A study of techniques for the assessment and control of performance in preparative scale liquid chromatographic separations

A thesis submitted for the degree of
Doctor of Philosophy
by
Joshua Matthew Philip King B.Eng (Hons.)

Department of Chemical and Biochemical Engineering
University College London

July 1994
At the end of the room were a couple of long tables smothered in, at the last count, six Macintosh computers. In the middle was the Mac II on which a red wire-frame model of his sofa was lazily revolving with a blue wire-frame model of his staircase, complete with banister rail, radiator and fuse-box details, and of course the awkward turn halfway up.

The sofa would start out spinning in one direction, hit an obstruction, twist itself in another plane, hit another obstruction, revolve round a third axis until it was stopped again, then cycle through the moves again in a different order. You didn’t have to watch the sequence for very long before you saw it repeat itself.

The sofa was clearly stuck.

Douglas Adams, "Dirk Gently's Holistic Detective Agency".
Abstract

In this study techniques have been developed to control the performance of preparative scale liquid chromatographic separations. Analysis of chromatograms has been carried out to determine how the individual components comprising the chromatogram overlap with one another. A mathematical model, consisting of one template function to describe the elution profile of each individual component, has been developed to describe the chromatogram and has been successfully applied to chromatographic data generated by a cation exchange of hen egg white. Parameters for the functions have been obtained by use of a directed search optimisation algorithm and results compared with those obtained by scanning SDS polyacrylamide gels of column fractions. The effect of the size of the difference between the actual and predicted chromatograms on the accuracy of the description of the separation gained has been investigated in terms of the subsequent effect on the quality of possible control decisions made.

Fuzzy logic techniques have been applied to interpret information obtained from chromatogram analysis in order to identify the position of the product within the overall chromatogram. Sources of failure related to either chromatographic phenomena or to mathematical features of the fuzzy method have been investigated.

Finally control actions which may be made using the information obtained from chromatogram analysis and product identification are investigated. In order to simplify the visualisation of control decisions the chromatograms have been transformed into fractionation diagrams, allowing the optimum fraction, in terms of purity and productivity, to be selected. Consideration is also given to when re-optimisation of separation is appropriate by calculating possible improvements to purity and productivity of the product fraction. This requires the use of the fractionation diagram and additional information gained off-line about the retention behaviour of individual components.
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Introduction</td>
<td>19</td>
</tr>
<tr>
<td>1.1 Introduction to liquid chromatography</td>
<td>20</td>
</tr>
<tr>
<td>1.1.1 Ion exchange chromatography</td>
<td>20</td>
</tr>
<tr>
<td>1.1.1.1 Adsorption equilibria</td>
<td>21</td>
</tr>
<tr>
<td>1.1.2 Affinity chromatography</td>
<td>22</td>
</tr>
<tr>
<td>1.1.3 Gel permeation chromatography</td>
<td>23</td>
</tr>
<tr>
<td>1.1.4 H.P.L.C - High performance liquid chromatography</td>
<td>23</td>
</tr>
<tr>
<td>1.2 Chromatographic output</td>
<td>24</td>
</tr>
<tr>
<td>1.2.1 Causes and effects of inefficiencies</td>
<td>26</td>
</tr>
<tr>
<td>1.2.1.1 Extra column effects</td>
<td>26</td>
</tr>
<tr>
<td>1.2.1.2 Intra column band broadening</td>
<td>27</td>
</tr>
<tr>
<td>1.2.1.2.1 Eddy diffusion</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1.2.2 Mobile phase mass transfer</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1.2.3 Stagnant mobile phase mass transfer</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1.2.4 Stationary phase mass transfer</td>
<td>28</td>
</tr>
<tr>
<td>1.2.1.2.5 Longitudinal diffusion</td>
<td>28</td>
</tr>
<tr>
<td>1.2.2 Band-broadening and scale-up</td>
<td>29</td>
</tr>
<tr>
<td>1.3 Chromatogram analysis: optimisation and quality criteria</td>
<td>30</td>
</tr>
<tr>
<td>1.3.1 Binary system criteria</td>
<td>30</td>
</tr>
<tr>
<td>1.3.2 Multicomponent criteria</td>
<td>31</td>
</tr>
<tr>
<td>1.3.3 Criteria for non-Gaussian peaks</td>
<td>35</td>
</tr>
<tr>
<td>1.3.4 Column efficiency</td>
<td>37</td>
</tr>
<tr>
<td>1.3.4.1 Measurement and calculation methods</td>
<td>38</td>
</tr>
<tr>
<td>1.3.4.2 Moment method</td>
<td>39</td>
</tr>
<tr>
<td>1.3.4.3 Asymmetric method</td>
<td>39</td>
</tr>
<tr>
<td>1.3.4.4 Exponentially modified Gaussian methods</td>
<td>40</td>
</tr>
<tr>
<td>1.3.4.5 Choice of method</td>
<td>40</td>
</tr>
<tr>
<td>1.3.5 Peak overlap detection by multidetector methods</td>
<td>40</td>
</tr>
<tr>
<td>1.3.5.1 Difference and Ratio methods</td>
<td>41</td>
</tr>
<tr>
<td>1.3.5.2 Derivative methods</td>
<td>43</td>
</tr>
<tr>
<td>1.3.5.3 Methods suitable for general use</td>
<td>44</td>
</tr>
</tbody>
</table>
1.3.6 Peak identification

1.4 Controllable parameters
1.4.1 Column loading
1.4.1.1 Flow rate
1.4.1.2 Analytical and preparative chromatography - sample size
1.4.1.2.1 Volume overload
1.4.1.2.2 Concentration overload
1.4.1.2.3 Comparison of modes
1.4.2 Elution
1.4.2.1 Isocratic elution
1.4.2.2 Gradient elution
1.4.2.3 Step change
1.4.2.4 Multisegment gradients
1.4.2.5 Flow rate
1.4.3 Strategies

1.5 Theories and Models of Chromatographic Separations
1.5.1 Plate theory
1.5.2 Rate theories

1.6 Expert Systems
1.6.1 The knowledge base
1.6.2 The inference engine
1.6.3 Expert systems for chromatography
1.6.3.1 CRISE - CRITERia SElection
1.6.3.2 SOS - System optimisation systems
1.6.3.3 Integrated packages
1.6.3.4 ECAT - Expert chromatography assistance team
1.6.3.5 Method validation
1.6.3.6 PREOPT
1.6.3.7 Dry-Lab
1.6.3.8 MABLAB

1.7 Scope of work
2 Deconvolution

2.1 Introduction to Deconvolution

2.2 Curve fitting

2.2.1 Models

2.2.1.1 The exponentially modified Gaussian function

2.2.1.2 The general exponential function

2.2.2 Objective functions

2.2.3 Constraining the problem

2.2.4 Optimisation Techniques

2.2.4.1 Box-Complex optimisation algorithm

2.2.4.2 Deconvolution failure

2.2.5 Maximum Acceptable Object Function Size

2.2.6 Theoretical Examination of Model Convergence

2.3 Experimental determination of deconvolution effectiveness

2.3.1 Choice of system

2.3.2 Apparatus

2.3.3 Method

2.3.4 Chromatographic method

2.3.4.1 Sample preparation

2.3.4.2 Separation conditions

2.3.5 Analysis of column fraction by SDS-PAGE

2.3.5.1 Sample preparation

2.3.5.2 Gel preparation

2.3.6 Band quantification by laser densitometry

2.3.7 Experimental Results

2.3.7.1 Chromatographic Data

2.3.7.2 Electrophoresis Data

2.3.8 Deconvolution Analyses of experimental data

2.3.8.1 Deconvolution of a single peak

2.3.8.2 Deconvolution of egg white chromatograms

2.3.8.3 Experimental Deconvolution results

2.3.9 Conversion of data to mass based units
2.3.9.1 Conversion of chromatographic absorbance profiles to mass based units 101
2.3.9.2 Conversion of electrophoresis data to mass based data 103
2.3.10 Comparison of deconvolution results with electrophoresis data 105

2.4 Conclusions 107

3 Product Peak identification 133

3.1 Introduction 134

3.2 Fuzzy Logic system parameter selection 135
3.2.1 Reference data set selection 135
3.2.2 Membership function selection 136
3.2.3 Matching techniques 138
3.2.3.1 Matching criterion 139
3.2.3.2 Possibility theory 140
3.2.4 Failure modes of fuzzy logic identification technique 140
3.2.4.1 Single Peaks 141
3.2.4.2 Overlapping peaks and combinations of peaks 142
3.2.4.3 Extra Peaks 143

3.3 Matching Worked Examples 143

3.4 Egg-white chromatograms (deconvolution) 151
3.4.1 Data Handling and Peak Integration 152
3.4.2 Fuzzy Logic System Parameters 152
3.4.3 Egg-white Separation Results 153

3.5 Conclusions 156
4 Control 164
4.1 Fraction Selection 166
4.1.1 Chromatogram Sub-fractions 166
4.1.2 Fractionation Diagrams 167
4.1.2.1 Construction of the Fractionation Diagram 168
4.1.2.2 Disadvantages of the Fractionation Curve 169
4.1.3 The Fractionation - Concentration Diagram 169
4.1.4 Selection of the Optimum Fraction 170

4.2 Effects of deconvolution on fraction curves 173
4.2.1 Effect of deconvolution upon concentration data 174
4.2.2 Implications of deconvolution accuracy on fraction selection 175

4.2.3.1 Selection of constraints for the determination of the optimum fraction 176
4.2.3.2 Selection of the optimum fraction 177
4.2.4 Egg-White chromatogram analyses 181
4.3 Conclusions 184

5. Discussion 195
5.1 Separation Performance 196
5.2 Integration of Analysis techniques 200
5.3 Baseline drift and noise 203
5.4 Comparison of techniques developed with other currently available software 204
5.5 Conclusions 206

6 Conclusions 209

7 Recommendations for Future Work 213
Appendices

<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Computer Programmes and Hardware</td>
<td>216</td>
</tr>
<tr>
<td>A2</td>
<td>Deconvolution Results</td>
<td>243</td>
</tr>
<tr>
<td>A3</td>
<td>Fuzzy Logic Matching Results</td>
<td>250</td>
</tr>
<tr>
<td>A4</td>
<td>Nomenclature</td>
<td>255</td>
</tr>
<tr>
<td>A5</td>
<td>References</td>
<td>263</td>
</tr>
</tbody>
</table>

List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The Langmuir Isotherm.</td>
<td>66</td>
</tr>
<tr>
<td>1.2</td>
<td>A Typical Chromatogram.</td>
<td>67</td>
</tr>
<tr>
<td>1.3</td>
<td>A schematic chromatogram showing the concept of capacity factor, k.</td>
<td>68</td>
</tr>
<tr>
<td>1.4</td>
<td>The Valley to Peak ratio and Peak Separation Criterion.</td>
<td>69</td>
</tr>
<tr>
<td>1.5</td>
<td>Two congruent peaks separated by 1.665 standard deviations.</td>
<td>69</td>
</tr>
<tr>
<td>1.6</td>
<td>Two incongruent coincident peaks.</td>
<td>70</td>
</tr>
<tr>
<td>1.7</td>
<td>The A:B ratio.</td>
<td>71</td>
</tr>
<tr>
<td>1.8</td>
<td>The proposed data analysis and control sequence</td>
<td>72</td>
</tr>
<tr>
<td>2.1</td>
<td>A chromatogram consisting of three peaks. It is not possible to determine the peak starts, ends or retention times because the peaks are overlapped.</td>
<td>109</td>
</tr>
</tbody>
</table>
Figure 2.2  The exponentially modified Gaussian function.

Figure 2.3  The General Exponential Function (\(a = 1.0\)).

Figure 2.4  The General Exponential Function (\(a = 10.0\)).

Figure 2.5  The General Exponential Function (\(a = 25.0\)).

Figure 2.6  Definition of constraints for optimisation using the General Exponential Functions.

Figure 2.7  Box-Complex Algorithm Flow Diagram.

Figure 2.8  The first Gaussian test chromatogram.

Figure 2.9  The fifth Gaussian test chromatogram.

Figure 2.10  The Bio-pilot Chromatography System.

Figure 2.11  An example of the complete chromatogram produced from the separation of egg white.

Figure 2.12  Two mini gels used to analyse the first egg white separation (table 2.5).
Figure 2.13 Deconvolution results of fitting a single exponentially modified Gaussian function and as a single General Exponential Function to objective function values shown.

Figure 2.14 The point differences between each model peak and the experimental peak shown in figure 2.20.

Figures 2.15 The first egg white chromatogram
- 2.17 (Deconvolution results).

Figure 2.18 The second egg white chromatogram.
(Deconvolution results).

Figure 2.19 The third egg white chromatogram.
(Deconvolution results).

Figure 2.20 Densitometry Calibration.

Figure 2.21 Extinction coefficients of egg white components.

Figure 2.22 Comparison of the peaks found by
- 2.24 deconvolution and those found by scanning SDS-PAGE gels for the first to third egg white chromatogram.

Figure 3.1 The elution order membership function for a reference set peak.
The measure rank is X.
Figure 3.2
(a) The match criterion is obtained from the overlap of area membership functions.
(b) A match is considered good if the match criterion is above the threshold.

Figure 3.3 The area data contained in table 3.1 may be represented by this chromatogram.

Figure 3.4 Peak area membership functions for the reference data set listed in table 3.1.

Figure 3.5 The overall match criterion for the matching of data in table 3.2 with data in table 3.5.

Figure 3.6 The overall match criterion for the fourth and fifth actual reference peaks in table 3.20 compared with each reference peak (table 3.17).

Figure 4.1 Flow diagram for the selection of the optimum fraction based on a productivity threshold.

Figure 4.2 The Fractionation Diagram.

Figure 4.3 The Fractionation - Concentration Diagram.

Figure 4.4 A tie line which represents three fractions with equal purification factors.

Figure 4.5 The Fraction Volume.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>Fractionation curves for the first Gaussian test chromatogram</td>
<td>190</td>
</tr>
<tr>
<td>4.7</td>
<td>Fractionation curves for the fourth Gaussian test chromatogram</td>
<td>191</td>
</tr>
<tr>
<td>4.8</td>
<td>The cumulative elution curve for the first Gaussian test chromatogram</td>
<td>192</td>
</tr>
<tr>
<td>4.9</td>
<td>The cumulative elution curve for the fourth Gaussian test chromatogram</td>
<td>193</td>
</tr>
<tr>
<td>4.10</td>
<td>The fractionation diagrams for the first, second and third egg white chromatograms</td>
<td>194</td>
</tr>
<tr>
<td>5.1</td>
<td>Baseline drift correction methods</td>
<td>207</td>
</tr>
<tr>
<td>A1</td>
<td>The calculation method for the matching criterion used by the fuzzy logic identification programme (A1.3). The two portions of the area of intersection are shown. The intersection point is defined by equation A1.2.</td>
<td>242</td>
</tr>
</tbody>
</table>
List of Tables

Table 1.1  Quality Criteria for Chromatograms. 32

Table 2.1  Box-Complex Algorithm Failure Modes. 83

Table 2.2  Physical Properties of Major Egg-White Components. 91

Table 2.3  Gradients used in cation exchange separation. 94

Table 2.4  Formulation of resolving and stacking gels. 95

Table 2.5  Formulation of running buffer, staining and destaining solutions. 95

Table 2.6  SDS-PAGE sample running buffer composition. 96

Table 2.7  Single Peak Deconvolution Results. 99

Table 2.8  Egg-White component extinction coefficients. 103

Table 2.9  Egg-White component dye binding characteristics. 104

Table 2.10  Peak to Valley Ratios for the Central peaks in the test and experimental chromatograms. 107

Table 3.1  Reference data set: computer generated chromatograms. 144

Table 3.2  Peak data giving a successful identification. 145
Table 3.3 Elution order matching criteria. 145
Table 3.4 Peak area matching criteria. 145
Table 3.5 Overall matching criteria. 145
Table 3.6 The transformed trial data set (table 3.2). 146
Tables 3.7 The transformed data set double peaks area matching criteria. 147
Table 3.8 Overlapped peaks elution order matching criteria matched as single peaks. 147
Table 3.9 Overlapped peaks elution order match criteria for double peaks. 148
Tables 3.10 Overall match Criteria for the transformed data set. 148
Table 3.11 Trial data set with an extra peak. 149
Table 3.12 Elution Order Match Criteria. 149
Table 3.13 Peak Area Match Criteria. 149
Table 3.14 Overall Match Criteria. 150
Table 3.15 Expected Egg-White Chromatogram Peak Area Data ( Reference Peak Data ). 153
Table 3.16 Deconvoluted Egg-White Chromatogram Peak Area Data ( trial Peak data. Objective function < - 183.6 ). 154
| Table 4.1 | Properties of the Fractionation - Concentration Diagram. | 170 |
| Table 4.2 | Optimum fractions for the test chromatograms. | 178 |
| Table 4.3 | Actual quality of optimum fractions selected from deconvolution solutions. | 179 |
| Table 4.4 | Optimum fractions for egg white chromatograms | 182 |
| Table 5.1 | Yields for maximum purities and Purities for maximum yield for test chromatograms (see section 2.2.6) | 197 |
| Table 5.2 | Purities at maximum yields with differing extinction coefficients. | 199 |
| Table 5.3 | A typical chromatographic operation sequence | 200 |
| Table 5.4 | Processing times for analysis techniques | 201 |
| Tables A2.1 | First to fifth Gaussian test - A2.5 chromatogram deconvolution results. | 244 - 245 |
| Tables A2.6 | First to third General Exponential Function test chromatogram - A2.8 deconvolution results. | 245 - 246 |
| Table A2.8 | Noisy Gaussian test chromatogram deconvolution results. | 246 |
| Table A2.9 | Noisy General Exponential Function chromatogram deconvolution results. | 247 |
Table A2.10 Experimental Deconvolution Results. 248

Table A3.1 Matching results for the 1st to 6th 251
- A3.6 egg white chromatogram. - 253
Acknowledgements

I should like to thank my supervisors Dr. N.J. Titchener-Hooker of University College London and Dr. A.C. Kenney of Drew Scientific, Chiswick, for their advice and encouragement during the course of this project.

I am also grateful to the SERC and OROS Instruments Ltd. for financing this PhD research project.

I should also like to thank Mr Martyn Vale and his colleagues in the Electronics Workshop for their help with configuration and installation of computer hardware and software, and the staff at OROS - particularly Howard Bates and Penny Vardy for their help with software development.

Financial support from Biotage UK Ltd. is acknowledged.
I am grateful to Dr. P.R. Levison of Whatman Specialty Products for helpful discussions and to Pharmacia for loan of equipment.

Finally I should like to thank the research staff and students at UCL for their advice and help.
Chapter 1.

Introduction.
1.1 Introduction to liquid chromatography

Liquid chromatography is a separation process in which soluble components are separated by adsorption onto or partition with a packed bed of solid material. There are several different types of liquid chromatography which fall into the categories - adsorption chromatography and partition chromatography. Adsorption chromatography includes ion exchange and affinity, while partition chromatography includes size exclusion (or gel permeation). For each type of chromatography a different type of packing or media is required because of the different separation mechanisms involved. Chromatography is widely used in the biotechnology industry. Due to the relatively high cost of chromatographic media it is generally one of the last unit operations. In addition particulates and other materials such as lipids, oils and fats cause column fouling and in large quantities blockage of the packed bed. This therefore means that all cells and cell debris must be removed by centrifugation or filtration.

Chromatography is particularly important in the pharmaceutical industry especially in the case of protein based drugs where most regulatory authorities require at least one chromatographic step to be included in any process (Sofer and Nystrom, 1989), as it is impossible to achieve the necessary purity using non-chromatographic methods.

1.1.1 Ion exchange chromatography

Separation in ion exchange chromatography (Engelhardt, 1979) relies on each of the components or component groups in the mixture having a different ionic charge. With protein mixtures the charge is dependent upon the pH of the solution and the isolectric point of the protein. The packing materials used for ion exchange have functional groups eg. groups containing COOH or NH$_4^+$ which can carry a charge attached to the surface. There are two types of exchangers: cation and anion exchangers. Anion exchangers exchange negatively charged ions with a positively charged packing. In the same way cation exchangers exchange positively charged ions with a negatively charged packing. Traditionally the most common material used has been cellulose substituted with diethylaminoethyl (DEAE-) or carboxymethyl (CM-) groups. More recently cross-linked agarose beads have been used which have the advantage of being more rigid and therefore allow high flow rates to be used without the risk of the bed collapsing.
Ion exchange chromatography is a multi-stage process. Firstly the column is equilibrated to the loading conditions i.e. the conditions at which the functional group and the protein or proteins to be bound are opposite in charge. Next the mixture is loaded onto the column and the column washed with buffer to remove any unbound proteins. The bound proteins are then eluted by changing the ionic conditions or by changing the pH. The change in ionic conditions or pH can be carried out isocratically i.e. with a step change to a constant ionic strength or by using a gradient i.e. a change between two ionic strengths over a set time. Each of these methods has a different effect on the separation of components, i.e. the performance of the separation. The output response of the column will also depend on the adsorption isotherm of the proteins to be separated.

1.1.1.1 Adsorption Equilibria
The equilibrium of the solutes between the mobile and stationary phase is often described using the Langmuir isotherm see equation 1.1 (Langmuir, 1916). Certain assumptions are made in this model:
(i) an equilibrium is established
(ii) only one particle or species adsorbs per site
(iii) adsorption is equally likely on all sites

\[
\frac{w}{c} = \frac{b c}{(1 + K c)} \tag{1.1}
\]

where \( w \) = the concentration of solute on solid (g solute adsorbed / g solid)
\( c \) = concentration of solute in the liquid (g solute / g solvent),
\( b \) = a constant related to the asymptotic value of the adsorbate loading ((g solute adsorbed.g solvent)/(g solute.g adsorbent)),
and \( K \) = an equilibrium constant (g solvent / g solute).

When \( Kc \gg 1 \), the isotherm is strongly favourable and when \( Kc < 1 \) the isotherm is nearly linear. (See figure 1.1.)
The exact shape of the equilibrium curve depends on the nature of the interactions between the free solute and the packing material. This model of equilibrium may be extended to multicomponent systems (where each component is given an adsorption isotherm to predict its behaviour) and can be used for the basis of predictive models for liquid chromatography, providing that solute-solute interactions are not significant.
In ion exchange chromatography the adsorption isotherm can be considered to be linear over a normal concentration range used in analytical chromatography.

The mechanism by which separation is obtained is one of reversible adsorption. Separation is possible since substances normally have different electrical properties and so are released at a different pH or salt concentration. In addition to the charge effect other types of binding may occur. These are mainly due to van der Waals forces and polar interactions. In ion exchange chromatography most of the components to be separated bind at the top of the column and because of this columns tend to be short and have a large cross sectional area. Once the desorption step has been completed the column must be regenerated i.e. any remaining protein must be removed and column re-equilibrated to the loading conditions. This is usually done by washing with a high salt concentration (higher than that used in the elution process) followed by washing to reduce the salt concentration to the value required for loading the column (re-equilibration).

1.1.2 Affinity Chromatography

Affinity chromatography, like ion exchange chromatography, is a reversible adsorption process (Hamilton and Sewell, 1977). However affinity chromatography is much more specific than ion exchange chromatography and the adsorption isotherm is strongly irreversible. In affinity chromatography binding occurs on ligands. The ligands are substances which have specific affinities e.g. an enzyme may be used to attach to its substrate or an antibody and its antigen. Co-enzymes are also used e.g. NAD'. If NAD' is used then for example all dehydrogenases which are NAD' dependent will be bound. This type of ligand which can bind many different substances is known as a general ligand. Others include NAD, NADP, and ATP. Cibacron Blue F3GA has been found to be a general ligand (KopperSchlager, Bohme, and Hofmann, 1982) which binds a variety of enzymes dependent on NAD or NADP even though it has no obvious structural similarity to NAD or NADP. Once the column has been loaded, and any unbound proteins washed through, elution must be carried out. This is done by passing free ligand through the column which competitively bind with the components bound to the column matrix or by passing a compound which has a greater affinity for the ligand than the bound material. More simply it may be achieved by altering the
pH or the salt concentration. Again column regeneration must be carried out.

1.1.3 Gel permeation chromatography
Gel permeation chromatography differs from both affinity and ion exchange chromatography since in theory no chemical or ionic adsorption takes place. It relies upon the size of molecules to be separated (Determann, 1968). In practice however some degree of adsorption usually takes place. The packing material consists of porous beads into which the species to be separated diffuse. The larger molecules are excluded from the pores because of their size and so travel down the column more quickly. Smaller molecules are retained within the column for a longer period as they have a longer path to travel ie. they are held up within the beads. Molecules which are medium sized only partially diffuse into the gel and can be said to be in a state of diffusion equilibrium. The components will elute in order of decreasing molecular size ie. largest first. The parameters which can be varied to improve resolution are the bead size, pore size, temperature (and hence viscosity), flow rate and the column length. By increasing the column length the retention time of those molecules which diffuse into the bead pores is increased in comparison to those molecules which are not. Reducing the bead size also improves the resolution as the interfacial area of the gel increases thus allowing an equilibrium to be achieved faster but causes a higher pressure drop.

1.1.4 H.P.L.C.
HPLC (high-performance liquid chromatography or high pressure liquid chromatography) is an ill defined term. It is normally used to describe liquid chromatography performed at high pressures with stationary phases typically less than 10 μm in diameter and pressures of 20 to 200 barg (Engelhardt, 1979). The stationary phases generally used for HPLC differ from the other types as soft gel packings cannot withstand the high pressures necessary in HPLC. Common stationary phases include silica gel, alumina and magnesium oxide as they are relatively polar materials with high specific areas. The mobile phase used is relatively non polar eg. heptane to tetrahydrofuran. The main advantage of HPLC is the decrease in separation time over low pressure chromatography. This is due to the small particles used which result in rapid mass transfer and hence the ability to employ faster flow rates.
The same separation mechanisms as in low pressure chromatography may be utilised in HPLC, i.e. size exclusion and ion exchange etc. One mechanism which is unique to HPLC is reversed phase chromatography where the stationary phase is less polar than the mobile phase and is usually only used in HPLC (Hamilton and Sewell, 1977). Chemically bonded octadecylsilane (ODS) is the most frequently used stationary phase. In adsorption chromatography water is one of the strongest elution media since it strongly interacts with the active centres in silica gel, so that adsorption of sample molecules becomes highly restricted and they are rapidly eluted as a result. The opposite is true of water in reverse phase chromatography - water cannot wet the non-polar alkyl groups and so gives the slowest elution rate. Because of this the greater the amount of water in the eluent the longer is the retention time. A similar list of relative affinities to the surface as with adsorption to silica may be written but with the degree of affinity depending on hydrophobicity.

1.2 Chromatography Output
The data produced from a chromatographic separation is typically in the form of a chart known as a chromatogram (see figure 1.2). Each peak on a chromatogram represents the response of a detector to a change in eluent composition. This is often the absorbance of ultraviolet radiation but other means of detection are available (see below). The peak represents a group of molecules which have identical retention properties under the particular separation mechanism in use but with a variance in accordance with the peak width. Since the separation may be due to one common functional group it is possible for each peak to contain a number of components i.e. more than one compound may have the same retention time. The composition of a peak can only be determined by other analytical techniques e.g. for proteins gel electrophoresis. The shape of the peak is governed by a number of factors. The shape of the adsorption isotherm is important. A linear relationship between the concentration in the stationary phase and the mobile phase gives rise to Gaussian peak shapes under ideal conditions. This is true for most isotherm types providing that the concentrations are sufficiently low (i.e. the solute concentrations are such that they correspond to the linear regions of the isotherm). Skewed peaks (i.e. non-symmetrical peaks) may be due to non-linear isotherms; Langmuir isotherms give rise to peak
tailing \(^1\) and anti-Langmuir isotherms cause fronting \(^1\) (Meyer, 1988). With non-linear isotherms unlike linear isotherms retention time will vary with sample concentration (Dose and Guiochon, 1990).

An ideal separation with small sample loadings would produce Gaussian shaped peaks which can be defined mathematically by the equation:

\[
h_i = \frac{A_i}{\sigma_i \sqrt{2 \pi}} \exp \left( -\frac{(V - V_i)^2}{2\sigma_i^2} \right)
\]

where \(h_i\) is the concentration of component \(i\) at an elution volume \(V, V_i\) is the retention volume and \(\sigma_i^2\) is the variance of the concentration. The parameter \(A_i\) is the area of the peak. Gaussian peaks are strictly theoretical (Foley and Dorsey, 1983) as inefficiencies are always present due to extra and intra column sources (see 1.2.1 below).

A more accurate description is the exponentially modified Gaussian peak, given by equation 1.3:

\[
h_i = A_i \exp \left( \frac{1}{\tau_i} - \frac{V - V_i}{\tau_i} \right) \text{erf} \left( \frac{1}{\sqrt{2}} \left( \frac{V_i}{\sigma_i} + \frac{\sigma_i}{\tau_i} \right) \right)
\]

In this equation an additional parameter \(\tau\) is introduced. This parameter describes the skew of the peak and is the decay constant of an exponential decay function which is convoluted with a pure Gaussian to produce the exponentially modified Gaussian function (Grushka, 1972). This model is both theoretically and experimentally justifiable since peaks are known to become modified by extra column effects, non-equilibrium mass transfer, and detector response lag. The use of the incorrect model may cause an error in the variance of up to 50% and an error in plate count of up to 100% (Foley and Dorsey, 1983).

A number of important measurements may be taken from the chromatogram. The simplest of these is the retention volume \(V_i\), the volume at which the peak maxima occurs.

---

\(^1\) The term tailing is used to describe skewed peaks when the end of the peak extends significantly beyond the end of an unskewed Gaussian peak. Fronting is the reverse situation where front of the peak is skewed.
Another measure of retention is the capacity factor $k_i$ (see figure 1.3). This is the retention volume relative to the retention volume of the unretained peak ($V_0$) (Schoenmakers, 1986).

$$k_i = \frac{V_R - V_0}{V_0} = \frac{V_{Ri}}{V_0} \quad (1.4)$$

The retention of any particular peak may also be measured relative to any other peak using the relative retention parameter.

$$\alpha_{ji} = \frac{V_{Rji}}{V_{Ri}} \quad (1.5)$$

where the subscript $i$ represents the first eluting peak and $j$ the last eluting peak of a pair. $\alpha$ is the parameter most directly related to selectivity and is often itself called selectivity.

It may be expressed in terms of capacity factor:

$$\alpha_{ji} = \frac{k_j}{k_i} \quad (1.6)$$

1.2.1 Causes of column inefficiency

The next section describes the causes column inefficiency ie. effects which give rise to skewed peaks.

1.2.1.1 Extra column effects

The length and volume of the tubing between the column end and the detector is of great importance. If the volume is too great this will lead to extra column band broadening. Band broadening is characterised by broad peaks with a lower than expected height. When peaks elute closely band broadening may give rise to a single peak, probably of an irregular shape, instead of a number of resolved peaks. Band broadening, whatever its cause (it may be caused by an increased sample size or an injection valve), may be expressed as a volume. The total effective volume may be expressed as:
\[(V_v)² = (V_p)² + (V_{ec})²\]  \hspace{1cm} (1.7)

where \(V_v\) is the effective total volume and \(V_p\) and \(V_{ec}\) are the volumes due to the particles and extra column effects respectively. If there are other contributions to band broadening such as an unsuitable injection valve or detector these may be included in the equation:

\[(V_w)² = (V_p)² + (V_{ec})² + (V_{inj})² + (V_{det})² + \ldots\]  \hspace{1cm} (1.8)

From the volume, \(V_v\), the peak variance \((\sigma^²)\) may be calculated which for a Gaussian peak is equal to \(V_v/4\).

1.2.1.2 Intra Column Band Broadening

The various sources of band broadening which occur within the column are discussed in the following sections. The relative importance of these is dependent upon the fluid velocity. The concept of reduced velocity (defined as \(v = ud_p/D_s\) where \(d_p\) is the particle diameter and \(D_s\) is the diffusion coefficient of the solute in the mobile phase) has been used to determine the overall effect of these causes of band broadening by relating them to the reduced plate height (reduced plate height is defined as \(h = H/d_p\). See section 1.3.4 for a definition of plate height). Various equations been developed to describe the relationship between reduced plate height and reduced velocity since Van Deemter et al, 1956 originally suggested such a relationship. These incorporate additional terms or take into account additional experimental data and these have been reviewed by Jonsson, 1987. One of the most widely used is that of Knox (Knox, 1986):

\[h = A\frac{1}{v^3} + B\frac{1}{v} + Cv\]  \hspace{1cm} (1.9)

Each of the three terms in equation 1.9 correspond to band broadening due to different physical mechanisms. The first of these is the contribution to band broadening caused by the tortuous nature of flow through a packed bed. The second describes the contribution to band broadening due to axial molecular dispersion. The final term describes the contribution from slow equilibration between the mobile and stationary phases.
1.2.1.2.1 Eddy Diffusion
Eddy diffusion is caused by the different microscopic flow-streams that flow between the stationary phase particles. As a result sample molecules take different paths through the column bed, depending on which flow-stream they are within. The size of band broadening is thus dependent on the path length and the width of the path: a narrow path leads to a high velocity while a wide path leads to a low fluid velocity.

1.2.1.2.2 Mobile Phase Mass Transfer
This refers to band broadening caused by the flow distribution in a flow-stream ie. a sample molecule close to a stationary phase particle will travel a shorter distance compared at a molecule at the centre of the flow. Again this results in band broadening.

1.2.1.2.3 Stagnant Mobile Phase Mass Transfer
With porous packings, the mobile phase contained within the pore is stagnant. Sample molecules will diffuse into and out of the pores. Due to a variance in the time spent by sample molecules in the pores, band broadening will occur since solute molecules of the same compound will take different times to pass through the column.

1.2.1.2.4 Stationary Phase Mass Transfer
If a sample molecule diffuses into a pore there is a probability that the molecule will penetrate the stationary phase or become attached to it in some way. Molecules which do so will move down the column more slowly in comparison to those which diffuse in and out of the pores without becoming attached.

1.2.1.2.5 Longitudinal Diffusion
Sample molecules in the stationary phase tend to diffuse randomly in all directions, whether the carrier solvent is moving or not. This will cause band broadening in addition to the effects mentioned above. This is not usually significant except at lower flow rates.
1.2.2 Band-broadening and Scale-up

Typically when a process chromatographic separation is being developed initially laboratory and pilot scale separations will be carried out to determine the correct operating conditions for full scale operation. Scaling up the separation whilst keeping the separation performance equivalent to the smaller scale operation requires knowledge of the contributions to band broadening discussed in section 1.2.1 above. Generally separations are scaled up by increasing the column cross-sectional area in proportion with the increase in flow rate desired whilst keeping the bed height constant (providing that the media will not collapse at wider column diameters due to lack of wall support). The ratio of media volume to gradient volume is also kept constant (Sofer and Nystrom, 1989). This keeps the linear flow rate (and reduced fluid velocity if the same media is used at both scales) constant producing the same number of plates available for the separation. Thus the separation within the column should be equivalent at both scales. However other factors may affect the performance at the larger scale due to additional band-broadening effects. An example of such an effect is the flow distribution system of larger scale equipment which is often less efficient in giving even distribution across the bed. Greater axial dispersion will also occur in the bed and band broadening will also be greater in the larger column end piece. It is also important to check that the distances between the column outlet, monitors, and fraction collectors give contributions which are equivalent, in terms of their band broadening, to those at the smaller scale. If there is a significant difference in band broadening between the two scales then it may be necessary to compensate for additional band broadening at the larger scale by reducing one of the factors contributing to the total peak widths. Two possible parameters which may be altered could be flow rate (i.e. reducing the contribution to band broadening from the intra-particle mass transfer resistances) or by reducing the sample size (i.e. reducing the contribution from overload). However both of these will reduce the productivity of the process which may well be undesirable.
1.3 Optimisation and Quality Criteria

Many criteria have been developed to describe the quality of a chromatographic separation. They use information obtained from a chromatogram such as peak maxima, peak minima, and points of inflexion to give a numerical value for the quality of the separation of mixture components. There are two basic types of criteria; those which compare the separation of a binary mixture and those which describe the separation of a multicomponent mixture.

1.3.1 Binary system criteria

The simplest and most well known measure of the separation of two peaks in a chromatogram is resolution, $R_s$. This can be defined in two ways (Snyder and Kirkland, 1979):

$$ R_s = \frac{2 (V_j - V_i)}{(w_j + w_i)} $$

where $V_i$ is the retention volume of component $i$ and $w_i$ is the peak width of the peak $i$.

Resolution is a convenient parameter to describe peak separation since it relates the mean peak width to the difference in retention volumes - two parameters which determine whether peaks overlap. A resolution, $R_s$ of 2.0, (ie. $w_j + w_i = V_j - V_i$) corresponds to complete peak separation.

An alternative method of calculation is given by:

$$ R_s = \frac{1}{4} \left( \frac{\alpha - 1}{\alpha} \right) \cdot \left[ \frac{k_j'}{1 + k_j'} \right] \cdot N^{1/2} $$

where $\alpha$ is the relative retention or selectivity factor between peak $i$ and peak $j(k_j' / k_i')$, $k_j'$ is the capacity factor of peak $j(t_j - t_i)$, and $N$ is the plate number of the column.

The relative retention $\alpha$ is itself a measure of separation between two peaks but whether the desired separation is achieved depends on the plate number of the column. $R_s$ and $\alpha$ are measures based on chromatography theory as they involve factors related to the retention of the components in the stationary phase. There are however other criteria which simply use properties of the chromatogram and are empirical. Examples of this type of criteria are valley-to-peak ratio, $V_v$,
and peak separation $P$ (Debets, Bajema and Doornbos, 1983). These measures are not affected by peak asymmetry. Peak-to-valley ratio is defined as the height of the valley between two peaks divided by the height of the smaller of the two peaks (see figure 1.4). If peaks are completely overlapped then $V$ is considered to be unity and if the peaks are resolved to the baseline then $V$ will be zero.

The peak separation, $P$, is defined as the depth of the valley between two peaks, below a straight line connecting the two peak maxima, divided by the height of the straight line above the baseline at the position of the valley (see figure 1.4). $P$ is said to be zero when the two peaks overlap completely and unity if they are resolved at the baseline. All of these criteria may provide a satisfactory description of the separation of two adjacent peaks. They have however been used, in the past, to describe multicomponent separations. This is achieved by calculating the value of the criterion for each pair of adjacent peaks. The best separation is considered to be the separation where the most poorly separated pair of peaks gives the highest value of the criterion. The usefulness of this method is questionable since these two component criteria only use information about adjacent peaks and information about the degree of separation of other peaks in the rest of the chromatogram is completely ignored (Debets, Bajema and Doornbos, 1983).

1.3.2 Multicomponent Criteria
Better solutions to the problem of how to judge the quality of a multicomponent separation have been suggested by Giddings (1960). In these criteria the separation of all pairs of peaks is taken into account. A list of multicomponent criteria developed in the last twenty years is shown in Table 1.1 (Debets, Bajema and Doornbos, 1983). Most of these are summations of two component criteria which are carried out in such a way that strongly overlapping peaks give a greater contribution to the criterion than well resolved peaks. Modifications have also been made to these criteria using weighting factors to enhance the separation of certain peak pairs and time factors where the time for the separation is taken into account.
Table 1.1 Quality Criteria for chromatograms

1. Total Overlap

\[ \Phi = \sum \exp(-2R_i) \]

\(R_i\) is the resolution of the ith pair of peaks

2. Chromatographic response function

(a)

\[ CRF = \sum_{i=1}^{k} \ln(p_i) \]

(b)

\[ CRF = \sum_{i=1}^{k} p_i \]

\(p_i\) is the peak separation of the ith pair of peaks

(c)

\[ CRF = \frac{1}{t} \sum_{i=1}^{k} p_i \]

\(t\) is the separation time

(d)

\[ CRF = \sum_{i=1}^{k} w_i \ln(p_i) \]

\(w_i\) is the weighting factor for the ith pair of peaks
(e)

\[ CRF = \sum_{i=1}^{k} \ln \left( \frac{p_i}{p_o} \right) - \alpha(t_o) \]

\( p_o \) is the desired value of peak separation. \( t_l \) is the retention time of the last peak.

(f)

\[ CRF = \sum_{i=1}^{k} \ln \left( \frac{p_i}{p_o} \right) + \beta(t_{max} - t_i) \]

\( t_{max} \) is the maximum allowed retention time.

3. Chromatographic optimisation function

\[ COF = \sum_{i=1}^{k} w_i \ln \left( \frac{R_i}{R_d} \right) + \beta(t_{max} - t_i) \]

\( R_d \) is the desired resolution.

4. Informing power

(a)

\[ P_{inf} = \sum_{i=1}^{k} \log(S_i) \]

(b)

\[ P_{inf} = \frac{1}{t} \sum_{i=1}^{k} \log(S_i) \]

\( S_i = (\Omega_{i-1,i} + \Omega_{i,i+1})^{1/2} \) where \( \Omega_{i-1,i} \) is the fractional overlap between peak \( i-1 \) and peak \( i \). \( t \) is the time or chromatography.

5. Separation number

\[ SN = \sum_{n=1}^{k} \log(p_n)^2 \]

\( p_n = x_n(2x_n - y_n)^{-1} \) where \( x_n \) is the real value or predicted value; whichever is larger; \( y_n \) is the smaller.
6. Product resolution

\[
\text{Prod.} R_s = \prod_{i=1}^{k} R_i
\]

\(R_i\) is the resolution of the \(i\)th pair of peaks

Debets et al (1983) reviewed these criteria by simulating chromatograms with overlapping Gaussian peaks produced by a computer program. Each of the criteria were compared with resolution for a range of peak separations. To find the peak maxima and minima a peak search routine was employed together with a standard integration package. They found that when peaks became resolved to better than \(R_s = 0.5\) the criteria improved quickly until baseline resolution was achieved. The criteria then levelled off to a final value. These criteria therefore provide a better estimate of quality than does resolution as the values of the criteria have finite limits unlike resolution which may become infinite, with increasing peak separation. Once two peaks are resolved at the baseline further separation is to no advantage and in fact on an economic basis it could be considered that further separation is disadvantageous since it will lengthen the separation time and so cost more with no increase in purity.

Debets et al (1983) also tested multicomponent criteria. Initially they tested these using a binary mixture and found that separation number was not a good measure as the response depended strongly on the way in which the data was scanned. Undesired perturbations occurred when two maxima were detected on the chromatogram. Also when two peaks were baseline resolved more irregularities occurred. Further tests were undertaken with four components and a varying solvent composition. The retention time of the components was assumed to change linearly with changes in solvent composition. They calculated each criterion for each selected solvent composition and plotted the results. All the criteria found two different solvent compositions for which the separation was an optimum but in these criteria time was not accounted for. Therefore in all cases the second of the two compositions was superior as it involved the shortest time.

All criteria required some prior information about the number of peaks if they were to give sensible answers. If such information was not given then optimum values were obtained corresponding to cases where the peaks strongly overlapped. All criteria require constraints or
corrections in their calculations. Also when peaks overlap strongly or when peaks become more than baseline resolved it may actually become impossible to calculate the criteria. None of the criteria were able to give an absolute value for the quality of the separation since a good separation of one pair of peaks may be compensated for by a bad separation of another pair of peaks. Weighting factors have been applied but are only useful if the peaks can be identified. When the retention time of components change with changing solvent composition peaks can only be identified with more complex systems.

In conclusion all the multicomponent criteria relate directly to chromatographic separation but lack any theoretical basis. Also if the elution order of the peaks were to change then the results become intractable. Identification of peaks may also be necessary to determine either which is the critical pair of peaks or so that the correct weighting may be applied when a multicomponent optimisation criterion is calculated.

1.3.3 Criteria for Non-Gaussian Peaks
Dose and Guiochon (1990) investigated the need for a measure of peak overlap for non-Gaussian peaks. This type of peak occurs when slow reaction kinetics or non-linear isotherms occur or when preparative loads are used where the column becomes overloaded. This may cause component bands to be compressed into small elution volumes which tend to displace other components. This is done to maximise local competition for sites and so minimise band overlap and leads to increasing productivity with a less than proportional loss of purity. In analytical chromatography the peaks are sufficiently close to a Gaussian shape to use criteria as described above in sections 1.3.1 and 1.3.2. However at larger scales of operation because of the effects stated above the peaks deviate significantly from Gaussian form. This means that Gaussian measures such as variance $\sigma^2$, resolution $R_s$, and even retention time become misleading. Dose and Guiochon (1990) suggested that a measure for non-Gaussian peaks should have the same desirable properties that resolution parameter $R_s$ possesses. These properties include: invariance such that the measure is not dependent of the graphical scales, uniqueness such that neither peak need be labelled first or second, continuity such that the measure and its derivative have no discontinuities relative to the location or widths of the peaks. They
proposed, \( Q \), as a general measure for overlap which is related to the quantum mechanical overlap integral and can be related to any peak shape. Equations were developed for Gaussian, Lorentzian, rectangular and triangular peaks. Chromatographic data was simulated and values of \( Q \) corresponding to degrees of peak overlap were calculated for each peak type. \( Q \) was found to be zero if the peaks were not overlapping and unity if the peaks overlapped and were of the exactly the same size and shape. Values in between represent different degrees of overlap. \( Q \) was defined by an equation of the form

\[
Q = \frac{\left[ \int C_A \frac{dC}{dt} \right]^2}{C_A^2 \int \frac{dC}{dt} \frac{dC}{dt} dt}
\]

Where \( C_A \) and \( C_B \) are the concentration of the two components at a given point on the chromatogram. The elution time is described by \( t \).

Preparative chromatography is often used in a sequential mode with each stage providing enrichment, i.e., partial resolution or reduction of peak overlap. In analytical chromatography one is concerned with the almost total resolution of all components for the purpose of quantitation or identification. Preparative chromatography therefore requires a different criterion from resolution since it is possible for a separation to have zero resolution but have a reasonably good fractional overlap. This is demonstrated by figures 1.5 and 1.6 which have equal overlap but figure 1.6 has a resolution, \( R_y \), of zero. This value of resolution would be of no use in analytical chromatography since a single peak would be observed and the amounts of the individual components could not be determined. However in preparative chromatography this may be acceptable since component enrichment has occurred. Fractional overlap is difficult to use since the peak shape must be known before applying the relevant formula and peak shapes are not available from a chromatogram. A process known as deconvolution must therefore be carried out in order to determine peak shapes.

Several alternative techniques have been developed to deconvolute chromatograms. The simplest but least accurate of these which are often used in chromatographic integration packages are the perpendicular drop method (the use of a vertical line drawn from the valley between the two peaks under examination to divide the chromatogram area between the two
components) and tangential skim method (where tangential lines are drawn on each peak at the points of greatest slope to represent each peak as a triangle). Papas and Tougas (1990) investigated these techniques and found that in general they were inaccurate but consistent for any given set of peaks. However, these methods are not suitable for determination of fractional overlap since they do not attempt to determine what peak functions the chromatogram consists of and so are only useful for estimating the amounts of components within the chromatogram. Several more sophisticated methods have been developed for chromatogram deconvolution. Curve fitting is mathematically the simplest of these. Cai and Wu (1991) and Jung and Shin (1986) used an exponentially modified Gaussian function to describe each peak function within a simulated chromatogram. The parameters for the peak functions were determined by minimizing the difference between the chromatogram and the sum of the exponentially modified peak functions used to describe the peak functions using an optimization algorithm. Vaidya and Hester (1984) applied a similar method but using a more general empirical peak model (the general exponential function – see section 2.2). In general curve fitting has the advantage of determining both the amounts of each component eluted and how the peaks overlap to produce the chromatogram. Other methods exist such as Fourier Transform techniques (Nelson, 1991), Kalman filters (Hayashi, Shibazaki and Uchiyama, 1987) and relaxation based iterative methods (Crilly, 1992). In general these methods only provide information regarding the position (i.e. retention volumes) and amounts of material eluted rather than how the peaks overlap and so are of no use in the determination of fractional overlap. The overlap of peaks may also be determined using diode array detectors (Fell et al, 1983) which monitor at several wavelengths simultaneously. However, these require expensive detectors in addition to the computational requirements of curve fitting. (Deconvolution of chromatograms is discussed further in Chapter 2.)

1.3.4 Column Efficiency

Many authors have suggested the concept of the chromatographic separation efficiency to monitor column condition. Biddlingmeyer and Warren (1984) reviewed the various methods available for calculating column efficiency. They stated that column efficiency alone cannot give
any indication of whether the column is able to separate a given mixture of compounds and that column efficiency is only a measure of how well the column has been packed and the degree of kinetic band broadening. Therefore it can be seen that efficiency and performance are two separate terms. Performance being the ability to carry out a particular separation - high efficiency alone will not guarantee this.

Several parameters affect the determination of column performance. These include eluent composition, viscosity, velocity, temperature, column length, packing type, packing size and measurement and calculation method chosen. It is therefore vital to state under what conditions efficiency has been calculated, as well as what method has been used. Many methods make no effort to remove any contribution obtained from the liquid chromatography system and so the value for efficiency obtained is the total efficiency for the whole system. One measure of particular interest is $H$ or HETP (height equivalent to a theoretical plate) as it adjusts for the length of the column. HETP (or $H$) reflects the ratio of column efficiency to column length whilst the other uses the inverse ratio. (Generally it is not used as chromatographers are conditioned to prefer a high value of $N$ rather than a low value of $H$!)

1.3.4.1 Measurement and calculation methods

In analytical chromatography peaks are often assumed to be Gaussian in shape and because of this $N$ is defined as follows

$$N = \frac{V_r^2}{\sigma^2} \quad (1.13)$$

Where $V_r$ is the retention volume and $\sigma^2$ is the variance in volume units. $\sigma^2$ can be written in terms of the peak width divided by a constant. The value of the constant depends upon where the width is measured on the peak, ie. 4%, 50%, or 100% of peak height. This can be incorporated into the above equation

$$N = \frac{a V_r^2}{w^2} \quad (1.14)$$

Where $a$ is a constant, the value of which depends on the percentage of the peak height at which the width, $w$, is measured.
Several approaches are available for the calculation of peak variance. The following is a list of different calculation methods: inflection method, width at half height, tangent method, height/area method, moment method and asymmetry-based methods. Biddlingmeyer and Warren (1984) tested each of the calculation methods for accuracy and consistency. The moments and asymmetric methods were found to be the most accurate. The simplest methods, ie. measuring the width at half height and measuring the width using tangential lines projected to the base line, were found to be the least accurate, but were the easiest to calculate.

1.3.4.2 Moment Method
In this method no assumptions are made about the peak shape. The characteristics of the peak are expressed in terms of statistical moments. Peak area is the zeroth moment, the first moment is the mean (for Gaussian peaks this is the peak maximum) and this occurs at the centre of mass. The second moment is the peak variance and the third and fourth moments are measures of skewness. The second moment may therefore be used in place of peak variance to calculate $N$.

1.3.4.3 Asymmetric Methods
Two types of asymmetric methods are available. The simpler of the two uses an empirical ratio of the peak widths either side of the peak maxima measured at 10% peak height. This ratio has the advantage of satisfying the chromatographer's intuitive idea of peak asymmetry and is easy to measure. It however lacks any theoretical basis, but this should not preclude its use.

The other method uses an exponentially modified Gaussian peak to represent peak skewness. This function consists of a Gaussian function combined with an exponential decay function. The total variance is:

$$\sigma^2_{\text{peak}} = \sigma^2 \sigma^2 + \tau^2$$

(1.15)

$\tau$ and $\sigma$ are not as easy to use as the other ratio and requires the use of a microcomputer for their calculation but have the advantage of being a fundamental measure of peak asymmetry.
1.3.4.4 Exponentially Modified Gaussian Methods

The use of exponentially modified Gaussian peaks has been investigated as a more accurate description of chromatographic peaks (Jeansonne and Foley, 1992, Foley and Dorsey, 1983, Grushka, 1972, and Yau, 1977). The exponentially modified Gaussian model may be justified since it is known that intra and extra column effects have the effect of skewing peaks causing the changes in variance produced by exponentially modified Gaussian peak functions. The use of the correct model is important - the use of the incorrect model may lead to great inaccuracies in plate counts and variances. The most frequently used formula for the number of plates for real peaks is:

\[ N_{\text{str}} = \frac{41.7(V_f/V_{f0})^2}{B/A + 1.25} \]  

(1.16)

This formula includes the asymmetry (using the B/A asymmetry ratio as shown in figure 1.7) of the peak in the calculation of the number of theoretical equilibrium stages (i.e., plates) corresponding to the operation of the column.

1.3.4.5 Choice of Method

Since there are many different methods of calculating efficiency which do not always give equivalent values it is important to decide why efficiency is to be measured before choosing which method to use. Generally methods which are least sensitive to peak asymmetry are the best choice but if the only objective is to monitor the efficiency of a column over a period of time then any of the methods may be useful. If comparisons are to be made with columns of different size or different packings then a more accurate method which gives a consistent result is required. With any evaluation of column efficiency the capabilities and limitations of the calculation methods used must be taken into account. When evaluating reported efficiencies contained within literature or commercial information it is difficult to make any direct comparisons due to the lack of any standardisation of experimental conditions.

1.3.5 Peak overlap detection by multi detector methods

The most frequently used detector in chromatography is the ultraviolet detector. Most of these detectors operate at fixed wavelength and
measure the absorbance of the eluent. Other more advanced UV detectors can detect at more than one wavelength - both dual and multiwavelength are available, the most sophisticated being those capable of scanning over a spectrum of wavelengths. These detectors are known as photodiode arrays. UV detectors are relatively insensitive to flow rate or temperature changes but the choice of mobile phase is restricted as they must be transparent at the wavelength at which the detector is operating (Hamilton and Sewell 1977).

Another type of detector is the refractive index detector. These work using a differential technique in which the sample's refractive index is compared with that of the eluent. Any substance which has a refractive index significantly different from the eluent may be detected but the method is very sensitive to changes in eluent and so is unsuitable for gradient elution unless solvents are chosen with identical refractive indices. It is also very temperature dependent (Engelhardt 1979).

Compounds may be detected using fluorescence. The process stream containing the substance to be detected is first excited by UV radiation and monitored for emission at another wavelength. Fluorescence detection is very specific but fluorescence may be suppressed or quenched by the presence of contaminants. The severe effect that such contaminants have on this detection system means that careful selection of column conditions must be carried out.

1.3.5.1 Difference and Ratio Methods

More than one detector may be used for detection. This has the advantage of providing more information about the separation eg. if the separation contains one substance which fluoresces then the degree of overlap may be measured by comparing the chromatogram obtained from UV with that from fluorescence.

Drouen, Billiet, and de Galan (1984) have used the absorbance ratio at two different wavelengths to provide more information about overlapping peaks. They defined the ratio, RAT, in the following way;
\[
\text{RAT} = \frac{A_1}{A_2} \quad \text{when } A_2 \geq A_1 > \text{threshold}
\]
\[
\text{RAT} = 2 - \frac{A_2}{A_1} \quad \text{when } A_1 > A_2 > \text{threshold}
\]
\[
\text{RAT} = 0.0 \quad \text{when } A_1 \text{ and } A_2 \leq \text{threshold}
\]
\[
\text{RAT} = -0.1 \quad \text{when } A_1 \leq \text{threshold and } A_2 \geq \text{threshold or}
\quad \text{when } A_2 \leq \text{threshold and } A_1 \geq \text{threshold.}
\]

Where \(A_1\) and \(A_2\) are the absorbances measured at the two different wavelengths at any given point in the chromatogram.

Initially they tested the ratio method using simulated fully-resolved peaks. They found that the ratio produced block shaped responses when the ratio was plotted against the elution time. They also showed that the ratiogram provided information when the simulated peaks were overlapped. When the two peaks were fully resolved two blocks were shown on the ratiogram corresponding to each of the peaks. Once peaks were overlapped the blocks combined with a step change from the first value to the second. As the degree of overlap increased the change from the first to the second became smoother until the two peaks coincided exactly when a single block occurred. These results were for perfect data i.e. no noise, or baseline offset, or time delay between signals. Baseline offsets may occur because of incompletely corrected background signal and may be variable or constant. They have the effect of distorting the blocks into asymmetrical shapes. This effect can be reduced by increasing the threshold value but this has the effect of making the purity check of the peak more difficult. Any distortion of the ratiogram caused by time delays may be removed by adjusting the detectors so that they are measuring at the same wavelength and determining the time delay between the two by comparing chromatograms. This time delay can then be taken into account. Tailing may also distort the ratiogram and lead to misinterpretations e.g. more components appear than are actually present. Again clearer data may be gained by increasing the threshold value but this reduces the amount of information available. The ability of this method to identify overlapping peaks also depends on the difference between the RAT of the two different peaks. The greater the difference the simpler the recognition of overlap. Drouen, Billiet and de Galan (1984) found that if the difference in the RAT of two components was less than 0.05 then the peak separation must be greater than 2\(\sigma\) for overlap to be identified. This is still better than single wavelength detection. The
selection of wavelengths for multichannel monitoring is important and it
should be made to gain as much information as possible. Mathematical
techniques (key set factor analysis (KSFA)) have been developed to
determine the optimum wavelength sets (Warren, Bidlingmeyer and
Drouen, Billiet, and de Galan (1985) extended their previous work on dual
wavelength detectors to multiwavelength detectors such as diode array
detectors which allow rapid scanning and a more detailed analysis of
overlapped peaks to be obtained by measuring the absorbance spectrum
of the eluent. This in turn gives more accurate analysis and
identification of severely overlapping components. However with proteins,
since their absorbance spectrums are often very similar, diode array
method would be essential (Carr et al, 1988).
A factor analysis method has been developed for the resolution of
unresolved peaks (Sakema, et al, 1990). Using this method estimates
can be made of the elution profiles of components eluting at the rising
or tailing edge of a peak. With unresolved three component peaks, peak
area of the three components were determined to within 10%. Methods
have been developed (Marr, Seaton, Clark and Fell, 1990) using modern
mutliwavelength detectors for examining the homogeneity, identity and
purity of peaks, based on work for stopped flow conditions (Ostojic, 1974).
Krstulovic, Rosie and Brown (1976) used the ratio of peak areas to
identify chromatographic peaks which were fully resolved but purity was
not considered as the method was purely for identification purposes.
Other methods have been devised for analysis of chromatograms measured
on two wavelengths. Li and Arrington (1979) used the difference between
chromatograms for the identification of components as well as
quantification of components in overlapped peaks. They were able to find
a relationship between the amount of substance present and the depth of
the valley caused by the subtraction of the chromatograms.

1.3.5.2 Derivative Methods
Grushka and Monacelli (1972) developed a method of analysis using the
second derivative. They tested this method on overlapped simulated
Gaussian peaks and on real chromatographic data. The method relies on
mathematical knowledge about the Gaussian function. The second
derivative of the function has two maxima and one minimum which are positioned at \( x = x_0 \pm \alpha (3)^{1/2} \) (the maxima) and \( x = x_0 \) (the minimum). Where \( x_0 \) is the Gaussian peak's centre of mass and \( \alpha \) is the standard deviation. The ratios of the maxima and minimum are independent of the peak height. The ratios of the second derivative maxima and minimums were defined as follows:

- \( R_1 \): maximum on front of peak / minimum
- \( R_2 \): maximum on tail of peak / minimum
- \( R_3 \): maximum on front of peak / maximum on tail of peak

A graph of \( R_1 \) and \( R_2 \) was obtained for various peak height ratios of overlapping peaks onto which contours of resolution were drawn. Once curves such as these are produced it is then possible to identify peaks if the peak width are known since the second derivative ratios will give information about the peak height and resolution. This method has the advantage of not being too dependent upon the exact shape of the chromatographic peaks but it is necessary that the data is of a good and reproducible quality and so may not be of general use. The main disadvantage of this method is that calibration separations must be carried out to detect overlapped peaks and the extension of this method to multicomponent systems would require a large database of information about overlap of each peak with each other peak, eg. a combination of four peaks would require information about six permutations of overlapping peaks. Measurement of absorbance ratios would require much less information in comparison - the value of the ratio for each pure peak which could be gained from one separation with all peaks fully resolved. Another disadvantage is that a numerical differentiation must be carried out which may not be accurate depending upon the quality of the data. Differentiation also causes magnification of noise in the signal. The limits of this method have been found to be dependent on the peak height ratio and the relative separation (Grushka and Israeli, 1990).

1.3.5.3 Methods suitable for general use

Ideally the method chosen for judging the quality of the separation should be general enough to be applied to separations based on the same mechanism and on other mechanisms. Fluorescence detection is heavily dependent upon properties of the components within the mixture to be separated and therefore will not be considered. It is also subject to
quenching. The most versatile method appears to be measurement of ultraviolet absorbance at two or more wavelengths as this allows a number of different methods of analysis:

(i) ratios of area
(ii) ratios of absorbance
(iii) slope analysis
(iv) difference analysis

It also allows the concentration of the solution in the detector flow cell to be determined using the Beer - Lambert law (see section 2.3.9).

The difference in absorbance spectra between different proteins is generally not that great and so a photodiode array would be required, together with superimposition techniques. (Carr et al, 1988 and Frank, Braat and Buine, 1987)

A combination of these methods could be used to gain as much information about the chromatogram as possible. For example ratios of absorbance and areas could be used for peak identification and difference analysis for quantification.

In addition to being general in terms of separation and mechanism any method should also be applicable to both overload and analytical conditions. Of the methods of analysis studied only the fractional overlap parameter $Q$ (Guiochon 1990) is applicable to non-Gaussian peaks.

### 1.3.6 Peak Identification

In order to judge the quality of a particular separation it is necessary to identify which peaks are the most important and from this information to assess the performance of the separation using a criterion listed above or another method. There are several methods for identification of peaks. They vary in complexity and the most appropriate one is dependent on the nature of the chromatographic data which is to be processed.

The simplest method is to compare the elution times of components peaks with a record of elution times from an ideal separation. This simple method has its disadvantages as elution times are subject to variations including variations in flow-rate, column temperature, injection techniques and column packing quality. The last of these problems is important as it is hard to pack a column consistently well. A very simple way to
overcome this is to normalise the elution times. Chilcote (1974) reviewed various methods of normalisation and tested normalisation techniques for gas chromatography. Elution times were found to be very reproducible even with a varying carrier gas flow rate. Curves were developed with the aid of one or two reference compounds to identify peaks of n-alkane from their elution times. Temperature however caused a big variation in elution time for this type of chromatography. Chilcote (1974) suggested that the ratio of absorbances at two different wavelengths would aid the identification of the reference peaks. The location and number of reference peaks was also said to be important. With fewer reference peaks identification was speeded up but the reference peaks needed to be close enough to the regions on the chromatogram which were of interest. It was also found that the reference peaks needed to be located in a position where the elution time varies by only a minimal amount. If the elution times vary greatly this type of identification system may become unsuitable. This may happen for example with the separation of proteins by HPLC or ion exchange. It occurs to a great degree with the separation of vitamins by HPLC.

Otto, Wegscheider, and Lankmayr (1988) suggested a peak tracking approach using fuzzy theory. This has the advantage of being able to identify overlapping peaks. Fuzzy theory was originally applied to component identification by spectral analysis which was used in order to account for the uncertainty of the data. In fuzzy theory the data is taken to be blurred curves of fuzzy sets. The characterisation of the fuzzy set is achieved by defining a membership function for the measurement of peak area with a range zero to one ie. \( m(x):Y \rightarrow [0,1] \) where \( Y \) is the absorbance set and \( x \) is a point on the \( x \) axis. The membership function used is of the form;

\[
m(z) = \left[ 1 - z^2 \right]^+ \text{ where } \left[v\right]^+ = \max(0,v) \text{ ie. the positive part of the function only.}
\]

The membership function is often described using the left-right representation. The LR representation of data is written:

\[
m_z(y;x) = \begin{cases} 
L \left( (a(x) - y)/a(x) \right) = L(x) \text{ for } y \leq a(x) \\
R \left( (y - a(x))/a(x) \right) = R(x) \text{ for } y \geq a(x) 
\end{cases}
\]

\[(1.17)\]
\( x \) is the variable for the channel i.e., absorbance etc. \( a_i(x) \) and \( b_i(x) \) are the left and right spreads of the membership function, respectively. \( a(x) \) represents the measured peak area. If the left and right hand spreads are symmetrical then the membership function may be written as a single function as in Chapter 3.

A membership function may also be defined for the elution rank of the peak which uses the difference between the trial and reference elution ranks to assess the similarity of peaks. An overall membership function may be determined using peak area data (determined from data obtained by monitoring at several wavelengths if available) and elution rank data. By combining membership functions in this way the identification is made more reliable as the combined data is less fuzzy.

Initially a reference run of chromatographic data is read into the computer system's memory. Each chromatographic separation will be compared to this set of data so it is important that this is a representative separation with all components fully resolved in their usual positions. Comparisons of peaks within the trial and reference peaks are made by determining to what degree the overall membership functions of the two peaks overlap. The size of the overlap is used to determine a matching criterion which may be used to judge whether the two peaks under analysis match one another. If peaks in the trial data set remain unmatched after comparison with all peaks in the reference data set then they may be compared to combinations of peaks in the reference data set. This allows for the possibility that peaks in the trial chromatogram may consist of more than one reference peaks due to peak overlap.

Otto, Wegscheider and Lankmayr (1988) developed a computer algorithm for comparing two chromatographic separations using fuzzy set theory. It was shown to work successfully, identifying peaks in a series of analytical separations of a single mixture operating under a variety of different separation conditions. This shows the method's ability to cope with variations in separation output. More recently this work has been further tested and applied to more than one wavelength (Lankmayr et al, 1989). Improvements were made to the algorithm to improve accuracy and reliability.

This method of identification is discussed further in chapter 3 but it can be seen that it is a sophisticated method which has been successfully applied to a number of different analytical scale separations.
1.4 Controllable parameters and modes of operation

In order to maximise productivity the column operating conditions should be kept under review to ensure that the system is operating under optimum conditions. However as well as productivity the purity must be kept in mind as this forms an important part of the separation quality.

1.4.1 Column Loading

The manner in which the column is loaded can have an effect on both the production rate and the resolution of the product from the contaminants in the mixture being separated. The key parameters which may be varied during the loading stage to control the separation performance are discussed in the following sections.

1.4.1.1 Flow rate

Flow rate may have an effect upon the productivity of the column. For example in ion exchange chromatography a high flow rate may be used as the equilibrium is generally chosen so that binding of the component or components of interest is highly favoured (Sofer and Nystrom, 1989). A relatively high flow rate may then be used without any significant loss of material. If however the column is equilibrated so that the components only just bind e.g. proteins close to their isoelectric point or affinity columns where the kinetics of the binding are slow, then the flow rate may have to be low in order that all the product material is bound. Obviously the higher the flow rate the greater the productivity but this must take into account the loss of any material due to the above reasons.

1.4.1.2 Analytical and Preparative Chromatography - sample size

Analytical chromatography and preparative chromatography differ in their objectives. Analytical chromatography is carried out with the objective of obtaining as much information as possible i.e. the process is optimised so that ideally each component is fully resolved. The aim with preparative chromatography is to maximize the production rate of a given compound or compounds at a specified purity within economic constraints. Full resolution of all the components is therefore not necessary — only those components which are to be produced need be resolved and this may only require partial resolution depending upon the purity criteria.
Therefore concepts used in analytical chromatography should be used carefully i.e. criteria such as resolution should be adapted to take these factors into account and are only applicable to the purity aspects unless processing time is incorporated into them. Probably the most important criterion in preparative chromatography is cost but since this is difficult to quantify it may be assumed to be proportional to the throughput. Knox (1986) stated that optimization of performance in preparative liquid chromatography implies seeking conditions which give the maximum throughput of material with specified purity and at a given yield within economic pressure drop limits. Guiochon and Katti (1987) suggested that it is not necessary to specify a yield if throughput is also specified since a yield of less than 100% is acceptable if the loss of material is compensated by an increase in production.

In order to maximise production preparative columns are operated under overload conditions. Overload refers to the amount of substance loaded - whether it is in terms of concentration or volume. In analytical chromatography only a small amount of the column capacity is used as the main aim is to fully resolve all components. Consequently in analytical chromatography it can be considered that individual components do not interact with themselves to any significant degree. Since preparative chromatography aims to increase the production rate it is therefore desirable to use as much of the available column capacity as possible.

1.4.1.2.1 Volume Overload

With volume overload (Guiochon and Katti, 1987) a large volume of material is loaded onto the column (in comparison with analytical separations). If the concentration of sample is kept within the linear part of the Langmuir isotherm and so remains as linear chromatography then the chromatogram produced is similar in shape to those produced under analytical conditions i.e. the peak shapes are approximately Gaussian in form but are skewed, and are much taller and wider than under non-volume overload conditions. Very large volume may lead to peaks with flat tops (due to concentration effects rather than detector overload) and Gaussian-type fronts and tails (Nicoud and Colin, 1990 and Knox and Pyper, 1986). By changing the degree of volume overload it is possible to control the degree of overlap and thus the performance of the separation both in terms of the purity and productivity achieved.
1.4.1.2.2 Concentration Overload
Concentration overload occurs when a sample of a higher than usual concentration is loaded onto the column. When this is carried out the equilibrium between the stationary and mobile phase becomes non-linear, i.e. the non-linear portion of the isotherm. The band profile broadens and the profile becomes unsymmetrical even if there is no volume overload. For Langmuir isotherms the profile becomes triangular with a nearly vertical front and a slanted tail but for anti-Langmuir and S shaped isotherms the triangular shape is reversed. For S shaped isotherms beyond the point of inflexion the shape becomes more complex (Snyder, Knox and Antle, 1987). Because of these distortions of the peak shape it is not possible to use the multi-component criteria described previously to judge separation performance. Concentration overload enhances the performance of a separation by increasing the amount of material that may be purified in one separation, i.e. increasing the productivity of the separation.

1.4.1.2.3 Comparison of modes
Guiochon and Katti (1987) state that concentration overload always results in a larger production than volume overload but ideally both should occur together. No limit on the degree of overload was given as this is dependent on the separation being undertaken. Clearly using both overload techniques allows the separation to be enhanced by increasing the amount separated by increasing the concentration and volume of the sample injected onto the column.

At very high volume or concentration overload conditions then displacement effects may occur and this may be undesirable. This would occur if saturation of binding sites occurred and one component displaced other components due to a greater affinity to the stationary phase. However this could enhance the separation by the removal of unwanted material during the loading stage or produce an elution of a product during the loading stage.

1.4.2 Elution
Elution is achieved by changing the interaction between the components to be separated and the matrix onto which they are absorbed eg. with ion exchange chromatography the ionic conditions of the buffer in which the column is operating are changed which may mean changing the salt concentration. Each compound will have a different affinity to the matrix
and will therefore elute at a different time allowing separation to occur. The change appropriate to separate the compounds has to be optimised as does the manner in which the change is carried out.

1.4.2.1 Isocratic elution
This method involves a constant mobile phase concentration of the chemical added to cause elution. In general it causes broader peaks than with gradient elution (see below) as the change is a much slower one and is equivalent to a very shallow gradient. Once isocratic elution has been tried unsuccessfully as a method of separation then there is little that can be done to fully resolve the components except by changing the mobile phase composition - either the substance or the concentration.

1.4.2.2 Gradient Elution
As the name suggests gradient elution involves a varying mobile phase concentration. The variation may be linear or curved with respect to time, although linear gradients are more usual. The start and finishing values on the gradient are important. The initial concentration corresponds to the concentration at which components of interest just begin to elute and the final being the concentration at which all components of interest can be considered to have been eluted. Making the gradient steeper (i.e. decreasing the time for the mobile phase to change from the initial value to the final value) makes the components elute more closely i.e. decreasing the overall resolution, and conversely making the gradient more shallow increases the resolution of components but only to a certain limit beyond which increasing the shallowness only leads to band broadening. The use of gradients allows separation to be carried out more quickly than isocratic elution, but with a good degree of selectivity.

1.4.2.3 Step Change
It may be desirable to run a separation using two isocratic separations linked together - the first being used to separate one component at low concentration of elution buffer and the second is used to separate a component at the high concentration of elution buffer. When the change occurs all the components which are eluted by concentrations between the starting and end values will elute together and so the size of the step change must be carefully controlled. This method may be used to speed up a process where components at the start and end of the separation are of interest - gradient elution would give fully or nearly full
resolution of the compound in the middle of the separation which are not required and thus lengthen the separation time. Its disadvantages are that it is impossible to detect what is eluted when the step occurs as many components will be eluted together.

1.4.2.4 Multisegment gradients

Noyes (1983) described a multisegment gradient program which avoids the lengthy calculations involved with complete mathematical optimisation. The program was based on the visual interpretation of isocratic versus composition relationships for the separation of a number of phenylthiohydrantoin (PTH) amino acids.

Issaq, McNitt, and Goldgaber (1984) reported on a method for the optimisation of a multisegment gradient program for the optimum resolution of all pairs of peaks in a programmed solvent liquid chromatography chromatogram. Their procedure involved programmed solvent experiments performed as a series of linear gradients between two solvents A and B for a time \( t_g \) or series of linear gradients with a fixed \( t_g \) but a variable final composition of B in A. For each pair of adjacent solutes the gradient which gives the best separation is then selected and the different linear segments are combined to give a multisegment gradient. The method is designed to produce the maximum resolution of all components within the mixture. The mechanism for the construction of the multisegment gradient from the optimum gradients from the individual peak pair is not stated and it is therefore difficult to know whether the final result is optimal (Schoenmakers, 1986). It appears that the goal of equally well resolved peaks over the whole chromatogram cannot be achieved unless the elution of a particular pair of peaks is only affected by the segment of the gradient specifically designed for it. Unless the different solutes pairs are very far apart in the chromatogram the resolution of a pair of peaks is likely to be affected by preceding segments of the gradient. If the peak pairs were very far apart in the chromatogram this would be far from an optimum separation.

1.4.2.5 Flow rate

As with loading elution flow rate is important. If it is too fast then turbulence will cause mixing of bands as well as the possibility of bed collapse due to the type of gel being used (i.e. soft gels such as those made from agarose used in low pressure systems are able to withstand much lower flow rates than more rigid packings used in HPLC such as
silica based packings). Increasing the flow rate will cause a decline in resolution (Sofer and Nystrom 1988) and an increase in band broadening (see section 1.2.1) which adversely affects separation performance.

1.4.3 Strategies
To optimise the separation on the basis of productivity and purity it will be necessary to use a combination of the methods for elution and loading described above. During washing and regeneration steps the flow rate should be the highest that can be withstood by the packing used as these stages take 2-3 column volumes.

The use of non-uniform gradients will also enhance separation speed as gradient and isocratic separation for the whole process will take longer than a series of step changes followed by isocratic separations where the full resolution of compounds is required.

The correct strategy (ie. what type and degree of overload is appropriate and exactly what type of elution profile is the optimum) for a given separation will be dependent upon the composition of the mixture being separated and will be different for each separation encountered and would have to be determined by experimentation determining trade offs between different parameters and their effects upon separation performance.

1.5 Theories and Models of Chromatographic Separations
There are two major theories used to describe and predict chromatographic separations. These are the plate theory and the rate theory which are described below.

1.5.1 Plate Theory
Plate theory models depict the column as a series of well mixed equilibrium stages. The process can be treated as a series of ideal mixed flow contactors. This is an adaptation of models used to describe tray distillation columns and extended to packed columns. The model clearly does not closely describe the actual physical processes occurring within the chromatography column since the column is a continuous equilibrium process - not a series of discrete equilibrium processes. However under certain conditions, described below, it provides a good approximation to the true situation. If a linear isotherm is assumed and if the sample size
is small, the solution of the material balance yields a Poisson distribution function. A long column produces a result which may be approximated to a Gaussian function. This model was first introduced by Martin and Synge (1941).

It is assumed in these models that the actual continuous flow process in which the equilibrium is attained is a stage of finite height before the solution volume is subjected to a series of equilibria. With a linear isotherm a binomial function is obtained.

The plate theories have a number of obvious problems associated with them eg. their inability to predict the number of stages or effective plate height and the failure to predict how a change in operating conditions will change the column performance. Also the plate heights are not constant for all compounds and the theory is only applicable where the isotherm can be assumed to be linear.

1.5.2 Rate Theories

The rate theory is fundamentally a set of material balances for each component, together with the necessary boundary and initial conditions. No assumptions are made regarding local equilibria or distribution. The continuity equation for the mobile phase is given by the equation:

\[ D_x \frac{\partial^2 C}{\partial x^2} - v \frac{\partial C}{\partial x} - R = \frac{\partial C}{\partial t} \]  \hspace{1cm} (1.18)

Where \( D_x \) is the axial diffusion coefficient, \( C \) is the adsorbate concentration, \( R \) is the mass transfer rate, \( v \) is the linear velocity and \( x \) and \( t \) are the space and time coordinates.

Many different versions of this model have been developed which use various expressions for \( R \), depending upon the mass transfer resistances present.

The practical use of either of the models for prediction of the performance of multicomponent separations is limited at present because of the inability to predict the interactions between compounds and the effect this may have on the adsorption isotherm ie. the equilibrium between one compound and the stationary phase is altered by the presence of other compounds. Also the parameters for models are only applicable at constant conditions eg. rate constants used in the rate models will vary with temperature, pressure and mobile phase composition.
The prediction of these changes is difficult - especially if the components interact with each other as well as with the stationary phase. If the isotherm is non-linear as is often the case with affinity chromatography (Yang and Tsao, 1982) these models cannot be used as no solutions are currently available. Thus alternative methods of predicting performance of the separation must then be used.

1.6 Expert Systems
In situations where no accurate model exists, such as multicomponent chromatographic separations, an expert system is often suitable. Expert systems rely on a combination of rules and a database. Every expert system consists of two parts: the knowledge base and the inference engine. Expert systems are particularly useful where models or rules are not well understood or not fully developed as they enable changes to be made and the consequences analysed.

1.6.1 The knowledge base
The knowledge base is a database that holds the specific information about a particular subject. It contains a list of objects with their associated rules and attributes. The object characteristics may be as simple as 'has' and 'has not' a particular attribute or more complicated involving numerical values and often a combination of attribute types will be used to describe an object.
For example a chromatogram may be analysed as a series of objects corresponding to the peaks. Each possible peak in the chromatogram will either exist or not. If it exists then it will have a series of numerical attributes corresponding to its elution time, peak area and other parameters. A chromatogram may also be defined as an object containing other objects, i.e. peaks as well as the operating conditions used to perform the separation. Thus the objects contain information describing the situation under analysis by the expert system, in this case the details of the chromatographic separation.
Examples of rules which may be applied will depend on the purpose of the expert system. Systems developed for use with chromatographic data have typically been used to assess the performance of separations, either to judge the success of an optimisation, the condition of a column, or the selection of equipment to be used based on a chromatogram (i.e. a set of objects. Examples of expert systems using these types of applications are
given in section 1.6.3). Rules applied to chromatographic data to assess separation performance may include rules about the number of peaks in the chromatogram, their skew, and the separation of the peaks (using an optimisation criterion such as those in section 1.3 to describe this).

### 1.6.2 The inference engine

The inference engine is the part of the system which attempts to apply information supplied to it to the rules and objects stored in the knowledge base. There are two types of inference engine: deterministic and probabilistic. Deterministic systems contain rules which can be answered with 100% certainty whereas probabilistic systems contain a degree of uncertainty. Though most real situations are probabilistic, many may be approximated by a deterministic system. Alternatively the system may be described as a fuzzy system in order to account for the degree of uncertainty in the system (see section 1.3.6 and Chapter 3). The manner in which rules are 'fired' by the inference engine is different from usual languages such as FORTRAN or C. Rather than having a fixed order for rules to 'fire', they are used according to the information available at any moment (ie. the object set available). The result of that rule may provide further information thus allowing further rules to be 'fired'. Often an explanation capability is built into the system allowing the order in which the rules were 'fired' to be examined thus enabling the reasoning (ie. how the final result was obtained from the rules and objects) of the system to be examined.

Expert system software may be written either using expert system shells or using a standard programming language. Expert system shells are special programming environments which allow the more rapid development of expert systems through the use of special tools which contain inference tools. Essentially they are environments which allow the knowledge base to be set-up and altered easily in comparison with usual languages where many lines of code would need to be altered, but they often tend to run more slowly (LISP based systems (a language often used in artificial intelligence) are especially slow due to 'garbage collection', a process where unwanted information is cleared out). Because of these factors expert system shells are often used for systems development and research before being coded in an imperative language such as PASCAL, FORTRAN or C.
1.6.3 Expert Systems for chromatography
A number of expert systems are available for use by chromatographers. These range from packages designed to carry out one task eg. mobile and stationary phase selection, to integrated packages claiming to be capable of selecting the mobile phase, optimising the selectivity and then optimising the column size and operating conditions to minimise separation times. These more advanced systems are still being developed. Also the applicability of the systems to all types of separation is not possible due to the large amount of knowledge which would have to be stored within such a system.

1.6.3.1 CRISE (Criterion SElection)
Peeters, Buydens, Massart and Schoenmakers (1988) developed an expert system for the selection of the best optimisation criterion for a particular separation. The decision is based on the following information:

(i) the relative importance of peaks (and the necessity of weighting factors)
(ii) non-idealities - asymmetric peaks, large solvent peaks, large peak height ratios etc.
(iii) expected changes in plate count
(iv) requirements of the separation ie. equal spreading and importance of each of the peaks
(v) factors to be optimised eg. analysis time, column length, particle size
(vi) parameters which are measurable

This system requires that the peaks are well resolved in all separations analysed so that the number of peaks remains constant and that criteria may be calculated between all peaks. This is acceptable for analytical separations but not for preparative separations. This system is of limited use since it only determines the type of optimisation criteria which best describes the separation not the value of the criteria which is optimum or how this optimum value can be attained.

1.6.3.2 SOS - System Optimisation System
An expert system has been developed (Schoenmakers, Dunand, Cleland, Musch, and Blaffert, 1988 and Schoenmakers and Dunand, 1989) for the
optimisation of the column size, particle size, column operating conditions and instrumentation. The system is only capable of optimising isocratic separations and may be consulted once the user has a chromatogram of separation in which the resolution is acceptable. The system requires certain information from the chromatogram:

For each peak: Retention time, peak height, peak width (or plate count), and asymmetric factor.

For the separation as a whole: the retention time of the unretained peak, the sample size and the noise.

The optimisation is carried out within certain system limits of resolution, signal to noise ratio, pressure drop and flow rate limits. The system uses a series of databases together with a number of equations relating flow rate to each parameter to optimise the separation producing a global optimum and advice.

One database contains information about the columns available to the user ie. column dimensions, particle size, porosity and permeability factor, as well as the packing type. Therefore for each column the pressure drop may be calculated and the effects of column changes may be predicted. Detector types are also recorded in a database together with information about detector cell volumes, relative sensitivities and noise factors.

The system optimises an already resolved chromatogram by increasing the flow rate within pressure drop and resolution limits for each column and packing type available. The signal to noise ratio is also important. The optimum is found by calculating the maximum flow rate allowed by each of the system limits ie. resolution, signal to noise ratio and pressure drop. The value chosen as the optimum is the lowest of these maximum values. This procedure is repeated for each combination of column and detector. The combination chosen is the one with the shortest analysis time. If the lowest flow rate was lower than the minimum required for the detector cell volume or pressure drop requirements then the particular combination of column and detector cannot be used.

The system also provides information about which parameter is limiting further improvement which may suggest to the user ways to overcome the problem eg. by re-assessing the goals or using a different column - packing combination.

This system at present only contains what is described by the authors as ' basic chromatographic knowledge '. Comprehensive and theoretical chromatographic knowledge is currently being gathered to extend the
knowledge base allowing gradient separations, wavelength optimisation and optimisation of separation with some peaks more important than others.

1.6.3.3 Integrated packages

Yuzhu, Peeters, Musch and Massart (1989) developed an expert system for the optimisation of ion pair liquid chromatography of twenty basic drugs. Only isocratic separation was considered. The method integrated current optimisation methodology into the system which was divided into four sections. The first section asks the user to identify the problem and ensures that the problem is relevant to the system. The second (the initial guess module) suggests mobile conditions for a first experiment or set of experiments. The third section called the adaptation module uses the results of the initial guess module. The fourth module uses formal optimisation techniques to improve the separation.

The system contains 87 rules obtained from an experimental study of ion-pair chromatography of twenty basic drugs. Experimental factors incorporated into the system included the equilibrium between the different species present (i.e. ion-pair reagent anion, eluent cation, solute cation and free binding site on the stationary phase), stationary phase selectivity, organic modifiers, buffer and pH, ionic strength and temperature. The system had a satisfactory success rate.

Schoenmakers, Peeters and Lynch (1990) have described the linking together of three pieces of chromatographic optimisation software and expert systems:

(1) CRISE
(2) Diamond - a selectivity optimisation package (not a traditional expert system)
(3) SOS - System optimisation system

Two of the packages (CRISE and SOS) have already been described above. Diamond is an optimisation package which relies upon stored data in the form of three dimensional response surfaces.

The criteria selection system was validated by applying it to ten literature reports which were selected to be representative of selectivity optimisation procedures. The selection of the system was found to be consistent with the decisions of a human expert.
Interactions between CRISE and Diamond were investigated and in particular the effect of the different criteria and the importance of each of the peaks. As would be hoped the system found a different optimum depending on the criteria selected by CRISE.

The system optimisation system was found to be able to halve the analysis time if the resolution of all peaks was required and reduce it by a factor of 3-5 if some peaks were said to be unimportant. Again as with other systems it is only really applicable to analytical separations but allows analysis of all three possible areas of chromatogram analysis to be carried out and allows the goal of the separation optimisation to be varied.

1.6.3.4 ECAT - (Expert Chromatography Assistance Team)

The ECAT system (Bach, Karnicky, and Abbott 1986) is being developed to choose between liquid and gas chromatography, to specify the column, mobile phase and detector, to decide on sample clean up prior to loading and to provide diagnosis of hardware problems.

From the description of the solutes to be separated the ECAT system can determine what stationary phase and mobile phase are suitable. This module contains 160 rules to cover a selection of commonly separated compounds. It also makes recommendations about guard columns.

A column diagnosis section of the program uses chromatographic parameters such as efficiency, asymmetry, selectivity and operating pressure to detect failures of the column. The following column failures can be detected by the system:

- Plugged column or frit
- Dissolution of the column bed
- Cleavage of the bonded phase
- Chemical alteration of Cn bonded phase
- Deactivation of Si adsorption sites by H₂O
- Loss of packing material
- Irreversible adsorption

Non column failures are also detected by the system:

- Too large an increase in injection volume
- Inadvertent change to strongly eluting injection solvent
- Inadvertent overloading

These last failures are sometimes wrongly assumed to be column failures due to a lack of operator knowledge and care over how exactly the column is operated.
1.6.3.5 Method Validation

A software package has been developed (Mulholland, VanLeeuwan and Vandeginste, 1989) to validate HPLC methods for reproducibility. The expert system requires information such as the number of injections, sample preparation times, numbers of users, and number of instruments. It uses rules based on the variance of peak areas height, and retention time to comment on the reproducibility and hence suitability as an HPLC assay. If the data under examination produces unacceptable results then the system will produce suggestions to overcome the non-reproducibility eg. changing process parameters such as flow rate or injection volume.

1.6.3.6 PRE OPT

Systems are available for the selection and optimisation of binary gradient separations. The PRE OPT package is an example of this type of system (Cela, Barroso, Viseras, and Perez-Bustamante, 1986). The system uses data gathered under isocratic conditions and uses a simplex optimisation routine to produce a binary gradient. The results should then be confirmed experimentally. The system has been tested on phenolic acids and aldehydes and the predicted results showed good agreement with experimental results. The system is limited to pre-optimisation of the gradient and no on-line monitoring is included.

1.6.3.7 DryLab

The DryLab system is a piece of software designed to aid chromatographic method development for isocratic and gradient elution modes in reversed phase, ion-pair and normal-phase HPLC (Dolan, Lommen, and Snyder, 1989 and Snyder, Dolan, and Lommen, 1989).

Retention models based on logarithmic relationships between mobile phase composition and capacity factor $k^{'}$ which may be applied under isocratic or gradient conditions are used to select mobile phase for a given separation. The system uses a small number of initial experiments using different mobile phase compositions to determine the parameters for the retention models. Then a series of simulated runs are carried out and from these the most suitable separation conditions can be selected. This system has been successfully applied to HPLC separations of small molecules. The main disadvantages of the system are related to the retention models used. These models rely upon the isotherm being linear (i.e. they are only applicable for small loadings). Also many compounds are not accurately modelled by this type of model i.e. a non-linear
relationship exists between log $k'$ and mobile phase composition. This type of model is typically applied to HPLC of small molecules and is subject to errors when applied to other types of separations such as ion-exchange of proteins (Dolan et al 1989).

1.6.3.8 MABL A

Kenney, Thompson and Harris (1989) described an automated system for controlling adsorption chromatography (more specifically affinity purification of monoclonal antibodies). This system applies a low level of artificial intelligence to a monoclonal antibody purification to take into account process variations and changes in process conditions for subsequent separations unlike other previously proposed automated systems which simply used elution volume (or time) or absorbance detection to determine a more constituent accurate product fraction position than possible by manual methods (Chase, 1986 and, Kenney and Chase 1987).

The system controls and monitors the performance of an adsorption chromatography column. Strictly due to its structure the system is not an expert system but a knowledge based system. It uses simple rules based on the judgements of peaks areas and shape.

When the system starts it carries out two separations for calibration. One overloads the column to determine the column capacity whilst the second determines which peak is required and at what pH the antibody is eluted.

The system requires a value for the concentration of antibodies in the feed stream. On initial runs the user's first concentration value is checked and a correction calculated if necessary. Load reduction of 10% occurs if the peak obtained is more than 90% of the expected value to prevent loss of antibody. If the peak is less than 70% of the expected value then fouling has probably occurred and cleaning is recommended.

A lower limit may be set and if this is reached packing replacement is recommended. Peak shape is also monitored by calculating and recording the column skewness factor $P$. A set value of $P$ (defined as $P'$) represents the extreme skewness value at which point remedial action should be taken. Poor peak shape results in deviation of $P$ which is caused by two factors:

1. If the pH at which elution is carried out is too high then the antibody will elute with a broad peak and an extended tail.
2. Lipid and denatured proteins may form causing column fouling. This will lead to uneven flow thus causing poor peak shape. It may be wholly
or partly corrected by cleaning. Before production runs the column will have been cleaned so pH is likely to be the problem. The pH is then reduced by 0.5 pH units but if another peak is eluted within 1 pH unit then this is not done. If after the pH is reduced there is an improvement but it is not within the P range a further reduction in pH is carried out.

As well as the above methods to maximise productivity other automated features have incorporated into the system eg. buffer and product level monitoring to prevent the column running dry and air sensors to prevent air entering the packed bed.

This system is at present the only knowledge based system reported in the literature which is specifically design to analyse preparative based chromatograms and to control a separation to obtain a product from the separation. The nature of the chromatogram obtained from the separation is relatively simple since it is an affinity separation which binds very few components other than the product. Additionally the system requires that peaks are virtually completely resolved for the position of the product fraction to be determined. Chromatograms where peaks are overlapped cannot be readily analysed since the positions of the peaks cannot be determined. Also where many more peaks occur in the chromatogram this makes the location of the product peak complex and is not accounted for in this system.

1.7 Scope of work

The primary aim of the project is to research methods for controlling preparative chromatographic separations using rapid post separation analysis in order to maintain product quality and productivity (as opposed to method development and separation optimisation systems for assay development). To date little work has been carried out in the field of monitoring and control of preparative chromatography (Kenney, Thompson, and Harris, 1989). The emphasis has been on the use of expert systems and other software to optimise separation methods at laboratory and pilot scale before scaling up (section 1.6.3). MABLAB (Kenney, Thompson and Harris, 1989) is the only system currently available which can select a product fraction from a chromatogram and this system requires well resolved peaks.
In developing methods suitable for the control of preparative separations it is recognised that preparative chromatograms are very different in nature to those for analytical separations. In a preparative separation not all portions of the chromatogram are fully separated. To separate a single component from a mixture it would be acceptable for all components preceding and following the product to be eluted as a single peak. Thus an idealised preparative chromatogram may be considered to consist of three peaks - two contaminant peaks surrounding a product peak. This type of chromatogram will be used as a model of a typical preparative chromatogram in this thesis.

The type of information that may be determined from a preparative mode chromatogram also differs from that determined in an analytical separation. Since the objective of a preparative separation is the production of material then it is necessary to determine from the chromatogram where the product fraction should be taken, i.e. where the material lies within the chromatogram. This requires that the product peak can be found and identified amongst all the contaminants and the start and end positions of the product peak be determined. Additionally if the purity and yield of the fraction are to be determined then the elution profiles of the individual components must be known so that the exact positions of the fraction starts and ends can be determined to give the desired product characteristics.

These requirements for the selection of the product fraction define further the techniques which will be researched to fulfil the stated aim of the thesis, namely analysis of separation conditions to maintain the productivity and purity of the product fraction, from separation to separation.

Since the type of chromatographic separation considered in this thesis operated in batch mode then the complete data describing the separation is not obtained until the end of the separation. The techniques which will be used in this thesis require the complete data describing a separation and hence the nature of control that may be applied to the separation will be at-line rather than on-line. Thus any information gained from the chromatogram analyses developed in this thesis may be used either to decide which column fractions should comprise the final product fraction for the current separation or where the product fraction should be for subsequent separations.
In simple terms a method to fulfil the aims of this thesis may be divided into three tasks:-

- determination of individual component elution profiles (deconvolution, Chapter 2).
- identification of the product and contaminants elution profiles, using the peak area data and elution order data determined by deconvolution (matching, Chapter 3).
- selection of the optimum product fraction, using the identified elution profiles for each component (Chapter 4).

This information is summarised in figure 1.8 which shows the proposed scheme of chromatographic data analysis for product fraction control. Initially deconvolution will be examined using computer generated and experimental data to determine the limits of the chosen deconvolution method and hence what types of chromatogram may be analysed (i.e., the combinations of peak separation and heights that can be analysed successfully by deconvolution). The degree of agreement between the true component elution profiles and those found by deconvolution will be examined. For computer generated chromatograms the true component profiles will be available but in order to confirm the success of deconvolution of experimental chromatograms it will be necessary to carry out an off-line analysis of fractions of the column eluent so that an alternative measurement of the elution profile of a component can be determined.

The reliability of the methods developed for the identification of peaks will be examined using peak area and elution order data determined from a number of chromatograms. This will be achieved using a variety of chromatographic data to enable the effects of variations in the peak characteristics caused by fluctuations in process conditions and the effects of variations caused by limitations of the deconvolution process to be assessed.

The effects of deconvolution accuracy and identification reliability on the selection of the optimum product fraction will be investigated. Finally techniques for the selection optimum fractions will be illustrated through case studies in which the requirements of high yield and high purity fractions will be studied.
Figure 1.1 The Langmuir Isotherm
Figure 1.2 A typical chromatogram. This is a chromatogram of a separation of a complex mixture (bovine liver extract) by ion-exchange in a 2.6x10cm column. (Sofer and Nystrom, 1989).
Figure 1.3 A schematic chromatogram showing the concept of capacity factor, $k$. 

$$k = \frac{t_{Ri} - t_0}{t_0}$$
Figure 1.4  a) V, The Valley to Peak Ratio  
b) P, Peak Separation Criterion
Figure 1.6 Two incongruent coincident peaks. $Rs = 0.0$
Peak overlap parameter (equation 1.12) = 0.50.
Ratio of standard deviations = 0.268

Figure 1.5 Two congruent peak separated by 1.665 standard deviations.
$Rs = 0.416$ and overlap parameter (equation 1.12) = 0.50
Both peaks have equal standard deviations.
Figure 1.7 Definition of asymmetry using the A:B Ratio
Figure 1.8 The proposed data analysis and control sequence showing the flow of data required into and out of the system.
Chapter 2.

Deconvolution.


2 Deconvolution

In this chapter techniques for deconvoluting chromatograms to determine the individual component's elution profiles contained within a chromatogram are investigated. In particular the situations in which the deconvolution algorithm will correctly predict the peak functions within a chromatogram are studied by using simulated chromatograms and experimental chromatographic data. Finally off-line data is used to provide information about actual peak positions with which to compare the data obtained by deconvolution.

2.1 Introduction to Deconvolution

In order to control any process there must be a means of monitoring it, i.e. making a reliable measurement of a relevant process parameter. Since fixed bed liquid chromatographic processes are typically batch separations the complete set of data describing the separation's performance is not available until the entire batch has been processed. This therefore makes it impossible to predict on-line, from the absorbance signal, the composition or purity of the eluted material unless a template chromatogram is available. Furthermore any template would only apply at one particular set of elution conditions and may so be of limited utility. The information and data obtained from a chromatographic separation are in the form of a chromatogram - typically a graph of ultra violet absorbance or refractive index of eluent versus elution volume or time. Any point on the chromatogram represents the interaction of a solute element with the mobile phase over the whole column and thus there is a time lag (approximately equal to the column void volume) between the interactions and measurements occurring.

In preparative separations baseline resolution is not usual since the productivity of the separation is important and this is effected by the degree of separation. Ideally there would only be one portion of the chromatogram which will be highly resolved i.e., the region around the product peak; in order to permit a high purity product to be obtained. The areas where contaminants are present should not be well resolved because the resolution of these components from one another is not needed and so causes an unnecessary increase in elution time and hence process costs.

If neither the product or neighbouring peaks are well resolved then accurate measurement of the peak areas and retention times will not be possible (figure 2.1). This is because the elution profiles of individual
components and the start and end of the mathematical functions which describe the component's elution profiles are not known. In order to obtain this information a process known as deconvolution (obtaining the mathematical functions which describe the elution profiles of the individual components) must be carried out. There are several methods of deconvolution available and each has its advantages and disadvantages. The methods vary in the complexity of the detection equipment used and the mathematics needed to obtain the functions describing the elution profiles of the individual components. The more complex methods, both in terms of detection equipment and mathematics, are suited to analytical separations where the number and amounts of components are unknown (Fell et al., 1983). Such methods use diode array detectors to obtain an absorbance spectrum at each elution volume or time analysed. For the monitoring of preparative chromatographic separations less complex curve fitting techniques may be used, since ideally the number of components will be known, as well as the amount of each component. The following section examines the use of such techniques.

2.2 Curve fitting techniques

It is necessary to have two pieces of information about the separation before curve fitting methods may be used. Firstly the number of components which the deconvolution process is expected to resolve must be known (this is not necessarily the same as the number of chemically different species since chemically similar species may co-elute and thus appear to be identical by chromatographic analysis). Secondly a model must be selected to describe the peak shape. Several models exist to describe peak shape. Ideally elution profiles should correspond to the Gaussian peak shape, equation 2.1.

\[ h_v = \frac{A}{\sigma \sqrt{2\pi}} \exp\left(\frac{-(V-V_p)^2}{2\sigma^2}\right) \]  

(2.1)

(See section 1.2 for a definition of parameters)

This type of peak rarely exists in reality due to inefficiencies caused by extra column mixing, non-ideal interactions between the solute and stationary phase, solute-solute interactions (especially at high solute loadings) and detector response time lag. To take into account these inefficiencies several other models have been developed which include the
exponentially modified Gaussian peak (Gladney, Dowden and Swalen, 1969, Equation 2.2) which is a Gaussian function combined with an exponential decay function, and the general exponential function (Vaidya and Hester, 1984, Equation 2.3).

\[ h_y = A \exp\left(\frac{1}{2} \frac{a^2}{\tau^2} \frac{V - V_g}{\tau}\right) \text{erf}\left(\frac{1}{\sqrt{2}} \frac{V_g}{\sigma + \tau}\right) \]  

(2.2)

\[ h_y = h_m \left(\frac{V - V_o}{V_m - V_o}\right)^{(b-1)} \exp\left(\frac{b-1}{a} \left[ 1 - \left(\frac{V - V_o}{V_m - V_o}\right)^a\right]\right) \]  

(2.3)

The exponentially modified Gaussian peak function (equation 2.2) has a theoretical basis (the exponential decay function has been shown to model inefficiencies due to detector response lag in gas liquid chromatography, McWilliam and Bolton, 1969) but the general exponential function, equation 2.3, is purely empirical.

The choice of model is critical in determining the success of the deconvolution process, since the model must be capable of describing the peak shapes produced in the separation.

2.2.1 Models

The following sections examine the application of two important models that can be used for deconvolution and discusses their relevance to the modelling of chromatographic data.

2.2.1.1 The exponentially modified Gaussian function

The exponentially modified Gaussian function, (equation 2.2) uses four parameters to describe a peak: \( A \), the peak area; \( V_g \), the centre of the unmodified Gaussian function; \( \sigma \), the standard deviation of the pure Gaussian function; and \( \tau \), the modifier (the decay constant in the decay function used to modify the Gaussian). For fixed \( A \), \( V_g \), and \( \sigma \), increasing the size of the modifier, \( \tau \), causes the observed maximum to be shifted to the right of the Gaussian maximum (see figure 2.2). The maximum also lies on the unmodified Gaussian curve. The area of the modified Gaussian is equal to that of the unmodified Gaussian.
The size of the tail is dependent on the modifier $\tau$. The shape of the front of the peak is not directly altered - instead $\tau$ and the height (or area) must be changed but this then changes the area and the tailing of the peak. This may cause the model to be unable to accommodate certain combinations of height and skew.

The exponentially modified Gaussian function was developed for the analysis of gas chromatograms (McWilliam and Bolton, 1969) and has been widely used (Dyson, 1990) - particularly to describe gas chromatograms and analytical HPLC separations of small molecules. The interactions between the solute and stationary phase in low pressure protein liquid chromatography will be significantly different from those present in gas chromatography. The relatively small molecules separated in gas chromatography will not interact in the same manner as proteins which probably bind at several sites. Thus the adsorption - desorption of proteins may give rise to a peak shape which could not be successfully modelled by the exponentially modified Gaussian function.

2.2.1.2 General exponential function

The general exponential function (equation 2.3) uses five parameters to describe peak shapes: $h$, the height of the peak at the maximum; $V_e$, the elution volume at which the peak starts; $V_r$, the elution volume of the peak maximum; and $a$ and $b$ which affect the peak shape ($0 > a < 25$ and $1 > b < 25$) and can be manipulated to change both the kurtosis (degree of curvature of the peak, eg. it can be varied from a Gaussian to a triangular peak) and degree of tailing of the peak. The peak can also be made to skew forwards and backwards, by manipulation of these parameters (see figures 2.3 to 2.5).

Comparing these functions shows that the general exponential function allows a wider range of different peak shapes to be modelled but requires one more parameter than the exponentially modified Gaussian function. In addition the general exponential function, unlike the exponentially modified Gaussian function, has no theoretical basis.

As can be seen from figures 2.3 to 2.5 generalisations on the effects of parameters $a$ and $b$ can be made. Firstly an increase in the value of $b$ increases the symmetry whilst reducing the effective peak width. The parameter $a$ however changes the peak skew from a tailing peak to a peak exhibiting fronting. A low $a$ leads to tailing, ie. the peak is distorted from a Gaussian shape so that the peak end is extended into higher
elution volumes, where as higher values lead to fronting (the opposite of tailing).

2.2.2 Objective functions

Once the peak model and the number of peaks \( j \) have been defined a theoretical chromatogram may be constructed from the sum of \( j \) peak functions chosen. The next task is to determine the parameters of each peak function. The peak function parameters are optimised so that the sum of the peak functions matches the experimental set of data under analysis as closely as possible. Generally a least squares minimisation is used when such an optimisation is carried out. However other objective functions may be more appropriate. These are discussed in the following paragraphs.

The most popular objective function is that of least squares minimisation. Here the sum of the differences is effectively a mean of the square of the differences between the experimental and model chromatograms. This means that a given value of this objective function may consist of a few relatively large differences or many relatively small differences. The former situation may be unacceptable especially if the deviations occur on or around the product peak function since this would cause a big difference between the actual elution profile and the elution profile obtained by deconvolution and so put calculations of the product purity and productivity significantly in error. In equation 2.4 of represents the means of the square of the differences and is therefore the objective function. The number of data points is \( i \) and \( j \) is the number of peak functions used to represent the chromatogram. \( f_j(a, b, c \ldots) \) represents the peak function of the \( j^{th} \) peak at the \( i^{th} \) data point. The parameters of the function are represented by \( a, b, c \ldots \). \( D_i \) is the \( i^{th} \) data point in the chromatogram.

\[
of = -\Sigma_0^i \left( D_i - \Sigma_0^j f_j(a, b, c \ldots) \right)^2 \tag{2.4}\]

An estimate of the difference in area between the actual and model chromatograms may then be made using this objective function (see section 2.2.5). In order to prevent large deviations occurring it may be necessary to consider other objective functions.
Rather than using the actual difference between chromatogram and model chromatogram the difference relative to the actual chromatogram value may be used in a 'relative difference' objective function. This would have the advantage of increasing the importance of large differences for small real chromatogram data point values. Either a sum or a maximum value may be used. If a sum is used then another objective function could give each data point on the chromatogram under analysis a weighting depending upon the perceived importance of the data point. This would be dependent upon the position of the data point relative to the product peak function. This therefore requires some knowledge of the positions of the peak functions. This is however exactly what the deconvolution process is supposed to determine, and so makes this method difficult to use.

An alternative method would be to consider the maximum difference. In this case instead of adding the square of the differences at all points on the chromatogram, the maximum difference (or relative difference) squared could be used as an objective function. This may be considered as a more effective objective function since it records the largest difference and is not affected by a large number of relatively small differences at other data points. For instance if a sum of the differences squared was used then a particular objective function value may contain one or more differences which are of a significant size but when taken together with other small differences give an overall objective function which is not large. If the maximum was recorded this problem would not occur.

The comparison of the objective functions above considers only the usefulness of the final value on convergence of the optimisation task. It does not consider the effect that using any of the functions may have on the effectiveness of the optimisation technique, ie, convergence may or may not occur and the objective function may or may not reach a suitably low value. For example the use of one objective function may introduce local minima which could cause the algorithm to converge to one of these points and thus fail to produce a model chromatogram which satisfactorily describes the actual chromatogram. Least squares minimisation has been widely used and known for its stability. For these reasons and the fact that a difference in area between the actual and model chromatogram may be estimated from it (see section 2.2.5), the least square difference objective function is used in this thesis.
2.2.3 Constraining the optimisation problem

In order to reduce the possible number of calculations in the optimisation process used to determine the peak function parameters for the model chromatogram, it is possible to define a range of values which the peak function parameters may physically have, and, to give the optimisation algorithm a starting point or range of starting points to constrain the optimisation problem.

Constraint information may be derived from the experimental chromatogram and procedures for extraction of such information is described below.

Certain of the peak function parameters may be constrained within limits. For example, the retention times of the peaks must lie between zero and the end of the chromatogram. For the general exponential function the chromatogram will constrain the maximum heights and the positions of the function maxima as shown in figure 2.6. For consistency of the general exponential function it is required that \( V_i > V_o \), i.e. the peak maximum has a larger elution volume than the peak start. A peak function's height \( h_i \) will always be below the observed height since an overlapping peak will increase the overall height. The minimum height constraint may be defined as zero or another value may be estimated from the chromatogram. The range of possible \( h_i \) values is given in equation 2.5, where the \( i^{th} \) peak is overlapped by a smaller \( i+1^{th} \) peak. This would be calculated as the difference between the observed peak height and the height of the taller of the two overlapping peaks (The largest possible contribution to the peak height from the overlapping peak).

\[
0 \leq h_{mi} \leq \text{observed peak height} - h_{mi+1} \tag{2.5}
\]

Thus the ranges for the \( V_0 \)'s as shown on figure 2.6 must be defined. This must be done subjectively by estimating where the peak functions are likely to start and where the maxima are likely to be. Choosing the constraining range for the start of each peak is the most difficult parameter to deal with for two reasons. Firstly the chromatogram contains relatively little information about where peak starts occur in comparison with peak function maxima. The first peak start however is simply defined as a single point, the start of the chromatogram portion under analysis. Secondly it has been found that if a large range is set for the peak maxima (which may be considered desirable since it is difficult to define clearly in what range this parameter will lie) this may
give rise to problems. A wide range for these parameters apparently introduces local minima to which the algorithm may converge. The global optimum or near global optimum is therefore not found. It is usually clear when this problem is occurring since the algorithm fails to converge to a satisfactory objective function value, indicating that the constraints must be investigated.

The first and third peaks’ observed maxima are shifted from the true peak function maxima by the overlapping of the central peak. Both are shifted towards the central peak and thus the actual peak function maximum for the first peak will lie between the chromatogram start and the observed peak maximum. For the last peak the actual peak maximum will lie between the observed peak maximum and the end of the chromatogram, ie.

$$0 < V_{m1} \leq V_{m1}^* \quad (2.6)$$

$$V_{m3}^* \leq V_{m3} \leq V_{end} \quad (2.7)$$

The problem may be further constrained by subjectively reducing the permissible range from chromatogram start and end. These constraint ranges are given in equations 2.6 and 2.7.

The observed shift in the central peak’s maximum is more difficult to analyse since it depends upon both of the neighbouring peaks. Since the effect of different peak shapes and heights overlapping on the shift in observed maxima is complex the best constraint that can be applied to the central peak’s actual maximum is that it will lie between the two valleys, or shoulders.

A summary of the constraints for the general exponential function are given in figure 2.6.

In the case of the exponentially modified Gaussian peak the centre of the first peak function will be between the start of the first peak and the observed maximum of the first peak. The standard deviation may be estimated by measuring the peak width towards the top of the peak and assuming an unmodified Gaussian model. This will tend to overestimate the standard deviation. The modifier will also be able to be constrained since the ratio of standard deviation to modifier will lie within defined limits. It is unwise to take exact values from the chromatogram since overlap will affect peak width but a range of values may be taken safely.
2.2.4 Optimisation Techniques

The selection of the model parameters which give the best approximation to actual data points then becomes an optimisation problem. It is important however to remember that this process must fit into the time scales available within a chromatographic separation. Since the optimisation process requires a complete chromatogram to analyse, on-line control of the separation using the current separation's data is not possible.

Two possible alternatives however are possible, either use data from previous separations to control the position of the product fraction in the current separation or use the analysis to decide which collected fractions should be pooled together to form the product fraction.

With either of these alternatives there is a significant period of time available to carry out the optimisation process. This time lasts from when the elution of material from the column is completed until the loading of the next separation is complete.

However not all of this time is available for deconvolution as other processes must be carried out such as identification (see chapter 3) and control decisions (see chapter 4) based on the information gained from the deconvolution. A more detailed discussion of the time required for all these processes is given in chapter 5.

The Box-Complex algorithm (Box, 1965) is used in this thesis and is described below.

2.2.4.1 Box-complex optimisation

The Box-complex technique (Box, 1965) is a directed search optimisation method similar to the Simplex method (Nelder and Mead, 1965) but with a constrained feasible space.

Initially a number, \( n \), of constraints (equal to the number of parameters) are set. For the \( i^{th} \) parameter:

\[
g_i \leq x_i \geq h_i \quad (2.8)
\]

These constraints define the feasible space. Initially \( n+1 \) (where \( n \) is the number of parameters to be optimised) vertices are randomly selected, eg. for the \( i^{th} \) vertex:

where \( r_j \) is a pseudo-random number.
\[ x_i = g_i + r_i(h_i - g_i) \quad (2.9) \]

The lowest point in the complex is then found and this is replaced by a point \( a > 1 \) from the centroid of the remaining points, the new point being co-linear with the old point and the centroid. If the new point is still the lowest it is replaced with a point half way between the old point and the centroid of the remaining points. If an explicit constraint is violated then the offending point is moved just inside the feasible space and then the point is modified in the same manner as for a repeating lowest point.

This procedure is repeated until either an arbitrary maximum number of iterations has been exceeded or the objective function is below a certain threshold (see figure 2.7).

One advantage of this method is that it does not require derivatives of the objective function. This makes the mathematics of the problem much simpler but makes the rate of convergence slower, since the rate of descent of the objective function towards zero is not used.

The Box-Complex optimisation and indeed any other optimisation method may fail to obtain an optimum solution to a given problem. This may be due to one of a number of reasons which are listed in table 2.1.

<table>
<thead>
<tr>
<th>Table 2.1 Box-Complex Algorithm Failure Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of problem</td>
</tr>
<tr>
<td>1 A minimum does not exist within the feasible space defined by the constraints.</td>
</tr>
<tr>
<td>2 The randomly chosen complex does not converge onto a global minimum.</td>
</tr>
<tr>
<td>3 The lowest point in the complex repeats after modification as the lowest point. This will occur if the objective function is concave, i.e. the objective function decreases towards the centroid.</td>
</tr>
</tbody>
</table>
2.2.4.2 Deconvolution failure

There are a number of situations where deconvolution methods may fail. All methods of deconvolution have a limit on the degree of peak separation they can accommodate before they begin to fail. With curve fitting this occurs when the overlapping peaks could each comprise a number of different combinations of peaks. This is described mathematically as ill conditioning, i.e., there is no unique solution. The degree of separation below which the method will fail is dependent on a number of factors including the relative heights of the overlapping peaks and the degree of skew, i.e., the exact peak shape. For example, a small peak and a relatively large peak separated by the same differences in retention time of elution volume, as two large peaks will exhibit ill conditioning at larger differences in retention time than the large pair.

The curve fitting method of deconvolution does not guarantee to find the global optimum, for instance, the method may select a set of parameters which is not the set which gives the lowest possible difference between the two chromatograms. Also, there is no guarantee that the method will, even if it does actually find the global minimum, produce peak functions which correspond to the actual elution profiles of the individual components. Such a situation may occur even if the problem is not ill conditioned. This is because the actual elution profiles may not correspond to the peak models used in the deconvolution process.

It is therefore necessary to determine under what conditions the particular deconvolution method will fail to produce the correct result and how accurate the result will be when the model converges to the desired objective function value. The accuracy will be determined by the degree of peak overlap and will also be affected by the size of the difference between the two chromatograms. The effects of these two factors will also need to be determined, as well as the most suitable and stable objective function. The following sections explore the effects of these factors on the convergence of deconvolution methods.

2.2.5 Maximum Acceptable Objective function size

Since the deconvolution process must be capable of dealing with chromatograms consisting of various peak separations and peaks of various heights it is necessary to describe the chromatogram so that an objective function size may be selected which is appropriate to a given chromatogram.
It is important to note that the significance of the size of the objective function will vary as a function of the relative height and separation of the peaks. For example a particular objective function value will be more significant for a chromatogram with a low mean data point height in comparison with a chromatogram with a high value.

The objective function selected (least squares minimisation) may be considered to be related to the area difference between the model and real chromatogram. The area difference therefore should be as small in relative to the area of the chromatogram. A simple measure of chromatogram area is the mean data point height since this multiplied by the total elution time gives the chromatogram area, ie:

\[ \text{Area} = \bar{x} \cdot \text{Elution Volume} \quad (2.16) \]

This does not give an indication of the peak separation or overlap but this may be estimated using the peak to valley ratio optimisation criterion (see section 1.3).

In this thesis an objective function value of 5.0% difference between model and real chromatograms was used as the maximum difference. Values greater than this gave unacceptable results and irreproducible results due to ill conditioning (ie. one objective function value may give rise to many different solutions).

The objective function (equation 2.4) may also be expressed in the following way:

\[ f = -n.(\Delta \bar{x})^2 \quad (2.17) \]

Where \( n \) is the number of data points, \( \Delta \bar{x} \) is the mean deviation between the model and actual chromatograms and \( f \) is the objective function value. Thus rearranging the mean deviation may be obtained;
\[ \Delta \bar{x} = \sqrt{-\left( \frac{f}{n} \right)} \quad (2.18) \]

\[ \Delta \text{area} = \Delta \bar{x} \cdot \Delta Y \quad (2.19) \]

This will tend to overestimate the size of the actual area difference as all differences will become positive when squared but this is deemed acceptable since it will always overestimate the degree of error. Thus a ratio of \( \frac{\Delta \bar{x}}{\bar{x}} \) (ie. \( \frac{\Delta \text{area}}{\text{Area}} \)) may be used to compare chromatograms of different area.

The following section will examine the deconvolution technique by applying it to various computer generated chromatograms with differing peak separation and peak height to determine what effect these factors have on the success of the process.

2.2.6 Theoretical examination of model convergence

In order to test the technique a series of chromatograms containing three peaks of various sizes was generated using the general exponential function and the Gaussian function (see Appendix A1.2). Some chromatograms included random noise at a level of between -1.5% and +1.5%. This value was chosen as Rossi (1988) stated that for reliable peak area measurement a signal to noise ratio of at least 60 is required which corresponds to a signal deviation of 1.7%. Above this level of noise even fully resolved chromatograms could not be integrated consistently. Typically noise is due to some local interference which may filtered. Random noise represents the worst possible case and for this reason is studied.

In general chromatograms were successfully deconvoluted, ie the peaks functions determined by the deconvolution process were in good agreement with those used to generate the chromatograms. No difference was observed between the results of the Gaussian based chromatograms and those generated using the general exponential function, despite the fact that this function uses three additional parameters and significantly
skewed peaks were used. No significant difference was found between the chromatograms containing noise and those without. According to Vaidya and Hester, 1984, peak areas are found to within 5% but only when there is adequate peak separation. This restriction was not reported by Vaidya and Hester whose chromatographic system consisted of two peaks. Their aim was only to determine the amounts of material in the two component mixture (for analytical purposes) rather than obtaining a detailed model chromatogram for analysis of preparative separations and so an analysis of the method to determine under which situations it would fail is not so critical.

All chromatogram deconvolutions carried out in this section took no longer than ten minutes (see Appendix 1 for hardware details).

An example of successful deconvolution is shown in figure 2.9. The features of this chromatogram which lend it to analysis by deconvolution are the tall central peak is relative to the surrounding peaks and only moderate overlap. These factors may be described by the peak to valley ratio (see section 2.3.10). Such a chromatogram is typical of a final stage in a chromatographic sequence where relatively small amounts of contaminant are present compared to the product.

When neighbouring and overlapping peaks did differ significantly or the peaks overlapped so that a shoulder appeared then unsatisfactory deconvolution performance occurred. An example of unsatisfactory deconvolution can be seen in figure 2.8 - peak functions determined by deconvolution are not in agreement with those used to generate the chromatogram - peak starts and ends differ greatly. This situation is less typical of a final stage chromatographic separation but may occur when the product yield is low or the extinction coefficient of the product is low in comparison with other components.

2.3 Experimental determination of deconvolution effectiveness

Having examined the capacity of the deconvolution technique to analyse certain synthesised data, typical of many chromatograms the effectiveness of the deconvolution technique was tested with data from real biological systems. The following sections describe this work.
2.3.1 Choice of Chromatographic system

To test the validity of the deconvolution technique on real data a chromatographic system had to be chosen. Before the system was selected the exact type of separation to be analysed, i.e. how many components would be present in the mixture to be examined, what type of chromatography was to be studied and what type of material was to be separated, had also to be determined. The number of components in the system is determined by where in a sequence of chromatographic steps the control and hence the deconvolution is to be carried out. The most appropriate application of the method will lie towards the end of a bioprocess sequence for several reasons. Firstly the number of components will be less at this position in a separation sequence and this will make deconvolution using curve fitting more practical, since if the number of components is too high then the number of variables which must be optimised will also be large, causing the process to become very time consuming.

Techniques have been devised to simplify complex chromatograms such as the 'nearest-neighbour' technique. This approach ignores all components in the chromatogram except the product and its neighbours which are difficult to separate from the product in subsequent the separation step. Thus the number of components which need be considered during design is reduced. This technique cannot simplify the deconvolution process since for deconvolution to work correctly all peaks must be included in the model chromatogram. Also the identity of peaks cannot be determined until after the deconvolution process has been carried out (see chapter 3), when further information about each peak is available and thus the product peak and its nearest neighbours cannot be determined and the simplification offered by the nearest neighbour approach cannot be exploited.

The more components that are present the more statistically likely it is that the solution to the deconvolution process will be ill conditioned in a given column of fixed capacity. Davis and Giddings (1985) have shown this using the statistical model of overlap where the observed number of peaks varies exponentially with the number of components. Expressed mathematically,
\[ p = \bar{m} e^{-\bar{m} n_c} \]  \hspace{1cm} (2.20)

where, \( \bar{m} \) is the number of components, \( p \) is the number of peaks and \( n_c \) is the column capacity (this is defined in terms of a number of peaks of a given width which may be separated and observed in a chromatogram of a separation).

For example assuming a fixed capacity of twenty peaks. When the number of components is twenty, the number of visible peaks is approximately seven but when the number of components is thirty the number of visible peaks is six. Therefore as the number of components increases the number of peaks visible decreases, ie. the amount of overlap increases. Thirdly if there are fewer components to be separated, the control of the separation becomes increasingly practical and potentially more beneficial. This is because chromatographic steps early in a sequence will be relatively crude separations which use a large amount of the column capacity, for economic reasons, ie. the process is operated as an adsorption / desorption process. If most of the capacity is used then there is little room to improve the separation quality or alter productivity by further separating the component peaks. Later chromatographic steps will tend to be used to provide very high purity products and these will require good resolution of the product peak from other components. This will necessitate that a large column capacity be used in comparison with the number of peaks (Davis and Giddings, 1985). The large column capacity should also give sufficient spare elution volume for the separation to be changed. Additionally if there are a large number of components to be separated then it is likely that the capacity factors of the components will be too similar for the separation to be changed significantly. Capacity factors are only large over a small range of modifier concentrations, ie. the separation is a desorption process and requires the use of a gradient rather than an equilibrium process that can be operated isocratically.

Finally the system selected should not consist an artificial mixture of components as such a mixture will not necessarily model a real process stream closely enough. An artificial mixture is unlikely to contain the series of trace components present in a real mixture and which would lead to aspects such as apparently noisy baselines and possible extra components appearing when the separation parameters are changed. Also the inclusion, within what appears to be one peak, of a trace compound
or compounds may cause the observed peak to deviate significantly from the chosen peak model.

All of the above possible problems may cause the convergence of the model to the correct solution to occur more slowly and may even prevent convergence to an acceptably low objective function value due to local minima or simply because the trace components prevent the model from accurately describing the chromatogram under analysis. It is important to test the deconvolution with a system exhibiting these problems as these difficulties may be expected to arise in process scale separations. For these reasons a real system was chosen.

Separation of hen-egg white by cation exchange at pH 6.0 was considered to be a suitable test system since the only three major components are bound to the column from approximately ten components. Additionally it has been thoroughly studied and it is well understood (Gilbert A.B, 1971 and OROS Instruments Ltd., 1990). Of the material bound to the stationary phase approximately 24% is lysozyme, 54% is ovotransferrin (also known as conalbumin), and 22% ovoglobulin.

These proportions of components (with their extinction coefficients) will however give rise to a chromatogram with significantly differing heights which will cause difficulties with deconvolution under certain circumstances (see section 2.2.6). Such a situation is not typical in many ways of preparative separations, especially if the central peak is taken as the product (ovoglobulin) and the other peaks are taken to be contaminants. Typically towards the end of a separation the product would constitute the majority of material present in the separation mixture, since other separation steps would have removed the majority of the contaminants. In this mixture ovoglobulin (the product) only constitutes 22% of the total. It should also be noted that the extinction coefficient of lysozyme is twice that of the other components thus making the lysozyme peak approximately twice as large as the ovoglobulin peak when there is an equivalent amount of each material in the two peaks. This makes the chromatogram produced during this separation appear more atypical than others with similar quantities of material present.

This type of chromatogram is studied in this thesis to fully determine exactly what the limits of the deconvolution technique are, rather than using a more straightforward chromatogram such as the test chromatogram shown above in figure 2.9. This chromatogram is more typical of preparative separations but deconvoluted with better accuracy under a wider range of peak separations.
The separation of these components from hen egg-white can be carried out by eluting from a cation exchanger with a sodium chloride gradient (concentration varied from 0 M to 1.5 M), with the bound components being eluted in the following order ovotransferrin, ovoglobulins, and finally lysozyme. The following table, table 2.2, shows properties of the bound components (Rhodes et al., 1959 and Gilbert, 1971).

An important fact worth noting is that the extinction coefficients are not equal which will therefore mean that peak areas obtained will not correspond directly to the masses of each compound loaded onto the column. This will have an effect on the deconvolution process since a point difference of a given size between the model and real chromatogram will not correspond to an equal mass difference at different points on the chromatogram.

### Table 2.2 Physical Properties of Major Egg-White Components

<table>
<thead>
<tr>
<th></th>
<th>Molecular Weights (Daltons)</th>
<th>pI</th>
<th>Extinction Coefficients at 280nm (AU mL mg⁻¹ cm⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovotransferrin</td>
<td>86000 - 76000</td>
<td>6.05</td>
<td>0.89</td>
</tr>
<tr>
<td>Ovoglobulin</td>
<td>45000 - 36000</td>
<td>5.80</td>
<td>0.90</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>17000 - 14000</td>
<td>11.00</td>
<td>2.02</td>
</tr>
</tbody>
</table>

The selection of the system described above assumes that the eluted material will lie within the detectable range so that measurement may be made. The degree of loading considered to be preparative could be an amount which produces milligrams of material (Nerbert, 1991). The nature of the product will affect the amount which must be produced in each separation. If comparatively large quantities of product are to be produced then saturation of the detection system may occur or the level of loading may be such that the eluted material reaches the non-linear portion of the detector's response. In such a situation it is not possible to quantify the amounts of material if the separated components are adequately separated or to apply a deconvolution technique if the components are not. If such a separation is to be analysed an alternative method, apart from deconvolution, must be used. Such an alternative could be the use of an analytical technique on-line to analyse the
composition of fractions of the column eluent. A suitable analytical technique could be an HPLC separation since fast automated systems are available. In such a situation the algorithms developed in this thesis would be applied to the latter analytical data.

2.3.2 Apparatus
A Biopilot System 3 (Pharmacia LKB Biotechnology, Uppsala, Sweden) was used to study the separation of major hen egg white proteins. The system uses a number of motorised valves which allow several columns to be attached to the system and to be automatically connected to the pumps. The automatic valves may also be used as a fraction collector and also allow several columns to be connected to the system.

Two pumps were used (high precision P-6000) and by using the LCC-500 Plus controller were configured to form both linear and stepwise gradients for the elution steps. A mixer unit was also used to smooth the gradient.

The samples were loaded manually into the superloop (50 mL) using a syringe before being loaded onto the column by the pumps. The ultraviolet absorbance at 280 nm was logged electronically using a PE-Nelson Intelligent interface Model 970 and Turbochrom software version 2.1 (Perkin-Elmer, Beaconsfield, England) running on an IBM PS/2 model 55SX computer (IBM Ltd., Portsmouth, England). Fractions were collected using a Pharmacia FPLC FRAC-100 fraction collector. A pre-packed Pharmacia Hiload 16/10 column, diameter 16 mm, bed depth 10 cm, volume 20 mL, packed with Pharmacia S-Sepharose High Performance cation exchanger was connected to the Biopilot system. Similar experiments were conducted using a Pharmacia FPLC with lower rated pumps (High Precision P-500).

2.3.3 Method
To test the deconvolution system a series of chromatographic separations had to be carried out and off-line measurements made to determine quantitatively the composition of the fractions taken throughout the separations. The type of off-line assay used for the measurement of the fraction composition should be considered. The method of detection used typically in a chromatographic system is based upon mass (ie., the ultraviolet absorbance, \( A = I \epsilon c \), where \( I \) is the optical path length, \( \epsilon \) is the extinction coefficient of the component and \( c \) is the concentration
of the component). It would therefore be inappropriate to use an assay based on activity since this would require that the protein be in an active form if it were an enzyme or it be in a form usable as a substrate, if it is a non-enzyme. This would mean that not all of the material detected by ultraviolet absorbance would necessarily be detected by an activity based assay.

Electrophoresis and densitometry would provide a suitable system since the absorbance of each band is measured and used to determine the amount of each band which if the electrophoresis has been successfully carried out will correspond to an individual component, and therefore is based on mass.

It is important that an estimate of the accuracy of the amounts determined by densitometry be found so that a valid comparison of the on-line analysis (ie. deconvolution) and the off-line analysis can be made, ie. whether or not the on-line analysis agrees with the off-line analysis within experimental error.

The size of the fractions will also affect the overall accuracy of the off-line analysis in terms of fraction selection. This is because the size of fractions will affect the spacing of data derived from analysis of the fractions about the composition of the chromatogram at certain points (see chapter 4).

It should be noted that the deconvolution will also be subject to uncertainty (see section 2.2.5).

2.3.4 Chromatographic method

The next sections describe the egg white separations carried out to examine the effectiveness of the deconvolution algorithm.

2.3.4.1 Sample preparation

The white of one hen egg was separated from its yolk and the supporting membranes were carefully removed. The egg white was then diluted with 200 mL of 5 mM sodium phosphate buffer at pH 6.0. This resulted in the precipitation of some components and so the resultant solution was then filtered using a Buchner funnel and Whatman No.1 filter paper. The solution was then stored at 277 K.

2.3.4.2 Separations

10 mL of diluted and filtered egg white were loaded onto the Hiload 16/10 column pre-packed with Pharmacia S-Sepharose HP, which had been equilibrated with 5 mM sodium phosphate buffer at pH 6.0. The unbound
material was then washed from the column using four column volumes of 5 mM sodium phosphate buffer, i.e., 40 mL at a flow rate of 2 mL min\(^{-1}\). Elution was carried out using a gradient of 0.2 M to 1.4 M sodium chloride in 5 mM sodium phosphate at pH 6.0. The volume over which the gradient was carried out was varied from ten column volumes to four column volumes of elution buffer, to produce a variety of different peak separations.

After elution had been carried the column was regenerated with 1.4 M sodium chloride / 5 mM sodium phosphate buffer to remove all bound material before re-equilibration with 5mM sodium phosphate buffer. (Table 2.3 lists the gradients used in separations carried out.)

Fractions of fixed volumes of the eluted material were collected and analysed by SDS polyacrylamide electrophoresis (section 2.3.6).

<table>
<thead>
<tr>
<th>Separation number</th>
<th>starting volume mL</th>
<th>finishing volume mL</th>
<th>starting concentration of NaCl M</th>
<th>finishing concentration of NaCl M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.0</td>
<td>150.0</td>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>100.0</td>
<td>130.0</td>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>100.0</td>
<td>106.0</td>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>4</td>
<td>100.0</td>
<td>106.0</td>
<td>0.3</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>100.0</td>
<td>106.0</td>
<td>0.3</td>
<td>1.4</td>
</tr>
<tr>
<td>6</td>
<td>100.0</td>
<td>106.0</td>
<td>0.3</td>
<td>1.6</td>
</tr>
</tbody>
</table>

2.3.5 Analysis by sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS–PAGE)

A Bio–Rad Mini–PROTEAN II slab cell (Bio–Rad Laboratories Ltd.) vertical slab electrophoresis instrument was used with an EPS 400/500 power unit (Pharmacia Ltd.). Discontinuous gels consisting a stacking gel and a resolving gel were used in all gel analyses. The composition of the gels
(Creighton, 1990) is shown in Table 2.4 and the composition of the running buffer, staining and destaining solution is shown in Table 2.5.

**Table 2.4 Formulation of resolving and stacking gels**

<table>
<thead>
<tr>
<th>Component</th>
<th>Resolving Gel 13.5%</th>
<th>Stacking Gel 4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrylamide/Bis (30%)</td>
<td>13.5 mL</td>
<td>2.7 mL</td>
</tr>
<tr>
<td>1.5 M Tris-HCl pH 8.8</td>
<td>7.5 mL</td>
<td>-</td>
</tr>
<tr>
<td>0.5 M Tris-HCl pH 6.8</td>
<td>-</td>
<td>5 mL</td>
</tr>
<tr>
<td>10% SDS</td>
<td>0.3 mL</td>
<td>0.2 mL</td>
</tr>
<tr>
<td>10% Ammonium persulphate</td>
<td>0.1 mL</td>
<td>0.1 mL</td>
</tr>
<tr>
<td>TEMED</td>
<td>0.02 mL</td>
<td>0.02 mL</td>
</tr>
<tr>
<td>Deionised water</td>
<td>8.6 mL</td>
<td>12 mL</td>
</tr>
<tr>
<td>Total volume</td>
<td>30 mL</td>
<td>20 mL</td>
</tr>
</tbody>
</table>

1. TEMED: N, N, N', N'- tetramethylethylenediamine

**Table 2.5 Formulation of running buffer, staining and destaining solutions**

<table>
<thead>
<tr>
<th>5X Running buffer stock pH 8.3</th>
<th>1L solution:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15g Tris base</td>
</tr>
<tr>
<td></td>
<td>72g Glycine</td>
</tr>
<tr>
<td></td>
<td>5g SDS</td>
</tr>
</tbody>
</table>

Staining solution

| 1L solution:                     |
| 417 mL water                     |
| 417 mL methanol                  |
| 167 mL acetic acid               |
| 1g Coomassie Blue G-250          |

Destaining solution

| 1L solution:                     |
| 600 mL water                     |
| 300 mL methanol                  |
| 100 mL acetic acid               |

2.3.5.1 Sample Preparation

Fractions obtained from the chromatographic separations were analysed using the Bradford protein assay (Bradford, 1979) to determine which had the largest concentration of protein. When this fraction had been
identified a volume of the fraction containing between 200 µg and 300 µg of protein was taken and adjusted to 1 mL in an Eppendorf tube. This volume was used with all the other fractions for consistency. 333 µl of 100%(w/v) trichloroacetic acid solution was then added to each fraction such that the final concentration of trichloroacetic acid was at least 25%(w/v). The fractions were then incubated at 277 K for at least 2 hours. The precipitated protein was then obtained by centrifugation at 13 000 rpm for 7 minutes. The supernatant was then carefully discarded. 1 mL of acetone / 5 mM HCl solution was added to the pellets of protein which were then re-suspended by vortexing.

The pellets were spun down as described above and again the supernatant was discarded. 1 mL of acetone was then mixed with each fraction. The pellets were recovered again as described above and dried in a speed vacuum desiccator (Savant, Speed vac sc 100). They were then dissolved in SDS-PAGE sample running buffer (Table 2.6) and if necessary adjusted to the correct pH with 1-2 µL of Tris base solution. Before being loaded onto the gels the samples were boiled for 1 to 3 minutes to ensure that the protein was completely denatured and spun at 13 000 rpm to ensure that any particulates were not loaded onto the gels.

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tris-HCl pH 6.8</td>
<td>62.5 mM</td>
</tr>
<tr>
<td>Glycerol</td>
<td>10 %</td>
</tr>
<tr>
<td>Mercaptoethanol</td>
<td>5 %</td>
</tr>
<tr>
<td>or DTT</td>
<td>10 mM</td>
</tr>
<tr>
<td>SDS</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Bromophenol Blue</td>
<td>0.05 %</td>
</tr>
</tbody>
</table>
2.3.5.2 Gel preparation
A 13.5% gel resolving gel was cast using the degassed solutions described in Table 2.4 above and allowed to polymerise for approximately 20 minutes (depending on the ambient temperature) before the 4% stacking gel was cast. This was allowed to polymerise. The wells were rinsed with running buffer before being attached to the running apparatus. Running buffer was then added to the upper and lower tanks and the samples loaded into the wells using a micro-syringe.
A current of 20 mA per gel was applied for approximately 1 hour 15 minutes or until the dye front reached the bottom of the gel.
On completion of the electrophoresis the gels were removed from the plates and the stacking gels removed. The resolving gels were then placed in staining solution (Table 2.5) for at least 1 hour and destained with destaining solution (Table 2.5). The destained gel was then scanned to determine the amounts of material in each band.

2.3.6 Band Quantification by Laser Densitometry
An LKB 2222-010 UltroScan XL laser densitometer was used to determine the amount of protein present in each band on the electrophoresis gels. The densitometer used a Helium-Neon laser at a fixed wavelength of 633 nm to determine the absorbance of the bands.

2.3.7 Experimental Results
The following sections give details of the chromatograms and electrophoresis data.

2.3.7.1 Chromatographic Data
Figure 2.11 shows an example of a complete chromatogram produced from the separations described in table 2.3. Until twenty minutes the peaks correspond to unbound material, ie. unwanted material. This portion of the chromatogram was not required for the deconvolution analysis. Portions of three chromatograms used in the deconvolution analysis are shown in figures 2.17 to 2.19.
In the fourth and fifth chromatograms (not shown) shoulders appear indicating that trace components are present which were not fully resolved in other chromatograms.
2.3.7.2 Electrophoresis Data

Figure 2.12 shows a photograph of the gels obtained for the first separation. The first band appears in the 4th track and this corresponds to Ovotransferrin, the first eluted peak. The second component appears in the first tracks of the second gel (fractions 11 and 12) and this corresponds to ovoglobulin. Finally lysozyme appears as a band towards the bottom of the second gel in fractions 12 and 13.

The identities of the components were confirmed using a molecular weight calibration of the 13.5% gels used in the electrophoresis analysis of the column fractions. The band corresponding to lysozyme ($R_f = 0.66$) was found to have a molecular weight of approximately 14900, the band corresponding to Ovoglobulin ($R_f = 0.24$) was found to have a molecular weight of approximately 45000 and finally the band corresponding to Ovotransferrin ($R_f = 0.1$) was found to have a molecular weight of 76000. These agree with those given by Gilbert (1971) and OROS Instruments (1990) (see table 2.2). A logarithmic relationship between molecular weights and $R_f$ values was found as described by See and Jackowski, 1989.

The data obtained by scanning this gel and those produced for the analysis of two other egg white separations are shown in figures 2.22 to 2.24 together with the peak function obtained by deconvolution for each of the chromatograms.

2.3.8 Deconvolution Analyses of experimental data

In this section experimental data are analysed using the deconvolution algorithm described in section 2.2.4.1.

In the following sections the correct peak model for protein peaks obtained during liquid chromatography will be determined which are then used in the deconvolution of the experimental chromatograms obtained above.

2.3.8.1 Deconvolution of a single peak

In order that the most appropriate model was used in the deconvolution programme both the exponentially modified Gaussian function (section 2.2.1.1) and general exponential function (section 2.2.1.2) were tested with a single protein peak.

Purified lysozyme, 0.06 g, (Sigma (Stock code:L6876), Poole, England) was dissolved in 50 mL of 5 mM sodium phosphate pH 6.0. 10 mL of this
solution was loaded onto the column (see section 2.3.2) and eluted using a sodium chloride gradient of 0 to 2M over 20 minutes at 2.0 mL min\(^{-1}\). The data collected from this separation was read into the deconvolution program (Appendix A1.1) but modified so that only one peak function contributed to the objective function. The algorithm was applied using the exponentially modified Gaussian and general exponential function to obtain the lowest possible objective function.

The results of the single peak deconvolution test are shown in table 2.7, and figures 2.13 and 2.14. Figure 2.13 shows the peak profiles of the experimental data and peak functions obtained using the above procedure. As can be seen from this figure the experimental peak is skewed to the left compared to a Gaussian function.

The general exponential function successfully models this shape of this skewed peak with an objective function value of -0.005. The exponentially modified Gaussian function however only achieves an objective function of -0.051 and does not skew to the left (It is not possible for this function to skew to the left see section 2.2.1.1). Instead a non-skewed peak is approximated to the experimental peak but with its maximum shifted to the left of the experimental peak maximum.

*Table 2.7 Single Peak Deconvolution Results*

<table>
<thead>
<tr>
<th></th>
<th>EMG Peak</th>
<th>GEX Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau = 0.928 ) ( min )</td>
<td>( V_0 = 49.38 ) ( min )</td>
<td>( V_0 = 42.00 ) ( min )</td>
</tr>
<tr>
<td>( \alpha = 0.670 ) ( min )</td>
<td>( a = 8.396 ) ( - )</td>
<td>( a = 16.10 ) ( - )</td>
</tr>
<tr>
<td>( A = 0.257 ) ( AU.min )</td>
<td>( b = 0.544 ) ( AU )</td>
<td>( h = 0.912 ) AU mins</td>
</tr>
<tr>
<td>Area = 0.912 AU mins objective function = -0.051 ( \Delta \bar{x}/\bar{x} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area = 0.918 AU mins objective function = -0.005 ( \Delta \bar{x}/\bar{x} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(For a definition of GEX (General Exponential Function) and EMG (Exponentially Modified Gaussian function) see section 2.2)
Figure 2.14 shows the difference between the experimental and model peak functions at each data point collected. The areas of the model peaks are both less than the experimental peak but as expected the exponentially modified Gaussian is greater in error than the general exponential function, due to its larger objective function value.

The point differences between the model functions and experimental peak are shown in figure 2.14. The areas under these curves give the area difference between the model and experimental peaks. As can be seen from this figure negative and positive area differences will cancel each other out when the area of the whole peak is calculated but when the chromatogram is analysed to determine the optimum fraction (see section 4.1.4) using fractions of the peak this will not occur - it is likely that there will be differences of only one sign.

This analysis shows the greatest agreement that can be achieved between the general exponential and exponentially modified Gaussian functions, and the experimental peaks, for a single peak, ie. there is no uncertainty caused by a neighbouring peak overlapping with it. The addition of extra peaks will cause additional uncertainties which will give rise to higher objective function values than possible for single peaks.

2.3.8.2 Deconvolution of egg white chromatograms

The data obtained from each of the experiments described in section 2.3.4 were converted to an ASCII format and read into the Box-complex programme (Appendix A1.1) in the chromatogram[] array. Since from the above analysis the general exponential function was found to give the best agreement with the experimental peaks this was used in the objective function.

The algorithm was then applied to each set of data until the lowest possible objective function was reached. The maximum time taken for convergence was always less than ten minutes which is approximately one third of the loading and elution time. Therefore this analysis could be carried between separations, as the length of this separation was approximately 50 minutes. The result validates the utility of the approach for at-line control and will be discussed in chapter 5.
2.3.8.3 Experimental Deconvolution Results

This section contains the results of the deconvolution analysis and a comparison with the data obtained from the SDS-PAGE analyses. The parameters obtained by deconvolution for each egg white chromatogram (corresponding to separations 1 to 6 in table 2.3) are given in appendix A2.

For the first chromatogram a series of three figures (figures 2.15 to 2.17) show and explain the format of the following figures for the other chromatograms. The first of this series of figures (2.15) shows the individual peaks obtained for the first chromatogram (table 2.3). The next figure (2.16) also shows the model chromatogram obtained by adding the three peak functions together. The final figure in the series (2.17) additionally shows the actual chromatogram (i.e., the experimental data). The optimisation algorithm attempts to minimize the difference between the actual chromatogram and the model chromatogram. The figures following this (2.18 and 2.19) show the data for two chromatograms (separations 2 and 3, table 2.3) in the same format as figure 2.17.

2.3.9 Conversion of data to mass based units

In order that the deconvolution results may be compared with the electrophoresis data, both must be put into comparable units. The most convenient method is to calibrate both methods so that they may be converted to concentration units.

2.3.9.1 Conversion of Chromatographic absorbance profiles to mass based units

In this work elution profiles were recorded digitally by the PE-Nelson interface at a certain sampling rate, typically one point per second, i.e., a measurement of the absorbance of the protein mixture was made once per second. If the solution within the detector flow cell contains only one protein then a concentration for that solution may be estimated, using the Beer-Lambert Law,

\[ A_i = \varepsilon_i I c_i \quad (2.21) \]

since if the flow cell optical path length \( I \) and the extinction coefficient, \( \varepsilon_i \), are known then the concentration, \( c_i \), can be found. If the volume of the flow cell is known then the mass of protein in the flow cell can be
determined. Thus an absorbance versus elution time basis may be converted to a mass versus elution time basis.

The same method may be used if the proteins in the chromatogram are not fully resolved, providing that the chromatogram has been successfully deconvoluted. This is because deconvolution process effectively determines from the total absorbance profile ie. at a point on the chromatogram:

\[ A_{TOTAL} = l \left( e_1 c_1 + e_2 c_2 + \ldots + e_n c_n \right) \] (2.22)

the absorbances due to the individual components, ie:

\[ A_1 = e_1 l c_1 \] (2.23)

\[ \ldots \]

\[ A_n = e_n l c_n \] (2.24)

The individual components may however not obey the Beer-Lambert law, ie. the relationship between absorbance and concentration may not be linear. If this is the case a polynomial would be required to describe the absorbance behaviour eg.

\[ A_1 = l \left( x_1 c_1^2 + y_1 c_1 + z_1 \right) \] (2.25)

\( x, y \) and \( z \) are constants found by fitting experimental data to equation 2.25. The \( z \) term is typically zero when the absorbance is measured in a spectrophotometer since the absorbance of the buffer is taken to be zero. With chromatographic detectors however the zero absorbance is taken to be the baseline value. Non-linearity of absorbance will cause the mass based peaks to have a different shape from their absorbance based peaks. The overall trace will also be different which will have an effect on the calculation of the optimum fraction.

Levison et al., (1989) observed that the absorbance of ovalbumin solution was non-linear from mass-balancing inconsistencies in data obtained from chromatograms. They noted that seemingly very similar chromatograms observed on an absorbance basis corresponded to very different concentration based elution profiles, after conversion using a polynomial such as equation 2.25. It is therefore necessary to determine the nature
of absorbance characteristics within the range of concentrations which will occur with the separation conditions under analysis.

**Determination of Egg-White components’ extinction coefficients**

A stock solution of between 1 mg mL\(^{-1}\) and 1.5 mg mL\(^{-1}\) was prepared with 5 mM sodium phosphate buffer at pH 6.0 (as used in the chromatographic separations) for each component studied (ovotransferrin, lysozyme, and ovoglobulins). A series of dilutions (with 5 mM sodium phosphate buffer) of these stock solutions were carried out. The absorbance at 280 nm was measured in a 1 cm X 1 cm quartz cuvette using a Beckman DU64 Spectrophotometer (Beckman, High Wycombe, Buckinghamshire, England). The extinction coefficients shown in table 2.8 were found by applying linear regression to the data shown in figure 2.21.

**Table 2.8 Egg-White component extinction coefficients**

<table>
<thead>
<tr>
<th>Component</th>
<th>(\varepsilon) (AU mL mg(^{-1}) cm(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovotransferrin</td>
<td>0.89</td>
</tr>
<tr>
<td>Ovoglobulin</td>
<td>1.02</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>2.02</td>
</tr>
</tbody>
</table>

**2.3.9.2 Conversion of electrophoresis data to mass based data**

Fractions collected from a column contain the total material eluted from the column in a given time period. When the fractions are analysed by electrophoresis, the absorbance of each band may be found. Each band on the gel corresponds to a particular component protein. Providing that the relationship between absorbance and concentration for each protein is known then the amount of each component in a fraction may be found. Since the fraction is collected over a known range of elution volume then the mean concentration of each component may be calculated.

**Gel scanning calibration and accuracy**

In order that the areas obtained by scanning gels may be expressed in terms of protein concentrations, a number of gels were run with known amounts of the components in each well, ovotransferrin, ovoglobulin and lysozyme.
5 μl of each sample of each component was loaded into a 13.5% SDS polyacrylamide gel (table 2.4) which was run as described in section 2.3.5 for column fraction analysis.

Figure 2.20 shows the relationship between scanned area and amount of component loaded onto the gels. Ovotransferrin and ovoglobulin both show a linear relationship between mass and area. Lysozyme however appears to show a plateau at higher concentrations. This may be due to a number of reasons, eg. limited capacity of the gel for the protein - dye complex, an inadequate staining times, loss of material in sample preparation, or precipitation of proteins at higher concentrations. All components show the same stoichiometric binding ratio of dye to protein, in the linear regions, (within experimental error) shown in table 2.9 in contrast to their extinction coefficients.

<table>
<thead>
<tr>
<th>Table 2.9 Egg-white components dye binding characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Ovotransferrin</td>
</tr>
<tr>
<td>Ovoglobulin</td>
</tr>
<tr>
<td>Lysozyme</td>
</tr>
</tbody>
</table>

If the peak shapes in concentrations units determined using the gel scanning are compared with those using the absorbance data and extinction coefficients it is observed that the concentrations determined by gel scanning are lower than those obtained from the absorbance data. This may be for more than one reason. In earlier peaks more than one band can be seen in the gel (see figure 2.12) this will contribute to the absorbance signal but not to the scanned area. Another reason may be that other proteins which contribute to the absorbance signal are not present in the resolving gel (either they remain in the stacking gel or they have passed off the end of the gel) and therefore do not contribute to the scanned area.

Because of the difficulties mentioned above with lysozyme calibration and the discrepancies with the total mass of material, it was decided not to convert the electrophoresis data into mass units since the conversion factors for all three components, within their linear regions, are similar. This means that the peak shapes determined by both methods will be comparable. Instead the scanned data may be compared with the
absorbance data to determine agreement between the peak starts, peak ends and peak maxima for the peaks obtained by gel scanning and those obtained by deconvolution.

2.3.10 Comparison of deconvolution results with electrophoresis data
In this section the results obtained by deconvolution are compared with those obtained by scanning the SDS PAGE gels. In particular the agreement between peak maxima starts and ends found by each technique are compared. Agreement of these three key points may only be found to an accuracy of the spacing of the column fractions collected during the egg white separations (ie. 5mL). A summary of these comparisons is given in table 2.10. This table shows peak to valley ratios for all chromatograms analysed in this chapter. Values of the ratio close to one correspond to a highly overlapping peak pair (ie. the peak height is equivalent to the height of the valley). All experimental and test chromatograms having both peak to valley ratios less than three fail to be deconvoluted accurately. However if only one is low then providing the other ratio is sufficiently large then deconvolution may be successfully applied (eg. the first, second and fifth egg white chromatogram).

Figures 2.22 to 2.24 show the peak functions found by deconvolution and those obtained by scanning the SDS PAGE gels for three of egg white separations described above. The points plotted represent the amount of a component determined by SDS PAGE analysis in a fraction plotted at the fraction centre. A more detailed discussion of the results and figures follows.

- In the first egg white chromatogram (figures 2.17 and 2.22) the three key points agree well despite the large difference in peak heights. This because of the separation between peaks is large enough to provide a sufficiently large amount of information about the central peak, so that deconvolution may be successfully applied.

- The second (figures 2.18 and 2.23) and third (figure 2.19 and 2.24) chromatograms contain peaks which are more heavily overlapped, resulting in higher valleys. For the second chromatogram the valleys are only slightly higher but this significantly affects the accuracy to which the peak functions may be found. Certain parts
of the chromatogram are found well, namely the characteristics of the larger peaks (ie. the first and third peaks).

- The third chromatogram is not well described by its model. This is due to the heavy overlapping of peaks (giving rise to ill conditioning of the problem) and this causes the central peak in the model chromatogram to be very heavily skewed - incorrectly extending into the third peak.

In the other chromatograms (not shown) it was found that deconvolution failed under the following situations:
- appearance of extra partially resolved peaks
- heavy overlapping resulting in no peak valleys.

Under these circumstances poor agreement with SDS-PAGE analysis was found.
Table 2.10 Peak to Valley Ratios for the Central peaks in test and experimental chromatograms

<table>
<thead>
<tr>
<th>Chromatogram</th>
<th>Height of central peak to first valley height</th>
<th>Height of central peak to second valley height</th>
<th>Satisfactory model chromatogram obtained?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Chromatograms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaussian 1</td>
<td>1.9</td>
<td>1.6</td>
<td>NO</td>
</tr>
<tr>
<td>Gaussian 2</td>
<td>9.8</td>
<td>440</td>
<td>YES</td>
</tr>
<tr>
<td>Gaussian 3</td>
<td>4.1</td>
<td>297</td>
<td>YES</td>
</tr>
<tr>
<td>Gaussian 4</td>
<td>3.3</td>
<td>97</td>
<td>YES</td>
</tr>
<tr>
<td>GEX 1</td>
<td>11.8</td>
<td>3.8</td>
<td>YES</td>
</tr>
<tr>
<td>GEX 2</td>
<td>3.0</td>
<td>100</td>
<td>YES</td>
</tr>
<tr>
<td>GEX 3</td>
<td>2.4</td>
<td>1.0</td>
<td>NO</td>
</tr>
<tr>
<td>Egg White Chromatograms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>6.3</td>
<td>1.1</td>
<td>YES</td>
</tr>
<tr>
<td>Second</td>
<td>4.3</td>
<td>1.1</td>
<td>IN PART</td>
</tr>
<tr>
<td>Third</td>
<td>1.5</td>
<td>1.1</td>
<td>NO</td>
</tr>
<tr>
<td>Fourth</td>
<td>2.3</td>
<td>1.1</td>
<td>NO</td>
</tr>
<tr>
<td>Fifth</td>
<td>4.4</td>
<td>1.3</td>
<td>YES</td>
</tr>
<tr>
<td>Sixth</td>
<td>1.1</td>
<td>1.0</td>
<td>NO</td>
</tr>
</tbody>
</table>

2.4 Conclusions

In this chapter a study of chromatogram deconvolution has been described. The limits of the method have been determined in terms of the chromatograms that may be analysed using the curve fitting technique. The exponentially modified Gaussian function did not satisfactorily describe the elution profiles for proteins. A better description of such peaks was obtained using the general exponential function.

For successful deconvolution to be carried out the chromatogram must have an adequate separation of the peaks for the particular range of heights within the chromatogram. The model and experimental chromatograms studied contain three peaks, the central peak being considered as the product. The degree of overlap of the first and final
peaks with the product peak determine whether a good model of individual peak elution profiles may be obtained. It has been found that if the height of the valley between peaks is too high (peak to valley ratio less than three) then this obscures information about both peaks and the true elution of profiles of the peaks cannot be found to an acceptable level of accuracy. If high valleys occur on both sides of the peaks then this further exacerbates the deconvolution problem. A high valley on one side of the central peak may be compensated for by a low valley on the other side of the peak. The importance of the valley height for determining whether successful deconvolution is possible is due to the fact that valley height indicates the degree to which neighbouring peak elution profiles extend into one another.

It has been shown that the appearance of extra peaks in the chromatogram due to resolution of minor components or contaminants affects whether a satisfactory model chromatogram may be obtained. When they do occur the deconvolution algorithm attempts to include the extra peaks within the following peak. This usually results in an excessively skewed peak. The degree of skew is dependent on how well resolved the extra peaks are (ie. over what elution volume in comparison with other peaks the extra peaks are spread).

Whilst a low objective function in comparison with the mean data point value is required for deconvolution an adequate separation of peaks is observed to be more critical in determining whether successful deconvolution can be carried out. Low values of objective function correspond to small area differences between model and actual chromatograms. Chromatograms which are deconvoluted to low objective function values but where the model chromatogram obtained is not a good prediction of the true peak functions are indicative of cases where the deconvolution process is subject to ill conditioning.

Having determined the individual peak functions by deconvolution it is then possible to analyse the chromatogram further to determine the optimum product fraction. In order to do this the first task is to determine which of the peaks determined by deconvolution corresponds to the product component and which to contaminants. The next chapter discusses a possible technique for peak identification using fuzzy logic.
Figure 2.1 A chromatogram containing three peaks. It is not possible to determine the peak starts ends or retention times because the peaks are overlapped.
Figure 2.2 The Exponentially Modified Gaussian Function
Figure 2.3 The General Exponential Function ( $a = 1.0$ )
Figure 2.4 The General Exponential Function ( $a = 10.0$ )
Figure 2.5  The General Exponential Function (a = 25.0)
Figure 2.6 Definition of constraints for the optimisation algorithm using the general exponential function.
Figure 2.7 Box-Complex Algorithm Flow Diagram

start

Generate point in initial complex
Complex of K points

check explicit constraints

check implicit constraints

move point in a distance delta inside the violated constraint

Evaluate Objective Function at each point

check convergence

Replace point with lowest function value by a point reflected through centroid of remaining points

is low point a repeater?

move point 1/2 distance in towards the centroid of the remaining points

no

no low point a repeater?

yes

initial complex generated?

yes

no
Figure 2.8 Peak functions obtained by deconvolution for the first Gaussian test chromatogram compared with the actual peak functions. The peak functions shown are for the lowest objective function obtained. It should be noted that the chromatogram is the sum of these functions.
Figure 2.9  Peak functions obtained by deconvolution for the fifth Gaussian test chromatogram compared with the actual peak functions. The peak functions shown are for the lowest objective function obtained. It should be noted that the chromatogram is the sum of these functions.
Figure 2.10  Bio-pilot Chromatography System
Figure 2.11 An example of a complete chromatogram produced from the separation of egg white. The chromatogram shows unbound material (0-20 mins) and peaks produced during elution (30-60 mins). This chromatogram describes the second separation in table 2.3
The two mini gels used to analyse the first egg white separation described in table 2.3. The fraction size is 2.5mL. Twenty fractions were taken starting at 72.5mL (see section 2.3.3).
Deconvolution results of fitting a single exponentially modified Gaussian function and a general exponential function to the experimental data obtained in section 2.3.8.1
Figure 2.14 This figure shows the point differences between each model peak and the experimental peak shown in figure 2.13
Figure 2.15  The first egg white chromatogram. This figure shows the individual peaks found by deconvolution at an objective function value of -72.3 $\Delta x / x = 0.2\%$.
Figure 2.16 The first egg white chromatogram. This figure shows the individual peaks found by deconvolution at an objective function value of $-72.3 \Delta x / x = 0.2\%$ and the model chromatogram (i.e., sum of individual peak functions).
Figure 2.17  The first egg white chromatogram. This figure shows the individual peaks found by deconvolution at an objective function value of -72.3 (Δx / x = 0.2%) and the model and actual chromatograms.
Figure 2.18 The second egg white chromatogram. This figure shows the individual peaks found by deconvolution at an objective function value of $180.0 \Delta x / x = 0.4\%$ and the model and actual chromatograms.
Figure 2.19 The third egg white chromatogram. This figure shows the individual peaks found by deconvolution at an objective function value of $183.6 \, \Delta x / x = 0.4\%$ and the model and actual chromatograms.
Figure 2.20 Densitometry Calibration

Scanned Area (AU*nm)

Mass Loaded in well (pL)

Ovotransferrin  Ovoglobulins  Lysozyme

Figure 2.20  Densitometry Calibration
Figure 2.21  Extinction coefficients of egg white components
Figure 2.22  Comparison of the peak starts, maxima and ends found by deconvolution and those found by scanning SDS PAGE gels for the first egg white (see table 2.5 and figure 2.17) separation.
Figure 2.23 Comparison of the peak starts, maxima and ends found by deconvolution and those found by scanning SDS PAGE gel for the second egg white (see table 2.3 and figure 2.18) chromatogram.
Figure 2.24  Comparison of peak starts maxima and ends found by deconvolution and those found by scanning SDS PAGE gels for the third egg white (see table 2.3 and figure 2.19) chromatogram.
Chapter 3.
Product Peak Identification.
3 Product Peak identification

In this chapter the second step in the control sequence shown in figure 1.8, that of peak identification, using a fuzzy logic technique will be discussed.

In particular this chapter addresses the application of the fuzzy logic technique to process chromatographic data - either to identify components present in a chromatogram or to identify the fact that deconvolution has not been completely successful in obtaining the elution profiles of the component peaks. (Under such circumstances it may be possible however for the technique to identify which peaks have merged causing deconvolution failure.) Finally the reliability of the identification technique will be investigated.

3.1 Introduction

Once the individual peak elution profiles have been found by deconvolution methods (see section 2.1) it is necessary to identify the elution profile which corresponds to the desired product of the separation. No processing of the data can be carried out before identification which requires the identity of the components, such as the use of extinction coefficients or normalisation of peak areas or elution volumes against for example the first peak. Normalisation using one particular peak assumes that its identity is known.

As discussed in section 1.3.6 there are several methods for identifying sets of data of which the fuzzy logic method is one of the most sophisticated. However despite its level of sophistication it can under some circumstances, fail to identify the peaks within a chromatogram correctly. It is important to study these situations both in terms of the mathematical reasons for this and in terms of the events which would physically lead to such situations, so that the suitability of this method's application to preparative chromatographic data may be tested. It is also necessary to describe and demonstrate how a fuzzy logic system could be implemented, practically, using chromatographic data.

Process situations must be considered which result in merging of peaks to a degree such that material appears to co-elute. As noted in the previous chapter this type of overlap causes problems with deconvolution (ie. when both peak to valley ratios are less than three deconvolution is not successful in describing the elution profiles of the individual components). In the case of co-elution the peak to valley ratio criterion is always exceeded and hence deconvolution is not successful. In such
a situation it is necessary to identify that this has occurred and which
component peaks have co-eluted. This will be examined within this
chapter using the fuzzy logic technique.

3.2 Fuzzy Logic system parameter selection
The following sections describe the selection of parameters necessary to
carry out identification using a fuzzy logic algorithm.

3.2.1 Reference data set selection
An important element of the fuzzy logic identification technique is the
reference set of data. For chromatographic data this will consist of a
retention order and peak area obtained by integrating the ultra-violet
absorbance trace (or trace obtained from another detection system such
as refractive index) corresponding to the particular peak.
Retention order, rather than elution volume was chosen as an
identification characteristic as values from separations carried out under
different separation conditions (eg. from varying the gradient conditions)
could be compared; as is possible with peak area. In a process control
system it be envisaged that gradient conditions could be altered to
maintain a consistent product and thus the use of elution or retention
order is advantageous.
The selection of the data is important in determining the success of the
method since if this data is not correct then the match cannot be carried
out reliably. There are several ways of selecting the peak area reference
data. A series of chromatographic separations can be carried out to
determine the elution order and the peak area. For this to be effective
the peaks should be well separated so that the individual peak elution
profiles are completely visible allowing the peak areas to be determined
accurately without the need to apply a peak deconvolution technique.
Using a peak deconvolution method would introduce extra uncertainty into
the quantification of the peak areas, ie. the size of the difference between
the real and model chromatograms (the objective function) would affect
the accuracy of area quantification. The separation will have to be
carried out using the same chromatographic system to that under control
so that the elution orders remain the same. The column loadings however
do not have to be the same as for the relevant process separation, and
under certain conditions it may be advantageous if they are lower when
determining the reference data set, since this may give rise to better
resolution. As long as the mixture to be separated is of the same composition and the amounts loaded during the calibration separation and the actual preparative separation are known, then the reference data set for the preparative data set may be determined. Absorbance characteristics, including any non-linearities will also be required for each component. This is so that the peak areas expected in the preparative separation can be estimated from the calibration separation. Alternatively the peak areas data may be determined by determining the composition of the mixture by another off-line method and providing the extinction coefficient of the individual components are known then the peak areas expected may be determined. It is also possible to define the reference data set purely empirically so that it matches peaks as expected by a particular individual. This would be achieved by carrying out a series of identifications for a number of data sets which are considered to be typical, changing the reference data set until the desired identification is achieved.

3.2.2 Membership function selection

The methods described above for reference data set selection and measurements made during the preparative separations determine values which are in fact, fuzzy (i.e. they may be subject to variation) but only a single value is obtained. Membership functions (see section 1.3.6) are used to describe the variations which may occur in a measurement by ascribing a level of probability (typically expressed as a value of between zero and one) to a range of values of the measurement. It is therefore necessary to determine the left and right hand spreads of the membership functions corresponding to values of the peak area measurement which correspond to non-zero probabilities. Typically this is done by estimating the expected errors in the peak area measurement which may be found in the standard way. This will define the range over which the membership function is non-zero but a function type must be chosen to describe the shape of the membership function. A parabola or Gaussian function are typically chosen for reasons of ease of use and symmetry (Kaufmann and Gupta, 1990, Otto, Wegscheider and Lankmayr, 1988).

The elution order is essentially a crisp measurement (i.e. the values measured correspond to discrete integer values) and other membership functions, such as triangular functions, are typically used for the description of the variation of this parameter. The function is shaped
such that the maximum (1.0) occurs at the point which corresponds to
elution order observed in the calibration runs and zero at zero elution
order and the maximum elution order (i.e., the number of peaks in the
mixture to be separated - see figure 3.1).

If two measurements are observed together in the trial data set then it
is possible for the fuzzy logic method to identify this. For area
measurements (and most other measurements in other applications) two (or
more if necessary) reference peak areas may be added together and
compared with a peak in the trial data set which is suspected of
consisting of one or more reference peaks. The membership function for
the composite peak is obtained simply, the area being equal to the sum
of the individual peak areas in the composite peak and the function
spread the sum of the individual membership function spreads. This
composite membership function is then treated in the usual way.

The combination elution order must be carried out in a different way
since it makes no sense to add together two orders. Instead either a
mean order or the highest or lowest elution order should be used. The
choice will have an effect upon the outcome of the identification. The use
of a mean or the highest elution order membership function may ignore
significantly different lower elution orders. This is important since it is
unlikely that peaks which do not neighbour one another will overlap. For
this reason the lowest elution order membership function is preferred.

The following paragraphs describe the membership functions used in the
identification system used in this thesis.

The membership function used for the following examples is a parabola (as
suggested by Kauffmann and Gupta, 1990). The function for $m_s$ is defined
as:

$$m_s = \frac{-(\text{area} - \text{area}_m)^2}{a} + 1 \quad \text{for} \quad (-\text{area}_m - \sqrt{a}) > \text{area} < (-\text{area}_m + \sqrt{a})$$

(3.1)

$$m_s = 0 \quad \text{for} \quad (-\text{area}_m - \sqrt{a}) < \text{area} > (-\text{area}_m + \sqrt{a})$$

(3.2)

Where $\text{area}$ is the area variable measured by integrating the elution
profile of a particular compound and $\text{area}_m$ is the area of the given peak
under investigation, $a$ is the parameter which defines the membership
function left and right hand spreads. The size of $a$ may be found
deciding upon the percentage deviation from $\text{area}_i$ at which the
membership function reaches zero. For example if $\text{area}_i$ was equal to 30.0
and a ±10.0% spread was to be used the value of $a$ would be 9.00 (10% of
30.0, squared).

A spread of ±10.0% is used in all examples in this chapter.

As stated in the equations above, the function will reach zero when $\text{area} = \text{area}_i ± /a$. All values of $m_j$ below zero are taken to be zero.

If after matching single trial peaks with single reference peaks there
remain some trial peaks and reference peaks which are un-matched then
combinations of reference peaks are used. The $\text{area}_i$ parameter is found
by adding together those of the reference peaks comprising this
composite peak.

The spread parameter, $a$, of the composite peak may be found as
described above and will be equal to the sum of the spreads of the peaks
comprising the composite peak. In contrast the elution order of peaks
within the reference data set is taken as a fuzzy value so that the effect
of changes in elution order may be quantified. The elution order
membership function has been defined at each trial peak $r_i$ (compared
with the $r_r$ peak in the trial data set) by equation 3.3:

$$m(t) = 1 - \frac{|r_i - r_r|}{\#\text{TRIAL} - 1}$$

(3.3)

Where $\#\text{TRIAL}$ is the number of peaks in the trial data set and $r_i$ and $r_r$
are the elution orders of the peak in the trial data set and the $r_{th}$ peak
in the reference set respectively.

As described in section 3.2.3 the match criterion for elution order is
simply the lower of the two elution order membership values for the two
peaks under comparison (i.e. the value for the reference peak since this
will always be equal to or less than that for the trial peak).

In comparison with area membership functions dealing with overlap elution
order requires that the mean or one or other of the elution order
membership functions is used when combining reference peaks. In this
implementation the lower of the elution order membership functions is
used.

### 3.2.3 Matching techniques

In order to test whether a good match exists methods must be available
to test whether the fuzzy numbers are similar and to quantify any
similarity. Two methods are available.

3.2.3.1 Matching Criterion

The first matching criterion is based upon the overlap between the two fuzzy sets representing the measurements. The area of the intersection of the two fuzzy sets is divided by the size of the larger of the two sets giving a number which may range between zero and one. This parameter takes into account the shapes of the functions, since the overlap (i.e. amount of intersection) of membership functions compare two numbers. If two fuzzy numbers \( m_A(x) \) and \( m_B(x) \) are compared then the match criterion is defined by:

\[
\frac{\text{match criterion}}{\text{area}} = \frac{\text{area} (m_A \cap m_B)}{\max(\text{area} m_A, \text{area} m_B)}
\]  

(3.4)

Thus if the membership functions form congruent figures the criterion may vary between zero and one, otherwise it will vary between zero and the ratio of the two areas.

Changing the spread of the membership functions will increase the match criterion between them (the width of the peaks becomes so great that numerically the area measurements become similar). Theoretically the widths may be increased without limit to increase the match criterion. However there is a limit above which it is unreasonable to increase the spread. This is because it defines the allowable deviation of the area measurement from the correct value. Often this is defined as a percentage of the area (or other measurement). This cannot reasonably exceed 100% and values above 20% may be considered exceptional. Also if high values are used the method loses selectivity since a larger range of peak area may produce a high match criterion, i.e. virtually any measurement will correspond to a good match.

For a pair of peaks to be considered a good match the membership function must be above a threshold level. The selection of the threshold value for the matching criterion is also important in determining the success of the method and to a certain degree has a similar effect as the size of the error used to determine the left and right spread of the membership functions as it also defines the region where a good match exists. The choice of threshold is usually purely empirical, but should take into account the size of the difference in area allowable between trial and reference peaks for a good match to be assumed. Thus a very low
threshold would reduce the selectivity of the method and many peaks could be incorrectly matched with a trial peak.

The threshold for the overall match criterion in this thesis was set at 0.75. If the match criterion between a trial peak and a reference peak(s) exceeds this value then a good match between the peaks is assumed. A value of 0.75 shows that 75% of the characteristics of the two peaks under comparison match.

### 3.2.3.2 Possibility theory

A second criterion which is less reliable than the matching criterion is the possibility theory. This is less realistic since it does not take into account the shapes of the functions. In this method the two fuzzy sets under comparison are transformed so that any overlapping regions take the lower of the two function values at each point x value within the set. Then the maximum of this transformed set is taken as 'the possibility law'.

Expressed mathematically this can be described by the following equation:

$$\text{poss}_H A = \bigvee_{x \in R} (m_A(x) \land h(x))$$

(3.5)

where $m_A(x)$ and $h(x)$ are the membership functions being compared. This criterion is not used in this work as it does not take into account the shape of the membership functions and is more complex to compute without any improvement in matching reliability (Dubois, 1988).

### 3.2.4 Failure modes of the fuzzy logic identification technique

The match criterion between two fuzzy numbers is determined by calculating the area of intersection of the two fuzzy numbers. (The match criterion may alternatively be expressed as a function of difference in peak area (or other measurement) versus match criterion value and so two different peak area values may be readily compared.)

The identification method uses a combination of peak areas and elution ranks to identify peaks. It assumes that the proportions and elution order do not vary significantly from separation to separation.

The following sections describe the ways in which this method can fail and the process situations which could cause these failures.

The analysis considers the effects of certain physical occurrences (eg. similar peak areas, extra peaks and overlapping peaks eg. peaks which...
are described by the deconvolution process as one single peak but actually consist of two) on the peak area membership function and peak elution order membership functions separately.

3.2.4.1 Single peaks

Following successful deconvolution (and peak integration) it is possible for the fuzzy logic method to fail if more than one peak exists in either the trial or data set with a similar area. These similar areas may be either due to contaminants entering the trial data or if for example two trial peaks exist which have areas which are above the matching threshold in two peak area membership functions in the reference data set. Additionally it may be due to imprecise deconvolution. Whether this will occur depends on the closeness of the reference peaks areas and the left and right hand spreads defined in the membership functions.

By sorting the peaks into size order it is possible to test whether the reference membership functions of peaks with a similar size are such that the match criterion functions for the comparison of the two individual trial peaks with one particular reference peak will overlap with each other. Of particular importance is the region above the threshold (see figure 3.2), since it is in this region that the method assigns a good match between trial and reference peaks. The probability of this difficulty arising is proportional to the size of the intersecting area above the threshold.

Also of importance is the overlapping of reference peak membership functions with each other since this could lead to the overlapping of match criteria described above. With the parabolic membership functions described in equation 3.1 this will occur if:

\[
    \left( \text{area}_{m1} + \sqrt{a_1} \right) > \left( \text{area}_{m2} + \sqrt{a_2} \right)
\]

(3.6)

Where \( \text{area}_{m1} > \text{area}_{m2} \) which are two peak area measurements.

Another problem which may also occur is the overlapping of trial membership functions with more than one reference peak membership function.

The matching algorithm used in this thesis takes the peak which gives the highest match function used as the best identification, but this is not necessarily always correct. A better solution may be to identify all peaks using all possibilities and to select the total identification set, i.e. the selection of all trial peaks and the corresponding reference peaks chosen
by the identification system, which has the highest mean match function value. However this would only be necessary in very complex systems where there are many peaks with similar characteristics, or where there is a possibility that difficulties with overlapping peaks may occur.

3.2.4.2 Overlapping peaks and combinations of peaks

Overlapping trial peaks with areas similar to that of a single peak in the reference data set will also cause difficulties for similar reasons to those for single peaks as mentioned in the previous section. In this context overlapping peaks refer to a peak which consists of two (or more peaks). Such a peak may occur where deconvolution has been applied and failed to produce an accurate description of the component elution profiles (ie. the deconvolution algorithm describes as one peak what is really two peaks).

Overlapping trial peaks may lead to unresolvable peak combinations ie. trial peak combinations which could be made from a number of different reference peak combinations. When such a situation occurs the method used within this thesis will, as in other situations, take the combination which gives the highest match function value. This may not give the correct identification. In order to check whether this is a problem analysis of match function intersections may be carried out. This is achieved by testing all combinations of two (or more) reference peaks with each other as suggested above for single reference peaks. This is only meaningful if identification has been carried out using single reference peaks, allowing those reference peaks which have been assigned to trial peaks, to be ignored when making combinations.

Non-overlapped peaks in either the trial or reference data set, which have a similar area to a combination of peaks in the trial data set, may also cause matching problems, since the matching routine must choose one or other of the combinations of peaks. At present the system attempts to identify trial peaks as single reference peaks before trying combinations of reference peaks. This is justified since single peak areas will be more certain (ie. less fuzzy) than combinations of peaks (see below), however it does not guarantee that the correct identification will be made. For example, if in a particular identification analysis there are two trial peaks with similar areas, one of which consists of two reference peaks and another which consists of one reference peak, identification may not be successful. This is because the single reference peak may give a match criterion above the threshold whereas the combination of
reference peaks may actually correspond to the correct match.

3.2.4.3 Extra peaks
If extra peaks occur then it is likely that deconvolution will fail, as the number of peaks in the model chromatogram used by the deconvolution technique will be too small. However if the extra peak was taken into account by the deconvolution process then the outcome of an identification will be that a number of peaks equal will remain unidentified. This number will equal the number of extra peaks but the unidentified peaks will not necessarily correspond to the extra peaks. The effect of extra peaks on the elution order match function must also be considered ie. the ability of the matching routine to match correctly when extra peak(s) occur between two identifiable peaks. At present the method uses a normalisation technique to match the elution rank start and end points in the trial and reference data sets ie. non-integer increments are used in the elution rank between the first and last peaks. Extra peaks may occur in two different situations, either before the trial peak under examination or after. The former of these two situations will potentially cause greater difficulties in identification since the trial peak itself will have an altered elution rank whereas in the later situation only those following it will have a different elution rank.

3.3 Matching Worked Examples
The following section shows by means of worked examples, the details, of how the fuzzy method works and may fail. Using computer generated chromatograms allows the complicating effects of differing peak areas, orders and numbers of trial peaks to be investigated without the additional effects of noise or errors arising from integration or deconvolution routines. A set of data was generated so that the potential limitations described in section 3.2.4 could be illustrated. The data is shown in table 3.1 and represented in figure 3.3.
Table 3.1: Reference data set: computer generated chromatograms

<table>
<thead>
<tr>
<th>Peak No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>30.0</td>
<td>50.0</td>
<td>5.0</td>
<td>33.0</td>
<td>35.0</td>
</tr>
</tbody>
</table>

( Arbitrary Units )

In the data set there are two possible sources of problems. These are of the first peak and the fourth peak, and the fourth peak and the fifth peak respectively and arise because these pairs have similar areas. Also the fourth and fifth peak are neighbouring one another and so will have similar elution order membership functions. The latter pair is potentially a more serious problem since in the calculation of the overall match criterion both area and elution order membership functions are used and so this makes it more likely that the overall match functions will overlap in the region above the threshold.

These potential problems may be seen in figure 3.4 showing the shape of the peak area membership functions for the reference data set. The first peak's membership function, with an area of 30.0, overlaps significantly with the membership function of the fourth peak (area = 33.0). Significant overlap also occurs between the membership functions of the fourth and fifth peaks.

The above analysis only includes the reference peak data and it is possible for trial peaks to occur which have a significant bearing upon this analysis. For example, even though the first and fifth peaks' membership functions do not overlap significantly it is possible that a trial peak may occur in a chromatogram which overlaps equally with these two peaks' membership functions. If such a situation occurred then clearly the order would become important and for this reason an elution order membership function is used in addition to the area membership function. Thus the separation between the peak membership functions is also important in determining whether incorrect matching may be achieved since if the membership functions are well spaced then it should not be possible for a trial peak to overlap with two reference peak membership functions simultaneously.

A set of data for a computer generated chromatogram which matches well with the reference data set is given in table 3.2 together with the corresponding individual elution order and area membership functions and overall matching functions in tables 3.3 to 3.5.
Table 3.2 Peak data giving a successful identification

<table>
<thead>
<tr>
<th>Peak No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>area (Arbitrary Units)</td>
<td>29.8</td>
<td>49.0</td>
<td>4.90</td>
<td>33.1</td>
<td>35.1</td>
</tr>
</tbody>
</table>

Table 3.3 Elution order matching criteria

<table>
<thead>
<tr>
<th>Reference Peaks No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1 Peaks No.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 3.4 Peak area matching criteria

<table>
<thead>
<tr>
<th>Reference Peaks No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1 Peaks No.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>*</td>
<td>*</td>
<td>0.57</td>
<td>0.27</td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>0.91</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>*</td>
<td>0.91</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>*</td>
<td>*</td>
<td>0.99</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>0.27</td>
<td>0.99</td>
<td>*</td>
<td>0.74</td>
<td>0.99</td>
</tr>
</tbody>
</table>

(* area membership functions do not overlap, \( m_j \) is effectively zero)

Table 3.5 Overall Matching Criteria

<table>
<thead>
<tr>
<th>Reference Peaks</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1 Peaks No.</td>
<td>0.99</td>
<td>0.38</td>
<td>0.25</td>
<td>0.41</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>0.38</td>
<td>0.96</td>
<td>0.38</td>
<td>0.25</td>
<td>0.13</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.38</td>
<td>0.96</td>
<td>0.38</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>0.41</td>
<td>0.25</td>
<td>0.38</td>
<td>0.99</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>0.14</td>
<td>0.13</td>
<td>0.25</td>
<td>0.38</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Table 3.5 above shows the overall match criteria for the comparison of the trial data set used in this section with the reference data set. As required, the match criterion is only higher than the threshold of 0.75 for the expected peaks (indicated by the bold figures). This signifies a successful identification. There are no possible ambiguities as the overall matching criteria do not overlap in the regions above the threshold. The successful match is depicted in figure 3.5. Each data set (represented by one marker) on this figure represents one reference peak match with each trial peak (represented on the x-axis). A successful and unambiguous match can be seen to have occurred since each data set is above the threshold at only one trial peak number (x-coordinate) and only one data set (i.e., reference peak) is above the threshold at each trial peak number, i.e., each trial peak matches with one reference peak only.

**Overlapping Peaks**

In order to test the methods ability to work with severely overlapped peaks the first set of trial data, listed in table 3.1, was transformed so that the first and second peak in the chromatogram were completely overlapped (table 3.6).

<table>
<thead>
<tr>
<th>Observed Peak No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformed Chromatogram</td>
<td>1+2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

The following overlapping peak pairs may potentially cause identification problems because they possess similar areas:
Reference peaks numbers 1 and 3, total area 34.7
Reference peaks numbers 3 and 4, total area 38.0
Reference peaks numbers 4 and 5, total area 40.0

Also unless the overlapping peaks coincides with the appearance of extra peaks there will be a reduction in the number of recorded peaks and this will in turn affect the elution order membership function. The results of the matching are shown below in table 3.7 which shows the matching of the composite peaks with all possible combinations of the reference peaks. It is important to note that table 3.7 shows the overlapped peak match as combinations of single reference peaks only. It is assumed that all other peaks in the trial chromatogram have already
been correctly matched as in the previous example.

Table 3.7 The transformed data set double peaks, area matching criteria

<table>
<thead>
<tr>
<th></th>
<th>Reference peaks</th>
<th>1+2</th>
<th>1+3</th>
<th>1+4</th>
<th>1+5</th>
<th>2+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1+2</td>
<td>0.94</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Reference 2+4</td>
<td>2+5</td>
<td>3+4</td>
<td>3+5</td>
<td>4+5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial 1+2</td>
<td>0.89</td>
<td>0.79</td>
<td>*</td>
<td>*</td>
<td>0.26</td>
<td></td>
</tr>
</tbody>
</table>

(* area membership functions do not overlap, \( m_x \), is effectively zero)

Table 3.7 shows that using peak area alone the overlapped peak is found to consist of reference peaks 1 and 2.

Since with overlapping peaks the number of trial and reference peaks is unequal, non-integer elution ranks must be used. The fact that a different trial peak in reality consists of two reference peaks has no effect upon the calculation of the elution order match criterion since this is only based on the position of peaks. When two (or more) reference peaks are combined the elution order match criterion is taken as the lowest order of the two peaks. The following table 3.8 shows the elution criteria of four trial peaks matched with five reference peaks as if they were all single peaks.

Table 3.8 Overlapped peaks elution order match criteria matched as single peaks

<table>
<thead>
<tr>
<th>Reference Peaks</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1 Peaks</td>
<td>1.00</td>
<td>0.67</td>
<td>0.33</td>
<td>0.00</td>
<td>(-0.33)</td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>1.00</td>
<td>0.67</td>
<td>0.33</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
<td>0.67</td>
<td>1.00</td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.33</td>
<td>0.67</td>
<td>1.00</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Taking the lower of two elution order match criteria for double peaks one can form table 3.9.
Table 3.9 Elution Order Match Criteria for Double Peaks

<table>
<thead>
<tr>
<th>Ref Peaks</th>
<th>1+2</th>
<th>1+3</th>
<th>1+4</th>
<th>1+5</th>
<th>2+3</th>
<th>2+4</th>
<th>2+5</th>
<th>3+4</th>
<th>3+5</th>
<th>4+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial Peaks 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.33</td>
<td>0.00</td>
<td>(-0.33)</td>
<td>0.33</td>
<td>0.33</td>
<td>(-0.33)</td>
<td>0.00</td>
<td>(-0.33)</td>
<td>(-0.33)</td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>0.67</td>
<td>0.33</td>
<td>0.00</td>
<td>0.67</td>
<td>0.33</td>
<td>0.00</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.67</td>
<td>0.67</td>
<td>0.33</td>
<td>0.67</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

It is worth noting that by taking the lower of the two elution order match criteria that some of the less likely combinations, such as peaks 1 and 3, 1 and 4, and 1 and 5 which consist of peak pairs which are not neighbouring, are virtually prevented from being selected as a good match. This is desirable since it is very unlikely that peaks which do not neighbour one another could combine to form a composite peak.

Combining the elution match criteria with the area match criteria gives the overall match criteria, listed in table 3.10.

Table 3.10 Overall Match Criteria for the transformed data set

<table>
<thead>
<tr>
<th>Reference</th>
<th>1+2</th>
<th>1+3</th>
<th>1+4</th>
<th>1+5</th>
<th>2+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1+2</td>
<td>0.80</td>
<td>0.17</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>Reference 2+4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td>Trial 1+2</td>
<td>0.17</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
</tr>
</tbody>
</table>

From the above tables it can be seen that the identification method correctly identified the overlapped peak as consisting of reference peaks 1 and 2.

It is also important to note that the overlapped peaks would not, in these cases, have been identified as single reference peaks, since the overlapped peak areas were, in general, too different from the single reference peaks.

**Extra Peaks**

The presence of extra peaks may cause problems. The elution ranks will be changed in a way similar to that of overlapping peaks, since there will be a different number of trial and reference peaks. As an example the trial data set in table 3.1 was used but with the addition of an extra peak.
after the second peak with an area of 15.0. This data set is listed in table 3.11.

Table 3.11 Trial data set with an extra peak

<table>
<thead>
<tr>
<th>Peak Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>area (Arbitrary Units)</td>
<td>29.8</td>
<td>49.0</td>
<td>15.0</td>
<td>4.9</td>
<td>33.1</td>
<td>35.1</td>
</tr>
</tbody>
</table>

The following tables (tables 3.12 to 3.14) show the elution order and area membership functions.

Table 3.12 Elution Order Match Criteria

<table>
<thead>
<tr>
<th>Reference data set</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Trial data Set</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>(-0.25)</td>
<td>0.00</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3.13 Peak Area Match Criteria

<table>
<thead>
<tr>
<th>Reference data set</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Trial data Set</td>
<td>0.97</td>
<td>*</td>
<td>*</td>
<td>0.56</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td>0.92</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>*</td>
<td>0.92</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5</td>
<td>0.56</td>
<td>0.56</td>
<td>*</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>0.23</td>
<td>0.23</td>
<td>*</td>
<td>0.75</td>
<td>0.99</td>
</tr>
</tbody>
</table>
As can be seen from table 3.14, the third extra peak which has been introduced into the data set has not been identified with any of the reference peaks. This is because the area membership function of this extra peak does not overlap with any of the trial peak membership functions, even at levels giving match criteria below the threshold level. For this to happen the extra peak would have to have an area within the range of one of the other peaks (see equation 3.5). The fifth trial peak has been matched to the fifth reference peak but should be matched to the fourth reference peak. The sixth trial peak is matched as the fifth reference peak (see figure 3.6). This is because the identification algorithm matches in peak number order. Hence the reference fifth peak will be unavailable for matching with the sixth trial peak and latter will remain unmatched unless a matching algorithm as described in section 3.2.4.1 is used which calculates and tests all identification sets and selects the set with the highest mean match criterion. The main reason for this difficulty is the similarity of the fourth and fifth reference peaks areas (see figure 3.4) and the fact that they are neighbouring. This results in the overall match criteria of the fifth trial matched with the fifth reference peak and the fourth trial matched with the fifth reference peak to overlap above the threshold. Also it should be noted that the overall match criteria of the trial peaks after the extra peak are lower than those in the 'successful identification' (table 3.4) data set. This is because the elution order membership functions are lowered by the introduction of an extra peak.
Conclusions

It can be concluded that peaks in a chromatogram may be identified successfully using peak area and retention number. It is also possible to predict if identification problems are likely to occur. Additionally the fuzzy logic technique has been shown to be able to identify whether a peak is comprised of more than one component and if so what the constituents of the overlapped peak are. If deconvolution has been applied to the chromatogram this may be used as a check on the accuracy of the deconvolution algorithm.

Using simulated data the limitations of the fuzzy logic method have been illustrated. These limitations arise from similarities in peak characteristics particularly when these occur in more than one characteristic (as would happen with neighbouring peaks having similar areas). Similar peaks may be contained within the separation mixture. If this is the case then it is possible to determine by an analysis of the match criteria in the reference data set what degree of similarity is necessary between peaks before matching difficulties will occur. By analysis of the reference chromatogram it is also possible to determine the degree of similarity required before a contaminant peak will cause identification problems by determining what peak areas and positions would overlap with the reference match criteria.

3.4 Egg White Separations

In this section the peak area and elution order data obtained from deconvolutions of data for the separation of raw hen egg-white are used in conjunction with the fuzzy logic identification technique to analyse the behaviour of the technique when applied to peak area data and elution order data obtained from deconvoluted chromatograms (see section 2.3 for Materials and Methods). As previously stated in sections 3.2.1 and 2.2.5 the process of deconvolution will introduce uncertainties into the determination of peak areas. The size of the uncertainty will be dependent on the size of the objective function (see section 2.2.5). The nature of the chromatogram under analysis (namely the relative heights of the peaks and the degree of overlap, see section 2.2.5 and 2.2.6) will also affect the accuracy to which the peak areas may be found. It is important to note that the identification technique must use the areas in
non-mass based units (i.e. AU.mL or mV.mL) since to convert to mass units would require that the components be identified so that the correct extinction coefficient could be applied to the raw peak data. In contrast to peak area peak elution orders should be relatively certain since these are discrete integer values which are easily determined whereas with peak area a precise knowledge of the individual component's elution profile is required.

3.4.1 Data Handling and Peak Integration
The detector signal was recorded digitally using a PE-Nelson Model 970 Intelligent Interface (Perkin-Elmer, Beaconsfield, England) and Turbochrom II software running on an IBM Model 55SX Computer (IBM Portsmouth, England). The chromatograms produced in these separations contain peaks which were too overlapped to be accurately integrated using the Turbochrom software (Papas and Tougas, 1990). Instead the peak area data for each egg white separation chromatogram (see section 2.3) was obtained by integrating each peak function separately (see appendix A1) using the parameters found by deconvolution (see section 2.3.).

3.4.2 Fuzzy Logic System Parameters Selection
In order that the matching routine may be applied it is necessary that a reference data set is available. This was determined from knowledge of the composition of hen egg-white determined by Gilbert (1971) and knowing that a sample volume of 10 mL was loaded the sample will contain 14.0 mg of Ovotransferrin, 7.0 mg of Lysozyme and 6.0 mg of Ovoglobulin. Converting these masses to an expected area measured on an absorbance basis, using the extinction coefficients determined in section 2.3.9.1, and millivolt basis\(^1\) gives the values shown in table 3.15.

\[^1\text{An output voltage of 100 mV from the UV detector corresponds to an absorbance of 2.0 AU.}\]
Table 3.15 Expected Egg White Chromatogram Peak Area Data (Reference peak data)

<table>
<thead>
<tr>
<th>Peak Number</th>
<th>Area ( AU . mL )</th>
<th>Area ( mV . mL )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.11</td>
<td>105.5</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>48.9</td>
</tr>
<tr>
<td>3</td>
<td>2.21</td>
<td>110.5</td>
</tr>
<tr>
<td>Total</td>
<td>5.30</td>
<td>264.9</td>
</tr>
</tbody>
</table>

3.4.3 Egg-white Separation Results

The next section describes the results of the egg white separations and the results of the identifications using the peaks found by deconvolution.

The parameters for the peak functions found by deconvolution are listed in table A2.11 and these parameters were integrated to produce the peak area data listed below in table 3.16. The deconvolution of each separation was carried out to a different degree of accuracy in terms of the preciseness to which the overall peak shapes were found - i.e., the positions of the peak maxima, starts and ends as well as the peak areas. (A discussion of the effectiveness of the deconvolution technique in determining the correct peak shapes for the egg white chromatograms is given in section 2.3.10.)

Clearly the ability of the deconvolution technique to determine accurately the peak areas will affect the subsequent ability of the fuzzy logic identification technique to identify accurately the peaks. The information determined by the identification routine may, as well as being used to identify the components in the separation, be used to comment on the effectiveness of the deconvolution algorithm. A low mean match criterion between a trial deconvoluted chromatogram and the reference chromatogram data indicating a poor result for both the matching algorithm and deconvolution algorithm.
Table 3.16 Deconvoluted Egg White Chromatogram Peak Area Data (trial peak data. Objective function < -183.6).

<table>
<thead>
<tr>
<th>Egg White Chromatogram Number</th>
<th>Peak Area (mV mL)</th>
<th>Total (1+2+3)</th>
<th>Actual Total Area¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>118.0</td>
<td>45.5</td>
<td>118.3</td>
</tr>
<tr>
<td>2</td>
<td>103.7</td>
<td>47.4</td>
<td>116.3</td>
</tr>
<tr>
<td>3</td>
<td>100.8</td>
<td>70.0</td>
<td>102.4</td>
</tr>
<tr>
<td>4</td>
<td>175.0</td>
<td>51.5</td>
<td>110.9</td>
</tr>
<tr>
<td>5</td>
<td>100.0</td>
<td>45.8</td>
<td>114.6</td>
</tr>
<tr>
<td>6</td>
<td>48.9</td>
<td>39.8</td>
<td>134.7</td>
</tr>
</tbody>
</table>

¹ Calculated by integrating the whole chromatogram

Such combinations of poor results may be due to a number of reasons. The composition of the mixture separated may change - either the amounts or the identities of the components. This may lead to a change in the total chromatogram area. In this situation it would be reasonable to expect a low overall match criterion since one or more of the areas of the peak functions comprising the chromatogram will be different. If the total area of the trial chromatogram is the same as the reference chromatogram but with only a low mean match criterion then one may assume that the deconvolution is inaccurate. Poor deconvolution may be because the constraints used in the deconvolution process were inappropriate (see section 2.2.4.1), the model used to describe the peak elution profiles is inappropriate (these should be checked first - see section 2.3.8.1), or the chromatogram is not suitable for analysis by deconvolution (see section 2.4).

The results of the identification of the deconvoluted peaks are listed in tables contained in appendix A3. The first, second, and fifth chromatogram pose no serious problems to the matching algorithm and the components are correctly matched to their template component in the reference data set. However three of the chromatograms listed in table 3.16 (numbers 3, 4 and 6) contain peak areas which differ significantly from areas in the other chromatograms. The reasons for these differences are discussed below together with comments on the subsequent effect.
upon the identification of the peaks.

The third egg white chromatogram contains a central peak which has an area approximately 25 mV.mL above the other reliable measurements of the central peak. The total area for this chromatogram is not significantly different from the total areas of most of the other chromatograms and this implies that if the area of the central peak (and hence that of the third peak) was correctly determined then the identification of the peaks would be more certain. (The reasons for the inaccurate deconvolution are discussed in sections 2.3.10 and 2.2.6.) The central peak in this chromatogram is matched with the central peak in the reference data set but only with an overall match criterion of 0.56. Both of the other peaks are matched with much higher overall match criteria (in excess of 0.9).

The fourth chromatogram has a total area well in excess of the others and the area of the first peak is much larger than in other chromatograms (even as a proportion of the total chromatogram). The peak shape obtained by deconvolution is not an un-reasonable one and a larger peak area would be expected since the peak height is higher than in other egg white chromatograms. This implies that this particular chromatogram may contain either a different proportion of the same components as in the other chromatograms or the chromatogram contains extra or different components. In order to determine which of these situations had occurred required further detailed analysis of the column eluent - particularly components not contained within the molecular weight range that was analysed by SDS PAGE (see section 2.3.5). In addition to these points indicating that the mixture may not be of exactly the same as in the other egg white chromatograms is the appearance of extra peaks in the chromatogram. The identification process gives a very poor match for the first peak as expected given its large area but good matches for the second and third peaks.

The sixth egg white chromatogram is subject to severe overlapping of peaks and consequently all peaks are very inaccurately determined and thus all peaks are poorly matched - particularly the first and second peaks. In fact all peaks are matched with significantly lower overall match criteria.
3.5 Conclusions
The matching algorithm described in this chapter has been shown to successfully identify component peaks within chromatograms using a reference or template chromatogram.

- Determination of the techniques limitations and process situations leading to failure.

The algorithm has been shown to be applicable to several types of chromatograms and its limitations illustrated using simulated chromatographic data. The limitations occur in the main where peak characteristics (namely peak area and elution order) in the reference and or trial data set are similar. This is particularly true when similarity exists between more than one characteristic. The fault which is most likely to cause the technique to produce an incorrect identification is the appearance of a peak of an additional peak in the trial chromatogram which has similar identification characteristics to a peak in the reference data set. Not only does this affect the identification of neighbouring peaks but all peaks in the chromatogram and may cause an incorrect identification. These problems with similarity of peak characteristics are exactly the same as those encountered by a human analysing such a problem. However, by using the fuzzy logic technique this allows a more quantitative analysis to be carried out resulting in a more consistent level identification being achieved.

The methods for determining how similar characteristics need to be before problems occur in the matching procedure have been described.

- Application to chromatograms requiring deconvolution.
Chromatograms containing severely overlapped peaks were also analysed successfully. The success of the analysis is dependent upon the degree of overlap and the relative heights of the peaks within the chromatogram. An important factor with the successful identification of the peaks is the accurate deconvolution of the chromatogram to obtain the peak functions describing the individual component elution profiles (see Chapter 2). This is required so that the amount
of a given component may be accurately determined (i.e., the peak area). Providing that this information is obtainable then the peaks within the chromatogram may be matched accurately with peaks in the reference chromatogram.

- Analysis of severe overlap:
  More severe overlapping has also been considered where the degree of overlap is such that no valley exists between the two peaks and in certain circumstances peaks have been successfully identified. The use of the technique to identify very heavily overlapped peaks (such as those obtained by poor deconvolution) and their composition has been investigated using simulated data.

The following chapter of this thesis considers the combined use of deconvolution data and peak identification techniques in the control of fraction selection for the optimisation of a chromatographic process, the third process listed on figure 1.8.
Figure 3.1 The elution order membership function for a reference peak. The measured rank is X.
Figure 3.2  (a) The match criterion is obtained from the overlap of membership functions
(b) A match is considered to be good if the match criterion is above the threshold.
Figure 3.3 The area data contained in table 3.1 may be represented by this chromatogram.
Figure 3.4  Peak Area Membership Functions for the Reference Data Set listed in table 3.1. The maxima of the membership functions correspond to the peak areas in table 3.1.
The overall match criterion for the matching of the data in table 3.2 with the data in table 3.6. Each data set shows the matching of each trial peak with a reference peak. This figure shows a successful identification since only one data set (i.e., only one reference peak) at each trial peak co-ordinate is above the threshold.
The overall match criteria for the fourth and fifth reference peaks in table 3.7 compared with each reference peak in table 3.1. This indicates that a problem occurs with the fifth trial peak since both the fourth and fifth reference peaks have overall match criteria above the threshold.
Chapter 4.
Control.
4 Control

Once the individual component elution profiles have been obtained (see section 2.1) and the components matched to a template chromatogram (see section 3.1), it is then possible to carry out control actions based on the information gained from these processes. The aim of the work in this thesis is to control preparative separations where the chromatography is carried out to produce material of a defined rather than to effect a high degree of resolution so that the composition of a mixture may be determined. Hence control actions will be instituted to maintain the purity of the product and the productivity (i.e., the amount of material produced per unit time or separation) of material produced. In such a process situation the issue of detector saturation may be critical, since as described in section 2.3.1 the analysis of chromatograms with detector saturation requires the use of an on-line or at-line analytical system such as HPLC. The chromatograms produced from each on-line analysis may then provide the source of data on which to base control decisions. Analysis of the information produced by the analytical system will, as for the process chromatogram, require some prior knowledge of the nature of the chromatogram produced by the at-line system. If the chosen analytical technique is a column-based chromatographic method then the information obtained will be in the form of a chromatogram with (provided the size of material loaded is determined correctly) linear detector response. The nature of this chromatogram may be different from the preparative chromatograms analysed in section 2.3.8. Knowledge of which components correspond to the nearest neighbour(s) on either side of the product is required since the elution order may not be the same as in the process chromatogram as either a different type of chromatography may be used or displacement effects may occur at a process scale due to high loadings used which may affect the elution order of some components. That withstanding the same generic approach as described in this thesis will apply to deconvolution, peak identification, and control.

Control actions which may be taken are the selection of the optimum fraction based on a defined set of productivity and purity criteria and secondly the re-optimisation of the separation should it not be possible to obtain a fraction with the desired productivity and purity. Since deconvolution and fuzzy logic identification require that the whole chromatogram be available for analysis on-line control using these techniques is not possible. The following actions may be taken at-line:
(i) The calculation of the position of the optimum fraction for subsequent separations. If the fraction selection action is to be applied on subsequent separations then it is essential that the variation in overall and component elution profiles from the \( n \)th separation to the \( n+1 \)th separation should not be significant for the action to be effective otherwise information about the position of the optimum fraction will not be applicable from separation to separation.

or

(ii) The selection of the optimum fraction(s) position for the current separation. For the fraction selection to be applied to the current fraction sub-fractions of column eluent must be collected at regular elution volumes. The decision on the optimum fraction(s) would then be applied by determining which sub-fractions should be pooled together to form the product fraction (Dista Products, 1992).

The following sections discuss the capacity to use the algorithms developed in this thesis to effect these two possible approaches to control.

4.1 Fraction Selection
This section considers methods available for the selection of the optimum fraction. Even after deconvolution and peak identification have been carried out further calculations must be performed to select the optimum fraction. Three possible calculation methods are available.

4.1.1 Chromatogram Sub-fractions
With the analysis of chromatographic sub-fractions (and all methods described in this thesis) either a minimum productivity or purity may be set. A series of iterative calculations are then carried out to determine the compositions of fractions of various sizes and with various starting and finishing points. The procedure for these calculations is given in the flowchart (figure 4.1), which is based on a minimum productivity threshold. A table of possible fractions is produced from these series of calculations. This table will include the start and end positions of the fractions, the purity of the fractions and the productivity obtained using each of the fractions. The fraction which is selected as the optimum is
the fraction which has the highest purity for the given productivity threshold.

Selection on the basis of purity is also possible where figure 4.1 is converted to give a flowchart based on a purity threshold (requiring that the terms productivity and purity and exchanged in the figure).

4.1.2 Fractionation Diagrams

An alternative way of visualising the above method is by making use of fractionation diagrams. This method was developed to describe fractional precipitation processes (Richardson et al 1990). The main advantages of this method over direct analysis of the chromatogram are the ability to determine rapidly the purity and yield that may be obtained from many different product fractions taken from the separation.

In the case of fractional precipitation where the goal is to isolate a product protein from other contaminating proteins, the fractionation diagram (see figure 4.2 for an example) is a plot of the product protein versus the total protein obtained by increasing the precipitant concentration i.e. the amount of product protein is related to the amount of total protein. This is an example of an application to a particular unit operation. More generally it is a plot of product substance versus the total material achieved by a separation process at different points during its operation.

For chromatographic separations the diagram may be constructed using the elution profiles of the individual components, found by deconvolution or by using off-line assays of the sub-fractions for the product protein and total protein. There are certain features of the construction of the diagram, described by Richardson et al (1990), which enable rapid analyses and therefore make this diagram suitable for the description of a range of separation processes.

The fractionation curve has a number of useful characteristics which have been described by Richardson, et al, 1990. The line AB on figure 4.2 represents the proportion of protein which is product in the fraction before the separation and the line CD represent the proportion of protein which is product in the fraction which starts at the point C and finishes at the point D.

The yield (and hence the productivity if the separation time is taken into account) may be determined from the curve and is equal to the length of the line FE on figure 4.2 for the fraction represented by the tie-line AB.
4.1.2.1 Construction of the Fractionation Diagram for Chromatographic Separations

As with the previous method of chromatogram analysis fractions of the chromatogram are analysed. If deconvolution data is used then by integrating the peak elution profiles of the product component and neighbouring contaminating peak at regular intervals along the chromatogram the amounts of product protein and total protein may be calculated from the start of the chromatogram to the chosen end point. If desired it is possible to simplify the number of peak functions used in the construction of fractionation diagrams for complex chromatograms, providing nearest neighbours can be identified (see chapter 2). If this is possible then the fractionation curve would be constructed using the nearest neighbour peak functions as the only contaminants. This diagram would however only give purification data based on the purification of the product from the nearest neighbours; no account would be taken of other components.

The areas obtained are then adjusted using the appropriate extinction coefficients to give masses of the product total proteins in the fraction analysed. Each fraction analysed corresponds to a point on the fractionation curve. If the fractions used to construct the fractionation diagram are taken at equally spaced volume (or time) units then unevenly spaced points will occur on the fractionation curve since the x-axis is in mass units. In order to obtain equally spaced points fractions should be taken which increase by equal peak area increments. This is particularly important when analysing the steep central portion of the fractionation curve and here points will be greatly spaced in comparison with other parts of the curve since this determines the maximum purification available from the given separation conditions. In terms of control the non-linear portions are important as these describe where the contaminants lie and mark the boundary with the linear region, which contains the highest purity material.

Construction of the curve in this way means that only bound material is included and any purification achieved during the binding process is therefore ignored. This amount will be determined by the operating conditions and is not a factor that can be controlled when determining the position of the optimum fraction (or by changing elution conditions to alter the separation performance).

Hence only purification achieved during elution which is the most relevant
to fraction selection is considered.

4.1.2.2 Disadvantages of the fractionation curve
There are two problems associated with the fractionation diagram. Firstly the fractionation curve only describes the total amount of contaminant present in comparison with the amount of product, not the actual composition of the fraction obtained. This may be important since it may be more critical to exclude one contaminant peak from a fraction than a second neighbouring contaminant peak. Richardson, (1987) developed an additional diagram (the so called enzyme-enzyme fractionation diagram) to overcome this problem by including two curves where each describes the fractionation of the product protein from the two possible contaminants. This is applicable to a chromatographic separation since a typical separation may be considered (and as used in this thesis) to consist of one product peak with two adjacent contaminant peaks.

The second problem is that the volume of the fraction cannot be determined from the fractionation curve. (This is not a problem in fractional precipitation since the product is in the solid phase.) The length of the tie-line representing the fraction does not give an indication of the fraction volume. This may be significant since a variation in concentration or volume may have undesirable effects upon subsequent separation stages. Generally a more dilute product fraction will cause an increase in the processing time of subsequent separation steps due to the larger product volume. For non-adsorption based separation processes this will also increase the subsequent process volumes and thus affect all unit operations in the same way. The issue of product volume is discussed further in section 4.1.4.

4.1.3 The Fractionation - Concentration Diagram
Within this thesis the fractionation diagram approach has been extended to three dimensions by adding the cumulative volume that has been eluted at the point where the fraction end is defined. This enables the concentration of each fraction to be determined.
The properties available from the diagram are shown in table 4.1 below.
Table 4.1 Properties of the Fractionation - Concentration Diagram

<table>
<thead>
<tr>
<th>Physical Property</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Product Protein (mg)</td>
</tr>
<tr>
<td>X</td>
<td>Total Protein (mg)</td>
</tr>
<tr>
<td>Z</td>
<td>Elution Volume (mL)</td>
</tr>
<tr>
<td>Y-Z gradient</td>
<td>Product Concentration (mg/mL)</td>
</tr>
<tr>
<td>Y-X gradient</td>
<td>Purification ratio(^1) (dimensionless)</td>
</tr>
<tr>
<td>X-Z gradient</td>
<td>Total Protein Concentration (mg/mL)</td>
</tr>
</tbody>
</table>

\(^1\) if the diagram were normalised this would be the purification factor.

An example of this diagram is given in figure 4.3. Although this figure contains all the possible information which may be determined to describe the separation it is not easy to visualise when represented in two dimensions - particularly when comparing two sets of data with similar characteristics. For these reasons in this thesis two figures are used to represent the data. These are the fractionation diagram and the cumulative elution of the product plotted against volume. Examples of these diagrams are shown in figures 4.6 and 4.8. Representing the data in this way still allows the use of all possible data describing a separation whilst enabling comparison of different separations or different deconvolution solutions.

The portion of the curve between points A and B on figure 4.3 (the concentration - fractionation diagram) show an increase in volume only. This corresponds to a baseline on the chromatogram from which this diagram was constructed. The portions of the curve between points B and C and points D and E correspond to parts of the chromatogram where contaminant components are eluted since only an increase in volume and total protein occurs. The central portion (points C to D) corresponds to the product peak as only the product protein and volume parameters increase significantly (The increase in total protein in the region is due in the main to the changes in product protein).

4.1.4 Selection of the Optimum Fraction

This section describes two possible methods of determining an optimum fraction based on either a constant yield or purification factor, since either of these parameters may be more important.

Richardson (1987) developed an algorithm for determining the maximum
purification factor for a given yield. The algorithm starts with the yield line (e.g., FE on figure 4.2) at zero on the total protein axis and the end of the tie line is at the point of the curve giving the desired yield. The purification factor for this fraction is calculated. The yield line is then moved by an increment in the total protein co-ordinate and the purification factor recalculated. This process is repeated until the set yield can no longer be reached. The optimum fraction will be the fraction with the highest purification factor.

If yield is to be maximised then the purification factor may be set at a constant value and the maximum yield obtainable from the fractionation curve determined. This is carried out by starting at the origin and extending a tie-line with a gradient equivalent to the purification factor set to the maximum total protein co-ordinate (i.e., the right-most position). Where this tie-line crosses the fractionation curve corresponds to the end point of a fraction with the desired purification factor. It is possible that more than one fraction could exist which has the same purification factor. This is shown in figure 4.4. Here three possible fractions exist with the same purification factor - one large fraction which may be split into two sub-fractions both with the same purification factor. Although the fraction of product protein to total protein will be equivalent for these three fractions the actual composition of contaminating components will be different. The fraction with the largest yield will clearly be a combination of the two sub-fractions, but this may not necessarily be the most desirable fraction.

Having calculated the yield for one starting point the tie-line is then moved along the total protein axis by a set increment and the process repeated. The yield is calculated for each fraction and the optimum fraction is defined as the fraction which gives the highest productivity for the preset purification factor.

The selection of the optimum fraction from the fractionation-concentration diagram (or the combination of two dimensional diagrams described in section 4.1.3) may be carried out in a similar way as for the fractionation curve (i.e., either a purification factor or a productivity may be set) but since an additional parameter has been introduced an extra optimisation criterion must be introduced. This additional optimisation criterion will either be a volume or a concentration limit which will affect both the yield and volume parameters together (since concentration is calculated from the product yield and fraction volume). Since it is likely that concentration changes will be the least important the selection of the
optimum fraction should use concentration as the least important parameter. There should however be a concentration or volume threshold, ie. a minimum (ie. a maximum volume) should be defined.

The range of possible fractions may be described by a volume shown in figure 4.5. The axes represent the volume of the fraction, the yield of the fraction and the purity of the product. The volume axis will run from zero to the total elution volume of the separation and the purity axis will run from zero to 100%. The yield axis will run from zero to the maximum possible yield. The cuboid shape indicates that at certain regions, ie. towards a corner of the cube fewer possible solutions exist with comparable characteristics. This is as would be expected since the corners represent the extreme conditions, ie. combinations of extremes of two parameters simultaneously. Towards the centre of the cube more potential solutions exist which may be seen by the greater size of the cube at this point. This shows that if more than one parameter is constrained to a higher value then the potential number of fractions with these characteristics is relatively small. Constraints which may be imposed on the quality and nature of the product fraction may also be added to the fractionation diagram as shown in figure 4.5 (planes indicated by the dotted lines). By plotting each of the fractions which can be obtained this diagram may be used as an aid in the selection of the best fraction. Generally corner A is the most desirable position and fractions which lie closest to this location should be selected as it corresponds to a combination of maximum purity and yield but minimum volume. The least optimum position is diametrically opposite (corner B) which corresponds to a high volume, and low yield and purity.

The relative importance of each of the parameters volume, yield, and purity will depend upon the aim and position of the separation in a sequence. The nearer the beginning of a separation the more important is the yield and the less important is the purity as the yield in early purification steps will affect the overall efficiency of the separation sequence. The purity may be gained in following steps but again at the expense of yield. The range of volumes of the fraction that would be acceptable is dependent upon the type of unit operations which follow. Unit operations which may typically follow an ion exchange chromatographic separation are gel filtration, freeze drying, and ultra filtration. All of these will be affected by variations in the volume of the product fraction of the preceding step. Of the unit operations listed gel filtration is particularly sensitive to increases in product fraction volume.
since the capacity of a gel filtration column will be rigidly fixed. If the product fraction increases above the column capacity then this will require another run to separate the excess volume since it cannot simply be added to the standard loading without significantly affecting the performance of the separation. This would increase the processing time greatly - in excess of the proportional increase in processing volume. Similarly reduction in process volumes at the end of a downstream process sequence is also important in terms of minimizing subsequent loads on finishing operations such as freeze-drying.

4.2 Effects of deconvolution accuracy on the fraction curve
As described in section 2.2.5 and 2.2.2 the accuracy of deconvolution will have an effect upon the accuracy to which peak areas are determined. Since fractions are subsequently selected using the peak areas as a basis, the deconvolution accuracy will have an important effect upon the actual purity and amount of product that is contained within the fraction selected. The consequences of deconvolution accuracy on subsequent fraction selection and the resultant fractionation diagrams are discussed in the following section.

As described in chapter two severely overlapping peaks and overlapping peaks with greatly differing heights cause the greatest errors in the determination of peak functions by deconvolution. From a study of the fractionation diagrams of the test chromatograms examined in chapter 2 the areas of the fractionation curves which are subject to the greatest error are caused by the same areas of the chromatogram which prove the most problematic in deconvolution, ie. those areas of the chromatogram which consist of more than one peak function - particularly where one peak is much taller than the other. In general the maximum yield is found correctly as is the maximum purification factor, the main exception to this is where a small central product peak is overlapped by relatively tall contaminant peaks. This is because the central peak in the model chromatogram may be subject to excessive fronting and tailing thus causing the area of the peak to be larger than in reality.

The initial and final curved regions of the diagram are the areas most susceptible to error as they correspond to areas of the chromatogram consisting of two or more component peaks (see figure 4.6).

Two example fractionation diagrams are shown in figures 4.6 and 4.7 the first is for a chromatogram with a high degree of overlap whilst the
second is for a chromatogram with relatively low overlap (both have similar peak heights). As can be seen the curve for the second chromatogram has been more accurately determined. The first was not deconvoluted satisfactorily (see chapter 2).

4.2.1 Effect of deconvolution upon Concentration Data
Unlike amount of product and total protein, volume does not require deconvolution to determine its value at any given point within the chromatogram and hence it is not subject to as great uncertainty as the previous parameters. However since the volume data may be used in conjunction with data concerning the relative amounts of product or total protein to obtain concentration data it is still important to consider the effects deconvolution will have on this parameter.

As described earlier (section 4.1.2) the concentration of product in a fraction is defined by the gradient of the tie line on the product and volume axes of the fractionation - concentration diagram or the gradient of the tie line on the cumulative product elution diagram. Thus the shape of the product peak predicted by the deconvolution algorithm is the only information which affects this since the data used to determine the concentration of a fraction are cumulative elution volume and cumulative eluted product mass at intervals throughout the chromatogram.

The major differences, as with the fractionation curves, are concerned with deviations arising from an inaccurate determination of yield due to problems associated with the deconvolution process as described above. The concentration data is less troublesome to obtain from the chromatogram since the volume term is available directly. Two example cumulative elution curves are shown in figures 4.8 and 4.9, the first is for a chromatogram with a high degree of overlap whilst the second is for a chromatogram with relatively low overlap.

4.2.2 Implications of accuracy deconvolution on fraction selection
This section will examine the implications of the deviations of the fractionation curve, obtained using data obtained by deconvolution, from the true curves.

For the fractionation curve generally there are three possible deviations which may occur. These are deviations in curvature at either end of the curve, non-agreement between the maximum slopes of the curves and non-
agreement between the maximum yields apparently obtainable.
This examination uses the methods of optimum fraction selection described
in section 4.1.4 in determining the effects of the different target yield
and purification factors. Also the effect of a concentration range or a
volume range as a third fraction selection criterion is studied.
The analysis will focus on fractionation curves since these combine yield
and purity which are the most important parameters. A number of test
chromatograms will be used to illustrate the problems mentioned above.
These are:

(i) A chromatogram with a high degree of overlap and a relatively small
central peaks, as an example of deviations in the apparent maximum yield
and deviation from the true curvature at the end of the curve.
(Chromatogram A, the first Gaussian test chromatogram in Appendix A2)

(ii) A chromatogram with the first two peaks relatively small compared to
the third peak as an example of deviation from the apparent yield only
(Chromatogram B, the fourth Gaussian test chromatogram in Appendix A2).

and

(iii) A chromatogram with low overlap and a relatively small central peak,
as an example of deviation from curvature at the start of the peak and
at the end of the peak
(Chromatogram B, the second general exponential function test
chromatogram in Appendix A2).

The analysis will use the following factors to judge the ability of the
optimisation technique to select the optimum product fraction using a
fractionation diagram for the lowest objective function (ie. the most
accurate solution) in comparison with actual curves:

(i) The purification factors and yields for the actual and observed
optimum product fractions.

(ii) The actual and observed purification factors and yields for the
fractions selected as optimum.
(iii) The start and end elution volumes for the observed and actual optimum fractions.

Using these criteria it will be possible to determine the difference between the actual and observed product fraction quality, and the difference between the apparent optimum fraction quality and the true optimum fraction quality (calculated using the actual peak functions).

4.2.3.1 Selection of Constraints for the determination of the optimum fraction
This section considers the numerical values of the optimisation criteria to which the product fraction will be optimised.

Purification Factor
The maximum purification factor (PF) will define the upper limit on this parameter. This lower limit is defined as 1.0 corresponding to no purification.

Yield
This criterion will lie between 60% and 100% of the apparent maximum yield (ie. the maximum yield which the fractionation diagram constructed using deconvolution data indicates is possible. This may differ from the true value).
Yields within this chapter are based on a maximum corresponding to the amount of product bound to the column, not the amount of product loaded onto the column.

Volume
The maximum possible fraction volume will be the elution volume of the whole chromatogram but it is desirable to reduce this and to impose a stricter limit to prevent the generation of excessive volumes which may have undesirable effects on subsequent separation stages (see section 4.1.4). In fact the fraction volume should not sensibly be significantly in excess of the width of the product peak determined from the
deconvoluted central peak.
The minimum volume will be related to the yield parameter since this will affect the concentration of the product. High concentrations may cause problems with high viscosity or may cause problems with concentration overload in subsequent chromatographic separation steps (see section 4.1.4).

Concentration
No concentration or volume constraints will be set. The concentration of the most optimum fractions determined will be compared.

4.2.3.2 Selection of the optimum fraction
The selection of the optimum fraction was carried out for each test chromatogram and the solution found by deconvolution at the lowest objective function value, using both methods of fraction selection (see section 4.1.4), ie. selection of a high purity and high yield fractions, using the programme given in Appendix A1.6. The high yield fraction was defined as a fraction which contained the whole of the product peak whilst minimizing the amount of other components, ie. a fraction with the maximum purification factor for maximum yield of the product.
Table 4.2 Optimum fractions for the test chromatograms

<table>
<thead>
<tr>
<th>Chromatogram / Deconvolution Solution</th>
<th>Fraction start (Arbitrary Volume units)</th>
<th>Fraction end (Arbitrary Volume units)</th>
<th>PF (-)</th>
<th>Yield (%)</th>
<th>Purity (%)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatogram A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>solution</td>
<td>8.38</td>
<td>13.50</td>
<td>8.3</td>
<td>90</td>
<td>81</td>
<td>High Yield</td>
</tr>
<tr>
<td>actual curve</td>
<td>7.63</td>
<td>13.50</td>
<td>7.8</td>
<td>90</td>
<td>60</td>
<td>High Yield</td>
</tr>
<tr>
<td>solution</td>
<td>8.63</td>
<td>13.38</td>
<td>8.8</td>
<td>89</td>
<td>86</td>
<td>High Purity</td>
</tr>
<tr>
<td>actual curve</td>
<td>9.38</td>
<td>12.38</td>
<td>12.3</td>
<td>89</td>
<td>94</td>
<td>High Purity</td>
</tr>
<tr>
<td>Chromatogram B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>solution</td>
<td>4.13</td>
<td>9.38</td>
<td>4.3</td>
<td>96</td>
<td>63</td>
<td>High Yield</td>
</tr>
<tr>
<td>actual curve</td>
<td>3.75</td>
<td>9.38</td>
<td>4.0</td>
<td>98</td>
<td>78</td>
<td>High Yield</td>
</tr>
<tr>
<td>solution</td>
<td>6.38</td>
<td>8.63</td>
<td>6.5</td>
<td>78</td>
<td>97</td>
<td>High Purity</td>
</tr>
<tr>
<td>actual curve</td>
<td>5.63</td>
<td>8.88</td>
<td>6.3</td>
<td>91</td>
<td>96</td>
<td>High Purity</td>
</tr>
<tr>
<td>Chromatogram C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>solution</td>
<td>6.00</td>
<td>8.00</td>
<td>3.3</td>
<td>93</td>
<td>50</td>
<td>High Yield</td>
</tr>
<tr>
<td>actual curve</td>
<td>6.00</td>
<td>8.00</td>
<td>3.3</td>
<td>93</td>
<td>50</td>
<td>High Yield</td>
</tr>
<tr>
<td>solution</td>
<td>6.38</td>
<td>7.88</td>
<td>6.5</td>
<td>80</td>
<td>96</td>
<td>High Purity</td>
</tr>
<tr>
<td>actual curve</td>
<td>6.38</td>
<td>7.75</td>
<td>6.5</td>
<td>84</td>
<td>96</td>
<td>High Purity</td>
</tr>
</tbody>
</table>

1 Calculated using the actual maximum yield for each chromatogram

Details of the optimum fractions which were determined for the actual test chromatograms and the best deconvolution solution are given in table 4.2. In general for the three chromatograms analysed, high yield fractions are found to the highest degree of accuracy (yield +/- 3.0% of actual maximum yield and purification factors +/- 6.0%).

The correct selection of the high yield fraction requires a good description of the product peak only and does not require a description of the surrounding peaks. Contrastingly the correct selection of a high purity fraction requires accurate information concerning all peaks in the chromatogram, so that a fraction containing mainly the product component with very little of any contaminant components, may be selected. This is clearly a more difficult task than just determining the central peak and as previously stated (see section 2.2.6) it is dependent on the nature of the chromatogram. For these reasons, ignoring chromatogram B which was
not correctly analysed by deconvolution, the yields and purifications factors of the optimum fractions found from the true and model curves differ by up to +/-30.0% for high purity fractions, compared to +/- 8.0% for high yield fractions (see table 4.2).

Table 4.3 shows the actual compositions and yields of the optimum product fractions determined using the deconvolution results, described in table 4.2. Additionally differences between the yield predicted for the fraction and the actual yield obtainable, and the predicted and actual purity are shown ($\Delta Y_1$ and $\Delta P_1$). The differences between the yield predicted and obtained from the optimum fraction determined from the actual fractionation curve is shown ($\Delta Y_2$). A similar difference is shown for purities $\Delta P_2$.

This data was calculated using the positions of the optimum product fractions found by deconvolution on the fractionation curve constructed using the actual peak functions.

Table 4.3 Observed and actual quality of optimum fractions selected from deconvolution solutions in table 4.2

<table>
<thead>
<tr>
<th>Test Chromatogram</th>
<th>Actual Yield ( % )</th>
<th>$\Delta Y_1$ (% )</th>
<th>$\Delta Y_2$ (% )</th>
<th>Purity ( % )</th>
<th>$\Delta P_1$ (% )</th>
<th>$\Delta P_2$ (% )</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromatogram A</td>
<td>90</td>
<td>0.0</td>
<td>0.0</td>
<td>60</td>
<td>21</td>
<td>0</td>
<td>High Yield</td>
</tr>
<tr>
<td>Chromatogram A</td>
<td>88</td>
<td>-1.7</td>
<td>26.0</td>
<td>84</td>
<td>2</td>
<td>-10</td>
<td>High Purity</td>
</tr>
<tr>
<td>Chromatogram B</td>
<td>98</td>
<td>-2.3</td>
<td>2.3</td>
<td>69</td>
<td>6</td>
<td>-9</td>
<td>High Yield</td>
</tr>
<tr>
<td>Chromatogram B</td>
<td>78</td>
<td>0.0</td>
<td>60.0</td>
<td>98</td>
<td>1</td>
<td>2</td>
<td>High Purity</td>
</tr>
<tr>
<td>Chromatogram C</td>
<td>93</td>
<td>0.0</td>
<td>0.0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>High Yield</td>
</tr>
<tr>
<td>Chromatogram C</td>
<td>82</td>
<td>-5.3</td>
<td>0.0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>High Purity</td>
</tr>
</tbody>
</table>

($\Delta Y_1$ corresponds to the percentage difference between the yield predicted from the deconvolution results and the actual yield for the fraction; $\Delta P_1$ is the corresponding difference for the purity.

$\Delta Y_2$ corresponds to the percentage difference between the yield predicted from the deconvolution results and the results for the best fraction obtained from the actual fractionation curve; $\Delta P_2$ is the corresponding difference for the purity.)

Though the actual positions and quality of the optimum high
yield fractions given in table 4.2 differ, the yield and purity values of
the optimum fractions obtained using the deconvolution results agree
reasonably well with the true values given in table 4.3 (see column $\Delta Y_1$
and $\Delta P_1$). The exception to this is the high yield fraction for the
chromatogram A (which was not satisfactorily deconvoluted).
In general the deconvolution results under-estimate the maximum yield
that may be obtained from the selected fraction (shown by negative
numbers in column $\Delta Y_1$). High purity fractions are not found so
accurately and may differ by up to 60.0% for reasons discussed earlier.
The purity of optimum fractions found by deconvolution and the actual
purity of such fractions also agree well (+/-6% or +/-3% column $\Delta P_1$).
Though chromatogram A which was not successfully deconvoluted again
shows a more substantial error.

As discussed above (section 4.1.4) the volume of the fraction may be an
important factor in determining the position of the optimum product
fraction. In determining the optimum fractions in table 4.2 the volume
parameter was not used in the selection process. However it was
observed that for the optimum fractions in the test chromatograms
analysed in this section few options with similar combinations of high
purity or high yield existed. When such options occurred volumes were
equal. In contrast, many fractions with both sub-optimal purity and
yields were available with a range of volumes (and hence concentrations).
Thus the volume of the fraction is determined by the separation operating
conditions rather than from the selection of the product cut. Therefore
the volume of the fraction is not an important parameter in the selection
of the optimum fraction. It should instead be a parameter taken into
account when the separation conditions are initially determined to achieve
the desired separation performance. If the volume of the product fraction
were too great then to operate the separation at its optimum would
require the operating conditions to be re-optimised so that the product
fraction may be contained within the allowable volume range.

In conclusion using computer generated chromatograms it has been
determined that providing the chromatogram has been deconvoluted
satisfactorily (see chapter 2) then high yield fractions may be selected
and the yield determined to within 5.5% of true values and purities to
within 6.0%. High purity fractions are more difficult to select as they
require information about all peaks in the chromatogram and are in
general not found so accurately. Volume should as a parameter when determining the initial separation conditions rather than a parameter to be used for control purposes when selecting the optimum fraction. Having proven the algorithm on synthesised data the next section will examine the approach with real chromatographic data.

4.2.4 Egg White chromatogram analyses
An analysis similar to that carried out for the test chromatograms was carried out for the egg white chromatograms. Only those chromatograms which were successfully deconvoluted and had their constituent peaks correctly identified were analysed - namely the first (figure 2.17), second (figure 2.18) and fifth (not shown) egg white chromatograms.

Construction of egg white fractionation diagrams
Fractionation diagrams were constructed as in section 4.2.1 for the test chromatograms using the programme in Appendix A1.5, the peak functions parameters found by deconvolution, and the extinction coefficients determined in section 2.3\(^1\). The optimum fractions, both for high yield and high purity, were selected using the programme described in Appendix A1.6.

Results
The fractionation diagrams obtained from the egg-white chromatograms (figure 4.10) are similar in nature, ie. they all exhibit similar maximum purification factors and yields and total chromatogram areas. The curve produced using data from the second egg white chromatogram differs the most. The main differences are a shallower curve, a higher yield and a high degree of curvature at the beginning and end of the fractionation curve. This is due to the high degree of overlap of the central peak with the third peak (see figure 2.18). These factors will affect the performance of the separation in terms of the selection of a high yield

\(^1\) The peak heights in Appendix A2.4 were converted to absorbance units by dividing by 50 AU.mV\(^{-1}\) (the detector range was set to 0.0 - 2.0 AU which resulted in a 0 to 100 mV output) and then to a mass by dividing by the length of the flow-cell (2mm) and the extinction coefficient of the component.
fractions which includes these sections of the fractionation curve (and chromatogram). However the selection of high purity fractions should be relatively unaffected since this type of fraction only uses central portions of the fractionation curve. The second and fifth egg white chromatogram contain extra visible peaks in addition to those peaks modelled. The optimum fractions for the three egg white chromatograms are shown in table 4.4.

Table 4.4 Optimum Fractions for egg white chromatograms (section 2.3)

<table>
<thead>
<tr>
<th>Fraction start (mL)</th>
<th>Fraction end (mL)</th>
<th>PF (-)</th>
<th>Observed Yield (mg)</th>
<th>Yield (%)</th>
<th>Purity</th>
<th>High Purity / High Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st egg white chromatogram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82.0</td>
<td>94.4</td>
<td>3.5</td>
<td>3.1</td>
<td>90</td>
<td>52</td>
<td>High Yield</td>
</tr>
<tr>
<td>84.0</td>
<td>92.3</td>
<td>6.5</td>
<td>2.3</td>
<td>88</td>
<td>96</td>
<td>High Purity</td>
</tr>
<tr>
<td>2nd egg white chromatogram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80.3</td>
<td>89.1</td>
<td>2.5</td>
<td>3.3</td>
<td>75</td>
<td>53</td>
<td>High Yield</td>
</tr>
<tr>
<td>82.8</td>
<td>87.9</td>
<td>4.8</td>
<td>2.2</td>
<td>51</td>
<td>100</td>
<td>High Purity</td>
</tr>
<tr>
<td>5th egg white chromatogram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69.9</td>
<td>79.1</td>
<td>3.5</td>
<td>3.9</td>
<td>89</td>
<td>73</td>
<td>High Yield</td>
</tr>
<tr>
<td>72.3</td>
<td>78.2</td>
<td>4.8</td>
<td>3.2</td>
<td>71</td>
<td>99</td>
<td>High Purity</td>
</tr>
</tbody>
</table>

To assess whether the correct optimum fractions had been determined a comparison of the fraction starts and ends was made with the data obtained by scanning SDS-PAGE gels. The data obtained by gel scanning cannot be used to assess the validity of the purity and yield values since the mass values are unreliable. Instead the results from SDS PAGE scanning are used to indicate the starts, maxima and ends of peak functions. This may be considered acceptable since the combination of a low objective function value (meaning a small difference between model and experimental chromatograms) together with agreement on the position of the peak function starts and ends should mean that the elution profiles of the peak functions are correct.

Also such SDS data may be used off-line to establish constraints for the deconvolution algorithm and the selection of the optimum fraction. The high yield fraction (described in table 4.4) for the first egg white chromatogram (figure 2.17) can be seen on figure 2.22 to include virtually
all of the fractions shown by the SDS-PAGE analysis to contain Ovoglobulins (the product material) but none which do not contain any Ovoglobulins.

• first egg white chromatogram
The position of the high yield fraction has been successfully determined (ie. all of the Ovoglobulins are within the selected fraction). The position of the high purity fraction for this chromatogram is also found correctly since it only includes fractions containing Ovoglobulin which do not contain significant amounts of other components, ie. only portions of the central product peak which do not overlap significantly with neighbouring peaks are selected for the high purity fraction.

• second egg white chromatogram
In comparison the position of the high yield optimum fraction (figure 2.18) is not determined precisely (see figure 2.23 for SDS-PAGE data). The high yield fraction is selected in accordance with the peak function found by deconvolution, ie. the start of the fraction occurs within the first peak at too low an elution volume and the fraction end occurs before the end of the peak function. This indicates that a poor high yield fraction has been selected which contains some sub-fractions with no product whilst excluding some containing the product material. In comparison the high purity fraction is found more satisfactorily since the positions of the first peak end and the third peak starts have been correctly determined by deconvolution. The reason for the inability to correctly select the optimum fraction is the appearance of extra peak which is highly resolved, in a valley.

• fifth egg white chromatogram
The position of the optimum fractions for the fifth egg white chromatogram were found correctly as for the first egg white chromatogram, despite having an extra peak present. This extra peak was not well resolved from the central peak and so made little impact on the model peak function for this peak.
4.3 Conclusions
In this chapter the fractionation diagram has been shown to be an effective aid in the selection of an optimum product fraction in chromatography - either for an optimum high yield or high purity. The successful determination of the product fraction is dependent upon deconvolution of the chromatogram under examination producing an accurate model of individual component elution profiles. This is dependent on the relative peak overlap and peak heights (see section 2.2.5 and 2.2.6). Additionally, for real systems, the presence of extra peaks (due to contamination or high resolution of minor resolution than in other separations) in the experimental chromatogram may adversely affect the accuracy of the fraction selection.

Accurate selection of an optimum high yield fractions is dependent on the accurate determination of the start and end of the product peak (by deconvolution) whereas the accurate selection of an optimum high purity fraction requires an accurate description of all peaks in the chromatogram so that relative amount of product and contaminant in the fraction may be calculated. The latter situation is more complex and thus requires a higher level of accuracy from the model chromatogram obtained by deconvolution.

The volume of the optimum fraction was found not to be a parameter which could be optimised (ie. for a given optimum purity or yield only one volume was available) when selecting the best product fraction. This is true of chromatographic separations where the product material is eluted in one peak. If it is necessary to change this parameter then it is required to re-optimise the separation conditions.

This chapter has demonstrated the utility of combining the techniques and algorithms developed within this thesis for the at-line control of chromatographic separations. Criteria necessary for successful fraction selection have been identified and tested with both synthesised and real chromatographic data. The control capability offered by the combination of deconvolution, fuzzy logic peak identification, and fractionation diagrams is significantly in advance of existing methods which fail to monitor or control the quality of the product obtained from separations.
Figure 4.1 Flow Diagram for the selection of the optimum fraction based on a productivity threshold.
Figure 4.2 The Fractionation Diagram
The Fractionation - Concentration Diagram.
The line from A to B represents an increase in volume only (ie. before the elution of any material starts). Between B and C, and D and E contaminant protein is eluted (an increase in all three parameters can be observed). The product protein is eluted in the central portion of the line (C to D, - the increase in total protein is mainly due to the increase in product protein).
Figure 4.4  A tie line which represents three fractions with equal purities and purification factors. The fractions are represented by the lines AB, BC and AC. Although they have equal purities the actual composition of the impurities is different.
Figure 4.5  The Fraction Volume
Figure 4.6

Fractionation curves for the first Gaussian test chromatogram at two objective function values and the fractionation curve for the actual peak functions.
Fractionation curves for the fourth Gaussian test chromatogram at two objective function values and the fractionation curve for the actual peak functions.

Figure 4.7
Figure 4.8 The cumulative elution of product versus elution volume for the deconvolution results and the actual chromatogram for the first Gaussian test chromatogram.
Figure 4.9 The cumulative elution of product versus elution volume for the deconvolution results and the actual chromatogram for the fourth Gaussian test chromatogram.
Figure 4.10 The fractionation diagrams for the first, second and fifth egg white chromatograms, constructed using deconvolution data.
Chapter 5

Discussion
5. Discussion

In this chapter the type of chromatographic separation which may be adequately analysed (and hence controlled) using techniques discussed and developed in previous chapters are examined - in particular the performance of these separations (ie. the purities and productivities which may be achieved).

The effects of linking the techniques developed in this thesis together will also be examined, as will the total time required for the analyses and how this would fit into an actual chromatographic process. Modifications which may be carried out to improve the techniques will also be considered.

A comparison with existing work in related areas is also given.

5.1 Separation performance

As discussed in the preceding chapters the nature of the chromatogram is critical in determining whether deconvolution will produce a model chromatogram which adequately describes the true individual component elution profiles and which may then be used for product identification and optimum fraction selection. The parameters which have been found to be important in determining whether successful chromatogram deconvolution is possible are the relative peak heights and the peak separation (see section 2.3.10). These may be expressed as a ratio of peak to valley height (P:V). The results from chapter 2 indicate that a P:V of at least four is required for one of the valleys whilst the ratio for the other valley should not be less than one, for production of an adequate model chromatogram. The size of this ratio for the two valleys in this type of model chromatogram give a measure of the peak overlap for the particular set of peak heights. By an examination of the relative peak overlaps, peak heights and P:V in those test chromatograms used in section 2.2.6 the possible quality of separation (ie. purification factor) that is obtainable from this type of separation may be determined. This examination is carried out in the following paragraphs.

The maximum productivities that may be obtained from any particular chromatogram are determined by the peak area of the product peak (for constant peak width). The purity that may be obtained however, is affected by the overlap of peaks ie. the size of the P:V ratio, in addition to relative peak height. As an indication of the performance which is
possible from separation, the purity available at maximum yield and the yield at 100% purity, may be used since they represent extremes of separation quality obtainable from a chromatogram.

Table 5.1 Yields for maximum purity and Purities for maximum yield for test chromatograms used section 2.2.6 and Appendix A2

<table>
<thead>
<tr>
<th>High Purity / High Yield</th>
<th>Yield ( % )</th>
<th>Purity ( % )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gaussian test chromatogram</td>
<td>46.8</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>8.2</td>
</tr>
<tr>
<td>2nd Gaussian test chromatogram</td>
<td>70.0</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>80.0</td>
</tr>
<tr>
<td>3rd Gaussian test chromatogram</td>
<td>54.0</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>49.2</td>
</tr>
<tr>
<td>4th Gaussian test chromatogram</td>
<td>86.0</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>35.6</td>
</tr>
<tr>
<td>1st General Exponential Function test chromatogram</td>
<td>60.0</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>29.2</td>
</tr>
<tr>
<td>2nd General Exponential Function test chromatogram</td>
<td>87.0</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>27.2</td>
</tr>
<tr>
<td>3rd General Exponential Function test chromatogram</td>
<td>45.5</td>
<td>&gt; 99.0</td>
</tr>
<tr>
<td>High purity</td>
<td>100.0</td>
<td>11.4</td>
</tr>
</tbody>
</table>

N.B. These figures assume equal extinction coefficients.

The data contained in table 5.1 shows the purities and yields attainable
from the test chromatograms when either maximum yield (productivity) or purity are required. The first Gaussian and third General Exponential Function test chromatograms (see table 2.12) are the only chromatograms in table 5.1 which are not satisfactorily deconvoluted. Such separations would be considered, in general, to be unsatisfactory since a low peak separation is achieved. In such cases the inability of the deconvolution algorithm to determine the individual component elution profiles from chromatograms such as the first Gaussian and third General Exponential Function test chromatograms is not a significant disadvantage since they represent inherently poor separations and would not in general be adopted in a process.

It should be noted that the figures in table 5.1 are calculated assuming equal extinction coefficients for all components. The biological system studied in previous chapters, hen egg white, contains components with differing extinction coefficients (see table 2.10) - one component, Lysozyme, has an extinction coefficient approximately twice that of others. If it is assumed that this difference is typical the effect on the purities that may be obtained can be examined for the high yield fractions. The results are displayed in Table 5.2.
Table 5.2 Purities at maximum yields with differing extinction coefficients

<table>
<thead>
<tr>
<th>Test Chromatogram</th>
<th>Purity at equal $\epsilon$ (table 5.1)</th>
<th>Purity with product peak $\epsilon - \epsilon * 2$</th>
<th>Purity with product peak $\epsilon - \epsilon / 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gaussian</td>
<td>8.2</td>
<td>4.3</td>
<td>14.1</td>
</tr>
<tr>
<td>2nd Gaussian</td>
<td>80.0</td>
<td>66.7</td>
<td>61.5</td>
</tr>
<tr>
<td>3rd Gaussian</td>
<td>49.2</td>
<td>32.6</td>
<td>70.0</td>
</tr>
<tr>
<td>4th Gaussian</td>
<td>35.6</td>
<td>21.7</td>
<td>52.5</td>
</tr>
<tr>
<td>1st General Exponential Function</td>
<td>29.2</td>
<td>17.1</td>
<td>45.2</td>
</tr>
<tr>
<td>2nd General Exponential Function</td>
<td>27.2</td>
<td>15.7</td>
<td>42.8</td>
</tr>
<tr>
<td>3rd General Exponential Function</td>
<td>11.4</td>
<td>6.0</td>
<td>18.6</td>
</tr>
</tbody>
</table>

The purity figures for the test chromatograms calculated assuming the product extinction coefficient is twice that of other components are lower than those given in table 5.1. Assuming that the product extinction coefficient is half that of other components give higher purities than those in table 5.1. These results are due to the respective increase and decrease in the masses of the product calculated.

Despite the increase in the relative amounts of contaminants caused by reducing the extinction coefficient of the product peak neither in the case of the first Gaussian or the third General Exponential Function test chromatograms does the purity of the high yield fractions exceed twenty percent. For the purity of these fractions to reach the levels of other chromatograms (approximately fifty percent) the product extinction coefficients would need to be eight times less than both the contaminant coefficients. With differing extinction coefficients the peak overlap is not
altered and so the yields at near 100% purity are not improved.

The techniques developed in this thesis are fully capable of dealing with realistic chromatographic separations. However some chromatograms do not lend themselves to analysis by deconvolution. Simple rules or properties of the chromatogram which must apply for successful deconvolution have been determined. In the small number of cases where test chromatograms are not deconvoluted satisfactorily it can be seen, from theoretical purities and yields, that they do not represent good separations and their performance does not take advantage of the high purification factors which are possible using chromatographic separations.

5.2 Integration of Analysis techniques

This section considers the integration of the techniques developed in previous chapters so that the feasibility of using them within a chromatographic sequence may be examined.

In an ion-exchange process, such as that used in section 2.3 to produce the egg white separations, several steps are involved (see Table 5.3). The majority of which do not involve the elution of any product material from the column.

| Step                  | Relative Time
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Column Equilibration</td>
<td>^2 column volumes</td>
</tr>
<tr>
<td>2. Sample Loading</td>
<td>variable</td>
</tr>
<tr>
<td>3. Column Washing</td>
<td>^2 column volumes</td>
</tr>
<tr>
<td>4. Sample Elution</td>
<td>^1-2 column volumes</td>
</tr>
</tbody>
</table>

Table 5.3 A typical chromatographic operation sequence

Of these, column equilibration and washing (steps 1 and 3)^1 take approximately two column volumes each. The relative time taken by the

---

1 Time is expressed here in terms of the number of column volumes of fluid which must be passed. It is assumed that a constant flow rate is used in all steps.
sample loading is dependent on the size of the volume of material loaded and the flow rate for optimum the binding of the product component. The time taken for the product elution step is usually approximately two column volumes. At least half of the separation time (steps 1 and 3) does not involve elution of product and thus production of elution profile data.

In this time data analysis of the previous chromatogram may be carried out so that the position of the optimum product fraction can be determined for subsequent chromatograms.

<table>
<thead>
<tr>
<th>Step</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Deconvolution</td>
<td>10.0</td>
</tr>
<tr>
<td>2. Peak Identification</td>
<td>1.4</td>
</tr>
<tr>
<td>3. Optimum Fraction selection</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>14.4</strong></td>
</tr>
</tbody>
</table>

**Table 5.4 Processing times for analysis techniques**

Table 5.4 shows typical analysis times (see Appendix A1 for hardware and software details) for each of the techniques discussed in previous chapters, for chromatograms containing three peaks (The number of peaks significantly affects the deconvolution and peak identification steps) which indicates that a total analysis time of approximately 14.4 minutes is required. The chromatographic system described in section 2.3.2 (column volume = ~20mL, flow rate = 2.0 mL min\(^{-1}\)) allows for this type of analysis since loading alone takes up to twenty minutes (see figure 2.11).

The integration of these techniques into other chromatographic processes depends on the equilibration and washing times. These are dependent on the column volume (typically these take 2 column volume each - Pharmacia, 1991). Maximum linear flow rates for typical soft gels are 60 to 600 cm h\(^{-1}\) (Pharmacia Sepharose CL-6B and monobeads, Pharmacia, 1991). HPLC packing would also tolerate linear flow rates of approximately 600 cm h\(^{-1}\).

\[
\text{Equilibration and Washing Time} = 4 \times \frac{\text{Column Volume}}{\text{Volumetric Flow Rate}} = 4 \times \frac{\text{bed depth}}{\text{Linear Flow Rate}}
\]
For the range of linear velocities given above, the corresponding equilibration times (measured in minutes) are between 0.4 and 4 times the chromatographic bed depth.

Thus the washing and equilibration times are dependent upon the packing stability and the depth of bed. If a more rigid packing were used for the egg white separation described in section 2.3 (bed depth = 10cm) then significantly faster equilibration washing and elution could be carried out to such an extent that insufficient analysis time would be made available.

If such a situation did occur then a range of improvements to the analysis could be considered under certain conditions. As shown in Table 5.4 the process of deconvolution is the most time consuming task. In Chapter 2 various optimisation techniques which may be used for deconvolution were considered. The Box-Complex (Box, 1965) technique used in this work has a number of advantages and disadvantages. The most significant advantage is that due to the generation of a random complex throughout the whole feasible space (defined by the constraints – see section 2.2.3) the method will find the global optimum. This is unlike other non-directed search methods (such the Levenberg-Marquardt technique, Marquardt, 1963) which require an initial value for each parameter (rather than a range). In such cases the solution obtained is sensitive to the actual initial values to such an extent that either a non-global solution may be reached or the alternative method may fail to converge. The advantage of a global optimum being obtained by Box-Complex optimisation is at the expense of computational time. This is because many objective function evaluations are required for chromatograms throughout the feasible parameter space. In the deconvolution of chromatograms the evaluation of the objective function is time consuming since it requires the calculation of the difference between the model and experimental chromatograms at each point recorded during elution.

Non-directed search methods will require fewer objective function evaluations since they use differentials to follow the direction of fastest descent towards an objective function value of zero and so will be faster.

If a particularly fast separation (such as an HPLC separation) was to be analysed then it is proposed that after an initial separation deconvolution would be carried out using the Box-Complex method. For subsequent separations a non-directed search method could be applied using the
original solution found by the Box-Complex method to provide a set of initial parameter values. This assumes that chromatograms do not change greatly from one separation to another which is a reasonable simplification.

If timing is crucial then re-coding by software engineers in a low level language to take advantage of the speed of low level machine instructions will reduce significantly the processing times required.

5.3 Baseline Drift and Noise
The problem of baseline drift in the integration and analysis of chromatographic data has been reviewed by Dyson, 1990. In general three techniques may be applied if the baseline drifts; firstly the use of a linear baseline defined as a straight line between two pieces of baseline before and after the peaks or secondly the use of a baseline for each peak defined as running from baseline to peak valley (Figure 5.1a), or peak valley to peak valley, or peak valley to baseline (Figure 5.1b). The second method is however not applicable when describing the overlap of peaks since by using an effective baseline between baseline and peak valley it completely ignores any area which may be overlapped.

A third technique may be used where an eluant causes drift (eg. due to UV absorbance, Figure 5.1c). Here a blank gradient is run on the system and a blank chromatogram recorded. This is subtracted from all subsequent separations to be analysed.

Of these techniques only the first and third is applicable to deconvolution since they both allow for peak valleys to occur which do not coincide with the baseline. In effect both these techniques are very similar - they involve linearising the chromatogram by accounting for a baseline which is a function of time. It would be very simple to include this within the deconvolution programme (Appendix A1) either by subtracting the baseline from the data points before analysis or by including the subtraction within the objective function.

As stated previously (section 2.2.6) a relatively noise-free signal is required for peak integration. For accurate peak integration a signal to noise ratio of at least 60 is required (Rossi 1988). Typically
chromatographic detection systems are capable of obtaining such ratios (Dyson 1990). The effect of the signal to noise ratio has been reported by Massart et al., (1988). When the signal to noise ratio exceeds 60 the error in peak area due to noise is less than that reported by Vaidya and Hester (1984) (+/-5%) when determining peak areas from deconvoluted peaks. This error is less than that found in this thesis (section 2.2.6). Truly random noise in chromatographic processes is rare. Noise is typically composed of several definable frequencies (eg. produced by pumps etc.) which may determined and removed using suitable mathematical techniques such as Kalman filtering (Jazwinski 1970). Other methods for filtering noise have been described by Dyson (1988) and rely upon electronic techniques such RC (capacitative) filters which are able to remove noise or spikes caused by electric motors or switches. Thus it can be seen that noise is not typically a problem in chromatographic analysis but if it should be so standard and well tried techniques are available to reduce it to an acceptable level.

In this thesis the effect of random noise has been studied. Providing it lies within certain ranges it does not cause a significant error in the determination of peak area.

5.4 Comparison of techniques developed with other software currently available
In this thesis a number of techniques have been described which allow the selection of a product fraction from a preparative chromatographic separation based on quantitative information. This has been achieved by analysing the chromatogram by deconvolution to determine individual peak elution profiles, matching of peaks with a template data set (using fuzzy logic) and finally analysis of fractions obtainable using fractionation diagrams. A model chromatogram is produced which gives the elution profile of each component and the identity of the individual components. By contrast with the work detailed in this thesis the majority of software previously developed for use with chromatographic data has centred on optimisation of analytical separations (see section 1.6). The methods of describing separation performance used are simple optimisation criteria (see section 1.3) to produce maximum peak separation (such as in PRE-
OPT, DryLaB, CRISE, SOS and Diamond) to allow quantification and identification of components present. Other systems comment on reproducibility (Mulholland VanLeeuwan and Vandeginste, 1989).

Of the software previously produced only MABLAB (Kenney, Thompson and Harris, 1989) has been designed to work with preparative systems (section 1.6). This however only considers the relatively simple case of monoclonal antibody production where only one major peak is present in the chromatogram and which may be quite easily resolved from others. Its main function was to inform the operator of changes in column performance and to adjust the load volumes as the column capacity changed over time. This system uses the valley (or baseline if fully resolved) to select the position of the product fraction and no quantitative information is used. This system does not identify which of the components present in the chromatogram correspond to the product and contaminants.

None of the systems discussed in the introduction (chapter 1) produces a model chromatogram which describes the elution of each component. The techniques developed and tested in this work provide an integrated capability for automated at-line analysis and control of preparative separations. No other software tools or combination of these provides the same capabilities.

In particular the major strengths of the methods described in this work are:

- accurate prediction of individual peak profiles enabling calculation of productivity and yield.
- automatic diagnosis of changes in elution order.
- an engineering framework for the control of the separation via the use of fractionation diagrams.

Each of these features make the methods unique. Furthermore the techniques developed in this thesis have been tested and proven for both simple cases typical of those found at the end of a
downstream process as well as more demanding cases where, by use of synthesised data, it has been possible to test the techniques under extreme conditions. Simple criteria for determining whether or not the developed methods will work successfully have been described.

5.5 Conclusions

In conclusion the techniques developed in this thesis have demonstrated their utility for automated selection of product fraction. In the small number of cases where chromatograms pose difficulties to analysis techniques these have been shown to represent poor chromatographic separations which do not take advantage of the high resolution of components possible with chromatographic separations. The technique is therefore sufficiently robust to be used in a process scheme as shown in figure 1.8.

It has been demonstrated that the analysis a chromatogram can be carried out at-line within a chromatographic process during the column equilibration and washing steps. The information thus obtained may then be used to determine the position of the product fraction in the following separation.

This form of retrospective control is quite consistent with the chromatographic process since changes in elution profile between repeat cycles is characteristically slow.

The techniques developed in this thesis subject to the same limitations of baseline drift and noise as standard integration techniques i.e. a definable baseline and a low noise signal. Standard filtering techniques could be applied in order to improve the signal quality should this be necessary. The combination of techniques discussed in this thesis has been shown to be an advancement over currently available software.
Figure 5.1 Baseline drift correction methods

(a) baseline linearised to baseline before and after peaks
(b) baseline set from valley to valley
(c) baseline linearised by subtracting a blank gradient
Chapter 6

Conclusions
6. Conclusions

Deconvolution

The successful deconvolution of chromatograms has been found to depend upon adequate peak separation for a given set of peaks. The ratio of peak height to valley height (P:V) has been found to be a good simple measure of overlap as it relates the valley height (a measure of peak overlap) to the peak height (a measure of peak area). In general for a good model chromatogram (consisting of three peaks) to be obtained by deconvolution P:V for one valley should be above whilst the other should be above four. If either or both of the valleys do not fall within these values then the chromatogram is ill conditioned as there is insufficient data available to determine the individual peak shapes. In addition to peak overlap, the correct peak model and number of peak functions in the model chromatogram have also been found to be important in ensuring successful deconvolution.

The importance of peak numbers has been shown to affect the reliability (see section 2.3) - particularly when resolution of minor components may occur. This restricts the method to systems where the number of peaks is known and is consistent.

A low objective function value (typically equivalent up to four percent of the mean chromatogram height) is necessary but not sufficient in determining whether a good model chromatogram is determined by deconvolution since as previously mentioned ill conditioning (due to high peak overlap).

Peak Identification

Fuzzy Logic has been shown to be an effective method of peak identification using peak area and elution rank as peak identification characteristics and a template data set. It has been tested with a variety of experimental systems - computer generated data and ion exchange separations of hen egg white containing overlapping peaks.

Limitations to the method have been identified and these relate to similarities between peak identification characteristics. These problems due to similarities in peak identification characteristics are similar to
those encountered by a human but fuzzy logic allows quantification to be introduced into the process giving a more consistent result. Methods for testing the template data for identification difficulties have been developed.

The identification of peaks within a deconvoluted chromatogram is dependent on successful deconvolution. The overall matching measure between the template and experimental data may be used as an additional measure of deconvolution success since if peak characteristics in the experimental data set do not match well with those in the template data set then they may have been incorrectly determined by the deconvolution process.

Control
The fractionation diagram approach has been applied to chromatographic separations for the selection of an optimum high yield or high purity product fraction. Accurate determination of the position of the optimum product is dependent on the correct identification of the product peak and the deconvolution producing an accurate model chromatogram (the conditions for this are discussed above).

The accurate selection of high yield fractions was found to be more reliably determined in comparison with high purity fractions. This may be assumed to be due the fact that high yield fractions require the accurate determination of the product peak only whereas the high purity fraction requires accurate determination of both product and contaminant peaks. The latter requires a more accurate model chromatogram and thus more accurate deconvolution.

Volume was found not to be a parameter which can be optimised when selecting the optimum fraction. This parameter is best optimised when the separation conditions are defined.
Integration of Analysis Techniques

Chromatograms which cannot be accurately deconvoluted by the techniques discussed in Chapter 2 may be considered to be very poor in terms of performance and the inability of the method to cope with such separations should not be considered a disadvantage.

The analysis techniques in this work have been shown to be fast enough to be included within the column washing and re-equilibration steps for most chromatographic processes. Methods for increasing the speed of these analysis methods have been identified for where this is not possible.
Chapter 7

Recommendations for Future Work
7. Recommendations for Future Work

The following areas for future work are suggested:

Extension of the deconvolution technique to allow for extra peaks. This could be accomplished by automatically changing the number of peak in the model chromatogram if the objective function could not be made to reach a suitably low value.

Investigate the constraints for skew parameters of the General Exponential Function so that the overlap limitations of chromatogram deconvolution may be reduced by gaining a better understanding of the actual peak skews that may occur for a given separation.

The use of upstream operation conditions and monitoring data so that chromatographic operating conditions may be controlled to give the desired product quality and the use of this data to give the deconvolution technique an alternative source for starting values.

The use of alternative methods for deconvolution including:

(i) non-directed search techniques (eg. Levenberg-Marquardt, 1963) to improve analysis times.

(ii) pre-optimisation manipulation of variable data to reduce the number of times the objective function needs to be evaluated (Shaw, 1993).

(iii) use of alternative peak models

and

(iv) use of speech-processing techniques to deconvolute peaks or account for changes from a template chromatogram.
The deconvolution technique could also be extended to an on-line method using a previous separation as a template to assess the deviation of the current separation from the template. Knowledge of the deviation from the template could be used to estimate the position of the product fraction.

Peak identification could be studied further using alternative characteristics to identify components. Two possible approaches are suggested. Firstly the elution of material could be monitored in an additional manner (thus giving an alternative peak characteristic) or secondly using peak characteristics such as peak skew or slope which may be calculated from the UV trace. Again upstream data may be used and could be incorporated as the template chromatogram.

Formal methods for the selection of the optimum product fraction could be developed.
This could include:

(i) the development of a suitable objective function which takes into account the relative weighting of purity and yield, using economic data to calculate the weightings.

(ii) the application of non-directed search optimisation techniques for the selection of the optimum fraction.

(iii) the development of an expert system to select the optimum fraction.

Interactions between chromatogram analysis methods could be further analysed. Particularly the use of solutions of one method as the starting point for the next technique applied or as a check on the reliability of the results obtained.

Further improvements could be identified by applying the techniques developed to real process-scale separations.
Appendix A1

Computer Programmes and Hardware
Appendix A1 : Computer Programmes and Hardware
This appendix contains listings of all computer programmes (written in C) used in this thesis and a description of the hardware used.

A1.1 Deconvolution : Box – Complex Algorithm
This programme is used in Chapter 2 to determine individual peak shapes from chromatograms. The flow diagram for the algorithm is given in figure 2.7 and a description of the algorithm is given in section 2.2.4.

/* Box-Complex method : M.J. Box, Computer J., No.8, p42-52, 1965. */

#include <stdio.h> /* standard input/output device */
#include <float.h> /* floating point header */
#include <math.h> /* maths header file */
#include <time.h> /* time header file */
#include <stdlib.h> /* standard library definitions */

#define NMAX 15 /* number of peaks X no of variables for peak */
#define KMAX 16 /* maximum number of points */
#define CMAX 4 /* maximum number of constraints */
#define BIGNUM 1e10 /* a large number to prevent rounding errors */
#define NUMPOINTS 200 /* number of points in the chromatogram */

/* PROTOTYPE DECLARATIONS */

extern void main(void);
static void check(int);
static void func(int);
static void consx(void);
static void slacks(void);
static void centroid(void);
static double gex(double, double, double, double, double);
static double rand0(int *);

/* VARIABLE DECLARATIONS COMMON TO ALL FUNCTIONS IN THE FILE */

double rate = 1.0; /* number of points per second */
double chromatogram[] = [0.003072, 0.004150, 0.005763, 0.007806, 0.010317, 0.013940, 0.000519, 0.004721, 0.002741, 0.001034, 0.000010];

double x[KMAX][NMAX];
double g[NMAX];
double h[NMAX];
double f[KMAX];
double xc[NMAX];

int kount;

int k = KMAX;
int n = NMAX;
int m = 4;
int it, itmax;
int lowest, lowest1, highest;
int limit;

double alpha = (double)1.5;
double beta = (double)0.75;
double delta = (double)0.001;

definition of function variable array
definition of two implicit constraints
definition of array of function values
definition of centroid points

definition of convergence criterion check
definition of number of variables
definition of number of points

definition of iteration counter and maximum number of its
definition of lowest, next lowest and highest function value
definition of max point used in calculation of centroid

definition of expansion coefficient

definition of desired minimum function value

definition of if constraint violated variable moved delta in

/* FUNCTION DECLARATION AND ARGUMENT DECLARATION */

void main (void)

{ int i;
FILE *fp;

itmax = 10000;

i = -10;
rand0(i);

f[lowest] = beta + 10.0;
while (fabs(f[lowest]) > beta )
{
consx();

if ((fp = fopen("testansr","w")) == NULL)
{
printf("can't open output file\n"); return;
}
printf(fp,"Result after %d iterations\n",it); /* output results */

/* find lowest point */
lowest = 0; /* set lowest point to zero */
for(i=1; i<k; i++)
{
  if ((f[lowest]-f[i]) > 0.0) /* if ith point less than lowest set lowest */
  {
    lowest = i;
  }
}
i = lowest;

printf(fp,"point %d function=%f \n",i,f[i]);

printf(fp,"hm=%f Vo=%f Vm=%f a=%f b=%f\n",x[i][0],x[i][1],x[i][2],x[i][3],x[i][4]);
printf(fp,"hm=%f Vo=%f Vm=%f a=%f b=%f\n",x[i][5],x[i][6],x[i][7],x[i][8],x[i][9]);
printf(fp,"hm=%f Vo=%f Vm=%f a=%f b=%f\n",x[i][10],x[i][11],x[i][12],x[i][13],x[i][14]);
fclose(fp); /* close output file */
return;
}

/*****************************************
/* FUNCTION DECLARATION AND ARGUMENT DECLARATION
*******************************************/

void consx(void) /* function to find the minimum using Box's complex method */
{
  int lowcenteq;
  int i; /* counter variable */
  int j; /* counter variable */
  int ran; /* set rand0 function not to reset sequence */
  int colcount;

  colcount = 0;
  ran = 0;
  strants(); /* generate random numbers for all points/vars */

  for(i=0; i<k; i++)
  {
    for(j=0; j<n; j++)
    {
      x[i][j]=g[i]+(rand0(4ran)*(h[j]-g[j]));
    }
    limit = i; /* set limit to current point being set */
    check(i); /* check point for constraint violations */
  }
}
lirait = k;

for (i=0; i<k; i++)
{
    func(i);
}

for(j=0; j<itmax; j++)
{
    /* find lowest point */
    lowest = 0;
    for(i=1; i<k; i++)
    {
        if ((f[lowest]-f[i]) > 0.0)
        {
            lowest = i;
        }
    }
    /* find highest point */
    highest = 0;
    for(i=1; i<k; i++)
    {
        if ((f[i]-f[highest]) > 0.0)
        {
            highest = i;
        }
    }
    /* check convergence */
    if ( fabs(f[lowest]) <= beta )
    {
        return;
    }

    /* replace point with lowest function value */
    centroid();
    for(i=0; i<NMAX; i++)
        x[lowest][i]=((1.0+alpha)*xc[i])-alpha*x[lowest][i];
    check(lowest);
    func(lowest);
lowestl = 0;
/* replace point if it repeats as lowest value */
for(i=1; i<KMAX; i++)
{
    if (([f[lowestl]-f[i]] > 0.0 ))
        lowestl = i;
/* record point with lowest function value */
}

while(lowestl==lowest)
/* check same point is still lowest */
{
    for(i=0; i<NMAX; i++)
    {
        x[lowestl][i]=(x[lowestl][i]+xc[i])/2.0;
        /* if its move half way towards the centroid */
    }
    check(lowestl);
    /* check constraints */
    func(lowestl);
    /* evaluate function */
    lowcenteq = 1;
    for(i=0; i<n; i++)
    {
        if( fabs(x[lowestl][i]-xc[i]) > delta ) lowcenteq = 0;
        /* check whether the lowest point has converged */
        /* the centroid */
    }
    if ((lowcenteq == 1) && (fabs(f[lowest]) > beta)) /* if it has print error and exit */
    {
        colcount++;
        printf("collapsed at centroid\n");
        return;
    }
    lowestl = 0;
    /* replace point if it repeats as lowest value */
    for(i=1; i<KMAX; i++)
    {
        if (([f[lowestl]-f[i]] > 0.0 ))
            lowestl = i;
        /* record point with lowest function value */
    }
}
return;
FUNCTION DECLARATION AND ARGUMENT DECLARATION

void centroid(void) /* function to find centroid of all points but */ /* the lowest point */
{
    int j, i; /* counter variables */

    for(j=0; j<n; j++)
    {
        xc[j] = 0.0; /* set all centroid variables to zero */
        for(i=0; i<limit; i++)
        {
            xc[j] = xc[j] + x[i][j]; /* add all equivalent variables at all points */
        }
        if(limit != 1)
            xc[j] = (xc[j] - x[lowest][j])/(limit-1.0); /* take mean of all points excluding the lowest */
        else
            xc[j] = xc[j] / 2.0; /* take mean of all points excluding the lowest */
    }

    return;
}

FUNCTION DECLARATION AND ARGUMENT DECLARATION

double rand0(int *idum) /*Returns a uniform random deviate between 0 and 1.0 using a system supplied routine rand(). */ /*Set idum to any negative value to initialise */ /* or reinitialise the sequence. */ /* From: Numerical Recipes in C */
{
    typedef long time_t;
    static double y, x, maxran = RAND_MAX*1.0, v[98];
    int seed;
    double dum;
    static int iff=0;
    int j;

    if(*idum < 0 || iff == 0) {
        iff=1;
        seed=time(0);
        srand((unsigned)seed);
        *idum=1;
        for (j=1; j<97; j++) dum=rand();
        for (j=1; j<97; j++) v[j]=(double)rand();
    }
}
y=(double)rand();
}
j=97.0*y/maxran;
if (j < 1 || j > 97) puts("out of range error");
y=v[j];
v[j]=rand();
return y/maxran;

/*******************************************************************************/
/* FUNCTION DECLARATION AND ARGUMENT DECLARATION */
/*******************************************************************************/

void func(int v)    /* function to evaluate function to be maximised*/
    /* In this case, the difference between the */
    /* chromatogram data points and sum of all the */
    /* GEX functions evaluated at all points on the */
    /* chromatogram using the current variables */
{
    double function; /* variable containing the function value */
    double time;    /* time value used in the evaluation of the GEX */
    double temp1, temp2, temp3, temp4; /* temporary variables */
    int j;        /* counter variables */

    function = 0.0;
    for(j=0; j<NUMPOINTS; j++)
    {
        time = j * rate;    /* calculate time at point on chromatogram */
        temp1 = gex(time,x[v][1],x[v][2],x[v][3],x[v][4])*x[v][0];    /* calculate value of GEX1 at current time point */
        temp2 = gex(time,x[v][6],x[v][7],x[v][8],x[v][9])*x[v][5];    /* calculate value of GEX2 at current time point */
        temp3 = gex(time,x[v][10],x[v][12],x[v][13],x[v][14])*x[v][10];    /* calculate value of GEX3 at current time point */

        temp4 = temp1 + temp2 + temp3;    /* combine three GEX functions */

        function = function +
            (-(BIGNUM*(chromatogram[j]))-(BIGNUM*temp4)) *    /* calculate difference between sum of GEXs */
            ((BIGNUM*(chromatogram[j]))-(BIGNUM*temp4))/(BIGNUM*BIGNUM);    /* and chromatogram data point at current time */
    }

    f[v] = function;    /* store result in array element */
    return;
}
/*******************************************************************************
/* FUNCTION DECLARATION AND ARGUMENT DECLARATION */
*******************************************************************************

void check(int q) /* check all constraints: explicit and implicit */
{
  /* explicit constraints */
  int ii; /* counter variable */
  label:
  for(ii=0; ii<n; ii++)
  {
    if (x[q][ii] < g[ii]) /* check lower constraint */
      {x[q][ii] = g[ii] + delta;
       goto label;}

    if (h[ii] < x[q][ii]) /* check upper constraint */
      {x[q][ii] = h[ii] - delta;
       goto label;}
  }
  return;
}

/*******************************************************************************
/* FUNCTION DECLARATION AND ARGUMENT DECLARATION */
*******************************************************************************

void straints( )
{
  /* lower limits */
  g[0] = 0.1;
g[1] = 0.0;
g[2] = ((0.42-0.4)*200.0)/22.4;
g[3] = 0.0;
g[4] = 1.0;
g[5] = 0.1;
g[6] = ((6.1-0.40)*200.0)/22.4;
g[7] = ((10.1-0.40)*200.0)/22.4;
g[8] = 0.0;
g[9] = 1.0;
g[10] = 0.1;
g[11] = ((11.7-0.40)*200.0)/22.4;
double gex(double V, double Vo, double Vm, double a, double b)
{
    /* Calculation of General Exponential Function */
    /* Function arguments as described in section 2.2.1.2*/

    double arg, temp1, temp2, temp3, temp4;

    if (V < Vo) return(0.0); /* Check that time (V) > peak start (Vo) */
    arg = (V-Vo)/(Vm-Vo);

    if(Vm < Vo) /* Check consistency of peak start and peak maxima */
    {
        printf("Error in gex - Vm < Vo");
        return(0.0);
    }

    temp1 = pow(arg, (b-1)); /* Calculate GEX function */
    temp2 = (1 - pow(arg, a)) * ((b-1)/a);
    temp3 = exp(temp2);

    temp4 = temp1 * temp3;

    if (temp4 < 0.0) temp4 = 0.0; /* Prevent negative values */

    return(temp4);
}
Variables:
All variables are described in the comments in the above listings.

Functions (Subroutines) descriptions:
main(): This function initialises the random number generator (Press, 1988) and calls consx(), the algorithm function until the target objective function is reached. Finally it prints the vertex of the complex which has the lowest objective function value to an output file.
consx(): Initially the constraints are set by the straits() function (which may if necessary be called each time a vertex is altered which would be necessary if the constraints were dependent upon vertex parameters). Next the vertices of the complex are randomly generated. The algorithm is then executed until either the target objective function value is reached or the complex collapses onto a common point.
centroid(): This function calculates the centroid of the complex.
rand0(): This function generates random numbers between 0 and 1 so that the vertices of the complex may be generated randomly.
func(): The objective function is evaluated in this function at a particular vertex.
check(): In this function altered (or newly generated) vertex parameters are checked (and if necessary changed according to the algorithm) against the constraints set in the straits function.
gex(): This function evaluates the General Exponential Function using a given vertex parameter.

Hardware
This programme was run on an IBM RS/6000 mainframe (Bloomsbury Computing Consortium Batch Service) using the UNIX operating (version AIX).
The programme typically takes 0.5 seconds per iteration (depending upon whether constraints are violated). The total execution time to achieve the desired objective function value is dependent upon this value and the initial randomly chosen starting position of the complex. It will also depend upon whether the complex collapses onto a common point and if it does how many times it does so. This cannot be predicted since it is dependent upon the random generation of the complex. Typically the programme may take up to 10 minutes to converge if 2000 iterations are required for a $\Delta \bar{x}/\bar{x} = 0.2\%$ (see section 2.2.5).
A1.2 Test Chromatogram Generation

The programme listed below is used to generate the Gaussian test chromatogram in section 2.2.6. The programme generates an ASCII output file with 200 points. The first point starts at the beginning of the first peak \( i = rt1 - 4*var \) and the last point at the end of the 3rd peak \( i = rt3 + 4*var \).

```c
#include<stdio.h> /* Standard i/o header */
#include<math.h> /* Maths header */

void main( void )
{
    FILE *fs;
    double i;
    double rtl, rt2, rt3;
    double var;
    double hi, h2, h3;
    double start, stop, inc;

    start = 0.4;
    stop = 22.8;
    rtl = 6.0;
    rt2 = 11.6;
    rt3 = 17.2;
    hi = 9.0;
    h2 = 2.0;
    h3 = 14.0;

    var = 1.4;
    inc = (stop-start)/200.0; /* Size of increment for 200 data points */

    if( (fs = fopen("test1.ref", "w")) == NULL ) /* Open Output file */
        printf("Cannot open file\n"); /* Print error message if this fails */
        return;

    for(i=start; i<stop; i=i+inc) /* Increment time variable from start to */
        /* stop by inc */
        fprintf(fs, "%f %f %f %f\n", i, (gauss(i,rtl,var)*hi), (gauss(i,rt2,var)*h2), (gauss(i,rt3,var)*h3)); /* Print time and individual functions */
        /* to the output file */

    fclose(fs); /* Close output file */

    return;
}
```

226
double gauss(double t, double rt, double var) /* Gaussian function, rt = retention time */
{ /* var = standard deviation of peak */
  return(exp((double)-(((t-rt)/var)*((t-rt)/var))/2));
}

The next programme is used to generate the General Exponential Function test chromatograms in section 2.2.6. The programme generates an ASCII file containing 200 data points between elution times (i) of 0.0 and 15.0.

```c
#include<stdio.h> /* Standard i/o header */
#include<math.h> /* Maths header */
#include<float.h> /* Floating point maths header */

extern void main(void);
static double gex(double, double, double, double, double);

void main(void)
{
  double i; /* Elution time variable */
  FILE *fp; /* Pointer to output file */
  if((fp=fopen("gtest302.prn","w"))==NULL) /* Open output file */
  {
    printf("cannot open output file\n"); /* Print error message if this fails */
    return;
  }

  for(i=0.0; i<15.0; i=i+0.075) /* Increment elution time by 0.075 */
  {
    printf(fp,"%f",i);
  }

  printf(fp,"%f",gex(i,20.659688*0.075,80.074227*0.075,2.439065,16.420665)*8.911385);
  printf(fp,"%f",gex(i,90.166228*0.075,126.414426*0.075,6.208355,2.997291)*1.945080);
  printf(fp,"%f",gex(i,113.835338*0.075,160.182052*0.075,4.158662,6.197101)*13.814451);

  fclose(fp); /* Close the output file */
  return;
}

double gex(double V, double Vo, double Vm, double a, double b) /* Calculation of General Exponential Function*/
/* Function arguments as in section 2.2.1.2 */
```
double arg, temp1, temp2, temp3, temp4;

if ( V < Vo ) return(0.0); /* Check that time (V) > peak start (Vo) */
arg = (V-Vo)/(Vf-Vo);

if(Vm < Vo)
{
    printf("Error in gex - Vm < Vo");
    return (0.0);
}

temp1 = pow( arg, (b-1) ); /* Calculate GEX function */
temp2 = (1 - pow( arg, a ) * ((b-1)/a);
temp3 = exp( temp2 );
temp4 = temp1 * temp3;

if ( temp4 < 0.0 ) temp4 = 0.0; /* Prevent negative values */
return( temp4 );

The 'Noisy' test chromatograms (section 2.2.6) were generated using the
programmes listed above but with the addition of the following multiplier
to the sum of peak functions which was written to the output file:
\* (1+((X*rand0(&j)/100)))
where X was the level of noise added.
The rand0() function (Press 1988) was as follows:
double rand0(int *idum)

returns a uniform random no between 0 and */
\*1.0 using a system supplied routine rand()*/
\* Set idum to any -ve value to initialise */
\* or reinitialise the sequence. */
\* From: Numerical Recipes in C */

typedef long time_t;
static double y,maxran=RAND_MAX+1.0,v[98];
int seed;
double dum;
static int iff=0;
int j;

if(*idum < 0 || iff == 0) {
    iff=1;
    seed=time(0);
    srand((unsigned)seed);
    *idum=1;
    for (j=1;j<97;j++) dum=rand();
    for (j=1;j<97;j++) v[j]=(double)rand();
y=(double)rand();


```c
j=1+97.0*y/maxran;
if (j < 1 || j > 97) puts("out of range error");
y=v[j];
v[j]=rand();
return y/maxran;
```

**Hardware**

Both these programmes were run on a 386SX IBM clone with a maths co-
processor operating under DOS.

**A1.3 : Fuzzy Logic Identification**

The programme listed below was used in Chapter 3 to calculate the elution
order, peak area, and overall match criteria.

```c
#include<stdio.h> /* Standard i/o header */
#include<math.h> /* Maths header */
#define NUM_TRIAL 3 /* Number of trial peaks -- constant */

void main( void );
double intfunc( double, double, double );
double areamemb( double, double );
double ordermemb( int, int );

void main( void )
{

FILE *fp; /* Pointer to output file */

double tempi, temp2;
static double area[][3] =
{ [-23.8571, -143.6975, -99.0960 ],
  [-23.9359, -125.8125, -89.9323 ],
  [-25.5286, -133.1040, -122.7818 ],
  [-24.5878, -129.6848, -100.7836 ],
  [-23.7818, -136.0667, -100.6376 ],
  [-24.1853, -140.3007, -104.7575 ],
  [-26.1068, -106.0222, -94.8536 ],
  [-25.5985, -107.8391, -93.1752 ],
  [-27.8717, -111.1228, -99.0960 ]},
  [-26.1611, -122.1706, -99.3354 ]);

/* Trial peak areas grouped in chromatograms. Three per chromatogram */
double aream[] =
{ -25.1615, -125.5305, -100.5195 };
/* Reference peak data */

int x, y, z; /* Peak & chromatogram indices */

fp = fopen ( "outputa", "w" ); /* Open output file */
```
```c
for ( x = 0; x < 10; x++ )
  /* Select each reference peak */
  { 
    fprintf(fp, "\n %dth data set\n",x+1);
    for ( y = 0; y < 3; y++ ) /* Select each chromatogram */
    {
      fprintf(fp,"\nmatching with %dth reference peak ( area = %f ) \n", y+1, aream[y]);
      for ( z = 0; z < 3; z++ ) /* Compare with each trial peak in */
        /* the xth chromatogram */
        { 
        temp1 = areamemb ( aream[y], area[x][z] ); /* Calculate area match criterion */
          if ( temp1 < 0.0 || temp1 > 1.0 ) temp1 = 0.0;
          temp2 = ordermemb ( ( y + 1 ), ( z + 1 ) ); /* Calculate order match criterion */
          fprintf( fp,"area = %f(%d)	areamf = %f	ordermf = if toverall = %f\n",area[x][z],z+1,tempi,temp2,(tempi+temp2)/2 ); /* Print results to the output file */
        } 
    }
  }
fclose( fp ); /* Close output file */
}
return;
}

double areamemb( double arearef, double areastrial ) /* Function to calculate the area */
  /* match criterion, using a parabolic mf */
{
  double atrial, aref; /* a parameter in equation 3.1 */
  double error; /* Fractional error in peak area */
  double a, b, c; /* Intermediates for calculation of */
  double start1, start2, stop1, stop2; /* Effective starts and ends of mf's */
  double intarea1, intarea2; /* Area of intersection of mf's parts 1+2 */
  double alltrial, allref; /* Total area of mf's */
  double totintarea, normarea; /* Total intersection area */
  double tempi; /* Largest mf area used for norm. */
```
error = 0.10;

if ( -arearef < -areatrial )
    |
    templ = arearef;
    arearef = areatrial;
    areatrial = templ;
    /\ Check that ref area > trial area /*
    /\ If not exchange */

aref = ( arearef * error ) * ( arearef * error );  /* Calculate a parameters */
atrial = ( areatrial * error ) * ( areatrial * error );

start1 = -arearef - sqrt( aref );  /* Calculate 1st mf part start */
a = (1.0 / aref) - (1.0 / atriial);
b = 2.0 * ((arearef/aref) - (areatrial/atrial));
c = ((arearef*arearef)/aref) - ((areatrial*areatrial)/atrial);

stop1 = ( -b - sqrt( (b*b) - (4.0*a*c) )) / (2.0*a);  /* Calculate 1st mf part end */
start2 = stop1;
stop2 = -areatrial + sqrt( atriial );

intarea1 = intfunc( aref, arearef, start1 ) - intfunc( atriial, arearef, start1 );
intarea2 = intfunc( atriial, areatrial, start2 ) - intfunc( atriial, areatrial, stop2 );

totintarea = intarea1 + intarea2;  /* Calculate Total and 2 part */
/* intersection areas */

allref = intfunc( aref, arearef, ( -arearef-sqrt(aref) ) ) - intfunc( aref, arearef, ( -arearef+sqrt(aref) ) );

allatrial = intfunc( atriial, areatrial, ( -areatrial-sqrt(atrial) ) ) - intfunc( atriial, areatrial, ( -areatrial+sqrt(atrial) ) );

if ( allref > allatrial ) normarea = allref;
else normarea = allatrial;

return( totintarea / normarea );  /* Normalise with larger mf area */
}

double intfunc( double sprd, double area, double x )  /* Function to calculate mf area */
{
    /\ from start until x */

double templ, temp2, temp3;

    templ = -( x * x * x ) / ( 3.0 * sprd );
    temp2 = -( area * x * x ) / sprd;
    temp3 = -( area * area * x ) / sprd;

    return( templ + temp2 + temp3 - x );
double ordermemb(int refrank, int trialrank)
{
    /* Func to calc order match criterion */
    double temp; /* with for refrank and trialrank */
    temp = (1 - fabs((double)refrank - (double)trialrank)/(NUM_TRIAL - 1));
    if (temp < 0.0 || temp > 1.0) temp = 0.0;
    return (temp);
}

Programme Description:
main() : This function contains the trial and reference chromatogram data. It also carries out the task of comparing each trial peak, in each trial chromatogram, with each peak in the reference set.
areamemb() : This function calculates the peak area match criteria. The function uses a parabolic membership function (section 3.2.2). The area of intersection is found in two parts. The first portion is found by integrating the larger peak area membership function between the function start and the intersection with the smaller peak area membership function. The second part is found by integrating the smaller peak from the intersection to the function end (See figure A1.1).
Variables a, b and c are coefficients of a quadratic function used to model intersection of the two membership functions. The quadratic function describing the point of intersection x is found by equating the two functions:
\[
\frac{(x + d_1)^2}{a_1} = \frac{(x + d_2)^2}{a_2}
\]
\[A1.1\]
Expanding and rearranging gives:
\[
x^2\left(\frac{1}{a_2} - \frac{1}{a_2}\right) + 2x\left(\frac{area_{m2}}{a_2} - \frac{area_{ml}}{a_1}\right) + \frac{area_{m2}^2}{a_2} + \frac{area_{ml}^2}{a_1} = 0 \quad A1.2
\]

intfunc() : The integration of the peak area membership functions are carried out in this function which assumes a parabolic peak area membership function (section 3.2.2).
ordermemb(): In this function the elution order match criterion is calculated between the given trial and reference ranks.

Hardware
This programme was run on a 386SX IBM clone with a maths co-processor operating under DOS.
The execution time for the system shown above i.e., ten trial chromatograms each with three peaks is approximately 85 seconds.

A1.4 Deconvoluted Peak Area Determination
This programme was used in section 3.3 to calculate the areas of individual peak elution profiles obtained by deconvolution. The programme uses the general exponential function parameters obtained by deconvolution to integrate the peak functions using a Gaussian - Quadratures integration subroutine (Press, 1988).

```c
#include<stdio.h> /* standard input / output header */
#include<io.h> /* special input / output header */
#include<float.h> /* floating point header */
#include<math.h> /* maths header */
#include<stdlib.h> /* standard library header file */

#define BIGNUM 1e5 /* a large number to prevent rounding errors */

static double qgaus(double(*)(double,double,double,double,double,double,double,double,double,double,double,void main(void):

```
vol = 33.333428;  
vml = 38.166738;  
a1 = 6.313147;  
b1 = 24.968281;  
hml = 0.849858;  

vo2 = 38.001975;  
vm2 = 40.500062;  
a2 = 1.445090;  
b2 = 2.179509;  
hm2 = 0.123772;  

vo3 = 40.000079;  
vm3 = 42.166767;  
a3 = 0.989805;  
b3 = 24.615135;  
hm3 = 0.756907;  

inc = (end - start) / 200.0;  /* calculate peak increment given */  /* start and end parameters */
peak1 = 0.0;  
peak2 = 0.0;  
peak3 = 0.0;  
k = 120.0;  

if(fp = fopen("eggfuzz6.prn","w")) == NULL)  /* open output file */  
  printf("Cannot open output file\n");  /* print error if this fails */  
  return;

for(stop = start; stop < end; stop = stop + inc)  /* Calculate area of fractions of peak */  /* between start and end */
  
  temp1 = qgaus(gex, start, stop, vol, vml, a1, b1) * hml * k;  /* Calculate the area of peak fractions */  
  temp2 = qgaus(gex, start, stop, vo2, vm2, a2, b2) * hm2 * k;  /* between start and stop using */  
  temp3 = qgaus(gex, start, stop, vo3, vm3, a3, b3) * hm3 * k;  /* the qgaus function */  
  peak1 = peak1 + temp1;  /* Calculate the running total */  /* of the peak functions */  
  peak2 = peak2 + temp2;  
  peak3 = peak3 + temp3;  

start = stop;  /* move new fraction to end of old */

fprintf(fp, "%f %f %f \n", peak1, peak2, peak3);  /* output the results to a file */
fclose( fp ); /* close output file */
return;

double ggaus(double (*func)(double, double, double, double, double), double a, double b, double c, double d, double e, double f)
    /* Function to integrate a function using Gaussian quadratures */
    /* from Numerical Recipes in C */
    /* a-lower integration limit, b-higher limit, c-Vo, d-Vm, e-a, f-b */
{
    int j;
    double xr, xm, dx, s;
    static double x[]={0.0, 0.1488743389, 0.4339753941,
                      0.8794095682, 0.8656336666, 0.9739065234};
    static double w[]={0.0, 0.2955242247, 0.2692667193,
                       0.2190863625, 0.1494513491, 0.06671344};
    xm = 0.5*(b+a);
    xr = 0.5*(b-a);
    s = 0.0;
    for (j=1; j<=5; j++)
        dx = xr*x[j];
        s += w[j]*(*func)(xm+dx, c, d, e, f) + (*func)(xm-dx, c, d, e, f);
    return s * xr;
}

double gex( double V, double Vo, double Vm, double a, double b)
    /* GEX Function - see above */
{
    double arg, temp1, temp2, temp3, temp4;
    if (V < Vo)
        return(0.0);
    arg = (V-Vo)/(Vm-Vo);

    if (Vm < Vo)
        printf("Error in gex\n");
        return(0.0);
    }
    temp1 = pow( arg, (b-1) );
    temp2 = (1 - pow( arg, a )) * ((b-1)/a);
    temp3 = exp( temp2 );
    temp4 = temp1 * temp3;
    if (temp4 < 0.0) temp4 = 0.0;
    return( temp4 );
}
A1.5 Fractionation Curve Calculation

This programme was used in chapter 3 to calculate the fractionation curves using the general exponential peak function parameters found by deconvolution in Chapter 2 for both the test chromatograms and the experimental chromatograms. The programme uses Gaussian-Quadratures subroutine (Press, 1988) to integrate the peak functions and takes the first and third peaks to be the contaminating peaks and the second to be the product component. The information obtained - the amount of product, total amount of protein and volume - up to the end of the portion of the fraction analysed, is written to an output file.

```c
#include<stdio.h>  // standard input / output header file
#include<io.h>     // input output header file
#include<float.h>  // floating point arithmetic header
#include<math.h>   // maths header file
#include<stdlib.h> // standard library header file

static double qgaus(double(*)(double,double,double,double,double), double, double, double, double,
                     double, double);  // Gaussian-Quadratures subroutine
static double gex(double, double, double, double, double);  // exponential peak function
extern void main(void);

void main()
{
    FILE *fp;    // pointer to the output file
    double start, stop, inc, end;  // start, end of fraction
    double temp1, temp2, temp3;  // end of chromatogram & increment
    double vol, vo2, vo3;        // peak starts
    double vml, vm2, vm3;        // peak maxima
    double a1, a2, a3;           // peak skew parameters
    double b1, b2, b3;           // peak skew parameters
    double hm1, hm2, hm3;        // peak height parameters
    double total;                // cumulative total mass of protein
    double prod;                 // cumulative mass of product

    start = 0.0;
    stop = 0.0;
    end = 25.0;

    vol = 3.3;
    vml = 6.0;
    a1 = 3.849;
    b1 = 24.98;
    hml = 8.82;
```
vo2 = 5.80;
v2m = 7.30;
a2 = 0.785;
b2 = 21.24;
hm2 = 1.99;

vo3 = 8.10;
v3m = 9.30;
a3 = 8.895;
b3 = 6.898;
hm3 = 13.89;

prod = 0.0;
total = 0.0;

inc = ( end - start ) / 200.0; /* calculate increment for 200 pts */
if((fp = fopen("ftg24.prn","w") == NULL) /* open output file */
  
  printf("Cannot open output file\n");
  return;
)

for( stop = start; stop < end; stop = stop + inc ) /* loop segment through chromatogram */

  temp1 = qgaus(gex, start, stop, vo1, vm1, a1, b1) * hml; /* calculate area of 1st peak segment */
  temp2 = qgaus(gex, start, stop, vo2, vm2, a2, b2) * hm2; /* calculate area of 2nd peak segment */
  temp3 = qgaus(gex, start, stop, vo3, vm3, a3, b3) * hm3; /* calculate area of 3rd peak segment */

  prod = prod + temp2; /* calculate cumulative total */
  total = total + temp1 + temp2 + temp3; /* calculate cumulative total */

  printf(fp, "%f %f %f \n", stop, prod, total); /* print results to output file */

  start = stop;

fclose(fp); /* close output file */
return;

double qgaus(double(*func)(double, double, double, double, double),double a,double b, double c, double d, double e, double f)
/* Gaussian Quadratures integration routine */
/* from Numerical Recipes in C, Press 1988 */
/* a-lower integration limit, b-higher limit, c-Vo, d-Vm, e-a, f-b */

int j;
double xr,xm,dx,s;
static double x[]={0.0,0.1488743389,0.4333953941,
0.6794095682,0.8650633666,0.97390652};
static double w[]={0.0,0.2955242247,0.2692667193,
0.2190663259,0.1494513491,0.06667134};
xm=0.5*(b+a);
xr=0.5*(b-a);
s=0.0;
for (j=1; j<=5; j++) {
    dx=xr*x[j];
    s += w[j]*(*func)(xm+dx,c,d,e,f)+(*func)(xm-dx,c,d,e,f);
}
return s * xr;

double gex( double V, double Vo, double Vm, double a, double b)
/* general exponential function */
/* see previous programmes */
/* for description */
{
    double arg, temp1, temp2, temp3, temp4;
    if( V < Vo )
        return(0.0);
    arg = (V-Vo)/(Vm-Vo);
    if( Vm < Vo )
    {
        printf("Error in gex\n");
        return(0.0);
    }
    temp1 = pow( arg, (b-1) );
    temp2 = (1 - pow( arg, a )) * ((b-1)/a);
    temp3 = exp( temp2 );
    temp4 = temp1 * temp3;
    if (temp4 < 0.0) temp4 = 0.0;
    return( temp4 );
}

A1.6 Optimum Product Fraction Selection
This programme uses the output file produced by the programme described in section A1.5 to determine the composition, yield and purification factor as well as information about the position of the product.
fraction under analysis. An initial purification factor is set and then all possible fractions with this particular purification factor are analysed, using a tie line to represent the product fraction. The tie line is started at one point on the curve and the end point is increased in the x coordinate (total protein) until it crosses the curve (ie. the difference between the tie line and curve becomes negative). Any fractions with greater than 60% of the maximum yield are recorded. The purification factor is then incremented and the process repeated until no fractions can be found at a particular purification factor.

```c
#include <stdio.h>

#define NUMPOINTS 200 /* Number of points on curve */
#define MOSOL -2 /* flag indicating no fraction */
#define MAXPF 13 /* maximum purification factor */

void main ( void );
void interpolate ( void );
void fracdata ( void );

double startx, starty, pf; /* start & end of fraction and pf */
double grada, gradb, ca, cb; /* gradient and intercept of tie line */
/* and two curve points */
double vol[] = { /* Fractionation curve volume data */
};
double prod[] = { /* Fractionation curve product data */
};
double total[] = { /* Fractionation curve total protein data */
};
double diff, prodc2, prodc1, totalc2, totalc1, volc1, volc2; /* difference between tie line and curve */
/* start and end points in all units of */
/* the tie line and curve */
int sign, lastsign; /* current and last sign */
double pf, yield, startvol, endvol;
double injurty, maxtot, maxprod;
int x, c, i, cl, c2; /* data point indices */
```
FILE *fp;
/
* pointer to output file */

void main ( void ) {
{
f = fopen("fracdata","w"); /* open output file */

fprintf(fp,"pf tvol1 tvol2 totalL total2 yield
total purity prodconc totalconc\n"); /* print output file heading */

maxtot = total[NUMPOINTS-1]; /* set maximum yields and total */
maxprod = prod[NUMPOINTS-1];
in_purity = maxprod / maxtot; /* set input purity */
for ( pf = 7.0; pf < MAXPF; pf = pf + 0.25 ) /* search from pf = 7.0 to MAXPF */{
{
for ( i = 8; i < NUMPOINTS; i++ ) /* search through data pts for line start */{
{
startx = total[i];
starty = prod[i];
lastsign = -1;

for ( c = i + 1; c < NUMPOINTS; c++ ) /* search through data pts for line end */{
{
cl = c - 1;
c2 = c;
diff = prod[c] - ( (pf*in_purity*total[c]) + prod[i] - (pf*in_purity*total[i]) ) ; /* calculate difference between curve & */
/* tie line */

if ( diff < 0.0 ) sign = -1; /* define sign */
if ( diff > 0.0 ) sign = +1;

if ( sign != lastsign ) break; /* if sign changes break & calculate */
/* product fraction information */
}
}
}
}

if ( c == ( NUMPOINTS - 1 ) || c == NUMPOINTS )
continue; /* check if end of curve is reached */
prodcl = prod[c1];
prod2 = prod[c2];
totalc1 = total[c1];
totalc2 = total[c2];

startvol = vol[i];
endvol = vol[c2];

if ((prod2 - prod[i]) > (0.6 * maxprod)) fracdata();  
   /* if yield is > 60% output data */
   
}

fclose(fp);
   /* close output file */

return;

}

void fracdata(void)
{

fprintf(fp, "%f\%f\%f\%f\%f\%f\%f\%f\n", pf, vol[i], vol[c2], total[i], totalc2, (prod2 - prod[i]),
   (totalc2 - total[i]), (prod2 - prod[i]) / (totalc2 - total[i]),
   (prod2 - prod[i]) / (vol[c2] - vol[i]),
   (totalc2 - total[i]) / (vol[c2] - vol[i]));
   /* output yield, fraction position, */
   /* composition etc. */

return;

}
Figure A1: The calculation method for the matching criterion used by the fuzzy logic identification programme (A1.3). The two portions of the area of intersection are shown. The intersection point is defined by equation A1.2.
Appendix A2

Deconvolution Results
Appendix A2: Deconvolution Results

This appendix contains the results of the deconvolutions of the chromatograms described in section 2.2.5 and section 2.3.

A2.1 Gaussian Test Chromatograms

Table A2.1 First Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta x/\bar{x}$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_\theta$</th>
<th>$V_\theta$</th>
<th>$V_\phi$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7862</td>
<td>2.2%</td>
<td>-1.497</td>
<td>1</td>
<td>8.99</td>
<td>0.5</td>
<td>5.9</td>
<td>2.649</td>
<td>6.526</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.89</td>
<td>8.5</td>
<td>11.3</td>
<td>0.322</td>
<td>10.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>13.7</td>
<td>12.1</td>
<td>17.2</td>
<td>2.721</td>
<td>6.039</td>
</tr>
</tbody>
</table>

(All $h_\theta$'s in arbitrary absorbance units. All $V_\theta$'s and $V_\phi$'s in arbitrary elution volume units)

Table A2.2 Second Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta x/\bar{x}$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_\theta$</th>
<th>$V_\theta$</th>
<th>$V_\phi$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>375</td>
<td>4.8%</td>
<td>-1.492</td>
<td>1</td>
<td>1.97</td>
<td>0.4</td>
<td>4.0</td>
<td>0.464</td>
<td>24.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>8.78</td>
<td>4.5</td>
<td>7.6</td>
<td>2.494</td>
<td>5.879</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1.82</td>
<td>10.6</td>
<td>13.1</td>
<td>1.914</td>
<td>4.237</td>
</tr>
</tbody>
</table>

(All $h_\theta$'s in arbitrary absorbance units. All $V_\theta$'s and $V_\phi$'s in arbitrary elution volume units)

Table A2.3 Third Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta x/\bar{x}$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_\theta$</th>
<th>$V_\theta$</th>
<th>$V_\phi$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6282</td>
<td>4.1%</td>
<td>-2.498</td>
<td>1</td>
<td>8.70</td>
<td>0.4</td>
<td>4.0</td>
<td>2.012</td>
<td>7.605</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>8.90</td>
<td>5.6</td>
<td>7.6</td>
<td>2.788</td>
<td>2.596</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>2.10</td>
<td>10.2</td>
<td>13.5</td>
<td>2.839</td>
<td>5.044</td>
</tr>
</tbody>
</table>

(All $h_\theta$'s in arbitrary absorbance units. All $V_\theta$'s and $V_\phi$'s in arbitrary elution volume units as in table 2.2)
### Table A2.4 Fourth Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>2900</th>
<th>$\Delta x / x$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_1$</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-1.444</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.03</td>
<td>0.4</td>
<td>4.0</td>
<td>0.866</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.91</td>
<td>5.6</td>
<td>7.6</td>
<td>3.324</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8.77</td>
<td>9.9</td>
<td>13.2</td>
<td>3.086</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_1$'s in arbitrary absorbance units. All $V_0$'s and $V_1$'s in arbitrary elution volume units)

### Table A2.5 Fifth Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>900</th>
<th>$\Delta x / x$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_1$</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-1.493</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2.09</td>
<td>2.6</td>
<td>4.0</td>
<td>1.900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9.09</td>
<td>3.6</td>
<td>6.6</td>
<td>2.508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.09</td>
<td>7.4</td>
<td>8.9</td>
<td>1.903</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_1$'s in arbitrary absorbance units. All $V_0$'s and $V_1$'s in arbitrary elution volume units)

### A2.2 General Exponential Function Test chromatograms

### Table A2.5 First General Exponential Function chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>764</th>
<th>$\Delta x / x$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_1$</th>
<th>$V_0$</th>
<th>$V_1$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-1.497</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8.68</td>
<td>0.3</td>
<td>6.0</td>
<td>12.619</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.00</td>
<td>4.2</td>
<td>8.9</td>
<td>4.440</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>13.97</td>
<td>8.2</td>
<td>12.0</td>
<td>8.269</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_1$'s in arbitrary absorbance units. All $V_0$'s and $V_1$'s in arbitrary elution volume units as in table 2.3 )
### Table A2.6 Second General Exponential Function chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta \frac{\Delta x}{\Delta x} = 8.7%$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_i$</th>
<th>$V_0$</th>
<th>$V_f$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2148</td>
<td>-1.499</td>
<td></td>
<td>1</td>
<td>8.68</td>
<td>1.0</td>
<td>6.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.79</td>
<td>5.2</td>
<td>7.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>13.65</td>
<td>8.0</td>
<td>9.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_i$'s in arbitrary absorbance units. All $V_0$'s and $V_f$'s in arbitrary elution volume units as in table 2.3)

### Table A2.7 Third General Exponential Function chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta \frac{\Delta x}{\Delta x} = 2.8%$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_i$</th>
<th>$V_0$</th>
<th>$V_f$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>604</td>
<td>-1.497</td>
<td></td>
<td>1</td>
<td>8.95</td>
<td>2.3</td>
<td>6.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.89</td>
<td>7.8</td>
<td>9.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>13.94</td>
<td>8.6</td>
<td>12.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_i$'s in arbitrary absorbance units. All $V_0$'s and $V_f$'s in arbitrary elution volume units as in table 2.3)

### A2.3 Noisy Test Data

### Table A2.8 Noisy Gaussian chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>$\Delta \frac{\Delta x}{\Delta x} = 2.1%$</th>
<th>$f$</th>
<th>peak number</th>
<th>$h_i$</th>
<th>$V_0$</th>
<th>$V_f$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2275</td>
<td>-1.494</td>
<td></td>
<td>1</td>
<td>9.05</td>
<td>0.4</td>
<td>6.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1.90</td>
<td>6.4</td>
<td>11.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>14.00</td>
<td>11.4</td>
<td>17.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(All $h_i$'s in arbitrary absorbance units. All $V_0$'s and $V_f$'s in arbitrary elution volume units as in table 2.3)
Table A2.9 Noisy General Exponential Function chromatogram deconvolution results

<table>
<thead>
<tr>
<th>iteration no</th>
<th>1972</th>
<th>$\Delta x/x$ = 4.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-1.492</td>
<td></td>
</tr>
<tr>
<td>peak number</td>
<td>$h_1$</td>
<td>$V_1$</td>
</tr>
<tr>
<td>1</td>
<td>8.99</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>1.97</td>
<td>5.0</td>
</tr>
<tr>
<td>3</td>
<td>13.99</td>
<td>8.5</td>
</tr>
</tbody>
</table>

(All $h$'s in arbitrary absorbance units. All $V_1$'s and $V_2$'s in arbitrary elution volume units as in table 2.3.)
### A2.4 Egg White Chromatograms

#### Table A2.10 Experimental deconvolution results

<table>
<thead>
<tr>
<th>Iteration</th>
<th>2090</th>
<th>( \Delta x / x = 0.2 % )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-72.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>29.21</td>
<td>69.24</td>
<td>80.97</td>
</tr>
<tr>
<td>2</td>
<td>6.50</td>
<td>83.18</td>
<td>92.66</td>
</tr>
<tr>
<td>3</td>
<td>45.82</td>
<td>92.84</td>
<td>96.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>2070</th>
<th>( \Delta x / x = 0.4 % )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-180.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>29.32</td>
<td>66.67</td>
<td>79.66</td>
</tr>
<tr>
<td>2</td>
<td>6.64</td>
<td>80.92</td>
<td>89.37</td>
</tr>
<tr>
<td>3</td>
<td>41.41</td>
<td>88.34</td>
<td>91.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>802</th>
<th>( \Delta x / x = 0.4 % )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-183.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>50.99</td>
<td>63.43</td>
<td>76.33</td>
</tr>
<tr>
<td>2</td>
<td>7.43</td>
<td>76.00</td>
<td>81.00</td>
</tr>
<tr>
<td>3</td>
<td>45.41</td>
<td>80.00</td>
<td>84.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>4022</th>
<th>( \Delta x / x = 0.3 % )</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>-183.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>76.04</td>
<td>63.43</td>
<td>69.02</td>
</tr>
<tr>
<td>2</td>
<td>6.89</td>
<td>71.89</td>
<td>79.78</td>
</tr>
<tr>
<td>3</td>
<td>60.72</td>
<td>78.34</td>
<td>82.72</td>
</tr>
</tbody>
</table>

(All \( h_n \)'s in mV. All \( V_0 \)'s and \( V_w \)'s in ml.)
### Table A2.10b Experimental deconvolution results

#### Egg White Separation 5

<table>
<thead>
<tr>
<th>iteration no</th>
<th>7671</th>
<th>( \Delta x / x = 0.4% )</th>
<th>f</th>
<th>peak number</th>
<th>( h_m )</th>
<th>( V_o )</th>
<th>( V_m )</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>-226.8</td>
<td></td>
<td>1</td>
<td>41.48</td>
<td>64.89</td>
<td>68.43</td>
<td>0.624</td>
<td>23.34</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>9.49</td>
<td>71.86</td>
<td>77.92</td>
<td>8.458</td>
<td>2.162</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>57.88</td>
<td>78.14</td>
<td>80.37</td>
<td>0.370</td>
<td>23.10</td>
</tr>
</tbody>
</table>

#### Egg White Separation 6

<table>
<thead>
<tr>
<th>iteration no</th>
<th>158</th>
<th>( \Delta x / x = 0.4% )</th>
<th>f</th>
<th>peak number</th>
<th>( h_m )</th>
<th>( V_o )</th>
<th>( V_m )</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>-172.8</td>
<td></td>
<td>1</td>
<td>25.43</td>
<td>75.02</td>
<td>77.70</td>
<td>0.565</td>
<td>23.52</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>12.81</td>
<td>76.76</td>
<td>80.40</td>
<td>1.388</td>
<td>7.149</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>44.21</td>
<td>78.46</td>
<td>84.77</td>
<td>10.03</td>
<td>4.011</td>
</tr>
</tbody>
</table>

(All \( h_m \)'s in mV. All \( V_o \)'s and \( V_m \)'s in mL.)
Appendix A3

Fuzzy Logic Matching Results
Appendix 3 Fuzzy Logic Matching Results

A3.1 Identification Results for the Egg White Chromatograms

Table A3.1 Matching Results for the 1st egg white chromatogram

<table>
<thead>
<tr>
<th>Matching with 1st Reference Peak (area = 105.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 118.0(1)</td>
</tr>
<tr>
<td>area = 45.5(2)</td>
</tr>
<tr>
<td>area = 118.3(3)</td>
</tr>
</tbody>
</table>

Matching with 2nd Reference Peak (area = 48.9)

| area = 118.0(1) | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |
| area = 45.5(2)  | areamf = 0.83 | ordermf = 1.00 | overall = 0.92 |
| area = 118.3(3) | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |

Matching with 3rd Reference Peak (area = 110.5)

| area = 118.0(1) | areamf = 0.85 | ordermf = 0.00 | overall = 0.42 |
| area = 45.5(2)  | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |
| area = 118.3(3) | areamf = 0.84 | ordermf = 1.00 | overall = 0.92 |

Table A3.2 Matching Results for the 2nd egg white chromatogram

<table>
<thead>
<tr>
<th>Matching with 1st Reference Peak (area = 105.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 103.7(1)</td>
</tr>
<tr>
<td>area = 47.4(2)</td>
</tr>
<tr>
<td>area = 116.3(3)</td>
</tr>
</tbody>
</table>

Matching with 2nd Reference Peak (area = 48.9)

| area = 103.7(1) | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |
| area = 47.4(2)  | areamf = 0.93 | ordermf = 1.00 | overall = 0.96 |
| area = 116.3(3) | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |

Matching with 3rd Reference Peak (area = 110.5)

| area = 103.7(1) | areamf = 0.85 | ordermf = 0.00 | overall = 0.43 |
| area = 47.4(2)  | areamf = 0.00 | ordermf = 0.50 | overall = 0.25 |
| area = 116.3(3) | areamf = 0.88 | ordermf = 1.00 | overall = 0.94 |
### Table A3.3 Matching Results for the third egg white chromatogram

<table>
<thead>
<tr>
<th>3rd data set, matching with 1st reference peak (area = 105.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 100.8(1)</td>
</tr>
<tr>
<td>area = 70.0(2)</td>
</tr>
<tr>
<td>area = 102.4(3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>matching with 2nd reference peak (area = 48.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 100.8(1)</td>
</tr>
<tr>
<td>area = 70.0(2)</td>
</tr>
<tr>
<td>area = 102.4(3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>matching with 3rd reference peak (area = 110.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 100.8(1)</td>
</tr>
<tr>
<td>area = 70.0(2)</td>
</tr>
<tr>
<td>area = 102.4(3)</td>
</tr>
</tbody>
</table>

### Table A3.4 Matching Results for the fourth egg white chromatogram

<table>
<thead>
<tr>
<th>4th data set, matching with 1st reference peak (area = 105.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 175.0(1)</td>
</tr>
<tr>
<td>area = 51.5(2)</td>
</tr>
<tr>
<td>area = 110.9(3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>matching with 2nd reference peak (area = 48.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 175.0(1)</td>
</tr>
<tr>
<td>area = 51.5(2)</td>
</tr>
<tr>
<td>area = 110.9(3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>matching with 3rd reference peak (area = 110.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 175.0(1)</td>
</tr>
<tr>
<td>area = 51.5(2)</td>
</tr>
<tr>
<td>area = 110.9(3)</td>
</tr>
</tbody>
</table>
### Table A3.5 Matching Results for the fifth egg white chromatogram

<table>
<thead>
<tr>
<th>5th data set, matching with 1st reference peak ( area = 105.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 100.0(1) areafm = 0.88 orderm f = 1.00 overall = 0.94</td>
</tr>
<tr>
<td>area = 45.8(2) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>area = 114.6(3) areafm = 0.81 orderm f = 0.00 overall = 0.40</td>
</tr>
<tr>
<td>matching with 2nd reference peak ( area = 48.9 )</td>
</tr>
<tr>
<td>area = 100.0(1) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>area = 45.8(2) areafm = 0.85 orderm f = 1.00 overall = 0.93</td>
</tr>
<tr>
<td>area = 114.6(3) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>matching with 3rd reference peak ( area = 110.5 )</td>
</tr>
<tr>
<td>area = 100.0(1) areafm = 0.77 orderm f = 0.00 overall = 0.38</td>
</tr>
<tr>
<td>area = 45.8(2) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>area = 114.6(3) areafm = 0.91 orderm f = 1.00 overall = 0.96</td>
</tr>
</tbody>
</table>

### Table A3.6 Matching Results for the sixth egg white chromatogram

<table>
<thead>
<tr>
<th>6th data set, matching with 1st reference peak ( area = 105.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>area = 48.9(1) areafm = 0.00 orderm f = 1.00 overall = 0.50</td>
</tr>
<tr>
<td>area = 39.8(2) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>area = 134.7(3) areafm = 0.44 orderm f = 0.00 overall = 0.22</td>
</tr>
<tr>
<td>matching with 2nd reference peak ( area = 48.9 )</td>
</tr>
<tr>
<td>area = 48.9(1) areafm = 1.00 orderm f = 0.50 overall = 0.75</td>
</tr>
<tr>
<td>area = 39.8(2) areafm = 0.53 orderm f = 1.00 overall = 0.76</td>
</tr>
<tr>
<td>area = 134.7(3) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>matching with 3rd reference peak ( area = 110.5 )</td>
</tr>
<tr>
<td>area = 48.9(1) areafm = 0.00 orderm f = 0.00 overall = 0.00</td>
</tr>
<tr>
<td>area = 39.8(2) areafm = 0.00 orderm f = 0.50 overall = 0.25</td>
</tr>
<tr>
<td>area = 134.5(3) areafm = 0.55 orderm f = 1.00 overall = 0.77</td>
</tr>
</tbody>
</table>
Appendix A4.

Nomenclature.
Appendix A4 Nomenclature

Chapter One: Introduction

$A_i$ - Area of peak $i$ ( mV or AU or mg/mL )

$A1$ - absorbance signal at wavelength 1 ( AU )

$A2$ - absorbance signal at wavelength 2 ( AU )

$A, B, C$ - constants in Van Deemter equation ( as appropriate )

$A/B$ - skew ratio: ratio of widths either side of peak maximum ( dimensionless )

$a(x)$ - measured peak area used in the membership function ( AU.mL or mV.mL )

$b$ - constant in Langmuir equation related to maximum adsorbate loading.

$C_A$ - concentration of component A in mobile phase ( mg/mL )

$c$ - the concentration of solute in the mobile phase ( g solute / g solvent )

$CRF$ - chromatographic response function ( dimensionless )

$d_p$ - stationary phase particle diameter ( m )

$D_i$ - diffusion coefficient of solute in the mobile phase ( m$^2$/s )

$D_x$ - axial diffusion coefficient ( mL.s / mg )

$h$ - reduced plate height ( dimensionless )

$H$ - height of a theoretical plate ( m )

$HETP$ - height of a theoretical plate ( m )
$h_i$ - height of peak$_i$ at a given elution volume 

                     ( mV or AU or mg/mL )

$K$ - Langmuir equilibrium constant

$k$ - number of peaks in multicomponent optimisation criterion 

                     ( dimensionless )

$k_j$ - capacity factor ( dimensionless )

$L, R$ - labels to left and right hand spreads, respectively, of membership functions

$m(x)$ - membership function ( dimensionless )

$m(z)$ - membership function ( dimensionless )

$m_a$ - membership function ( dimensionless )

$N_{ST}$ - actual number of plates for real skewed peaks ( dimensionless )

$P$ - Peak separation ( dimensionless )

$P_{inf}$ - informing power ( dimensionless )

$R$ - mass transfer rate ( mg / mL.s )

$RAT$ - ratio of two absorbance signal ( dimensionless )

$R_s$ - resolution parameter describing the separation between two peaks 

                     ( dimensionless )

$SN$ - separation number ( dimensionless )

$t$ - total separation time ( s )

$V_i$ - retention volume of peak$_i$ ( mL )

$V_{Ri}$ - retention volume of peak$_i$ ( mL )

$V'_{Ri}$ - retention volume of peak$_i$ relative to the retention volume of the
unretained peak ( mL )

\( V_0 \) - retention volume of unretained peak ( mL )

\( V_e \) - effective total dispersion volume ( mL )

\( V_p \) - dispersion volume due to particles ( mL )

\( V_{ec} \) - dispersion volume due to extra column effects ( mL )

\( V_{inj} \) - dispersion volume due to an injection valve ( mL )

\( V_{det} \) - dispersion volume due to the detector ( mL )

\( v \) - reduced fluid velocity ( dimensionless )

\( V \) - valley to peak ratio ( dimensionless )

\( w_i \) - weighting of peak pair \( i \) in a multicomponent optimisation criterion ( dimensionless )

\( w_i \) - width of peak \( i \) ( mL )

\( w \) - the concentration of solute on solid packing ( g solute / g solid )

\( w_{0.1} \) - width of peak at 10% of maximum height ( mL )

\( x \) - distance down column ( m )

\( a_i \) - left hand spread of membership function ( AU.mL or mV.mL )

\( a_{ji} \) - the selectivity of peak \( i \) relative to peak \( j \) ( dimensionless )

\( \beta_i \) - right hand spread of membership function ( AU.mL or mV.mL )

\( \sigma_i \) - standard deviation of Gaussian peak \( i \) ( mL )

\( \tau_i \) - exponentially modified Gaussian decay constant ( mL )
\( \phi \) - total overlap (dimensionless)

\( \Omega \) - fractional peak overlap (dimensionless)

Chapter Two: Deconvolution

\( A \) - the area of the peak function (mL.mV)

\( A \) - ultra-violet absorbance (AU)

\( a \) - a general exponential peak function skew parameter (dimensionless)

\( \Delta \text{Area} \) - the difference in area between the actual and model chromatograms (mV.mL)

\( b \) - a general exponential peak function skew parameter (dimensionless)

\( c_i \) - the concentration of the \( i^{\text{th}} \) component in the flow-cell (mg mL\(^{-1}\))

\( D_i \) - the \( i^{\text{th}} \) data point in a chromatogram

\( f \) - the deconvolution objective function (dimensionless)

\( g_j \) - lower constraint bound for peak function parameter (as appropriate)

\( h_j \) - upper constraint bound for peak function parameter (as appropriate)

\( h_v \) - the height of a peak function at an elution volume of \( v \) (mV).

\( h_{ni} \) - the height of the \( i^{\text{th}} \) peak function (mV).

\( h_{ni}^* \) - the height of the \( i^{\text{th}} \) peak function observed on the chromatogram (mV).
$J(x)$ - Jacobian matrix of first derivatives (as appropriate)

$l$ - path length of detector flow cell (m)

$m$ - the number of components in the mixture analysed by the statistical theory of overlap (dimensionless)

$n$ - the number of data points used to describe a chromatogram (dimensionless)

$n_c$ - the capacity of the column analysed by the statistical theory of overlap (number of peaks)

$p$ - the number of peaks predicted to be observed on a chromatogram by the statistical theory of overlap (dimensionless)

$r_i$ - random number between zero and one (dimensionless)

$R(x)$ - matrix of objective function coefficients (as appropriate)

$S(x)$ - Hessian matrix of second derivatives (as appropriate)

$V_{i\delta}$ - the centre of a Gaussian peak function (mL)

$V_{oi}$ - the start of the $i^{th}$ general exponential function (mL)

$V_{ai}$ - the position of the $i^{th}$ general exponential function maximum (mL)

$V_{ai}^{*}$ - the observed position of the $i^{th}$ general exponential function maximum (mL)

$V_{end}$ - the end of the chromatogram (ie. the last recorded data point) (mL)

$x_i$ - peak function parameter (as appropriate)

$x_{l}$ - first estimate of parameter $x$ (as appropriate)
\( x_2 \) - second estimate of parameter \( x \) (as appropriate)

\( \bar{x} \) - the mean height of a chromatogram (mV)

\( \Delta \bar{x} \) - the mean difference in height between the actual and model chromatograms (mV)

\( \Delta \bar{x}/\bar{x} \) - the ratio of mean difference in height to mean height of chromatograms (dimensionless)

\( \Delta Y \) - the spacing between data points (mL)

\( \varepsilon_i \) - the extinction coefficient of the \( i^{th} \) component in the flow cell (AU mL mg\(^{-1}\) cm\(^{-1}\))

\( \lambda_i \mu_i \) - damping parameters for Gauss-Newton and Levenberg-Marquardt optimisation methods (dimensionless)

\( \sigma \) - standard deviation of a Gaussian peak function (mL)

\( r \) - the modifier (a decay constant) used to the pure Gaussian function in the exponentially modified Gaussian function (mL).

Chapter Three: Fuzzy Logic Identification

\( a \) - the spread parameter of the peak area membership function (dimensionless)

\( \text{area}_a \) - the centre of the peak area peak membership function (dimensionless)

\( m_A(x) \) - fuzzy number A of measurement parameter \( x \) (dimensionless)

\( m_B(x) \) - fuzzy number B of measurement parameter \( x \) (dimensionless)

\( m_i \) - elution order match criterion (dimensionless)

\( m_s \) - area membership function (dimensionless)
$r_i$ - the rank of the $i$th peak (dimensionless)

#$REF$ - the number of peaks in the reference chromatogram (dimensionless)

#$TRIAL$ - the number of peaks in the trial chromatogram (dimensionless)
Appendix A5.

References.
Appendix A5: References


