $X_2$ without adjustment for flow rate (Chandwani et al., 1997). This plot reveals the presence of six distinct and compact clusters of apparently very similar chromatograms.](image1)

**Figure 4.9.** Scores plot for PC1 and PC2 in the gross model for the one at a time data set $X_2$ without adjustment for flow rate (Chandwani et al., 1997). This plot reveals the presence of six distinct and compact clusters of apparently very similar chromatograms.

![Scores plot of the first two components for the gross (transformed) model for $X_2$. The plot shows the qualitative positions of clusters of similar operating conditions. There are 9 distinct clusters as predicted by the experimental design from Table 4.3. There appear to be two qualitative outliers; $F^*$ should be a member of cluster $F$, and $C^*$ which should be a member of cluster $C$.](image2)

**Figure 4.10.** Scores plot of the first two components for the gross (transformed) model for $X_2$. The plot shows the qualitative positions of clusters of similar operating conditions. There are 9 distinct clusters as predicted by the experimental design from Table 4.3. There appear to be two qualitative outliers; $F^*$ should be a member of cluster $F$, and $C^*$ which should be a member of cluster $C$. 

104
Performing PCA on the gross (transformed) dataset $\mathbf{X}_2$ resulted in a three principal component model accounting for 88.33% of the variation. [PC1 accounted for 73.52% of the variation, PC2 13.55% and PC3 1.31%]. Figure 4.10 shows the scores plot for the first two principal components for the gross (transformed) model for $\mathbf{X}_2$. Each of the operating eight operating conditions (Table 4.2) and the centre point runs form the expected nine distinct clusters. Chromatograms centre, T and t which were indistinguishable in the untransformed scores plot (Figure 4.9) are now clearly distinguishable in the scores plot using the transformed model. The qualitative outliers $F^*$ and $C^*$ are indicated and should belong to clusters F and C respectively.

4.5.4. Analysis of data set $\mathbf{X}_1$ with adjustment for sample load (after flow rate transformation)

It was decided to adjust the area under each profile to a constant value of equal to that of the mean area of the centre point chromatograms. Since the area under the profile is proportional to the mass loaded onto the column, this particular method of adjustment for load was valid and should enable the PCA method to focus on subtle peak shape shifts which can be a vital indicator of changes in column performance. Figure 4.11 shows how all 72 profiles have been adjusted. When PCA was performed on these pre-processed data (after centring), PC1 accounted for 88.55%, PC2 8.77% and PC3 1.98% of the variation within the data.

The scores plot for the first two of these components is displayed in Figure 4.12. This plot yielded no supplementary diagnostic information compared with when the analysis was performed with only a time to volume transformation (Figure 4.5). It is concluded that adjustment of chromatograms to a constant load did not provide any supplementary useful diagnostic information.
Figure 4.11. All 72 chromatograms adjusted to constant area to reduce the influence of the mass loaded in the analysis. These chromatograms have previously been transformed from time to volume to account for the dominating effect of flow rate.

Figure 4.12. Scores plot for the first two principal components for data set XI following transformations to account for both flow rate and mass load effects.
4.5. Conclusions

Principal component analysis (PCA) has been used for the exploratory diagnosis of chromatographic separations. Variations in flow rate result in significant positional shifts of the chromatograms. Use of a pre-processing stage to transform the chromatograms so that they are represented on a volumetric basis, instead of time, was included prior to performing PCA. Following transformation, PCA was able to isolate correctly the effects of the minor flow rate and temperature variables which were not immediately evident from the scores plots of the gross model and which were masked by the dominating influence of flow rate generated using the pre-transformed data set. Adjustment of chromatogram areas to account for sample load size did not lead to improved diagnosis from PCA in this study.

Investigations of subtle effects was achieved by either creating a disjoint principal component model or by magnifying the clusters on the gross model scores plot. Both methods were observed to perform better in terms of quality of diagnostic information obtained when using non-transformed data. The ability to detect subtle changes between chromatograms obtained at slightly different operating conditions and of detecting outliers make PCA ideal for use in process chromatography where changes in column performances are often small but can be indicative of deteriorating column performance. In conclusion, the main benefits of the time-to-volume pre-processing stage are as follows:

- The PCA is able to detect differences between the actual shapes of chromatograms. Without the transformation, the PCA only served to isolate the trivial effect of baseline positions of chromatogram peaks inherent of the flow rate used to generate them;
- Subtle differences in chromatograms can now be identified such as those resulting from a small change in temperature;
- Outliers are much more easily identifiable.

The sample load adjustment was not beneficial and the suggestion is that this should not be performed.
In the next chapter the techniques developed for improved diagnosis will be applied in the analysis of data obtained from the separation of erythromycin which form the experimental core of the thesis. Additional methods will also be developed so as to gain maximal information from the raw data sets.
5. The Use of Principal Component Analysis for the Modelling of High Performance Liquid Chromatography

5.1. Summary

5.2. Introduction

5.2.1. The Modelling of Liquid chromatography

5.2.2. Principal Component Analysis

5.3. Materials and methods

5.3.1. The Separation Method

5.3.2. The Erythromycin Samples

5.3.3. The Modular HPLC system

5.3.4. Column Cleaning Procedure.

5.3.5. Generation of data sets at the 3 scales of operation.

5.3.6. Data pre-processing

5.3.6.1. Retention time transformation

5.3.6.2. Centring of chromatograms

5.4. Results and Discussion

5.4.1. Analysis of the full factorial data set for the 25 cm column (with retention time adjustment)

5.4.2. Selection of subsets of chromatograms - the framework for scale-up predictions

5.4.3. PCA of subsets of chromatograms (adjusted for retention time effects).

5.4.4. Analysis of chromatograms derived from columns of different lengths

5.4.4.1. Retention times

5.4.4.2. Principal component scores

5.4.5. Relationship between principal components and chromatographic profiles

5.4.6. Prediction of chromatographic profiles

5.4.6.1. Goodness of fit

5.5. Conclusions
Chapter 5. The Use of Principal Component Analysis for the Modelling of High Performance Liquid Chromatography

5.1. Summary

Principal component analysis (PCA) was used to analyse the behaviour of a chromatographic separation as its scale increased. Three 4.6 mm diameter columns identical in every respect except for column length (25, 15 and 5 cm), were used to generate the data from a test system based on the reversed-phase HPLC separation of crude erythromycin on a polystyrene matrix (PLRP 1000) having a particle diameter of 8 μm and a pore diameter of 100 nm. The species were separated with an isocratic solvent composed of 45/55 acetonitrile/water at about pH 7. An experimental design technique was used to investigate the effects of four process variables (load volume, load concentration, temperature and pH of buffer) on the chromatogram shapes. Following appropriate pre-processing of the chromatographic data, subsets of critical chromatograms were selected which sufficiently characterised the entire data set. From this subset, the corresponding runs were performed on the different sized columns and principal component models were generated for each. At 5 and 15 cm a single principal component was sufficient to characterise all the variance in the chromatograms which the range of process variables introduced, but at 25 cm two principal components were required, particularly to characterise the chromatograms with small loads. Excellent correlations were observed between the first principal components at the three scales. The possibility of predicting the separations on the 25 cm column from an analysis of the separations observed at 5 cm was investigated. The study revealed that good predictions could be made at high loads (>92%), but the model was not effective at low loads because of the need to incorporate a second principal component which was not defined by the range of variables applied to the 5 cm column.
5.2. Introduction

5.2.1. The Modelling of Liquid chromatography

Chapter 2 provided an in-depth summary of the difficulties experienced in the modelling of liquid chromatography. A brief resume of these problems is presented here.

Traditionally, mathematical models have been used to model chromatographic behaviour. These models are based mainly on the mass balance or continuity equation

\[
\frac{\partial C_i}{\partial t} + F \frac{\partial q_i}{\partial t} + u \frac{\partial C_i}{\partial z} = D_{li} \frac{\partial^2 C_i}{\partial z^2} 
\]

(1)

\(C_i\) = local concentration of species i in the mobile phase, mol/m³
\(q_i\) = local concentration of species i in the stationary phase, mol/m³
\(z\) = axial dimension, m
\(D_{li}\) = axial dispersion coefficient of i in the mobile phase (this includes contribution to axial dispersion due to molecular diffusion and non-homogeneity of flow), m²/s
\(F\) = phase ratio = \((1-\varepsilon)/\varepsilon\) = volume stationary phase/volume mobile phase.

Each species in the system requires its own mass balance including each species in the mobile phase.

There are several fundamental assumptions for the validity of equation (1). These are detailed in Guiochon et al (1994) and include: an incompressible mobile phase; constant packing quality across the width of the column (radially homogeneous); constant viscosity of mobile phase; constant \(D_{li}\); no solvent adsorbed; isothermal operation.

Equation (1) can be modified and extended to obtain solutions. For example an ideal system (assumes infinitely efficient column) would have no axial dispersion and so the RHS diffusive term would equal zero. Other, more realistic models include the Equilibrium-Dispersive model and the Lumped Kinetic models (see Guiochon et al, 1994). These models account for non-linear phenomena such as: mass transfer effects (e.g. intraparticular diffusion); adsorption/desorption kinetics; equilibrium

111
thermodynamics; eddy diffusion. All of these effects are important and are extremely complex to model.

Adsorption isotherms are used to provide the equilibrium data for concentrations of species in the mobile and stationary phases $C_i$ and $q_i$. There are many types of types of adsorption isotherm, which can be extremely complex. These are well-documented in Guiochon et al (1994) and include: Langmuirian, bi-Langmuirian, Freundlich, Fowler and S-shaped isotherms. The shapes of these isotherms reflect the binding ability as the concentration of a pure species increases on a particular stationary phase under constant conditions such as temperature. The Langmuirian isotherm has the general form

$$q_i = \frac{a_i C_i}{1 + b_i C_i}$$

(2)

Where $a$ and $b$ are constants. As the complexity of a separation increases, i.e. as the number of species in the mixture increases, the equilibria become much more difficult to model because of the competition for binding sites. There have been attempts to construct competitive isotherms. Felinger and Guiochon (1996) used a competitive Langmuirian model of the form

$$q_i = \frac{a_i C_i}{1 + \sum_{j=1}^{n} b_j C_j}$$

(3)

in which the binding ability of species $i$ was related to the concentrations and binding abilities of the $n$ total species in the mixture.

The main problems of modelling chromatography (specifically non-linear chromatography) are summarised below:

- The mass balance equation (1) is a complicated 2$^{nd}$ order partial differential equation and relies on several fundamental assumptions.
- There are many phenomena occurring simultaneously which further serve to provide complications. These include adsorption/desorption kinetics, intraparticular mass- transfer, equilibrium thermodynamics and eddy diffusion.
• Complex adsorption isotherms are required which generally require pure species for their measurement.
• Multi-species separations are currently impossible to model accurately in their entirety.

This chapter investigates the utility of using PCA for the modelling of chromatography which does not use the assumptions of any physico-chemical phenomena. This approach has been applied to a complex chromatographic separation of crude erythromycin. The chapter further examines the possibility of using PCA to correlate behaviour between different sized columns which would have potential for scale-up purposes.

5.2.2. Principal Component Analysis

Chapter 3 presented the theory behind Principal Component Analysis. The software used throughout this study uses the NIPALS algorithm (section 3.6.1) to calculate the principal components.

PCA has been used effectively in several chromatographic applications. Meglen (1991) used it for data base mining, and Jokinen (1994) as a tool to provide quantitative information, via scores plots, about the current status of industrial processes. Malmquist and Danielsson (1994) analysed entire chromatograms for peptide mapping as a quality control aid for recombinant protein production. Statheropolous et al (1996) used PCA for resolving coeluting substances in doping control analyses. Finally Chandwani et al (1997) used the scores plots to diagnose quickly whether batches of SEC chromatograms were within a process specification.

Pate et al (1998) (Chapter 4) extended these ideas to examine how the quality of diagnostic information could be improved by the appropriate pre-processing of the chromatogram data. This study revealed that, before implementing PCA, a transformation from the conventional time basis to a volumetric basis is necessary when chromatograms are generated with different flow rates.
There are no reported applications of PCA being used as a tool for modelling or scaling-up of chromatography. The main focus in this chapter is to predict the nature of chromatograms which would be obtained using a 25 cm column from data at a 5 cm scale covering a 25-fold range of sample loads. A secondary aim of the study is to identify the minimum quantity of information necessary for satisfactory predictions from small to large scale.

5.3. Materials and methods
5.3.1. The Separation Method
The mechanism used for the separation is reverse phase chromatography which is the most common technique used in high performance liquid chromatography. The following method was the basis for the reversed-phase HPLC separation of erythromycin. It was based on that outlined by Paesen et al. (1993). Further method development at UCL enabled the removal of 2-methyl-2-propanol from the mobile phase. The pH was reduced to 7.0 from the recommended 9.0. These adjustments led to a reduction of the run time from 80 minutes to 15 minutes.

- **Columns:** PLRP-S [poly(styrene-dininylbenezene)] 8 µm particle diameter, 1000 Å pore size (Polymer Laboratories, Church Stretton, Shropshire, England). Dimensions of the three columns: 25 cm×4.6 mm, 15 cm×4.6 mm and 5 cm×4.6 mm.
- **Mobile Phase:** 45% acetonitrile (‘Hipersolv™‘ far UV grade for HPLC, BDH Laboratory Supplies, Poole, UK), 50% water, 5% 0.2M potassium phosphate buffer at pH 7.0 which comprised potassium dihydrogen orthophosphate (KH₂PO₄), (Fisons Scientific Equipment, Loughborough, Leics., UK) and potassium phosphate (K₂HPO₄), (Sigma Chemicals, St. Louis, MO, USA). 0.2 M of both salts were dissolved in water. The KH₂PO₄ had a pH of around 9 and the K₂HPO₄ of around 4.3. The required pH was set using a Mettler-Toledo pH meter by mixing the two buffers. The water was de-ionised using a Purite water purification system. The mobile phase was filtered using 0.2 µm Nylon 66 filters (Phenomenex, Macclesfield, UK) and degassed using compressed helium (BOC) before introduction into the HPLC system.
- **Flow rate:** 1.2 mL/min
• **Sample volume:** 100 μL.

• **Sample concentration:** 10 mg/mL.

• **Temperature:** 64 °C ± 0.1 °C.

• **Run time:** 15 minutes at flow rate of 1.2 mL/min.

• **UV Wavelength:** 215 nm.

These conditions define the 'centre point' or process template runs for the 25 cm column. Variations from these setting are presented in Table 5.1.

<table>
<thead>
<tr>
<th>Level</th>
<th>C/(mg/mL)</th>
<th>T/(°C)</th>
<th>P/(pH units)</th>
<th>V/(μL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25 cm</td>
<td>15 cm</td>
<td>5 cm</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>5</td>
<td>60</td>
<td>6.75</td>
<td>50</td>
</tr>
<tr>
<td>-</td>
<td>7.5</td>
<td>62</td>
<td>6.88</td>
<td>75</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>64</td>
<td>7.00</td>
<td>100</td>
</tr>
<tr>
<td>+</td>
<td>15</td>
<td>66</td>
<td>7.12</td>
<td>150</td>
</tr>
<tr>
<td>+2</td>
<td>20</td>
<td>68</td>
<td>7.25</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 5.1. Table showing the values of each variable at their respective levels. The volume of sample loaded onto the column is proportional to the size of the column.¹

5.3.2. **The Erythromycin Samples**

Erythromycins are a sub-group from the macrolide family of antibiotics which are well-established antimicrobial agents in both clinical and veterinary medicine. These agents can be administered orally and are generally used to treat infections in the respiratory tract, skin and soft tissues, and genital tract caused by gram-positive organisms [Kirk and Othmer, 1991]. The macrolides are very structurally diverse and are produced by the fermentation of soil micro-organisms and are divided into three families according to the size of the aglycone ring which can be 12, 14 or 16-membered. The erythromycins are 14-membered ring macrolides. Erythromycin A is the most widely used macrolide antibiotic and was the principal product found in culture broths of *Streptomyces erythreus*.

Erythromycins are unstable in acidic conditions which makes them susceptible to degradation in the stomach. Acid-stable salts and/or esters which are relatively water-insoluble have been employed to protect erythromycin during passage through the

¹ Usually the sample volume increases in proportion to column diameter. The validity of increasing sample with column length can be slightly misleading (H. Colin from private consultation).
stomach, to increase bioavailability and to decrease the variability of oral absorption. The empirical formula of Erythromycin A is C_{37}H_{67}NO_{13} which corresponds to a molecular weight of 733.5. It is basic having a pK_a value of 8.6. It exists in two allotropic modifications which melt at 135-140 and 190-193 °C respectively [Kirk and Othmer, 1996].

The principal use of erythromycin is in group A beta-haemolytic streptococcal, staphylococcal and pneumococcal infections. It is also active against many atypical mycobacteria. It often acts as a substitute in individuals with streptococcal or pneumococcal infections who are hypersensitive to penicillin. It is widely used to treat A.I.D.S patients because it is effective against a wide range of infections. Usual dosage expressed in terms of erythromycin base is 250-500 mg every six hours in adults although in severe cases this may be doubled. Adverse reactions to oral administration may occasionally be experienced which may include nausea, vomiting and diarrhoea.

The semi-purified erythromycin used in this study was supplied by Abbott Laboratories, N.Chicago, IL, USA. It’s potency was quoted as 0.779 mcg/mg and hence it was necessary to handle the material with caution. The samples of erythromycin that were loaded on the column were all dissolved in the mobile phase (45% ACN, 50% water, 5% buffer). The most concentrated sample (20 mg/mL) was difficult to dissolve and involved reducing the size of the erythromycin particles slightly in a mortar and pestle followed by mixing using a magnetic flea and stirrer for about two hours. All the erythromycin samples (both dissolved samples and the crude powder) were kept in a refrigerator at 4 °C prior to use. The samples were filtered into 2mL sample vials (Kimble vials, Sigma Chemical Co., St. Louis, MO, USA) using 0.2 µm syringe filters (Phenomenex, Macclesfield, UK) before positioning in the autosampler.

5.3.3. The Modular HPLC system
In this chapter, PCA was employed on chromatographic data from the crude erythromycin separation at three different sizes of column length. The HPLC system used was a Beckman modular system (Beckman Instruments Inc., CA, USA)
comprising an Autosampler 507, Solvent Module (pumps) 126 and a programmable ultraviolet (UV) Detector Module 166. The column used for the separations was temperature-controlled to ± 0.1 °C via a thermostat. All of these mechanical components were controlled using an IBM PS/2 personal computer. Figure 5.1 is a schematic of the analytical HPLC system. The software used for the control and operation was the System Gold® (Beckman Instruments Inc.).

The 507 Autosampler was fitted with sample loops of various size depending on the sample volume being investigated for a particular run. The autosampler enables consecutive separations to be performed without having to inject manually each time.

Figure 5.1. A schematic of the Beckman analytical HPLC system comprising UV detector, autosampler and solvent delivery module all linked to an IBM personal computer.
The 126 Solvent Module comprises two pumps each of which controls up to four solvents. The use of two pumps is especially of benefit when performing gradient elutions. The separations performed in this thesis were isocratic in nature so one pump sufficed. In practice however each pump was used operating at 50% capacity to improve accuracy. This also spreads the wear on each of the pumps. Initial calibration of the two pumps was performed at the start of the experiments. Flow rate checks were subsequently performed periodically to confirm that the pumps’ performance remained within the tolerance guidelines given by the manufacturer. This check involved pumping solvent into a measuring cylinder at a flow rate of 1 mL/min for 10 minutes. The pumps were tested individually and were deemed to be within tolerance if the volume collected was between 9.8 and 10.2 mL (+/- 2%).

The UV Detector Module 166 is capable of operating at any wavelength in the range 150 nm to 600 nm. The maximum absorbance of erythromycin is 190 nm. At this wavelength however, acetonitrile is strongly absorbed. For this reason a wavelength used was 215 nm.

The run times were 15 minutes, 9 minutes and 4 minutes for column lengths 25 cm, 15 cm and 5 cm respectively. The UV profiles used to provide the data for the PCA comprised 300 absorbance values for each run. The following process variables were investigated using an experimental design approach: load concentration, C; load volume, V; column temperature, T and pH of potassium phosphate buffer, P.

5.3.4. Column Cleaning Procedure.
Due to the opportunity for build up of non-eluted material to occur after successive runs, it was deemed necessary to clean the column periodically. This was to ensure that high back-pressures were not experienced and also to contribute to the reproducibility of the experiments. A thorough cleaning procedure also prolongs the life of columns and leads to the efficiencies not being greatly diminished with time. Manufacturers usually offer cleaning guidelines which are column-specific. For the erythromycin separation on the polystyrene columns, the following cleaning procedure was used:
50% v/v Methanol (far UV grade for HPLC, BDH Laboratory Supplies, Poole, UK) 49.5% v/v water and 0.5% v/v 85% phosphoric acid. This mixture was run through the column (after filtering and de-gassing) at the specified flow rate for about 10 column volumes. The column was taken to be clean when a constant zero absorbance value had been achieved for about 10 minutes. Erythromycin is a base and the role of the phosphoric acid was to aid the rapid removal of erythromycin left on the column from previous runs. The column was cleaned after approximately every five runs except when the column had been overloaded. In this instance, the column was cleaned after every two runs. Full equilibration of the column was implemented before each separation after cleaning.

5.3.5. Generation of data sets at the 3 scales of operation.

In order to investigate the ability of PCA to predict changes between different scales of operation, an experimental design approach was used to generate a series of chromatograms at the 25 cm scale. The two-level full factorial design also incorporated one-at-a-time changes to each of the variables at four levels. Centre point runs were included so that the design investigated 5 levels of each variable: '+2' and '+' were the high settings; '-2' and '-' were the low settings and '0' was the centre point setting. The total number of runs using this design was 36 and each was performed three times giving 108 chromatograms in total. This design matrix is given in Table 5.2.

The order of the runs was randomised to guard against systematic bias. When performing runs at the different scales, the only variable which was changed was load volume (V) which was adjusted in direct proportion to the volume of the column. The range of sample loads applied to each of the 3 columns is given in Table 5.3.
Table 5.2. Table showing the levels of the variables for the full factorial data set using the 25 cm column. The design includes one-at-a-time changes of each variable at four levels (+, -, +2 and -2). Four centre point runs are also included. Each of the 36 runs were performed three times giving a total of 108 chromatograms.

<table>
<thead>
<tr>
<th>Run</th>
<th>C</th>
<th>V</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>cvtp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2.</td>
<td>cvtP</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>3.</td>
<td>cvTp</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>4.</td>
<td>cvTP</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>5.</td>
<td>cVtp</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>6.</td>
<td>cVtP</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>7.</td>
<td>cVTp</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8.</td>
<td>cVTP</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>9.</td>
<td>Cvtp</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10.</td>
<td>CvtP</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>11.</td>
<td>CvTp</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>12.</td>
<td>CvTP</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>13.</td>
<td>CVtp</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>14.</td>
<td>CVtP</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>15.</td>
<td>CVTp</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>16.</td>
<td>CVTP</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>17.</td>
<td>c</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18.</td>
<td>C</td>
<td>+</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19.</td>
<td>V</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>20.</td>
<td>V</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>21.</td>
<td>t</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>22.</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>23.</td>
<td>p</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24.</td>
<td>P</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25.</td>
<td>cc</td>
<td>-2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26.</td>
<td>CC</td>
<td>+2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27.</td>
<td>vv</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>28.</td>
<td>VV</td>
<td>0</td>
<td>+2</td>
<td>0</td>
</tr>
<tr>
<td>29.</td>
<td>tt</td>
<td>0</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>30.</td>
<td>TT</td>
<td>0</td>
<td>0</td>
<td>+2</td>
</tr>
<tr>
<td>31.</td>
<td>pp</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32.</td>
<td>PP</td>
<td>0</td>
<td>0</td>
<td>+2</td>
</tr>
<tr>
<td>33.</td>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34.</td>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35.</td>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36.</td>
<td>centre</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.3. Table showing the range of both sample volumes and mass of material added to each of the three columns.

<table>
<thead>
<tr>
<th>Column length / (cm)</th>
<th>Sample volume range / (μL)</th>
<th>Sample mass load range / (mg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10-40</td>
<td>0.1-0.45</td>
</tr>
<tr>
<td>15</td>
<td>30-120</td>
<td>0.3-1.35</td>
</tr>
<tr>
<td>25</td>
<td>50-200</td>
<td>0.5-2.25</td>
</tr>
</tbody>
</table>

120
5.3.6. Data pre-processing

The data pre-processing in this chapter follows the approach of Pate et al, 1998 but the chromatograms are also adjusted to account for the minor (but significant) retention time shifts which occur due to changes in the temperature and pH variables. A similar pre-processing technique was also used by Andersson and Hamalainen (1994) to match the retention times and shapes of test peaks onto related target peaks.

5.3.6.1. Retention time transformation

An increase in either temperature or pH caused the species of the crude erythromycin to be eluted later (the peaks shifted to the right). The combination of high settings for both of these variables was seen to cause even greater shifts. Since the PCA would detect these time shifts, it was expected that to account for them before the analysis would reduce the number of principal components needed to describe the data. Accordingly the chromatograms were stretched or compressed so that the main erythromycin A peak in each was lined up to match the mean retention time (timed from point of injection) for the corresponding peak of the centre point chromatograms (1.32, 3.60 and 4.60 minutes for 5, 15 and 25 cm columns respectively). This was done by stretching or compressing the time axis for each chromatogram using a spreadsheet in Microsoft Excel, so that:

\[
\frac{RTA_{test}}{RTA_{centre}} = \frac{t_{x0}}{t_{x0}'}
\]

Where:
- \( t_{x0} \) = original time value for each of the 300 absorbance values, s
- \( t_{x0}' \) = new time value for each of the 300 absorbance values in any chromatogram, s
- \( RTA_{centre} \) = Mean Retention time for erythromycin A (centre point chromatograms), s
- \( RTA_{test} \) = Retention time for erythromycin A for test chromatogram, s

A splining software package (Curve Expert 1.1 for Windows) was then used to return each chromatogram to the original time increments, necessary before implementing PCA. Figure 5.1b shows all 108 chromatograms prior to retention time adjustment and Figure 5.1c shows all 108 chromatograms after the same adjustment.
Figure 5.1b. All 108 Chromatograms for the 25 cm column prior to retention time adjustment.

Figure 5.1c. All 108 Chromatograms for the 25 cm column after retention time adjustment.

Figure 5.2 shows how this adjustment was effected for two chromatograms at the 5 cm scale. Figure 5.2a) shows the chromatograms before retention time adjustment and Figure 5.2b) shows the same chromatograms after adjustment.
5.3.6.2. Centring of chromatograms

Before implementing the PCA on any data set, it was necessary to centre the data by subtracting the mean centre point chromatogram from every chromatogram within the data set. This locates the position of the centre point chromatograms around zero on the scores plots which allows easier comparisons to be made between chromatograms. Centring also effectively reduces some of the variance within the data set (see section 3.5).

The principal component analysis was performed using SIRIUS for Windows 1.2 (Pattern Recognitions A/S, Bergen, Norway) which uses the NIPALS algorithm [Wold (1978) and section 3.6.1] as the basis for the extraction of the principal components.

5.4. Results and Discussion

5.4.1. Analysis of the full factorial data set for the 25 cm column (with retention time adjustment)

Centring the data to the centre chromatogram (as opposed to the mean of all the chromatograms), served to reduce the total amount of possible variance. It was seen that the maximum variance using this method of centring was around 81%. The principal component analysis for the full factorial data set with retention time adjustment yielded a PC model from which the first three principal components accounted for 66.93%, 6.41% and 4.08% of the variance (77.43% total). The loadings profiles for the first two components are shown in Figure 5.3, and the scores plot for the first two components in Figure 5.4 (Figure 5.4 shows the mean scores value for each set of conditions). The loadings profile for PC3 (not shown) is similar to baseline noise on the chromatogram, and was not a useful principal component.
The scores plot shows a definite trend in the chromatograms with PC1 scores increasing as the weight of the sample applied to the column increases. The PC2 scores group the chromatograms with respect to the other variables and to particular combinations of sample concentration and volume.
Figure 5.3. The loadings profiles for the first two principal components of the complete data set (with retention time adjustment). PC1 represents the majority of the variance (66.53%) within the data set and is visually similar to a chromatogram. PC2 plays a more minor role, only accounting for 6.55% of the variance.

Figure 5.4. Scores plot for the first two PCs in the data set with RT adjustment. This plot shows a definite trend of the chromatograms with respect to the mass of material loaded onto the column, the larger loads having high scores in PC1 and the smaller loads scoring low in PC1. PC2 serves to isolate the chromatograms with respect to the other variables and to how the mass of material was loaded in terms of combination of concentration and volume. The mean scores values for each setting are shown.
5.4.2. Selection of subsets of chromatograms - the framework for scale-up predictions

For PCA to be a useful tool it must be able to predict the trends in the principal components from as small a group of chromatograms as possible. This is particularly important for scale-up where any large-scale tests which are included in the analysis will be expensive in material and resource. To discover if this was possible, subsets of chromatograms were selected and analysed separately. The extracted principal components were then used to predict the features known to be present in the full data set.

<table>
<thead>
<tr>
<th>Subset of Chromatograms</th>
</tr>
</thead>
<tbody>
<tr>
<td>centre</td>
</tr>
<tr>
<td>cc</td>
</tr>
<tr>
<td>cvTP</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>cVtp</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>CC</td>
</tr>
<tr>
<td>CVTP</td>
</tr>
</tbody>
</table>

Table 5.4. The critical subset of chromatograms from which scale-up predictions were made. 9 sets resulted from the adjusted data set (27 chromatograms) representing one for each sample load setting and a set of centre chromatograms.

In the first instance chromatograms were chosen which represented the different sample loads (Table 5.4) and which appeared in different “mass-loaded” clusters in the original analysis (Figure 5.4). (Table 5.3 shows the range of sample loads both in terms of mass and volume which were applied to each column). The selection of chromatograms was also made so as to provide a good spread of retention times, thus ensuring that the timing of the eluted peaks could be predicted as well as the chromatographic separation itself. Together with the centre chromatogram a group of nine resulted, each being performed in triplicate (27 chromatograms from the original 108).
5.4.3. PCA of subsets of chromatograms (adjusted for retention time effects).

The maximum possible variance from the PCA of the selected subset was seen to be around 85% for the same reason as was explained in section 5.4.1. The first three principal components extracted from the selected subset of chromatograms accounted for 70.73%, 9.88%, and 2.57% of the variation (83.18% total). The scores values for the remaining 81 chromatograms which had not been included were calculated from the loadings of the subset of chromatograms using an iterative technique in Microsoft Excel. This calculates a score for the principal component which when combined with the loading, minimises the difference from the actual chromatogram.

Comparing these scores with the original values derived from the model using all 108 chromatograms reveals striking similarities. This is clearly evident in Figure 5.5 which shows a plot of the PC1 scores derived from an analysis of all 108 chromatograms (y axis) against the scores derived from the analysis of the 27 chromatograms comprising its subset (x axis).

![Figure 5.5](image.png)

*Figure 5.5. A comparison between the PC1 scores generated using just 9 groups from 36 (27 out of 108 chromatograms) (x-axis) and between the actual PC1 scores generated using all 108 chromatograms. The scores show excellent agreement around line y = x, indicating that the chromatographic separation of all 108 chromatograms are adequately described by a much smaller subset.*
The correlation is an excellent fit to the line \( y = x \) (the actual best fit line has equation \( y = 1.040x - 0.025 \), coefficient of regression \( r = 0.995 \)), which indicates that all of the 108 chromatograms in the full factorial data set can adequately be described using only a fraction of the original data. Similar results were obtained for PC2 except that there is more scatter about the line \( y = x \) (actual line \( y = 1.003x - 0.040 \), \( r = 0.989 \)).

It is concluded that accurate principal component models can be obtained using only a fraction of the original number of runs. It is naturally desirable to use as few runs as possible at a large scale of operation as the basis for predicting the shapes of chromatograms. However if nine groups of chromatograms (27 in total) are representative of the original 36 groups (108 in total) what further reduction is possible without detracting from the analysis? To answer this smaller subsets comprising seven, five and just three groups (9 chromatograms) was selected (Table 5.5) for separate analysis.

<table>
<thead>
<tr>
<th>3 cluster model (9 chromatograms)</th>
<th>5 cluster model (15 chromatograms)</th>
<th>7 cluster model (21 chromatograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cvTP</td>
<td>cvTP</td>
<td>cvTP</td>
</tr>
<tr>
<td>cent</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>CVTP</td>
<td>cent</td>
<td>cent</td>
</tr>
<tr>
<td>C</td>
<td>cvTP</td>
<td>C</td>
</tr>
<tr>
<td>CVTP</td>
<td>C</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CVTP</td>
</tr>
</tbody>
</table>

Table 5.5. Table showing the clusters selected to comprise the three, five and seven cluster subsets.

The PCA of the smallest set generated PC1 correlations similar to those described in Figure 5.5. The loadings for PC1 derived from all 108 chromatograms and from the smaller subsets are very similar (Figure 5.6). This is most beneficial in minimising the number of large-scale experiments which the analysis requires. The major discrepancies appear on either side of the main peak. The fine structure becomes more clearly defined when larger numbers of chromatograms are used to derive the loading.
Figure 5.6. The PC1 loadings profiles for the 25 cm column obtained using all 36, 9 and 3 groups (108, 27 and 9 chromatograms).

As before the scores for the remaining 99 chromatograms in the original set of 108 were calculated from the loadings derived from the subset of 9. The same iterative technique in Excel was used. The PC1 scores so derived correlate well with the scores derived from the analysis of the full set of 108 indicated by the $y = x$ plot in Figure 5.7 (actual line $y = 1.031x - 0.060$, $r = 0.999$).

Figure 5.7. A comparison between the PC1 scores generated using just 3 groups from 36 (9 out of 108 chromatograms) (x-axis) and between the actual PC1 scores generated using all 108 chromatograms.
5.4.4. Analysis of chromatograms derived from columns of different lengths

5.4.4.1. Retention times

The main focus of this study was the prediction of chromatograms expected from columns at a 25 cm scale from those obtained at 5 cm. A 15 cm column was also used to provide an intermediate scale.

The chromatograms from the columns of different lengths contain valuable information regarding the retention times of the separated species obtained under various operating conditions. This data is most useful in establishing the correlations between the retention times at the different scales of operation, and it is easily obtained before the data is pre-processed (section 5.3.3). These relationships for the main erythromycin A peak in the unadjusted chromatograms at the 5 cm scale and for the peaks at both the 15 and 25 cm scales are shown in Figure 5.8. From this plot it is possible to predict the expected retention time at the larger scales from the 5 cm data. The times were correlated using least squares linear approximations.

![Figure 5.8. Retention time correlations between 5 and 15 cm columns and 5 and 25 cm columns.](image)

5.4.4.2. Principal component scores

PCA models generated for the 5 cm and 15 cm columns resulted in the following amount of explained variance.
5 cm column.     PC1 94.85%, PC2 0.87%, PC3 0.80% (total 96.46%).
15 cm column.    PC1 92.39%, PC2 2.32%, PC3 0.81% (total 95.52%).

The percentage variances explained with 3 PCs for both these columns are higher than for the 25 cm column (77.43%). This is because the centre chromatograms for the smaller columns are very similar to the mean chromatograms.

In each case the analysis is derived from a reduced set of nine groups (27 chromatograms) rather than from the full set of 36 groups used to analyse the performance of the 25 cm column. Almost all of the variance from these shorter columns is associated with PC1; in both cases the loadings of PC2 and PC3 appear as a noisy background.

The loadings for PC1 at the three scales of operation are only comparable if the retention times are adjusted to account for the different column lengths. After this adjustment the loadings for the 5 and 15 cm columns are similar in shape but they differ somewhat from the loading for the 25 cm column (Figure 5.9).

![Figure 5.9](image)

**Figure 5.9.** The PC1 loadings profiles of the nine cluster models for the 25 cm, 15 cm and 5 cm columns. To allow easier comparability, the loadings at the 5 cm and 15 cm scale have been stretched to align with the main peak of the 25 cm loadings.
A further analysis of a biased set of chromatograms from the 25cm column where the quantity of sample was small (0.50 - 1 mg) revealed a PC1 loading (Figure 5.10) whose profile, apart from a small “negative” peak at about 5min, more closely resembled those from the 5 and 15 cm columns. The profile of PC1 apparently changes in a manner which reflects the increasing quantity of sample loaded onto columns of uniform diameter.

![Figure 5.10. The PC1 loadings profile for the model using only the low sample loads.](image)

5.4.5. Relationship between principal components and chromatographic profiles

The previous section may imply that the underlying chromatographic separation is unaffected by the length of the column provided that the quantity of sample applied does not exceed 1mg. It is important to avoid the conclusion that the loadings themselves represent that underlying separation. The principal components are derived from a particular factorial matrix (see Table 5.2) in which not only the quantity of sample, but also pH and temperature were varied. All of these factors, and the pre-processing which includes the subtraction of the centre point chromatogram, affect the analysis. As one of the factors is varied independently the trends which are
visible in the principal components are not the same as those which would be visible in the chromatograms themselves.

The data pre-processing and the weighting inherent in the matrix algebra also affect the extraction of the principal components so that PC1 derived from an unbiased set of data is representative of the more heavily loaded chromatograms. This effect of the quantity of sample on PC1 makes the role of PC2 in defining the shape of the chromatograms from the 25cm column of particular importance. The PC2 loading (Figure 5.3) and its significantly negative score when low sample volumes and concentrations are applied to the column (Figure 5.4) will adjust the shape of the PC1 loading as the sample load changes. It can be concluded that the absence of PC2 from the description of the 5 and 15 cm columns is further evidence that these columns were run within a range of sample concentrations and volumes which did not affect the chromatographic performance.

5.4.6. Prediction of chromatographic profiles

If the assumption is made that the process variables affect the chromatographic separation in the same way regardless of scale, it can be deduced that correlations can be made between the shapes of chromatograms at different scales but at otherwise related conditions. When the PCs are compared, they reflect similar features of the chromatograms at each scale. For a particular set of conditions the score for PC1 at the 5cm scale is related to its score at 15 and at 25cm. The PC1 scores data in Figure 5.11a) were correlated using least squares quadratic correlations. The correlation between the 5 and 15 cm columns is noticeably more linear than that between the 5 and 25 cm columns.

Starting from the correlations in these scores plots it is possible to derive the size and shape of chromatograms at one scale of operation from those at another. For example a separation could be performed at the 5 cm scale. After pre-processing, the PC1 scores value would be calculated. The scores value for the corresponding separation at the 25 cm scale would then be predicted from the correlation curve in Figure 5.11a).
When combined with the PC1 loading at the 25 cm scale, the prediction of the chromatogram shape at 25 cm scale would be made.

Figure 5.11a). PC1 scores correlations from 5 cm to both 15 and 25 cm column length using a nine cluster model. The 5 to 15 cm correlation is more linear than the 5 to 25 cm.

Figure 5.11b). PC1 scores correlations from 5 cm to 25 cm column using a 3 cluster model
Finally the small adjustment to the retention time (predicted from the correlations in Figure 5.8) would be included to effect the final prediction (Figure 5.12).

![Chromatogram CC at 25 cm predicted from 5cm data using the three and nine cluster models.](image)

**Figure 5.12.** Chromatogram CC at 25 cm predicted from 5cm data using the three and nine cluster models.

### 5.4.6.1 Goodness of fit

To establish how well the models were able to make predictions from the 5 cm to the 25 cm scale, a goodness of fit criterion was applied to each predicted chromatogram. A sum of squares of errors (SSE) was performed on the difference between the actual chromatogram and that predicted from the various models. This value was then compared with the maximum sum of squares of errors (SSE<sub>max</sub>) which results from the actual chromatogram when compared with zero data.

\[
\text{Goodness of fit \%} = 100 \left( \frac{\text{SSE}_{\text{max}} - \text{SSE}}{\text{SSE}_{\text{max}}} \right)
\]  

(5)

The goodness of fit for each of the 9 chromatogram groups for the 25 cm column are shown in Table 5.5.

The prediction does demand that the loadings for the principal components at the 25cm scale are known in advance. However the loadings from the three cluster model
do resemble those obtained from the full data set (Figure 5.6). Although the correlation between the scores associated with PC1 at the 5 and 25 cm scales is necessarily less exact for this three cluster model (Figure 5.11b) than it is for the full set of nine clusters (Figure 5.11a), it is apparent from Table 5.6 and Figure 5.11 that the 3 cluster model does predict the chromatogram with reasonable accuracy. This is particularly true for sample loads greater than 1 mg.

<table>
<thead>
<tr>
<th>Mass loaded (mg)</th>
<th>% fit after PC1(9) (3 cluster model)</th>
<th>% fit after PC1(15) (5 cluster model)</th>
<th>% fit after PC1(21) (7 cluster model)</th>
<th>% fit after PC1(27) (9 cluster model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cvTP</td>
<td>0.50</td>
<td>36.5</td>
<td>32.2</td>
<td>46.0</td>
</tr>
<tr>
<td>cc</td>
<td>0.5625</td>
<td>46.9</td>
<td>41.6</td>
<td>60.3</td>
</tr>
<tr>
<td>c</td>
<td>0.75</td>
<td>77.3</td>
<td>78.0</td>
<td>79.5</td>
</tr>
<tr>
<td>p</td>
<td>1.00</td>
<td>89.7</td>
<td>89.2</td>
<td>89.8</td>
</tr>
<tr>
<td>cent</td>
<td>1.00</td>
<td>99.9</td>
<td>97.0</td>
<td>97.6</td>
</tr>
<tr>
<td>cVTP</td>
<td>1.125</td>
<td>95.7</td>
<td>95.4</td>
<td>94.9</td>
</tr>
<tr>
<td>CC</td>
<td>1.50</td>
<td>92.7</td>
<td>96.9</td>
<td>95.4</td>
</tr>
<tr>
<td>CVTP</td>
<td>2.00</td>
<td>92.70</td>
<td>93.2</td>
<td>97.0</td>
</tr>
<tr>
<td>CVTP</td>
<td>2.25</td>
<td>99.6</td>
<td>99.6</td>
<td>97.8</td>
</tr>
</tbody>
</table>

Table 5.6. The goodness of fit from 3, 5, 7 and all 9 cluster models for chromatograms at 25 cm scale predicted using a IPC model prediction from 5 cm data.

The model is much less capable of predicting the chromatograms with a low sample load (0.75 mg and below). The reason for this lies in the fact that only one principal component is being used to make these predictions. Whereas PC2 has been seen to make only a small contribution to the overall models in these experiments, its contribution to the chromatograms with a small load at the 25 cm scale appears to be very significant. When the mass of sample applied to the column is large PC1 dominates the analysis of the chromatogram. Its score ranges between 2 and 3, while that for PC2 lies between -0.5 and -0.8. However at the low loads the PC1 score is much smaller, lying between -1.0 and -0.5 and is in the same range as the PC2 score (Figure 5.4).

From the data presented here it is impossible to predict the PC2 score required to ensure a good fit for the chromatograms obtained with a low sample load on the 25 cm column. This is because at the 5 cm and the 15 cm scales PC1 was sufficient to explain almost all of the variance found in the chromatograms over the range of separation
conditions explored. No second component was generated from these shorter columns on which to base the necessary correlation with the 25cm data. Moreover at the 25cm scale the scores for PC2 are not obviously related in any direct fashion either to PC1 or to the sample load (see Figure 5.4). The belief is that this problem can be overcome by investigating the chromatographic effect of a wider range of sample loads at each size of the column.

5.5. Conclusions

In this chapter principal component analysis (PCA) has been used to predict chromatogram shapes at different scales of operation. The predictions were made without the need for mathematical models of the adsorption process which can be inaccurate. With effective pre-processing of the basic chromatograms, especially their scaling to align the main product peak, a single principal component was sufficient to explain most of the variance within the data sets obtained at each scale.

Correlations between the PC1 scores at the different scales of operation were obtained from subsets of chromatograms in the full factorial data sets. These subsets also provided loadings for PC1 which approximate to those obtained for the full data sets; for a comparison of chromatograms at the 5 and 25cm scales, three clusters of data (9 chromatograms) provided an adequate representation of the full data set of 36 clusters (108 chromatograms). From the predicted PC1 scores at the different scale, the expected shape of the corresponding chromatogram at this scale could be calculated from the loading of the relevant principal component. Then from the correlation of the retention times between the scales, it was possible to alter the length of the chromatogram to obtain the completed prediction. A flow diagram of the basic steps of the analysis is given in Figure 5.13.

For chromatograms at a 25cm scale with a high sample load, accurate predictions were observed using a single principal component model. However a single principal component model could not predict chromatograms with a low sample load at this scale. The full description of the variance found in the chromatograms at 25cm requires two principal components, and the influence of the second component is
particularly strong where the sample load is small. This contrasts with the data for the 5 & 15 cm columns where the full range of effects of temperature, pH, flow rate, sample volume and load require only one component for their full description.

In these experiments the range of the variables studied at these smaller scales was insufficient to encompass fully the variation required to make the necessary prediction at the larger scale. In particular the sample loads at the small scale were insufficient to define the second principal component from which a broader range of predictions could be made.

Subsequent results chapters involve increasing the sample loads to establish the range over which the models are valid. Such models require different techniques of pre-processing and have been shown (Chapters 6 and 7) to yield a greater number of principal components. These techniques will be extended to discover whether it is possible to predict the shape of chromatograms based on packing materials with variations in porosity and particle size, and over a much wider range of column geometries (specifically using progressively wider columns), culminating with predictions at an industrial preparative scale of operation. To date the analysis is incomplete, and it lacks any theoretical structure, but the belief is that it does offer a practical alternative to approaches based on an analysis of isotherms.
Figure 5.13. A flow-sheet showing the main stages in the use of PCA for the prediction of chromatograms from 5 cm to 25 cm scale. The shaded sections are the stages involved with reducing the amount of data required at the 25 cm scale.
6. The use of Principal Component Analysis for the modelling and scale-up of non-linear chromatography.

6.1. Summary
6.2. Introduction
6.3. Materials and methods
6.3.1. The test HPLC system
6.3.2. Columns
6.2.3.1. HPLC Column Packing
6.4. Results and Discussion
6.4.1. An overview of the chromatograms from the three columns.
6.4.2. Extraction of a PCA model for the CG71 methacrylate column
6.4.2.1. Subtraction of the mean chromatogram
6.4.2.2. Extraction of PC1
6.4.2.3. Extraction of PC2 and PC3
6.4.2.4. Correlations between the amount of sample chromatographed and the values in the scores matrices
6.4.3. Reconstruction of CG71 chromatograms
6.4.4. Extraction of PC models for the CG300 and the PLRP1000-S polystyrene columns
6.4.5. Alternative models
6.4.5.1. Model based on scaling of a mid-range chromatogram
6.4.5.2. The use of differentials of the mean chromatogram
6.4.6 Comparison of Principal Components extracted from different columns
6.4.6.1. Loadings
6.4.6.2. Scores
6.4.7. Correlations between Principal Components from different chromatograms.
6.5. Conclusions
6. The use of Principal Component Analysis for the modelling and scale-up of non-linear chromatography.

6.1. Summary

The use of Principal Component Analysis (PCA) has been used in this chapter for the modelling of non-linear chromatography in overload conditions. Three different columns formed the basis of the study with dimensions ranging from 5cm×4.6mm to 15cm×7.5mm and stationary phases ranging from 8µm polystyrene to 35µm methacrylate. A tenfold range of crude erythromycin samples were loaded onto each column in direct proportion to their bed volumes. The same isocratic separation of 45/55 v/v acetonitrile/water was used in each case. Principal component models were generated for the data which were able to model excellently complete chromatograms in the absence of mathematical models and physico-chemical data. Correlations between sample size and the principal component scores were made for each which were remarkably consistent for the three column types despite the different column geometries and stationary phases. The belief is that such correlations offer considerable potential for modelling of non-linear chromatography as well for scale-up predictions.

6.2. Introduction

The use of Principal Component Analysis (PCA) was successfully used in chapter 5 for the prediction of chromatography profiles for the separation of erythromycin. The study enabled accurate predictions to be made from a 5cm×4.6mm to a 25cm×4.6mm column with identical stationary phases. It is the aim of this chapter to extend the techniques developed in chapter 5 for the modelling and subsequent scale-up of chromatograms from columns of different diameters and stationary phases.

In chapter 5 chromatograms with high sample loads were modelled accurately using a single principal component model. However a single principal component model could not predict chromatograms with a low sample load at this scale. The full description of the variance found in the chromatograms at 25cm scale required at least two principal components, and the influence of the second component was
particularly strong where the sample load was small. This contrasted with the data for
the 5cm column where the full range of effects of temperature, pH, flow rate, sample
volume and load required only one component for their full description.

In these experiments the range of the variables studied at these smaller scales was
insufficient to encompass fully the variation required to make the necessary
prediction at the larger scale. In particular the sample loads at the small scale were
insufficient to define the second principal component from which a broader range of
predictions could be made. It is the aim of the work presented in this chapter is to
develop principal component models from chromatograms resulting from a much
wider range of sample loads (encompassing true overload conditions) applied to
columns with different geometries and packing materials. From these models,
relationships will be established in an attempt to ascertain whether the principal
components are able to reflect similar chromatogram features regardless of column
geometry and packing material used. It was decided to maintain constant pH and
temperature (as is the case with adsorption isotherms) throughout this study since
these variables may have served to complicate the analysis in chapter 5. This may
have prevented relationships being developed for the more minor principal
components.

Unfortunately the term “loading” applied to the Principal components has a verbal association with the chromatographic load.
For this reason this Chapter refers to the “amount of sample chromatographed” to describe the latter, and the term “loading” is
used only as it applies to the Principal Component Analysis. In addition, the term “component” may be ambiguous with
components or molecular species in the chromatographic mixture. For this reason the components in the mixture will be
referred to as “biomolecules” or “species”.

6.3. Materials and methods

6.3.1. The test HPLC system

The same isocratic reversed-phase separation of erythromycin used in chapter 5 was
also used in this chapter. The same Beckman HPLC System was used in conjunction
with the same Beckman System Gold software as described in Chapter 5. The
separation method is as follows:
Mobile Phase: The identical 45/55 v/v Acetonitrile/water solvent described in Chapter 5. The pH was constant at 7.00 throughout.

Flow rate: Column 1. 1.2 mL/min; Columns 2 and 3 1 mL/min.

Sample volume range: 0.9-9% of bed volume for each column.

Sample concentration: 20 mg/mL.

Temperature: 64 °C ± 0.1 °C.

Run times: Column 1. 6 minutes; column 2. 28 minutes; column 3. 80 minutes.

UV Wavelength: 215 nm.

6.3.2. Columns:

1. PLRP1000-S [poly(styrene-divinylbenzene)], 8 μm particle diameter, 1000 Å pore size, dimensions 5cm × 4.6mm. (Polymer Laboratories, Church Stretton, Shropshire, England).

2. CG 300 [poly(styrene-divinylbenzene)], 35 μm particle diameter, dimensions 15 cm × 4.6mm. Column packed with Shandon HPLC column packer (Runcorn, Cheshire, UK).

3. CG 71, (methacrylate ester), 35 μm particle diameter, dimensions 15 cm × 7.5 mm.

The loose CG300 chromatographic support, and the CG71 column were kindly supplied by ToSOHaas GmbH (Stuttgart, Germany).

6.2.3.1. HPLC Column Packing

Column 2 (and, subsequently other columns) was packed at UCL using a Shandon HPLC Column packer (Shandon Southern Products Ltd., Astmoor, Runcorn, Cheshire).

Description

The Shandon column packer is designed for the rapid and efficient packing of stainless steel HPLC columns from analytical to semi-preparative scale. The unit
consists of a piston pressure intensifier pump, a stainless steel column reservoir and 4 solvent vessels.

The flow for the packing procedure is triggered using two pistons. A large piston draws nitrogen gas (supplied via nylon tubing) and pushes on a smaller piston which draws the packing solvent from a reservoir. There is a relationship between the gas pressure and the solvent pressure. For this particular column packer, the relationship is:

\[ \text{Solvent pressure} = 60 \times \text{Gas Pressure} \]  \hspace{1cm} (1)

Therefore the gas pressure is set for the desired packing solvent pressure. The maximum pressure allowed is 10 bar gas pressure (for a maximum solvent pressure of 600 bar).

**Operation**

Before the packing procedure commences, the whole system is cleaned with methanol. The empty column is then attached to the column reservoir. The packing material is made into a homogeneous slurry before pouring into the column reservoir. The volume of the slurry is usually about 1.5 times the volume of the empty column to make sure the column is packed efficiently i.e. with no voids.

The pressure of nitrogen gas is then applied in compliance with the desired solvent packing pressure. For the CG300 packing material, the manufacturers recommended a solvent pressure of 60 bar which meant the gas pressure was 1 bar.

Once the gas pressure has been applied, the packing procedure commences with loud pump pulsations which gradually slow down before finally stopping altogether. The pulsations only last a matter of a few seconds and when they stop the packing is complete. For safety reasons, it is necessary to turn all switches off. The column is then disconnected and the end fittings screwed in. The surplus packing material is then pushed out with water and the reservoir is cleaned.
The packed columns

It is good practice to run mobile phase at low flow rates through the newly-packed columns for several hours after packing. Good quality columns (tested initially with acetone and subsequently with the erythromycin separations) were observed with column 2 and for other column packed in the course of this thesis.

As was the case in Chapter 5, the UV profiles used to provide the data for the PCA comprised 300 absorbance values for each run. The masses of material loaded onto each column are given in Table 1. Each run was performed in duplicate and the order of the runs randomised to guard against systematic bias.

The principal component analysis was performed using SIRIUS for Windows 1.2 (Pattern Recognitions A/S, Bergen, Norway) which uses the NIPALS algorithm (Wold, 1978) as the basis for the extraction of the principal components.

<table>
<thead>
<tr>
<th>Mass loaded onto COLUMN 1 5cm×4.6mm PLRP1000 (8μm)/mg</th>
<th>Mass loaded onto COLUMN 2 15cm×4.6mm CG 300 (35μm)/mg</th>
<th>Mass loaded onto COLUMN 3 15cm×7.5mm semiprep CG71 (35μm)/mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.16</td>
<td>0.46</td>
<td>1.2</td>
</tr>
<tr>
<td>0.32</td>
<td>0.92</td>
<td>2.4</td>
</tr>
<tr>
<td>0.64</td>
<td>1.84</td>
<td>4.8</td>
</tr>
<tr>
<td>0.8</td>
<td>2.26</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>3.38</td>
<td>9</td>
</tr>
<tr>
<td>1.34</td>
<td>3.74</td>
<td>10</td>
</tr>
<tr>
<td>1.6</td>
<td>4.54</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 6.1. Table of mass loaded onto each column type. Each mass loaded was proportional to the bed volume of each column. All samples had concentration of 20 mg/mL.

6.4. Results and Discussion

6.4.1. An overview of the chromatograms from the three columns.
The effect of increasing the amount of sample chromatographed on the separation of the biomolecules in the erythromycin mixture depends on the support (Fig. 6.1a-c). The two polystyrenes (columns 1 and 2) are roughly similar to one another in their performance. If the first small identifiable peaks which elute at 0.50 min from the PLRP-S (Fig. 6.1a) and at 2.0 min from the CG300 (Fig. 6.1b) are taken as unretained peaks then the total porosity of the columns are 0.72 and 0.8 respectively. As the amount of sample increases the main product peak of erythromycin A itself elutes progressively later from both columns. The effect is more pronounced on the PLRP-S
whose capacity factor for this peak increases from 1.5 to 2.3 as the peak trails back from a fixed leading edge (Fig 6.1a). The increase for the CG300 is smaller, from 2.1 to 2.5, and the peak shape is less asymmetric (Fig 6.1b). The data imply a slightly concave isotherm for the adsorption of the main peak onto both polystyrenes over the range studied (cf. Siedel-Morgenstem and Guichon, 1993 and Guiochon et al, 1994).

In contrast the methacrylate CG71 (Fig. 6.1c), for which we estimate a rather low porosity of 0.61, shows the effects expected of overload on a convex Langmuir isotherm (see Golshan-Shirazi and Guiochon, 1988) of the main peak. The main peak elutes earlier as the amount of the sample increases. Over this range the capacity factor for the peak maxima decreases from 5.9 to 3.5 and the back edge is common for each sample. Peaks next to the main product peak which exhibit baseline separation when the amount of sample is small, become shoulders to the main peak as it increases in size so that they blend together and co-elute. The curious effect of a sharp early peak when the amount of sample is greatest is a reproducible feature of the chromatography. It is suspected that this is a feature of displacement chromatography (Guiochon et al, 1994), but we have not yet identified the species in the peak.

![Figure 6.1a. All chromatograms for column 1 (5cm×4.6mm PLRP, 8 μm). The overloading of this column results in a shifting to the right of the main product peak, indicating a slight concave isotherm.](image-url)
Figure 6.1b. All chromatograms for column 2 (15cm×4.6mm CG300, 35 μm). At low loads there appears to be a slight shifting to the left of the main peak as the sample load increases. At the higher loads the main peak shifts to the right with increasing sample load, indicating a slight concave isotherm.

Figure 6.1c. All chromatograms for column 3 (15cm×7.5mm CG71, 35 μm). The main product peak undergoes large retention time shifts to the left as mass load increases. This indicates a distinct convex isotherm of the main peak.
6.4.2. Extraction of a PCA model for the CG71 methacrylate column

Models with three principal components were developed for each column. This is exemplified in detail for the chromatography on the CG71 support, whose behaviour is the most complex of the three.

6.4.2.1. Subtraction of the mean chromatogram

The analysis of each set of data first calculates a mean chromatogram (Fig. 6.2a) which is subtracted from each of the chromatographic traces (Fig. 6.1c). This eliminates the most obvious relationship between all of the samples so that the analysis itself detects the more subtle changes. The analysis then proceeds as described in section 3.6.1 with each of the residual traces (Fig. 6.2b) being considered as a row of the matrix whose value is to be minimised with the progressive extraction of the principal components.

6.4.2.2. Extraction of PC1

The analysis next extracts a single “loadings” matrix (the row matrix in Fig. 2.1), and an associated “scores” matrix (the column matrix in Fig. 2.1). Together these minimise the value of the residual matrix.

![Figure 6.2a. The mean chromatogram of the CG71 column data (column 3).](image)
Although the "loadings" matrix has no physical significance it does have the superficial attributes of a chromatographic trace (Fig. 6.3a). Each value in the associates "scores" matrix is a scaling factor which is applied to the "loadings" matrix before the result is subtracted from the appropriate matrix row, in effect the "chromatographic trace" generated by the removal of the mean chromatogram. The result (Fig. 6.3b) is a new set of residuals whose total value the principal component analysis had minimised. Together the "loadings" matrix and the "scores" matrix make up the first principal component (PC1). It describes nearly 84% of total the variance of the mean-centred data (Fig. 6.2b).

6.4.2.3. Extraction of PC2 and PC3
The analysis then extracts the loadings and scores matrices of a second principal component (PC2) from the matrix which describes the residuals in Fig. 6.3b. Again the resulting loadings matrix and the residuals can be represented as if they were chromatographic traces (Fig. 6.4a&b). In this instance the combined removal of PC1 and PC2 describes over 95% of the variance left after the removal of the mean.
Figure 6.3a. The PC1 loadings profile for the CG71 column (column 3).

Figure 6.3b. The residuals after extraction of PC1 for the CG71 column (column 3).
Figure 6.4a. The PC2 loadings profile for the CG71 column (column 3).

Figure 6.4b. The residuals after extraction of PC2 for the CG71 column (column 3).
The analysis continued to extract a third principal component (PC3) after which less than 2% of the variance remained unexplained (Fig. 6.5a & b).

Figure 6.5a. The PC3 loadings profile for the CG71 column (column 3).

Figure 6.5b. The residuals after extraction of PC3 for the CG71 column (column 3).
6.4.2.4. Correlations between the amount of sample chromatographed and the values in the scores matrices

As might be predicted, the values in the scores matrix of each principal component are strongly correlated with the amount of sample used to generate each of the chromatographic traces which form the rows of the original matrix (Fig. 6.6). The correlations are surprisingly well ordered with

- PC1 scores showing a linear correlation (1st order polynomial),
- PC2 scores, a quadratic correlation (2nd order polynomial), and
- PC3 scores, a cubic correlation (3rd order polynomial).

The form of the correlations was computed using Microcal Origin 4.1. (Microcal Software, Northampton, MA, USA).

![Figure 6.6](attachment:figure66.png)

**Figure 6.6.** Scores against mass loaded for the CG71 column (column3). The relationship for PC1 is linear (correlation coeff $r = 0.99$), PC2 is quadratic ($r = 0.95$) and PC3 is cubic ($r = 0.80$)
6.4.3. Reconstruction of CG71 chromatograms

The chromatograms predicted by the PC models are derived using an analytical method in reverse. The product of the loadings matrix and the score value appropriate to each sample is calculated for each PC (Fig. 6.7). The results are all added to the mean chromatogram, in this instance to reconstruct the chromatogram predicted for the 9mg sample.

Figure 6.7. A diagrammatic representation showing the reconstruction of a chromatogram using PCA. The contribution for each PC is the product of the scores and loadings. The sum of these products are added to the mean chromatogram to complete the prediction. This example shows the reconstruction of a chromatogram with 9mg of sample applied to the CG71 column.

To establish how well the models were able to describe the actual data, the goodness of fit criterion (section 5.4.6) was applied to each predicted chromatogram.

Figure 6.8 shows the reconstruction of all the chromatograms from column 3 (CG71). They look very similar to the original data, and the analysis accurately models the timing of the small peak which appears as a shoulder, or at high loads as an inflection,
on the leading edge of the main peak (Fig 6.9). The major discrepancies appear in the prediction of the sharp spike ahead of the main peak at the highest load, and at the lowest sample load (1.2mg), where the residual variance after fitting the three PCs is over 16% (Table 6.2).

**Figure 6.8.** All chromatograms from the CG71 column (column 3) reconstructed using a 3PC model. The chromatograms are visually very similar to the original set (Fig 2c).

**Figure 6.9.** The retention times of the shoulder of the main peak for the CG71 column (column 3). The actual retention times are compared with those predicted from a 3 PC model.
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 6

<table>
<thead>
<tr>
<th>% fit after pc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2mg</td>
</tr>
<tr>
<td>2.4mg</td>
</tr>
<tr>
<td>4.8mg</td>
</tr>
<tr>
<td>6mg</td>
</tr>
<tr>
<td>9mg</td>
</tr>
<tr>
<td>10mg</td>
</tr>
<tr>
<td>12mg</td>
</tr>
</tbody>
</table>

Table 6.2. Mean goodness of fits for how a 3 PC model describes the column 3 (CG71) data. The amount of total variance unexplained for a 3PC model of the CG71 data is only 1%.

The spike in the single chromatogram of the highest load is an anomaly which is correctly represented, however it also predicts, quite reasonably that it should appear at a smaller scale at some intermediate loads. The fact that it is anomalous is clear from its appearance as a negative peak at the lowest loads (Fig. 6.8). This is a feature of the chromatography on the CG71 column which is unexplained.

The raw data for the 1.2mg sample from the CG71 column (column 3, Fig. 6.1c) and the residuals which remain as the analysis proceeds (Figs. 6.3b, 6.4b, 6.5b) are indicated with the dotted lines. Despite its small initial size, a disproportionate amount of the 1.2mg chromatograms still remains, even after the extraction of PC3. The first three PCs, despite being able to model the majority (99%) of all the data, do not combine to give a complete description of the 1.2mg chromatogram (Fig. 6.10). The minor PCs (PC4 onwards) evidently carry some important information which are required to model these low loads successfully. The 1.2mg chromatogram is probably within the limits of linear chromatography. For this reason, the necessity to model this sample size accurately is of lesser importance than for true overload conditions.
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 6

Figure 6.10. Figure showing actual with predicted chromatograms for the lowest sample load (with poorest fit) from column 3. The goodness of fit is 83.1%. The residual here (containing all PCs from PC-4 onwards) is significant.

Figure 6.11a. The loadings profiles for PLRP1000 column (column 1).
6.4.4. Extraction of PC models for the CG300 and the PLRP1000-S polystyrene columns

The chromatograms derived from the polystyrene columns were also subjected to PCA. The PC loadings derived from both columns (Fig. 6.11 a&b) are quite similar to one another, both in shape and scale. The scores values are once again strongly correlated with the amount of sample used to generate the chromatograms. Surprisingly the correlations for both columns are ordered in the same way as before, with the PC1 scores showing a linear, the PC2 scores a quadratic and the PC3 scores, a cubic correlation (Fig. 6.12a&b).

The reconstructed chromatograms are very similar to the originals. The fit leaves very small residuals unexplained even at the lowest loads (Fig. 6.13).
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 6

Figure 6.12a. Scores against mass loaded for PLRP1000 column (column 1).

Figure 6.12b. Scores against mass loaded for CG300 column (column 2). Both of these plots are strikingly similar to the scores observed for the CG71 column (Fig 6.5).
A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

Chapter 6

Figure 6.13. The reconstruction of the 0.46mg chromatogram compared with the actual chromatogram for column 2.

6.4.5. Alternative models

6.4.5.1. Model based on scaling of a mid-range chromatogram

Other simpler models derived directly from the chromatograms are less effective at modelling the effect of the increasing load. Even where the peak shapes are fairly symmetrical a simple scaling of a mid-range chromatogram (eg from Fig. 6.1b) will obviously fail to model any changes in retention times, capacity factors and peak shapes which the PCA is able to do. (Fig 6.14). In this case the same shaped chromatogram is being scaled. This is analogous to a single component PCA model being used explaining 100% of the data.

6.4.5.2. The use of differentials of the mean chromatogram

Another model derived from a set of first and second order differentials of a mid-range chromatogram, which bear a superficial resemblance to the shape of the PC loadings themselves is equally ineffective. The PCA is much more effective in extracting the underlying shapes and their changes as the sample loads increases. Figure 6.15 shows the differentials of a mid-range chromatogram from the CG71 set
Figure 6.14. The set of chromatograms for column 2 each predicted from a simple scaling of a mid-range chromatogram. This method is analogous to a single PC model being used which explains 100% of the data. This prediction method will clearly fail to model any change in retention times, capacity factors and peak shapes which the PCA is able to do.

Figure 6.15. The differentials of a mid-range chromatogram from the CG71 set of chromatograms (Fig 6.2c). This technique of modelling the separation was also ineffective.
6.4.6. Comparison of Principal Components extracted from different columns

6.4.6.1. Loadings

It is hardly surprising, given the difference in the shape of the original chromatograms, that the shapes of the PC loadings derived from the CG71 methacrylic ester column (Figs. 6.3a, 6.4a, 6.5a) are very different from those derived from the polystyrene columns. Their absolute size in comparison to the original chromatograms is also much larger.

Comparisons of the ratios of the area of the PC1 loading to the area of the mean chromatogram were made for each of the three sets. The PLRP-S column gave a ratio of 0.286, the CG300 a ratio of 0.391 and the CG71 a ratio of 0.613. The differences in these figures are indicative of the degree of non-linearity experienced moving from low to high loads. It is clear that the CG71 column with the highest of these ratios exhibits the most non-linear behaviour of the three columns (see Fig. 6.1c).

The PC loadings possess no direct physical significance. The belief is, however that they strongly reflect the features of the chromatograms which change with increasing load. PC1 is dominated by the size of the sample load applied, but the more nearly linear the separation, the greater is the percentage variance explained by PC1 (Table 6.3).

<table>
<thead>
<tr>
<th></th>
<th>% VARIANCE for COLUMN 1</th>
<th>% VARIANCE for COLUMN 2</th>
<th>% VARIANCE for COLUMN 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>84.1</td>
<td>90.8</td>
<td>83.9</td>
</tr>
<tr>
<td>PC2</td>
<td>11.6</td>
<td>7.8</td>
<td>11.5</td>
</tr>
<tr>
<td>PC3</td>
<td>3.3</td>
<td>1.1</td>
<td>2.9</td>
</tr>
<tr>
<td>total</td>
<td>99.0</td>
<td>99.7</td>
<td>98.3</td>
</tr>
</tbody>
</table>

Table 6.3. The percentage variance associated for the first 3 PCs for each column.

PC2 appears to reflect the changing retention times of the peaks with increasing load. For the chromatography on the CG71 column (Fig 6.1c) the isotherm is convex. The major peak shifts to the left with increasing load and the maxima of the PC2 loading lies to the right of the PC1 maxima (cf. Figs. 6.3a & 6.4a). In contrast, for the PLRP-S and CG300 (columns 1 and 2, Fig. 6.1a and 6.1b) the isotherms are concave and the main peaks shift to the right with increasing load. Here the maxima of the PC2
loadings are to the left of the PC1 maxima (Fig. 6.11a and 6.11b). PC3 appears to be a peak broadening function.

6.4.6.2. Scores

One of the most striking features of all three plots (Figs. 6.6 and 6.12 a&b) which relate the PC scores to the amount of sample chromatographed is that when PC1 score equals zero, PC3 is also zero. This also corresponds to the amount of sample at which is the PC2 score is at a maximum. Thus at a unique amount of sample for each column

\[
\text{PC1 score} = \text{PC3 score} = \frac{d(\text{PC2 score})}{d(\text{mass})} = 0 \tag{2}
\]

The sample amounts at which this phenomenon occurs are 0.88mg, 2.45mg and 6.5mg respectively for columns 1, 2 and 3. These all represent 55% of the maximum amount chromatographed on the three columns. In each case at these amounts of sample the actual chromatograms comprise the numerical mean plus a proportion of the PC2 loading defined by its maximum score. This is a surprising result given the very different nature of the three sets of chromatograms.

6.4.7. Correlations between Principal Components from different chromatograms.

When the scores obtained for PC1 from the three columns are compared at equivalent amounts of sample chromatographed the values are correlated in linear fashion (Fig. 6.16). This is an obvious result of the fact that the PC1 scores are all themselves a linear function of the amount of sample chromatographed. What is remarkable is that there is also a linear correlation between the PC2 and the PC3 scores when they are compared in the same way (Fig. 6.17a & b). The correlation coefficients are all 0.9 or greater except for the PC3 scores derived from the CG300 column. This presumably reflects the fact that PC1 and PC2 explain over 98% of the total variance in the original CG300 chromatograms leaving its PC3 component (1.1% variance) as little more than random noise. In contrast for the chromatography on the PLRP-S1000 and the CG71 columns PC3 accounts for about 3% of the variance with about 95-96% explained by PC1 and PC2 (Table 6.3).
The linear correlation between the scores of the principal components implies that they are all related to the amount of sample chromatographed by what is fundamentally the same set of polynomials except for a linear transformation of the x (amount of sample chromatographed) and y (PC score) axes. The scale of the transformation is different for each Principal Component.

Presently these relationships have only been explored for this particular separation of erythromycins over a ten-fold-range in the amount of sample chromatographed. If the relationships were to be confirmed over a wider range the implications for the scale up of a chromatographic system are obvious. Apparently one would only need sufficient data to set the scale of the axes in order to relate the PC scores to the amounts of sample.

Figure 6.16. PC1 scores for columns 2 and 3 predicted from column 1 (PLRP1000). The correlation coefficients are 0.998 and 0.997 for columns 2 (CG-300) and 3 (CG-71) respectively.
Figure 6.17a. PC2 scores for columns 2 and 3 predicted from column 1 (PLRP1000). The correlation coefficients are 0.92 and 0.96 for columns 2 (CG-300) and 3 (CG-71) respectively.

Figure 6.17b. PC3 scores for columns 2 and 3 predicted from column 1 (PLRP1000). The correlation coefficients are 0.603 and 0.94 for columns 2 (CG-300) and 3 (CG-71) respectively.
The refinement of any model would then come from the better definition of the shape of the PC loadings which is dependent on the number of chromatograms recorded, and their spacing across the range of sample amounts chromatographed. Some extrapolation outside of that range might be possible, but it is suspected that its value would be limited by an imprecise definition of the required PC loadings.

6.5. Conclusions

Underlying the correlations which the PCA has revealed is the variety of isotherms which describe the interaction of the biomolecules in the crude sample of erythromycin with the three chromatographic supports, and the equations of mass transport which influence their separation. While this physical chemistry can model the separation of the biomolecules in general terms it is not possible to describe the behaviour of them all as they interact with one another on each of the supports. Nor is it easy to find a convenient theoretical model which is easily applied to the scale up of the chromatographic process. In this respect PCA is at an advantage because it requires no underlying assumptions about the mechanism of the chromatography, nor any data defining the isotherms or the mass transport of the individual biomolecules. We believe that the use of PCA offers considerable potential for the rapid and reliable modelling of chromatographic processes and can achieve greater accuracy than the traditional mathematical approaches.

The erythromycin sample is not an ideal material for these experiments because its properties limit the sample range over which chromatographic data can be easily collected. It has a rather low absorbance at 215nm which effectively limits the minimum convenient load. It is also rather insoluble in the 45% acetonitrile and this places an upper limit on the separations. However the results taken together range across two different chemical supports (methacrylate and polystyrene) in columns whose diameter ranges from the analytical scale (4.6mm) to the semi-preparative (7.5mm). The particle diameters range from 8 to 35µm and the amount of sample chromatographed from 0.2 to 2mg/mL bed. This is not a broad range but it is sufficient to demonstrate three very different chromatographic separations. All are well modelled by PCA.
There is too little data to speculate on the generalities of the correlations and transformations which we have found, but they do reinforce the view that PCA has an important role to play in the analysis and modelling of chromatographic separations. They also suggest to us that the PCs may represent real models of the adsorption process, albeit very different from the usual direct analysis based on isotherms and mass transport. It is not yet known how far it is possible to extrapolate the PCA models beyond the ranges from which the PC loadings are extracted, but even the chemical models must rely on internally consistent isotherms and mass transport properties. In chapter 7 it is intended to test the model over a range of column diameters up to 60 mm each packed with the same chromatographic support (CG300).
Chapter 7. The use of PCA for the Modelling of Pilot-Scale Chromatograms.

7.1. Summary

7.2. Materials and Methods.

7.2.1. Overview

7.2.2. Preparation of Stationary Phase Slurry

7.2.3. Column Packing

7.2.4. Sample Injection

7.2.5. Column Unpacking

7.2.6. Problems Associated with Commissioning of the Prochrom HPLC System

7.2.1. The Columns and their Separation Methods

7.2.2. The Sample Loads

7.3. Results and Discussion

7.3.1. An overview of the chromatograms from the two semi-preparative columns

7.3.2. An overview of the chromatograms from the Pilot (Prochrom) Scale Column

7.3.1. Extraction of PCA models for the 2 semi-preparative columns

7.3.1.1. Loadings for the 2 semi-preparative columns

7.3.1.2. Scores for the 2 semi-preparative columns

7.3.4. Extraction of a PC model for Column 7.3 (12cm×60mm pilot scale)

7.3.4.1. Loadings Profiles for the 60mm Diameter Pilot Scale Column

7.3.4.2. Scores for the 60mm Diameter Pilot Scale Column

7.3.5. Prediction of all 4 chromatograms using the 3 PC model.

7.3.5.1. The Significance of PC3

7.3.6. The Prediction of a Test Chromatogram within the range of Sample Loads for the Pilot-Scale column.

7.4. Conclusions
Chapter 7. The use of PCA for the Modelling of Pilot-Scale Chromatograms

7.1. Summary

It was seen in Chapter 6 that despite different column geometries and stationary phases, very similar principal component scores patterns were obtained as the sample size increased (see Figures 6.6 and 6.12). Three further columns formed the basis of the results detailed in this chapter. These three columns each had the same stationary phase, 35μm methacrylate (CG300) which was introduced in Chapter 6. The diameters of the columns used were 7.8mm, 10mm and 60mm respectively whilst maintaining a constant bed length of 15cm. As far as possible, a tenfold range of erythromycin samples were applied to each column in direct proportion to their bed volumes (as in Chapter 6). The same isocratic separation using 45%/55% acetonitrile/water was used. Patterns in the principal component scores were obtained which were of a very similar nature as to the ones obtained in Chapter 6. The results provide further strong evidence that the principal component analysis is a very powerful technique for the modelling and scale-up of non-linear chromatography.

7.2. Materials and Methods.

7.2.1. Overview

The preparative scale chromatography system used in this study was manufactured by Prochrom (Champigneulles, France). The system consisted of a stainless steel column (maximum length 90cm, internal diameter 6cm) mounted on a tripod skid; a solvent delivery module with a positive displacement pump; a UV detector (Beckmann Instruments inc., CA, USA), and a personal computer (100MHz with 16MB RAM, Hewlett Packard) used for data collection and control purposes. The column, skid and pump were all positioned in a flame-proofed (zone 2) process area. The detector and personal computer were both positioned in a nearby no-flameproof control room. Figure 7.1 shows a schematic for the Prochrom system.
Figure 7.1. A schematic for the preparative chromatography (Prochrom) system. The solvent area is a flame-proof area containing the column and associated modules. The control room contains the computer and detector which contain potentially spark-causing electronic components.

The skid held the column in place and contained the facilities for packing the column. The column is packed vertically-upwards using dynamic axial compression, whereby a piston is driven by a hydraulic jack mounted within the skid. The hydraulic jack (Haskel, CA, USA) operated the piston during the packing procedure and maintained the pressure on the piston during column operation. This was to help ensure homogeneity of the packed bed. A flow distributor was also used at the column inlet to create as much plug-flow as possible through the packed bed.

The solvent module contained a diaphragm positive-displacement pump (LEWA, Leonberg, Germany) which was designed to deliver a constant flow rate of the mobile phase. The manufacturers quote an expected flow rate fluctuation of ±1%. Greater fluctuations than this were seen in practice, up to ±10%. The main reason for this appeared to be the fact that the pump had been inactive for about 8 years prior to use which may have caused fouling of the hydraulic oil (see section 7.2.6). The solvent module also contained a selection of valves which were used e.g. for small or large
sample selection and for the collection of eluent fractions. The system was generally operated manually although software was written in LabView (National Instruments, Austin, Texas, USA) to control the switching of solenoid valves (see Figure 7.1) for gradient purposes and also for data collection of the UV absorbance profiles.

7.2.2. Preparation of Stationary Phase Slurry

A one litre of an Amberchrom® resin (CG300) was kindly supplied by TosoHaas (Stuttgart, Germany). Section 6.2.3.1 outlined the packing procedure for the analytical and semi-preparative columns. The Amberchrom® resins supplied by TosoHaas belong to a group of macroporous synthetic polymeric packing materials. The CG300 packing used throughout this study was a polystyrene and had a mean particle diameter of 35μm. CG300 is widely used for the separation of molecules such as peptides, proteins, nucleic acids, antibiotics and vitamins.

The CG300 resin was supplied in a 20% solution of ethanol in water. Prior to packing, it was important that the packing material was left to settle for about twenty four hours. The resulting bed height was then adjusted to about 80% of the total height (packing plus solvent) prior to packing into the Prochrom column.

7.2.3. Column Packing

The piston of the Prochrom unit was raised into the calculated position. For a column of 15cm length\(^1\), it was decided to raise the piston to a height of 20cm as it was deemed that the column would compress about 25% when packed (in compliance with manufacturer’s instructions). The slurry was well-mixed prior to transferral into the column in order to ensure as much bed homogeneity as possible. The column was then secured at the top using a clamp and the piston pressure applied.

It was very important that the correct piston pressure was applied to the bed during packing. The CG300 matrix is polystyrene and the manufactures recommend that a pressure of no more than 15 bars should be applied. The required piston pressure was applied which was accompanied by a cycling noise which continues throughout

---

\(^1\) After compression, the bed attained a constant height of 12 cm.
packing. As the packing procedure reached completion, the cycling noise died away and when no additional piston noises were heard, it was decided that the column had been packed. The column was tested with several acetone samples. A column was deemed well-packed when a Gaussian (symmetrical) trace was achieved.

7.2.4. Sample Injection
Prior to sample injection, the column was equilibrated in the same fashion as the smaller columns i.e. by passing the mobile phase through for about 15 column volumes. The samples were injected through sample loops of varying size via a six-way valve (London Valves and Fitting Company, W. Sussex, UK). After passing through the column, the effluent was split using a non-return valve, sending about 95% to waste (or fraction collection) and about 5% to the UV detector in the control room.

7.2.5. Column Unpacking
Column unpacking was performed several times during the commissioning of the Prochrom unit. It was a very simple procedure whereby the clamp and end fitting were removed from the top of the column and then piston pressure applied. The packing material was removed as a single plug, ready for re-hydration and subsequent use. The inside of the column could then be cleaned in preparation for future use.

7.2.6. Problems Associated with Commissioning of the Prochrom HPLC System
There were several problems encountered during commissioning and early use of the Prochrom HPLC system. These problems are summarised below and should be noted by future users of the system.

- The Prochrom system was inactive for 8 years before its initial use for this research. Obviously, mechanical problems and inaccuracies were inevitable following such a period of inactivity.
- The diaphragm pumping system did not always give a reproducible flow rate within the manufacturer's specification of ±1%. This value seemed to fluctuate daily giving a more realistic value of ±10%. The manufacturers also
recommended that the hydraulic oil be changed after a maximum of 2 years. Changes to the oil did result in more reliable flow rates. It was deemed important to check regularly the flow rates during runs.

- There were occasions where severe pressure drop fluctuations were observed across the packed bed. This was indicative of the presence of channels within the packed bed. When this situation was observed, it was decided that the column needed to be repacked.
- The CG300 packing material was very susceptible to swelling when changes in the mobile phase compositions were made. For this reason, the cleaning procedure had to very carefully monitored. During cleaning, 85%/15% v/v of acetonitrile to water was required. Under normal isocratic conditions (for the erythromycin separation) the ratio used was 45%/55% and so very gradual gradients were required to increase the concentration of acetonitrile from 45% to 85% and then back to 45% ready for column equilibration.
- It became evident during the early trials that the chromatograms were severely drifting. This is usually indicative of using a mobile phase of poor quality. For this reason, it was decided to regularly de-gas all mobile phases both prior to and during separations. Furthermore, sealing the bottles containing the mobile phase from the environment appeared to reduce the amount of baseline drift observed in the chromatograms.

7.2.7. The Columns and their Separation Methods

All of the three columns used for the experiments in this chapter were packed with the CG300 polystyrene matrix which was described in section 6.3.2. The dimensions of the three columns were as follows:

- **Column 7.1.** 15cm x 7.8mm,
- **Column 7.2.** 15cm x 10mm,
- **Column 7.3.** 12cm x 60mm.

It was originally decided keep the bed length constant at 15cm. It was difficult to pack to an exact height on the largest of the 3 columns, the 60mm diameter Prochrom
column. This was because as the piston moved up into the column during packing, the distance moved was difficult to assess. The column also continued to undergo further compression during packing until it attained its constant height of 12 cm.

The same isocratic separation method using 45% acetonitrile (described in section 6.3.1), was used for each of the three columns. The only exception was the temperature at which the Prochrom column operated. This column operated at room temperature (20-23°C) as opposed to 64°C for the rest of the columns. This change in temperature at pilot scale would naturally have a thermodynamic effect which would be reflected in the k' value. The flow rates and run times for the three columns were as follows:

- **Column 7.1.** 1.00 mL/min, run time 72 minutes
- **Column 7.2.** 1.64 mL/min, run time 72 minutes
- **Column 7.3.** 70 mL/min, run time 45 minutes.

### 7.2.8. The Sample Loads

As the size of columns were increased, there were slight constraints placed upon the range of sample loads applied to the columns. This was because of the limitations in the range of sample loop sizes available at the semi-preparative scale (7.8 mm and 10 mm diameter columns). At the pilot (Prochrom) scale (60 mm diameter) there were constraints placed on both the sample loop sizes available and the number of runs which could be done economically. For the 2 semi-preparative scale columns, 6 sample sizes were applied to each column in proportion to the bed volumes. These were performed in duplicate, resulting in 12 chromatograms each. At the pilot scale (60 mm diameter) only 4 runs were performed (without duplication) with increasing sample loads. A fifth run was also performed to test the predictive ability of the PCA approach. The sample loads applied to the three columns are given in Table 7.1. Once again, as in Chapter 6, all erythromycin samples had a concentration of 20 mg/mL.
Table 7.1. Table of mass loaded onto each column type. All samples had concentration of 20 mg/mL.

The range of sample volumes applied to each column as a percentage of the bed volumes are shown in Table 7.2.

Table 7.2. The bed volumes for the 3 columns together with the range of sample volumes applied to each column

7.3. Results and Discussion

7.3.1. An overview of the chromatograms from the two semi-preparative columns.
The two sets of chromatograms from the semi-preparative columns (7.8mm and 10mm diameter) packed with the CG300 material are given in Figure 7.2. The CG300 matrix exhibits a slight concave isotherm with respect to the main product peak (section 6.4.1). The chromatography for both of these semi-preparative columns is, not surprisingly, very similar although the highest load chromatograms from the 15cm×7.8mm column displays evidence of detector overload.
Figure 7.2a. All chromatograms for column 7.1 (15cm×7.8mm CG300, 35μm). The overloading of this column results in a shifting to the right of the main product peak, indicating a slight concave isotherm.

Figure 7.2b. All chromatograms for column 7.2 (15cm×10mm CG300, 35μm).
7.3.2. **An overview of the chromatograms from the Pilot (Prochrom) Scale Column**

The four initial runs performed on the pilot scale column are given in Figure 7.3. Because of the fluctuations in flow rate (see section 7.2.6), it was decided to convert the chromatogram data from the conventional time basis to a volumetric basis (see also Chapter 4). Although the concave isotherm typical of the CG300 matrix is evident, the chromatography at this scale of operation is clearly quite different from the other columns packed with the CG300 material. For example it appears that all of the impurities which tended to elute earlier than the main peak on the smaller columns, now elute simultaneously as a poorly separated shoulder to the left of the main peak.

It is difficult to observe from Figure 7.3 how the chromatography changes as the sample load is increased from the lowest to the highest sample load. Figure 7.4 shows the lowest sample load (95.3 mg) on an expanded scale. The separation is better resolved at the lowest sample load than at the higher sample loads. e.g. there is a distinct shoulder which follows the main peak at the smallest load which becomes more and more blended in with the main product peak as the sample load in increased.

![Figure 7.3. All 4 chromatograms for column 7.3 (12cm×60mm CG300, 35μm).](image-url)
7.3.3. Extraction of PCA models for the 2 semi-preparative columns

The first three principal components for the data from the two semi-preparative columns were extracted in the same way as described in section 6.4.2. The amount of variance associated with the 3 PCs for the two columns are as shown in Table 7.3.

| % VARIANCE for      | % VARIANCE for      |
| COLUMN 7.1          | COLUMN 2            |
| (15cm×7.8mm)        | (15cm×7.8mm)        |
| PC1                 | 92.71               | 95.53               |
| PC2                 | 5.74                | 2.98                |
| PC3                 | 1.17                | 1.22                |
| total               | 99.62               | 99.72               |

Table 7.3. The percentage variance associated for the first 3 PCs for both semi-preparative columns.

The high variances associated with PC1 for both of these columns indicate that the chromatographic separations are more linear for the two semi-preparative CG300 column than for any of the 3 columns detailed in Chapter 6 (the columns in Chapter 6 were 5cm×4.6mm 8µm PLRP polystyrene; 15cm×4.6mm 35µm polystyrene CG300 and 15cm×7.5mm 35µm methacrylate CG71).
7.3.3.1. Loadings for the semi-preparative columns

The loadings profiles for the first three PCs for Column 7.1 (15cm×7.8mm CG300) are shown in Figure 7.5. The shapes and positions of the loadings are very similar to those for the 15cm×4.6mm CG300 column (Figure 6.11b). Once again the peak of the PC2 loading is to the left of PC1 (the main product peak shifts to the right with an increase in sample load). The loadings profiles for Column 7.2 (15cm×10mm) were (not surprisingly) very similar to those from the 7.8mm diameter column and are not shown here.

7.3.3.2. Scores for the semi-preparative columns

The patterns in the scores values with sample size for the semi-preparative columns were remarkably similar to the ones obtained for the three columns in Chapter 6 (see Figures 6.6 and 6.12). The scores values against sample size plot for the 7.8mm diameter column is given in Figure 7.6. The scores plot for the 10mm diameter column (column 7.2) was very similar and is not shown.

![Figure 7.5. The loadings profiles for column 7.1 (15cm×7.8mm)](image)
Figure 7.6. Scores against mass loaded for column 7.1 (15cm×7.8mm). The relationship for PC1 is linear (correlation coeff r = 0.995), PC2 is quadratic (r = 0.900) and PC3 is cubic (r = 0.881).

Once again PC1 scores were linear, PC2 scores followed a quadratic and PC3 scores followed a cubic as the sample size increased.

7.3.4. Extraction of a PC model for Column 7.3 (12cm×60mm pilot scale)

The first 3 principal components were extracted in the usual way on the 4 chromatograms from the pilot scale data (Figure 7.3). Table 7.4 gives the amount of variance associated with each PC.

<table>
<thead>
<tr>
<th></th>
<th>% VARIANCE</th>
<th>for COLUMN 7.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(12cm×60mm pilot scale)</td>
<td></td>
</tr>
<tr>
<td>PC1</td>
<td>91.46</td>
<td></td>
</tr>
<tr>
<td>PC2</td>
<td>7.67</td>
<td></td>
</tr>
<tr>
<td>PC3</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>100.00</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.4. The percentage variance associated for the first 3 PCs for Column 7.3 (60mm diameter pilot scale).
7.3.4.1. Loadings Profiles for the 60mm Diameter Pilot Scale Column

The loadings profiles for the first three principal components of Column 7.3 are given in Figure 7.7. Similar patterns were observed with respect to the shapes and positions of the loadings profiles as with the other CG300 columns (Figures 7.3 and 6.11b).

7.3.4.2. Scores for the 60mm Diameter Pilot Scale Column

The scores values against sample size plot for the 60mm diameter pilot scale column is given in Figure 7.8. Although there were only 4 chromatograms used to generate the PCA model, the points for PC1 and PC2 follow excellent linear and quadratic relationships respectively (r=0.99992 and r=0.99999 respectively). However because of the small number of points, a cubic relationship was impossible to derive for PC3.

![Figure 7.7](image-url) The loadings profiles for the first 3 PCs for Column 3 (60mm diameter pilot scale)
7.3.5. Prediction of all 4 chromatograms using the 3 PC model.

In Chapter 6 (section 6.4.3) it was shown that 3 principal components were able to model excellently the whole range of the CG71 chromatograms (Figure 6.8). The 4 chromatograms from the pilot scale were reconstructed in the same way and these are shown in Figure 7.9. Applying the same goodness-of-fit criteria first used in section 5.4.6, each of the 4 predicted chromatogram resulted in a 100% goodness-of-fit. The predicted set from Figure 7.9 can be visually compared with the original set in Figure 7.3.

From the scores correlations with mass loaded given in Figure 7.6, direct predictions over the whole range of mass loads were only possible for PC1 and PC2 scores. For this reason, the prediction of PC3 scores was only possible using data from the smaller columns.
Figure 7.9. The 4 pilot-scale chromatograms predicted using a 3 PC model. All chromatograms had a goodness-of-fit of 100%.

7.3.5.1. The Significance of PC3

The amount of variance associated with PC3 was very small only accounting for 0.87% (Table 7.3). Despite this low value it was established that PC3 was an important principal component for inclusion at the low sample loads. Figure 7.10a shows the prediction of the lowest load chromatogram (sample load of 95.3 mg) with the exclusion of PC3. Figure 7.10b shows the prediction of the same chromatogram with PC3 included. These two figures illustrate how important PC3 is in describing the low sample loads. As the sample load was increased the need for PC3 became negligible (PC3 was seen to be negligible above sample load of about 150 mg).

Because of the importance of PC3 in predicting the low sample loads, it was necessary to make a prediction for the PC3 scores correlation over the range of applied sample loads. Section 6.4.7 showed in detail how the scores could be predicted at larger scales where there are fewer data from those at smaller scales where data are more readily available.
Figure 7.10a. The prediction of the lowest sample load chromatogram from the pilot-scale with the exclusion of PC3. The prediction is shown with the actual chromatogram.

Figure 7.10b. The prediction of the lowest sample load chromatogram from the pilot-scale with PC3 included. The prediction is shown with the actual chromatogram and has a goodness-of-fit of 100%.
7.3.6. The Prediction of a Test Chromatogram within the range of Sample Loads for the Pilot-Scale column.

A fifth chromatogram was obtained by running an erythromycin sample of mass load 600mg. The predicted chromatogram was obtained using the scores correlations shown in Figure 7.6 for PC1 and PC2 scores together with a prediction for PC3 based on the 5cm PLRP1000 column (see Chapter 6, Figure 6.17, although it was suspected that the PC3 contribution would be negligible with this sample size). The test chromatogram is shown in Figure 7.11 together with the original 4 chromatograms.

![Column Volumes](image)

**Figure 7.11.** The 600 mg test chromatogram shown together with the 4 original pilot-scale chromatograms.

Using the 3PC model based around the 4 original chromatograms, the values of PC1 and PC2 scores were 1.828 and 0.385 respectively (interpolated from Figure 7.8). The PC3 scores value was predicted to be 0.11 based from the scores correlations data (Figure 6.17). Combining these scores values with the loadings derived for the 4 original chromatograms (Figure 7.7) and adding in the mean chromatogram, a prediction was obtained for the 600 mg chromatogram. The predicted chromatogram
and actual chromatogram are shown in Figure 7.12. The prediction is excellent, the goodness-of-fit being in excess of 99%.

![Figure 7.12. The predicted chromatogram at 600mg based on a 3PC model compared with the actual chromatogram obtained with 600mg sample load.](image)

### 7.4. Conclusions

The potential of using PCA for the modelling of true pilot scale (60mm diameter) chromatographic separations has been shown in this chapter. The chromatography observed at the pilot scale showed some similarities to the chromatography at the two semi-preparative scales e.g. the positions of the peaks with respect to the amount of column volumes and the similar shifting to the right of the main erythromycin A peak as sample load increased. There were also differences such as poorer resolution of the impurities in front of the main peak. The Principal Component models generated for each set of chromatograms once again yielded very similar patterns in the PC scores to those obtained in Chapter 6, thus reinforcing the possibility of the patterns being generic.

A potential drawback of the PCA approach at pilot scale is caused by the lack of control of the flow rate variable. The pump tended to fluctuate a little each day.
(although it remained constant whilst in use) so that the chromatograms obtained on
days where the flow rate was comparatively fast contained peaks with earlier elution
times to those obtained when the flow rate was slower. A simple transformation from
time to volume was made, although the flow rate needed to be checked regularly
throughout each run.

The potential of the predictive ability of the PCA approach was demonstrated with the
prediction of a test chromatogram within the original range of sample loads.

The following Chapter seeks to establish whether or not the patterns in the PC scores
are likely to hold for any chromatographic separation. PCA is applied to a set of
single-species chromatograms generated using a finite difference solution to the mass-
balance equation.
Chapter 8. The use of Finite Difference Techniques

8.1. Introduction

It has been shown throughout Chapters 6 and 7 that the PCA models generated for the erythromycin separation were remarkably consistent despite differences in the stationary phases and column geometries. The plots of sample load against PC1, PC2 and PC3 scores appeared to exhibit linear, quadratic and cubic relationships respectively.

However because of the fact that only up to 7 sample loads were used, the exact nature of the relationships with respect to PC2 and PC3 scores could not be established. The question arose during the analysis whether or not the PC2 and PC3 scores relationships were better described by discontinuous, linear functions (Chapters 6 and 7). To answer this question experimentally, one would need to perform many more experiments at very small sample load increments within the specified range. Furthermore because of the regular patterns in the scores experienced across the range of geometries and stationary phases, the question arose whether other separations displayed similar patterns in the PC scores with increasing sample load.

In an attempt to answer both of these questions, it was decided to generate some artificial chromatogram data using a numerical solution of the mass balance equation (Chapter 2, equation 1).

\[
\frac{\partial C}{\partial t} + F \frac{\partial q}{\partial t} + u \frac{\partial C}{\partial z} = D_n \frac{\partial^2 C}{\partial z^2} \tag{1}
\]

8.2. The Principle of Finite Difference Methods

It is impossible to solve the mass balance equation (1) using conventional analytical techniques (Guiochon et al, 1994). It is a second order partial differential equation in time and distance and such equations generally require a numerical solution, usually involving finite difference methods.

The principle of the finite difference methods consists of replacing the continuous plane \((z, t)\) by a grid obtained by dividing the space and time into a number of small,
equal segments (of size \( h \) for space and \( \tau \) for time) and replacing each partial differential term in (1) by a finite difference term.

For simplicity the mass balance equation can be re-written as

\[
\frac{\partial G(C)}{\partial t} + \frac{\partial C}{\partial z} = \frac{D_a}{u} \frac{\partial^2 C}{\partial z^2} \tag{2}
\]

Where \( G(C) = \frac{(C + F q)}{u} \). The first order terms

\[
\frac{\partial C}{\partial z} \quad \text{and} \quad \frac{\partial G(C)}{\partial t}
\]

can be replaced by one of the following finite differences

- a forward finite difference: \( \frac{G_{n+1} - G_n}{\tau} \) and \( \frac{C_{n+1} - C_n}{h} \) respectively
- a backward finite difference: \( \frac{G_n - G_{n-1}}{\tau} \) and \( \frac{C_n - C_{n-1}}{h} \) respectively
- or a central finite difference: \( \frac{G_{n+1} - G_{n-1}}{2\tau} \) and \( \frac{C_{n+1} - C_{n-1}}{2h} \) respectively.

Similarly the second order term

\[
\frac{\partial^2 C}{\partial z^2} \tag{4}
\]

can be replaced by a central finite difference: \( \frac{C_{n+1} - 2C_n + C_{n-1}}{h^2} \).

There are several ways to combine the different finite differences in order to solve the mass balance equation. In seeking a solution it was important to find stable solutions and ones that minimises truncation errors. The error associated with each step can, if not minimised, lead to diverging solutions that quickly reach infinity.

### 8.3. An Example involving Finite Difference Techniques

The following example illustrates how the finite difference approach can be used to predict the shapes of elution profiles as the sample size increases. The example is for
a pure component and uses the ideal model which assumes no band broadening. Thus equation (2) becomes

$$\frac{\partial G(C)}{\partial t} + \frac{\partial C}{\partial z} = 0$$  \hspace{1cm} (5)

It was decided to use the forward-backward scheme where the first term is replaced with a forward finite difference and the second term by a backward finite difference. Hence the equation (5) becomes

$$\frac{C_{n+1}^i - C_n^i}{h} + \frac{G_n^i - G_{n-1}^i}{\tau} = 0$$  \hspace{1cm} (6)

which can easily be solved for $C_{n+1}^i$ as follows

$$C_{n+1}^i = C_n^i + \frac{h}{\tau} (G_n^i - G_{n-1}^i)$$  \hspace{1cm} (7)

Equation (7) allows the calculation of the concentration at the new space position $n+1$, knowing the concentration at the previous space position $n$. This method calculates the band profiles along the column with time increments equal to $\tau$. The elution profile at each sample size is the history of concentrations at $z = L$ (the length of the column).

Two sets of chromatograms were generated, one set derived from a slight Langmuirian isotherm and one set from a slight anti-Langmuirian isotherm. Figure 8.1 and 8.2 show the Langmuirian set and anti-Langmuirian set respectively. Notice that the anti-Langmuirian set is the reverse of the Langmuirian set. The profiles of the Langmuirian set broaden and shift to the left with increasing load. The anti-Langmuirian set broaden and shift to the right with increasing load.

Principal Component Analysis was performed on both sets of chromatograms. The amount of variance associated with the first three PCs were identical for both sets and were as follows.

PC1 96.26%, PC2 3.57% and PC3 0.16% (99.99% total).
Figure 8.1. A set of single-component elution profiles generated from a Langmuirian isotherm using the finite difference approach.

Figure 8.2. A set of single-component elution profiles generated from an anti-Langmuirian isotherm using the finite difference approach.
The loadings profiles for Langmuirian and anti-Langmuirian sets are given in Figures 8.3 and 8.4 respectively.

**Figure 8.3.** The loadings profiles of the first 3 PCs for the Langmuirian set of chromatograms generated using the finite difference method.

**Figure 8.4.** The loadings profiles of the first 3 PCs for the anti-Langmuirian set of chromatograms generated using the finite difference method.
The loadings profiles for the anti-Langmuirian set are identical to those for the Langmuirian set in reverse. It was shown in chapters 6 and 7 that the maxima PC2 loadings lie to the right of the PC1 loadings for convex isotherms whilst the opposite is true to concave isotherms. The scores for the first 3 PCs are shown in Figures 4.5 (Langmuirian set) and 4.6 (anti-Langmuirian set). These two plots are identical for both isotherms.

These two plots bear a striking resemblance to the plots derived from the erythromycin separation using columns with widely differing geometries and stationary phases (see chapter 6 and 7). PC1 scores are once again linear. PC2 and PC3 scores bear strong similarities to the quadratic and cubic relationships previously proposed. However when attempting to fit these curves, it became apparent that, with these particular examples, PC2 was not a precise quadratic and PC3 not a precise cubic.

![Figure 8.5. The scores for the first three PCs with increasing sample load for the Langmuirian isotherm generated using the finite difference approach.](image-url)
Figure 8.6. The scores for the first three PCs with increasing sample load for the Langmuirian isotherm generated using the finite difference approach.

Once again, as seen with the erythromycin separation (chapters 6 and 7), PC1 scores intersect PC3 scores when PC2 scores are a maximum. The PC2-PC3 intersection point in this example though, was slightly removed from zero scores.

8.4. Discussion about the Finite Difference Approach

It has been shown in this chapter that chromatograms derived from a finite difference solution to the mass-balance equation produce similar results to the erythromycin separations when the application of PCA is made. The method was used for a simple pure component using the ideal solution to the mass-balance equation. Although a single component is unrealistic in practice, the results of the analysis provide further evidence that PCA models for different chromatographic separations appear to isolate similar features. PC1 scores show a linear relationship with load. PC2 scores show a relationship of similar nature to a quadratic and PC3 scores show a relationship similar to a cubic. The analysis in the previous sections has provided considerable evidence that the PC2 and PC3 scores are continuous functions and not discontinuous ones which may have been inferred from the experimental data of the erythromycin separations.
More chromatogram data is required before generalisations can be made with regards to effect of the Principal Component Analysis. Although the finite difference approach has given some useful indicators, solutions are only possible for one or two molecular species.
Chapter 9. Overall Conclusions and Discussion

The overall objective of the work reported in this thesis was to make use of Principal Component Analysis for the modelling of High Performance Liquid Chromatography. The thesis aims were outlined at the end of Chapter 1 and are briefly as follows.

- To use PCA as a tool for the modelling of a complex separation under realistic, non-linear overload conditions.
- To examine the role data pre-processing has on the quality of the information derived from the PCA.
- To conduct the separations on a range of columns of different sizes and stationary phases to examine whether there were any relationships within the principal components which could be used for predictive purposes.

This chapter provides a brief overall discussion and conclusions of the preceding major chapters.

The introduction to the thesis compared and contrasted analytical and preparative scales of operation were. Some of the widely used types of chromatography were also highlighted. Principal Component Analysis was introduced which included an in-depth literature review of how PCA has been applied in the past to chromatographic data. This conclusions from the literature survey showed that PCA had been used widely in the past to analyse chromatographic data. However it had generally been used for classification purposes and not for the modelling of entire separations.

Chapter 2 was a theory chapter on the traditional approach to the modelling of chromatography. The mass-balance equation was described which is used as the basis for the mathematical approach. In order to solve the mass-balance equation, adsorption isotherm data is required which link together the concentration of pure sample in solution and the concentration adsorbed onto the stationary phase. This chapter provided an in-depth description of adsorption isotherms both for single component and multi-component systems. Descriptions as to how adsorption isotherms are generated were also described in chapter 2. The main conclusions from
this chapter were that mathematical models have only had real success in predicting binary separations. As such, realistic biological separations containing many molecular species could not be modelled effectively using the traditional mathematical approach.

Chapter 3 provided a thorough description of the theory behind Principal Component Analysis. A simplified example was given on how PCA was to be used in the thesis to handle chromatographic data. Seven samples of increasing sample size were described in the example using 11 absorbance values at regular time increments.

Although classification was not used in this thesis, a section was also given on how PCA is used for classification. Other multivariate statistical techniques like Partial Least Squares were also summarised.

Chapter 4 was the first results chapter in this thesis. The data used were derived from Chandwani’s thesis (1995) who used size exclusion chromatography to separate three proteins; a monomer, a dimer and a trimer. The main conclusions from chapter 4 were as follows:

- When chromatograms are generated at different flow rates, the transformation of all chromatograms from a time to a volumetric basis is necessary prior to implementing PCA.
- PCA was able to distinguish some of the subtle changes in chromatograms not visible to a human operator. These changes included those resulting from a small change in temperature e.g. from 23°C to 28°C.
- Normalisation of all chromatograms to constant area was made but did not add to the quality of the diagnostic information derived from the PCA.

Chapter 5 was the first results chapter based on the erythromycin separation. Three columns identical in every respect except for length (25cm, 15cm and 5cm) were used to separate erythromycin samples which were added in direct proportion to the size of each column. Experimental design techniques were used to investigate the effects of
4 process variables load concentration, load volume, temperature and pH on the ease of scale-up on the separation. The main conclusions were as follows:

- Since changes in the four process variables caused positional shifts, a pre-processing stage was necessary prior to PCA in order to provide maximum diagnostic information. Essentially this pre-processing stage involved stretching or compressing the time axis of each of the chromatograms so that the main erythromycin A peak was common to every run.
- Only a fraction of the runs were required on the largest (25cm length) column to obtain a significantly accurate PCA model.
- Accurate predictions (over 92% fit) were achieved across the scales
- With only a single PC model it was possible to predict across the three scales of operation. This was due to the small (about 4-fold) range of sample loads applied to each column.
- The pH and temperature variables were considered to have contributed unnecessary confusion to the analysis.

Chapter 6 extended the methods developed in Chapter 5. Since it was perceived that the temperature and pH variables were creating unnecessary confusion in the analysis, it was decided only to vary sample size. A constant concentration of 20 mg/mL was maintained throughout and sample volume of between 1 and 10% bed volume was applied to each of three columns thus achieving true overload conditions. The 3 columns used were each of different size and had different stationary phases. The smallest was a 5cm×0.46cm column packed with 8μm polystyrene particles. The largest was a 15cm×0.75cm column packed with 35μm methacrylate particles. PCA models were developed for each column. The main conclusions were as follows:

- PC1 scores increased linearly with increasing sample size.
- PC2 scores varied with a curve similar to a quadratic as sample size was increased.
- PC3 scores varied with a curve similar to a cubic as sample size was increased.
- The patterns were remarkably consistent regardless of column size or stationary phase used.
• Very accurate predictions of chromatograms were made using the PC models (>98% fit).
• Minimal runs were required on the larger columns whilst maintaining highly accurate predictions.

Chapter 7 likewise used three columns operating with the same degree of overload as the basis for the experimental work. This time the type of stationary phase was maintained (35μm polystyrene) for each column type whilst the column size increased from 15cm×0.78cm to 12cm×6cm. Principal component models were generated for each column. The main conclusions were as follows:

• The same patterns for PC1, PC2 and PC3 scores were observed for the three columns (PC1 linear, PC2 quadratic and PC3 cubic).
• Because of the similarities, there is potential for scale-up.
• It was deemed that the runs on the pilot scale (12cm×6cm) column were not as reproducible due to the lack of control flow rate control. This was remedied by taking the flow rate regularly throughout each run and subsequently converting all chromatograms from a time to a volumetric basis.
• A test pilot-scale chromatogram, within the original range of sample was obtained. This was compared to one predicted at the same sample load. The prediction was excellent which confirmed that the PCA models generated are predictive within the limits of the sample sizes.

Chapter 8 looked at performing PCA on chromatogram data generated using a finite difference solution to the mass-balance equation. PCA was performed on two sets of data, one following a concave isotherm, the other a convex isotherm. The main conclusions were as follows:

• Very similar patterns in the PC1, PC2 and PC3 scores were observed for both sets of generated data as for the erythromycin separation (chapters 6 and 7).
• The patterns in PC2 and PC3 were continuous functions, not discontinuous one which may have been inferred from data in Chapters 6 and 7.
Overall this result lends considerable weight to the conclusion that the behaviour in PC1, PC2 and PC3 observed in this thesis are true reflections of trends in chromatographic data as sample loads are increased. As such they merit further research and analysis.

The belief is that the overall aims of this thesis, which were outlined at the end of Chapter 1 and again at the beginning of this chapter, have been achieved. PCA has been demonstrated to be a very powerful tool for the modelling and analysis of chromatographic traces, especially when the increasing sample size is resulting in the changing of the chromatography from linear to non-linear conditions.

In order for PCA to be effective, similar regions e.g. peak positions must be compared by the method. For this reason, the role of data pre-processing prior to PCA is very important. This was shown in Chapters 4 and 5 where pre-processing was an essential part of the analysis.

In Chapter 6 where only sample size was changed, a pre-processing stage was not performed since the shifts in peak positions as sample size increased were deemed to be a realistic phenomenon for the PCA to model. However any user of the PCA method must understand what type and how much (if any) pre-processing is required and tailor it to suit a particular problem.

Chapter 6 developed a set of strikingly consistent patterns in the PC scores which could well be generic for any chromatographic separation. The benefits of these patterns would obviously help with predictions at larger scales where the number of runs should ideally be minimised. These ideas were confirmed in Chapter 7 where only 4 chromatograms were used as the basis for a model at pilot scale. The results from the finite difference approach (Chapter 8) show the likelihood of the genericity of the PCA approach.

The main benefits of using PCA to model chromatographic separations are summarised as follows:
• The method is fast and accurate

• PCA does not rely on physical data such as isotherms which are not always readily available.

• Entire complex separations can be modelled. A complex mathematical model could handle only 2 peaks, or at the most 3 well-resolved peaks.

• The method appears to be internally predictive—very accurate predictions were obtained within the specified range of sample loads.

• Because of the similarities in the PC scores across different columns, there is great potential for scale-up.

The main disadvantages of the PCA technique are summarised below:

• The method relies on reproducible chromatograms i.e. the user would need to generate a set of consistent and reliable chromatograms.

• Some runs are required at every scale of operation to develop PC models.

• Extrapolation outside the range of sample loads does not appear to be possible unless the loadings at larger scale can be modified from data at smaller scales.

The following chapter makes recommendations for areas of future study.
Chapter 10. Future work

The potential of using PCA to model complex chromatographic separations has been shown throughout this thesis. However this potential needs to be exploited and extended. The following recommendations are made for subsequent work in this area:

- The test separation used throughout this study was that of crude erythromycin. The separation was a complex one and was designed to mimic the type of separation which would be obtained following fermentation and primary downstream processing. The recommendation is that other, realistic test separations be examined in order to establish whether or not similar patterns in the scores values are obtained as the sample size is increased and to confirm the generality of the approach.

- Erythromycin is not particularly soluble in the mobile phase used in these experiments (45% Actonitrile/50% water/5% phosphate buffer). For this reason a maximum sample concentration of 20 mg/mL only was possible. The recommendation is that future test separations use a crude sample which is more soluble than erythromycin. This will allow examination of a greater range of overload conditions and possibly provide a greater challenge for the PCA method.

- Volume overload was applied throughout the work in this thesis whilst maintaining sample concentration. The recommendation is that future studies also examine concentration overload with constant sample volumes.

- Reversed-phase HPLC was used throughout this study. The recommendation is that future studies use the techniques described in this thesis on other types of liquid chromatography such as ion-exchange chromatography and hydrophobic interaction chromatography.

- Fractionation diagrams are now being developed in the Department of Biochemical Engineering at UCL. It would be beneficial if the PCA techniques developed in this thesis were combined with such diagrams in order to provide a quantitative measure of the purities and yields obtained when different sample sizes are separated.
• It was shown in Chapter 6 that the PCA technique was able to predict the timing of a minor contaminant to the main product peak. This demonstrated the potential of PCA for control purposes i.e. for the automatic location of the timing cut-points which would enable the collection of maximum yield. This area requires further investigation.

• It is strongly suspected that the principal component scores and loadings profiles reflect some aspects of the adsorption properties of the species in the separating mixture. The PC1 scores followed a linear relationship with increasing sample size, PC2 scores a quadratic and PC3 scores a cubic. The recommendation is that future studies attempt to answer the question of why the Principal Component Analysis results in the formation of these patterns.

• It has been established that the methods developed allow chromatogram predictions by interpolation within the specified range of sample loads. The recommendation is that some research is made to establish whether or not it is possible to extrapolate the shapes of the loadings profiles at larger scale based on small scale data.

• There were many problems encountered when commissioning the large scale HPLC system (6cm diameter Prochrom column), mainly due to the length of time it had been inactive. The recommendation is that time is spent to gain experience in how best to operate the system. At present the system is partly automated. It would naturally be of benefit if more of the system e.g. the pumping and fraction collection facilities were computer-controlled.
Appendix 1. Column Efficiencies

This appendix provides information of the efficiencies of each of the columns used throughout this thesis. The main product peak, erythromycin A was used to evaluate the efficiencies. It was observed that the CG 71 methacrylate column provided lower efficiencies than the PLRP1000 or CG300 polystyrene columns.

A1.2. Introduction
It was explained in section 2.3 that realistic, non-ideal chromatography columns have a finite efficiency. This is to say that as the components travel through the column they will experience mass transfer effects, adsorption/desorption kinetics and other smaller effects such as eddy diffusion. These result in band broadening i.e. the widening of the elution peaks. High efficiency columns are those which give rise to tall, narrow peaks whilst low efficiency columns provide smaller, broader peaks.

The efficiency, $N$, of a chromatographic column was defined in section 1.4.2 (equation 4) as

$$N = 5.54 \left( \frac{t_1}{w_{1/2}} \right)^2$$

(1)

Where $t_1$ is the retention time of the peak of interest and $w_{1/2}$ is the width of that peak at half its height. The efficiency is expressed as the number of theoretical plates required to provide the resulting peak. The value of $N$ is often expressed as a number of plates per metre of column bed. So if a 25cm column had 250 plates, its efficiency would be expressed a 1000 plates per metre. The above equation is only strictly valid for symmetrical peaks. Some of the peaks in this thesis were not symmetrical so a few minor inaccuracies may have resulted.

The peak which is used to provide the efficiency measurement is generally a non-binding material, usually acetone. However for the work in this appendix, the main product peak, erythromycin A is used for the efficiency calculations since this gives a
more relevant measure of the separation capabilities of each column for the particular target molecule. Figure A1.1 shows how the parameters $t_1$ and $w_{1/2}$ are evaluated.

![Graph](image)

**Figure A1.1.** A diagram showing how parameters $t_1$ and $w_{1/2}$ are calculated. This example is for the CG71 column.

### A1.3. The effect of flow rate

It is known (see Guiochon et al, 1994) that band broadening (and hence efficiency) is a function of the flow rate of the mobile phase. The classic Knox equation relates $H$ (height of a theoretical plate = $1/N$) to flow rate $u$ as follows

$$H = Au^{1/3} + B/u + Cu$$

where $A$, $B$ and $C$ are all constants for a particular separation system. Figure A1.2 shows a typical plot from a Knox equation.

There is a minimum value of $H$ (and thus a maximum value of efficiency $N$) which corresponds to the optimum flow rate. This value for the optimum flow rate is only a rough guide to the operating flow rate. For example the optimum flow rate derived from such a plot may be too low to be economical. The Knox equation plot does,
however provide guidelines as to the range of flow rates which would provide particularly low efficiencies and therefore should be avoided.

![Diagram of Knox Equation](attachment:image.png)

**Figure A1.2.** A typical Knox Equation plot relating height of a theoretical plate $H = \frac{L}{N}$ with flow rate of mobile phase $u$.

A1.4. Discussion.

Optimisation of parameters such as flow rate has not been an aim of this thesis. One flow rate setting was used for each set of experiments. The same flow rate of 1.2 mL/min was used for all the PLRP1000 runs in Chapter 5. For the experiments in Chapters 6 and 7, the flow rates were as follows:

Chapter 6.  
1.2 mL/min for PLRP1000 5cm × 4.6mm  
1 mL/min for CG 71 15cm × 7.5mm  
1 mL/min for CG 300 15cm × 4.6mm  

Chapter 7.  
1 mL/min for CG 300 15cm × 7.8 mm  
1.64 mL/min for CG 300 15cm × 10 mm  
70 mL/min for CG 300 12cm × 60 mm.\(^1\)

\(^1\)The flow rate on the large-scale Prochrom fluctuated from day to day between 70 and 80 mL/min.
Usually when chromatography is scaled up, the flow rate is increased in proportion to the increase of the ratio of cross-sectional areas. Both the CG71 and CG300 matrices from TosohHaas were soft, non-robust materials. For this reason, the recommended maximum back pressures were very low compared with many other matrices used in reversed-phase HPLC. Thus constraints were imposed on many of the flow rates used throughout this thesis. It was only when the CG300 15cm × 10 mm column was used, could its operating flow rate be scaled from the previous size (the 15 cm × 7.8 mm column, with flow rate of 1 mL/min). The back pressure associated with a flow rate of 1.64 mL/min \((10^7/7.8^2 \times 1)\) for this column was permissible.

### A1.4.1. Values of N

Masses of crude erythromycin were loaded onto each column in direct proportion to the bed volume. The same mobile phase of 45/55 acetonitrile/water was used throughout and each sample had concentration of 20 mg/mL. Table A1.1 gives the efficiencies of the columns used in this thesis.

<table>
<thead>
<tr>
<th>Column</th>
<th>Mass loaded (mg)</th>
<th>Flow rate (mL/min)</th>
<th>N (plates)</th>
<th>N (plates per metre)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5cm × 4.6mm</td>
<td>0.15</td>
<td>1.20</td>
<td>294</td>
<td>5870</td>
</tr>
<tr>
<td>PLRP1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15cm × 4.6mm</td>
<td>0.45</td>
<td>1.20</td>
<td>641</td>
<td>4280</td>
</tr>
<tr>
<td>PLRP1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25cm × 4.6mm</td>
<td>0.75</td>
<td>1.20</td>
<td>1292</td>
<td>5170</td>
</tr>
<tr>
<td>PLRP1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15cm × 7.5mm</td>
<td>1.20</td>
<td>1.00</td>
<td>222</td>
<td>1480</td>
</tr>
<tr>
<td>CG71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15cm × 4.6mm</td>
<td>0.46</td>
<td>1.00</td>
<td>216</td>
<td>1440</td>
</tr>
<tr>
<td>CG300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15cm × 7.8mm</td>
<td>1.32</td>
<td>1.00</td>
<td>445</td>
<td>2970</td>
</tr>
<tr>
<td>CG300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15cm × 10mm</td>
<td>2.16</td>
<td>1.64</td>
<td>498</td>
<td>3320</td>
</tr>
<tr>
<td>CG300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12cm × 60mm</td>
<td>95.6</td>
<td>70</td>
<td>144</td>
<td>1200</td>
</tr>
<tr>
<td>CG300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A1.1. Table showing the efficiencies of the columns used in this thesis. The flow rates and sample loads are also given. All sample concentrations were 20 mg/mL and the same mobile phase used throughout (45% Acetonitrile).
The figures displayed show the mean values of 2 runs. The columns were tested at the beginning and end of every set of runs, and no significant change in the efficiencies were observed (values of N were within ± 2%).

The 3 PLRP1000 columns have the highest efficiencies with value ranging between 4300 (15cm column) and 5900 plates per metre (5cm column).

The 3 CG300 at analytical (4.6mm diameter) and semi-preparative scale (7.8mm and 10mm diameter) show marked discrepancies. The smallest (4.6mm diameter) has a very low efficiency of only 1440 plates per metre, whereas the largest 2 have efficiencies of around 3000 plates per metre. This is probably explained by the fact that the flow rate used for the 4.6mm diameter column is nowhere near its optimum value. The flow rates for the 7.8mm and 10mm diameter columns were in proportion to their cross-sectional areas. The suggestion is that if the flow rate for the 4.6mm diameter column were also in proportion (= 0.35mL.min), then the efficiency would be more comparable. However the much reduced flow rate would prove disadvantageous from the point of view of the total run time. The large-scale Prochrom column gave an efficiency of about 1200 plates per metre. This was a low value compared with the smaller analytical columns. This low figure is expected at large-scale, indeed Subramanian (1991) suggests that preparative-scale columns can only realistically attain a value for N of 1/3 of the value at analytical scale.

The CG71 column had a low efficiency of 1480 plates per metre, which is indicated by the short and wide peak (Figure A1.1).

Values of N between the different columns are not strictly comparable. This is because one does not know how far these values are removed from the maximum possible efficiency (see Know equation plot, Figure A1.2). For this reason, any future efficiency checks must be made with the flow rates and sample sizes specified in Table A1.1.
Appendix 2. Review of the use of Artificial Intelligence in Bioprocessing

The applications of AI in bioprocess control, as with all other forms of control, have been mainly concerned upstream sections, principally fermentation. This highlights the lack of integration in bioprocesses between upstream and downstream operations. The following subsections attempt to explain the fundamentals of neural networks, fuzzy logic and expert systems which have been used as an alternative to multivariate statistical techniques in bioprocess control.

A2.1. Neural Networks

Artificial neural networks (ANNs) which attempt to model the human brain allowing computers to 'learn' became a practical possibility when Minsky invented the Perceptron in 1969. The ability for computers to learn, adapt, associate and generalise is obviously a very powerful phenomenon. Kohonen (1984) defines an ANN as a 'massively parallel interconnected network of simple (usually adaptive) elements and their hierarchical organisations which are intended to interact with the objects of the real world in the same way as the biological nervous system does'. It is this ability to learn and adapt which makes them suitable for process control. Neural Networks can also be useful for their ability to generalise e.g. when an ANN is subjected to a signal for the first time, the network will use information about similar and related inputs to make an educated guess as to the corrective action required [Barnard (1992)]. The most widely used ANN in process engineering is the feedforward network developed by Rummelhart and McLelland (1986). It is the ability to capture process non-linearities and to adjust dynamically to environmental and time-variant changes which is of most benefit in process engineering [Morris et al (1991)].

Chandwani's thesis (1995) contains an in-depth description of neural networks. Basically a neural network consists of three layers: The input layer receives information from the outside world, the hidden layer(s) are responsible for processing information and the output layer presents the processed information to the outside world. Each processing element consists of a combining function and a transfer
function. The transfer function is a relationship between input and output, common forms being a step, linear and sigmoidal relationship. The combining function combines inputted transfer functions (often a summation) before outputting the result to the next element in the sequence. A diagram of a neural network with three elements per layer is given in Figure A2.1.

Transfer functions are often tailored to a specific process [Willis et al (1991)] and it is this ability to tailor the transfer function which make ANNs attractive for process control. The applications of neural networks are very wide, ranging from learning about speech dynamics [Bengio and De Mori (1992)] to process control [Leigh (1992)].

Figure A2.1. A (feedforward) artificial neural network with 3 elements per layer.

In bioprocessing, there have been a large number of applications of ANNs, a common one being the estimation of biomass [Di Massimo et al (1991) and Morris et al (1991)]. Morris et al developed a neural network for the estimation of biomass concentration in a fed batch penicillin fermentation. The study revealed that one of the applications of the neural network was in fault detection. The paper also explains that neural networks have similar advantages for control that statistical methods like PCA in that they are able to cope with non-linearities and even proposes that the combination of neural networks with PCA could provide an even more powerful data handling technique.
ANNs have been used in fermentation monitoring because of their ability to take on-line information and give quality information about the fermentation. Cooney et al (1991) used all available data such as oxygen uptake rate and temperature to provide inputs. The subsequent outputs provided information such as nutrient limitation. Linko and Zhu (1992) developed a feedforward network to control a glucoamylase fermentation, the control mechanism being the use of the network to estimate and predict the state variables on the basis of available on-line measured data.

There have been very few applications of neural networks in chromatography. Rowe (1994) used a neural network to classify a series of chromatograms each with 100 points which meant that the input layer contained 100 neurons. The resulting network showed great promise for rapid diagnosis of peak profiles but the disadvantages proved too great. These include the vast computational power needed to train the network with such a large number of networks. Pre-processing functions were needed to extract resolution and retention times criteria, the use of which may be too time consuming for practical applications. Also the network was unable to cope with time-variant shifts, further illustrating a problem with using ANNs.

A2.2. Fuzzy Control Systems

Fuzzy logic was first developed by Zadeh in the 1960s to model linguistics truth tables which have non-binary truth values. Binary logic traditionally requires a 'true' or 'false' response, but fuzzy logic is less restrictive and offers an alternative, suggesting the degree of truth to which a particular expression is true or false. Fuzzy logic is viewed as being at a high-level supervisory control level where human operators interact with the process especially in the determination of set points and physical variables [Spinrad and Ray (1990)]. At this level, mathematical models and traditional controls are difficult to develop which makes fuzzy systems ideal as an interface between operator and managerial levels. Fuzzy systems are usually easy to develop and the reasoning of the human operator can be incorporated into the control system.
Fuzzy logic has been used in many disciplines including risk analysis, economic forecasting and weather forecasting. It has been used in fermentation monitoring and control together with other techniques like neural networks [Otto (1990), Shimizu and Ye (1995)] and expert systems [Siimes et al (1992), Shimizu and Ye (1995)]. Siimes et al (1992) developed LAexpert, an expert system using fuzzy reasoning to cope with on-line, real time handling of uncertainties in process knowledge, measurements and diagnosis. An expression evaluator calculated specific rates for growth and substrate and product formation at different physiological states during cultivation from measured data. These specific rates were then compared with target rates stored in a database. If differences outside set tolerances were observed, the reasons for faults were analysed on the basis of knowledge from a ‘knowledge network’ and of fuzzy membership functions of the process variables.

The applications of fuzzy logic in chromatographic data interpretation have been confined mainly to peak tracking. To account for the imprecisions in peak area, the change of peak elution order and the uncertainties from overlapping peaks - all of which contribute to the recognition of peaks, the method used by Otto et al (1988) was based on fuzzy theory which is able to cope with uncertainties and imprecise data. This study was centred around the imprecise peaks resulting from the HPLC separations of vitamins using binary and ternary mobile phases. With the ternary mobile phase, correct peak identification required two-channel monitoring using a UV absorbance and a detection meter. The developed method operated using on-line peak tracking even if unknown samples were to be analysed.

The potential of fuzzy logic to handle uncertainties in measurement was exhibited by King in his PhD thesis (1993) who used fuzzy logic to identify optimal cut points. The method involved using exponentially modified Gaussian model for a peak and could not be used for rapid analysis. Despite having potential for tight control and monitoring of product quality, quantitative monitoring of chromatography using fuzzy logic is still only a possibility [Nugent and Olson (1990)] and will not be adopted in this study.
A2.3. Expert systems and knowledge bases

An expert system is a computer program capable of human-like expert reasoning and decision making within a special domain of expertise [Linko (1988)]. In process engineering, they are able to contain the knowledge and experience of human specialists into a software package that is able to control, monitor, troubleshoot and recommend corrective actions. A knowledge base is central to an expert system which is designed to continuously capture and update all of the expert’s knowledge. Knowledge bases contain mainly qualitative information which are stored as rules and provide a platform for the interfacing between the process and supervisory levels. For these rule-based expert systems, the knowledge base consists of facts and rules acquired from a human expert and the associated knowledge base will be comprised of possible outcomes, of which a single decision will be arrived at. There are also tandem or hybrid decision systems where the knowledge base incorporates a model together with additional facts based on the expert’s reasoning skills. In addition, there are situation-action systems in which the knowledge base consists of a network of situation-action rules. They allow data to be collected from an ongoing process and a set of decisions is generated which can be used for control.

Expert systems have been used in fermentation control for fault diagnosis, state recognition and aeration control. Karim and Halme (1988) developed a real-time expert system for monitoring and analysing, diagnosis of faults and malfunctions and on-line control of fermentations. Their methods were applied to a batch fermentation in which the enzyme α-amylase was produced using Bacillus Subtilis. Pokkinen et al (1992) used LAexpert with a fuzzy expert system shell for the diagnosis of microbial activity during lactic acid fermentations. The system proved successful in detecting faults not deducible from off-line analytical data. Hubert et al (1994) developed a knowledge-based expert system to facilitate the development of fully automated methods for the solid-phase extraction of relatively hydrophobic basic drugs from plasma, coupled with their determination by HPLC. Ren et al (1994) designed and implemented a computer supervised knowledge-based system with an on-line machine vision sensing system. Besides the normal data acquisition and process monitoring, the system software allowed users to set process recipes at any stage and
to define process control strategies. Using measured process variables combined with cellular information from the sampling and vision system, complete bioprocess control and automation was achieved.

Pestka and Thompson (1990) applied a knowledge-based system to ion-exchange and affinity chromatography. Rules written into the software were the basis for the analysis of the chromatographic performance and were modified by an expert system familiar with the typical variances which may occur in the feedstock. Characterisation experiments were performed on the column to determine optimum operating conditions and if the resulting separation of the characterisation feedstock was inadequate, the separated feedstock was recycled with an adjusted gradient in order to improve resolution. The knowledge-based system developed diagnosed that the proposed estimate of protein concentration was incorrect and altered it accordingly to minimise product loss. This system highlighted the strength of AI to control chromatographic separations. An expert system with the ability to resolve overlapping peaks in HPLC was developed by Conti et al (1991), the time taken being the order of minutes. Pirogov et al (1995) made use of expert systems to develop a piece of software which allows chromatograms to be generated using several mathematical functions and their superposition. The program allowed chromatogram shape estimation to be influenced by human experience and allowed the option of adding these generated chromatograms to a database.

This appendix has shown the utility of AI applications in some aspects of bioprocessing. Most applications have been in fermentation and the few downstream application such as for chromatography, have been met with limited success. This is mainly due to the complexities associated with such approaches.
Appendix 3. References


A Practical Investigation into the use of Principal Component Analysis for the Modelling and Scale-up of High Performance Liquid Chromatography

References


