NEAR INFRARED SPECTROSCOPY
TECHNIQUE FOR BIOPROCESS
MONITORING AND CONTROL

Ken S. Y. Yeung

September 1997

Thesis submitted for the degree of Doctor of Philosophy
from the University of London

Centre for Advanced Instrumentation System
Mathematical and Physical Sciences
University College London
Gower Street
London
W1CE 6BT
U.K.
For my parents

and

Zoë
Abstract

The thesis described an application of near infrared (NIR) spectroscopy to a specific characterisation of a complex biological process stream. Its use is for the monitoring and control of a downstream process in which alcohol dehydrogenase (ADH) is to be recovered from a broth of homogenised yeast. Other than the product ADH, the yeast homogenate consists of contaminants such as cell debris, protein, nucleic acids, lipids and others. While the first two represented a large fraction of the total contaminant and could be removed sufficiently by centrifugation and fractional precipitation, respectively. The remains residuals of contaminants often poses a fouling problem to the final stage of hydrophobic interaction chromatography (HIC) purification. The fouling has been minimised by the inserting a flocculation process at the early stage of recovery so that the problem contaminants are removed and the overall recovery is enhanced.

A low budget instrument has been developed for sensitivity in the region of the NIR spectrum (from 1900 to 2500 nm) where preliminary work found distinctive NIR signatures from cell debris, protein and RNA in the yeast homogenate. The instrument is configured as a scanning spectrophotometer. Multivariate calibration technique with partial least squares (PLS) has been used for the three mentioned contaminants calibration.

Two types of samples are used for calibrating the NIR instrument. In one case samples are prepared by adding materials representative of the contaminants to clarified yeast homogenate. In the other samples are taken from the process stream after flocculation and floc removal. In the former case the contaminant levels are well known but are outside the range of interest; in the latter there is uncertainty of analysis of contaminant level but the calibration is in the range of interest. The NIR instrument together with the calibration models has demonstrated potential in rapid monitoring of contaminants.
Acknowledgements

I would like to thank my supervisors most sincerely, Mrs Nina Thornhill and Prof. Mike Hoare for their continue advise, encouragement, guidance and stimulation, without which I would not have been able to complete the project. I would also like to thanks my other supervisor Mr. Tom Williams for his help and advise on the designing of a Near Infrared spectrophotometer for this project.

Thanks also to Prof. David Miller and Dr. Tony Alunut for their advises. Many thanks to all those within the Department of Chemical and Biochemical Engineering, the Electrical and Electronics Engineering Department for their help throughout my time at UCL. I am particularly grateful to Mr Jeetendra Vaghjiani, Dr. Mark Bulmer, Dr. Claire Turner, Dr. Malcolm Gregory, Mr George Habib, Mr. Cyrille Alheritiere who give me a lot of insight to biochemical engineering.

I would also like to acknowledge the support of the Engineering and Physical Sciences Research Council, Department of Trade Industry and Sira Limited for this project.
Table of Contents

ABSTRACT 3
ACKNOWLEDGEMENTS 4
TABLE OF CONTENTS 5
LIST OF TABLES 13
LIST OF FIGURES 14
ABBREVIATIONS 20
NOMENCLATURE 22

1. INTRODUCTION 25

2. NEAR INFRARED (NIR) TECHNIQUES 29
2.1 Summary 29
2.2 Introduction 30
2.3 NIR in alcohol dehydrogenase (ADH) recovery process 31
   2.3.1 Bakers yeast 31
   2.3.2 Recovery process by fractional precipitation 33
   2.3.3 Enhanced recovery process with selective flocculation 35
   2.3.4 Role of NIR monitoring and control to optimise flocculation 36
2.4 Applications of near infrared spectroscopy 38
   2.4.1 Near infrared analysis in pharmaceuticals 39
   2.4.2 At-line and on-line NIR monitoring of bioreactors 40
2.5 Theory of near infrared spectroscopic technique 42
   2.5.1 Near infrared and infrared absorption 42
      2.5.1.1 The harmonic oscillation of molecule vibration characterised by quantum mechanic model 43
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Tables of Contents

2.5.1.2 The overtone bands due to anharmonic oscillation 48
2.5.2 Physics of the interaction of radiation with matter 51
  2.5.2.1 Transmission of radiation 52
  2.5.2.2 Absorption measurements 53
  2.5.2.3 Absorbance and analyte concentration 54
  2.5.2.4 Analyte concentration determined from overlapping absorption peaks 57
2.6 NIR instruments 59
  2.6.1 Optical principles in NIR instruments 60
  2.6.2 Dispersive systems 61
    2.6.2.1 Monochromator and spectrograph instruments 61
    2.6.2.2 Acousto-optically tuneable filters 63
  2.6.3 Non-dispersive 65
  2.6.4 Non-thermal source 68
  2.6.5 Sample presentation 69
2.7 Conclusion 70
  2.7.1 NIR monitoring system to enhance flocculation control and hence optimise early stage of ADH recovery process 70
  2.7.2 NIR monitoring in some of the relevant industries 72
  2.7.3 The theory of NIR spectroscopic techniques 73
  2.7.4 NIR instruments 74

3. NEAR INFRARED CALIBRATION 76
3.1 Summary 76
3.2 Introduction to the calibration experiment 77
3.3 Practical needs for multivariate calibration 78
  3.3.1 Chemical interference in samples 81
  3.3.2 Physical interference in samples 81
  3.3.3 Mistakes and unexpected errors 82
3.4 Partial least squares (PLS) calibration based on principal component analysis (PCA) 83
  3.4.1 Multiple linear regression (MLR) and its problems 84
  3.4.2 Data reduction with principal components analysis (PCA) 85
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.2.1 Matrix representations and projection</td>
<td>86</td>
</tr>
<tr>
<td>3.4.2.2 Factors and principal component analysis (PCA)</td>
<td>87</td>
</tr>
<tr>
<td>3.4.3 Principal component regression</td>
<td>91</td>
</tr>
<tr>
<td>3.4.4 Partial least squares</td>
<td>93</td>
</tr>
<tr>
<td>3.5 Prediction ability, validation and the number of principal components used in calibration model</td>
<td>95</td>
</tr>
<tr>
<td>3.5.1 Prediction ability</td>
<td>95</td>
</tr>
<tr>
<td>3.5.1.1 Prediction errors</td>
<td>96</td>
</tr>
<tr>
<td>3.5.1.2 Confidence Intervals</td>
<td>97</td>
</tr>
<tr>
<td>3.5.2 Validation of calibration model in practice</td>
<td>98</td>
</tr>
<tr>
<td>3.5.3 Using prediction error and selection of the number principal components in a calibration model</td>
<td>99</td>
</tr>
<tr>
<td>3.6 Conclusion</td>
<td>102</td>
</tr>
<tr>
<td>3.6.1 Needs of multivariate calibration</td>
<td>102</td>
</tr>
<tr>
<td>3.6.2 Principal components analysis and partial least squares solutions</td>
<td>103</td>
</tr>
<tr>
<td>3.6.3 Prediction ability, validation method and choosing calibration model</td>
<td>105</td>
</tr>
<tr>
<td>4. A BUDGET NIR SPECTROPHOTOMETER FOR BIOPROCESS</td>
<td>107</td>
</tr>
<tr>
<td>4.1 Summary</td>
<td>107</td>
</tr>
<tr>
<td>4.2 Introduction</td>
<td>108</td>
</tr>
<tr>
<td>4.3 Spectrophotometer</td>
<td>109</td>
</tr>
<tr>
<td>4.3.1 Design layout</td>
<td>109</td>
</tr>
<tr>
<td>4.4 Diffraction gratings</td>
<td>112</td>
</tr>
<tr>
<td>4.4.1 Linear dispersion in spectrographs</td>
<td>113</td>
</tr>
<tr>
<td>4.4.2 Grating order and order sorting filter</td>
<td>114</td>
</tr>
<tr>
<td>4.4.3 Blazed gratings and efficiency</td>
<td>116</td>
</tr>
<tr>
<td>4.4.4 Diffraction grating stray light</td>
<td>118</td>
</tr>
<tr>
<td>4.4.5 Choice of holographic grating</td>
<td>118</td>
</tr>
<tr>
<td>4.5 PbS detector system</td>
<td>120</td>
</tr>
<tr>
<td>4.5.1 PbS detectors and figures of merit</td>
<td>122</td>
</tr>
<tr>
<td>4.5.2 PbS detectors bias supply voltage</td>
<td>124</td>
</tr>
</tbody>
</table>
4.5.3 Temperature 126
4.5.4 Noise and noise compensations 127
  4.5.4.1 Modulation techniques 128
  4.5.4.2 Gated Integrator 129
4.5.5 Analogue to digital conversions 130
4.6 Acquiring the NIR spectrum 131
  4.6.1 Acquiring the NIR spectrum from the spectrographs output 131
  4.6.2 Spectral smoothing 132
    4.6.2.1 Moving point averages 133
    4.6.2.2 GCV/GA spline 133
    4.6.2.3 Comparison of moving average, interpolating cubic spline and GCV/GA spline in spectral smoothing 134
4.7 Specification performance of the spectrophotometer and practical aspects 135
  4.7.1 Wavelength calibration 136
  4.7.2 Approach to referencing and baseline correction 137
  4.7.3 Reproducibility of the spectrophotometer 141
    4.7.3.1 Wavelength reproducibility 141
    4.7.3.2 Measurements reproducibility 142
  4.7.4 Practical aspects of using the spectrophotometer 143
    4.7.4.1 Electrical noise 143
    4.7.4.2 Vibration 145
4.8 Calibrating the spectrophotometer for determination of protein and RNA in yeast homogenate using PLS method 145
  4.8.1 Materials 146
  4.8.2 Methods 146
    4.8.2.1 Clarified yeast homogenate 146
    4.8.2.2 Protein calibration and validation samples 147
    4.8.2.3 RNA calibration and validation samples 147
    4.8.2.4 NIR spectrum 148
    4.8.2.5 Partial least squares and cross validation modelling 148
  4.8.3 Calibration and validation of protein PLS model 148
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Tables of Contents

4.8.3.1 NIR spectra of protein in clarified yeast homogenate 148
4.8.3.2 PLS modelling of protein in clarified yeast homogenate calibration samples 149
4.8.3.3 Validation of PLS protein calibration model using fresh and aged samples 150
4.8.3.4 Identification of the effect of spectral differences between fresh and aged samples 152

4.8.4 Calibration and validation of RNA PLS model 153
4.8.4.1 NIR spectra of RNA in clarified yeast homogenate 153
4.8.4.2 PLS modelling of RNA in clarified yeast homogenate calibration samples 154
4.8.4.3 Validation of PLS RNA calibration model 155

4.9 Conclusion 157
4.9.1 Design and optimisation of a budget NIR spectrophotometer 157
4.9.2 Performance of the in-house budget spectrophotometer 159
4.9.3 Using the spectrophotometer in conjunction with PLS for quantitative analysis of bioprocess contaminants 161

5. CALIBRATION AND PREDICTION OF YEAST HOMOGENATE SAMPLES 162

5.1 Summary 162
5.2 Introduction 163
5.3 Calibration strategy 164
5.3.1 Ideal contaminants calibration 165
5.3.2 Add-Back contaminants calibration 166
5.3.3 Process stream samples calibration & validation 168

5.4 Materials and methods 169
5.4.1 Standard materials 169
5.4.2 Clarified yeast homogenate and washed debris 169
5.4.3 Add-back contaminant samples 170
5.4.4 Process stream samples 170
5.4.5 Analytical Data 171
5.4.6 Assays for biological contaminants 171
5.4.7 NIR Spectroscopy & Multivariate analysis 172
5.5 Results and discussion 173
5.5.1 Measurements and measurements variability 173
  5.5.1.1 Optical density measurements with *add-back* calibration samples 173
  5.5.1.2 Protein assays on *add-back* calibration samples 175
  5.5.1.3 RNA assays on *add-back* calibration samples 176
  5.5.1.4 Optical density of two set of process stream calibration samples 178
  5.5.1.5 Protein assays of two sets of process stream calibration samples 179
  5.5.1.6 RNA assays of two sets of process stream calibration samples 180
5.5.2 PLS calibration models using *add-back* calibration samples 182
  5.5.2.1 NIR spectra of the *add-back* calibration samples 182
  5.5.2.2 Calibration of cell debris (optical density) using *add-back* calibration samples 183
  5.5.2.3 Calibration of protein using *add-back* calibration samples 184
  5.5.2.4 Calibration of RNA using *add-back* calibration samples 186
5.5.3 PLS calibration models using *process stream* samples 188
  5.5.3.1 NIR spectra of two sets of *process stream* calibration samples 188
  5.5.3.2 Calibration of cell debris (optical density) by process stream samples 190
  5.5.3.3 Calibration of protein using *process stream* calibration samples 191
  5.5.3.4 Calibration of RNA using *process stream* calibration samples 192
5.5.4 Process monitoring using both *add-back* and *process stream* calibrated PLS models 194
  5.5.4.1 Prediction of cell debris (optical density) contaminants in flocculation processes 194
  5.5.4.2 Prediction of protein contaminants in flocculation processes 196
  5.5.4.3 Prediction of RNA contaminants in flocculation processes 197
5.5.5 Reasons for poor prediction by the *add-back* calibrated models in process monitoring 199
  5.5.5.1 Interpretation of PLS models 199
  5.5.5.2 Effect of particle size on the NIR spectra 201
5.5.6 Establishing control set point 203
### 5.6 Conclusion

- 5.6.1 Application and calibration methods 204
- 5.6.2 Measurements and measurement variability 206
- 5.6.3 PLS calibrations using *add-back* contaminant samples 207
- 5.6.4 PLS calibrations using *process stream* samples 209
- 5.6.5 Process monitoring and control set point 210
- 5.6.6 Differences between *add-back* and *process stream* samples 211

### 6. Conclusion and Recommendation 213

- 6.1 Conclusion 213
- 6.2 Recommendation 215
  - 6.2.1 NIR process monitoring within the ADH recovery process 215
  - 6.2.2 NIR Instruments for future ADH recovery processes 216
  - 6.2.3 Calibration experiments 221
- 6.3 Final Review 222

### A 1. The Fabry-Perot Interferometer 224

### A 2. Multiple Linear Regression in Spectroscopy Applications 228

- A 2.1 Introduction 228
- A 2.2 Problem of MLR with real spectroscopic data 229
- A 2.3 Summary 231

### A 3. Computing Algorithm for Principal Component Analysis (PCA) and Partial Least Squares (PLS) 232

- A 3.1 The NIPALS algorithm for PCA 232
- A 3.2 Orthogonalised PLS algorithm for one Y-variable 234
  - A 3.2.1 PLS calibration 234
  - A 3.2.2 Full prediction 237
  - A 3.2.3 Short prediction 237
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 4. Comparison of the in-house built spectrophotometer with two commercial spectrophotometers</td>
<td>238</td>
</tr>
<tr>
<td>A 5. Identification of NIR response to biological contaminants using a commercial spectrophotometer</td>
<td>240</td>
</tr>
<tr>
<td>A 5.1 Introduction</td>
<td>240</td>
</tr>
<tr>
<td>A 5.2 Material and methods</td>
<td>241</td>
</tr>
<tr>
<td>A 5.3 Results</td>
<td>242</td>
</tr>
<tr>
<td>A 5.3.1 NIR signature of debris in clarified yeast homogenate</td>
<td>242</td>
</tr>
<tr>
<td>A 5.3.2 NIR signature of protein in clarified yeast homogenate</td>
<td>243</td>
</tr>
<tr>
<td>A 5.3.3 NIR signature of RNA in clarified yeast homogenate</td>
<td>244</td>
</tr>
<tr>
<td>A 5.4 Summary</td>
<td>245</td>
</tr>
<tr>
<td>A 6. Concentration of contaminants in single contaminant calibration samples</td>
<td>246</td>
</tr>
<tr>
<td>A 7. Composition of the add-back calibration samples</td>
<td>248</td>
</tr>
<tr>
<td>A 8. Optical density and assay results</td>
<td>250</td>
</tr>
<tr>
<td>A 8.1 Optical density at 650 nm of all calibration and validation samples</td>
<td>250</td>
</tr>
<tr>
<td>A 8.2 Protein assays of all calibration and validation samples</td>
<td>252</td>
</tr>
<tr>
<td>A 8.3 RNA assays of all calibration and validation samples</td>
<td>253</td>
</tr>
<tr>
<td>A 9. The affect of polyethyleneimine (PEI) in yeast homogenate</td>
<td>254</td>
</tr>
<tr>
<td>A 9.1 Introduction</td>
<td>254</td>
</tr>
<tr>
<td>A 9.2 Materials and Methods</td>
<td>254</td>
</tr>
<tr>
<td>A 9.2.1 Materials</td>
<td>254</td>
</tr>
<tr>
<td>A 9.2.2 Methods</td>
<td>255</td>
</tr>
<tr>
<td>A 9.3 Results and discussions</td>
<td>255</td>
</tr>
<tr>
<td>A 9.4 Summary</td>
<td>256</td>
</tr>
<tr>
<td>References</td>
<td>257</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>General statistics of contaminants after yeast disruption.</td>
<td>32</td>
</tr>
<tr>
<td>2.2</td>
<td>Regions of near, mid and far infrared spectrum.</td>
<td>51</td>
</tr>
<tr>
<td>4.1</td>
<td>Suggestion on applications for holographic or ruled grating.</td>
<td>119</td>
</tr>
<tr>
<td>5.1</td>
<td>Statistics of the PLS calibrations in add-back experiments.</td>
<td>207</td>
</tr>
<tr>
<td>5.2</td>
<td>Statistics of the calibrations in process stream experiments.</td>
<td>209</td>
</tr>
<tr>
<td>5.3</td>
<td>Statistics on the prediction of flocculation process given by add-back and process stream calibrated PLS models.</td>
<td>210</td>
</tr>
<tr>
<td>A 4.1</td>
<td>Comparisons of the in-house built spectrophotometer with two commercial spectrophotometers.</td>
<td>238</td>
</tr>
<tr>
<td>A 6.1</td>
<td>Protein stock and final known concentration of protein in calibration samples.</td>
<td>246</td>
</tr>
<tr>
<td>A 6.2</td>
<td>RNA stock and final known concentration of RNA in calibration samples.</td>
<td>247</td>
</tr>
<tr>
<td>A 7.1</td>
<td>Composition of the add-back calibration samples.</td>
<td>248</td>
</tr>
<tr>
<td>A 7.2</td>
<td>The expected protein and RNA concentrations in the add-back calibration samples.</td>
<td>249</td>
</tr>
<tr>
<td>A 8.1</td>
<td>Optical density at 650 nm of all add-back calibration samples.</td>
<td>250</td>
</tr>
<tr>
<td>A 8.2</td>
<td>Optical density at 650 nm of two sets of process stream calibration and two sets of process stream validation samples.</td>
<td>251</td>
</tr>
<tr>
<td>A 8.3</td>
<td>Protein assays of all add-back calibration samples.</td>
<td>252</td>
</tr>
<tr>
<td>A 8.4</td>
<td>Protein assays of two sets of process stream calibration samples and two sets of process stream validation samples.</td>
<td>252</td>
</tr>
<tr>
<td>A 8.5</td>
<td>RNA assays of all add-back calibration samples.</td>
<td>253</td>
</tr>
<tr>
<td>A 8.6</td>
<td>RNA assays of two sets of process stream calibration samples and two sets of process stream validation samples.</td>
<td>253</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Illustration of the bakers yeast composition.</td>
<td>32</td>
</tr>
<tr>
<td>2.2</td>
<td>Flow diagram of a traditional process for intracellular protein (alcohol dehydrogenase) recovery.</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>Flow diagram of the alcohol dehydrogenase recovery process; including the flocculation stage and the configuration of NIR monitoring.</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>Illustration of the energy of a diatomic molecule undergoing simple harmonic motion and anharmonic vibration.</td>
<td>45</td>
</tr>
<tr>
<td>2.5</td>
<td>Illustration of the quantum theory of absorption and radiation.</td>
<td>46</td>
</tr>
<tr>
<td>2.6</td>
<td>Interaction of radiation with matter.</td>
<td>51</td>
</tr>
<tr>
<td>2.7</td>
<td>Graphical statement of the Beer Lambert law showing the effect of deviations due to polychromatic radiation, chemical interaction and scatter, reflection.</td>
<td>54</td>
</tr>
<tr>
<td>2.8</td>
<td>A plot of tangent and baseline methods for background correction.</td>
<td>57</td>
</tr>
<tr>
<td>2.9</td>
<td>Illustration of simultaneous determination of two components in a mixture.</td>
<td>59</td>
</tr>
<tr>
<td>2.10</td>
<td>The three categories of optical methods used in NIR instruments.</td>
<td>60</td>
</tr>
<tr>
<td>2.11</td>
<td>Configuration of the monochromator used for absorption measurements.</td>
<td>61</td>
</tr>
<tr>
<td>2.12</td>
<td>The diagram illustrates the section of diffraction grating.</td>
<td>62</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.13</td>
<td>An illustration of the principle operation of acousto optical tuneable filters.</td>
<td>64</td>
</tr>
<tr>
<td>2.14</td>
<td>Schematic diagram of Michelson interferometer and interferogram using a moving mirror.</td>
<td>66</td>
</tr>
<tr>
<td>3.1</td>
<td>Illustration of calibration modelling.</td>
<td>77</td>
</tr>
<tr>
<td>3.2</td>
<td>Illustration of prediction using calibrated model.</td>
<td>77</td>
</tr>
<tr>
<td>3.3</td>
<td>A plot of the ideal calibration arrangement.</td>
<td>78</td>
</tr>
<tr>
<td>3.4</td>
<td>The diagram illustrates non-selective instrumental measurement for the analyte and non-linear instrument response.</td>
<td>79</td>
</tr>
<tr>
<td>3.5</td>
<td>Illustration of the traditional selectivity enhancement prior to measurements.</td>
<td>79</td>
</tr>
<tr>
<td>3.6</td>
<td>Illustration of the selectivity enhancement by multivariate calibration.</td>
<td>80</td>
</tr>
<tr>
<td>3.7</td>
<td>Row and column space matrices.</td>
<td>86</td>
</tr>
<tr>
<td>3.8</td>
<td>Projection of y down on the space spanned by the vectors $x_1$ and $x_2$.</td>
<td>87</td>
</tr>
<tr>
<td>3.9</td>
<td>A plot of the matrix $X$ in column space with illustration of the first eigenvector.</td>
<td>88</td>
</tr>
<tr>
<td>3.10</td>
<td>Illustration of the first and second eigenvector of $X'X$.</td>
<td>89</td>
</tr>
<tr>
<td>3.11</td>
<td>Spectral data of multiple samples presented in a matrix format.</td>
<td>91</td>
</tr>
<tr>
<td>3.12</td>
<td>Conceptual illustration of prediction error as function of number of principal components included in the calibration model.</td>
<td>100</td>
</tr>
<tr>
<td>4.1</td>
<td>The design layout of the in-house NIR spectrophotometer.</td>
<td>110</td>
</tr>
<tr>
<td>4.2</td>
<td>Mirror configuration for reflectance measurements.</td>
<td>111</td>
</tr>
<tr>
<td>4.3</td>
<td>Diagram of the lead sulfide PbS detector mounted on a stage.</td>
<td>112</td>
</tr>
<tr>
<td>4.4</td>
<td>Illustration of light diffracted from a grating and higher order responses.</td>
<td>115</td>
</tr>
<tr>
<td>4.5</td>
<td>An illustration of litrow condition for a single groove of a blazed grating.</td>
<td>116</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>Efficiency profile of a blazed ruled grating and a non-blazed holographic grating.</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>Schematic diagram of the complete PbS detection system.</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>A plot of the normalised detectivity ($D^*$) of various detectors.</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>Diagram of a typical lead sulfide detector packaging.</td>
<td></td>
</tr>
<tr>
<td>4.10</td>
<td>Schematic of a photoconductive detector and simple drive circuit.</td>
<td></td>
</tr>
<tr>
<td>4.11</td>
<td>Manufacture's recommendation on maximum detector voltage vs. detector area and resistance for detector temperature of 298°K.</td>
<td></td>
</tr>
<tr>
<td>4.12</td>
<td>Typical spectral detectivity at temperature and noise vs. frequency.</td>
<td></td>
</tr>
<tr>
<td>4.13</td>
<td>Manufacture's data on PbS normalised detectivity and optimum operating condition.</td>
<td></td>
</tr>
<tr>
<td>4.14</td>
<td>An illustration of the repetitive signal and detection &quot;windows&quot; to reduce noise detection.</td>
<td></td>
</tr>
<tr>
<td>4.15</td>
<td>Illustration of the discrete scanning of a water transmittance spectrum.</td>
<td></td>
</tr>
<tr>
<td>4.16</td>
<td>Comparison of three point moving average, interpolating cubic spline, GCV/GA spline spectral smoothing.</td>
<td></td>
</tr>
<tr>
<td>4.17</td>
<td>The spectral transmittance response of water from a commercial spectrophotometer and the in-house built spectrophotometer.</td>
<td></td>
</tr>
<tr>
<td>4.18</td>
<td>A plot of improved measurements in low NIR energy by the in-house spectrophotometer.</td>
<td></td>
</tr>
<tr>
<td>4.19</td>
<td>A plot of spectral reproducibility within and across three days in the in-house NIR spectrophotometer.</td>
<td></td>
</tr>
<tr>
<td>4.20</td>
<td>Illustration of the transmission variation caused by chopper frequency and the affect of excessive electrical noise in the in-house NIR spectrophotometer.</td>
<td></td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>4.21</td>
<td>NIR spectra of twelve protein in yeast homogenate calibration samples.</td>
<td></td>
</tr>
<tr>
<td>4.22</td>
<td>Loadings ($p_1$) plot of the decomposition of the X-data matrix.</td>
<td></td>
</tr>
<tr>
<td>4.23</td>
<td>A plot of the predicted protein against the known protein concentration in a PLS calibration of single contaminants in yeast homogenate.</td>
<td></td>
</tr>
<tr>
<td>4.24</td>
<td>A spectral differences plot of fresh and 24 hours protein calibration samples.</td>
<td></td>
</tr>
<tr>
<td>4.25</td>
<td>The NIR transmission spectra of eight RNA calibration samples.</td>
<td></td>
</tr>
<tr>
<td>4.26</td>
<td>The loadings plot of the 3 principal components used in the PLS calibration model for RNA.</td>
<td></td>
</tr>
<tr>
<td>4.27</td>
<td>A plot of the predicted RNA against known RNA concentration in a PLS calibration of single contaminants in yeast homogenate.</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Illustration of the ideal calibration and process monitoring contaminants concentration region.</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>A plot of the OD measured in all add-back calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>A plot of protein concentration in all add-back calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>A plot of RNA concentration in all add-back calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>A plot of OD measured in process stream calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>A plot of protein concentration in process stream calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>A plot of RNA concentration in process stream calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>The transmission spectra of all 36 add-back calibration samples.</td>
<td></td>
</tr>
<tr>
<td>5.9</td>
<td>A plot of the predicted OD against the measured OD in add-back calibration samples.</td>
<td></td>
</tr>
</tbody>
</table>
5.10 A plot of the predicted protein against the expected protein in *add-back* calibration samples 185
5.11 A plot of the predicted RNA against the expected RNA in *add-back* calibration samples. 187
5.12 NIR transmission spectra of *process stream* calibration samples. 189
5.13 A plot of the predicted OD against the measured OD in *process stream* calibration samples. 190
5.14 A plot of the predicted protein against the measured protein in *process stream* calibration samples. 192
5.15 A plot of the predicted RNA against the expected RNA in *process stream* calibration samples. 193
5.16 Process monitoring of optical density by measurements, *add-back* and *process stream* calibrated PLS models. 195
5.17 Process monitoring of protein by measurements, *add-back* and *process stream* calibrated PLS models. 197
5.18 Process monitoring of RNA by measurements, *add-back* and *process stream* calibrated PLS models. 198
5.19 A loading $p(\lambda)$ plot of both optical density *add-back* and *process stream* PLS model and the mean spectra of both sets of calibration samples. 201
5.20 Illustration of the PEI feed control set point justified by predicted contaminants concentration. 203
6.1 An illustration of the transflectance (transmittance) and reflectance NIR measurements with optical fibre measuring probe. 217
6.2 NIR reflectance measurement with optical fibre probe. 218
6.3 Illustration of two methods of NIR measurements to three process stream locations. 219
**A 1.1** A diagram illustrates the Fabry-Perot (FP) interferometer uses fringes generated in transmitted light following multiple reflections in the gap between two parallel plates.

**A 1.2** Illustration of transmission of Fabry-Perot for three reflectivities.

**A 1.3** An illustration of piezo-electrically scanned Fabry-Perot.

**A 5.1** A plot of the absorption and 2nd derivative spectra of clarified yeast homogenate and debris samples recorded by a commercial spectrophotometer.

**A 5.2** A plot of the absorption and 2nd derivative spectra of clarified yeast homogenate and protein samples recorded by a commercial spectrophotometer.

**A 5.3** A plot of the absorption and 2nd derivative spectra of clarified yeast homogenate and RNA samples recorded by a commercial spectrophotometer.

**A 9.1** Normalised transmission spectra of distill water, 2%w/v PEI in distilled water and various concentration of PEI in yeast homogenate.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC</td>
<td>analogue to digital converter</td>
</tr>
<tr>
<td>ADH</td>
<td>alcohol dehydrogenase</td>
</tr>
<tr>
<td>AOTF</td>
<td>acousto-optical tuneable filters</td>
</tr>
<tr>
<td>BLIP</td>
<td>background limited in performance</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CYH</td>
<td>clarified yeast homogenate</td>
</tr>
<tr>
<td>FACSS</td>
<td>Federation of Analytical Chemistry and Spectroscopy Societies</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>FFT</td>
<td>fast Fourier transform</td>
</tr>
<tr>
<td>FP</td>
<td>Fabry-Perot</td>
</tr>
<tr>
<td>FT-IR</td>
<td>Fourier transform infrared</td>
</tr>
<tr>
<td>FT-NIR</td>
<td>Fourier transform near infrared</td>
</tr>
<tr>
<td>GA</td>
<td>genetic algorithm</td>
</tr>
<tr>
<td>GCV</td>
<td>generalised cross validation</td>
</tr>
<tr>
<td>HIC</td>
<td>hydrophobic interaction chromatography</td>
</tr>
<tr>
<td>HPLC</td>
<td>high performance liquid chromatography</td>
</tr>
<tr>
<td>InGaAS</td>
<td>indium gallium arsenide</td>
</tr>
<tr>
<td>IR</td>
<td>far infrared</td>
</tr>
<tr>
<td>LED</td>
<td>light emitting diodes</td>
</tr>
<tr>
<td>MIR</td>
<td>middle infrared</td>
</tr>
<tr>
<td>MLR</td>
<td>multiple linear regression</td>
</tr>
<tr>
<td>MSC</td>
<td>multiplicative scatter correction</td>
</tr>
<tr>
<td>MSE</td>
<td>mean square error</td>
</tr>
<tr>
<td>NEP</td>
<td>noise equivalent power</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>NIPALS</td>
<td>non-linear iterative partial least squares</td>
</tr>
<tr>
<td>NIR</td>
<td>near infrared</td>
</tr>
<tr>
<td>OD</td>
<td>optical density</td>
</tr>
<tr>
<td>PbS</td>
<td>lead sulfide</td>
</tr>
<tr>
<td>PC</td>
<td>principal component</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>principal component regression</td>
</tr>
<tr>
<td>PEI</td>
<td>polyethylenemine</td>
</tr>
<tr>
<td>PLS</td>
<td>partial least squares</td>
</tr>
<tr>
<td>QH</td>
<td>quartz halogen</td>
</tr>
<tr>
<td>rms</td>
<td>root mean square</td>
</tr>
<tr>
<td>RMSEP</td>
<td>root mean square prediction</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>SEP</td>
<td>standard error of prediction</td>
</tr>
<tr>
<td>SNV</td>
<td>standard normal variate</td>
</tr>
<tr>
<td>TE</td>
<td>thermal electric</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>YH</td>
<td>yeast homogenate</td>
</tr>
</tbody>
</table>
Nomenclature

a  acceleration of the mass from equilibrium 
\(a\)  the absorptivity 
\(\alpha\)  a constant for a particular molecule 
A  absorbance 
\(A'\)  observed absorbance 
B  a matrix of regression coefficients 
b  thickness through the sample 
c  concentration of molecules in sample 
c  velocity of light in vacuo 
\(C_n\)  the correction coefficient 
D  angle of diffraction normal to the grating 
D*  normalised detectivity 
\(\Delta f\)  the electrical bandwidth in semiconductors detector 
E  a matrix of errors associated in the model of X matrix 
\(E\)  energy 
\(E_2-E_1\)  energy at level 2 and 1 respectively 
\(E_D\)  the dissociation energy 
\(\Delta E\)  transition energy between \(E_2\) and \(E_1\) 
\(\varepsilon_{ij}\)  an element in the E matrix 
F  a matrix of errors associated in the model of Y matrix 
f  chopping frequency 
f  focal length 
\(F\)  force (Hook’s law)
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control  

Nomenclature

**G**  
a matrix of errors associated in the model of X matrix in PLS model

**I**  
angle of the light ray incident on adjacent grooves at I to the grating normal

**I_{pr}**  
the total number of predicting samples

**i^{th}**  
the row number (samples) number in the X or Y data matrix

**j^{th}**  
the column number; spectral variables (wavelength) in X matrix or analyte concentration in the Y matrix

**k**  
the force constant or Boltzman constant

**λ**  
wavelength of the travelling wave

**λ_B**  
blaze wavelength

**m**  
mass

**m**  
order of diffraction

**N**  
any odd number used in moving point averages

**N**  
number of energy pulses

**n**  
refractive index

**n_1**  
the number of molecules in the excited state

**n_2**  
the number of molecules in the ground state

**h**  
Plank’s constant

**P**  
loading matrix for X matrix

**p**  
loading vector for X matrix

**P'(PP')^{-1}**  
the generalised inverse (GI) of P matrix

**P_A**  
radiant power absorbed in the sample

**P_O**  
total radiation power incident on the sample

**P_R**  
radiant power reflected by the sample

**P_S**  
the power of stray radiation

**P_T**  
radiant power transmitted out of the sample

**Q**  
loading matrix for Y matrix

**q**  
loading vector for Y matrix

**R_L**  
load resistance in the detector circuit
**Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control**

### Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rm&lt;sub&gt;n&lt;/sub&gt;</td>
<td>the master reference spectrum of distilled water</td>
</tr>
<tr>
<td>R&lt;sub&gt;n&lt;/sub&gt;s</td>
<td>the reference spectrum of distilled water before each measuring sample</td>
</tr>
<tr>
<td>s</td>
<td>distance of the object</td>
</tr>
<tr>
<td>s'</td>
<td>distance of the image</td>
</tr>
<tr>
<td>S&lt;sub&gt;n&lt;/sub&gt;'</td>
<td>baseline corrected spectrum</td>
</tr>
<tr>
<td>S&lt;sub&gt;n&lt;/sub&gt;</td>
<td>the spectrum of the measuring sample</td>
</tr>
<tr>
<td>δ</td>
<td>retardation distance</td>
</tr>
<tr>
<td>T</td>
<td>absolute temperature</td>
</tr>
<tr>
<td>T</td>
<td>score matrix for X matrix</td>
</tr>
<tr>
<td>t</td>
<td>score vector for X matrix</td>
</tr>
<tr>
<td>t</td>
<td>displacements time</td>
</tr>
<tr>
<td>T</td>
<td>transmittance</td>
</tr>
<tr>
<td>U</td>
<td>score matrix for Y matrix</td>
</tr>
<tr>
<td>u</td>
<td>score vector for Y matrix</td>
</tr>
<tr>
<td>ν</td>
<td>the frequency of a harmonic wave</td>
</tr>
<tr>
<td>ω</td>
<td>blaze angle in the grating</td>
</tr>
<tr>
<td>X</td>
<td>a data matrix of the sample spectra</td>
</tr>
<tr>
<td>X'</td>
<td>the transpose of X matrix</td>
</tr>
<tr>
<td>X&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>the inverse of X matrix</td>
</tr>
<tr>
<td>X'X</td>
<td>the covariance matrix of X</td>
</tr>
<tr>
<td>Y</td>
<td>a data matrix of analytes concentration in the samples</td>
</tr>
<tr>
<td>ŷ</td>
<td>the predicted concentration of analyte</td>
</tr>
<tr>
<td>y&lt;sub&gt;ij&lt;/sub&gt;</td>
<td>the concentration data in i&lt;sup&gt;th&lt;/sup&gt; row and j&lt;sup&gt;th&lt;/sup&gt; column of Y matrix</td>
</tr>
</tbody>
</table>
1. Introduction

In bioprocessing operations, process monitoring provides important process parameters for maintaining successful operation. For feedback or feedforward bioprocess controls, the controller responses rely upon process monitoring to provide relevant information. Due to the recent advances in NIR spectrophotometer and sophisticated multivariate calibration technique the technology of on- or at-line process monitoring systems have been promoted for bioprocess operations. Several uses of NIR for the upstream fermentation operation have been demonstrated to provide important process parameters.

The thesis describes research on NIR spectroscopy technique for bioprocess monitoring. This is a use of NIR spectroscopy to enhance downstream process operation where surveys have shown the technique has remained novel to downstream process. The primary research involved is: identifying the location of use of NIR monitoring system in order to maximise its effectiveness, determining the required monitoring bioprocess parameters, design and building of a budget NIR spectrophotometer and using multivariate analysis to calibrate the NIR spectrophotometer for three prime bioprocess parameters. The first two outline the research target and the remaining are the primary research activities.

Traditionally, most of the process parameters are obtained from off-line analysis. This can be inaccurate and time consuming so that real time control of the process is difficult to implement. The result of this research has demonstrated that the NIR monitoring provided sufficient process information and has potential in real time process control. This was because the technique is fast and need not require any sample preparation.
In the second chapter of this thesis, a literature survey of the applications in bioprocessing and pharmaceutical industries is given and the chosen downstream process for this research is introduced. This is an alcohol dehydrogenase (ADH) recovery process from bakers yeast homogenate. The recovery process involves flocculation, two-cut fractional precipitation and hydrophobic interaction chromatography (HIC) processes. It has been identified that NIR monitoring should be used in the early stage of recovery by flocculation process. Monitoring of three key bioprocess parameters, contaminants such as cell debris, protein and ribonucleic acid, provides valuable information that can benefit both flocculation control and further downstream operations. Also included in this chapter are the theory of NIR spectroscopic technique and the various types of NIR spectrophotometer. The former illustrated that the NIR spectral absorption responses are weak and highly overlapping signal, therefore a multivariate calibration technique is used (described in Chapter 3). The introduction of various types of NIR spectrophotometers has found that a dispersive grating instrument is simple to implement hence has lower building cost. It provides better photometeric measurement than those observed in non-dispersive instruments, therefore a dispersive grating instrument is used in this research.

The third chapter has highlighted the problem of univariate and simple multiple linear regression calibration methods. Therefore the need of a sophisticated multivariate calibration technique is apparent. Partial least squares (PLS) is selected for all NIR spectroscopic multivariate calibration in this research and the theory of PLS is described in details. The measure of PLS analysis error of the NIR spectrum and the PLS model validation technique used are described. Also the procedure to select an optimum PLS model is given.

The preliminary research has identified both the bioprocess parameters and the location of NIR monitor to be employed, these are three different process contaminants and the flocculation process respectively. Because of the result from the preliminary research and the cost budget reasons, a dedicated NIR spectrophotometer was designed and built to achieved sufficient performance for the subsequent research. This is a dispersive grating spectrophotometer that produce NIR spectrum of wavelength 1900 to 2500 nm. The spectrophotometer employed a
Single element lead sulfide detector and spectrum is achieved by scanning the dispersive spectrum of the grating. The detailed theory of grating spectrophotometer and the design to implementation has been included in Chapter 4. A particularly important research in this instrumentation is that there is a constraint of limited sensitivity by the detector. The work has found that the overall sensitivity can be improved by utilising optical throughput and a different referencing material. Specification of the NIR spectrophotometer together with the performance on quantitative analyses of two single contaminants in process samples are given. These are the concerned contaminants found in process stream (protein and RNA) and the demonstrations included the use of PLS multivariate calibration. In both cases, the results of multivariate calibration have shown that the low budget NIR spectrophotometer provided sufficient resolution for quantitative analyses of bioprocess contaminants.

Use of PLS and the low budget NIR spectrophotometer have demonstrated successful quantitative analyses of single contaminants. This has been extended further to meet the research target of NIR spectroscopic monitoring of multiple contaminants. In Chapter 5, two methods of NIR calibration are shown. These are known as the add-back contaminants and process stream samples for calibration experiments. The former is a calibration experiment based on the full factorial design where samples are composition of the three mentioned contaminants. These calibration samples provided maximum variation in the NIR spectra, thus a relatively robust PLS calibration is expected. Also this method of calibration gave the opportunity to test both the in-house spectrophotometer and the functionality of PLS for multiple contaminants calibration. However, due to technical difficulties in selective removal of contaminants from the process samples the contaminants are added back into the process samples so that the full factorial design could be satisfied, hence the term add-back contaminants calibration.

In Chapter 5 the successful PLS calibration using add-back contaminants and process stream samples are shown. In the add-back calibration where due to the nature of add-back contaminants the calibration range of the contaminants are actually outside the range of interest. However, in the process stream calibration samples are similar to those obtained from a process stream. Therefore the PLS
calibration models have the corrected concentration range of the concerned contaminants although the process stream PLS calibrations are less robust compared to the full factorial add-back calibrations. It is shown in chapter 5 that the prediction of contaminants in two flocculation processes by the PLS process stream calibrated models are better than those predicted by the add-back calibration models. It is concluded that process stream calibration models should be used for process monitoring of contaminants in the flocculation process. The illustration of control setpoint for the flocculation process is also given.

Finally in chapter 6, the conclusion of the thesis and recommendations for future research are given. These are in the aspects of NIR instrumentation and planning of calibration experiments and evaluating difficulties.
2. Near infrared (NIR) techniques

2.1 Summary

In this chapter, the objectives of near infrared spectroscopic technique for downstream bioprocess monitoring are presented. The process studied here is representative of the recovery of intracellular protein in the alcohol dehydrogenase (ADH) from *Saccharomyces cerevisiae* (bakers yeast). Such recovery processes benefit from inserting a flocculation stage at the beginning of the purification process. This allows rapid removal of many contaminants which severely affect subsequent high resolution purification steps (for example fine debris, colloidal protein and nucleic acids). In order to control and optimise the flocculation process, NIR monitoring has an important role to identify and quantify the contaminants in the process so that the flocculation process can be adjust accordingly.

The used of NIR spectroscopy for monitoring bioprocess is surveyed. It has been found that the technique is beginning to be used in fermentation processes. Other than food and agricultural industries who are the dominant users of NIR, pharmaceuticals have also started the use of NIR for process monitoring. The theory of NIR spectroscopy is given and the optical principles operation in today’s commercially available spectrophotometers are also presented.
2.2 Introduction

In the study of ADH recovery process, important process information is often obtained by off-line assays. The results of the off-line assays are generally available hours later and after the process events; by then the assays are redundant for making any decision to affect the process. In the past, only restricted on-line or at-line information were available (e.g. optical density), this confined the process operating windows and reduced the product yield. In order to enhance the bioprocess a rapid monitoring technique for identification of the contaminants levels has been regarded as a priority. With recent advances in optical technology and sophisticated computing power, NIR spectroscopy is potentially ideal for rapid monitoring in bioprocess (Yu, 1992, Belchamber, 1996). The work in the thesis addresses the use of NIR spectroscopic technique for monitoring contaminants levels. This has been applied to the early stage of the ADH recovery process where flocculation can be controlled with the aid of NIR monitoring results.

This chapter presents the ADH recovery process in detail. The emphasis is on the use of NIR for monitoring and its application for enhancing flocculation control. The recent usage of NIR measurements in bioprocess and pharmaceuticals are surveyed and discussed. The theory of NIR absorption by chemicals described by classic mechanical and quantum mechanics are presented. Also the optical principles such as dispersive, non-disperive and non-thermal implemented in NIR instruments are shown.
2.3 NIR in alcohol dehydrogenase (ADH) recovery process

In this section, the bio-process selected for exploration of NIR spectroscopic monitoring technique is described. This is a downstream process operation and the task is to recover the alcohol dehydrogenase enzyme (ADH) from yeast homogenate. The section introduces the medium bakers yeast and the basic recovery process of ADH. The flocculation process employed for enhancing the recovery process is then presented. Lastly, the potential improvements which maybe made using NIR spectroscopy measurements in the control of the flocculation process are described.

2.3.1 Bakers yeast

Bakers yeast used for baking and brewing is also called *Saccharomyces cerevisiae* which translated means “the sugar fungus of beer”. Yeast which is classified with fungi has an oval shape and has a semi-permeable cell wall. This allows certain water soluble substances to pass either into or out of the cell. In this way food materials can pass into the cell and metabolites such as carbon dioxide and alcohol leave it. Each individual cell measures about 6 μm in length and there are approximately 15,000 million cells in 1g.

The make-up of an individual cell is illustrated in figure 2.1. Inside the yeast cell is a jelly like mass called cytoplasm, into which is sited a single nucleus. The nucleus is responsible for cell reproduction and carries the chromosomes which determine the heredity properties of the cell. The vacuole holds in solution organic and inorganic reserve foods materials and the mitochondria produce and store energy and can be likened to a charged battery. Also contained within the cytoplasm, but not shown here, are ribosomes. Ribosomes are responsible for protein synthesis.
In the following section, the ADH recovery process is described. This involves disruption of the yeast cell and releases all the components within. Other than ADH all of the components are regarded as contaminants. Generally, the contaminants are summarised and have the overall statistics given by M. Bulmer (1992) as shown in table 2.1.

<table>
<thead>
<tr>
<th>Contaminants</th>
<th>Approximate Percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polysaccharide</td>
<td>42</td>
</tr>
<tr>
<td>Protein</td>
<td>40</td>
</tr>
<tr>
<td>DNA</td>
<td>2</td>
</tr>
<tr>
<td>RNA</td>
<td>5</td>
</tr>
<tr>
<td>Lipid</td>
<td>7</td>
</tr>
<tr>
<td>Ash</td>
<td>4</td>
</tr>
<tr>
<td>Others</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.1 General statistics of contaminants after yeast disruption.
2.3.2 Recovery process by fractional precipitation

Alcohol dehydrogenase (ADH) is an enzyme that is produced intracellularly from bakers yeast fermentation process. The recovery of enzymes is generally performed by the sequential application of relatively few procedures. The methodology and process operations have proved sufficient for the isolation of many enzymes to varying degrees of purity. The complete operation reflect costs, scale and the process performance (Bonnerjea et al., 1986). The ADH enzyme is located within the yeast cell medium, then rupture of the yeast cells will be necessary. The most common release mechanism is by mechanical rupture and the technique adopted here is known as homogenisation. This has the advantage in cost and ease of scale-up.

![Flow diagram of a traditional process for intracellular protein (ADH) recovery.](image)

In the ADH recovery process studied here, the source material was the readily available packed yeast. Therefore the first operation involved re-suspension of the packed bakers yeast into buffer solution (see figure 2.2). The yeast cells are then
disrupted by the homogeniser, typically operated at 500 bar with 5 passes to achieve 90 percent breakage. The yeast homogenate is first clarified by centrifugation which separates the wanted enzyme along with other soluble component such as nucleic acid, other protein and also fine debris from the bulk of the cell debris. The next stage of the recovery process is the removal of protein as well as the residual cell debris. This is achieved with precipitation processes. Various methods exist for precipitating proteins (Bell *et al.*, 1982) the most common being salting-out using ammonium sulphate as the precipitant. Precipitation can also be used to fractionate selectively a target protein from a protein mixture comprising many different contaminating species including other proteins as well as non-proteinaceous contaminants such as lipids and nucleic aids.

Fractional protein precipitation (figure 2.2) is known as a *two-cut* process. By adjustment of the precipitant concentration, it is possible to achieve selective precipitation. In the first cut unwanted low solubility proteins are removed in the precipitate phase with only a small amount of the desired protein product being precipitated. In the second cut the majority of the desired protein with some contaminating protein is recovered as the precipitate phase. The position of the cuts are made so that an acceptable level of purification is obtained and at a high process yield (Scopes, 1982). Finally, the clarified yeast homogenate is entered into a hydrophobic interaction chromatography (HIC) column where ADH is extracted.

In the following section, the addition of a flocculation process is described which is used to enhance removal of contaminants in the early stage of recovery.
2.3.3 Enhanced recovery process with selective flocculation

In the cell disruption process by homogenisation the cells are broken to release the contents. This produces a biocolloidal suspension containing a substantial fraction of polydisperse sub-micron particles and also releases the cytoplasmic contents (nucleic acid, proteases etc.). It has been reported by Cordes (Cordes, 1986) that microorganisms contain from 1 to 25% nucleic acid on a dry mass basis and mostly in the form of RNA. The lipid content has also been identified by Milburn et al (Milburn et al, 1990) to approximately 10% of cells. These can cause various problems in bio-processing, the most common being the increase in viscosity associated with the release of nucleic acids, and the poor clarification of homogenates by centrifugation due to the colloidal nature of the cell wall debris.

The ADH enzyme is present as a small fraction of the total protein. Therefore, it is essential that the early stages of downstream process operations remove as much of the contaminants as possible, such as nucleic acids, unwanted protein and cell wall materials. The primary objective of initial recovery stages is to obtain well-clarified supernatant which will have a direct impact on any subsequent high resolution stages.

The purification by centrifugation usually gives poor performance in the separation of bio-colloids. This is because the matter in colloidal suspensions tends to remain separate and discrete. One practical method of dealing with colloidal suspensions is to alter their surface chemistry in order to induce coagulation, which makes the resultant flocs easier to separate. This is achieved with a flocculating agent Polyethylenemine (PEI) adding to the yeast re-suspensions and creating larger flocs (Bulmer, 1992). Polyethyleneimine (PEI) has been shown by Atkinson (Atkinson and Jack, 1973) to produce effective selective precipitation of nucleic acids.

Hence, there are two distinct areas of bioprocessing in which polyethyleneimine (PEI) has been employed as a selective flocculant. The first area is in the selective precipitation of unwanted bio-colloidal material (cell debris) in order to facilitate in the clarification of yeast homogenates. The second area is the co-precipitation of
nucleic acids and the enzymes involved in nucleic acid metabolism. The advantages of such co-precipitation are the rapid concentration of certain enzymes whilst the rest of the protein remain in solution. Therefore in the ADH recovery process, PEI has a major use in selective flocculation of cell debris, nucleic acids, lipids and unwanted protein material from yeast homogenates. This gives a significant improvement in removal of contaminants (by centrifugation) and reduces column fouling (Milburn et al, 1990). The flocculation of nucleic acid material also reduces the solution viscosity and increases the density difference between the suspension and floc, thus aiding in the recovery by centrifugation.

2.3.4 Role of NIR monitoring and control to optimise flocculation

It has been understood that the flocculation process aids removal of unwanted contaminants in the early stage of purification. However, the ability to identify any residual contaminants during the recovery process would give significant insight into the process operation. This requires on-line or at-line measurements of contaminant levels throughout the recovery process. For example if the contaminants levels are high at the early stage of purification then the stream should be stopped from going onto the subsequent process or the high resolution chromatography, since the column used can be easily fouled by contaminants.

Previously in the ADH recovery process, at-line measurements of cell debris and off-line protein assays were achieved within a time constraint. But due to time consuming factors in the measurements of other important contaminants such as nucleic acid and lipids (up to 24 hours), these important contaminants measurements are not available to benefit the process operation in real time. Therefore, there is a necessity to develop a method of rapid monitoring in order to enhance the monitoring of the recovery process.
In the ADH recovery process, rapid NIR monitoring in the flocculation process is extremely useful to acquire quantitative information of contaminants (see figure 2.3). These are contaminants such as cell debris, protein, nucleic acid and lipids. The information available can be fed back to a controller which then decides the amount of PEI flocculant induced to the process. Therefore, the NIR monitoring technique allows for closed loop control of the flocculation process which permits an operator or controller to optimise the flocculation process. Under-flocculation limits the effectiveness of purification, while over-flocculation tends to restabilise contaminants in suspension. Furthermore in the case of over flocculation, excess flocculating agent (PEI) remaining in the process stream can also foul the column. Therefore it is important to established a control set point for the flocculation process, in order to achieve an optimum operating window.

Fig 2.3 Flow diagram of the alcohol dehydrogenase (ADH) recovery process; including the flocculation stage and NIR monitoring by sampling off the main process stream.

The technique can also be used further downstream in the purification process, i.e. the two-cut precipitation can be further optimised when contaminants information are gained. In process control and optimisation, process measurements are the key
inputs for these operations. Due to lack of process monitoring methods, the control of the recovery process has previously only been performed with minimum at-line measurements (cell debris by optical density) and ADH yield. However for the reasons given above, the use of NIR spectroscopic technique will facilitate information to maximise the yield of ADH and the removal of the contaminants.

2.4 Applications of near infrared spectroscopy

NIR is one of the most successful techniques for materials measurements and offers a combination of speed and precision measurements. A textbook by Osborne et al. (1993) has covered the practical NIR spectroscopy in the field of food and beverage analysis. Simple NIR analysers are able to produce quantitative information such as moisture, fat, protein contents in flour within seconds. While NIR analysers have been heavily employed in these sectors, they have also been used in many other applications such as tobacco, wool, textiles, baked products, petrochemicals industries. A handbook of near infrared analysis by Burns and Ciurczak (1992) has summarised these applications.

With modern generation of monochromator, diode array detectors and acousto-optically tuneable filters used in NIR instruments (section 2.6) and together with high speed computers and sophisticated multivariate calibration technique (chapter 3), as well as the above mentioned industries, bioprocessing and pharmaceutical analysts became interested in the applications of near infrared spectroscopy. As will be seen, there are reports in the monitoring of bioreactors, but the survey of literature has not shown any reports of the use of NIR in downstream recovery and extraction. It is believed that the work reported in this thesis is breaking new ground.
2.4.1 Near infrared analysis in pharmaceuticals

In the pharmaceutical industry uses of NIR for on-line monitoring have been in progress. Due to the competitive industry and since many of the NIR monitoring units are on provisional trials, little has been published so far. However, the European Pharmaceutical Review publication by P. Hailey from Pfizer Central Research (Hailey, 1996) has point out that the NIR spectroscopy could have a major role in the measurement of pharmaceutical manufacturing. He has suggested that throughout the manufacture process, traditional laboratory based testing to ensure the product specification will be feasibly replaced by NIR spectroscopy (i.e. raw material or excipient identification at the commencement of a drug product). It was mentioned in the same paper that NIR was used to characterise the blending process of excipient in a vessel. Aldridge (1996) reported at the 23rd FACSS conference in September 96 that on-line NIR was used to determine the homogeneity and provided a far greater understanding of the blending process. At this conference, Q. Wang from Brucker Optics demonstrated an NIR instrument that was able to determine concentration of aspirin in the tablets in three different ways; NIR diffuse reflectance measurements made on the surface of the tablet as well as through the original blister packages, and transmittance measurements of whole tablets.

These are just some indications of the pharmaceutical industry uses of NIR measurement in manufacturing process and analyses of products quantitative. Similar agreements were also presented in the same European Pharmaceutical Review publication by Brutsche (Brutsche, 1996). There, concern was reported about the approval of NIR spectroscopy from the FDA. However, it was indicated that research and development on the new methods is ongoing at various FDA field laboratories and suggested that FDA will accept this methodology for synthesis of active ingredients, seeing NIR as alternative to present chromatographic methods. FDA Center for Drug Analysis Director Thomas Layloff predicts that NIR, in particular, will become a predominant methodology in the quality control lab within three to five years (Brutsche, 1996).
2.4.2 At-line and on-line NIR monitoring of bioreactors

Fermentation processes in bioreactors are often run for hours or even days of operation. To ensure these processes operate efficiently, real-time monitoring of biomass, substrates, nutrients, and products is essential. Present analytical methods such as absorption of UV radiation are useful in the analysis of protein and nucleic acid solutions. However, the turbid nature of fermentation broths causes problems for traditional spectroscopic techniques because of light scattering by the cells and reduced spectral signal related to composition. Other methods such as high-performance liquid chromatography (HPLC) are time-consuming and require the use of environmentally unacceptable solvents. Compared to other regions of the infrared spectrum, the absorptions that appear in the NIR region are weak and can be used to investigate samples that are highly absorbing and strongly light scattering. These features in NIR have advantages for on-line and at-line monitoring.

Researchers at the Center for Process Analytical Chemistry at the University of Washington have developed a fibre optic unit for a portable computer system which is capable of acquiring the short wave (700 - 1100 nm) NIR spectrum (Lysaght et al., 1991). They reported on its use in the on-line analysis of the anaerobic batch fermentation of *Saccharomyces cerevisiae* (Cavinato et al., 1990). The spectra were collected by diffuse reflectance, thus the measurements were non-invasive. The calibrated system has shown the capability of measurements of ethanol concentration from 0 - 15 % (w/w) with standard error of 0.42 % (w/w). In the work reported by Brimmer and Hall (Brimmer and Hall, 1993) using NIR diffuse reflectance measurements in 1100 to 2500 nm, nutrient levels (e.g. oil) in a fermentation process were determined directly with no sample pretreatment. There the accuracy of the NIR measurements was affected by highly scattering medium in the fermentation, but from a combination of statistics and spectroscopic knowledge the problem was overcome.
NIR spectroscopy was used in at-line control and fault analysis of an industrial high cell density *escherichia coli* fermentation (Macaloney *et al.*, 1994). There, at-line NIR provided the opportunity for rapid, simultaneous measurements of carbon and nitrogen nutrients, byproduct and biomass to improved process control. In this fermentation, the carbon nutrient was fed in an open loop, feed forward control was implemented to maintain carbon limited growth. But such schemes are not robust if disturbances occur, i.e. due to failure in feed system. With at-line NIR measurements of the important parameters provided a method for stabilising and improving the feed forward control by feeding the relevant parameters information to the controller. Also reported in this work, NIR analysis was also successfully used to identify bioreactor faults by detecting a pH controller failure. Another successful application is a lactic acid fermentation process reported by Vaccari (Vaccari *et al.*, 1994). There lactic acid, glucose and biomass concentration were determined by NIR spectroscopy using a special sample presentation technique.

Summarising the above, NIR spectroscopy has been demonstrated for both at-line and on-line measurements of key parameters in a bioreactor process. These clearly have benefit to bioprocess operation, as well as fault detection, relevant process information can improve the process control and hence improve the production yield. Another major advantage of NIR measurements is that it can be a non-invasive technique and without any sample preparation. In section 2.6.5 later, it is shown that using an NIR instruments together with fibre optics probe, several measurements can be multiplex to one instrument thus providing process information from various location of the process or measurements of various position in a bioreactor.
2.5 Theory of near infrared spectroscopic technique

2.5.1 Near infrared and infrared absorption

The near infrared and infrared absorption are concerned with the absorption of electromagnetic radiation. Electromagnetic radiation can be considered as a simple harmonic wave. The properties which undulate are interconnected electric and magnetic fields and it is these which interact with matter to give rise to a spectrum. Any simple harmonic motion has the properties of the sine wave defined by the equation 2.1 and this has also been known as the basic equation of wave motion.

\[ y = A \sin (2\pi vt) \]  \[2.1\]

where \( y \) is the displacement with a maximum value \( A \), this is dependent on the frequency of a harmonic wave (\( v \)) and the displacement time.

There is another property of the wave is the distance travelled in a complete cycle, known as the wavelength (\( \lambda \)). It is necessary to express the equation [2.1] in terms of variation of displacement with distance instead of with time. This involves substitution \( t = \lambda / c \) where \( \lambda \) is the distance of the wave travelled in time \( t \) at velocity \( c \). The velocity \( c \) is known as the velocity of light in vacuo and is a universal constant. The wavelength may be defined as;

\[ \lambda = c / \nu \]  \[2.2\]

The wave model given may be used to explain many properties of electromagnetic radiation but it fails to account for phenomena associated with the absorption or
emission of energy. Therefore it is necessary to view electromagnetic radiation as a stream of discrete particles called photons with an energy proportional to the frequency of the radiation.

2.5.1.1 The harmonic oscillation of molecule vibration characterised by quantum mechanic model

Consider a mechanical model of molecular vibration, a mass $m$, attached to one end of a spring and the other end of which is fixed. The force of gravity is constant and will therefore only influence the equilibrium point and not the motions of the masses about that point. The disturbance of the mass along the axis of the spring results in a motion which is described by Hooke’s law. This states that the restoring force $F$ exerted by the spring is proportional to the distance $y$ that it has travelled from the equilibrium position;

$$F = -ky$$  \[2.3\]

where $k$ is a constant called the force constant. The acceleration $a$ of the mass from equilibrium is

$$a = \frac{d^2y}{dt^2}$$  \[2.4\]

Therefore by application of Newton’s second law of motion, $F = ma$

$$\frac{md^2y}{dt^2} = -ky$$  \[2.5\]
One solution of the differential equation [2.5] is

\[ y = A \sin (\alpha t) \]  \hspace{1cm} [2.6]

where \( \alpha = \sqrt{k/m} \). After one period of motion, \( y \) will return to its initial value and the sine wave will repeat each time \( \alpha t \) is increased by \( 2\pi \), from which it may be deduced that

\[ \nu = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \]  \hspace{1cm} [2.7]

Substituting \( \alpha = \sqrt{k/m} = 2\pi \nu \) into equation [2.6] gives equation [2.1]. Therefore, electromagnetic waves and mechanical oscillators may be described in the same terms and the significance of spectroscopic measurements lies in the association between the frequency of radiant energy and the frequencies of molecular motions. For a diatomic molecule, for example, equation [2.7] is modified to accommodate a description of a spring system consisting of two masses \( m_1 \) and \( m_2 \), by substituting the reduced mass \( \mu = (m_1 m_2)/(m_1 + m_2) \) for \( m \),

\[ \nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}} \]  \hspace{1cm} [2.8]

This is a good approximation of the vibration of a chemical bond, with \( m_1 \) and \( m_2 \) the masses of the two atoms and \( k \) being the force constant for the chemical bond. Using this simple mechanical model it is possible to explain many spectral observations in the infrared spectrum.
Fig 2.4 The energy of a diatomic molecule undergoing simple harmonic motion (dotted curve) and anharmonic vibration (solid curve).

The potential energy diagram for a harmonic oscillator is shown as the dotted curve in figure 2.4. If considering the mass and spring model, the energetics of the system undergoes cyclic conversion from potential energy to kinetic energy. At the equilibrium position, potential energy may be considered to be zero, but as the spring is compressed or stretched by a small amount $dy$ it increases by an amount $dE$.

Hence,

$$dE = -F dy \quad [2.9]$$

Combining equations [2.3] and [2.9] gives

$$dE = ky dy \quad [2.10]$$

Integrating equation [2.10] from $y = 0$ to $y$, gives

$$E = ky^2/2 \quad [2.11]$$

The above is the equation of a parabola, with $y$ the displacement from equilibrium. At the turning point of the motion corresponding to maximum amplitude $A$, the
kinetic energy is zero and the total energy is potential energy. As the spring is compressed or stretched, the energy reverts to kinetic and decreases parabolically to zero at the equilibrium position.

\[ \Delta E = h\nu \]  

where \( h \) is a constant known as Plank's constant. The significance of equation [2.12] is that if a beam of radiation containing a wide range of frequencies is directed onto a molecule in energy state \( E_1 \), energy will be absorbed from the beam and a transition to energy state \( E_2 \) will occur. A detector place to collect the radiation after its interaction with the molecule will show that the intensity of the radiation has decreased only at frequency \( \nu = \Delta E/h \). All other frequencies are undiminished in
intensity and in this way an absorption spectrum is produced (figure 2.5). In practice, since the number of energy levels for any molecule is infinite and there are many possible transitions. Therefore, a spectrum from the simplest of molecules would be very complex if it were not for the selection rules (see later).

A molecule in space may possess many forms of energy such as vibrational energy due to the periodic displacement of its atoms from their equilibrium position and rotational energy by virtue of bodily rotation about its centre of gravity. Absorption of infrared radiation is largely confined to molecular species for which energy differences exist between different vibrational and rotational states. The energy required to cause a change in rotational states is however, very much smaller than for vibrational states and rotational absorption bands may only be observed in the case of gases. Therefore, for the study of infrared spectra of solid and liquid samples, only vibrational motions need to be considered.

Vibrational energies, like all other molecular energies, are quantised and the allowed vibrational energies for any particular system may be found by solving a series of partial differential equations known as the quantum mechanical wave equations. Solutions of these equations assuming a simple harmonic oscillator are found for energy levels.

$$E = (\nu + 0.5)h\nu$$  \[2.13\]

where \(\nu\) is the vibrational quantum number which may have the value 0, 1, 2, etc. From equation [2.13], it may be deduced that the lowest vibrational energy level \(E_0\) when \(\nu\) takes the value zero equals \(h\nu/2\). Therefore, a molecule can never have zero vibrational energy; i.e. the atoms can never be completely at rest relative to each other. \(E_0\) depends only on the strength of the chemical bond and the atomic masses and the prediction of \(E_0\) is the basic difference between wave mechanical and classical approaches to molecular vibrations. Promotion to the first excited state (\(\nu = 1\)) thus requires absorption of radiation of energy, \((3h\nu/2)-(h\nu/2)= h\nu\); the frequency \((\nu)\) of radiation that will bring about this change is identical to the vibration frequency of the bond defined by equation [2.8], therefore
From equation [2.13] it can be seen that $\Delta E$ given by equation [2.14] is the energy associated with the transition between any pair of adjacent levels.

**Selection rules**

1. Quantum theory indicates that the only allowed vibrational transitions are those in which $\nu$ changes by one ($\Delta \nu = \pm 1$).

2. Spectral bands will only be observed if the vibration interacts with the radiation. Therefore, the vibration due to radiation is electromagnetic in origin, such interaction depends upon the existence of an electric moment across the vibrating bond. Thus, homonuclear diatomic molecules, for example, do not exhibit vibrational absorption bands. It is sufficient, however, for polar bonds to be present such that molecular vibration causes a temporarily induced dipole moment.

### 2.5.1.2 The overtone bands due to anharmonic oscillation

The quantum mechanical treatment of a harmonic oscillator explains the observed infrared absorption bands due to fundamental modes of molecular vibration. However, it does not explain the presence of overtone bands in the NIR. These bands arise from transitions where $\Delta \nu$ is $\pm 2$, $\pm 3$ etc. and so are forbidden by the selection rule 1) above. The answer to this anomaly is that real molecules and real bonds do not obey exactly the laws of simple harmonic motion. As two atoms approach one another, Coulombic repulsion between the two nuclei causes the potential energy to rise more rapidly than the harmonic approximation predicts, and when the interatomic distance approaches that at which dissociation occurs, the potential energy levels off (figure 2.4). It has been illustrated by the dotted curve in figure 2.4 that the success of harmonic model stems from the fact that the two curves are almost identical at low potential energies. An empirical function, which fits the solid curve in figure 2.4 to a good approximation is

$$E = E_D(1 - e^{-\alpha v})^2$$

[2.15]
where α is a constant for a particular molecule and $E_D$ is the dissociation energy. When equation [2.15] is used to solve the wave mechanical equations, then the solution for anharmonic oscillator becomes;

$$E = (v + 0.5)h\nu - (v + 0.5)^2h\nu x - (v + 0.5)^3h\nu x^2 - \ldots$$  \[2.16\]

where $x, x', \ldots$ are known as anharmonicity constants which are small and positive and of decreasing magnitude. However, for small values of $v$, the third term and beyond in equation [2.16] may be ignored and

$$E = (v + 0.5)h\nu - (v + 0.5)^2h\nu x = h\nu(1 - x(v + 0.5))(v + 0.5)$$  \[2.17\]

The above equation is similar to the simple harmonic oscillator equation [2.13] if $v$ was replaced by

$$v = v[1 - x(v + 0.5)]$$  \[2.18\]

Thus the anharmonic oscillator behaves like the harmonic oscillator but with an oscillation frequency which decreases steadily with increasing $v$. $E_0$ is now $(h\nu/2)(1 - 0.5x)$. The energy associated with a transition from $v$ to $v + \Delta v$ is;

$$\Delta E = h\nu[1 - (2v + \Delta v + 1)x]$$  \[2.19\]

and the selection rules are $\Delta v = \pm 1, \pm 2, \pm 3, \ldots$. Thus, they are the same as for the harmonic oscillator but with the additional possibility of larger jumps. Although predicted by theory, however, these are in practice of rapidly diminishing probability and normally only bands due to $\Delta v = \pm 1, \pm 2, \pm 3$ at the most have observable intensity. Furthermore, at room temperature nearly all the molecules in a particular sample exist in the lowest energy level ($v=0$). This follows from the Maxwell-Boltzmann law which describes the proportion of molecules in an excited state $n_1/n_2$, where $n_1$ is the number of molecules in the excited state and $n_2$ the number of molecules in the ground state, as an exponential function
where $k$ is a universal constant known as the Boltzmann constant and $T$ is the absolute temperature. At room temperature $\Delta E = h\nu \gg kT$.

From the above discussion it is apparent that the three most important transitions in IR spectroscopy are

i) $v = 0 \rightarrow v = 1; \Delta v = +1$, thus $\Delta E = h\nu(1 - 2\alpha)$

ii) $v = 0 \rightarrow v = 2; \Delta v = +2$, thus $\Delta E = 2h\nu(1 - 3\alpha)$

iii) $v = 0 \rightarrow v = 2; \Delta v = +3$, thus $\Delta E = 3h\nu(1 - 4\alpha)$

To a good approximation since $\alpha = 0.01$ the three bands lie very close to $v$, $2v$ and $3v$. The line near $v$ is called the fundamental absorption, while those near $2v$ and $3v$ are called the first and second overtones respectively. The lowest wavelength at which absorption of radiation at fundamental vibration frequencies occurs is at 2500 nm and the region between 700 to 2500 nm is defined as NIR. It is in this NIR region that absorption at overtone frequencies occurs. In addition to overtone bands, combination and difference bands with frequencies that are the sums or differences of multiples of their fundamental frequencies,

$$v_{\text{combination}} = n_1 v_1 \pm n_2 v_2 \pm \ldots$$ [2.21]

where $n_1, n_2, \ldots$, are positive integers. Combination bands are of very low probability unless they arise from no more than two vibrations involving bonds which are either connected through a common atom or multiple bonds. Difference bands, which are due to absorption by molecules residing excited vibrational states, are of very low probability at room temperature as a consequence of equation [2.20].

Although the infrared region of the electromagnetic spectrum extends from 700 to $10^6$ nm, it has been commonly divided into three region of near, middle and far infrared (see table 2.2).
<table>
<thead>
<tr>
<th>Spectrum</th>
<th>Transitions characteristic</th>
<th>Wavelength (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near Infrared (NIR)</td>
<td>Overtone combinations</td>
<td>700 to 2500</td>
</tr>
<tr>
<td>Middle infrared (MIR)</td>
<td>Fundamental vibrations</td>
<td>2500 to $5 \times 10^4$</td>
</tr>
<tr>
<td>Far infrared (IR)</td>
<td>Rotations</td>
<td>$5 \times 10^4$ to $10^6$</td>
</tr>
</tbody>
</table>

Table 2.2 Regions of infrared spectrum

2.5.2 Physics of the interaction of radiation with matter

When monochromatic radiation interacts with a sample (biochemical in our application), it may be absorbed, transmitted or reflected (figure 2.6). The law of energy conservation requires that the total radiation power incident on the sample ($P_o$) must equal the sum of the radiant power absorbed ($P_A$), transmitted ($P_T$) and reflected ($P_R$).

$$P_o = P_A + P_T + P_R \quad [2.22]$$

The significance of equation [2.22] in spectrophotometry is that $P_A$ may be deduced by measuring $P_o$, $P_T$ and $P_R$. And if the experiment is ideally arranged then either $P_T$ or $P_R$ is zero.

![Fig 2.6 Interaction of radiation with matter](image-url)
2.5.2.1 Transmission of radiation

Propagation of radiation through a medium other than a vacuum takes place through temporary polarization of the particles contained in the medium. The polarization arises from interaction of the electric vector of the radiation with electrons which causes them to oscillate with respect to their nuclei. If there is not any absorption taking place, the energy required for polarization is remitted unaltered. Therefore the frequency of the radiation is unchanged.

However, the time lapse between retention and remission of the polarization energy causes the velocity of the radiation to decrease and the ratio of the velocity in vacuo to that in the medium is called the refractive index \((n)\) of the medium. The whole process of transmission is thus a stepwise one in which the oscillating particles act as intermediates. For particles which are small compared with the wavelength of the radiation, destructive interference ensures the propagation of the beam along its original path, i.e. electromagnetic radiation travels in straight lines. When the radiation crosses the boundary between two media of different refractive index, however, an abrupt change in the direction of propagation occurs (figure 2.6) and this is known as refraction. This deviation is defined by Snell's law:

\[
\frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1} \quad [2.23]
\]

where \(n_1\) is the refractive index of the less dense medium, \(n_2\) that of the denser medium and \(\theta_1, \theta_2\) are the angles of the beam from the normal in each medium.
2.5.2.2 Absorption measurements

The mechanism of absorption of NIR and IR radiation has been discussed in section 2.5.1. It is now necessary to consider the quantitative aspect of absorption. Analytical absorption spectroscopy is concerned with measuring the relative amounts of radiant energy absorbed at each frequency and from these measurements inferring the amounts of various substances in a mixture since different materials absorb at different frequencies and exhibit different intensities of absorption. The attenuation of the transmitted radiation by an absorbing sample is described by the Beer-Lambert law. This states that the fraction \( \frac{dP}{P} \) of radiant energy \( P \) absorbed by an infinitesimal thickness, of sample is proportional to the number of molecules \( dn \) in that thickness, or

\[
- \frac{dP}{P} = k \ dn \quad [2.24]
\]

Integrating through the sample gives

\[
- \ln \left( \frac{P_T}{P_o} \right) = k \ n \quad [2.25]
\]

where \( P_o \) is the power of the incident and \( P_T \) that of the transmitted radiation, and \( n \) is the number of molecules in the path of the beam.

Since \( n \) is proportional to the concentration \( (c) \) of molecules in the sample and the thickness \( (b) \) through which the radiation passes,

\[
\log \left( \frac{P_o}{P_T} \right) = abc \quad [2.26]
\]

where the ratio of \( P_T \) to \( P_o \) has been inverted to remove the negative sign, and the logarithm is now based to 10. The constant \( a \) is called the absorptivity and if \( c \) is expressed in mol l\(^{-1}\) and \( b \) in cm, \( a \) is given the special symbol \( \varepsilon \) mol\(^{-1}\) 1cm\(^{-1}\) and called the molar absorptivity. A more rigorous derivation of the Beer-Lambert law is given by Swinehart, 1972.
Experimentally, the fraction of radiation \( (P/P_0) \) transmitted by the sample is measured and this is called the transmittance \( (T) \). This has also been known as the ratio technique (see chapter 4). In practice, the transmittance is converted to the absorbance \( (A) \) which is defined by

\[
A = \log \frac{1}{T} = \log \left( \frac{P_0}{P} \right) \tag{2.27}
\]

### 2.5.2.3 Absorbance and analyte concentration

The absorbance equation [2.27] can be represented by equation [2.28];

\[
A = abc \tag{2.28}
\]

Equation [2.28] is employed to determine the concentration of a sample from its measured absorbance. It is necessary to have an accurately defined sample thickness and to determine the value of the absorptivity using a series of samples of known concentration. In situations where the Beer-Lambert relationship rigorously holds, a plot of absorbance against concentration will result in a straight line through the origin with slope \( ab \) (figure 2.7). Since \( b \) is known, \( a \) may be determined easily. The linear relationship between absorbance and path length at a fixed concentration of absorber has no known exceptions, but deviations from the direct proportionality of absorbance with concentration for fixed path length are frequently encountered, see figure 2.7.

![Graphical statement of the Beer Lambert law](Fig 2.7) Graphical statement of the Beer Lambert law \( A = abc \) showing the effect of deviations due to polychromatic radiation, chemical interaction (a) and scatter, reflection (b)
The major causes of non-adherence to the Beer-Lamber law are as follows;

1. Complete absence of reflection is assumed, i.e. \( P_R = 0 \) in equation \([2.22]\). If reflection occurs, then \( P_A \) cannot be deduced by measuring \( P_T \) since some of the radiant power has been lost other than by absorption.

2. The relationship is a limiting one for low concentration because absorptivity is altered by electrostatic attractions between the molecules at concentrations (Kubelka, 1948).

3. Complete absence of scatter is assumed. The presence of scatter causes an increase in effective path length hence an apparent increase in \( A \).

4. Monochromatic radiation is assumed since absorbance changes with wavelength to produce an absorption spectrum. The value of the absorptivity will be determined for a given wavelength which should be a peak maximum. For a broad peak the variation of absorbance with wavelength will be small at the maximum. Therefore the measurement will not be seriously dependent on wavelength accuracy of the instrument.

Another factor related to monochromaticity of the radiation is contamination with stray radiation arising from reflections from various surfaces in the instrument. This radiation may differ greatly in wavelength from that of the principal radiation and may not have passed through the sample. When measurements are made in the presence of stray radiation, the observed absorbance \( A' \) is given by

\[
A' = \log\left(\frac{P_0 + P_s}{P_T + P_s}\right)
\]  
[2.29]

where \( P_s \) is the power of the stray radiation.
5. Apparent deviations are encountered as a result of chemical interactions such as association, dissociation or reaction of the absorbing species with the solvent or other solutes.

Although the Beer-Lambert law forms the basis for quantitative absorption spectroscopy, adherence to it by a given system cannot be assumed. In addition, as a consequence of various interferences in the measurements it is not always plausible to assume a priori, that the line should go through the origin. Therefore, a more realistic representation of the relationship between absorbance and concentration is

\[ A = A_0 + abc \]  

[2.30]

where \( A_0 \) is the intercept. In spite of background absorption and non-linear behaviour it is often still possible to described the curve, at least over a limited range by equation [2.30].

In the bioprocessing applications where there may be biomass present, then the Beer-Lambert law may not hold for two reasons. Firstly, the path length \( b \) ceases to be a constant because of the effect of scattering. This problem may however be overcome by the selection of measurement and reference wavelengths at which scatter is constant. Secondly, since some of the radiation undergoes diffuse reflectance, \( \log 1/T \) does not represent the attenuation of the beam by absorption.

In order to obtain a true reading for the desired analyte, it is often necessary to correct the reading for background absorbance (i.e. water absorption band in NIR or scatter effect in general spectroscopy). This background correction may be achieved in a number of ways, the principal ones being double beam spectrophotometry, the cell-in, cell-out method and the baseline method. The first two methods involve the use of the solvent in a matched cell in a reference beam, or in the cell read immediately after the sample, as a reference absorber. \( A \) may then be taken as \( \log(P_{\text{solvent}}/P_{\text{solution}}) \).
In many cases, a more general procedure is required and this often involves baseline correction. In simple situations, the absorbance due to the background interference merely raises the overall baseline by a constant amount. This background is equal to the minimum absorbance in the spectrum and all that is necessary is to subtract this reference reading from that at the desired peak maximum, i.e. corrected absorbance = $A_{\text{max}} - A_{\text{min}} = \Delta A$. Usually the spectrum of the interference alone is unobtainable so an average linear background has to be assumed even if the absorbance of the interference is not equal at $A_{\text{max}}$ and $A_{\text{min}}$ (figure 2.8a). If information about the spectrum of the contaminant is available, then it may be desirable to make a correction using a wavelength other than $A_{\text{min}}$. If the background is linear but not constant, the tangent method (figure 2.8a) may be employed. This consists of drawing a tangent across the base of the peak but is only valid if the tangent remains reasonably constant with changing composition across this base.

The approach used in the present work is the cell-in, cell-out method and this is explained in section 4.7.2.

### 2.5.2.4 Analyte concentration determined from overlapping absorption peaks

If the background has a definite absorption spectrum with a peak which overlaps with the peak of interest, then it becomes necessary to resort to more complex correction techniques. These techniques may also be extended to analysis of multicomponent
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

systems (see Chapter 4). The corrections depend on the law of additivity which states that the absorption of radiation by one species is unaffected by the presence of other materials, whether this absorbs or not. In other words, the observed absorbance at one given wavelength is a linear sum of the products of the individual concentrations \(c_i\) and absorptivities \(a_i\) for each of the species present;

\[
A = b \sum a_i c_i \tag{2.31}
\]

To illustrate the procedure, consider a two component mixture \(p + q\) in which \(p\) has a maximum at wavelength \(\lambda_1\), and \(q\) has a maximum at wavelength \(\lambda_2\) (figure 2.9). The absorbances of the mixture \(A_1\) and \(A_2\) at the two wavelengths \(\lambda_1\) and \(\lambda_2\) may be described by the two equations

\[
A_1 = a_{1p}bc_p + a_{1q}bc_q \tag{2.32}
\]

\[
A_2 = a_{2p}bc_p + a_{2q}bc_q \tag{2.33}
\]

The four absorptivities \(a_{1p}, a_{1q}, a_{2p}, a_{2q}\), the path length \(b\) and the absorbances \(A_1\) and \(A_2\) are all experimentally determinable. Therefore the simultaneous equation [2.32] and [2.33] may be solved for \(c_p\) and \(c_q\), the concentrations of \(p\) and \(q\) in the mixture. However, the above method is only valid if the Beer-Lambert law is followed by both components. The Beer-Lambert law has been shown to be valid for the NIR spectrophotometric analysis of mixtures of amides (Krikorian and Mahpour, 1973).

The use of simultaneous equation analogous to [2.32] and [2.33] for the estimation of a number of components is possible in principle. However, this becomes of increasing practical difficulty. For example, assume that it is desired to determine the concentrations of four substances in a mixture. It now becomes necessary to make at least four experimental measurements of absorbance values. Furthermore, these measurements must be made at wavelengths at which the absorbance of each is mainly due to a different one of the four compounds, i.e. it is of no use to measure
four bands entirely or principally to the same compound. Therefore the mathematical methods of multivariate analysis are applied in this thesis as reported by chapter 3. The key feature is that complete spectra are measured rather than just absorbances at few wavelengths.

![Absorbance](image)

Fig 2.9 Simultaneous determination of two components in a mixture

### 2.6 NIR instruments

It has been previously said that the NIR absorptions are primary due to overtones and combination bands. These ultimately require instruments that are highly sensitive and have good spectral reproducibility. Sensitivity is achieved with modern solid state detector fabrication giving quality detection while spectral reproducibility can be established with careful attention to instrument design.

NIR instruments can be divided into three categories, these are dispersive, non-dispersive systems and non-thermal sources. These categories are described briefly in the following. The instrument used in this project is a dispersive system, and will be further discussed in chapter 4. The use of fibre optical process monitoring is also introduced such that idea of use can be visualised for further possibilities in process monitoring.
2.6.1 Optical principles in NIR instruments

The technique employed by which light is spectrally selected or modulated defines the optical operation of an NIR instrument. Instruments are separated into three categories, these are dispersive, non-dispersive and non-thermal techniques (see figure 2.10). In the dispersive and non-dispersive systems, these generally offer broad band spectrum measurements. For these systems, thermal radiation produced by an incandescent filament is used, frequently quartz halogen (QH) lamps are used.

In non-thermal systems, 'cold' sources are used where wavelength selection is inherent in the source's spectrally narrow emitting range. Non-thermal instruments are generally expensive compared to the previous two systems.

---

Fig 2.10 The three categories of optical methods used in NIR instruments. The instrument used in the project was a grating spectrophotometer with scanning arrangement.
2.6.2 Dispersive systems

In a dispersive system the constituent wavelengths of light are separated spatially. Figure 2.11 illustrates the concept of a dispersive system and indicates that the measuring samples in a transmittance configuration.

![Diagram](image)

Fig 2.11 Configuration of monochromator used for absorption measurements.

2.6.2.1 Monochromator and spectrograph instruments

The classical dispersing element is the prism. However, the prism is an inefficient arrangement with low and nonlinear dispersion. Most modern spectrophotometers use a grating monochromator and can be configured as figure 2.11.

There are two primary types of diffraction gratings, that are either manufacture by use of a ruling engine by burnishing grooves with a diamond stylus or holographically with the use of interference fringes generated at the intersection of two laser beams. The ruled gratings are plano or concave and possess grooves each parallel with the next. Holographic grating grooves are either parallel or of unequal distribution in order that system performance can be optimised. Holographic gratings are generated on plano, spherical, toroidal, and many other surfaces.

A diffraction grating has a substrate usually made of optical material with a large number of parallel grooves ruled or replicated in its surface and overcoated with a reflecting material such as aluminium. The grating equation may be derived by
supposing a section through the grating surface, normal to the ruling direction as a sawtooth pattern as shown in figure 2.12.

![Figure 2.12 A section of diffraction grating with sawtooth pattern](image)

Light rays A and B, of wavelength $\lambda$, incident on adjacent grooves at angle $I$ to the grating normal are shown. These are diffracted by the grating and have a diffracted angle $D$ to the grating normal. The path difference between the diffracted rays A' and B' rays can be seen to be:

$$a \sin I + a \sin D$$  \[2.34\]

Note also that if angle $D$ is on the opposite side of the grating normal from $I$, then it is of the opposite sign.

Summing of the rays A' and B' results in destructive interference if this path difference is equal to any odd multiple of half the wavelength, and constructive interference if the path difference is equal to any multiple of the wavelength $\lambda$.

$$a(\sin I + \sin D) = m\lambda$$  \[2.35\]

where $m$ is an integer, and is the order of diffraction. Thus light of different wavelengths can be spatially separated, since each wavelength has constructive interference at different angle $D$. This is known as the basic grating equation.
In a monochromator, the design is such that the input light will enter the grating via an entrance slit, diffracting off the grating and exit via an output slit. Light of other wavelengths is absorbed. At any grating setting only a very small range of angles around \( D \), ideally one wavelength passes through the monochromator at a time. But in a spectrograph, an output slit is not necessary, but rather a wide range of angles \( D \) are looked at simultaneously. Therefore a range of wavelengths which satisfy the grating equation for the range of angles \( D \) are presented. The result is that the output is a long strip of over which the various wavelength are dispersed in a geometry. The output spectrum of a spectrograph therefore is linearly dispersed such that a diode array detector can be used for spectrum measurement. This has the advantage over the tradition scanning (monochromator) which requires a period to scan the required spectrum, while the diode array detector can acquired the broad band spectrum instantaneously.

Cost and simplicity dictated the use of a NIR spectrophotometer using a spectrograph for this project. The detailed configuration is presented in chapter 4 where the compromises inherent in the use of a scanning system are also discussed.

2.6.2.2 Acousto-optically tuneable filters

The acousto-optical tuneable filters (AOTF) is another promising technology for fast NIR spectral measurements. The principle of acousto-optic diffraction in an anisotropic crystalline medium. As shown in figure 2.13, the broad-band light is directed onto a TeO\(_2\) crystal. On one surface of the specially cut and polished crystal is bonded to an acoustic transducer. The acoustic transducer is a piezoelectric material driven by 1-4 Watts of radio frequency (RF) coupled into the transducer. The high frequency (30 - 200 MHz) acoustic waves induce index of refraction waves in the acousto-optical material. The waves travel through the crystal in a very short time (typically 20 - 30 \( \mu \text{sec} \)) and the waves fill the crystal, interacting with the broad-band light travelling through the crystal. The angles of the crystal axis, the relative angles of the broad-band light, and the acoustic wave are defined in the crystal design to optimise the light-acoustic wave interaction. The result of the interaction is the
splitting of the broad band light into three beams. The centre beam is the unaltered white light travelling through the crystal. The TeO₂ material has virtually no absorption from the visible spectrum up to 5 μm in the infrared region. The two new beams generated by the acoustically excited crystal are monochromatic and orthogonally polarised. These beams can be used as monochromatic light sources for spectral measurements.

Fig 2.13 Principle of operation of acousto-optical tuneable filters

The primary advantage of AOTF optics is that the wavelength is electronically selected without the delays associated with mechanical scanning monochromators. The electronic wavelength selection allows a very high duty-cycle because almost none of the time is wasted between wavelength switching. In comparison with a scanning instruments the advantage is not only that the scanning rate is orders of magnitude faster but also the wavelength access can be random. AOTF can have a working apertures of up to 10 x 10 mm and efficiencies are comparable with interference filters. The wavelength resolution depends on the interaction length in the crystal, is fixed at the time of manufacture and subsequently cannot be changed. Crystals can readily be designed for the usual range of NIR requirements. Besides the speed and efficiency of wavelength selection, the acousto-optical tuneable spectrometers (AOTS) are much smaller than grating monochromators.

The first report on demonstrating the use of a complete computerised spectrophotometer for NIR analytical purposes based on AOTF was presented in
1986 (Kemeny et al., 1986). Since then, companies (Bran + Luebbe, Westinghouse and Infrared Fiber Systems) have also demonstrated AOTF based NIR analyser (Ciurczak, 1990). The acousto-optic materials, their theory and devices have been available for some time. The reason for their not being applied in commercial NIR instruments was that there were many detailed engineering tasks to be solved before the devices and systems could show true optical performance comparable to the highest grade research instrument. It was because of these reasons and also the cost of implementing an AOTF spectrophotometer, this new technology was not used for this project.

### 2.6.3 Non-dispersive

The most common non-dispersive system is a interferometric system. Many variants of these devices have been constructed but few have gained commercial significance. They are primary divided into two groups, two-beams and multiple-beams systems and they are commonly referred to Michelson interferometer and the Fabry-Perot interferometer. Since the Michelson interferometer is more used in the spectrometer, it is described briefly. The mechanism of Fabry-Perot interferometer is given in the Appendix A1.

The two-beam interferometer was designed in 1891 by Michelson (Michelson, 1891). The Michelson interferometer creates the conditions for optical interference by splitting light into two beams and then recombining them after a path difference has been introduced. The system uses two mirrors that are mounted perpendicularly to each other and between them is a beam splitter, fixed at an angle of 45°. The latter divides the light equally between beams and also serves to recombine them after they have been reflected back by the two mirrors. This is illustrated in figure 2.14a.

In this arrangement, half of the light is reflected back in the direction of the source and the other half can be collected by the detector. One mirror is fixed, whilst the other can be moved strictly parallel with the beam of radiation.
If the radiation source is a well-collimated beam of monochromatic light, the travelling distance from the beam splitter to the fixed mirror and back is $2 \times OA$. Likewise the distance to movable mirror and back is $2 \times OB$. Therefore the path difference between the two beams will be $2(OA-OB)$. This optical path difference is called retardation ($\delta$).

Zero path difference, or $\delta = 0$, occurs when $OA = OB$, at the point where both mirrors are equidistant from the beam splitter. The two beams interfere constructively because they are perfectly in phase and emerge with maximum intensity (see figure 2.14b). If the moving mirror is moved out by one-quarter of the wavelength, the
retardation is twice this, l/2. The two waves then interfere destructively and the intensity received by the detector will be zero (figure 2.14c). As the mirror is moved out again by a further l/4 the wavetrains will combine constructively again to give another maximum intensity (figure 2.14d).

The linear motion to the moving mirror will cause a sinusoidal intensity variation at the detector. The corresponding electrical frequency will be proportion to the velocity of the mirror. If light of another wavelength is sent through the system it will be modulated in the same manner, but the electrical frequency will be different because the distance traveled between successive maxima will be different.

For the described monochromatic radiation, a sine wave was obtained. In practice, the radiation source will almost always be polychromatic and the resulting interferogram will be a sum of sine waves, whose individual amplitudes represent the intensities of corresponding spectral elements. A typical interferogram is shown in figure 2.14a and a frequency decomposition, by taking a Fourier transform, is required to represent the spectrum. The Fourier when implemented numerically in a digital computer the fast Fourier transform (FFT) algorithm is used (Brigham, 1974).

In theory a perfect interferometer can provide very high spectral resolution. It also measures all spectral elements simultaneously allowing the light to be imaged onto one detector. In practices, there are limitations due to optical alignment and scanning, phase error, beam divergence and dynamic range, good details to these problems are given by Griffiths and de Haseth (1986).

The reasons a Michelson interferometer was not selected for this project was because it is ideal for resolving high spectral information but can be photometrically poor on weak absorption features. Therefore, if only a limited number of spectral signals need to be measured then most of the advantages of the interferometer will be wasted. In the NIR spectroscopy, the spectral response from species are broad due to overtones and combination vibrational effect, and also since the absorption in NIR are generally weak.
2.6.4 Non-thermal source

The previous described methods have selected or modulated energy produced by a broad emitting source. However, there are other types of emitter, classified as non-thermal, that emit radiation from a much narrower range of wavelengths down to individual emission lines (i.e. close to single wavelength). Example of these are discharge lamps, light emitting diodes (LED), laser diodes and lasers. The latter three are valuable NIR sources.

The advantage of the non-thermal source is their high energy efficiency; most of the energy consumed appears as emitted radiation over a narrow range of wavelengths. Power consumption is much reduced lowering thermal dissipation, and compact, even battery power portable instruments are possible.

For recording of NIR spectrum, the well established tuneable dye laser can be employed but it only provides a limited range of the NIR spectrum emission. More recently, F-centre lasers have been introduced. These are made with heavy alkali-halides, doped with lithium or sodium. They operate similarly to dye lasers and work over 0.6 to 4μm, having a high power output of the order of watts. One of the drawbacks with tuneable lasers is that the device has to be cooled. The use of lasers is still novel so experience and device reliability has to be improved. Generally, laser spectrophotometers have not yet made an impact in the field of NIR absorption measurements.
2.6.5 Sample presentation

In most of the traditional laboratory spectroscopy analysis of chemical or biological samples, glass cells are used to hold the sample for measurements (see figure 2.11) and this is the method adapted in the research work presented here. In monitoring of process, this requires bringing the measuring sample from the process into the instruments. Often samples are pumped from the bioreactor or process stream and samples are loaded into the measuring instrument using a flow-cell, this is an at-line system. In the work presented here, however, the samples were collected manually.

The use of fibre optic probes has made remote measurement possible. Instead of bringing the sample into the light absorption measurement, light is taken from the instrument to bioreactor or process stream via a fibre optic probe. After interaction with the measuring sample, the modified light spectrum is forwarded to the instrument by another fibre. Again this is a non-invasive technique and this remote on-line method has the advantages that previously mentioned in section 2.4.2.

Another possibility using fibre optics is that several measurements can be multiplexed into a single NIR instrument. This can be achieved with a optical switch or known as multiplexer. An example has been demonstrated by Aaljoki et al.(Aalijoki et al., 1994) where an NIR instrument equipped with fibre optics and multiplexer was used for refinery process and pilot monitoring. There on-line measurements of reformate octane numbers, densities and total aromatics were carried out with one NIR spectrophotometer. Automation of the instrument used in this project for remote measurement or for monitoring several process stream may be developed at a later stage, but has not been attempted in the present project.
2.7 Conclusion

This chapter has described the alcohol dehydrogenase (ADH) recovery process and how NIR techniques in monitoring of contaminants can be used to support the control of flocculation. A survey on NIR monitoring system used in bioprocesses and pharmaceutical areas has been presented. The theory of NIR spectroscopy and the different types of NIR instruments available was also given. The following are the conclusions to each of these topics.

2.7.1 NIR monitoring system to enhance flocculation control and hence optimise early stage of ADH recovery process

- The ADH purification process from yeast homogenate can be divided into three stages; flocculation to enhance removal of cell debris, protein and most important of all are contaminants such as nucleic acids and lipids; fractional precipitation which removes most of the protein; lastly purification by HIC to yield product.

- It is known that contaminants such as lipids and nucleic acid cause an increase in viscosity in the yeast homogenate and reduce the effectiveness of centrifugation throughout the recovery process.

- The excess contaminants in the final stage of process stream can foul the column and the replacement can be very costly.
• Thus there is a need to monitor the activity of the contaminants and if contaminant level are excessive then the stream should be cut off from the expensive downstream process equipment.

• Current operations have been based on at-line and off-line measurements of cell debris and protein. Measurements of contaminants such as nucleic acids and lipids required tedious and long hours of assays. Thus the vital contaminant information is not available during process operations.

• The subsequent precipitation and HIC purification processes can benefit if the initial removal of contaminant by flocculation and centrifugation has been maximised.

• A flocculatant polyethyleneimine (PEI) has been used to produced flocculation in the early stage of recovery. Depending on the concentration of PEI introduced to yeast homogenate, removal of contaminants can be maximised.

• The NIR monitoring system is introduced to measure contaminants after the first centrifugation. It has the potential to identify any excess contaminants as well as the PEI remaining in the process stream, and to alert subsequent processes.

• The contaminant levels measured by NIR can be utilised by the controller that operates on the PEI feed. The contaminant information allows the controller to adjust the flocculant used. Therefore a closed loop control of the flocculation process can be formed, allowing optimisation of the whole flocculation process.

• Eventually, NIR monitoring can also be used in the precipitation stage of purification, giving further optimisation of protein removal.
2.7.2 NIR monitoring in some of the relevant industries

- The survey has shown that the traditional use (food and agriculture industries) of NIR has been extended to many other areas. There are uses and active research in NIR monitoring in both pharmaceutical and bioprocess operations.

- In pharmaceutical production, NIR monitoring has been used for raw material identification, thus ensuring substances used are at their required standard. Since the majority of these products are expensive, it is necessary to make every effort to safeguard these processes.

- NIR absorption and diffusion reflectance measurements have also been used for quantitative analyses of products, for example, to determine the concentration of aspirin in the tablets within original package.

- NIR offers quick measurements without any sample pretreatment. There are ongoing researches in the use of NIR for pharmaceutical industries but one of the requirements is to obtain the FDA approval of this new analytical technique. It has been predicted that NIR will become a predominant methodology in the quality control lab within three to five years.

- The survey has also shown uses of NIR around bioreactors. These bioreactors often have long hours of operation and require constant monitoring of many key bioprocess parameters, i.e. measurements of carbon and nitrogen nutrients, byproduct and biomass.

- NIR monitoring has been used in fault detection, for instance in detecting a pH controller failure during process operation.
• At-line measurements using the NIR technique can produce accurate results in a very short time which traditional wet chemical assays have difficulty to achieve. These results assist control of expensive fermentation and improve product yield.

• On-line measurements can be benefit from use of a fibre optic probe. This offers non-invasive measurements, and the measuring instrument can be isolated from process environment. This reduces any risk of an operator working in a hazardous location.

2.7.3 The theory of NIR spectroscopic techniques

• The near infrared spectrum has been defined from 700 to 2500 nm. The phenomenon of infrared absorption has been described by a basic wave equation and the harmonic oscillation of molecule vibration characterised by quantum mechanic models.

• Spectral absorption in NIR is due to overtones (anharmonic oscillation) and combination bands of C-H, N-H and O-H vibrations observed in the longer wavelength of the spectrum (mid-infrared, 2500 to $5 \times 10^4$ nm). These absorptions are less intense than the fundamental bands in mid-infrared and are also difficult to assign directly to functional groups of the sample as can be done with the fundamental bands.

• The response in the NIR spectrum contains highly overlapping absorptions which are complex to interpret.

• These disadvantages, however, become advantages in certain quantitative applications. The weak absorptions are linear over a wide dynamic range and samples can be measured in reflectance or transmission without any dilution, hence eliminating the use of solvent or any other agents.
• NIR is sensitive to both chemical and physical effects and provides rich information that is important for measuring process performance. NIR is therefore particularly well suited to at-line and on-line measurements.

2.7.4 NIR instruments

• There are fundamentally three different types of NIR instruments; dispersive, non-dispersive and non-thermal systems.

• Most dispersive systems are based upon diffraction grating. A monochromator is used to separate white light radiation into a narrow line spectrum and a detector is used to record the separated line spectrum, a spectrum is recorded by scanning of all separated lines. This spectrum separation process can either take place before or after the interaction with the measuring sample.

• A combination of diffraction grating and diode array detector has the benefit of reducing moving parts in NIR spectrophotometer. This is particularly welcome because it offers near instant spectrum recording and better reproducibility (see chapter 4).

• An alternative emerging dispersive instrument is based on acousto-optically tuneable filter (AOTF). The output diffracted light is controlled by low radio frequency drive signal into a transducer (TeO₂ crystal). The primary advantage of AOTF optics is that the wavelength is electronically selected without the delays associated with mechanical scanning monochromators. Also it offers a large aperture and allows higher light throughput for detection.

• Non-dispersive systems are commonly known as Fourier Transform-Infrared or Near infrared (FT-IR, FT-NIR). These are based on the Michelson interferometer that uses a moving mirror to produce an interferogram. If the source of energy is
monochromatic, the interferogram will be sinusoidal. If the source of energy is polychromatic, the interferogram will be complex. The interferogram is Fourier transformed in a computer and the transmission spectra can be calculated. The process is fast and records the complete spectrum. Although, it is ideal for resolving spectral information it can be photometrically poor on weak absorption features such as those found in the NIR region.

- Non-thermal source devices are discharge lamps, light emitting diodes, laser diodes and lasers. The latter three offer valuable NIR sources and they all emit a narrow range of wavelengths. Tuneable lasers are also becoming available but these are still expensive and reliability has yet to be improved. Therefore they are less used in NIR spectroscopy applications. However, the focused laser energy would certainly be appropriate in measuring biological substances, particularly around a bioreactor where high energy throughput is required to penetrate across the long optical pathlength of the reactor.

- Of the three optical systems described, dispersive instruments are easiest to implement, but until wider support of AOTF commercial spectrophotometer manufacture is available, the combination of diffraction grating and diode arrays detector systems is still very successful.

- Sample presentation has always been another important issue in NIR measurements. Traditionally for biological materials, glass cells are used for off-line measurements and flow-through cells are used for at-line purposes. These generally have on optical pathlength of 1 to 10 mm. An alternative approach is the use of fibre optics for both transmission and diffusion reflectance measurements. This is much preferred because the complete measurement process is non-invasive and provides the opportunity to process information online.
3. Near Infrared Calibration

3.1 Summary

In near infrared spectroscopy, the NIR instrument requires calibration before it can be used for quantitative measurements. There is a variety of calibration methods used for near infrared spectroscopy techniques which all aim to overcome the same basic difficulties. The main difficulties are due to the complex NIR spectra, in which any peak of interest is generally overlapped by one or more interfering peaks, and the strong dependence of absorption on the scattering properties of the sample, particularly on the particle size variations. Some of these problems are highlighted at the beginning of this chapter.

The variety of approaches to calibration stems from its empirical nature. This chapter will first show the simplest form of multivariate calibration by multiple linear regression. The complications found in this calibration are presented, emphasising the need of multivariate calibration based on principal component analysis (PCA). For this reason, the partial least squares (PLS) calibration method has been adapted for the research work here. The description of PLS is given here after the presentation of PCA and principal component regression (PCR). This is because PLS is an improved version of PCR.

In this chapter, the method of validation of multivariate calibration is also given. And the measures of confidence in PLS used in practice are derived.
3.2 **Introduction to the calibration experiment**

A set of calibration samples should be collected first. These should be representative of the population of samples which it is desired to analyse in future with the instruments. The samples are then analysed by the reference method for the constituent of interest and by the NIR instrument. For example, to calibrate for protein in a chemical sample, a set of twenty samples with a range of protein contents previously determined by assays might be used. These samples are also presented to a near infrared (NIR) spectrophotometer and its absorbance spectra are recorded for each sample. The protein contents form an independent or reference data matrix (Y-data) and the spectra construct a data matrix (X-data). The purpose of calibration phase is to produce a model that relates the NIR spectra to the values obtained by the assays reference method (see figure 3.1). In the prediction phase, the NIR spectrum of unknown samples (X-data) are taken as the input (see figure 3.2), then the calibration model predicts the unknown protein contents (Y-data) based on the X-data.

![Figure 3.1](image)

*Fig 3.1 During the calibration stage, a calibration model is established based on a representative set of corresponding measurements of both X- and Y-variables.*

![Figure 3.2](image)

*Fig 3.2 Once the calibration model has been established, this model can be used to predict (unknown) Y-values (concentration values) from the measurements of new X-variables (absorbance spectrum).*
### 3.3 Practical needs for multivariate calibration

In this section, the problems of univariate calibration are highlighted. These are problems primary due to interference which affects spectral data. This may be significantly overcome by using multivariate calibration. The following are classical problems found in traditional chemical analysis with used of univariate calibration, these are simple illustrations used by Marten and Næs (Marten and Næs, 1989).

In traditional wet chemical analysis it is often difficult to obtain ideal measurements (figure 3.3), it is even more difficult with biochemical analysis. In the ideal chemical measurement, a sample is analysed in a high-precision instrument producing a measurement that is linearly related to the concentration of the analyte and not affected by other analytes in the measuring sample.

In such circumstances, calibration would be simple using a univariate calibration equation (section 2.5.2.4) or multiple linear regression (section 3.4).

![Figure 3.3 The ideal calibration arrangement](image)

In practice, however, the analysis may be affected by chemical and physical interference in the samples themselves and experimental interference arising in the measurement process. Examples of interferences are given in section 3.3.1 and 3.3.2. In addition to the usual random measurement noise, by non-linearities often create...
additional problems. Thus if there is a calibrated model, the accuracy of measurements would rely heavily on the minimum interference stated (see figure 3.4).

![Diagram](image)

Fig 3.4 The instrumental measurement is not selective for the analyte, and the instrument response is non linear.

Conventionally, the interference has to be removed physically, to ensure linearity to the instrument scales (figure 3.5). But in many cases this is a expensive process or physically impossible. An example is the total removal of cell debris or nucleic acids from yeast homogenate (section 5.3.2). Sometimes one is interested in quantifying several to the mutually interfering constituents, making the removal of interesting constituents illogical. Traditionally, these problems maybe improved by cleaning, standardising and diluting each sample (symbolised by a filtering stage in the figure below) before measurement of $x$.

![Diagram](image)

Fig 3.5 Traditional selectivity enhancement prior to measurements.
However, calibration may still be limited to a narrow linear range. Thus when using the calibrated model for further measurements of unknown sample, the data from the new samples should lie near the linear range.

Using multivariate calibration, interference and individual non-linearities properties represent less of a problem. They are most easily dealt with if their influence on the instrument response is known in advance and physical removal of their influence is not necessary. Their effects are instead modelled mathematically, using data from representative calibration samples with sufficient variability (figure 3.6).

Fig 3.6 Selectivity enhancement by multivariate calibration.

The cleaning, standardising and dilution is replaced by mathematical modelling of multichannel measurements $x_1, x_2, x_3, \ldots$ (i.e. absorbance measurements at several wavelengths, as shown in figure 3.6). This process removes interference effects and extends the linear range of the calibration.

It is the intention in the following sections to discuss the above interference. These will briefly address the issues of chemical and physical interference, and also mistakes and unexpected errors. The work reported later will illustrated the problems arising in the first two interferences, while the latter have also shown up in the laboratory work included here as would occurred in almost any lab practice.
3.3.1 Chemical interference in samples

Chemical interferences are described as a form of systematic error in the quantitative determination of a certain analyte. These errors are due to the presence of other chemical constituents which influencing the instrument response of the desired analyte. The majority of chemical samples and especially those of biologicals are comprised of mixtures of several different chemical constituents. Many of the constituents may be sufficiently similar to provide instruments responses that overlap. This is especially true in the NIR spectral region where absorptions are derived from combination and overtones and therefore overlapping of spectral responses are expected.

3.3.2 Physical interference in samples

Physical interferences in the samples are systematic errors caused by physical effects in the quantitative determination of a chemical constituent. Irrelevant physical phenomena in the samples can affect the measured signal strongly. To get the desired measurement, the physical effect can either be kept constant or compensated mathematically.

An example of physical effect is turbidity in the sample, i.e. biomass found in the biochemical samples. Traditionally, turbidity would normally be removed by filtration or centrifugation prior the measurement. With multivariate calibration the physical interference can instead be corrected for mathematically and even modeled. Hence there is no need for sample purification prior to the measurements.

Another example of physical interference is the effect of temperature to the measuring samples. Unless the temperature is the same for all samples, or the effect of temperature is compensated, then the prediction of chemical composition in biochemical or chemical application from the NIR spectra would be imprecise. In
multivariate calibration it is possible to correct for temperature effects, by using sample temperature as an extra predicting variable or by letting the calibration program model its effect as another 'unidentified interference' from the spectra themselves. In the work presented here, the latter modelling approached is used but also the temperature of the samples was maintained constant by exposing the samples to ambient room temperature in an air conditioned room at a fixed period before measurements.

3.3.3 Mistakes and unexpected errors

There are many other possible errors that cannot be estimated at the time of first calibration. This might be gradual or abrupt instrument drift over time. For example, changing hardware components in the instruments such as the spectrophotometer lamp source can lead to effects that cannot be removed completely by standardising the instrument. In complicated cases univariate calibration cannot correct for such effects, but by multivariate calibration it is possible to build a 'bridge' across such discontinuities.

Human mistakes and even fraud should also be taken into consideration. No matter how conscientious the analyst is, there is always a probability of unexpected errors, either due to an unexpected interference or unnoticeable instrument trouble. With multivariate calibration many types of sample anomalies or instrument problems can automatically be detected as outliers. Human mistakes and fraud can also be revealed by these methods.
3.4 Partial least squares (PLS) calibration based on principal component analysis (PCA)

In the previous section, it was suggested that improved calibration and robust prediction are accomplished by use of multivariate data. NIR spectroscopy is able to provide vast amount of spectral information by scanning through many wavelengths. A simple use of multivariate data for calibration is the multiple linear regression (MLR) technique which is described below (section 3.4.1). However, it is necessary to ensure only the relevant information in the spectral data is correlated to the desired analyte concentration. Since the spectrum information may be use to account for system variation (e.g. cause by noise or drift in the measuring instruments), the calibration model must have the ability to correlate spectral information for analytes and discriminate against other interference. Because MLR is not ideal to satisfy the requirement in practical calibration modelling, an alternative multivariate calibration technique such as partial least squares (PLS) is used.

Partial least squares (PLS) regression is a multivariate data analytical technique designed to handle intercorrelated regressors. It is based on H. Wold’s general PLS principle (Wold, 1983), in which multivariate systems analysis problems were solved by a sequence of simple least squares regressions. It was first applied successfully to NIR data by H. Martens and S.A. Jensen (Martens and Jensen, 1983). This technique is based on principal components analysis which has the ability to reduce spectral variables into a few basic profiles called principal components. Important spectral information is capture by the principal components, and their correlation to the analyte yields a better calibration model.

In later subsection, principal component regression (PCR) is discussed prior the partial least squares (PLS) method. PCR provides the basic concept of calibration
model using a principal component regression technique. The analogy to partial least squares technique is an extension of PCR, therefore it is constructive to derive PLS by understanding of PCR.

3.4.1 Multiple linear regression (MLR) and its problems

In multiple linear regression, $Y$ is estimated from the linear combination of variables in $X$ which minimizes the errors in reproducing $Y$. The relationship of analyte concentration $Y$ with the NIR spectral response $X$ can be written as follows.

\[ Y = B X + E \quad [3.1] \]

where $B$ is a matrix of regression coefficients and $E$ is a matrix of errors associated with the MLR model. The regression coefficients $B$ is estimated by linear regression, and the error given by equation 3.2, 3.3 is minimized.

\[ \text{Err} = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \hat{y}_{ij})^2 \quad [3.2] \]

\[ \text{Err} = \sum_{i=1}^{I} \sum_{j=1}^{J} \epsilon_{ij}^2 \quad [3.3] \]

In the above equation, $y_{ik}$ is the actual concentration data in the $i^{th}$ row and $j^{th}$ column of $Y$ matrix, and $\hat{y}_{ij}$ is the MLR estimated concentration using the $B$ coefficients, and $\epsilon_{ij}$ is a element in the $E$ matrix.

However, the MLR approach is only ideal if the model is applied to a well behaved system. This implies that system has linear response, does not suffer from interference, no analyte-analyte interactions, low noise and data matrix is not collinear. Since the previous sections have pointed out that majority of these requirements are unlikely to occur in practice, thus problems will be encountered with application of the MLR method to NIR spectra. This is because a MLR model
attempts to use all of the variance in the $X$ matrix to correlate to $Y$ matrix, thus including duplicate information to model $Y$. If the model is applied to a new sample, the model will assume that the correlation established between the calibration $X$ and $Y$ matrices also exits in that new sample. But if the model was built using irrelevant information in the $X$ matrix, the assumption would not be correct to all cases.

An example of this problem in MLR calibration and prediction is given in Appendix A.2. This demonstrates that during calibration, the model incorporated irrelevant information and subsequent predictions produced misleading analyte results. It was for this reason and also because the nature of spectral data ($X$ matrix) gives a highly collinear matrix, a partial least squares (PLS) technique based on principal component analysis of the $X$ matrix was used.

### 3.4.2 Data reduction with principal components analysis (PCA)

It has been pointed out by D.A. Burns and E.W. Ciurczak (Burns and Ciurczak, 1992) that the initial mathematical foundation for what is now called principal components analysis (PCA) can be traced as far back as 1901 given by K. Pearson, then further developed into the modern use PCA by H. Hotelling in 1933 (Hotelling, 1933). Some applications to NIR analysis are found in work of I.A. Cowe (Cowe, 1985, 1988), H. Mark (Mark, 1986), H. Martens (Martens, 1988) and there are many others.

Before the report of principal components analysis, it is necessary to introduce important concepts. The following illustrates a matrix representation and the concept of projection of vectors, then the use of PCA in spectroscopic data ($X$ matrix) is described.
3.4.2.1 Matrix representations and projection

In a low dimensional matrix where the number of rows and columns are small, graphical representation of the matrix can be simple. For example the following $2 \times 3$ matrix $X$ can be represented graphically as $X$ in row space or as $X$ in column space.

$$X = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 2 & 5 \end{bmatrix}$$

Row space is the space formed with the rows of $X$ as the axes (figure 3.7a) and hence it is two dimensional. In figure 3.7a, three points representing a corresponding column in the matrix in this two dimensional space. Similarly, the column space (figure 3.7b) has two points in the three dimensional space.

Projection is an important concept used in principal component method. The projection is of either a point or a vector onto a vector or plane. Each of these can be

Fig 3.7 a) The matrix $X$ shown in row space, b) The matrix $X$ shown in column space.
seen as being the perpendicular shadow of one object onto another. The illustration in figure 3.8, the result of projecting a vector y co-ordinate by 3 dimension on a plane of 2 dimension.

![Diagram](https://via.placeholder.com/150)

Fig 3.8 Projection of y down on the space spanned by the vectors $x_1$ and $x_2$

### 3.4.2.2 Factors and principal component analysis (PCA)

The concepts of factors forms the basis of partial least squares (PLS) and principal component regression (PCR) model construction. These factors, also known as the principal components, are defined as any linear combination of the original variables in $X$ or $Y$ matrix (Sharaf, 1986). It can be shown that given $J$ factors for $I \times J$ matrix $X$, one can also represent the variables in $X$ as a linear combination of these same $J$ factors. It is important to determine factors that represents useful (analyte response) and also small features (such as noise components) in the $X$ matrix and model can be based on significance of factors. For example, the information in the $I \times J$ matrix $X$ can be expressed as $I \times J$ matrix $X^{\text{new}}$, where the columns of $X^{\text{new}}$ are linear combinations of the original columns in $X$. The primary advantage is that if a particular column in $X$ is not useful, then small weight is given to the factor representing that column. Similarly, if the column in $X$ is useful then the factor will be given a larger weight. The following examples uses a simple $5 \times 2$ matrix $X$ to illustrate the role of principal components.
In principal component analysis, the columns of $X$ are often mean-centred and scaled. The process of mean-centring requires subtracting the average of the column from each element of that column. The scaling provides equal weights to each variable, achieved by dividing each element by the variance of the column. The PCA is then performed on the covariance matrix $X'X$ formed from the mean-centred and scaled $X$ matrix. The first eigenvector corresponding to the largest eigenvalue is, by definition, the direction in the space defined by the columns of $X$ that describes the maximum amount of variation in the samples. This is represented in figure 3.9, the data and the direction of the first eigenvector where the space defined by $X$ is a plane (2 dimensional).

$$X = \begin{bmatrix}
2 & 4 \\
1 & 2 \\
0 & 0 \\
-1 & -2 \\
-2 & -4
\end{bmatrix}$$

Fig 3.9  The matrix $X$ is plotted in column space with illustration of the first eigenvector.
In this illustration, all of the variation in the data has been described by one eigenvector. The samples all fall on a line in the column space of $X'X$, thus all of the variation lies in one direction. With PCA, if all of the variation in the samples cannot be explained by one eigenvector then a second or subsequent eigenvector will be found. The second eigenvector is perpendicular or orthogonal to the first and this describes the maximum amount of residual variation that was not described by the first eigenvector. In figure 3.10 is another demonstration of projecting a $20 \times 2$ data matrix $X$ onto two eigenvectors in $X'X$. The direction of the first eigenvector describes the maximum amount of variation in the data set. The samples in column space fall within a two dimensional clusters, and the first eigenvector has described to the major axis of the cluster.

![Image](image.png)

**Fig 3.10** A $20 \times 2$ matrix $X$ is plotted in column space with illustration of the first and second eigenvector of $X'X$, the eigenvectors are orthogonal to each other.

The first factor or principal component is a linear combination of the original variables and can be written as follows;

$$Xp = t \quad [3.4]$$
where $X$ can be the spectroscopic data, $p$ is the first eigenvector of $XX$, and $t$ is the score vector, which is the projection of $X$ onto the first eigenvector. The score is a principal component for the columns of $X$ because it can be written as linear combination of the columns. If $J$ eigenvectors are calculated and used to form an $I \times J$ matrix $T$ of scores. Then the original variables in $X$ can be expressed as the linear combination of the scores;

$$X_{PP'} = TP'$$  \[3.5\]
$$X = TP'(PP')^{-1}$$  \[3.6\]

where $P'(PP')^{-1}$ resulted from linear algebra is called the generalised inverse (GI) of $P$. In principal components analysis, the element in the GI of $P$ are called the loading in principal component. The loadings have numerical range from -1 to 1 and are the cosines of the angles between the eigenvector and the variable axes (Marten and Næs, 1989). For example, if the eigenvector is orthogonal or near orthogonal to the variable, then small loading is obtained [$\cos(90°) = 0$]. On the other hand if the angle between the eigenvector and the variable is zero or near zero, then loading would be maximum since $\cos(0°) = 1$.

In the example used in figure 3.9, the two eigenvectors lie in the same space as the original variables, they can be used as a new set of axes for the matrix $X$. The new axes also can be viewed as a rotation of the original axes, i.e. the transfer from the old to the new axes can be achieved by rotating the original axes. These new axes are more useful for describing the variation of the samples.

So far, all the illustrations have been with a matrix that can be represented in a two dimensional space. In more complicated situation the matrix can be of higher dimensionality than two and it is also certainly the case for spectroscopic data where $X$ matrix contains hundreds of columns of data (see figure 3.11). In such a case, the process of determining eigenvectors continues until all of the variation in the samples is described. The maximum possible dimensionality of the samples equals the number of columns in $X$. However, it often occurs in spectroscopic data that fewer dimensions are required. This is because there is collinearity between the variables.
of spectroscopic data. This collinearity means that some of the rows in the original matrix are correlated to the other rows and therefore contain redundant information. In this cases, PCA can describe the variation in \( X \) using fewer dimensions than the number of columns in \( X \).

\[
X = \begin{pmatrix}
\text{Wavelength} \\
\vdots \\
\text{Spectrum no.}
\end{pmatrix}
\]

Fig 3.11 Spectral data of multiple samples presented in a matrix format, each row in the matrix is one spectrum and each column is the spectrum data at one particular wavelength.

PCA abstracts influences and relevant information in a given matrix (space) and describes key relationships by principal components (subspace). This can be visualised by considering two balloons floating across a room. The column space of the room is three dimensional. However, the relative position of one balloon to the other can be described by two dimensional information. This characteristic of the principal component analysis is exploited to yield more informative models in PCR and PLS, as are described in the following.

### 3.4.3 Principal component regression

In calibration model building, an attempt is made to correlate a set of analyte concentration data (\( Y \) matrix) to columns of spectroscopic variables (\( X \) matrix). The first step is to determine the eigenvectors or principal components for the \( X \) matrix as the eigenvectors of \( X'X \). The aim is to redefine the variables using a small number of principal components where the first principal component is the linear combination of the original variables that points in a direction that is best correlated to all of the columns in row space. It is also the direction in column space that best describes the variation in the samples.
In principal component regression (PCR) the $X$ matrix is re-expressed in term of a much smaller score matrix $T$ which results from projecting $X$ onto the eigenvectors. A generalised $P$ loading is obtained, see equation 3.7. The $E$ matrix is the residual of the unexplained data variation by the principal components. The second step of PCR is to regress the $Y$ matrix onto the score matrix as $TQ' = Y$. The loading $P$, and the matrix of regression coefficients, $Q$, together form the PCR model (see equation 3.8).

$$X = TP' + E \quad \text{[3.7]}$$

$$Y = TQ' + F \quad \text{[3.8]}$$

In PCR only the significant variation in the $X$ data, as captured in $T$, is used in regression. In MLR, by contrast, the whole of the $X$ matrix is used (equation 3.1). Note, however, that for an $I \times J$ matrix $X$ where all of the $J$ eigenvectors are used to form $P$, PCR will yield results identical to MLR.

The advantages of PCR lies in its data compression process by concentrating the information in $X$ into fewer principal components than $J$. The first principal component is the most informative and each subsequent principal component explains less of the information contained in $X$. In calibration model, some minor principal components may be eliminated from the model because they are mostly describing the noisy information of $X$ matrix, and have no significance to the concerned analyte ($Y$ matrix).

The above descriptions show the advantages of PCR over MLR in many situation. The method using principal components allows data compression which avoids collinearity problems, and a calibration model based on regression of $X$-$Y$ matrix from relevant information in the principal components thus has the ability to discriminate against noise.
3.4.4 Partial least squares

The method behind PLS is similar to that in PCR. However the modelling procedure has simultaneously estimates the factors in both \( \mathbf{X} \) and \( \mathbf{Y} \). These factors are used to define a subspace in \( \mathbf{X} \) that is suited to model \( \mathbf{Y} \). In PCR, the rotation defined by the eigenvectors was used to find a subspace in \( \mathbf{X} \) that subsequently was used to model \( \mathbf{Y} \). In PLS, the principal components are chosen to describe the variables in \( \mathbf{Y} \) as well as in \( \mathbf{X} \). This is achieved by using the columns of the \( \mathbf{Y} \) matrix to estimate the principal components for \( \mathbf{X} \). At the same time, the columns of \( \mathbf{X} \) are used to estimate the factors for \( \mathbf{Y} \). The resulting models are

\[
\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \quad [3.9]
\]

\[
\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} \quad [3.10]
\]

where \( \mathbf{T} \) and \( \mathbf{U} \) are the scores of \( \mathbf{X} \) and \( \mathbf{Y} \), respectively, and the \( \mathbf{P} \) and \( \mathbf{Q} \) are the loadings. As equation 3.7, 3.8, \( \mathbf{E} \) and \( \mathbf{F} \) in the above are the residuals associated with modelling \( \mathbf{X} \) and \( \mathbf{Y} \) in the PLS model.

In PLS, the \( \mathbf{T} \) factors are not optimal for estimating the columns of \( \mathbf{X} \) as they were in PCR, but instead are rotated so as to simultaneously describe the \( \mathbf{Y} \) matrix. The error \( \mathbf{E} \) is larger in PLS than in PCR, but the total error expressed in both \( \mathbf{E} \) and \( \mathbf{F} \) is smaller. In the ideal situation, the sources of variation in \( \mathbf{X} \) are exactly equal to the sources of variation in \( \mathbf{Y} \), and the factors for \( \mathbf{X} \) and \( \mathbf{Y} \) are identical. In practice, \( \mathbf{X} \) varies in ways not correlated to the variation in \( \mathbf{Y} \), and therefore \( \mathbf{t} \neq \mathbf{u} \) where \( \mathbf{t} \) and \( \mathbf{u} \) are factors for \( \mathbf{X} \) and \( \mathbf{Y} \) matrices respectively. However, when both matrices are used to estimate factors, the factors for \( \mathbf{X} \) and \( \mathbf{Y} \) matrices have the following relationship.

\[
\mathbf{U} = \mathbf{B}\mathbf{T} + \mathbf{\epsilon} \quad [3.11]
\]

where the \( \mathbf{B} \) gives the inner relationship between \( \mathbf{U} \) and \( \mathbf{T} \) and is used to calculate subsequent factors if the intrinsic dimensionality of \( \mathbf{X} \) is greater than one.
The relationship given in equation 3.11 is then used to describe the model for $Y$. Instead of using equation 3.10, equation 3.11 maybe substitute into this expression to give:

$$Y = b' T Q' + G$$

[3.12]

where $G$ is the residual unexplained model.

Therefore, the analysis of data using PLS can be summarised as the determination of factors in $X$ and $Y$ using all of the information available. The final PLS model consists of the score matrices $T$ and $U$ that are linearly related with a coefficient, $B_f$, describing the relationship for each of the $L$ factors.

In prediction, the model is used as follows. The spectrum of the samples ($X$ matrix) is known. $P$ is known, enabling calculation of the score matrix $T$. Thus, since $B$ and $Q$ are known from the calibration model, the unknown concentration $Y$ can be determined.

All the multivariate calibration and prediction used in the work reported in this thesis are performed in a commercial software package Unscrambler (Unscrambler, Camo A/S, 1994). The algorithms used in Unscrambler for PCA, PLS are described by H. Martens and T. Næs (1989) and these are listed in Appendix A3.
3.5 Prediction ability, validation and the number of principal components used in calibration model

This section defines the prediction error in the multivariate calibration model and the concept of prediction ability estimated by confidence intervals is explained. The technique of cross validation is used during all calibration model building, this method is described in comparison with external and internal model validation. Lastly, there is a discussion of how to determine the optimum number of principal components used in a calibration model. These considerations ultimately will set the quality of the calibration model for future prediction.

3.5.1 Prediction ability

The prediction ability should give indication of how well the calibrated model can predict future unknown samples. In the following section, the expressions that are often used to defined the error in prediction are given (Martens and Næs, 1989). These provide an estimation of the error in prediction, as well as being used in model validation and choosing optimum principal components for the calibration model.

Also needed is an understanding of how confidently one can rely on the prediction results. This is given by the confidence interval, also known as the uncertainty limits in predictions. The method chosen for this estimation is provided by commercial multivariate calibration software (Unscrambler, 1994) which has been used throughout all calibration model reported in this thesis.
The prediction error can be defined in many different ways. The most common statistical expression is the mean square error defined as follows.

$$\text{MSE} = \frac{\sum_{i=1}^{I_{pr}} (\hat{y}_i - y_i)^2}{I_{pr}} \quad [3.13]$$

where $\hat{y}$ is the predicted $y$ and $I_{pr}$ is the total number of predicting samples.

It can be expected that the error of prediction is directly related to the model error, i.e. to the residual $E$ and $G$ matrix in PLS, and estimation of MSE from these model error is though to be possible. However, the statistical theory has not yet been fully developed and such derivations of MSE are not available (Martens and Naes, 1989) except in some very simple cases.

Therefore, the estimation of MSE is commonly based on the real comparison of data $y$ and $\hat{y}$ from a limited number of samples from either the calibration set (internal validation) or from a separate test or prediction set (external validation). But this also involves statistical distribution considerations. For example, in order to obtain a good estimate of the average prediction ability, the set of test objects on which it is based must be representative for the whole population of future unknown samples, otherwise the estimated MSE can be very misleading.

One alternative representation of prediction error is based on the absolute error ($|\hat{y} - y|$), rather than squared error. For non-statisticians, the squared prediction error MSE may be difficult to interpret physically (e.g. MSE gives units of 'squared Mmol' or 'squared O.D. units'). The square root of the estimated MSE may then be preferable because this is measured in the same unit as $y$ itself. The square root of MSE is denoted RMSE (root mean square error) or RMSEP (root mean square error of prediction).
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Chapter 3

RMSEP = \sqrt{\frac{I_p}{I_{pr}} \sum_{i=1}^{I_p} (\hat{y}_i - y_i)^2} \quad [3.14]

The next description of prediction error is known as the standard error of prediction (SEP) which is the unit used throughout the result presentation in this thesis. SEP is usually equal to or smaller than RMSEP, however, in cases of few prediction samples and small bias then the SEP may be larger than RMSEP due to different denominators.

\[
SEP = \sqrt{\frac{I_p}{I_{pr}} \sum_{i=1}^{I_p} (\hat{y}_i - y_i - \text{Bias})^2} \quad [3.15]
\]

\[
RMSEP^2 = SEP^2 + BIAS^2 \quad [3.16]
\]

BIAS is interpreted as the average difference between \( \hat{y} \) and \( y \) in the prediction set.

\[
BIAS = \frac{\sum_{i=1}^{I_p} (\hat{y}_i - y_i)}{I_{pr}} \quad [3.17]
\]

3.5.1.2 Confidence Intervals

Another way of assessing the quality of predictions rather than by MSE, RMSEP and SEP is by using confidence intervals for the unknown quantities. In the work reported here, the estimated confidence limits to each prediction are given by Unscrambler multivariate calibration software (Unscrambler, 1994). The estimation of confidence intervals given are dependent on:

- The model error: The Y-residual in the calibration process (E and G matrix)
- The X-measurements errors in the prediction samples
The number of calibration samples
It was pointed out by Unscrambler that the estimation technique of the confidence limit is not a standard formula that can be found in PLS or PCR theory, but an empirically found relationship that has given satisfactory indications on the uncertainty in predictions for large range of applications. It can be interpreted as weighted variance of both X residuals and Y residuals, together with some heuristic quantities that inflate the uncertainty prediction when the number of samples is small.

3.5.2 Validation of calibration model in practice

In multivariate calibration, the calibrated model can be validated in three different ways:

i. Internal validation

ii. External validation

iii. Cross validation

Internal validation concerns validation from the calibration data themselves and an assessment based on internal validation is not as good as using this calibrated model to predicted new unknown samples. Ideally, the predictive ability can only be assessed by testing on new samples. As internal validation uses the same data both for model building and testing, thus this may lead to serious under-estimation of data variation and over-estimate of its predictive ability. It is therefore not usually used for validation purposes.

In external validation, the calibrated model is validated by new testing samples external to that of calibration samples. Such external prediction testing is conceptually the easiest of the empirical validation methods to explain. But it is not the validation to be recommended in most routine calibration work, because it is usually rather wasteful and expensive since it requires a large and representative set of test samples in order to give relevant and reliable estimates of future prediction ability. It is unwise to use just a few of the samples for testing because the prediction
ability from few, imprecise or an unrepresentative test sample may give a meaningless result. Therefore it is more economical to use all the available data both for calibration and for prediction testing.

The complications in both internal and external validation has been overcome by a method known as cross validation (Stone, 1974; Snee, 1976; Wold, 1978). This method of validation is used throughout all PLS calibration in this research. The cross validation is a better internal validation method. Like the external validation approach it seeks to validate the calibration model on independent test samples, but contrary to the external validation it does not waste data for testing only.

In full cross validation one repeats the calibration $I$ times, each time treating one $j^{th}$ part of the whole calibration set as prediction samples. In the end, all the calibration samples have been used as prediction samples and the estimated MSE can be calculated. Since full cross validation is based on repeated calibration which my be somewhat time consuming for computation, an important alternative is to perform cross validation by segmenting the calibration set into $M$ (where $M < I$) segments and then calibrating only $M$ times, each time validating the model by $1/M$ part of the calibration set.

3.5.3 Using prediction error and selection of the number principal components in a calibration model

One important factor in the calibration model is how to choose the number of dimensions (principal components) to be used in subsequent predictions.

The aim of multivariate calibration with PLS is to reduce the prediction error by modelling chemical and physical interference phenomena that would otherwise destroy concentration determinations ($Y$ matrix) from spectral data ($X$ matrix). Each interfering phenomenon requires an independent principal component in the calibration model. The prediction error is composed of two main contributions, these
are interference error (underlying bias) and the estimation error (see figure 3.12). The former is the systematic error due to un-modelled interference in the spectral data and the latter is caused by random measurement noise of various kinds.

![Graph showing Underfitting, Optimum, Overfitting with Error of prediction vs Number of principal components used](image)

**Fig 3.12** Conceptual illustration of prediction error as function of number of principal components included in the calibration model, the arrow suggest the optimum point of the calibration model.

The two contributions to the prediction error illustrated in figure 3.12 have shown opposite profiles as the number of principal components increases. Provided that the calibration samples are sufficiently representation for the new samples to be predicted, the interference error should decrease with more principal components because increasing number of interferences are being modelled. However, the statistical uncertainty error increases at the same time, this is due to the increased number of independent model parameters estimated from the available data. This is an important observation because only a limited number of independent parameters can be estimated with high precision from a given set of calibration data. Use of too many principal components can lead the model to adapt itself to random error in the calibration set. In theory, the minimal prediction error is obtained when the remaining interference error and the uncertainty error balance each other. If the model has incorporated too few phenomena (principal components) it is called
underfitting and modelling too many phenomena is called overfitting (Martens and Naes, 1984).

In practice, the technique used to compute the optimum number of principal components used in the calibration model is either by prediction testing or internal validation for calibration data. In this way one can draw a graph of MSE on the basis of the number of principal components used in the calibration model. A suggestion is then to select the number of principal components which gives the lowest MSE. Alternatively, it is usually advantageous and more robust not to be too restricted. If for instance a similar value of MSE is observed at a number of principal components less than is suggested by the lowest MSE, then this number of principal components should be selected (Osten, 1988). That is, if the minimum in figure 3.12 is broad and flat the arrow for the optimum should be moved to the left. The effect is to trade off a slightly higher MSE value against a simpler model with fewer principal components. In all of the multivariate calibrated reported in this thesis, the optimum number of principal components used in PLS models have been selected based on the above suggestion.
3.6 Conclusion

This chapter has described the needs of multivariate calibration for correlation of spectroscopic data to analyte concentration. The chosen partial least squares (PLS) calibration technique that is based principal component analysis (PCA) was presented in detail. The practical aspects of measurement error in prediction together with the theory behind validation of the calibration model and the choice of the number of principal components used in the calibration model have been discussed. The following are the conclusions.

3.6.1 Needs of multivariate calibration

- An NIR instrument requires pre-calibration before it can be used to predict analyte concentrations for given NIR spectra.

- The complex overlapping signal derived from NIR spectrum make simple univariate calibration impossible. This overlapping property may be further complicated by chemical and physical interferences in the measuring samples. Therefore multivariate calibration has been used.

- The weak signal intensity occurring in NIR spectra may suffer from instrument noise. Simple univariate calibration does not separate the relevant information and noise thus a misleading result may be given from this calibration model.

- To improved the chance of success in measurement by the simple univariate calibration method, samples require some pre-treatment such that interferences are minimised. However, this is not often possible and the result may be limited to a narrow range of concentrations.
• As with any calibration, multivariate calibration requires a sufficient population of samples such that the analyte in these samples has a similar distribution of variation as in the future prediction samples.

• Multivariate calibration uses all available spectral data, and compensates for interference or system noise. It has the ability to provide better predictions for the future measuring samples.

3.6.2 Principal components analysis and partial least squares solutions

• Multiple linear regression (MLR) models Y (analyte concentrations) and X (spectral data) using a least squares criterion. All of the X variables are used to model Y.

• The success of MLR models for future prediction of Y from X depends on well behaved systems, i.e. linear responses, no interfering signals, no analyte-analyte interactions and low noise. But in practical situation, these well behaved systems are unavailable and in NIR spectroscopy signals are often overlapping and hence interfering each other. Thus it is impractical to use MLR for NIR calibration modelling.

• An MLR model depends on the inversion of the covariance matrix X'X, but the spectroscopic data matrix is often collinear. Therefore the inversion may be impossible or mathematically unstable to perform.
The decomposition of $X$ into principal components by principal component analysis (PCA) has the advantage that most variations in the spectra in $X$ are described by the first few principal components. The higher order principal components may describe only noise. Thus in a principal component method these uninformative principal components are deleted for model building.

A calibration model based on principal components uses a score matrix $T$ and some factors. The score indicates which spectra in the $X$ matrix are responsible for most of the variation in the $X$ data matrix. The factors are linear combinations of the original variables in the $X$ matrix. It can be shown that given $J$ factors for $I \times J$ matrix $X$, one can also represent the variables in $X$ as a linear combination of these same $J$ factors.

It is important to determine factors that represent useful (analyte response) and also small features (such as noise components) in the $X$ matrix and the calibration model can be based on significance of factors.

Principal component regression (PCR) first models $X$ by a score matrix $T$, then regression of the score $T$ and analyte concentration $Y$ is achieved.

Principal components are orthogonal to each other, thus the problem of collinearity is eliminated. This also implies that fewer factors are used to describe $X$ matrix.

The partial least squares (PLS) method is an improvement over PCR for regression model building. It differs from PCR by using the variation in the $Y$ matrix during the decomposition of $X$ matrix into principal components. By balancing the $X$ and $Y$ information the method reduces the impact of large, but irrelevant $X$ variations in the calibration modelling.
3.6.3 Prediction ability, validation method and choosing calibration model

- In prediction, the prediction error can be given by the mean square error (MSE) which is often used by statisticians because the square features give relatively higher emphasis to large errors than to the smaller errors.

- The root mean square prediction (RMSEP) and the standard error of prediction (SEP) are also used. Generally, these are easier to adapt than MSE because these have the same units as those in the predicting values, i.e. error in terms of 'Mmol' or 'g L\(^{-1}\)'. The SEP is approximately equal to RMSEP unless there is a model bias, and SEP is used as a measure of error of prediction in the rest of the work reported.

- The confidence interval defines the confidence boundary for the prediction results. There are different ways of deriving this interval and the method suggested by a commercial multivariate calibration software (Unscrambler) has been adapted in all multivariate calibrations reported here.

- There are three ways to validate any multivariate calibration model. These are the internal, external validation and cross validation techniques. Only cross validation technique is used in all calibration reported in this thesis.

- Internal validation is based on testing of the calibrated model with its calibration samples. There are doubts about how robustly this calibration would perform in prediction of new samples.
• External validation uses new samples for testing and overcomes the problem in the internal validation method. However, it demands sufficient new samples that have been measured by traditional chemical analysis. In practice, this is time consuming and it does not make the best used of the available calibration data.

• Cross validation gives a compromise between the problems in internal and external validation. This is accomplished by omitting some calibration samples during calibration modelling and used the omitted calibration samples for test. The validation is repeated several times, so that all calibration samples have been used for validation. This technique has the benefits of external validation because calibration models were established without some of the calibration samples, but calibration samples have not been wasted for validation purposes only.

• The number of principal components incorporated in a calibration model should have a direct relationship to the number of influences in the data. Typically the first few principal components describe most information of the data set, while higher orders of principal components tend only to describe noise. Therefore it is necessary to find the optimum number of principal components to be used in the calibration model.

• The optimum number of principal components can be determined by plotting the model MSE of prediction as a function of the number of principal components used. A general suggestion is that the first local minimum of MSE determines the number of principal components to be used in the calibration model. This suggestion has been adapted in all calibrations reported in this work.
4. A budget NIR spectrophotometer for bioprocess monitoring

4.1 Summary

This chapter describes the design and implementation of a budget NIR spectrophotometer. The preliminary investigation of bioprocess contaminants in the previous chapter has provided a guideline to the design specification. The spectrophotometer is based on a holographic grating spectrograph and a lead sulfide detector system. A commercial spectrophotometer is expected to cater for a range of customer requirements and would, for example, cater the full NIR wavelength range 700 to 2500 nm. Our requirements were specifically focused on the range 1900 to 2500 nm, which the in-house spectrophotometer achieved. In order to achieve sufficient NIR energies detection of the desired bioprocess contaminants, a different referencing technique is used. This has allowed a higher light throughput through 2 mm of turbid bioprocess material and limiting the detection window only to the concerned signal level. This has given the spectrophotometer detection error of ±0.22 percent. The in-house built spectrophotometer is compared with two commercial available spectrophotometers, this is shown in Appendix A4.

The implemented spectrophotometer has been calibrated for quantitative analysis of protein and RNA contaminants in clarified yeast homogenate. This has been achieved using partial least squares (PLS) models and two series of protein and RNA calibration samples. With both the spectrophotometer and the calibrated PLS model, quantitative analysis of samples with similar protein and RNA concentration to those used in the calibration samples can be achieved with precision of ±1.53 g L⁻¹ and 0.49 g L⁻¹ respectively.
4.2 Introduction

In this chapter, the design of a NIR spectrophotometer is presented and the emphasis on the selection of main components used in the spectrophotometer are given. These are diffraction grating, lead sulfide detector and techniques that can minimise noise detection and maximise signal detection. Preliminary survey using loaned instrument from Perstrop Analytical (NIRS6500, Berkshire, UK), this has provided a guideline of specification for the design of NIR spectrophotometer. In the preliminary survey, spectra acquired in all experiments used sampling cell with 1 mm pathlength, spectra acquired at every 2 nm between 1900 to 2500 nm, and absorbance range was within 3 Au. The detail of the preliminary survey is given in the Appendix A5.

The in-house spectrophotometer is built from individual components and formed a key stage of the project. The task involved optical design, mounting of components, precision alignment of optical paths, carpentry to build a robust enclosure and install sufficient ventilation. The total cost of components used was seven times less than the purchase of a commercial NIR spectrophotometer (NIRS6500, Perstrop Analytical, Berkshire, UK) which was used in the preliminary survey.

The completed spectrophotometer is tested and the results are also presented. There are two levels to this test. Firstly, the spectral reproducibility of a sample has been monitored for several days of operation and the standard error of measurement is identified. Secondly, using the spectrophotometer in conjunction with partial least squares (PLS) calibration has facilitated quantitative analysis of bioprocess contaminants.
4.3 Spectrophotometer

The NIR spectrophotometer is based on a diffraction holographic grating unit and a single element lead sulfide detector. The original design was such that it can be configured for reflectance as well as transmittance measurement. However, only the transmittance configuration has been fully tested and used for contaminants measurements. The following section describes the design layout of the spectrophotometer and the fundamental components to the spectrophotometer are discuss in the subsequent subchapters.

4.3.1 Design layout

The original design of the spectrophotometer was such that it will allow both transmittance and reflectance measurements. This has been achieved with the arrangement shown in figure 4.1.

A 20 Watts tungsten halogen lamp is used as the energy source in this spectrophotometer. It is capable of emitting stable ultra violet, visible and NIR energies. The output from the tungsten halogen lamp, i.e. the image of the lamp filament, is collected by a large concave mirror $A$. This reflects and focuses the lamp filament onto a silica glass cell via a mechanical chopper.
Fig 4.1 The design layout of a NIR spectrophotometer capable of both transmittance and reflectance measurements. The figure is shown in transmittance measurement configuration where the output of the set-up optics is focused directly to the sample cell. Reflectance measurements can be achieved by lowering the set-up optics plane to align the light ray to the reflectance light path.

The large concave mirror $A$ has a diameter of 101.6 mm collects most of the irradiance from the lamp’s filament. The imaged filament distance from the concave mirror $A$ is calculated from Gaussian conjugation equations 4.1 (Guide for spectroscopy, 1994) based on the mirror 150 mm focal length ($f$) and lamp distance ($s$) of 250 mm. The distance of mirror $A$ to the sampling cell is given by equation 4.1, $s = 375$ mm.

$$s' = \frac{fs}{f + s} \quad [4.1]$$

where $f = 150$ mm, $s = -250$ mm (sign convention in the Gaussian conjugation equation).

$$s' = \frac{150 \times (-250)}{150 + (-250)} = 375 \text{ mm}$$

The described units are mounted on a aluminium plane and this formed the set-up optics to the spectrophotometer. The set-up optics output is a focused image of the
filament in the tungsten halogen lamp. The measuring sample is held in the glass cell. Therefore the focused filament can be sent directly to the glass cell for transmittance measurements. Alternatively reflectance measurement can be achieved, by positioning the set-up optics plane such that the light path for transmittance measurements meets the reflectance light path. This required the use of a small concave mirror B to re-focus the filament image onto the other side of the cell, and a piece of non-reflecting ceramic material is placed on the left of the cell to eliminate energy loss through the back. The small concave mirror has a diameter of 25.4 mm and focal length of 50 mm to give $s_B=115$ mm and $s_B'=88.5$ mm.

Fig 4.2 Position of a concave mirror B for reflectance measurements. The mirror collects the set-up optics light output (the lamp filament) and reflects this onto the sample cell. The lamps filament appearing on the right hand side of the cell is used for reflectance measurements.

The filament image appearing on the right hand side of the sample cell, either by penetration through the sample from the left side for transmittance measurements or reflected by the sample for reflectance measurements is collected by another large concave mirror C. This has the same optical arrangement as mirror A and the filament image is focused onto the entrance slit of the image spectrograph. The entrance slit serves as a energy input control to the spectrograph. The output from the spectrograph is a linear spectrum at wavelength between 1100 to 2500nm that spans across 24.7 mm. A single element lead sulfide PbS detector is mounted on a stepper motor driven stage. This allows the transportation of the detector across the output spectrum of the spectrograph to collect the NIR spectrum as illustrated in figure 4.3. An order sorting filter is placed in the front of the PbS detector to eliminate unwanted energies emitted in shorter wavelengths.
Fig 4.3 The lead sulfide PbS detector is mounted on a stage driven by stepper motor. The arrangement is set to detect the linear dispersed NIR spectrum of 1100 to 2500 nm over a distance of 24.7 mm output from the image spectrograph. This is achieved by moving the detector across the presented spectrum.

### 4.4 Diffraction gratings

The selected spectrograph in the spectrophotometer has a holographic grating. This provides a linearly dispersed spectrum at the output. A single PbS detector mounted on a stepper motor driven stage is used to collect the NIR spectrum. It is also suitable for upgrade to a linear diode array detector for fast spectral collections by placing the diode array directly across the output port of spectrograph. The selected spectrograph has covered the specification required determined in the preliminary investigation (Appendix A5). It provides spectral output 1100 to 2500 nm and is capable of delivering spectral information every 0.47 nm. Previously in Chapter 2 the basic theory of diffraction gratings has been given. In the following, the theory and practical extension of rule and holographic grating are discussed. These included the linear dispersion, bandpass, spectral resolving power, blazed grating, and stray light. Finally, a summary on the two types of gratings are presented which highlights the reasons for choosing the holographic grating.
4.4.1 Linear dispersion in spectrographs

Referring back to section 2.6.2.1, the design of a monochromator is such that the input light will enter the grating via an entrance slit, diffract off the grating and exit via an output slit. At any grating setting only a very small range of angles around $D$, ideally one wavelength passes through the monochromator at a time. But in a spectrograph, an output slit is not necessary, but rather a wide range of angles $D$ are look at simultaneously (see figure 2.12). Therefore a range of wavelengths which satisfy the grating equation for the range of angles $D$ are presented. The result is that the output is a long strip of spectrum over which the various wavelength are dispersed in a geometry.

Linear dispersion defines the extent to which a spectral interval is spread out across the focal field of a spectrometer and is expressed in nm/mm. For example, consider two spectrometers, one instrument disperses a 0.1 nm spectral segment over 1mm while the other takes a 10 nm spectral segment and spreads it over 1mm. The fine spectral detail would be more easily identified in the first instrument than the second. The second instrument demonstrates low dispersion compared to the higher dispersion of the first. Linear dispersion is associated with the ability of an instrument to resolve fine spectral detail.

In a spectrograph and monochromator, the angle $I$ in the grating equation is fixed and if differentiated with respect to wavelength;

\[
\frac{m}{a} = \frac{(\sin I + \sin D)}{\lambda} \tag{4.2}
\]

\[
\frac{\delta D}{\delta \lambda} = \frac{m}{a \cos D} = \frac{\sin I + \sin D}{\lambda \cos D} \tag{4.3}
\]

$\delta D/\delta \lambda$ is the angular dispersion or the change of diffraction angle corresponding to a small change in wavelength. It is greater for smaller groove spacing, $a$, (greater number of lines per millimetre) and larger orders, $m$, and larger diffraction angles, $D$. 
The linear dispersion at the exit slit of a spectrograph varies with focal length, $f$, and the diffraction angle $D$ and is the product of the focal length and angular dispersion:

$$\frac{\delta x}{\delta \lambda} = f \times \frac{\delta D}{\delta \lambda} = \frac{f \times m}{a \times \cos D} \quad [4.4]$$

Commonly, the term *reciprocal linear dispersion* is used which gives the wavelength dispersion in mm/nm of slit width.

### 4.4.2 Grating order and order sorting filter

An example of a first order spectrum from 1000 to 2500 nm spread over a focal field in a spectrograph configuration is illustrated in figure 4.4 below. From the earlier basic grating equation [2.35] with a grating of given groove density and for a given value of incident angle $I$ and diffraction angle $D$, then the product of grating order $m$ and wavelength $\lambda$ would be constant. Therefore, if the diffraction order $m$ is doubled, $\lambda$ is halved.

$$m \lambda = \text{constant} \quad [4.5]$$

If a light source emits energies from 600 nm to 2500 nm, then at the physical location of 1500 nm in first order (figure 4.4) wavelengths of 750, 500 and 375 nm will also be present and available to the same detector. In order to monitor only light at 1500 nm, filters must be used to eliminate the higher orders.
The selected spectrograph in the spectrophotometer produced output spectrum of 1100 to 2500 nm. In practice lower order wavelength also appears at its output, therefore an order sorting filter has been added. To ensure good elimination from lower order across the output wavelength can be complicated by its bandwidth. This is to provide good suppression on the second order response for 2500 nm would required a filter to eliminate energy from 1250 nm. However, this filter would also suppress the first order emission of 1250 nm. Thus, ideally two filters with different pass band are necessary to avoid complications from higher order interference in a broad spectrum output spectrograph.

The preliminary investigation in this project found that only 1900 to 2500 nm NIR spectrum is required. An order sorting filter that had transmission for 1900 nm and above was chosen to eliminate second order interference.
4.4.3 Blazed gratings and efficiency

In the case where monochromatic light reaches a grating, a fraction of this is diffracted into other orders. The fraction diffracted into the first order determines the grating efficiency. Blazed gratings are manufactured to produce maximum efficiency at designated wavelengths. A grating can be described as *blazed at 250 nm* or *blazed at 1 micron*, by appropriate selection of groove geometry.

A blazed grating is one in which the grooves of the diffraction grating are controlled to form right triangles with a *blaze angle*, $\omega$, as shown in figure 4.5. Blazed grating groove profiles are calculated for the Littrow condition where the incident and diffracted rays are autocollimated (i.e., $I = D$). The input and output rays are therefore propagated along the same axis. In this case at the *blaze* wavelength $\lambda_B$,

\[
\sin I + \sin D = m \ n \ \lambda_B \quad [4.6]
\]

\[
\omega = I = D \quad \text{where} \ \omega = \text{blaze angle}
\]

\[
2 \sin \omega = m \ n \ \lambda_B \quad [4.7]
\]

For example, the blaze angle ($\omega$) for a 1200 g/mm grating blazed at 250 nm is 8.63° in first order ($m = 1$).
The efficiency of a diffraction grating is measured in the Littrow configuration at a given wavelength.

\[
\text{% Absolute efficiency} = \frac{\text{energy out}}{\text{energy in}} \times 100 \quad [4.8]
\]

In figure 4.6a and 4.6b two typical efficiency curves of a blazed, ruled grating and a non-blazed, holographic grating are shown respectively. As a general approximation, for blazed gratings the strength of a signal is reduced by 50% at two-thirds the blaze wavelength, and 1.8 times the blaze wavelength.

![Efficiency curves](image)

Fig 4.6 (a) left and (b) right. Efficiency profile of a blazed ruled grating and a non-blazed holographic grating. These have shown that blazed rule grating provided higher efficiency at the blazed wavelength.

The efficiency of a blazed grating and its responses to grating order can be summarised to the following points.

i. A grating blazed in the first order is equally blazed in the higher orders. Therefore, a grating blazed at 600 nm in the first order is also blazed at 300 nm in the second order and similar for the third and higher order.

ii. Efficiency in higher orders usually follows the first order efficiency curve.

iii. For a grating that is blazed in the first order, the maximum efficiency for each of the subsequent higher orders decreases as the order \( m \) increases.

iv. The efficiency also decreases the further off-Littrow (angle \( I \neq D \)) the grating used.
The efficiency of a non-blazed holographic grating does not necessarily mean that it is less efficient. Figure 4.6b showing the efficiency curve for an 1800 g/mm sinusoidal grooved holographic grating with good efficiency between 500 to 700 nm. General holographic grating has a flatter efficiency response when compared to ruled grating.

### 4.4.4 Diffraction grating stray light

Light other than the wavelength of interest reaching a detector is referred to as stray light. This stray light often includes one or more elements of scattered light. Scattered light can be produced by either; randomly scattered light due to surface imperfections on any optical surface, and focused stray light due to non-periodic errors in the ruling of grating grooves.

If the diffraction grating has periodic ruling errors, a ghost, which is not scattered light will be focused in the dispersion plane. Ghosts are focused and imaged in the dispersion plane of the spectrograph. Generally, a holographic grating has stray light up to a factor of ten times less than that of a classically ruled grating. Also holographic gratings do not show any ghosts because there are no periodic ruling errors and, therefore often represent the best solution to ghost problems.

### 4.4.5 Choice of holographic grating

Holographic gratings are created using an interferometric hologram and sometimes an etching process. They have been known as the best gratings, but this is not necessarily so. Sinusoidally grooved gratings produce very little scattered light but have low, flat efficiency curves, although they are generally quite broad. Blazed holographic gratings use etching during the interferometric process, or an ion gun to form a blaze angle in a secondary process. The former does not produce strong
Neel Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Chapter 4

blazing, and while the latter produces high efficiencies at the blaze wavelengths, light scatter is increased due to the formation of micro structures along the edges of the grooves. Blazed ruled grating can be achieved but ghosts and subsequent stray light intensity are proportional to the squares of order and groove (Guide for spectroscopy, 1994). Thus careful consideration has to be made when using ruled gratings in higher order or with high groove density. Table 4.1 has list the applications for both holographic and ruled gratings.

<table>
<thead>
<tr>
<th>Holographic gratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Grating is concave, i.e. spectrograph.</td>
</tr>
<tr>
<td>ii. laser light is present, e.g. Raman, laser fluorescence.</td>
</tr>
<tr>
<td>iii. High groove density should be 1200 g/mm or more (up to 6000 g/mm) for use in near ultra-violet, visible, and near infrared spectrum.</td>
</tr>
<tr>
<td>iv. UV below 200 nm down to 3 nm.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ruled gratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Infrared application above 1.2 mm, if an ion-etched holographic grating is unavailable.</td>
</tr>
<tr>
<td>ii. Very low groove density, e.g., less than 600 g/mm.</td>
</tr>
</tbody>
</table>

Table 4.1 Suggestion on applications for holographic or ruled grating used.

A spectrograph and hence a holographic grating was chosen in the design of the NIR spectrophotometer. The primary reason for spectrograph was because a diode array PbS detector was one of the future improvement which will allow instant collection of the complete spectrum and elimination of a moving part of stepper motor in the spectrophotometer. This requires a linear dispersion diffraction grating and hence a holographic grating spectrograph was used.

The preliminary investigation on specification of the NIR spectrophotometer for the biological application determined the spectral resolution requirement. This had suggested spectral measurements of every 2 nm and a useful NIR response between 1900 to 2500 nm. The selected spectrograph CP14-2021 holographic grating (Instrument S.A. Ltd, Middlesex, UK) has a groove density of 120 g/mm, cover wavelength range 1100 to 2500 nm, average dispersion of 56.7 nm/mm and length of
spectrum of 24.7 mm. This spectrograph is capable to deliver spectral information at every 0.47 nm (56.7 / 120). When spectrographs are used with a diode array detector or with a single detector stepping across the output spectrum, the resolution is limited by the step distance in the scanning process. For scanning of every 2 nm in the spectrum, the detector has to travelled a distance of 35.3 μm across the spectrograph’s linear output.

The ideal blaze wavelength would be between 1900 to 2500 nm for the required bioprocess application. However, due to availability on holographic grating with the desired spectral range, and cost budget reason at the time of use, a non-blazed spectrograph was chosen.

### 4.5 PbS detector system

The lead sulfide detector has been chosen for the NIR instrument because it provides good detectivity (defined in section 4.5.1) for wavelength ranges from 1000 to 2500 nm, and particularly so in 1900 to 2500 nm. Therefore it is ideal for the required bioprocess application. The detector used is a single element and has an active area of 3×3 mm. In order to increase the detectivity, a two stage thermal electric (TE) cooler is used. The detector was supplied together with temperature controller, bias voltage power supply and amplifier circuits that are integrated to a single unit (IRI 2700 PbS, TC-328 controller, Graseby Infrared, Orlando, USA). The detector is required to operate with chopped optical signal, and incorporates AC coupled electronic amplification circuits. This was because the 1/f noise caused by the bias current makes them unsuitable for DC operation and the AC coupled electronics suppresses this noise. The following figure illustrates the detection system implemented.
The constant optical energy is chopped by a mechanical rotating chopper wheel as also seen in the design layout earlier (figure 4.1). This modulates the optical source providing pulsing energy for detection. The lead sulfide detector generally has lower noise with higher modulation frequency, however, the detection response reduces as modulation frequency increases. The modulation frequency has been optimised between detector noise, signal detectivity and avoid problems from the natural modulation frequencies of noise sources (50 Hz mains power source and its harmonics). A chopped or modulation frequency of 375 Hz has been chosen.

The detector is biased by ±60 volts DC and a two stage thermal electric (TE) cooler is build into the detector package. During operation it is set to operating at a temperature of 263°K (-10°C). The heat generated from the cooler is dissipated by a heat sink which is mounted on the back of the detector package. The advantage of employing a TE cooler is discussed in section 4.5.3.

The output from the PbS detector is AC coupled electrical signal that has amplitude corresponded to the measuring NIR energy. The signal is pre-amplified and
amplified in the supplied detector electronics. This has a low output impedance with maximum ±10 volts output. The output from this complete detection system is then feed into a data acquisition board (AT-MIO-16XE-50, National Instrument, Berkshire, UK) and interfaced to a IBM compatible 80486 personal computer. The PC will act as a signal integrator that converts optical pulse energy signal to a numerical units before further data analysis.

In the following sections, the basic theory of detection system and the practical aspects of the selected detector system will be described. Firstly, a brief review of figures of merit of a detector are given, then the characteristics of the lead sulfide detector which led to its use are highlighted. Followed are the temperature characteristics of the detector and compensation techniques. Then there is a review of detector noise and the specific detection system adapted in order to minimise noise detection.

4.5.1 PbS detectors and figures of merit

Detectors are described by certain figures of merit. The figures of merit are usually functions of wavelength and temperature and may be affected by detector size, modulating frequency, bias voltage and the gain of any internal amplifier. Three commonly used figures of merit are given below.

**Noise equivalent power (NEP)**

NEP is the radiant flux in watts necessary to given an output signal equal to the r.m.s. noise output from the detector. The flux may be either continuous or sinusoidally modulated.

The response is assumed to be linear down to the noise level. NEP values should be stated at a specified wavelength, modulation frequency, detector area, temperature and detector bandwidth. Detector bandwidth is usually chosen as 1 Hz and NEP is frequently quoted in watts Hz\(^{-1/2}\). NEP is the most commonly used version of Noise Equivalent Detector Input (P\(_N\)).
Detectivity (D)

Detectivity, D is the reciprocal of NEP. This gives a figure of merit which is larger for more sensitive detectors.

Normalised Detectivity (D*)

For most detector DA_D^{1/2} is constant, i.e. the detectivity varies inversely with the square root of the area of the detector. This is because the electrical noise power is usually proportional to the detector area A_D, and current or voltage, which provide a measure of that noise, are proportional to the square root of power. Similarly, because most detector noise is white noise, and the white noise power is proportional to Δf (where Δf is the electrical bandwidth), the noise signal is proportional to Δf^{1/2}.

D* is defined to allow comparison of different types of detectors independent of the detector area and bandwidth.

\[
D^* = D \left( A_D^{1/2} \Delta f^{1/2} \right) = D \left( A_D \Delta f \right)^{1/2}\ 
\]

\[
D^* = \frac{A_D \Delta f}{NEP} \ 
\]

The units of D* are cm Hz^{1/2}W^{-1}. Since D*, like NEP, is a function of wavelength and modulation frequency, it is common to see the normalised spectral responsivity D*(λ), or D*(λ, f, Δf). Figure 4.8 illustrates the D* of a number of detectors.

In comparison with other detectors (see figure 4.8), lead sulfide PbS is ideal for this application of NIR spectroscopy because its detectivity response covers the wavelengths of interest and it is for this reason PbS detector is used in this spectrophotometer. The PbS normalised detectivity (D*) is function of modulation frequency and operating temperature. In the following sections, the selection of bias voltage, operating temperature and modulation frequency are presented.
Fig 4.8 Normalised detectivity ($D^*$) of various detectors are presented. This emphasizes PbS has the appropriate detectivity at the wavelength of interest.

4.5.2 PbS detectors bias supply voltage

PbS detector is fabricated by chemically depositing polycrystalline film on a quartz substrate. Gold electrodes are plated to the edges of the film to provide electrical contact and the whole assembly is sealed inside a package with an appropriate window, quartz or sapphire, as illustrated in figure 4.9.
A lead sulfide detector is a form of photoconductive detector which absorbs incident photons to produce free charge carriers. These change the electrical conductivity of the detector. A bias voltage is applied (figure 4.10b), causing a current to flow which is proportional to the photon irradiance.

Fig 4.10 a) Schematic of a photoconductive detector. b) PbS simple circuit arrangement for conversion of NIR energy to electrical signal before pre-amplifier.

A bias voltage of ±60 volts and loading resistance of 1 MΩ (figure 4.10b) have been used. This followed the detector manufacture’s data specification for the 3×3 mm active detection area (figure 4.11).

Fig 4.11 Manufacture’s recommendation on maximum detector voltage vs detector area and resistance for detector temperature of 298°K (25°C).
4.5.3 Temperature

PbS detectors have some temperature dependence in their noise and detectivity characteristics. More reproducible results can be achieved with stabilising the thermal environment of the detector. Also the PbS has much better detectivity when cooled because dark currents (response to dark) diminish a factor of 2 for every 5 to 20°C, depending on the detector characteristics.

![Graph of spectral detectivity at different temperatures](image)

![Graph of noise vs frequency](image)

Fig 4.12 a) Typical spectral detectivity at temperature. b) Example of noise vs frequency as a function of detector temperature.

The PbS detector used has included two stages of thermal electric (TE) cooling and a thermistor sensor inside of the housing. The thermistor output signal is fed into an external closed loop cooling controller unit. This controls the TE cooling so that the detector’s operating temperature is maintain constant. The heat being generated by the TE coolers also protects the windows from condensation. The temperature characteristic of the PbS used is plotted in the above figure 4.12. The PbS current setting in the spectrophotometer is 263°K (-10°C), and it is control by a hybrid controller. A heat sink unit is also mounted at the back of the PbS detector package to minimise heat build up generated by the cooler. The operating temperature of the detector has been set to 263°K (-10°C) throughout the project. At this setting, heat dissipation has been well managed for long hours of operation under normal room temperature.
4.5.4 Noise and noise compensations

There are various types of noise that can limit the detectivity. These are well known and commonly defined in the following. Two methods of noise compensation are given in the subsequent sections. These are modulation and gated integrating techniques.

**Shot noise**
This is due to the discrete nature of radiation, which is composed of photons arriving randomly in time. Absorbed photons produce photoelectrons at random intervals, and this variation in current appears as noise. This noise can be generated by actual desired signal photons or, in case of very low signal and very low noise detectors, by background photon flux. When that occurs, the detector system is said to operate in a Background Limited In Performance (BLIP) mode.

**Generation recombination noise**
This is seen in photoconductors in which the absorbed photons produce both positive and negative charge carriers. Some of the free carriers may recombine before they are collected. Thermal excitation may generate addition carriers. Both the generation and recombination occur randomly, resulting in noise fluctuations in the output current.

**Johnson Noise or thermal noise**
It is caused by the random motion of carriers in a conductor. The result is fluctuations in the internal resistance of the detector, or in any resistance in series with the terminals of the detector.

**Flicker or 1/f Noise**
This is not well understood. It occurs in detectors such as photoconductors which require a biasing current. Its magnitude is proportional to $1/f^B$ where $B$ is usually between 0.8 and 1.2.
4.5.4.1 Modulation techniques

It is very difficult to control noise characteristics, but in practice an AC technique is applied to suppress noise detection. This is by converting the direct NIR energy signals to pulses of energy known as modulating the signal. Frequently, an AC coupled amplifier with a band pass filter centred on the modulation frequency helps increase the signal to noise ratio significantly. The limits to this band pass filter are controlled by the stability of the modulating signal.

In the PbS detection system used here, the built in AC coupled pre-amplifier electronics incorporates a band pass filter. The manufactured electronics has a band pass frequency from 10 Hz to 1 kHz, which allows flexibility on selecting the optimum modulation (chopping) frequency for the detector responses.

The optimum modulation frequency is also dependent upon the detector's operation temperature 263°C (-10°C). In order to avoid problems from the natural modulation frequencies of noise source (50 Hz mains power source), the modulation frequency of 375 Hz has been used. This combination operates the detector close to the optimum detectivity (figure 4.13).
Fig 4.13 Manufacture’s data on PbS normalised detectivity. This shows optimum detectivity requires balance between modulation frequency and operating temperature. The modulation frequency used in the spectrophotometer detection system is 375 Hz to achieve an optimum detectivity when detector is operating at 263°K (-10 °C).

4.5.4.2 Gated Integrator

This is an operation which the integration period is gated to the window of signal, this is illustrated in figure 4.14. The integration only takes place when the signal is present and the noise contribution which would be accumulated in the signals absence has been ignored, therefore the signal to noise ratio is improved.

Fig 4.14 Repetitive signal and detection “windows”. Noise measurement is reduced when the detection period is limited to the presence of good signal.
The gated integrator process take place in the PC and the process is repeated for \( N \) pulses (where \( N = 375 \), with the chopper running at 375Hz for 1 second). In theory, this leads to signal to noise ratio improvement of \( N^{1/2} \) if the noise is of the white variety, Johnson or shot. This is because the integrated signal contribution increases as \( N \), while noise contribution increases only as \( N^{1/2} \).

### 4.5.5 Analogue to digital conversions

The output electrical signal from the detector system is fed into a data acquisition board via a signal conditioner. The signal conditioner is a simple potential divider unit that reduces the ±15 volts output from the detector to ±10 volts. This provides a suitable input voltage range for the data acquisition board to capture the detector's output.

The data acquisition board AT-MIO-16XE-50 (National instrument, Berkshire, UK) has a 16 bits analogue to digital converter (ADC), this is capable of representing the detector signal by 65536 \( (2^{16}) \) divisions. Thus the quantisation error (error between the digital representation and the actual detector signal) is negligible compared to the noise accumulated through the whole detection system. The sampling frequency has been maximised to 20 kHz which is the upper limit in the data acquisition board. Since the modulated optical frequency (chopper frequency) is 375 Hz, the sampling frequency at 20 kHz is higher than the minimum required sampling frequency stated by the Nyquist sampling frequency theorem. This defined sampling must be at more than twice the rate of the maximum frequency component in the sampling signal.

The complete analogue to digital conversion process provided digital representation to the actual detector signal. Both errors introduced by the digital quantisation and the sampling process have been considered negligible. Thus the analogue to digital conversion process is achieved without loss of any detail in the original detected signal.
4.6 Acquiring the NIR spectrum

In the previous sections, emphasis has been on the preparation of arranging the NIR energy to interact with the measuring samples, choosing the spectrograph to separate the NIR energy into a linear disperse spectrum, and to optimise signal to noise ratio on the detection system. The following will discuss how the NIR dispersed spectrum presented at the output of the spectrograph is collected.

4.6.1 Acquiring the NIR spectrum from the spectrographs output

The NIR spectrum is collected by moving the PbS detector across the spectrum in discrete steps. This allows precise location of the detector in the spectrum by the step counter (i unit shown in the figure 4.15). The detector is set stationary for 1 second while measuring (gated integrating) the infrared signal. The duration of the stationary interval and the longer integration time has given better signal to noise performance, even though the detector has a response time in the order of 0.5 to 1 ms. The timing sequence of the stepping process is illustrated in the figure below.

![Discrete scanning of a water transmittance spectrum](image)

Fig 4.15 This illustrates the discrete scanning of a water transmittance spectrum. At ith wavelength, the detector is at stationary position and integrates the incoming NIR signal. The detector is then moved to next distance of the dispersed spectrum to integrate the next or i+1 th wavelength.
This approach provided good stable detection but has a long spectrum collection time. This was because the stepper motor is required to be turned on and off at each discrete point of detection.

It was pointed out that the NIR spectrum should be collected at every 2 nm (section 4.2) for the chosen bioprocessing application. However, because the discrete method is a slow process, in order to reduce the spectrum scanning time, spectral data is collected at every 4 nm. This was necessary when large volumes of sample are required to be processed within time constraints during the calibration experiments, see chapter 5.

4.6.2 Spectral smoothing

The spectral smoothing process is applied immediately after recording of the NIR spectrum. Although various precautions have been mentioned on collecting the NIR spectra, smoothing can remove further noise or any disturbance during recordings of the spectrum. Spectral smoothing is primarily concerned with reduction in higher frequency ripple noise. The corresponding lower frequency noise (e.g. instrument drift during the scanning of spectrum) is more difficult to suppress because it may resemble the real information in the spectra. Therefore smoothing is necessary to remove as much noise as possible in the spectrum without excessively degrading important information. Two popular methods for smoothing spectra are moving point average and spline.

Throughout all spectral smoothing in this project, a spline function that uses the generalised cross validation (GCV) and genetic algorithm (GA) method by Thornhill (Thornhill et al, 1994) is used. This has the advance of choosing both the degree of spline and minimising its parameters at the expenses of computation time. This GCV/GA spline smoothing has provided better smoothing compared to moving average or conventional cubic spline functions. The following describes briefly both
near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Chapter 4

smoothing techniques and an example is given which indicates the advantage of using GCV/GA spline over the others mentioned.

4.6.2.1 Moving point averages

Moving point average smoothing performs an average of an odd number of sequential points, replacing the centre point with the average, i.e. the spectral data $x_j$ at each wavelength $j = 1, 2, \ldots, J$ is replaced by averages of itself and its neighbouring points from $j-D$ to $j+D$.

$$x_j = \frac{\sum_{-D}^{+D} x_{j+d}}{N} \quad [4.11]$$

The denominator $N$ in the equation is any odd number of sequential points and is equal to $2D+1$. The moving point average smoothing process usually starts on the left end of the spectrum (at the shortest wavelength) and moves to the right one data at a time until the right end of the spectrum is reached. Note that the centre point of the interval is replaced with average, $D$ points are not included in each end of the smoothed spectrum. The detailed properties of moving point average are given by Rabiner and Gold (Martens, 1989).

4.6.2.2 GCV/GA spline

The spline technique is based upon the assumption that in small intervals most functions can be fitted by lower degree polynomials. Therefore it divides the spectrum into segments and fits polynomials to each of these segments under the restriction that the resulting combined polynomial is a continuous function. The polynomial segments join at the points where all the derivatives, except the highest, are continuous. For example, a cubic spline has a discontinuities only in the third derivative at the segment joints. It has been recognised that the smoothing of the spline approximation avoids the following of random noise in the data. An adjustable smoothness factor was suggested by Reinsch (1967). This is achieved by
balancing between the fidelity of the spline to the data, and its roughness indicated by the values of the higher derivatives.

It was found by Craven and Wahba (1979) that an optimum choice of smoothness factor by generalised cross validation (GCV) eliminated the need for a fidelity constraint. The GCV technique generates a set of splines from the spectral data, but some spectral data are neglected during generalisation of the spline. The spline is then used to predict the neglected data values. The spline function used in this project, developed by Thornhill et al (Thornhill, 1994) for process measurement data, chooses both the degree of the spline and its parameters by minimising the GCV with a genetic algorithm (GA).

4.6.2.3 Comparison of moving average, interpolating cubic spline and GCV/GA spline in spectral smoothing

In this work the use of an interpolating spline and moving average method have been compared to the smoothing spline method. In figure 4.16 below, smoothing to a NIR water transmission spectrum between 1900 to 2500 nm has been performed by moving average, interpolating cubic spline and GCV/GA spline functions. An artificial disturbance has been added to the original measured spectrum in 2200 nm, this simulated a noise as if due to an electrical spike while recording the NIR spectrum. The figure has shown that moving point average has spread the noise to its neighbouring points and cubic spline has not shown any reduction, and only the GCV/GA spline have efficiently suppressed this disturbance. It was really the ability to suppress sudden disturbances in the spectra that led to the decision to use the GCV/GA spline function for all spectral smoothing.
4.7 Specification performance of the spectrophotometer and practical aspects

The implementation of the spectrophotometer has been fully discussed in the previous sections. It is necessary to study the performance of the spectrophotometer. The following reports both the wavelength calibration of the spectrophotometer and the specification of the NIR transmission measurements. Also described are some practical aspects that one needs to be aware of during the collection of the NIR spectrum.
4.7.1 Wavelength calibration

The design and in-house built spectrophotometer is required to collect NIR spectra between 1900 to 2500nm for the biological chemical analysis. In this spectrophotometer, the spectrum is presented at the spectrograph output and its exact wavelength output can be determined by the geometry given by the manufacturer.

Generally, interference filters are used on a commercial spectrophotometer for wavelength calibration. These filters should have well defined wavelength with sharp pass bands. Here in order to minimise cost of interference filters, the wavelength calibration was referenced by the water absorption spectrum measured from a commercial spectrophotometer (NIRS6500, Perstrop Analytical, Berks, UK). The transmission spectrum between 1900 to 2500 nm has been chosen to compared with the in-house build spectrophotometer.

The nature of this work is to use the NIR response from the interesting bioprocess contaminants within the predetermined wavelength range (1950 - 2500 nm) and to apply multivariate analysis. Thus a calibration is needed that is suitable for the purpose. Exact wavelength calibration is not needed, but rather the spectra must be shown to be reproducible.

The desired NIR wavelength range has been identified by two distinct spectral response in the water transmission spectrum. Figure 4.17 has shown that the transmission peak of water at 2210 nm by the in-house spectrophotometer is matched with the recording from the Perstrop spectrophotometer. The next distinctive spectral response of water observed by the Perstrop is the minimum transmission at 1980 nm. A similar minimum transmission is also observed by the in-house spectrophotometer at 1950 nm. The two spectrophotometers do not agree on this minimum point of transmission, literature by G.J. Kemeny (Kemeny, 1992) has reported the minimum water transmission was observed at 1950 nm which agrees with the in-house built spectrophotometer whereas the Perstrop gave a minimum at 1980 nm. This comparison has shown that the detector has been correctly aligned to
Near Infrared Spectroscopy Technique for Bioprocess Monitoring and Control

Chapter 4

record a realistic NIR spectrum. Hence the recorded NIR spectrum between 1900 to 2500 nm is fit for the purpose and satisfies the specification requirement.

![Graph showing spectral transmittance response of water from commercial and in-house built spectrophotometers.]

Fig 4.17 The spectral transmittance response of water from a commercial spectrophotometer and the in-house built spectrophotometer described here. This confirmed that the in-house spectrophotometer is capable of recording the desired NIR spectrum (1900 to 2500 nm).

4.7.2 Approach to referencing and baseline correction

In most commercial spectrophotometers, resultant measurements are usually expressed in unit of percentage of transmittance or absorbance (Au) defined in equation 2.27. This is a ratio technique that compensates drifts from an imperfectly stable optical source. In order to achieved this ratio, a reference spectrum is required to be measured prior to sample measurements. Traditionally, for absorbance or transmittance the reference spectrum would be the measurement of the direct optical source, i.e. open path measurement. Therefore, both transmittance and absorbance would describe only the measuring samples.
Fig 4.18 a) The tungsten halogen lamp is set to produce low light energy output, this does not saturate the detection input (see open path transmission spectrum). At this low light energy, the 1 mm pathlength transmission spectrum of water is relative small and the 2 mm pathlength transmission spectrum of water is relatively close to a dark spectrum measurement. b) The lamp output is set to produce higher light energy output, the 1 mm pathlength transmission spectrum of water has saturated the detection input. The 2 mm pathlength transmission spectrum has utilised the whole detection limits and hence maximises the signal to noise ratio in of spectrophotometer.

A tungsten halogen lamp is used as the optical source in the NIR spectrophotometer. In practice, the lamp output is adjusted for two reasons, these are that the direct lamp output (i.e. $P_O = P_T$, equation 2.27) should not saturate the detector while also providing sufficient optical energy for the measuring applications. The former could be achieved by reducing the light power from the tungsten halogen lamp such that the maximum open path transmission does not saturate the detector input (e.g. the normalised transmission should be less than unity). However, this suggestion is complicated by the high NIR absorption in water and the requirement on the detection of low concentration of biological components in a very turbid background, e.g. protein and RNA at 26 and 12 g L$^{-1}$. This problem is illustrated in figure 4.18a,
the spectrophotometer has been configured to satisfy the former criteria by low input light energy, but the 1 mm pathlength transmission spectra of water has only approximately 7 percent of the peak light transmission (see inset window of figure 4.18a). This implies that transmission spectra of bioprocess samples used in this project would suffer small signal to noise from this low input light energy configuration.

Alternatively, the spectral responses from these bioprocess samples may be enhanced by increasing the optical source radiation thus allowing greater light energy to penetrate through water and the turbid samples. Figure 4.18b has illustrated the affect of the higher input light energy used throughout this project. There, the 1 mm pathlength transmission spectrum of water has been increased such that it even saturates the detection limit. In fact in input light energy has been increased so that a 2 mm pathlength transmission of water is within the detection limit where previously in figure 4.18a, it has similar transmission to dark measurements. This method is effective since it increases the signal to noise in measuring bioprocess samples and thus improves the resolution of detecting low concentration of the contaminants of interested in this project. This method has however saturated the detector input and therefore the measurement on the direct optical source cannot be used for reference. The total absorption (Au) unit which is commonly used in chemical analysis is not feasible. Therefore another form of reference material or sample is needed for the reference spectrum.

A 2 mm pathlength transmission spectrum of distilled water had been used as a reference throughout all experiments. This was because it has a similar transmission profile to the biological sample of interest. A reference using distilled water provides a guideline to adjust the output radiation from the optical source. It has the advantage that the reference spectrum of water fits into the detection window and does not saturate the detector input (figure 4.18b). By providing greater NIR energy from the light source, it is now possible to increase the 1 mm pathlength of the sample used in the preliminary survey to 2 mm. This again increases NIR response to the bioprocess samples by the longer path of interaction. The overall signal to
noise is therefore improved since the NIR spectra has been moved from the noisy region (figure 4.18a) to the signal detection region (figure 4.18b).

A transmission spectrum of the distilled water was recorded and used as a master reference. In subsequent measurements, a reference spectrum of distilled water is recorded before every measurement of the biological sample. The differences between the master reference and the new reference spectrum are calculated to give a spectrum of correction coefficient. This spectrum of correction coefficients is then used for adjusting any baseline shift in the subsequent spectrum, i.e. the spectrum of the bioprocess sample.

\[ C_n = \frac{R_m}{R_s} \]  \[ \text{[4.12]} \]

where  
\( R_m \) is the master reference spectrum of distilled water  
\( R_s \) is the reference spectrum of distilled water before each measuring samples  
\( C_n \) is the correction coefficient  
\( n \) is wavelength at \( \lambda, \lambda+1, \lambda+2, \ldots \)

\[ S_n' = S_n \times C_n \]  \[ \text{[4.13]} \]

where  
\( S_n \) is the spectrum of the measuring sample  
\( S_n' \) baseline corrected spectrum

The baseline corrected spectrum of the measuring sample \( S_n' \) is then used for all subsequent multivariate analysis. The continued collection of reference spectra \( R_s \) prior to sample measurement allows studies of day to day variability from the spectrophotometer. Any sudden variation within the spectrophotometer that has affected spectrum recording can be immediately identified and adjustment may be made to the spectrophotometer. Examples of these variations are discussed in later section 4.7.4 on practical aspects in using the spectrophotometer.
4.7.3 Reproducibility of the spectrophotometer

An examination of spectral reproducibility studied the variability of spectra measured from the same sample. The baseline correction technique described above is designed to eliminate systems drift. In practice, however, there are always differences between spectra of the same sample, these are in wavelength and the measurement of NIR radiation differences. Therefore it is important to identify the variability of the spectrophotometer and to ensure the variations are within limits acceptable for the proposed analysis.

4.7.3.1 Wavelength reproducibility

The wavelength reproducibility defines how well the spectrophotometer can measure a particular wavelength every time. The wavelength variation in the spectrophotometer is regarded as negligible. This is because that the high precision stepper motor stage used has an error of approximately 1 μm for the full 16 mm travelled. The spectrograph output has an average dispersion of 56.7 nm/mm and in all cases spectral information is recorded coarsely by every 4 nm (distance of 70.5 μm). Therefore it is expected that the wavelength reproducibility is sufficient for the work here.
4.7.3.2 Measurements reproducibility

The reproducibility of water transmission spectra for three days are studied. In each day, ten transmission spectra of distilled water are recorded with baseline corrected by the referencing technique (section 4.7.2). The spectrophotometer had undergone a complete turn off and turn on cycle between days of scanning. The spectra are normalised with maximum transmission equalled to unity, the average of ten spectra from each day are shown (figure 4.19) together with the spectrum of standard deviation found within each day of scanning. In this figure, the highest standard deviation from the 3 days of scanning is 0.22%. The standard deviations are generally larger in the region of high transmittance, as expected.

Overall the spectrophotometer has shown good spectral reproducibility and no systematic drift was found. The achieved acceptable band on standard deviation is drawn to 0.22% within a day of measurements. The inset figure (4.19) shows the
variability between spectra from different days is less than the variability within one
day of scanning.

4.7.4 Practical aspects of using the spectrophotometer

During the operation of the spectrophotometer, it has been found that electrical noise
and vibration can upset the spectrum measurements. The former is caused by a
nearby electrical appliance generating a large electrical noise to the mains supply line
which runs to the spectrophotometer. The latter is caused by vibration or weight
loading to the spectrophotometer but this can be avoided with careful attention. Both
cases are discussed in the following subsections.

4.7.4.1 Electrical noise

There are subsystems in the spectrophotometer which are affected by electrical
disturbance. These are the chopper controller unit and the PbS detector system that
includes bias voltage supply, thermal electric cooler controller and amplifiers. Both
units are subject to drifting, for example the chopper manufacture's data specified
that the stability is 0.2%/hour. Since this is small the baseline correction technique
can overcome this drifting effect. The chopper and cooler controls have employed
closed loop controllers and if any excessive noise excites the controller, the controller
deviates from the set point (i.e. the set point for chopper is 375 Hz or cooler
temperature -10°C). The duration of this disturbance depends on the controller's
response time. If this happens during recording of a spectrum then a change to the
spectrum can be observed.

In the chopper controller, the variation happens frequently and set point may not be
regained for several minutes. The variations could be up to ± 2 Hz and the
contribution of this to a single wavelength measurement could be up to 0.75% as
illustrated in figure 4.20. Note that this error is greater than the standard deviation of
0.22% defined in the previous section.
The thermal electric cool controller in the PbS detection system, is generally less sensitive to electrical noise. But if a large spike appears in the mains supply, the controller would deviate from the control set point and may fail to return to the set point. The overall effect of this is illustrated below with two water spectra, one taken under normal detector operation and the other when the PbS system is out of set point control. It also seems that when the cooler’s controller is out of set point, the noise appearing at the output of the amplifier subsystem circuits is enhanced probably because they are part of the same electronic subsystem. It was decided that this excess of disturbance to the output was not acceptable. It is recommended that after a disturbance of this type the detector electronics should be reset by switching off and back on before any further spectral measurements are made.

Fig 4.20 a) Illustrates the transmission variation with chopper frequency for distilled water. Transmissions at 2214 nm are recorded for 5 different chopper frequencies and 12 measurements at each frequency. These have shown that the transmission decreases as chopper frequency reduces and deviation of 0.0075 (0.75%) can be observed for ± 2 Hz variation about 375 Hz chopper frequency. b) Transmission spectra of distilled water after an excessive electrical noise to the detector power supply affected the detector's closed loop control. The noise level output from the detector amplified increased and the dark current also increased.
4.7.4.2 Vibration

The NIR energy is focused onto a small 0.5 x 8 mm entrance slit. During any vibration around the surroundings of the spectrophotometer, the filament may not be accurately focused onto the entrance slit. Hence the detector will also respond directly to this vibration. Generally in order to achieve good stable spectral recording, it is necessary to keep the spectrophotometer isolated from any vibration. A second source of mechanical interference was because the chassis used in the spectrophotometer is not rigid. Any loading or stress applied to and around the spectrophotometer can tilt the optical path and move the filament image off the entrance slit and therefore NIR energies affect the detector output signal. It is recommend that the instrument be mounted on an isolated and sturdy bench that is standing on a solid ground.

4.8 Calibrating the spectrophotometer for determination of protein and RNA in yeast homogenate using PLS method

The spectrophotometer is individually calibrated for two different bioprocess contaminants before investigation of multiple contaminants calibration in the next chapter. The purpose is to identify the performance of quantitative analysis when the spectrophotometer is working in conjunction with a partial least squares (PLS) model described in chapter 3. The bioprocess contaminants chosen are protein and RNA in clarified yeast homogenate which are contaminants found in the alcohol dehydrogenase recovery process and required for process monitoring.

The range of concentrations in these samples have been designed such that they are related to the concentration range of interest in the process. However, there is an offset level because the samples used are prepared by adding a contaminant to
clarified yeast homogenate which already has some level of contaminants. The calibration of protein is at concentration 26 to 52 g L\(^{-1}\) and RNA at concentration 11.5 to 23 g L\(^{-1}\). The results have shown that the spectrophotometer and calibrated PLS models can successfully determine both protein and RNA at the stated concentration with standard deviations of 1.53 g L\(^{-1}\) and 0.49 g L\(^{-1}\) respectively.

4.8.1 Materials

Standards used were yeast ribonucleic acid (RNA, highly polymerised), bovine serum albumin (BSA, fraction V), both supplied by Sigma Chemical Ltd (Poole, Dorset, UK).

4.8.2 Methods

4.8.2.1 Clarified yeast homogenate

The yeast homogenate was prepared from packed bakers yeast, *Saccharomyces cerevisiae* supplied by Distillers Company Ltd. (Sutton, Surrey, UK). The bakers yeast was re-suspended in 100 mM phosphate buffer, pH 6.5, to a final cell concentration of 500 g wet packed weight L\(^{-1}\). The yeast suspension was disrupted using a high pressure homogeniser (Model Lab 40; APV Gaulin, APV, Crawley, Sussex, UK) for two discrete passes at 1200 bar g and maintained at approximately 4°C by cooling. Following homogenisation the homogenate was clarified using a centrifuge (Beckman, J2-MI) at 16,000 rpm for 0.3 h at 4°C. Finally, a volume of 40 ml of clarified yeast homogenate was pipetted out from the cell debris and the less dense lipid layer and stored at 4°C prior to usage.
4.8.2.2 Protein calibration and validation samples

Stock solutions of protein were prepared from standard materials. BSA was made up to a final concentration of 52 g L\(^{-1}\) with phosphate buffer (100mM, pH 6.5). The calibration samples had a composition of clarified yeast homogenate of 500 g L\(^{-1}\) (wet pack weight) bakers yeast and diluted with protein stock between 0 to 52 g L\(^{-1}\). The clarified yeast homogenate alone contained 52 g L\(^{-1}\) of protein and therefore the protein concentration range in the calibration samples were from 26 to 52 g L\(^{-1}\). The sample preparation is designed to maintain all other contaminants in calibration samples to the same level. The composition of twelve protein calibration samples is listed in Appendix A6, table A6.1.

There were two sets of validation samples, both were obtained from the same stock as the calibration samples. The first set of validation samples were used fresh from preparation and the second set of validation samples were stored at 4°C for approximately 24 hours before scanning of their NIR spectra.

4.8.2.3 RNA calibration and validation samples

Stock solutions of RNA were prepared from standard materials. RNA was made up to a final concentration of 23 g L\(^{-1}\) with phosphate buffer (100mM, pH 6.5). The calibration samples had a composition of clarified yeast homogenate of 500 g L\(^{-1}\) (wet pack weight) bakers yeast and diluted with RNA stock between 0 to 23 g L\(^{-1}\). The clarified yeast homogenate contains 23 g L\(^{-1}\) of RNA and therefore the net RNA concentration range in the calibration samples were from 11.5 to 23 g L\(^{-1}\) and all other contaminant levels are maintained at the same level. The composition of eight RNA calibration samples is listed in Appendix A6, table A6.2. The calibration samples were then divided into two sets, one set was used for calibration and the other set was used for validation. The NIR spectra of these validation samples were collected from fresh. There was no 24 hours storage samples for validation because it has been known that RNA in the samples would degrade with time.
4.8.2.4 NIR spectrum

The NIR spectra were obtained using the in-house built spectrophotometer, described in the previous sections. The three external control settings; lamp input voltage, detectors temperature and chopper frequency have been selected to 5 volts, -10°C and 375 Hz respectively. This scanning spectrophotometer was set to collect NIR radiation between 1900-2500 nm. The spectral data were collected at every 4 nm with a total scan time of 240 s.

4.8.2.5 Partial least squares and cross validation modelling

An IBM 486 compatible PC running at 33MHz was used for all analyses. All spectra were baseline corrected using a pre-scanned reference spectrum. The transmission spectra were then smoothed by a GCV/GA spline routine (Thornhill et al, 1994). All partial least squares calibrations and predictions were carried out by a commercial software Unscrambler version 5.0 (CAMO, Trondheim, Norway). In order to make use of all available spectra for calibration, cross validation (Martens and Næs, 1989) has been applied throughout. This optimises the PLS model for all but a few calibration spectra; the omitted spectra are used for validating the PLS model.

4.8.3 Calibration and validation of protein PLS model

4.8.3.1 NIR spectra of protein in clarified yeast homogenate

The 12 normalised transmission spectra of the protein calibration samples are shown in figure 4.21. In the spectra, protein response can be found in the region of the 2200 nm peak transmission. This has shown that the transmission decreases as protein in the calibration samples increases.
Fig 4.21 NIR spectra of twelve protein in yeast homogenate calibration samples. The calibration samples have added protein into clarified yeast homogenate, the total protein concentration range from 26 to 52 g L$^{-1}$ while other contaminants are maintained to the same level. The protein has responded to the NIR in the 2200 nm transmission peak and shown that the transmission decreased with protein concentration increased.

4.8.3.2 PLS modelling of protein in clarified yeast homogenate calibration samples

A PLS model is established using the twelve spectra of the calibration samples in the X-data, and the reference analytic Y-data are the known protein concentrations that were defined in the sample preparation. The result from the cross validation technique has selected a single principal component for this protein PLS calibration model.

The decomposition of the X-data by PLS has found the loadings spectrum related to the protein concentrations (figure 4.22). This loading spectrum clearly shows that the maximum loading at 2200 nm agrees with the previous observation in the transmission spectra of all protein samples (figure 4.21). There the strongest NIR signature is located in the same region in the NIR transmission spectrum.
Fig 4.22 a) Loadings ($p_1$) from the decomposition of the X-data matrix. This has high loading in the 2200 nm region which indicated NIR response to protein.

4.8.3.3 Validation of PLS protein calibration model using fresh and aged samples

The calibrated protein model is used to back predict the calibration samples (figure 4.23). This shows satisfactory protein predictions on all calibration samples when correlated to their known concentration. The protein model is also tested with the spectra of the fresh and aged validation samples, set $a$ and $b$ respectively. These spectra were collected immediately after calibration samples and 24 hours after preparation. The predictions of fresh validation samples ($a$) have again shown good agreement with their known protein concentrations and also matching the predicted protein found in the calibration sample set. In the 24 hours aged validation samples ($b$), the predictions are less accurate in some of the lower protein samples. These less accurate predictions have a high associated error bar, which indicates larger uncertainty in these prediction are expected by the PLS model. The prediction differences in the calibration and validation ($a$) samples are small for majority of
predictions. The reason for the larger differences found in aged validation \((b)\) samples is due to physical changes in the samples after the 24 hours delay.

The average uncertainty limits found in prediction of calibration, validation \(a\) and \(b\) sample set are 0.967, 1.395 and 1.645 \(g\ L^{-1}\) respectively. The standard error of prediction in all three set of samples is 1.532 \(g\ L^{-1}\). The conclusion is that both the spectrophotometer and PLS calibrated model provide satisfactory calibration, and the prediction accuracy is approximated to 1.53 \(g\ L^{-1}\).

![Figure 4.23](image)

Fig 4.23 Using a PLS calibrated model, protein concentration in the calibration and in the two sets of validation samples are predicted. The PLS model required a single principal component and it is established using the NIR spectra of the 12 calibration samples in X-data and the known amount of protein concentrations in the reference Y-data. The calibration samples are clarified yeast homogenate with various additions of protein. These twelve calibration samples have protein concentration ranges from 26 to 52 \(g\ L^{-1}\) while other contaminants in the clarified yeast homogenate are maintained to the same level. The two sets of validation sample are prepared from the same batch of samples used in calibration but the spectra were recorded immediately after (set \(a\)) and approximately 24 hour after (set \(b\)) recording of the calibration samples.
4.8.3.4 Identification of the effect of spectral differences between fresh and aged samples

The spectral difference between the fresh calibration samples and 24 hours aged samples (i.e. the ‘b’ validation spectra) are studied. This identifies any spectral differences that may have caused by physical changed in the samples. The figure 4.24 have shown that the aged samples 2 to 6 have larger negative spectral differences than the acceptable band (±0.22%) of spectrophotometer measurement defined in section 4.7.3. These samples 2 to 6 have also been predicted by the PLS model with larger protein concentration than that found in fresh samples. In sample no. 8, prediction has found lower protein than that in fresh sample and it is reflected by a positive spectral differences.

![Spectral differences between normalised fresh and 24 hours later samples](image)

Fig 4.24 The spectral differences of fresh and 24 hours protein calibration samples. The acceptable band of standard deviation of measurement error in the spectrophotometer is also shown. The 24 hours aged samples no. 2 to 6 have shown have the most significant differences with the calibration samples. These have differences greater than the acceptable band. The results suggest there are differences between fresh and aged samples, and the PLS model would give less accurate prediction on aged samples.
The conclusion is that spectral differences have been observed between fresh and aged samples. Therefore the confidence on using the PLS model established from fresh calibration samples and applied to predict fresh samples would be higher, whereas the confidence would be reduced when using the same model to predict aged samples.

4.8.4 Calibration and validation of RNA PLS model

4.8.4.1 NIR spectra of RNA in clarified yeast homogenate

There are 8 RNA calibration samples with concentration ranging from 11.5 to 23 g L\(^{-1}\) of RNA in clarified yeast homogenate and other contaminants have been maintained to the same levels. The NIR transmission spectra of these RNA calibration samples are shown in figure 4.25. It has been found that the RNA contaminant also has a signature in the 2215 nm transmission window. Like those observed with protein, the NIR transmission decreased as RNA increased, however this decrease is not consistent for all calibration samples. Since the variation in the transmission is small the inconsistent transmission with RNA concentration could have been complicated by the error of the spectrophotometer measurements that is indicated by the error bar shown in the inset figure.
Fig 4.25 The NIR transmission spectra of eight RNA calibration samples. These calibration samples have RNA concentration ranging from 11.5 to 23 g L\(^{-1}\) in clarified yeast homogenate. RNA has shown an NIR signature in 2215 nm region and the transmission is highest with low RNA concentration and lower vice versa. The small figure, there are inconsistent NIR responses to RNA concentration indicating that measurements may be complicated by the spectrophotometer sensitivity.

4.8.4.2 PLS modelling of RNA in clarified yeast homogenate calibration samples

The eight NIR spectra of the RNA calibration samples are used in the X-data matrix for PLS calibration model, the reference analytic Y-data are the known RNA concentration in these eight calibration samples. In the established PLS RNA model, 3 principal components are required to describe the relationship between the spectra and their RNA concentrations. The three loadings spectra in this PLS calibration are shown in figure 4.26.

In the first principal component loading plot, the loadings are large in the region about 2215 nm (figure 4.25) and somewhat similar to the loading plot that was observed in protein model (figure 4.22). However, there is a positive loading in \(p_1\) at
about 2090 nm which suggest the RNA signature has an influence at that wavelength. Other loading in $p_2$, $p_3$ have suggested the whole of the collected NIR spectrum 1900 to 2500 nm is required to establish the RNA calibration model.

![Graph]

Fig 4.26 The loadings of the 3 principal components used in the PLS calibration model for RNA are shown. Loading $p_1$ have the similar spectral feature those to observed in PC1 protein model, expect that there is a positive loading in the 2090 nm region. The loadings in the second and third principal components suggested the presence of RNA gives spectral features throughout the NIR spectrum.

### 4.8.4.3 Validation of PLS RNA calibration model

The model is used to back predict the RNA concentration in its calibration samples and the results are compared against their known RNA concentration in figure 4.27. The figure has also included the prediction of RNA in the validation samples and these have shown that the calibrated PLS model has satisfactory predictions of both set of samples. The near unity slope value of the regression line in the figure has indicated the overall predictions are balanced with their known RNA concentration. The standard error of predictions to both sets of samples is 0.488 g L$^{-1}$. 


The conclusion is that the in-house spectrophotometer together with the use of PLS model has provided satisfactory calibration and prediction. Since the standard error of prediction to both sets of samples is 0.488 g L\(^{-1}\), this gives a guideline of RNA prediction accuracy to ± 0.49 g L\(^{-1}\).

![Graph showing RNA prediction concentration against known RNA concentration](image-url)

**Fig 4.27** The RNA predicted concentration are plotted against their known RNA concentration. The regression line in the figure is representing the correlation of all predictions to the known RNA concentrations. The PLS model required three principal components and it is established using the NIR spectra of the 8 calibration samples in X-data and the known amount of RNA concentrations in the reference Y-data. The figure has shown satisfactory prediction to both back predicted calibration samples and the validation samples.
4.9 Conclusion

This chapter has described three phases of implementation of a budget NIR spectrophotometer for bioprocess monitoring. These are the design and optimisation of a budget NIR spectrophotometer, definition of the spectrophotometer’s detection performance and the use of partial least squares (PLS) in conjunction with the spectrophotometer for analysis of single bioprocess contaminants. The following are the conclusions to each one of the three phases.

4.9.1 Design and optimisation of a budget NIR spectrophotometer

- The design specification of the in-house build spectrophotometer has guidelines defined by the preliminary investigation of NIR signatures of the bioprocess contaminants of interest. These guidelines were identified by a high quality commercial spectrophotometer (NIRS 6500, Perstrop Analytical, Berks, UK).

- The design of the spectrophotometer was under a budget constraint. Therefore, the majority of components used are inexpensive and the instrument’s performance has been focused towards the application, rather than being more general purpose. For example, instead of diode array detectors, a scanning spectrophotometer is configured by used of a single element PbS detector and a stepper motor driven stage. While the former can record a spectrum within a few seconds, the scanning spectrophotometer required 240 seconds.
The spectrophotometer is capable of both transmittance and reflectance measurements. However, only transmittance measurements have been tested and applied for all subsequent work.

Although the design has been targeted for a budget spectrophotometer it allows the option for future improvement. For example, moving parts in the spectrophotometer can be minimised and the spectrum collection time reduced by replacing the single element PbS detector and the stage with PbS diode array detector.

A spectrograph using a holographic grating has been selected for the spectrophotometer. This was because its linear dispersion output can accommodate a diode array detector when upgrade is decided. Although the holographic grating is not blazed at any wavelength it offers low stray light when compared to a ruled grating.

A low cost PbS detector is implemented. This has a signal to noise ratio of approximately 450:1 (estimated from 0.22% error in section 4.7.3) comparing to an well expensive high performance spectrophotometer with signal to noise ratio of 1000:1.

For the application of bioprocess contaminants monitoring, the limited signal to noise ratio from a low cost PbS detector and detector electronics have made conventional ratio technique or baseline correction unfeasible. In order to enhance the NIR signal of interest, both the optical source energy and sample pathlength have been increased such that the full scale spectral variation falls into the detection window of the detector. The increased in optical source energy can saturate the detector, so the method of recording of the distilled water spectrum for use as the reference spectrum for baseline correction was used in the subsequent sample measurements.
• The overall cost of the components used in the spectrophotometer built is seven times less than the purchase of a commercial diffraction grating NIR spectrophotometer.

4.9.2 Performance of the in-house budget spectrophotometer

• The implemented spectrophotometer records a spectrum in the range 1900-2500 nm. This has been identified by comparison of water transmission spectrum measured from this spectrophotometer and a commercial spectrophotometer (NIRS6500, Perstrop Analytical, Berks, UK).

• A complete spectrum scan requires 480 seconds with spectral data collected at every 2 nm. In order to reduce the scanning time, the spectral data are collected at every 4 nm, halving the scanning time (240 seconds) for all subsequent work.

• The recorded sample spectrum is baseline corrected with a reference spectrum of distilled water that is scanned before each sample.

• A reference spectrum is recorded before recording of every sample spectrum. Therefore, the minimum time required to collect both reference and sample spectra would be $2 \times 240$ seconds plus appropriate sample exchange time. In practice the total time required 8 to 9 minutes.

• The quality of the recorded spectrum is affected by excessive electrical supply noise or vibration to the spectrophotometer. It is necessary to observed if any significant change has appeared in the recorded spectrum. Alternatively, if the affected spectrum is used in PLS predictions, then outliers would be suggested by PLS.
• The final spectrum is smoothed by a GA/GCV spline function to give the 1900-2500 nm spectral data at every 6 nm. This is necessary to remove high frequency ripple appearing on the spectrum.

• Spectral reproducibility studies have found the maximum standard deviation in energy measurement within a day is 0.22% and 0.44% between days of measurements. Wavelength reproducibility is sufficient since a high precision stage has been used.

• The comparison of the in-house built spectrophotometer with two commercial instrument is given in Appendix A 4.
4.9.3 Using the spectrophotometer in conjunction with PLS for quantitative analysis of bioprocess contaminants

• Two PLS calibrated models have been established with the spectra recorded from the in-house build spectrophotometer. These are models for quantitative analysis of protein and RNA in clarified yeast homogenate and these have demonstrated that the NIR spectrophotometer is sensitive enough to detected the contaminants of interest.

• The PLS models are calibrated by samples with protein contaminants in clarified yeast homogenate at concentration ranged from 26 to 52 g L\(^{-1}\) and RNA contaminant in clarified yeast homogenate at concentration ranged from 11.5 to 23 g L\(^{-1}\).

• The spectrophotometer together with the PLS models can predict protein and RNA contaminants within the calibrated concentration range. The standard error of prediction are ± 1.53, ± 0.49 g L\(^{-1}\) for protein and RNA respectively.

• The calibration models are achieved with fresh calibration samples. In quantitative analysis of new samples, accuracy and small uncertainty of prediction can only be maintained with fresh samples. The aged samples have different spectral characteristic compared to the fresh samples. Thus the contaminants prediction in the aged sample may not be a true representation of the contaminants level.
5. Calibration and prediction of yeast homogenate samples

5.1 Summary

The chapter describes an application of near infrared (NIR) spectroscopy to characterisation of a complex biological process stream. Its use is for the monitoring and control of a process in which alcohol dehydrogenase (ADH) is purified from homogenised bakers yeast (chapter 2). This is achieved by selective removal of contaminants by flocculation. The homogenate contains cell debris, protein, RNA and many other contaminants. An instrument has been developed (chapter 4) for sensitivity in the region of the NIR spectrum (from 1900 to 2500 nm) where preliminary work found NIR signatures from cell debris, protein and RNA. Calibration models have been derived using principal component methods for concentrations of debris, RNA and protein such as would be found after the flocculation process.

Two types of samples are used for calibrating the NIR instrument (Yeung et al., 1996). In one case samples were prepared by adding materials to the homogenised yeast representation of the contaminants. In the other samples were taken from the process stream after flocculation and floc removal. In the former case the contaminant levels are well known but are outside the range of interest, in the latter there is uncertainty of analysis of contaminant level but the calibration is in the range of interest. The NIR instrument together with the calibration models has demonstrated potential in rapid monitoring of contaminants.
5.2 Introduction

This chapter presents the use of a low cost NIR spectrophotometer (chapter 4), developed for laboratory studies for rapid analysis of bioprocess stream, and its potential use for on- or at-line monitoring of complex bioprocesses. The recovery of a yeast intracellular enzyme alcohol dehydrogenase (ADH) from yeast cell homogenate by a selective flocculation process has been chosen for this exploration (Salt et al, 1995). The ADH is an intracellular protein produced in *Saccharomyces cerevisiae* (bakers yeast). In the recovery process, bakers yeast has to be disrupted to release the ADH. This process also releases other unwanted contaminants into the process stream. These include cell debris, protein, RNA, DNA and lipid all of which considerably affect the performance of subsequent high resolution purification operations. It is the purpose of the recovery process to remove these contaminants (Atkinson and Jack, 1973) in the early stage of recovery. If NIR monitoring is able to identify and determine the amount of residual contaminants in a process stream this would give two immediate advantages. Firstly, NIR can be an indicator for the next or final stage of a recovery process. For example if the contaminant levels are high then the stream should be stopped from going onto a high resolution chromatography, since the columns used can be easily damaged by contaminants. Secondly, NIR gives important information for potential control of the process to maximise contaminants removal. This would allow optimisation of the whole recovery process.

Two multivariate calibration experiments using the partial least squares method are described. Each experiment provided calibrations of cell debris, protein and RNA in clarified yeast homogenate. In the previous chapter, the developed spectrophotometer has been shown in successful calibration of single contaminant analysis. In this chapter, the first calibration experiment shows that the developed spectrophotometer is capable for use in multiple contaminants analysis. This is achieved with calibration samples defined by a full factorial experiment design.
Then the second calibration experiment used real flocculation process stream samples. These calibration samples covered the concentration of contaminants of interest. In these two experiments, calibrations are accomplished with partial least squares (PLS) technique (chapter 3).

Both calibration experiments are designed for analysis of contaminants in a 250 g L\textsuperscript{-1} (wet weight pack) yeast homogenate flocculation process. However, the nature of the first calibration experiment have made it less effective for process monitoring while the second calibration experiment is more successful.

In the following, the two calibration experiments known as the add-back and process stream calibration are described. The PLS models established from the two calibrations are used on process monitoring of contaminants and the results are compared with conventional assays. Then the selection of optimum process control set point is illustrated.

5.3 Calibration strategy

The selection of calibration samples is one of the challenges in setting up NIR analysis on any constituent or contaminant in a bioprocess (Marten and Næs, 1989). These calibration samples should preferably be representative of the samples to be analysed in the future. Too few samples will not provide a good distribution and too many will result in unnecessary effort in performing the calibration assays. The task is to quantify the amounts of contaminants or contaminants representatives which are cell debris, protein and RNA in the process stream.

In this work, two sets of calibration samples are used for establishing calibration models on cell debris, protein and RNA. In the first set of calibration samples, controlled levels of contaminants are added back into the clarified yeast homogenate so that contaminants in the calibration samples are defined. These samples are termed add-back samples. Based upon this technique a full factorial design of
calibration samples is prepared. The second set of calibration samples are samples similar to those obtain during the flocculation process and are termed process stream samples. Assays on protein and RNA of these calibration samples are necessary in order for calibration modelling. However, it is important to note that assays are very time consuming and can be inaccurate.

The size of calibration samples is also required to be considered. This is mainly dependent upon the total number of calibration samples that can be handled within each day. It is important that samples do not suffer from ageing (see section 4.8.3.3) to eliminate any unaccountable variability in the calibration samples. In Chapter 4, the specification of the NIR spectrophotometer showed the scan time of 9 minutes are required for collection of reference and sample spectra. These restricted the maximum calibration samples to approximately 40 samples for each day. It has also been found that ageing does not become a problem until after 24 hours. This ensured the collection of up to 40 NIR spectra a day with no significant ageing effect on the calibration samples.

### 5.3.1 Ideal contaminants calibration

Cell debris, protein and RNA are the primary contaminants for PLS calibration. The ADH recovery process uses for 250 g L$^{-1}$ w/w yeast homogenate and NIR monitoring is used in the first instance for clarified yeast homogenate after flocculation and floe removed (as shown in figure 2.3).

In an ideal calibration, the NIR should calibrate for contaminants from zero to maximum possible contaminants level. Since the first clarified yeast homogenate only had reduced amount of cell debris and lipid reduced (see section 5.4.2), thus the two other contaminants of interest are in their highest concentration. For example, in 250 g L$^{-1}$ (wet packed weight yeast) yeast homogenate contains approximately 26 g L$^{-1}$ (dry weight) of protein and 11.5 g L$^{-1}$ (dry weight) of RNA. Thus calibration
samples should therefore include a range of samples with zero protein to above 26 g L\(^{-1}\) protein. Similar calibration applied for both cell debris and RNA.

![Diagram](image)

Fig 5.1 Illustrated the ideal calibration and process monitoring contaminants concentration region. The calibration region used in add-back contaminants experiment is shown twice the concentration of the ideal case. Note that the calibration samples in the add-back experiments has the same variation of contaminant level as in the ideal case.

### 5.3.2 Add-Back contaminants calibration

Cell debris, protein and RNA are the primary contaminants for NIR calibration. In an ideal calibration, the NIR should calibrate for contaminants from zero to the maximum expected level (as defined previously). NIR absorption peaks are highly overlapping and there is no unique peak for each contaminant, therefore PLS calibration with NIR spectra is required. Since the spectral response for one contaminant may be influenced by another contaminants, the calibration modelling requires a mixture of composition from the three primary contaminants in the
calibration samples. A full factorial design experiment is adapted in the *add-back* calibration.

However, other than for debris it is often difficult to achieve selective and accurate removal of any one contaminant from the yeast homogenate. Instead contaminants are added back to the clarified yeast homogenate from which cell debris has been reduced of the yeast homogenate. The amounts of added protein and RNA are designed to increase the contaminants present by a similar value as those present in the clarified homogenate. For example, the amount of added protein ranges from 0 to 26 g/L. Thus the differences between the maximum and the minimum individual type of contaminants in the calibration samples are similar to those in the process. However, a base level of contaminant is already present in the clarified homogenate. Hence while the contaminant levels are very well defined, their concentrations are outside the range of interest for the process (see figure 5.1). Also the contaminants added back are not identical in chemical or physical conformation to the contaminants naturally in the homogenate. For instance, three concentrations of cell debris contaminant, three of protein and four of RNA have been chosen in this full factorial *add-back* experiment.

Although this *add-back* contaminants calibration does not calibrate the NIR spectrophotometer for analysis of ideal process samples, the design of the experiment does keep the variation in the contaminants similar to the variation in the ideal case. Therefore this calibration is also used to assess the prospect of the in-house build spectrophotometer for measurement of ideal process samples, as well as a good experiment to examine PLS calibration of multi-contaminants.

An attempt was made to use a calibrated NIR based on this *add-back* contaminants calibration to determine the analysis of some previously unseen process samples. Since these samples have contaminants in the lower concentration, the prediction of contaminants required back-extrapolation from the calibration model. The results of these analysis are shown in the later section on *process monitoring* (section 5.5.4).
5.3.3 *Process stream* samples calibration & validation

A second set of calibration samples, called the *process stream* samples, are similar to those found in a flocculation process. Here, flocculated samples are prepared on laboratory scale. These are samples that contain the ideal contaminants concentration for calibration. Samples are obtained from those found in a flocculation process. The flocculation process involved controlled feed of Polyethyleneimine (PEI) into the process stream. Here, flocculated samples are prepared on laboratory scale. These flocculated samples are spun to removed the contaminants in the solid, and the remaining supernatants are used for calibrations.

Additional *process stream* samples are prepared, these are used as validation samples for testing of the NIR spectrophotometer in conjunction with the two types of calibration; *add-back* contaminants and *process stream* samples calibrations. There are total of four sets of *process stream* samples and they are similarly prepared. The first two sets are used for calibration and they are denoted as *process stream* ‘A’ and *process stream* ‘B’ in later sections. The remaining two sets of *process stream* samples are used for validation and they are denoted as ‘V1’ and ‘V2’ validation samples.
5.4 Materials and methods

5.4.1 Standard materials

Standards used were yeast ribonucleic acid (RNA, highly polymerised), bovine serum albumin (BSA, fraction V), both supplied by Sigma Chemical Ltd (Poole, Dorset, UK). The assay chemicals were orcinol, ferric chloride, perchloric acid. These were all supplied by BDH Chemicals Ltd (Poole, Dorset, UK). Polyethyleneimine was supplied by Fluka Chemicals (Dorset, UK).

5.4.2 Clarified yeast homogenate and washed debris

The yeast homogenate was prepared from packed bakers yeast, *Saccharomyces cerevisiae* supplied by Distillers Company Ltd. (Sutton, Surrey, UK). The bakers yeast was re-suspended in 100 mM phosphate buffer, pH 6.5, to a final cell concentration of 500 g wet packed weight L⁻¹. The yeast suspension was disrupted using a high pressure homogeniser (Model Lab 40; APV Gaulin, APV, Crawley, Sussex, UK) for two discrete passes at 1200 bar g and maintained at approximately 4°C by cooling. Following homogenisation the homogenate was clarified using a centrifuge (Beckman, J2-MI) at 16,000 rpm for 0.3 h at 4°C. Finally, a volume of 40 ml of clarified yeast homogenate was pipetted out from the cell debris and the less dense lipid layer and stored at 4°C prior to usage. The cell debris pellet was washed by re-suspending the pellet in phosphate buffer by vortexing and re-centrifuging at 16,000 rpm for 0.3 h at 4°C. The supernatant and lipid layer were discarded and the pellet of cell debris was mixed to homogeneity. This was then re-suspended in phosphate buffer to give a concentration of 150 mg mL⁻¹ (wet weight) and it was used as cell debris contaminants in the add-back calibration experiment.
5.4.3 Add-back contaminant samples

Stock solutions of protein and RNA were prepared from standard materials. BSA was made up to a final concentration of 156 mg mL\(^{-1}\) with phosphate buffer (100mM, pH 6.5). RNA was made up to a final concentration of 69 mg mL\(^{-1}\) with the same phosphate buffer (100mM, pH 6.5), and cell debris stock is as previously described. Using the clarified yeast homogenate and the three stock of suspensions, 36 different calibration samples were prepared. These are the add-back calibration samples and the composition of their contaminants was based on a full factorial design experiment to give a maximum variation for calibration modelling. The combination of contaminants added to clarified yeast homogenate for each calibration sample is listed in the Appendix A7 table A7.1, and table A7.2 is the expected contaminants concentration in these samples.

5.4.4 Process stream samples

Polyethyleneimine (PEI) was diluted into phosphate buffer (100mM, pH 6.5) at 2\% w/v. The un-clarified yeast homogenate described above was diluted 1:1 with phosphate buffer to give homogenate equivalent to 250 g wet packed weight yeast L\(^{-1}\). The PEI stock was added to this homogenate to cause flocculation. Four sets of samples were prepared and each set contained ten samples with various extents of flocculation. The PEI stock concentrations used for each set of samples are 0, 2, 6, 10, 15, 20, 25, 30, 35 and 40 percents (volume/volume). Two sets of these process stream samples are used for calibration. The remaining two sets were used for model validation purposes during process monitoring experiments.
5.4.5 Analytical Data

In the add-back calibration experiment, analytical data of all 36 calibration samples are obtained in two ways. Firstly, cell debris (turbidity) was ascertained by measuring the supernatant absorption of an appropriate dilution of the samples at 650nm against a buffer blank. This is referred to as the optical density (OD) measured at 650nm. Secondly for both protein and RNA, these were taken to be the expected concentrations in the add-back calibration samples (section 5.4.3 or Appendix A7, table A7.2).

5.4.6 Assays for biological contaminants

Protein was measured using the dye-binding method of Bradford (Bradford, 1976).

RNA was assayed using a method based on the orcinol assay (Bulmer, 1992, Munro & Fleck, 1966). The method was adapted for yeast as follows. Orcinol reagent was prepared by dissolving orcinol 3 g L\(^{-1}\) in concentrated hydrochloric acid to which was added ferric chloride (10% w/v, 1 mL). The orcinol reagent was stored in the dark and used fresh. Samples (100 μL) were precipitated with 60% perchloric acid (100 μL) in an Eppendorf tube and stored at 4°C for 24 h and then centrifuged (13,500 g, 0.12 h). The supernatant (100 μL) was mixed with sodium hydroxide (NaOH, 2M, 100 μL) and incubated (2 h, 37°C). Orcinol reagent (800 μL) was added to the samples which were then placed in boiling water for 0.3 h, then cooled and centrifuged (13,000 rpm, 0.6 h). The samples were read against a reagent (orcinol) blank at 665nm.

Cell debris (turbidity) was ascertained by measuring the supernatant absorption of an appropriate dilution of the samples at 650nm against a buffer blank. This optical density measurement has a measurement error of 6%.
All protein and RNA assays were performed in triplicate, the resultant standard deviation being 4% for protein and 9.5% for RNA. Numerical results of the assays are listed in the Appendix A8.

5.4.7 NIR Spectroscopy & Multivariate analysis

The NIR spectra are obtained using the in-house built spectrophotometer described in chapter 4. The scanning procedures are the same as that described in section 4.8.2.4. Each spectrum is recorded and smoothed for wavelength 1900 to 2500 nm and data interval at every 6 nm.

Multivariate analysis used partial least squares (PLS) routine and the cross validation technique in all calibration (chapter 2). The complete procedure is also used in section 4.8.2.5.
5.5 Results and discussion

5.5.1 Measurements and measurements variability

Before the start of any PLS calibration, concentrations have to be identified. These are used as the reference analytic data (Y-matrix) for the PLS calibration models. Optical density measurements and assays have been performed on all *add-back* and *process stream* calibration samples. In the following, all measurements and assays are represented. Some practical measurements variability in these calibration samples are also observed. It has been found that in the *add-back* calibration experiments, the expected rather than the measured protein and RNA concentration data should be used in the PLS calibration. This was because the assays on the *add-back* calibration samples are less consistent at higher concentration of contaminants (i.e. samples with contaminants added), while assays on the *process stream* samples (i.e. samples without contaminant added) are more consistent. Since the expected concentration data have smaller errors than the assay results therefore the expected concentration data are used in the *add-back* calibration. However, in *process stream* calibration, only assays results are available to PLS calibration. Assays results are more consistent in the lower concentration of contaminants as those in the *process stream* calibration samples, so there is no problem with using the assay result in the *process stream* calibration.

5.5.1.1 Optical density measurements with *add-back* calibration samples

In all of the cell debris quantitative analyses, optical density (OD) measured at 650 nm is used to give indication on the level of cell debris concentration. The optical density measured for all of the 36 *add-back* calibration samples are shown with their calibration samples identification in figure 5.2. The small error bar attached to each
measurement is the averaged standard deviation of ±0.05 Au found in all triplicate measurements. The cell debris concentration of these calibrations samples is well defined by the factorial experiment design. The first group of twelve samples should have the lowest optical density, while the next and last group of twelve samples should have medium and high optical density measurements respectively. The figure has shown that the average optical density of these three groups are 0.58, 1.05 and 1.55 Au. Since the sample preparation is such that the cell debris within each group are constant, the standard deviation observed within each group gives an indication of the measurement and sample preparation variability in optical density. In the three groups of optical density measured, their standard deviation are 0.035, 0.0613 and 0.069 Au. These groups of standard deviation are illustrated as error bands to their group averaged optical density.

Fig 5.2 Optical density (OD) measured for all 36 add-back calibration samples and they are shown with their sample identification number. The error bar associated with all measurements is the average standard deviation found in triplicate measurements of optical density. The calibration samples have three groups of optical densities, the average optical density and standard deviation to each group are indicated by a ♦ symbols and error bands. From the error bands, outliers are observed to be samples no. 2, 13, 27, and 35.
Using the error band, calibration samples with outlier optical density measurements are identified. These are sample no. 2 in the first group, 13 in the second and 27, 35 in the last group of optical density. There are two possible causes for these outliers. Outliers may have been caused by sample preparation error, that is the calibration samples do not contained the cell debris concentration as expected to their group, and therefore the reported optical density is the true measurement of the calibration. Alternatively, outliers are due to optical density measurement error. Later in section 5.5.2.1, the prediction of optical density in these add-back calibration samples suggested that the outliers in figure 5.2 are due to optical density measurement errors. This confirmed that the cell debris sample preparation for these add-back calibration samples was well controlled.

### 5.5.1.2 Protein assays on add-back calibration samples

The result from the triplicate Bradford assays on each one of the 36 add-back calibration samples are shown in figure 5.3. The small error bar associated with each result indicates the average standard deviation of ±0.5 g L\(^{-1}\) found in triplicate assays. As the result of the full factorial experimental design, it is expected to see 9 clusters of samples (figure 5.3) and there are 3 clusters with the same group of low protein, and the other 6 clusters should divided into medium and high protein concentrations groups.

In figure 5.3, each clusters averaged protein concentration and standard deviation are shown by a ◊ symbol and error band. Clusters 2, 3 and 6 have shown a very large standard deviation within the clusters and the best protein group separation are clusters 4, 5 and clusters 7, 8, 9. The expected protein concentration in the low, medium and high protein groups are 26, 39 and 52 g L\(^{-1}\) (Appendix A7, table A7.2). If these are compared to clusters 7, 8 and 9 which have mean protein concentration 39.2, 55.7 and 68.2 g L\(^{-1}\), it can be seen that there is an offset between the expected and assayed protein by greater than 13 g L\(^{-1}\). This offset could have been simply caused by either under estimation of the expected protein concentration or there is an offset in the calibration of the Bradford assays.
Since the protein sample preparation in the *add-back* calibration experiment is thought to have been consistent, the expected protein concentration will be used in the later PLS calibration.

![Protein concentration of all 36 add-back calibration samples established by Bradford assays are shown with their sample identification number. The error bar associated with all measurements are the average standard deviation found in the triplicate measurements. There are 9 clusters and each represents 4 calibration samples that should have the same levels of protein. Three different levels of protein concentration are expected at 26, 39 and 52 g L\(^{-1}\) of protein. The average protein concentration and standard deviation within each cluster are shown with \(\Diamond\) symbols and error bands. The average clusters protein suggested that there is an offset between the expected and the assayed protein. Also, the high standard deviation in the clusters suggested that the expected protein concentration rather than the assayed concentration should be used for PLS calibration.](image)

**Fig 5.3** Protein concentration of all 36 *add-back* calibration samples established by Bradford assays are shown with their sample identification number. The error bar associated with all measurements are the average standard deviation found in the triplicate measurements. There are 9 clusters and each represents 4 calibration samples that should have the same levels of protein. Three different levels of protein concentration are expected at 26, 39 and 52 g L\(^{-1}\) of protein. The average protein concentration and standard deviation within each cluster are shown with \(\Diamond\) symbols and error bands. The average clusters protein suggested that there is an offset between the expected and the assayed protein. Also, the high standard deviation in the clusters suggested that the expected protein concentration rather than the assayed concentration should be used for PLS calibration.

### 5.5.1.3 RNA assays on *add-back* calibration samples

The RNA concentration found in the *add-back* calibration samples by orcinol assays are shown in figure 5.4. Because the assays are time consuming and complicated, only a single assay is taken for each calibration sample. In this figure, all the calibration samples that have the same expected RNA concentration are shown
grouped together. These are separated into 4 groups and each have a different level of RNA concentration defined by the factorial experimental design.

Fig 5.4 The RNA concentration in the 36 add-back calibration samples assays by the orcinol method are shown by △ symbols. All 36 calibration samples have been grouped to 4 different levels, in each level samples are corresponded to a RNA concentration defined by the full factorial experimental design listed in table A7.2. The averaged of the assayed RNA in each individual level are shown by ○ symbols, and the expected RNA concentration in each level is shown by ♦ symbols.

The average concentration and standard deviation within each group are calculated, shown by ○ symbols and error bands. There is good agreement with the expected RNA concentration (♦ symbols) in the first three levels. But the standard deviations found in calibration samples in the three higher levels of RNA are large. It was because of this large standard deviation that the expected RNA concentration data are used in PLS calibration model later.

The standard deviation in each of the four levels of RNA assays are shown in figure 5.4. An increase in the standard deviation is shown as the amount of added RNA contaminant increases. These results have suggested that the orcinol assay yielded larger inconsistency at higher concentration of RNA but the assay is acceptably consistent in lower level of RNA contaminants.
5.5.1.4 Optical density of two set of process stream calibration samples

The optical density of the calibration samples taken from two process streams ‘A’ and ‘B’ (section 5.3.3) are measured (figure 5.5). In both sets of the process stream calibration samples, the optical density shows that the cell debris is gradually decreased. These have shown the cell debris is removed by PEI flocculation and clarified by centrifugation. The removal of cell debris increases further as PEI concentration is between 5 to 15 percent, then the cell debris in both processes has reached virtually minimum.

![Graph showing optical density vs PEI concentration](image)

**Fig 5.5** As a measure of cell debris concentration, the optical densities are measured for two sets of process stream samples A and B and shown by ■ and ▲ symbols respectively. The solid and dash lines are used to purely to illustrate the cell debris removal according to the PEI concentration increase. In both process stream, the calibration samples started with high optical density and gradually reduced. At PEI concentration between 5 to 15 percent, the optical density reduced more rapidly and cell debris in the calibration samples has reached the minimum at about 20 percent of PEI concentration. Then the removal of cell debris became less effective with higher PEI concentration. The error bar associated with each measurement is the standard deviation in triplicate optical density measurements. The removal of cell debris in the two process streams have shown consistent results.
5.5.1.5 Protein assays of two sets of process stream calibration samples

The protein assays by the Bradford method for the two sets of process stream calibration samples are shown in figure 5.6. In the process stream samples set ‘A’, the protein in the clarified yeast homogenate is at 24.8 g L\(^{-1}\) when PEI concentration is zero. This is actually very close to the expected protein concentration 23 g L\(^{-1}\) in this clarified yeast homogenate without any flocculation. Then the protein gradually decreased until PEI has reached 20 percent (v/v). There are indications in both sets of calibration samples that protein level increases from 35 percent (v/v) of PEI concentration. This effect is thought to have been caused by over-flocculation and shows the removal of protein contaminant is less effective at such PEI concentration.

Similar protein removal is also found in process stream samples ‘B’. There is one outlier in these process stream calibration samples which was later verified as being due to assay error. Note that the dashed line (used for illustration of protein removal) has been shown deliberately without the influence of the outlier since other than the single outlier, all assays have shown consistent measurements of protein.
Fig 5.6 The measured protein concentration of the two sets of process stream calibration samples A and B are shown by □, △ symbols respectively. The error bars represent the average standard deviation found in the triplicate Bradford assays for these measurements. The solid and dashed lines are purely to highlight the protein contaminant removal with increase in PEI concentration. Both process streams show that the protein concentration decreased as PEI increases 0 to 15 percent. Above 20 percent of PEI the protein removal becomes constant. There is an outlier in the second set of process stream calibration samples. This was later confirmed as being due to assay error. Overall, these protein reduction in the two process streams are as expected and the results are consistent enough for calibration purposes in the absence of any other information about protein concentration.

5.5.1.6 RNA assays of two sets of process stream calibration samples

The removal of RNA contaminants in two sets of process stream calibration samples are shown by the results of their orcinol assays (figure 5.7). The RNA contaminants decreases with increasing PEI concentration. This has illustrated the importance of using PEI flocculating reagent to enhance contaminants removal. The two initial process stream samples have RNA contaminant at 13.5 and 12.2 g L\(^{-1}\). These are
relatively close to the expected 11.5 g L\(^{-1}\) of RNA and that the higher error associated to these assays should also be considered.

The RNA assays for the two sets of process stream samples have shown the expected trend in removal of contaminants. The consistency of these assays have shown that the results can be use for PLS calibration.

![Graph showing RNA concentration against PEI concentration for process streams A and B.](image)

Fig 5.7 The measured RNA concentration in the two sets of process stream calibration samples A and B are shown by ■, ○ symbols respectively. The error bar indicated the high error in triplicate orcinol assays, the solid and dash lines are used purely to highlight that the reduction in RNA contaminant is function of PEI concentration. Both process streams have shown immediate removal of contaminant from 0 to 15 percent of PEI and further PEI increased shown less effective RNA removal. Overall, the RNA reductions in the two process streams are as expected and the results are consistent enough for calibration purposes in the absence of any other information about RNA concentration.
5.5.2 PLS calibration models using add-back calibration samples

The spectra of the add-back calibration samples are used for PLS calibration. This exercise has a primary task to identify whether the spectra collected from the in-house built spectrophotometer are sufficient in quality for calibration of multiple contaminants. The following sections show the spectra of all add-back calibration samples and the results of all the PLS calibration. These demonstrate that satisfactory calibrations are achieved for all three cell debris (characterised by optical density), protein, and RNA contaminants. Thus the spectrophotometer and the calibrated PLS models are suitable for quantitative analyses of multiple contaminants.

5.5.2.1 NIR spectra of the add-back calibration samples

The NIR transmission spectrum of all 36 add-back calibration samples are shown in figure 5.8. The spectra can be seen to be separated into three groups. These are groups of samples with low, medium and high concentration of cell debris. The samples with low cell debris are samples no. 1-12, medium cell debris are samples no. 13-24 and high debris with samples no. 25-36 (see Appendix A7).

The peak transmissions from each of the three groups are shown in the right windows of figure 5.8. In the top window, spectra of the first twelve calibration samples are seen to be further divided into three levels. These three levels corresponded to the three protein concentration which are defined in the full factorial design. NIR responses to these protein samples have shown that transmission decreased as protein concentration increased. These have shown the same NIR energy responses found from the protein calibration in the previous chapter (see figure 4.21). In the middle and lower window, the separations in spectra for the three levels of protein are not so obvious. This is because of the increasing addition of cell debris to these calibration samples. Also in the lower window where cell debris concentration in the calibration samples are at their highest, the transmission corresponding to different protein levels
are not necessarily in order. This may be due to variability in the cell debris of the samples masking the effect of protein concentration. Such complications need a sophisticated modelling technique, hence PLS calibration has been used.

Fig 5.8 The transmission spectra of all 36 add-back calibration samples. The spectra of samples with low, medium and high concentration of cell debris are grouped together and are shown in the three windows on the right hand side. In each of the three windows are calibration samples 1-12, 13-24 and 25-36 which corresponded to the 3 bands of spectra shown in the main figure. The effect of protein to NIR spectra can be easily identified for samples with low cell debris. But finding of direct relationship of protein in the higher cell debris or RNA concentration to NIR spectra are not obvious and hence PLS method is used.

5.5.2.2 Calibration of cell debris (optical density) using add-back calibration samples

The optical density (that characterises cell debris) PLS model is established using the 36 NIR spectra of the calibration samples in the add-back experiment (X-matrix), and together with their measured optical density (OD) at 650 nm used for the Y reference data. This model only required one principal component in the X spectral
data for correlation with the OD measurements (Y-matrix). The use of this model to back-predict the OD of the calibration samples is shown in figure 5.9. The figure has shown that samples 2 and 13 are outliers since the PLS model has given a predicted OD for samples 2 and 13 that is the same as the others in their groups of low and medium cell debris levels respectively. Thus the outliers are deduced to have been in the measurement errors; that is, the deviations from the calibration line are attributed to measurement errors. Overall, the correlation shows that the PLS calibration model gives satisfactory predictions of cell debris concentration as accounted by OD.

![NIR spectra of the add-back calibration samples](image)

**Fig 5.9** NIR spectra of the *add-back* calibration samples (X-matrix). Together with the use of their measured OD (Y-matrix), a PLS calibration model is achieved by one principal component. In this PLS calibration model cross validation technique is used. The model is then used to back-predict OD of all *add-back* calibration samples. The predicted OD are plotted against their measured OD, also the uncertainty limits given by PLS predictions are indicated by the error bars. The results have shown satisfactory prediction that gave a correlation coefficient of 0.982 and standard error of prediction (SEP) 0.0751 Au.

### 5.5.2.3 Calibration of protein using *add-back* calibration samples

The *add-back* protein model is established from the 36 NIR spectra of the calibration samples in the *add-back* experiment as X-matrix, and Y reference analytic data are
the expected protein concentrations that are defined by the *add-back* contaminants experiment. It is better to use expected concentrations than the protein assays because the sample preparation error is considered to be smaller than the variability in the Bradford assay. Variability of the protein assay has been discussed by Dehghani (Dehghani *et al.*, 1995) in detail.

Four principal components are used in the *add-back* calibration model. The first principal component accounts for the cell debris level which has a significant effect across the NIR spectrum. The remaining components account for the variation in protein while compensating for varying amounts of RNA. The protein *add-back* calibrated model is used to back-predict the calibrated samples (figure 5.10).

![Statistics](image-url)

Fig 5.10 The protein concentration have been calculated to give the *expected* values. These are based on the knowledge of protein present in the clarified yeast homogenate and the amount of protein added to these calibration samples. Using the *expected* protein concentrations (Y-matrix) and the NIR spectra (X-matrix), a cross validated PLS protein model with 4 principal components is created. The model is then used to back-predict protein of all *add-back* calibration samples. The predicted protein are plotted against their *expected* values, also the uncertainty limits given by PLS predictions are indicated by the error bars. The results have shown satisfactory prediction that gave a correlation coefficient of 0.891 and standard error of prediction (SEP) 4.35 g L\(^{-1}\).
5.5.2.4 Calibration of RNA using add-back calibration samples

The RNA model is established using $X$-matrix formed by the NIR spectra of the add-back calibration samples and $Y$ reference analytic data which are the expected RNA concentrations. There are five principal components used in this add-back calibrated model. Using the PLS add-back calibrated model, the back-predicted RNA contaminant in the calibration samples and their uncertainty limits are compared with the expected concentrations (figure 5.11).

There are fewer points shown on figure 5.11 than the number of samples because during the calibration, obvious outliers that prevented PLS model building were identified and removed. It is presumed that the outliers have been caused by sample preparation error because the standard RNA material used is very difficult to dilute during the preparation of RNA stock solution (section 5.4.3). The outliers represented more than quarter of the calibration samples. It is necessary to removed these from calibration modelling in order to avoid over influence by these outliers.
Fig 5.11 The expected RNA concentrations on all calibration samples have been calculated based on the knowledge of RNA presented in the clarified yeast homogenate and the amount of RNA added to each calibration samples. Using these expected RNA concentrations (Y-matrix) and NIR spectra of calibration samples (X-matrix), a cross validated PLS RNA add-back calibration model using 5 principal components is established. The back-predicted RNA concentration of the calibration samples together with their uncertainty limits are plotted against their expected values. In this calibration, 10 outliers were identified in the PLS calibration model. Thus for statistical reasons, it was necessary to remove these outliers in order to prevent over influencing the calibration model. The correlation coefficient on this prediction is 0.937 and the standard error of prediction (SEP) is 0.877 g L$^{-1}$. 

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>RNA</td>
</tr>
<tr>
<td>Sample Set</td>
<td>add-back</td>
</tr>
<tr>
<td>No. Of Samples</td>
<td>26</td>
</tr>
<tr>
<td>No. Of P.C.</td>
<td>5</td>
</tr>
<tr>
<td>Offset</td>
<td>1.554</td>
</tr>
<tr>
<td>Slope</td>
<td>0.901</td>
</tr>
<tr>
<td>S.E.P.</td>
<td>0.877</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.937</td>
</tr>
</tbody>
</table>
5.5.3 PLS calibration models using process stream samples

Process stream calibration models used samples that are similar to those obtained in a real flocculation process. This is ideal for general calibration since all calibration samples are representative to those for future predictions. Provided enough representative samples are available for calibration, then the calibration experiment is rather straightforward. This is because in a real process the samples would be easy to access by sampling off a process stream.

In the work reported here, the two sets of process stream 'A' and 'B' calibration samples and further two more sets of validation samples 'V1' and 'V2' used were prepared in a laboratory. This allowed good reproducibility of the process and would make comparison of result easier in the later sections. Once the PLS calibration models using process stream samples are established, these models are restricted to quantitative analyses of a similar process. By contrast, the add-back model was calibrated using the full factorial experiment and is capable of yielding better prediction of sample with any composition of the three contaminants.

Nevertheless, the use of process stream samples for the PLS calibrations have achieved satisfactory calibration for the application. The following text reports the spectra of the two sets of process stream calibration samples and the three PLS calibrations on cell debris (by optical density), protein and RNA contaminants.

5.5.3.1 NIR spectra of two sets of process stream calibration samples

The NIR transmission spectra of the two sets of process stream calibration samples are shown in figure 5.12. There are total of twenty calibration samples and all have shown large spectral response in the peak transmission region. The close up studies of all calibration samples in the peak transmission region are shown in the right window of figure 5.12. The two sets of process stream samples have identification marked by 'A', 'B' and the percentage of PEI feed used in the samples are marked
together beside each spectrum. It can be seen that using the sample identification and referring back to figure 5.5, the higher the PEI concentration used in the process stream sample the lower the cell debris (optical density) is found in that sample. Therefore in figure 5.12, as expected and previously seen in figure 5.8, the lower cell debris samples have greater NIR energy transmission. However, these spectra suggested that there are cell debris variations within calibration samples which are difficult to identify the presence of protein and RNA (as seen in the inset figure 5.8). The visual inspection here cannot identify any significant differences between add-back and process stream calibration samples other than the transmission peak levels that could have been altered by cell debris concentrations.

These spectra are used for PLS calibrations of protein, RNA and cell debris (optical density) and results are shown in the followings sections.

Fig 5.12 The NIR transmission spectra of two sets of process stream calibration samples. The right window shows the spectra variation according to process stream samples in the peak transmission region 2195 to 2225 nm. These spectra suggest that there are cell debris variation within calibration samples which are difficult to identify the presence of protein and RNA. Note that the 'A' and 'B' letters indicated which set of process stream samples and the numbering represents the PEI concentration used in flocculation.
5.5.3.2 Calibration of cell debris (optical density) by process stream samples

The two sets of *process stream* calibration samples are scanned using the NIR instrument and OD are measured (see figure 5.5), these formed the X and Y matrices respectively. A cross validated PLS calibration model is established using one principal component. Using this *process stream* calibration, the optical density of the two sets of calibration samples are back-predicted and compared with the measured OD (figure 5.13).

![Statistics](image)

Fig 5.13NIR spectra of the *process stream* calibration samples (X-matrix) and together with the use of their measured OD (Y-matrix), a PLS calibration model is achieved by one principal component. In this PLS calibration model cross validation technique is used. The model is then used to back-predict OD of all *process stream* calibration samples. The predicted OD are plotted against their measured OD, also the uncertainty limits given by PLS predictions are indicated by the error bars. The results have shown satisfactory prediction that gave a correlation coefficient of 0.981 and standard error of prediction (SEP) 0.057 Au. The significant outlier in the *process stream* B sample is suspected to have been caused by pipette error during sample presentation to the NIR spectrophotometer. The horizontal error bars indicate the standard deviation of the optical density measurements.
The results of OD predictions have shown satisfactory correlation coefficient of 0.981 between the predicted and measured ODs. There is one outlier appearing in the process stream 'B' samples. If referring back to figure 5.12, the spectrum of this outlier sample marked B15 (process stream 'B' sample that has 15% of PEI) has lower NIR transmission when compared to a similar samples spectrum A15. It is expected that both A15 and B15 should have a similar level of cell debris, but their NIR spectrum have shown otherwise. Therefore it has been suspected that the outlier was due to pipette error during sample presentation to the NIR spectrophotometer.

Nevertheless, the OD prediction for both set of process stream calibration samples have shown satisfactory results and the standard error of prediction is 0.057 Au.

5.5.3.3 Calibration of protein using process stream calibration samples

The NIR spectra of two sets of process stream samples and their assayed protein concentrations formed the X and Y-data matrices to the PLS calibration, respectively. This cross validated PLS protein calibration model required 3 principal components and the model is used to back-predict the protein concentration of these calibration samples (figure 5.14). A satisfactory correlation coefficient of 0.915 has been achieved with just one large outlier which has previously verified as being an assay error. The standard error of prediction found in these predictions is 2.34 g L\textsuperscript{-1}. 
Fig 5.14 Protein assays (Y-matrix) and NIR spectra (X-matrix) of two sets of flocculation process stream samples are collected for PLS calibration. With cross validation technique, a calibration model with 3 principal components is established. Using this model to back-predict the calibration samples, the predicted protein and uncertainty limits are plotted against the measured concentration. This has an overall correlation coefficient of 0.915 and a standard error of prediction (SEP) 2.34 g L\(^{-1}\). The horizontal error bars indicate the standard deviation of protein assays. The one significant outlier with process stream B samples has previously verified as being due to assay error.

5.5.3.4 Calibration of RNA using process stream calibration samples

The NIR spectra of two sets of process stream samples and their assayed RNA concentration formed the X and Y-data matrices to the PLS calibration, respectively. The cross validated PLS calibration model required 3 principal components to correlate the spectra with the RNA concentrations.

The process stream calibrated model is used to back-predict the RNA concentration in these calibration samples, and satisfactory predictions to the two sets of process stream calibration samples are achieved (figure 5.15). Notice that the uncertainty
limits estimated by the PLS for each calibration samples are relatively large. This is because the NIR response to this low concentration of RNA contaminants is small and the responses are influenced by the spectrophotometer instrument noise. Therefore, given under the circumstances, the RNA prediction are expected to be less accurate. Nonetheless, the conclusion is that the calibrated model has given sufficient correlation between predicted and measured result. The correlation coefficient here is 0.919 and the standard error is prediction of 1.36 g L\(^{-1}\).

![Statistics](image)

Fig 5.15 Two sets of flocculation process stream samples are used for RNA calibration. These samples are RNA assayed (Y-matrix) and together with their NIR spectra (X-matrix), a cross validated PLS calibration model using 3 principal components is created. Using this process stream model to back-predict these calibration samples the predicted RNA and the uncertainty limits are plotted in the above figure. Horizontal errors are used to indicated the standard deviation in RNA assays and for clarity only a few of the horizontal error bars are shown. Despite the large assay errors, a satisfactory correlation coefficient of 0.934 and standard error of prediction (SEP) of 1.164 g L\(^{-1}\) are achieved.
5.5.4 Process monitoring using both *add-back* and *process stream* calibrated PLS models

Previously, PLS models for quantitative analyses of optical density, protein and RNA contaminants have been presented. These models are created either by *add-back* contaminants samples or using *process stream* samples. The former is initially used to test the ability of the spectrophotometer and PLS calibration models for multiple contaminants and success of this has been shown. The latter calibration models are designed for quantitative analyses of contaminants in a flocculation process. In the following, comparisons are made on the two types of calibration models to identify the level of cells density (by optical density), protein and RNA contaminants in two flocculation processes. These are validated by traditional OD measurements and assays.

The overall results have found that the calibration model created from *process stream* samples are better in all three quantitative contaminant analyses. Thus it is concluded that the *process stream* PLS model should be used for further analysis of contaminants in the process stream.

### 5.5.4.1 Prediction of cell debris (optical density) contaminants in flocculation processes

The removal of cell debris in two flocculation processes is shown in figure 5.16. The extent of cell debris removal is given by optical density measurements, together with the predictions given by both *add-back* and *process stream* calibrated PLS models.

In the first process (figure 5.16a), the predicted optical density by the *process stream* calibrated PLS model shows small uncertainty limits and good agreement to the actual optical density measured. The predictions on the process samples by the *add-back* calibrated PLS models have also shown small uncertainty limits and good agreement to measurements in the first few process samples, but subsequent
predictions from the *add-back* model have a large offset from the actual measurements and large uncertainty limits. These findings are repeated in the second process (figure 5.16b).

The lack of prediction performance in both cases by the *add-back* calibrated model arises because the lower optical density prediction required back extrapolation from *add-back* calibrated models. There are also other differences between the *add-back* and *process stream* spectra that contribute to the errors in back-extrapolation (see section 5.5.5). The conclusion is that for process monitoring, the PLS model established from *process stream* samples should be used for accurate quantitative analyses.

![Graph showing optical density vs PEI concentration](image)

Fig 5.16 a, b) show the reduction in cell debris characterised by optical density (OD). These profiles are typically found in the flocculation process. As PEI feed concentration increases more cell debris is removed from the yeast homogenate. The error bars on the measured results represented the standard deviation typically found in OD measurements, and the error bars on the prediction results indicated the uncertainty limits defined by the PLS predictions. The results show that the prediction from the *process stream* model match the measurements well, but that predictions from the *add-back* calibration do not match well.


5.5.4.2 Prediction of protein contaminants in flocculation processes

The protein contaminants in the two sets of flocculation processes stream are shown in figure 5.17. In both processes, protein contaminants are assayed and compared to the predictions by add-back and process stream calibrated PLS models. According to the assays result, the removal of protein contaminants in the process stream is similar to that expected (see also figure 5.6). The initial protein contaminants are at their highest without the use of PEI flocculating reagent. As the PEI concentration increased the protein contaminants reduced. When PEI is approximately at 20 percent (v/v), the optimum removal of protein is reached and a further increase in PEI is no longer effective.

In the figure 5.17a, all protein predictions by the process stream calibrated model have small offsets, but the prediction and assays removal profile have shown good and satisfactory agreement. Similar results have been achieved by the process stream calibrated model in the second flocculation process (figure 5.17b).

The protein predictions given by the add-back calibrated model in the second flocculation have only shown good agreement with the assay in the initial process samples, i.e. PEI concentration at 0 percent (v/v). Then the prediction accuracy started to fall and the uncertainty error increase. This loss of performance in prediction by the add-back model is very much similar to that accounted in the optical density prediction by the add-back model. The conclusion are that the add-back calibrated model is not suitable for process monitoring, while process stream models predictions have shown satisfactory result, therefore it is appropriate for protein quantiative analyses in process operation.
Fig 5.17 a, b) shown the removal of protein contaminants during two flocculation processes. The error bar from the measured results represents the standard deviation typically found in the Bradford assays. The error bar on the prediction results indicated the uncertainty limits defined by the PLS predictions. The results show that the prediction from the process stream model match the measurements well, but that predictions from the add-back calibration do not match well and also have large error bars.

5.5.4.3 Prediction of RNA contaminants in flocculation processes

The monitoring of RNA contaminants in two flocculation processes are shown in figure 5.18. In both processes, RNA contaminants are assayed and compared to the predictions by add-back and process stream calibrated PLS models. The lines in the figure are used to highlight the removal of RNA contaminants during the flocculation processes. These are only shown for the assays and predictions by the process stream calibrated model. The illustration line for the add-back predictions are not shown (figure 5.18) because of the inconsistent prediction results. The uncertainty limits predicted in the two sets of flocculation samples from the add-back calibrated model are large and are not included in order to enhance the clarity in this figure.
In the monitoring of both flocculation processes, only the *process stream* calibrated model has predicted RNA contaminants similar to that determined by assays. In fact, the predicted RNA are closer to that of the expected RNA concentrations in the process stream than the assays result. For example, the expected RNA from the beginning of the process is 11.5 g L\(^{-1}\) while the assays have indicated 17.5 g L\(^{-1}\) in both processes and the predictions are 12.5 g L\(^{-1}\) which is closer to the expected RNA concentration. Therefore, the predictions have given more realistic RNA contaminants levels than the assays.

Overall, these have demonstrated that the *add-back* model is not calibrated to predict *process stream* samples. However, the similar profiles between the measured and *process stream* model prediction have demonstrated that this model is reliable for process monitoring of RNA contaminants.

![Graph showing variation of RNA contaminants](image)

Fig 5.18 a, b) show the variation of RNA contaminants in two flocculation processes. The error bars in these figures indicated the mean standard deviation in RNA assays and the uncertainty limits defined by the PLS predictions. These figures have demonstrated similar RNA removal characteristic between assayed and predicted RNA from the *process stream* model. They differ from each other only by an offset. On the other hand predictions from the *add-back* PLS model have not shown any of the expected result. The only indication is that there is high RNA contaminants with low PEI concentration and vice versa.
5.5.5 Reasons for poor prediction by the add-back calibrated models in process monitoring

In the previous section, it was shown that the prediction of contaminants using the add-back calibration was not satisfactory. The following describes the reasons for the weakness in prediction of process stream samples by add-back models. This can be understood by a mathematical interpretation of both add-back and process stream PLS models. The discussions suggests that particle size distribution in the two types of samples is the main cause of differences in these models.

5.5.5.1 Interpretation of PLS models

The calibration models for optical density provides insight into the means by which PLS enables predictions of OD, protein and RNA from the NIR spectra. For OD, only one principal component is needed for the reconstruction. The optical density model is of the form:

\[ X_i(\lambda) - \bar{X}(\lambda) = t_i p(\lambda) + e_i(\lambda) \]  \hspace{1cm} [5.1]

For the add-back model \( X(\lambda) \) is the \( i \)th spectrum in the add-back experiments, \( \bar{X}(\lambda) \) the mean of all the add-back spectra, \( p(\lambda) \) is the loading for the principal component (shown in figure 5.19a, b) and \( t_i \) is the score for the \( i \)th spectrum. Hence, the deviation of each NIR spectrum from the mean is proportional to \( p(\lambda) \), apart from a small error term \( e_i(\lambda) \). The analysis shows that this form of model fits the add-back data with a precision of 0.075 Au for OD, 4.35 g L\(^{-1}\) for protein and 0.88 g L\(^{-1}\) for RNA.
In order to use PLS to predict a value of OD from the NIR signature, the following model is used. Here $\text{OD}_i$ is the measured optical density of the $i$th sample, $t_i$ is the score, $q$ is the loading for the OD data, and $g_i$ is the residual error.

$$
\text{OD}_i - \overline{\text{OD}} = t_i q + g_i
$$

[5.2]

During calibration of the PLS model the aim is to use the spectra $X_i(\lambda)$ and measurements $\text{OD}_i$ to find $\overline{\text{OD}}$, $\overline{X}(\lambda)$, $p(\lambda)$ and $q$. These are then the parameters of the calibration model.

The calibration parameters may then be used to predict the optical density $\text{OD}_j$ for a new sample by evaluating a new score value, $t_j$, from the new spectrum $X_j(\lambda)$. Overall, the effect is to find the unknown $\text{OD}_j$ to give the best fit of the spectrum to the expression:

$$
X_j(\lambda) - \overline{X}(\lambda) = \frac{\text{OD}_j - \overline{\text{OD}}}{q} p(\lambda)
$$

[5.3]

Plots of $p(\lambda)$ for the add-back and process stream models are shown in figure 5.19a. As expected they are similar to one another showing that the underlying shapes of the NIR spectra in the two experiments are similar. However they are not exactly the same. Plots of $\overline{X}(\lambda)$ are shown in figure 5.19b. As expected, these are different in amplitude because the samples were prepared in different OD ranges (see figure 5.1, showing add-back and process steam regimes). Likewise the parameters $\overline{\text{OD}}$ of the add-back and process stream models are different from one another, at 1.03 and 0.27 respectively.

The parameters $q$ for the two cases are 0.127 and 0.257; this difference is significant as it indicates a difference in relationship between the intensity of the NIR spectral response and the OD values. Both the NIR transmission and the OD have been observed to change when trace amounts of PEI were added to a sample (PEI by itself
does not present a significant NIR signature, see Appendix A9). It is also significant that the cell debris contaminant used in the add-back calibration samples has a larger particle size distribution than those in the process stream samples. These observation suggest that the flocculating effect of residual PEI in the process stream causes a difference in particle size that is responsible for the difference in OD and NIR transmission. Such effect also accounts for the differences in the protein and RNA calibration models.

5.5.5.2 Effect of particle size on the NIR spectra

The explanation for the difference in the PLS models is thought to be the particle size of the material in the samples. Transmission is a function of particle size as well as of the concentration of material in the sample. It depends on a quantity known as the extinction coefficient which represents the ratio between the scattering cross section of the particles and the geometric cross section. The extinction coefficient is influenced by diffraction effects when the particles in suspension are of the same
order as the wavelength of the light, and is a function of particle size (Kerker, 1969). Therefore the relationship between the optical density of a sample and the concentration of particles is constant only as long as the particle size does not change. However, it has been suggested that the particle size distribution in the process stream samples has been altered by PEI flocculation and further complicated by the larger particle size used in cell debris contaminants for the add-back calibration samples. Appendix A9 gives an unambiguous demonstration of the particle size effect to the NIR spectra that is caused by residuals amount of PEI added to the yeast homogenate.

The extinction coefficient is also relevant to the transmission of the NIR radiation through the samples. As figure 5.19a shows, the peak in the PLS loading spectrum for the process-stream samples is shifted to a slightly longer wavelength compared with that for the add-back samples. The NIR instrument scans across a range of wavelengths (1900-2500nm) and the extinction coefficient therefore varies throughout the scan. An explanation for the shifted peak in the process stream samples is that the change of particle size has moved the maxima and minima of the extinction coefficient to different NIR wavelengths. NIR transmission through the process stream samples would be attenuated at some wavelengths and increased at others compared with the add-back samples, thus altering the underlying shape of the NIR spectra.

The conclusion from the above discussions is that the particle size has an influence on NIR transmission measurements with turbid samples. An accurate PLS model can be derived provided the particle size is constant, for instance within the add-back sample set or the process stream set. However, the PLS model developed for one particle size cannot be extended to samples having a different particle size. To do that, one would need a measurement of the particle size to use as a variable in the PLS model, or alternatively to compute a correction factor via the extinction coefficient.
5.5.6 Establishing control set point

The flocculation process previously shown can be optimised with the calibrated NIR spectrophotometer. Using the PLS process stream calibrated models and NIR spectra, the variations in all three prime contaminants during the whole of the flocculation process can be monitored frequently. It has been known that over feeding of PEI can reverse the efficiency of contaminants removal and excessive PEI residual would remain in the process stream. The use of calibrated NIR system can enable tracking on the maximum removal of individual contaminant and selection of the appropriate PEI feed control set point. Therefore the calibrated NIR system can enhance understanding of the activities of the contaminants during the process and also allow process optimisation. Figure 5.20 shows a suitable control set point.

![Graph showing control set point](image)

Fig 5.20 PEI feed control set point justified by prediction of contaminants using the calibrated NIR system. Predicted OD, protein and RNA on one set of process stream samples are showed with symbols ■, • and ♦ respectively. The control set point can be selected to the region at which the three contaminants have approached their first minimum concentration together.
5.6 Conclusion

This chapter has reported two different calibration experiments. These are the *add-back* contaminants and *process stream* samples PLS calibrations with spectra recorded by the in-house spectrophotometer. Both of these have been used in flocculation processes monitoring. The three important contaminants of cell debris (characterised by optical density), protein and RNA concentration are sufficiently quantified by predictions from the *process stream* calibrated models. It has been suggested that information about these contaminants would be an advantage for process control and would aid process optimisation. The following lists the conclusions for work presented in this chapter.

### 5.6.1 Application and calibration methods

- Two types of calibration samples are used for calibrating the in-house low cost NIR spectrophotometer. These are known as the *add-back* and *process stream* calibration samples.

- Once the NIR spectrophotometer is calibrated, its use will be for quantitative analyses of three different contaminants in a flocculation process. These are cell debris, protein and RNA contaminants.
An ideal and comprehensive set of calibration samples should follow a full factorial design which will give an robust calibration. This requires the variation of concentration in the calibration samples similar to those obtained in the future analysis samples and that all combination of mixtures (contaminants) are included in the calibration samples (hence a full factorial design). The latter case is important because the NIR signatures are weak and highly overlapping.

The ideal set of calibration samples cannot be achieved because selective removal of contaminants from yeast homogenate is very difficult. Therefore, the contaminants are added back into the clarified yeast homogenate. This simplified all preparation and implementation of full factorial design calibration samples. Calibration samples prepared in this way are called add-back calibration samples.

The drawback of add-back calibration samples is that there are background levels of contaminants in the samples. Thus calibration samples are outside the range of interest. Also the contaminants added back are not identical in chemical or physical conformation to the contaminants naturally in the homogenate. Therefore, the add-back calibration samples may not be representative of real process samples.

The add-back calibration is suitable to test the NIR spectrophotometer and PLS calibration for multiple contaminants analysis. As this is achieved, the calibrated models are used for quantitative analyses of contaminants in process stream.

Alternatively, calibration samples that are similar to those obtained in process stream are used. These have the advantage that the calibration samples are representative of those analysed in the future. But due to difficulties in selective removal of contaminants, the full factorial design cannot be followed. These calibration samples are known as the process stream calibration samples.
5.6.2 Measurements and measurement variability

- Optical density measured at 650 nm is used to characterise cell debris concentration in all samples. The standard deviation observed is 0.05 Au for measurements between 0 to 1.8 Au units.

- Bradford assays are used to determine protein concentrations in all samples. The standard deviation found in protein assays is 2 g L\(^{-1}\) for measurements between 0 to 75 g L\(^{-1}\) of protein in yeast homogenate. Bovine Serum Album (BSA, fraction V), supplied by Sigma Chemical Ltd has been used for Bradford assays standard and all add-back calibration samples.

- Orcinol assays are used to determine RNA concentrations in all samples. The standard deviation found in RNA assays is 3.5 g L\(^{-1}\) for measurements between 0 to 33 gL\(^{-1}\) of RNA in yeast homogenate. Yeast ribonucleic acid (RNA, highly polymerised), supplied by Sigma Chemical Ltd has been used for Orcinol assays standard and also all add-back calibration samples.

- The optical density and two assays results found on all process stream calibration samples are used for the PLS calibration models.

- In add-back contaminants calibration samples, only optical density measurements were used in PLS calibration. The expected protein and RNA concentrations of all calibration samples were used for the rest of PLS calibration experiments. The protein and RNA assays are not used because the error associated in the add-back calibration samples are large while the expected concentrations data are free from assays error.
• It has been found that the Orcinol assay technique is not suitable for quantifying RNA added to the yeast samples. This was because in the add-back experiments, the technique has failed to produce consistent assays on the calibration samples with external RNA added.

5.6.3 PLS calibrations using add-back contaminant samples

• Using all 36 add-back samples for PLS calibrations, the table below summarised the statistic of all three contaminants PLS calibrations.

<table>
<thead>
<tr>
<th></th>
<th>No. of principal components</th>
<th>correlation coefficient (r)</th>
<th>standard error of prediction (SEP)</th>
<th>No. of outliers</th>
<th>Plot Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD</td>
<td>1</td>
<td>0.982</td>
<td>0.0751 Au</td>
<td>3</td>
<td>5.9</td>
</tr>
<tr>
<td>Protein</td>
<td>4</td>
<td>0.891</td>
<td>4.351 g L(^{-1})</td>
<td>2</td>
<td>5.10</td>
</tr>
<tr>
<td>RNA</td>
<td>5</td>
<td>0.937</td>
<td>0.877 g L(^{-1})</td>
<td>10*</td>
<td>5.11</td>
</tr>
</tbody>
</table>

Table 5.1 Statistics of the three PLS calibrations in add-back calibration experiments. Note that *indicated outliers were removed from PLS calibration.

• Cell debris (characterised by OD) has the greatest influence on the recorded NIR spectrum. Only a single principal component is required for a successful PLS calibration.

• The standard error of prediction in OD is similar to that the standard deviation observed in actual OD measurements.

• Both the protein and RNA have small complicated NIR signatures in the spectrum, and calibrations are more complex. Thus the PLS calibrations required 4 and 5 principal components for protein and RNA calibrations respectively.
• The standard error of prediction (SEP) in protein is larger than the standard deviation in protein assays, while the SEP in RNA is less than the standard deviation in RNA assays. Some outliers had to be removed from the RNA prediction.

• The population of outliers in the calibration samples can influence the establishment of PLS calibration model.

• In both OD and protein calibrations, only a small number of outliers were observed after the calibrations. Because the outliers were a small population compared to the calibration samples, therefore the PLS models were not influenced.

• In RNA calibration, 10 outliers are suspected to have been caused by sample preparation error because the experimental technique of RNA addition is difficult. Since these represented a large population relative to the calibration samples, calibration was not achieved with these outliers present. As a result, these outliers are removed in order to achieved the RNA PLS calibration.

• Overall, the PLS add-back contaminants calibration has demonstrated the success of PLS calibration of multiple contaminants using the NIR spectra captured from the in-house spectrophotometer.
5.6.4 PLS calibrations using process stream samples

- Using two sets of process stream samples for PLS contaminants calibrations, the table below summarised the statistics of all three contaminants PLS calibrations.

<table>
<thead>
<tr>
<th></th>
<th>No. of principal components</th>
<th>correlation coefficient (r)</th>
<th>standard error of prediction (SEP)</th>
<th>No. of outliers</th>
<th>Plot Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD</td>
<td>1</td>
<td>0.981</td>
<td>0.057 Au</td>
<td>2</td>
<td>5.13</td>
</tr>
<tr>
<td>Protein</td>
<td>3</td>
<td>0.915</td>
<td>2.343 g L^{-1}</td>
<td>1</td>
<td>5.14</td>
</tr>
<tr>
<td>RNA</td>
<td>3</td>
<td>0.919</td>
<td>1.358 g L^{-1}</td>
<td>1</td>
<td>5.15</td>
</tr>
</tbody>
</table>

Table 5.2 Statistics of the three PLS calibrations in process stream calibration experiments.

- All three contaminants calibration by PLS modelling are successful. These used all OD measurements and assays result as the referencing data (Y-matrix).

- The effect of cell debris on the NIR spectrum of the calibration samples is large, and calibration of OD has only required a single principal component. Both protein and RNA have smaller NIR signal responses and are more complicated for calibration. These required three principal components for each successful calibration.

- The standard error of prediction for both OD and protein are similar to the standard deviation by measurement and assay, while the standard error of prediction for RNA contaminant is less than the standard deviation of the RNA assay. This suggested that predictions of contaminants using PLS calibrated models are better than traditional assays.
• In all process stream calibration, only a few outliers were identified and PLS calibration models were not influenced by these small populations of outliers.

• Overall, the process stream samples calibration have demonstrated successful PLS calibration of multiple contaminants using the NIR spectra captured from the in-house spectrophotometer.

5.6.5 Process monitoring and control set point

• Two sets of process stream flocculation samples are used for process monitoring. Both add-back and process stream calibrated PLS models predicted the three important contaminants concentrations. The statistics of predictions by these two models to these two sets of flocculation process samples are given below.

<table>
<thead>
<tr>
<th></th>
<th>Correlation coefficient (r)</th>
<th>Standard error of prediction (S.E.P.)</th>
<th>Offsets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Add-back</td>
<td>Process</td>
<td>Add-back</td>
</tr>
<tr>
<td>OD</td>
<td>0.985</td>
<td>0.984</td>
<td>0.03 Au</td>
</tr>
<tr>
<td>Protein</td>
<td>0.774</td>
<td>0.938</td>
<td>8.18 g L(^{-1})</td>
</tr>
<tr>
<td>RNA</td>
<td>0.218</td>
<td>0.899</td>
<td>8.03 g L(^{-1})</td>
</tr>
</tbody>
</table>

Table 5.3 Statistics on prediction of two flocculation process contaminants given by add-back and process stream calibrated PLS models.

• In process monitoring, (also see table above on correlation coefficient and offset figures) PLS calibration models by process stream samples have given much better predictions than add-back calibrated models.
• The PLS predicted uncertainty limits for all three contaminants predictions by add-back calibrated models are large. This is because the level of contaminants observed in the flocculation process are outside the range covered by the add-back calibration experiments. Therefore the add-back models have foreseen problems in prediction of process samples by giving larger uncertainty limits.

• The quality of contaminants prediction given by the process stream calibrated PLS models are sufficient to provide information for process control (see figure 5.19) to a control set point. Thus the removal of contaminants and the flocculation process can be optimised.

5.6.6 Differences between add-back and process stream samples

• Both add-back and process stream PLS calibration models are studied. The calibration models use the loading spectrum, $p$, and the parameter, $q$, and the measuring samples NIR spectrum to predict the contaminants concentration ($Y$-matrix). It has found that the loading spectra, $p$, and the parameter $q$ are different in the add-back and process stream calibration of optical density (OD).

• As the loading spectra, $p$, and the parameter $q$ are different in the two OD calibration models, the same performance in prediction of process stream samples by the add-back calibration models should not be expected.

• Also the add-back calibration model for protein and RNA required 4 and 5 principal components in their models respectively, and since their first principal components are not the same as those in process stream samples. Therefore these add-back protein and RNA models are not expected to give accurate prediction of the process stream samples.
The reasons for the differences in calibration parameters of the add-back and process stream calibration models are thought to have been caused by particle size distribution differences and also the effect of PEI residuals in the process stream samples.

Particle size distribution in the two types of calibration samples are different. This is because the cell debris contaminant used in the add-back calibration is from the first separation of cell debris in the yeast homogenate (see section 5.4.2) and this generally has the largest particle size in the yeast homogenate. Therefore, the add-back calibration samples have larger particle size distribution than those in process stream samples.

Particle size distribution in the process stream depends on PEI concentration used in the flocculation. Since PEI creates flocs and enhances the removal of cell debris from the centrifugation, thus this also has an impact on the particle size distribution.

Both the optical density and NIR spectral measurements are affected by the particle size distribution in the measuring samples. The extent of this effect is governed by the extinction coefficient which is also wavelength dependent.

The extinction coefficient also explained the reason why the loading spectrum, \( p \), of the add-back calibration samples is shifted when compared to the loading spectrum of process stream samples. This is because NIR response depends upon the wavelength of measurements.
6. Conclusion and Recommendation

6.1 Conclusion

The advantages of NIR are that absorption is linear over a wide range and samples can be measured without any preparation or dilution. The NIR spectroscopy technique for quantitative analysis has been implemented in many food and agricultural industries and is currently moving toward applications in pharmaceutical and bioprocess areas. The thesis has demonstrated the novel use of NIR to aid a downstream bioprocess. This is an alcohol dehydrogenase (ADH) recovery process where process monitoring by NIR is eventually used to aid process control.

As part of the research, an NIR spectrophotometer has been developed using a linear dispersion grating and a single element lead sulfide detector. A scanning process is required to record the desired spectrum (1900 to 2500 nm). The detector has temperature control but it has low signal to noise ratio compared to one found in an advanced commercial spectrophotometer. This problem has been overcome by maximising the optical throughput at the expense of saturating the detector during standard open-path light measurement. This has therefore prevented direct light reference reading that is required in order to obtain absorbance (Au) unit measurement of the subsequent sample measurements. Instead direct output voltage of the detector is measured and the transmission spectrum is baseline corrected by a pre-scanned reference. It has been shown that the method provided sufficient signal in the NIR spectrum for quantitative analysis of biological contaminants.

A multivariate calibration technique based on principal components is used. This is partial least squares (PLS) technique that has been described in the thesis. A PLS calibration model decomposes the most relevant spectral information into a set of
underlying profiles called principal components and correlates these with the contaminant concentration. The number of principal components incorporated into the PLS model are chosen such that noise in the spectral data is rejected but so that enough principal components are present to explain all the influences in the calibration model. Cross validation technique has been used throughout all PLS calibration modelling. This method self-validates the PLS model and uses calibration samples efficiently.

In a comparison of add-back and process stream methods of multiple contaminants calibrations, PLS calibration models derived from process stream have given satisfactory prediction of contaminants level in process monitoring. It is concluded that the use of the NIR spectrophotometer with process stream calibration models is suitable for monitoring of the flocculation process. On the other hand, the prediction of process stream contaminants by the add-back calibration models has given unsatisfactory results. These problems have been explained by the differences of particle size distribution between the process stream and add-back calibration samples and an un-modelled effect due to residual quantities of a flocculating agent.

However, the full factorial design in the add-back calibration experiment has shown that robust multiple contaminants calibration can be achieved. If the problems of particle size distribution and un-modelled contaminants can be solved then the add-back calibration experiment would be extremely useful in practice. This is because the NIR instrument can be pre-calibrated in the laboratory by this add-back approach before it is applied to at-line or on-line process monitoring in the process area.

In summary, the thesis has demonstrated that a low budget spectrophotometer built to a specific NIR spectrum (1900 to 2500 nm) can be used for process monitoring of a flocculation process. Although the total sampling time of a process sample is 9 minutes (including a reference measurement and followed by several seconds of PLS predictions) before producing contaminants information, this is a more complete and faster method than the traditional assays. Hence with NIR spectroscopic technique of process monitoring is possible, rapid monitoring leading to a recommendation for its use in process control.
6.2 Recommendation

Recommendation to any future can be separated into three parts of research. These are; further use of NIR process monitoring and control within the ADH recovery process, improvements in NIR instrumentation, and improvements in calibration and modelling. The following sections address each one of these topics individually.

6.2.1 NIR process monitoring within the ADH recovery process

The thesis has demonstrated that NIR process monitoring can be use to aid flocculation process control. Also process information generated from the NIR can benefit the subsequent 'two-cut' precipitation or even the final stage of purification in the hydrophobic interaction chromatography (HIC, see figure 2.3). For example, if excessive residual RNA remained in the process then this may be cut off before reaching the HIC to prevent fouling of the column, or the protein level predicted by NIR may affect the control decision within the 'two-cut' precipitation process. These examples suggest uses for the current NIR monitoring in feed forward control.

NIR monitoring may also be applied after the 'two-cut' precipitation or HIC process to provide more accurate process information for the precipitation control and to determine the purity of ADH product from the output of HIC. While the former allows further optimisation of recovery process, the latter defines the quality of ADH product at the end of the recovery process. Hence a complete recovery process can be envisaged which is under constant monitoring at all times.

For such applications it is expected that further calibration of the NIR spectrophotometer will be required. The calibration samples will be representative of those found in 'two-cut' precipitation and HIC outputs, and the contaminants will be
present at lower levels. Therefore it is first necessary to ensured the NIR spectrophotometer is sensitive enough to detect these lower level of contaminants. Also expected is that calibration for these areas of monitoring are simpler because the process stream samples are becoming more and more pure. Thus the problem of particle size distribution is eliminated and less complex spectra should be achieved for the individual contaminants which will be easier to identify by multivariate calibration. In the following, the idea of implementing multiple NIR process monitoring system by use of a single NIR spectrophotometer is developed.

6.2.2 NIR Instruments for future ADH recovery processes

The existing budget NIR spectrophotometer has been designed to give the option of reflectance measurements. Research in the reflectance measurement of the contaminants of interest and repeats of multivariate calibration should be studied. This is important for scaling up the downstream process where the concentration of the yeast homogenate and the turbidity would increase. Due to higher light scattering from such samples weaker NIR energies will be detected when the spectrophotometer is operated in transmittance measurements. In reflectance measurements, the higher light scattering returns greater energies to the detector thus this allows better signal to noise ratio in measurements. Also whenever remote measurements by optical fibre probe is to be implemented, a decision of using either transmittance or reflectance measurement can be based upon comparison of laboratory spectrophotometer tests.
In consideration of sample presentation to a spectrophotometer, the transmittance and reflectance measurements by optical fibre probe can be achieved by the following illustrated configurations (figure 6.1). Imagine that the process stream sample flows across a specially designed rectangular glass tubing, the optical light source is carried forward to the sample by the inner optical fibre. The light energy is transmitted across the cross section of the specially designed tubing and is reflected back to the measuring outer optical fibre probe by a mirror. This enables the measured transmitted light energy to return to the remote spectrophotometer for spectral measurements. (Note that the total transmission optical pathlength is twice the cross section width of the tubing and this arrangement has also been known as transflectance configuration). In reflectance measurements, the mirror used in transmittance is replaced by a non-reflecting material (i.e. ceramic), thus only reflected light is collected by the outer optical fibre (see figure 6.1). In a reflectance measurement, the specially designed tubing can be eliminated. Instead only an optical window will be need in the process pipe in order to allow infrared energy to interact with the process samples (as illustrated in figure 6.2).
In case of further use of NIR spectrophotometer in other parts of the ADH recovery process, more sensitive detection of contaminants is thought to be necessary because lower concentration of contaminants will be anticipated. This can be achieved by further increase optical light throughput to the measuring sample, thus allowing an increase in the pathlength to produce a stronger optical signal for detection. However, excessive increase in optical source energy may cause build up of heat to the biological samples and create problems in the samples. Alternatively, a more sensitive detector and detector amplifying circuits have to be investigated.

Previously in section 6.2.1, it was suggested that NIR monitoring can be used in two other locations of the ADH recovery process. It is possible to use one spectrophotometer for all process monitoring. This can be achieved by pumping samples from different part of the process to a flow cell in the spectrophotometer (see figure 6.3). Samples may be presented to the spectrophotometer either by flow through or stop flow. An investigation of which method will provide better rapid process monitoring has yet to be identified since the stop flow method will give better signal for detection but is slow, and vice versa for the flow through method. However, it is thought that the need for regular washing of the flow cell is inevitable. Alternatively, many points of measurement can be multiplexed to an NIR spectrophotometer.
spectrophotometer (see figure 6.3), whereby the measuring point is connected to the spectrophotometer by the optical fibre multiplexer. This is a quicker method because the transfer of light from the process into the spectrophotometer is at the speed of light while the transfer of samples through a pump system can be a time consuming process. Also the optic fibre approach does not require regular washing of a flow cell in the spectrophotometer.

![Diagram of NIR Spectrophotometer](image)

Fig 6.3 Illustration of two methods of NIR measurements to three process stream locations. In the first method (upper arrangement), samples and wash reagent are fed into the NIR spectrophotometer. These are controlled by four switching valves and pump unit such that only a single process stream location is monitored by the NIR spectrophotometer at a time. In the second method (lower arrangement), optical fibre probes are attached to each process stream location. With use of multiplexer unit, only one optical fibre probe is connected to the NIR spectrophotometer at a time for measurements.
The current spectrophotometer required approximately 9 minutes to record the desired spectrum. If several points of measurements are loaded into one spectrophotometer either by flow cell or optical fibre multiplexer, then the overall measurement times may be excessive for rapid control of the process. Thus, faster spectrum record time is required by the NIR spectrophotometer. This can be achieved by replacing the existing single element lead sulfide detector with diode array detectors or by use of an acoustic-optically tuneable filters (AOTF) instrument described in chapter 2. In both cases, recording of spectrum can be achieved in seconds, and because no mechanical moving parts are necessary spectral reproducibility would also improved.
6.2.3 Calibration experiments

It was found that problems with *add-back* calibration could be attributed to the particle size of the debris components. In the future preparation of the *add-back* calibration experiments, selection of debris contaminants that have similar particle size to those found in the *process stream* sample should be established. This should ideally make the spectra of the *add-back* and *process stream* similar to each other, and overcome the problem of spectral shifting. It is expected that this revised calibration model will produce better prediction of cell debris (measured by optical density) for process monitoring. The *add-back* calibration samples can further include other residual contaminants which were previously unaccounted (i.e. DNA, lipids). Although this would increase the number of calibration samples, but the *add-back* method has the advantage that the analyte information is defined and does not require assays of the calibration samples.

It is also expected that the *add-back* calibration approach can be used for calibrating the NIR instrument for monitoring of the 'two-cut' precipitation and HIC process. This is because clearer process samples are expected and calibrations are less affected by particle size distribution problem.

It is recommended that modelling of the particle size distribution may be linked to the pre-processing of the NIR spectrum. The technique would depend on the extinction coefficient of the measuring samples which in turn is function of wavelength. The idea is to produce similar spectra that have the same contaminants regardless of differences in their particle size distribution. There are spectrum pre-processing techniques such as multiplicative scatter correction (MSC) (Martens et al, 1993), standard normal variate (SNV) (Barnes et al, 1989), zeroing of the constant term in Fourier representation (Burns and Ciurczak, 1992) which all can be adapted specifically to minimise any particle size distribution problem encountered.
6.3 Final review

- NIR spectroscopy is based on the absorption of electromagnetic radiation in the near infrared region (700 to 2500 nm) by chemical bonds such as -OH, -CH, and -NH. The absorption spectrum has highly overlapping peaks due to combination and overtone bands. In additional to that, the spectrum is further complicated in biological applications because these contain many different chemical species, thus the use of NIR technique to biological processes in general is challenging.

- An NIR spectrophotometer has been developed using a linear dispersion grating and a single element lead sulfide detector. The inexpensive in-house built spectrophotometer is dedicated for monitoring of ADH downstream process. Hence, the spectrophotometer restricted to the NIR spectrum of region 1900 to 2500 nm and transmission measurements. Appendix 4 summaries the performance of the in-house spectrophotometer and gives a comparison with commercial instruments.

- Research has determined that the in-house built spectrophotometer is suitable for transmission measurements, but the detector used in the spectrophotometer suffers from lower signal to noise ratio compared to a commercial spectrophotometer. This problem has been overcome by maximising the optical throughput and thus increases sample presentation pathlength to 2 mm (usually ≤1 mm pathlength cell is used in a commercial spectrophotometer).

- The NIR spectrophotometer calibration for process monitoring has been achieved by pre-selection of the NIR spectral region for all multivariate analysis. The multivariate analysis is based on partial least squares (PLS) technique to establish calibration models and predictions.

- It was thought the idea of add-back calibration method described in section 5.3.2 would be useful since the NIR spectrophotometer can be pre-calibrated in the
laboratory before it is applied to at-line or on-line process monitoring. However, due to the particle size differences in the add-back calibration and process stream samples it is recommended to use the calibration models established from process stream samples, as described in section 5.3.3.

- The calibrated NIR spectrophotometer can be use for monitoring of the ADH recovery process whereby quantitative information on cell debris, protein and RNA can be generated seconds after the acquisition of the NIR spectrum. The spectrum acquisition time requires 9 minutes but the quantities of the three contaminants can be given at once, which still gives an advantage over the traditional assays.

- The concept of the NIR spectroscopic technique for downstream process monitoring has been proven. The quantitative information of cell debris, protein and RNA contaminants generated by the NIR technique are valuable data to aid the alcohol dehydrogenase (ADH) recovery process and control of the process.

- It is recommended for future improvement that an optical fibre probe may be used for on-line measurements. Depending upon the concentration of yeast homogenate used, the optical fibre probe can be configured for transflectance or reflectance measurement. Also, the use of NIR technique can be extended further downstream in the ADH recovery processes for close monitoring of the contaminate residuals. In either case, a higher signal to noise spectrophotometer may be required since the use of optical fibre reduces the signal quality for detection and in further downstream of the recovery process the concentrations of contaminants are small, therefore in both cases increases in signal to noise of the spectrophotometer may be required.
A 1. The Fabry-Perot interferometer

The theory of the Fabry-Perot interferometer is based on optical interference that involves multiple reflection. Two examples of this are the coloured patterns observed in light reflected from a thin film of oil on water and those seen on a soap bubble surface.

The Fabry-Perot (FP) interferometer uses fringes generated in transmitted light following multiple reflections in the gap between two parallel plates that have been coated with dielectric reflecting layers.

In figure A 1.1, the transmitted energy travelled through the first plate but a fraction of energy is reflected. The remaining energy is further transmitted through the second plate while a fraction is reflected back to the first plate, the cycle of internal reflection between the two plates and the output energy are as illustrated in figure A1.1. A phase difference corresponding to the extra path in the gap is introduced on each successive transmission.
\[ \delta = 2\pi\lambda(nd\cos\theta) \]  

By summing the output energy series to infinity, the output amplitude can be expressed as:

\[ Ae^{i\theta} = att + att^2 e^{i\delta} + att^4 e^{i2\delta} + \ldots \]  

[A1.2]

and by multiplying the amplitude function by its complex conjugate to derive the transmitted intensity, we obtain the equation

\[ I_t = \frac{I_0}{1 + \frac{4r^2}{(1-r^2)}\sin^2\left(\frac{\delta}{2}\right)} \]  

[A1.3]

This known as the Airy function and represents the instrumental profile of the Fabry-Perot. In order to scan the interferometer, the separation of the two plates is modified, hence altering the transmission phase difference and equation [A1.3]. Figure A1.2 shows the transmission of the Fabry-Perot as phase difference is increased (Osborne, 1993). The Airy function has been plotted for three different reflectivities of the coating on the plates. The higher the reflectivity, the greater the effective number of reflections and the narrower the width of the function.

![Fig. A 1.2 Transmission of Fabry-Perot for three reflectivities](image)

If the effective number of reflections between the two plates is \( k \), then \( 2\pi dk \) will be the total optical path. This is the same as the optical path difference introduced in both Michelson and grating instruments.
The order of the Fabry-Perot is defined by the value of $m$. For a very high values of $m$, extreme resolution can be obtained. For low values of $m$ the Fabry-Perot behaves much like an interference filter. As with the grating spectrophotometer, order sorting is accomplished using an optical filter. The resolution is limited by the finite size of the entrance aperture. Like the Michelson, the instrument has circular symmetry and a high resolution luminosity product.

The Fabry-Perot interferometer has similar problem to the Michelson that is an extreme sensitivity to vibration and misalignment. Previously, the instrument was dedicated to vibration-free areas in research laboratories. Scanning was even undertaken without moving the plates by varying the refractive index of the gas in the gap and surroundings (Hindle and Reay, 1967).

Today, Fabry-Perot interferometer has been brought into the industrial environment. The interferometer is based on a piezo-electrically alignment. Figure A1.3 is a schematic of a piezo-electrically aligned and scanned system. Tiny capacitors, made by depositing metallic zones on the edge of the plates, are used to measure the plate separation to very high accuracy.

![Piezo-electrically scanned Fabry-Perot](image)

Fig. A 1.3 Piezo-electrically scanned Fabry-Perot
The separation of the two plates is measured at three points around the periphery. From this, signals are generated to servo-control both the gap parallelism and separation. The ceramic stacks which expand and contract with applied voltage, are optically contacted onto the plates, forming a very rigid construction. With a tight servo loop, this design can maintain alignment through strong knocks and vibrations.

The FP can be scanned rapidly through 0.5-1 octaves of spectrum when used in low order. The spectrum is recorded directly and with high luminosity. A disadvantage of the instrument is its instrumental profile which has wide wings, that is broad at the base. A second disadvantage is when the device is used at high resolution, there is the small inter-order spacing.
A 2. Multiple linear regression in spectroscopy applications

A 2.1 Introduction

This appendix highlights one of the problems associated with the use of the MLR calibration method for NIR spectroscopic data. This is the technique that utilizes all available spectral data (X-matrix) to calibrate for the analyte concentrations (Y-matrix). However, in majority of cases, it is not all the spectral data have a direct relationship with the analyte concentration. In such circumstances, the irrelevant spectral data shows a smaller impact during calibration, but the result of this calibration model could have a significant effect on future predictions. Situation

The approach of MLR has been described briefly in Chapter 3. In the following, a simple example is adapted from the literature of K. Beebe and B.R. Kowalski (1987) to demonstrate the problem of MLR.
A 2.2 Problem of MLR with real spectroscopic data

This model chosen solely according to this criterion attempts to use all of the variance in the X matrix, including any irrelevant information, to model Y. When the resulting model is then applied to a new sample the model will assume that the correlation found between the calibration X and Y matrices also exists in that sample. Because the model was built using irrelevant information in the X matrix, this assumption will not be true. Unfortunately, even noise has a very high probability of being used to build the model. The following example illustrates this point. Consider the following X and Y matrices;

\[
X = \begin{bmatrix}
75 & 152 & 102 \\
63 & 135 & 82 \\
96 & 218 & 176 \\
69 & 157 & 124 \\
\end{bmatrix}
\quad
Y = \begin{bmatrix}
2 & 7 & 5 \\
4 & 3 & 3 \\
9 & 12 & 3 \\
6 & 8 & 2 \\
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
-0.71 & 0.55 & 0.48 \\
0.42 & -0.41 & -0.24 \\
-0.08 & 0.28 & 0.05 \\
\end{bmatrix}
\]

MLR is used to determine the B matrix, as described in Chapter 3. A standard measure of the effectiveness of the model is the value of Err as presented in section 3.4.1. For this example Err = 0.49, and it has assumed that the matrix B is an accurate estimate of the true model. The coefficients in B, therefore, will closely approximate the true relationship between the variables in X and Y.

To illustrate how a MLR calibration method can be inappropriate, a column of random numbers ranging from zero to 100 was added to the X matrix. The addition of this column is analogous to the inclusion of a wavelength in near infrared analysis that has
no useful information for describing the three analytes ($c_1$, $c_2$ and $c_3$) of the samples. 

The resulting matrix $X_2$ and MLR model are

$$
X_2 = \begin{bmatrix}
75 & 152 & 102 & 91 \\
63 & 132 & 82 & 36 \\
96 & 218 & 176 & 74 \\
69 & 157 & 124 & 51
\end{bmatrix}
$$

and $Y$ remained as;

$$
Y = \begin{bmatrix}
2 & 7 & 5 \\
4 & 3 & 3 \\
9 & 12 & 3 \\
6 & 8 & 2
\end{bmatrix}
$$

The new $B$ coefficient is

$$
B_2 = \begin{bmatrix}
0.71 & 0.18 & 0.42 \\
-0.42 & -0.19 & -0.20 \\
0.24 & 0.20 & 0.03 \\
-0.12 & 0.03 & 0.01
\end{bmatrix}
$$

The new model found from the regression of the same $Y$ as in the first example onto this new matrix $X_2$ is given in $B_2$. In this new model, the error $\text{Err} = 0.07$, where the smaller value implies that this second model is more effective at modelling $Y$. In order to evaluate these results, note that when $X$ is multiplied by $B$, the $j$th column of $X$ is always multiplied by the $j$th row of $B$. The importance of variable $j$ to the model can therefore be considered by examining the $j$th row of $B$. The nonzero column in the fourth row of $B_2$ reveal that a variable consisting of random numbers (column number four of $X_2$) is chosen as a significant contributor to the calibration model. In an ideal analysis, this random variable would have been ignored in the construction of the model. Furthermore, the inclusion of this column in $X_2$ has changed the estimated model coefficients so that they no longer represent the true model. The upper 3 x 3
portion of $B_2$, which represents the model for the first three variables in $X_2$, is not equal to $B$, which represents the true model for the same variables. The addition of a column of random numbers has resulted in a model that appears to be better, in that it is more effective at reproducing $Y$, and yet it does not describe the true relationship. This is because MLR uses all of the matrix $X$ to build the model, regardless of whether or not it is relevant in describing the true model. Therefore, an erroneous model can be derived and subsequently used to predict the analytes of new samples. Thus MLR alone often will generate misleading models with subsequent errors in future predictions.

### A 2.3 Summary

- Multiple linear regression estimates $Y$ from $X$ and find the linear combination of variables in $X$ that minimises the errors in reproducing $Y$ (section 3.4.1).

- MLR method is an ideal calibration method for well-behaved systems. This is system with linear responses, no interfering signals, no analyte to analyte interactions, low noise and no collinearities.

- Using irrelevant spectroscopic data may produce smaller error in calibration but misleading results in future predictions.

- Although it seems possible to eliminate using irrelevant data for calibration by careful inspection of the spectral data, but this is a very time consuming and complex method when the relevant spectroscopic data is small as found in NIR responses. Alternative regression techniques that are based on principal components of the spectroscopic data are more effective.
A 3. Computing algorithm for principal component analysis (PCA) and partial least squares (PLS)

The following are the non-linear iterative partial least squares (NIPALS) algorithm (Wold, 1966) for PCA and the orthogonalised PLS algorithm, both are employed in Unscrambler software used in this project.

A 3.1 The NIPALS algorithm for PCA

The algorithm extracts one factor at a time. Each factor is obtained iteratively by repeated regression of \( X \) on scores \( \hat{t} \) to obtain improved \( \hat{p} \), and of \( X \) on these \( \hat{p} \) to improved \( \hat{t} \). The algorithm proceeds as follows:

1. Pre-scale the \( X \)-variables to ensure comparable noise-levels. Then centre the \( X \)-variables, e.g. by subtracting the calibration means \( \bar{x}' \), forming \( X_0 \). Then for factors \( a = 1, 2, \ldots, A \) compute \( \hat{t}_a \) and \( \hat{p}_a \) from \( X_{a-1} \):

2. **Start:**

   Select start values, e.g. \( \hat{t}_a \) = the column in \( X_{a-1} \) that has the highest remaining sum squares.

   Repeat points i) to v) until convergence.

3. i) Improve estimate of loading vector \( \hat{p}_a \) for this factor by projecting the matrix \( X_{a-1} \) on \( \hat{t}_a \), i.e.
\[ \hat{p}_a = (\hat{t}_a \hat{t}_a)^{-1} \hat{t}_a \cdot X_{a-1} \]

ii) Scale length of \( \hat{p}_a \) to 1.0 to avoid scaling ambiguity:

\[ \hat{p}_a = \hat{p}_a (\hat{p}_a \cdot \hat{p}_a)^{0.5} \]

iii) Improve estimate of score \( \hat{t}_a \) for this factor by projecting the matrix \( X_{a-1} \) on \( \hat{p}_a \):

\[ \hat{t}_a = X_{a-1} \hat{p}_a (\hat{p}_a \cdot \hat{p}_a)^{-1} \]

iv) Improve estimate of the eigenvalue \( \hat{\tau}_a \):

\[ \hat{\tau}_a = \hat{t}_a \cdot \hat{t}_a \]

v) Check convergence: If \( \hat{\tau}_a \) minus \( \hat{\tau}_a \) in the previous iteration is smaller than a certain small pre-specified constant, e.g. 0.0001 times \( \hat{\tau}_a \), the method has converged for this factor. If not, go to set i).

Subtract the effect of this factor:

\[ X_a = X_{a-1} - \hat{t}_a \hat{p}_a \cdot \]

and go to Start for the next factor.
A 3.2 Orthogonalised PLS algorithm for one $Y$-variable

A 3.2.1 PLS calibration

C 1  The scaled input variables $X$ and $y$ are first centred, e.g.

$$X_0 = X - \bar{X}'$$ and $$y_0 = y - \bar{y}'$$.

Choose $A_{\text{max}}$ to be higher than the number of phenomena expected in $X$.

For each factor $a = 1, \ldots, A_{\text{max}}$ perform steps C 2.1 - C 2.5:

C 2.1 Use the variability remaining in $y$ to find the loading weights $w_a$, using least squares and the local 'model'

$$X_{a-1} = y_{a-1}w_a' + E$$

and scale the vector to length 1. The solution is

$$\hat{w}_a = c X_{a-1} y_{a-1}$$

where $c$ is the scaling factor that makes the length of the final $\hat{w}_a$ equal to 1, i.e.

$$c = (y_{a-1}'X_{a-1}X_{a-1}'y_{a-1})^{0.5}$$

C 2.2 Estimate the scores $\hat{w}_a$ using the local 'model'
The least squares solution is (since $\hat{w}_a \hat{w}_a = 1$)

$$\hat{t}_a = X_{a-1} \hat{w}_a$$

C 2.3 Estimate the spectral loading $p_a$ using the local 'model'

$$X_{a-1} = \hat{t}_a p_a + E$$

which gives the least squares solution

$$\hat{p}_a = X'_{a-1} \hat{t}_a / \hat{t}_a \hat{t}_a$$

C 2.4 Estimate the chemical loading $q_a$ using the local 'model'

$$y_{a-1} = \hat{t}_a q_a + f$$

which gives the solution

$$\hat{q}_a = y'_{a-1} \hat{t}_a / \hat{t}_a \hat{t}_a$$

C 2.3 Create new $X$ and $y$ residuals by subtracting the estimated effect of this factor:

$$\hat{E} = X_{a-1} - \hat{t}_a \hat{p}_a$$

$$\hat{f} = y_{a-1} - \hat{t}_a \hat{q}_a$$

Compute various summary statistics on these residuals after a factors, summarising $\hat{e}_{ik}$ over objects $i$ and variables $k$, and summarising $\hat{f}_i$ over $i$ objects.
Replace the former \( X_{a-1} \) and \( y_{a-1} \) by the new residuals \( \hat{E} \) and \( \hat{f} \) and increase \( a \) by 1, i.e. set

\[
X_a = \hat{E} \\
y_a = \hat{f} \\
a = a + 1
\]

C 3 Determine \( A \), the number of valid PLS factors to retain in the calibration model.

C 4 Compute \( \hat{b}_0 \) and \( \hat{b} \) for \( A \) PLS factors, to be used in the predictor

\[
\hat{y} = 1 \hat{b}_0 + X \hat{b} \quad \text{(optional, see P4 below)}
\]

\[
\hat{b} = \hat{W} (\hat{P}' \hat{W})^{-1} \hat{q}
\]

\[
\hat{b}_0 = \bar{y} - \bar{x}' \hat{b}
\]
A 3.2.2 Full prediction

For each new prediction object \( i = 1, 2, \ldots \) perform steps P1 to P3, or alternatively, step P4.

P 1  Scale input data \( x_i \) like for the calibration variables. Then compute

\[
x_i' = x_i - \bar{x}'
\]

where \( \bar{x} \) is the centre for the calibration objects.

For each factor \( a = 1, \ldots, A \) perform steps P 2.1 - P 2.2.

P 2.1  Find \( \hat{t}_{i,a} \) according to the formula in C 2.2 e.e.

\[
\hat{t}_{i,a} = x_i' \hat{w}_{a}
\]

P 2.2  Compute new residual \( x_i,a = x_i,a-1 - \hat{t}_{i,a} \hat{p}_a \).

If \( a < A \), increase \( a \) by 1 and go to P 2.1. If \( a = A \), go to P 3.

P 3  Predict \( y_i \) by \( \hat{y}_i = \bar{y} + \sum_{a=1}^{A} \hat{t}_{i,a} \hat{q}_a \).

Compute outlier statistics on \( x_{i,A} \) and \( \hat{t}_i \).

A 3.2.3 Short prediction

P 4  Alternatively to steps P1-P3, find \( \hat{y} \) by using \( \hat{b}_0 \) and \( \hat{b} \) in C 4, i.e.

\[
\hat{y}_i = \hat{b}_0 + x_i' \hat{b}
\]
A 4. Comparison of the in-house built with two with commercial spectrophotometers

Two commercial spectrophotometers NIRS6500 (Perstrop Analytical, Maidenhead, UK) and Brimrose AOTF (Brimrose, Hemel Hempstead, UK) have been selected for comparison with the in-house built spectrophotometer.

<table>
<thead>
<tr>
<th>Spectrophotometer</th>
<th>In-house built</th>
<th>NIRS6500</th>
<th>Brimrose AOTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost (approx.)</td>
<td>£5000</td>
<td>£70,000</td>
<td>£39,000</td>
</tr>
<tr>
<td>Type</td>
<td>Grating</td>
<td>Grating</td>
<td>AOTF</td>
</tr>
<tr>
<td>Detector</td>
<td>PbS</td>
<td>PbS</td>
<td>InGaAs</td>
</tr>
<tr>
<td>Spectral range</td>
<td>1900-2500 nm</td>
<td>400-2500 nm</td>
<td>1100-2300 nm</td>
</tr>
<tr>
<td>Data interval</td>
<td>6 nm</td>
<td>2 nm</td>
<td>2 nm</td>
</tr>
<tr>
<td>Acquisition time</td>
<td>540 secs</td>
<td>90 secs</td>
<td>8 secs</td>
</tr>
<tr>
<td>Sample presentation</td>
<td>Cell</td>
<td>Cell or fibre probe</td>
<td>fibre probe</td>
</tr>
<tr>
<td>Mode</td>
<td>Transmittance</td>
<td>Transmittance</td>
<td>Transmittance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reflectance</td>
<td>Reflectance</td>
</tr>
<tr>
<td>Moving parts</td>
<td>yes</td>
<td>yes</td>
<td>none</td>
</tr>
<tr>
<td>Signal to noise</td>
<td>approx. 450:1</td>
<td>approx. 1000:1</td>
<td>approx. 1000:1</td>
</tr>
</tbody>
</table>

Table A 4.1 Technical comparisons of the in-house built spectrophotometer with two commercial spectrophotometers.

1 In-house built spectrophotometer can acquire spectrum from 1100-2500 nm but it will require a longer linear stage to transport the PbS detector across the full output port of the grating.
In-house built spectrophotometer can also provide spectral data points at every 2 nm but it is restricted by the acquisition time. For instance, if data interval is provided at every 3 nm instead of 6 nm (540 seconds) then the total acquisition time will be approximately 1080 seconds.

In-house built spectrophotometer can also be used for reflectance measurements, but it will require careful optical alignment whenever changing the mode of operations.

Signal to noise ratio is estimated from the maximum unit of absorbance that spectrophotometer can be measured. Both NIRS6500 and Brimrose can measure up to 3 Au unit without reaching of saturation, e.g. 1000:1.
Appendix 5

A 5. Identification of NIR response to biological contaminants using a commercial spectrophotometer

A 5.1 Introduction

The contaminants of interest are protein, RNA and cell debris found in first clarified yeast homogenate. This short experiment has shown that NIR responses from these contaminants are identified by a commercial spectrophotometer (NIRS6500, Perstrop Analytical, Berkshire, UK). The protein and RNA samples used in this experiment are synthetic, prepared by increasing contaminant levels in the clarified yeast homogenate with standard material. Then the NIR spectra of clarified yeast homogenate is compared with the synthetic samples. For cell debris contaminants identification, clarified homogenate is compared with further flocculated and clarified samples, since the flocculation aid removal of further cell debris and thus differences in the flocculated samples are expected.

The results have found that NIR signature of cell debris is presented throughout the complete NIR spectrum, appropriate protein and RNA signatures are identified between 1900 to 2500 nm.
A 5.2 Material and methods

Preparation of clarified yeast homogenate

Bakers yeast (The Distillers Co. Ltd., Sutton, Surrey, U.K.) was resuspended in KH₂PO₄ buffer (100mM, pH 6.5) at concentration of 250g wet pack weight L⁻¹. The yeast suspension was disrupted using a Gaulin high-pressure homogeniser (Lab 40, APV, Lubeck, Germany), for 2 discrete passes at 1200 bar(g) and maintained at approximately 4°C. The pH was adjusted to 6.5 with HCl (4M).

Polyethyleneimine (PEI) (Fluka) 2%w/v was prepared in deionised water. PEI stock solution (2%w/v, 100 mL) was added to yeast homogenate (900mL) and stirred for 300s. The homogenate was clarified at 10,000g for 0.5L (J2MI, Beckman Instruments, Highwycombe, Bucks). The clarified homogenate was decanted off the cell debris and less dense lipid layer was removed. It was then stored at 4°C prior to the addition of the following reagents.

Adding contaminants to clarified yeast homogenate

Debris; addition of 500 mg washed debris into 5 ml of clarified homogenate. The washed debris was obtained from preparation of clarified yeast homogenate.

Protein; addition of 1g bovine serum albumin (BSA, Fraction V, Sigma Ltd) into 5 ml of clarified homogenate.

RNA; addition of 250 mg ribonucleic acid (RNA, BDH) into 5 ml of clarified homogenate.

NIR spectrum

All NIR spectra of prepared samples were collected using a sophisticated commercial spectrophotometer (NIRS6500, Perstrop Analytical, Berkshire, UK). The scanning specifications were; 1900 to 2498 nm, 2 nm data interval and the final spectrum was averaged of 32 scans. The sample was stored in a 1 mm pathlength glass cell and loaded to the spectrophotometer’s sampling compartment for measurements.
A 5.3 Results

A 5.3.1 NIR signature of debris in clarified yeast homogenate

The absorbance spectrum of both clarified yeast homogenate and clarified yeast homogenate with added debris samples are shown in figure A4.1a. These spectra have shown that the debris signature is presented across the recorded spectrum (1900 - 2500 nm) and the absorption has increased in the debris sample. The second derivative (figure A4.1b) spectra of these samples have shown very little difference between the two samples. This indicated that absorption of the sample with added debris is increase by an offset. The noisy regions shown in debris spectra were due to spectrophotometer input limitation (i.e. the light absorption by the sample was too high, which gave too little optical energy for spectrophotometer to detect).

Fig. A 5.1 a) The absorption spectra of clarified yeast homogenate and debris samples. b) Their second derivative spectra. Both spectra were recorded by a commercial spectrophotometer (NIRS6500, Perstrop analytical).
A 5.3.2 NIR signature of protein in clarified yeast homogenate

The absorbance spectrum of both clarified yeast homogenate and protein samples are shown in figure A4.2a. These spectra have shown that the protein signature was presented in the indicated NIR spectrum (1900 - 2500 nm). The second derivative (see figure A4.2b) spectra of these samples have further confirmed the NIR signatures of protein.

![Graphs showing absorption spectra and second derivative spectra for clarified yeast homogenate and protein](image)

Fig. A 5.2 a) The absorption spectra of clarified yeast homogenate and protein samples. b) The second derivative spectra. Both spectra were recorded by a commercial spectrophotometer (NIRS6500, Perstrop analytical).
A 5.3.3 NIR signature of RNA in clarified yeast homogenate

The absorbance spectrum of both clarified yeast homogenate and RNA samples are shown in figure A4.3a. These spectra have shown that the RNA signature was presented in the indicated NIR spectrum (1900 - 2500 nm). The second derivative (see figure A4.3b) spectra of these samples have further confirmed the NIR signatures of RNA.

![Absorbance/2nd Derivative Spectra](image)

Fig. A 5.3 a) The absorption spectra of clarified yeast homogenate and RNA samples. b) Their second derivative spectra. Both spectra were recorded by a commercial spectrophotometer (NIRS6500, Perstrop analytical).
A 5.4 Summary

- NIR signatures of debris, protein and RNA contaminants in clarified yeast homogenate have been identified.

- NIR signatures were found in 1900 to 2500 nm.

- Presents of debris in clarified yeast homogenate caused scattering effects and reduced light transmission through samples.

- The second derivative spectrum have been used to confirm NIR signatures of protein and RNA.
A 6. Concentration of contaminants in single contaminant calibration samples

Samples were prepared for single contaminants calibration using PLS and the in-house built spectrophotometer in section 4.8. Different concentrations of contaminants stock solutions were prepared and each was added to 500 g L⁻¹ clarified yeast homogenate to create the final calibration samples. The concentrations of protein, RNA stock solutions and the final concentration of contaminants in these calibration samples are listed in table A5.1 and A5.2 respectively. Note that the estimated protein and RNA contaminants in the 500 g L⁻¹ clarified yeast homogenate are 52 and 23 g L⁻¹ and these concentrations were used to calculated the known concentration in the final calibration samples.

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>Protein stock concentration (/g L⁻¹)</th>
<th>Final known concentration (/g L⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>26.0</td>
</tr>
<tr>
<td>2</td>
<td>2.6</td>
<td>27.3</td>
</tr>
<tr>
<td>3</td>
<td>5.2</td>
<td>28.6</td>
</tr>
<tr>
<td>4</td>
<td>10.4</td>
<td>31.2</td>
</tr>
<tr>
<td>5</td>
<td>15.6</td>
<td>33.8</td>
</tr>
<tr>
<td>6</td>
<td>20.8</td>
<td>36.4</td>
</tr>
<tr>
<td>7</td>
<td>26.0</td>
<td>39.0</td>
</tr>
<tr>
<td>8</td>
<td>31.2</td>
<td>41.6</td>
</tr>
<tr>
<td>9</td>
<td>36.4</td>
<td>44.2</td>
</tr>
<tr>
<td>10</td>
<td>41.6</td>
<td>46.8</td>
</tr>
<tr>
<td>11</td>
<td>46.8</td>
<td>49.4</td>
</tr>
<tr>
<td>12</td>
<td>52.0</td>
<td>52.0</td>
</tr>
</tbody>
</table>

Table A 6.1 Protein stock concentration and final known concentration of protein in calibration samples.
### Table A 6.2 RNA stock concentration and final known concentration of RNA in calibration samples.

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>RNA stock concentration (/g L(^{-1}))</th>
<th>Final known concentration (/g L(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>11.500</td>
</tr>
<tr>
<td>2</td>
<td>1.15</td>
<td>12.075</td>
</tr>
<tr>
<td>3</td>
<td>2.30</td>
<td>12.650</td>
</tr>
<tr>
<td>4</td>
<td>4.60</td>
<td>13.80</td>
</tr>
<tr>
<td>5</td>
<td>6.90</td>
<td>14.950</td>
</tr>
<tr>
<td>6</td>
<td>9.20</td>
<td>16.100</td>
</tr>
<tr>
<td>7</td>
<td>11.5</td>
<td>17.250</td>
</tr>
<tr>
<td>8</td>
<td>13.8</td>
<td>18.400</td>
</tr>
<tr>
<td>9</td>
<td>16.1</td>
<td>19.550</td>
</tr>
<tr>
<td>10</td>
<td>18.4</td>
<td>20.700</td>
</tr>
<tr>
<td>11</td>
<td>20.7</td>
<td>21.850</td>
</tr>
<tr>
<td>12</td>
<td>23.0</td>
<td>23.000</td>
</tr>
</tbody>
</table>
A 7. Composition of the *add-back* calibration samples

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>C.Y.H. 0.15ml</th>
<th>Debris 0.15ml</th>
<th>BSA 0.15ml</th>
<th>RNA 0.15ml</th>
<th>Sample Number</th>
<th>C.Y.H. 0.15ml</th>
<th>Debris 0.15ml</th>
<th>BSA 0.15ml</th>
<th>RNA 0.15ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45ml</td>
<td>A X m</td>
<td></td>
<td></td>
<td>25</td>
<td>0.45ml</td>
<td>C X m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&quot;</td>
<td>A X n</td>
<td></td>
<td></td>
<td>26</td>
<td>&quot;</td>
<td>C X n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>&quot;</td>
<td>A X o</td>
<td></td>
<td></td>
<td>27</td>
<td>&quot;</td>
<td>C X o</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>&quot;</td>
<td>A X p</td>
<td></td>
<td></td>
<td>28</td>
<td>&quot;</td>
<td>C X p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>&quot;</td>
<td>A Y m</td>
<td></td>
<td></td>
<td>29</td>
<td>&quot;</td>
<td>C Y m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>&quot;</td>
<td>A Y n</td>
<td></td>
<td></td>
<td>30</td>
<td>&quot;</td>
<td>C Y n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>&quot;</td>
<td>A Y o</td>
<td></td>
<td></td>
<td>31</td>
<td>&quot;</td>
<td>C Y o</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>&quot;</td>
<td>A Y p</td>
<td></td>
<td></td>
<td>32</td>
<td>&quot;</td>
<td>C Y p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>&quot;</td>
<td>A Z m</td>
<td></td>
<td></td>
<td>33</td>
<td>&quot;</td>
<td>C Z m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&quot;</td>
<td>A Z n</td>
<td></td>
<td></td>
<td>34</td>
<td>&quot;</td>
<td>C Z n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>&quot;</td>
<td>A Z o</td>
<td></td>
<td></td>
<td>35</td>
<td>&quot;</td>
<td>C Z o</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>&quot;</td>
<td>A Z p</td>
<td></td>
<td></td>
<td>36</td>
<td>&quot;</td>
<td>C Z p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>&quot;</td>
<td>B X m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>&quot;</td>
<td>B X n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>&quot;</td>
<td>B X o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>&quot;</td>
<td>B X p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>&quot;</td>
<td>B Y m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>&quot;</td>
<td>B Y n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>&quot;</td>
<td>B Y o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>&quot;</td>
<td>B Y p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>&quot;</td>
<td>B Z m</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>&quot;</td>
<td>B Z n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>&quot;</td>
<td>B Z o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>&quot;</td>
<td>B Z p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where

- A = 0.15 mL of buffer
- B = 75 g L\(^{-1}\) of washed debris
- C = 150 g L\(^{-1}\) of washed debris
- X = 0.15 mL of buffer
- Y = 78 g L\(^{-1}\) of BSA
- Z = 156 g L\(^{-1}\) of BSA
- m = 0.15 mL of buffer
- n = 13.6 g L\(^{-1}\) of RNA
- o = 40.8 g L\(^{-1}\) of RNA
- p = 68 g L\(^{-1}\) of RNA

Table A 7.1 Composition of *add-back* calibration samples for factorial design experiments. Note C.Y.H. stands for clarified yeast homogenate at the concentration of 500 g L\(^{-1}\) and buffer refers to the addition of 100 mM phosphate buffer at 6.5pH as the controls.
<table>
<thead>
<tr>
<th>Sample Number</th>
<th>Expected Protein (g L^-1)</th>
<th>Expected RNA (g L^-1)</th>
<th>Sample Number</th>
<th>Expected Protein (g L^-1)</th>
<th>Expected RNA (g L^-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>11.5</td>
<td>25</td>
<td>28</td>
<td>11.5</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>13.8</td>
<td>26</td>
<td>28</td>
<td>13.8</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>18.4</td>
<td>27</td>
<td>28</td>
<td>18.4</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>23.0</td>
<td>28</td>
<td>28</td>
<td>23.0</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
<td>11.5</td>
<td>29</td>
<td>39</td>
<td>11.5</td>
</tr>
<tr>
<td>6</td>
<td>39</td>
<td>13.8</td>
<td>30</td>
<td>39</td>
<td>13.8</td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>18.4</td>
<td>31</td>
<td>39</td>
<td>18.4</td>
</tr>
<tr>
<td>8</td>
<td>39</td>
<td>23.0</td>
<td>32</td>
<td>39</td>
<td>23.0</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
<td>11.5</td>
<td>33</td>
<td>52</td>
<td>11.5</td>
</tr>
<tr>
<td>10</td>
<td>52</td>
<td>13.8</td>
<td>34</td>
<td>52</td>
<td>13.8</td>
</tr>
<tr>
<td>11</td>
<td>52</td>
<td>18.4</td>
<td>35</td>
<td>52</td>
<td>18.4</td>
</tr>
<tr>
<td>12</td>
<td>52</td>
<td>23.0</td>
<td>36</td>
<td>52</td>
<td>23.0</td>
</tr>
<tr>
<td>13</td>
<td>28</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>28</td>
<td>13.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>18.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>28</td>
<td>23.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>39</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>39</td>
<td>13.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>39</td>
<td>18.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>39</td>
<td>23.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>52</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>52</td>
<td>13.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>52</td>
<td>18.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>52</td>
<td>23.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A 7.2 Based on the estimated background of 52 g L^-1 protein and 23 g L^-1 RNA in 500 g L^-1 clarified yeast homogenate, the expected protein and RNA concentrations in the add-back calibration samples are calculated.
A 8. Optical density and assay results

A 8.1 Optical density at 650 nm of all calibration and validation samples

<table>
<thead>
<tr>
<th>Samples No.</th>
<th>O.D. 650nm (Au)</th>
<th>Samples No.</th>
<th>O.D. 650nm (Au)</th>
<th>Samples No.</th>
<th>O.D. 650nm (Au)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.556</td>
<td>13</td>
<td>0.900</td>
<td>25</td>
<td>1.509</td>
</tr>
<tr>
<td>2</td>
<td>0.665</td>
<td>14</td>
<td>1.089</td>
<td>26</td>
<td>1.508</td>
</tr>
<tr>
<td>3</td>
<td>0.602</td>
<td>15</td>
<td>1.104</td>
<td>27</td>
<td>1.432</td>
</tr>
<tr>
<td>4</td>
<td>0.581</td>
<td>16</td>
<td>1.072</td>
<td>28</td>
<td>1.623</td>
</tr>
<tr>
<td>5</td>
<td>0.591</td>
<td>17</td>
<td>1.096</td>
<td>29</td>
<td>1.544</td>
</tr>
<tr>
<td>6</td>
<td>0.600</td>
<td>18</td>
<td>1.061</td>
<td>30</td>
<td>1.546</td>
</tr>
<tr>
<td>7</td>
<td>0.595</td>
<td>19</td>
<td>1.062</td>
<td>31</td>
<td>1.582</td>
</tr>
<tr>
<td>8</td>
<td>0.534</td>
<td>20</td>
<td>1.038</td>
<td>32</td>
<td>1.561</td>
</tr>
<tr>
<td>9</td>
<td>0.568</td>
<td>21</td>
<td>1.090</td>
<td>33</td>
<td>1.457</td>
</tr>
<tr>
<td>10</td>
<td>0.570</td>
<td>22</td>
<td>1.135</td>
<td>34</td>
<td>1.515</td>
</tr>
<tr>
<td>11</td>
<td>0.536</td>
<td>23</td>
<td>0.990</td>
<td>35</td>
<td>1.685</td>
</tr>
<tr>
<td>12</td>
<td>0.598</td>
<td>24</td>
<td>1.028</td>
<td>36</td>
<td>1.577</td>
</tr>
</tbody>
</table>

Table A 8.1 Optical density at 650 nm of all *add-back* calibration samples
<table>
<thead>
<tr>
<th>PEI %v/v</th>
<th>Process stream ‘A’</th>
<th>Process stream ‘B’</th>
<th>PEI (%v/v)</th>
<th>Process stream ‘V1’</th>
<th>Process stream ‘V2’</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.734</td>
<td>0.679</td>
<td>0</td>
<td>0.719</td>
<td>0.682</td>
</tr>
<tr>
<td>2</td>
<td>0.693</td>
<td>0.653</td>
<td>2</td>
<td>0.682</td>
<td>0.638</td>
</tr>
<tr>
<td>6</td>
<td>0.602</td>
<td>0.599</td>
<td>6</td>
<td>0.608</td>
<td>0.623</td>
</tr>
<tr>
<td>10</td>
<td>0.487</td>
<td>0.408</td>
<td>10</td>
<td>0.487</td>
<td>0.383</td>
</tr>
<tr>
<td>15</td>
<td>0.078</td>
<td>0.066</td>
<td>15</td>
<td>0.078</td>
<td>0.031</td>
</tr>
<tr>
<td>20</td>
<td>0.117</td>
<td>0.010</td>
<td>20</td>
<td>0.060</td>
<td>0.017</td>
</tr>
<tr>
<td>25</td>
<td>0.136</td>
<td>0.030</td>
<td>25</td>
<td>0.058</td>
<td>0.014</td>
</tr>
<tr>
<td>30</td>
<td>0.014</td>
<td>0.010</td>
<td>30</td>
<td>0.014</td>
<td>0.011</td>
</tr>
<tr>
<td>35</td>
<td>0.023</td>
<td>0.021</td>
<td>35</td>
<td>0.024</td>
<td>0.026</td>
</tr>
<tr>
<td>40</td>
<td>0.024</td>
<td>0.047</td>
<td>40</td>
<td>0.023</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table A 8.2 Optical density at 650 nm of two sets of process stream calibration samples and two sets of process stream validation samples. The concentrations of PEI used in flocculating the samples are also listed.
### A 8.2 Protein assays of all calibration and validation samples

<table>
<thead>
<tr>
<th>Samples No.</th>
<th>Protein (g L(^{-1}))</th>
<th>Samples No.</th>
<th>Protein (g L(^{-1}))</th>
<th>Samples No.</th>
<th>Protein (g L(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.55</td>
<td>13</td>
<td>37.84</td>
<td>25</td>
<td>37.99</td>
</tr>
<tr>
<td>2</td>
<td>32.27</td>
<td>14</td>
<td>39.29</td>
<td>26</td>
<td>39.16</td>
</tr>
<tr>
<td>3</td>
<td>36.64</td>
<td>15</td>
<td>39.35</td>
<td>27</td>
<td>39.54</td>
</tr>
<tr>
<td>4</td>
<td>44.51</td>
<td>16</td>
<td>37.19</td>
<td>28</td>
<td>40.10</td>
</tr>
<tr>
<td>5</td>
<td>31.76</td>
<td>17</td>
<td>56.29</td>
<td>29</td>
<td>56.03</td>
</tr>
<tr>
<td>6</td>
<td>31.62</td>
<td>18</td>
<td>56.17</td>
<td>30</td>
<td>54.47</td>
</tr>
<tr>
<td>7</td>
<td>45.76</td>
<td>19</td>
<td>57.21</td>
<td>31</td>
<td>56.12</td>
</tr>
<tr>
<td>8</td>
<td>34.76</td>
<td>20</td>
<td>54.64</td>
<td>32</td>
<td>56.23</td>
</tr>
<tr>
<td>9</td>
<td>63.83</td>
<td>21</td>
<td>75.04</td>
<td>33</td>
<td>72.08</td>
</tr>
<tr>
<td>10</td>
<td>47.44</td>
<td>22</td>
<td>59.64</td>
<td>34</td>
<td>69.05</td>
</tr>
<tr>
<td>11</td>
<td>60.96</td>
<td>23</td>
<td>32.80</td>
<td>35</td>
<td>64.69</td>
</tr>
<tr>
<td>12</td>
<td>69.35</td>
<td>24</td>
<td>69.84</td>
<td>36</td>
<td>66.82</td>
</tr>
</tbody>
</table>

Table A 8.3 Protein assays of all add-back calibration samples

<table>
<thead>
<tr>
<th>PEI %v/v</th>
<th>Protein (g L(^{-1}))</th>
<th>PEI %v/v</th>
<th>Protein (g L(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process stream 'A'</td>
<td>Process stream 'B'</td>
<td>Process stream 'V1'</td>
</tr>
<tr>
<td>0</td>
<td>27.29</td>
<td>32.06</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>26.68</td>
<td>31.28</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>23.80</td>
<td>28.43</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>20.60</td>
<td>21.69</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>16.00</td>
<td>18.46</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>14.22</td>
<td>17.79</td>
<td>20</td>
</tr>
<tr>
<td>25</td>
<td>14.07</td>
<td>16.67</td>
<td>25</td>
</tr>
<tr>
<td>30</td>
<td>13.66</td>
<td>15.44</td>
<td>30</td>
</tr>
<tr>
<td>35</td>
<td>13.87</td>
<td>25.89</td>
<td>35</td>
</tr>
<tr>
<td>40</td>
<td>15.06</td>
<td>17.04</td>
<td>40</td>
</tr>
</tbody>
</table>

Table A 8.4 Protein assays of two sets of process stream calibration samples and two sets of process stream validation samples. The concentrations of PEI used in flocculating the samples are also listed.
A 8.3 RNA assays of all calibration and validation samples

<table>
<thead>
<tr>
<th>Samples No.</th>
<th>RNA (g L(^{-1}))</th>
<th>Samples No.</th>
<th>RNA (g L(^{-1}))</th>
<th>Samples No.</th>
<th>RNA (g L(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.21</td>
<td>13</td>
<td>12.30</td>
<td>25</td>
<td>13.58</td>
</tr>
<tr>
<td>2</td>
<td>13.40</td>
<td>14</td>
<td>16.31</td>
<td>26</td>
<td>14.27</td>
</tr>
<tr>
<td>3</td>
<td>21.41</td>
<td>15</td>
<td>21.52</td>
<td>27</td>
<td>19.67</td>
</tr>
<tr>
<td>4</td>
<td>23.13</td>
<td>16</td>
<td>22.72</td>
<td>28</td>
<td>25.18</td>
</tr>
<tr>
<td>5</td>
<td>14.30</td>
<td>17</td>
<td>12.07</td>
<td>29</td>
<td>13.11</td>
</tr>
<tr>
<td>6</td>
<td>10.61</td>
<td>18</td>
<td>16.93</td>
<td>30</td>
<td>15.04</td>
</tr>
<tr>
<td>7</td>
<td>17.00</td>
<td>19</td>
<td>18.94</td>
<td>31</td>
<td>23.29</td>
</tr>
<tr>
<td>8</td>
<td>15.22</td>
<td>20</td>
<td>29.62</td>
<td>32</td>
<td>24.57</td>
</tr>
<tr>
<td>9</td>
<td>12.30</td>
<td>21</td>
<td>13.11</td>
<td>33</td>
<td>12.77</td>
</tr>
<tr>
<td>10</td>
<td>10.97</td>
<td>22</td>
<td>14.69</td>
<td>34</td>
<td>12.15</td>
</tr>
<tr>
<td>11</td>
<td>18.66</td>
<td>23</td>
<td>18.74</td>
<td>35</td>
<td>16.43</td>
</tr>
<tr>
<td>12</td>
<td>31.93</td>
<td>24</td>
<td>26.34</td>
<td>36</td>
<td>29.12</td>
</tr>
</tbody>
</table>

Table A 8.5 RNA assays of all add-back calibration samples

<table>
<thead>
<tr>
<th>PEI (%v/v)</th>
<th>RNA (g L(^{-1}))</th>
<th>PEI (%v/v)</th>
<th>RNA (g L(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process stream ‘V1’</td>
<td>17.87</td>
<td>Process stream ‘V2’</td>
<td>17.83</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>15.52</td>
<td>Process stream ‘V2’</td>
<td>15.74</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>13.81</td>
<td>Process stream ‘V2’</td>
<td>13.03</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>12.29</td>
<td>Process stream ‘V2’</td>
<td>10.20</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>10.20</td>
<td>Process stream ‘V2’</td>
<td>7.97</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>6.58</td>
<td>Process stream ‘V2’</td>
<td>8.15</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>8.41</td>
<td>Process stream ‘V2’</td>
<td>7.82</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>5.01</td>
<td>Process stream ‘V2’</td>
<td>5.73</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>5.39</td>
<td>Process stream ‘V2’</td>
<td>7.27</td>
</tr>
<tr>
<td>Process stream ‘V1’</td>
<td>6.22</td>
<td>Process stream ‘V2’</td>
<td>7.48</td>
</tr>
</tbody>
</table>

Table A 8.6 RNA assays of two sets of process stream calibration samples and two sets of process stream validation samples. The concentrations of PEI used in flocculating the samples are also listed.
A 9. The effect of Polyethyleneimine (PEI) in yeast homogenate

A 9.1 Introduction

The aim is to study the significant spectral variation in yeast homogenate caused by low concentration of Polyethyleneimine (PEI). It has been known that addition of PEI into yeast homogenate creating flocculation, hence, flocs and the size of these flocs depend on the concentration of PEI. An experiment has been conducted and found that very low concentration of PEI causes a significant decrease in transmission of NIR energy.

A 9.2 Materials and Methods

A 9.2.1 Materials

The yeast homogenate was prepared from packed bakers yeast, *Saccharomyces cerevisiae* supplied by Distillers Company Ltd. (Sutton, Surrey, UK). The bakers yeast was re-suspended in 100 mM phosphate buffer, pH 6.5, to a final cell concentration of 150 g wet packed weight L\(^{-1}\). The yeast suspension was disrupted using a high pressure homogeniser (Model Lab 40; APV Gaulin, APV, Crawley, Sussex, UK) for two discrete passes at 1200 bar g and maintained at approximately 4\(^{\circ}\)C by cooling.
Various samples were prepared by adding different concentration of Polyethleneimine (PEI, Fluka Chemicals, Dorset, UK) to yeast homogenate. These samples have 0.1, 0.2, 0.3 and 0.4%w/v PEI in the yeast homogenate.

Distilled water and 2%w/v of PEI were also prepared for NIR spectral scans.

**A 9.2.2 Methods**

The NIR spectra are obtained using an in-house build spectrophotometer described in chapter 4. The scanning procedures are the same as that described in section 4.8.2.4. Each spectrum is recorded and smoothed for wavelength 1900 to 2500 nm and data interval at every 6 nm.

**A 9.3 Results and discussions**

In figure A8.1a, the normalised transmission spectra of distilled water and 2%w/v PEI in distilled water are shown. This has shown that the transmission peak of 2%w/v PEI in distilled water has decreased by only 4.5% compared to the peak of distilled water. The normalised spectrum of yeast homogenate is shown in figure A8.1b together with the spectra of various yeast homogenate samples that had different concentration of PEI added. The figure has shown the transmission of the yeast homogenate with 0.1%w/v of PEI has been significantly reduced and similarly with the subsequent samples of higher PEI concentrations. Note that the concentration of 2%w/v PEI is beyond the range used in the process monitoring experiments, for example, only up to 35 to 40% of the 2%w/v PEI are used in processes since the optimum flocculation is achieved below these concentrations (see 5.5.6).

Since figure A8.1a has shown that 2%w/v of PEI alone led to a 4.5% decreased in transmission at 1920 nm. Then the significant decrease (figure A8.1b) in the transmission of yeast homogenate with 0.1%w/v PEI by 40% has shown the effect of
floculation and flocs size. There the larger the flocs in the yeast homogenate, hence the particle size distribution, higher the light scattering occurred, therefore reducing the overall NIR transmission.

![Diagram](image)

Fig. A 9.1 Normalised transmission spectra of distilled water and 2% w/v PEI in distilled water. b) Normalised spectra of yeast homogenate and various concentrations of PEI in homogenate.

### A 9.4 Summary

Flocculation in yeast homogenate is generated by adding PEI. The flocs depended on the extend of flocculation which results in the increase in particle size distribution.

The transmission region at 1920 nm of 2% w/v PEI in distilled water is reduced by about 4.5%, while comparing the transmission of yeast homogenate and yeast homogenate with only 0.1% w/v of PEI has shown a significant 40% reduction in NIR transmission.

This significant reduction of transmission in the flocculated yeast homogenate is due to higher scattering affected by larger particle size distribution.


References


Pearson K, 1901, On lines and planes of closest fit to systems of points in space, Phil. (Ser. 6) 2: 559-572.


