The University of Hull

Development and Application of Calibration Free Techniques for Reaction Profiling

Being a Thesis submitted for the degree of

Doctor of Philosophy

in the

University of Hull

By

Selena Elizabeth Richards MChem (Hons)

April 2007

Declaration

The work submitted in this thesis was carried out in the Department of Chemistry, University of Hull between October 2001 and August 2004. Except where indicated by references, the work in this thesis is original and has not been submitted for any other degree.

Selena Richards

2007

Table of Contents

TABLE OF CONTENTS **3**

ACKNOWLEDGEMENTS **5**

ABSTRACT **6**

ABBREVIATIONS **7**

MATHEMATICAL NOTATION **10**

GLOSSARY **12**

PROLOGUE **14**

RATIONALE 17

AIMS OF THESIS 19

I INTRODUCTION AND TOOLS **22**

I.1 PRELIMINARIES 23

I.2 CALIBRATION FREE ANALYSIS 29

I.3 MODEL VALIDATION 64

II PROCESSES UNDERSTUDY **69**

II.1 EXPLORATORY AND QUANTITATIVE ANALYSIS OF THE CATALYSED ASYMMETRIC TRANSFER HYDROGENATION REACTION OF A PROCHIRAL IMINE 70

II.2 MULTI-WAY PENALTY ALTERNATING LEAST SQUARES 107

II.3 QUALITATIVE ANALYSIS OF THE BASE CATALYSED ESTERIFICATION REACTION OF ACETIC ANHYDRIDE USING NWAY P-ALS 117

II.4 QUANTITATIVE ITERATIVE TARGET TRANSFORMATION FACTOR ANALYSIS 139

II.5 EXPLORATORY ANALYSIS OF SIMULATED HPLC-DAD USING QITTFA 153

II.6 SEMI-QUANTITATIVE ANALYSIS OF CALIBRATION SAMPLES FROM THE BP VINYL ACETATE PROCESS 171

III CONCLUSIONS **190**

III.1 GENERAL CONCLUSIONS 191

IV REFERENCES 194

V FURTHER WORK 208

VI APPENDIX 209

Acknowledgements

I would like to thank my supervisor Dr Walmsley for his support and guidance throughout my PhD. I would like to extend my gratitude to Dr Flåten, who greatly assisted in developing my programming skills in MATLAB®. I would like to thank Prof. Blackmond and Miss Ropic for access to the catalysed asymmetric transfer hydrogenation data and Dr Becker and Dr Lynch for assistance with the vinyl acetate calibration problem. I would like to thank Prof. Tauler and Dr Juan for the opportunity to visit the Universitat de Barcelona, Solution Equilibria and Chemometrics group and Prof. Gemperline for the opportunity to visit East Carolina University and for their support and guidance. Finally, I would like to acknowledge the EPSRC and CPACT for financial sponsorship.

Abstract

In this technological information age, dimension reduction methods are key because they enable the almost instantaneous extraction of relevant information from large complex data sets. This is particularly crucial within the process analytical environment where "right-first-time" and "just-in-time" approaches push the technological and economic persuasions of a manufacturing culture. In line with this paradigm, mathematical tools which decompose highly complex multivariate and multicomponent measurements into their lowest dimensionality without the need of *a priori* knowledge have been used to provide intelligence regarding different processes. This intelligence includes the pure spectra and concentration profiles of the reaction constituents, by-products and short lived intermediates. These tools are known as calibration free techniques (CFT) and in this thesis they have been developed and applied to complex academic and industrial problems, which include the rhodium catalysed asymmetric transfer hydrogenation reaction of a prochiral imine, the pyridine catalysed esterification reaction of acetic anhydride and the vinyl acetate monomer process. These chemical systems are typically deficient in *a priori* information leading to the generation of a chemical or dynamic process model. The application of CFTs are favourable because they do not require *a priori* information to provide intelligence regarding the reaction constituents which may lead to a reduction in process cost and increased efficiency of the manufacturing process.

Abbreviations

ALS	Alternating Least Squares
CATHy	Catalysed Asymmetric Transfer Hydrogenation Reaction
CFT	Calibration Free Techniques
CLS	Classical Least Squares
CPAC	Center for Process Analytical Chemistry
CPACT	Centre for Process Analytics and Control Technology
Csel	Equality constraint Concentration
DAD	Diode Array Detection
ee	Enantiometic excess
EFA	Evolving Factor Analysis
EPSRC	Engineering and Physical Sciences Research Council
FA	Factor Analysis
FDA	Food and Drug Administration
FID	Flame Ionisation Detector
FNNLS	Fast Non-Negative Least Squares
FSW-EFA	Fixed Size Window- Evolving Factor Analysis
FTIR	Fourier Transform Infrared
GC	Gas Chromatography
GRR	Generalised Ridge Regression
GUI	Graphical User Interface
HELP	Heuristic Latent projective Analysis
HPLC	High Performance Liquid Chromatography
HPLC-DAD	High Performance Liquid Chromatography-Diode Array Detection
ILS	Inverse Least Squares
IR	Infrared
ITTFA	Iterative Target Transformation Factor Analysis
LC	Liquid Chromatography
LOF	Lack-of-fit
LS	Least Squares
MCR	Multivariate Curve Resolution
MCR-ALS	Multivariate Curve Resolution - Alternating Least Square
MIR	Mid-infrared

MLR	Multiple Linear Regression
MSEP	Mean Squared Error of Prediction
NIR	Near Infrared
NMR	Nuclear Magnetic Resonance
NNC	Non-Negativity constraint Concentration
NNLS	Non-Negative Least Squares
NNS	Non-Negativity constraint Spectra
NWAY P-ALS	Multi-way Penalty-Alternating Least Squares
OPA	Orthogonal Projection Analysis
PAC	Process Analytical Chemistry
P-ALS	Penalty-Alternating Least Squares
PAT	Process Analytical Technology
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principle Component Regression
PLS	Partial Least Squares
QITTFA	Quantitative Iterative Target Transformation Factor Analysis
RE	Relative Error
RMSEP	Root Mean Squared Error of Prediction
RMSPE	Root Mean Square Prediction Error
RR	Ridge Regression
SEC	Size Exclusion Chromatography
SEP	Standard Error of Prediction
SFA	Subwindow Factor Analysis
SFC	Supercritical Fluid Chromatography
SIMPLISMA	Simple to-use interactive Self modelling Mixture Analysis
SMCR	Self Modelling Curve Resolution
SNV	Standard Normal Variate
SVD	Singular Value Decomposition
T-SIMPLISMA	Transposed - Simple to-use interactive Self modelling Mixture Analysis
TTFA	Target Transformation Factor Analysis
UV	Ultraviolet
UV-visible	Ultraviolet-visible

VAM	Vinyl Acetate Monomer
WFA	Window Factor Analysis

Mathematical Notation

Much of the work in this thesis has involved the use of matrices, vectors and scalars. Throughout the text the following mathematical notation is followed. A notations list is given in the glossary.

Scalar quantities (i.e., numbers) are represented by lowercase letters, i.e., *a, b, c, x, y* and *z*. Vectors (i.e., one dimensional arrays of numbers) are symbolised by bold, lowercase letters, i.e., **s**, **t**, **u**, **x**, **y** and **z**. Superscript **T** implies a transposition of a column vector to a row vector. Row vectors are denoted by a superscript **T**, i.e., $\mathbf{s}^{\mathbf{T}}, \mathbf{t}^{\mathbf{T}}, \mathbf{u}^{\mathbf{T}}, \mathbf{x}^{\mathbf{T}}, \mathbf{y}^{\mathbf{T}}$ and $\mathbf{z}^{\mathbf{T}}$. Subscripts are used to characterise matrices, vectors and scalars. Numbers and lowercase letters are used as subscripts. Bold, uppercase letters or enclosures in square brackets [ ] signify matrices. Matrix transformation, whereby rows and columns are interchanged are denoted by a superscript **T**, consistent with the vector notation.

Based on this notation:

b_{ij} is a scalar.

$\mathbf{b}_j = \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \end{bmatrix}$ is the *j*th column vector.

$\mathbf{b}_i^{\mathbf{T}} = \begin{bmatrix} b_{ij} & \ldots\ldots & b_{ij} \end{bmatrix}$ is the *i*th column vector.

$\mathbf{B} = [\mathbf{B}] = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$ is a matrix.

$\mathbf{B}^{\mathrm{T}} = [\mathbf{B}]^{\mathrm{T}} = \begin{bmatrix} b_{11} & b_{21} & b_{31} \\ b_{12} & b_{22} & b_{32} \end{bmatrix}$ is the transposed matrix.

The ˆ, called "hat" above a quantity signifies an estimated (or calculated) quantity.

$\mathbf{D}$ is the matrix containing the measured data, d_{ij}.

$\hat{\mathbf{D}}$ and $\tilde{\mathbf{D}}$ represent different estimations of $\mathbf{D}$.

$\mathbf{D}^{*}$ is an estimation of $\mathbf{D}$ based on a reduced factor space.

Another specialised notation that will be used consistently throughout the text is:

$\mathbf{X}^{+} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the pseudoinverse of $\mathbf{X}$.

$\|r\| = \left(\sum r_i^2\right)^{1/2} = (\mathbf{r}'\mathbf{r})^{1/2}$ is the norm of vector $\mathbf{r}$.

$\|\mathbf{R}\| = \left(\sum\sum r_{ij}^2\right)^{1/2}$ is the norm of matrix $\mathbf{R}$.

Glossary

All symbols used in the text are defined in the notation list given below.

Symbol	Description
n	Number of mixture spectra / samples
m	Number of variables in the spectrum / sample
t	Number of experiments
i	Row / sample index
j	Column / variable index
k	Pure variable index
$\mathbf{D}$	$(n \times m)$ bilinear multi-component and multivariate instrumental measurements
$\mathbf{C}$	Matrix of component concentration profiles
$\mathbf{S}$	Matrix of component spectral profiles
A	Absorbance
c	Analyte concentration
I_0	Incident radiation
I	Transmitted radiation
a	Absorptivity
l	Pathlength through the absorbing medium
ε	Molar absorptivity (mol dm^{-3} cm^{-1}). If c (mol dm^3) and l(cm)
λ	Wavelength (nm)
$\bar{v}$	Wavenumber (cm^{-1})
T	Transmittance
$\mathbf{y}$	Analytes quantitative (concentration) information
$\mathbf{b}$	Regression coefficients (ILS)
a_0	Intercept determined by the regression (CLS)
a_1	Regression coefficient determined by the regression (CLS)
b_0	Intercept determined by the regression (ILS)
b_1	Regression coefficient determined by the regression (ILS)
$\mathbf{E}$	$(n \times m)$ Error (residual) matrix of $\mathbf{D}$
$\mathbf{C}_0$	Initial estimate of the concentration profiles
$\mathbf{S}_0$	Initial estimate of the spectral profiles
$\mathbf{C}_{nit}$	Final concentration solution at maximum number of iterations
$\mathbf{S}_{nit}$	Final spectral solution at maximum number of iterations
nit	Maximum number of iterations
$\mathbf{R}$	$(nc \times nc)$ Rotation Matrix
q	Scalar
$\mathbf{T}_0$	Top sub-matrix (EFA)
$\mathbf{B}_0$	Bottom sub-matrix (EFA)
$\mathbf{T}$	PCA Scores matrix (eigenvalue decomposition)
$\mathbf{P}$	PCA loadings matrix (eigenvalue decomposition)

$\mathbf{f}$	Purity spectrum (SIMPLISMA)
μ	Mean value
σ	Standard deviation
δ	Constant (offset) to correct for noise (SIMPLISMA)
ω	Weight factor (SIMPLISMA)
$\underline{n}$	Number of factors
nc	Number of independent factors / chemical components included in decomposition
$\mathbf{t}_{\mathbf{in}}$	Scores of the target spectrum, **in**
in	Target spectrum
out	Output target spectrum
$\mathbf{s}_{norm}$	Normalised spectrum according to equation 26 or 27
$\mathbf{1}$	A vector of ones
$\mathbf{C}_{\mathbf{sel}}$	Concentration selectivity matrix
$\mathbf{C}_{\mathbf{ALS}}$	The LS estimated concentration matrix
$\mathbf{x}$	Estimated concentration values in the $\mathbf{c}_{\mathbf{ALS}}$ vector
$\underline{\mathbf{o}}$	Unknown concentration values in the $\mathbf{c}_{\mathbf{ALS}}$ vector
$\underline{\mathbf{z}}$	$\underline{z}$ -scores from SNV correction
$\underline{e}$	Prediction errors from a quantitative model
$\underline{d}$	Difference between predictive errors from two competing quantitative models (differences per case)
$\underline{\bar{d}}$	Mean of differences per case
T	Test statistic
T_{obs}	Evaluation data
S	Evaluation set
$\underline{m}$	Number of randomised trials
$\mathbf{D}_r$	$(n \times mt)$ Row-wise augmented measurement matrix
$\mathbf{D}_c$	$(nt \times m)$ Column-wise augmented measurement matrix
$\mathbf{D}_t$	$(t \times nm)$ Tube-wise augmented measurement matrix
$\mathbf{C}_c$	Column-wise augmented concentration matrix
$\mathbf{g}$	Predefined values for constraints (NWAY P-ALS)
$\mathbf{H}$	Constraints matrix (NWAY P-ALS)
φ	The penalty factor weighting (NWAY P-ALS)
IN	Needle matrix
$\mathbf{U}$	Column-orthonormal singular vectors from SVD
$\mathbf{V}$	Row-orthonormal singular vectors from SVD
$\mathbf{\Lambda}$	Diagonal matrix of singular values from SVD
$\mathbf{W}$	Scores matrix
$\mathbf{L}$	Loadings matrix
α and β	Factor Scaling coefficients
$\mathbf{w}_{in}$	Scores of the target spectrum, **in**
$\mathbf{Z}_{\mathbf{s}}$	Needle output spectral matrix (spectra)
$\mathbf{Z}_{\mathbf{c}}$	Needle output spectral matrix (concentration)
$\mathbf{l}_{in}$	Loadings of the target spectrum, **in**

Prologue

Chemometrics is "*a chemical discipline that uses mathematics, statistics and formal logic a) to design or select optimal experimental procedures; b) to provide maximum relevant chemical information by analysing chemical data and c) to obtain knowledge about chemical systems"* [1]. This chemical discipline forms part of the fundamental strategy for exploratory chemical analysis and monitoring and control of a vast array of chemical processes within analytical chemistry.

At the industrial level chemometrics coincides with the Process Analytical Technology (PAT) initiative. PAT is a system of designing, analysing and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes with the goal of ensuring final product quality. Within PAT, analytical is viewed broadly to include chemical, physical, microbiological, mathematical and risk analysis conducted in an integrated manner [2]. Chemometric tools are suited to PAT application because they provide effective and efficient means for acquiring information to facilitate process understanding, develop risk-mitigation strategies, to achieve continuous improvement and share information and knowledge.

Tools which are defined as PAT tools include any technological developments which enable scientific, risk managed, pharmaceutical developments, manufacture and quality assurance and these tools can be categorised as;

1. Multivariate data acquisition and analysis tools
2. Modern process analysers or process analytical chemistry tools
3. Process and endpoint monitoring and control tools

4. Continuous improvement and knowledge management tools.

An appropriate contribution of some, or all of these tools may be applicable to a single unit operation, or to an entire manufacturing process and its quality assurance.

The PAT initiative has received support from the Food and Drug Administration (FDA) Science Board and the Advisory Committee for Pharmaceutical Science.

Process Analytical Chemistry (PAC) was initiated in the late 1970s and is a subsidiary field of PAT and deals specifically with control and optimisation of the performance of a chemical process in terms of capacity, quality, cost, consistency and waste reduction [3, 4]. The standing of PAC as an academic discipline has grown enormously in recent times, mainly catalysed by the creation of the Center for Process Analytical Chemistry (CPAC) at the University of Washington in 1984 [5]. The industry/academic collaboration continues to thrive with around 50 industrial funding partners. Similarly, the creation of the Centre for Process Analytics and Control Technology (CPACT) in 1997 at the Universities of Hull, Strathclyde and Newcastle is also an industry/academic collaboration. Other academic institutions are active in PAC to some extent small specialist groups have been established particularly in The Netherlands, and at other academic institutes in the UK.

The move towards PAC has been fuelled by two developments. Firstly increasing international competitiveness within the chemical industry has led to the widespread adoption of “right-first-time” and “just-in-time” approaches to manufacturing and quality. This has placed the emphasis of building quality into all stages of the process, increased manufacturing flexibility, reduced inventory and improved control of processes. Secondly, during the past decade advances in analytical chemistry and in particular the development of the microcomputer and improved algorithms for data

handling, have enabled almost instantaneous generation of information from large complex measurement matrices [3].

Rationale

In this thesis multivariate analytical PAT tools, known as CFT, have been developed and applied to complex academic and industrial problems. These complex problems are confined to chemical systems which evolve in a systematic, non-random way as a function of time, pH, temperature, etc. These chemical systems are typically deficient in *a priori* information leading to the generation of a chemical or dynamic process model. Such information may include the number of reacting chemical constituents, their evolutionary profiles, their identity, material and energy balances, heat and mass transfer considerations and known reaction kinetics. In some cases, although enough chemical information is available to establish a chemical model, changes in the external conditions may deteriorate the predictive capabilities of the model, i.e., fluctuations in temperature for predictive modelling using isothermal reaction kinetics. In another case, the acquisition of neat reference spectral data for a target analyte may be difficult under the specified reaction conditions, i.e., the vaporisation of a high density organic compound at a temperature and pressure not appropriate for complete vaporisation, which could impede the development of a calibration model for that analyte.

In the past these problems were not tackled and were simply avoided because too many uncontrollable variables influenced the data. In such circumstances advanced chemometric approaches known as calibration free modelling techniques are required because they do not rely on *a priori* information to reveal the underlying chemical model directly from the multivariate measurements. A mathematical decomposition is used to deconvolve the two-way signals from instrumentally unresolved multi-component mixtures into single specie component spectra and compositional profiles from evolutionary systems. This information can be used to answer the most fundamental questions in a chemical problem, i.e., how many factors influence the

observable? What is the nature of these factors in terms of physically significant parameters?

Therefore, CFT can play a significant role within industry and academia because data of great complexity can be investigated, large quantities of data can be analysed using standard computer programs, data can be simplified and interpreted in useful ways and many types of problems can be studied.

The only premise for the decomposition is that the total response is a linear additive signal of each component, i.e., the elements of the measurement matrix, **D** must be a linear sum or combination of the product terms **C** and **S**. Therefore, the measurement must be of the form, see equation 1.

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{T} \qquad \textbf{Equation 1}$$

In common practice, these premises are naturally satisfied by two-way data obtained from multivariate measurements, such as near-infrared (NIR), mid-infrared (MIR), or Fourier transform infrared (FTIR) on mixtures with varying composition.

Therefore, the ultimate aim of CFT is to determine:

1. The number of absorbing components (i.e., reagents, intermediates, products)
2. The evolutionary profile of each component in the mixture (i.e., for purity determination or the prediction of the optimum reaction endpoint)
3. The spectrum of each component (i.e., for identification of each absorbing component).

Aims of Thesis

The feasibility of the application of CFT was shown using several different academic and industrial problems. The first problem was the exploratory and quantitative investigation of the rhodium catalysed asymmetric transfer hydrogenation (CATHy) of a prochiral imine. Here CFT were applied to find an alternative approach to chromatographic analysis of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline and 1-methyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline monitored using in-situ FTIR. The prochiral imine and chiral amine were quantified using high performance liquid chromatography (HPLC). However, an alternative approach was required because the sampling time, including pre-sampling preparation was approximately twice as long as the reaction (~1hr). The foreseeable advantages of applying CFT to this data were: a) the ability to determine the concentration and pure spectral profile of the target analytes from extremely overlapped FTIR spectroscopic data without previous calibration information, which would remove the need for constant sampling and free operator time; b) the reduced time required to obtain the quantitative information; and c) the determination of other reaction constituents not previously identified through chromatographic analysis, which could shed light on the number of reaction constituents, their identities and their evolutionary profiles for contribution towards mechanistic studies.

Secondly, to develop a new Multi-way Penalty Alternating Least Squares (NWAY P-ALS) function to enable optionally hard constraints (no deviation from predefined constraints) or soft constraints (small deviations from predefined constraints) to be applied through the application of a row-wise penalty least squares function. The significant benefits of this method were a) reduced distortion of resolved profiles, b) reduction in the number of active constraints at convergence which reduced the model

lack-of-fit, and c) a reduced impact of noise and non-ideal response on constraints which lead to improved results.

NWAY P-ALS was applied to the multi-batch data acquired from the base catalysed esterification reaction of acetic anhydride. The aim of the study was to resolve the concentration and spectral profiles of 1-butanol with the reaction constituents. The benefits of using the NWAY P-ALS approach included the reduction of the number of active constraints at the solution point, whilst the batch column-wise augmentation allowed strong constraints in the spectral profiles and resolved rank deficiency problems of the measurement matrix. The results were validated by comparing the percent yield of 1-butyl acetate determined by gas chromatography (GC) for each batch. The NWAY P-ALS results were also compared with the multi-way multivariate curve resolution alternating least squares (MCR-ALS) results using hard and soft constraints to determine whether any advantages had been gained through using the weighted least squares function of NWAY P-ALS over the MCR-ALS resolution.

A new calibration free strategy was developed, with the aim of producing starting estimates which approximated the true solution from two-way measurement data to enable MCR-ALS to converge to the correct solution (particularly in cases where no *a priori* information existed for the pure constituents). A tool such as this was required because the quality of the initial estimates and application of constraints had been found to be integral to the success of the resolution procedure. Normally, the MCR-ALS solution can be improved by the addition of *a priori* information in the initial estimates and/ or in the constraints. However, in cases where no *a priori* information exists, an exploratory tool which could produce starting estimates which approximate the actual

solution was required Quantitative Iterative Target Transformation Factor Analysis (QITTFA) was developed as a solution to this problem.

QITTFA was used to resolve the pure spectrum of vinyl acetate monomer (VAM) with the rest of the reaction constituents in the vapour state from the two-way calibration mixture data collected from a British Petroleum (BP) process NIR analyser on the vinyl acetate plant. This was especially important for VAM because neat VAM tended to condense at the specified reaction conditions. The MCR-ALS resolution initialised from the QITTFA starting estimates was compared to the MCR-ALS resolution initialised from a more traditional exploratory tool, Simple to use Interactive Self-modeling Mixture Analysis (SIMPLISMA) in order to demonstrate the robustness and reliability of the QITTFA approach.

I Introduction and Tools

I.1 Preliminaries

I.1.1 Modern Process Analysers

Chemical process measurements consist primarily of traditional engineering variables such as temperature, pressure, flow rate and electrical power because they can be measured easily. These measurements are generally used to calibrate instrumental responses in order to qualitatively model the plant in-line operation for process control [2, 3, 6-9] etc. However, due to developments in fibre optics, fibre optic interfacing and the advent of micro-computers, there has been a tremendous development of instrumentation which can be used on-line or at-line. This additional information provides better estimates of chemical composition throughout the reaction for quality control [2-4, 6, 8-15].

In particular, the use of on-line optical spectroscopic techniques such as NIR [3, 8, 16-18], MIR [3, 8, 19-24], FTIR [3, 8, 25-28], Raman [3, 8, 29-33] and ultraviolet-visible (UV-visible) [3, 8, 34-38] are of interest for reaction monitoring. This is because they not only provide physical process parameters, such as temperature, pressure, flow rate and liquid level, but also molecular parameters relating to component concentrations, molecular structure and composition of process constituents. Also they allow real-time control during a manufacturing process. Both the physical process parameters and molecular parameters can be used in process control, quality control, industrial hygiene, safety or for other value-adding purposes.

Applications of optical spectroscopy for process analysis continues to grow steadily because process spectrometers now have a higher standard of automation, ruggedness, and simplicity in order to withstand harsh manufacturing environments (i.e., humidity,

corrosion, temperature fluctuations). This has resulted in large amounts of spectroscopic data being produced that reflect the optical properties of both process stream and batches in real time. The ease with which *in situ* spectroscopic measurements can be made when coupling the spectrometer with fibre optics makes process spectroscopy a much more effective choice over chromatography for process analysis [39, 40]. Drawbacks associated with on-line chromatographic analysis in batch process analysis include sample processing and instrumental maintenance, that effect the system stability and reproducibility. Quenching the reaction during sample preparation destroys reactive intermediates so the composition of the sample analysed does not accurately reflect the composition of the reaction mixture, which can be a great limitation for exploratory analysis of synthetic processes. Additionally, the limited number of data points that can be determined using the current systems may result in a less than ideal representation of the concentration profile of a specific constituent. Nevertheless, once in operation chromatographic methods can provide good quantitative information with high selectivity and acceptable accuracy, precision and sensitivity.

Spectroscopic measurements have great potential on-line because under the right conditions they can be used as truly non-invasive monitoring techniques. Information relating to the chemical composition and molecular structures is extracted from the spectroscopic data using chemometric dimensional reduction methods. These methods remove correlated or redundant information (e.g., heteroscedastic noise, caused by drifting baselines, fluctuations in the surrounding environment on the reaction vessel) from the measurement data and retains the essential information. Chemometric dimensional reduction methods are particularly well suited for improving the effectiveness of process spectroscopic analytical techniques through two main

functions; (1) extracting a wealth of useful information from convoluted measurements, and (2) facilitating the automation of *in-situ* analytical techniques.

Optical absorbance spectroscopic datasets, such as ultraviolet (UV), visible and infrared (IR) are suitable for dimensional reduction because the basis of an absorption method for quantitative determination of absorbing species is expressed as a linear sum or combinations of the product terms. This is given in the relationship between the absorbance, A, and the analyte concentration, c, known as Beer's Law (Bougier-Beer-Lambert-Law), given in equation 2.

$$A = \log_{10} \frac{I_0}{I} = acl \qquad \textbf{Equation 2}$$

It is related in a logarithmic way to the ratio of the intensity of the incident radiation, I_0 and the intensity of the radiation transmitted through the sample solution I, a is a proportionality constant called the absorptivity and l is the pathlength through the absorbing medium. When concentration is expressed in moles per litre and l is in centimetres, the proportionality constant is called the molar absorptivity and is given the symbol ε. Thus the absorbance, A, is related to the molar concentration, c, times the pathlength, l, in centimetres times the molar absorbptivity, ε, see equation 3.

$$A = \varepsilon cl \qquad \textbf{Equation 3}$$

where ε has the units of mol dm^{-3} cm^{-1}. The desired parameter in spectroscopy is absorbance, but it cannot be directly measured. Thus, a UV-visible spectrophotometer compares the intensity of the transmitted radiation with that of the incident UV-visible radiation. An IR spectrometer records IR spectra as a plot of the wavelengths, λ, or wavenumbers, $\bar{\nu}$, of absorbed radiation against the intensity of absorption in terms of

transmittance, *T,* or absorbance, *A*, see equations 4-5. Presently, the wavenumber unit is used because it is directly proportional to energy.

$$T = \frac{I_0}{I} \qquad \textbf{Equation 4}$$

$$A = \log_{10} \frac{1}{T} \qquad \textbf{Equation 5}$$

For a multi-component spectroscopic system, the total response at a specific wavenumber (at a constant light pathlength) is the linear additive signal of each chemical constituent, provided no interactions occur among the various species, see equation 6.

$$A_\lambda = c_1 \varepsilon_{\lambda 1} + c_2 \varepsilon_{\lambda 2} + \ldots\ldots + c_{nc} \varepsilon_{\lambda, nc} = \sum_{k=1}^{nc} c_k \varepsilon_{\lambda k} \qquad \textbf{Equation 6}$$

where A_λ is the absorbance at wavelength, λ, ε_λ is the molar absorbtivity, mol dm^{-3} cm^{-1}, for the *kth* component at wavelength lambda and *c* is the concentration, mol dm^{-3}, for the *kth* component.

It must be stressed that Beers Law is only strictly applicable to dilute solutions, where the interactions between the absorbing particles are insignificant. Deviations can also be seen as a consequence of associations, dissociations or reaction of the absorbing species with the solvent. Instrumental effects come from the use of polychromatic radiation (generally not significant) and stray light due to instrumental imperfections. The effect of these is to reduce the measured absorbance [8, 41].

I.1.2 Multivariate Data Acquisition and Analysis Tools

Factor analysis based methods, such as principal component analysis (PCA) multivariate calibration analysis [11, 42-48] and CFT (otherwise known as self

modelling curve resolution (SMCR) [11, 48-74] are linear models which can be applied to compress and extract relevant information from the measurements. The ultimate goal of PCA is to decompose a bilinear measurement matrix into its component parts (under the two constraints of orthonormality and maximum variance) to reveal patterns and trends within the data, the goal of multivariate calibration is the establishment of a calibration model from multivariate measurements allowing the quantitative determination of the analyte in the presence of unknown interferents or in a complex chemical matrix, even if the analyte signal selectivity is poor (i.e., prediction of the expensive measurement). CFT, on-the-other hand, resolve multi-component and multivariate measurement matrices into pure factors, such as spectral profiles, time profiles, and pH profiles, for individual species with no *a priori* knowledge of the system. This allows the qualitative and semi-quantitative determination of the reaction constituents, including unknown interferents

Traditional exploratory and quantitative tools, such as chromatography and multivariate calibration are used extensively within the process analytical environment for reaction monitoring and process control because once in operation they can provide good quantitative information with high selectivity and acceptable accuracy, precision and sensitivity. However, if there are reactions or processes for which it is not possible to prepare mixtures of known composition, due for instance, the absence of isolated reference material, stability issues and where the preparation of such samples are time consuming and expensive, CFT are much more favourable because no assumptions are made concerning the underlying chemical or physical model and these techniques are relatively simple to use. This has been shown by recent application within many diverse fields such as environmental studies [75-80], kinetic reactions [71, 81-86],

quantitative analysis [73, 79, 87-91], peak purity assessments [49, 61, 92-95], characterisation of batch reactions and on-line reaction monitoring [21, 96-103].

I.2 Calibration Free Analysis

I.2.1 Principles of Calibration Free Methodologies

The aim of CFT is to determine:

1. The number of absorbing components (i.e., reagents, intermediates, products)
2. The evolutionary profile, **C** of each component in the mixture
3. The spectrum, **S** of each component

The main steps used to generate and alternatively estimate the pure spectra **S** and concentration profiles **C** are outlined in box 1, and consists of an optional preprocessing of the raw data to remove any extraneous factors caused by instrumental artefacts. A selection of some initial starting point followed by iterative refinement with constrained alternating least squares steps until convergence to a stationary solution and validation of the model using an external method.

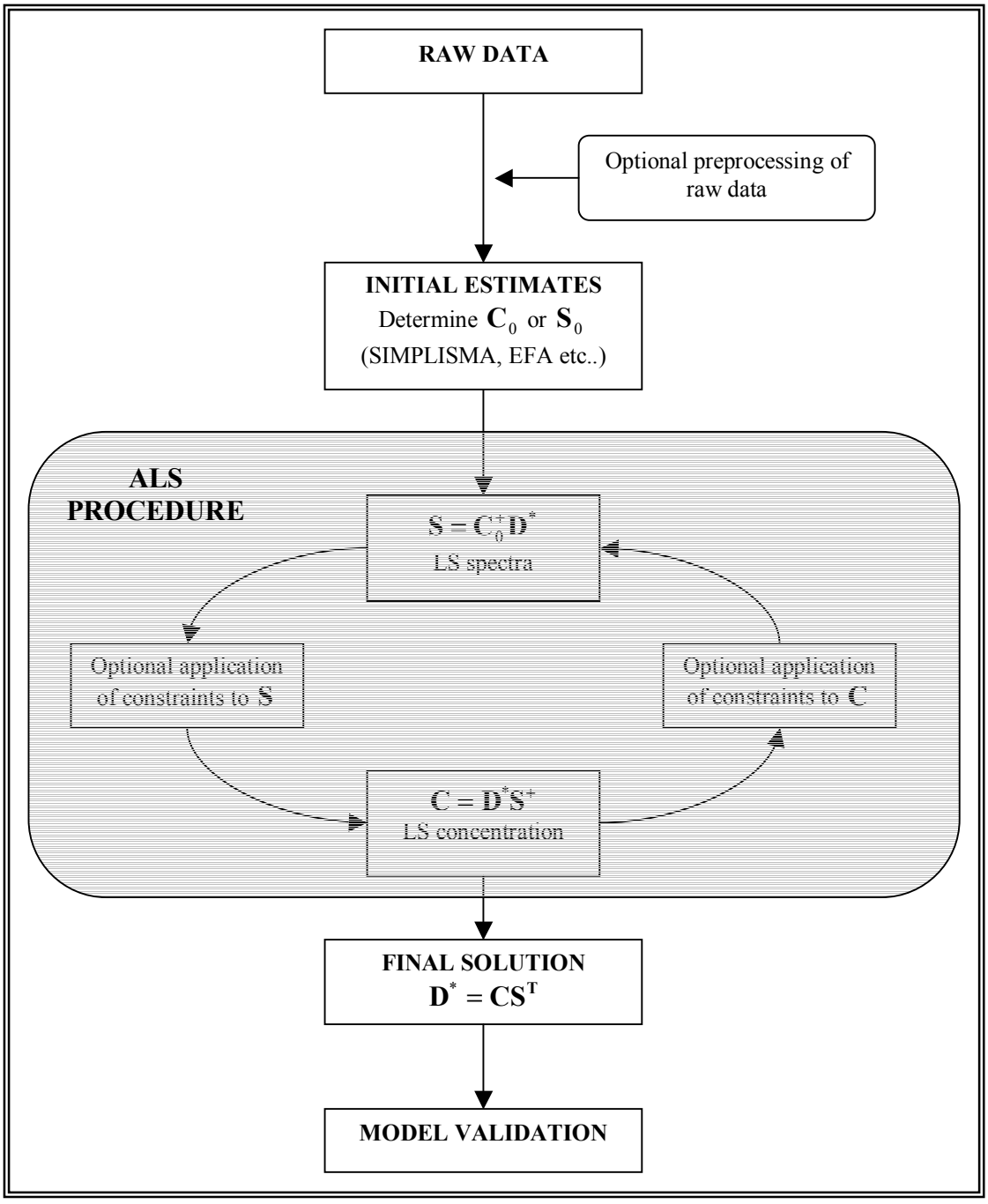
I.2.1.1 Alternating Least Squares

1. An initial solution or "estimate" is selected by any number of popular exploratory tools (see section I.2.3.2). If good information is available about the concentration profiles, an initial estimate of the concentration profiles, $\mathbf{C}_0$, may be used or alternatively, spectra, $\mathbf{S}_0$, or variables from the measurement matrices themselves which are hypothesised to be good approximations of pure component spectra or pure variables.

2. The initial starting point $\mathbf{C}_0$ or $\mathbf{S}_0$, seldom obeys the constraints imposed, and *constrained ALS* steps are used to fit the initial unconstrained solution producing better "constrained" estimates.

a. Given some initial estimate of **C**, find **S** such that **S** minimises an error function, i.e., $\left\|\mathbf{D}-\mathbf{C}\mathbf{S}^{\mathbf{T}}\right\|$ subject to constraints on **S** such as **S** > 0, etc.

b. Given some least squares estimate of **S**, find **C** such that **C** minimises an error function, i.e., $\left\|\mathbf{D}-\mathbf{C}\mathbf{S}^{\mathbf{T}}\right\|$ subject to constraints on **C** such as **C** > 0, etc.



Box 1. Flow chart of SMCR methodology

Thus, an estimate of the unknown species spectra is given by least squares, which is simply, $\mathbf{S} = \mathbf{C}_0^{+}\mathbf{D}^{*}$ and the new estimation of the concentration profile is given by $\mathbf{C} = \mathbf{D}^{*}\mathbf{S}^{+}$. Where $\mathbf{D}^{*}$ is estimated based on reduced factor space.

The purposed of constraints during fitting is to ensure that the original starting solution converges smoothly and monotonically to the desired result. Several published papers describe in detail the mechanism for solving these types of constrained problems in a least square manner that ensures monotonic convergence, the most widely used being the non-negative least squares (NNLS) method of Lawson and Hansen [104], adapted by Bro and de Jong [105]. Despite this, the direct substitution approach; in which the ALS estimate is substituted with the constrained ALS estimate, does not result in a least square solution, but is commonly used. One of the reasons for its popularity among SMCR practitioners is that it is fast and convenient. Further research completed by Van Benthen *et al.* [106] revealed that for a particular set of simulations with various levels of noise and magnitudes of offsets, the solutions using equality constrained least square procedure versus the equality constrained substitution approach were not substantially different. Although the most important factor was not being able to predict when the results would be discrepant.

3. Each application of the two least squares steps (2a and 2b) produce a better estimate of the constrained concentration profiles and the constrained component spectra; thus a simple iterative refinement process is used whereby these two alternating steps are repeated until no further improvement in the estimates of $\mathbf{C}_{nit}$ or $\mathbf{S}_{nit}$, is observed or the maximum number of iterations, *nit*, is reached. This is the basis of

the multivariate curve resolution-alternating least squares (MCR-ALS) routine used by R. Tauler *et al.* [107].

4. If reference information is available, the resolved profiles can be validated against independent external reference measurements.

Therefore, the basic principles of calibration free analysis or SMCR is to seek a bilinear model that gives the best fit, in the sense of least squares, to the two way data **D**. In other words, SMCR estimates pure variables **C** and **S** that minimise the following error criteria **E**, see equation 7.

$$\mathbf{E} = \left\| \mathbf{D} - \mathbf{C}\mathbf{S}^{\mathbf{T}} \right\| \qquad \textbf{Equation 7}$$

The most commonly used error criterion is the squared difference between **D** and $\mathbf{CS}^{\mathbf{T}}$, though some techniques use weighted error [108] or normalised squared error [109].

The minimisation of equation 7 over **C** and **S** cannot guarantee a unique solution to the pure variables, i.e., there are many solutions of **C** and **S** which reproduce the data with the same fit quality. Put another way, the correct reproduction of the original data matrix can be achieved by using response variables differing in shape (rotational ambiguity) or in magnitude (intensity ambiguity) from the true solution, without changing the residual associated with the model [59, 68, 110, 111].

The basic equation associated with resolution methods, (see equation 1) can be transformed as shown in equations 8-10;

$$\mathbf{D} = \mathbf{C}\left(\mathbf{R}\mathbf{R}^{-1}\right)\mathbf{S}^{\mathbf{T}} \quad \textbf{Equation 8}$$

$$\mathbf{D} = \left(\mathbf{C}\mathbf{R}\right)\left(\mathbf{R}^{-1}\mathbf{S}^{\mathbf{T}}\right) \quad \textbf{Equation 9}$$

$$\mathbf{D} = \hat{\mathbf{C}}\hat{\mathbf{S}}^{\mathbf{T}} \quad \textbf{Equation 10}$$

Where $\hat{\mathbf{C}} = \left(\mathbf{C}\mathbf{R}\right)$ and $\hat{\mathbf{S}} = \left(\mathbf{R}^{-1}\mathbf{S}^{\mathbf{T}}\right)$ describes the **D** matrix as correctly as the true **C** and $\mathbf{S}^{\mathbf{T}}$ matrices do, though $\hat{\mathbf{C}}$ and $\hat{\mathbf{S}}^{\mathbf{T}}$ are not the true solutions. The rotational ambiguity problem indicates that a resolution method can potentially provide as many solutions as rotation matrices, **R** can exist, i.e., infinite unless **C** and **S** are forced to obey certain conditions. In a hypothetical case with no rotational ambiguity, the basic resolution model could still be rewritten as shown in equations 11-12:

$$\mathbf{D} = \sum_{i=1}^{n}\left(\frac{1}{q_i}\mathbf{c}_i\right)\left(q_i\mathbf{s}_i^{\mathbf{T}}\right) \quad \textbf{Equation 11}$$

$$\mathbf{D} = \hat{\mathbf{C}}\hat{\mathbf{S}}^{\mathbf{T}} \quad \textbf{Equation 12}$$

Where q_i are scalars. The concentration profiles of the new $\hat{\mathbf{C}}$ matrix would have the same shape as the real ones, but being q_i times smaller, whereas the spectra of the new $\hat{\mathbf{S}}^{\mathbf{T}}$ matrix would be shaped like the **S** spectra though q_i times more intense. This is known as the intensity ambiguity.

A thorough discussion of constraints that are required to reduce or eliminate either the rotational ambiguity or intensity ambiguity is discussed below:

I.2.1.2 Constraints

There are various constraints which can be imposed during the ALS procedure in order to reduce or eliminate the rotational or intensity ambiguity. Such constraints include the i) non-negativity, ii) normalisation, iii) closure, iv) unimodality and v) selectivity, see steps 2-3 of the SMCR methodology, section I.2.1. If the constraints selected are

characteristic (or fulfilled) by the measurements, the constraints can be perceived as the driving force of the iterative process to the correct solution, with the aim of ensuring fast monotonical convergence [68, 72].

I.2.1.2.1 Non-negativity

This is probably the most commonly used constraint in curve resolution since the initial work of Lawton and Sylvestre [50]. This constraint assumes that the concentration estimates can only be positive or zero $(\mathbf{C} \geq 0)$ and in many spectroscopies, spectral values can only be positive or zero $(\mathbf{S} \geq 0)$. The application of this constraint will reduce the rotational or intensity ambiguity, but it will not constrain the problem sufficiently to give unique solutions [50, 58, 111].

I.2.1.2.2 Normalisation

It is possible to limit the size of the concentration or the spectral profiles using appropriate normalisation and closure constraints. In Tauler's MCR-ALS procedure, the pure spectral profiles $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{nc}$, can be optionally normalised to either length or height [107]. By the constant length normalisation procedure each pure spectrum is normalised to a constant Euclidean norm, by dividing the square root of the sum of squares with the absorbance values, see equation 13. In the second normalisation procedure the spectral profiles are normalised by dividing the maximum intensity with the absorbance values, see equation 14. This ensures that the signal height or maximum intensity of the spectral profile is equal to a constant value.

$$\mathbf{s}_{norm,k} = \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \qquad k = 1,2,...,nc \qquad \textbf{Equation 13}$$

$$\mathbf{s}_{norm,k} = \frac{\mathbf{s}_k}{\max(\mathbf{s}_k)} \qquad k = 1,2,...,nc \qquad \textbf{Equation 14}$$

Application of this constraint will reduce the intensity ambiguity, but it will not lead to the unique solution [58].

I.2.1.2.3 Closure

The closure constraint is applied to the rows of the pure concentration profiles. By this constraint the sum of the elements of each row of the concentration matrix is equal to a known constant. A chemical example of this may be a reaction-based system, in which a mass balance equation is obeyed by the concentration profiles of the species present in the system.

For example, a matrix **C** with closure is;

$$\mathbf{C} = \begin{bmatrix} 1.0 & 0.0 \\ 0.7 & 0.3 \\ 0.5 & 0.5 \\ 0.6 & 0.4 \end{bmatrix}$$

Where in each row of **C** the numbers add up to 1. This can be written in matrix notation as;

$$\mathbf{C}\mathbf{1}_2 = \begin{bmatrix} 1.0 & 0.0 \\ 0.7 & 0.3 \\ 0.5 & 0.5 \\ 0.6 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{1}_4$$

Where the symbol $\mathbf{1}_{nc}$ is used to indicate a $(nc \times 1)$ vector of ones. Note that despite the closure, the number of independent components in **C** is two. If the matrix **C** is column-

mean centred, the rank is reduced by one [112, 113]. The closure constraint can be formulised as given in equation 15;

$$\mathbf{C}\mathbf{1}_{nc} = \mathbf{1}_{n} \qquad \textbf{Equation 15}$$

Where $\mathbf{1}_{n}$ is the number of rows (samples) in the measurement matrix. Imposing the closure constraint does not solve the rotational ambiguity, however, it has been shown that the intensity ambiguity is removed if closure constraints are imposed, provided that **C** has full column rank [114].

I.2.1.2.4 Unimodality

This constraint is typically applied to either elution profiles or concentration profiles in reaction-based systems. This constraint assumes that only one peak maxima exists in each concentration profile [115]. This constraint can be applied to ensure *(a) vertical* unimodality (the classic unimodality), *(b) horizontal* unimodality or *(c) average* unimodality [107, 116].

The common steps in each unimodal constraint are:

1. Determination of the maximum intensity in the concentration profile
2. Suppression of the left local maxima
3. Suppression of the right local maxima

In *(a)* the *vertical* (classical) unimodal constraint, the secondary maxima are eliminated by setting the non-unimodal elements equal to zero, in *(b)* the *horizontal* unimodal constraint, the secondary maxima are reduced by setting the non-unimodal element equal to the nearest element keeping the unimodal condition. In *(c)* the *average* unimodal constraint, the secondary maxima are reduced by setting the non-unimodal elements to the average of the nearest element keeping the unimodality condition and

the non-unimodal element. The values adjacent to the concentration maxima can be weighted to allow small departures from unimodality, to compensate for noisy peaks normally associated with minor compounds. The application of this constraint will reduce the rotational or intensity ambiguity, but it will not constrain the problem sufficiently to give unique solutions [114].

I.2.1.2.5 Equality

If a concentration (e.g., zero concentration) or spectral intensity (e.g., baseline offset, slope or pure component spectra) is known, this allows equality constraints to be introduced. Equality constraints in the best case can alleviate rotational ambiguity and in the worst case reduce rotational ambiguity. The known concentration or spectral estimates are set to be invariant along the iterative process. Following this concept the knowledge of a profile does not need to be complete to be used. Equality constraints, however, are sometimes too strong during the optimisation and in many circumstances it is not possible to know whether the values are precise. Under these circumstances an inequality constraint bounded by an upper threshold can be defined, which assumes that a particular species does not exist at appreciable concentrations or that it does not contribute to the signal in an appreciable way [58, 68].

I.2.1.2.6 Selectivity and local rank

Selectivity [111] and local rank constraints [117] refer to the fact that in certain windows or regions in the data matrix **D**, a particular species is known to exist while others are known not to exist, i.e., spectral and concentration windows where only one component is present. In all circumstances selectivity and local rank constraints have a tremendous effect of narrowing considerably the band of feasible solutions, eventually collapsing them into unique solutions. For instance, in a reaction based system it is a

common situation that some of the species are not present at the beginning or at the end of the reaction process. Similarly, spectroscopic data may not absorb a component in a particular spectral range. This information can be used in the selectivity constraint to define the spectral regions or concentration windows where only one species exists. Local rank constraints identify regions where species are non-existent. Evolutionary factor analysis methods, such as EFA can be used to define zero component regions. Difficulties may arise because of noise or low contribution from analytes in the signal and therefore, defining the beginning and the end of the concentration for the particular analyte may be difficult.

I.2.2 SMCR Research Hypothesis

The use of constraints during the ALS procedure predetermines a SMCR research hypothesis, i.e. *There exists an unconstrained bilinear model with unimodal, non-negative pure component concentration profiles and pure component non-negative spectral profiles that fits the data matrix of measurements obtained from the evolving system*. To determine whether the hypothesis is true or false the hypothesis is normally tested by iteratively fitting a constrained model until convergence is achieved. If the proposed research hypothesis is correct, the resultant SMCR solution would contain *no active constraints*. However, *active constraint* are often present in the solution because of noise, non-ideal chemical response and non-ideal spectroscopic response. Active constraints may improve the model interpretability at the expense of an increase in the model lack-of-fit. In some cases, the lack-of-fit can be so severe that it call into question the validity of the original research hypothesis.

In such cases, a soft constrained solution may be sought. Soft constraints refers to a situation where a natural constraints such as non-negativity are not strictly enforced, i.e.

small deviation from non-negativity allowed. Conversely, hard constraints refers to a situation were the constraints are strictly enforced, i.e. no deviation from condition. For many datasets, it has been observed that the use of hard constraints leads to final solutions with many active constraints [72]. These types of solutions may exhibit distorted composition and spectral profiles. Compared to hard constraints, SMCR models with soft constraints often have fewer active constraints which minimises the impact of noise and non-ideal response and hence lowers model lack-of-fit, and are more likely to fit the original research hypothesis.

The soft constrained solution can be implemented using a least squares penalty alternating least squares algorithm, called penalty Alternating Least Squares (P-ALS).

I.2.2.1 Penalty Alternating Least Squares

The P-ALS methods uses the same four steps (outlined in section I.2.1.1); however, the two alternating least squares problems are solved in a row-wise fashion with least squares penalty functions to implement constraints. Bro showed that finding the optimum least squares estimate of individual rows in **S** (row-wise estimation of **S**) or column-wise estimation of $\mathbf{S}^{\mathbf{T}}$) can be performed independently for rows in the same mode [72]. Thus steps 2a and 2b outlined in section I.2.1.1. become:

a) Given some initial or intermediate estimate of **C** for every *j*, find $\mathbf{s}_j$ such that $\mathbf{s}_j$ minimises $\left\|\mathbf{d}_j - \mathbf{C}\mathbf{s}_j^{\mathbf{T}}\right\|$ subject to constraints on $\mathbf{s}_j$, such as $\mathbf{s}_j = \mathbf{g}_j$, where $\mathbf{g}_j$ is a vector of constraints, defined later.

b) Given some least squares estimate of $\mathbf{S}^{\mathbf{T}}$ for every *i*, find $\mathbf{c}_i$ such that $\mathbf{c}_i$ minimises $\left\|\mathbf{d}_i - \mathbf{S}\mathbf{c}_i^{\mathbf{T}}\right\|$ subject to constraints on $\mathbf{c}_i$, such as $\mathbf{c}_i = \mathbf{g}_i$

The global solution to **S** is obtained by solving the row-wise subproblem j times. The procedure is illustrated schematically in figure 1a. By transposing the problem, the same algorithm can be used to solve the row-wise estimation of $\mathbf{C}^{\mathbf{T}}$ as shown in figure 1b.

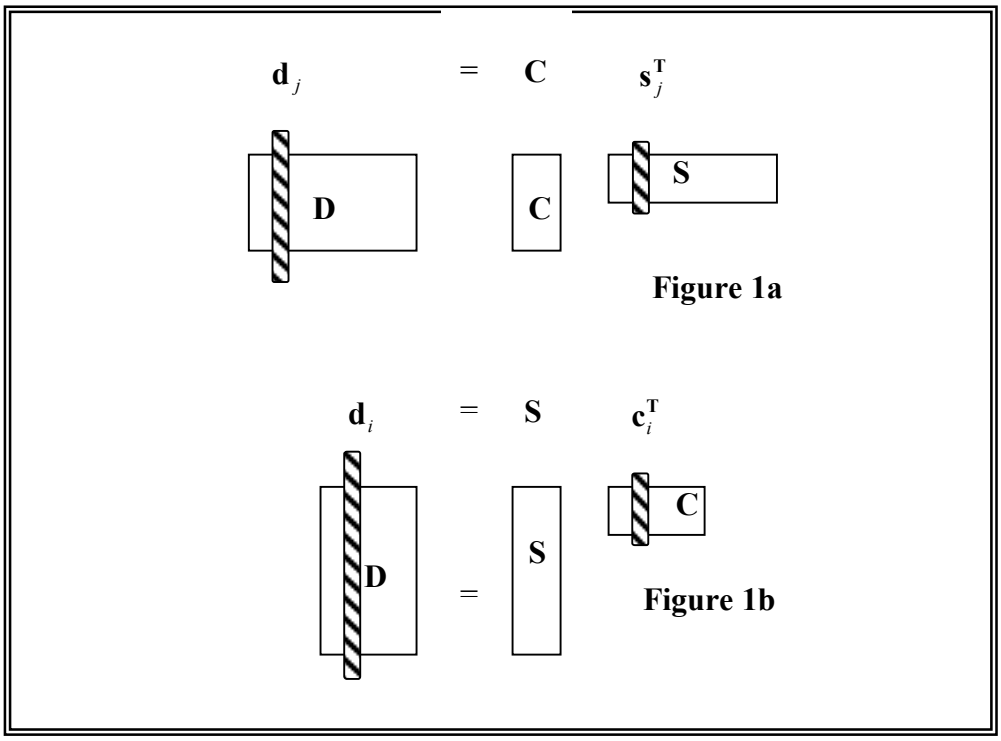


Figure 1a. Schematic illustration of the row-wise fitting algorithm for finding rows of S. Figure 1b. Schematic illustration of the row-wise fitting algorithm for finding rows of C.

Details of how the penalty function can be used to construct equality constraints[106], non-negativity constraints, closure constraints[106] and unimodality constraints using the row-wise method of Bro[72] is given below.

I.2.2.1.1 Equality Constraints

Approximate equality constraints can be implemented for least squares problems **y=Xb** via penalty functions [72, 106]. Notice that both row-wise P-ALS subproblems (a) and (b) described can be generalised to this form. Suppose the equality constraints $b_i=g_i$ are desired, where the elements g_i are desired, where the elements g_i represent the desired

goals for selected coefficients b_i of **b**. By augmenting **y** with **g** and **X** with a matrix, **H**, of appropriately positioned ones and zeros, one or more coefficients of the normal least squares solution vector, *b*, can be forced to conform to the desired values, g_i. The model is written as shown in equation 16;

$$\begin{bmatrix} \mathbf{y}_j \\ \varphi\mathbf{g} \end{bmatrix} \cong \begin{bmatrix} \mathbf{X} \\ \varphi\mathbf{H} \end{bmatrix} \mathbf{b}_j^{\mathbf{T}} \qquad \textbf{Equation 16}$$

The penalty function weighting factor ,φ, can be adjusted to small values when soft constraints are desired, or it can be set very large to give hard constraints. To properly weight problems of different sizes and different measurement scales, the penalty function weighting factor,φ, can be adjusted relative to the norm of **X**; for example, $0.01\times\text{norm}(\mathbf{X})$ for soft constraints or $10\times\text{norm}(\mathbf{X})$ for hard constraints.

If one or more pure component spectra are known *a priori*, this information can be included as equality constraints on $\mathbf{S}^{\mathbf{T}}$ in step (a) of the P-ALS method. If reference data is available for concentrations, say for example, aliquots of a reaction mixture are analysed at selected time intervals by HPLC, these sparsely sampled reference data can be used as equality constraints on **C** in step (b) of the P-ALS method. To implement non-negativity, unimodality, closure and equality constraints during P-ALS the following procedures are followed:

I.2.2.1.2 Non-negativity Constraints

A simple modification of P-ALS steps 2a and 2b can be used to impose approximate non-negativity constraints using the least squares penalty functions and equality constraints. First the standard unconstrained least squares problem is solved. The resulting coefficients, b_i, are inspected. For any coefficient, b_i, that are negative an

equality constraint, $b_i = 0$ is set. With a large penalty value, the resulting solution converges to same result that would be obtained using hard constraints, such as those obtained from algorithms NNLS or FNNLS. An example illustrating the implementation of non-negativity constraints is shown below. First the unconstrained least squares solution to a sample problem is given in equation 17.

$$\mathbf{y} = \begin{bmatrix} 0.9218 \\ 0.7382 \\ 0.1763 \\ 0.4057 \end{bmatrix} = \begin{bmatrix} 0.4451 & 0.8462 & 0.8381 \\ 0.9318 & 0.5252 & 0.0196 \\ 0.4660 & 0.2026 & 0.6813 \\ 0.4186 & 0.6721 & 0.3795 \end{bmatrix} \hat{\mathbf{b}} \begin{bmatrix} 0.2660 \\ 0.8163 \\ -0.0595 \end{bmatrix} \qquad \textbf{Equation 17}$$

Noting that coefficient b_3 is negative, the non-negativity constrained least squares problem is solved as shown below with equality constraints and penalty weighting function φ=10, giving the approximate non-negativity constrained solution $\mathbf{b}_{\text{P-ALS}(\varphi:10)} = [0.2749 \quad 0.7645 \quad -0.0003]$, see equation 18. If hard constraints are imposed by use of algorithm NNLS, the solution $\mathbf{b}_{\text{NNLS}} = [0.2749 \quad 0.7643 \quad 0]$is obtained. This solution compares favourably to the solution obtained with penalty weight φ=100, $\mathbf{b}_{\text{P-ALS}(\varphi:100)} = [0.2749 \quad 0.7643 \quad -3\times10^{-6}]$. If soft constraints are desired, a smaller penalty weight, such as φ=1.0,can be used producing the solution $\mathbf{b}_{\text{P-ALS}(\varphi:1)} = [0.2721 \quad 0.7809 \quad -0.0189]$.

$$\mathbf{y} = \begin{bmatrix} 0.9281 \\ 0.7382 \\ 0.1763 \\ 0.4057 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 0.4451 & 0.8462 & 0.8381 \\ 0.9318 & 0.5252 & 0.0196 \\ 0.4660 & 0.2026 & 0.6813 \\ 0.4186 & 0.6721 & 0.3795 \\ 0.0 & 0.0 & 10 \end{bmatrix} \hat{\mathbf{b}} \begin{bmatrix} 0.2749 \\ 0.7645 \\ -0.0003 \end{bmatrix} \qquad \textbf{Equation 18}$$

I.2.2.1.3 Closure Constraints

In some cases such as batch reactions studies or chemical equilibrium studies, the principles of mass balance can be invoked such that the sum of all or selected species concentration profiles should equal a constant. Closure constraints can be implemented with equality constraints in the manner shown by Van Benthem, Keenan and Haaland [106] using $\mathbf{H} = [1 \quad 1, \quad ..., \quad 1]$ and $g = [1]$. Augmenting the previous example with closure constraints with a penalty weighting function φ=10 gives the following results with $\text{sum}(\mathbf{b}_{\text{P-ALS}}) = 1.0005$, given in equation 19. This example also illustrates that many different combination of constraints can be solved in one step.

$$\mathbf{y} = \begin{bmatrix} 0.9218 \\ 0.7382 \\ 0.1762 \\ 0.4057 \\ 0.0 \\ 10 \end{bmatrix} \mathbf{X} = \begin{bmatrix} 0.4451 & 0.8462 & 0.8381 \\ 0.9318 & 0.5252 & 0.0196 \\ 0.4660 & 0.2026 & 0.6831 \\ 0.4186 & 0.6721 & 0.3795 \\ 0.0 & 0.0 & 10 \\ 10 & 10 & 10 \end{bmatrix} \hat{\mathbf{b}} \begin{bmatrix} 0.2450 \\ 0.7645 \\ -0.0004 \end{bmatrix} \qquad \textbf{Equation 19}$$

For some chemical systems, a weighted sum of several species would be expected to give a constant, for example, as in the reaction $2\text{A} \rightarrow \text{B} \rightarrow \text{C}$. Here the appropriate constraint would be $\mathbf{H} = [1/2 \quad 1 \quad 1]$ and $g = [1]$.

I.2.2.1.4 Unimodality Constraints

Further modification of P-ALS step 2b can be used to impose approximate unimodality constraints using least squares penalty function and equality constraints. In a fashion similar to the non-negativity constraint implementation described above, the standard unconstrained least squares problem is solved first. The resulting concentration profiles are inspected as a function of time to find the global peak maximums, one for each

profile. Searching in both the forward and reverse directions from the global peak maximums of each constituent, unimodality constraints must be added at time $i+1$, $g_{i+1,j} = c_{i,j}$, if secondary local maximums are encountered, for example, $c_{i,j} < c_{i+1,j}$ or $c_{i,j} > c_{i-1,j}$. An example illustrating the implementation of unimodality constraints is shown below. First, the unconstrained least squares solution to a sample problem is given in equation 20.

$$\mathbf{y} = \begin{bmatrix} 0.5534 \\ 0.2920 \\ 0.8580 \\ 0.3358 \end{bmatrix} = \begin{bmatrix} 0.1970 & 0.4796 & 0.2625 \\ 0.9913 & 0.4960 & 0.1863 \\ 0.7120 & 0.2875 & 0.9171 \\ 0.8714 & 0.0609 & 0.1233 \end{bmatrix} \hat{\mathbf{b}} \begin{bmatrix} 0.0766 \\ 0.4085 \\ 0.7832 \end{bmatrix} \qquad \textbf{Equation 20}$$

Suppose it is expected that a peak maximum occurs at b_2. Noting the coefficients b_3 in the unconstrained solution is too large, the unimodal constrained least squares problem is solved, as shown below, with the equality constraint $b_3 = 0.4085$ and the penalty weighting function φ=10, giving the approximate unimodal constrained solution, $\mathbf{b}_{\text{P-ALS}(\varphi:10)} = [0.1720 \quad 0.5908 \quad 0.4102]$. After several iterations of the ALS algorithm, the result converges to $\mathbf{b}_{\text{P-ALS}(\varphi:10)} = [0.1408 \quad 0.5311 \quad 0.5323]$.

I.2.3 Exploratory Analysis Tools

I.2.3.1 Number of components

When applying SMCR to "black" systems, i.e., when the concentration of the constituents and the spectra of the constituents are unknown, and perhaps even the number of the constituents is unknown, the first step is to estimate the chemical rank. The correct estimation of the number of chemical components in the system is crucial for the correct resolution. If there were no measurement noises and other pitfalls from measurements in the data, the mathematical rank (the number of independent components and or factors in two-dimensional data) and chemical rank (the number of chemical components in unknown mixtures) should be the same. The determination of the mathematical rank of a noise-free matrix is a trivial task. One can reduce the matrix to row-echelon form by Gaussian elimination and count the number of non zero rows. However, determining the chemical rank of an experimental data matrix is a difficult task because of the following factors: (i) the presence of measurement noise and their non-assumed distributions, (ii) heteroscedasticity of the noise; and (iii) collinearity in the measurement data, i.e., if the chemical species do not vary independently within a mixture, the rank of the measurement matrix will be different to the number of chemical species. For example, in an overall second order kinetic reaction, **A** + **B**→ **C** (first order in constituents **A** and **B**), both reagents would be consumed at the same rate, such that the concentration profiles of **A** and **B** are mathematically indistinguishable. In this instance the overall rank of the system is less than the number of chemical constituents. This system is said to be rank deficient, i.e., the number of independent components is less than the number of chemical species. Rank deficiency may also be caused by data pretreatment (e.g., mean centering, autoscaling, differentiation). To circumvent this

problem the measurement matrix is augmented either by using multiple process runs or by adding known amounts of absorbing species already present in the mixture [112].

Exploratory tools such as FA or principal component analysis (PCA) can be used to estimate the chemical rank of two-way data because they can decompose the matrix into several independent and orthogonal components [1, 11, 48]. The number of independent and orthogonal principal components (PC) will correspond to the number of independent chemical species in the mixture. The mathematical formula of PCA is expressed elsewhere [1, 48]. Methods used for estimating the number of independent components include PCA [1, 48], the Scree test [48], Malinowski's F-Test [11], and a number of calibration free exploratory tools, such as SIMPLISMA [64, 118-121], Evolving Factor Analysis (EFA) [62, 122, 123] etc.

I.2.3.2 Exploratory Methods

Calibration free methodologies can be divided into two general categories; those which produce boundaries of solutions and those which produce single solutions for each component. In their seminal paper, Lawton and Sylvestre [50] introduced SMCR for two-component mixtures. The algorithm based on PCA generated boundaries of valid solution meet three criteria;

1. All pure component *spectral* values are non-negative
2. All pure component *concentration* values are non-negative
3. All pure component spectra must lie within the subspace spanned by two eigenvectors in the spectra space.

A constraint of unit-area spectra was also applied to resolve ambiguities of scale. It was further indicated in the paper that if there were channels in which only one component

gave a response, single solutions rather than boundaries could be obtained. The original SMCR method proposed by Lawton and Sylvestre was restricted to two components. Further extensions of the method to three components was completed by Ohta in 1973 [124], Sasaki *et al.* in 1983 [125] and Borgen and Kowalski in 1985 [51]. However because the algorithms were quite complex, difficult to program and computationally intensive, SMCR methods that produced single solution were favoured and as a result developed in parallel with boundary search methods. These single solution methods can be divided into unique methods and rational resolution methods.

Unique methods can be identified as those methods which try to pick up a unique resolution in which the factors for single species are uniquely defined according to the mathematical principles involved. A characteristic feature of unique resolution techniques is to exploit information in local feature regions such as selective regions or zero concentration regions to reduce or eliminate rotational and intensity ambiguities in the solution. The drawbacks of these methods is the accurate determination of the feature regions through exploratory local rank analysis, often the solution obtained tend to be dependent on the experience of the analyst [59]. Unique methods include EFA [62, 122, 123], Fixed-Sized Window–Evolving Factor Analysis (FSW-EFA) [126], Heuristic Latent Projection Analysis (HELP) [127-129],Window Factor Analysis, (WFA) [130] and Subwindow Factor Analysis (SFA) [66].

Rational techniques such as, Iterative Target Transformation Factor Analysis (ITTFA) [49, 131], ALS [67, 70], SIMPLISMA [64, 118, 120, 121, 132], Orthogonal Projection Analysis (OPA) [95, 97, 133, 134] can be classified by those techniques which aim at finding a rational resolution in which the factors for single species do not violate the generic prior knowledge, such as non-negativity, unimodality, etc. Rational resolution

may produce a set of feasible solutions and the accuracy of the solutions depends on the correlation or collinearity among the pure profiles underlying the two-way data. Rational resolution methods tend to produce solutions which approximate the true solution very well, if the correlation amongst the chemical constituents is not severe. Rational resolution methods can be distinguished in the way in which the initial estimates are determined or in the optimisation algorithm to iteratively improve the solution [59].

The methods of EFA, SIMPLISMA and ITTFA are described below. These tools were used in the investigation of some of the processes under study (see Chapter 2 and 3).

I.2.3.2.1 EFA

EFA, developed by Maeder *et al.* [62, 123], is a unique resolution method, that was originally developed for the resolution of overlapping chromatographic peaks. However, this technique can be applied to any chemical system which evolves in a systematic non-random way as a function of time, pH, temperature, etc. These systems are called evolutionary systems and are normally complex systems. Examples of such systems include partially resolved chromatographic peaks eluting as a function of time [79, 92, 135, 136].

There are two EFA methods which can be used to estimate the evolutionary profiles. Method (1) is the iterative approach, in which evolutionary profiles are estimated through successive local rank analysis of sub-windows within a measurement matrix. ALS regression is used to compute the normalised spectra of the analytes. Method (2) is a non-iterative approach, in which the zero component regions of the estimated evolutionary profiles derived from the initial EFA analysis (see Method 1) are used to

determine the rotation matrix **R**. The 'real' evolutionary profiles are computed and a least square calculation is used to compute the normalised spectra of the analytes.

Typically in curve resolution analysis the initial estimate of the evolutionary profile is determined using the iterative approach (Method 1). To reduce the number of feasible solutions a selectivity condition can be included for the generation of the 'real' evolutionary profiles in the form of zero component regions. To complete this zero component regions are determined from the evolutionary profiles derived from method 1 and this information with the evolutionary profile are used in the non-iterative procedure. An overview of both the iterative and non-iterative procedures is given in figure 2 and described below.

Iterative EFA

Datasets can either be factor analysed column-wise or row-wise (depending on the information required). The rows tend to follow a logical sequence, such as response according to time and this property can be exploited to locate the pure row factors in the matrix. This is illustrated on a four component chromatogram given in figure 3. The compounds are present in well defined time windows, e.g., compound **A** is present in window $t_1 - t_4$, compound **B** is present in window $t_2 - t_5$, compound **C** is present in window $t_3 - t_7$ and compound **D** is present in window $t_6 - t_8$. Factor analysis is applied in succession to the sub-matrices $\mathbf{D}_i$ formed by the first 1,2, …,*i* ,…,*n* spectra (formed by adding rows from an initial top sub-matrix, $\mathbf{T}_0$ to the bottom sub-matrix $\mathbf{B}_0$). In the example the rank of these matrices increase from one to four as schematically shown in figure 3. By analysing the ranks as a function of the number of added rows, time windows are derived where one, two or three etc. significant PCs are present. By plotting the eigenvalue, variance described by each eigenvector (or the log of the

eigenvalue) against the number of rows added, it is possible to derive (from the significant eigenvectors), the number of new components being produced for each sub-matrix.

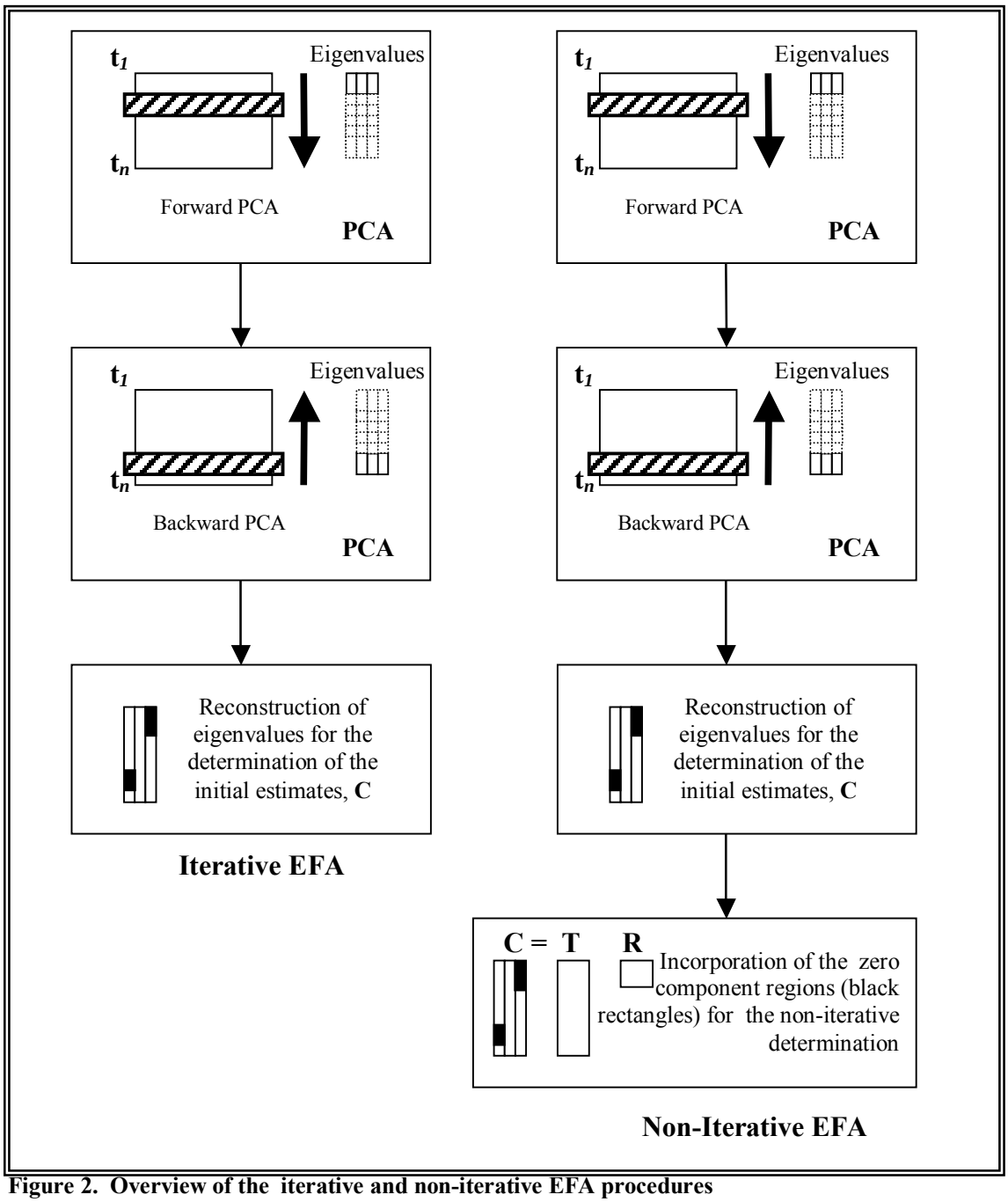


Figure 2. Overview of the iterative and non-iterative EFA procedures

This is shown in figure 4b and the chromatographic elution of components **A-D** is shown in figure 4a. Figure 4b can be interpreted as follows. For each sub-matrix a new significant eigenvector appears each time a new compound is introduced into the

spectra, thus at $t = t_1, t_2, t_3$ and t_5. This is observed from an increase in eigenvalue. Naturally the *i*th eigenvalue is not a direct measure of the concentration of the *k*th species.

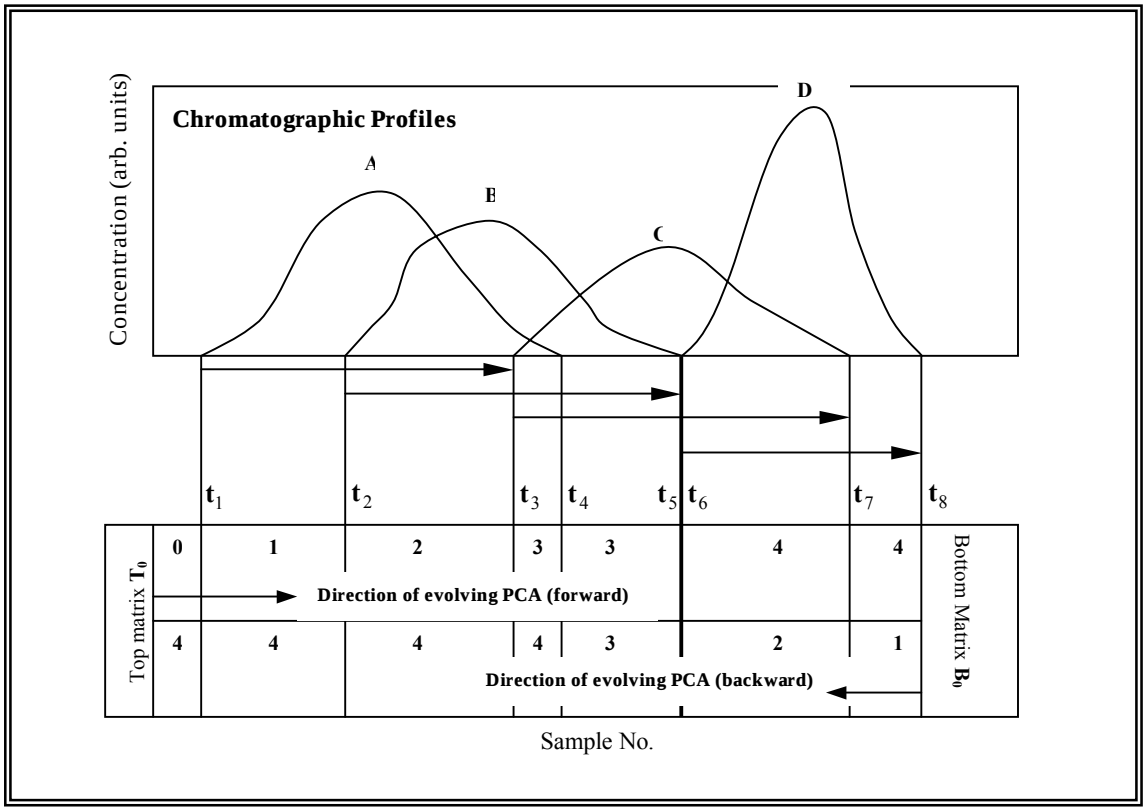


Figure 3. Time windows in which four compounds are present. The rank of the data matrices are formed by adding rows to a top matrix T_0 (from the top to the bottom matrix) or by adding rows to a bottom matrix B_0 (from the bottom to the top).

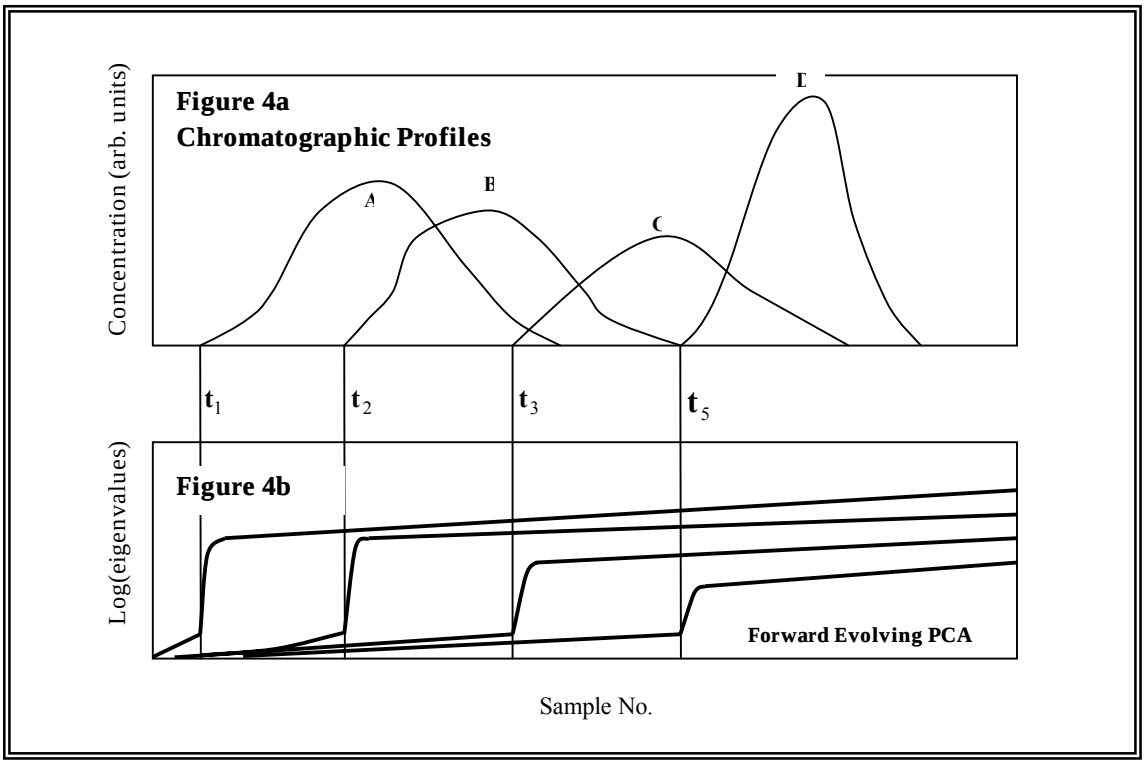


Figure 4. Forward Evolving PCA. Figure 4a. The model chromatographic profiles. Figure 4b. Eigenvalues calculated when adding rows to T_0 (forward evolving PCA). The forward EFA plot indicates the elution of new components at windows t_1, t_2, t_3 and t_5.

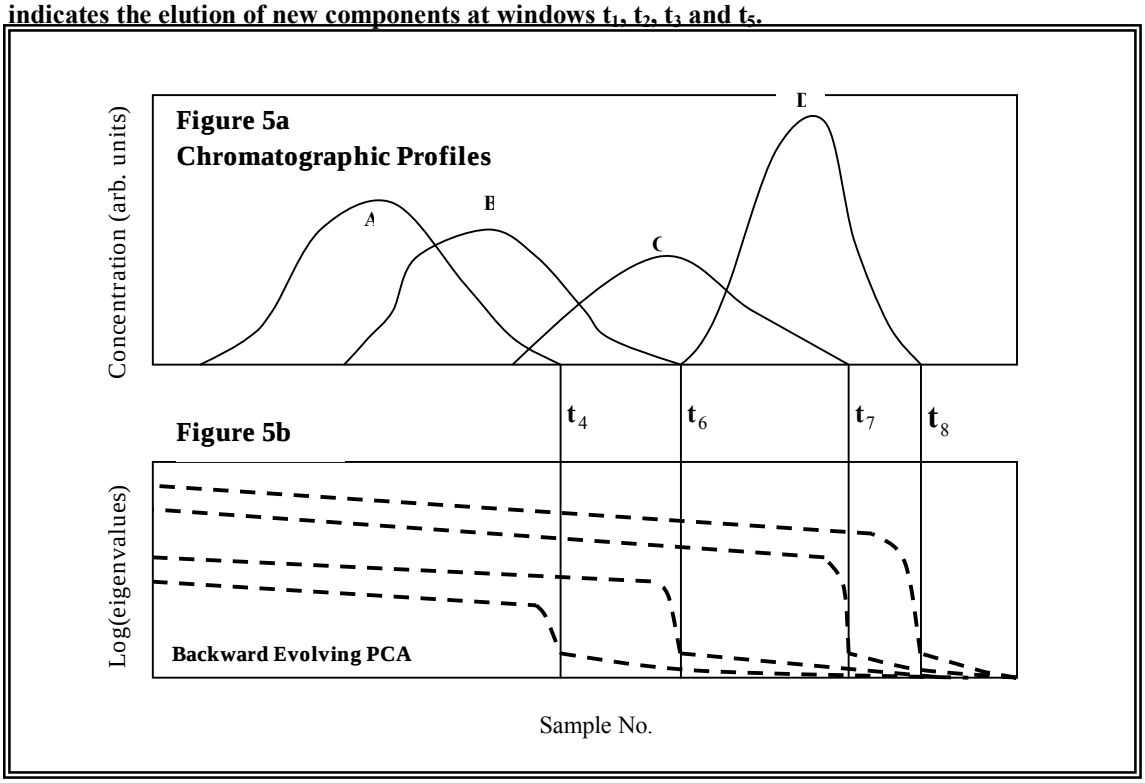


Figure 5. Backward Evolving PCA. Figure 5a. The model chromatographic profiles. Figure 5b. Eigenvalues calculated when adding rows to B_0 (backward evolving PCA). The backward EFA plot indicates the disappearance of components at windows t_4, t_6, t_7 and t_8.

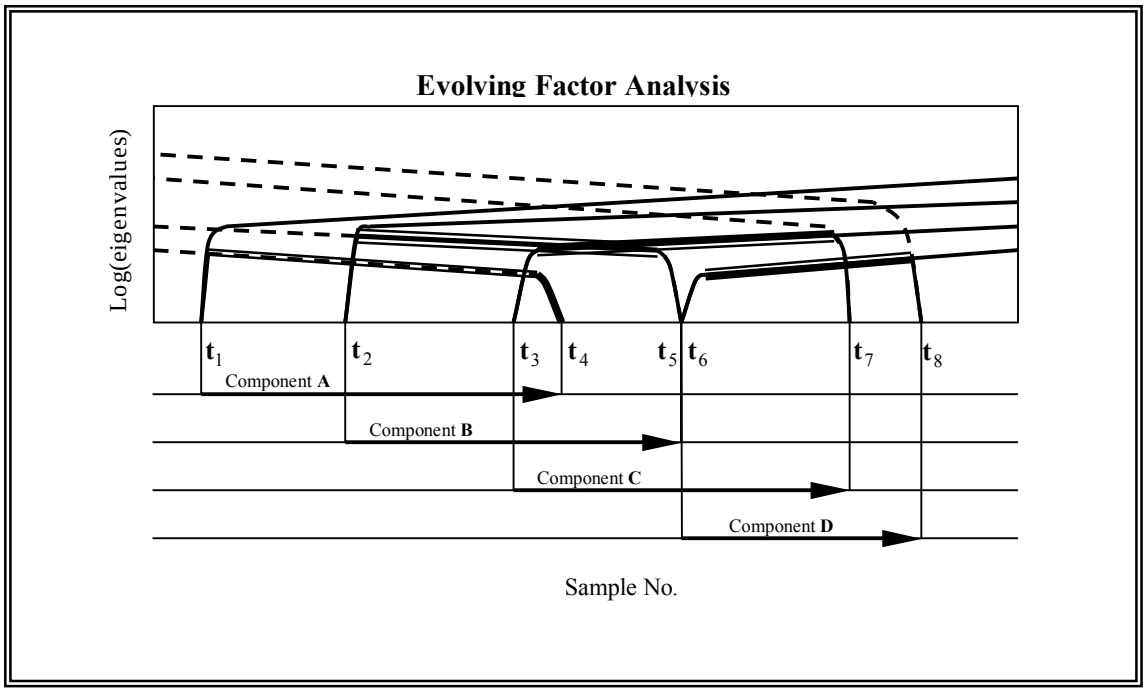


Figure 6. Reconstructed concentration profiles from the combination of figures 4b and 5b. The concentration window for the *k*th component is defined by the rise of the *i*th eigenvalue in the forward EFA plot(__) and the (*n*+1-*i*)th eigenvalue in the backward EFA plot (---).

The eigenvalues are strongly dependent on the dissimilarity of the absorption spectra and the elution profiles [62]. Nevertheless, from figure 4b it is not yet possible to derive the compound windows, as this plot indicates the appearance of a new compound but not its disappearance. Therefore, a second analysis is completed in the reverse order, i.e., one starts from the bottom sub-matrix, $\mathbf{B}_0$, and rows are added to this sub-matrix to the top sub-matrix, $\mathbf{T}_0$. A similar plot is shown in figure 5b of the reverse analysis. New factors appear at $t = t_1, t_2, t_3$ and t_5 in the forward EFA analysis and disappear at $t = t_4, t_6, t_7$ and t_8 in the backward EFA analysis. The compound windows are found by connecting the rising part of the *i*th forward curve with the falling part of the (*n*+1-*i*)th backward curve. This results in a rough estimate of the concentration window of the *k*th component. The compound regions are found by connecting the first appearing compound with the last appearing compound, shown in figure 6. All the resulting curves are arranged into the columns of the concentration matrix **C**. These

evolutionary profiles are abstract representations of the true concentration profiles. These concentration estimates are used to initialise the ALS procedure, see step 1 and 2 of the SMCR methodology, section I.2.1.1.

Non-Iterative EFA

The non-iterative calculation of the evolutionary profiles is described below. The evolutionary profiles and the zero component regions derived from the iterative EFA procedure are used in the non-iterative procedure. The rotation matrix **R** is calculated using the concentration windows determined from EFA analysis (see above). The important factor for resolution with EFA is the so-called zero-concentration window. An analytes zero–concentration window is defined as the part of the data matrix where the analyte does not contribute to the signal, but all other analytes do. All analytes must have a good zero-concentration window if the non-iterative EFA is to succeed. In this context, the term 'good' implies that the other analytes contribute significantly to the signal in the zero-concentration window when compared to the contribution from the noise.

If we concentrate on the evolutionary profiles, where $\mathbf{C}$ in $\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} = \mathbf{T}\mathbf{R}\mathbf{R}^{-1}\mathbf{P}^{\mathbf{T}}$ can be written as equation 21.

$$\mathbf{C} = \mathbf{T}\mathbf{R} \qquad \textbf{Equation 21}$$

Where $\mathbf{T}$ and $\mathbf{P}^{\mathbf{T}}$ are the independent eigenvectors.

Once the rotation matrix **R** is found the resolved concentration profiles **C** are computed. Use of equation $\mathbf{S}^{\mathbf{T}} = \left(\mathbf{C}^{\mathbf{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathbf{T}}\mathbf{D}$ resolves the spectra. Figure 7 illustrates how **R** is found. The grey areas symbolise the zero-concentration window for each of the three

analytes present. Column k in **R** represents the rotation vector that takes us from the abstract evolutionary profiles **T**, which is the scores matrix of **D**, to the real evolutionary profiles $\mathbf{c}_k$. This means that we can solve the rotation problem separately for each analyte as indicated in figure 7b and equation 22.

$$\mathbf{c}_k = \mathbf{T}\mathbf{r}_k \qquad \textbf{Equation 22}$$

Equation 22 is represented in figure 7b. The crucial idea here is that the grey areas of $\mathbf{c}_k$ is a linear combination of the grey areas of $\mathbf{T}$. To get figure 7c the grey areas of the vectors are employed. $\mathbf{c}_k^0$ is the zero vector and therefore is a homogeneous system of equations with the obvious trivial solution $\mathbf{r}_k = 0$. The rank of $\mathbf{T}^0$ however is only $nc-1$ as the contribution of component $\mathbf{c}_k$ is eliminated. Therefore, this homogeneous system of equations has a non-trivial solution; one element of $\mathbf{r}_k$ can be chosen freely and the rest of $\mathbf{r}_k$ is calculated by a simple linear regression. $\mathbf{c}_k$ is determined by application of equation 22. This procedure is repeated in turn for all *nc* components, thus yielding the complete concentration matrix **C** which at the end of the calculation is used to compute the spectral matrix.

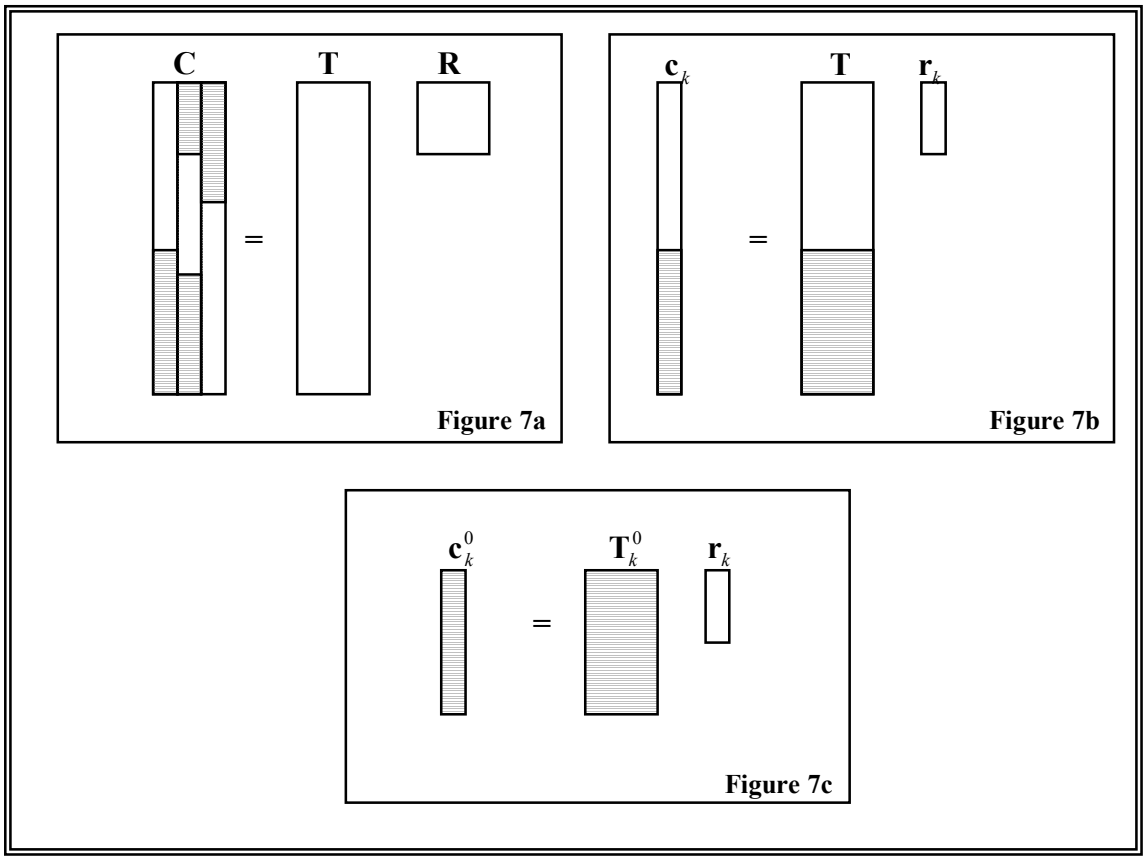


Figure 7. Finding the rotation matrix of R using EFA. Figure 7a. The scores of T are rotated into the evolutionary profiles C by means of R. The grey areas designate the zero-concentration window of the analytes. Figure 7b.The rotation step is completed independently for each analyte. The *k*th analytes concentration profile is found by rotating the score by means of the *k*th column in R. Figure 7c. Solving the rotation equation. $\mathbf{c}_k^0$ is the zero component vector of $\mathbf{c}_k$ and $\mathbf{T}_k^0$ is the part of the scores that correspond to the analytes zero-component window. The rotation vector, $\mathbf{r}_k$, that rotates the scores into the evolutionary profiles can now be found.

EFA is not dependent upon selective regions, which means that even systems with complete overlap in both directions can be solved [136]. It is important to note that EFA identifies independent factors not constituents. Reactants that co-vary would not be split out as a separate factor [17].

I.2.3.2.2 SIMPLISMA

SIMPLISMA [64, 118-121], on the other hand, provides the number of components and the purest concentration or spectral profiles directly from the measurement matrix, based upon a purity criterion. SIMPLISMA forms part of the *rational resolution*

method or *pure column or row factor analysis* methods developed by Windig and Guilment [64].

SIMPLISMA is a pure-variable based method[64] and is used to select the pure components from the original measurement matrix. This means that it is assumed that every component in the mixture under study has a variable, which has a finite intensity for that particular component, and that the variable has a zero intensity for all other components in the mixture. In contrast to most SMCR methods, SIMPLISMA does not use PCA to resolve the data. It is based on the evaluation of the relative standard deviation (σ_j/μ_j) of the columns (wavelength) *j* on **D**. This yields a standard deviation spectrum, referred to as the purity spectrum, **f**. The *pure variable* is basically the variable having the maximum ratio of the standard deviation and the mean of all the intensities. The corresponding concentration profile at this pure variable is used as an initial estimate for this component.

$$f_j = (\sigma_j/\mu_j) \qquad \textbf{Equation 23}$$

In equation 23, f_j represents the purity value at the *j*th variable index. The mean of the column vector at variable *j*, μ_j is calculated as given in equation 24;

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} d_{i,j} \qquad \textbf{Equation 24}$$

and σ_j represents the standard deviation of variable *j*, as shown in equation 25;

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_{i,j} - \mu_j)^2} \qquad \textbf{Equation 25}$$

Where $d_{i,j}$ is the intensity at the *j*th variable, for the *i*th case. Problems may arise for variables with a low noise range intensity. In order to correct for this, the purity is redefined by the addition of a constant (offset), δ, to the denominator of equation 23, as shown in equation 26;

$$f_j = \frac{\sigma_j}{\mu_j + \delta} \qquad \textbf{Equation 26}$$

If δ has a relatively low value with respect to μ_j (i.e., for a high value of μ_j), the effect will be negligible, but for low values of μ_j (i.e., in the noise range), the effect is that this noise correction term will make the purity value lower, which is what is required to correct for noise. Typical values for δ range from 1-5% of the maximum of μ_j. The % is selected through sequentially increasing the percent deviation by a percentage mark and evaluating the SIMPLISMA solution. A good estimate of the constituent profile would appear to be smooth, whereas a noisy profile may appear to be rugged.

All the f_j values are plotted in the form of a spectrum, a so-called *purity-spectrum*, **f**, in which the wavelengths with the highest intensity represents the first pure variable. The visualisation of the purity of the variables facilitates the detection of pure variables caused by unwanted features in the datasets such as noise.

The determination of the *next pure variable*, $f_{j,k}$, where *k* is the *k*th pure variable that will be selected from the *k*th purity spectrum, is different to the determination of the *first pure variable* (wavelength) $f_{j,1}$, because the weight factor, $\omega_{j,k}$ is calculated using a determinant-based function; *this ensures that the second pure variable is selected*

orthogonal to the first (or is least uncorrelated to the first pure variable). The general formula for the purity spectrum with the determinant-based function is given in equation 27.

$$f_{j,k} = \omega_{j,k} \frac{\sigma_j}{\mu_j + \delta}$$ **Equation 27**

$\omega_{j,k}$, is a weight factor, which is added to correct for previously chosen variables. Variables which are highly correlated with the previously selected pure variable are down weighted, i.e., they have a value for $\omega_{j,k}$ close to 0, while variables that are dissimilar to the previously selected pure variable have a high value for $\omega_{j,k}$ close to 1. All the f_j values for the *k*th pure variable are plotted in a purity-spectrum, $\mathbf{f_k}$, in which the wavelength with the highest intensity represent the next pure variable. This procedure is repeated for the all the *nc* pure variables to be determined.

An alternative procedure is to evaluate the purity of the rows *i* instead of the purity of the columns *j,* this is known as the T-SIMPLISMA approach [77], given in equation 28.

$$f_{i,k} = \omega_{i,k} \frac{\sigma_i}{\mu_i + \delta}$$ **Equation 28**

The rows with the highest purities are estimates of the row factors. The corresponding spectral profile at this pure row factor is used as an initial estimate for the *k*th pure component.

The concentration estimates determined by the SIMPLISMA approach or the initial spectral estimates determined by the Transposed - Simple to-use interactive Self modelling

Mixture Analysis (T-SIMPLISMA) approach can be used to initialise the ALS procedure, see step 1 and 2 of the SMCR methodology, section I.2.1.1.

I.2.3.2.3 ITTFA

The final exploratory method described is ITTFA. This method is similar to target transformation factor analysis (TTFA) [11, 63], the only difference being that target factor analysis is completed on a series of input vectors, which have been constrained to a known property of the data such as non-negativity. ITTFA is generally applied when good candidate spectra are not available. It was first introduced by Hope *et al.* in environmetrics [137], and Gemperline [49, 61] and Vandeginste *et al.* [131] in chromatography.

Method TTFA

The method of ITTFA is based on TTFA. By TTFA each candidate spectrum is tested individually on its presence in the mixture. The targets are tested in the space defined by the significant PCs of the data matrix. Therefore, TTFA begins with PCA of the data matrix **D** of the measured spectra. Any row of **D** can be written as shown in equation 29.

$$\mathbf{d}_i = \mathbf{t}_i^* \mathbf{P}^{*\mathbf{T}} + \mathbf{e}_i \qquad \textbf{Equation 29}$$

Where $\mathbf{d}_i \ (1 \times m)$ is the *i*th row of **D**, $\mathbf{t}_i^* \ (1 \times nc)$ is the scores of the *i*th row of **D** for the *nc* significant eigenvectors, $\mathbf{P}^{*\mathbf{T}} \ (nc \times m)$ is the loadings of the *nc* significant eigenvectors and $\mathbf{e}_i \ (1 \times m)$ is the error associated with the *i*th row of **D**.

Each mixture spectrum is a linear combination of the *nc* significant eigenvectors. Equally, the pure spectra are linear combinations of the first *nc* PCs. A target spectrum

taken from the library can be tested on this property. If the test passes, the spectrum or target may be one of the pure factors. This is completed as described below;

The first step is to calculate the scores $\mathbf{t}_{\mathbf{in}}^{*}$ of the target spectrum, **in,** to be tested by solving equation 30:

$$\mathbf{t}_{\mathbf{in}}^{*} = \mathbf{in}\mathbf{P}^{*}\left(\mathbf{P}^{*\mathrm{T}}\mathbf{P}^{*}\right)^{-1} = \mathbf{in}\mathbf{P}^{*} \qquad \textbf{Equation 30}$$

These scores give the linear combination of the PCs that provides the best estimation (in a least squares sense) of the target spectrum. How good that estimation is can be evaluated by calculating the sum of squares of the residuals between the re-estimated targets or output target (from its scores) and the input target. The output target **out** is equal to $\mathbf{t}_{\mathbf{in}}^{*}\mathbf{P}^{*\mathrm{T}}$. The overall expression for TTFA is given in equation 31.

$$\mathbf{out} = \mathbf{in}\mathbf{P}^{*}\mathbf{P}^{*\mathrm{T}} \qquad \textbf{Equation 31}$$

If the difference between **out** and **in** $\left(\|\mathbf{out} - \mathbf{in}\|\right)$ can be explained by the variance of the noise, the test passes and the target is possibly one of the pure factors.

Method of ITTFA

In the absence of good candidate targets to be tested by TTFA, one defines initial targets of the *nc* pure components which are gradually improved until the test passes. The initial targets can be found using the uniqueness test [49, 61, 69] which is performed by constructing a vector of zeros with a single element set to a value 1. The test is performed for each row (sample) in the original data matrix:

$$\mathbf{in}_1 = (1,0,0,...,0,0,0)$$
$$\mathbf{in}_2 = (0,1,0,...,0,0,0)$$
.
.
.
$$\mathbf{in}_n = (0,0,0,...,0,0,1)$$

Each test vector can be thought to approximate a very narrow Gaussian or skewed Gaussian distribution at a particular retention time. When the retention time represented by **in** corresponds to the retention time of a real component, a local minima is observed in the sum of the squares of the difference between the test vector, **in**, and the predicted vector **out**. The local minimum indicates that the very narrow Gaussian test peak is a better approximation of the real elution profile at the selected retention time. The test is repeated at each of the retention times represented in the raw data matrix so that *nc* local minima may be found, each one corresponding to the retention time of one of the *nc* real components. When more than *nc* minima are found, only the *nc* smallest are selected. This is the so-called needle search.

An initial input target $\mathbf{in}_1$ is projected into space defined by the eigenvectors. The target is tested by inspecting the output vector $\mathbf{out}_1$ as described in TTFA. If the input and output vector are significantly different then the loadings of the output vector can be inspected to ensure that specific requirements are met for the data, such as the non-negativity of the response for elution profiles, or removal of secondary maxima or shoulders in peaks. The data is constrained by correcting negative values and removing certain noise from the data. The consequence of the application of the constraint is that vector $\mathbf{out}_1$ is lifted from the plane and rotated over a smaller angle, giving vector $\mathbf{in}_1'$, to obtain results which are closer to the true factors, than $\mathbf{in}_1$.

The input vector $\mathbf{in}_1'$ is regarded as a new target spectrum. TTFA is applied and a second output vector $\mathbf{out}_1'$ is produced. The overall result is the rotation of $\mathbf{out}_1'$ to $\mathbf{in}_1''$ additional constraints are placed on the loadings until the differences between the input and the output converges to zero, which indicates a stable solution. This solution indicates a pure factor, a second factor may be found by repeating the procedure with a new target input spectrum.

A good solution can be obtained when appropriate constraints are formulated and when good target spectra are available.

The main steps of ITTFA are summarised below:

1. Calculate the significant PC from the data matrix
2. Choose an initial target $\mathbf{in}_1$
3. Project $\mathbf{in}_1$ in the space defined by the eigenvectors by applying equation 31. An output target $\mathbf{out}_1$ is obtained.
4. Evaluate the correlation between the input and output target; if the correlation is larger than a specified value, the procedure converges to a factor. Repeat the procedure with another initial target, until all pure factors are estimated.
5. Otherwise adapt the projection target by applying constraints. This gives a new target to be tested. Return to step 3.

This method has proved to be quite successful, although practitioners have realised the importance of the quality of the initial iterative vector to ensure the success of the resolution and to ensure that the algorithm converges quickly and accurately [81].

I.3 Model Validation

Validation is defined here as ensuring a suitable model is obtained. This covers many diverse areas such as data preprocessing and pretreatment methods, validation of MCR-ALS solution, validation of multivariate calibration analysis and comparative validation of various calibration methodologies using significance testing.

I.3.1.1 Preprocessing and pretreatment methods

The data pretreatment methods optionally applied in the calibration free analysis include truncation, minimum offset, zero average offset, normalisation and standard normal variate (SNV).

Truncation

Truncation removes uncorrelated and uninformative regions from the measured matrix without compromising the bilinearity assumption. In the spectral dimension, truncation allows the removal of wavelength regions which have high amounts of noise or uncorrelated and uninformative regions. Similarly the truncation in the concentration dimension allows the removal of spectral samples measured during times which should not be included in the calibration free data analysis. For example, sampled pre-reaction mixture spectra of a multi-stage reaction may need be separated into individual stages and calibration free analysis attempted separately on each stage.

Offset Methods

The offset methods (the *minimum offset method* and the *zero average offset method*) enable small baseline adjustments to reduce rotational ambiguity by removal of a predefined offset. The minimum offset method searches for the minimum value in each

spectrum and subtracts it from the corresponding value, to ensure that the minimum value in each spectrum is set to zero. If the measured data contains negative values, this method would remove all negative values in the data. This would enable the application of the non-negativity constraint in the ALS optimisation, to reduce the number of feasible solutions. The zero average offset method ensures that the average value in a predefined region which is supposed to be zero, is set to zero. In each spectrum, the average row value of the selected range is subtracted from the corresponding spectrum [138].

Normalisation

In the application of SMCR methods it has been shown that with a specific normalisation, (the profiles are assumed to have either unit norm, which is usually taken as the Euclidean norm or 1-norm in accordance with the specific SMCR method [56, 59, 109]) and under a general non-negativity constraint, the data points in the two-way data **D** are contained in the simplex whose vertices are constituted by the pure components [51, 52, 139]. This has important consequences for reducing the intensity ambiguity and locating the pure variable, and for the development of non-algebraic SMCR techniques. However, this aspect of SMCR analysis (i.e., determining pure variables by locating the vertices of a simplex) is beyond the scope of this thesis.

SNV

An effective preprocessing method is the use of SNV. This type of standardisation works by considering each spectrum, $\mathbf{d}_i$, as a set of n observations and calculating their $\underline{z}$-scores, (not to be confused with needle output spectrum, **z**), given in equation 32:

$$\underline{\mathbf{z}}_i = (\mathbf{d}_i - \mu_i)/\sigma_i$$ **Equation 32**

It has the effect of removing an overall offset by subtracting the mean spectral reading, μ_i, and it corrects for differences affecting the overall variation. SNV effectively reduces sample-to-sample intensity differences and preserves spectral shape without the need of any user-defined parameter. In various settings it has been found to be an effective preprocessing method [48]. However, absolute quantitative information may be lost after SNV transformation due to the scaling operation and thus SNV is primarily used with qualitative methods.

I.3.1.2 Validation of MCR-ALS Solution

To determine whether a stable MCR-ALS solution has been obtained the lack of fit (LOF), which gives a measure of the relative fit quality between the experimental data and ALS reconstructed data, is assessed [107]. The MCR-ALS solutions converge once the LOF (%) is within the defined experimental error, given in equation 33. Where d_{ij} is the experimental absorbance at the sampled point, i, and the wavenumber, j, and $\hat{d}_{ij}$ is the ALS calculated absorbance for that element.

$$LOF(\%) = 100\sqrt{\frac{\sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2}}$$ **Equation 33**

I.3.1.3 Validation of Multivariate Calibration Models

The multivariate calibration models are validated using the regression statistics stipulated below. The regression statistics are substantiated just for the error in the test set. However, the regression statistics for the calibration are calculated analogously using the values from the training set.

The slope and offset of the regression gives an indication of the accuracy of the calibration models. The correlation coefficient between the reference concentrations, y_i, and predicted concentration, $\hat{y}_i$, are calculated to determined whether a linear relationship exist between y_i and $\hat{y}_i$. A correlation value of plus one, represents a perfect positive correlation, a value of zero means that there is no correlation. The Root Mean Square Error of Prediction (RMSEP) is a measure of the accuracy of prediction. The sum of the prediction error for all, *n* samples for the test set is calculated to assess the predictive capabilities of the calibration model. The RMSEP is measured in the same units as y_i, given in equation 34.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad \textbf{Equation 34}$$

The Standard Error of Prediction (SEP) is a measure of the precision of prediction, given in equation 35. The BIAS (absolute deviation from $\bar{y}$) tracks the systematic prediction error, given in equation 36. The Relative Error (RE %) is similar to the LOF calculation, but gives a measure of the fit quality between the predicted and reference concentrations, given in equation 37.

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i - BIAS)^2}{n-1}} \qquad \textbf{Equation 35}$$

$$BIAS = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \qquad \textbf{Equation 36}$$

$$RE(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2}} \qquad \textbf{Equation 37}$$

II Processes Understudy

II.1 Exploratory and Quantitative Analysis of the Catalysed Asymmetric Transfer Hydrogenation Reaction of a Prochiral Imine

II.1.1 Introduction

This study was completed in collaboration with Melanie Ropic and Professor Donna Blackmond, The Catalyst Group, The University of Hull, UK.

Aim

The aim of the study was to use calibration free techniques to characterise the asymmetric transfer hydrogenation reaction of prochiral imine, 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline, without the explicit use of the underlying chemical model linked to it. This avoided error caused by the assumption of a wrong model and allowed the presence and modelling of chemical components, which were not realised from prior quantitative analysis. Secondly, CFT were applied to provide an alternative method for the HPLC analysis of the prochiral imine and chiral amine; 1-methyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline. This was essential as the HPLC method was time consuming (the HPLC workup was approximately twice as long as the reaction (~1hr)). For explorative analysis, the HPLC method was limited to target analytes, as intermediates and by-products were frequently lost in the work-up procedure.

Introduction

The catalysed asymmetric transfer hydrogenation reaction is an increasingly common industrial reaction which is used for the in-situ hydrogenation of prochiral imines to produce enantiomerically pure chiral amines. Enantiomerically pure chiral amines are of increasing commercial value in the fine chemical and pharmaceutical areas in view of

their application as resolving agents, chiral auxiliaries/chiral bases and catalysts for asymmetric synthesis. However, chiral amines often possess pronounced biological activity in their own right, and hence are in significant demand as intermediates for pharmaceuticals and agrochemicals in an expanding market where revenues due to chiral technology are expected to reach US$19.9 billion by 2009 [140].

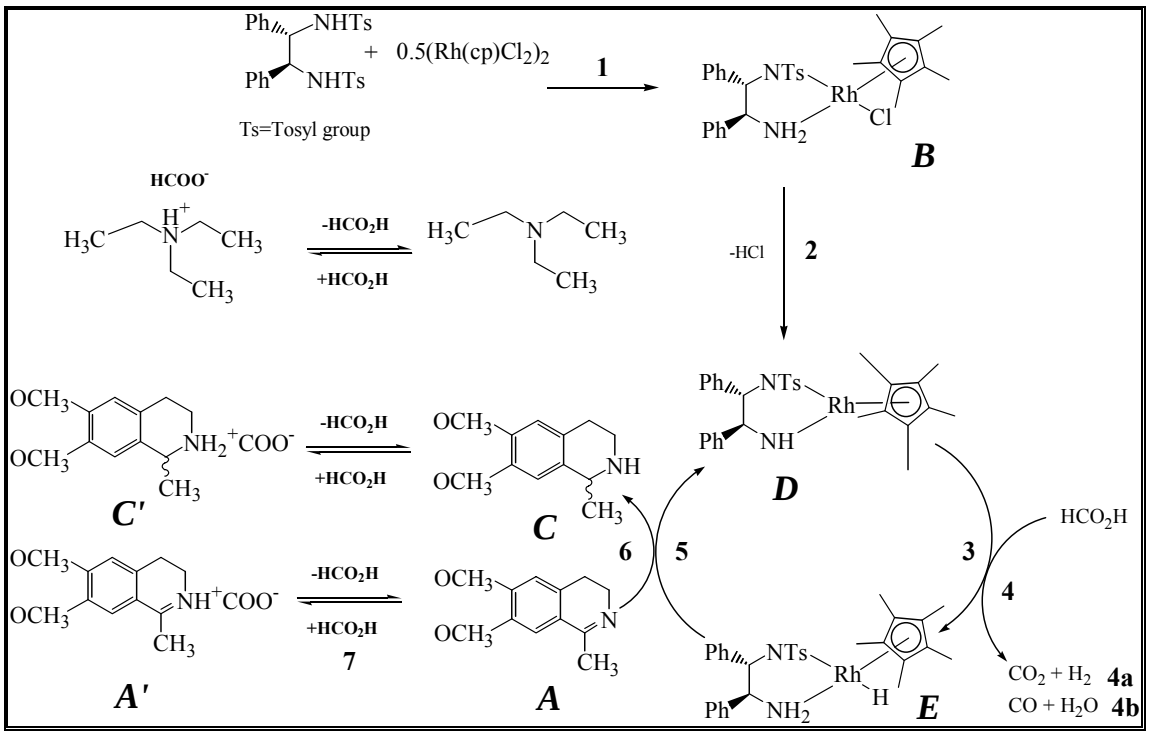


Figure 8. Noyori and Hashiguchi's [141] proposed reaction mechanism for the asymmetric hydrogenation of imines.

The catalysed asymmetric transfer hydrogenation (CATHy) has been shown to be a beneficial generic method both economically and technically for the reduction of C=N and the saturation of C=C and C=O linkages [141, 142]. It is operationally very simple and requires non-hazardous organic molecules. It has been shown to be a powerful alternative to asymmetric hydrogenation using molecular hydrogen with chiral Ru(II)-bisphosphane catalysts due to its practical simplicity and the possibility of using accessible and robust ligands [143]. Noyori and coworkers [141, 142] found that CATHy reactions were particularly useful for the asymmetric enantioselective reduction

of cyclic imines giving amines with 90–97% enantiomeric excess (ee) (using Ru-CATHy), which opened a new general route to natural and synthetic isoquinoline alkaloids as well as a convenient preparation for chiral amines [141, 142].

Nevertheless, a complete investigation of the CATHy mechanism is required (from an industrial perspective) to monitor and predict the evolution of specific chemical constituents within the reaction to aid control. Uematsu *et al*. [142] proposed the most feasible and detailed mechanism for this reaction, shown in figure 8, with Casey [144] and Yi [145] affirming the likelihood of steps 3, 5 and 6 of the stepwise hydrogen transfer. In this mechanism in-situ formation of the catalyst precursor from the reaction of $(Rh(cp)Cl_2)_2$ and 1,2-diphenyl-1-tosyl-2-aminoethane, ***B,*** is followed by a reaction to remove HCl to give the active catalyst species, ***D***. It is suggested that formic acid adds irreversibly to form the metal hydride, which undergoes concerted transfer of the hydride and the N-H proton to afford the amine product, ***C***. In the study completed by M. Ropic *et al.* [146, 147], no erosion of ee was observed, supporting the evidence of the irreversibility of the cycle. However, the potential for catalyst poisoning in side reactions that form carbon monoxide has been noted, cited in [146], see step 4b of the stepwise hydrogen transfer. One of the special features of the imine-formic acid system is the acid-base equilibria that exists peripheral to the catalytic cycle, where both imine, ***A,*** and amine, ***C,*** are shifted to the protonated imine and amine respectively. Nuclear magnetic resonance (NMR) studies completed by M. Ropic *et al.* [146, 147], showed that under the reaction conditions of 1M formic acid/0.4M triethylamine in methanol, both imine and amine were strongly shifted to the aminium ***A'*** and iminium salts ***C'*** respectively.

In this study, CFT were used to resolve the concentration of imine and amine as an alternative to HPLC analysis and to reveal the evolutionary profiles of species not identified through chromatographic analysis.

II.1.2 Experimental

II.1.2.1 Reaction Conditions

The FTIR data was collected by M. Ropic and S. Richards.

Two batches were run, FTIR(I) and FTIR(II), that were identical except for the background collection method.

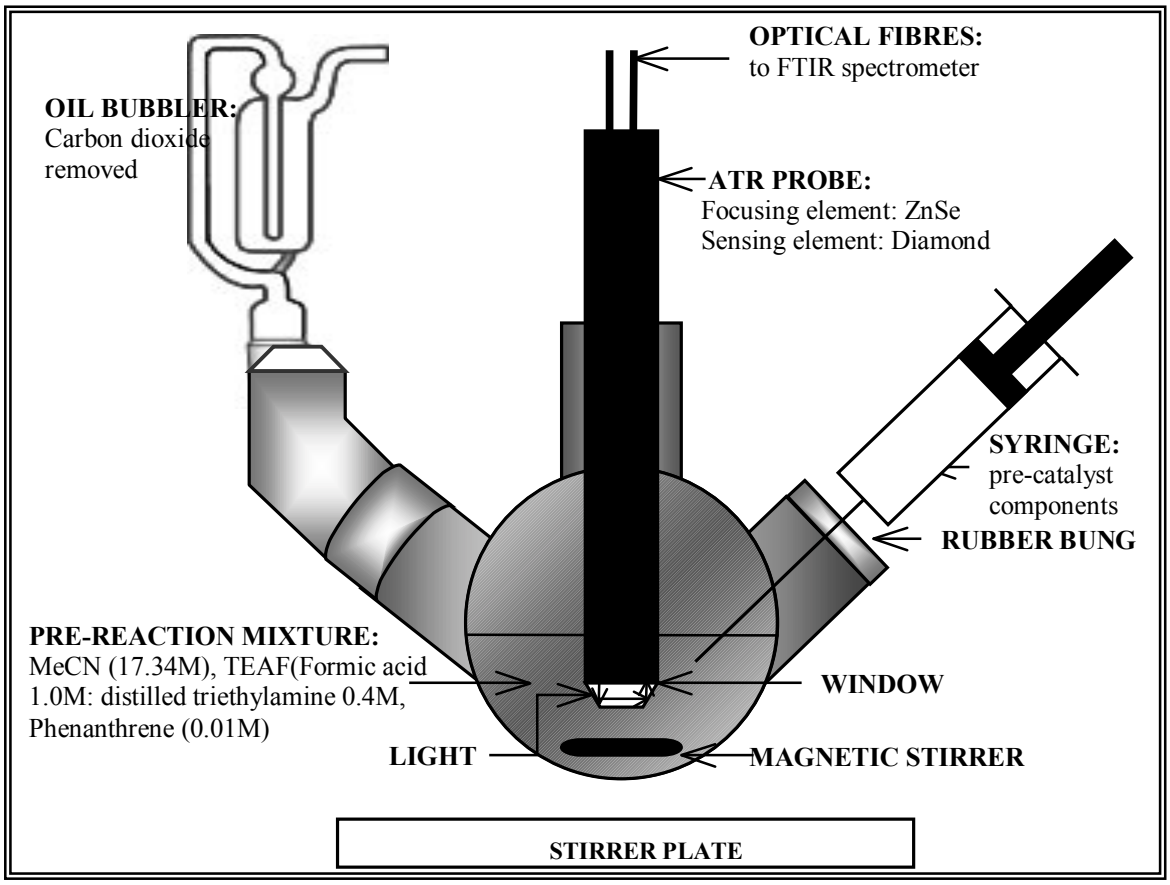


Figure 9. CATHy reation set-up

Negative FTIR(I) - Background subtraction of Triethylamine and formic acid (TEAF)

The reaction solvent, acetonitrile [17.34M; Fisher Chemicals], hydrogen donating reagent TEAF [distilled formic acid 1.0M: distilled triethylamine 0.4M; both Fisher Chemicals] and the HPLC internal standard phenanthrene [0.01M; Aldrich] were placed in a 250ml round bottom flask, which was constantly stirred using a magnetic stirrer and a ReactIR background spectrum of the components was acquired, shown in figure 9. TEAF was subtracted from each of the measured reaction spectrum to prevent the imine signal being swamped. Three sample spectra of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline (imine) [0.25M: Acros Organics] was acquired after imine was added to the reaction mixture and the reaction was initiated by injecting the pre-catalyst components into the round bottom flask [0.0005M, dichloro(pentamethylcyclopentadienyl) rhodium(III) dimer; Strem Chemicals: 0.001M (1*R*,2*R*)-(-)-*N*-*p*-tosyl-1,2-diphenylethylenediamine (TsDPEN); supplied by Avecia, Huddersfield, UK].

Non-Negative FTIR(II)-No background subtraction of TEAF

A second experiment was performed to remove the negative absorbance caused by the subtraction of TEAF from the measured spectra, in order to increase the number of constraints in the calibration free analysis. The reaction constituents subtracted from the measured spectra were the reaction solvent; acetonitrile and the internal standard; phenanthrene. Three sample spectra of the 1-methyl-3,4-dihydroisoquinoline and TEAF, were taken before initiating the reaction by the addition of the pre-catalyst components, to define the start point of the reaction.

HPLC Reference Analysis

Nine aliquots of the reaction mixture were taken and quenched [2M sodium hydroxide (BDH GPR) and dichloromethane (Fisher Chemicals)] at time intervals 0, 1, 2, 5, 10, 15, 30, 45 and 60 minutes for analysis by HPLC [Daicel Chemical Industries, Ltd., Chiralcel OD Analytical column 0.46cm ID x 25cm, chiral stationary phase; cellulose tris(3,5-dimethylphenylcarbamate), mobile phase; Hexane: isopropyl alcohol (IPA), 96: 4]. HPLC samples were worked up by addition of 2M NaOH and dicholomethane (DCM). DCM was dried over $MgSO_4$ before filtering through cotton wool and diluting with MeOH to a total volume of 2ml.

II.1.2.2 Data Acquisition Parameters

The reactions were monitored for an hour using the ASi ReactIR 1000, because it allowed bond changes to be continuously observed during the reaction. The resolution of the instrument was $8cm^{-1}$, and a total of 869 points were recorded in the spectral direction. The sample spectra were recorded every 30 seconds for the first 5 minutes, followed by every 2 minutes for the remaining time, a total of 39 samples were recorded. This was to ensure a higher sampling frequency during the kinetically-controlled stage of the reaction. Each spectrum was recorded at a predefined time using an average of 32 scans. The data was converted to the correct format for data processing in the MATLAB6p5® (The Math Works, Inc) environment using the RECATIR software and an in-house written file conversion program.

II.1.3 Results and Discussion

The aim of this study was to use CFT to characterise the asymmetric transfer hydrogenation reaction of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline, without the explicit use of the underlying chemical model and secondly, to provide an alternative

method for the HPLC analysis of the prochiral imine and amine; 1-methyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline.

II.1.3.1 Pure Spectra

The reference spectra of the reaction constituents were acquired to validate the results of the calibration free analysis. The neat spectra of formic acid and distilled triethylamine were acquired both separately and as a mixture. The neat spectra of imine and carbon dioxide were acquired prior to data acquisition (acetonitrile in the background). The amine spectrum was acquired from a previous synthesis [147]. The traditional notation of *v* for a stretching mode and δ for a bending mode is used in the spectral figures. The common group frequencies for imine and amine were identified and are tabulated in table 1. Characteristic group frequencies for imine are 1675-1600 cm^{-1} for >C=N-C and 1700-1575 cm^{-1} for C=N stretch in substituted imine shown in figure 10. For secondary amines, characteristic group frequencies are 3700-3000 cm^{-1} for N-H stretch, 3500-3100 cm^{-1} for CH-NH-CH stretch and 1650-1500 cm^{-1} for NH bend in secondary amine, shown in figure 11. The characteristic group frequencies of carbon dioxide are 2348-2336 cm^{-1} for the asymmetric stretch of O=C=O and 683-648 cm^{-1} for the degenerate bending of O=C=O, shown in figure 12a. The characteristic bands of formic acid are 3300-2500 cm^{-1} for the O-H stretch, the peak ~1730 cm^{-1} is due to the C=O stretch in saturated carboxylic acid, the peak 1340-1150 cm^{-1} is due to the C=O stretch and 830-670 cm^{-1} is due to the OH deformation, shown in figure 12b. The characteristic functional group frequencies of triethylamine are 1210-1150 cm^{-1} for the CN stretch of the tertiary amino.

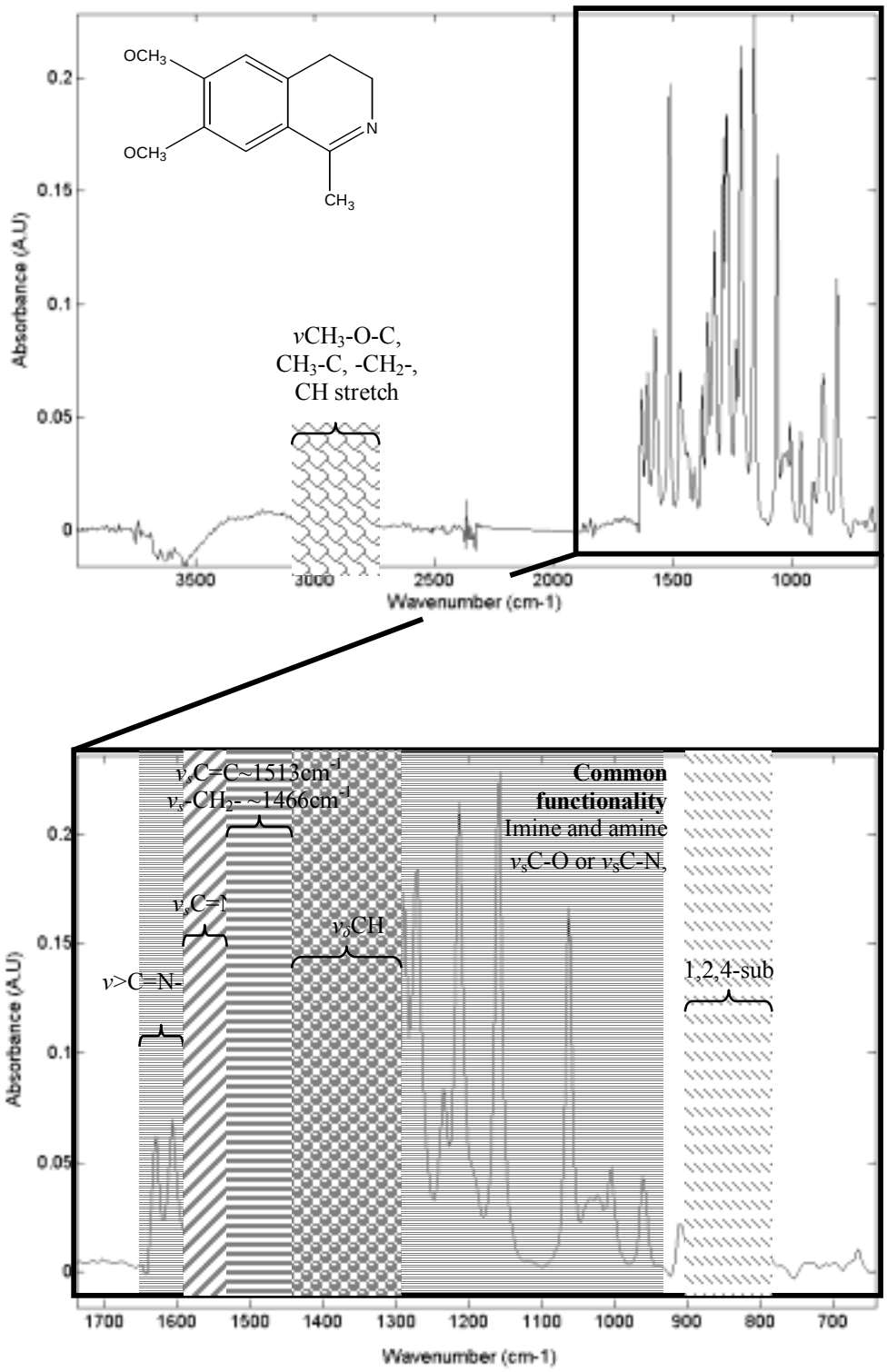


Figure 10. The neat spectrum of 1-methyl-6,7-dimethoxy-3,4-dihydrisoquinoline.

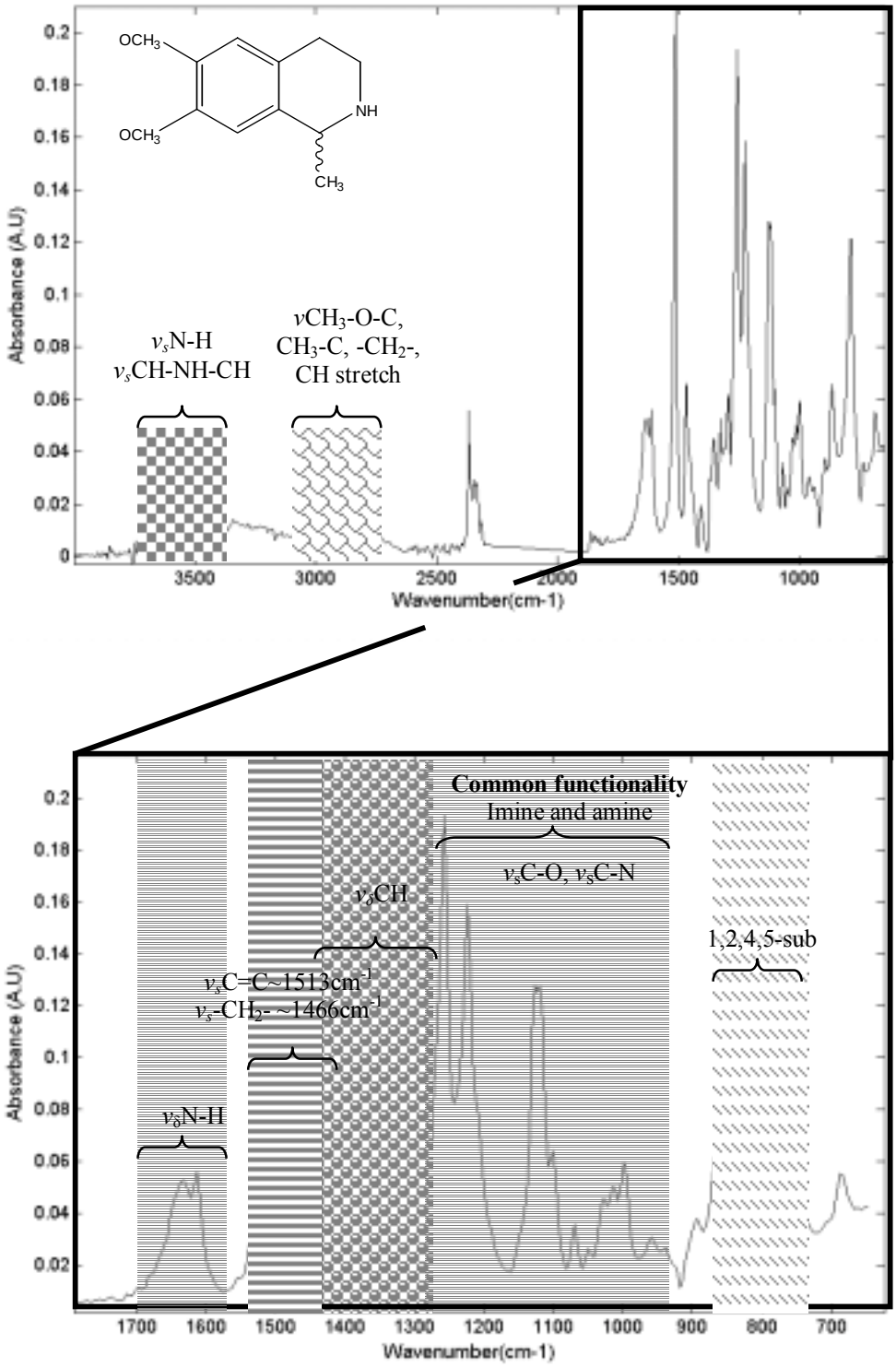


Figure 11. The neat spectrum of 1-methyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline

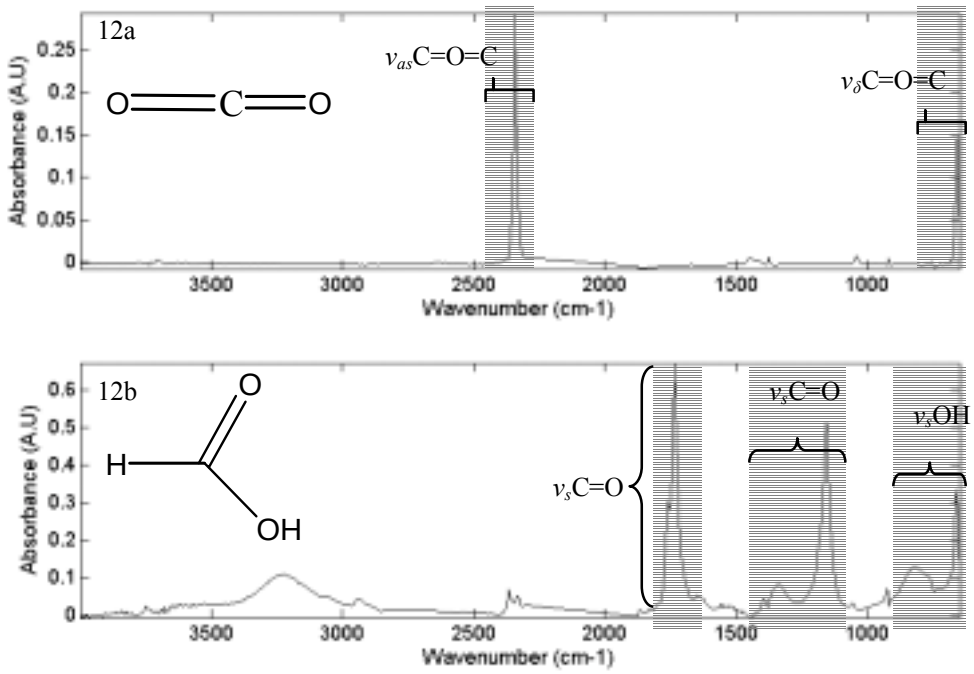


Figure 12a-b. The neat spectrum of carbon dioxide and formic acid respectively

Functional Group	Bond Movement	Frequency range cm^{-1}
Ether or Alkyl	CH_3-O-H / CH_3-C / -CH_2- / CH stretch	3200-2800 / 3000-2840 / 2950 –2800 / 3300-2700
Substituted Aromatics	C=C	1590-1575, 1465-1440
Alkyl	-CH_2-	1475-1425
Alkyl	CH bends	1475-1300
Ether / 2º Amine	C-O stretch / C-N stretch	1300-900
Substituted Aromatics	CH out of plane deformations	900-860 / 860-800 if 1,2,4 –substituted

Table 1. Common functional group frequencies of imine and amine

II.1.3.2 Exploratory Analysis of the CATHy Reaction

The CATHy reaction was completed using the TsDPEN ligand because it was recognised to be the optimal ligand for the rhodium(III) catalysed asymmetric hydrogenation [148]. Formic acid was used as the hydrogen source to reduce the need

for high pressure, from H_2, which may affect the enantio-selectivity and increase the threat of an explosion. Acetonitrile was used as the solvent because very rapid reaction rates were observed compared to isopropyl alcohol, ether and acetone in which the reaction did not reach completion [146].

II.1.3.3 Data Pretreatment

The region of data where diamond absorbs 3188.4 - 2763.8cm^{-1} was removed from both the *negative* FTIR (I) dataset and the *non-negative* FTIR (II) dataset. Initially, the resolution of imine and amine were optimised based on the baseline correction methods, which was essential as the analysis using no pretreatment or a soft baseline correction method, such as zero average offset, resulted in poor resolution of the spectral and concentration profiles.

Negative FTIR(I) – Background subtraction of TEAF

The *negative* FTIR(I) dataset was baseline corrected using a standardisation method, known as SNV. It was found that the SNV corrected measurement matrix produced excellent resolution when initial estimates close to the actual solution were used to initialise the ALS optimisation, for typical fitting see appendix (1.1.1 to 1.1.3). The standardisation was completed by considering each spectrum as a set of observations and then calculating the $\underline{z}$ -scores [48].

Non-negative FTIR(II) – No background subtraction of TEAF

The *non-negative* FTIR(II) dataset was baseline corrected using the minimum offset method, rather than the SNV method, as in this case this method was found to give a better solution when initial estimates close to the actual solution was incorporated into the resolution.

FTIR Reaction Profiles

Initial observations of the *negative* FTIR(I) dataset, shown in figure 13 showed several regions of interest. The first is the negative absorbance emanating from regions **A, B** and **C** 660.9–718.2cm^{-1}, 1141.4-1225.1cm^{-1} and 1692.4-1767.3cm^{-1} respectively.

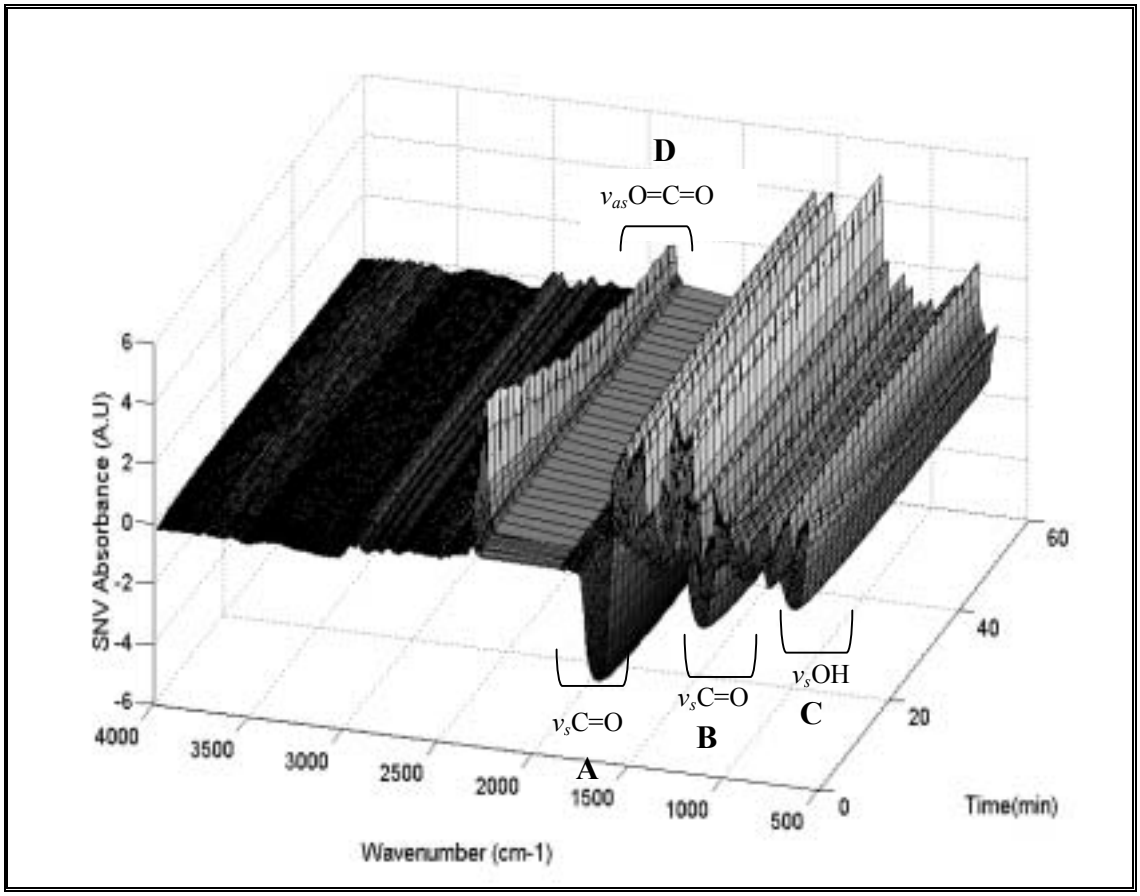


Figure 13. Negative FTIR(I) reaction profile – Background subtraction of TEAF. The negative absorbance of regions A, B and C are representative of formic acid. Region D is representative of the asymmetric stretch of O=C=O in carbon dioxide.

These frequencies were comparable to absorption bands found in the neat spectrum of formic acid taken prior to the analysis, see figure 12b. The negative response intensified as the reaction proceeded, which was due to the decomposition of formic acid to form by-products carbon dioxide and hydrogen. Carbon dioxide is a desirable by-product, because it is highly stable and virtually non-basic. Being a gas carbon

dioxide escapes from the solution into the air because of the high pressure of the gas over the solution. As such, peak **D** increased as carbon dioxide was produced and decreased as it escaped from the solution. It was difficult to distinguish any other characteristic group frequencies from the *negative* FTIR(I) reaction profile because many of the reaction constituents contained the same functional groups at similar vibrational frequencies and so yielded severely overlapped peaks.

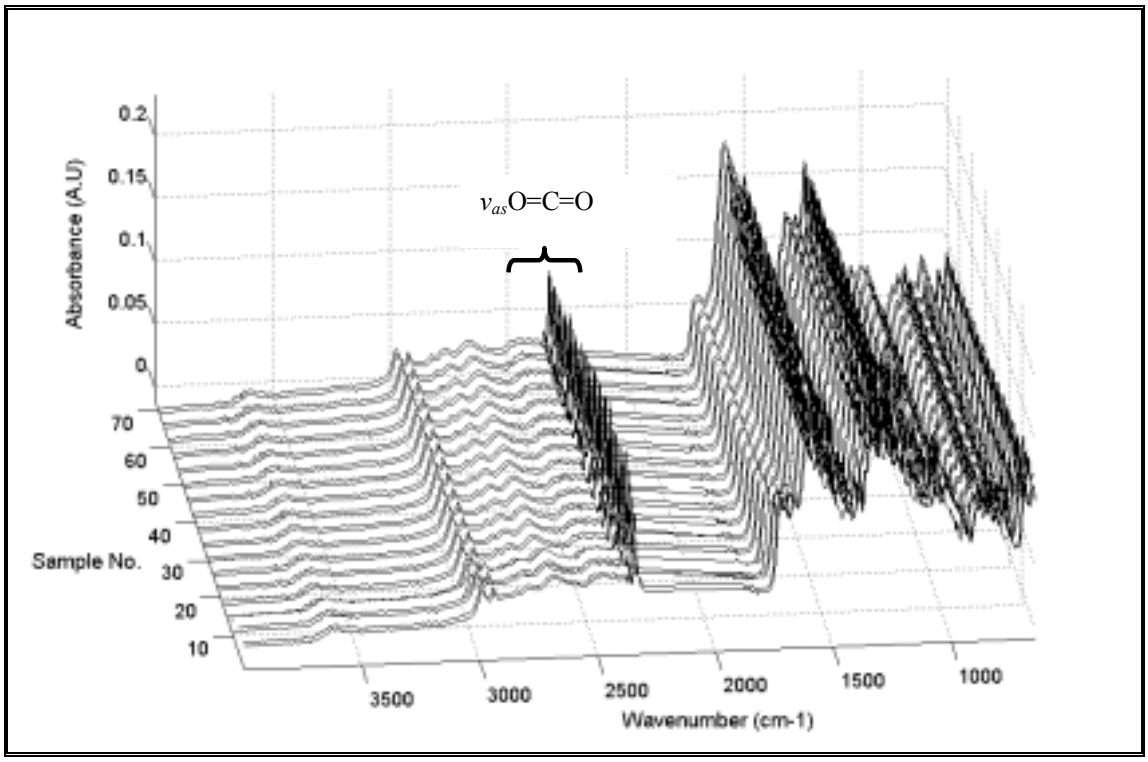


Figure 14. Non-negative FTIR (II) reaction profile – No background subtraction of TEAF.

The reaction profile of the *non-negative* FTIR(II) dataset is given in figure 14. It is possible to distinguish the asymmetric stretch of O=C=O in carbon dioxide and some common group frequencies of the prochiral imine and amine.

II.1.3.4 Preliminary Analysis

PCA was used to determine the number of components to include in the MCR-ALS analysis, through visual inspection of the scores and loadings. The exploratory tools used to validate the number of components and generate the initial estimates of the

independent reaction constituents for MCR-ALS were SIMPLISMA, EFA and the needle spectral estimates derived from the MCR-ALS concentration profiles (see section II.1.3.5).

Principal component analysis

Principal component analysis of the SNV corrected *negative* FTIR(I) reaction profile revealed structured variance in the scores and loading plots of the first three principal components, shown in figures 15-19. Characteristic peaks of formic acid and amine were present in the PC1 loading plot, see figure 15. Imine and carbon dioxide were present in the PC2 loading plot, see figure 16 and the characteristic functional group frequencies of carbon dioxide was present in the PC3 loading plot, shown in figure 17. It was clear that carbon dioxide varied independently within the mixture because it was resolved separately. The fourth PC loading plot was likely to be due to the atmospheric absorption of carbon dioxide in the sample spectrum, rather than a reaction constituent because it contained the asymmetric stretch of O=C=O, but not the degenerate bending of O=C=O (683-648cm^{-1}), see figure 18. The scores plot of PC1, 2 and 3 were relatively smooth, whereas the fourth profile was erratic in nature and as such was attributed to a noise component, figure 19.

Therefore, three independent reaction constituents were resolved although at least five chemical components were expected, which were imine, formic acid, carbon dioxide, amine, and triethylamine substantiated from the proposed Noyori [4] reaction, (see figure 8, step 4a). The catalyst components were unlikely to be resolved as the molecular changes occurred at concentrations lower than the limit of detection. The data is rank deficient because the chemical rank, i.e., the number of reacting chemical components is greater than the mathematical rank of the matrix. No sample outliers

were observed from the first three PCs. Approximately ~97% of the variance was attributed to PC 1, ~3% to PC 2, and ~0.3% of the variance was attributed to PC 3.

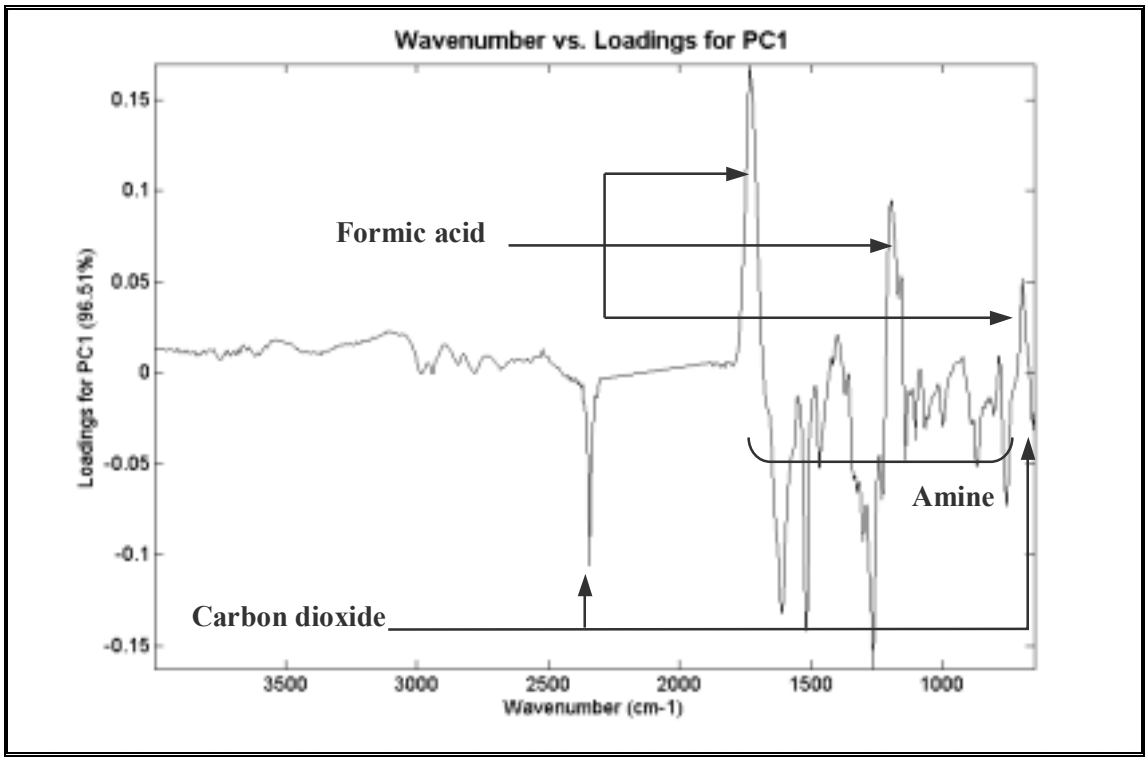


Figure 15. Loadings for the First PC of the negative FTIR (I) reaction profile

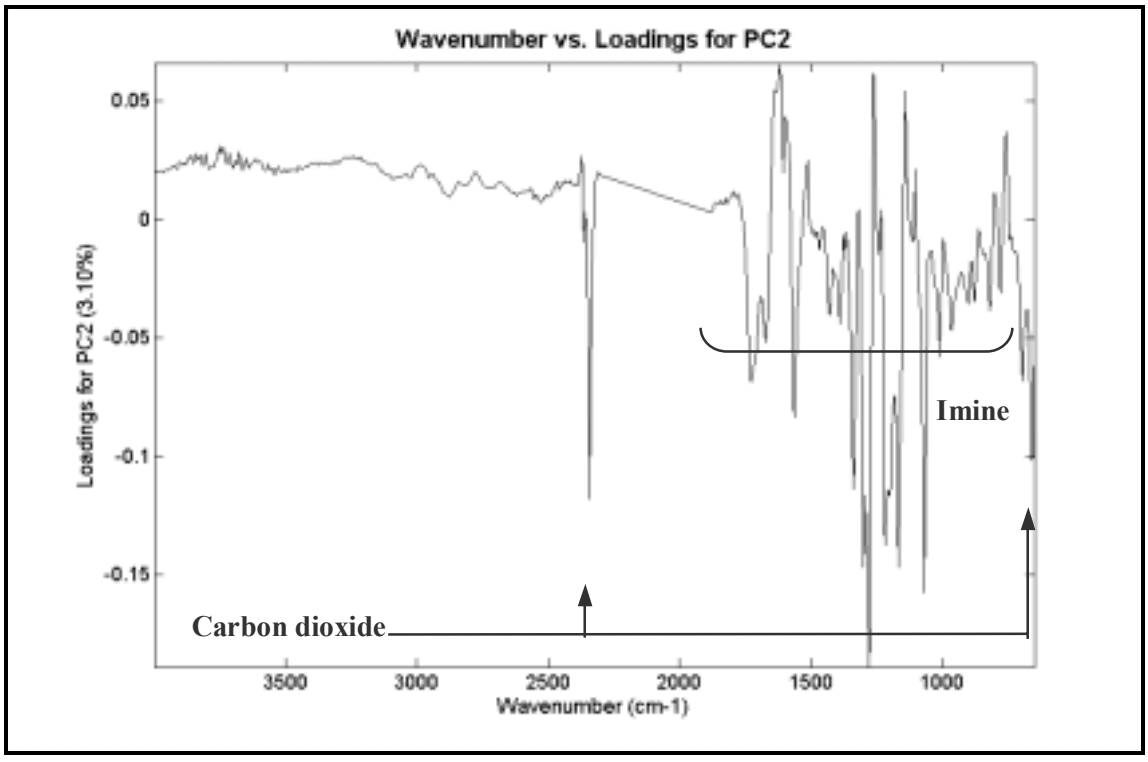


Figure 16. Loadings for the Second PC of the negative FTIR (I) reaction profile

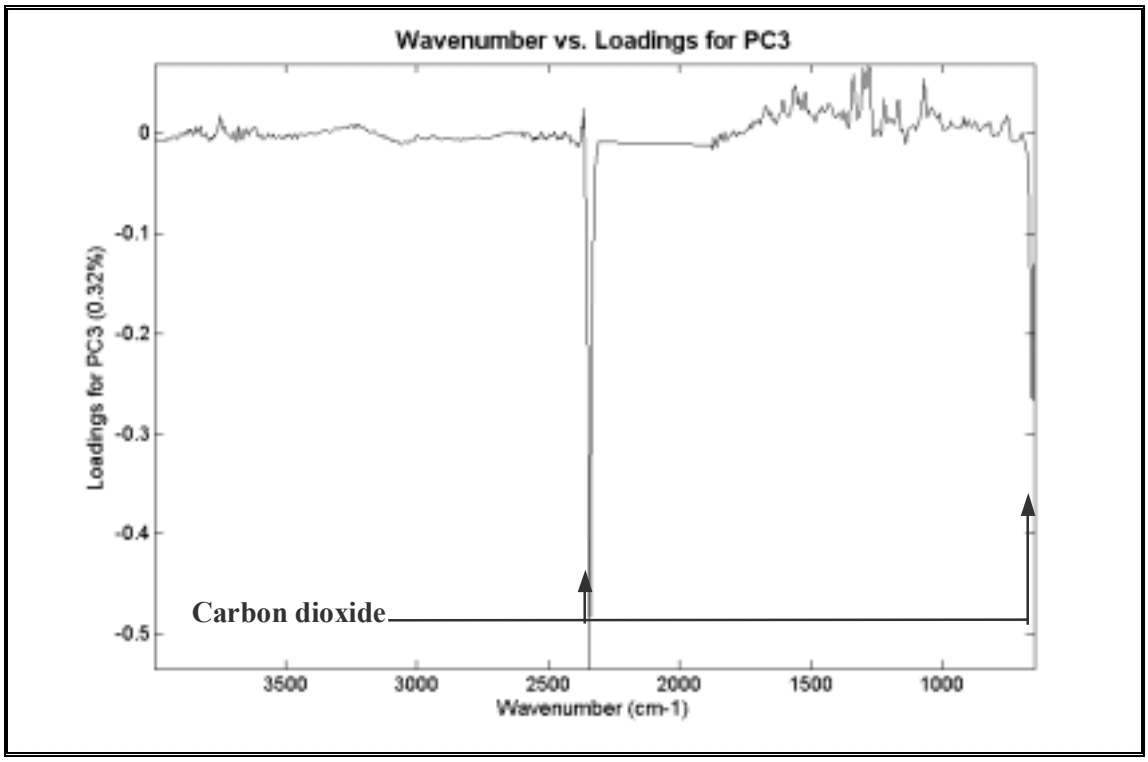


Figure 17. Loadings for the Third PC of the negative FTIR (I) reaction profile

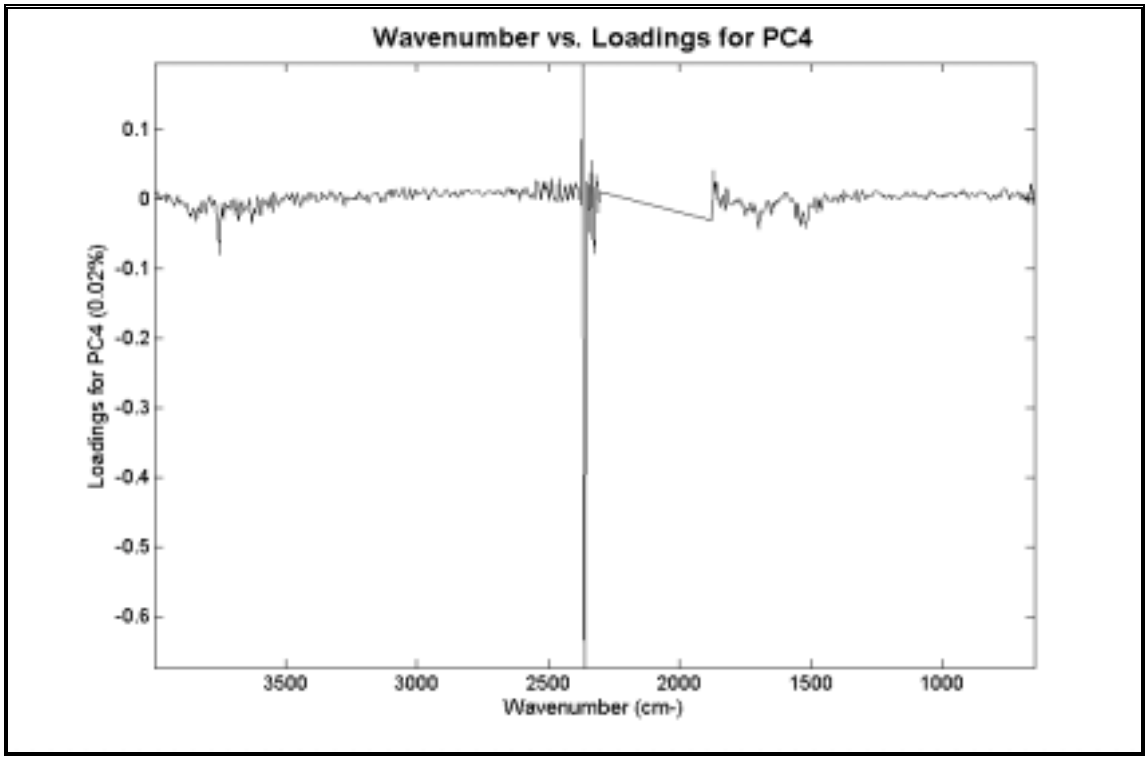


Figure 18 Loadings for the Fourth PC of the negative FTIR (I) reaction profile

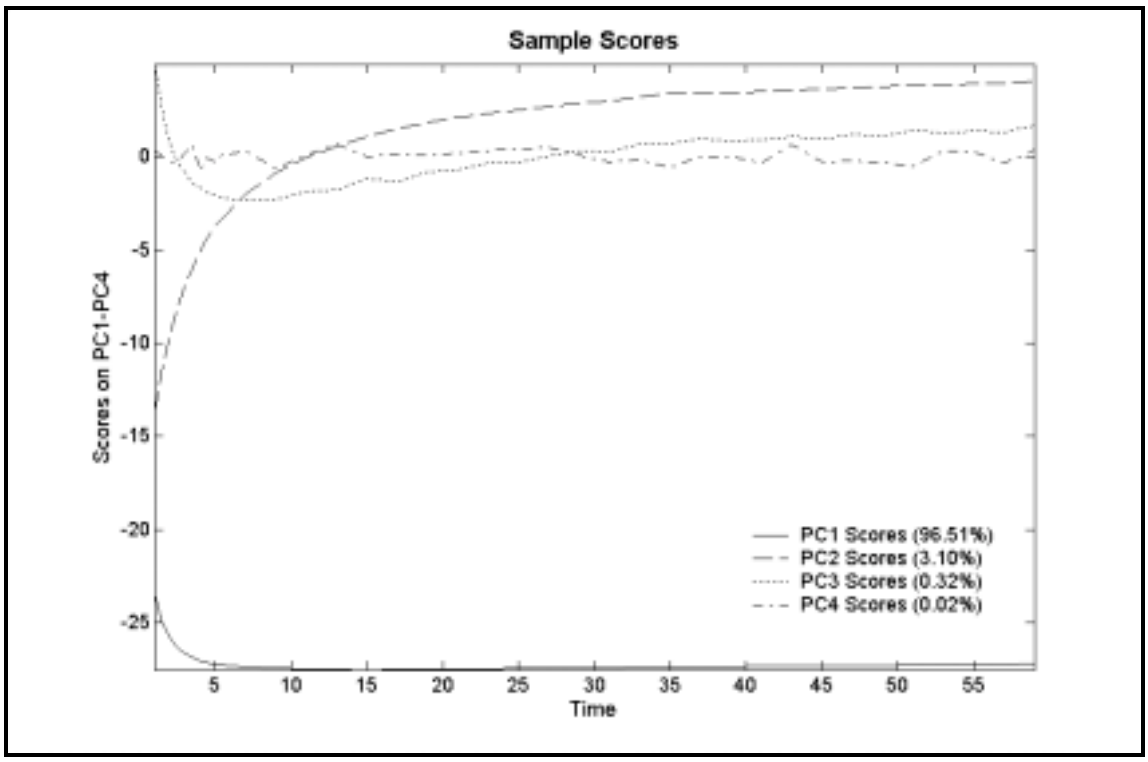


Figure 19. Scores on the First to Fourth PC of the negative FTIR(I) reaction profile

PCA analysis of the *non-negative* FTIR(II) reaction profile also revealed three independent components. The loadings patterns of the first principal component differed to the *negative* FTIR(I) loadings profile because the formic acid peaks were not present, shown in appendix 1.2.1. The first loading plot contained the characteristic functional group frequencies of amine and the mixture spectrum of triethylamine and formic acid, the second loading plot contained the common functional group frequencies of imine and amine and the third loading plot contained characteristic group frequencies of carbon dioxide, with a small contribution from imine and amine, as shown in appendix 1.2.1. The plot of each of the singular vectors versus time revealed evolutionary profiles similar to those described in the PCA analysis of the *negative* FTIR(I) reaction profile. Approximately ~99% of the variance was attributed to the first PC, ~0.2% was attributed the second PC and ~0.02% was attributed to the third component.

As noted above, three independent constituents were present in both the *negative* FTIR(I) and the *non-negative* FTIR(II) measurement matrices. All components apart from triethylamine were identified from the loadings plot. Triethylamine was not observed because of the severe overlap of (1) the asymmetric and symmetric stretch (2970-2950cm^{-1} and 2880-2860cm^{-1}) and bend (1470-1430cm^{-1} and 1380-1370cm^{-1}) of the CH methyl group, (2) the asymmetric and symmetric stretch (2935-2915cm^{-1} and 2865-2845cm^{-1}) and bend (1485-1445cm^{-1}) of the CH methylene group, and (3) the CH_3-C functional group frequency (3000-2860cm^{-1}) in triethylamine, with the methyl groups present in imine and amine. Severe overlap of the characteristic CN stretch (1210-1150cm^{-1}) of the tertiary amino present in triethylamine with both the aryl-O stretch (1270-1230cm^{-1}) present in imine and amine and the C-O (1300-900cm^{-1}) present in imine, amine and formic acid obscured the characteristic peaks of triethylamine.

The scores and loadings profiles obtained from PCA are abstract representations of the concentration profiles and spectral profiles, whereas the goal of calibration free analysis is to produce the true concentration profiles and spectral profiles. Realistic starting estimates of the components was required for the initialisation of ALS.

II.1.3.5 Initial Estimates

SIMPLISMA

SIMPLISMA was used to determine the pure spectral profiles of the independent components. The inputs were the measurement matrix and δ, a noise correction factor. The value of δ varied between 1% and 15% of the maximum mean spectrum. Larger correction factors from (6-15%) were used because the solution obtained using a lower

correction 5% of μ_j were noisy. This did not have any noticeable effect on the solution.

The negative values in the *negative* FTIR(I) data were zeroed before SIMPLISMA was applied. This transformation was necessary because the 'mixture behaviour' utilised in the algorithm, assumes positive spectra and positive concentration [121]. However, this transformation was not necessary for the analysis of the *non-negative* FTIR(II) data.

Analysis of negative FTIR(I)

The first pure spectrum and standard deviation spectrum determined from the transformed *negative* FTIR(I) data using a 15% noise correction factor is given in figure 20. The standard deviation spectrum is plotted to give clues to the identification of the pure factor. Variables with a high standard deviation can be attributed to pure factors. The first purity spectrum is a linear combination of all the reaction constituents. The maximum intensity in the first purity spectrum is found at 1215.9cm^{-1} in the common group frequency range of imine and amine. Variables which were correlated to the first pure variable virtually disappear in the second standard deviation spectrum, shown in figure 21. The maximum intensity in the second purity spectrum is found at 2343.0cm^{-1} in the characteristic group frequency range of carbon dioxide, shown in figure 21. The maximum intensity in the third purity spectrum was 1165.7cm^{-1} and was located in the common group frequency range of imine and amine, shown in figure 22. The three purest variables selected were wavenumbers 1215.9cm^{-1}, 2343.0cm^{-1} and 1165.7cm^{-1}. Of the three variables chosen, variable one and variable three were chosen from regions of low selectivity in the spectral domain, whereas the second variable was chosen from a region of relatively high spectral selectivity corresponding to the asymmetric stretch of O=C=O.

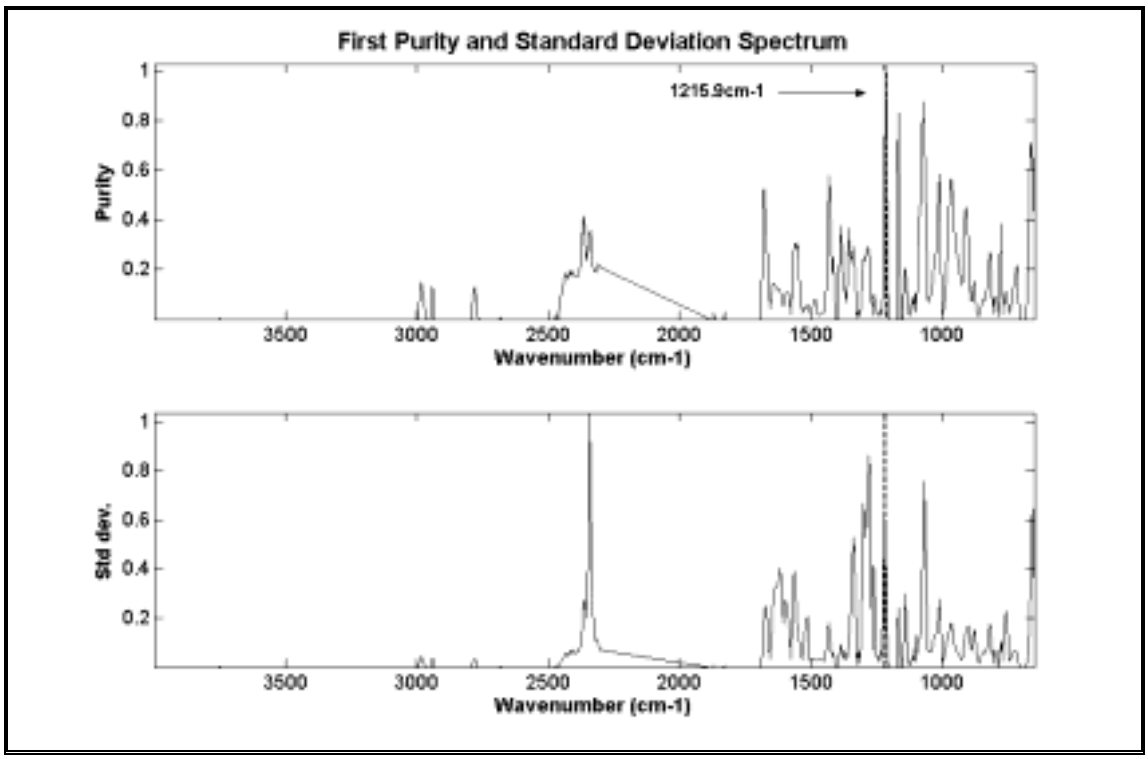


Figure 20. The first purity and standard deviation spectrum obtained from the transformed negative FTIR(I) reaction profile.

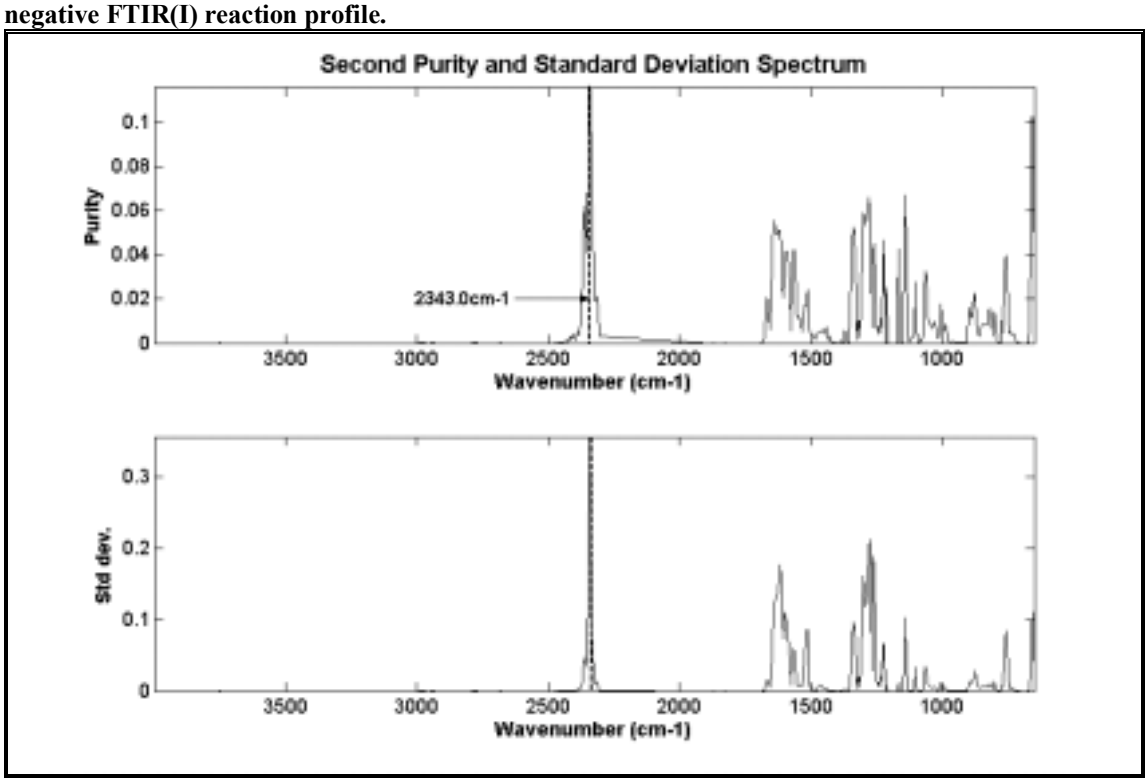


Figure 21. The second purity and standard deviation spectrum obtained from the transformed negative FTIR(I) reaction profile.

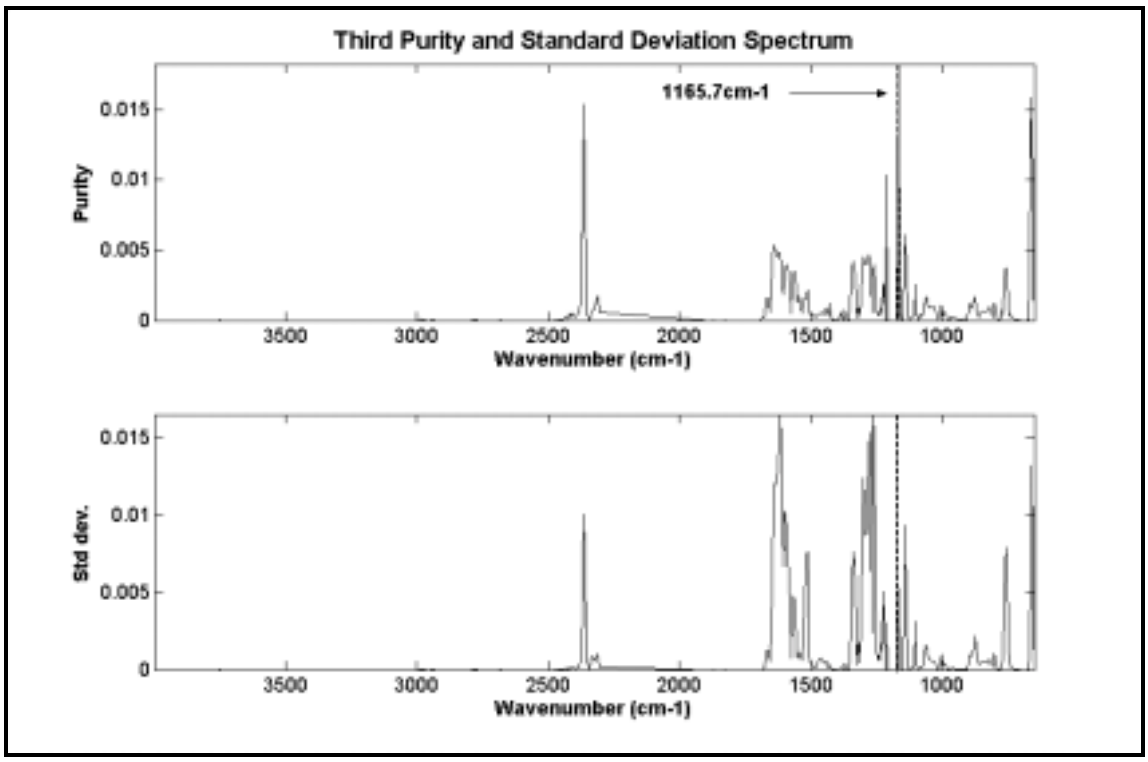


Figure 22. The third purity and standard deviation spectrum obtained from the transformed negative FTIR(I) reaction profile.

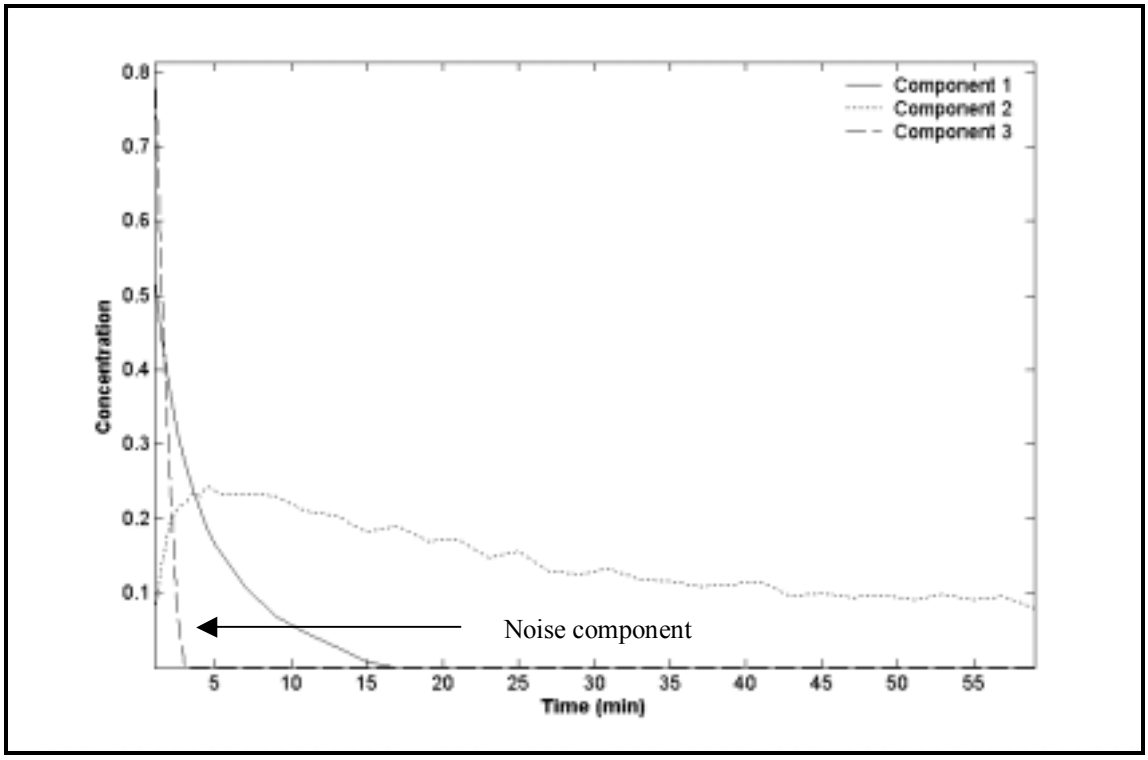


Figure 23. The SIMPLISMA concentration estimates obtained from the transformed negative FTIR(I) reaction profile.

The concentration estimates are shown in figure 23, they were plotted to establish whether the pure variables contained structured variance or were attributed to noise. The concentration profiles of the first component decreased with time, and the majority of change occurred between 0-15 minutes. This profile was representative of a reagent. The second concentration profile increased for the first 5 minutes and slowly declined over the remaining time. This was indicative of an intermediate component. The third concentration profile decreased with respect to time and the greatest change occurred over the first 2 minutes of the reaction before levelling off. This component was attributed to noise, therefore, increasing the noise correction factor to 15% did not improve the solution.

Analysis of non- negative FTIR(II)

SIMPLISMA was applied to the non-negative FTIR(II) dataset using a noise correction factor of 15%. The maximum intensity of the first purity spectrum was 1069.2cm^{-1}, chosen in the characteristic group frequency range of amine. The corresponding concentration profile was smooth and representative of a reagent which was consumed throughout the reaction. The standard deviation spectrum also contained characteristic group frequencies of amine. The second variable was 2343cm^{-1} selected in the characteristic group frequency range of the asymmetric stretch of O=C=O. The corresponding concentration profile was relatively smooth and represented an intermediate component. The third variable selected was 2362.3cm^{-1} which had a very low standard deviation value. The corresponding concentration profile was rugged which confirmed that this factor was a noise component.

To reduce the effect of noise in the SIMPLISMA analysis, the analysis was repeated on both reaction profiles using correction factors of up to 15% to compensate for low

intensity variables (appendix). However, this did not improve the selection of key factors from the measurement matrices. The dimension of the measurement matrices (*negative* FTIR(I) and the *non-negative* FTIR(II)) were not reduced prior to selection because the pure variables were selected from within the fingerprint region of the FTIR spectrum.

Through this analysis it was found that the concentration profiles resolved from the *negative* FTIR(I) and the *non-negative* FTIR(II) datasets were not appropriate for MCR-ALS analysis because in each case a noise component was resolved rather than the evolutionary profiles expected for the reaction constituents.

In the next section, EFA was applied to the *negative* FTIR(I) and *non-negative* FTIR(II) dataset to extract the evolutionary profiles of the three independent components from the reaction profiles, based upon local rank analysis of successive sub-matrices to obtain better initial estimation of the independent components.

EFA

Analysis of negative FTIR(I)

EFA was applied to the *negative* FTIR(I) dataset. The inputs required for the data analysis were (1) the measurement matrix, (2) the number of rows in the dataset ,and (3) the number of components (factors). There were 37 rows and three components.

The $\log_{10}$ of the forward and backward EFA singular values is shown in figure 24. The forward EFA analysis revealed two components in the first 1.5 minutes and a third component appeared at 2 minutes.

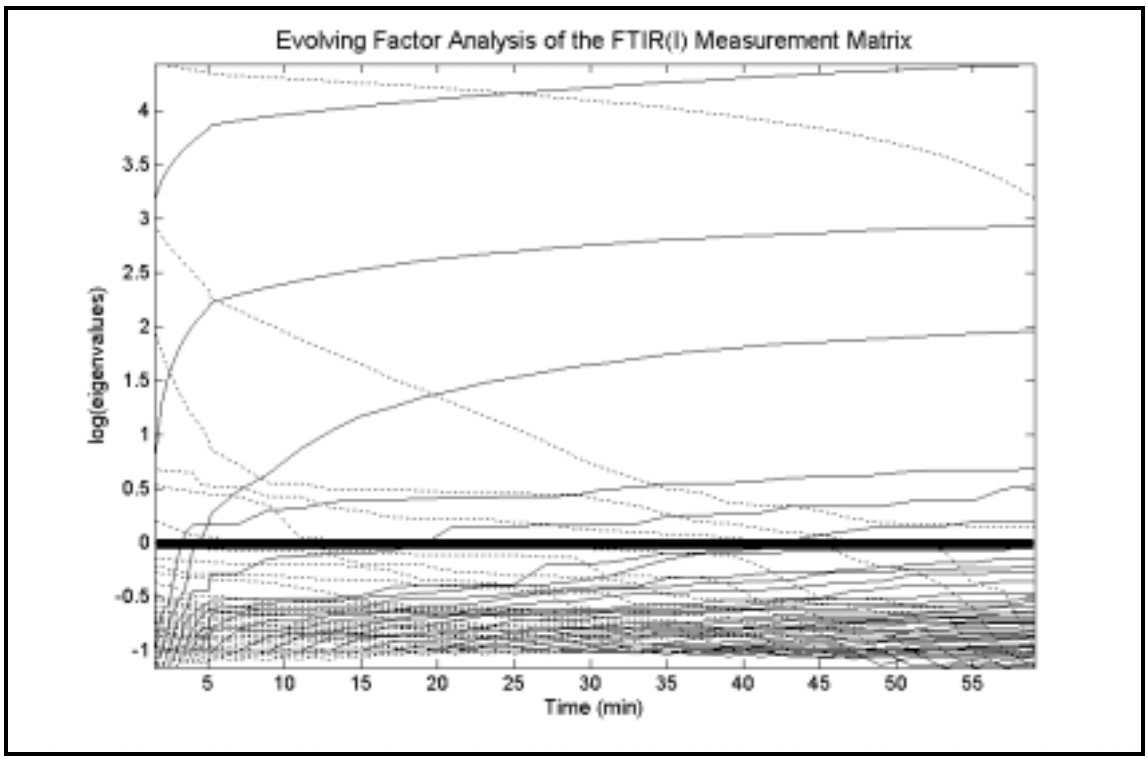


Figure 24. Forward and backward EFA analysis of negative FTIR(I) reaction profile.

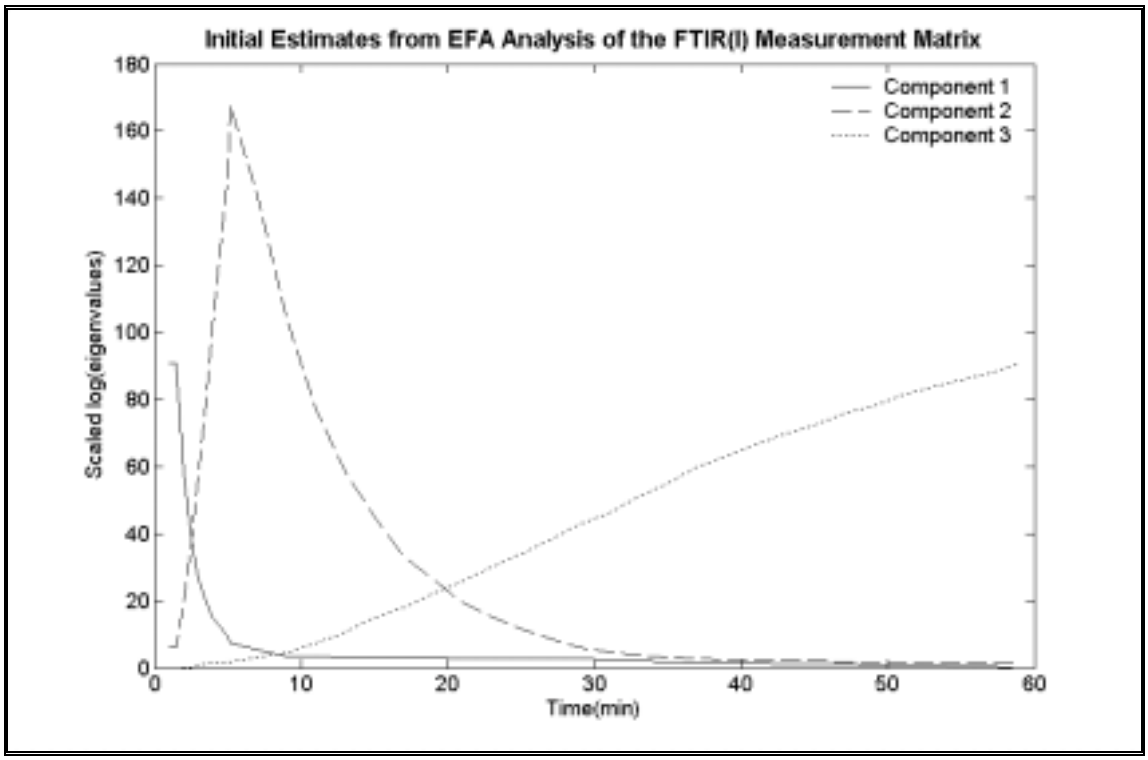


Figure 25. Singular vectors determined by EFA of negative FTIR(I)-TEAF reaction profile.

The backward analysis showed the disappearance of component one at the 57th minute, components two and three were present throughout the remaining reaction time. The

arranged and cleaned profiles are shown in figure 25. In the course of the reaction there was no time where either no components or just one component was present. However, a zero component constraint could be applied in the constrained ALS procedure because component three was not present in the first two sampled time points.

Analysis of non-negative FTIR(II)

The EFA analysis was completed using the same inputs as the *negative* FTIR(I) data. The evolutionary profiles resolved resembled a reagent, product and intermediate species. The time profiles (window of existence) were similar to the *negative* FTIR(I) data.

Generation of Needle Initialisation Spectra

Needle spectra for use in MCR-ALS were generated through a correlation analysis of MCR-ALS results obtained using EFA initialisation. The Pearson correlation coefficients (of the concentration profiles determined from MCR-ALS with EFA initialisation) were calculated with the column data at every measured wavelength which had a correlation $r^2 > 0.800$, a value of one was set in the needle spectra for that component. This technique was used to generate needle spectrum for initialisation of a second MCR-ALS step by following this procedure. Wavelengths in the original measurements that tended to respond uniquely for pure concentration components were identified and included in their corresponding needle spectra.

Analysis of negative FTIR(I)

The needle spectrum of the first component represented imine because characteristic and common group frequencies associated with imine predominated the spectrum. The

spectrum of the second needle profile contained four groups of needles; two groups were located in the characteristic functional group frequencies of carbon dioxide, and the second two groups were located in common functional group frequencies of imine and amine. The third needle spectrum contained characteristic group frequencies associated with amine, formic acid and carbon dioxide.

Analysis of non-negative FTIR(II)

The first needle spectrum contained common group frequencies of imine and characteristic functional group frequencies of imine and formic acid. The second needle spectrum contained characteristic functional group frequencies of carbon dioxide and the third needle spectrum contained common group frequencies of amine and characteristic group frequencies of amine and formic acid.

II.1.3.6 MCR-ALS Analysis of the FTIR data

Data Analysis

The starting estimates determined from EFA and the needle spectra were used to initialise the ALS procedure. A flow chart of the optional constraints in the multivariate analysis is given in figure 26 and a list of the experiments completed using the different initial estimates and constraints for both datasets are tabulated in tables 2-4.

The three optional constraints applied in these experiments were; non-negativity in the concentration profiles, non-negativity in the spectral profiles and equality in the concentration profiles. The spectral profiles in each of the experiments were normalised to height (see equation 14). The error was measured using the LOF between successive iterations, the convergence criteria was 0.1% and the maximum number of iterations was set to 100.

Validation

The FTIR samples measured at 1, 2, 5, 15, and 45 minutes, which coincided with the HPLC samples were used to validate the MCR-ALS resolution. The reaction samples were scaled between zero and one in the Root Mean Square Prediction Error (RMSPE) calculation.

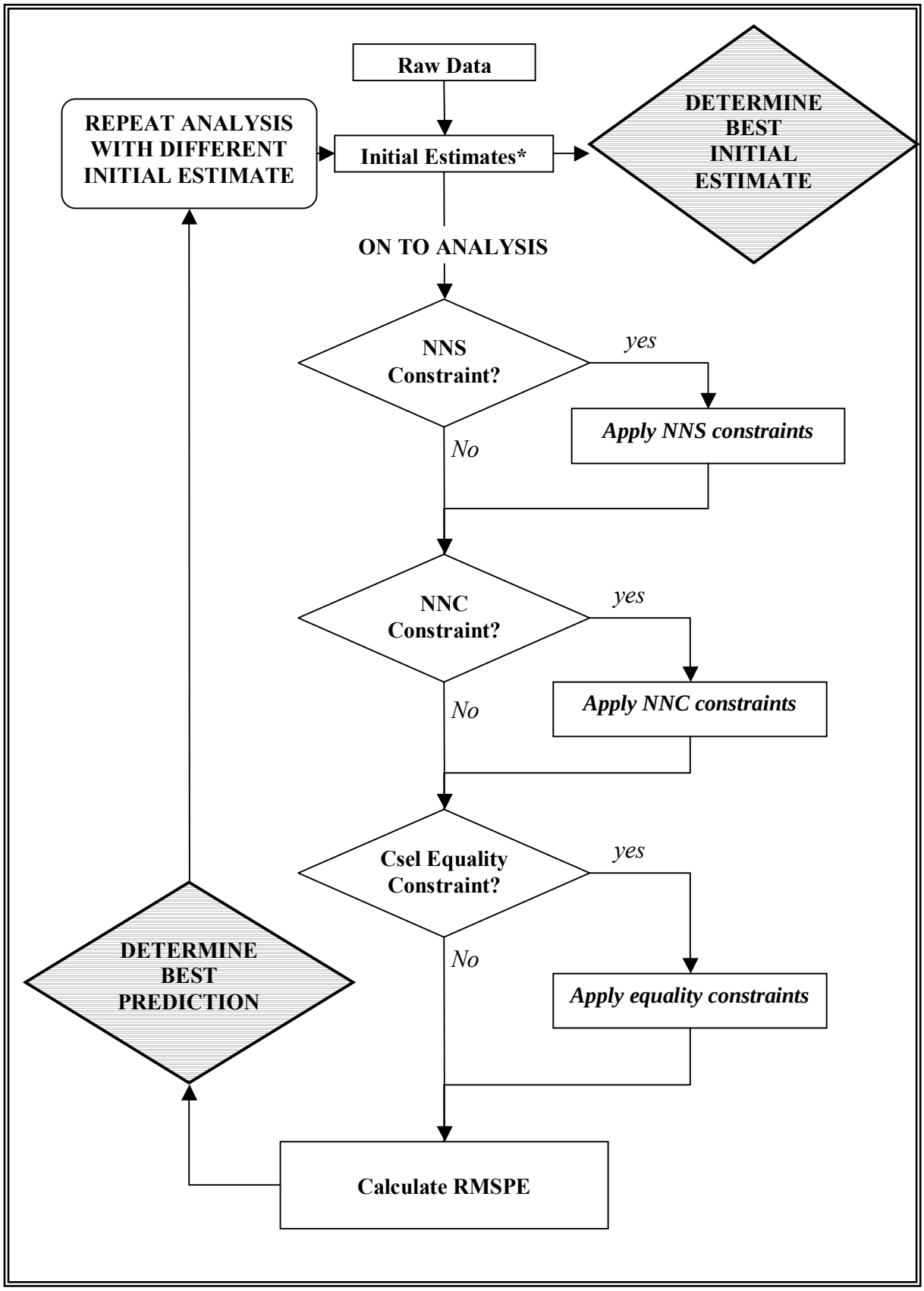


Figure 26. Flow Chart of Multivariate Analysis
***Initial Estimates: a)EFA Evolutionary profiles b)Needle Spectra obtained from correlation constraint c) Reference Spectra**

Constraints: NNS – Non-negativity constraint spectra, NNC – Non-negativity constraint concentration, Csel – Equality constraint concentration

Expt. No.	Data	Starting Estimates	Concentration	Spectra (Normalised)
1	FTIR (I)	EFA	None	None
2	FTIR (I)	EFA	NNC	None
3	FTIR (I)	EFA	NNC and Csel	None
4	FTIR (II)	EFA	None	None
5	FTIR (II)	EFA	NNC	NNS
6	FTIR (II)	EFA	NNC and Csel	NNS

Table 2. Exploratory analyses completed on FTIR(I) and FTIR(II) datasets using no *a priori* knowledge in the MCR-ALS resolution.

Expt. No.	Data	Starting Estimates	Concentration	Spectra (normalised)
7	FTIR (I)	Needle Estimate	None	None
8	FTIR (I)	Needle Estimate	NNC	None
9	FTIR (II)	Needle Estimate	None	None
10	FTIR (II)	Needle Estimate	NNC	NNS

Table 3. Exploratory analyses completed on FTIR(I) and FTIR (II) datasets using no *a priori* knowledge in the MCR-ALS resolution.

Expt. No.	Data	Starting Estimates	Concentration	Spectra (Normalised)
11	FTIR (I)	Pure Spectra (I,C,A)	None	None
12	FTIR (I)	Pure Spectra (I,C,A)	None	NNS CO_2
13	FTIR (I)	Pure Spectra (I,C,A)	None	NNS CO_2 and Imine
14	FTIR (I)	Pure Spectra (I,C,A)	None	NNS CO_2 and Amine

Table 4. Exploratory analyses completed on FTIR (I) using *a priori* knowledge in the MCR-ALS resolution. I(Imine), C(Carbon dioxide), A(Amine).

II.1.3.6.1 MCR-ALS Analysis of the FTIR measurement matrices using exploratory profiles

EFA Initialisation

Experiments 1-3 in table 5 gives the results of the MCR-ALS analysis of the *negative* FTIR(I) dataset, initialised from the EFA concentration estimates. The MCR-ALS experimental design for these experiments are given in table 2. In each of the experiments the RMSPE of the imine was ~0.10 and the RMSPE of amine was ~0.30. Neither the imine concentration profile nor the amine concentration profile were predicted particularly well. Addition of the non-negativity constraint in the concentration profiles showed no real improvement to the results, i.e., no active constraints were present in the solution, although the number of iterations required for convergence increased. It is most likely that the addition of the constraint moved the solution further away from the local minima, thereby increasing the time for convergence. The addition of the zero component constraints, increased the LOF, but the number of iterations required for convergence decreased in comparison to experiment 2. This suggests that the constrained solution was closer to a local minima, but the increased LOF suggests that the proposed model was incorrect. The resolved spectral profiles of the three components using EFA did not exhibit good agreements with the expected spectral profiles. This is because the predicted imine and amine spectrum contained the characteristic formic acid peaks. The predicted carbon dioxide profile contained characteristic vibrational frequencies from both formic acid and amine. Each of the predicted spectral profiles tended to be highly correlated to formic acid because of linear dependency amongst the constituents. Similarly in the MCR-ALS analysis of the *non-negative* FTIR(II) data, (the MCR-ALS experimental design is

given in table 2 (experiments 4-6) and the results for these experiments are given in table 5), each of the predicted spectra contained the characteristic and common group frequencies of imine, amine and carbon dioxide.

Expt. No.	Data	Initial Estimates	No iterations	LOF (%)	RMSPE Imine	RMSPE Amine
1	FTIR(I)	EFA	2	2.5	0.13	0.34
2	FTIR(I)	EFA	26	2.5	0.12	0.35
3	FTIR(I)	EFA	6	4.9	0.15	0.35
4	FTIR(II)	EFA	2	0.7	0.09	0.35
5	FTIR(II)	EFA	2	0.7	0.08	0.33
6	FTIR(II)	EFA	100	0.9	0.03	0.34

Table 5. Results of the MCR-ALS analysis using the EFA starting estimates.

The probable reason for the large difference between the predicted concentration and expected concentration and the predicted spectral profiles and expected spectral profiles was due to the highly collinear EFA starting estimates.

Needle Spectra Initialisation

The results of the MCR-ALS analysis of the *negative* FTIR(I) and the *non-negative* FTIR(II) datasets are given in table 6 (see experiments 7-10). The concentration profiles resolved from the MCR-ALS analysis of the *negative* FTIR(I) dataset without constraints is shown in figure 27. The concentration profiles were comparable to the HPLC data (see experiment 7). The use of the needle spectral estimates to initialise the ALS optimisation noticeably reduced the RMSPE of imine and amine. Despite the improved selectivity of the needle spectral estimates compared to SIMPLISMA and EFA starting estimates, the resolved spectral profiles of imine and amine each contained identical characteristic and common functional group frequencies attributed to imine, amine and formic acid, see appendix 1.3.1. The resolved spectral profile of carbon dioxide also contained contribution from imine and amine contained as well as the correct functional group frequencies, i.e., the asymmetric stretch and degenerate

bending of O=C=O. The addition of the non-negativity constraint in the concentration profiles (experiment 8) did not improve the solution.

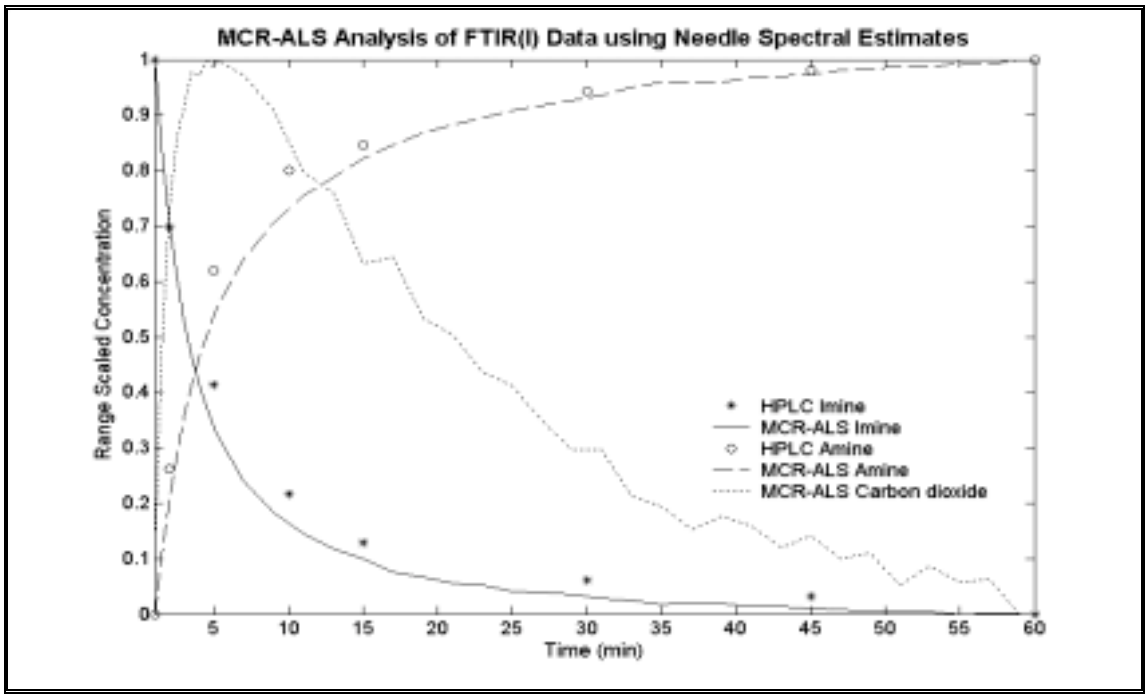


Figure 27. Concentration profiles resolved from the MCR-ALS analysis of the FTIR(I)-TEAF measurement matrix using needle spectral estimates (Expt. 7), see appendix 1.3.1 for spectral profiles

Similarly, the results of the MCR-ALS analysis of the *negative* FTIR(II) dataset produced excellent concentration profiles, comparable to the HPLC data, (see experiment 9). However, each spectral profile contained contributions from imine, amine and carbon dioxide, see appendix 1.3.2 for concentration and spectral profiles. The addition of the non-negativity constraint in both the concentration and spectral direction only increased the RMSPE of amine (see experiment 10). The best solution; in terms of the lowest RMSPE values were obtained using experiment 7 and 9. The higher LOF was due to incorrectly predicted spectral profiles.

Expt. No.	Data	Initial Estimates	No iterations	LOF (%)	RMSPE Imine	RMSPE Amine
7	FTIR(I)	Needle	3	2.5	0.04	0.05
8	FTIR(I)	Needle	9	6.2	0.05	-
9	FTIR(II)	Needle	2	0.7	0.03	0.06
10	FTIR(II)	Needle	10	0.9	0.03	0.34

Table 6. Results of the MCR-ALS analysis using the needle spectral estimates

Summary

Overall the concentration profiles resolved from MCR-ALS analysis of the *negative* FTIR(I) and *non-negative* FTIR(II) datasets were agreeable to the HPLC results using the needle spectra initialisation with no constraints applied during the ALS optimisation (see experiments 7 and 9, table 6). However, the resolved spectral profiles did not exhibit good agreement with the expected profiles. Therefore, the results from MCR-ALS analysis could not be used as an alternative to the HPLC analysis for either datasets because the spectral profiles resolved were incorrect.

The spectra and concentration profiles resolved using the EFA initial estimates exhibited differences to the expected profiles. This was because the starting estimates did not approximate the true solution. The SIMPLSMA approach was found to be unreliable because noise components were resolved in each case, i.e., with the *negative* FTIR(I) data and the *non-negative* FTIR(II) data rather than the structured independent components. The SIMPLISMA approach was also observed to be unreliable by Chew *et al.* [74], particularly when a large but unknown number of species were present and their component spectra were highly overlapping. They attributed this to the conditions of resolution which assume the number of observable species present and the so-called "pure-wavelength" for each species.

The *non-negative* FTIR(II) dataset which was acquired in order to increase the number of constraints in the MCR-ALS analysis, i.e., the addition of the non-negativity constraint in the spectral dimension, did not result in any real improvement. Therefore, no further analysis was completed with this dataset.

In the next section the pure spectra of imine, amine and carbon dioxide were used to initialise the MCR-ALS procedure, because it has been observed that favourable initial

estimates, i.e., those which approximate the true solution offer great assistance to cause the concentration and spectral vectors to converge fast and uniquely [74, 81].

II.1.3.6.2 MCR-ALS Analysis of the *negative* FTIR(I) measurement matrix using *a priori* knowledge

Excellent resolution of the concentration and spectral profiles were obtained using the neat spectrum of imine, amine and carbon dioxide to initialise the ALS procedure, when no constraints were applied in either the concentration or spectral direction (see table 7, experiment 11, and figure 28) and when only the non-negative constraints were applied in the spectral profile of carbon dioxide (see experiment 12).

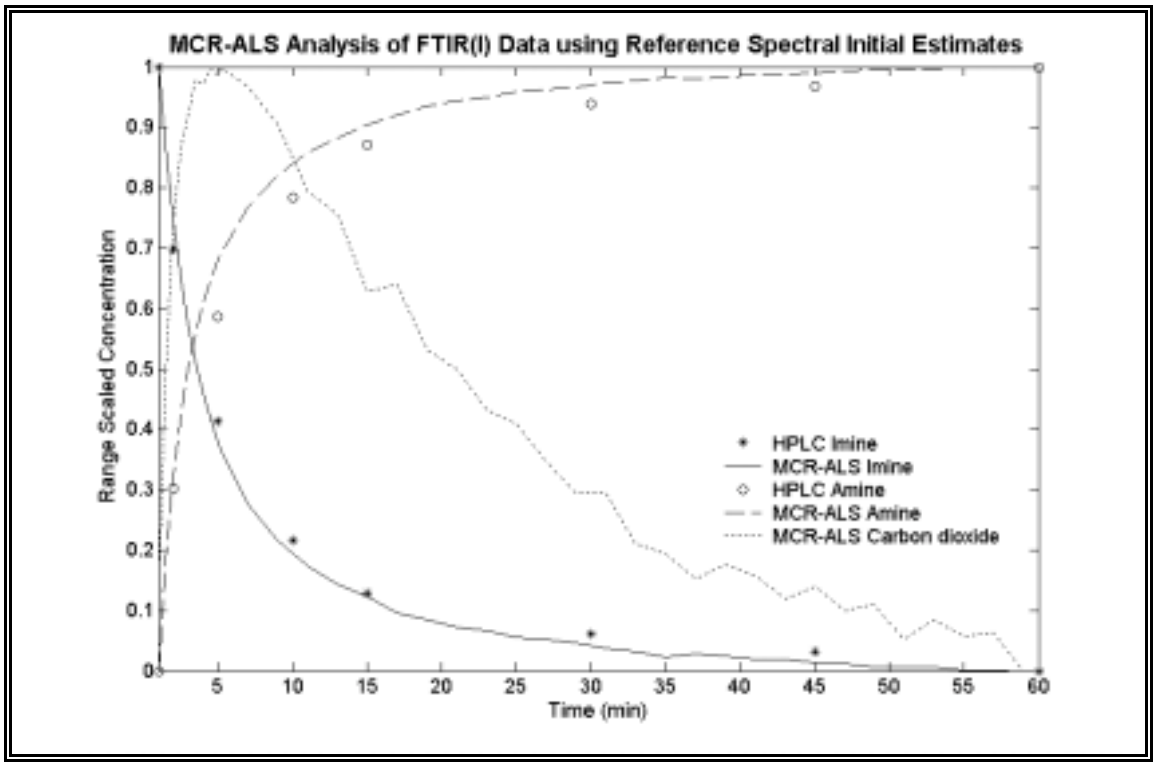


Figure 28. Concentration profiles resolved from the MCR-ALS analysis of the negative FTIR(I) measurement matrix using the reference spectra (experiment 11).

The predicted spectral profiles resolved from MCR-ALS using the resolution conditions from experiments 11 and 12 were comparable to 1) the reference spectra of imine, 2) amine with significant contamination with formic acid, and 3) carbon dioxide.

However, there were a small number of spectral differences between the neat and predicted spectral profiles.

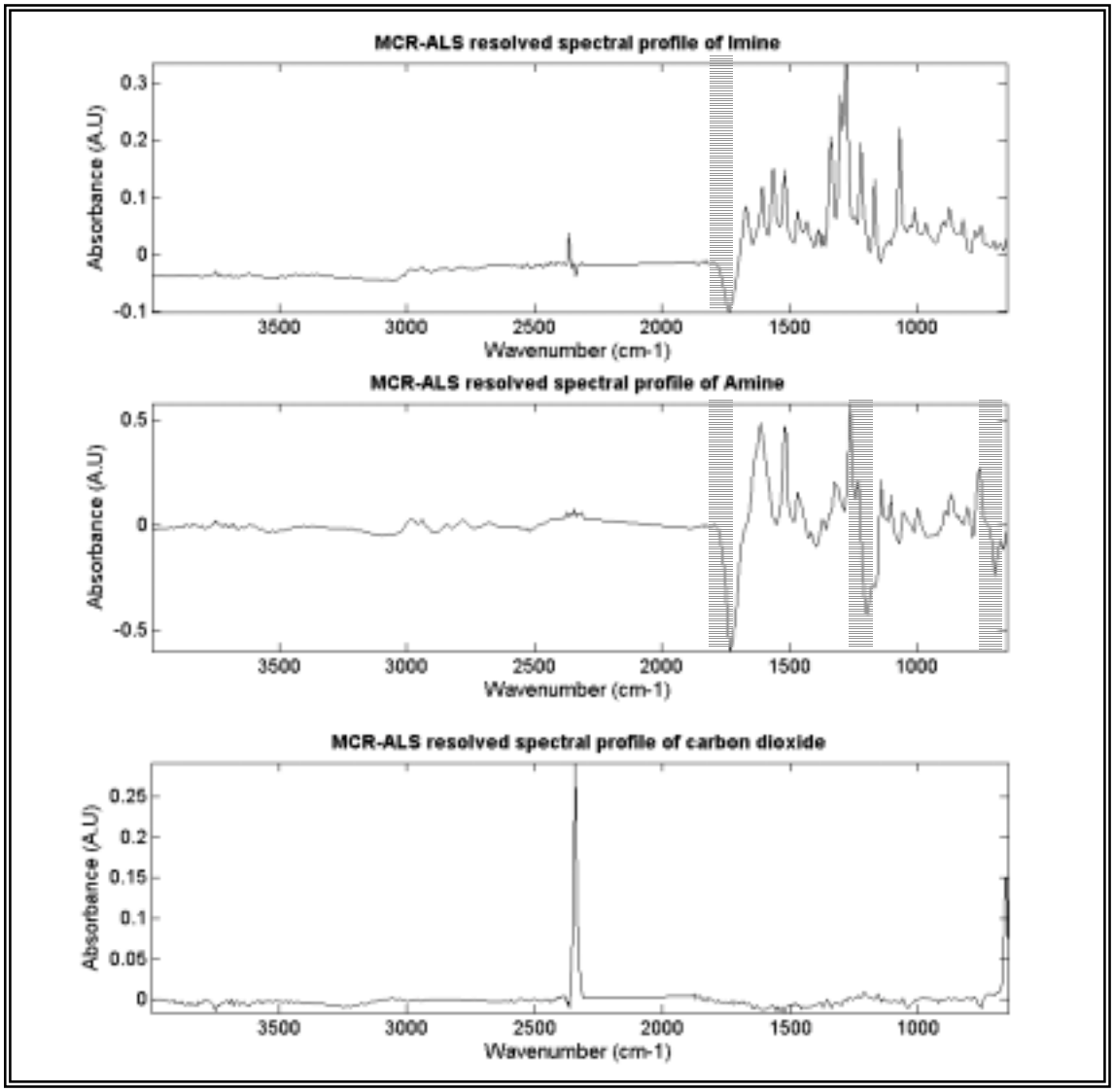


Figure 29. MCR-ALS resolved spectral profile of imine, amine and carbon dioxide. No constraints applied in the spectral or concentration direction, see experiment 11.

The resolved spectrum of imine contained a negative peak between 1800-1750cm^{-1}, which is indicative of a carbonyl or imino group, highlighted in figure 29. To determine whether this peak was from formic acid the characteristic functional group frequencies of the C-O-H stretch and O-H deformation were searched for in the resolved imine spectrum. As peaks attributed to the C-O-H stretch and O-H deformation were not

found, this extra peak was attributed to slight contamination of formic acid in the spectrum.

The resolved spectrum of amine is linearly combined with formic acid because it contained negative absorbance at the characteristic vibrational frequencies of formic acid, i.e., C-O, C-O-H, and O-H groups, 1780-1680cm^{-1}, 1230-1140cm^{-1} and 750-650cm^{-1} respectively, highlighted in figure 29. Formic acid and amine are collinear as formic acid is consumed (decreased negative signal as the reaction proceeds) at the same rate that amine is produced, and as a result the concentration profiles could not be mathematically separated. The concentration profile of amine was comparable to the assay data collected and the concentration profile of formic acid was resolved. The resolved spectral profile of carbon dioxide contained all the characteristic functional group frequencies, although slight differences in the solution persisted between 3450-3100cm^{-1} and 1800-800cm^{-1}.

Experiments 13 and 14 showed no improvements over the results obtained using no constraints in the ALS resolution or when non-negative constraints were applied in the spectral profile of carbon dioxide (see experiments 11 and 12). In fact, the LOF appreciated substantially in both the MCR-ALS solution using the non-negative constraint in the spectral profiles of carbon dioxide and imine with slight contamination from formic acid (experiment 13) and the MCR-ALS solution using the non-negative constraints in the spectral profiles of carbon dioxide and linear combined amine and formic acid (experiment 14), caused by the truncation of the formic acid peaks during the ALS optimisation.

Using MCR-ALS with the resolution conditions of experiments 11 and 12 it was possible to successfully resolve the concentration profiles of imine and linear combined

amine and formic acid to provide an alternative approach to HPLC analysis, as well as the concentration profiles of carbon dioxide which could not be realised from the chromatographic analysis.

Expt. No.	Data	Initial Estimates	No iterations	LOF (%)	RMSPE Imine	RMSPE Amine
11	FTIR(I)	Pure spectra	3	2.5	0.02	0.04
12	FTIR(I)	Pure spectra	2	4.6	0.03	0.03
13	FTIR(I)	Pure spectra	2	19.9	0.04	0.05
14	FTIR(I)	Pure spectra	2	56.9	0.03	0.01

Table 7. Results of the MCR-ALS analysis of the negative FTIR(I) dataset using the neat spectra of imine, amine and carbon dioxide.

II.1.4 Conclusion

The combination of in situ FTIR with MCR has been used successfully to determine the spectral and concentration profiles of imine, linear combined formic acid and amine, and carbon dioxide in the complex CATHy reaction using the resolution conditions found in experiments 11 and 12. The resultant pure spectra and concentration profiles of the analytes of interest ensured that this technique could be used as an economically viable and convenient replacement to the currently used HPLC method. The additional information from MCR with respect to the carbon dioxide could be used to notify batch operators of dangerous levels of carbon dioxide in the reactor. The FTIR method developed enabled the reaction to be monitored in situ and eliminated the need for constant sampling. The combined approach offers significant advantages in cases where viewing complex experimental data by eye is problematic.

Future Work

Future work would include breaking the rank deficiency of the dataset through chemical perturbation of the system and partial kinetic modelling of the data could be applied to give clues about the reaction mechanism.

II.2 Multi-way Penalty Alternating Least Squares

II.2.1 Introduction

In this study an extension to the P-ALS approach [72] is introduced, called multi-way P-ALS (NWAY P-ALS)[149], in which optionally hard constraints or soft constraints are applied. The novel aspect of this work is that it is the first time penalty ALS constraints are used in conjunction with multi-way constraints to enable the application of soft constraints during the multi-way P-ALS analysis. The limitation of P-ALS approach is that when multiple batches are available; it is not possible to take advantage of the optional multi-way constraints that can be applied in either the spectral or concentration profiles.

In the NWAY P-ALS routine a row-wise penalty least squares function is used to implement the constraints. The advantage of this technique is that it incorporates all the advantages of the P-ALS approach. In addition, species common to two or more experiments gain a second order advantage, especially if there are more active constraints in the multiple experiments [111]. The multi-way constraints that can be employed include i) common species must have the same spectrum in all matrices ii) common species must have concentration profiles with equivalent shape in all data matrices iii) one or more regions of zero concentration components may be present [111]. A detailed description of the NWAY P-ALS approach employed herein is given below. Note this approach combines the P-ALS method outlined in ref. [111] with multi-way constraints. This is a novel marriage of algorithms and the view taken is to determined whether it is a useful and feasible methodology.

II.2.2 Methodology

A dataset sized $n \times m \times t$ with at least one order in common, can be unfolded in three different directions: along the row space (n), along the column space (m) and along the third direction of the cube (t), also called the tube space.

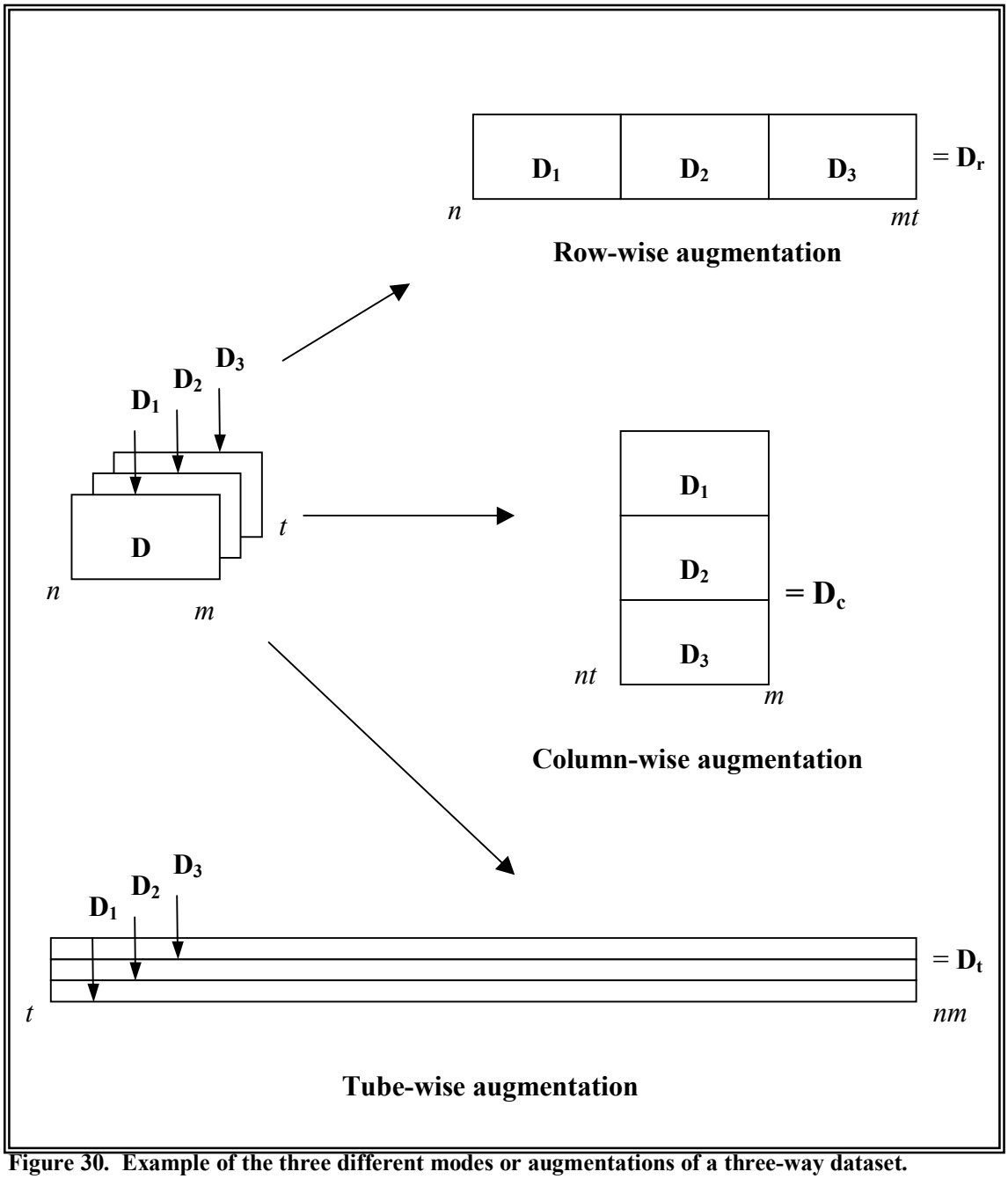


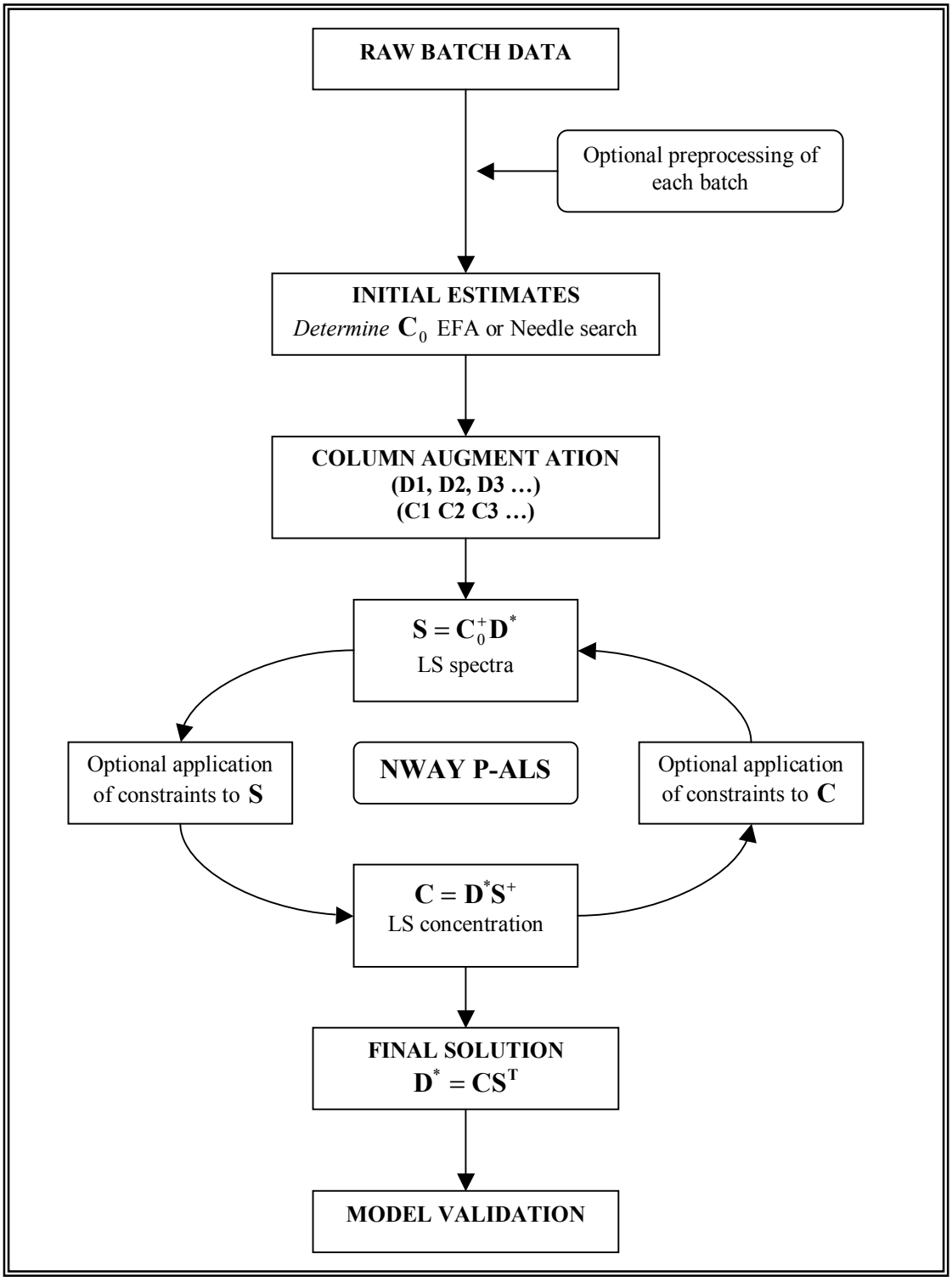
Figure 30. Example of the three different modes or augmentations of a three-way dataset.

The three unfolding procedures give a row-wise augmented matrix, $\mathbf{D}_r(n \times mt)$, a column-wise augmented matrix, $\mathbf{D}_c(nt \times m)$, and a tube-wise augmented matrix, $\mathbf{D}_t(t \times nm)$, respectively, shown in figure 30. In typical spectroscopic applications, rows in the data matrix represent mixture spectra recorded as a function of time (n), columns represent wavelengths (m) and tubes represent different experiments (t).

In chemistry the most common structure of a three-way dataset is bilinear, rather than trilinear. A three-way dataset can be classified as trilinear only if the dataset can be represented as the sum of the product of triads of vectors in rows, columns and tubes. This implies that the set of basic vectors or rows, columns and tubes have equal rank. Normally multi-way chemical data do not have equal rank in each dimension owing to the underlying chemical process. For example, a chemical process monitored by NIR at several different experimental conditions may have different shaped kinetic profiles which increase the rank of the row matrix. Alternatively, the instrument may have sample-to-sample variation in the response profiles, such as retention time shifts or shape changes in different HPLC-DAD runs [150].

Row-wise augmentation can be applied if two or more types of measurement are made in which the row space is common, such as simultaneous spectroscopic acquisition from multiple detectors. Column-wise augmentation can be applied when multiple batches have a common column space, i.e., multiple batches run of the same IR spectrometer. Here, the NWAY P-ALS function is presented for the column-wise augmentation of the measurement matrices. The bilinear decomposition of the augmented matrix, $\mathbf{D}_c$, would result in an augmented concentration matrix, shown in figure 31a. Each matrix contains the concentration profile of the common species in each batch, $\mathbf{C}_c$, and a

common spectral matrix, **S**. This leads to an ALS algorithm with the following two steps, a) and b), and is diagrammatically outlined in box 2.



Box 2. Flow chart of NWAY P-ALS methodology

a) Given some initial or intermediate estimate of $\mathbf{C}_c$ for every *j*, find $\mathbf{s}_j$ such that $\mathbf{s}_j$ minimises $\left\|\mathbf{d}_{cj} - \mathbf{C}_c\mathbf{s}_j^{\mathrm{T}}\right\|$ subject to constraints on $\mathbf{s}_j$, such as $\mathbf{s}_j = \mathbf{g}_j$, where $\mathbf{g}_j$ is a vector of constraints, defined later.

b) Given some least squares estimate of $\mathbf{S}^{\mathrm{T}}$ for every *i*, find $\mathbf{c}_{ci}$ such that $\mathbf{c}_{ci}$ minimises $\left\|\mathbf{d}_{ci} - \mathbf{S}\mathbf{c}_{ci}^{\mathrm{T}}\right\|$ subject to constraints on $\mathbf{c}_{ci}$, such as $\mathbf{c}_{ci} = \mathbf{g}_i$

Steps a) and b) are repeated in an iterative fashion to alternatively constrain **S** and $\mathbf{C}_c$ until the starting solutions converge smoothly and monotonically to the desired result. The row-wise fitting algorithm for finding rows of **S** and $\mathbf{C}_c$ is illustrated in figures 31 a-b for the column-wise augmentation.

The least squares problem $\mathbf{d}_{cj} = \mathbf{C}_c\mathbf{s}_j^{\mathrm{T}}$, is solved for $\mathbf{s}_j^{\mathrm{T}}$ with the adaptation of the equality constrained weighting method described by Lawson and Hanson [104, 106], using a vector of constrained values and a constraint matrix **g** and **H** respectively. The model is written as shown in equation 38;

$$\begin{bmatrix}\mathbf{d}_{cj}\\ \varphi\mathbf{g}\end{bmatrix} \cong \begin{bmatrix}\mathbf{C}_c\\ \varphi\mathbf{H}\end{bmatrix}\mathbf{s}_j^{\mathrm{T}} \qquad \textbf{Equation 38}$$

where, $\mathbf{d}_{cj}$ is a column vector of the augmented measurement matrix, $\mathbf{C}_c$ is the initial estimate of the concentration profiles for each matrix and $\mathbf{s}_j^{\mathrm{T}}$ is an estimate of the absorptivity of each component at a specific wavelength.

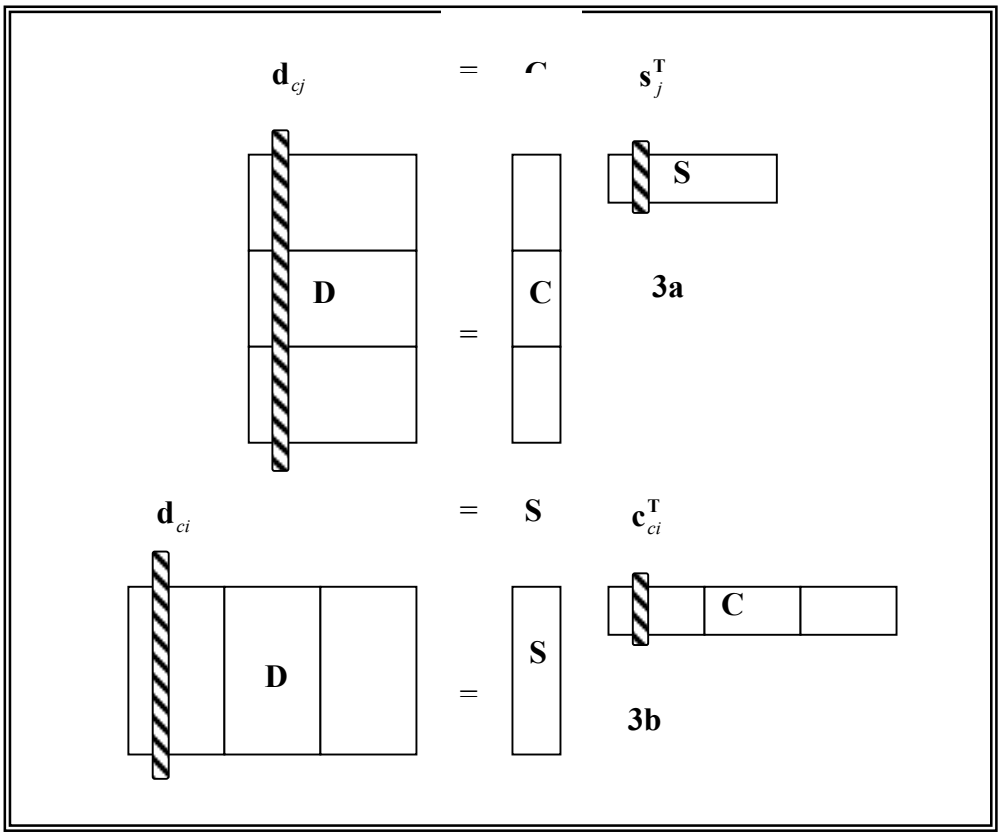


Figure 31a Schematic illustration of the row-wise fitting algorithm for finding rows of S. Figure 31b Schematic illustration of the row-wise fitting algorithm for finding rows of C.

The constraint matrix is a binary matrix of ones and zeros, which contain the coordinates for the application of the respective constraints. The number 1 specifies where the constraint is applied for that specific component and 0 where the constraint is not applied. The penalty factor weighting function, φ, can be adjusted; this allows flexibility in the implementation of the constraint, such that a reduced weight can be applied in the analysis of noisy data and a larger weight can be applied in the analysis of clean data. For data of different measurement scales and sizes, φ, can be adjusted relative to the norm of $\mathbf{C}_c$ i.e. $0.01\times norm(\mathbf{C}_c)$ for soft constraints or $10\times norm(\mathbf{C}_c)$, for hard constraints.

A detailed description of the application of the equality constraint, non-negativity constraints, unimodality constraints and closure constraint using the single matrix P-ALS is described in the references [72]. An example of how the non-negativity and

equality constraints are applied simultaneously in a multi-way case for a three component model is given below to show how it is possible to apply several types of constraints simultaneously due to the structure of the model. The unconstrained least squares solution is shown below for a column-wise augmentation of two matrices.

$$\mathbf{d}_{c1} = \begin{bmatrix} 0.7060 \\ 0.1365 \\ 0.4430 \\ 0.3932 \\ \hline 0.2775 \\ 0.2247 \\ 0.5989 \\ 0.0741 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \mathbf{d_1} \\ \\ \\ \\ \mathbf{d_2} \end{matrix} \qquad \mathbf{C}_c = \begin{bmatrix} 0.4451 & 0.8462 & 0.8381 \\ 0.9318 & 0.5252 & 0.0196 \\ 0.4660 & 0.2026 & 0.6813 \\ 0.4186 & 0.6721 & 0.3795 \\ \hline 0.1870 & 0.4796 & 0.2625 \\ 0.9913 & 0.4960 & 0.1863 \\ 0.7120 & 0.2875 & 0.9171 \\ 0.8714 & 0.0609 & 0.1233 \end{bmatrix} \begin{matrix} \\ \\ \\ \mathbf{c_1} \\ \\ \\ \\ \mathbf{c_2} \end{matrix} \qquad \hat{\mathbf{s}}_1^{\mathbf{T}} = \begin{bmatrix} -0.0159 \\ 0.2663 \\ 0.5819 \end{bmatrix}$$

The first column vector from the augmented matrix, $\mathbf{D}_c$ and the augmented matrix, $\mathbf{C}_c$ is used to solve the first column vector of $\mathbf{S}^{\mathbf{T}}$. Suppose an equality constraint is required in coefficients s_1 and s_3, such that $s_1 = g_1 = 0.0580$ and $s_3 = g_3 = 0.4821$ (values can be reference values or derived from an external reference method) and a penalty weighting function $\varphi = 10$. The weighted values are place in the constraint matrix $\varphi\mathbf{g} = \begin{bmatrix} 10 \times 0.0580 & 10 \times 0.4821 \end{bmatrix}$ and the penalty value is placed in the constraint matrix **H**, where the constraint is to be applied as given below. As a positive equality coefficient is applied in s_1, there is no need for a non-negativity constraint.

$$\mathbf{d}_{c1} = \begin{bmatrix} 0.9218 \\ 0.7382 \\ 0.1765 \\ 0.4057 \\ \hline 0.5534 \\ 0.2920 \\ 0.8580 \\ 0.3358 \\ \hline 0.5800 \\ 4.8210 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \} \varphi \mathbf{g} \qquad \mathbf{C}_c = \begin{bmatrix} 0.4451 & 0.8462 & 0.8381 \\ 0.9318 & 0.5252 & 0.0196 \\ 0.4660 & 0.2026 & 0.6813 \\ 0.4186 & 0.6721 & 0.3795 \\ \hline 0.1870 & 0.4796 & 0.2625 \\ 0.9913 & 0.4960 & 0.1863 \\ 0.7120 & 0.2875 & 0.8171 \\ 0.8714 & 0.0609 & 0.1233 \\ \hline 10 & 0 & 0 \\ 0 & 0 & 10 \end{bmatrix} \} \varphi \mathbf{H} \qquad \hat{\mathbf{s}}_1^{\mathrm{T}} = \begin{bmatrix} 0.0570 \\ 0.2696 \\ 0.4829 \end{bmatrix}$$

The least squares result after the first iteration is shown. As a hard constraint was imposed the constrained values are close to the predefined constrained values, $\hat{\mathbf{s}}^{\mathrm{T}}_{1(\varphi:10)} = [0.0570 \quad 0.2696 \quad 0.4829]$. A harder constraint $\hat{\mathbf{s}}^{\mathrm{T}}_{1(\varphi:100)} = [0.0580 \quad 0.2693 \quad 0.4821]$ would have resulted in constrained values closer to the predefined equality constraints stipulated for coefficients s_1 and s_3. A smaller penalty would have resulted in a smaller deviation from the unconstrained values, such that slight deviations are allowed from the equality constrained coefficients, i.e. $\hat{\mathbf{s}}^{\mathrm{T}}_{1(\varphi:0.1)} = [-0.0153 \quad 0.2665 \quad 0.5808]$.

In order to test for the presence of active constraints, once the NWAY P-ALS has converged, it is useful to turn off all constraints and estimate the concentration profiles and spectral profiles using a conventional unconstrained least-squares method. If the unconstrained solution is a good match with the constrained solution, one may deduce that there were relatively few active constraints, that they had a minimal impact on the final solution, and the research hypothesis is fulfilled. On the other hand, when large differences are observed between the final constrained and unconstrained solutions, one may conclude there are strong active constraints that exert a significant influence on the results, thus the original research hypothesis is not fulfilled. Constrained models with

many influential active constraints suggest the SMCR model is inappropriate for one or more reasons. Some examples might include non-linear response or other significant deviations from the bilinear model, i.e., the wrong number of components was used in the SMCR model, or the shapes of the profiles estimated by SMCR are a poor match to the true underlying profiles.

Software Implementation of NWAY P-ALS

GUIPRO is a MATLAB program with a graphical user interface which has been designed to enable non-experts to easily apply pre-processing steps and perform data analysis using a variety of curve resolution methods. The NWAY P-ALS GUI in GUIPRO is described below and pictured in figure 32.

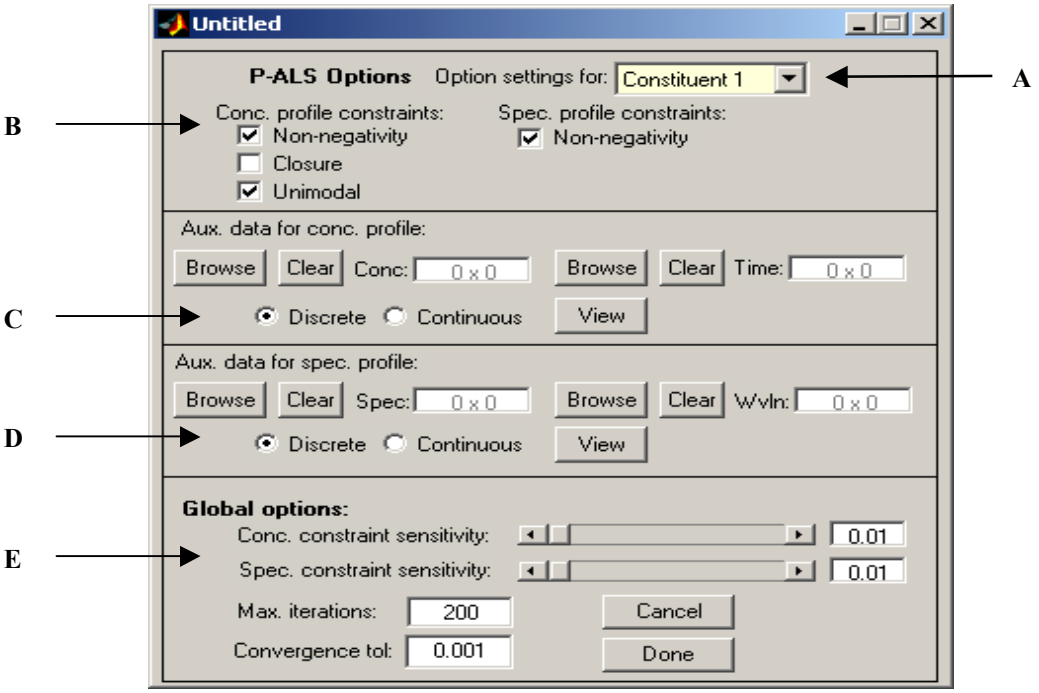


Figure 32. Screen shot of the NWAY P-ALS options window for the first constituent. Figure 32a. The pull down menu can be used to change the constraint options for the different constituents. Figure 32b. Optional non-negativity, closure and unimodality constraints can be applied in the concentration profile of constituent one, through checking the appropriate boxes. Optional non-negativity constraints can be applied in the spectral profile of constituent one, through checking the appropriate box. Figure 32c-d. Sparse or complete auxiliary concentration data with the corresponding time axis can be selected by pressing the browse button and uploading the appropriate files into the NWAY P-ALS analysis for equality constraints. This option is also available for the spectral auxiliary data. Figure 32e. The penalty factor extends from 0.01 to 20 and can be set by typing a value in the text box or sliding the bar adjacent to the text box. The

maximum number of iterations and convergence tolerances can be set by typing a value in the appropriate text box. The "Done" button is pressed when all the constraints are set to start the calculation.

The matrices to be loaded for NWAY P-ALS analysis should first be individually pre-treated and saved using compatible pre-treatment options. Pre-treatment options include selection of sub-matrices of spectra and concentration profiles, baseline correction, removal of outliers, normalisation, and estimation of the number of PCs. Once an initial pre-treated sub-matrix (individual data matrix) has been loaded into the GUIPRO environment, the user can select the NWAY P-ALS function from the drop down menu, under N-way methods. Up to 10 pre-processed sub-matrices can be loaded into the GUIPRO environment. Initial guesses of the concentration profiles are required to start the NWAY P-ALS algorithm. These can be determined using a variety of methods. The first method simply uses initial estimates defined in the analysis of individual data matrices. The second method uses the needle search technique of single batches while the third method uses EFA estimates of the single batches. The options available in the constrained least squares optimisation for the different concentration profiles include non-negativity constraints, closure constraints, and unimodality in the concentration profiles. Non-negativity constraints can also be applied in the spectral profiles. Auxiliary data in the form of reference spectra or reference concentration values can be uploaded into the NWAY P-ALS environment and used as equality constraints for selected spectral profiles or concentration profiles, respectively. Penalty factors for the constraints can be set by inputting values directly in the text boxes, or by using a sliding bar. The tolerance and the convergence criteria can be predefined by the user.

II.3 Qualitative Analysis of the Base Catalysed Esterification Reaction of Acetic Anhydride using NWAY P-ALS

II.3.1 Introduction

This study was completed in collaboration with Prof. Paul Gemperline, East Carolina University, USA.

Aim

The aim of the study was to use NWAY P-ALS to resolve the concentration and spectral profiles of 1-butanol with the reaction constituents of the base catalysed esterification reaction of acetic anhydride. The benefits of using the NWAY P-ALS approach include the reduction of the number of active constraints at the solution point, whilst the batch column-wise augmentation allowed strong constraints in the spectral profiles and resolved the rank deficiency. The NWAY P-ALS solution was validated by comparing the percent yield of 1-butyl acetate determined by GC for each batch. The results were also compared with the multi-way MCR-ALS results using hard and soft constraints to determine whether any advantage had been gained through using the weighted least squares function of NWAY P-ALS over the MCR-ALS resolution.

Introduction

The base catalysed esterification reaction of acetic anhydride was studied because it was expected to provide a relatively simple reaction mechanism to study for the multi-batch analysis using NWAY P-ALS. The proposed reaction mechanism for the catalysed esterification is as follows. An activated complex of acetic anhydride with pyridine is formed in the pre-reaction mixture from nucleophilic attack of pyridine on the carbonyl carbon atom of acetic anhydride, giving a tetrahedral intermediate. The intermediate

reacts with 1-butanol to form 1-butyl acetate, acetic acid, and liberating the catalyst, pyridine, as shown in figures 33-34.

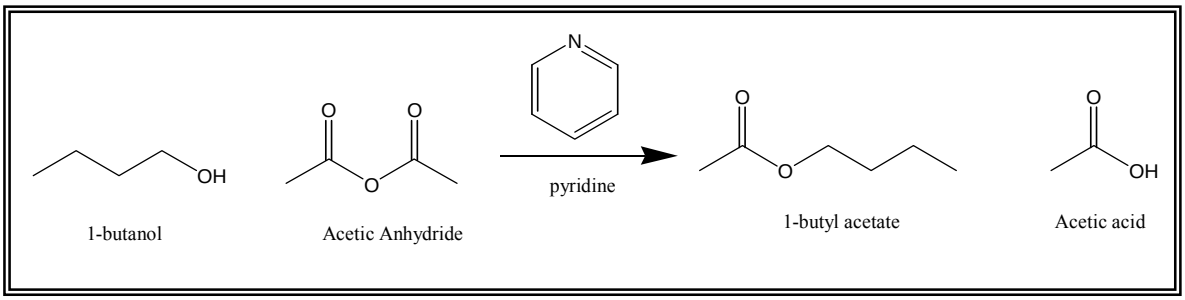


Figure 33. Base catalysed esterification of acetic anhydride

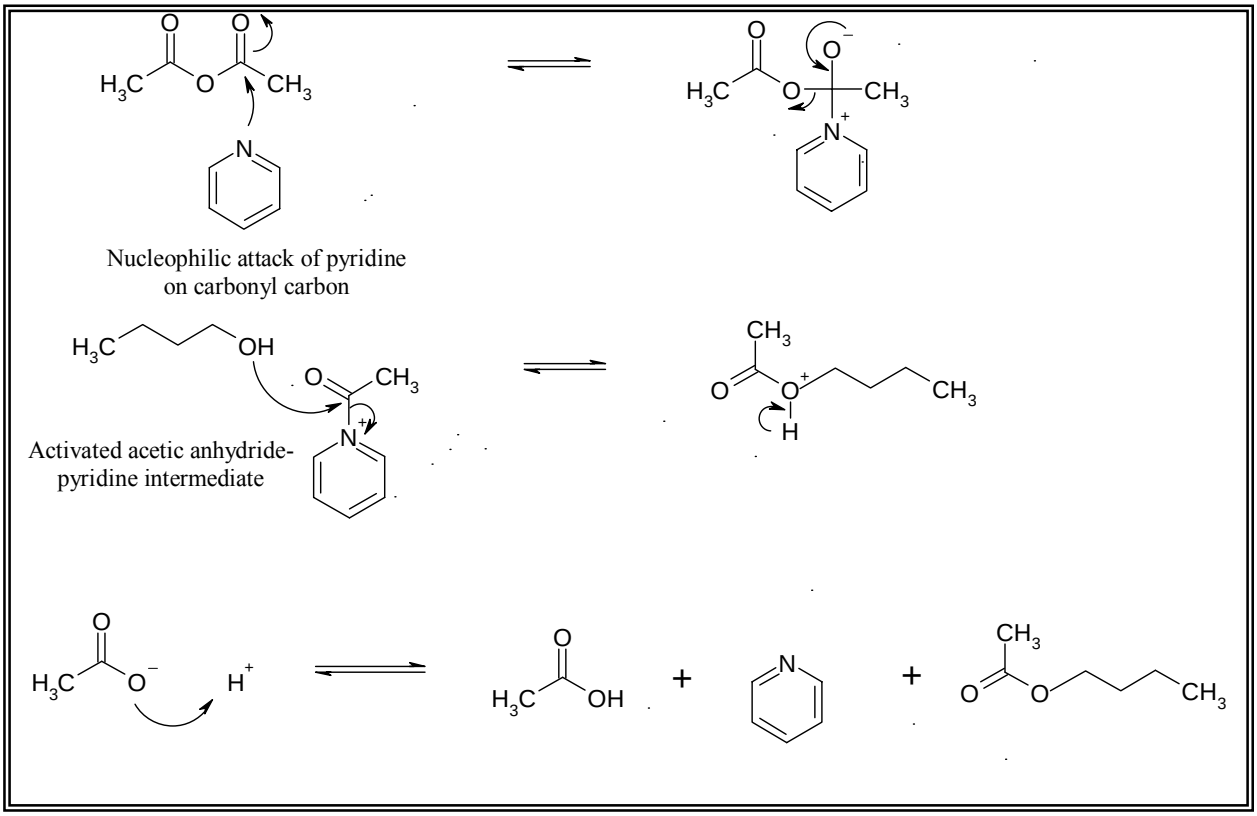


Figure 34. Postulated reaction mechanism of the base catalysed esterification of acetic anhydride

In a previous study completed by R. Miller [151], the concentration of 1-butyl acetate was predicted using several chemometric approaches CLS, PLS and P-ALS as well as an empirical approach, kinetic fitting. The study concluded that the kinetic fitting approach gave the best prediction for 1-butyl acetate because the underlying chemical model of the system was known and *a priori* information relating to the starting conditions was available. The predominant factors for the reduced prediction

capabilities of the remaining methods were attributed to the rank deficiency of the solution for both CLS and P-ALS, the large number of feasible solutions produced from P-ALS and the design and implementation of the PLS model. To improve the P-ALS results sequential batch analysis of the multiple experiments using NWAY P-ALS was applied. This approach was expected to reduce the number of feasible solutions and break the rank deficiency in the measurement matrices.

A unique feature of the P-ALS algorithm is the ability for user control implementation of constraints such that weighted constraints, i.e., soft constraints which allow small deviations from the constrained values, can be applied during P-ALS. Advantages of the implementation of soft constraints during P-ALS include reduced distortions of resolved profiles, reduction of the number of active constraints at convergence which reduces the model lack-of-fit, and a reduced impact of noise and non-ideal response on constraints which lead to improved results. The NWAY P-ALS approach also gains a second order advantage especially if there are more active constraints in the multiple experiments. The multi-way constraints that can be employed include i) common species must have the same spectrum in all matrices, ii) common species must have concentration profiles with equivalent shape in all data matrices, and iii) one or more regions of zero components may be present [111].

In this study NWAY P-ALS is applied to the sequential batch data to reduce the number of feasible solutions, to break the rank deficiency in the measurement matrix, to improve the prediction capabilities of the P-ALS analysis and to determine whether any advantage had been gained using the weighted least squares function of NWAY P-ALS over MCR-ALS resolution.

II.3.2 Experimental

The NIR data was collected by Robert Miller, East Carolina University.

II.3.2.1 Reaction Conditions

Three reactions of the pyridine catalysed esterification of acetic anhydride were completed at equimolar ratios of anhydrous 1-butanol (46.3ml, Aldrich Milwaukee, WI), to acetic anhydride (47.6ml, Aldrich Milwaukee, WI), with different pyridine catalyst concentrations, (4ml, 2ml, 8ml, Aldrich Milwaukee, WI). All reactions were performed in an auto-MATE (H.E.L. Inc., Lawrenceville, NJ computer controlled reactor system). Process conditions, including reaction temperature, jacket temperature and agitation were controlled by WinISO® software from H.E.L. running on 333 MHz Pentium II computer. A custom glass reactor (75 mL) designed to accept a bundle fiber-optic NIR transflectance probe was used. The temperature of the reactions were thermostated with a recirculating Lauder RM6 Heater/chiller at 30˚C. For each run, the reactor was allowed to equilibrate until stable temperature was maintained.

The reactor vessel of the AutoMATE reactor system was charged with the respective volume of acetic anhydride and pyridine. Nine pre-reaction spectra were acquired initially to measure the two reagents in the absence of 1-butanol, so as to partially break rank deficiency in estimated concentration profiles. Between acquisition of the ninth and tenth spectrum, the pre-measured aliquot of anhydrous 1-butanol (BuOH, Aldrich) was charged into the reaction manually to initiate the reaction. Spectral acquisition occurred every 30 seconds for ~1.5hours.

II.3.2.2 Data Acquisitions

Near-IR background spectra were acquired using a FOSS-NIRSystems model 6500 spectrophotometer, fitted with a transflectance bundle fiber-optic probe (Silver Springs, MD). The probe gap was set to 0.5 mm giving an effective pathlength of 1.0 mm. Spectra were acquired over the range of 1100 to 2498 nm, at a resolution of 2nm and ratioed against an air blank recorded prior to the start of the reactions. After the start of a reaction, spectra were recorded every 30 seconds for about 1.5 hours, by averaging 10 scans using the VISION data acquisition software from FOSS-NIRSystems. MATLAB6p5® (The Math Works, Inc) was used to complete all data processing.

II.3.2.3 Validation

The chromatographic analyses were carried out on a Hewlett-Packard GC (Model: GC-6890), equipped with a split/splitless injection port (the split injection port was used). The fused capillary column (HP-35MS), 30.0 m x 0.25 mm I.D, was connected to a flame ionisation detector (FID). The initial temperature was set to 30°C, with a temperature ramp of 10°C/min, the final temperature being 120°C. The complete run took 12 minutes. Helium was used as the carrier gas with a flow rate of 20mL/min. The data acquisition, data analysis and instrument control was carried out using HP Chemstation software.

GC standards were prepared by the following procedure. A Mettler AT400 digital balance was used to accurately weigh each of the chemicals; 1-butanol (Fisher Scientific), 1-butyl acetate (Fisher Scientific) and methanol (Fisher Scientific). The final working standards were prepared over a weight ratio range of 1:1, 2:1, 3:1, 5:1 and 7:1 of 1-butyl acetate:1-butanol. The stock solutions were diluted with methanol (10.0g).

Approximately 1 μL aliquots from each standard was injected onto the GC column. Quantification was performed by measuring the ratio of the peak height of 1-butanol and 1-butyl acetate and the percent weight ratio of 1-butanol by 1-butyl acetate. A linear calibration curve was constructed, see R. Miller's thesis [151].

II.3.3 Results and Discussion

II.3.3.1 Aim

The first objective of this study is to illustrate the use of NWAY P-ALS to resolve the concentration and spectral profiles of the reaction constituents of the base catalysed esterification reaction of l-butanol with acetic anhydride. The second objective is to compare the results of the multi-way NWAY P-ALS method with the multi-way multivariate curve resolution Alternating least squares (MCR-ALS) [67, 70] results using both hard and soft constraints to determine whether any advantages had been gained through the application of the weighted least squares function of NWAY P-ALS over the MCR-ALS resolution.

II.3.3.2 Research Hypothesis

The SMCR research hypothesis for these experiments, states: *There exists an unconstrained bilinear model with unimodal, non-negative pure component concentration profiles and pure component non-negative spectral profiles of acetic anhydride and 1-butanol that fits the data matrix of measurements obtained from the evolving system.*

II.3.3.3 Reaction Profiles

The data was column-wise augmented and the segment between 2200 to 2498nm was deleted as non-linear detector response was observed above 2200nm due to high levels

of stray light. Spectra were baseline corrected by subtracting the average absorbance from 1100 to 1124nm where no significant NIR absorbance was observed. From the reaction spectral profiles it was possible to observe a non-linear shift of the 1150-1200nm peak to 1180-1230nm upon the addition of 1-butanol to the pre-reaction mixture, shown in figure 35. The broad O-H stretching overtone peak of 1-butanol in the region from 1400 nm to 1600 nm appeared to coalesce into a new sharp peak at 1390-1470nm (discussed in section II.3.3.4), suggesting the extensive O-H bonding in 1-butanol was disrupted to a significant degree by possible molecular association (hydrogen bonding) between 1-butanol and acetic anhydride.

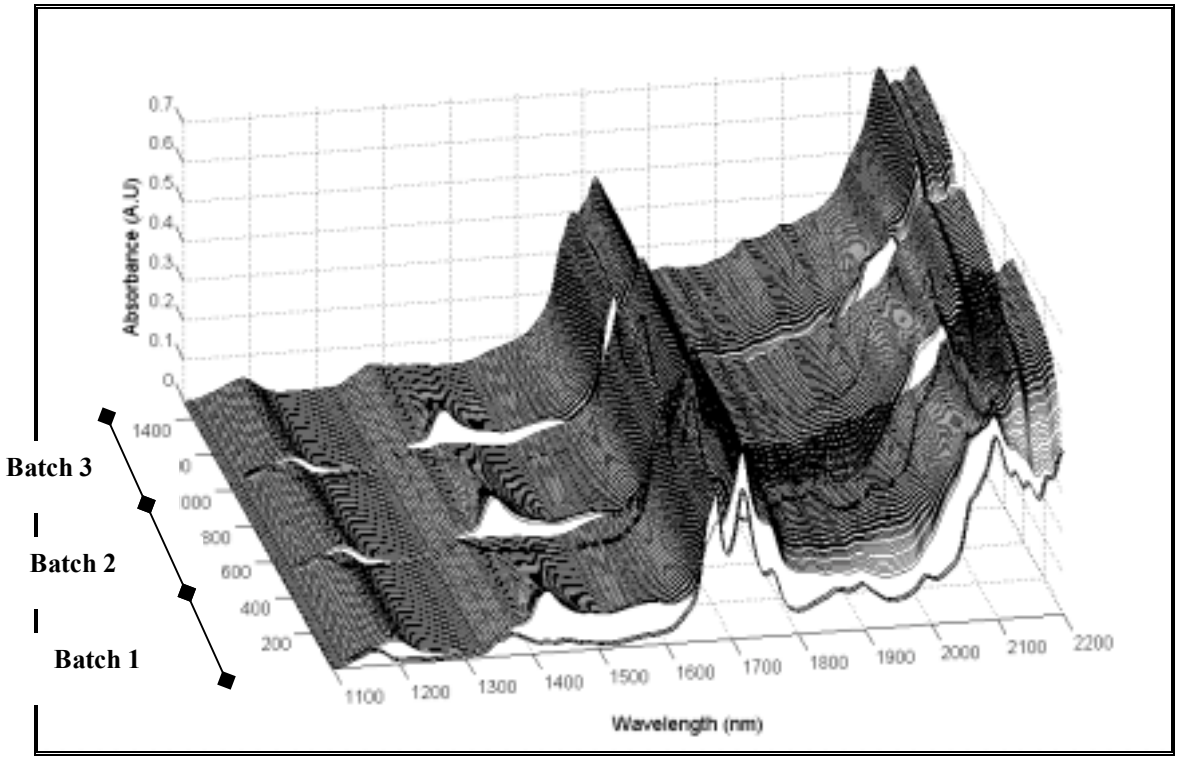


Figure 35. Column-wise augmentation of the NIR batches of the pyridine catalysed esterification reaction of acetic anhydride

The number of principal components was estimated using an F-Test for each batch. Three components were identified representing 1-butanol, acetic anhydride and a pseudo product; which was a linear combination of the two products; 1-butyl acetate and acetic acid both of which were formed at identical rates. Initial estimates of the

concentration profiles were resolved from the original single batch analysis using EFA. The EFA initial estimates were column augmented in the appropriate order to initialise the alternating least squares procedure.

In the NWAY P-ALS analysis, non-negativity and unimodality constraints were applied in the concentration profiles of all constituents. Pure component reference spectra of 1-butanol and acetic anhydride were loaded for use as equality constraints for the estimation of the spectral profiles of 1-butanol and acetic anhydride because it was defined in the research hypothesis that these components were present in the SMCR solution. Non-negativity constraints were applied in the spectral profiles of all three constituents. In cases where hard constraints were desired, the penalty value was set to 20 and for soft constraints the penalty value was set to 1. The convergence criteria was $1e^{-9}$ (relative change in residual sum of squares from one iteration to the next) and the maximum number of iterations was 500.

II.3.3.4 Hard vs. Soft NWAY P-ALS Constraints

The calculated spectra of the resolved constituents, acetic anhydride, 1-butanol and the pseudo-product, using hard and very soft penalty functions in the NWAY P-ALS calculation are given in figures 36a-b respectively. After convergence, the unconstrained solutions were also computed and superimposed on the constrained solutions to test for the presence of influential active constraints.

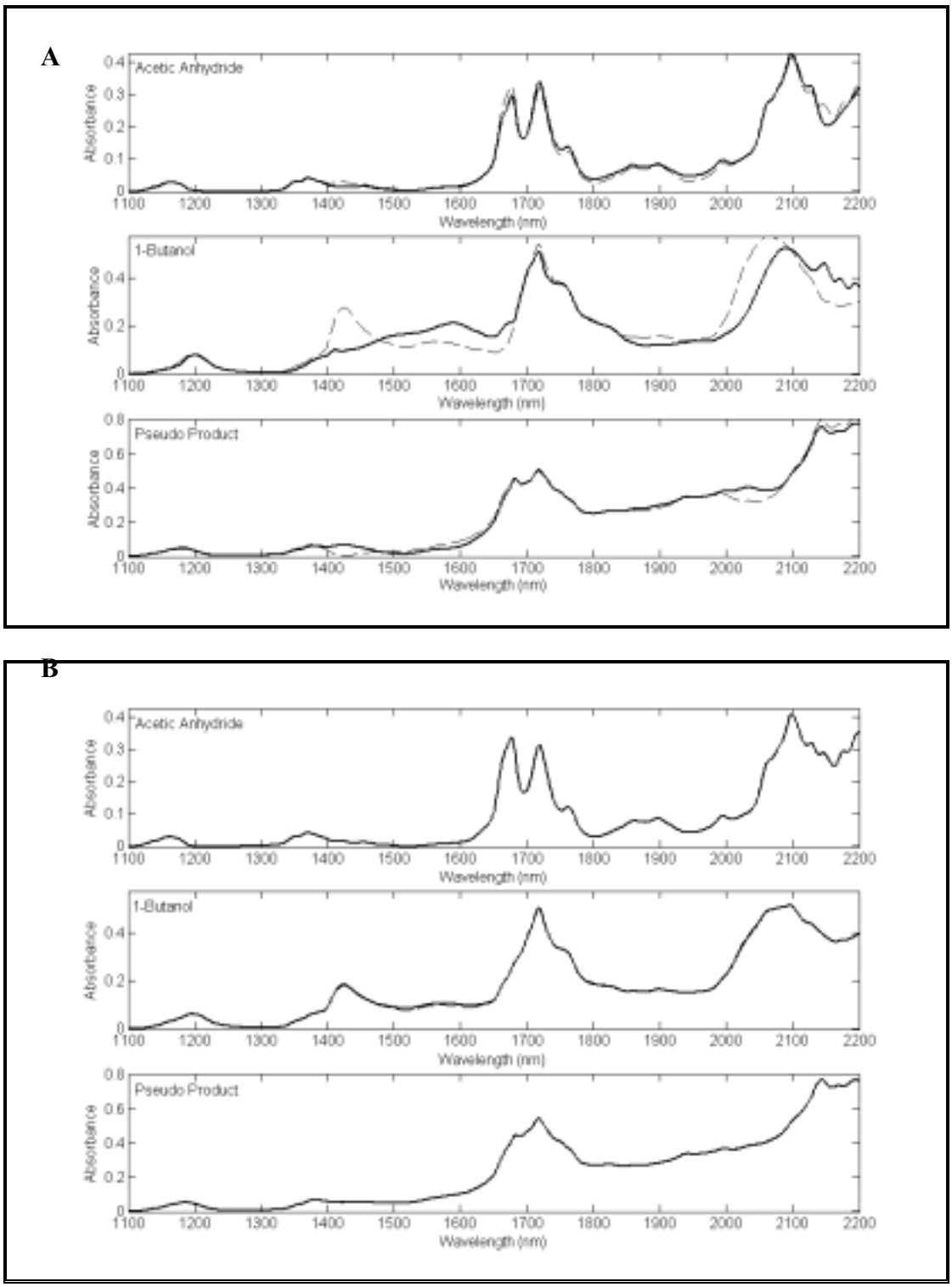


Figure 36a-b. Resolved spectral profiles. Figure 37a. Hard NWAY P-ALS options and Figure 37b. Soft NWAY P-ALS options. The constrained profile (bold line), unconstrained profile (dashed line).

Comparison of the *spectral profiles* for the constrained and unconstrained solutions using hard constraints are shown in figure 36a. Comparison of the unconstrained spectrum of 1-butanol to the hard constrained spectrum reveals significant differences,

indicating the presence of many, active influential constraints in the hard constrained solution. Comparison of the unconstrained solution and the soft constrained solution shown in figure 36b reveals there are no influential active constraints.

On the other hand, both the hard unconstrained profile (figure 36a) and the soft constrained profile (figure 36b) of 1-butanol contain two distinct differences from the constrained profile (figure 37a), which were initially thought to be due to water contamination in the reagent [82]. The presence of water was ruled out in a separate experiment by measuring the NIR spectrum of 1-butanol dried over molecular sieves. It was speculated that the addition of 1-butanol to acetic anhydride caused the formation of a new complex between 1-butanol and acetic anhydride, which in turn disrupted the extensive hydrogen-bonding network in neat 1-butanol. In hydrogen-bonded alcohols, there is a broad peak in the 1460-1600 nm region, this peak was attributed to the first overtone of the O-H stretching vibration [152, 153], shown in the constrained solution. L G Weyer and S-C Lo [153], and other workers report that the series of peaks between 1410 nm and 1600 nm are due to the first overtone of the O-H stretching vibration of different hydrogen bonded aggregates. In dilute solutions using a non-hydrogen bonding solvent, aliphatic alcohols have a first overtone peak at about 1410nm, corresponding to non-hydrogen bonded O-H stretching vibrations. A peak at this position can be observed in the unconstrained spectral profile of 1-butanol, although it is significantly broader than the first overtone of the non-hydrogen bonded O-H stretching vibration reported by other workers [152, 153]. Other supporting evidence determined by Czarnecki *et al.* and Weyer *et al.* [152, 153], found that the hydrogen bonded peak at 1570 nm decreases in intensity with temperature whereas the non-hydrogen bonded peak at 1410 nm increases in intensity, which supports the theory that it is due primarily to non-hydrogen bonded O-H overtones. It has been shown that in

the mid-IR, a diffuse O-H association band related to the deformation of the O-H bond occurs at 1420 cm^{-1} which accounts for the 2100 nm NIR band. This diffuse band is reported to disappear in dilute solutions of alcohols in the mid-IR, where hydrogen bonding does not occur [152, 153]. In the NIR spectrum, the peak at 2100nm becomes narrower and is shifted towards a shorter wavelength in dilute solution of non-hydrogen bonding solvent or with increasing temperature.

Further analysis to determine whether the peak at 1410 nm was formed by a complex formation between 1-butanol and acetic anhydride was completed by Gemperline *et al.* [154]. Initially 5ml of 1-butanol was added to acetic anhydride without the catalyst present. From the reaction profile it was possible to observe the appearance of a peak at ~1410 nm, which suggested that the extensive hydrogen bonding in 1-butanol was disrupted, as a complex was formed between 1-butanol and acetic anhydride. Therefore, the research hypothesis concerning the estimated spectrum of 1-butanol was inaccurate as both the complexed 1-butanol and uncomplexed 1-butanol was present at the beginning of the reaction. A pseudo spectrum of these species was calculated by adding the spectrum of 1-butanol with the spectrum of 1-butanol obtained from the original soft NWAY P-ALS analysis. This pseudo spectrum was used as an equality constraint in place of the neat 1-butanol spectrum.

Furthermore, it was postulated that there were at least four components in the NWAY solution because the concentration of pyridine varied in the batches. This was confirmed by visual inspections of the PCA scores plot of the column-wise augmented measurement matrices, see appendix 1.4.1. Structured variance was observed in the first four principle components. A new research hypothesis was tested; *There exists an unconstrained bilinear model with unimodal, non-negative pure component*

concentration profiles and pure component non-negative spectral profiles of a) acetic anhydride, b) linear combination of complexed 1-butanol and uncomplexed 1-butanol, and c) pyridine that fits the data matrix of measurements obtained from the evolving system. Therefore the pure spectrum of acetic anhydride, pyridine and a linear combination of complexed 1-butanol and uncomplexed 1-butanol were loaded for use as equality constraints. Starting estimates for the four component's resolutions were obtained using EFA of the single batches. The initial estimates from the three batches were automatically column augmented in the NWAY P-ALS algorithm in the appropriate order to initialise the ALS procedure. The hard penalty equality constraint was set to 20 and the soft penalty equality constraint was set to 1. The convergence criteria and the maximum number of iterations were as specified previously.

In this case the application of soft constraints enabled the correct resolution of the spectral and concentration profiles of acetic anhydride, linear combination of complexed and uncomplexed 1-butanol and the linear combination of 1-butyl acetate and acetic acid. This can be observed in figure 37b and figure 38b respectively. The hard constrained profile is shown in figure 37a and figure 38a respectively. For the hard constraint, the correct spectral profiles of the reaction constituents were resolved, i.e., the correct functional groups were present in each of the unconstrained profiles. However, there were minor deviations between the unconstrained and constrained profiles of the pseudo 1-butanol spectrum and the unconstrained and constrained profiles of pyridine. The main deviations in the pseudo 1-butanol spectrum occurred in the spectral ranges between 1400-1600 nm and ~2000-2200 nm which were attributed to the disruption of the extensive hydrogen bonding network in neat 1-butanol caused by complex formation between 1-butanol and acetic anhydride and in the pyridine spectrum between 1700-2100nm.

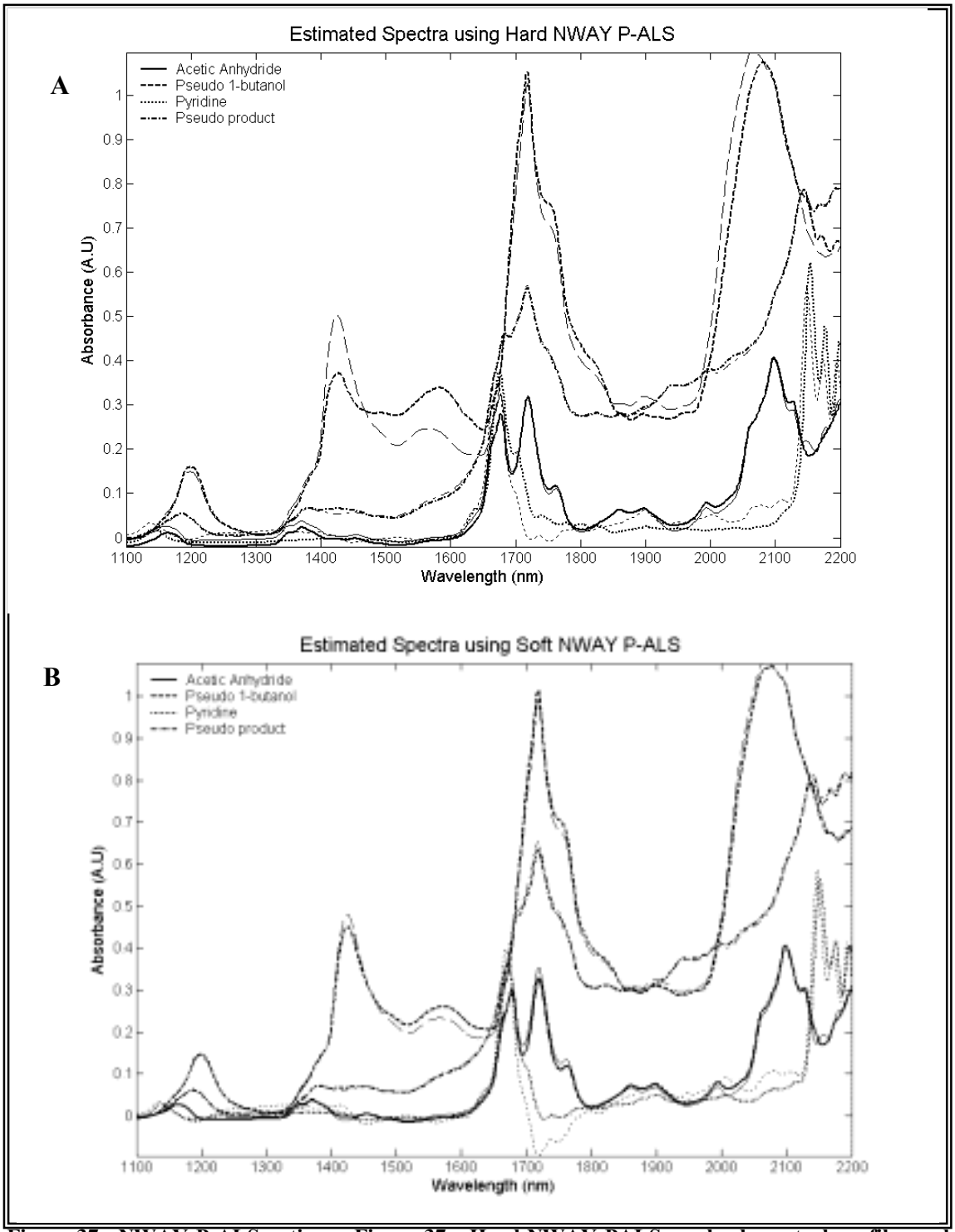


Figure 37. NWAY P-ALS options. Figure 37a. Hard NWAY-PALS resolved spectral profiles and Figure 37b. Soft NWAY P-ALS resolved spectral profiles. The constrained profile (bold line), unconstrained profile (normal line).

The soft constrained spectral profiles shown in figure 37b contained less active constraints in the pseudo 1-butanol spectrum, although the deviations in the pyridine

spectrum persisted. However, from observation it was possible to realise overall that deviations between the constrained and unconstrained spectra were minimal.

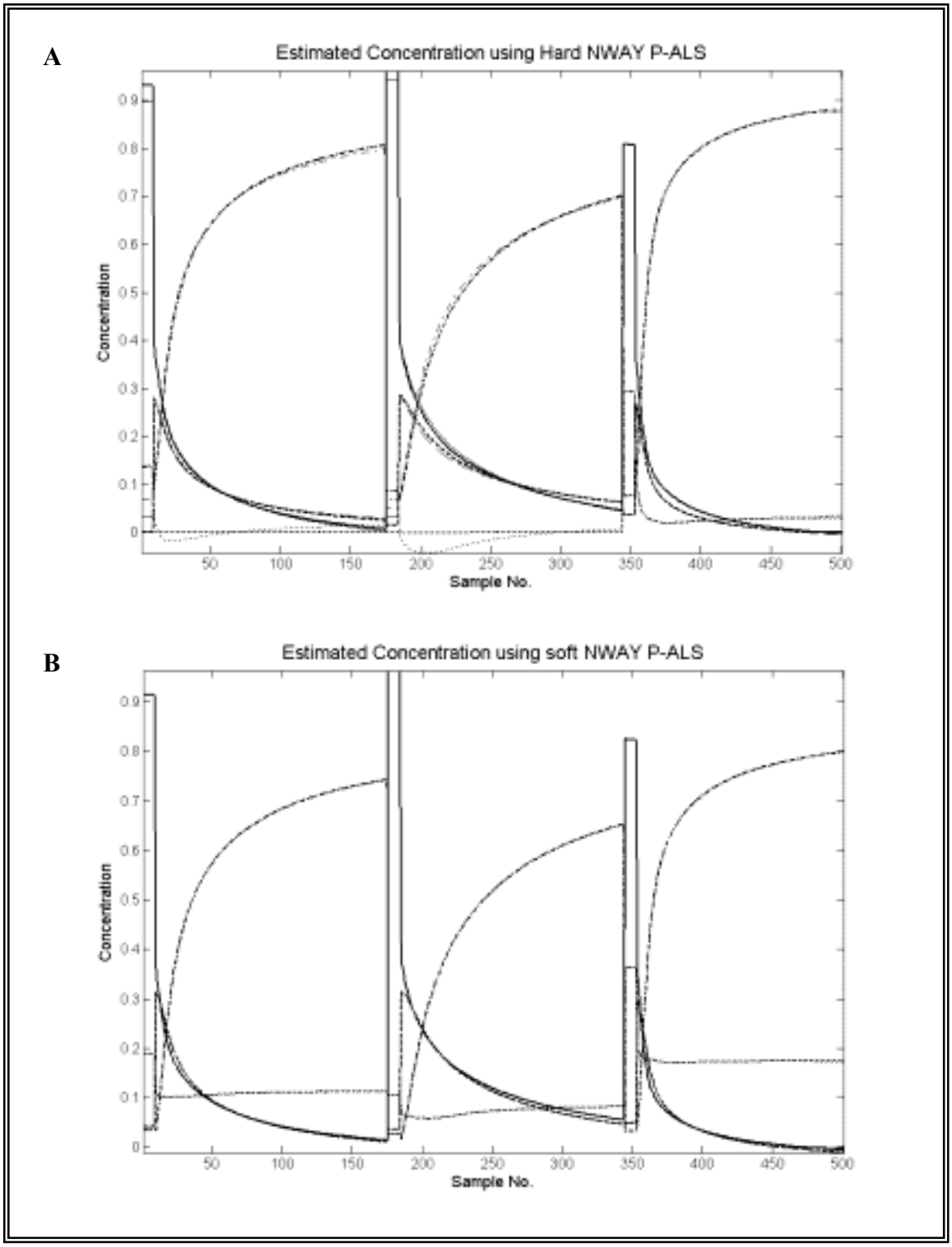


Figure 38. NWAY P-ALS options. Figure 38a. Hard NWAY-P-ALS resolved concentration profiles and Figure 38b. Soft NWAY P-ALS resolved concentration profiles. The constrained profile (bold line), unconstrained profile (normal line).

In this application the real benefit of using less stringent constraints was readily observed in the resolved concentration profiles of acetic anhydride, pseudo 1-butanol spectrum and pyridine. Comparing figure 38a and figure 38b for the hard and soft constrained solution respectively, it was possible to observe that the soft constrained reagent profiles were consumed at approximately the same rate, whereas the hard constrained reagent profiles did not give as good a solution. The concentration profile resolved for pyridine using the hard constraint was also incorrect because the hard constraints forced a non-negative solution with unrealistic concentration profiles. This is shown in figure 38a where the constrained solutions of batches 1 and 2 show a zero concentration profile and batch 3 shows an increase of concentration. The soft constrained profiles on-the-other hand produce the correct concentration profiles of pyridine in each of the batches, which were constant and non-negative in batch 1 and 3, although batch 2 shows a slight positive incline of concentration. The RMS error between the reconstructed matrix and the original matrix was calculated and it was found that less error was found in the soft constrained solution as more variance was described than in the hard constrained solution (RMS 8.94×10^{-6} and 3.72×10^{-5} respectively). The NWAY P-ALS algorithms converged smoothly and monotonically, however, the hard constrained solution converged faster than the soft constrained solution.

The percent yield of 1-butyl acetate quoted from R. Miller's thesis and calculated from the NWAY P-ALS results for the hard and soft constrained profiles are given in table 8. The average percent yield deviation from the GC results was $8\pm12\%$ for the hard constrained profile and $7\pm2\%$ for the soft constrained profile. The percent yield obtained for the hard constrained profiles of batches 1 and 2 had an error margin of $2\pm2\%$, and showed that these results were close to the GC percent yield data for the

respective batches. However, the third batch produced the greatest error in the prediction of the 1-butyl acetate and is responsible for the marked increase in the percent deviation. On the-other-hand the error produced in the soft constraints were more equally distributed across the three batches, reducing the error margin in the final result.

Batch No	GC (%)	Hard NWAY P-ALS(%)	Soft NWAY P-ALS(%)
1	87	87	82
2	81	77	72
3	89	109	97

Table 8. Comparison of the GC percent yield with the percent yield obtained from hard and soft NWAY P-ALS analysis

As the percent deviation in the GC data was unknown it was impossible to determine whether the results were within the acceptable range for the solution. It was postulated that the GC results were slightly imprecise because of a possible side reaction occurring during the work-up procedure. This side reaction was postulated to occur when the reaction was quenched using methanol, causing the equilibrium to be shifted to the left-hand-side, reducing the expected concentration of 1-butyl acetate [151].

Nevertheless, from the data available, it was concluded that the solution obtained from the soft NWAY P-ALS solution contained less error than the hard NWAY P-ALS solution and was closer to the percent yield GC data calculated for 1-butyl acetate.

Summary

It has been shown that the use of soft multi-way constraints greatly assisted in the resolution of the correct concentration and spectral profiles of the reagents and product profiles. Additionally, the application of the soft multi-way constraints reduced the number of active constraints whilst the batch augmentation allowed strong constraints in

the spectral profiles and the breaking of the rank deficiency to resolve the concentration and spectral profile of pyridine. Here, the importance of understanding the effects of intermolecular interactions on the pure component spectrum was integral to the correct application of equality constraints in the NWAY P-ALS resolution.

In the following section the results of the soft constrained four component NWAY P-ALS resolution was compared with the results of the multi-way MCR-ALS analysis using both hard and soft constraints.

II.3.3.5 Comparison of the resolution using NWAY P-ALS and multi-way MCR-ALS

Aim

The aim of this analysis was to compare the results of the soft NWAY P-ALS analysis with the results of multi-way MCR-ALS analysis using both hard and soft constraints to determine whether any advantage had been gained using the weighted least squares approach.

Hard MCR-ALS vs. Hard NWAY P-ALS

The research hypothesis specified for the four component resolution using NWAY P-ALS was tested for the four component resolution using multi-way MCR-ALS. In the multi-way MCR-ALS analysis the EFA starting estimates were used to initialise the multi-way MCR-ALS procedure. The hard options used for the constraints imposed in the concentration profile were non-negativity and average unimodality with a tolerance of 1.00 in each constituent profiles [107]. The hard constraints imposed in the spectral profiles were non-negativity in each constituent profile and equality constraints in the

spectral profiles of acetic acid, pyridine and pseudo 1-butanol. The convergence criteria was set to $1e^{-9}$ and the maximum number of iterations was 500.

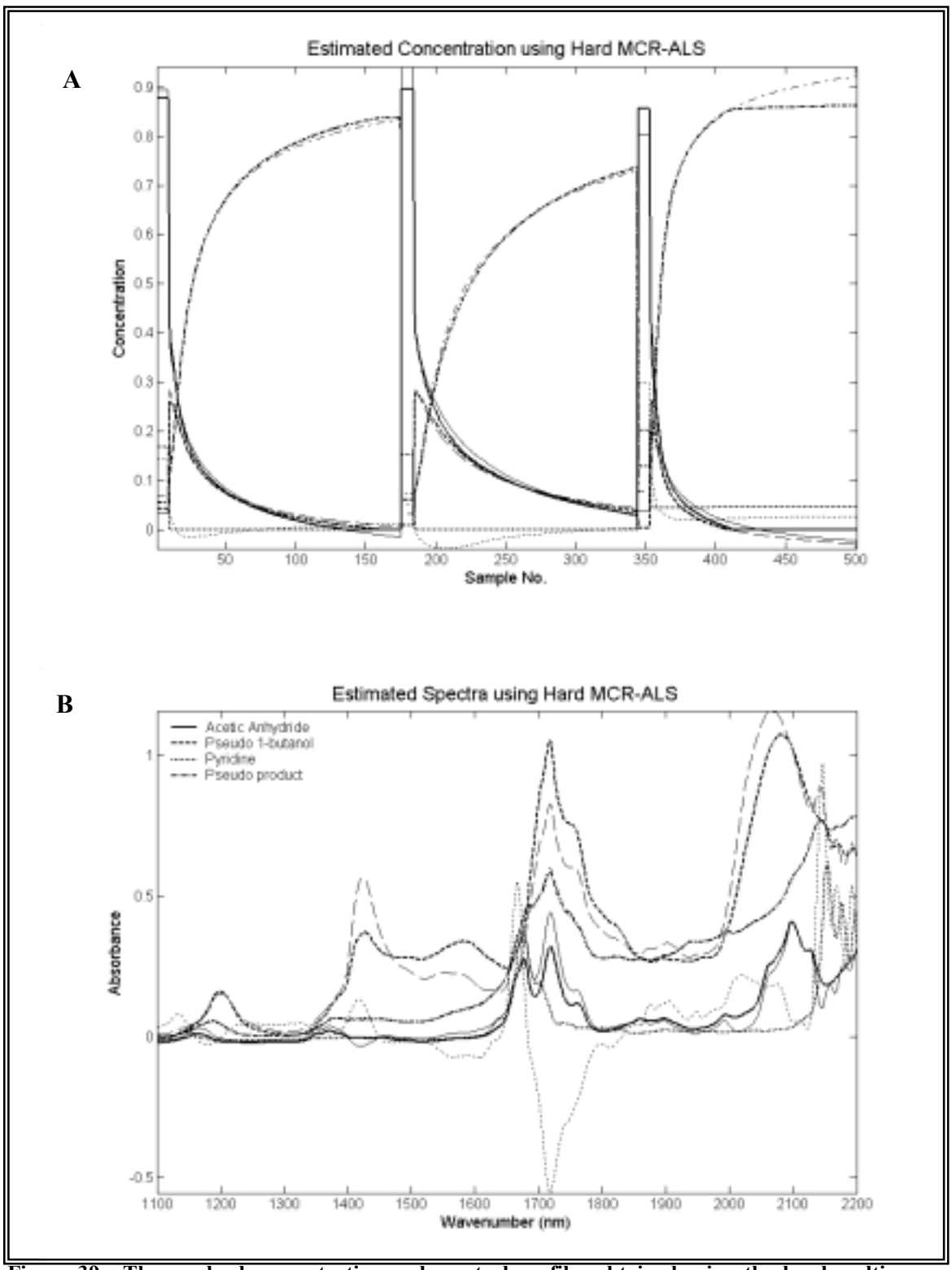


Figure 39. The resolved concentration and spectral profiles obtained using the hard multi-way MCR-ALS options. Figure 39a. The resolved concentration profiles. Figure 39b. The resolved spectral profiles. Constrained profile (bold line), unconstrained profile (normal line).

The hard multi-way MCR-ALS analysis produced good estimations of the reagent concentration profiles, shown in figure 39a. The reagent concentration profiles were observed to be consumed at approximately the same rate. Active constraints were found in the concentration profiles of acetic anhydride and pseudo 1-butanol in batches 1 and 3, 1-butyl acetate in batch 3 and pyridine in batches 1-3 due to the application of the non-negativity constraint. The effect of this was unrealistic estimates of each constituent where the non-negativity constraint was applied. This was also reflected in the flat concentration profiles of 1-butyl acetate predicted for batch 3, shown in figure 39a. The hard constrained multi-way MCR-ALS spectral profiles were similar to the hard constrained NWAY P-ALS spectral profiles, shown in figure 40b. The spectral profile of the pseudo 1-butanol spectrum contained minor deviations from the unconstrained solution (as discussed in the four component hard NWAY P-ALS resolution), and the spectrum of acetic anhydride contained minimal active constraints. However, the constrained pyridine spectral profile contained influential active constraints, which were particularly noticeable between 1374-1250nm, 1523-1645nm, 1684-1835nm and 1823-2115nm. The first three peaks in the unconstrained profiles spectra were attributed to pseudo 1-butanol, i.e., the unconstrained pyridine spectrum was contamination by pseudo 1-butanol. It was postulated that the inclusion of these peaks in the solution by relaxing the constraints in the multi-way MCR-ALS would produce an erroneous solution. This was tested using the soft multi-way MCR-ALS options.

The RMS error between the reconstructed matrix and the original matrix was calculated. It was observed that slightly less error was found in the hard NWAY P-ALS approach than in the hard MCR-ALS solution (RMS 3.72×10^{-5} and 4.92×10^{-5} respectively). The percent yield deviation from the GC result was less than the hard NWAY P-ALS

solution ($8 \pm 4\%$). The deviation in the multi-way MCR-ALS was less because the zero concentration of pseudo 1-butanol and acetic anhydride in the final batch reduced the expected 1-butyl acetate concentration, see unconstrained solution, figure 39a. No real advantage had been observed though employing the hard multi-way MCR-ALS approach over the hard NWAY P-ALS approach in this application.

Soft MCR-ALS vs. Soft NWAY P-ALS

The soft options for the constraints imposed in the concentration profiles were non-negativity and average unimodality with a tolerance of 1.05 in each constituent profile[107]. The soft options for the constraints imposed in the spectral profiles were non-negativity in each constituent profile and the *less than* equality constraint in the spectral profiles of acetic anhydride, pyridine and the pseudo 1-butanol. The same convergence criteria and the maximum number of iterations as specified previously were used in the analysis.

The application of the soft options in the MCR-ALS resolution resulted in the correct resolution of acetic anhydride spectrum, pseudo 1-butanol spectrum and pyridine with less active constraints. However, the concentration profiles of acetic anhydride and pseudo 1-butanol were not resolved well. The reagents were consumed at approximately the same rate, however the mole fraction of the reagents, after 1-butanol was added to the pre-reaction mixture, were markedly different, producing concentration profiles which did not approximate the true solution. Slightly less error was found in the soft constrained MCR-ALS solution than in the soft constrained NWAY P-ALS solution (RMS 8.78×10^{-6} and 8.94×10^{-6} respectively). However, the deviation from the GC percent yield data was greater than the soft NWAY P-ALS approach. Reduced error was found in the soft constrained MCR-ALS solution

compared to the hard constrained MCR-ALS solution (RMS 4.92×10^{-5} and 8.78×10^{-6} respectively).

The application of the soft MCR-ALS constraints did not improve the soft NWAY P-ALS solution. However, the application of soft constraints in the multi-way MCR-ALS resolution enabled a better resolution of the reaction constituent spectral profiles with less active constraints.

II.3.4 Conclusion

To summarise, it was not possible to resolve the correct concentration and spectral profiles for the reaction constituents using the hard multi-way MCR-ALS method and no real advantage was observed from the application of the hard NWAY P-ALS method over the hard multi-way MCR-ALS method. However, employing the soft constraints in the MCR-ALS resolution enabled a reduction in the number of active constraints in the spectral profiles, in comparison to the hard constraints in the MCR-ALS resolution and the reduction of the RMS.

Overall the best solution was obtained using the soft NWAY P-ALS options. The reagent and product concentration and spectral profiles of acetic anhydride, pyridine, linear combination of complexed and uncomplexed 1-butanol and linear combination of 1-butyl acetate and acetic acid were resolved from the multi-batch measurement matrix. The percent yield of 1-butyl acetate predicted using the soft NWAY P-ALS approach was closest to the percent yield calculated using GC, although from the available data it was not possible to determine whether the results were within an acceptable range. The increased success in the NWAY P-ALS resolution of the reaction constituents over the multi-way MCR-ALS approach was attributed to the increased flexibility and hence control in the application of constraints. Additionally, the application of the soft

NWAY P-ALS approach enabled a reduction in the number of active constraints in the solution and a smooth, monotonical convergence.

II.4 Quantitative Iterative Target Transformation Factor Analysis

II.4.1 Introduction

In the studies discussed so far, the quality of the initial estimates and application of constraints has been integral to the success of the resolution procedure, especially when the application of constraints produced local minima or divergence (NWAY P-ALS, Chapter II.2) or in the case where the variables found using the traditional exploratory tools were far from pure, i.e., they contained sizeable contributions from other components in the mixture (CATHy, Chapter II.1). Traditionally the improvement of the solution from the constrained ALS procedure has been executed by the addition of *a priori* information to the initial estimates and / or constraints. However, there are cases where this information is impossible to acquire under the process conditions or the information simply does not exist. Therefore, new exploratory tools are required to find refined starting estimates close to the actual solution in order to reduce the error in the MCR-ALS solution.

In this work a new type of rational SMCR strategy, based on ITTFA and SIMPLISMA, called Quantitative Iterative Target Transformation Factor Analysis (QITTFA)[155] has been developed. QITTFA was developed in order to produce initial estimates which approximate the true solution in the absence of selectivity for the individual constituents. QITTFA incorporates both a noise reduction procedure and generic constraints in the production of initial estimates. The ALS solution is improved in cases where no selectivity for individual constituents exist or where the application of constraints produce local minima or divergence. No *a priori* information regarding the process, such as pure spectral features or calibration information is required in the QITTFA procedure. Therefore, the advantage of this algorithm is attributed to two

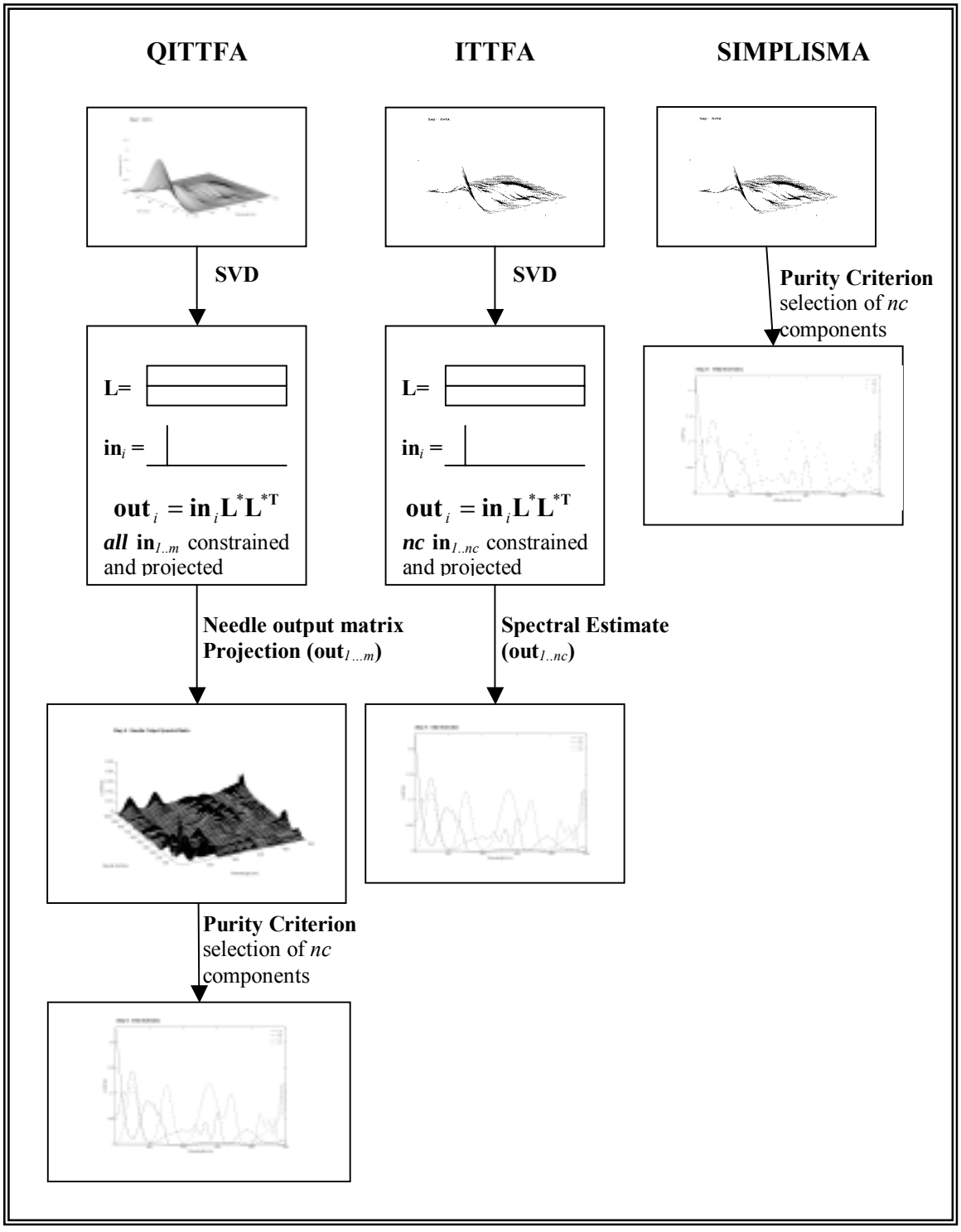
novel features, which are: 1. Each spectrum (or concentration profile) in the solution space can be repeatedly constrained and projected using generic constraints; the constrained estimates are often closer to actual solution; and 2. *Absence of unstructured variance (noise)* in the solution space, from which the initial estimates are determined.

The difference between the ITTFA and QITTFA procedure lie in the selection of the pure spectrum or initial estimates for alternating least squares. In the ITTFA procedure (using the needle search algorithm to select the starting estimates) each needle is projected into the space defined by the eigenvectors *once,* and the correlation between the input and output target is evaluated. The number of significant eigenvectors are retained to start the iteration procedure. In the QITTFA procedure each needle vector is projected into the space defined by either the reduced *column* orthonormal vector or the reduced *row* orthonormal vector (see later) and *all* output target spectra are retained and refined using generic constraints. SIMPLISMA is then used to select the most pure variables from the refined output target spectra.

In conventional SIMPLISMA applications, SIMPLISMA is used to select the pure components from the original measurement matrix. This method is known as the conventional SIMPLISMA approach, which is a pure-variable based method [64]. This means that it is assumed that every component in the mixture under study has a variable, which has a finite intensity for that particular component, and that the variable has a zero intensity for all other components in the mixture. A major limitation is that in the absence of pure regions for each component this technique is not suitable. To overcome this limitation a strategy was developed, in which SIMPLISMA was used to select pure components from inverted *non-negative* second derivative spectra [118, 121]. However, it was later recognized, that the second derivative approach could not

resolve well components, which were characterized by broad spectral features. A strategy combining the use of conventional and inverted *non-negative* second derivative spectra was formulated [156]. The second derivative spectra were used for a component, which was characterized, by narrow spectral feature (sharp peaks) and the conventional spectra was used for a component, which was characterized by broad spectral features. A limitation of the combined approach was that it was not very clear from the intermediate results whether one had to deal with a component with broad spectral feature or narrow spectral features. This lack of clarity in the character of the components to be extracted complicated the combined approach. The stepwise maximum angle calculation (SMAC) algorithm was introduced to enhance the intermediate results so it was easier to determine the character of the component (i.e. broad or narrow band) to be being extracted [157]. However, the limitation of both of the approaches (combined strategy and SMAC) is one still needs to subjectively characterize the component spectrum in order to select the correct approach. Therefore a method which enables the resolution of pure components in both the presence and absence of selectivity and components of differing spectral characteristics (narrow or broad spectral features) is required. QITTFA is also presented as a solution to this problem. In the QITTFA procedure, SIMPLISMA is used to select the most pure variables from the refined output target spectra. The advantages of the selection from the refined output target spectra are (a) the variance contribution due to noise is markedly reduced and (b) the refined target output spectra span the solutions space defined by the independent components. This enables the resolution of pure components in both the presence and absence of selectivity and components of differing spectral characteristics (narrow or broad spectral features) can be resolved from the

refined target output spectral matrix. The differences between the QITTFA, ITTFA and SIMPLISMA methodologies are in box 3.



Box 3. Outline of the QITTFA, ITTFA and SIMPLISMA methodologies

II.4.2 Methodology

In QITTFA procedure the pure variables are selected from a noise reduced solutions matrix, called the *needle output spectral matrix*, rather than either the original matrix as in the conventional SIMPLISMA approach or the second derivative measurement matrix as in the second derivative SIMPLISMA approach. The calculation of the needle output spectral matrix is described below.

II.4.3 QITTFA procedure

QITTFA is an automatic procedure for the determination of initial estimates from a measurement matrix. The user inputs required for QITTFA are (a) the data matrix, **D**, where each row represents a sample spectrum and each column represents the total response, the latter of which is a linear additive signal of each chemical constituent (at a constant pathlength), (b) the estimated number of chemical components (*nc*), (c) a correction factor ( ), and (d) the maximum number of iterations (*nit*). The outputs of QITTFA are (a) the needle output spectral matrix (**Z**), and (b) the initial estimates (either $\mathbf{S}_0$ or $\mathbf{C}_0$). See appendix C for MATLAB routines. The main steps of QITTFA are given in table 9.

Steps	Procedure
1	Aggregation of user inputs for QITTFA analysis
2	Generation of needle spectra
3	Generation of needle output spectra
4	Optional application of constraints to the needle output spectra
5	Production of needle output spectral matrix
6	Selection of initial estimates from the needle output spectral matrix

Table 9. Steps of the QITTFA procedure.

In the following section a description of the QITTFA procedure is given, using the simulated HPLC-DAD data (see Experimental Chapter II.5.2.1).

1. Aggregation of user inputs for QITTFA analysis

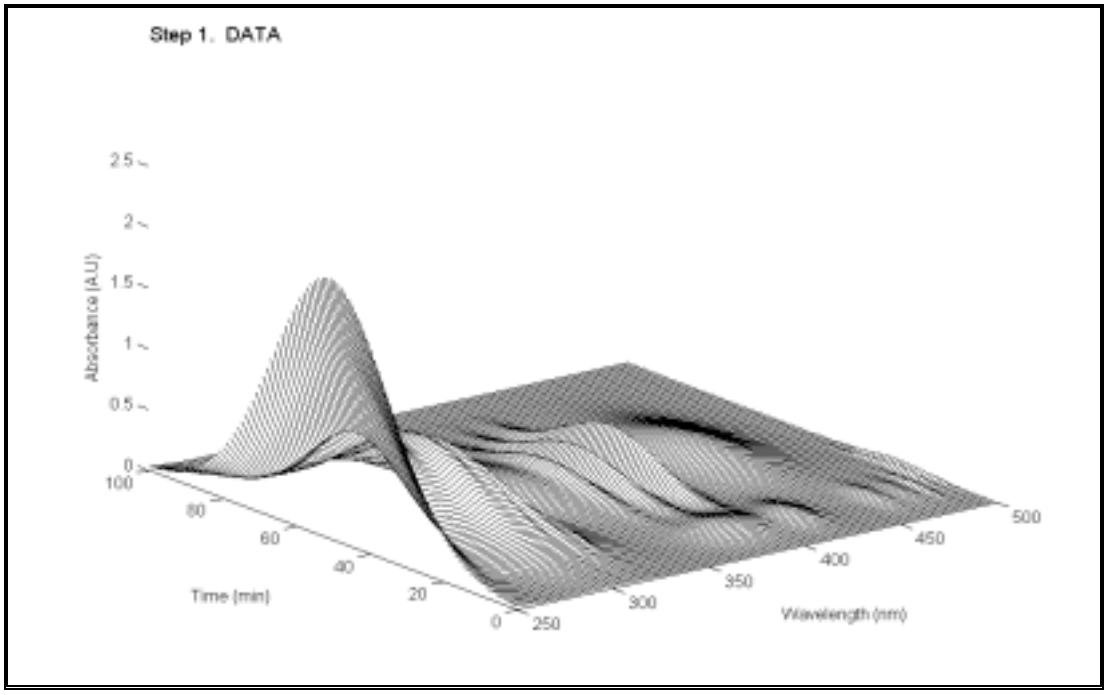


Figure 40. Simulated HPLC-DAD data

The measurement matrix for the simulated HPLC-DAD data is shown in figure 40. The measurement matrix has 500 spectral channels and 100 samples. The user inputs for $QITTFA_s$ are the data, four components, 1% correction factor and maximum iterations of 100.

2. Generation of Needle Spectra

The needle spectra are the starting estimates from which QITTFA generates the pure profiles (the needle spectra are not unique to curve resolution and have been used over several years to provide starting estimates in the absence of *a priori* information). Due to the dual space of the measurement matrix, QITTFA can either be applied in the spectral or concentration dimension, abbreviated to $QITTFA_s$ and $QITTFA_c$ respectively. For a measurement matrix, **D**, of size $(n \times m)$, with *n* absorption spectra in each row at *m* wavelengths. The initial estimates for the spectral domain ($QITTFA_s$),

would have *m* needle spectra $\mathbf{in}_j\ (1\times m)$. Each needle spectrum has a spike equal to unit length one, at a single position, which is unique to each needle spectrum and the remaining elements are equal to zero. The *m* needle spectra would be expressed as follows:

$\mathbf{in}_1$ = (1, 0, 0, ….,0)

$\mathbf{in}_2$ = (0, 1, 0, ….,0)

$\mathbf{in}_m$ = (0, 0, 0, ….,1)

The size of the needle matrix, $\mathbf{IN_s}$, is $(m\times m)$.

If QITTFA$_c$ is applied to the measurement matrix, $\mathbf{D}\,(n\times m)$, *n* needle spectra $\mathbf{in}_i\,(1\times n)$ would be generated, with a spike equal to unit length one and the remaining elements equal to zero, and the size of the needle matrix $\mathbf{IN_c}$, is $(n\times n)$.

The properties of the needle spectra are that they are linearly independent, and so represent *n* or *m* independent components. The response at each variable is maximised to a unit length, i.e., one, so that the variance contribution from each needle spectrum is the same. A typical needle spectrum is shown in figure 41.

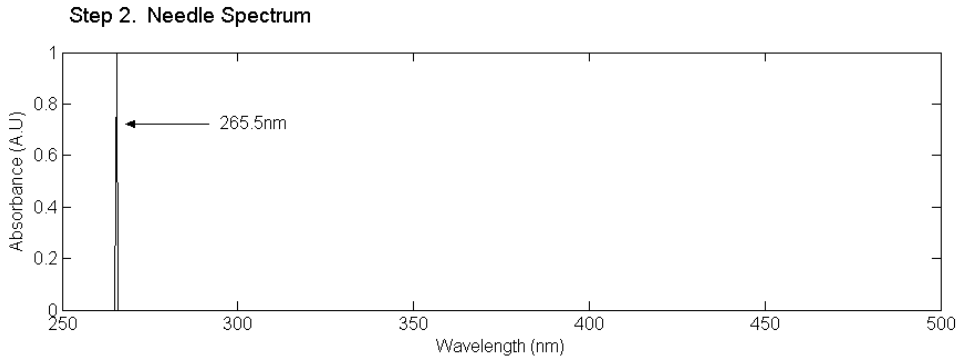


Figure 41. Needle spectrum with spike at 265.5nm

3. Generation of Needle Output Spectra

Singular value decomposition (SVD) is used to decompose the measurement matrix. The SVD of measurement matrix **D** is defined by equation 39, where **D** is the measurement matrix, **U** is the $\mathbf{U}(n \times \underline{n})$ column-orthonormal singular vectors and **V** is the $(m \times \underline{n})$ row–orthonormal singular vectors and is a $(\underline{n} \times \underline{n})$ diagonal matrix of singular values whose elements are arranged in decreasing value. The dimension of $\underline{n}$ can be at most equal to or smaller than the dimensions *n* or *m*.

$$\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathbf{T}} \qquad \textbf{Equation 39}$$

The solutions space is defined by the estimated number of chemical species. The chemical rank is equal to the mathematical rank (i.e., the number of independent components) when no noise is present in the data and the chemical species are independent. The mathematical rank of the data can easily be determined by the reduction of the matrix to row-echelon form by Gaussian elimination, and is equal to the number of non-zero rows. The determination of the chemical species is more complex because of the presence of measurement noise and their non-assumed distribution, heteroscedasticity and collinearity of the measurement matrix [11, 59]. Once the number of components have been determined, the reduced singular vectors are written as shown in equation 40, where $\mathbf{U}^{*}(n \times nc)$, $\mathbf{V}^{*\mathbf{T}}(nc \times m)$ and $\mathbf{\Lambda}^{*}(nc \times nc)$ are the independent singular vectors and **E** is the $(n \times m)$ residual matrix, (*) refers to the reduced singular vectors and *nc* refers to the number of independent components which are retained in the model.

$$\mathbf{D} = \mathbf{U}^{*}\mathbf{\Lambda}^{*}\mathbf{V}^{*\mathbf{T}} + \mathbf{E} \qquad \textbf{Equation 40}$$

4. Optional Application of Constraints

The needle output spectral matrix is constructed to store the refined estimates for each of the projected and constrained needle spectra. The formation of the needle output spectral matrix in the QITTFA$_s$ procedure is described below.

Each row spectrum in the needle spectral matrix, $\mathbf{IN}(m \times m)$, represents a spectral estimate. The needle spectra are each tested within the space defined by the reduced row-orthonormal basis vectors (loadings) of the data matrix, see equation 40. A row spectrum of **D** would be written as, equation 41, where $\mathbf{d}_i$ is a $(1 \times m)$ vector, $\mathbf{w}^*$ is a $(1 \times nc)$ vector, $\mathbf{L}^{*\mathbf{T}}$ is a $(nc \times m)$ vector and $\mathbf{e}_i$ is a $(1 \times m)$ vector. Each row represents either a mixture or pure spectrum, which is a linear combination of the singular vectors; which span the space of all the components. The first needle spectrum, $\mathbf{in}_1$, is projected into the basis vectors defined by the reduced loadings. The scores $\mathbf{w}^*_{in_1}$ of this needle spectrum, $\mathbf{in}_1$, is calculated by solving equation 42 using equation 43. The scores, $\mathbf{w}^*_{in_1}$ of the needle spectrum gives the linear combination of singular vectors which best describe the initial needle spectrum, as shown in figure 42.

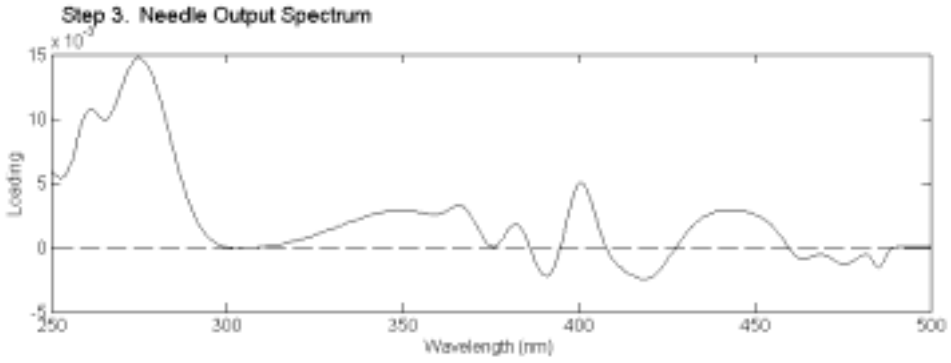


Figure 42. Needle output spectrum calculated for the needle spectrum with spike at 265.5nm

$$\mathbf{d}_i = \mathbf{w}^* \mathbf{L}^{*\mathbf{T}} + \mathbf{e}_i \qquad \textbf{Equation 41}$$

$$\mathbf{in}_1 = \mathbf{w}^*_{in_1}\mathbf{L}^{*\mathbf{T}} + \mathbf{e}_1 \qquad \textbf{Equation 42}$$

$$\mathbf{w}^*_{in_1} = \mathbf{in}_1\mathbf{L}^*\left(\mathbf{L}^{*\mathbf{T}}\mathbf{L}^*\right)^{-1} = \mathbf{in}_1\mathbf{L}^* \qquad \textbf{Equation 43}$$

The needle output spectrum $\mathbf{out}_1\left(\mathbf{out}_1 = \mathbf{w}^*_{in_1}\mathbf{L}^{*\mathbf{T}}\right)$, described by the least squares estimate $\mathbf{w}^*_{in_1}$, may contain some sort of deviation from the generic characteristics of the system, i.e., negativity. To correct for this, a non-negativity constraint can be applied to the loadings vector $\mathbf{out}_1$, to ensure that the solution contains no negative signals, (if this is a fulfilment of the measured system). This constraint is applied by setting all negative values to zero, prior to the proceeding projection. The constrained needle output spectrum $\mathbf{in}'_1\left(\mathbf{in}'_1 = \mathbf{w}^*_{in_1}\mathbf{L}^{*\mathbf{T}} + \mathbf{e}_1\right)$ is then projected into the reduced singular vector space and the needle output spectrum $\mathbf{out}'$ is assessed. This procedure is repeated until the difference between $\mathbf{in}'_1$ and $\mathbf{out}'_1$ is representative of the noise or the maximum number of iterations has been exceeded, as shown in figure 43. The overall expression is given in equation 44.

$$\mathbf{out}_1 = \mathbf{in}_1\mathbf{L}^*\mathbf{L}^{*\mathbf{T}} \qquad \textbf{Equation 44}$$

This is repeated for all the needle spectra $\mathbf{in}_1$, $\mathbf{in}_2$,…, $\mathbf{in}_m$.

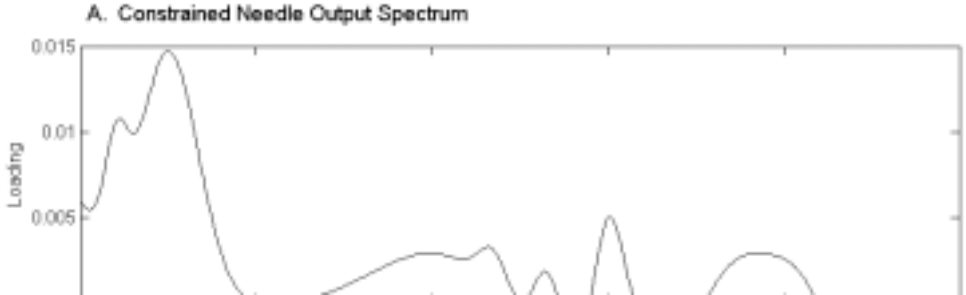


Figure 43. Constrained needle output spectrum. Figure 44a. Non-negative constrained needle output spectrum. Figure 44b. Final non-negative constrained needle output spectrum

5. Production of needle Output Matrix

The constrained needle output spectra for each projected needle spectra are arranged sequentially into an $\mathbf{Z_s}\,(m \times m)$ needle output spectral matrix, as shown in figure 44.

$\mathbf{z}_1\ (1 \times m) = \mathbf{out_1} = \mathbf{in_1}\ \mathbf{V}^*\ \mathbf{V}^{*\mathbf{T}}$

$\mathbf{z}_2\ (1 \times m) = \mathbf{out_2} = \mathbf{in_2}\ \mathbf{V}^*\ \mathbf{V}^{*\mathbf{T}}$

.

.

.

$\mathbf{z}_n\ (1 \times m) = \mathbf{out}_n = \mathbf{in}_n\ \mathbf{V}^*\ \mathbf{V}^{*\mathbf{T}}$

The $\mathbf{Z_c}\ (n \times n)$ needle output spectral matrix, representative of the concentration profiles, is calculated in a similar way to the needle output spectral matrix representative of the spectral profiles. The needle output spectrum $\mathbf{in}_1\,(1 \times n)$ is projected into the reduced column-orthonormal vectors (scores), see equation 45. The loadings which best describe the initial needle spectrum is described by equation 45 and solved using

equation 46. The overall expression of QITTFA$_c$ for concentration estimate is expressed in equation 47. An additional constraint to QITTFA$_c$ for the concentration domain include the average unimodality constraint [116].

$$\mathbf{in}_1 = \mathbf{W}^{*}\mathbf{l}_{in_1}^{*\mathbf{T}} + \mathbf{e}_1 \qquad \textbf{Equation 45}$$

$$\mathbf{l}_{in_1}^{*\mathbf{T}} = \left(\mathbf{W}^{*\mathbf{T}}\mathbf{W}^{*}\right)^{-1}\mathbf{W}^{*\mathbf{T}}\mathbf{in}_1 = \mathbf{W}^{*\mathbf{T}}\mathbf{in}_1 \qquad \textbf{Equation 46}$$

$$\mathbf{out}_1 = \mathbf{W}^{*\mathbf{T}}\mathbf{W}^{*}\mathbf{in}_1 \qquad \textbf{Equation 47}$$

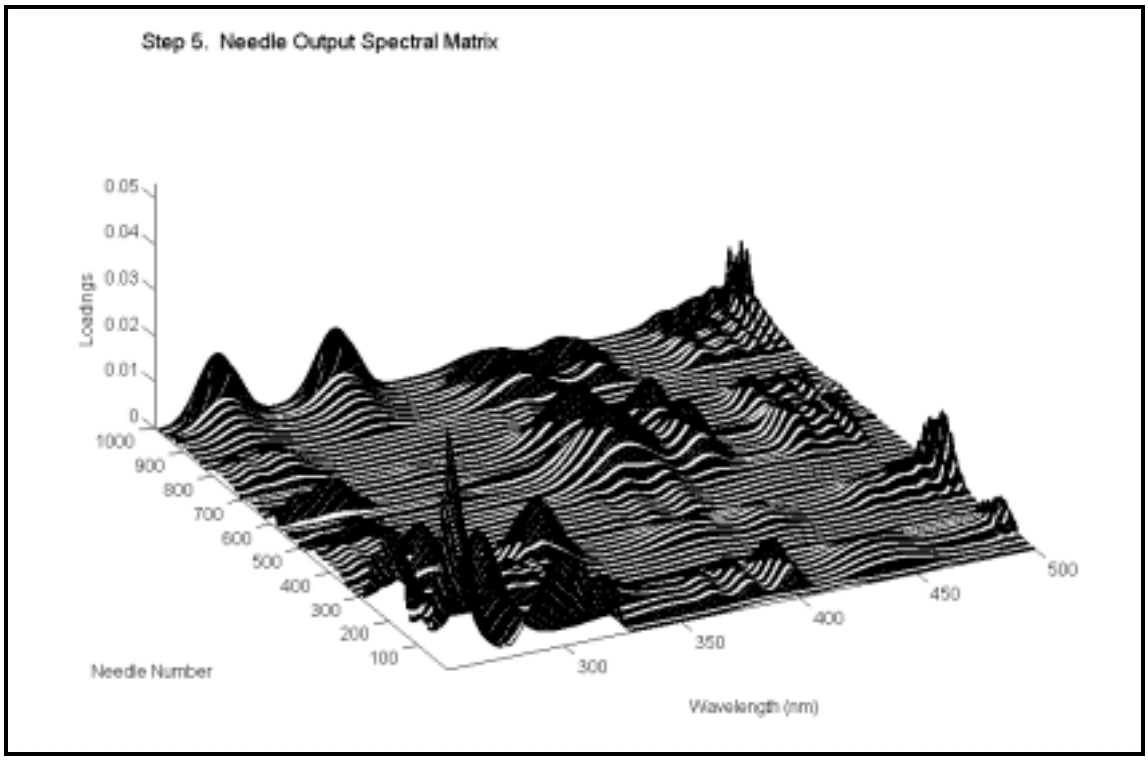


Figure 44. Needle output spectral matrix

The refined estimates in the needle output matrix, $\mathbf{Z_s}$ or $\mathbf{Z_c}$, are often closer to the actual solution than randomly generated initial estimates, because the initial estimates are defined from the reduced singular vector space, i.e., the space from which the majority of noise has been removed. The variance contribution from each of the components in the needle output matrix are maximised, which makes it easier to resolve minor

components and the needle output spectra can be optionally constrained with characteristic constraints prior to ALS analysis.

6. Selection of Initial Estimates from Solutions Matrix

The method of selecting initial estimates from one or the other $\mathbf{Z_s}$ or $\mathbf{Z_c}$ matrix is described below.

For the selection of the initial spectral estimates from the $\mathbf{Z}_s$ matrix, the needle output spectra are arranged in the needle output spectral matrix, $\mathbf{Z_s}$ such that each row represents a refined spectral estimate for the determination of the pure spectral profiles. SIMPLISMA is used to determine the purest needle variables from the needle output spectral matrix $\mathbf{Z_s}$ to provide the initial estimates of the spectra, shown in figure 45.

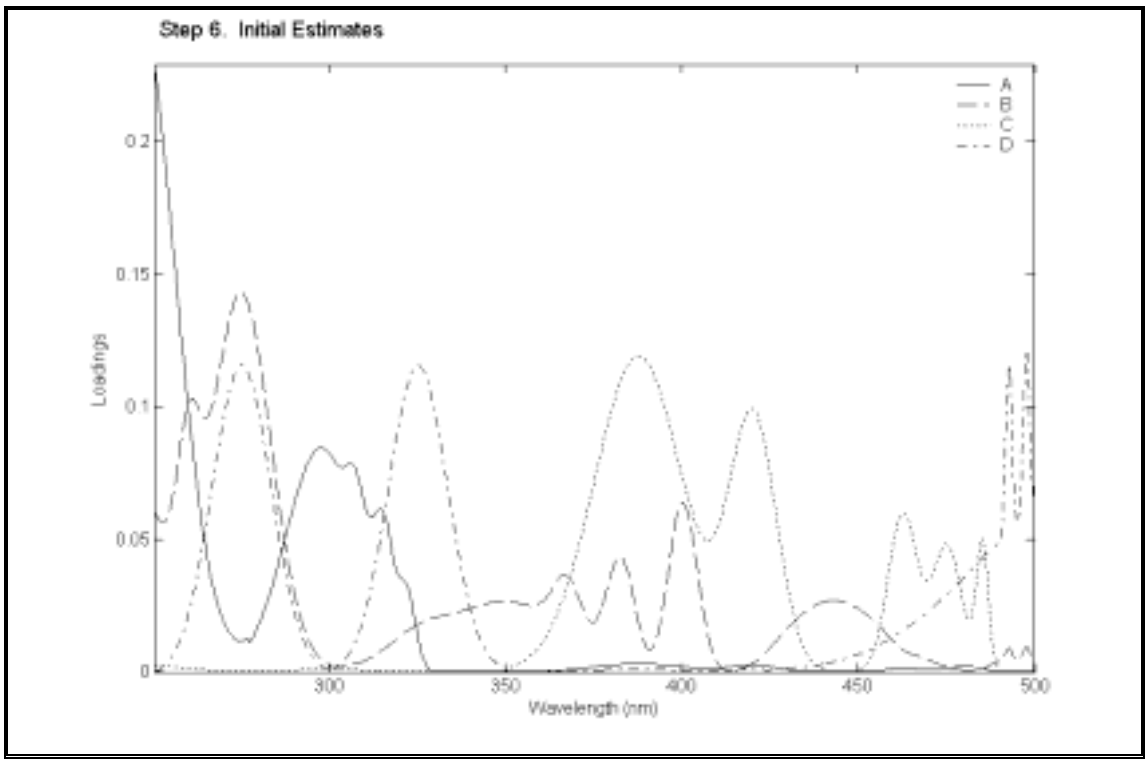


Figure 45. Initial spectral estimates selected from needle output matrix

For the selection of the initial concentration estimates from the $\mathbf{Z}_c$ matrix, the needle output concentration profiles are arranged in the needle output spectral matrix, $\mathbf{Z}_c$ such that each column represents a refined concentration estimate for the determination of the pure concentration profiles. SIMPLISMA is used to determine the purest needle variables from the needle output concentration matrix $\mathbf{Z}_c$ to provide the initial estimates of the concentration.

Either the spectra or concentration estimates can be used to initialise the ALS procedure. The advantage of the determination of the initial estimates from one or the other needle output matrix, $\mathbf{Z}_s$ or $\mathbf{Z}_c$, is that the majority of the variance contribution due to noise has been removed. Secondly, SIMPLISMA selects sub-matrices from one or the other $\mathbf{Z}_s$ or $\mathbf{Z}_c$ matrix that spans the space defined by the independent variables, which allows each component to be easily resolved.

The performance of the QITTFA$_c$ method was initially compared with its counter-part exploratory tool, EFA, using a noise free simulated dataset (simulated HPLC-DAD (I)) to highlight the importance of the generation of accurate initial estimates. In the second comparison, the determination of the key variables from the needle output spectral matrix (QITTFA$_s$) was compared with the determination of the key variables from the original matrix (SIMPLISMA). The analysis was completed in order to find out whether the construction of the needle output spectral matrix improved the selection of the pure variables and hence the final MCR-ALS solution. This analysis was completed with the simulated HPLC-DAD (II) dataset.

II.5 Exploratory Analysis of Simulated HPLC-DAD using QITTFA

II.5.1 Qualitative Analysis of the Simulated HPLC-DAD (1) Data

The simulated (I) HPLC-DAD data consisted of 51 spectra following a pseudo elution of four analytes (**A-D**), shown in figure 46. The data was generated as part of an example used for resolution of multi-batch measurement matrices using MCR-ALS [107]. No selectivity exists in the spectral direction and fully selective conditions exist in the concentration dimension. Combinations of these two conditions are normally required for good resolution. The spectra have 96 spectral channels from 200-295nm (resolution of 1nm) and the size of the data is 51×96. The residence times of species **A-D** were 13-31 minutes, 21-41 minutes, 31-47 minutes and 4–26 minutes respectively.

MATLAB6p5® (The Math Works, Inc) was used to complete all data processing.

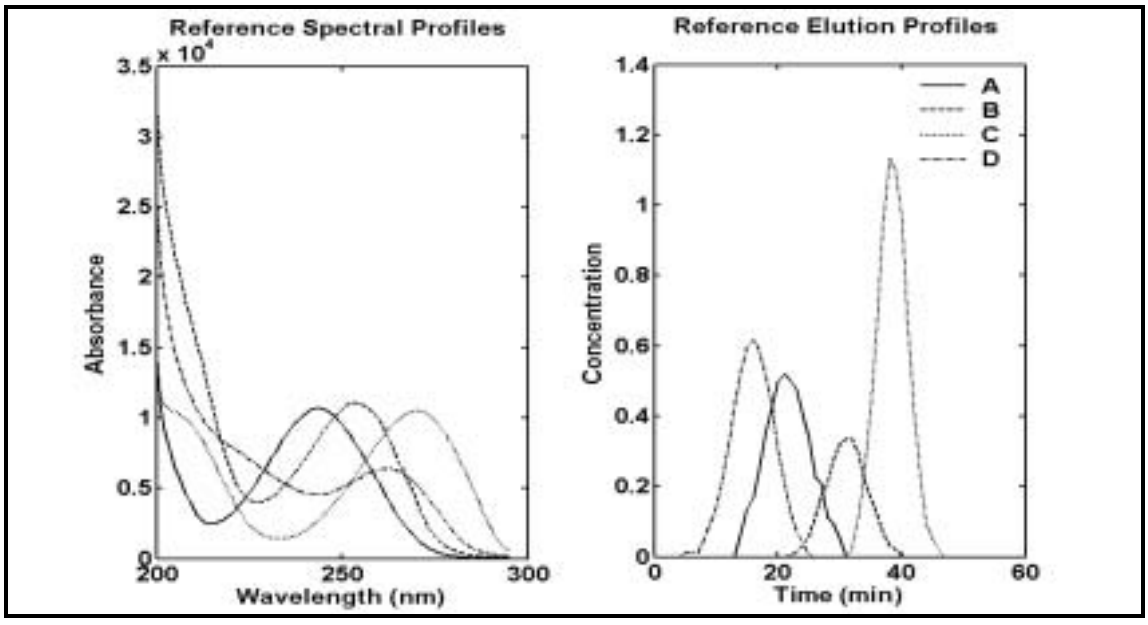


Figure 46. The reference spectra and elution profiles of species A-D.

II.5.1.1 Application of QITTFA$_c$ Concentration to the Simulated HPLC-DAD (I) Data

II.5.1.2 Simulated HPLC-DAD (I) Data

No observable unique features could be distinguished from preliminary observations of the surface plot of the simulated HPLC-DAD (I) dataset, shown in figure 47. The MCR-ALS solution generated from the QITTFA$_c$ and EFA starting estimates were expected to be close to the actual solution because regions of low local rank (i.e., a rank of 1 or 2) existed in the concentration direction for a good resolution.

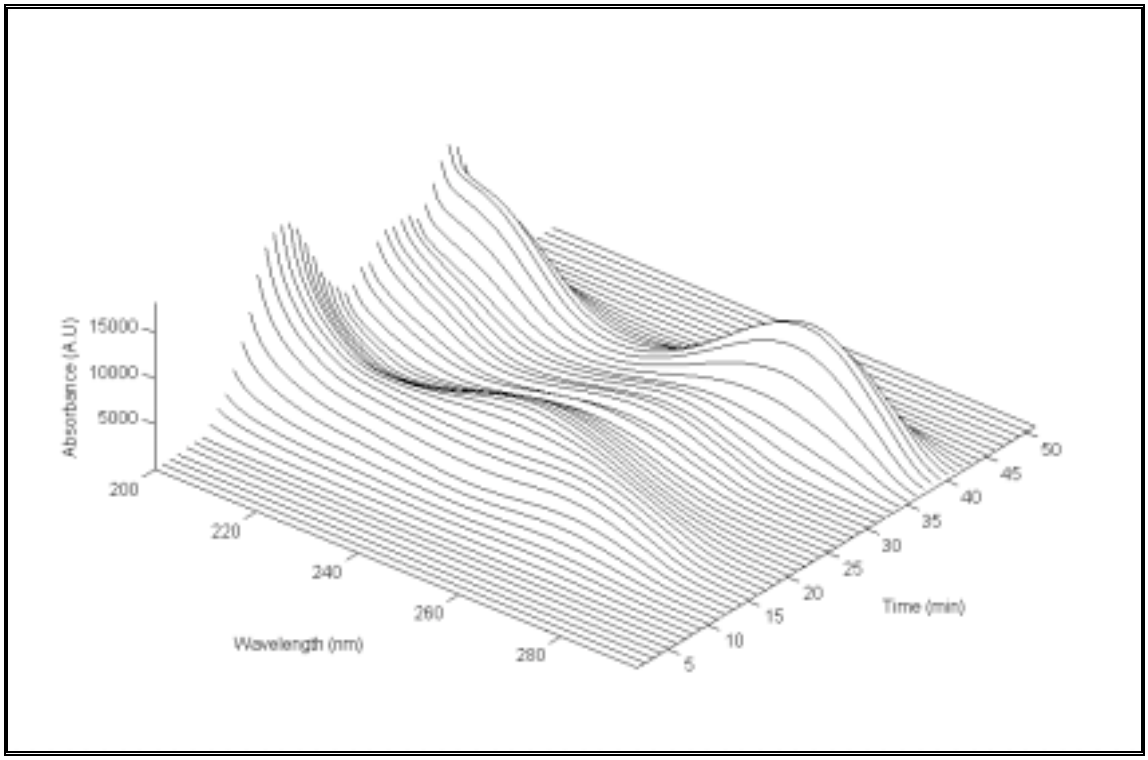


Figure 47. Simulated HPLC-DAD (I) data. No selectively is present in the spectral profiles and local rank conditions are present in the elution profiles.

II.5.1.3 Initial Estimates

In the QITTFA$_c$ procedure fifty-one independent needle vectors representative of elution profiles were automatically generated (input vectors) and sequentially projected

into the reduced scores space, to obtain the output solution matrix, $\mathbf{Z_c}$, shown in figure 48.

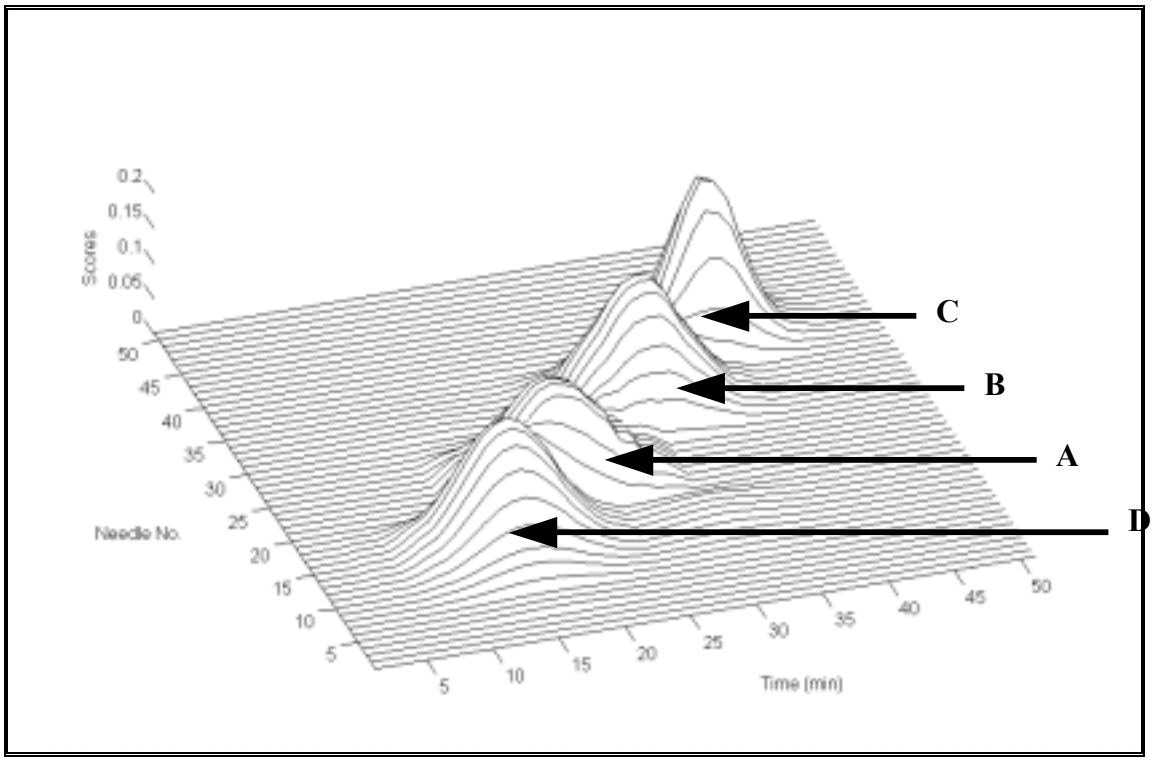


Figure 48. The output matrix, Z_c contained fifty-one possible solutions, one for each needle spectrum projected and constrained using non-negativity and average unimodality constraint within the reduced scores space.

The constraints imposed during the QITTFA$_c$ procedure were non-negativity and the average unimodality constraint (see section I.2.1.2.4), because the solutions were expected to have absorbance values equal to or greater than zero and Gaussian shaped profiles. The solution matrix, $\mathbf{Z_c}$, contained fifty-one possible solutions, one for each needle vector projected. The surface plot of the solution space is similar to a contour plot of a DAD; however, the absorbance is plotted as a function of the retention time and the needle variable.

The percent variance of the principal components and the rank of both the simulated HPLC-DAD (I) data and the needle output spectral matrix, $\mathbf{Z_c}$ were calculated and are shown in table 10. The rank of the simulated data and the needle output spectral matrix,

$\mathbf{Z_c}$, was four. The variance was spread more evenly across the PCs of the needle output spectral matrix. The condition number of the HPLC-DAD (I) data and the needle output spectral matrix ($\mathbf{Z_c}$) were calculated. The condition number of a matrix measures the sensitivity of the solution of a system of linear equations to errors in the data. It gives an indication of the accuracy of the results from matrix inversion and the linear equation solution.

% Variance	**PC 1**	**PC 2**	**PC 3**	**PC 4**	**PC 5**
HPLC- DAD (I)	90.24	8.43	0.92	0.41	0.00
$\mathbf{Z_c}$	40.95	30.39	19.10	9.56	0.00

Table 10. The percent variance distribution of PCs 1-5, calculated for the simulated HPLC-DAD (I) data and the needle output spectral matrix.

The condition number was calculated by dividing the maximum singular value with the smallest singular value, for each measurement matrix. The condition number of the simulated HPLC-DAD (I) data was 14.91 and the needle output spectral matrix was 2.07, which indicated that the quality of the needle output spectral matrix and hence the final least squares solutions had improved over the original data, in terms of the sensitivity to error and accuracy of the linear equation solution.

The needle variables were selected from the solutions matrix. The maximum intensity in the first purity spectrum was found at 38 minutes, which coincided with the t_{max} of component **C**. The second pure variable which was least correlated to the first component, was located in the second purity spectrum at 30 minutes. The second pure needle variable was close to the t_{max} of constituent **B**. The third purest needle variable was located at 17 minutes in the third purity spectrum and coincided with components **A** and **D**, however, the pure needle variable was closer to the t_{max} of component **D** than **A**. The final pure needle variable was found at 21 minutes. Thus, the pure needle variables selected for components **A-D** were 21, 30, 38 and 17 respectively. The

reference elution profiles are shown in figure 49a and the initial estimates selected from the solutions matrix are shown in figure 49b. Good starting estimates of the elution profiles were obtained from the needle output spectral matrix, because it was assumed that the principal axes (basis vectors) coincided with the independent components. The difference between the QITTFA$_c$ profiles and the reference profiles were due primarily to differences in the scaling of the component profiles. The evolutionary profiles estimated using the EFA approach are given in figure 49c. The plot of the log eigenvalue versus time in the forward direction revealed new independent components at 5, 13, 22 and 32 minutes while the plot of the backward analysis showed the disappearance of the independent components at 26, 31, 41 and 48 minutes. The combined forward and reverse profiles from EFA revealed the windows of existence of the four independent components. It was observed that the EFA profiles did not closely represent the pure elution profiles because of difference between the shapes of the profiles, but the QITTFA$_c$ elution profiles closely represented the pure elution profiles.

The elution windows predicted for components **A-D** using QITTFA$_c$ and EFA are tabulated in table 11. Deviation from the reference data was shown in the elution profile of the QITTFA$_c$ constituent **D**, which had a weak tail peak between 25-35minutes, highlighted in figure 49b.

Elution windows (min)	**A**	**B**	**C**	**D**
Reference	13-31	21-41	31-47	4–26
QITTFA$_c$	13-31	21-40	31-47	4-35
EFA	13-31	22-41	32-48	5-26

Table 11. Elution windows of components A-D, predicted using QITTFA$_c$ and EFA compared to the reference elution windows.

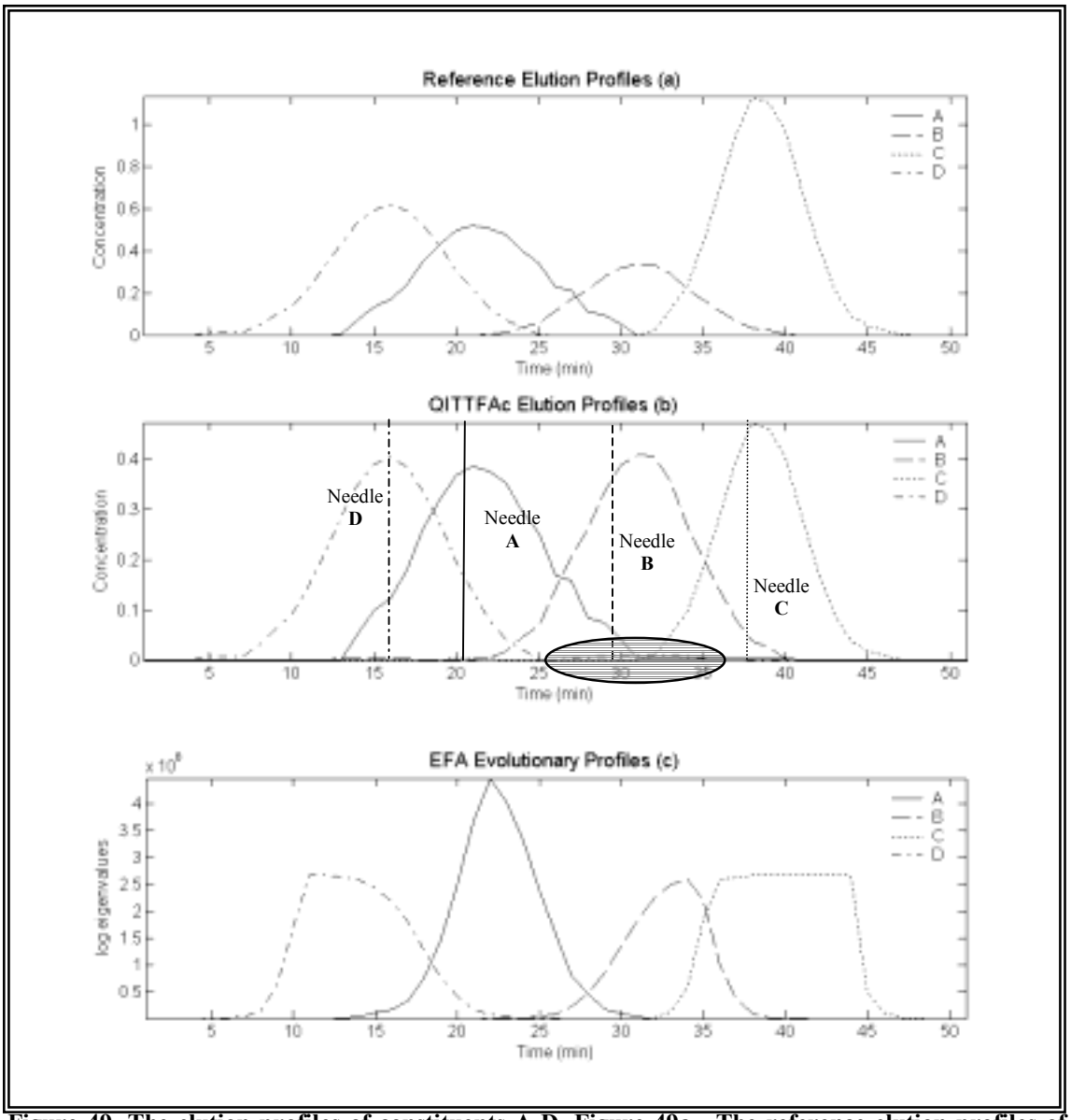


Figure 49. The elution profiles of constituents A-D. Figure 49a. The reference elution profiles of constituents A-D. Figure 49b. The QITTFA$_c$ elution profiles; selected from the needle output spectral matrix, rank 4. Figure 49c. The EFA estimates of the elution profiles.

II.5.1.4 MCR-ALS

The MCR-ALS solution obtained with the QITTFA$_c$ and EFA starting estimates are shown in figure 50. During the ALS optimisation the concentration profiles were constrained with non-negativity and vertical unimodality constraints. The spectral profiles were normalised to a height of one and non-negativity constraints were applied. The model error was measured using the relative error between successive iterations.

The convergence criterion was 0.1% of the difference between the MCR-ALS reconstructed matrix and the PCA constructed matrix. The maximum number of iterations was set to 100. The resolved spectral and concentration profiles were compared to the reference profiles by scaling the response of each component between 0 and 1. The reference and predicted profiles were overlaid for visual comparison and the relative error (RE %) which gives a measure of the quality of fit between the predicted and reference concentration and spectral profiles was determined.

The resolved elution profiles generated from the QITTFA$_c$ starting estimates showed no observable deviations from the reference elution profiles, see figure 50a. This was also observed in the resolved spectral profiles generated from the QITTFA$_c$ starting estimates, see figure 50c. On the-other-hand, the resolved elution profiles obtained using the EFA estimates showed observable deviations in the prediction of the elution profiles of constituents **A**, **B** and **D**. This was particularly noticeable between 13-18 minutes in component **A**, 21-32 minutes and 34-45 minutes in component **B** and 15-25 minutes in component **D**, highlighted in figure 50b. The resolved spectral profiles obtained using the EFA starting estimates did not agree with the reference spectral profiles of the constituents **A**-**D**, which is highlighted in figure 50d and is reflected in the percent RE of each of the constituents, see table 12.

	Elution Profiles		***Spectral Profiles***	
%RE	**QITTFA$_c$**	**EFA**	**QITTFA$_c$**	**EFA**
A	0.40	4.00	0.24	12.97
B	0.12	9.08	0.22	0.84
C	0.00	0.03	0.05	2.92
D	0.15	10.38	0.00	1.57

Table 12. The percent RE calculated for the prediction of the elution and spectral profiles for each constituent. The solutions were obtained from the constrained ALS procedure using the QITTFAc and EFA starting estimate.

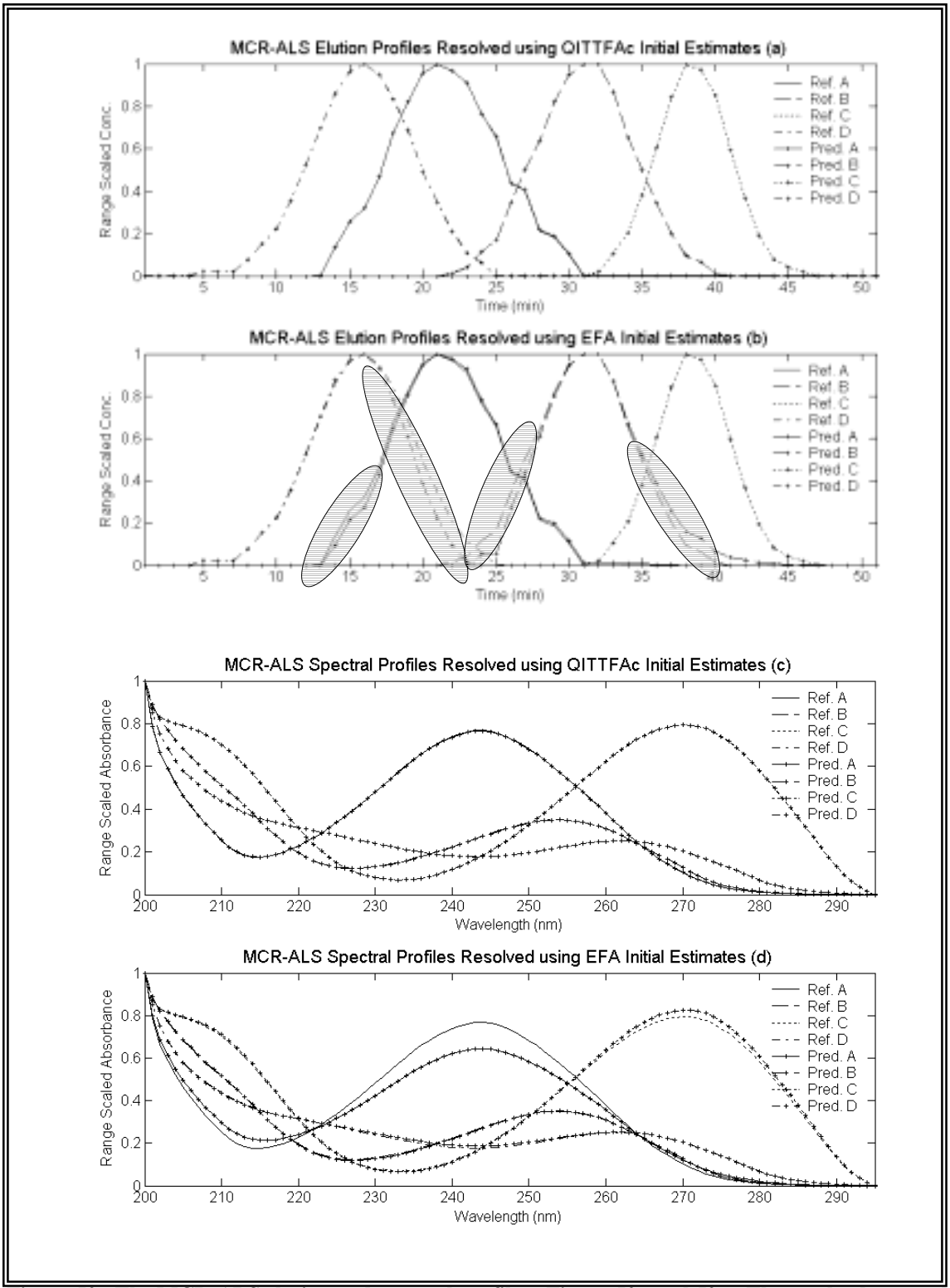


Figure 50. The MCR-ALS elution and spectral profiles. Figure 50a and 50c - The resolved spectral and elution profiles using the QITTFA$_c$ initial estimates overlaid with the reference elution profiles for constituents A-D respectively. Figure 50b and 50d - The MCR-ALS elution and spectral profiles resolved using the EFA estimates overlaid with the reference spectral profiles for constituents A-D respectively.

The percent RE of each of the constituents concentration and spectral profiles resolved using the QITTFA$_c$ starting estimates varied between 0-0.50%. The best QITTFA$_c$ result was found for the resolved spectral profile of constituent **D** which contained no error. The percent RE of each of the constituents concentration and spectral profiles resolved using the EFA starting estimates varied between 0.03% and 12.97%. Therefore, using the QITTFA initial estimates, which closely represented the actual solution, produced results with little or no deviation from the true solution.

II.5.1.5 Summary

The solutions obtained using the QITTFA$_c$ starting estimates closely resembled the actual elution profiles. The evolutionary profiles determined using EFA analysis were abstract representations of the concentration profiles, and as such the starting estimates were not as accurate as those determined using the QITTFA$_c$ procedure. Here, the importance of the generation of accurate initial estimates approximating the true solution has been shown, as the error in the predicted concentration and spectral profiles using the EFA starting estimates were greater than those obtained with the QITTFA$_c$ starting estimates.

II.5.2 Qualitative Analysis of the Simulated HPLC –DAD (II) Data

II.5.2.1 Simulation(II) – Partial Selectivity in spectral direction

Simulated dataset (II) consisted of 100 HPLC-DAD spectra following a pseudo elution of the target analyte, and three by-products, which eluted with the main product, species **A-D**, shown in figure 51. The spectra have 500 spectral channels from 250.5 - 500.0nm (resolution equal to 0.5nm) and the size of the data is 100×500. The chromatographic profiles and the spectral profiles were created using Gaussian functions.

MATLAB6p5® (The Math Works, Inc) was used to complete all data processing.

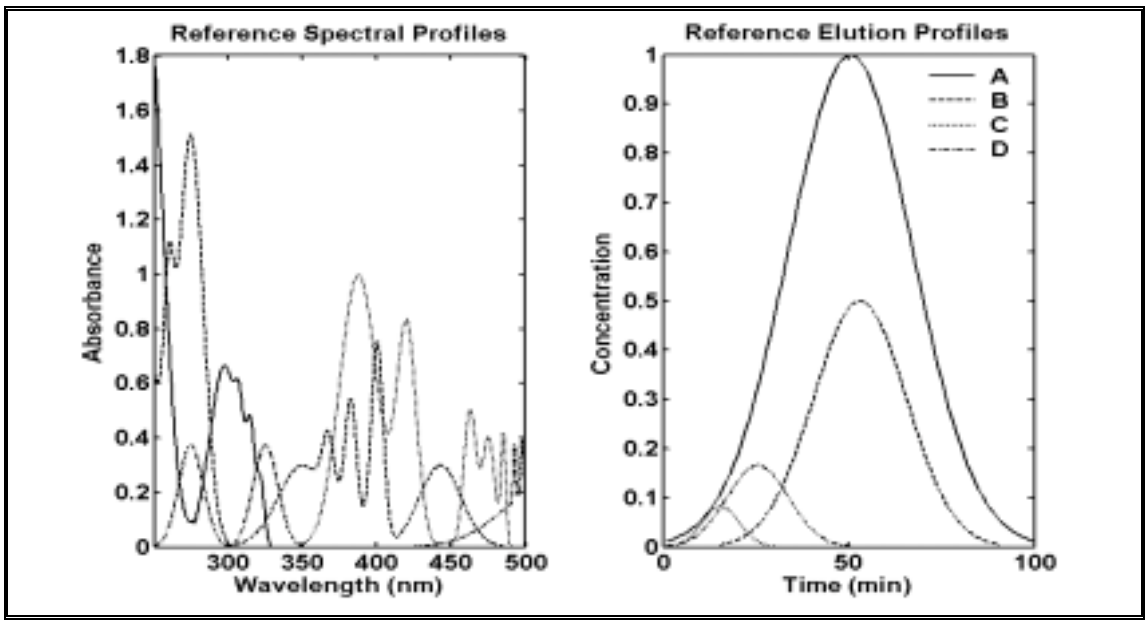


Figure 51. The reference spectra and elution profiles of species A-D.

The reaction profile of the simulated HPLC-DAD data (II) is shown in the schematic overview of the QITTFA$_s$ procedure, step 1 (Chapter II.4.2).

II.5.2.2 Initial Estimates

In the QITTFA$_s$ procedure a reduced loadings space of rank 4 (number of independent components) was used to obtain the needle output spectral matrix, $\mathbf{Z_s}$. The constraint imposed during the QITTFA$_s$ procedure was non-negativity, as the solutions were expected to have absorbance values equal to or greater than zero. The output matrix, $\mathbf{Z_s}$, contained five hundred possible spectral solutions, one for each needle vector projected. The four purest spectra were extracted from the $\mathbf{Z_s}$ matrix at needles 10, 273, 174 and 31, the original needle input spectra at these needles had a spike at wavelength variables 255.0nm, 386.5nm, 377.0nm and 265.0nm respectively. The needle spectra were representative of components **A**, **C**, **D** and **B** respectively. In figure 52a, the output of the selected needle spectra are overlaid with their corresponding reference spectra to determine the selectivity of the key variables, based upon absorptivity. SIMPLISMA was applied to the simulated HPLC-DAD (II) data and the four purest elution profiles were selected at wavelength variables 275.0 nm, 498.0 nm, 417.5 nm and 300.0 nm, which were representative of components **B**, **D**, **C** and **A** respectively. In figure 52b, the key variables are overlaid with their corresponding reference spectra.

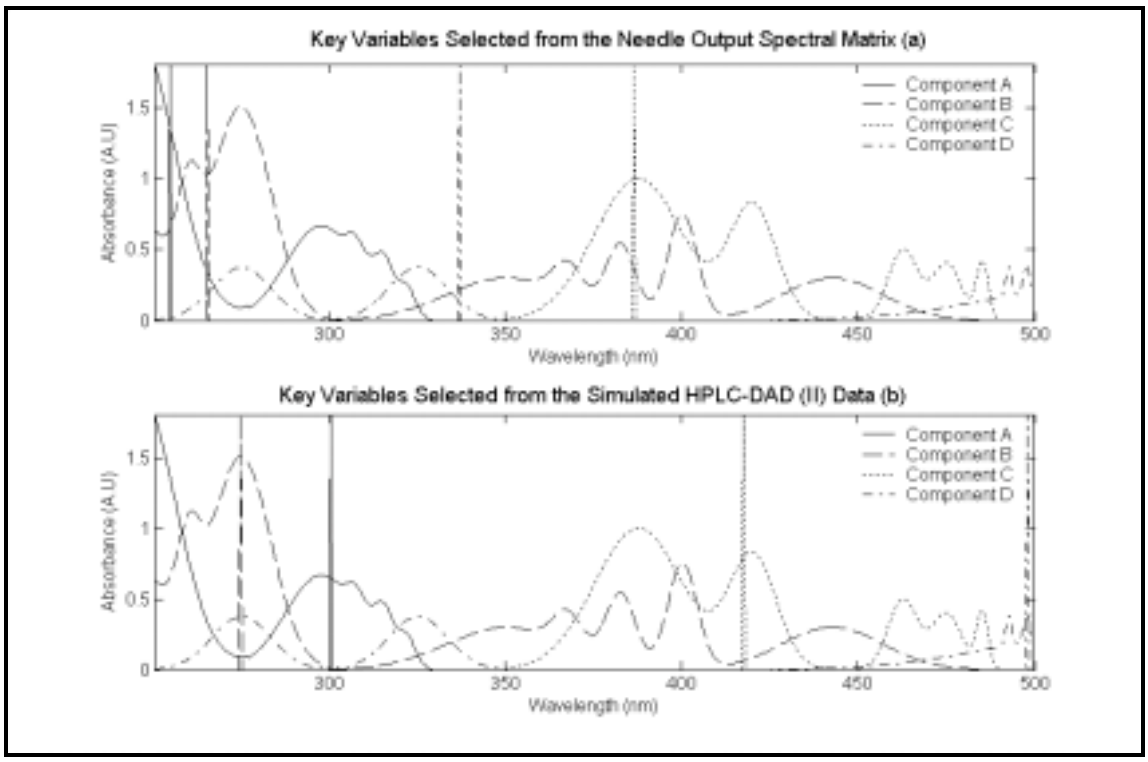


Figure 52. The key variables selected from the needle output spectral matrix and the original HPLC-DAD (II) data. Figure 52a. The key variables selected from the needle output spectral matrix. Figure 52b. The key variables selected from the original HPLC-DAD (II) using the SIMPLISMA analysis.

Purity (max)	**P1**	**P2**	**P3**	**P4**
QITTFA$_s$	1.907	1.279	1.092	0.443
SIMPLISMA	0.964	0.393	0.066	0.002

Table 13. The purity value of the key variables (P1-P4) determined from the purity spectra 1-4 calculated from the needle output spectral matrix and the simulated HPLC-DAD (II) data.

The key variables selected from the needle output spectral matrix differed to the key variables selected from the original matrix in two aspects. Firstly, the QITTFA$_s$ key variables had higher purity values than the SIMPLISMA key variables. Secondly, the pure variables were easier to select from the needle output spectral matrix because the variance contribution for each spectrum in the matrix had been maximised, whereas the purity of the original data was largely dependent on the variance contribution of each of the constituents in the original matrix (see table 13).

The initial QITTFA$_s$ spectral estimate and the elution profiles estimated using the SIMPLISMA analysis are shown in figure 53. It is clear that the correct key variables were selected from the needle output matrix as the initial QITTFA$_s$ estimates were similar to the reference spectra in terms of shape. The final key variable selected from the original matrix using the SIMPLISMA procedure was incorrect. This was because the final purity spectrum had a relatively low signal-to-noise ratio, so it was difficult to distinguish the key variable from the final purity spectrum. As such, the wrong key variable was selected for constituent **C**.

II.5.2.3 MCR-ALS

The MCR-ALS solutions obtained using the QITTFA$_s$ and SIMPLISMA initial estimates are given in figures 54-55. During the ALS optimisation the concentration profiles were constrained with non-negativity and vertical unimodality constraints. The spectral profiles were normalised to a height of one and non-negativity constraints were applied. The model error was measured using the lack of fit between successive iterations. The convergence criterion was 0.1% for the difference between the MCR-ALS reconstructed matrix and the PCA constructed matrix. The maximum number of iterations was set to 100. The resolved spectral and concentration profiles were compared to the reference profiles by scaling the response of each component between 0 and 1. The reference and predicted profiles were overlaid for visual comparison and the determination of the relative error (RE %).

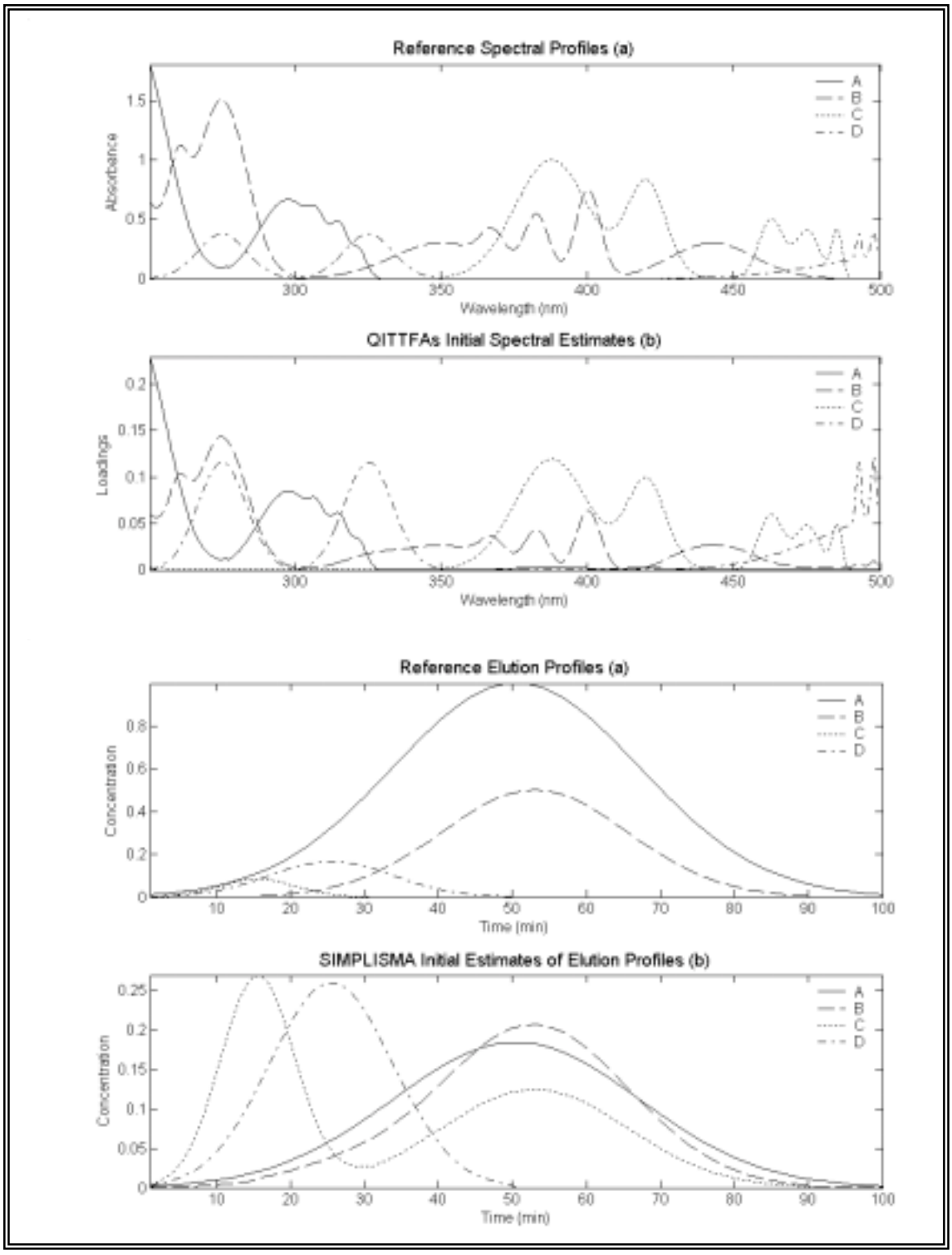


Figure 53. The initial estimates of the spectral profiles and elution profiles. Figure 52a. The reference elution profiles of constituents A-D. Figure 53b. The QITTFA$_s$ elution profiles; selected from the reduced solution space of rank 4. Figure 53c. The SIMPLISMA estimates of the elution profiles.

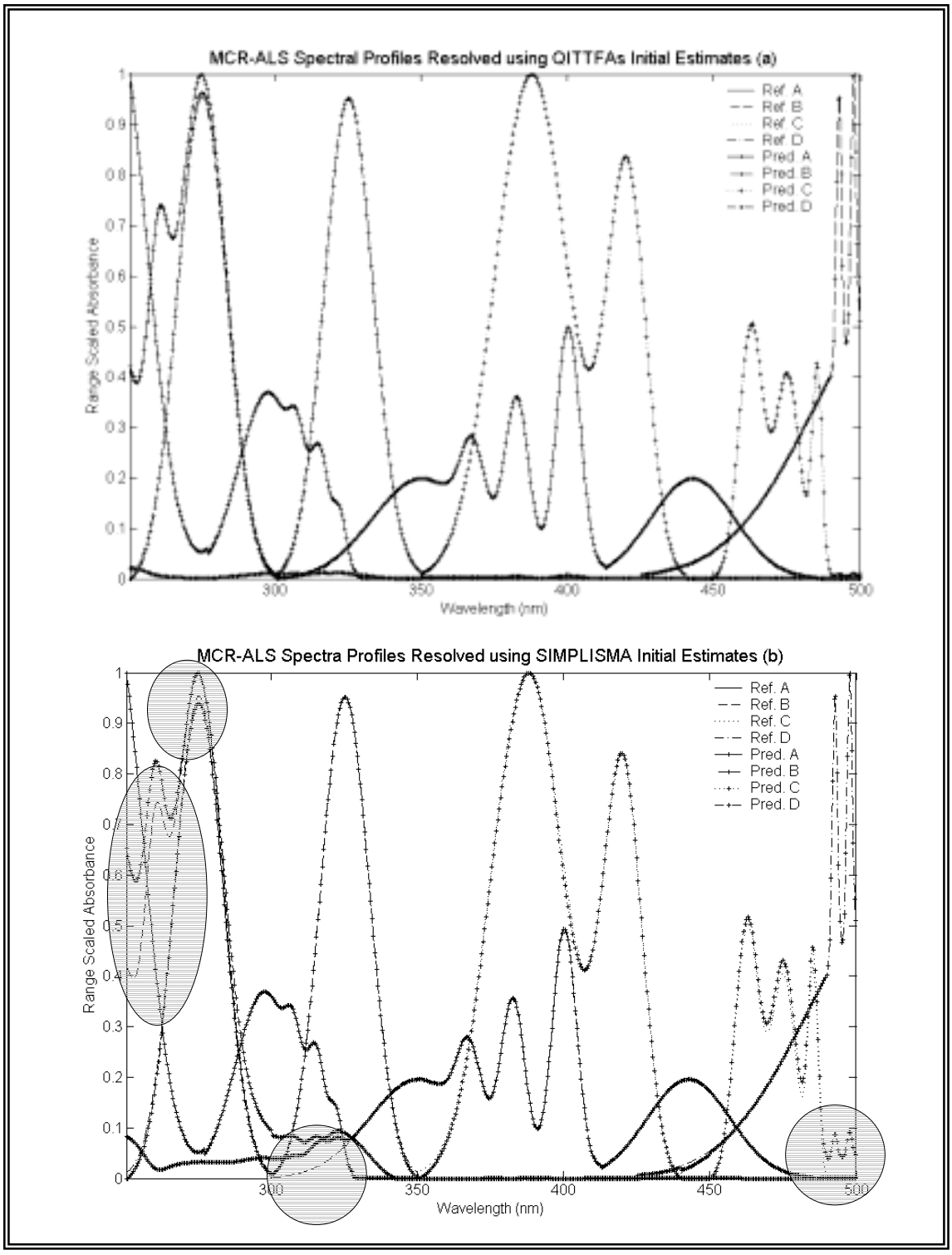


Figure 54. The MCR-ALS resolved spectral profiles. Figure 54a. The MCR-ALS spectral profiles resolved using QITTFA$_s$ initial estimates overlaid with the reference spectral profiles for constituents A-D. Figure 54b. The MCR-ALS spectral profiles resolved using SIMPLISMA initial estimates overlaid with the reference spectral profiles for constituents A-D.

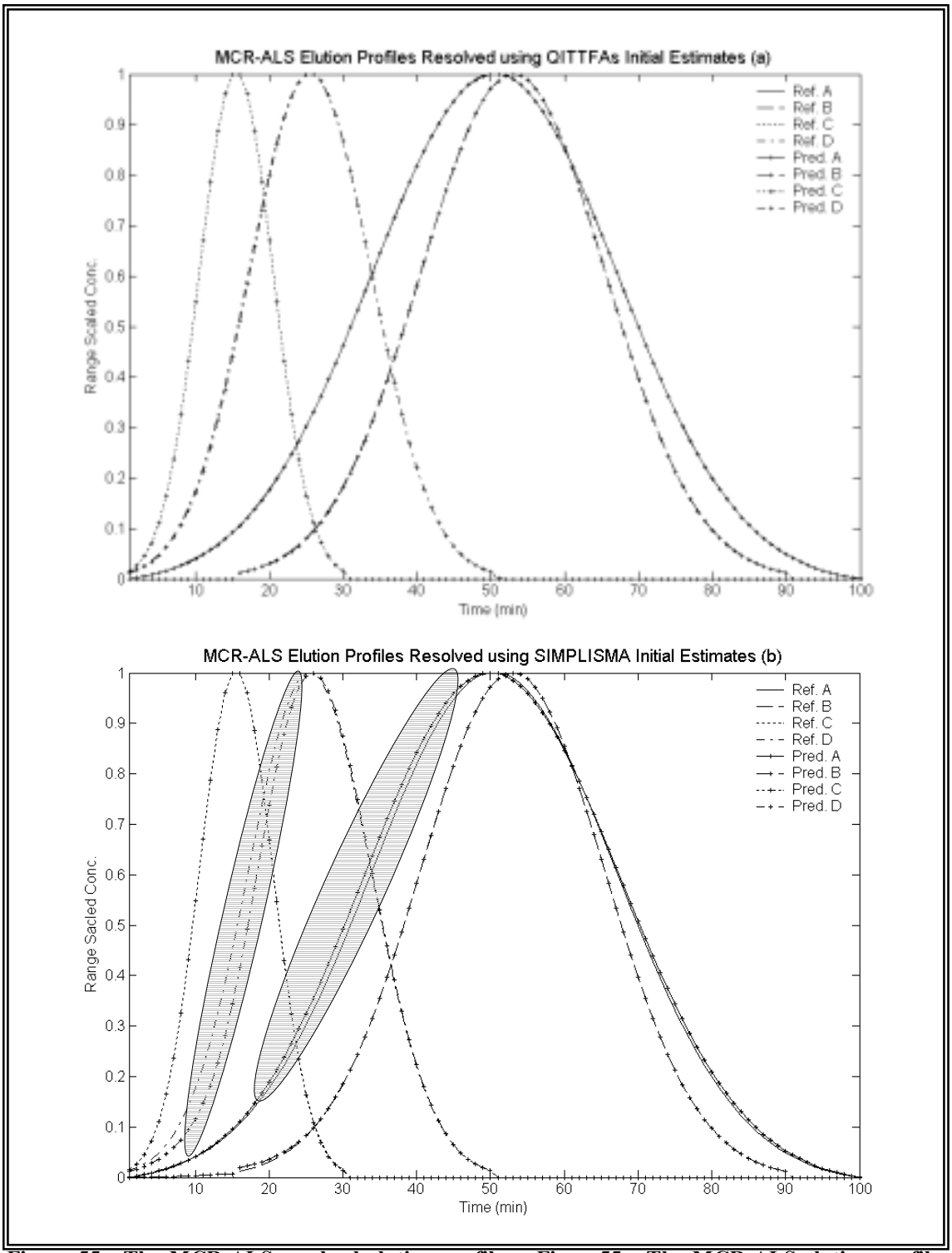


Figure 55. The MCR-ALS resolved elution profiles. Figure55a. The MCR-ALS elution profiles resolved using QITTFA$_s$ initial estimates overlaid with the reference spectral profiles for constituents A-D. Figure 55b. The MCR-ALS elution profiles resolved using SIMPLISMA initial estimates overlaid with the reference spectral profiles for constituents A-D.

The percent RE are given in table 14.

The MCR-ALS spectral and elution profiles resolved using the QITTFA$_s$ initial estimates were comparable to the reference data. Good agreement was observed in the

overlay spectra. The MCR-ALS spectral and elution profiles obtained with the SIMPLISMA initial estimates were not as accurate as the MCR-ALS solutions obtained using the QITTFA$_s$ initial estimates. Slight differences between the resolved spectral profiles and the reference spectral profiles of components **B** and **C** were found see figure 54b (shaded areas). This was expected as the key variables selected for each component had low purity values. The concentration profiles resolved using the SIMPLISMA estimates contained errors that were particularly noticeable in the concentration profiles of constituents **A** and **D**. This was not expected as the key variables selected for these constituents were relatively pure (i.e., the purity values of the variables were high, see table 13).

	Spectral Profiles		***Elution Profiles***	
%RE	**QITTFA$_s$**	**SIMPLISMA**	**QITTFA$_s$**	**SIMPLISMA**
A	0.46	0.10	0.07	2.47
B	0.44	13.65	0.27	0.44
C	1.44	9.19	0.02	0.25
D	1.05	1.46	0.85	8.32

Table 14. The percent RE of the elution profiles and the spectral profiles for each constituent. The solutions were obtained from the constrained ALS procedure using the QITTFA$_s$ and SIMPLISMA starting estimate.

II.5.2.4 Summary

In this example the SIMPLISMA starting estimates did not approximate the actual solution very well. This had an adverse effect on the quality of the final MCR-ALS solution, and as such the MCR-ALS solutions obtained using the SIMPLISMA initial estimates contained greater error. These results have indicated that the QITTFA$_s$ routine is superior to the SIMPLISMA method for the determination of the initial estimates using the simulated HPLC-DAD (II) dataset and this example has highlighted

the importance of using initial estimates, which closely approximate the true solution to gain a successful resolution.

II.5.3 Conclusions

The MCR-ALS solution obtained using the QITTFA$_c$ and QITTFA$_s$ starting estimates in examples II.5.1 and II.5.2 contained less error than its traditional counterpart methods using simulated noise-free datasets. This is because the needle output spectra were constrained using constraints which were characteristic of the data, which meant that the singular vectors were not abstract representations of the solutions. Secondly, the singular vectors were not scaled, so the selection of the initial estimates from the needle output spectral matrix was not based on its variance contribution.

In the following study, a quantitative MCR-ALS approach was developed using accurate initial estimates determined from QITTFA and a correlation constraint (regression constraint) for the calibration of an industrial process.

II.6 Semi-Quantitative Analysis of Calibration samples from the BP Vinyl Acetate Process

II.6.1 Introduction

This study was completed in collaboration with Dr Edo Becker and Dr Alasdair Thomson, Spectroscopic On-Line Analysis, Program Development, BP Chemicals Ltd, Saltend Lane, UK.

Aim

The aim of the study was to determine whether the pure spectrum of VAM in the vapour state could be resolved using CFT from the two-way NIR mixture data collected from the vinyl acetate plant. Previously, the pure spectrum of VAM was not available because neat VAM tended to condense at the specified reaction conditions. The new exploratory tool, QITTFA, (see section II.4) was employed in order to resolve the pure spectrum of VAM. The resolution using the QITTFA starting estimates was compared to the resolution using a traditional exploratory tool, SIMPLISMA, in order to demonstrate the robustness and reliability of the approach.

Introduction

The manufacture of vinyl acetate is an industrially significant process. In 1999 the total world demand for vinyl acetate was approximately 4 million metric tonnes per annum. This number has increased steadily to almost 5 million metric tonnes over the last 4 – 5 years. Forty percent of VAM production is currently controlled by two major manufacturers, BP Chemicals and Celanese, although there are many producers with smaller capacities [158]. The majority of vinyl acetate is used in the manufacture of poly vinyl alcohol for textiles, textile fibres, adhesives, emulsions and photosensitive coatings [159].

In the VAM reaction, ethylene, pure oxygen, and acetic acid are converted into the vinyl acetate monomer product. Water and carbon dioxide are by-products. The reaction takes place as follows, see equations 48-49;

$$C_2H_4 + CH_3COOH + \tfrac{1}{2}\, O_2 \rightarrow CH_2 = CHOCOCH_3 + H_2O \qquad \textbf{Equation 48}$$

$$C_2H_4 + 3O_2 \rightarrow 2CO_2 + 2H_2O \qquad \textbf{Equation 49}$$

The industrial manufacture of VAM was first developed by Wacker via the vapour phase reaction of acetic acid and acetylene during the early 1930s, by the use of the Wacker $PdCl_2$ catalyst with $CuCl_2$ and O_2 [160], see equations 50-51.

$$Pd^0 + 2CuX_2 \rightarrow PdX_2 + 2CuX \qquad \textbf{Equation 50}$$

$$2CuX + 2HX + \tfrac{1}{2}O_2 \rightarrow 2CuX_2 + H_2O \qquad \textbf{Equation 51}$$

Virtually all VAM was produced by this technology until the early 1960s when the advent of selective transition metal oxidation catalysts enabled the replacement of acetylene by ethylene as the feedstock. The VAM production using the ethylene based routes became more popular because of the lower raw material cost, which translated into a lower cost product. In the early 1960s several liquid phase ethylene based production processes were developed and commercialised. Between the late 1960s and the early 1970s all of the liquid phase VAM plants were shut down primarily due to the unexpected corrosion problems that necessitated expensive equipment modifications. The chemistry of vapour phase ethylene acetoxylation to vinyl acetate was discovered around 1960. In less than a decade, fixed bed, vapour phase ethylene acetoxylation VAM manufacturing became the process of choice and further decreased product cost. The fluid-bed process is another method of improving the VAM manufacture and reducing production costs.

BP is the industrial leader in the development of fluid-bed processing for the manufacture of vinyl acetate. The continuous reactor fluid-bed at BP Chemicals produces ~250,000 tonnes of vinyl acetate per annum. LEAP® is the registered trademark referring to the new method of manufacturing vinyl acetate developed by BP. This development was driven by the need to reduce manufacturing cost. The advantages of the fluid-bed reactor over the fixed bed reactor is that it is easier and cheaper to construct. Due to engineering constraints most world-scale fixed bed plants utilise two reactors. For the fluid-bed, only a single bed is required leading to a major cost saving. In addition, the nature of the fluid-bed allows the feeds to be processed in a completely different manner, which eliminates the number of pieces of equipment [161].

II.6.2 Experimental

The calibration data was collected by Edo Becker, Alasdair Thomson, Dave Lightowlers and Ian Taylor-Hayward, BP Chemicals, Hull.

II.6.2.1 Reaction Conditions

Three hundred and thirty nine NIR calibration standards were prepared in the plant as mixtures of five organic components, ethylene (BOC Gases Ltd), carbon dioxide (BOC Gases Ltd), water (demineralised), acetic acid (BP Chemicals final product) and vinyl acetate (BP Chemicals final product). An automated mixing system, "stealth trolley", was designed specifically to make-up the calibration standards in the SPECAC NIR, Typhoon T13 gas cell, using evaporators and mass flow controllers. Each sample took ~20 minutes to prepare and was introduced into the gas cell at 120˚C. The partial pressure of each constituent and the total pressure for each calibration sample differed from sample-to-sample to mimic varying process conditions.

II.6.2.2 Data Acquisition

The spectra were acquired using a BOMEM MB-155 and MB-160 fitted with a TE-cooled InAs detector. Each spectrum was recorded using the average of 32 scans, the spectral region was 9998.2- 4497.7cm^{-1} and the resolution was 7.7cm^{-1}. Reference spectra of the vaporised pure components, ethylene, water, carbon dioxide and acetic acid were obtained prior to the analysis. MATLAB6p5® (The Math Works, Inc) was used to complete all data processing.

II.6.3 Results and Discussion

In the final investigation, the aim was to resolve the pure spectrum of VAM in the vapour state from the two-way calibration mixture data collected from a BP process NIR analyser on the vinyl acetate plant using CFT. The pure spectrum of VAM was not available because neat VAM tended to condense at the specified reaction conditions. In this study QITTFA was applied in order to resolve the pure spectrum of VAM. The MCR-ALS resolution initialised from the QITTFA starting estimates was compared to the MCR-ALS resolution initialised from SIMPLISMA, in order to demonstrate the robustness and reliability of the approach.

II.6.3.1 Reaction Profile

The NIR data consists of five components, ethylene, carbon dioxide, water, acetic acid and vinyl acetate. The first 160 NIR spectroscopic samples acquired are shown in figure 56. The traditional designation of ν for a stretching mode and δ for a bending mode is used in the figure. There were several regions of interest in the NIR spectra, representative of three of the five chemical constituents. The first overtone of the OH group associated with monomer acetic acid is apparent at ~6994cm^{-1}. The first overtone of the asymmetric C-H stretch from ethylene cover 6200-6090cm^{-1} and the first

overtone of the symmetric C-H stretch appear at 5993cm^{-1} together with the combination bands with relatively strong absorbance appearing at 4760cm^{-1}. The combination bands from the asymmetric and symmetric stretching modes of water appear at ~5150 and ~6900cm^{-1}.

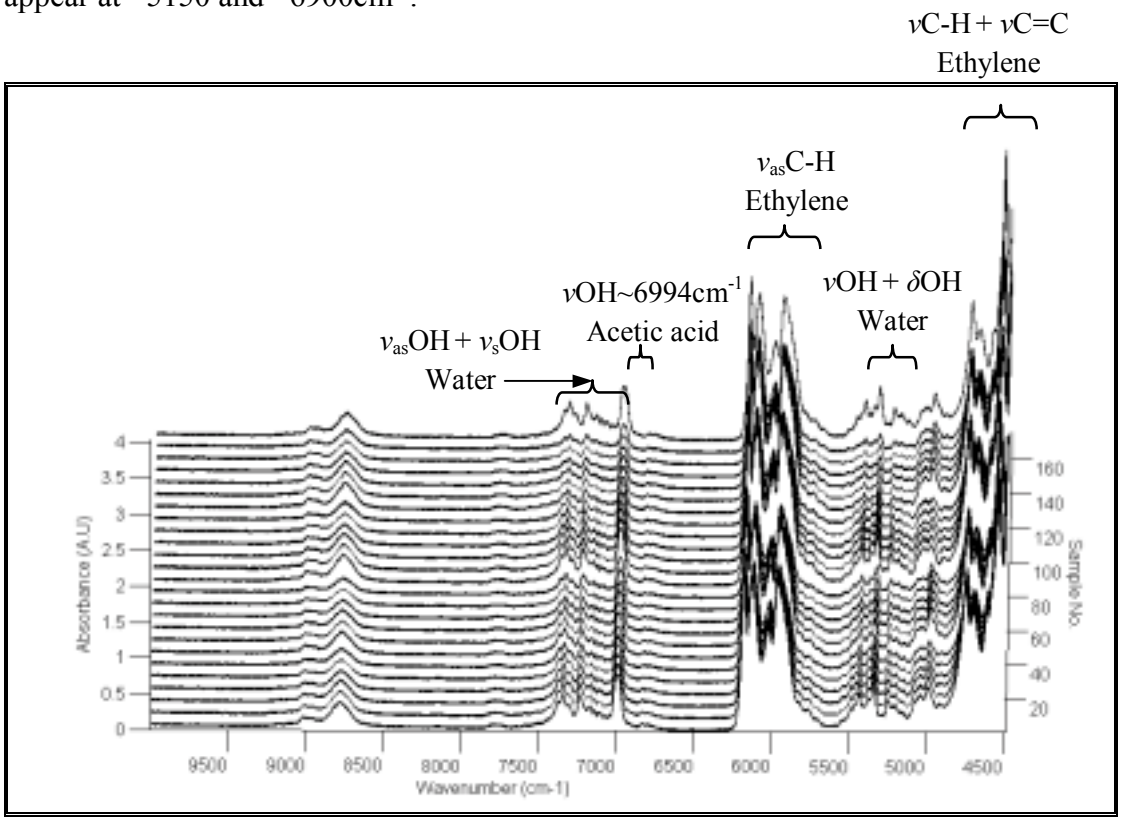


Figure 56. The first 160 calibration samples acquired from the industrial NIR process analyser.

Selective regions for vinyl acetate and carbon dioxide were not identified as they coincided with characteristic functional group frequencies of ethylene, acetic acid and water.

II.6.3.2 Qualitative Data Analysis

For each investigation the mixture's NIR calibration spectra were baseline corrected using the minimum-offset method (removal of negative absorbencies). Sample 168 was removed as an outlier because it had an abnormally high absorption. The two methods employed to determine the starting estimates were QITTFA and SIMPLISMA.

QITTFA$_s$ analysis was repeated using 4, 5 and 6 components to describe the reduced space and capture the structured variation. In each analysis the needle output spectra were constrained with non-negativity constraints. The QITTFA$_s$ initial estimates were used to initialise the MCR-ALS procedure. In a separate analysis, SIMPLISMA was applied to resolve the chemical constituents. The analysis was repeated using 4, 5 and 6 components. The SIMPLISMA initial estimates were used to initialise the MCR-ALS procedure.

In both applications (QITTFA followed by MCR-ALS analysis and SIMPLISMA followed by MCR-ALS analysis) the spectral profiles were normalised to a height of one. Non-negativity constraints were not applied in the spectra or concentration profiles because the predicted vinyl acetate spectrum contained the first overtone of the OH group associated with monomer acetic acid (~6994cm^{-1}). This occurred in both the MCR-ALS analysis initialised using the QITTFA$_s$ starting estimates and the MCR-ALS analysis initialised using the SIMPLISMA starting estimates. The convergence criterion was 7% (chosen to maintain data quality) and the maximum number of iterations was set to 100. The predicted and the reference concentration profiles were scaled between zero and one for visual comparison.

II.6.3.3 Initial Estimates

Initial QITTFA Estimates

The preliminary investigation was completed using four components, which were chosen to describe the reduced space and to capture the structured variation. The needle output spectra were constrained with a non-negativity constraint. Initial spectral estimates obtained from the needle output spectral matrix were ethylene, water, acetic

anhydride and vinyl acetate monomer. Each spectrum contained the absorption bands associated with the characteristic overtones and combination bands of the functional groups present in each molecule. Carbon dioxide could not be resolved by simply increasing the number of components in the needle output spectral matrix by one. A subspace of six components was used to account for additional noise and five components were selected from the needle output spectral matrix. The results of the MCR-ALS analysis using five components was not correct because the spectrum of vinyl acetate was contaminated with the first overtone of the OH group associated with monomer acetic acid. The best solution was obtained when six components were used to describe the reduced loadings space and six components were incorporated into the MCR-ALS resolution. The sixth component is strong baseline component which is slightly contaminated by residual variance from water and acetic acid. The sixth spectrum and the corresponding ALS concentration profile is shown in figure 57.

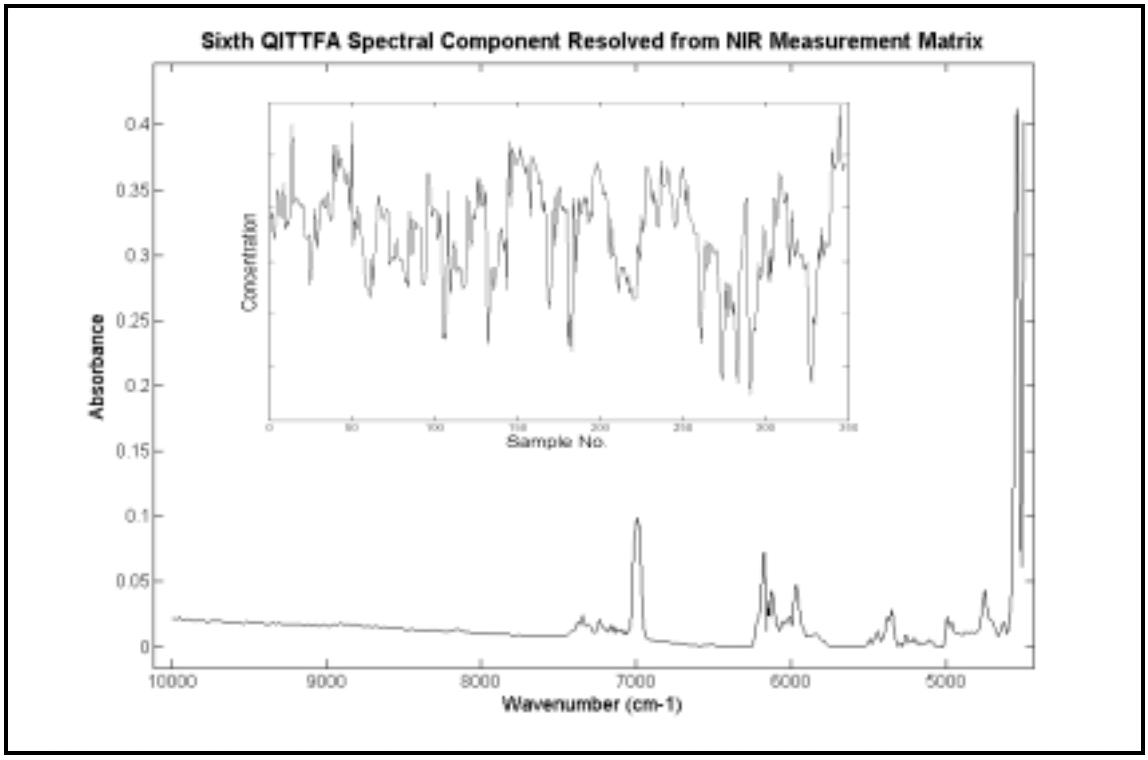


Figure 57. Sixth QITTFA spectral component resolved from the NIR measurement matrix.

The sixth component spectrum contained peaks at the characteristic functional groups of water, acetic acid and ethylene. A rising baseline was also observed between 10 000-7500cm^{-1}. Consequently, the corresponding concentration profile resolved did not vary according to any design parameters. This component, although independent, was putatively attributed to structured variance in the data which was uncorrelated to the concentration profiles of the reaction constituents. Thus, the initial spectral estimates resolved from the needle output spectral matrix were ethylene, water, acetic acid, carbon dioxide, vinyl acetate monomer and a probable structured noise component. Each spectrum contained the correct absorption bands associated with the functional groups present in each molecule, apart from carbon dioxide which contained slight contribution from a methyl group between 6190-6110cm^{-1} in the spectrum.

Initial SIMPLISMA Estimates

Six components were selected from the original NIR measurement matrix using the SIMPLISMA procedure because preliminary MCR-ALS analysis using five components resulted in spectra and concentration profiles which deviated substantially from the true solution. The sixth spectrum and concentration profile resolved using SIMPLISMA followed by MCR-ALS resolution is given in figure 58. The sixth component spectrum contained a strong contribution from ethylene, as well as water and acetic acid. However, the rising baseline observed in the sixth component spectrum using the QITTFA initial estimate was not observed in the respective profile. The corresponding concentration profile contained structured variance which was not too dissimilar from the concentration profile expected for ethylene. The strong similarity between the ethylene concentration profile and the sixth component concentration profile was attributed to the strong contribution of ethylene in the sixth component

spectrum (evident in figure 58). In this application the sixth component resembled a combination spectrum of ethylene, water and acetic acid.

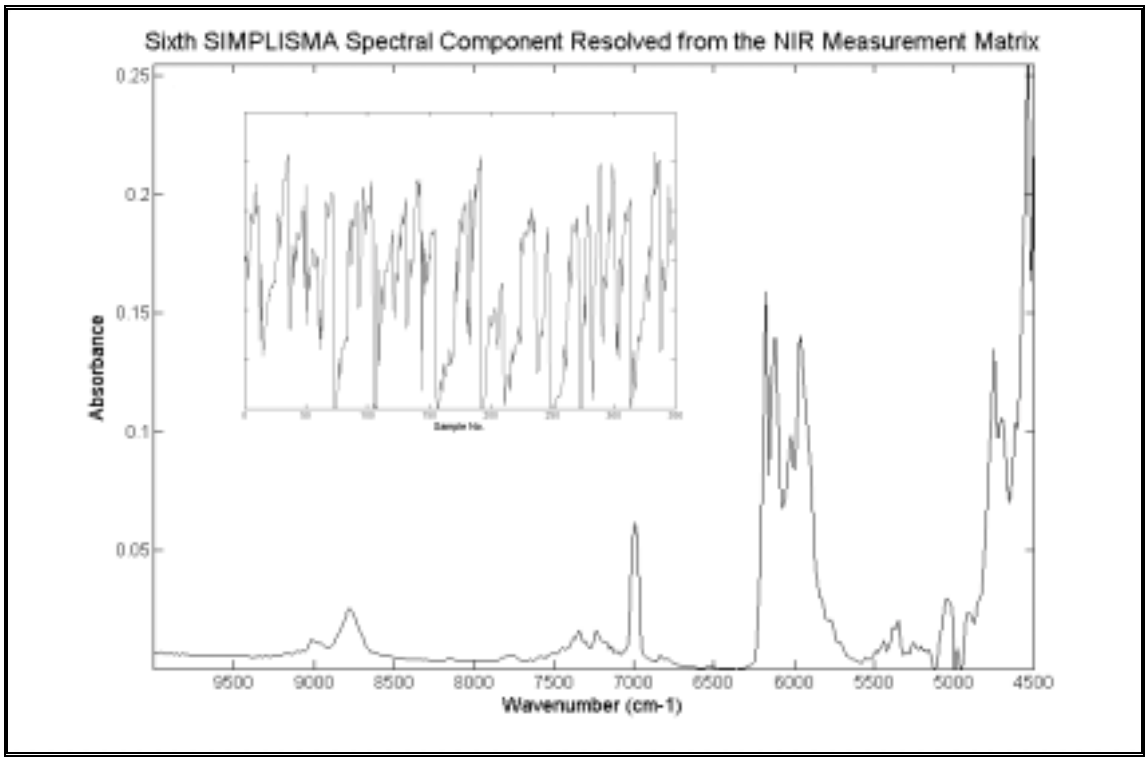


Figure 58. Sixth component resolved from NIR measurement using SIMPLISMA initial estimates to initialise MCR-ALS.

The SIMPLISMA *concentration* estimates were used to initiate the MCR-ALS procedure instead of the SIMPLISMA *spectral* profiles because each of the spectral profiles resolved contained; a) the free stretching first overtone of the OH group associated with acetic acid, b) the first overtone of the asymmetric C-H stretch and the first overtone of the symmetric C-H stretch as well as the combination bands from ethylene, and c) the combination bands of the asymmetric and symmetric modes of water. The lack of selectivity in the sample direction was the probable cause of the linear combination spectra resolved from the measurement matrix.

For both the QITTFA$_s$ and SIMPLISMA analysis, six components were required to initialise the MCR-ALS procedure. The rank of the needle output matrix was six and the rank of the original matrix was equal to the number of sample spectra. The

condition number was calculated based on the significant singular values from the needle spectral matrix. The condition number of the needle output spectral matrix was 3.1 and the original matrix was 67.8, which indicated that the quality of the needle output matrix and hence the final least squares solution had improved over the original data, in terms of sensitivity to error and accuracy of the linear equation solution.

II.6.3.4 MCR-ALS

The MCR-ALS resolved spectral profiles obtained using the QITTFA$_s$ initial estimates are shown in figures 59-63. The resolved MCR-ALS spectrum of ethylene contained the correct absorption bands associated with the functional groups, however, slight intensity differences persisted between 9300-8600cm^{-1}, 6230-5850cm^{-1} and 5100-4550cm^{-1}, highlighted in figure 59. The carbon dioxide spectrum contained the correct absorption bands associated with the functional groups, apart from two incorrect bands 6200-6100cm^{-1} and 4620-4500cm^{-1}, shown in figure 60. The water and acetic acid spectral profiles were predicted accurately, shown in figures 61 and 62, respectively. The vinyl acetate spectrum contained the correct absorption bands, i.e., the first overtone and combination bands from CH groups, and the first and second overtones from an alkene vinyl group are present in the predicted vinyl acetate spectrum, shown in figure 63. Overall the spectral profiles resolved from the NIR data using the QITTFA$_s$ initial estimates closely resembled the actual solution.

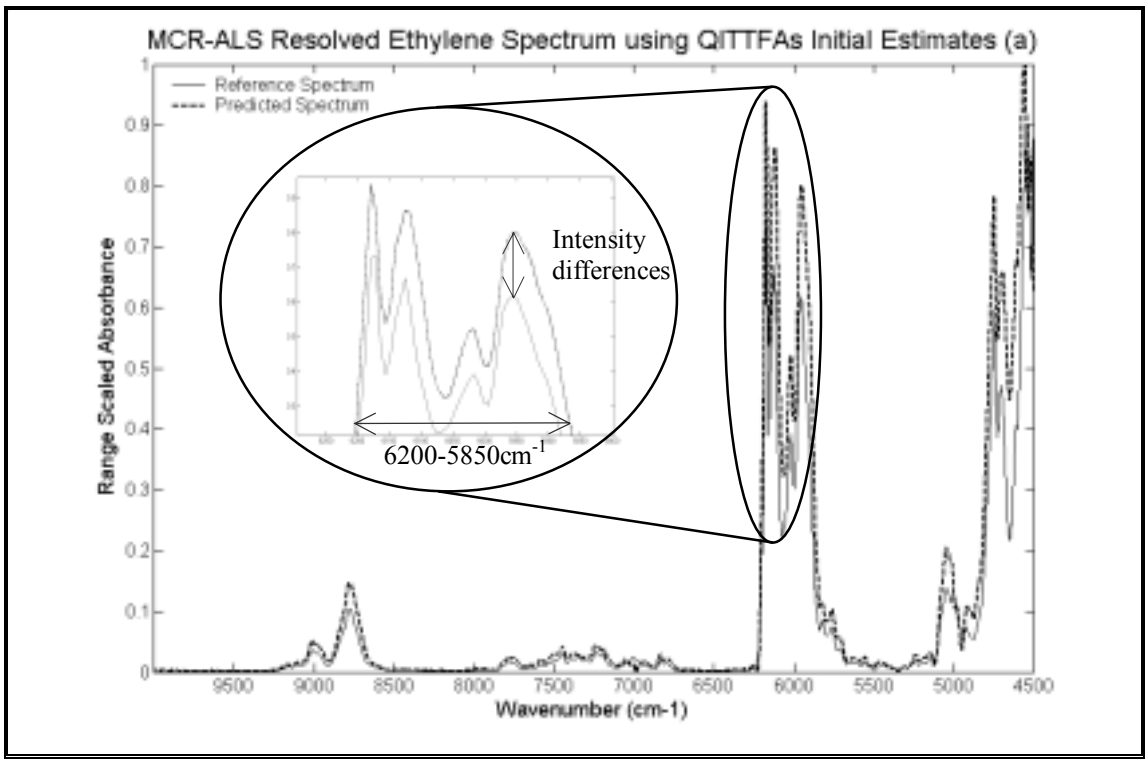


Figure 59. The MCR-ALS ethylene spectrum resolved using QITTFA$_s$ initial estimates overlaid with the reference spectrum of ethylene.

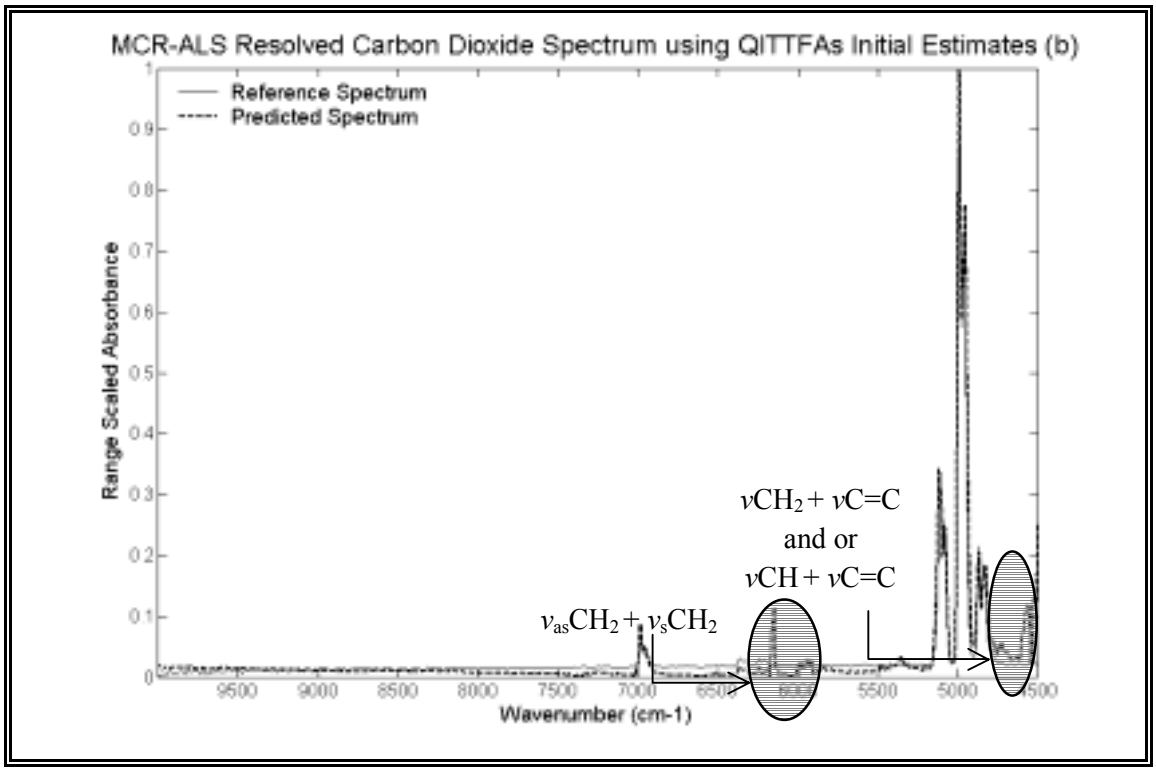


Figure 60. The MCR-ALS carbon dioxide spectrum resolved using QITTFA$_s$ initial estimates overlaid with the reference spectrum of carbon dioxide.

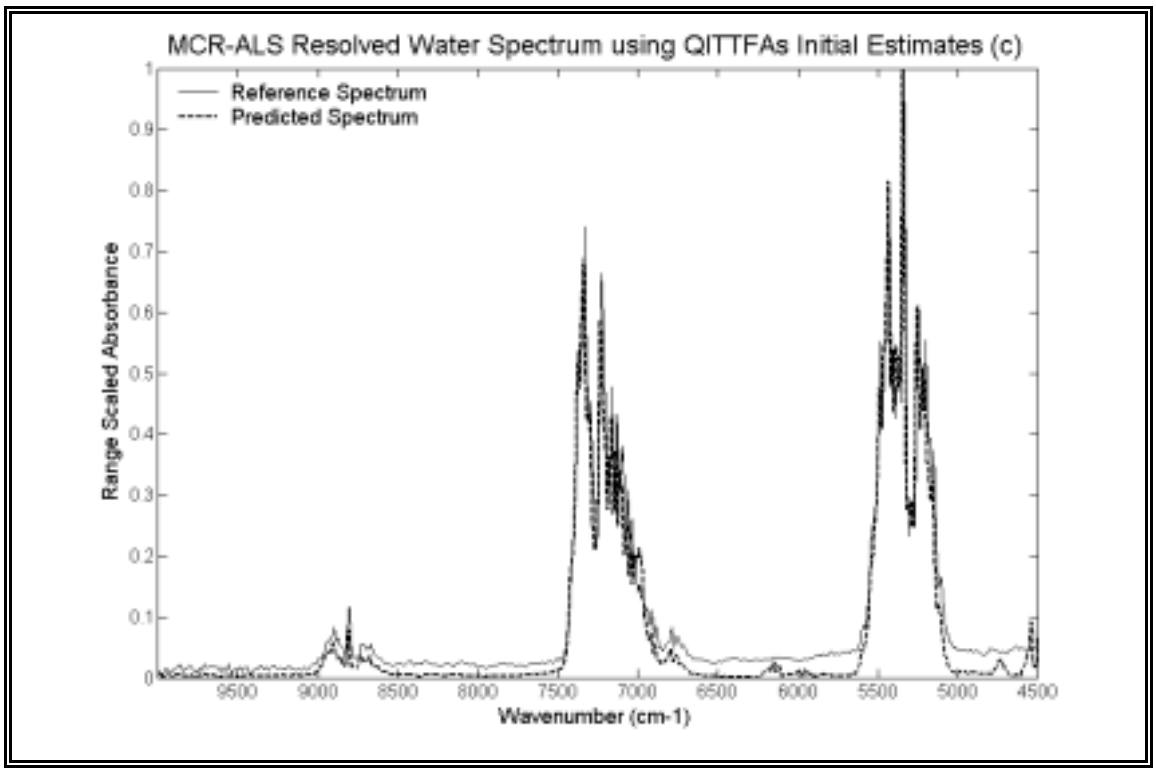


Figure 61. The MCR-ALS water spectrum resolved using QITTFA$_s$ initial estimates overlaid with the reference spectrum of water.

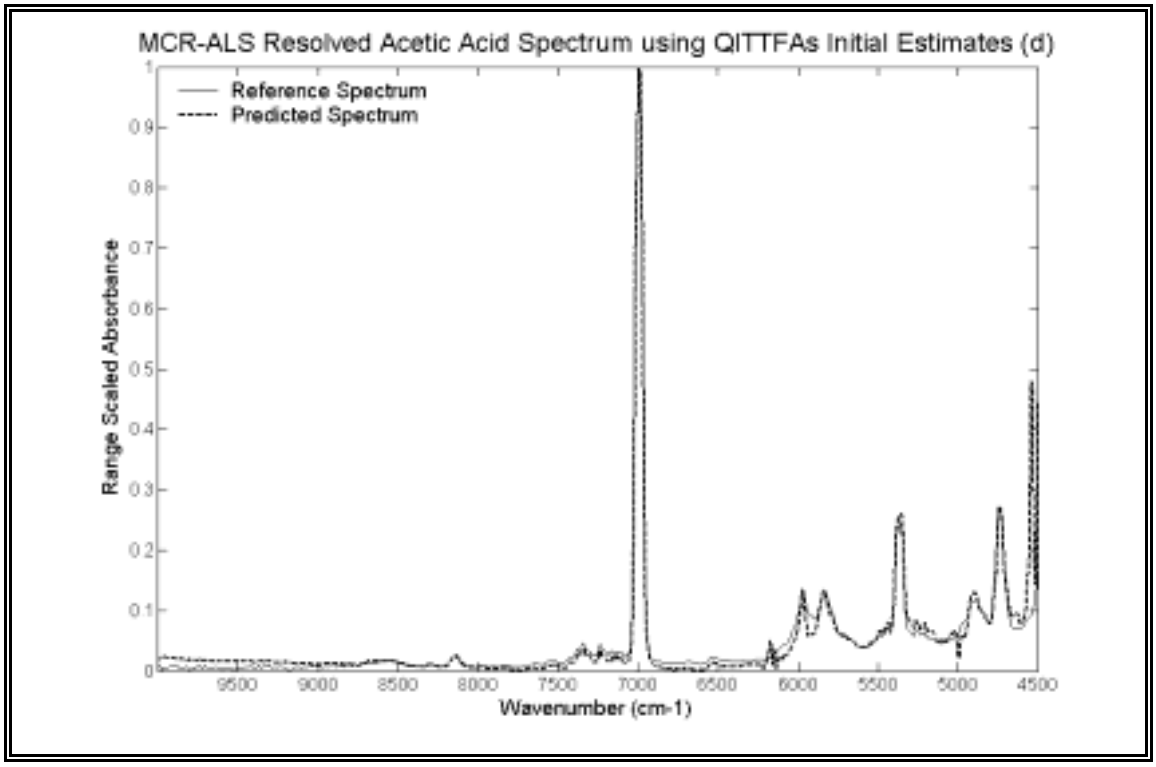


Figure 62. The MCR-ALS acetic acid spectrum resolved using QITTFA$_s$ initial estimates overlaid with the reference spectrum of acetic acid.

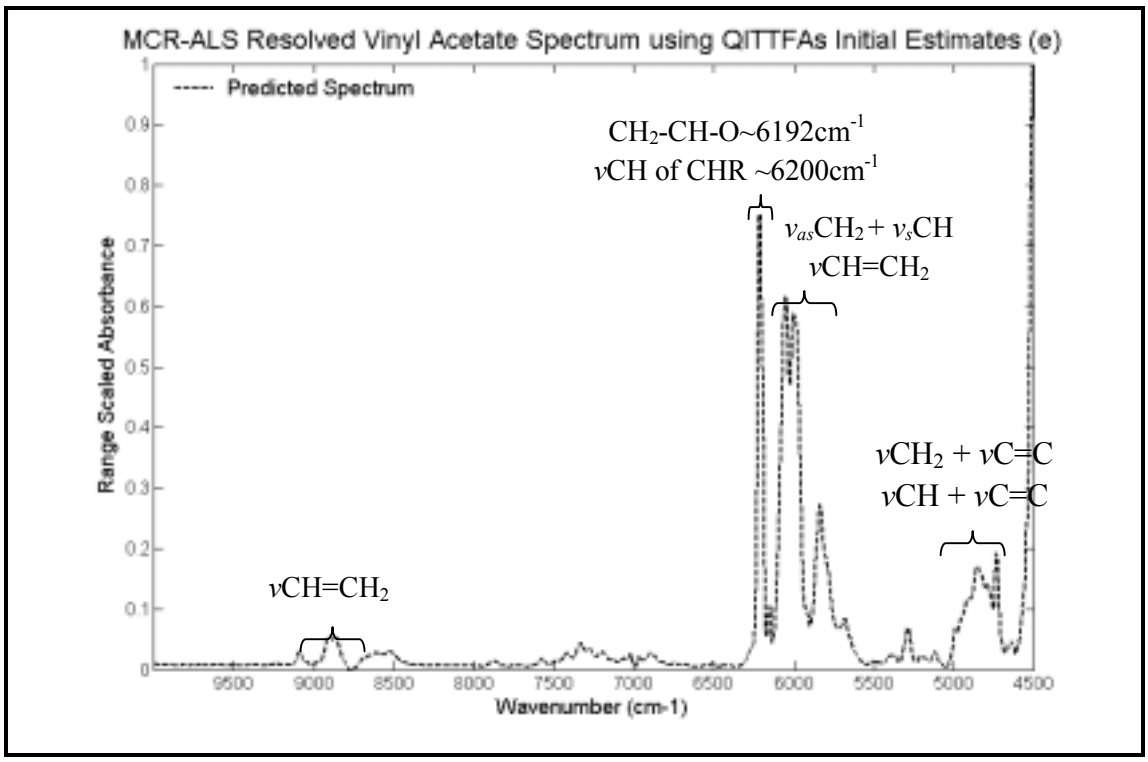


Figure 63. The MCR-ALS vinyl acetate spectrum resolved using QITTFA$_s$ initial estimates overlaid with the reference spectrum of vinyl acetate.

The MCR-ALS spectra obtained using the SIMPLISMA initial estimates did not exhibit good agreement with the expected spectral profiles. The resolved ethylene spectral profile was similar to the spectral profile resolved using QITTFA$_s$ initial estimates, i.e., slight intensity differences persisted in the region of the spectrum associated with first overtones of the asymmetric and symmetric stretch of the C-H groups and combination bands. The resolved spectral profile of carbon dioxide deviated from the expected spectral profile, and this was observed through a rising baseline (10,000-7500cm^{-1}), 2) and the presence of a) the first overtone of the OH stretch and OCO bending from acetic acid or vinyl acetate (7450-7000cm^{-1}), and b) asymmetric stretching of C-H from ethylene or first overtone from -CH_3 in acetic acid (~5990-5930cm^{-1}). The baseline artefact resolved in the QITTFA analysis (component 6) was not separated from the measurement matrix using SIMPLISMA starting estimates. This baseline artefact seems to have contaminated the predicted spectrum of carbon dioxide, which resulted in

the incorrect prediction of this spectrum. The resolved water profile contained differences from the expected spectra profile between 9000-8600cm^{-1} which was possibly due to contamination from ethylene or acetic acid. The acetic acid profile contained the correct absorption bands associated with functional groups between 6000-4000cm^{-1}, and slight deviations from the reference profile were observed between 7530-7040cm^{-1}, 5600-5400cm^{-1} and 5300-5020cm^{-1}. The resolved vinyl acetate spectral profiles contained the correct absorption bands associated with the functional groups.

The MCR-ALS resolved concentration profiles obtained using the QITTFA$_s$ starting estimates are shown in figures 64-68. The resolved MCR-ALS ethylene concentration profile was comparable to the reference data. Samples 167-228 were not synchronised which was attributed to experimental error in sampling the data. The carbon dioxide profile followed the same trend as the reference concentration data, although there were slight intensity differences in parts of the profile. The greatest discrepancy was for samples 278-301, where the actual shape of the concentration profiles differed from the reference data. The water concentration profile contained slight differences in intensities, shown at samples 12-22, 49-56, 105-119 and 228-234. Nevertheless, the correct profile was obtained for this component. The acetic acid profile contained slight shape differences for sample 50-90, although the general shape of the concentration profile was consistent with the reference data. The shape of the vinyl acetate profile was in accordance with the reference data, but slight intensity differences persisted.

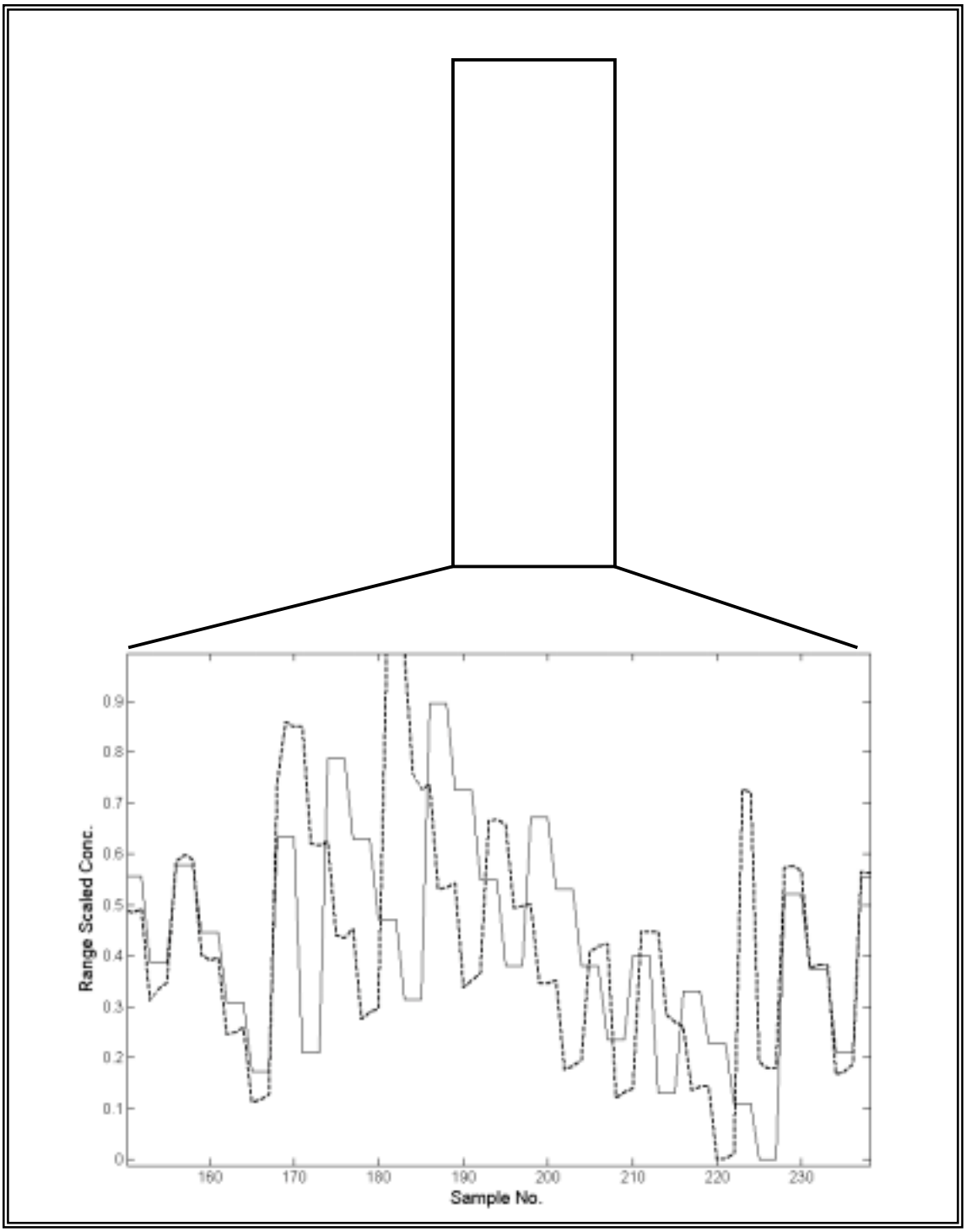
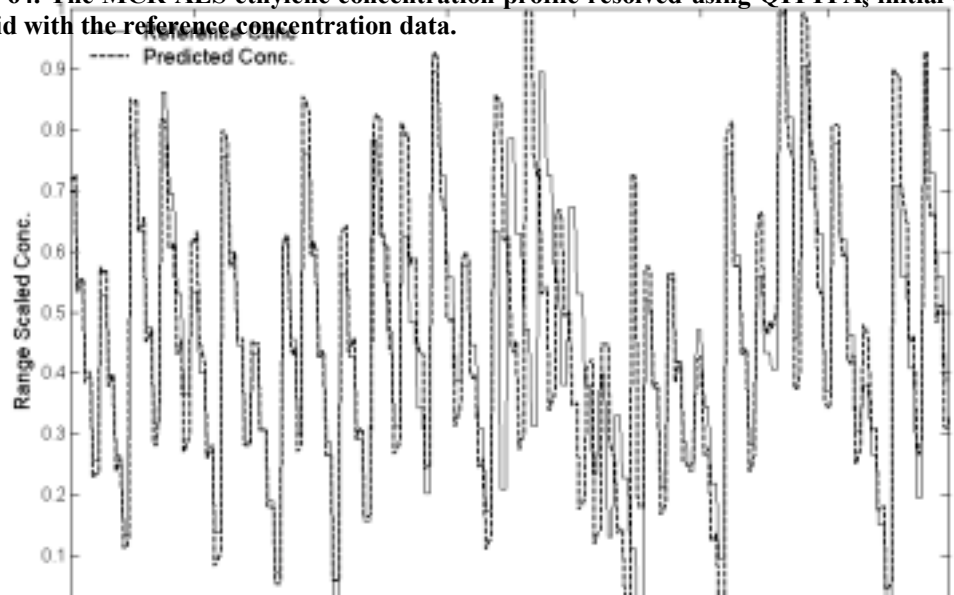


Figure 64. The MCR-ALS ethylene concentration profile resolved using QITTFA$_s$ initial estimates overlaid with the reference concentration data.



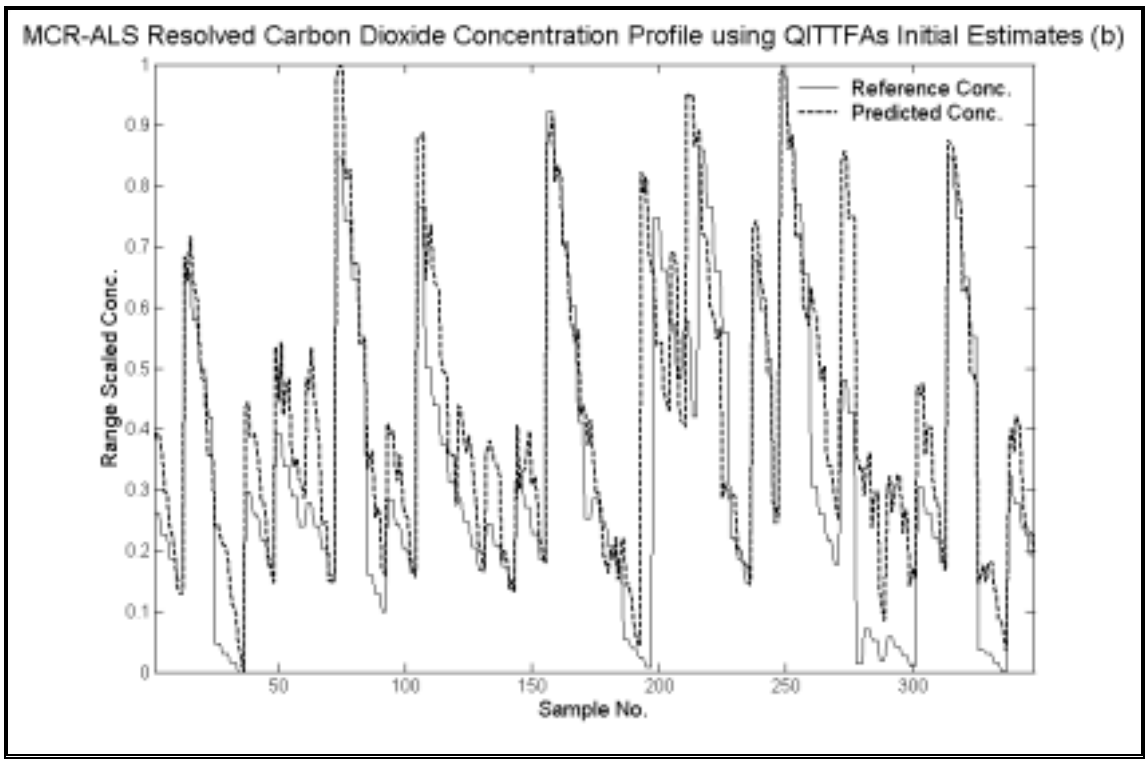


Figure 65. The MCR-ALS carbon dioxide concentration profile resolved using QITTFA$_s$ initial estimates overlaid with the reference concentration data.

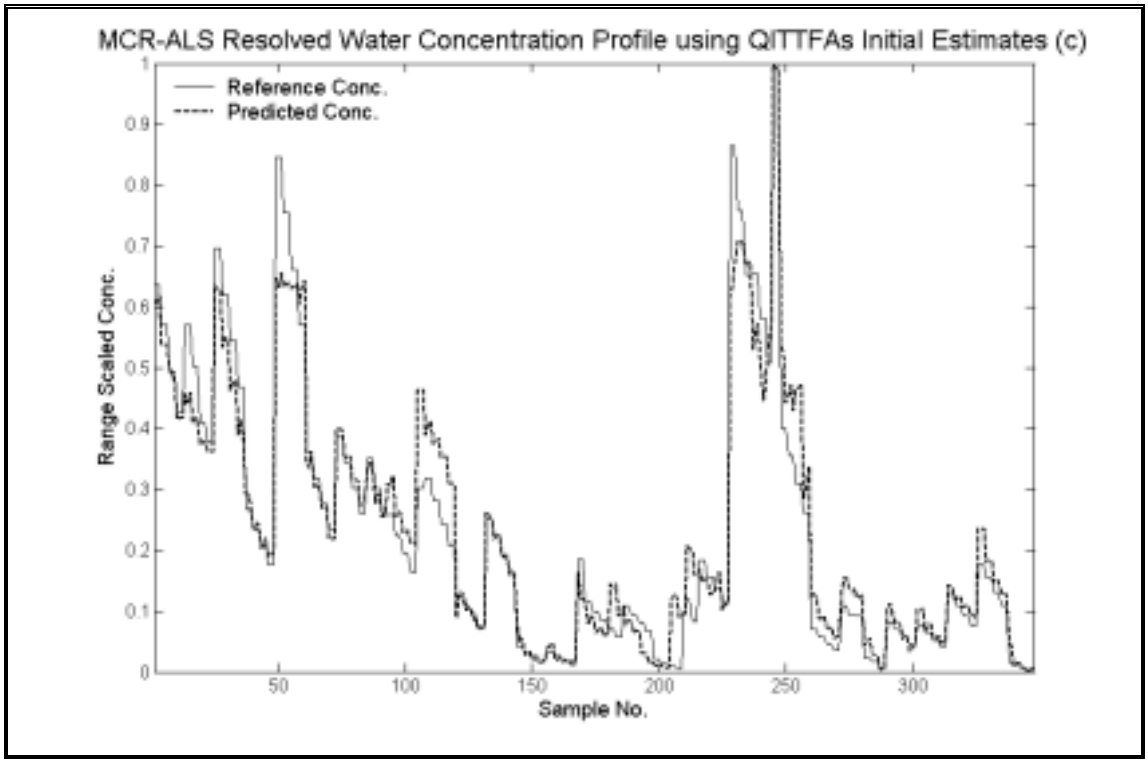


Figure 66. The MCR-ALS water concentration profile resolved using QITTFA$_s$ initial estimates overlaid with the reference concentration data.

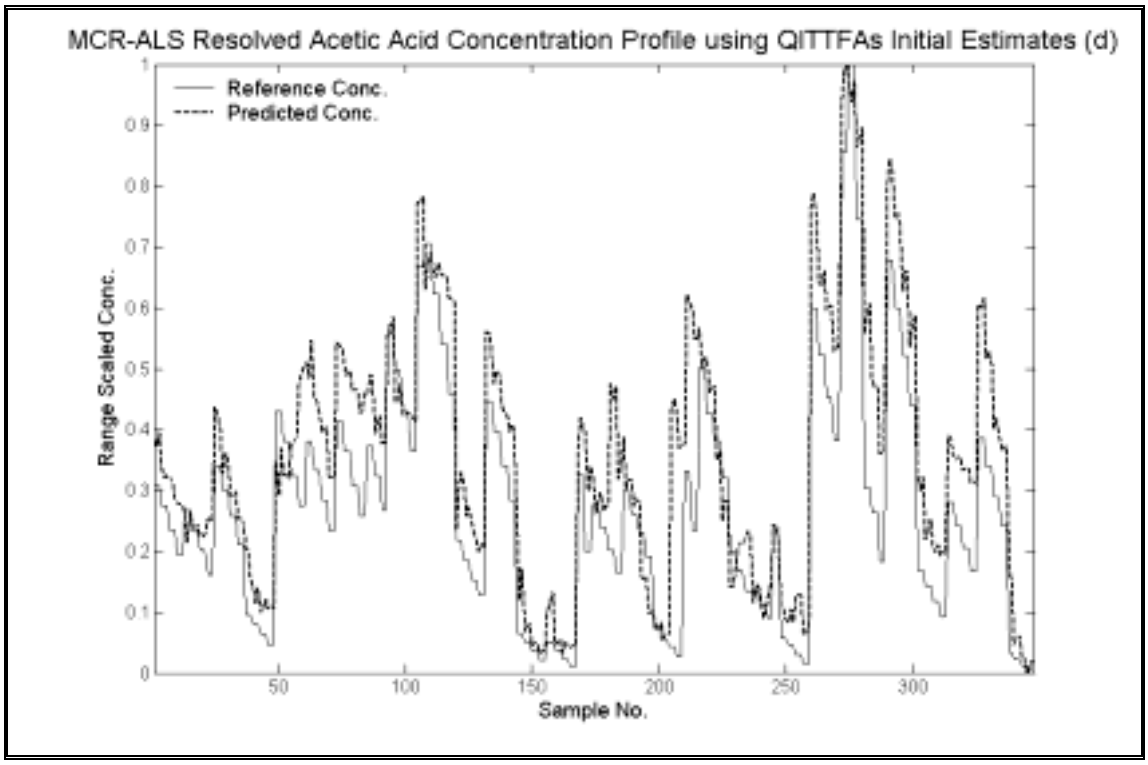


Figure 67. The MCR-ALS acetic acid concentration profile resolved using QITTFA$_s$ initial estimates overlaid with the reference concentration data.

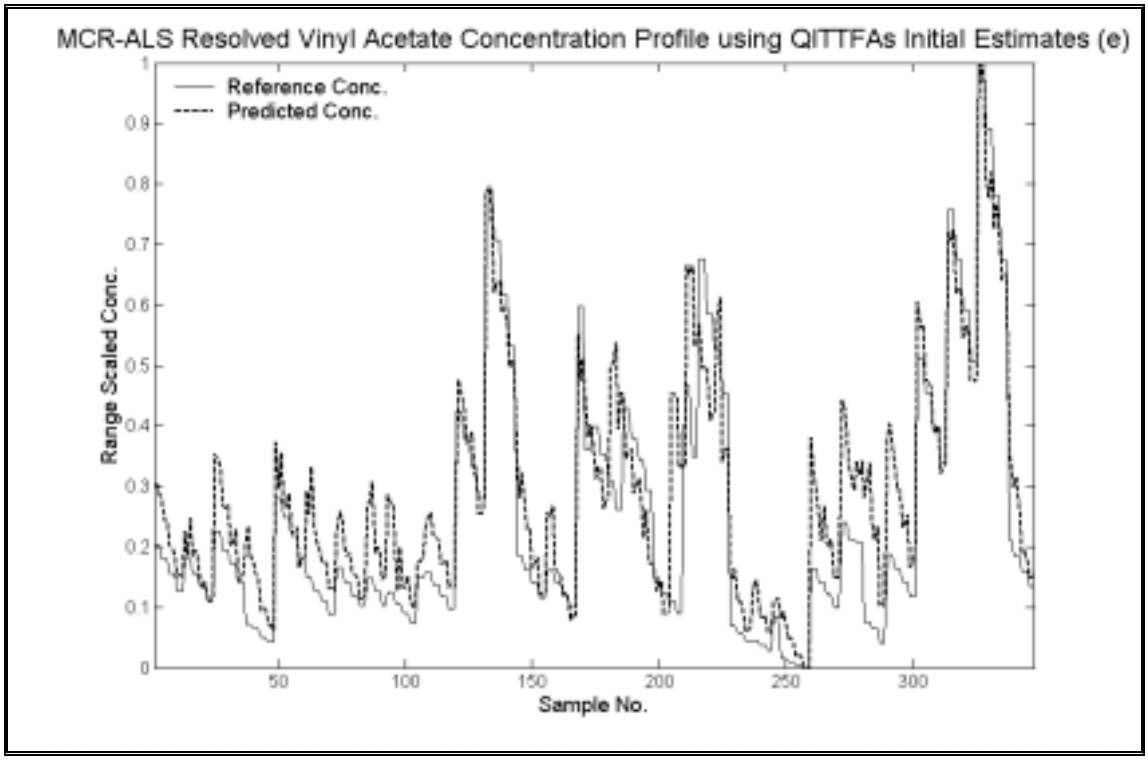


Figure 68. The MCR-ALS vinyl acetate concentration profile resolved using QITTFA$_s$ initial estimates overlaid with the reference concentration data.

The MCR-ALS resolved concentration profiles obtained using the SIMPLISMA initial estimates were worse than those obtained using the $QITTFA_s$ initial estimates. The resolved MCR-ALS ethylene concentration profile contained similar types of errors as the ethylene concentration profiles resolved using the $QITTFA_s$ initial estimates. There were severe errors in the resolved concentration profile of carbon dioxide. High concentrations were predicted well, but low relative concentrations were incorrect. There were shape differences in the carbon dioxide profile, possibly due to the similarity of the concentration profile of this component with acetic acid. The water concentration profile contained more obvious resultant errors as the actual shape and direction of the step changes within the data differed from the reference concentration values. The acetic acid concentration profile was not consistent with the reference data. The vinyl acetate concentration profile contained more deviation from the reference data than the resolved profile obtained using the $QITTFA_s$ initial estimates, which is reflected in the percent RE and RMSPE, see table 15.

	Concentration Profiles			
	$QITTFA_s$		**SIMPLISMA**	
	RMSPE	**%RE**	**RMSPE**	**%RE**
C_2H_4	0.137	27.16	0.137	27.03
CO_2	0.164	39.05	0.293	69.69
H_2O	0.057	17.67	0.124	38.41
CH_3COOH	0.118	35.33	0.174	52.00
VAM	0.094	29.01	0.135	41.55

Table 15. The RMSPE and percent RE of the elution profiles for each constituent. The solutions were attained from the constrained ALS procedure using the $QITTFA_s$ and SIMPLISMA starting estimates.

II.6.4 Conclusion

Here it was shown that when good initial estimates were available it was possible to resolve the spectrum of vinyl acetate in the vapour state from the mixture measurement matrix with no *a priori* information. The four reaction constituents; ethylene, acetic acid, water and carbon dioxide (a by-product) were also resolved simultaneously from the mixture measurement matrix with no *a priori* information. The QITTFA starting estimates approximated the actual solution and the baseline artefact was separated from the mixture. The success of the application was attributed to two novel features of the QITTFA algorithm, which are: 1. Each spectrum (or concentration profile) in the solution space can be repeatedly constrained and projected using generic constraints; the constrained estimates are often closer to actual solution; and 2. *Absence of unstructured variance (noise)* in the solution space, from which the initial estimates are determined. SIMPLISMA, on-the-other hand, did not provide estimates which approximated the actual solution and it was not possible to separate the baseline artefact. Therefore, it has been shown that if SIMPLISMA is used to select the initial estimates from the needle output spectral matrix rather than the original matrix the solutions are improved where there is a lack of selectivity and noise is present in the measurement matrix.

III Conclusions

III.1 General Conclusions

Several multivariate calibration free strategies have been developed and applied to analyse and predict the concentration and spectral profiles of reagents, products and intermediate constituents using either no *a priori* or very little *a priori* knowledge relating to the chemical or physical properties of the system. Significant contributions from this research include;

1. The quantitative determination of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline, 1-methyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline, carbon dioxide and formic acid from a catalysed asymmetric transfer hydrogenation reaction.

2. The development and application of a NWAY P-ALS function to the base catalysed esterification reaction of acetic anhydride. The rank deficiency of the measurement matrix was broken and the pure spectral and concentration profile of acetic anhydride, pseudo 1-butanol, pyridine and a linear combination spectrum of 1-butylacetate and acetic acid was resolved with the implementation of the soft NWAY P-ALS constraints.

3. The development and application of a rational resolution exploratory tool called QITTFA. It was shown that QITTFA out-performed its counterpart tools, SIMPLISMA and EFA. Secondly, QITTFA was also applied to a complex industrial problem to resolve the pure spectrum of VAM in the gaseous state.

The most important conclusion from the work reported here was the use of initial estimates, which approximate the true solution, and appropriate constraints to get good resolution of the reaction constituents. A number of applications of different origin and

nature have shown that calibration free analysis can give good results pertaining to the pure component spectra and concentration profiles of the reaction constituents, from which it is possible to infer the underlying chemical model to enable reaction monitoring, process control, end point determination or purity assessments. Improved alternatives to existing techniques have been developed and in some cases the new techniques have allowed the resolution of reaction constituent concentrations and spectral profiles which could not otherwise have been obtained using other techniques.

The ultimate advantage of calibration free analysis is its ability to identify the reaction constituents, the number of independent constituents and their evolutionary profile without *a priori* information, which may provide huge economic savings and increased insights for reaction monitoring and process control. The weakness of the technique is centred around the non-uniqueness of the solution and this may present a challenge to the widespread adoption of the techniques, particularly in applications where it is not possible to eliminate the rotational and intensity ambiguity in the solution through the addition of *a priori* information in the form of constraints or initial estimates. This is primarily a result of the rotational and intensity ambiguity and without external validation methods or some chemical knowledge, it is not possible to know whether the results make chemical sense or not. Therefore, to increase the confidence in the results and aid interpretation, solutions should be presented alongside the feasible solutions space to alert users of the ambiguity in the solution.

Calibration free analysis can be beneficially utilised within many fields such as analytical chemistry for developing fast and cheap calibration methods for a variety of chemical analytes. Such methods may well reduce the cost and use of additional chemicals. In food analysis, for the determination and quantification of compounds

which may degrade the food quality, this would lead to improved quality of food, and hence greater profitability. In biomedical applications for identification and classification of biomarkers for the early diagnosis of certain conditions, such as cancer, diabetes, obesity etc. This method would save lives and reduce spiralling health costs. Other industries include, bioprocesses, environmental analysis, pharmaceutical etc.

This thesis is a contribution to the maturing field of calibration free analysis; of which the principles were introduced in the 1970s. The multitude of problems that have been shown to be handled efficiently with calibration free analysis holds promise for future work. Getting a grasp of complex situations and data is a limiting factor for any sound problem solution in science and technology. Calibration free analysis may help here.

IV References

1. MASSART, D. L., VANDEGINSTE, B. G. M., BUYDENS, L. M. C., JONG, S. D., LEWI, P. J. & MEYER-VERBEKE, J. S. (1997) *Data Handling in Science and Technology 20A* (Amsterdam, The Netherlands, Elsevier Science B.V).

2. FDA Process Analytical Technology (PAT) Intiative (CDER, www.fda.gov/cder/OPS/PAT.htm).

3. MCLENNAN, F. & KOWALSKI, B. R. (1995) *Process Analytical Chemistry* (London, Blackie Academic & Professional).

4. CALLIS, J. B., ILLMAN, D. L. & KOWALSKI, B. R. (1987) Process Analytical Chemistry, *Analytical Chemistry*, 59, 625A-637A.

5. NEWMAN, A. R. (1990) The Center for Process Analytical Chemistry, *Analytical Chemistry*, 62, 965A-976A.

6. MILLER, C. E. (2005) *Process Analytical Technology* (Oxford, Blackwell Publishing Ltd).

7. KELLNER, R. (1998) *Analytical chemistry: The approved text to the FECS curriculum, analytical chemistry* (Weinheim; Chichester: Wiley-VCH, c1998).

8. CHALMERS, J. M. (2000) *Spectroscopy in Process Analysis* (Sheffield Academic Press).

9. GUNNELL, J. J. & VAN VUUREN, P. (2000) Process Analytical Systems: A vision for the future, *Journal of Process Analytical Chemistry*, 6.

10. WORKMAN, J., CREASY, K. E., DOHERTY, S., BOND, L., KOCH, M., ULLMAN, A. & VELTKAMP, D. J. (2001) Process Analytical Chemistry, *Analytical Chemistry*, 73, 2705-2718.

11. MALINOWSKI, E. R. (1991) *Factor Analysis in Chemistry* (USA, John Wiley & Sons).

12. WORKMAN, J., VELTKAMP, D. J., DOHERTY, S., ANDERSON, B. B., CREASY, K. E., KOCH, M., TATERA, J. F., ROBINSON, A. L., BOND, L., BURGESS, L. W., BOKERMAN, G. N., ULLMAN, A. H., DARSEY, G. P., MOZAYENI, F., BAMBERGER, J. A. & GREENWOOD, M. S. (1999) Process Analytical Chemistry, *Analytical Chemistry*, 71, 121R-180R.

13. WORKMAN, J., KOCH, M. & VELTKAMP, D. (2005) Process analytical chemistry, *Analytical Chemistry*, 77, 3789-3806.

14. BEEBE, K. R., BLASER, W. W., BREDEWEG, R. A., CHAUVEL, J. P., HARNER, R. S., LAPACK, M., LEUGERS, A., MARTIN, D. P., WRIGHT, L. G. & YALVAC, E. D. (1993) Process Analytical-Chemistry, *Analytical Chemistry*, 65, R199-R216.

15. BLASER, W. W., BREDEWEG, R. A., HARNER, R. S., LAPACK, M. A., LEUGERS, A., MARTIN, D. P., PELL, R. J., WORKMAN, J. & WRIGHT, L. G. (1995) Process Analytical-Chemistry, *Analytical Chemistry*, 67, R47-R70.

16. BIJLSMA, S., LOUWERSE, D. J. & SMILDE, A. K. (1998) Rapid estimation of rate constants of batch processes using on-line SW-NIR, *Aiche Journal*, 44, 2713-2723.

17. SCHERZER, T., MULLER, S., MEHNERT, R., VOLLAND, A. & LUCHT, H. (2005) In-line determination of the conversion in acrylate coatings after UV curing using near-infrared reflection spectroscopy, *Nuclear Instruments & Methods in Physics Research Section B-Beam Interactions with Materials and Atoms*, 236, 123-129.

18. DUMITRESCU, O. R., BAKER, D. C., FOSTER, G. M. & EVANS, K. E. (2005) Near infrared spectroscopy for in-line monitoring during injection moulding, *Polymer Testing*, 24, 367-375.

19. MCGILL, C. A., FERGUSON, R. H., DONOGHUE, K., NORDON, A. & LITTLEJOHN, D. (2003) In-line monitoring of esterification using a miniaturised mid-infrared spectrometer, *Analyst*, 128, 1467-1470.

20. AMRHEIN, M., SRINIVASAN, B., BONVIN, D. & SCHUMACHER, M. M. (1996) Inferring concentrations on-line from near-infrared spectra: Nonlinear calibration via mid-infrared measurements, *Computers & Chemical Engineering*, 20, S975-S980.

21. SASIC, S., OZAKI, Y., OLINGA, A. & SIESLER, H. W. (2002) Comparison of Various Chemometric Evaluation Approaches for On-Line FT-NIR Transmission and FT-MIR/ATR Spectroscopic Data of Methyl Methacrylate Solution Polymerisation, *Analytica Chimica Acta*, 452, 265-276.

22. POLLARD, D. J., BUCCINO, R., CONNORS, N. C., KIRSCHNER, T. F., OLEWINSKI, R. C., SAINI, K. & SALMON, P. M. (2001) Real-time analyte monitoring of a fungal fermentation, at pilot scale, using in situ mid-infrared spectroscopy, *Bioprocess and Biosystems Engineering*, 24, 13-24.

23. ROYCHOUDHURY, P., HARVEY, L. M. & MCNEIL, B. (2006) At-line monitoring of ammonium, glucose, methyl oleate and biomass in a complex antibiotic fermentation process using attenuated total reflectance-mid-infrared (ATR-MIR) spectroscopy, *Analytica Chimica Acta*, 561, 218-224.

24. RUCKEBUSCH, C., SOMBRET, B., FROIDEVAUX, R. & HUVENNE, J. P. (2001) On-line mid-infrared spectroscopic data and chemometrics for the monitoring of an enzymatic hydrolysis, *Applied Spectroscopy*, 55, 1610-1617.

25. LIN, Z. H., ZHOU, L. L., MAHAJAN, A., SONG, S., WANG, T., GE, Z. H. & ELLISON, D. (2006) Real-time endpoint monitoring and determination for a pharmaceutical salt formation process with in-line FT-IR spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis*, 41, 99-104.

26. POLLANEN, K., HAKKINEN, A., REINIKAINEN, S. P., RANTANEN, J., KARJALAINEN, M., LOUHI-KULTANEN, M. & NYSTROM, L. (2005) IR spectroscopy together with multivariate data analysis as a process analytical tool for in-line monitoring of crystallization process and solid-state analysis of crystalline product, *Journal of Pharmaceutical and Biomedical Analysis*, 38, 275-284.

27. MAO, Z. X., DEMIRGIAN, J., MATHEW, A. & HYRE, R. (1995) Use of Fourier transform infrared spectrometry as a continuous emission monitor, *Waste Management*, 15, 567-577.

28. HORVATH, A., DE SMET, K., ORMEROD, D., DEPRE, D., PEREZ-BALADO, C., GOVAERTS, T., VAN DEN HEUVEL, D. & SCHORPION, I. (2005) Development of the one-carbon homologation of a 4-methylcoumarin assisted by in-line FTIR, *Organic Process Research & Development*, 9, 356-359.

29. REIS, M. M., ARAUJO, P. H. H., SAYER, C. & GIUDICI, R. (2004) Comparing near infrared and Raman spectroscopy for on-line monitoring of emulsion copolymerization reactions, *Macromolecular Symposia*, 206, 165-178.

30. REIS, M. M., ARAUJO, P. H. H., SAYER, C. & GIUDICI, R. (2004) Development of calibration models for estimation of monomer concentration by Raman spectroscopy during emulsion polymerization: Facing the medium heterogeneity, *Journal of Applied Polymer Science*, 93, 1136-1150.

31. WANG, C., VICKERS, T. J., SCHLENOFF, J. B. & MANN, C. K. (1992) Insitu Monitoring of Emulsion Polymerization Using Fiberoptic Raman-Spectroscopy, *Applied Spectroscopy*, 46, 1729-1731.

32. SCHUSTER, K. C., EHMOSER, H., GAPES, J. R. & LENDL, B. (2000) On-line FT-Raman spectroscopic monitoring of starch gelatinisation and enzyme catalysed starch hydrolysis, *Vibrational Spectroscopy*, 22, 181-190.

33. SETAREHDAN, S. K. (2004) Modified evolving window factor analysis for process monitoring, *Journal of Chemometrics*, 18, 414-421.

34. GURDEN, S. P., WESTERHUIS, J. A. & SMILDE, A. K. (2002) Monitoring of batch processes using spectroscopy, *Aiche Journal*, 48, 2283-2297.

35. WESTERHUIS, J. A., GURDEN, S. P. & SMILDE, A. K. (2000) Spectroscopic monitoring of batch reactions for on-line fault detection and diagnosis, *Analytical Chemistry*, 72, 5322-5330.

36. LANGERGRABER, G., FLEISCHMANN, N., HOFSTAEDTER, F. & WEINGARTNER, A. (2004) Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy, *Water Science and Technology*, 49, 9-14.

37. THURSTON, T. J., BRERETON, R. G., FOORD, D. J. & ESCOTT, R. E. A. (2004) Principal components plots for exploratory investigation of reactions using ultraviolet-visible spectroscopy: application to the formation of benzophenone phenylhydrazone, *Talanta*, 63, 757-769.

38. GEMPERLINE, P. J., YANG, Y. & BIAN, Z. H. (2003) Characterization of subcritical water oxidation with in situ monitoring and self-modeling curve resolution, *Analytica Chimica Acta*, 485, 73-87.

39. GUILLEMIN, C. L. (1994) The Deferred Standard, *Process Control and Quality*, 6, 9-25.

40. ENGELL, S. & TOUMI, A. (2005) Optimisation and control of chromatography, *Computers & Chemical Engineering*, 29, 1243-1252.

41. MARK, H. & WORKMAN, J. (1998) Linearity in calibration - Act II, scene I, *Spectroscopy*, 13, 18-21.

42. NAES, T., ISAKSSON, T., FEARN, T. & DAVIES, T. (2002) *A User-Friendly Guide to Multivaraite Calibration and Classification* (NIR Publications).

43. BEEBE, K. R. & KOWALSKI, B. R. (1987) An Introduction to Multivariate Calibration and Analysis, *Analytical Chemistry*, 59, 1007A-1017A.

44. THOMAS, E. V. (1994) A Primer in Multivariate Calibration, *Analytical Chemistry*, 66, 795A-804A.

45. LORBER, A., FABER, N. M. & KOWALSKI, B. R. (1997) Net analyte signal calculation in multivariate calibration, *Analytical Chemistry*, 69, 1620-1626.

46. GELADI, P. & KOWALSKI, B. R. (1986) Partial Least-Squares Regression - A Tutorial, *Analytica Chimica Acta*, 185, 1-17.

47. MARTENS, H. & NAES, T. (1989) *Multivariate Calibration* (Guildford, Biddles Ltd).

48. VANDEGINSTE, B. G. M., MASSART, D. L., BUYDENS, L. M. C., JONG, S. D., LEWI, P. J. & SMEYERS-VERBEKE, J. (1998) *Data Handling in Science and Technology - volume 20B* (Netherlands, Elsevier).

49. GEMPERLINE, P. J. (1984) A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis, *Journal of Informatics and Computer Science*, 24, 206-212.

50. LAWTON, W. H. & SYLVESTRE, E. A. (1971) Self Modeling Curve Resolution, *Technometrics*, 13, 617-633.

51. BORGEN, O. S. & KOWALSKI, B. R. (1985) An extension of the multivariate component-resolution method to three components, *Analytica Chimica Acta*, 174, 1-26.

52. BORGEN, O. S., DAVIDSEN, N., MYNGYANG, Z. & OYEN, O. (1986) The Multivariate N-component resolution problem with minimum assumptions, *Mikrochimica Acta*, 2, 1-6.

53. KIM, B. M. & HENRY, R. C. (1999) Extension of Self Modeling Curve Resolution to Mixtures of More Than Three Components Part 2: Finding The

Complete Solution, *Chemometrics and Intelligent Laboratory Systems*, 49, 67-77.

54. KIM, B. M. & HENRY, R. C. (2000) Extension of Self Modeling Curve Resolution to Mixtures of More Than Three Components Part 3: Atmospheric Aerosol Data Simulation Studies, *Chemometrics and Intelligent Laboratory Systems*, 52, 145-154.

55. SYLVESTRE, E. A., LAWTON, W. H. & MAGGIO, M. S. (1974) Curve Resolution using a postulated reaction, *Technometrics*, 16, 353-368.

56. RAJKO, R. & ISTVÁN, K. (2005) Analytical solution for determining feasible regions of self modeling curve resolution (SMCR) method based on computer geometry, *Journal of Chemometrics*, 19, 448-463.

57. GEMPERLINE, P. (1999) Computing the range of feasible solutions in self-modeling curve resoltion algorithms, *Analytical Chemistry*, 71, 5398-5404.

58. TAULER, R. (2001) Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, *Journal of Chemometrics*, 15, 627-646.

59. JIANG, J.-H., LIANG, Y. & OZAKI, Y. (2004) Principles and Methodologies in Self Modeling Curve Resolution, *Chemometrics and Intelligent Laboratory Systems*, 71, 1-12.

60. LEGER, M. N. & WENTZELL, P., D (2002) Dynamic Monte Carlo self-modeling curve resolution method for multicomponent mixtures, *Chemometrics and Intelligent Laboratory Systems*, 62, 171-188.

61. GEMPERLINE, P. J. (1986) Target Transformation Factor Analysis with Linear Inequality Constraints Applied to Spectroscopic-Chromatographic Data, *Analytical Chemistry*, 58, 2656-2663.

62. MAEDER, M. & ZUBERBUHLER, A. D. (1986) The Resolution of Overlapping Chromatographic Peaks by Evolving Factor Analysis, *Analytica Chimica Acta*, 181, 287-291.

63. HOPKE, P. K. (1989) Target Transformation Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, 6, 7-19.

64. WINDIG, W. & GUILMENT, J. (1991) Interactive Self Modelling Mixture Analysis, *Analytical Chemistry*, 63, 1425-1432.

65. KVALHEIM, O. M. & LIANG, Y. Z. (1992) Heuristic Evolving Latent Projections - Resolving 2-way multicomponent data.1.Selectivity,Latent-projective graph, Datascope,Local rank, and Unique resolution, *Analytical Chemistry*, 64, 936-946.

66. MANNE, R., SHEN, H. & LIANG, Y. (1999) Subwindow Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, 45, 171-176.

67. KARJLAINEN, E. J. (1989) The spectral reconstruction problem use of alternating least sqaures regression for unexpected spectral components in two dimensional spectroscopies, *Chemometrics and Intelligent Laboratory Systems*, 7, 31-38.

68. JUAN, A. D. & TAULER, R. (2003) Chemometrics Applied to Unravel Multicomponent Processes and Mixtures, *Analytica Chimica Acta*, 500, 195-210.

69. JUAN, A. D., BOGAERT, V. D., SANCHEZ, F. C. & MASSART, D. L. (1996) Application of the Needle Algorithm for Exploratory Analysis and Resolution of HPLC-DAD data, *Chemometrics and Intelligent Laboratory Systems*, 33, 133-145.

70. TAULER, R. & CASASSAS, E. (1992) Spectroscopic Resolution of Macromolecular COmplexes using Factor-Analysis - Cu(II)-Polyethyleneimine System, *Chemometrics and Intelligent Laboratory Systems*, 14, 305-317.

71. JUAN, A. D., MAEDER, M., MARTINEZ, M. & TAULER, R. (2000) Combining Hard- and Soft-Modelling to Solve Kinetic Problems, *Chemometrics and Intelligent Laboratory Systems*, 54, 123-141.

72. GEMPERLINE, P. J. & CASH, E. (2003) Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions, *Analytical Chemistry*, 75, 4236-4243.

73. ANTUNES, M. C., SIMÃO, J. E. J., DUARTE, A. C. & TAULER, R. (2002) Multivariate Curve Resolution of Overlapping Voltammetric Peaks: Quantitative Analysis of Binary and Quaternary Metal Mixtures, *Analyst*, 127, 809-817.

74. CHEW, W., WIDJAJA, E. & GARLAND, M. (2002) Band-Target Entropy Minimization (BTEM): An advanced Method for Recovering Unknown Pure Component Spectra. Application to the FTIR Spectra of Unstable Organometallic Mixtures, *Organometallics*, 21, 1982-1990.

75. LACORTE, S., BARCELO, D. & TAULER, R. (1995) Determination of Traces of Herbicide Mixtures in Water by Online Solid-Phase Extraction Followed by Liquid-Chromatography with Diode-Array Detection and Multivariate Self-Modeling Curve Resolution, *Journal of Chromatography A*, 697, 345-355.

76. CRUZ, B. H., DÍAZ- CRUZ , J. M., ARIÑO, C., TAULER, R. & ESTEBAN, M. (2000) Multivariate Curve resolution of Polarographic Data Applied to the Study of the Copper-Binding Ability of Tannic Acid, *Analytica Chimica Acta*, 424, 203-209.

77. GOSSART, P., SEMMOUD, P., RUCKEBUSCH, C. & HUVENNE, J.-P. (2003) Multivariate Curve Resolution applied to Fourier Transform Infrared Spectra of Macromolecules: structural Characterisation of the acid form and the salt form of humic acids in interaction with lead, *Analytica Chimica Acta*, 477, 201-209.

78. SÁNCHEZ, F. C., RUTAN, S. C., GILL GARCÍA, M. D. & MASSART, D. L. (1997) Resolution of Multicomponent Overlapped Peaks by the Orthogonal Projection

Approach, Evolving Factor Analysis and Window Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, 36, 153-164.

79. TAULER, R., LACORTE, E. & BARCELÓ, D. (1996) Application of Multivariate Self-Modelling Curve Resolution to the Quantitation of Trace Levels of Organophosphorous Pesticides in Natural Water from Interlaboratory Studies, *Journal of Chromatography A*, 730, 177-183.

80. BATONNEAU, Y., LAUREYNS, J., MERLIN, J. C. & BREMARD, C. (2001) Self-modeling mixture analysis of Raman microspectrometric investigations of dust emitted by lead and zinc smelters, *Analytica Chimica Acta*, 446, 23-37.

81. ZHU, Z.-L., CHENG, W.-Z. & ZHAO, Y. (2002) Iterative Target Transformation Factor Analysis for the Resolution of Kinetic-Spectral Data with an Unknown Kinetic Model, *Chemometrics and Intelligent Laboratory Systems*, 64, 157-167.

82. GEMPERLINE, P. J., PUXTY, G., MAEDER, M., WALKER, D. S., TARCZYNSKI, F. & BOSSERMAN, M. (2004) Calibration-Free Estimates of Batch Process Yields and Detection of Process Upsets Using in Situ Spectroscopic Measurements and Nonisothermal Kinetic Models, *Analytical Chemistry*, 76, 2575-2582.

83. BEZEMER, E. & RUTAN, S. (2002) Resolution of overlapped NMR spectra by two-way multivariate curve resolution alternating least squares with imbedded kinetic fitting, *Analytica Chimica Acta*, 459, 277-289.

84. BEZEMER, E. & RUTAN, S. C. (2001) Multivariate Curve Resolution with Non-Linear Fitting of Kinetic Profiles, *Chemometrics and Intelligent Laboratory Systems*, 59, 19-31.

85. SHAMSIPUR, M., HEMMATEENEJAD, B., AKHOND, M., JAVIDNIA, K. & MIRI, R. (2003) A study of the photo-degradation kinetics of nifedipine by multivariate curve resolution analysis, *Journal of Pharmaceutical and Biomedical Analysis*, 31, 1013-1019.

86. BIJLSMA, S., BOELENS, H. F. M., HOEFSLOOT, H. C. J. & SMILDE, A. K. (2000) Estimating reaction rate constants: comparison between traditional curve fitting and curve resolution, *Analytica Chimica Acta*, 419, 197-207.

87. TAULER, R., SANCHEZ, F. C. & MASSART, D. L. (1996) Validation of Alternating Least Squares Multivariate Curve Resolution for Chromatographic Resolution and Quantitation, *Trends in Analytical Chemistry*, 15, 279-286.

88. IZQUIERDO-RIDORSA, A., SAURINA, J., HERNANDEZ-CASSOU & TAULER, R. (1997) Second-order multivariate curve resolution applied to rank-deficient data obtained from acid-base spectrophotometric titrations of mixtures of nucleic bases, *Chemometrics and Intelligent Laboratory Systems*, 38, 183-196.

89. SASIC, S. (1998) Quantitative Analysis of Overlapped Raman Spectra by Target Factor Analysis and Evolving Factor Analysis, *Analyst*, 123, 1193-1197.

90. SASIC, S., ANTIC-JOVANOVIC, A., KUZMANOVIC, M. & JEREMIC, M. (1999) Quantitative Analysis of the Raman Spectra of Mixtures of Weakly Interacting Components by Factor Analysis Methods, *The Analyst*, 124, 1481-1487.

91. SAURINA, J. & TAULER, R. (2000) Strategies for solving matrix effects in the analysis of triphenyltin in sea-water samples by three-way multivariate curve resolution, *Analyst*, 125, 2038-2043.

92. BYLUND, D., DANIELSSON, R. & MARKIDES, K. E. (2001) Peak Purity Assessment in Liquid Chromatography-Mass Spectrometry, *Journal of Chromatography A*, 915, 43-52.

93. LINCOLN, D., FELL, A. F., ANDERSON, N. H. & ENGLAND, D. (1992) Assessment of Chromatographic Peak Purity of Drugs by Multivariate Analysis of Diode-Array and Mass Spectrometric Data, *Journal of Pharmaceutical and Biomedical Analysis*, 10, 837-844.

94. BRAEKELEER, K. D., JUAN, A. D. & MASSART, D. L. (1999) Purity Assessment and Resolution of Tetracycline Hydrchloride Samples Analysed using High-Performance Liquid Chromatography with Diode Array Detection, *Journal of Chromatography A*, 832, 67-86.

95. SANCHEZ, F. C., TOFT, J., BOGAERT, V. D. & MASSART, D. L. (1996) Orthogonal Projection Approach applied to Peak Purity Assessment, *Analytical Chemistry*, 68, 79-85.

96. GEMPERLINE, P. J., ZHU, M., CASH, E. & WALKER, D. S. (1999) Chemometric Characterisation of Batch Reactions, *ISA Transactions*, 38, 211-216.

97. BRAEKELEER, K. D., MAESSCHALCK, R. D., HAILEY, P. A., SHARP, D. C. A. & MASSART, D. L. (1999) On-Line Application of the Orthogonal Projection Approach (OPA) and the Soft Independent Modelling of Class Analogy Approach (SIMCA) for the Detection of the End Point of Polymorph Conversion Reaction by Near Infrared Spectroscopy (NIR), *Chemometrics and Intelligent Laboratory Systems*, 46, 103-116.

98. QUINN, A. C., GEMPERLINE, P. J., BAKER, B., ZHU, M. & WALKER, D. S. (1999) Fiber-optic UV/visible composition monitoring for process control of batch reactions, *Chemometrics and Intelligent Laboratory Systems*, 45, 199-214.

99. TAULER, R. & KOWALSKI, B. R. (1993) Multivariate Curve Resolution Applied to Spectral Data from Multiple Runs of an Industrial Process, *Analytical Chemistry*, 65, 2040-2047.

100. RICHARDS, S., ROPIC, M., BLACKMOND, D. & WALMSLEY, A. (2004) Quantitative Determination of the Catalysed Asymmetric Transfer Hydrogenation of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline using In-situ FTIR and Multivariate Curve Resolution., *Analytica Chimica Acta*, 519, 1-9.

101. MA, B., GEMPERLINE, P. J., CASH, E., BOSSERMAN, M. & COMAS, E. (2003) Characterizing batch reactions with in situ spectroscopic measurements, calorimetry and dynamic modeling, *Journal of Chemometrics*, 17, 470-479.

102. GOURVENEC, S. & MASSART, D. L. (2004) Orthogonal projection approach (OPA) and related methods in process monitoring, *Analytical and Bioanalytical Chemistry*, 380, 373-375.

103. WALKER, D., PURDY, K. & TARCZYNSKI, F. (1999) UV/vis spectroscopic reaction optimiszation requiring no a-prior knowledge or calibration to determine reaction rates, *Electro Optic Integrated Optic and Electronic Technologies for Online Chemical Process Monitoring*, 353, 26-33.

104. LAWSON, C. L. & HANSON, R. J. (1974) *Solving Least Squares Problems* (Englewood Cliffs, New Jersey, Prentice-Hall).

105. BRO, R. & DE JONG, S. (1997) A Fast Non-Negative Constrained Least Squares Algorithm, *Journal of Chemometrics*, 11, 393-401.

106. VAN BENTHEM, M. H., KEENAN, M. R. & HAALAND, D. M. (2002) Application of Equality Constraints on Variables during Alternating Least Squares, *Journal of Chemometrics*, 16, 613-622.

107. TAULER, R. & JUAN, A. D. http://www.ub.es/gesq/mcr/mcr.htm.

108. PAATERO, P. (1997) Least squares formulation of robust non-negative factor analysis, *Chemometrics and Intelligent Laboratory Systems*, 37, 23-35.

109. JIANG, J. H., LIANG, Y. Z. & OZAKI, Y. (2003) On simplex-based method for self-modeling curve resolution of two-way data, *Chemometrics and Intelligent Laboratory Systems*, 65, 51-65.

110. SPJOTVOLL, E., MARTENS, H. & VOLDEN, R. (1982) Restricted Least-Squares Estimation of the Spectra and Concentration of 2 Unknown Constituents Available in Mixtures, *Technometrics*, 24, 173-180.

111. TAULER, R., SMILDE, A. K. & KOWALSKI, B. R. (1995) Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *Journal of Chemometrics*, 9, 31-58.

112. AMRHEIM, M., SRINIVASAN, B., BONVIN, D. & SCHUMACHER, M. M. (1996) On the rank deficiency and rank augmentation of the spectral measurement matrix, *Chemometrics and Intelligent Laboratory Systems*, 33, 17-33.

113. PELL, R. J., SEASHOLTZ, M. B. & KOWALSKI, B. R. (1992) The Relationship of Closure, Mean Centering and Matrix Rank Interpretation, *Journal of Chemometrics*, 6, 57-62.

114. SMILDE, A. K., HOEFSLOOT, H. C. J., KIERS, H. A. L., BIJLSMA, S. & BOELENS, H. F. M. (2001) Sufficient conditions for unique solutions within a certain class of curve resolution models, *Journal of Chemometrics*, 15, 405-411.

115. BRO, R. & SIDIROPOULOS, N. D. (1998) Least squares algorithms under unimodality and non-negativity constraints, *Journal of Chemometrics*, 12, 223-247.

116. JUAN, A. D., HEYDEN, Y. V., TAULER, R. & MASSART, D. L. (1997) Assessment of New Constraints Applied to the Alternating Least Squares Method, *Analytica Chimica Acta*, 346, 307-318.

117. MANNE, R. (1995) On The Resolution problem in Hyphenated Chromatography, *Chemometrics and Intelligent Laboratory Systems*, 27, 89-94.

118. WINDIG, W. (1994) The Use of Second Derivative Spectra for Pure-Variable Based Self-Modeling Mixture Analysis Techniques, *Chemometrics and Intelligent Laboratory Systems*, 23, 71-86.

119. WINDIG, W. (1997) Spectral data Files for Self-Modeling Curve Resolution with Examples using the Simplisma Approach, *Chemometrics and Intelligent Laboratory Systems*, 36, 3-16.

120. WINDIG, W. & HECKLER, C. E. (1992) Self-Modelling Mixture Analysis of Categorized Pyrolysis Mass Spectral Data with the SIMPLISMA Approach, *Chemometrics and Intelligent Laboratory Systems*, 14, 195-207.

121. WINDIG, W. & STEPHENSON, D. A. (1992) Self-Modelling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data Using the SIMPLISMA Approach, *Analytical Chemistry*, 64, 2735-2742.

122. GAMPP, H., MAEDER, M., MEYER, C. J. & ZUBERBUHLER, A. D. (1985) Calculation of equilibrium constants from multiway spectroscopic data, III. Model-free Analyisis of Spectrophotometric and ESR titrations, *Talanta*, 32, 1133-1139.

123. MAEDER, M. (1987) Evolving Factor Analysis for the Resolution of Overlapping Chromatograhic Peaks, *Analytical Chemistry*, 59, 527-530.

124. OHTA, N. (1973) Estimating Absorption-Bands of Component Dyes by Means of Principal Component Analysis, *Analytical Chemistry*, 45, 553-557.

125. SASAKI, K., KAWATA, S. & MINAMI, S. (1983) Constrained Non-Linear Method for Estimating Component Spectra from Multicomponent Mixtures, *Applied Optics*, 22, 3599-3603.

126. KELLER, H. R. & MASSART, D. L. (1991) Peak Purity Control in Liquid Chromatography with Photodiode Array Detection by Fixed Size Moving Window Evolving Factor Analysis, *Analytica Chimica Acta*, 246, 379-390.

127. LIANG, Y.-Z. & KVALHEIM, O. M. (1993) Heuristic Evolving Latent Projections: Resolving Hyphenated Chromatograhic Profiles by Component Stripping, *Chemometrics and Intelligent Laboratory Systems*, 20, 115-125.

128. KVALHEIM, O. M. & LIANG, Y.-Z. (1992) Heuristic Evolving Latent Projections: Resolving Two-Way Multicomponent Data. 1. Selectivity, Latent Projective Graph, Datascope, Local Rank, and Unique Resolution, *Analytical Chemistry*, 64, 936-946.

129. LIANG, Y.-Z., KVALHEIM, O. M., KELLER, H. R., MASSART, D. L., KIECHLE, P. & ERNI, F. (1992) Heuristic Evolving Latent Projections: Resolving Two-Way Multicomponent Data. 2. Detection and Resolution of Minor Constituents, *Analytical Chemistry*, 64, 946-953.

130. MALINOWSKI, E. R. (1992) Window Factor-Analysis - Theoretical Derivation and Application to Flow-Injection Analysis Data, *Journal of Chemometrics*, 6, 29-40.

131. VANDEGINSTE, B. G. M., DERKS, W. & KATEMAN, G. (1985) Multicomponent Self-Modeling Curve Resolution in High-Performance Liquid-Chromatography by Iterative Target Transformation Analysis, *Analytica Chimica Acta*, 173, 253-264.

132. WINDIG, W. (1997) Spectral Profiles for Self-Modelling Curve Resolution with Examples using the SIMPLISMA Approach, *Chemometrics and Intelligent Laboratory Systems*, 36, (3-16).

133. BRAEKELEER, K. D. & MASSART, D. L. (1997) Evaluation of the Orthogonal Projection Approach (OPA) and the SIMPLISMA approach on the Windig Standard Spectral Data Sets, *Chemometrics and Intelligent Laboratory Systems*, 39, (127-141).

134. BRAEKELEER, K. D., JUAN, A. D. & MASSART, D. L. (1999) Purity Assessment and Resolution of Tetracycline Hydrochloride Samples Analysed using High-Performance Liquid Chromatography with Diode Array Detection, *Journal of Chromatography A*, 832, (67-86).

135. TOFT, J. (1995) Evolutionary Rank Analysis Applied to Multidetectional Chromatographic Structures, *Chemometrics and Intelligent Laboratory Systems*, 29, 189-212.

136. GRUNG, B. & KVALHEIM, O. M. (1995) Resolution of multicomponent profiles with partial selectivity. A comparison of direct methods, *Chemometrics and Intelligent Laboratory Systems*, 29, 75-87.

137. HOPKE, P. K., ALPERT, D. J. & ROSCOE, B. A. (1983) Fantasia - a Program for Target Transformation Factor-Analysis to Apportion Sources in Environmental-Samples, *Computers & Chemistry*, 7, 149-155.

138. GEMPERLINE, P. & CASH, E. (1998) User's Guide for GUIPRO, pp. 9 (Greenville, East Carolina University).

139. GRANDE, B.-V. & MANNE, R. (2000) Use of convexity for finding pure variables in two-way data from mixtures, *Chemometrics and Intelligent Laboratory Systems*, 50, 19-33.

140. TURNER, N., FOTHERINGHAM, J. & SPEIGHT, R. (2004) Novel Biocatalyst technology for the preparation of chiral amines, *Innovations in Pharmaceutical Technology Journal*, 4, 114-122.

141. NOYORI, R. & HASHIGUCHI, S. (1997) Assymetric Transfer Hydrogenation Catalysed by Chiral Ruthenium Complexes, *Accounts of Chemical Research*, 30, 97-102.

142. UEMATSU, N., FUJII, A., HASHIGUCHI, S., IKARIYA, T. & NOYORI, R. (1996) Asymmetric Transfer Hydrogenation of Imines, *Journal of American Chemical Society*, 118, 4916-4917.

143. MOHAR, B., VALLEIX, A., DESMURS, J. R., FELEMEZ, M., WAGNER, A. & MIOSKOWSKI, C. (2001) Highly enantioselective synthesis via dynamic kinetic resolution under transfer hydrogenation using Ru(eta 6-arene)- N-perfluorosulfonyl-1,2-diamine catalysts: a first insight into the relationship of ligand's pK(a) and the catalyst activity, *Chemical Communications*, 24, 2572-2573.

144. CASEY, C. P., SINGER, S. W., POWELL, D. R., HAYASHI, R. K. & KAVANNA, M. (2001) Hydrogen Transfer to Carbonyls and Imines from a Hydroxycyclopentadienyl Ruthenium Hydride: Evidence for Concerted Hydride and Proton Transfer, *Journal of American Chemical Society*, 123, 1090-1100.

145. YI, C. S. & HE, Z. (2001) Transfer Hydrogenation of Carbonyl Compounds Catalyzed by A Ruthenium-Acetamido Complex:nEvidence for a Stepwise Hydrogen Transfer Mechanism, *Organometallics*, 20, 3641-3643.

146. BLACKMOND, D. G., ROPIC, M., STEFINOVIC, M., HODGES, G. G. R. & BLACKER, A. J. (2006) Kinetic Studies of the Asymmetric Transfer Hydrogenation of Imines with Formic Acid Catalysed by Rh-Diamine Catalyst, *Organic Process Research & Development*, 10, 457-463.

147. ROPIC, M. (2006) The Kinetic Mechanism of the Catalysed Asymmetric Transfer Hydrogenation of 1-Methyl-6,7-dimethyoxy-3,4-dihydroisoquinoline *Chemistry* (Hull, The university of Hull).

148. CROSS, D. J., KENNY, J. A., HOUSON, I., CAMPBELL, L., WALSGROVE, T. & WILLS, M. (2001) Rhodium Vesus Ruthenium: Contrasting Behaviour in the Asymmetric Transfer Hydrogenation of alpha-substituted acetophenones, *Tetrahedron: Asymmetry*, 12, 1801-1806.

149. RICHARDS, S., MILLER, R. & GEMPERLINE, P. (2008) Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Penalty Alternating Least Squares (P-ALS) Extension to Multi-way Problems, *Applied Spectroscopy*, 62.

150. JUAN, A. D. & TAULER, R. (2001) Comparison of three-way resolution methods for non-trilinear chemical data sets, *Journal of Chemometrics*, 15, 749-772.

151. MILLER, R. (2003) Comparison of the Chemometric Methods to Solve Rank deficiency in the Esterification Reaction of !-butanol by Acetic Anhydride *Department of Chemistry*, pp. 92 (Greenville, East Carolina University).

152. WORKMAN, J. J. (1996) Interpretive Spectroscopy for Near Infrared, *Applied Spectroscopy Reviews*, 31, 251-320.

153. WEYER, L. G. & LO, S.-C. (2002) Spectra-Structure Correlation in the Near-Infrared, in: Griffiths, P. R. (Ed.) *Handbook of Vibrational Spectroscopy* (Chichester, John Wiley & Sons).

154. GEMPERLINE, P. (2004) Base catalysed esterification of acetic anhydride, in: Richards, S. (Ed.) (Personal Communication, Greenville).

155. RICHARDS, S. & WALMSLEY, A. D. (2007) Quantitative Iterative Target Transformation Factor Analysis, *Journal of Chemometrics*, 22, 63-80.

156. WINDIG, W., ANTALEK, B., LIPPERT, J. L., BATONNEAU, Y. & BREMARD, C. (2002) Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach, *Analytical Chemistry*, 74, 1371-1379.

157. WINDIG, W., GALLAGHER, N. B., SHAVER, J. M. & WISE, B. M. (2005) A new approach for interactive self-modeling mixture analysis, *Chemometrics and Intelligent Laboratory Systems*, 77, 85-96.

158. NEXANT (2000) Fluidized Bed Vinyl Acetate Process (PERP Program, http://www.chemsystems.com/search/docs/abstracts/98S3abs.pdf).

159. BP, C. (2006) Vinyl Acetate (BP Chemicals, http://www.bp.com/sectiongenericarticle.do?categoryId=9003635&contentId=7006749).

160. HOSOKAWA, T. & MURAHASHI, S. I. (1990) New Aspects of Oxypalladation of Alkenes, *Accounts of Chemical Research*, 23, 49-54.

161. BP, C. (2006) BP Chemicals Announces Major Advance in VAM Technology (BP Chemicals, http://www.bp.com/genericarticle.do?categoryId=2012968&contentId=2000889).

V Further Work

VI Appendix

Table of Contents

TABLE OF CONTENTS I
1.1 CATHY DATA PRETREATMENT I
1.2 PRINCIPLE COMPONENT ANALYSIS VII
1.3 NEEDLE SPECTRAL INITIALISATION VIII
1.4 NWAY P-ALS XII

1.1 CATHy Data Pretreatment

Referenced from section II.1.3.3, pg. 80

Baseline Correction	Data	Initial Estimate	No. Iterations	LOF(%)	RMS PE Imine	RMSPE Amine
None	FTIR (1)	Pure spectra	5	2.46	0.04	0.72
Zero average offset	FTIR (1)	Pure spectra	30	2.36	0.05	0.74

Table 1. Results of MCR-ALS analysis of the negative FTIR(1) dataset using different baseline correction methods and the neat spectra of imine, amine and carbon dioxide as starting estimates for ALS. The experimental conditions for ALS resolution is given in experiment 11 (pg 98).

1.1.1 No Baseline Correction

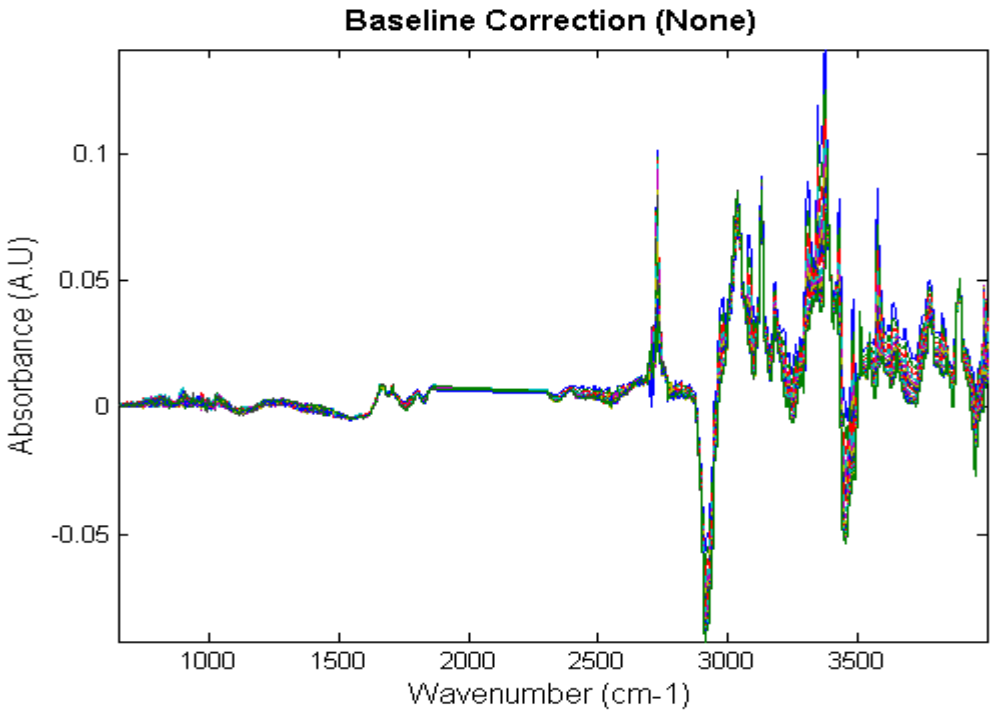


Figure 1. Non-negative FTIR(I). No baseline correction applied to FTIR profiles

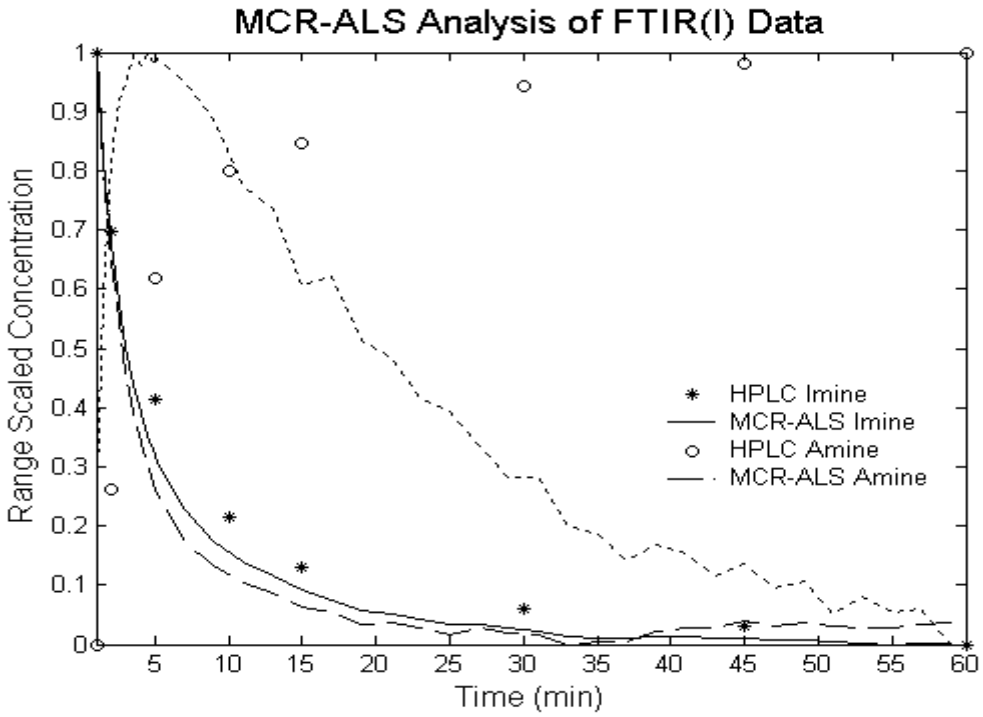


Figure 2. MCR-ALS resolution of the concentration profiles determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with no baseline correction. The amine (MCR-ALS) concentration profile is predicted incorrectly

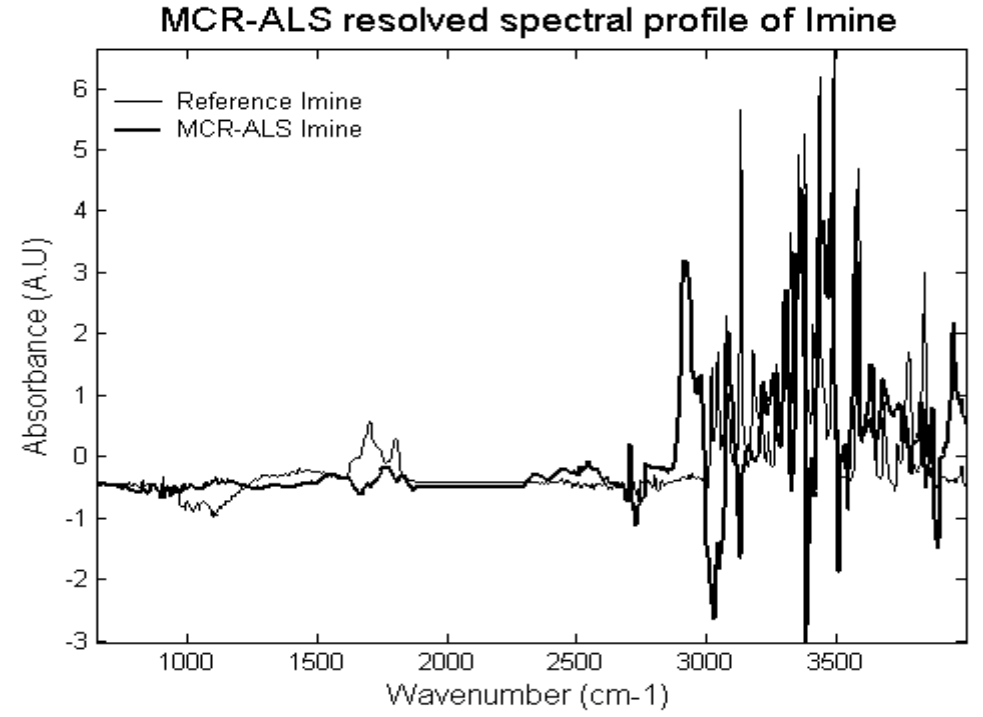


Figure 3. MCR-ALS resolution of the imine spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with no baseline correction. The imine MCR-ALS profile contains contribution from amine.

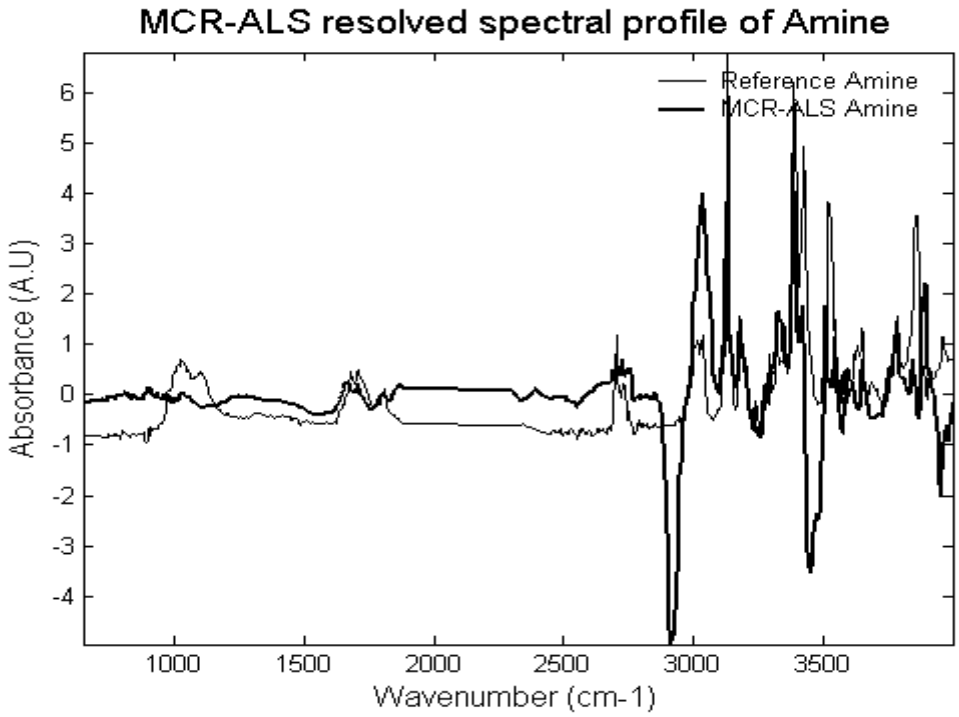


Figure 4. MCR-ALS resolution of the amine spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with no baseline correction. The amine MCR-ALS profile contains contribution from formic acid.

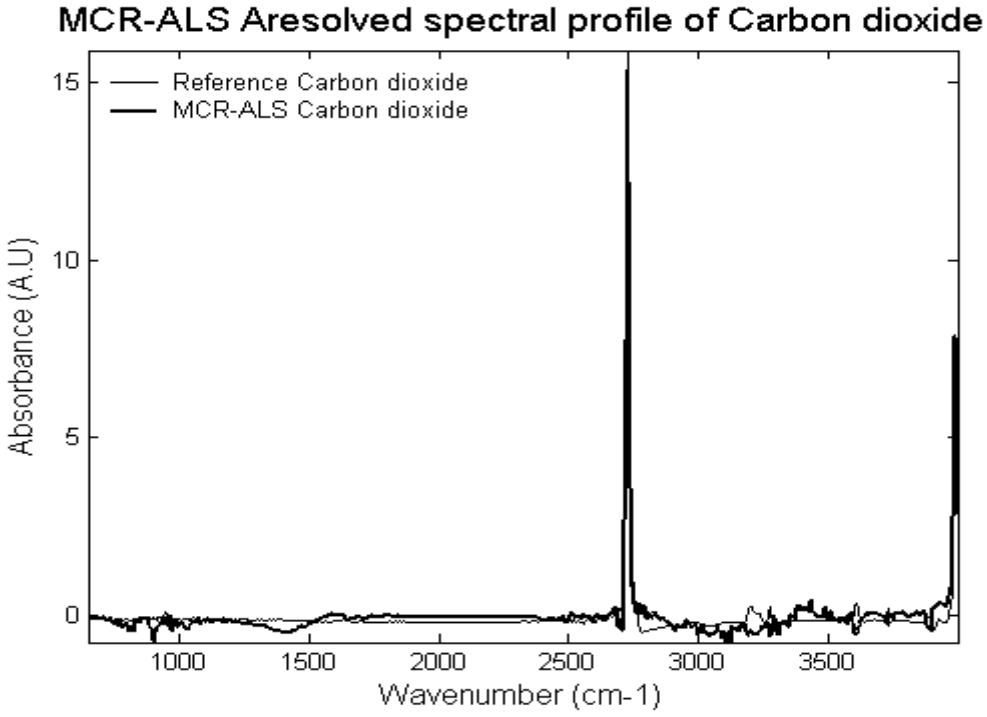


Figure 5. MCR-ALS resolution of the carbon dioxide spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with no baseline correction. Good prediction of the carbon dioxide spectrum.

1.1.2 Zero Average Offset

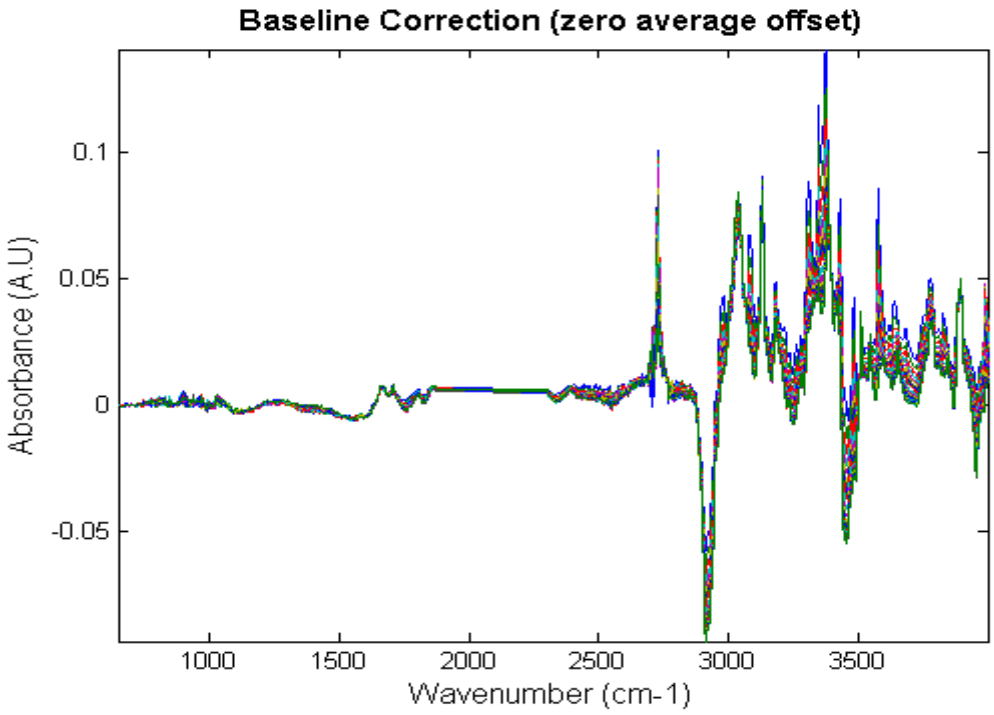


Figure 6. Non-negative FTIR(I). Zero average offset (3917-3998 cm^{-1}) applied to FTIR profiles

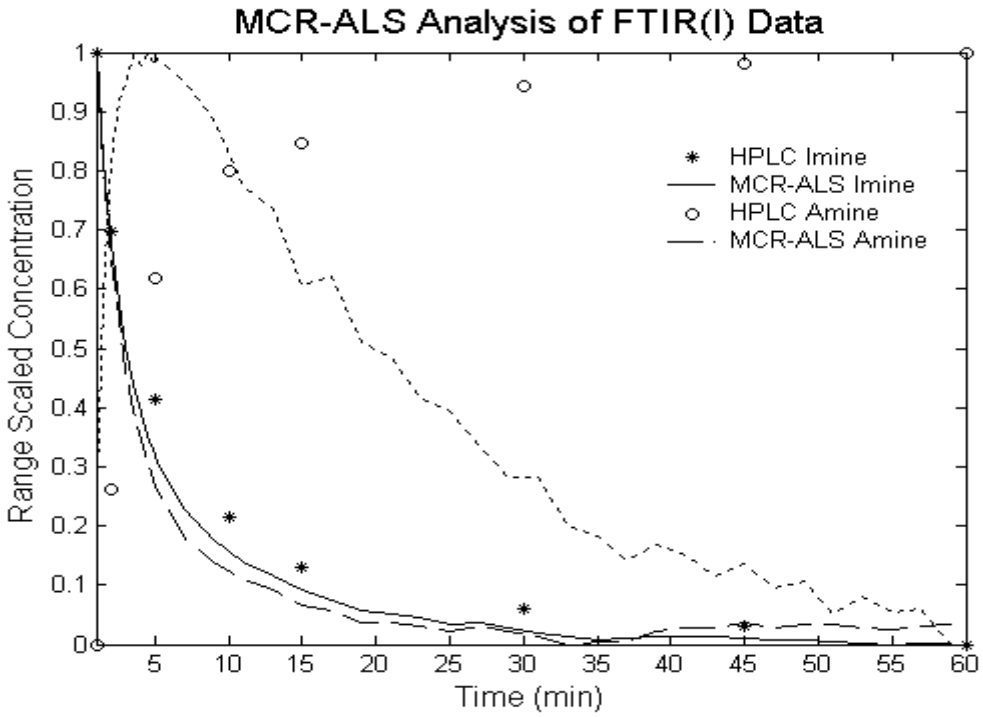


Figure 7. MCR-ALS resolution of the concentration profiles determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with zero average offset correction (3917-3998 cm^{-1}). The amine MCR-ALS profile is predicted incorrectly

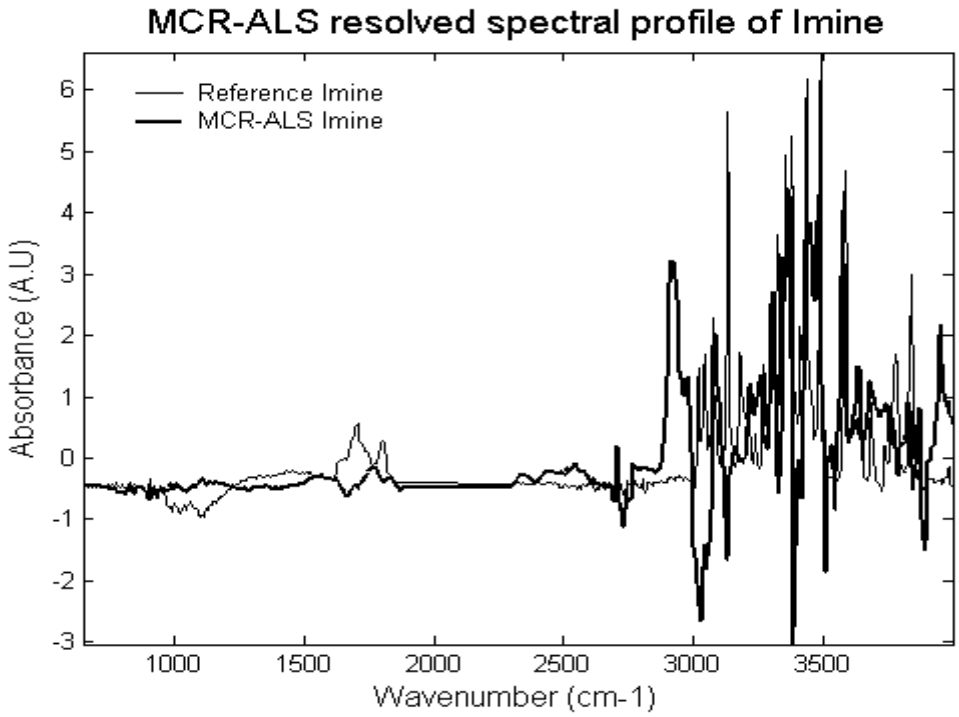


Figure 8. MCR-ALS resolution of the imine spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with zero average offset correction (3917-3998 cm^{-1}. The imine MCR-ALS profile contains contribution from amine.

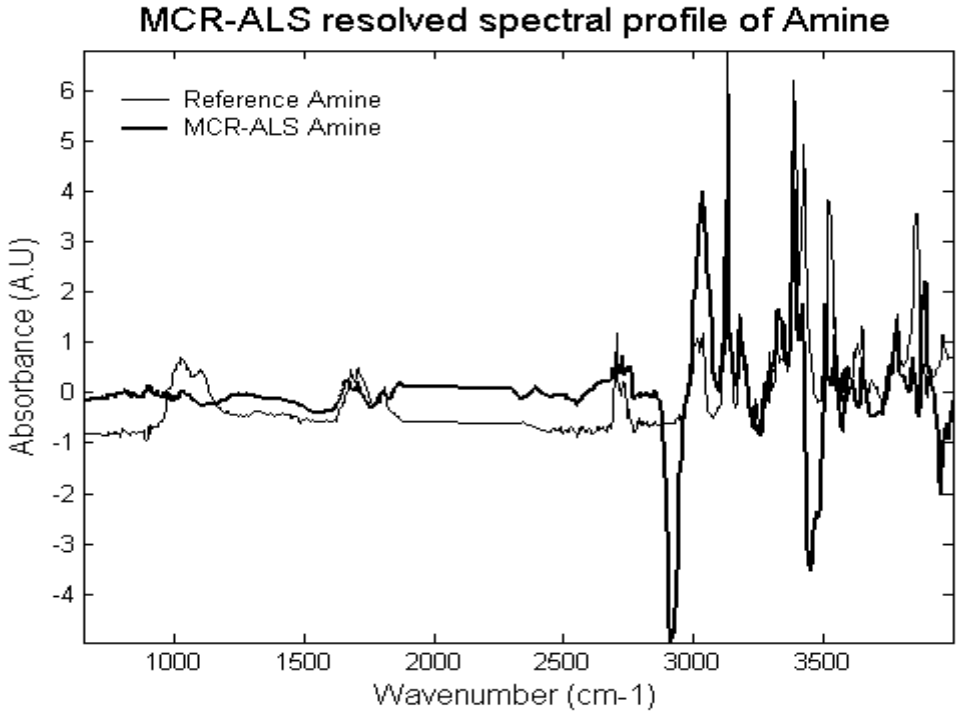


Figure 9. MCR-ALS resolution of the amine spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with zero average offset correction (3917-3998 cm^{-1}. The amine MCR-ALS profile contains contribution from formic acid.

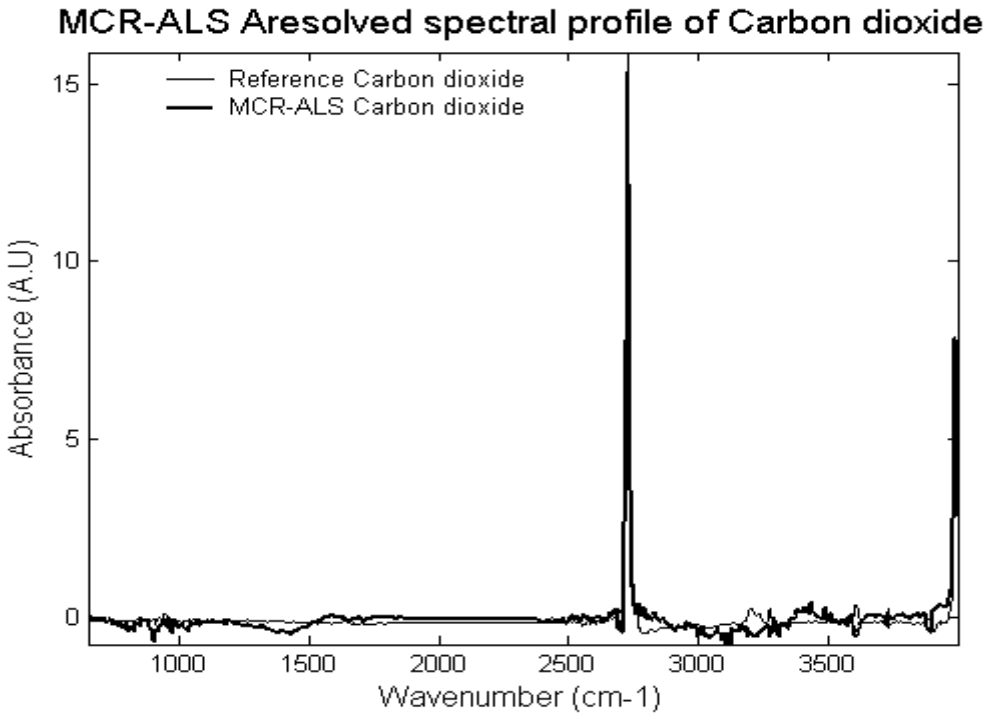


Figure 10. MCR-ALS resolution of the carbon dioxide spectral profile determined using experiment 11 (pg. 98) MCR-ALS protocol and the FTIR profiles with zero average offset correction (3917-3998 cm^{-1}). The carbon dioxide profile is predicted correctly

1.1.3 Minimum Offset method

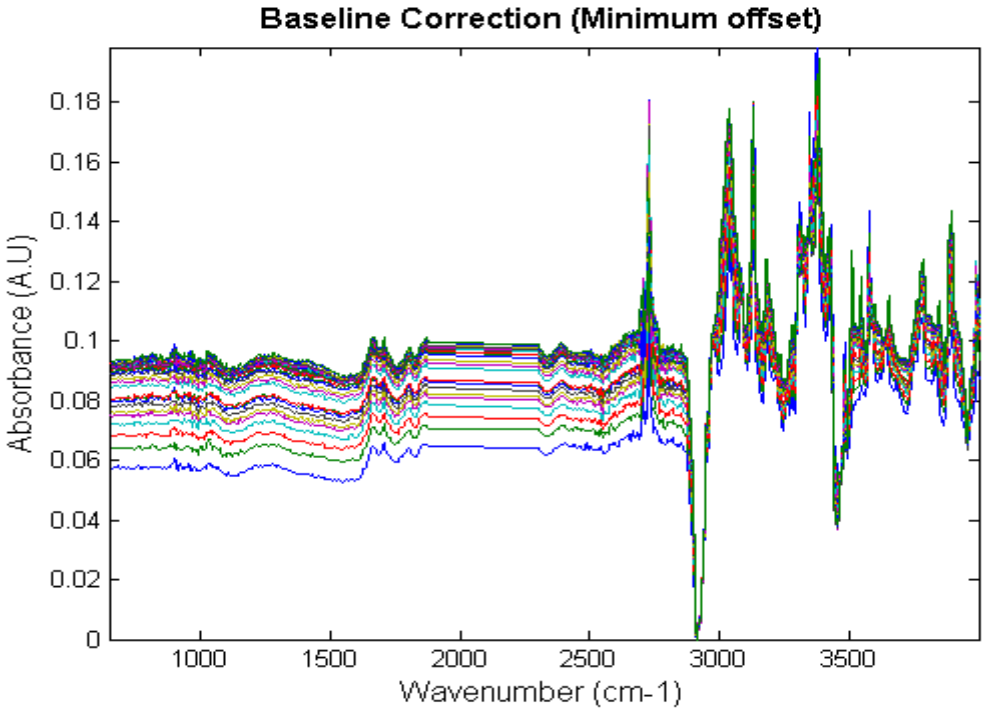


Figure 11. Non-negative FTIR(I). Minimum offset applied to FTIR profiles, which not maintained the original structure of the data, which is evident from the baseline shift between 800-2500 cm^{-1}. No further analysis completed with this measurement matrix.

1.2 Principle Component Analysis

1.2.1 FTIR(II) PCA Scores and Loadings plot

Referenced from section II.1.3.4, pg. 86

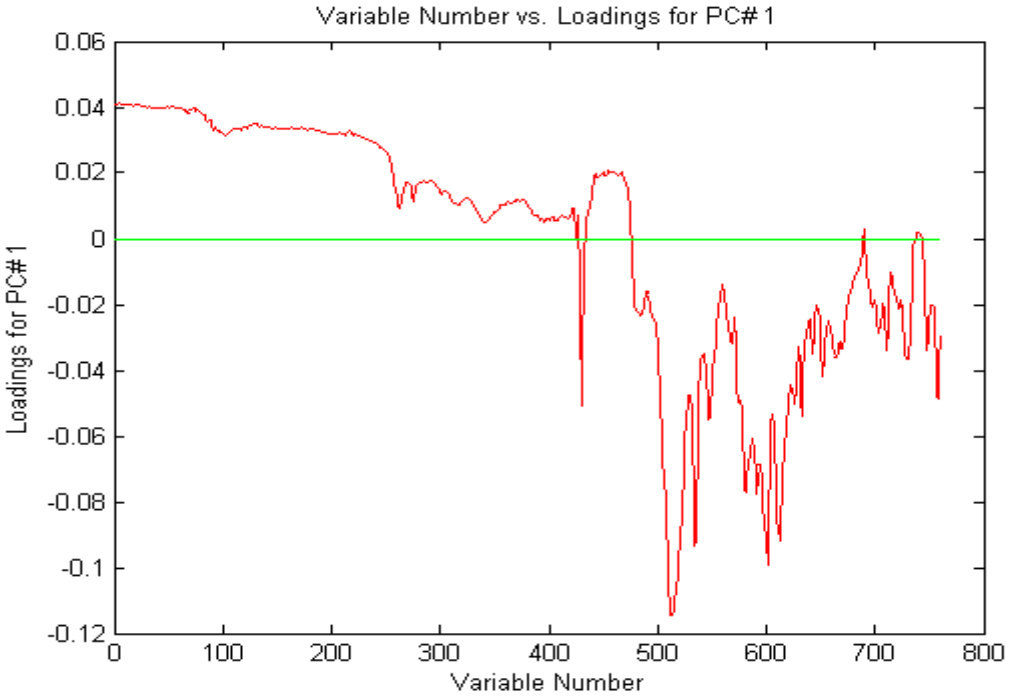


Figure 12. PCA loadings calculated from the FTIR (II) data. The first loading plot contained the characteristic functional group frequencies of amine and the mixture spectrum of triethylamine and formic acid,

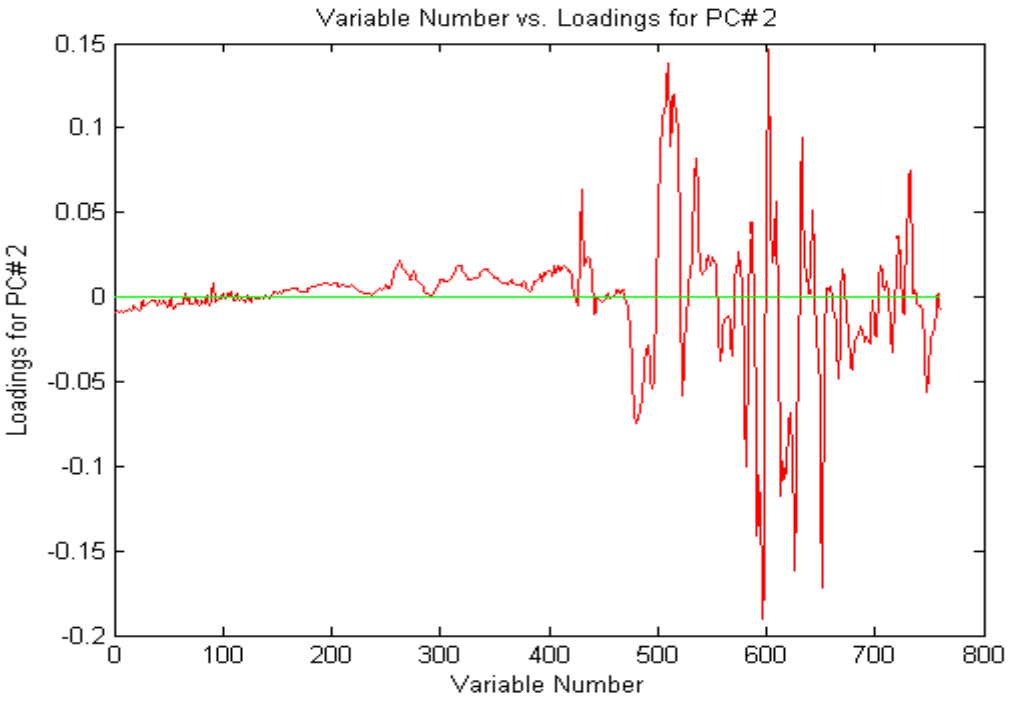


Figure 13. PCA loadings calculated from the FTIR (II) data. The second loading plot contained the common functional group frequencies of imine and amine

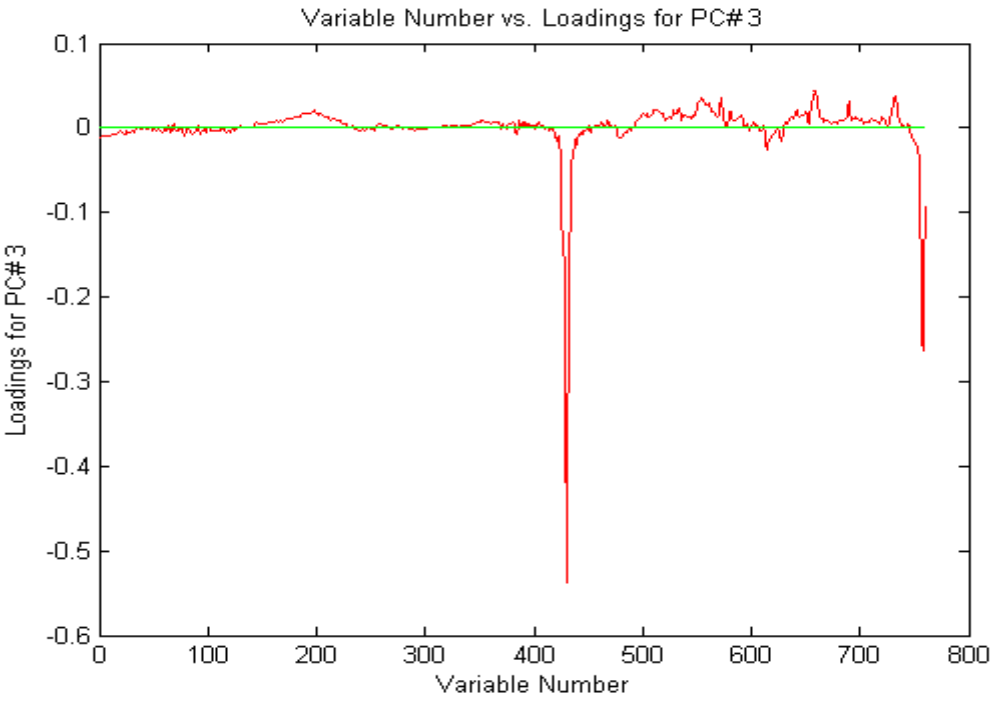


Figure 14. PCA loadings calculated from the FTIR (II) data. The third loading plot contained characteristic group frequencies of carbon dioxide, with a small contribution from imine and amine

1.3 Needle Spectral Initialisation

1.3.1 FTIR(I)

Referenced from section II.1.3.6.1, pg.100

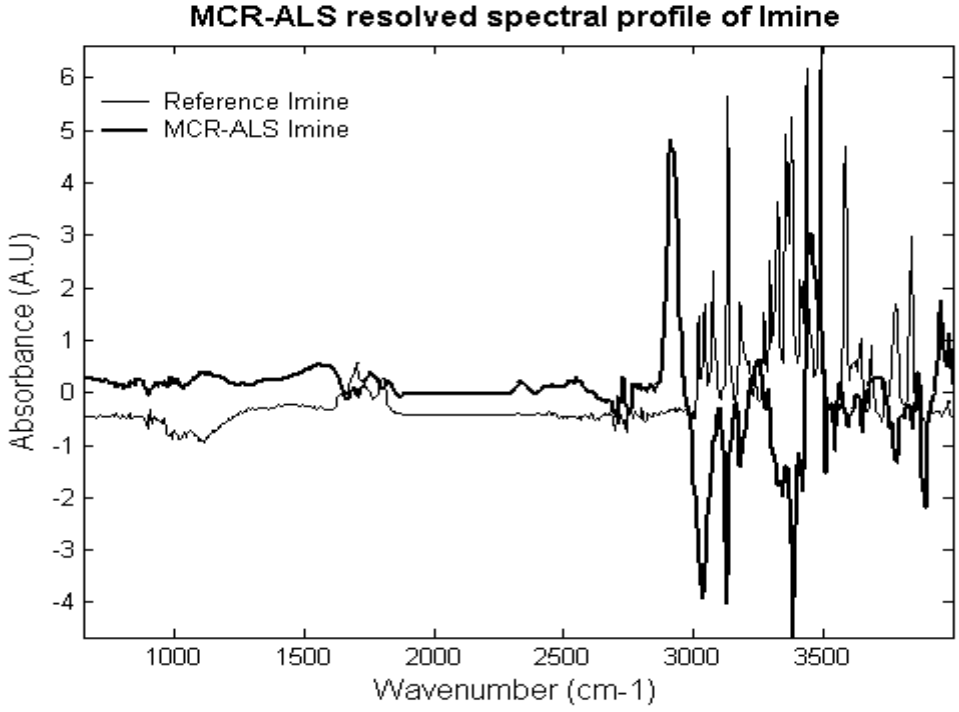


Figure 15. MCR-ALS analysis of the FTIR(I) data using the needle spectral estimates (expt 7). The imine spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of imine contains characteristic and common functional group frequencies attributed to imine, amine and formic acid

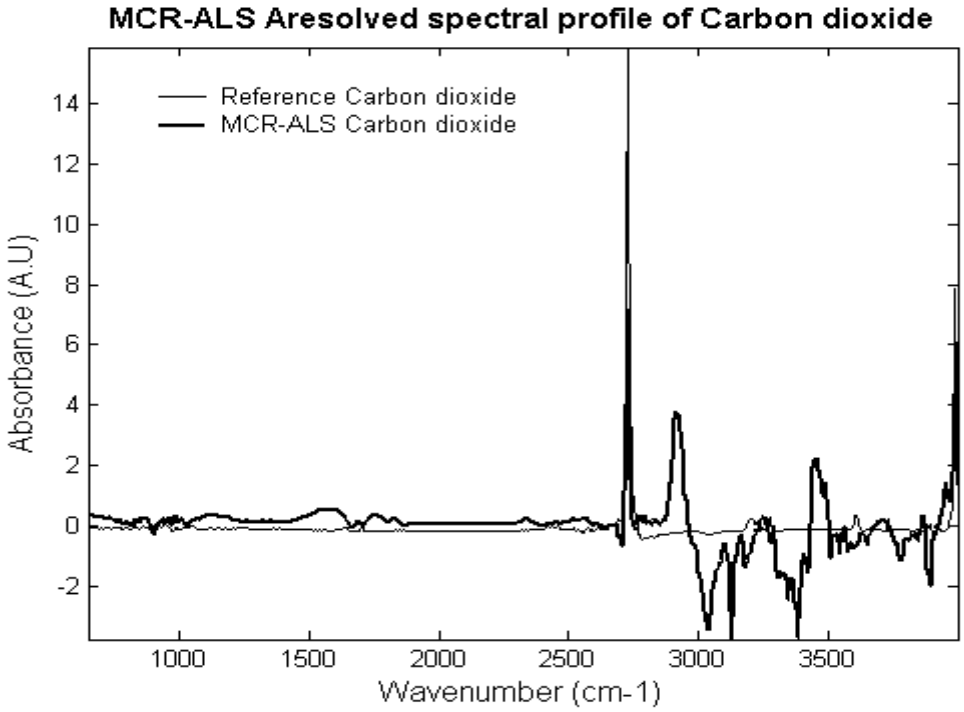


Figure 16. MCR-ALS analysis of the FTIR(I) data using the needle spectral estimates (expt 7). The carbon dioxide spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of carbon dioxide contains characteristic and common functional group frequencies attributed to imine, amine and formic acid

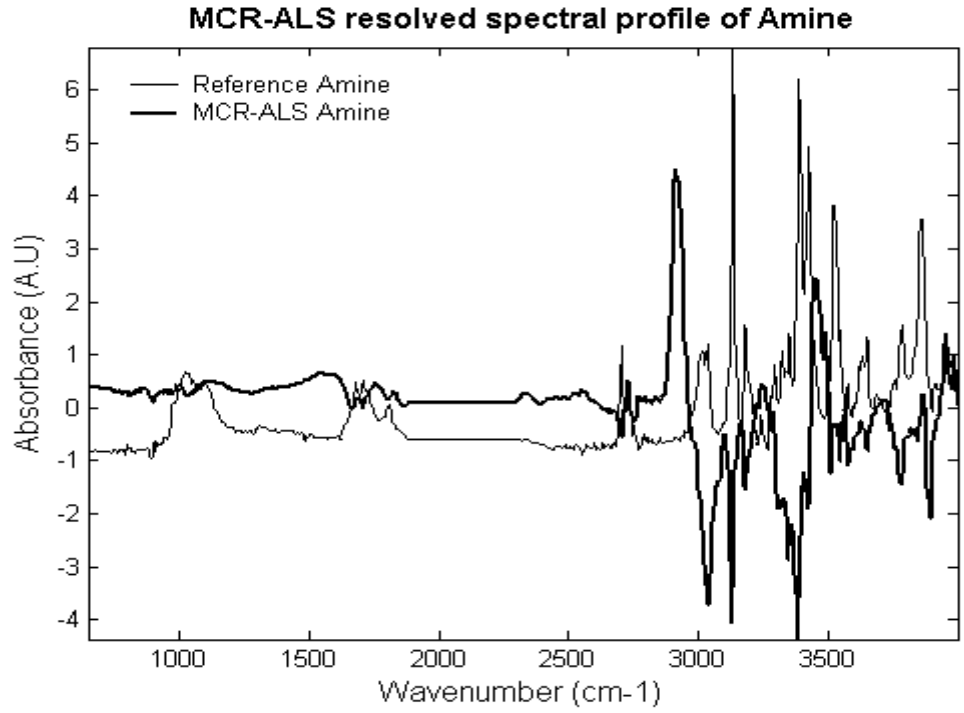


Figure 17. MCR-ALS analysis of the FTIR(I) data using the needle spectral estimates (expt 7). The amine spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of amine contains characteristic and common functional group frequencies attributed to imine, amine and formic acid

1.3.2 FTIR(II)

Referenced from section II.1.3.6.1, pg.101

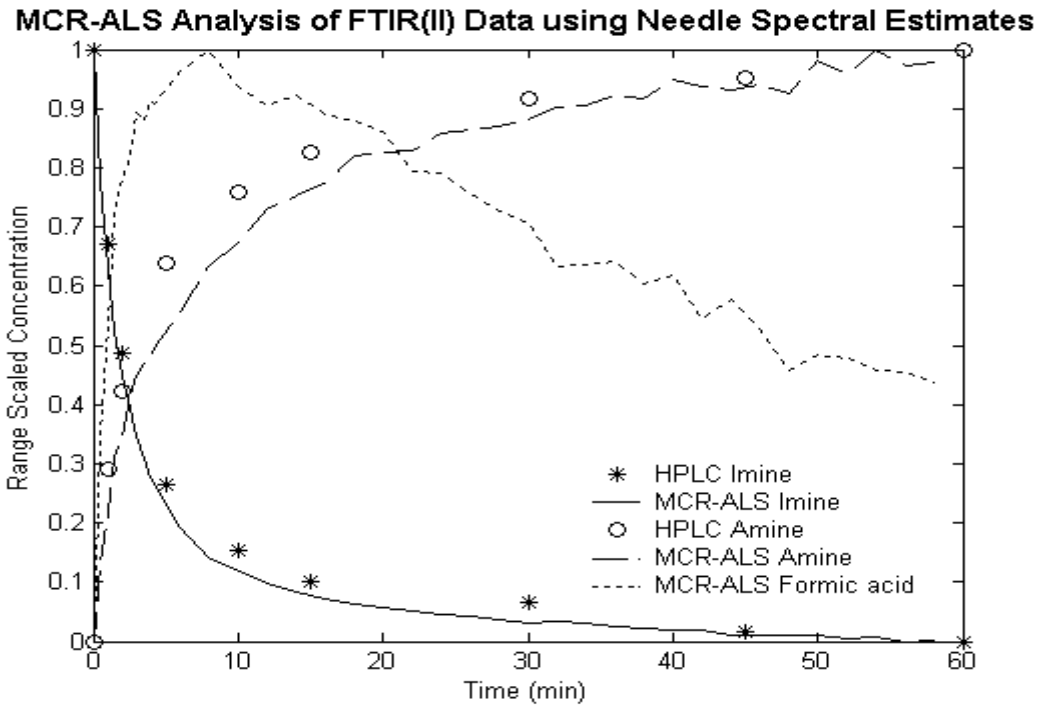


Figure 18. Concentration profiles resolved from the MCR-ALS analysis of the FTIR(II) measurement matrix using needle spectral estimates (Expt. 9). Good prediction of the imine and amine concentration profiles

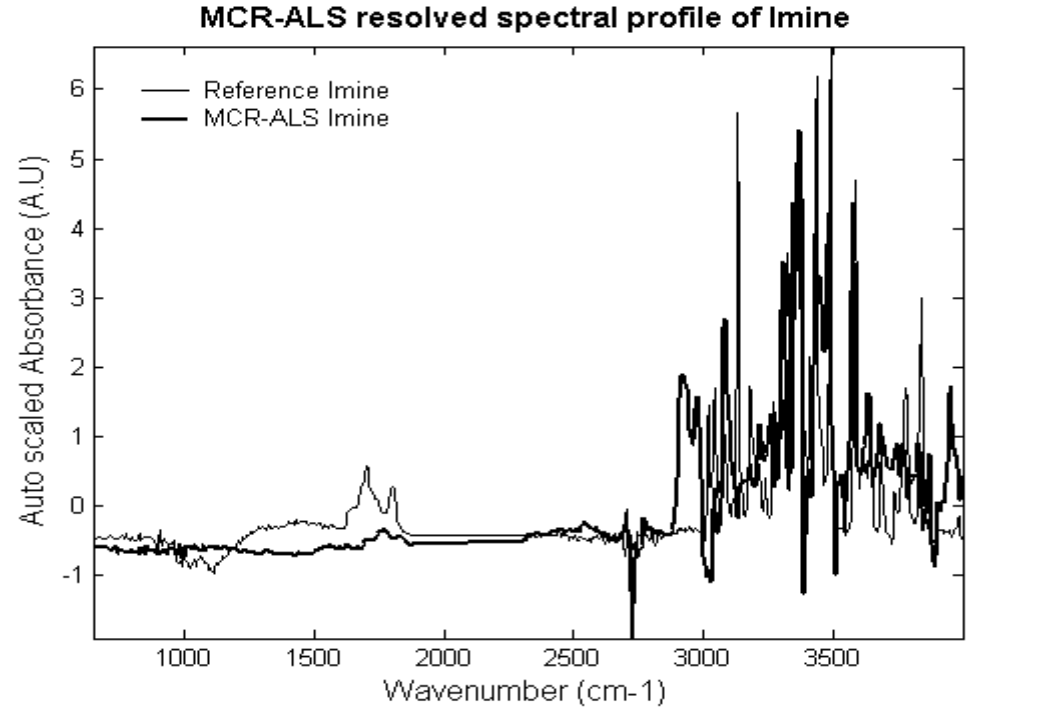


Figure 19. MCR-ALS analysis of the FTIR(II) data using the needle spectral estimates (expt 9). The imine spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of imine contains contributions from amine.

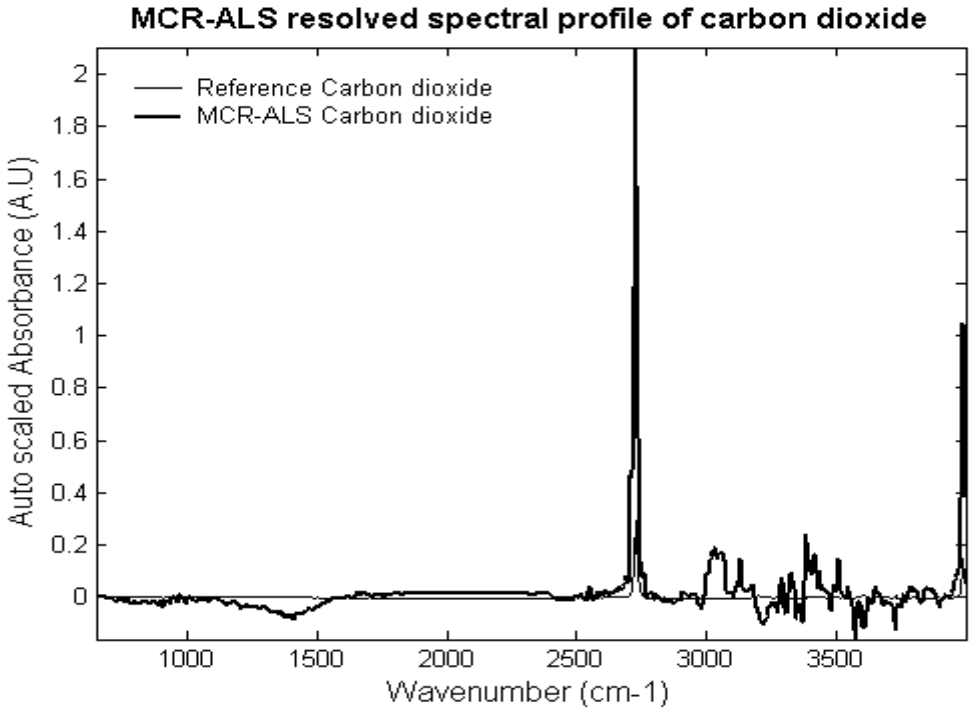


Figure 20. MCR-ALS analysis of the FTIR(II) data using the needle spectral estimates (expt 9). The carbon dioxide spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of carbon dioxide contains contributions from imine and amine.

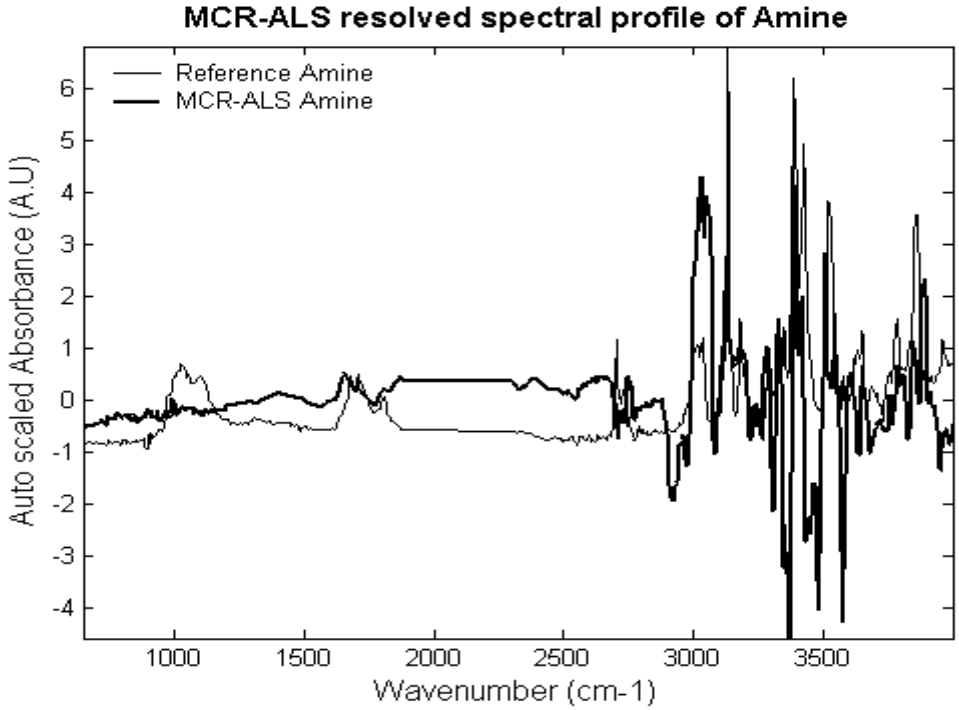
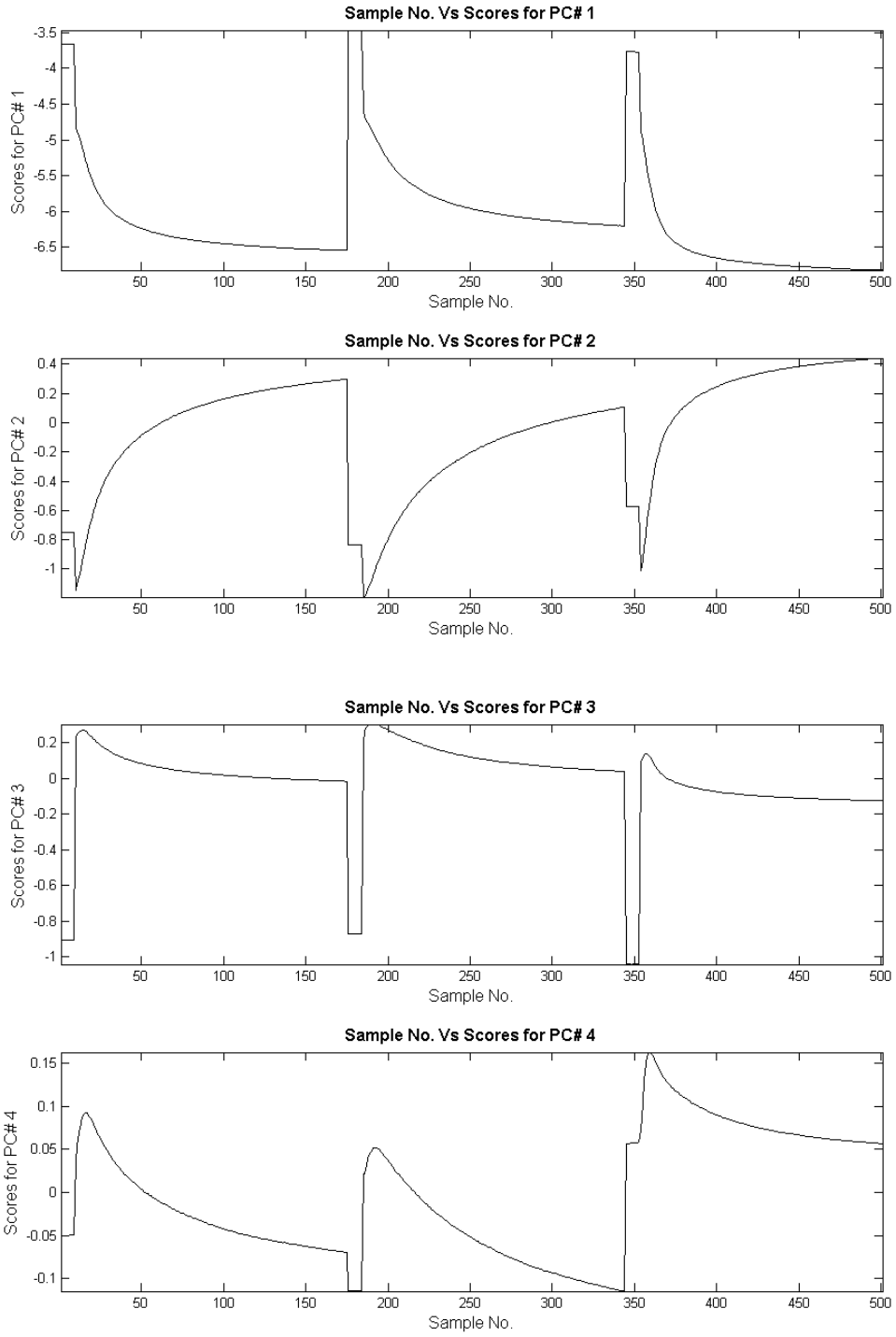


Figure 21. MCR-ALS analysis of the FTIR(II) data using the needle spectral estimates (expt. 9). The amine spectral profile resolved using MCR-ALS analysis. The resolved spectral profiles of amine contains contributions from imine and amine

1.4 NWAY P-ALS

1.4.1 PCA scores plots

Referenced from section II.3.3.4, pg.127



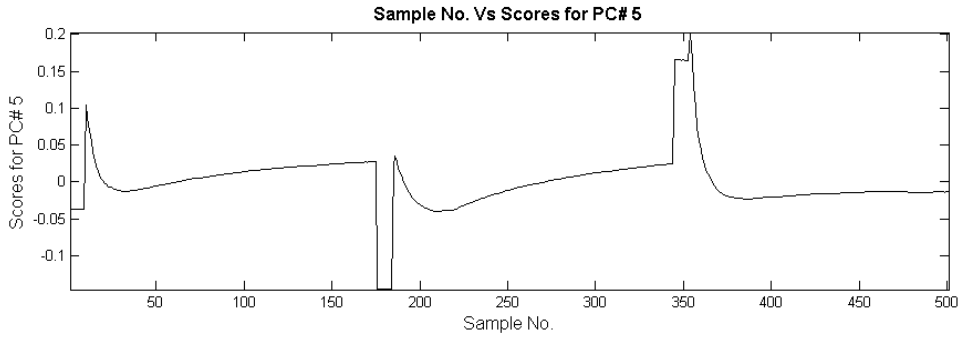


Figure 22. PCA analysis of the column-wise augmented NIR measurement matrix. Visual inspection of the PCA scores confirms structural variance in each of the four PC. The fifth component varies quite dramatically over the three batches.

Principle component number	Eigenvalue of Cov(X)	% Variance Captured by this PC	% Variance captured Total
1	3.80e+001	99.41	99.41
2	1.58e-001	0.41	99.98
3	6.09e-002	0.16	99.99
4	5.18e-003	0.01	99.99
5	1.48e-003	0.00	100.00

Table 2. PCA variance statistics calculated for the five PC's.

TITLE

A novel approach to the quantification of industrial mixtures from the Vinyl Acetate Monomer (VAM) process using Near Infrared spectroscopic data and a Quantitative Self Modeling Curve Resolution (SMCR) Methodology.

AUTHORS

Selena Richards[1*], Edo Becker[2], Romá Tauler[3], Anthony Walmsley[4]

SUBMITTED: Chemometrics and Intelligent Laboratory Systems

ABSTRACT

This work demonstrates the application of a new quantitative self modeling curve resolution (SMCR) approach for the simultaneous qualitative and quantitative recovery of reaction constituents; ethylene, acetic acid, water, vinyl acetate monomer and carbon dioxide from the BP Chemicals Vinyl Acetate Monomer (VAM) process. A cheaper, easier and faster method for the calibration of the VAM process was required because the current calibration procedure is time consuming and expensive. A quantitative SMCR strategy which uses a correlation constraint (regression constraint) during the Alternating Least Squares (ALS) procedure was used to quantify each reaction constituent. Starting estimates for ALS were determined using Quantitative Iterative target Factor analysis (QITTFA) and the NIR spectroscopic data. Vinyl acetate could not be vaporised, therefore QITTFA was selected to provide starting estimates approximating the true solution in the absence of selectivity and *a priori* knowledge. The results were compared to a well-established multivariate calibration method; Partial Least Squares (PLS) using a non-parametric statistical randomisation test for multivariate calibration models and the model reference error (relative error (RE)) for the prediction of each constituent. It was concluded that the quantitative SMCR procedure could be used to quantify ethylene 9.06% (RE), acetic acid 19.30% (RE), water 13.77% (RE) and carbon dioxide 30.46% (RE) within the defined relative error margin. The advantages of the new approach were a ~90% reduction in the calibration time, ~90% reduction in the number of training samples required for the calibration and the simultaneous recovery of the reaction constituent spectral profiles. Therefore this quantitative SMCR strategy could be used for reactions or processes for which it is not possible to prepare mixtures of known composition, due to the absence of isolated reference material, stability issues and where the preparation of such samples are time consuming and expensive.

Keywords: Vinyl Acetate Monomer (VAM), Correlation Constraint, Quantitative Iterative Target Transformation Factor Analysis (QITTFA), NIR Spectroscopy, Multivariate Calibration, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS).

INTRODUCTION

Process Analytical Control (PAC) is required to control and optimise the performance of a chemical process in terms of capacity, quality, cost, consistency and waste reduction[1]. Typically on-line process analysers, such as Mid Infrared (MIR), Near Infrared (NIR), Raman, Fourier Transform Infrared (FTIR) etc., provide not only physical process parameters, such as temperature, pressure, flow rate and liquid level, but also molecular parameters relating to component concentrations, molecular structure and degree of reaction[2]. Due to the complexity of some process samples and robustness of the analytical instruments, one is focused on finding relationships between the cheap measurements which are easy to acquire, and measurements, which are either expensive or labour intensive. Therefore the goal is to find good relationships to predict the expensive measurements rapidly and with high accuracy, from the cheaper ones[3]. Factor analysis based methods, such as multivariate calibration methods[3-7] and Self Modelling Curve Resolution methods (SMCR)[8, 9] are linear models which can be applied to compress and extract relevant information from the measurements. The ultimate goal of multivariate calibration methods is the establishment of a calibration model from multivariate measurements allowing the quantitative determination of the analyte in the presence of unknown interferents or in a complex chemical matrix, even if the analyte signal selectivity is poor. SMCR methods decompose unresolved multi-

component and multivariate measurement matrices into pure factors, such as spectral profiles, concentration profiles, pH profiles, for individual species with no *a priori* knowledge of the system. The requirement for the application of curve resolution techniques is that the data should be at least, bilinear, i.e. the elements of the measurement matrix, **D** must be a linear sum or combination of the product terms **C** and **B**. Therefore the measurement must be of the form given in equation 1,

$$\mathbf{D} = \mathbf{C}\,\mathbf{B}^{T} + \mathbf{E} \qquad \textbf{Equation 1}$$

Where $\mathbf{D}\,(n \times m)$, has *n* absorption spectra in each row at *m* wavelengths. $\mathbf{C}\,(n \times nc)$ is the concentration profiles, for the *nc* components and $\mathbf{B}^{T}\ (nc \times m)$ is the corresponding spectral profiles for the pure components and $\mathbf{E}\,(n \times m)$, is the error matrix associated with the decomposition. Such measurements include spectroscopic measurement of multi-components systems, i.e. Infrared (IR), Near Infrared (NIR), Raman, or two-way data from hyphenated chromatographic methods with multi-channel detection. Recent applications of SMCR techniques include quantification of trace analytes[10], peak purity assessments[11], characterisation of batch reactions[12, 13] and on-line reaction monitoring[14] see references[15, 16] for comprehensive reviews.

In this paper, the feasibility of an alternative approach to multivariate calibration is presented. A cheaper, easier and faster method for the calibration of the vinyl acetate process was required because the current calibration procedure requires over 300 hundred calibration samples and each sample takes approximately 20 minutes to create. Currently there are no commercially available calibration mixtures of the gases (ethylene, acetic acid, carbon dioxide, water and vinyl acetate). Consequently this procedure is time consuming and expensive. To address this issue a quantitative SMCR strategy was applied to the gaseous NIR mixtures to simultaneously resolve the reaction constituents. The quantitative procedure uses a correlation (regression) constraint developed by Antunes *et al* [17] and an exploratory approach called Quantitative

Iterative Target Transformation Factor Analysis (QITTFA)[18]. QITTFA was developed to provide initial estimates approximating the true solution. No *a priori* information regarding the process, such as key spectral features or calibration information was required. In a previous study it was found that QITTFA improved the performance of Simple-To-Use-Interactive Self-modeling Mixture Analysis (SIMPLISMA) [19-23] in cases where pure variables did not exist or where the contribution of these components were low. In addition components of differing spectral characteristics shapes (narrow or broad spectral features) were resolved, without *a priori* knowledge of the shapes of the constituents[18].

The advantage of the quantitative SMCR approach are (a) marked reduction in the number of training samples (b) marked reduction in sampling time and (c) simultaneous resolution of qualitative information i.e. the pure spectral profiles from the mixture measurement matrix. The prediction ability of the quantitative SMCR models were compared with PLS models using a non-parametric statistical randomised test and the model reference error (relative error (RE)) for the prediction of each constituent. A description of the quantitative SMCR procedure is outlined below and given in figure 1.

THEORY

Scalars, including elements of vectors and matrices, are indicated by lower case italics, i.e. x, y and z. Vectors by bold lowercase characters, i.e. $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. Bold capitals are used for two-way matrices i.e. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$. The letters n and m are reserved for indicating the dimension of the first and second mode of the two-way matrix and i and j are used as indices for each of these modes.

Initial estimates of the pure spectral profile of each constituent were determined from the NIR spectral profiles of the multi-component mixture using QITTFA. The inputs required for the QITTFA procedure are the number of components, a noise correction factor, the maximum number of iterations and the selection of constraints.

The initial estimates were used to initialise the ALS procedure. Constrained ALS steps are used to fit the initial estimates ($\mathbf{B}^{\mathbf{T}}$) producing better "constrained" estimates. An estimate of the unknown species concentration profile is given by least squares, which is simply, $\mathbf{C} = \mathbf{D}^{*}\mathbf{B}^{+}$ and the new estimation of the spectra is given by $\mathbf{B} = \mathbf{C}_{0}^{+}\mathbf{D}^{*}$. Where $\mathbf{D}^{*}$ only contains the correlated variance due to the independent components. Constraints are used to obtain chemically meaningful solutions and include non-negativity constraints[24], selectivity constraints [25], unimodality constraints[26], correlation constraints[17] etc. The correlation constraint is a regression constraint, which enables quantitative analysis to be performed on unresolved mixtures. The correlation constraint is applied to the concentration profiles of known components. To implement the correlation constraint, the concentration values in matrix **C** for a particular analyte is divided into two subsets; those corresponding to the training subset used to build the calibration model and those corresponding to the test subset. Concentration values calculated for the training sample subset are then correlated with their known 'true' concentration values, using classical least squares and the best regression line are obtained in each case. Using both the linear equation and the concentration values obtained from MCR-ALS for the test sample subset, the predicted concentration values are obtained. These values are regressed against their true values and validated using the regression statistics described in the proceeding section. A graphical description of the correlation constraint is presented in figure 2, and is further expounded in the following text.

The Correlation Constraint

Initialisation

The concentration selectivity matrix $\mathbf{C}_{\mathbf{sel}}$ $(r \times nc)$ contains sparse known (training) concentration values for the target analyte (known concentration values

shown as black rectangles). The optimum number of known concentration samples required to build a good model using this procedure has not been investigated. The known concentration values in the vector $\mathbf{c}_{sel}$ should span the concentration range for the target analyte.

Least Squares

The QITTFA$_s$ initial estimates of the spectral profiles are used to initiate the ALS procedure, where, $\mathbf{B}^{+}\ (c \times nc)$ is the pseudo inverse of the QITTFA$_s$ spectral estimates, the estimated concentration matrix; $\mathbf{C}_{ALS}\ (r \times nc)$ is the concentration calculated from the least squares estimate, equation 2.

$$\mathbf{C}_{ALS} = \mathbf{D}\mathbf{B}^{+} \qquad \textbf{Equation 2}$$

Correlation Constraint

In **step A**, the known concentration values (**k**) in the $\mathbf{c}_{sel}$ vector corresponding to the target analyte is regressed against the corresponding estimated concentration values (**x**) in $\mathbf{c}_{ALS}$ vector of the target analyte using classical least squares, see equations 3-4.

$$\mathbf{k} = b_1\,\mathbf{x} + b_0 \qquad \textbf{Equation 3}$$

$$\mathbf{p} = b_1\,\mathbf{u} + b_0 \qquad \textbf{Equation 4}$$

Where b_1 is the regression coefficient and b_0 is the intercept. In **step B**, the model, defined in equation 4 is used to predict the concentration in the unknown (test) samples (**u**) in the $\mathbf{c}_{ALS}$ vector. In **step C**, the target analyte concentration vector is $\mathbf{c}_{ALS}\ (r \times 1)$ is updated with the predicted values (**p**) and the known values (**k**). In **step D** the new $\mathbf{C}_{ALS}$ matrix is used to predict the spectral profiles.

Regression Statistics

The calibration models were assessed using the regression statistic stipulated below, in each case the quantitative SMCR model was compared with the PLS model to determine whether the results were significantly different. The slope and offset of the

regression gives an indication of the accuracy of the models. The correlation coefficient between the reference concentrations, c_i and predicted concentration, $\hat{c}_i$ was calculated to determined whether a linear relationship existed between c_i and $\hat{c}_i$. A correlation value of plus one, represents a perfect positive correlation, a value of zero means that there is no correlation. Here, due to the complexity of the process data analysed, typical correlation greater than 0.990 was not expected. The Root Mean Square Error of Prediction (RMSEP) is a measure of the accuracy of prediction. The sum of the prediction error for all (*N*) samples for the training set was calculated to assess the future predictive properties of the calibration model and the predictive capabilities of the test set. RMSEP is measured in the same units as c_i, equation 5.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{c}_i - c_i)^2}{N}}$$ **Equation 5**

The Standard Error of Prediction (SEP) is a measure of the precision of prediction, equation 6. The bias (absolute deviation from $\bar{c}$) tracks the systematic prediction error, equation 7. The Relative Error (RE %) is similar to the lack of fit calculation, but gives a measure of the fit quality between the predicted and reference concentration, equation 8.

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{c}_i - c_i - bias)^2}{N-1}}$$ **Equation 6**

$$BIAS = \frac{\sum_{i=1}^{n}(\hat{c}_i - c_i)}{N}$$ **Equation 7**

$$RE(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(c_i - \hat{c}_i)^2}{\sum_{i=1}^{n} c_i^2}}$$

Equation 8

Convergence Criteria for ALS

The convergence criterion for ALS was assessed using the lack of fit (*lof*), which gives a measure of the relative fit quality between the experimental data and ALS reconstructed data. The solutions converge once the *lof* (%) is within the defined experimental error, equation 9. Where, d_{ij} is the experimental absorbance at the sampled point *i*, and the wavenumber *j*, and $\hat{d}_{ij}$ is the ALS calculated absorbance for that element.

$$lof(\%) = 100\sqrt{\frac{\sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2}}$$

Equation 9

Significance Testing

To determine whether the predictive ability of the quantitative SMCR models were comparable to the PLS1 models, a significance test was completed using the randomised t-test, based on Monte Carlo simulation[27]. The MSEP (mean squared error prediction) was used as a simple indicator for the predictive abilities of models. Each test was completed using a two-sided randomised t-test, 199 randomised trials, permitting a *p* value of $p = 0.005$. In a randomised trials if a significance value is below the significance level, i.e. $p = 0.005$, there is a significant difference between the prediction accuracy of the two methods. If a significance value is above this level, there is *no* significant difference between the prediction accuracy of the two methods.

EXPERIMENTAL

Data Acquisition

The spectra were acquired using a BOMEM MB-155 and MB-160 fitted with a TE-cooled InAs detector. Each spectrum was recorded using the average of 32 scans, the spectral region was 9998.2- 4497.7cm^{-1} and the resolution was 7.7cm^{-1}. Reference spectra of the vaporised pure components, ethylene, water, carbon dioxide and acetic acid were obtained prior to the analysis.

One hundred and sixty nine NIR calibration standards were prepared in the plant as mixtures of five organic components, ethylene (BOC Gases Ltd), carbon dioxide (BOC Gases Ltd), water (demineralised), acetic acid (BP Chemicals final product) and vinyl acetate (BP Chemicals final product). An automated mixing system, "stealth trolley", was designed specifically to make-up the calibration standards in the SPECAC NIR, Typhoon T13 gas cell, using evaporators and mass flow controllers. Each sample took ~20 minutes to prepare and was introduced into the gas cell at 120˚C. The partial pressure of each constituent and the total pressure for each calibration sample differed from sample-to-sample to mimic varying process conditions.

MATLAB6p5® (The Math Works, Natick, MA, USA) was used to complete the programming and calculations relating to QITTFA, MCR-ALS and PLS. The MCR-ALS algorithm can be obtained from (http://www.ub.es/gesq/mcr/mcr.htm). The PLS analysis was completed using the MATLAB PLS Toolbox (Eigenvector Research, Manson, WA, USA).

Data Pretreatment

The NIR spectroscopic samples acquired are given in figure 3. The data were baseline corrected using the minimum-offset method (removal of negative

absorbencies). Sample 168 was removed as an outlier because it had an abnormally high absorption.

Quantitative Analysis

In the quantitative investigation separate quantitative models were built to predict the quantities of ethylene, vinyl acetate, water, carbon dioxide and acetic acid in the calibration samples. The data set was divided into two subsets, one training set and one test set to allow for block validation of each model.

Training Set

Forty training samples were used in the calibration of each model (i.e. ethylene, carbon dioxide, water, acetic acid and vinyl acetate monomer). The concentration vectors were each ordered from the smallest to largest concentration, and five samples were selected every 50 samples along the ordered concentration range. The training samples were chosen such that they spanned the concentration range. Identical training samples were used to build both the corresponding constituent quantitative SMCR model and the corresponding constituent PLS models. The resulting predicted values from each model were regressed against the known reference values in order to assess the prediction capabilities of the models.

Test Set

The remaining one hundred and nineteen test samples were used to assess the performance of each model, that is the quantitative SMCR models and the PLS models for the prediction of ethylene, carbon dioxide, water, acetic acid and vinyl acetate monomer. Identical test sets were used to validate the corresponding constituent in the quantitative SMCR models and in the corresponding constituent PLS models. The

concentrations of test samples were estimated using both the quantitative SMCR models and the PLS models. The resulting predicted values were regressed against the known values in order to validate the models.

Reference values

The partial pressure for each constituent in the calibration sample was measured in bar absolute. The (%) volume for each constituent was calculated as a percentage of the total pressure. Here, it is important to note that the reference data was the partial pressure (which is a misnomer due to the molecular interactions) of each gas delivered into the mixing chamber. The relationship between pressure, P (atm), volume V (dm^3) and temperature, T (K) is approximate as the gases and gas mixtures are non-ideal. If the gaseous mixtures were ideal and under standard temperature and pressure, it would be possible to express the pressure as a function of PVT using the equation, $P = \left(\frac{n}{V}\right)RT$, and for mixtures $P = RT\sum\left(c_{nc}\right)$. The results which are presented here assume ideal behaviour. The inherent errors in the prediction of the individual constituents were acceptable for the process under study. This calibration technique also corrects for the presence of acetic acid dimers, which are also known to be present under these conditions.

Determination of Initial QITTFA Spectral Estimates

In the application of $QITTFA_s$, six components were chosen to describe the reduced space and to capture the structured variation. The sixth component contained a strong baseline component which was mixed with a spurious signal at approximately ~4500 cm^{-1}. This component was attributed to structured variance in the data, which was uncorrelated to the concentration profiles of the reaction constituent. In each analysis the needle output spectra were constrained with non-negativity constraints. An

offset value of 1% was used to select the pure variables from the output matrix and the maximum number of iterations was 500.

Qualitative SMCR Models

The QITTFA$_s$ spectral estimates were used to initiate the ALS procedure. During constrained ALS the spectral profiles were normalized to a height of one. Non-negativity constraints were applied in the each of the component spectra apart from vinyl acetate, as previous analysis showed that the non-negatively constrained vinyl acetate spectrum contained the free stretching first overtone of the OH group associated with monomer acetic acid. The concentration profiles were constrained with the non-negativity constraint. The convergence criterion was 7% (relative change of the LOF from one iteration to the next) and the maximum number of iterations was 500. The concentration profiles were scaled between 0 and 1 and the predicted test samples for ethylene, carbon dioxide, water, acetic acid and vinyl acetate monomer were regressed against the scaled known reference values.

Quantitative SMCR Models

The QITTFA$_s$ spectral estimates were used to initiate the ALS procedure. In the application of MCR-ALS the spectral profiles were normalised to a height of one and the correlation constraints were applied in the concentration profile. The aforementioned convergence criterion was maintained (see qualitative analysis). The training and test set used in MCR-ALS with correlation constraints are described above. However, the training and test set have a slightly different meaning with respect to the normal use of the respective terms. Here, the training set refers to the subset of known concentration values which were used in the MCR correlation constraint (**k**). The test set refers to the subset of concentration which were predicted by the resulting model

(**p**), see correlation constraint. The resulting $\mathbf{c}_{ALS}$ values from the training set and test set for each of the five calibration models were regressed against the known reference values in order to validate the models and assess the prediction capabilities of the models.

PLS1 models

Initially the analysis was completed on the mean centred NIR data and compared to the analysis with no mean centring, the comparison of the residual error for the mean centred and not mean centred data for each model was not significant. The results for the analysis with no mean centring is presented and compared to the resolution with no constraint and the correlation constraint. The NIR spectroscopic data was divided in to a training set and test set (see above). The contiguous block, cross validation method with 7 points was used to determine the number of latent variables to include in each model, see number of components, (N.o.C). In cases where this was difficult to decide, the error in the calibration and validation of the model was calculated, and the model which gave the smallest error was selected.

Reference Method Error

The original BP calibration model reference error (relative error) for the prediction of each constituent are give in table 1. No actual concentration values are reported in the results and discussion because this information is confidential.

RESULTS AND DISCUSSION

There were several regions of interest in the NIR spectra, which is shown in figure 3, representative of three of the five chemical constituents. The first overtone of the OH group associated with monomer acetic acid apparent at ~6994 cm^{-1}. The first

overtone of the asymmetric C-H stretch from ethylene appears at 6149 cm^{-1}, showing a PQR structure covering 6200-6090 cm^{-1} and the first overtone of the symmetric C-H stretch showing PR structure at 5993 cm^{-1} together with the combination bands with relatively strong absorbance appearing at 4760 cm^{-1}. The combination bands from the asymmetric and symmetric stretching modes of the water molecule appear at ~5150 and ~6900 cm^{-1}. Selective regions for vinyl acetate and carbon dioxide were not identified..

Initial Estimates

The QITTFA initial spectral estimates and the corresponding reference spectra are given in figure 4. The resolution of the initial estimates was not a trivial task, due to the severe overlap of bands from functional groups in different molecules, such as (1) the first overtones and combination bands from hydrocarbon and carbon-carbon double bond in ethylene and vinyl acetate, (2) the carbonyl groups in carbon dioxide and (3) vinyl acetate and the combination bands from hydroxyl present in water and acetic acid. The QITTFA initial estimates are comparable to the reference data available for ethylene, water and acetic acid. The carbon dioxide spectrum contained the correct functional groups apart from asymmetric stretching of $=CH_2$ (6190-6110 cm^{-1}). No reference data was available for vinyl acetate. The correct functional group, i.e. the first overtones and combination bands from CH groups bending from vinyl acetate are present in the predicted vinyl acetate spectrum.

Qualitative SMCR Analysis

The ALS spectral profiles of the reaction constituents ethylene, water, acetic acid, vinyl acetate and carbon dioxide calculated using the initial QITTFA spectral estimates and the NIR spectra in the ALS procedure are given in figure 5. Each of the constituent spectral profiles contained the correct absorption bands associated with the functional groups. The relative concentration profiles determined from the ALS model

with the five separate quantitative SMCR models (for ethylene, carbon dioxide, water, acetic acid and vinyl acetate monomer) were compared to the five separate PLS models. The results are discussed below.

Quantitative SMCR Analysis

Ethylene

Training Samples

The calibration model for ethylene determined using the correlation constraint in the ALS optimisation was compared with the calibration model determined using the PLS1 model, see the regression statistics in table 2. The application of correlation constraints in the ALS regression resulted in perfect predictive modelling, i.e. no error was found in the correlation calibration model. This result is not representative of the predictive capabilities of the correlation constraint, since calibration modelling is usually a question of local approximation to some unknown, more or less non-linear function, therefore all calibration models must be expected to contain some degree of model error. However, since the correlation calibration estimates were generated through a local linear model to approximate the regression coefficients, perfect predictive abilities were gained because the algorithm converged just after the $\mathbf{c}_{ALS}$ vector was updated with the calibration samples from the concentration values from **k** and **p**. The prediction of the ethylene test samples was not effected because they were determined from the model.

Test Samples

The ethylene results obtained for the qualitative SMCR analysis of the NIR data using ALS without the correlation constraint were compared to the quantitative SMCR analysis using the correlation constraint, see tables 3 and 4 for the regression statistics

and *p*-values of the two-tailed randomised t-test respectively. The first difference observed between the two methods was the increased LOF in the constrained solution. This was attributed to the addition of reference concentration information in the constrained ALS procedure. The slight reduction in the model fit quality suggest that active constraint were present in the solution. Although the addition of the correlation constraint reduced the RE to 9.06%, the solution with no constraints was less biased and therefore more accurate and more precise (RMSEP = 0.15, SEP = 0.15 and BIAS = -0.01) than the constrained solution (RMSEP = 0.07, SEP = 0.07 and Bias = 0.03).

The ethylene quantitative results obtained from the correlation constraint was compared to the PLS1 model determined for ethylene. The validation samples predicted using the correlation constraint were more precise, but less accurate than the PLS1 solution, due to the greater bias in the correlation constraint model. Only 3 latent variables could be used in the PLS calibration to predict ethylene concentration. Incorporating more latent variables in the PLS1 model increased the error of prediction. The two-tailed randomised t-test was revealed that there were no significant differences between the concentration profiles predicted using the correlation constraint and the PLS1 model given in table 3. The quantitative model determined for ethylene using the correlation constraint was acceptable in the pre-defined range for this component. In this example the correlation constraint was comparable to the PLS1 model.

Water

Training Samples

The MCR-ALS resolution of the concentration profiles for water using the correlation constraint converged in 9 iterations, which indicated that the initial estimate was further away from the optimum solution under the specified convergence criteria. The quantitative correlation results were compared with the quantitative PLS1 results.

The results indicated that the quantitative SMCR model and the PLS1 calibration model were both accurate. Quantitative SMCR model; (RMSEP = 0.09 and SEP = 0.09 and BIAS = 0.00) and PLS1 model; (RMSEP = 0.06 and SEP = 0.06 and BIAS = 0.00). However, the PLS1 model was slightly more precise, see SEP in table 2. Thus the PLS1 calibration had slightly better predictive properties than the correlation constraint. The correlation coefficients obtained from both models indicated that there was a strong positive correlation (correlation coefficient: 0.968; quantitative SMCR and 0.986; PLS1) between the reference and predicted samples.

Test Samples

The qualitative model for water determined from the MCR-ALS application with no constraints was better than the PLS1 model and correlation constraints model, in terms of the accuracy and precision of the concentration profiles, RMSEP, SEP and Bias are shown in table 2. The results indicated that it was possible to qualify the constituent with *no priori* information as the model did not deviate exceptionally from the conditions of resolution and the initial estimate closely represented the true profile. The predictive abilities of the quantitative PLS and correlation constraint model, were comparable. The quantitative SMCR model was less accurate, but was more precise than the PLS1 model. The two-tailed randomised t-test was revealed that there were no significant differences between the concentration profiles predicted using the correlation constraint and the PLS1 model, this is shown in table 3. The quantitative model determined for water using the correlation constraint was acceptable in the pre-defined range for this component.

Acetic acid

Training Samples

The ALS resolved profiles converged after 12 iterations, which suggested that the initial estimates were further away from the optimal resolved profiles under the

specified convergence criteria. The results indicated that the quantitative SMCR model and the PLS1 calibration model were both accurate. Quantitative SMCR model; (RMSEP = 0.09 and SEP = 0.09 and BIAS = 0.00) and PLS1 model; (RMSEP = 0.05 and SEP = 0.05 and BIAS = 0.00), but the PLS1 model was slightly more precise, see SEP in table 2.

Test Samples

The samples predicted using no constraints were not as accurate as the constrained solution and the PLS model. The addition of the correlation constraint reduced the relative error by ~16%. Comparing the PLS1 model to the quantitative SMCR model showed that the predictive ability, i.e. the accuracy and precision of both models were similar. This suggest that both models determined are very close to the best model that could be determined for this constituent. It is proposed that the prediction error may be due to the molecular interactions of acetic acid with water and dimerisation, however further investigation needs to be completed. The two-tailed randomised t-test was revealed that there were no significant differences between the concentration profiles predicted using the correlation constraint and the PLS1 model, see table 3. The quantitative model determined for acetic acid using the correlation constraint was an acceptable calibration model in the pre-defined range for this component.

Vinyl Acetate

Training Samples

The predictive property of the correlation constraint was compared to the PLS1 model. The MCR-ALS solution with the correlation constraint converged in 13 iterations, which suggest initial estimates were slightly further away from the optimum

ALS solution, this may be due to slight scaling ambiguity in the spectral resolution. The precision of the PLS1 model was better than the correlation constraint, although both methods were accurate.

Test Samples

The qualitative determination of the vinyl acetate concentration profiles with no constraint were more precise than the correlation constraint and PLS1 model, but was less accurate due to the greater bias. The addition of the correlation constraint in the MCR-ALS resolution reduced the relative error by 3.3%. Generally the predictive ability of the correlation constraint was slightly lower than the PLS1 model. Although the two-tailed randomised t-test revealed that there were no significant differences between the concentration profiles predicted using the correlation constraint and the PLS1 model. The quantitative model determined for vinyl acetate using the correlation constraint was not an acceptable calibration model in the pre-defined range for this component, see tables 2-3.

Carbon dioxide

Training Samples

The calibration model determined for carbon dioxide using the correlation constraint converged in 12 iterations. The correlation constraint and PLS1 model were both accurate (Bias = 0.00 for both models), but the PLS1 model was a little more precise, see the SEP in table 2.

Test Samples

The relative error of prediction for the resolution in which no constraints were applied was low, RE 41.7%. The addition of the correlation constraint improved the prediction by ~10%. Comparing the regression statistic for the correlation constraint

model with the PLS1 model, it was possible to observe a difference in the accuracy of prediction between the two methods. The PLS1 model was less bias and hence more accurate than the correlation constraint, see RMSEP, SEP and bias in table 2. The two-tailed randomised t-test was revealed that there were no significant differences between the concentration profiles predicted using the correlation constraint and the PLS1 model. The quantitative model determined for carbon dioxide using the correlation constraint was an acceptable calibration model in the pre-defined range for this component.

CONCLUSION

The quantitative SMCR procedure could be used to quantify ethylene 9.06% (RE), acetic acid 19.30% (RE), water 13.77% (RE) and carbon dioxide 30.46% (RE) within the defined relative error margin. The advantages of the new approach were a ~90% reduction in the calibration time, ~90% reduction in the number of training samples required for the calibration and the simultaneous recovery of the reaction constituent spectral profiles. Therefore this quantitative SMCR strategy could be used for reactions or processes for which it is not possible to prepare mixtures of known composition due to the absence of isolated reference material, stability issues and where the preparation of such samples are time consuming and expensive. The %RE of the vinyl acetate model was outside the acceptable range for use on the plant. Therefore, further investigations need to be completed to determine how this model could be improved. Nevertheless, the overall economic gain of using the quantitative SMCR strategy more than adequately compensates any further refinement to the vinyl acetate model. The resolution and identification of by-products and intermediates from the vinyl acetate process in the absence of selectivity and *a priori* information using the

QITTFA spectral estimates provides a useful tool to study complex, information deficient processes and reactions to enable efficient monitoring and control.

ACKNOWLEDGEMENTS

Alasdair Thomson from BP is acknowledged for giving access to the NIR data. Funding from CPACT and EPSRC/DTI link program are appreciated.

REFERENCES

1. MCLENNAN, F. & KOWALSKI, B. R. (1995) *Process Analytical Chemistry* (London, Blackie Academic & Professional).
2. CHALMERS, J. M. (2000) *Spectroscopy in Process Analysis* (Sheffield Academic Press).
3. NAES, T., ISAKSSON, T., FEARN, T. & DAVIES, T. (2002) *A User-Friendly Guide to Multivaraite Calibration and Classification* (NIR Publications).
4. BEEBE, K. R. & KOWALSKI, B. R. (1987) An Introduction to Multivariate Calibration and Analysis, *Analytical Chemistry*, 59, 1007A-1017A.
5. THOMAS, E. V. (1994) A Primer in Multivariate Calibration, *Analytical Chemistry*, 66, 795A-804A.
6. HAALAND, D. M. & THOMAS, E. V. (1988) Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information, *Analytical Chemistry*, 60, 1193-1202.
7. HAALAND, D. M. & THOMAS, E. V. (1988) Partial Least-Squares for Spectral Analyses. 2. Application to Simulated and Glass Spectral Data, *Analytical Chemistry*, 60, 1202-1208.
8. GEMPERLINE, P. J. (1984) A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis, *Journal of Informatics and Computer Science*, 24, 206-212.
9. LAWTON, W. H. & SYLVESTRE, E. A. (1971) Self Modeling Curve Resolution, *Technometrics*, 13, 617-633.
10. TAULER, R., SANCHEZ, F. C. & MASSART, D. L. (1996) Validation of Alternating Least Squares Multivariate Curve Resolution for Chromatographic Resolution and Quantitation, *Trends in Analytical Chemistry*, 15, 279-286.
11. BYLUND, D., DANIELSSON, R. & MARKIDES, K. E. (2001) Peak Purity Assessment in Liquid Chromatography-Mass Spectrometry, *Journal of Chromatography A*, 915, 43-52.
12. MA, B., GEMPERLINE, P. J., CASH, E., BOSSERMAN, M. & COMAS, E. (2003) Characterizing batch reactions with in situ spectroscopic measurements, calorimetry and dynamic modeling, *Journal of Chemometrics*, 17, 470-479.
13. GEMPERLINE, P. J., ZHU, M., CASH, E. & WALKER, D. S. (1999) Chemometric Characterisation of Batch Reactions, *ISA Transactions*, 38, 211-216.
14. RICHARDS, S., ROPIC, M., BLACKMOND, D. & WALMSLEY, A. (2004) Quantitative Determination of the Catalysed Asymmetric Transfer Hydrogenation of 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline using In-situ FTIR and Multivariate Curve Resolution., *Analytica Chimica Acta*, 519, 1-9.
15. JUAN, A. D. & TAULER, R. (2003) Chemometrics Applied to Unravel Multicomponent Processes and Mixtures, *Analytica Chimica Acta*, 500, 195-210.
16. JIANG, J.-H., LIANG, Y. & OZAKI, Y. (2004) Principles and Methodologies in Self Modeling Curve Resolution, *Chemometrics and Intelligent Laboratory Systems*, 71, 1-12.
17. ANTUNES, M. C., SIMÃO, J. E. J., DUARTE, A. C. & TAULER, R. (2002) Multivariate Curve Resolution of Overlapping Voltammetric Peaks: Quantitative Analysis of Binary and Quaternary Metal Mixtures, *Analyst*, 127, 809-817.
18. RICHARDS, S. & WALMSLEY, A. D. (2007) Quantitative Iterative Target Transformation Factor Analysis, *Journal of Chemometrics*, 22, 63-80.
19. WINDIG, W. & GUILMENT, J. (1991) Interactive Self Modelling Mixture Analysis, *Analytical Chemistry*, 63, 1425-1432.

20. WINDIG, W. (1994) The Use of Second Derivative Spectra for Pure-Variable Based Self-Modeling Mixture Analysis Techniques, *Chemometrics and Intelligent Laboratory Systems*, 23, (71-86).
21. WINDIG, W. (1997) Spectral Profiles for Self-Modelling Curve Resolution with Examples using the SIMPLISMA Approach, *Chemometrics and Intelligent Laboratory Systems*, 36, (3-16).
22. WINDIG, W. & HECKLER, C. E. (1992) Self-Modelling Mixture Analysis of Categorized Pyrolysis Mass Spectral Data with the SIMPLISMA Approach, *Chemometrics and Intelligent Laboratory Systems*, 14, (195-207).
23. WINDIG, W. & STEPHENSON, D. A. (1992) Self-Modelling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data Using the SIMPLISMA Approach, *Analytical Chemistry*, 64, (2735-2742).
24. BRO, R. & DE JONG, S. (1997) A Fast Non-Negative Constrained Least Squares Algorithm, *Journal of Chemometrics*, 11, 393-401.
25. TAULER, R., SMILDE, A. K. & KOWALSKI, B. R. (1995) Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *Journal of Chemometrics*, 9, 31-58.
26. JUAN, A. D., HEYDEN, Y. V., TAULER, R. & MASSART, D. L. (1997) Assessment of New Constraints Applied to the Alternating Least Squares Method, *Analytica Chimica Acta*, 346, 307-318.
27. VOET, H. V. D. (1994) Comparing the Predictive Accuracy of Models using a Simple Randomisation Test, *Chemometrics and Intelligent Laboratory Systems*, 25, 313-323.

FIGURE and TABLE LABELS

Figure 1 Quantitative SMCR procedure for the simultaneous qualitative and quantitative recovery of reaction constituents from the vinyl acetate monomer process.

Figure 2 Implementation of the Correlation constraint. **1. Initialisation** - The $\mathbf{C}_{sel}$ matrix contains sparse known concentration values for the target analyte (black rectangles). **2. Least squares** – The starting estimate are regressed against the data matrix to estimate the $\mathbf{C}_{ALS}$ concentration profiles for each constituent. **3. Correlation Constraint – Step A**: The **k** and **x** vectors are regressed using the classical least squares. **Step B**: The model determined in step one is used to predict **p** for the unknown values, **u**. **Step C**: The $\mathbf{C}_{ALS}$ matrix is updated with **p** and the known values **k**. Step four: $\mathbf{C}_{ALS}$ is used in the least squares estimate of **S**.

Figure 3 The NIR spectra of the five component gas mixture. Regions of interest include the first overtone of the OH group associated with monomer acetic acid is apparent at ~6994 cm^{-1}. The first overtone of the asymmetric C-H stretch from ethylene appears at 6149 cm^{-1}, showing a PQR structure covering 6200-6090 cm^{-1} and the first overtone of the symmetric C-H stretch showing PR structure at 5993 cm^{-1} together with the combination bands with relatively strong absorbance appearing at 4760 cm^{-1}. The combination bands from the asymmetric and symmetric stretching modes of the water molecule

appear at ~5150 and ~6900cm^{-1}.

Figure 4 The initial spectral estimates of the reaction constituents ethylene, water, acetic acid, vinyl acetate and carbon dioxide calculated using the QITTFA routine and the NIR spectra, shown in Figure 4, labelled A-E respectively. The NIR spectral profiles of the neat vaporised samples of the reaction constituents shown in Figure 4, labelled F-J respectively.

Figure 5 The MCR-ALS spectral profiles of the reaction constituents ethylene, water, acetic acid, vinyl acetate and carbon dioxide calculated using the initial QITTFA spectral estimates and the NIR spectra in the ALS procedure.

Table 1 The original BP calibration model reference error (%RE) for the prediction of the reaction constituents.

Table 2 The regression statistics for the qualitative SMCR models, quantitative SMCR models and the PLS1 models quoted for each of the VAM reaction constituents

Table 3 Two sided tests, 199 iterations in each randomised t-test. *p* values are for comparison with PLS1 model for each constituent

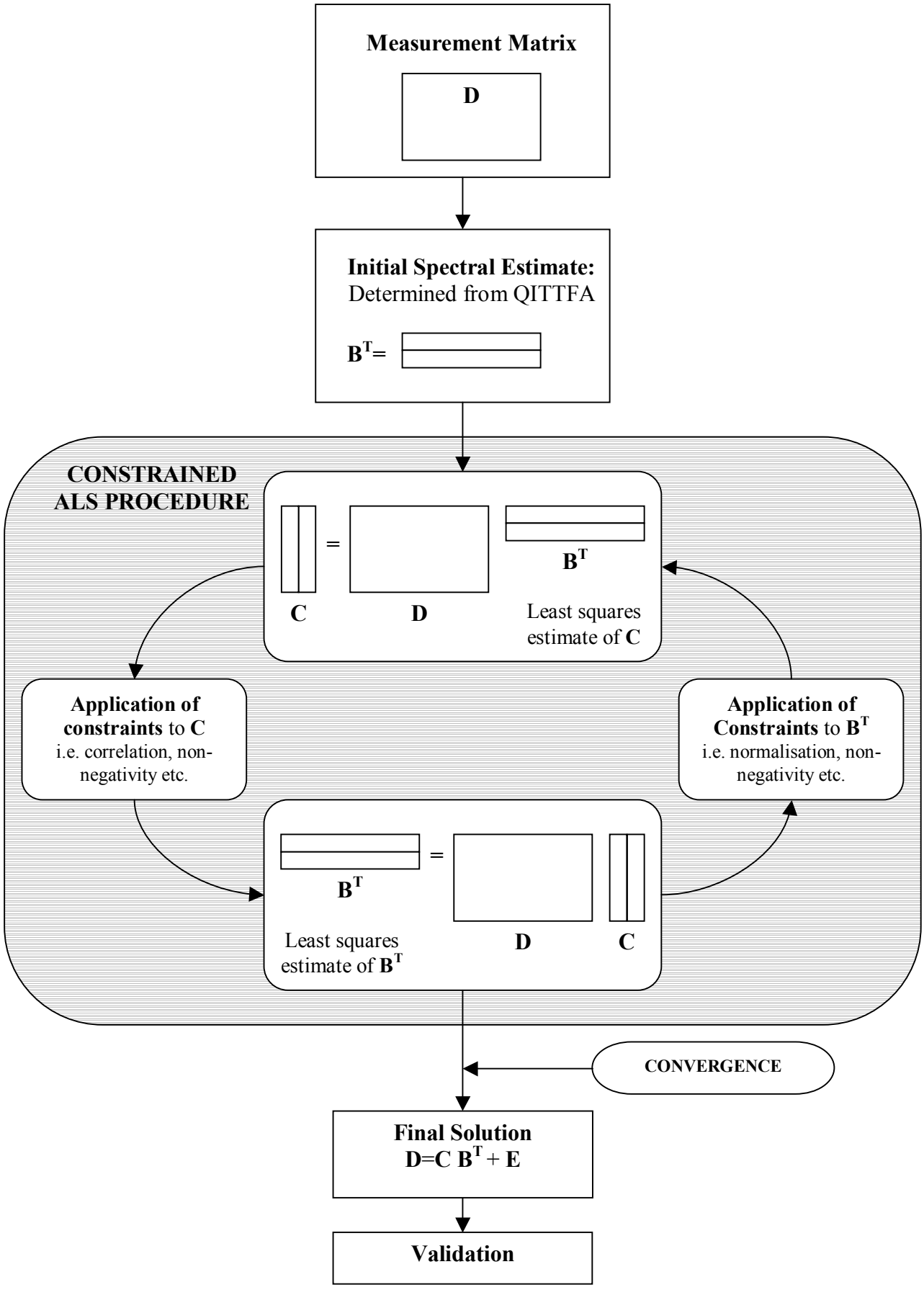
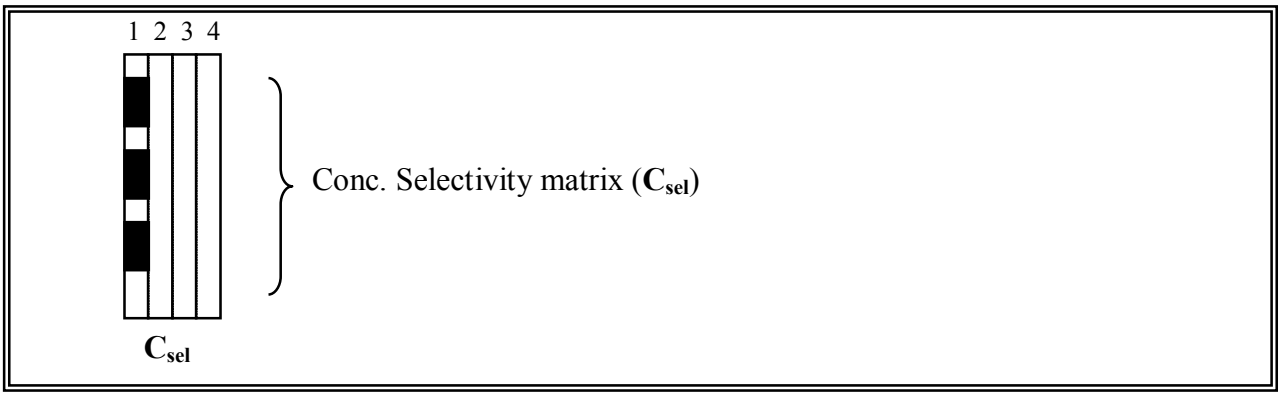
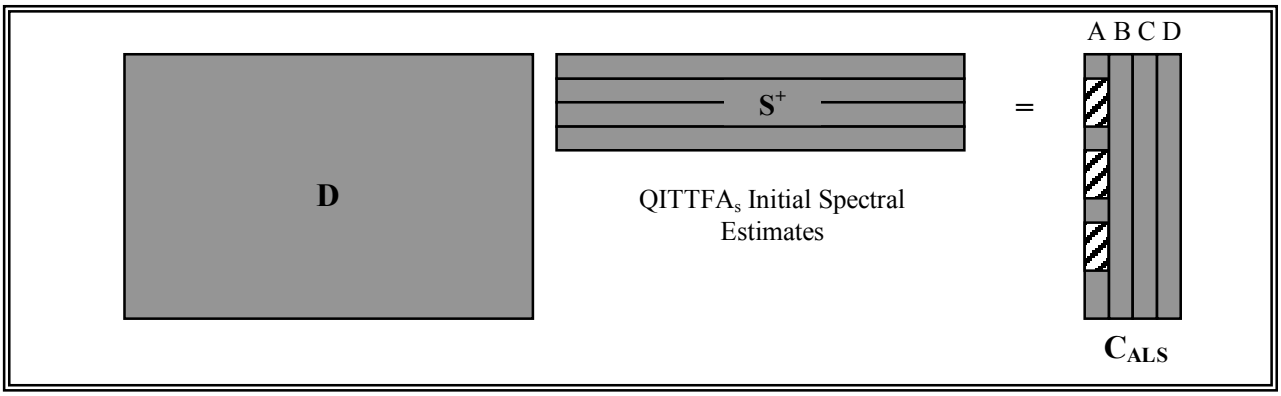


Figure 1. Quantitative SMCR procedure for the simultaneous qualitative and quantitative recovery of reaction constituents from the vinyl acetate monomer process

1. Initialisation:



2. Least Squares:



3. Correlation Constraint:

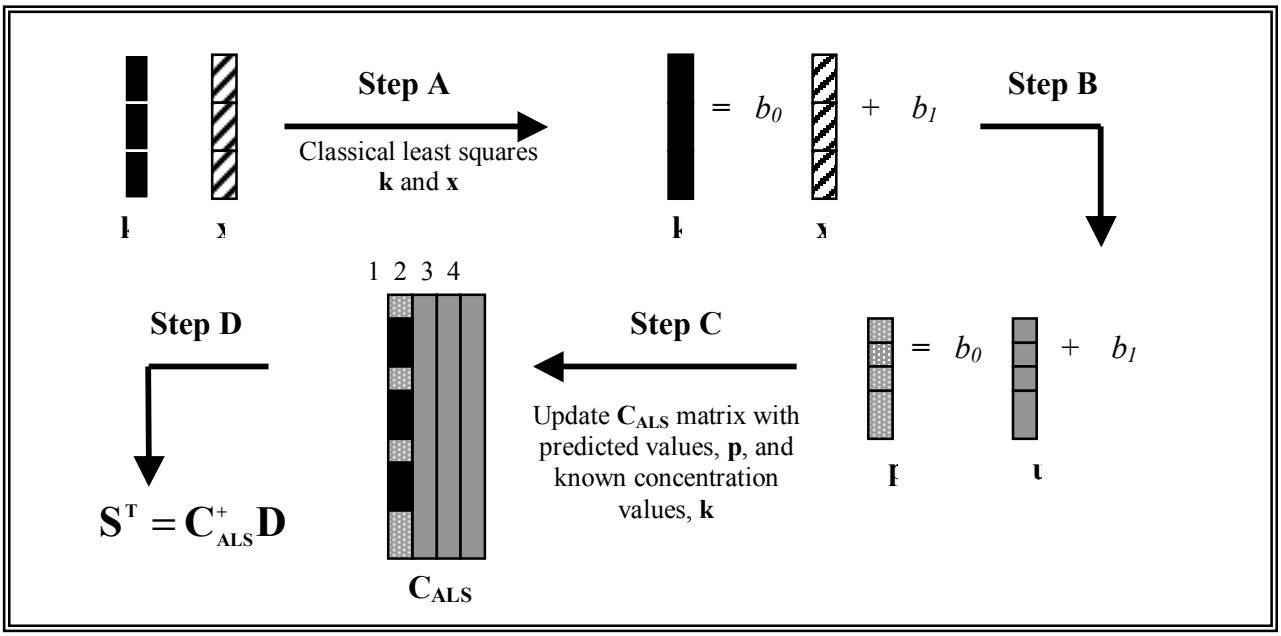


Figure 2. Implementation of the Correlation constraint. **1. Initialisation** - The $\mathbf{C_{sel}}$ matrix contains sparse known concentration values for the target analyte (black rectangles). **2. Least squares** – The starting estimate are regressed against the data matrix to estimate the $\mathbf{C_{ALS}}$ concentration profiles for each constituent. **3. Correlation Constraint** – **Step A**: The **k** and **x** vectors are regressed using the classical least squares. **Step B**: The model determined in step one is used to predict **p** for the unknown values, **u**.

Step C: The $\mathbf{C}_{\mathbf{ALS}}$ matrix is updated with **p** and the known values **k**. Step four: $\mathbf{C}_{\mathbf{ALS}}$ is used in the least squares estimate of **S**.

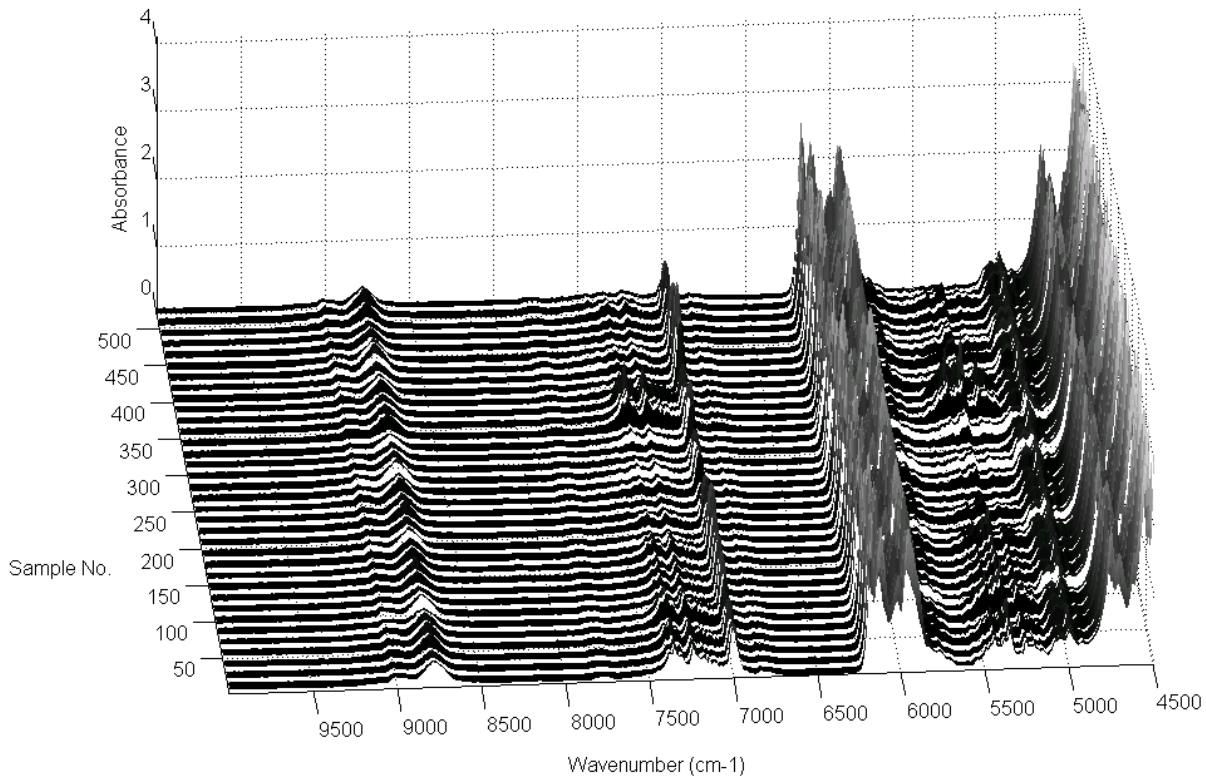


Figure 3. The NIR spectra of the five component gas mixture. Regions of interest include the first overtone of the OH group associated with monomer acetic acid is apparent at ~6994 cm^{-1}. The first overtone of the asymmetric C-H stretch from ethylene appears at 6149 cm^{-1}, showing a PQR structure covering 6200-6090 cm^{-1} and the first overtone of the symmetric C-H stretch showing PR structure at 5993 cm^{-1} together with the combination bands with relatively strong absorbance appearing at 4760 cm^{-1}. The combination bands from the asymmetric and symmetric stretching modes of the water molecule appear at ~5150 and ~6900cm^{-1}.

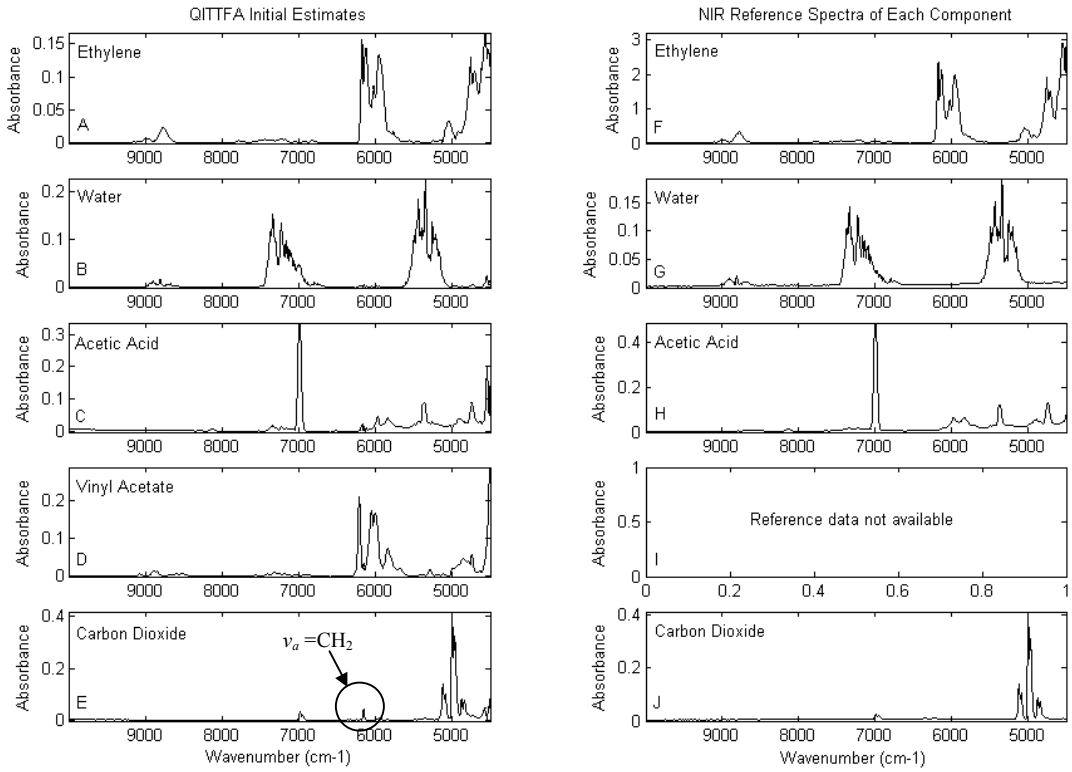


Figure 4. The initial spectral estimates of the reaction constituents ethylene, water, acetic acid, vinyl acetate and carbon dioxide calculated using the QITTFA routine and the NIR spectra, shown in Figure 4, labelled A-E respectively. The NIR spectral profiles of the neat vaporised samples of the reaction constituents shown in Figure 4, labelled F-J respectively.

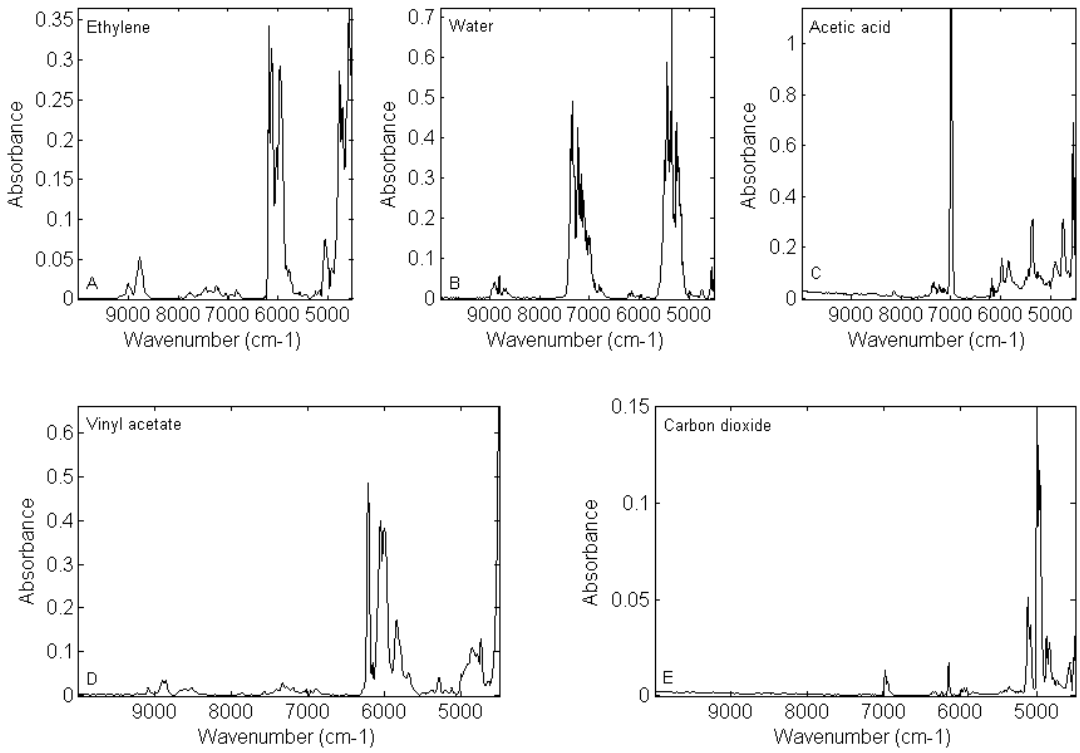


Figure 5. The MCR-ALS spectral profiles of the reaction constituents ethylene, water, acetic acid, vinyl acetate and carbon dioxide calculated using the initial QITTFA spectral estimates and the NIR spectra in the ALS procedure.

	Ethylene	Water	Acetic Acid	Vinyl Acetate	Carbon dioxide
% RE	10 ± 2	15 ± 2	20 ± 2	15 ± 2	30 ± 2

Table 1. The original BP calibration model reference error (%RE) for the prediction of the reaction constituents.

	Method	N.o.C	LOF	Nit	RMSEP	SEP	Bias	RE%	Slope	Offset	Cor
Ethylene											
Training	Correlation	6	4.40	3	0.00	0.00	0.00	0.00	1.000	0.00	1.000
	PLS	3	N.A	N.A	0.76	0.77	-0.03	12.75	0.613	2.27	0.842
Test	No Constraint	6	1.57	3	0.15	0.15	-0.01	29.46	0.857	0.07	0.783
	Correlation	6	4.40	3	0.53	0.53	0.08	9.06	0.748	1.37	0.863
	PLS	3	NA	N.A	0.59	0.59	-0.04	10.11	0.652	2.04	0.822
Water											
Training	Correlation	6	1.57	9	0.09	0.09	0.00	17.53	0.936	0.02	0.968
	PLS	4	N.A	N.A	0.06	0.06	0.00	11.43	0.974	0.01	0.986
Test	No Constraint	6	1.57	3	0.05	0.05	0.00	14.63	0.896	0.02	0.982
	Correlation	6	1.57	9	0.06	0.06	0.03	13.77	0.944	-0.01	0.983
	PLS	4	N.A	N.A	0.07	0.07	0.02	16.16	0.947	0.00	0.973
Acetic Acid											
Training	Correlation	6	1.57	12	0.09	0.09	0.00	14.78	0.936	0.03	0.967
	PLS	6	N.A	N.A	0.05	0.05	0.00	8.28	0.977	0.01	0.990
Test	No Constraint	6	1.57	3	0.11	0.08	-0.08	35.03	1.028	0.07	0.912
	Correlation	6	1.57	12	0.09	0.09	-0.01	19.30	0.976	0.02	0.923
	PLS	6	N.A	N.A	0.09	0.09	0.01	18.67	0.907	0.03	0.920
Vinyl Acetate											
Training	Correlation	6	1.57	13	0.11	0.12	0.00	20.11	0.910	0.04	0.954
	PLS	5	N.A	N.A	0.03	0.03	0.00	6.02	0.994	0.00	0.996
Test	No Constraint	6	1.57	3	0.09	0.08	-0.04	26.95	0.816	0.09	0.926
	Correlation	6	1.57	13	0.11	0.11	0.02	23.67	0.848	0.04	0.915
	PLS	5	N.A	N.A	0.07	0.07	0.00	15.72	0.988	0.00	0.964
Carbon dioxide											
Training	Correlation	6	1.57	12	0.27	0.28	0.00	19.09	0.885	0.14	0.941
	PLS	4	N.A	N.A	0.15	0.16	0.00	10.73	0.964	0.04	0.982
Test	No Constraint	6	1.57	3	0.17	0.14	-0.08	41.17	0.764	0.16	0.820
	Correlation	6	1.57	12	0.38	0.37	-0.10	30.46	0.810	0.30	0.831
	PLS	4	N.A	N.A	0.36	0.36	-0.04	28.23	0.910	0.14	0.860

Table 2. The regression statistics for the qualitative SMCR models, quantitative SMCR models and the PLS1 models quoted for each of the VAM reaction constituents

Randomised T-test	Ethylene	Water	Acetic acid	Vinyl acetate	Carbon dioxide
p values	0.075	0.025	0.665	0.005	0.140

Table 3. Two sided tests, 199 iterations in each randomised t-test. p values are for comparison with PLS1 model for each constituent