DEVELOPMENT OF 2D- AND 3D- BTEM FOR PATTERN RECOGNITION IN HIGHER-ORDER SPECTROSCOPIC AND OTHER DATA ARRAYS

GUO LIANGFENG

(B.Eng.)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY DEPARTMENT OF CHEMICAL & BIOMOLECULAR ENGINEERING NATIONAL UNIVERSITY OF SINGAPORE

2006

ACKNOWLEDGEMENT

I am forever grateful to my supervisor, Prof. Marc Garland, who has patiently provided me with invaluable guidance and great encouragement in all the areas related to my research. His passionate, vital ideas and assistance has inspired me throughout my graduate studies. I sincerely thank him for the support and concern that he has given throughout my research work.

I also extend my thanks to the staff in the Chemical & Environmental Department for their help in this project. I wish to thank my colleagues for their generous help and invaluable comment. Especially, I am deeply indebted to them who greatly help me in my fulfilments of my research. I would like to thank Dr Chen Li, Dr. Effendi Widjaja, Dr. Chew Wee, Dr. Li Chuanzhao, Mr. Zhang huajun, Mr. Ayman Daoud Allian, Mr. Karl Irwin Krummel, Mr. Martin Tjahjono, Ms. Gao Feng, Ms. Zhao Yangjun and Ms. Cheng Shuying. I would also like to give my special gratitude to Dr. Effendi Widjaja, for sharing his time, knowledge. Also I would like to thank the administrative staff in our department-especially Mr. Boey, Mr. Mao Ning, Ms. Jamie and many others.

I would like to thank Peter Sprenger (Bruker Biospin) for his collaboration in the NMR studies in Singapore and at Bruker Biospin AG in Zurich, Switzerland. I am also grateful to Dr. Fethi Kooli, Dr. Anette Wiesmat and many others in the Institute of Chemical and Engineering Sciences (ICES in Singapore) for their collaboration. Thanks would be given to Prof. Stanford who provided the samples for my Power XRD study. My research has been made possible only with their invaluable contributions.

My family has been significantly supporting me all these years. I am indebted to my parents for their love and care, and to my wife, in particular, for her constant support and encouragement throughout my research work. The support and encouragement from my good friends are also gratefully acknowledged.

Finally, I am grateful for the scholarship and resources that the National University of Singapore (NUS) had provided during my study.

TABLE OF CONTENTS

SUMMARY

Both pure component spectral reconstruction from spectroscopic data arrays and chemical system identification are important steps in exploratory chemometric studies. Various methods and techniques have been reported in the literature. In recent years, the use of simultaneous multiple 1D spectroscopies as well as higher order spectroscopies i.e. 2D and 3D data, has become quite common in the chemical sciences. The resulting data is often very complex and the size of the data arrays can be huge. Very few if any feasible algorithms/methods have been devised for treating very large scale spectroscopic data arrays, *particularly for* recovering pure component spectra without the use of any *a priori* information. In this thesis a model-free spectral reconstruction method for large scale and particularly higher dimensional data sets is developed. A variation on the concept of entropy minimization is used to deconvolute the signals.

As a starting point for the present studies, the 1D-BTEM algorithm[i] was extended, and with some modification, it was successfully applied for the first time, to sets of acoustic data and solid state powder X-ray diffraction data. After further modifications, it was applied to non-reactive and reactive ^{1}H-^{13}C-^{19}F-^{31}P 1D NMR spectroscopic data.

Subsequently, a higher dimensional entropy minimization method based on the BTEM and related techniques were developed for very large scale arrays. Starting from computer simulated experiments, the algorithms were tested. Then they were successfully applied to various sets of 2D images, both black and white, as well as color. They were then successfully

[i] Widjaja, E.; Li, C.; Garland, M. Organometallics, 2002, 21, 1991-1997.

implemented on 2D spectroscopic data, in particular, 2D NMR spectroscopic data (COSY and HSQC) and 2D fluorescence spectral data sets. The performance of these proposed novel methods, both with simulated and real experimental mixture spectral data is very good. The pure component images/spectra were recovered from mixture data with very little *a priori* information what-so-ever. This means there was no assumption made about the number of pattern present, nor the characteristics of the patterns. Also the relative concentrations of the constituents were obtained. The ideas for 2D entropy minimization were successfully extended to 3D, and 3D patterns were extracted.

Starting from the known concept of 1D target transformation for pattern analysis, the concepts of 2D and 3D target transformation are introduced. The mathematical procedures needed are developed.

The present developments represent a significant step forward for very complex blind source separation problems (inverse problems with multiple sources). The ability to obtain accurate deconvolution with no assumptions what-so-ever, opens many possibilities. Indeed, a vast range of different types of 2D spectroscopic mixture data and 3D spectroscopic mixture data can now be analyzed in the future. Also, the present development promotes system identification in the chemical sciences (both non-reactive and reactive systems), and sets detailed in-situ spectroscopic studies of reactive systems on a much more firm basis. This will certainly lead to more accurate mechanistic and kinetic models.

NOMENCLATURE

Abbreviations

ca.	circa (approximately)
*corr*2	2D correlation coefficient between two matrices
ALS	Alternating Least Square
AR	Alternating Regression
BSS	Blind Source Separation
BTEM	Band-Target Entropy Minimization
CANDECOMP	CANonical DECOMPosition
COT	cyclooctene
COW	Correlation Optimized Warping
COSY	^{1}H-^{1}H Correlation Spectroscopy
DECRA	Direct Exponential Curve Resolution Algorithm
DMC-SMCR	Dynamic Monte Carlo SMCR
DTLD	Direct TriLinear Decomposition
DTW	Dynamic Time Warping
EA	Evolutionary Algorithm
EEM	Emission/Excitation matrix
EFA	Evolving Factor Analysis
EPR	Electron Paramagnetic Resonance
Eq.	Equation
FID	Free Induction Decay
FT	Fourier Transform
FTIR	Fourier Transform Infra Red
GA	Genetic Algorithm
GC	Gas Chromatography
GRAM	Generalized Rank Annihilation Method
HELP	Heuristics Evolving Latent Projections
HMBC	Heteronuclear Multiple Bond Correlation
HMQC	Heteronuclear Multiple Quantum Correlation
HPLC	High Performance Liquid Chromatography

HSQC	Heteronuclear Single Quantum Correlation
ICA	Independent Component Analysis
ICES	Institute of Chemical and Engineering Sciences
INADEQUATE	Incredible natural abundance double quantum transfer experiment
IPCA	Interactive Principal Component Analysis
IR	Infra Red
ITTFA	Iterative Target-Testing Factor Analysis
KSFA	Key Set Factor Analysis
LBBL	Lambert-Beer-Bouguer-Law
LC	Liquid Chromatography
LC-DA-UV	Liquid Chromatography Diode Array-UV
LC-DAD	Liquid Chromatography – Diode Array Data
MCR	Multivariate Curve Resolution
MEM	Maximum Entropy Method
MESS	Minimization of Entropy with Spectral dis-Similarity
MS	Mass Spectrometry
MS-MS	Tandem mass spectrometers
NIPALS	Non-Linear Iterative Partial Least-Square
NIR	Near Infra Red
NLP	NonLinear Program
NMF	Non-negative Matrix Factorization
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect spectroscopy
OPA	Orthogonal Projection Approach
PAGA	Peak Alignment by a Genetic Algorithm
PAH	Polycyclic Aromatic Hydrocarbon
PARAFA	PARAllel RActor analysis
PC	Principal Component
PCA	Principal Component Analysis
PGSE	Pulsed Gradient Spin Echo
PLF	Partial Linear Fit
PMF	Positive Matrix Factorization

RAFA	Rank Annihilation Factor Analysis
RGB	Red, Green and Blue
ROESY	Rotational Nuclear Overhauser Effect spectroscopy
SA	Simulated Annealing
SIMCA	Soft Independent Modeling of Class Analogy
SIMPLISMA	Simple-to-use interactive self-modelling mixture analysis
SMCR	Self-Modelling Curve Resolution
SVD	Singular Value Decomposition
SVD-SM	Singular Value Decomposition with Self-Modeling Method
TOCSY	Total correlation Spectroscopy
TOF-SIMS	Time-of-flight secondary ion mass spectrum
TTFA	Target Transformation Factor Analysis
UV	Ultraviolet
VCD	Vibrational Circular Dichroism
VIS	Visible
WFA	Window Factor Analysis
XRD	X-ray Diffraction

Symbols

$a_{s\times\nu}$	pure component spectral estimates for *s* chemical species
$\hat{a}_{1\times\nu}$	estimated 1D pure spectrum with ν channels of wavenumbers
$\hat{a}_{s\times\nu}$	estimated pure component spectral estimates for s chemical species
$\hat{a}_{m\times n}$	estimated 2D spectrum with *m* rows and *n* columns
$A_{q\times\nu}$	spectroscopic data matrix with *q* spectra and ν channels of wavenumbers
$\underline{A}_{q\times m\times n}$	a 3-way array composed of *q* mixture spectra (each with size of *m*-by-*n*)
C_i	loading for component *i* result from PCA decomposition
C_{qxs}	concentration matrix for s species in *q* samples
$\hat{C}_{q\times 1}$	estimated concentrations profile for one species in *q* samples
$\hat{C}_{qxs}$	estimated concentration matrix for *s* species in *q* samples
E	error and noise term
F_i	matrix-formatted component *i* result from PCA decomposition
F_{obj}	objective function value
H	information entropy
H^{2D}	entropy of 2D spectrum
I	unit matrix
N_i	moles of each species in one sample
P	penalty function
Q	emission/excitation matrix of fluorescence spectrum
$\underline{Q}$	a 3-way array composed of a series of fluorescence EEM spectra
R	rotational matrix
U	matrix of left singular vectors
V^T	transposed matrix of right singular vectors
X	unknown array
Y	unknown array

Greek Letters

ε	experimental error
γ_a	penalty coefficient to ensure positivity of spectral estimate
γ_c	penalty coefficient to ensure positivity of concentration
Σ	diagonal singular values matrix
Δa	difference between the test spectrum a and the estimated vector $\hat{a}$ in TTFA

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

There are countless problems encountered in science, in which there are imbedded patterns in the observed data set, but the experimentalist does not know how many patterns there are nor what the patterns may look like. In the pure and applied mathematics literature these are often referred to as inverse problems (Sabatier, 1978). In the electrical engineering literature, they are often referred to as blind-source separation problems (Jutten and Hérault, 1991; Cardoso, 1997). In the chemical sciences the term spectral deconvolution is often used (Brown *et al*., 1996).

Finding a proper model that describes significant dependencies between variables is an essential first step to untangle the data. Superpositions of patterns result when m individual sources are instantaneously mixed, contaminated with noise E, and the n resulting superpositions are observed. A simple formulation is given below:

$$Y = f(X) + E \tag{1.1}$$

where X denotes a series of sources. For example, these sources could be time series signals $X(t)$ such as acoustics in electrical engineering, or these sources could be electromagnetic spectra of constituents in the chemical sciences. In the latter case, the source represents a series of intensity measurements along the wavelength or frequency $X(v)$. Here, f denotes an unknown function which maps the m dimensions of sources to n dimensions of observations. The really interesting, intricate, and difficult work is to invert

the experimental observations *Y* and recover both the function *f* and all the sources *X* as precisely as possible – preferably with no *a priori* information about the system.

In the modern chemistry laboratory, large observation/data sets can be routinely obtained from sophisticated analytical instruments (particularly spectrometers), manipulated and stored. The common bottleneck in the chemical sciences today is the full analysis and utilization of the spectroscopic data.

Chemometrics, a relatively new and separate branch of chemistry, is a data analysis methodology with the application of mathematical, statistical and logical methods to elucidate the concealed information embedded inside the observable data set (Wold, 1995). The revealed information commonly forms the basis for new understanding of the studied system for the chemist or chemical engineer.

If a chemist or chemical engineer has a reactive system, and has appropriate analytical instrumentation, there are some basic questions that can be asked in almost all cases. These include (1) how many observable species are present and what are their spectra[i] (2) how many observable reactions are present and what are the reaction stochiometries (3) what are the physico-chemical parameters associated with the observable species[ii] and (4) what are the physico-chemical parameters associated with the observable reactions[iii]? The answers for the above questions provide very detailed system identification models for the system i.e., algebraic model, thermodynamic model, kinetic model, etc. At the present moment, the most important point to note is the need to solve Part (1) at the outset. In other words, the determination of the observable species present is of primary importance. It should be clear that the solution to Part (1) is a difficult *inverse*

[i] From bulk spectroscopic measurements
[ii] Requires additional bulk density, refractive index, dielectric measurements, etc
[iii] Requires additional bulk density, bulk calorimetric measurements, etc

problem and represents a special case of Eq. 1.1, where each species has its own unique spectral pattern. A robust solution to Eq. 1.1 in order to solve Part (1) without any *a priori* information would be very important for the chemical sciences. In part, a robust solution is difficult to obtain, since spectroscopic signals are inherently non-stationary. In other words, the pure component spectra (patterns) are non-constant.[iv]

Over the past few decades, quite a lot of work has focused on spectroscopy and the reconstruction of pure component spectra from multi-component mixtures. Numerous self-modeling curve resolution methods are now available for spectroscopic data. For example, iterative target transformation factor analysis (ITTFA) (Gemperline, 1984, 1986; Vandeginste 1985), multivariate curve resolution and alternating least squares method, (MCR-ALS) (Tauler *et al*., 1991; Tauler 2001), simple to use interactive self-modeling mixture analysis (SIMPLISMA) (Windig, 1991, 1997), and heuristic evolving latent projection (HELP) (Kvalheim and Liang, 1992). Most of these methods deal with the general 1-dimensional (1D) spectroscopic data set. Recently, some methods/algorithms were extended to the analysis of large scale multi-way spectroscopic data set. A family of methods have been developed to treat such data sets where a trilinear structure is assumed: direct trilinear decomposition (TLD) (Sanchez and Kowalski, 1990), parallel factor analysis (PARAFAC) (Carroll and Chang, 1970), TUCKER3 (Tucker, 1966; Kroonenberg and de Leeuw, 1980) and also MCR-ALS (Tauler *et al*., 1998; de Juan and Tauler, 2001). In all of the above examples, some sort of *a priori* information is needed, or some sort of severe restriction in the scope of the method exists.

[iv] The term non-stationary is extensively used in the physical literature to denote signals whose mean and standard deviation change. For problems in the chemical sciences, non-stationary spectra are ubiquitous. They arise due to a convolution of physical and instrumental effects, and are known to effect electromagnetic spectra from the radio wave (Nuclear Magnetic Resonance) to X-ray diffraction.

Thesis Objective

Over the past few years, our research group has developed a very robust algorithm for treating 1D spectroscopic data (solving Eq 1.1 and Part (1)), which does not require any *a priori* information what-so-ever. *The primary objective of the present thesis is to develop and successfully test an algorithm which is applicable to higher dimensional problems, where the patterns are matrices X(v×v) or even tensors* $\underline{X}$ *(v×v×v) instead of vectors x(v).* This would considerably extend the scope of problems that can be treated in the chemical sciences. Here, it is important to note that NMR (Nuclear Magnetic Resonance) is the most important spectroscopic tool in the chemical sciences and that 2D and 3D NMR are of incredible importance for understanding structural and dynamic molecular problems.

During the course of this PhD thesis, I first worked with the groups' 1D algorithm and extended its scope. Then a new higher dimensional pattern recognition algorithm was successfully developed and tested without requiring any *a priori* information what-so-ever.

Outline of this Thesis

The organization of this thesis is summarized as follows.

Chapter 2 provides a broad review of recent and related literature pertinent to this multi-disciplinary thesis. This review covers chemometrics, self-modeling curve resolution, chemometric techniques for high dimensional data and NMR spectroscopy. A brief review of numerical optimization algorithms is also included.

Chapter 3 can be considered as an introductory tutorial to the fundamental concepts, mathematics and methodologies that will be needed and used in chemometric data analysis. Data pretreatment and data enhancement are also covered.

Chapter 4 As a starting point, this chapter is devoted to the 1D spectroscopic problem. The group's advanced spectral reconstruction algorithm named Band-Target Entropy Minimization (1D-BTEM) is introduced. I successfully applied it to solve four (4) sets of group data from different types of homogeneous catalytic hydroformylation. After some modification, it was successfully applied for the first time, to sets of acoustic data and solid state powder x-ray diffraction data. After further modifications, it was applied to non-reactive and reactive ^{1}H-^{13}C-^{19}F-^{31}P NMR data (in collaboration with Bruker AG Switzerland).

In **Chapter 5,** the theoretical and mathematical foundations of 2D-BTEM and a more general 2D EM method are developed and proposed. The necessary mathematical manipulations are described and the higher dimensional target transformation technique is discussed.

Chapter 6 applies the tools from chapter 5 to simulated 2D spectral data to make sure that the algorithm works. Then a real problem from image processing is successfully treated.

In **Chapter 7**, 2D-BTEM is further tested and applied to several real experimental systems. In particular it is applied to both COSY and HMQC NMR data sets (in collaboration with ICES and Bruker Singapore). Also another important type of 2D pattern, fluorescent excitation-emission-matrix (EEM) data is successfully treated.

Chapter 8 describes the theoretical and mathematical foundations of 3D entropy minimization method and its applications.

The final **Chapter 9** provides a retrospective discussion and suggests some possible future works that could be endeavored from the present study.

All computational work was implemented on a NT workstation with 2GB RAM and 2 Xeon processors running MATLAB 6.5[v].

[v] MATLAB, Mathworks. http://www.mathworks.com/

Chapter 2

Literature Review

This chapter provides an overview of the theoretical background and literature relevant to this study and presents a theoretical framework for the research. The outline of chapter 2 is as follows. Section 2.1 gives a brief introduction and the development of chemometric studies. Section 2.2 reviews the various chemometric techniques used in quantitative spectroscopy. In section 2.2.1 the progress and development of various self-modeling curve resolution techniques are discussed. Section 2.2.2 reviews chemometric techniques for higher dimensional data analysis. Section 2.2.3 reviews chemometric techniques for NMR data analysis. In Section 2.3, numerical optimization algorithms used in analytical chemistry applications are reviewed. At the end, in section 2.4, there is a summary of this chapter.

2.1. What is Chemometrics?

Chemometrics has been evolving into a separate discipline within chemistry for more than three decades. The terminology “Chemometrics” was coined by S. Wold in 1971(Brereton, 1990). Chemometrics is a chemical discipline that applies mathematical, statistical and logical methods to elucidate the concealed phenomena and reveal information embedded in the observations or experimental data set. And for the chemist or chemical engineer, the revealed information forms the basis for considerably better understanding of the system. It is fair to say that chemometrics is the tool that bridges the gap between chemical data and chemical knowledge by investigating and extracting

information from the data. Chemometrics heavily relies on the use of mathematical models and applies the most widely used multivariate calibration and pattern recognition techniques to solve data analysis problems in the chemical sciences. In the early years, chemists borrowed some basic methods which originally developed in other fields such as statistics, electrical engineering, and psychology where very complex data sets are encountered and sophisticated analytical tools are needed. Today, many new methods are being developed within the chemometrics community itself.

After 30 years of rapid development, various important topics in chemometrics today include (Einax, 2004): "Descriptive statistics, planning and evaluation of sampling, experimental design and optimization, signal detection and univariate signal processing, calibration, multivariate signal processing, multivariate data analysis, geostatistical methods, time series analysis, soft modeling, laboratory information and management systems, library search and expert systems, analytical quality assurance, process analysis and optimization." Detailed reviews of the methodologies and practice of data analysis in chemistry have appeared in the biennial "Fundamental Reviews" issue of the journal *Analytical Chemistry* (Brown et. al., 1988, 1990, 1992, 1994, 1996; Lavine, 1998, 2000, 2002; Lavine and Workman, 2004).

2.2. Chemometrics in Quantitative Spectroscopy

There are various chemometric methods used in processing and interpreting spectroscopic data. It covers data calibration, the data acquisition and signal enhancement, feature selection and extraction, pattern recognition, cluster analysis and other multivariate calibration techniques. Due to the scope of the thesis, this chapter will focus on self-modeling curve resolution techniques.

2.2.1. Self –Modelling Curve Resolution

Self-modeling curve resolution (SMCR) comprises a family of chemometric techniques which target the reconstruction of pure component spectra from mixture spectroscopic data. Even though there are already many attempts to resolve the components in complex spectroscopic data sets (Wallace, 1960; Blackburn, 1965), the new term SMCR first appeared when Lawton and Sylvestre (1971) resolved a two-component system measured by UV/Vis spectroscopy in 1971. Although only applicable for a two-component system, this pioneering work inspired further studies by Ohta (1973) and Borgen *et al.* (1985, 1987). During the next two decades, significant progress was made by several research groups. Ritter *et al*.(1976) proposed a method to determine the number of components in chromatography-mass spectrometric data, and similar work also was done by Davis *et al*. (1974). SMCR analysis was successful implemented in infrared spectroscopy and the number of components in a mixture was predicted even in case where the spectra of the individual compounds were very similar (Rasmussen, 1978). In the 1980s, the *information* entropy concept was introduced into SMCR method by Sasaki and co-workers (Sasaki *et al*., 1983, 1984; Kawata *et al*., 1985). Later, Kawata *et al*. applied its extension to multispectral images data (1987, 1989). They minimized the entropy function with non-negativity constraints to search for pure component spectral estimates.

As a new discipline, chemometric techniques have experienced continuous rapid development along with their applications. In recent developments, many research groups have applied SMCR to spectroscopic studies of complex chemical kinetic and equilibrium systems (Bijlsma *et al*., 1998, 1999, 2000; Forland *et al*., 1996; Libnau *et al*., 1995; Nodland *et al*., 1996). At the same time, a number of self-modeling curve resolution

methods were made available for spectroscopic data analysis applications: Key set factor analysis (KSFA) (Malinowski, 1982), iterative target transformation factor analysis (ITTFA) (Gemperline,1984, 1986; Vandeginste *et al*., 1985), evolving factor analysis (EFA) (Maeder, 1987; Keller and Massart, 1992), window factor analysis (WFA) (Malinowski, 1992), multivariate curve resolution and alternating least squares method ,(MCR-ALS) (Tauler *et al*., 1991, 2001), simple to use interactive self-modeling mixture analysis (SIMPLISMA) (Windig, 1991; Windig and Stephenson 1992), orthogonal projection approach (OPA) (Sanchez *et al*., 1994, 1996b), heuristic evolving latent projection(HELP) (Kvalheim and Liang, 1992), SAFER (Kim, 1989), interactive principal component analysis (IPCA) (Bu and Brown, 2000), Dynamic Monte Carlo SMCR (DMC-SMCR) (Leger and Wentzell, 2002), singular value decomposition with self-modeling method (SVD-SM) (Steinbock *et al*., 1997; Zimanyi *et al*., 1999; Zimanyi, 2004).

Also non-negativity is a natural condition for many spectroscopic applications. Methods based on this property are positive matrix factorization (PMF) (Paatero and Tapper, 1994), non-negative matrix factorization (NMF) (Lee and Seung, 1999), etc.

There is another independent category of techniques developed from the signal processing field and which comes under the name of blind source separation (BSS) and within this, the most common method is independent component analysis (ICA). Blind source separation consists in extracting independent sources from superimposed signals, by manipulation of the statistical independence between sources/components. Most studies have been focused on linear systems which have some close analogs to spectroscopic data analysis in chemometrics. ICA tools have been applied to some

chemical data analysis problems (Chen and Wang, 2001; Ladroue *et al*., 2002; Ren *et al*., 2004; Stogbauer *et al*., 2004; Shao *et al*., 2004; Simonetti *et al*., 2005).

For older reviews of SMRC methods and chemometrics studies, one can refer to the contributions from Gemperline (1989), Hamilton and Gemperline (1990), Sanchez *et al*. (1996a), Mobley *et al*. (1996), Workman *et al*. (1996), Bro *et al*. (1997). More recently, reviews by de Juan *et al*. (2003) and Jiang *et al*. (2004) provided some further descriptions of SMCR methodologies.

Even though SMCR has been widely applied in chemometrics; there are still some ubiquitous problems that have not been fully addressed. (1) Non-stationarity (or nonlinearity) is the major obstacle when applying SMCR techniques to spectroscopic data, where Beer-Lambert law is not observed[i]. Therefore, a bilinear model is only locally valid and not globally valid. Data pretreatment and signal enhancement may help to some degree to correct this problem. (2) Secondly, the correct estimation of the number of components present in the systems is another very difficult quantity to determine. The experimentalist unfortunately faces the problems of unknown concentration matrix, unknown spectral matrix, unknown error component and unknown number of species all at the same time. Effort has been invested in solving this problem (Chen *et al*., 1999, 2001); and it shows that determining the number of components in the real experimental data matrix really is a hard task. (3) The inverse problem is normally ill-posed in other ways as well, for example, due to ill-conditioning and this may significantly deteriorate the performance of the self-modeling. This problem arises particularly, in the case when

[i] Several phenomena can cause a deviation from Beer-Lambert law. The two most common causes are 1. changes in temperature or pressure which induce spectral changes and 2. changes in concentrations which induce spectral changes (changes in solvation induce absorbance peaks shifting, band shape changes).

there are minor components and their contribution is small compared to the other components present, and when the noise signal contribution is significant in comparison with the minor component. In these situations, self-modeling methods may fail to predict the correct results accurately.

For more detailed discussions of SMCR technique, see section 4.1 in chapter 4.

2.2.2. Chemometric Techniques for Higher Dimensional Data Analysis

Most of chemometric tools, especially the SMCR methods, are designed to deal with 1D spectroscopic data. However, 2D spectroscopic data, which is obtained as an analytical response in matrix-format rather than a vector, is becoming much more common in today's analytical laboratory. A real need exists for the development of chemometric techniques for 2D data.

It should be noted that not all 2D formatted data is equivalent from any analysis viewpoint. Some 2D formatted data has more structure and can be factorized into the product of 2 vectors. The most common example is luminescence data (excitation-emission-matrices). Other 2D formatted data has less structure and has to be treated as a whole. Common examples are some 2D NMR and even photographs. Clearly, an analysis that can treat the less structured data would represent a more robust generalized way of solving the problems. A method/solution that can treat the less structured data will also be able to treat the more structured data.

The matrix-formatted measurement of 2D luminescence of a dilute solution, is the prototype for bilinear data which can be factorized into a row and a column. When dealing with the bilinear 2D data, there is also a theoretical "second-order advantage" which

means the accurate and reliable discrimination of the analyte can be performed in the presence of unknown interferents (Sanchez *et al*., 1987; Ramos *et al*., 1987). There are families of rank annihilation methods targeting at the resolution of such 2D bilinear data and they play an important role in the high-dimensional data analysis (Ho *et al*., 1980, 1981; Ramos *et al*., 1987; Millican and Mcgown, 1990, Faber *et al*. 2001a, 2001b).

The rank annihilation factor analysis (RAFA) was proposed by Ho *et al*. in 1978 (1978). Later it was modified into an efficient chemometric technique based on the eigenanalysis (rank analysis) for the two-way data and it is often applied to quantitatively analyze a system with unknown interferents (Lorber, 1984, 1985). But RAFA suffers from a serious deficiency, namely, that it needs a pure standard with known concentration. Sanchez and Kowalski fixed this deficiency and developed GRAM (the generalized rank annihilation method) algorithm, a general extension of RAFA and applied it to liquid chromatography diode array-UV (LC-DA-UV) data (Sanchez and Kowalski, 1986, Sanchez *et al*., 1987) and pulsed gradient spin echo (PGSE) NMR data (Antalek and Windig, 1996).

Besides GRAM, Sanchez *et al*. (1990) suggested a tensorial resolution: Direct Trilinear Decomposition. All the above are eigen-problem based methods. Another main method is the family of alternating least-square (ALS) methods which are more flexible but more numerically expensive. And these ALS methods also can be constrained with some criteria, such as non-negative, unimodality, and column-wise orthogonality. The two major significant families are PARAFA (PARAllel RActor analysis)/CANDECOMP (CANonical DECOMPosition) (Carrol and Chang 1970; Harshman and Lundy, 1996) and TUCKER3 (Tucker, 1966) series. Smilde has reviewed various TUCKER unfolding

schemes and PARAFAC modeling, and offered a discussion of the history and applications of higher-order analysis (Smilde, 1992). As an extension of PMF, namely, PMF3, a weighted nonnegative least-square algorithm for three-way factor analysis was proposed (Hopke *et al*., 1998; Paatero, 1997), and the property of nonnegativity is achieved by posing a logarithmic penalty. Such higher-order analysis also encounters many difficulties inherented from the trilinear form, including ambiguity of the correct model size (number of factors involves in the system), model mismatch and the interference by noise.

2.2.3. Chemometric Techniques for NMR Data Analysis Studies

As the most important tool in the chemical science, NMR spectroscopy has been of long interested in chemical analysis, pharmaceutical analysis (Lepre *et al*., 2004)[ii], biomedical analysis, especially metabonomic studies (Lenz *et al*., 2004; Holmes and Antti, 2002). In bioinformatics studies, the complex NMR data are treated by cluster analysis and other pattern recognition techniques, which are implemented to identify, e.g. diagnostic compounds. Normally these would involve chemometric techniques, such as, soft independent modeling of class analogy (SIMCA), and K-nearest neighbor analysis. Other chemometric techniques are also used in more general chemical science studies. Most of this work falls into the category of signal enhancement (Lin and Hwang, 1993; Koehl, 1999) and multivariate linear calibration methods (Schulze and Stilbs, 1993). However, few of them are related to the application of SMCR methods on mixture NMR data. In 1996, Antalek and Windig, applied one of the variations of generalized rank annihilation method (GRAM), namely, DECRA (direct exponential curve resolution algorithm) to directly resolve PGSE NMR mixture data; and later extended to magnetic

[ii] Also other articles in the same thematic issue: Chem. Rev. Vol.104, 2004

resonance images (Antalek and Windig, 1996; Windig *et al*., 1999). Xie *et al*. (1998) re-investigated these three NMR spectral data sets using a least-square approach called Positive Matrix Factorization (PMF). In 2002, based on the DECRA, Pedersen *et al*. (2002) proposed a method, SLICING, for the decomposition of low-field pulsed NMR data.

In 2001, Vives *et al*. (2001) applied the MCR-ALS method to study ^{1}H NMR data. In the same group, MCR-ALS approaches have been implemented for the analysis and resolution of simple transformed NMR data (Joaquim *et al*., 2003).

NMR spectra of complex mixtures possess unique features, such as, phase problems and sample-to-sample variability in peak positions due to the different chemical environment surrounding the analyte, for example, the different concentrations of molecules. Therefore, proper preprocessing tools would be extraordinarily important for NMR data analysis. Unfortunately, very little literature has reported the study of NMR data preprocessing and subsequent SMCR data analysis. Except for specific studies on PGSE NMR data sets (Huo *et al.*, 2003, 2004), the application of spectral resolution algorithms in the decomposition of general NMR data to obtain pure component spectra and concentrations is seldom reported (Brekke and Kvalheim, 1996; Alam and Alam, 2005).

2.3. Optimization Methods

In most of the SMCR methods, the kernel of data analysis consists of two parts, one is the definition of the objective function related to the curve solution, and the other is the search of the optimal solution. Many SMCR methods differ primarily by the philosophical approach taken and hence the formulation of the equations which will be

used to resolve the spectra. However, there can also be significant differences in the numerical approaches that can be taken to find optimal solutions. In most cases, the function is nonlinear; a nonlinear program (NLP) problem exists in analysis. Normally the NLP problem can be formulated as

$$\text{Min } F(x) \quad \text{subject to } g(x) = 0 \text{ and } h(x) >= 0 \qquad (2.1)$$

A solution is found by adjusting or searching a set of parameters to minimize [iii] a chosen cost function, F(x), subject to one or more of constraint functions related to $g(x)$ and $h(x)$.

Global optimization is an area with great theoretical challenges and has a broad range of scientific, engineering applications. The objective of global optimization is to find the best solution for a nonlinear problem which may possess a multitude of local optima. In practice, systematic searches can be performed to find the local optima; a typical example is gradient-based methods, which are very efficient to find the local optima but not global optimum, since it will be easily trapped in the area around its starting point. To overcome this difficulty, many algorithms were proposed to find ways to move beyond local optima to the global solution without using an exhaustive search. Tabu search, Genetic/Evolutionary Algorithm (GA/EA), Simulated Annealing (SA) are the most common methods that have been used in chemometrics and they have countless applications in analytical science and chemical engineering. Excellent reviews about these optimization methods can be found in references (Shaffer and Small, 1997; Lavine and Moores, 1999; Leardi, 2001). In the present study, GA and SA are used for solving the entropy minimization problems. Also they are often used to solve curve-fitting problems.

[iii] It is known that the many problems come naturally in a minimization form. A maximum problem can always be converted into a minimization problem by changing the sign of the objective.

Simulated annealing (Metropolis *et al*., 1953; Kirkpatrick *et al*., 1983; Siarry, 1997) is a stochastic optimization method based on the Monte Carlo sampling technique. The method starts from an initial point and takes an iterative strategy. During its annealing procedure, some steps which produce the "worse" points can also be temporarily accepted according to a Boltzman-like probability function, which depends on a pseudo-temperature parameter. At higher temperature, more steps are accepted and more state spaces are searched. However, when the temperature drops, the search range shrinks and the focus of the search shifts to improving the quality of the current solution, in other words, the increasing partition of "better" points is favored. The numerical cooling procedure mimics physical annealing and enables SA to efficiently search the solution space instead of being trapped at local optima (Brooks and Morgan, 1995).

Genetic algorithm (GA) (Holland, 1975; Goldberg, 1989) is a global optimization technique which is based on the analogy of iteratively selecting the best genes and best solutions among a population of evolving candidate solutions through modification of their gene pool and guided by evolutionary mechanisms ("survival of the fittest"). Selecting a parent from the previous generation by evaluation of "fitness", GA forms the next generation by applying genetic operators, such as mutation and/or recombination (crossover). Without the need to use complicated differential equations, genetic algorithm is often more attractive than gradient search methods even if the response space of the function is very complicated.

Compared to other methods, the strength of GA and SA methods is that they are especially useful for functions with discontinuities and multimodality since they do not make use of the gradient. They also perform well when there are lots of local optima to avoid and they do not rely on the quality of the initial guess. On the other hand,

substantial computation efforts are normally needed for GA and SA. They are less efficient than a gradient-guided method if a gradient function is available and the function is well-behaved. Also parameters such as mutation crossover rates for GA and cooling factor for SA should be fine-tuned to the particular application. Otherwise, it would either converge prematurely or never converge to a global minimum at all. Nevertheless, GA and SA are strongly favored when solving many difficult optimization problems in analytical chemistry due to their convenience.

It is unrealistic to expect to find one general optimization tool which can solve every kind of nonlinear problem. The choice of an appropriate algorithm would depend on the specific characteristic of the targeted function. For a well-behaved function and gradient available function, gradient-based solvers would be the more efficient and suitable algorithm. But in high-dimensional problems where gradients are typically not available, random search method are favored. Also a few studies have compared stochastic methods, and general guidelines about when to use GA and SA are far from complete (Lucasius *et al*., 1994; Hörchner and Kalivas, 1995).

2.4. Summary

In modern chemical science laboratories, a huge amount of data can be produced. In particular, a significant portion of this data comes from spectrometers, which can acquire many megabytes and even gigabytes of 1D, 2D, and even 3D data daily. Chemometric tools play an important part in analyzing such data. A lot of effort has been invested in resolving pure component spectra using SMCR techniques. Global optimization plays a key role in achieving numerical solutions to many of the problems.

Chapter 3

Data Manipulation in Spectroscopy

As introduced in chapter 2, chemometrics is a field of data analysis that involves the application of various multivariate analysis methods to chemical data sets. In practice, after considerations about experiment design, the experimental data is collected with spectroscopic instruments. It is clear that the success of data analysis heavily depends on the quality of the data. After data acquisition, the first step is often a data pre-processing procedure. The reason is that spectra, as well as other types of analytical measurements (and their data arrays), often contain both random and systematic errors and noises, and artifacts of various types. Often the unwanted signals or variations should be eliminated or at least reduced. The corrected or pre-processed data is then more suitable for other chemometric techniques. A wide variety of tools are employed for this purpose in exploratory data analysis. In this chapter, a detailed exposition of the concept and mathematical principles related to some data manipulations needed in this study is provided.[i]

---

[i]There is some specific nomenclature used in higher dimensional analysis, i.e. the terms "dimension" and "way". The dimension of a data array refers to the number of coordinates needed to specify a point inside the array and refer to "dimensionality" also. For example, an $n \times m$ array is two-dimensional array, while a tensor ($n \times m \times k$) array is three-dimensional array. However, in some literature, an $n \times m$ array is described to be *n* points in an *m* dimensional space, in another word, the data is m-dimensional. In order to avoid confusion, in this thesis, the traditional nomenclature is adopted and the $n \times m$ data array would be regarded as 2D (two-dimensional or two-way) data. By analogy, an ($n \times m \times k$) array is three-dimensional, and is also called three-way data.

3.1. Different Types of Measurements in Multivariate Analysis

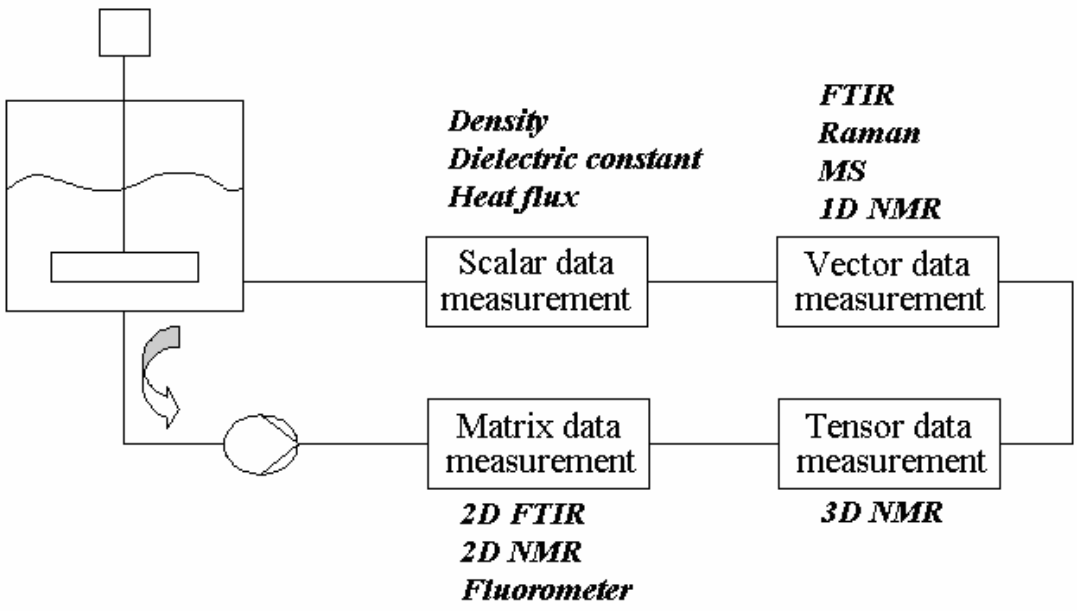


Figure 3.1. A batch reaction with four kinds of on-line measurements according to the dimension of the individual measurements

In chemical and engineering science, data are collected from all kinds of detectors. Figure 3.1 represents, to some extent, a general liquid phase reaction carried out in a batch reactor in fine chemical or pharmaceutical research. The reaction is carried out with some sorts of on-line measurements. Ideally, in a very modern and well equipped research laboratory, one would like to make as many simultaneous and sophisticated measurements as possible, in order to properly and thoroughly investigate the system.[ii] The inversion of such data leads to a model of the system (system identification).

[ii] Such a system like Figure 3.1 is presently being completed in Prof Garland's *Advanced Reaction Engineering, Process Analytics and Chemometrics* lab at the Institute of Chemical and Engineering Sciences in Singapore (ICES). The scalar instruments include (1) a densitometer (Anton Paar, Austria) (2) a refractometer (Dr Kernchen GmBh, Germany) (3) a dielectric cell (Scientifica, Princeton NJ) and (4) an ultra-sensitive flow through calorimeter (Thermometrics, Sweden). The vector instruments include (1) Far Infrared (Bruker, Germany) (2) Mid Infrared (Bruker, Germany), (3) Raman Optical Activity (Biotools, USA). Facilities for MS, 1D, 2D and 3D NMR are available in the neighboring lab.

As shown in Figure 3.1, meaningful on-line measurements can be classified into four categories according to the data format of the individual measurements, e.g. scalar, vector, matrix and tensor. However, these individual measurements can be grouped in various ways and this leads to one-way, two-way, three-way, four-way data etc.

One-Way Data:

All measurements in the engineering sciences are related to the physical phenomenon or response (radiation, density, electric current, and mass). Quite a lot of thermo-physical property measurement data, such as density (ρ_v), heat capacity at constant pressure (C_p), heat capacity at constant volume (C_v), thermal conductivity (h) and surface tension (σ) are simple scalar measurements. If a series of measurements of temperature (scalar data) in a reaction system are collected with a temperature detector (such as a thermometer), the data set is considered one-way data since the scalar data is only collected in one direction (a time series) denoted as $T(t)$, over time. Also if a sensitive heat flow calorimeter is used, simultaneous heat flux measurements Q(t) are collected over time. Other forms of one-way data can be time series measurements of viscosity (μ) or dielectric constant (ε).

Two-Way Data:

As in the above system, a digitized FTIR spectrometer can be employed to measure the mixture spectra of the liquid phase at specific intervals in time. A set of vector-formatted data are collected along time which forms a "two-way" data set, denoted as A (v, t). This data set is a matrix A of absorbance with one way along wavenumber (v)

and a second way along time (*t*). Analogously, a matrix *I* (*m/e*[iii], *t*) of intensity measurements can be obtained from an on-line MS (mass spectrometer) and a matrix *I* (*ppm*[iv], *t*) from an on-line NMR spectrometer. This kind of two-way data are formed by time-series measurements of vector-formed data. These examples are perhaps the most common.

Hyphenated two-way data sets also exist, and these represent another type of 2-way data. Good examples are GC-MS, HPLC-MS, GC-IR, HPLC-UV and GC-GC. All of these examples represent series of spectra collected at even intervals of time during a chromatographic run. This yields a matrix indexed by wavenumber from the spectra in one direction and by time from the chromatography in another direction. All of the mentioned examples arise from hyphenated instruments which consist of two instruments in series (Albert, 2002).

Yet another kind of two-way data exists and this type is often referred to as a 2D spectroscopy. A 2D spectroscopy arises from excitation at one frequency followed by changes at other frequencies. Examples are represented by two-dimensional infrared (2D IR) spectroscopy (Zhao and Wright, 2000; Ge and Hochstrasser, 2001; Brixner *et al*., 2005), and two-dimensional Raman (2D Raman) (Tanimura and Mukamel, 1993) which can be denoted as *A* (ν_1,ν_2). Other more common examples are two-dimensional NMR spectroscopy (2D NMR Spectroscopy) (Wider *et al.*, 1984, Ernst *et al*., 1987) and 2D excitation emission fluorometry (Patonay, 1987).

[iii] m/e refers to "mass to charge ratio", which is the x-axis of a mass spectrum.

[iv] In NMR spectroscopy, the chemical shift is defined as $\delta = 10^6 \times \frac{\nu_{signal} - \nu_{reference}}{\nu_{reference}}$ (ν: frequency of resonance), so that its value can be independent of the magnetic field strength. The scale is made more manageable by expressing it in *parts per million* (ppm).

Three-Way Data:

Three-way can be formed by collecting two-way data along a third direction. An ordinary chemical example could be 2D excitation-emission matrix fluorescence spectra collected for a series of discrete physical samples. Another example could be chromatographic separation followed by any kind of 2D spectroscopy – again for a series of discrete physical samples. Finally, a series of 2D spectra collected in time is also an example of a three-way data set. The development of second-order (i.e. GC-MS or 2D NMR) and higher-order (i.e. the GC-MS-MS) instrumentation has benefited modern analytical chemistry by providing two and multidimensional data arrays, which greatly enhances chemical identification.

Analogous to the 2D NMR, the one type of 3D NMR data can be constructed by combining two kinds of 2D NMR experiment which consists of correlating the various nuclei either through scalar coupling (COSY, TOCSY, HMQC, and HSQC) or through space (NOESY) and spreading this overlapping along the third chemical shift axis by combining two 2D experiments.

Another type of 3D NMR experiment is the so called triple resonance experiment. The most common type is found in the bio-molecule studies where a ^{1}H-^{13}C-^{15}N triple resonance is detected. An individual 3D triple resonance NMR measurement is also a three-way data array.

Four-Way Data:

It is obvious that more elaborate analytical instruments increase the dimension of the data sets collected. Actually, there is no limit to the number of "ways" in which the data set form. Therefore, four-way data can be constructed in many ways.

For the purpose of this thesis, there is only one type of four-way data which is particularly important, and this is formed from a set of 3D triple resonance NMR measurements.

Univariate, Bivariate, Multivariate Data

The number of variables measured also provides a way to classify the data. Depending on the number of variables measured simultaneously, we obtain three different categories of data: (a), univariate data: measurements with only one variable per observation. (b), bivariate data: measurements with two variables per observation. (c), multivariate data: measurements with many variables per observation. A multivariate data set consists of several variables recorded for a number of objects or samples. The most obvious example of multivariate data is spectroscopic data where a spectrum is recorded at hundreds of different frequency channels on a single sample instantaneously.

3.2. Data Pretreatment and Data Enhancement

As already mentioned, it is apparent that the success of data analysis heavily depends on the quality of the data. Many available chemometric techniques are based on the linearity of the data (for example, bilinear model for PCA or trilinearity for PARAFAC and TUCKER3). But unfortunately this important prerequisite for linearity is seldom adequately met in the real world and the measurement is always deteriorated by experimental error or noise. It is an important step to try to "clean up" the data set prior to the further data analysis. Suitable pre-processing of the data would be very useful in improving the quality of the treated data. These pre-processing methods include, but are not limited to, signal enhancement (signal filtering, signal smoothing, signal restoration).

Also outlier detection and spectral alignment are indispensable techniques in many data processing procedures. A brief discussion of these methods follows.

3.2.1. Outlier Detection

The term outlier normally refers to an observation that lies an abnormal distance from other similar observations in a data set. This is normally attributed to infrequent random events (i.e. a power fluctuation) or even mistakes. The presence of outlier has a detrimental effect on the data analysis since an outlier can easily bias estimations of a spectrum and make the model fail, distort the result and obscure their prediction. In statistics, outliers can be dectected by checking with range of the "inner fences" and "outer fences", whose definition relates to the value of quartiles. An observation beyond the inner fence would be labeled a mild outlier, and a point beyond an outer fence is considered an extreme outlier. Hotelling's T^2 test which is based on the squared Mahalanobis distance is another method for detecting outliers (Jackson, 1991). Other more sophisticated treatments for outlier detection exist (Walczak, 1995; Singh, 1996).

3.2.2. Data Filtering

Signal-to-noise enhancement of 1D signal can be accomplished by several well-developed methods.

3.2.2.1. Time Averaging/ Ensemble Averaging Method

Ideally, signals can be distinguished from noise by the fact that the noise is not reproducible, whereas, the genuine signal is at least partially reproducible. Therefore the signal to noise ratio can be improved by repeating the successive set of measurements and

adding up all this data point-by-point. The price for the improvement of signal-to-noise-ratio is the time needed for repeated measurements. Even though the ensemble averaging method is one of the most powerful methods for improving signals, it seems that it is not suitable for measurements which are slow on the reaction time scale. Trying to reproduce the signal as many times as possible, seems impractical for continuously reacting systems.

3.2.2.2. Moving Average Algorithm

The moving average method is often used to reduce the effects due to random variation. The simplest form of moving average simply replaces each data value with the average of neighboring values. Mathematically it can be implemented by convolving the untreated spectrum with a box-shaped function of $2 \times m + 1$ points with their values all equal to $1/(2 \times m + 1)$, where m denotes the number of points before and after the target data point (where the average is being performed). More details can be found in the useful references (Tomita and Tsuji, 1977; Wells, 1986).

3.2.2.3. Savitzky–Golay Smoothing Method

Savitzky–Golay smoothing method is a popular smoothing method in the field of chemistry (Savitzky and Golay, 1964). Instead of simply averaging, Savitzky–Golay smoothing method can be thought of as a generalized moving average method. It performs a least-square polynomial fit of a small set of consecutive data and replaces the central point with the new smoothed calculated point of the fitted polynomial curve by using a set of pre-computed weighting coefficients. The polynomial order and the number of points

used to compute each smoothed output value are two parameters which should be tuned for specific applications.

3.2.3. Fourier Transformation and Wavelet Transformation

Instead of filtering the data in the time domain, filtering can be implemented in the frequency domain. Fourier transformation (FT) is the most common transformation procedure in spectroscopic data analysis. In the frequency domain, one can truncate the noise part and back-fill them with zeros to remove the high-frequency components (noise). A signal enhancement will then result when the inverse Fourier transformation is calculated. Briefly speaking, signal smoothing is performed by FT by removing completely the noise frequency components, while the information bearing frequency components are still retained (this assumes that the signal has a lower frequency component than the noise). It is noted that finding a suitable digital filter is very critical for the success of filtering process. High-frequency noise may be removed by a low-pass filter, but if the filter is too soft, unfiltered non-linearity or noise would mess up later processing steps, meanwhile, if over-filtered, some essential information will be lost or the spectrum will be distorted and further other nonlinearities will be introduced into the system.

As an analogue of Fourier transform, wavelet transform replaces the sinusoidal waves of Fourier transform by a family of functions which are generated by translation and dilation of a wavelet. By setting the threshold[v], one is able to retain the primary contributions of the signal and successively denoise the spectra. The whole spectrum can

[v] The purpose of setting threshold is to keep those coefficients that are sufficiently large, and to re-set the rest of the small coefficients to zero.

be reconstructed using these coefficients by an inverse wavelet transformation. A detailed and practical introduction can be found in MATLAB [vi] Wavelet Toolbox, tutorials (Alsberg *et al*., 1997; Jetter *et al*., 2000) and references (Meyer, 1993; Kaiser, 1994; Vetterli and Herley, 1992).

3.2.4. Maximum Entropy Method (MEM)

Since the Band-Target Entropy Minimization method is based on the entropy concept from information theory, some theoretical background information about the MEM algorithm would help to clarify the entropy concept. Since both MEM and BTEM belong to "spectral reconstruction", misunderstanding easily arises (Lavine and Workman, 2004) and the very significant and important differences have to be made clear.

The maximum entropy method is regarded as a variation on the Bayesian method. This approach was pioneered by Jaynes (1957) and applied to statistical physics in 1957. MEM is designed to extract as much information from "one" measurement as possible. According to the maximum entropy method principle, the estimate that has the maximum entropy would be the *single* best candidate within all possible spectral reconstructions consistent with this *one* measurement. (In contrast, BTEM aims to reconstruct *all* the embedded pure spectra within an *entire set* of mixture data.) Maximum-entropy method has been successfully used in a variety of scientific fields, including NMR spectroscopy, X-ray crystallography, fluorescence, astronomical imaging, and digital image restoration (Laue *et al*., 1986; Gilmore *et al*., 1990, 1993).

In information theory, entropy refers to lack of order. So it is fairly plausible that the maximum entropy method seeks the unbiased solution which has minimum structure

[vi] MATLAB, MathWorks Inc. MATLAB Reference Guide, 1995.

(order) that remains consistent with the data and the prior knowledge. The entropy of a spectrum with equal probability of measuring a random noise at all the wavelengths reaches the maximum value – so pure noise with no structure has a maximum entropy. The entropy of the re-constructed spectrum will decrease if the contribution of noise decreases and the entropy will increase if the contribution of noise increases.

In practice, the Maximum entropy method is approached by maximizing the entropy of the estimate while minimizing the discrepancy between the estimate and the raw observations/data. The Cambridge algorithm, one implementation of MEM method which was originally developed for applications in astronomy (Skilling and Bryan, 1984), consists of the following steps. First, an initial trial spectrum is made if no prior knowledge is available, and this trial object is blurred with a blurring function to get an estimated one. The residuals obtained by subtracting the estimated one from the raw data are used to modify the initial guess to give a new estimated spectrum. And the new estimate is again blurred and new residuals computed. During the process of the modification to the next estimate, all the negative values are set to zeros. This procedure is repeated until the variance of the residuals is reduced to be almost the same level of noise as the original measure spectrum (this can be realized with, for example, a chi-square test). Since there would be many sets of candidate spectra which would generate similar residuals to the original noise, we need a simple guideline to make the choice. Entropy maximization is used to select the minimal structure (maximum entropy) from the all accepted spectra. On the other hand, a "fitting "constraint is imposed to make sure that the calculated result will not depart from the experiment mean by more than one standard deviation. Therefore, the MEM algorithm comes into a "tug of war" between the minimizing the "fit" and the maximizing entropy in practice.

Maximum-entropy reconstruction is a powerful method for spectrum analysis. It is frequently used as an alternative to classical methods based on the discrete Fourier transform (Sibisi, 1983; Sibisi *et al*., 1984; Jones and Hore, 1991). It is worth repeating that MEM is applied to an *individual* spectrum. Statistically, MEM gives us a more uniform or broad distribution. Details of the MEM can be found in many references (Cornwell and Evans, 1985; Skilling, 1989; van Smaalen *et al*., 2003).

3.2.5. Alignment

There are various undesired variations associated with the ordinate values in spectroscopic data matrices, which may deleteriously affect subsequent data analyses. Sometimes severe non-linearities due to shifting peak positions and changing peak shapes occur. For example, it has been reported that temperature has a strong influence on the position and intensity of near infrared (NIR) spectral absorption bands and thus affect the predictive ability of the associated calibration model (Delwiche *et al*., 1992; Thygesen and Lundqvistt, 2000; Blanco and Valdes, 2004). It is known that shifts in X-ray diffraction data occur due to calibration/alignment experimental error (Jenkin and Snyder, 1996). Also effects of laser frequency shift happen in Raman spectroscopy (Swierenga *et al*., 1999). The difficulties encountered in aligning the mass spectral peak of time-of-flight secondary ion mass spectrum (TOF-SIMS) (Zheng *et al*. 1995) imply a similar situation.

Since spectral non-linearities are almost ubiquitous for spectroscopy, proper pretreatment is needed to “clean” the spectroscopic data. This non-stationary characteristic caused by constant peak shifting can be corrected by pre-processing the data with a re-alignment algorithm. The nonlinearity caused by peak shifting should be eliminated, or at least partially corrected before data analysis.

For alignment, several methods have been proposed. Two kinds of well-established warping algorithms: DTW and COW, have received considerable attention for chromatographic data, spectroscopic data, and gene sequence data. DTW (dynamic time warping) was borrowed from the electronic community which initially devised it for aligning frequency spectra of speech (Myers and Rabiner, 1981). In general, DTW is a method designed to achieve an optimal alignment between two given sequences which are "warped" non-linearly to match each other. In 1998, Nielsen *et al*. (1998) introduced an algorithm called piecewise linear correlation optimized warping (COW) of chromatographic profiles. It begins with the selection of the target spectrum, which is representative of the whole set of spectra. The target spectrum is divided into several segments which can be either of equal or different length depending on specified features in the data set. Then the predetermined spectral segment is stretched or shrunk to match the corresponding segment in the target spectrum. The measurement of match/mismatch is the correlation coefficient and its variance. COW was adopted to reduce chromatographic variation on multi-way models and significantly facilitated the modeling (Bylund, 2002). Both of them were devised to correct the position shifts prior to consequential modeling. DTW works through the signals element by element until all the spectra are aligned. The measurement/evaluation is measured by the cumulative distance between the peaks, which means that this method would be sensitive to peak height difference. Tomasi *et al*. (2004) studied both COW and DTW as preprocessing for chromatographic data and discuss the connection between the two algorithms.

Other general alignment methods include: PAGA (Peak alignment by a genetic algorithm) method (Forshed *et al*., 2003) which aligns a spectrum to a corresponding reference part by sideways movement via a GA optimizer, Partial linear fit (PLF) method

(Vogels *et al*., 1993) which minimizes the difference between shifted result and reference by trying each possible relevant combination of spectral segment size and movement size.

Brown and Stoyanova proposed an automatic phase correction and frequency shift of one single NMR resonance peak in a series of spectra by using PCA, and this algorithm was further improved (Witjes *et al.*, 2000; Stoyanova and Brown, 2002; Stoyanova *et al.*, 2004). But it is still assumed to be applicable to spectrally isolated peaks.

3.3. Data Decomposition

Multivariate data analysis is related to a family of data analysis techniques which aims at investigating the patterns of relationships between several variables simultaneously. Multivariate data analysis deals with all kinds of different measurements. And it is designed to uncover significant relationships among the variables of the samples. Most of the time, PCA and SVD are the primary tools for multivariate data analysis. Decomposition will factorize the data set from many variables to a few factors which can express the main information and facilitate the variance analysis in the data.

3.3.1. Principle Component Analysis (PCA)

Principal component analysis forms the basis for multivariate data analysis which allows us to explore patterns in data, similar to exploring patterns in psychometric data (Wold *et al*., 1987; Jackson, 1991). An early description of PCA was made by Cauchy in physics in 1829 and it was further developed by Pearson (1901) and Hotelling (1933).The main purpose of PCA is to project the data from a higher dimensional space into a lower dimensional space, therefore concentrated onto a few underlying latent variables which capture most of the information of the data. These new bases in the subspace are also

called components ("latent variables", "regression factors", or "factors"). So a large number of variables can be substituted by a small number of new latent factors needed to reproduce the original data matrix.

The PCA can be considered a general framework for rank reduction and data compression which re-express a noisy and garbled data set by some new and most meaningful *basis*, meanwhile, redundancy and small noise variabilities are removed. So PCA also addresses data redundancy reduction with decrease in dimensionality, leading to more parsimonious, more robust models. This separation of structure and noise can be utilized for data compression. Statistically, PCA finds lines, planes and hyperplanes in the K-dimensional space that approximate the data as well as possible in the least-square sense.

In practice, a simple way to calculate principal component begins with the covariance matrix of the data set. By solving eigenvalues and corresponding eigenvectors of the covariance matix and ordering the eigenvectors in the order of descending eigenvalues (largest first), an ordered orthogonal basis is obtained. For more details, see two excellent discussions of the issues given by Joliffe (1986) and Jackson (1991).

3.3.2. Limitation of Principle Component Analysis (PCA)

A primary assumption of this method is that the data itself is valid and all data is suitable for the model. The "garbage in and garbage out" rule definitely applies when messy data are provided. The quality of the data unequivocally affects the result of processing. Also another drawback of classical PCA is that it is not robust to the presence of "outliers" which are common in real data. Some "robust PCA" algorithms have been proposed to remedy this weakness (Croux and Haesbroeck, 2000). Also one property of

PCs (principle components) is their orthogonality, which may not be necessary or even desired in some real applications. Therefore other factorization methods like independent component analysis (ICA) and Non-negative matrix factorization (NMF) have been developed.

As an essential technique for data analysis, PCA still plays a fundamental and important role in many areas of chemometrics. A lot of articles discuss PCA and different aspects in detail (Horst, 1992). PCA decomposition can be calculated directly using Singular Value Decomposition (SVD), or iteratively using the Non-linear Iterative Partial Least Squares (NIPALS) (Wold *et al*., 1987) algorithm.

3.3.3. Singular Value Decomposition (SVD)

There are more robust and reliable PCA techniques. SVD (Deprettere *et al*., 1988; Berry, 1992; Berry *et al*., 1993) is a general mathematical technique used to extract principle components of complex mixture spectra. Even though they have the same purpose, there are some differences between PCA and SVD. SVD is the more robust and efficient process (Golub and Kahan, 1965; Golub and Reinsch, 1970). SVD has a lot of applications, including data compression and visualization, micro-array data analysis, and control problems (Karlsmore *et al*., 1994; Romo *et al*., 1995; Alter *et al*., 2000; Wall *et al*., 2001; Yeung *et al*., 2002; Schmidt *et al*., 2003).

Let A denote a $m \times n$ matrix of real-valued data and where without loss of generality $m \leq n$. The observation matrix A can be decomposed into three parts: $A_{m\times n} = U_{m\times m}\, \Sigma_{m\times n}\, V^T_{n\times n}$, where U is an $m \times m$ matrix, Σ is an $m \times n$ diagonal matrix, and V^T is also an n x n matrix, where the superscript T denotes the standard transpose. All

of U and V are orthonormal, in other words, $UU^T = VV^T = I_m$ (I denotes the unit matrix with size m). The columns of U are called the *left singular vectors*, u_k, and form an orthonormal basis for the principal *scores*, so that $u_i \cdot u_j = 1$ for $i = j$, and $u_i \cdot u_j = 0$ otherwise. The rows of V^T contain the elements of the *right singular vectors*, v_k, and form an orthonormal basis for *loadings*. The middle diagonal matrix Σ contains the *singular values* of matrix A, and the square of its diagonal elements represent the amount of information corresponding to each principal component. By convention, the ordering of the singular vectors is determined by high-to-low sorting of singular values, with the highest singular value in the upper left index of the Σ matrix. Note that for a square, symmetric matrix A, singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem. For references on the mathematics and computation of SVD, see the work of Gentle (1998) and Golub and Van Loan (1996).

3.3.4. Number of Components

In some SMCR methods, the "statistical determination" of the number of components is one of the most important preliminary steps for the result. If the wrong number of components is chosen, realistic pure spectra and all other related results may be erroneous. PCA is generally used to determine the number of pure spectra present in the data. Cattell proposed a measure of indicating the contribution of each major component by the use of a "scree" graph that is simply a plot of eigenvalue versus PCs (Cattell, 1966). Other different methods include: LEV (log eigenvalue) plot (Craddock and Flood 1969, Farmer, 1971), Malinowski's IND (Malinowski, 1977, 2002) and REV methods (Malinowski, 1987); F-test (Faber and Kowalski, 1997), the permutation test (Dijksterhuis

and Heiser, 1995) and leave-one-out cross-validation (Wold, 1978; Eastment and Krzanowski 1982). However, normally the result of numbers of components from different methods is not consistent. A recent reference can be found in the study where more that ten methods for determining the number of significant components were tested (Wasim and Brereton, 2004).

Actually, it is not at all surprising that there are problems with the "statistical determination" of the number of pure spectra present. The basis for most if not all the tests is the assumption that the spectra are stationary and that the spectroscopic matrix represents a linear system. In most if not all cases, this is not possible. So the starting assumption of linearity possesses a problem. This issue will be addressed a number of times in this thesis.

3.4 Hyphenated Data Analysis

Simultaneous spectroscopic data are now more common with the increasing need for hyphenated measurements in chemistry. The complementary information provided by independent spectroscopic data will facilitate the interpretation of an unknown/partly known chemical mixture. Several kinds of structural determination spectroscopes can be combined and used for the identification process. An example is the application of FTIR-VCD (vibrational circular dichroism) to identify chiral components in a mixture system. If huge amounts of data are collected for the same samples, separate analysis of each data set causes loss of the intrinsic relations (correlations) between the observations taken from the same system. Extracting such information from the whole data set leads to better qualitative and quantitative understanding on each of the interesting components.

The first problem faced with really big arrays will be the manipulation of large scale data sets, including the singular value decompositions required when the number of floating point operations is enormous. Fortunately, several software packages have been proven to be efficient tools for solving large-scale eigenvalue problems (Lehoucq, 1998; Larsen, 1998; Berry 1992). MATLAB has also adopted the ARPACK package in MATLAB 6.0 for calculating a few singular values and vectors for a matrix with large dimension *n*.

3.5. Multi-Way Data Analysis and High Dimensional Decomposition

It is common practice to have a single matrix data set with *m* rows and *n* columns. But in a number of cases, the data must be collected in a three-mode data "box", which has *m* rows, *n* columns, and *k* slices. As already noted, even higher dimensional data sets are possible. A decomposition on the data is necessary in many applications and make the encoding easy to interpret. For these purposes, two major significant families exist, namely, are PARAFA and TUCKER3 series. These two algorithms relate to alternating least-square (ALS) methods, and they can used with constraints, such as non-negative, unimodality, orthogonality etc. PARAFAC and TUCKER3 are regarded as complementary to the PCA algorithm.

3.5.1. PARAFAC/CANDECOMP Model

Harshman(1970)and Carrol and Chang (1970) independently developed an easy-to-interprete model for fitting an *n*-linear model to an *n*-way array. It is called PARAFAC (*PARallel FACtor analysis*) by Harshman, whereas Carrall and Chang named their method

CANDECOMP (canonical decomposition). A three way PARAFAC model can be formulated as the following algebraic equation:

$$x_{ijk} = \sum_{n}^{N} a_{in} b_{jn} c_{kn} + e_{ijk} \tag{3.2}$$

The elements a_{in} , b_{jn},and c_{kn} are the decomposed triads.

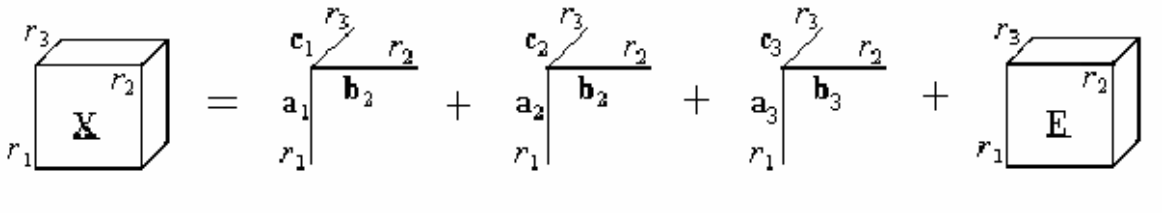


Figure 3.2. A three-component PARAFAC/CANDECOMP model.

3.5.2. The Tucker3 Model

A more general analysis tool for three-mode factor analysis based on reducing the dimensionality of all the three modes to extract the information in the three-way data was proposed by Tucker (1966). All these modes are symmetrically treated and defined as:

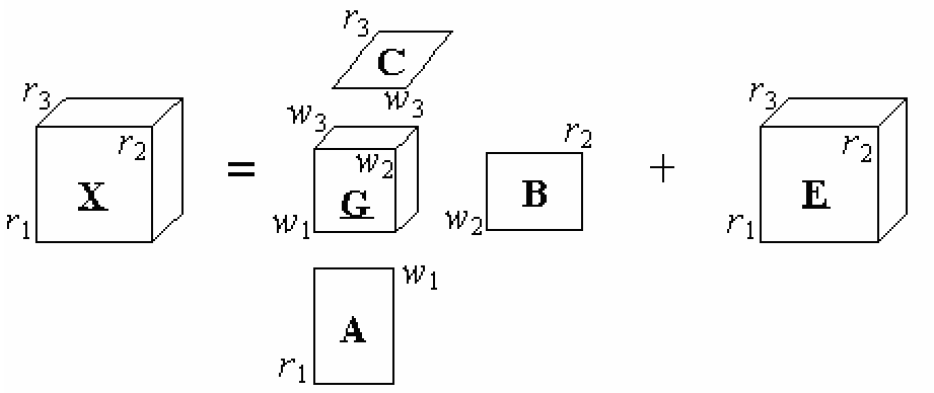


Figure 3.3. A TUCKER 3 model.

$$x_{i.j.k} = \sum_{l=1}^{w_1}\sum_{m=1}^{w_2}\sum_{n=1}^{w_3} a_{i.l} b_{j.m} c_{k.n} g_{l.m.n} + e_{i.j.k} \qquad (3.3)$$

where w_1, w_2 and w_3 denote the degraded dimensionalities of the component spaces for the three modes respectively. And *a, b, c* are the elements of the extracted component matrices *A, B, C*. The three-way array *G* is the so-called core whose elements *g* denoting interaction between these triads. Kroonenberg and de Leeuw (1980) suggested an alternating least squares fitting for the TUCKER3 model. The TUCKER3 model allows for extraction of *different* numbers of factors in each of the modes which are different with the PARAFAC model. The most difficult part of TUCKER's model is how to interpret the core matrix elements from which essential information on the interaction between modes can be obtained.

3.5.3. Comparison

One major difference of the TUCKER3 model compared with the PARAFAC model is the presence of the core array $\underline{G}$ (w_1, w_2, w_3). Fig.3.3 illustrates that the TUCKER3 model is a weighted sum of all possible outer products (i.e., triads), where the weight of the outer product among the *i*th factor from *A*, the *j*th factor from *B* and the *k*th factor from *C* is determined by element g_{ijk} of the core. The elements of *G* represent the interaction between these component triads.

The PARAFAC can be regarded as TUCKER3 by imposing the constraints that the core array $\underline{G}$'s elements $g_{w1,w2,w3}$ =1,when w_1= w_2= w_3, and equate zero otherwise. The heavy constraint results in the simplest format together with the expected loss of fit. For trilinear data or almost trilinear data, PARAFAC is preferred for easy interpretation. But as

for TUCKER3, the analyst will not only face the complexity of interpretation of the result, but will also be frustrated with the ambiguity of rotation.

The PARAFAC model always yields a unique solution, whereas the TUCKER3 model does not. It is certain that there are an infinite number of different solutions *A*, *B, C* and $\underline{G}$ that fit the $\underline{X}$ equally well even when we fix the values of *w1, w2, w3*. Mathematically it can be explained by the fact that one solution of TUCKER3 can be transformed by adding a rotation factor into their mode matrixes resulting in a new and different solution without any loss in fit.

The profiles derived by the PARAFAC model are often left unconstrained for chemical applications. However, the factors in the TUCKER3 model are often constrained to be orthogonal since the resulting core is easier to interpret and the model requires much less time for computation. Also some efforts has been invested in simplifying the core array in a parsimonious sense by obtaining a fixed number of zeros in the core by appropriate rotation (Kier, 1998) or by the maximization of the leading squared core entry in TUCKER3 (Henrion, 2000).

3.5.4. The Discussion of Multi-Way System Analysis

In the application to a real system, many unexpected difficulties will be met. The first and most important thing is how to determine the optimal number of factors which is extracted from the data set with maximum information. Recently Chen *et al*. (2001) proposed an algorithm named ADD-ONE-UP, which performs the PARAFAC decomposition repeatedly, and thus determines the factor number when some pre-specified degree of fit is achieved. This method suffers from an expensive computation cost.

Once the factor number problem is resolved, the next concern will be focused on the choice of model and algorithm. Depending on their degree of trilinearity, 3-way modeling methods can be classified into two groups: (1) direct trilinear decomposition and PARAFAC, and (2) TUCKER3 and MCR-ALS. Recently some algorithms, such as PARAFAC2 (Kiers *et al*., 1999), PARATuck2 (Harshman and Lundy, 1996), have been proposed to deal with non-trilinear data while keeping PARAFAC-like uniqueness in the solution. The computation cost of performing these decompositions will differ.

In practice, the prerequisite of trilinearity seldom can be met. For example, non-linear detector/instrumental responses may introduce deviations from trilinearity. In hyphenated experiments with LC or GC the retention time shifts are sometimes not so reproducible. Also for fluorescence EEM data, even though the signal of the analyte follows a strict trilinear mode, background signals which consist of first and second order scattering do not (Esteves da Silva, 2002).

3.5.5. Multi-Way Analysis with Unfolding

The intuition-driven way of tackling a higher-dimensional array is unfolding. Unfolding creates a hierarchy of each mode with slices along different modes which breaks the n-dimensional structure in the data set. The illustration diagram is shown in Figure 3.4.

Then conventional chemometric methods can be applied to the resulting matrix after unfolding. The MCR-ALS method has been applied to unfolded three-way data sets (de Juan and Tauler, 2001).

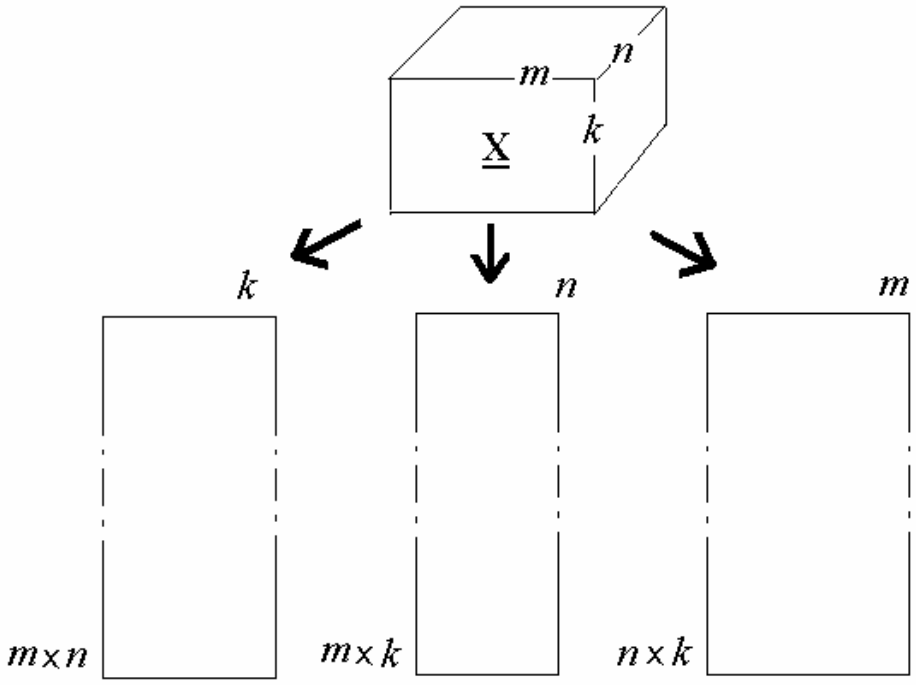


Figure 3.4. A three-mode data set and the three kinds of unfolding

3.6. Summary

In this chapter, different types of measurements encountered in multivariate analysis are mentioned. Also many methods available for the effective preprocessing are discussed. But it is known that it is still hard to decide which method is suitable for a particular application since it depends on the nature of the specific data in question. Not discussed in this chapter are other known pretreatment methods, such as spectral pre-conditioning, baseline correction etc., which will not be needed to solve any of the problems presented in this thesis.

Data decomposition, including multi-way data analysis and high dimensional decomposition were the main focus of this chapter. With decomposition, one can obtain

the singular values and the right singular vectors of the mixing structure, which could in principle be further transformed into spectral estimates by meaningful transformations. The indeterminacy of rotation is the big challenge to pure spectra recovery. PCA breaks down the observation matrix into a series of orthogonal vectors, but does not achieve a realistic solution. Depending on *a priori* knowledge and spectral features, some of these constraints can be interactively chosen for specific cases to break this ambiguity. It would be most meaningful if PCA/SVD can be used as a starting point and spectral estimates predicted with little or no a priori information.

In the next chapter, the group's new 1D entropy-based curve resolution method will be introduced and extended. Then in the following chapters pattern recognition for 3-way and 4-way data will be developed.

Chapter 4

1D Minimum-Entropy Based Pure Component Spectral Reconstruction

In this chapter, a relatively new methodology for pure component spectral reconstruction based on entropy concept will be presented. Beginning with an introduction of the concept of entropy and a necessary review of SMCR methods, the new pure component spectral reconstruction method based on minimum-entropy is described and extended. It is applied to several real chemical reaction projects and very accurate pure component spectral reconstructions are obtained. It is also successfully applied to the blind source separation problem in acoustics.

4.1. Entropy Minimized Spectral Reconstruction – Algorithm

4.1.1. Concept of Entropy

In 1948, Claude Shannon (1948) published the seminal paper "A mathematical theory of communication" which laid the foundation of information theory. The concept of *Shannon's entropy* plays a central role in information theory. This entropy refers to the *measure of uncertainty* which provides a way to estimate the average minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols. The Shannon entropy equation is shown in Eq. 4.1.

$$H = -\sum_{i=1}^{N} p_i \log p_i \tag{4.1}$$

In the Shannon entropy equation, p_i is the probability distribution of a given symbol. The entropy of a random variable is defined in terms of its probability distribution and can be shown to be a good measure of randomness or uncertainty. In the spectroscopic field, the

technique most related to entropy concept is the Maximum Entropy Method. (Details of MEM were discussed as a signal restoration method in chapter 3, section 3.2.4.) It is worth noting that entropy is a common concept in many fields, including thermodynamics, pattern classification, information theory, signal processing, etc. In the literature, there are different types of entropy expressions used for various purposes in various fields. Classical information entropy describes information-related properties for a given signal. In the signal processing field, Shannon's entropy can be viewed as a measure of the degree of randomness of the observed variables. Therefore, random noise has a very high entropy.

Conceptually, the approach taken in this chapter is very simple. A set of observations will be searched in order to find the simplest irreducible patterns. These patterns will have a very low entropy. The big challenge is to find a numerical way to achieve this.

4.1.2. Entropy Minimized Spectral Reconstruction

4.1.2.1. General Bilinear Model

In chapter 1, the X term in Eq. 1.1 represents the input of the process; Y is the observation of the process function f, with the inference of E. In a totally unknown system, the solution to the Inverse Problem seems impossible since little is known about the quantities on the right hand side of Eq. 1.1, neither the function f nor the inference E. Inverse problems are often ill-posed, which means the process of recovering the unknown parameters is very sensitive to errors in the measured response. And a problem is considered to be ill-posed, if it does not meet one or more of the following criteria (Hadamard, 1902): (1) for each set of data, there exists a solution, (2) the solution is

unique and (3) the solution depends continuously on the data. Also it is worthy to note that in some cases, there are more unknown parameters than the number of limited observations in the problem.

Generally speaking, the practical solution of an inverse problem involves the recovery of interesting sources which possess some physically/chemically meaningful characteristics. In chemistry, excitations might be electro-magnetic radiation, heat, phonons, electrical potential etc. and these induce a response. The response has valuable information and the inverse problem attempts to recover this physical/chemical information.

Fortunately in many chemical studies, the measured properties of the system can be approximated by a linear model at least in some specific range. In spectroscopy, the simplified model can often be achieved by assuming the source has a well defined energy, the effects are additive, other convolution effects are negligible. However, caution must be heeded since the linear model should not be applied everywhere without consideration, for example, the Beer-Lambert Law is only true for limited ranges in the composition space; the exact range of solutions must be determined experimentally. Beyond this range, the direct use of the model based on Beer-Lambert Law will be erroneous.

In perhaps the simplest case, Eq 1.1 can be presented as a very simple instantaneous linear model, which holds for any point/element, and can be represented in the form of a set of linear equations:

$$Y = BX + E \tag{4.2}$$

or

$$y_i = \sum B_{ij} x_j + e_i \tag{4.3}$$

where, E denotes the undesired error and noise, which could be ignored or eliminated in a well-designed experiment. If, in Eq. 4.2, the E term drops, a simplified form: $Y = BX$ is available. This equation coincidently has the same form as the Beer-Lambert Law equation which represents the relationship between concentration of a compound in solution and the spectroscopic absorption of the solution.

The Beer-Lambert law of the absorption in a mixture sample can be regarded as a linear combination of the individual component spectrum with their concentration. The Beer-Lambert Law states that the absorbance, A, of a species at a particular wavenumber of electromagnetic radiation, ν, is proportional to the concentration, C', of the absorbing species and to the optical path length, L, of the electromagnetic radiation through the sample containing the absorbing species where a is the molar absorptivity of the absorbing species (Eq 4.4).

$$A(v) = C'L \cdot a(v) \tag{4.4}$$

$$A(v) = C \cdot a(v) + \varepsilon \tag{4.5}$$

In practice, the Beer-Lambert Law is normally formulated in its bilinear form given by Eq. 4.5 where $C = C'L$. Also in the real world, the measured data are often corrupted by experimental noise and/or error ε (Eq. 4.5). These errors may originate from many sources, i.e., the use of dirty cuvettes, poorly mixed solutions, and instrumental errors. It is worthy to note that in the real data, the bilinear form of Eq. 4.5 is only locally valid. If the pure component spectra a(ν) are non-stationary (they are not always strictly constant), then the solution of Eq 4.5 forcing a(ν) to be constants, leads to both random and systematic error in the error term (Garland, 1997).

4.1.2.2. Self Modeling Curve Resolution Methods

The pure component spectral reconstruction is actually an inverse problem where one wants to determine $a(\nu)$ with the only knowledge of $A(\nu)$, given only the bilinear model (Eq. 4.5). The problem is implicitly ill-posed because one does not know *a priori* how many species are even present.

It is no surprise that there will be many solutions for the above inverse problem mathematically since three kinds of ambiguity related to the solution of the problem, namely, scale ambiguity, rotation ambiguity and order ambiguity exist. Order ambiguity is trivial in spectroscopic analysis, since it is meaningless to judge the order of the components in the recovered procedure. Scale ambiguity also is not so important since it can be circumvented by adding mass balance constraints or some other *a priori* information, such as, a reference component. Alternately, without loss of generality, the spectra can be assumed to have unit norm, which is usually taken as Euclidean norm or L^1-norm according to the specific SMCR method and the scale ambiguity is taken out of the consideration. However, the rotation ambiguity is still the core problem for the SMCR methods. As shown in Eq. 4.6, if C and $a(v)$ are the true solution of the problem, it is easy to find more than one no-singular matrix R, which will also make the product $-CR$, and the product $-R^T a$ satisfied Eq. 4.6. This phenomenon is so-called rotation ambiguity.

$$A(v) = CR \cdot R^T a(v) + \varepsilon \tag{4.6}$$

One group of self-modeling curve resolution methods integrates some type of generic knowledge concerning pure variables in the spectroscopic data in order to break the rotation ambiguity problem and obtain unique resolution. These methods normally exploited the local rank analysis which confines the feasible solution to a desirably small

region. SIMPLISMA (Windig, 1988; Windig et. al., 1990), one popular method for self-modeling, uses close examination of pure variables to help finding suitable rotations to generate solutions. Finding a pure variable that has intensity contribution from only one of the components in the mixture is the basic principle of SIMPLISMA. The apparent limitation of the method is that the solutions are subject to uncertainty unless each species has a signature that is independent of the others at one or more of the wavelength channels in the multivariate data set. Other similar methods require the identification of pure variables include EFA (Maeder , 1987) and WFA (Malinowski, 1992).

To break rotational ambiguity, several constraints have been proposed to narrow the feasible range for solutions. A group of SMCR algorithms seek to break the ambiguity by seeking a best fit of the original data set, in a least-square or weighted least-square sense. Typical examples are Multivariate Curve Resolution and Alternating Least Squares method (MCR-ALS) (Tauler *et al*., 1991, 2001), Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; Paatero, 1997), Iterative Target-Testing Factor Analysis (ITTFA) (Gamperline, 1984) and Alternating Regression (AR) (Karjalainen, 1989). Unfortunately the minimization of error between the regression model and real data while optimizing both *C'* and *a* may not guarantee a unique solution. Also the final solutions of iterative regressions often rely on the quality of initial estimates.

Heuristic evolving latent projections (HELP) (Kvalheim and Liang, 1992; Liang *et al*., 1992; Keller *et al*., 1992; Liang and Kvalheim, 2001) was initially proposed in the field of liquid chromatography with photodiode array detection. It targeted to resolve two-way bilinear multi-component data into spectra and chromatograms of the pure constituents. Since there is no substantial information about the compositions in the system, the utilization of their time series correlations along the chromatographic direction

is important. For the liquid chromatography with diode array detection, during a chromatographic run, data are collected as a matrix indexed by wave-number of the spectra in one direction, and by time from chromatography in another direction. In this case, the additional correlation offers an additional constraint that helps to remove ambiguity in the rotated solution. Tracking of the number of significant principal components/or the rank over the time axis permits an estimate of the number of species that vary over the data set. Proper rotation yields both the spectral response and the chromatogram, and gives estimates of pure responses and relative concentration of analytes. Malinowski discussed this approach in more detail (1991). In the HELP algorithm, the so-called zero-component regions and local rank analysis technique are used.

All these abovementioned algorithms were designed to be applied to any experiment whose outcome is a continuous curve that is a sum of unknown, non-negative, linearly independent functions. However, for a specified system, we can manage to impose suitable constraints for this inverse problem. There are several examples of SMCR methods now available to separate a set of complex spectral curves under certain constraints. Generic knowledge about the bilinear system is proposed to narrow the feasible range for solution. The most often used constraints are a characteristic of the spectra and the concentration profile, such as non-negativity of concentrations and absorptivity estimates, unimodality of concentrations and spectra, selectivity (presence in parts of the experiment of only some of the species), spectra closure (the sum of concentrations of particular species remains constant). Also the constraint of forcing the shapes of the spectra to be as distinct as possible is favored. Besides the aforementioned means, involving some hard-model component, stoichiometric and kinetic characteristics

are also suitable as constraints to the curve resolution, for example, assuming a first order kinetic reaction carried out in the system (Saurina *et al*., 1998) or fitting rate constants of distinct chemical model reactions (Neuhold and Maeder, 2002). In PGSE NMR spectroscopic data analysis, it was assumed that each component decays with an exponential profile (Stilbs *et al*., 1996).

Depending on the *a priori* knowledge and spectral features, some of these constraints can be interactively chosen for specific cases. However, only for specific cases can we manage to impose suitable constraints for this inverse problem. We should note that for a totally blind system, the constraints, such as closure, spectral shape, and stoichiometric and kinetic characteristics are totally unavailable.

4.2. Historical Perspective and Developments of BTEM

The Entropy concept was applied and a new algorithm for self-modeling curve resolution method was developed by Sasaki *et al*. (1983, 1984) They used an optimization objective which focused on the non-negativities of absorptivities and required a minimization of Shannon's entropy. In these seminal papers, the pure spectral reconstruction was obtained by minimizing the entropy of the second derivative of the spectra estimate. Sasaki *et al*. first used entropy minimization for chemical component spectral reconstruction – but the quality was modest and only two-three components could be recovered in their papers.

A number of advances were made by others in Prof Garland's group for pure component reconstruction before the original BTEM algorithm was developed (Zeng and Garland, 1997, 1998; Pan *et al.*, 1999, 2000; Widjaja, 2002; Chew, 2003). The newly developed algorithm of BTEM has been successfully applied to resolution of various

spectroscopic data, such as FTIR, (Chew *et al.*, 2002; Widjaja *et al.*, 2002; Widjaja and Garland, 2002; Chen *et al.*, 2003; Widjaja *et al.*, 2003), Raman (Sin *et al*., 2003; Ong *et al*, 2003), NMR (Widjaja, 2005), and MS (Zhang, 2002).

The BTEM algorithm in this thesis started with the two important premises that pure component spectra have simplest patterns in the data set and that due to non-stationary/non-linear behavior, statistical tests for the number of species present is of limited use. The ideas are fashioned together with a novel spectral band-targeting strategy for one-at-a-time spectral recovery and signal enhancement. Also simplified spectral measures (which differ form the original entropy formulae) have been used recently to expedite computation, particularly for the analysis of very large data sets. However, the name "entropy minimization" has been retained due to the original problem formulation and the fact that the goal remains a search for spectral simplicity. Further discussions about the close relationship between entropy minimization and pattern recognition (Watanabe, 1981), relationship between entropy minimization and the principle of simplicity (Kapur, 1993) can be found elsewhere.

The most attractive and important characteristic of BTEM is its ability to retrieve the extremely weak signals of the pure component spectra from trace-level species which are impossible to retrieve using other existing methods (Li *et al*., 2002, 2003). Some mild non-linearity problems due to shifting band position and changing band shape can be solved by BTEM, which is capable of recovering most of the real chemical signals contained in absorbance data. The primary utility of BTEM comes from (1) the fact that no *a priori* estimate of the number of species present is needed, (2) considerable noise reduction can be obtained and (3) its goal-oriented approach; the user targets a single spectral feature of interest, and the algorithm returns the full-range deconvoluted spectrum.

4.3. Entropy Minimization Method: BTEM

The first procedure in BTEM is performing singular value decomposition on the set of spectroscopic data $A_{q\times\nu}$ (spectroscopic data matrix with q spectra and ν channels of wavenumbers).

$$A_{q\times\nu} = U_{q\times q}\,\Sigma_{q\times\nu}\,V^{T}_{\nu\times\nu} \tag{4.7}$$

Assuming that a diligent experimental design has been carried out and that a sufficiently large number of spectra were acquired, then the first few vectors in V^T (perhaps as many as 50-100) should have localized signals, and the remaining vectors in V^T are more-or-less random noise. This means that most of the physically meaningful information is captured in the first j vectors.

The experimentalist then identifies local features in the first 10-20 right singular vectors which are "targeted" one-at-a-time by BTEM. This strategy would ensure that each feature is retained in the final reconstruction by BTEM. Not only the feature is retained, the entire function $f(x)$ associated with the feature is recovered – without any *a priori* knowledge. The primary numerical manipulation is associated with the transforming of the abstract right singular vectors, V^T into pure component spectral estimates, $\hat{a}$ by proper linear combination of these basis vectors.[i]

$$\hat{a}_{1\times\nu} = t_{1\times z}V^{T}_{z\times\nu} \tag{4.8}$$

The transformation is governed by an optimization for the elements of t. In this optimization, the global minimum value of the proposed objective function corresponds to the final estimate of the pure component spectral estimate, $\hat{a}$. GA or SA is used

[i] A forerunner of BTEM was MESS (Minimization of Entropy with Spectral dis-Similarity) (Chen *et al*., 2003). In this problem, a so called *square problem* is solved where only s basis vectors are taken and a T_{sxs} is optimized simultaneously. There are also other SMCR approaches where a square transformation must be solved.

throughout this thesis. Statistically speaking, entropy can be expressed as disorder, therefore minimizing the entropy of the transformed eigenvectors means maximizing the spectrum simplicity. In other words, minimizing entropy localizes the spectral information around the major bands and maximizes the number of zero elements in the spectrum (Widjaja, 2002). Therefore, the objective function consists of one entropy term *H* which is defined by Eq. 4.10 and one penalty term *P*.

$$F_{obj} = H + P \tag{4.9}$$

$$H = -\sum_{v} h_v \ln h_v \tag{4.10}$$

In Eq. 4.10, h_v is a discrete probability distribution function described as the derivative amplitude of the estimated pure spectrum in a L^1 norm (Eq. 4.11).

$$h_v = \frac{\left|\frac{d^m \hat{a}_v}{dv^m}\right|}{\sum_{v}\left|\frac{d^m \hat{a}_v}{dv^m}\right|} \tag{4.11}$$

where the superscript *m* denotes the degree of spectral differentiation. A penalty *P* is imposed for guaranteeing the non-negativity of spectra and concentrations.

$$P(\hat{a}_{1\times v}, \hat{C}_{q\times 1}) = \gamma_a F_1(\hat{a}_{1\times v}) + \gamma_c F_2(\hat{C}_{q\times 1}) \tag{4.12}$$

where

$$F_1(\hat{a}_{1\times v}) = \sum(\hat{a}_{1\times v})^2 \quad \forall \quad \hat{a}_{1\times v} < 0 \tag{4.13}$$

$$F_2(\hat{C}_{q\times 1}) = \sum(\hat{C}_{q\times 1})^2 \quad \forall \quad \hat{C}_{q\times 1} < 0 \tag{4.14}$$

γ_a and γ_c are penalty factors set to the penalty function, which provide soft constraints when tuned accordingly. It should be noted that the maximum of absorbance in the range/window of the targeted band is normalized to be unity before the entropy calculation.

The choice of band target windows is pragmatic. Since if the bands move during the whole set of mixture data (non-stationarity),i.e. collected during reaction, we need to have a big enough band-target window to make sure that the band is within this window. Thus for complex infrared data, with a lot of overlap, but the bands move ±1, then the window is usually about 2 wavenumbers wide. If there are no bands maxima are close together in this targeted window, then a very big window can be used, perhaps 5 or even greater.

After extensive search for all possible spectra by targeting the interested features, the final estimate of the pure component spectra, $\hat{a}_{s\times v}$, can be used for calculating the corresponding concentration expectation, $\hat{C}_{qxs}$, with Eq. 4.15, where the superscript "+" denotes the Moore-Penrose pseudo inverse[ii].

$$\hat{C}_{q\times s} = A \cdot \left(\hat{a}_{s\times v}\right)^{+} \tag{4.15}$$

With the powerful method BTEM, not only the pure component solute spectra can be resolved, but subsequently, further algebraic system identification of complex reaction systems can be carried out. Further details of BTEM are available in Widjaja's thesis (2002). Finally, it should be noted that in many applications, the penalty function in $\hat{C}_{q\times 1}$ (Eq 4.14) can be safely omitted. Sufficiently good spectral estimates can be obtained with a penalty function for non-negativity of spectra alone.

For completeness, it is necessary to note that Banerjee (1991) suggested the similarity of using derivative minimization to produce pure component spectra, but an algorithm was not suggested. Also Volkov (1996) tried an approach by adding some

[ii] The classical inverse is restricted to square and non-singular matrices. Moore-Penrose pseudo inverse is the best approximate solution to the problem of inverting a rectangular matrix.

penalty terms including the similarity term, the spectral curve length term and the overall curvature term of trial component spectra vectors. This approach utilized a square method[iii] and therefore could not account for non-linear spectral contributions. Karjalainen (1990) tried to use entropy to be a complement to the AR (alternating regression) algorithm in the GC/MS data analysis.

4.3.1. Discussion

There are many important differences between BTEM and the other SMCR techniques. One important difference between BTEM with other SMCR methods is the strategy of solving one spectrum a time. Another importance difference is the idea of targeting the interested band. By exhaustive search, all the spectra would be obtained with the all interested feature retained in the result.

This approach is obviously simple and convenient, especially, in those cases where only one interesting solution is needed, instead of the whole spectral system. Band-target strategy would make it easy to target the interesting band out without solving the whole system first. The BTEM algorithm offers many advantages over traditional SMCR methods by solving for several local solutions (one spectrum at one run) instead of solving the perplexing global solution (obtaining all species' spectra in one run),

In practical terms, this approach simplifies/lightens the burden of optimization of the nonlinear objective function, especially, when we deal with the large scale data sets, which possess a large number of measurements or where each measurement consists of a large number of variables. As we know, the decrease in the number of variables will dramatically decrease complexity, and this would expedite the optimization problem. On

[iii] see footnote i

the other hand, if the number of species increases rapidly, a supremely arduous optimization exists due to the big number of variables involved.

Another distinction of BTEM is the utilization of right singular vectors by making use of the SVD technique. As discussed in chapter 3 section 3.3.3, SVD is a robust and reliable PCA technique. One purpose of using of the SVD technique, is to project the data from a higher dimensional space into a lower dimensional space, therefore concentrating onto a few underlying latent variables which capture most of the information of the data. When the number of samples is much bigger than the number of sources, it is fairly reasonable to do analysis with just a few latent variables instead of the huge set of original data. A more important issue is that the data are made "cleaner" by the SVD procedure by producing the reduced latent factors and discarding the noisy factors. And this makes it possible for BTEM to tolerate the minor nonlinearities inside the original spectral data and find the minor components.

The band-targeting technique makes deliberate and explicit use of the right singular vectors by targeting each feature inside. Such local features are normally regarded as having little or no physical meaning at all. By targeting the features in the right singular vectors, these features are remained in the final reconstruction. Band-targeting technique integrates the user's interest with a goal-oriented approach. It also takes advantage of spectroscopic knowledge of the chemist i.e. in example 4.4.1.c (infra vida) the unusual event of ketone formation during hydroformylation is readily confirmed by targeting the localized feature at circa 1725 cm^{-1} which implies a "ketonic" type carbonyl vibration.

4.4. Applications of BTEM to Real Chemical Reaction Systems

A number of real spectroscopic data sets from Prof. Garland's group, NUS, as well as from other laboratories at ICES had been investigated.

4.4.1. The Data Sets from the Hydroformylation Reactions of Alkenes

The hydroformylation reaction of alkenes is one of the most important homogeneous catalytic processes worldwide (Cornils and Herrmann, 1996; Van Leeuwen, 2000). However, due to the complexity of the catalytic mechanisms, the systems are still the object of many studies. Detailed knowledge of the catalytic mechanisms, in particular the possibility of identification of intermediates and obtaining their concentrations, is an important step towards creating good kinetic models for reaction engineering purposes. (Garland, 1989, 1991a, 1991b, Feng and Garland, 1999)

(a) Hydroformylation of COT (cyclooctene) with $Co_2(CO)_8$.

This data set was generated by L. Susithra in 2000 (before BTEM algorithm was developed). These mixture data were measured under isobaric and isothermal conditions using *in situ* high-pressure infrared spectroscopy. Further analysis of the data set was needed. Without spectral preconditioning, estimated pure component spectra of $Co_2(CO)_8$, $Co_4(CO)_{12}$, $HCo(CO)_4$, RCHO, and COT from the reaction mixture data were obtained by the BTEM (some of result are shown in Figure 4.1). Satisfyingly, some of these reconstructed results are highly consistent with their corresponding *in situ* infrared pure spectrum dissolved in solvent after the subtraction of background signals. The BTEM analysis and further analysis were consistent with the preliminary analysis by Susithra

(1999). A kinetics / reaction engineering manuscript is in preparation which will include the BTEM results.

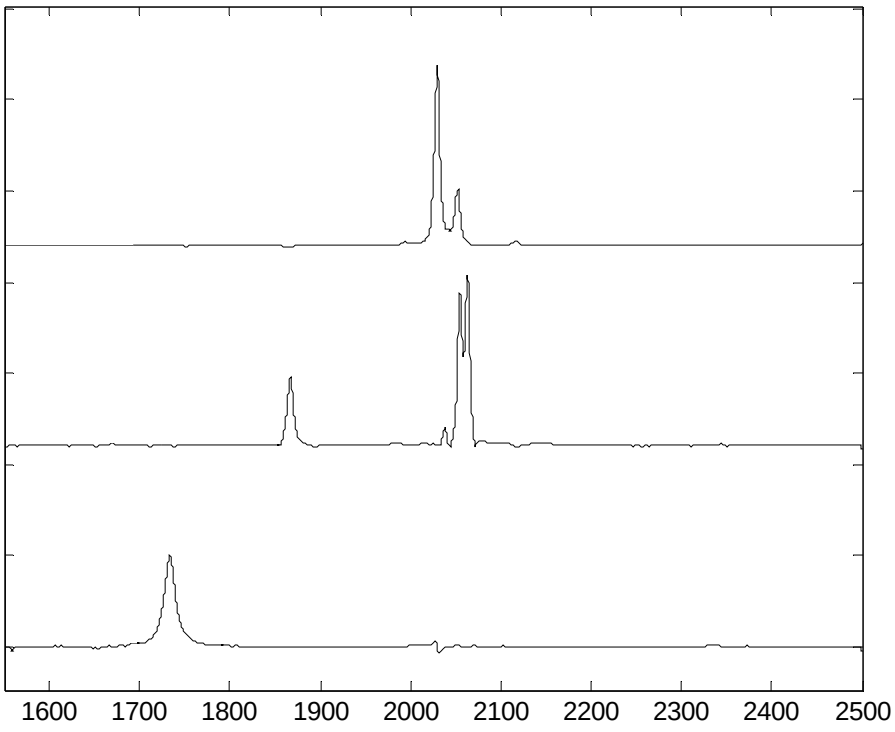


Figure 4.1. The estimated infrared spectra of $HCo(CO)_4$, $Co_4(CO)_{12}$ and RCHO.

(b) Hydroformylation of (COT) cyclooctene with $Rh_4(CO)_{12}$

COT has potentially undesirable reactivity patterns with $Rh_4(CO)_{12}$. Some new pure component spectra were reconstructed from the non-preconditioned data – spectra not obtained in the previous analysis using routine subtracting method (Liu, 1999). Not only the pure component spectra of background water, hexane, carbon dioxide, dissolved CO and the reagent, namely cyclooctene (COT), the product aldehyde ($C_8H_{15}CHO$), the catalyst precursor $Rh_4(\sigma\text{-}CO)_9(\mu\text{-}CO)_3$ and intermediate $RCORh(CO)_4$ were recovered. An unforeseen finding was a species with a spectrum with vibrations centered at ca. 1707.8

cm^{-1}, 1759.4 cm^{-1} which is consistent with the formation of a ketone as a byproduct (shown in Figure 4.2). The organometallic intermediate, $HRh(CO)_4$ and $Rh_4(\sigma\text{-}CO)_{12}$ were reconstructed as well. This analysis helps to understand the reaction system and a detailed kinetics/ reaction engineering manuscript was submitted. This data set was generated by Liu Guowei in 2000 (before BTEM algorithm was developed). The results have been published (Liu *et al*., 2006). A reprint can be found in Appendix A.

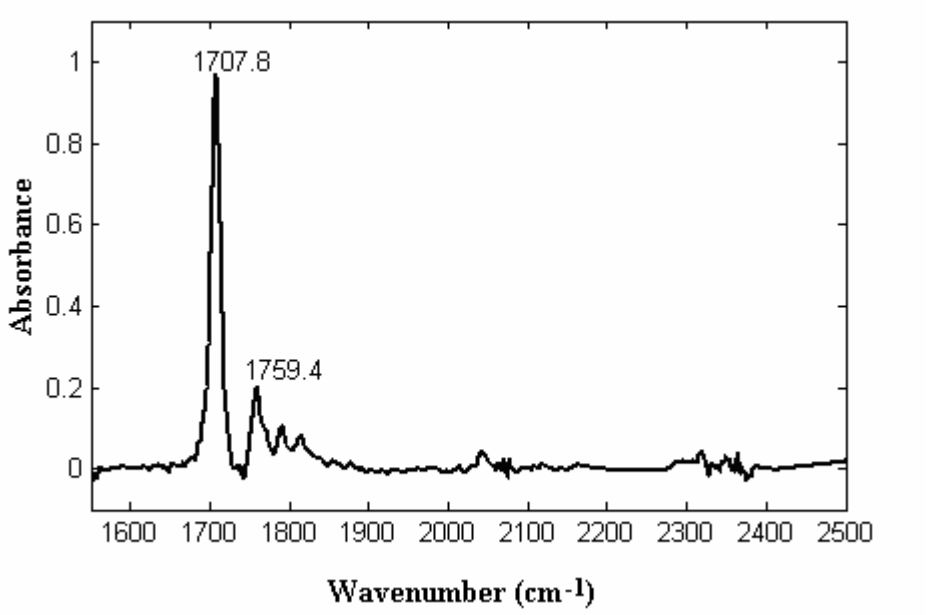


Figure 4.2. The estimated infrared spectrum of byproduct – ketone.

(c) Homogeneous hydroformylation of ethylene catalyzed by $Rh_4(CO)_{12}$.

Rhodium by far is the most active metal in the homogeneous catalyzed hydroformylation reaction and previous studies show that spectroscopic analysis of the hydroformylation of ethylene is extremely difficult to understand (Liu and Garland, 2000). The purpose of this study was to re-examine the unmodified rhodium catalyzed hydroformylation of ethylene with the BTEM. This experiment was carried out by Li Chuanzhao in 2002. Without preconditioning, the data was decomposed by SVD and

reconstructed with BTEM. Although there was severe spectral overlapping in the mixture spectra, some new pure component spectra were identified by BTEM.

The numerical results showed that the expected organometallic species $RCORh(CO)_4$ and $Rh_6(CO)_{16}$ are reconstructed well. Two unexpected spectrum were found also: one spectrum with vibrations centered at ca.1695.2, 2017.6, 2040.2 and 2090.4cm^{-1}, another mainly centered at ca.1641.6 and 1695.4 cm^{-1}.These new species are identified as $CH_3CH_2CORh(CO)_3(C_2H_4)$ and ketone (Figure 4.3). This is an important finding since no organometallic species from this class have ever been observed before. The results have already been presented at a conference and published in *Organometallics* (Li *et al*., 2004a). A reprint can be found in Appendix B.

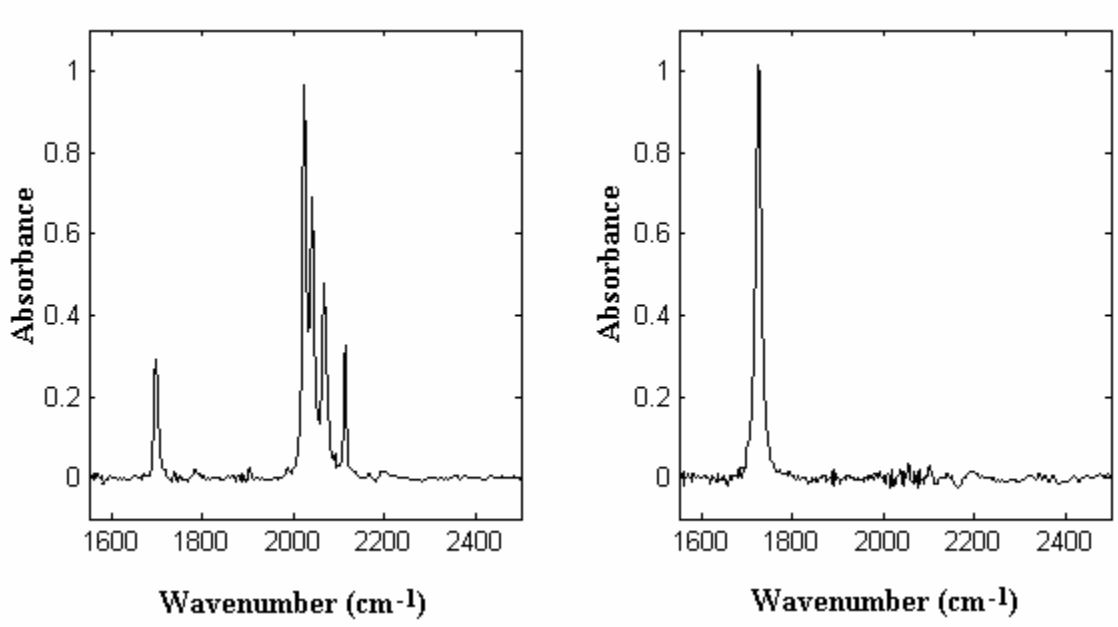


Figure 4.3. The estimated infrared spectra of new species: $CH_3CH_2CORh(CO)_3(C_2H_4)$ (left) and ketone (right).

(d) Hydroformylation of 4-Vinyl Pyridine catalyzed by $Rh_4(CO)_{12}$.

Many of the primary classes of rhodium organometallic carbonyl species postulated to exist have now been found. One class, namely alkyl rhodium tetracarbonyls

$RRh(CO)_4$ have not been identified yet. Very recently in Italy, it was shown that 4-vinyl pyridine undergoes considerable hydrogenation to 4-ethyl pyridine under hydroformylation conditions (Lazzaroni *et al*., 2002). An elaborate experimental design was performed and a series of *in-situ* spectra were collected by Li Chuanzhao in our group. I performed singular value decomposition of the spectra which is followed by BTEM analysis. Several organometallic species were recovered, but the expected acyls and alkyls could not be found. The reaction undergoes very fast deactivation (loss of precursor from solution), and we are not able to model the system kinetics. One possible explanation is that this reaction could be heterogeneously catalyzed on metallic rhodium formed by decomposition of the $Rh_4(CO)_{12}$ precursor. Rhodium metal is known to be able to perform room temperature hydrogenations – it is very active.

(e) A new species involved in hydroformylation $RhRe(CO)_9$.

The group found that solutions of rhodium and rhenium carbonyls are much more active that solutions of rhodium carbonyls alone. Li Chuanzhao performed an experiment where he contacted rhodium and rhenium carbonyls under hydrogen and carbon monoxide in n-hexane. In-situ spectra were taken using FTIR spectroscopy. The data set was deconvoluted using BTEM. A new pattern was found. Further vibrational group theoretic and other analysis indicated that it belonged to a previously unknown organometallic species $RhRe(CO)_9$. These results appeared in *Organometallics* (Li, *et al*., 2004b). A reprint of this paper is given in Appendix C

4.4.2. NMR Data Sets

Entropy minimization in an alternate form was successfully applied to PGSE NMR data by Widjaja (2005). Given this encouraging result, we then tried to apply BTEM to resolve the pure component spectra of more common NMR data sets.

A set of experimental ^{1}H NMR data (10 spectra) were collected from mixtures of 4 chemical components (Prof. Leong Weng Kee's group in Chemistry Department, NUS). The untreated data were analyzed by BTEM and the results were poor. The reconstructions are quite far away from true pure spectra. One reason for this result is the severe non-linearity found in NMR (non-stationary signals - change in absorption frequency). This shifting has been a serious impediment to the use of chemometric methods on NMR data.

A numerical experiment was carried out on a small region (401 channels) of the spectra instead of the whole region. A small range of NMR data with mild overlapping was pre-treated in order to re-align the peaks in the same data channels. BTEM was preformed after the singular value decomposition of the aligned data. Two major components were obtained and compared with the reference spectra (Figure 4.4). The reconstructed spectrum (1) is almost the same with the reference spectrum. The left side of the second constructed spectrum was noisy and the mismatch with the reference may be due to in-exact pretreatment of the data. This experiment shows that if a proper and robust alignment method could be developed, BTEM could be successfully applied to a wide range of NMR data. Further investigations about NMR data and the development of NMR data analysis are illustrated in section 4.5.

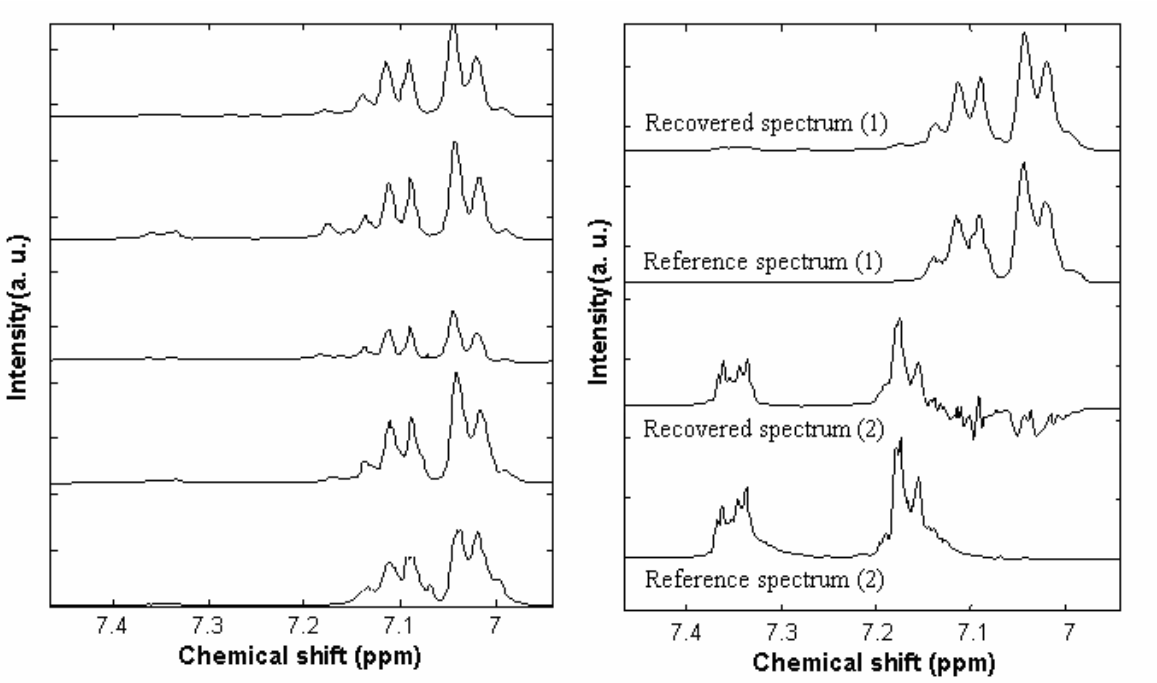


Figure 4.4. The first five ^{1}H-NMR mixture spectra (left) and resolved pure component and their references (right).

4.4.3. XRD Data Set

Before this analysis, all of the applications of BTEM were limited to liquid phase analysis. In principle, if a linear model still holds, there is no significant difference between a liquid phase analysis and a solid phase analysis. One obvious common type of analysis in material science is powder XRD (X-ray diffraction). In this technique, powdered crystalline materials are measured, and the intensities of the X-rays are measured as a function of the angle of diffraction. Such information provides information regarding the phase, crystallinity, crystal structure etc.

Pure component spectral reconstruction from a set of XRD data (12 spectra) were preformed with BTEM. The analysis provided the right prediction that 5 components were present. Outstanding pure component XRD patterns were obtained for all 5 components ($Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$, and PbO) present as indicated by comparison

with authentic references. These results have implications for a large variety of intrinsically inseparable multi-component mixtures encountered in material science research. These include un-reactive as well as reactive systems, and *ex-situ* as well as *in-situ* studies, involving organic, inorganic and even metallic/alloy components. Initial tests suggest that BTEM may be well suited for recovering even trace component diffraction patterns present and hence greatly aiding material characterization. Dr. Fethi Kooli at ICES conducted the experiment in 2003 and I performed numerical data analysis. Full details of the study of XRD data can be found in Journal: *Analytica Chimica Acta* (Guo *et al*., 2004). A reprint is provided in Appendix D

4.4.4. Entropy Minimization and Sound Source Separation Application

4.4.4.1. Introduction

The classical problem associated with Blind Source Separation (BSS) is the so-called “Cocktail Party Problem” which has long been of interest in electrical engineering (Haykin, 2003), acoustical engineering (Bronkhorst, 2000) and cognitive sciences (Conway *et al*., 2001; Gaetz *et al*., 1998). It describes a typical scene at a party where several sound sources are located in the room: voice of woman, man, children, foreign language, music, and even noises from surroundings. Amazingly we can focus on the voice from one specific person among a mixture of conversations and background noises. This elaborate auditory source separation problem is easily solved by the human brain, but proved to be a tough and complex problem in digital signal processing. In the BSS study for sound separation, the only input for analysis is the recorded signal from several microphones in a room and the goal is to identify and extract the voices/sounds of each individual speaker (sources) from the recorded signals.

A number of approaches have been proposed for the solution of the BSS problem, typically, for the sound source separation problem (Plumbley *et al*., 2002; Makeig *et al*., 1997). And independent components analysis (ICA), a computational method for separating a multivariate signal into components with the assumption of the mutual statistical independence of source signals has a wide practical application (Cardoso, 1998; Vigario, 1997).

The above-mentioned situation is an excellent analogue to the mixture analysis of chemical signals which consist of different constituents. In this section, a simulation of mixture sounds was studied. The objective of the analysis is to find the original acoustic signals from the mixture signals via Entropy-Minimization method.

4.4.4.2. Experiment Section

A simulated sound data set (5-by-10000) was generated by mixing three digitized sound waves (with 10000 channels each) with random coefficient matrix (5 by 3) created by MATLAB function "rand"[iv]. The data set is shown in Figure 4.5. The pure sound waves are also shown in the Figure 4.6. The mixture sound waves are rather similar to each other since there are no highly localized features that stand out – there is a certain lack of a characteristic structure.

[iv] The "rand" function generates random numbers uniformly distributed in the range from zero to one.

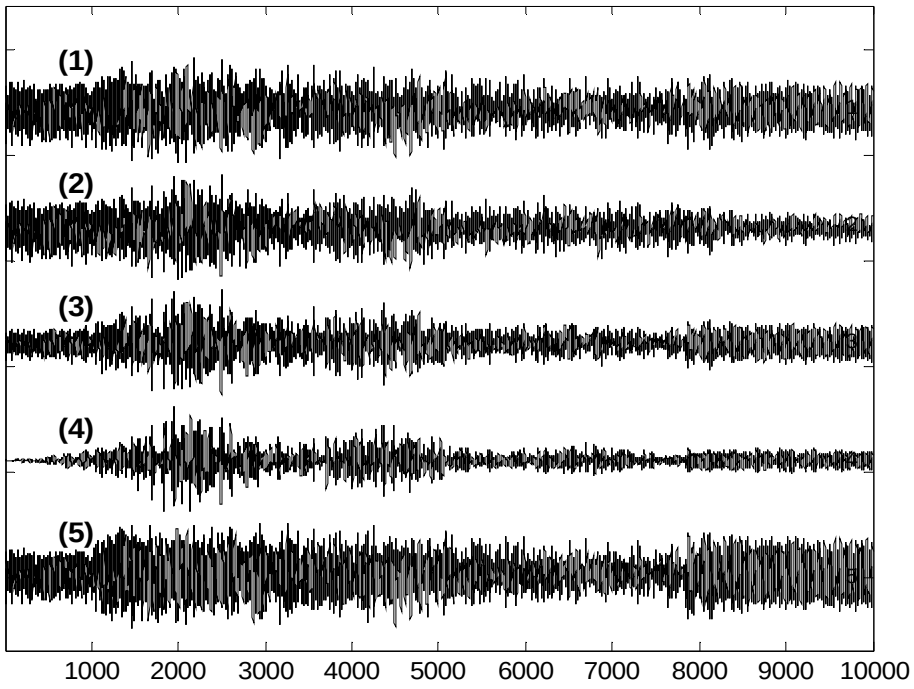


Figure 4.5. The sound waves of the five experimental mixtures (shown in channels).

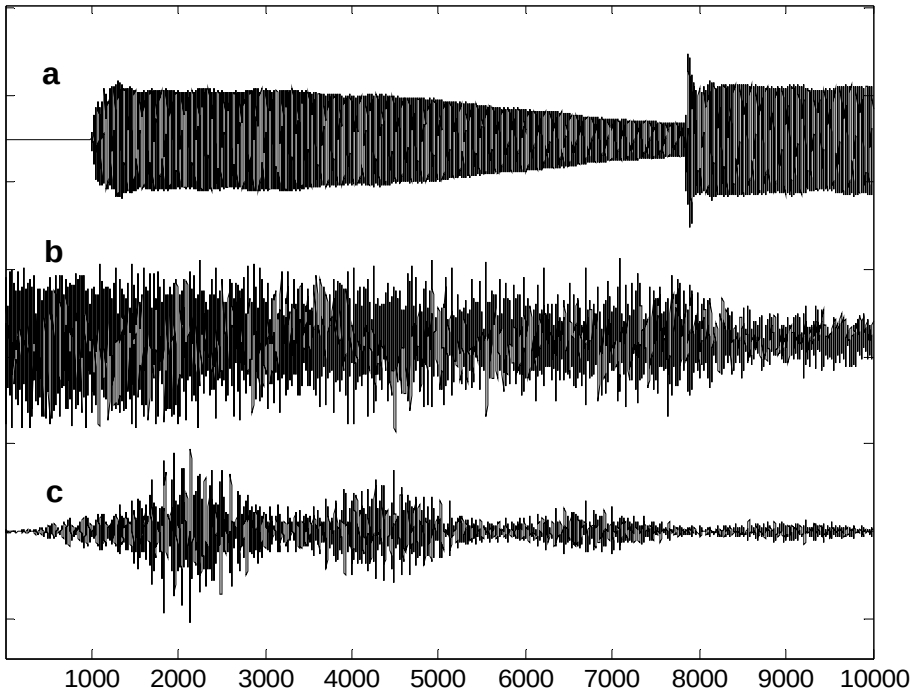


Figure 4.6. The sound waves of three pure sources (shown in channels).

4.4.4.3. Entropy Minimization with Dissimilarity Constraints

Singular value decomposition (SVD) of the spectra was performed. From Figure 4.7, it is apparent that the fourth V^T vector is close to the homoscedastic noise where the noise is almost constant in magnitude between variables. The real diagonal elements of the mixture matrix, called the singular values, are ordered in descent: 21.21, 9.05, 5.72, 0.00, and 0.00. Since no noise was added to the data, almost all the information concentrates in the first three vectors; no useful information would be imbedded in the last two vectors. Therefore it makes sense that only the first three V^T vectors were used in the transformation or reconstruction.

Quite different from the normal chemical spectrum which is a sum of a series of Gaussian or Lorentzian band shapes, the signals of a sound wave are sinusoidal waveforms without any obviously different features. In this case, it is hard to search for a band to target in BTEM. So instead of a band targeting method, entropy minimization is used without targeting. The resolution of the pure sound source still can be achieved by solving Eq. 4.9. The objective function F_{obj} includes the entropy term H along with a newly defined penalty function P' (infra vida).

It is apparent that without a similarity criterion, only one sound will be resolved. So the first step is to recover the first pure source sound, where P' is set to zero, and the global entropy minimum is sought. In each subsequent search, the penalty function is formulated such that only sources dissimilar to all previous results are admissible. This effect can be achieved by adding a dissimilarity penalty function, P', to the reconstruction objective function. Four types of spectral dissimilarity functions were investigated, namely angle function, Euclidean inner product, determinant of covariance matrix, and

condition number(Widjaja, 2002). In this study, the Euclidean inner product was used due to its simplicity. The Euclidean inner product between two *normalized* pure component spectra, *a* and *b*, is calculated as follows.

$$\langle a,b \rangle = \sum_{i=1}^{v} a_i . b_i \tag{4.16}$$

If vectors *a* and *b* are similar, the Euclidean inner product will produce a large value and tend to be unity when they are identical. In practice, the vectors *a* and *b* would be the estimated vector during the current optimization and a previously determined source vector(s) respectively. The penalty function prevents any identical reconstruction from occurring in subsequent optimizations which will help to produce a set of distinct or unique pure sounds by maximizing the dissimilarities.

Two objective functions were tested. In the first trial, the first derivative entropy and dissimilarity penalty functions are used, but in the second trial, second derivative and integration of area under the curve are the main part of the objective function. The dissimilarity term is added as a penalty. Both objective functions work with almost identical results. Here, only reconstruction results using the entropy objective function are presented in Figure 4.8. We notice that the amplitudes, sign and the ordering of the signals may not be the same as the original sources. But by adjusting the scale, we can “hear” that they are the same as the original sources.

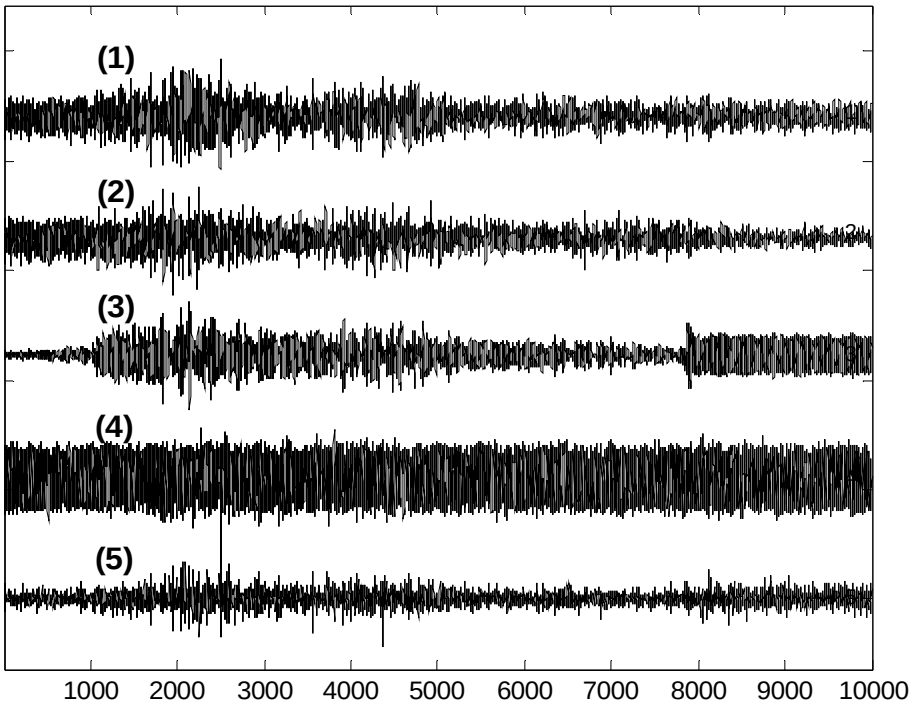


Figure 4.7. Plot of the five right singular vectors obtained from the SVD of the mixture sounds. The last two vectors contain primarily noises.

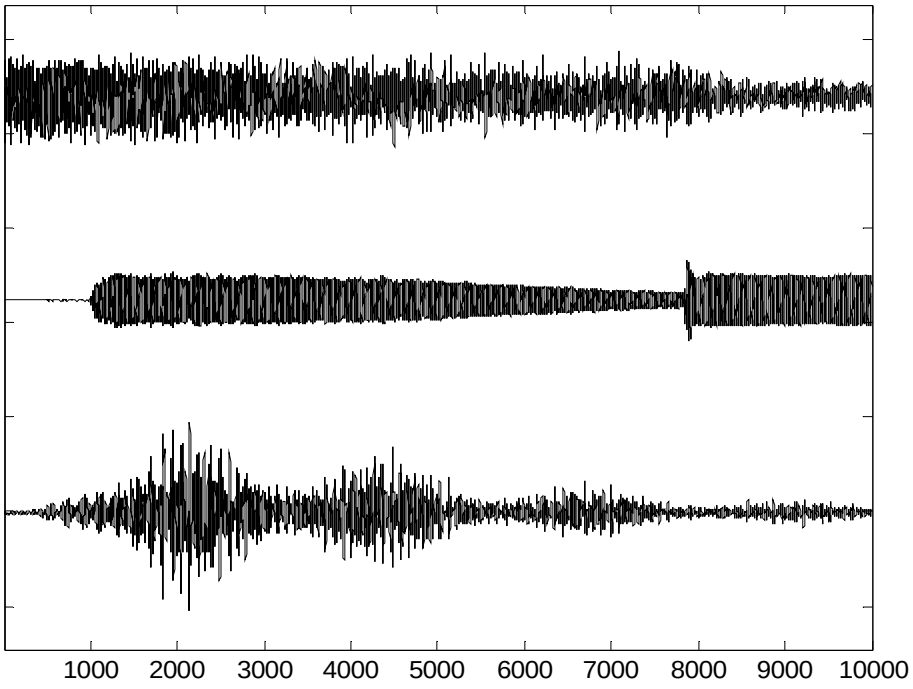


Figure 4.8. The reconstruction results using the entropy objective function.

4.4.4.4. Fourier Analysis and Band-Targeting Entropy Minimization

It is apparent that one big difference between acoustics and spectroscopy is that the former is a time series. As we know, sound is vibration that propagates through air. Although there is no obvious feature such as a band can be used for identification in the oscillations of waves, our brains sort the audio signals by frequencies. In the same spirit, Fourier transform would be a suitable tool to transform the acoustics from time domain to frequency domain. With this idea, the mixture data are transformed into a new set of data via Fourier transformation. BTEM is then implemented afterward.

In Figure 4.9, the FT result of the five sound mixtures is shown. In the FT domain there are not so many sinusoidal signals appearing. Five right singular vectors are shown in Figure 4.10. The last two V^T vectors are still not physically important. Only the first three significant V^T vectors are used in the reconstructions. We notice that most of band features are located on the right side of the plots.

A close examination finds that there are a few bands around index number 8000 to 9500. After an exhaustive search using all the feature bands, only three patterns are obtained by the BTEM approach. They resulted from targeting bands: 8605 ± 2, 8804 ± 2 and 9049 ± 2. In Figure 4.11 the extrema of the targeted bands are labeled. The inversed Fourier transformation of the reconstruction results are shown in Figure 4.12.

It is possible that other transforms such as the wavelet transform, could work as well or perhaps a little better in such applications. Since acoustic data analysis is not the primary aim of my thesis, other transformations were not further pursued.

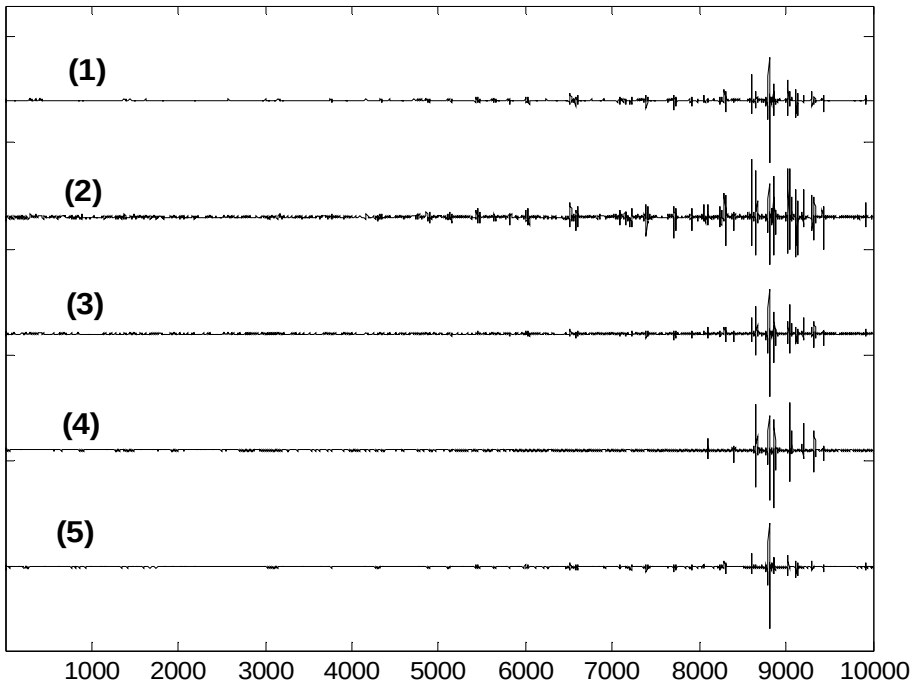


Figure 4.9. The Fourier transformation result of the five mixture sound waves.

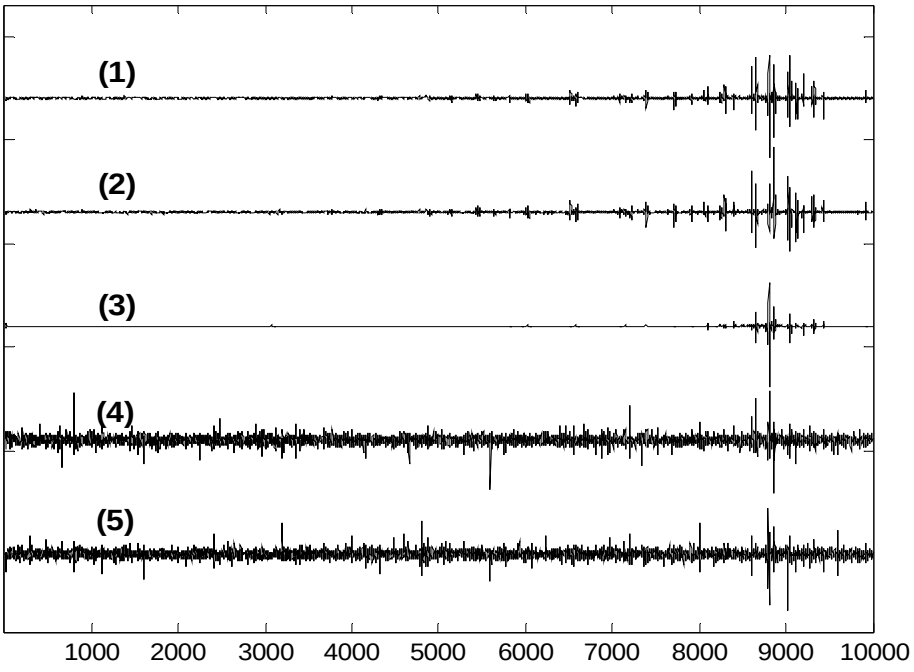


Figure 4.10. Plot of the five right singular vectors of V^T obtained from the SVD of the Fourier transformed mixture sounds. The last two vectors contain primarily noise.

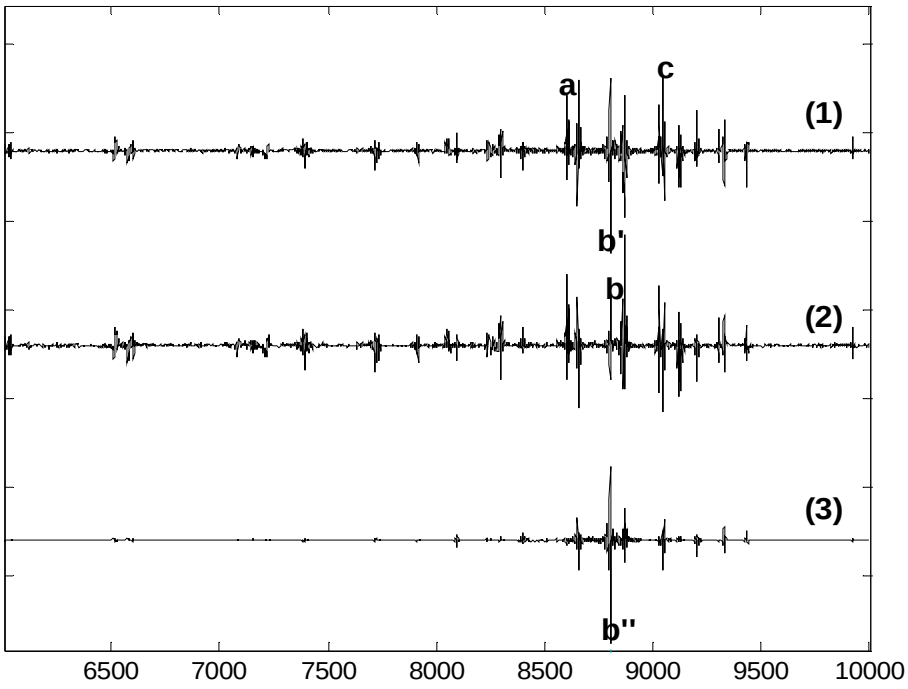


Figure 4.11. Plot of the first three right singular vectors of V^T obtained from the SVD of the Fourier transformed mixture sounds. Letters a-c indicate different peaks subsequently targeted by BTEM. Letter b, b' and b'' indicate the same peaks appear in different V^T vectors.

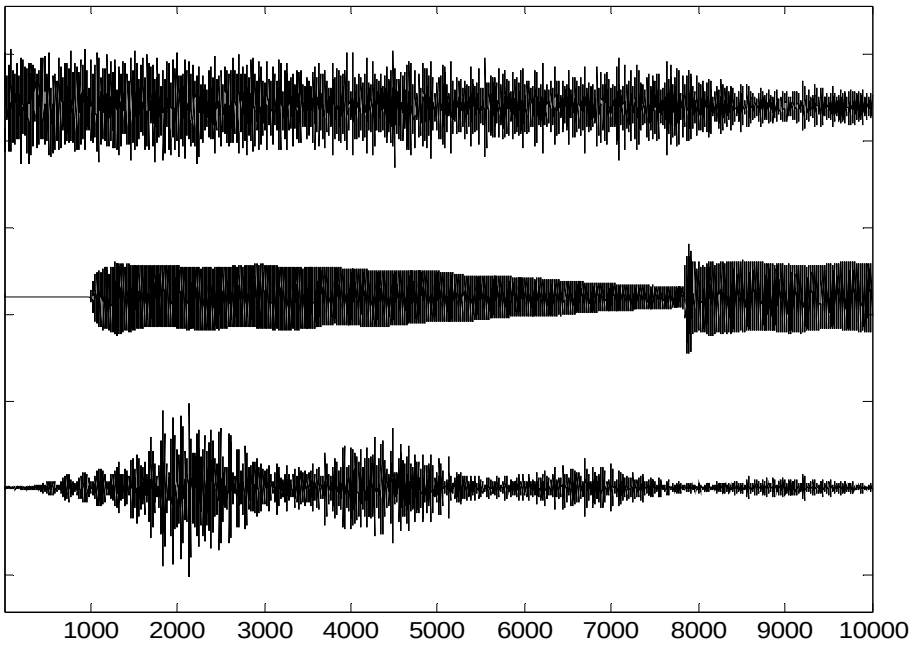


Figure 4.12. Reconstruction result of three sound patterns by BTEM and Fourier analysis.

4.4.4.5. Discussion

1. Entropy minimization: It is important to note that three different strategies can be attempted to obtain the curve resolution, namely, BTEM, Entropy Minimization with dissimilarity constraints (section 4.4.4.3.) and MESS(Chen *et al*., 2003a). It is obvious that BTEM is characterized by its band-targeting technique. The discrepancy between the Entropy Minimization with dissimilarity constraints and MESS lies in the fact the MESS is used for simultaneous resolution of *s* species by rotating *s* basis right singular vectors; meanwhile, the former is still one-at-a-time approach.

In this simple and noiseless case, since there are only three components inside, we can specify this fact, and with this knowledge we can apply MESS. However, in most real data, there are more severe nonlinearities imbedded. Hence, in most real situations, it is quite troublesome to predict the number of components precisely.

2. Two object functions (1)2^{nd} derivative and area and (2) entropy based on 1^{st} derivative were implemented in section 4.4.4.3. The qualities of the reconstructions were outstanding, but it was observed that it is more practical to use objective function (1) than entropy. The main reason lies in the fact that the response surface of the entropy function is fairly complex. From Table 4.1, we can see the objective function value of each pure component. It is apparent that depending on the measure used, the sources are either very similar to one another or very different from one another. Since the sources are much different from one another when using the 2^{nd} derivative method, this may explain in part why the 2^{nd} derivative method provided faster convergence to the correct solutions.

Table 4.1. The values of the two types of objective functions. The variation between 2nd derivative values of different sources is much larger than their entropy value.

Sound waves	**Original sources**		**Recovered result with similarity penalty**	
in Fig. 4.5.	2nd derivative	1st derivative entropy	2nd derivative	1st derivative entropy
1st	5.6195e+002	8.9145e+000	5.6419e+002	8.9188e+000
2nd	1.7549e+003	8.8337e+000	1.7634e+003	8.8354e+000
3rd	1.8640e+002	8.6472e+000	1.8903e+002	8.6641e+000

3. It is known that FT is a linear transform. The bilinear model is not affected by the FT operation. Since FT is a reversible procedure, the resolved pattern will be extracted from the transformed data and easily be converted back into the time domain via the inverse FT. FT converts the complex time domain signal into the simple and explicit spectrum in frequency domain, which we are accustomed to seeing in FT-IR and NMR spectroscopy. Band-targeting is easier to be implemented with FT data in this case since localized signals arise in the frequency domain.

4.5. Application to 1D Nuclear Magnetic Resonance Spectroscopic Data

NMR spectroscopy is an absorption spectroscopy involving the absorption of radio frequency electromagnetic waves and plays an important role in the structural determination of a wide variety of organic and inorganic species. As an unrivaled tool to be used to analyze complex systems, the NMR technique is now widely used in various areas such as physics, chemistry, biology, material science, etc. Also as a complementary tool to infrared spectroscopy, NMR has been employed in the study of catalysis. Both qualitative and quantitatively information on complex reaction systems can be obtained. Routinely ^{1}H, ^{13}C, ^{19}F, ^{31}P and ^{29}Si-NMR are employed. Similar to FT-IR, Raman and

other spectroscopes, the initial basic assumption is that NMR spectra obey a bilinear form, where a generalized Beer-Lambert law holds.

A method based on Minimization of Entropy (MESS) was successfully applied to PGSE NMR data by Widjaja (2005). In his study, by specifying *a priori* that there are 3 species present, all the pure component spectra could be extracted. PGSE is quite a special technique, since the concentration of analytes does not change but one nevertheless obtains a series of spectra with different spectral contribution from each component. Since the analyte concentrations do not change, the pure component spectra in each spectrum are very similar. Non-stationary effects are dramatically reduced. So PGSE experiments are more well-posed than typical NMR experiments.

In Widjaja's study, MESS outperformed DECRA (direct exponential curve resolution algorithm) (Windig, 1997; Windig *et al*., 1997, 1999a, 1999b, 1999c) and PMF (positive matrix factorization) (Xie *et al*., 1998). To my knowledge, compare with vibration spectroscopy, relatively little effort has been invested in the NMR data analysis with multivariate data technique (Antalek and Windig, 1996; Antalek, 2002; Stilbs *et al*., 1996; Schulze and Stilbs, 1993; Dyrby *et al*., 2005). Again, the reason is probably the very bad non-stationary effects including phase, lineshape and line position. Below, BTEM is applied to resolve pure component spectra of more typical and more commonly encountered NMR data sets.

In this section, a series of non-reactive mixture NMR data and a reactive ^{13}C NMR data set are analyzed. All of these data were measured in collaboration with a Bruker scientist, Peter Sprenger, at the Bruker headquarters in Switzerland (Bruker Biospin AG., Zurich) in July 2004.

4.5.1. Study of 1D NMR Mixture Data with Four Chemical Components

4.5.1.1. Experimental: Materials and Sample Preparation

All $^{1}H/^{13}C/^{19}F/^{31}P$ NMR spectra were recorded on a Bruker AVANCE 400 spectrometer (400 MHz) in the FT mode at 295K. The instrument was equipped with 5 mm z-gradient $^{1}H/^{13}C/^{19}F/^{31}P$ quad-nucleus probe.

Four chemical species were used in this non-reactive experiment. These were (a): chloroform-D ($CDCl_3$, deuterochloroform) (99.96 atom % D, Aldrich), (b): 2,5-dimethyl-2,4-hexadiene (96%, Aldrich), (c): tris(pentafluorophenyl)phosphine (97%, Aldrich) ; (d): ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate (97%, Aldrich) . And Table 4.2 shows that not all compounds contained all elements.

Table 4.2. The elements contained in (a), chloroform-D, (b), 2,5-dimethyl-2,4-hexadiene, (c), tris(pentafluorophenyl)phosphine and (d) ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate.

Chemical	Element			
	H	**C**	**P**	**F**
a		o		
b	o	o		
c		o	o	o
d	o	o	o	o

Ten solutions containing the four chemical components were also prepared according to Table 4.3. Reference solutions were prepared by dissolving each pure component in the solvent (chloroform-D) individually (from number 11 to 14 in Table 4.3).

Table 4.3 Composition of chloroform-D (a), 2,5-dimethyl-2,4-hexadiene (b), tris(pentafluorophenyl)phosphine (c) and ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate (d) in the ten mixtures and four reference samples.

	Mixture No.										**Reference No.**			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
a(µl)	700.0	850.0	750.0	800.0	750.0	850.0	800.0	900.0	950.0	700.0	900.0	800.0	800.0	800.0
b(mg)	125.0	50.0	100.0	75.0	25.0	50.0	25.0	125.0	25.0	50.0	0	100.0	0	0
c(mg)	50.4	51.0	75.8	50.6	123.5	74.9	49.0	74.0	26.0	101.0	0	0	75.0	0
d(mg)	25.8	50.4	48.0	98.4	73.2	23.0	124.0	24.5	48.6	75.1	0	0	0	100.0

4.5.1.2. Methodology of Data Pretreatment

It was observed that the peaks drifted from spectrum to spectrum. This was most serious for ^{1}H-NMR spectra (Figure 4.13), but the other nuclei also showed some drifting. Therefore, it is clear that the untreated data does not obey a bilinear model. One reason is the spectral shifting (change in absorption frequency) due to the different surrounding chemical environments (concentrations). This shifting is probably the main reason that SMCR techniques have not been seriously applied to NMR data. NMR data with mildly overlapping spectra should be treated by alignment and other re-adjustment techniques. BTEM can be implemented after the singular value decomposition of these pre-processed data.

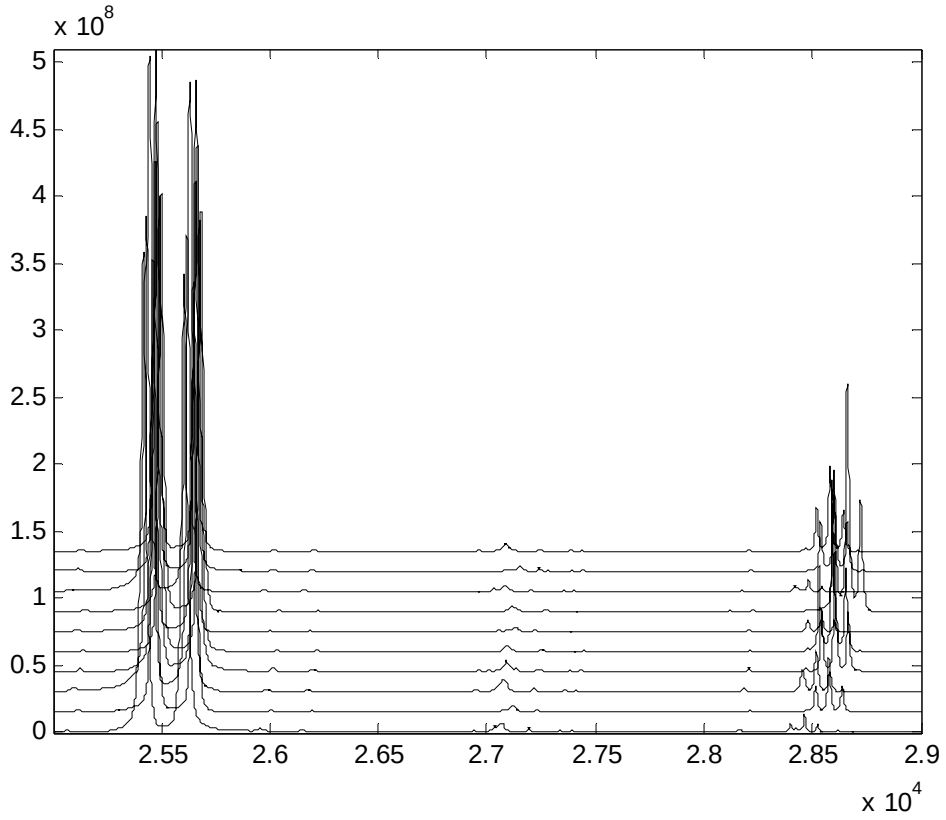


Figure 4.13. Example of the unsystematic drift of each peak in ^{1}H-NMR spectra taken from the ten random four-component solutions.

Indeed, this non-stationary characteristic of the data should be corrected by pre-processing the data with a re-alignment algorithm. But the drift is not constant along the whole spectral range, which makes the alignment process more complicated than normal. Furthermore, the irregular peak shapes also make alignment more difficult. A five-step procedure was proposed. The alignment of the spectra was obtained by the following algorithm.

Step1. The large regions that contain the significant peaks were chosen and the rest were discarded.

Step2. Further sub-divided the large regions with many bands, into smaller regions. Each of these smaller regions represented a segment which could be individually controlled/aligned.

Step3. For each segment, the alignment would be optimized by permitting a constant shift. First, we chose one representative spectrum from our ten mixtures – this was our fixed "reference" spectrum. The remaining nine spectra were compared one-by-one with the "reference" spectrum. The targeted spectrum was shifted in a systematic way until a maximum inner product was achieved. At this point the "reference" pattern and shifted vector were aligned. Figure 4.14 shows the re-aligned segments in the ^{1}H-NMR spectra for one band. There is little doubt that the peaks are aligned well if all peaks in the spectra are symmetric about their peak centers.

Unfortunately, in NMR, especially in ^{13}C-NMR, the frequency dispersion of chemical shift appears in a board range, and the resolution of the peak is low. In other words, often the peak only consists of several points and the peak is asymmetric.

In Figure 4.15, the result of alignment in a single peak ^{13}C-NMR segment according to the procedures stated above is shown. But it is obvious that peaks in the top row (a) are not well aligned. It is clear that the higher one should shift more to the left. Row (b) in Figure 4.15 shows the result of moving the higher spectrum to left side by exactly one channel. This move has over-compensated. Even one channel move is too much, that is to say, precise alignment is impossible to achieve with low resolution data.

Step4. The linear interpolation method was applied to overcome the alignment problem due to the low resolution – and this can only be done by increasing the number of the data channels used. It is obvious that the interpolation does not distort the peak heights and area under the peak (row (c) in Figure 4.15). As a result, the original ^{13}C-NMR

spectrum had a digital resolution of approximately 1.3 points/Hz and was increased to 5.4 points/Hz after interpolation.

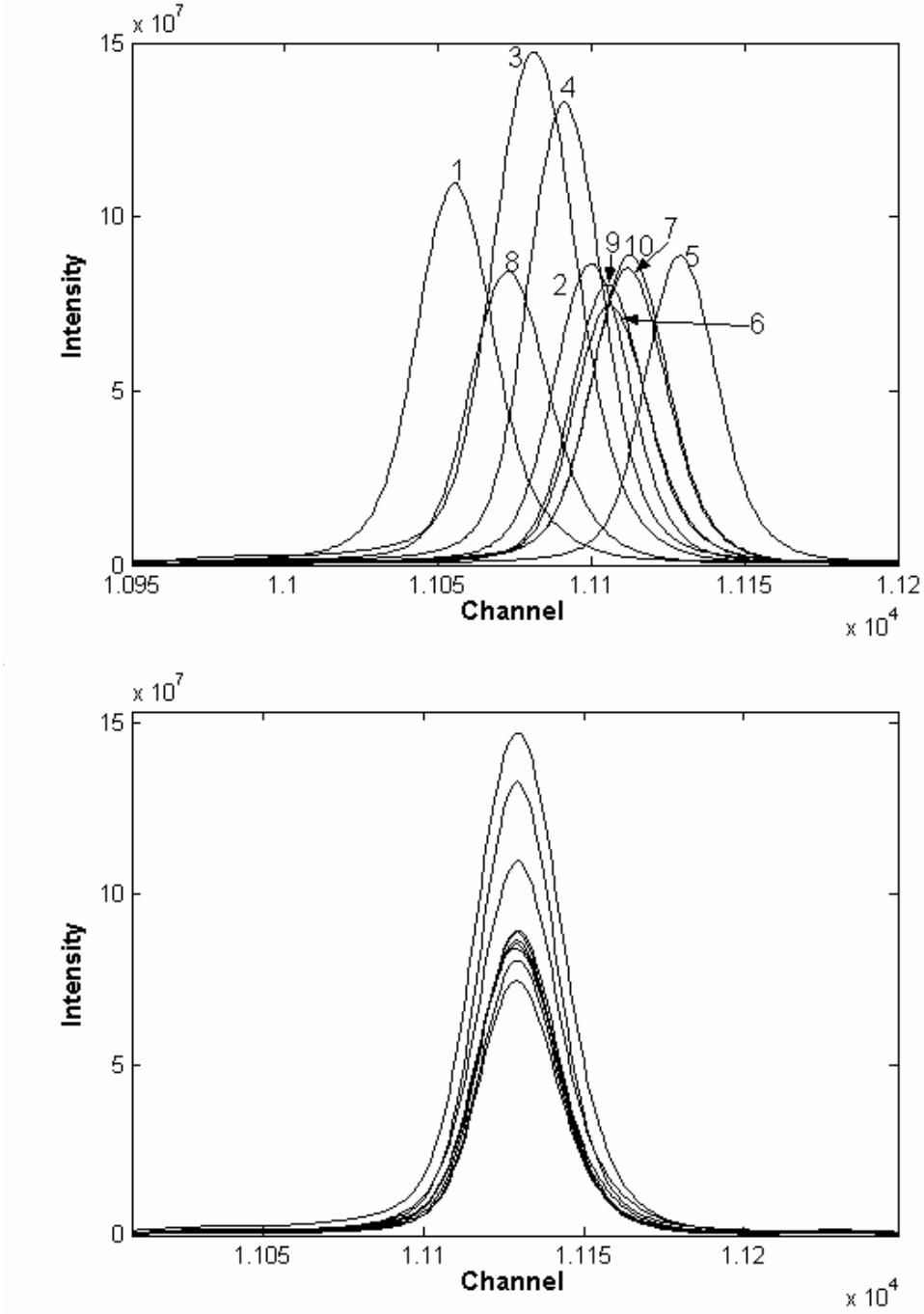


Figure 4.14. The result of alignment. Upper figure: the stack plot of ten mixture ^{1}H-NMR spectra around in peak *s* (Figure 4.17),Bottom figure: spectra after alignment, the index of spectra from top to bottom is 3, 4, 1 ,10, 5, 2, 7, 8, 9, 6.

Step5. After interpolation, (Figure 4.15c) the two peaks still were not symmetric in shape.[v] Accordingly, a smoothing technique would help to alleviate or remove the artifacts. The result of a smoothing approach with the Savitzky-Golay method (discussed in chapter 3, section 3.2.2) is shown in row (d) in Figure 4.15. These two peaks now have similar shape and coinciding peak centers. Of course, other smoothing procedures could be implemented by using other methods and the primary criterion might be to achieve the least possible distortion.

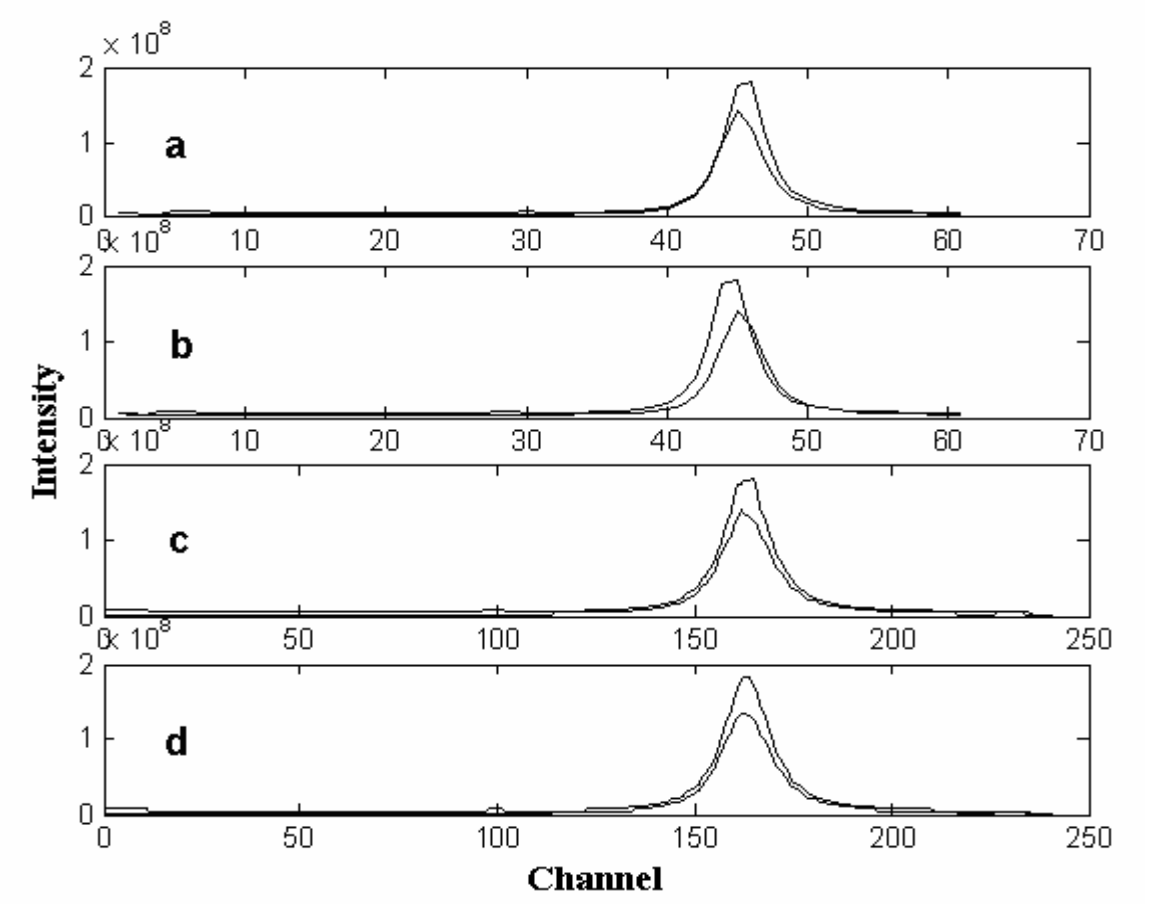


Figure 4.15. The alignment difficulty due to the asymmetric peak in ^{13}C-NMR. (a) the result of left shift, (b) the result of right shift, (c) the alignment result after interpolation, (d) the alignment result after interpolation integrated with smoothing. Note: the top two figures have circa 60 channels of data. The bottom two figures have circa 4×60 channels to facilitate interpolation.

[v] The severe non-linearities may not only originate from the shifting peak positions, but also from the changing shapes of peaks belonging to the same components presented in the different mixture samples.

4.5.1.3. Result

^{1}H-NMR Data Set

Data arrangement: The ^{1}H-NMR data from the 10 mixtures were collected in a matrix ***D*** with size of 10 × 32K(32768), in which each spectrum contained 32K (chemical shifts) readings from -125.84Hz to 369.68Hz. An original multi-component ^{1}H-NMR spectrum is shown in Figure 4.16. The NMR software MestReC, that is available at http://www.mestrec.com/, was used for display / plotting of the spectrum. Further plots having the same format were produced using the same software.

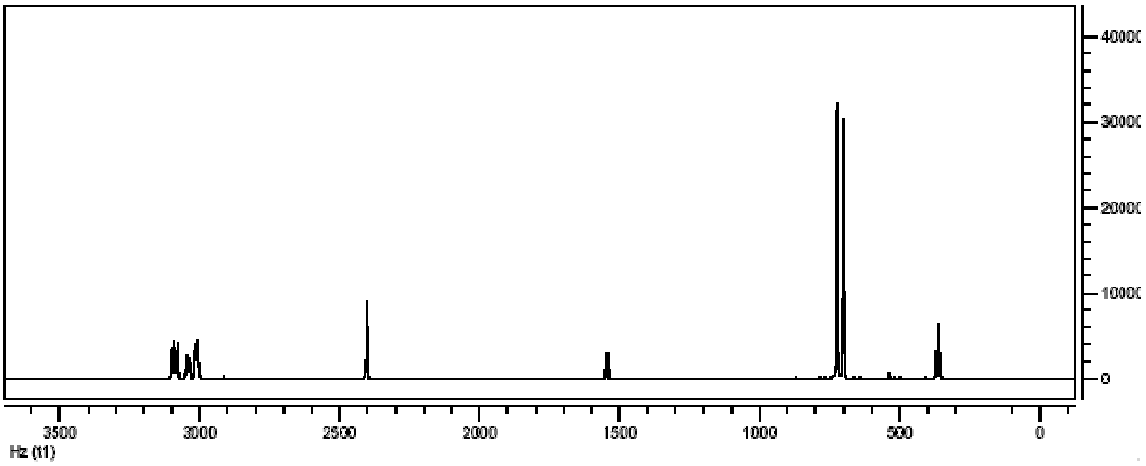


Figure 4.16. One spectrum of the mixture ^{1}H-NMR (in Hz)

In Figure 4.17, all ten mixture ^{1}H-NMR spectra (refer to Table 4.3) are shown. After careful examination of the ten spectra, shifts in some characteristic peaks were clearly discernable. One specific segment was plotted previously in Figure 4.14 and discussed.

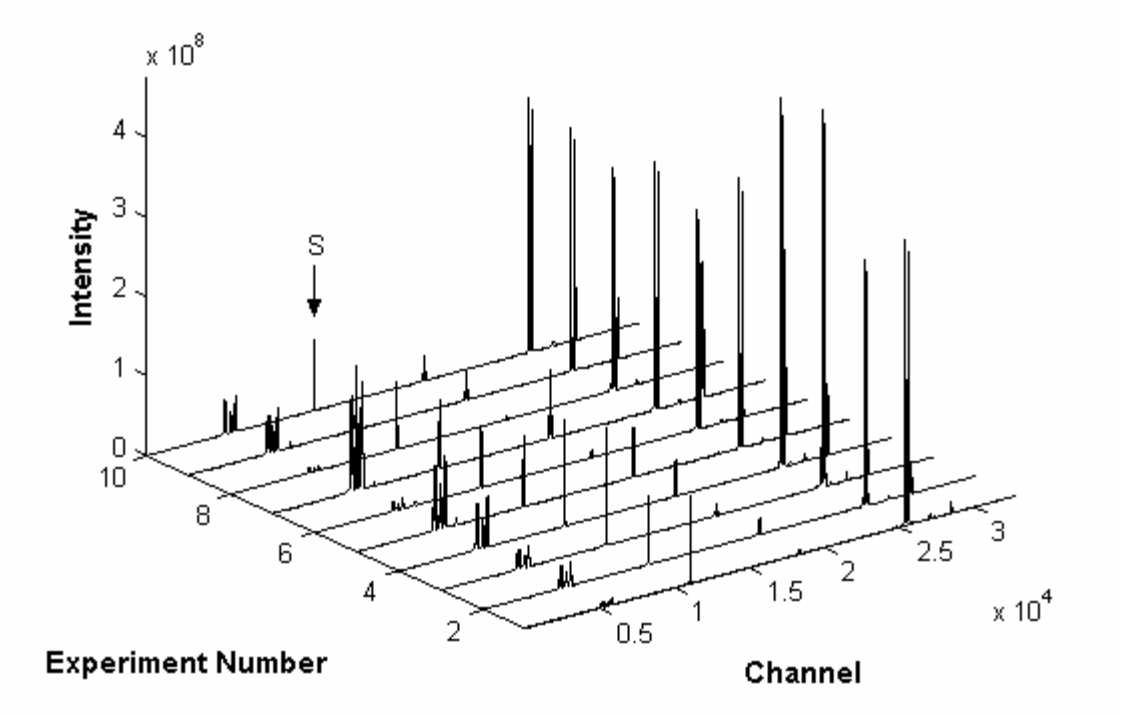


Figure 4.17. Ten original ^{1}H-NMR mixture spectra (reformatted with data channels and not chemical shifts in ppm)

Data manipulation: Due to the irregular shifts in the ^{1}H-NMR spectra, the interesting regions were selected and segments re-aligned individually. The re-aligned data were then filtered using the Savitzky-Golay method. The final pre-treated data were consolidated into a single matrix. SVD was performed on the pre-treated data matrix, yielding the two orthonormal matrices U and V^T, and the diagonal singular value matrix Σ. And the right singular vectors in V^T are needed for the BTEM analysis. Both components possessing hydrogen were obtained and these were compared with the reference spectra. The reconstructed spectra are almost the same as the reference spectra. Some regions which should be zeros in the reconstructed spectra are actually noisy (i.e., see 4.17d from 5000-

10000 data channels). This is due to changing band shapes, inexact shift alignment etc.

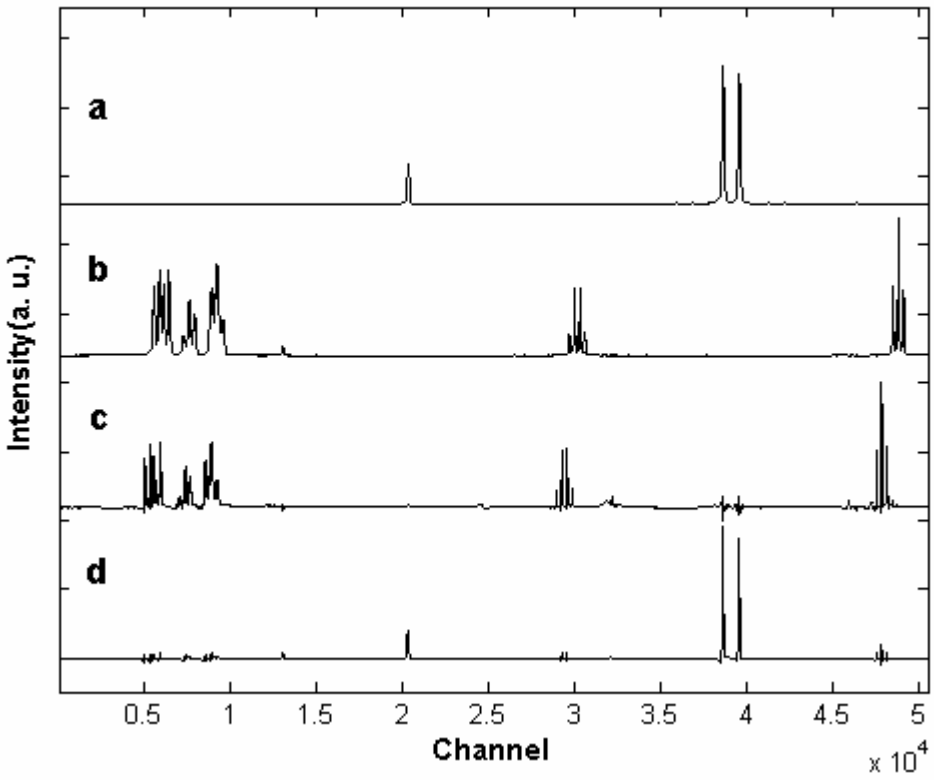


Figure 4.18. The reference ^{1}H-NMR spectra (a and b) and the recovered spectra (c and d) via BTEM. (a) and (d), 2,5-dimethyl-2,4-hexadiene. (b) and (c), ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate

^{13}C-NMR Data Set

Data arrangement: The ^{13}C-NMR data from the 10 mixtures were collected in a matrix ***D*** with size of 10 × 32768, where each spectra contained 32K (chemical shifts) readings from -1958.43 Hz to 22080.04 Hz.

A multi-component ^{13}C-NMR spectrum is shown in Figure 4.19. In Figure 4.20, all ten mixture ^{13}C-NMR spectra are shown.

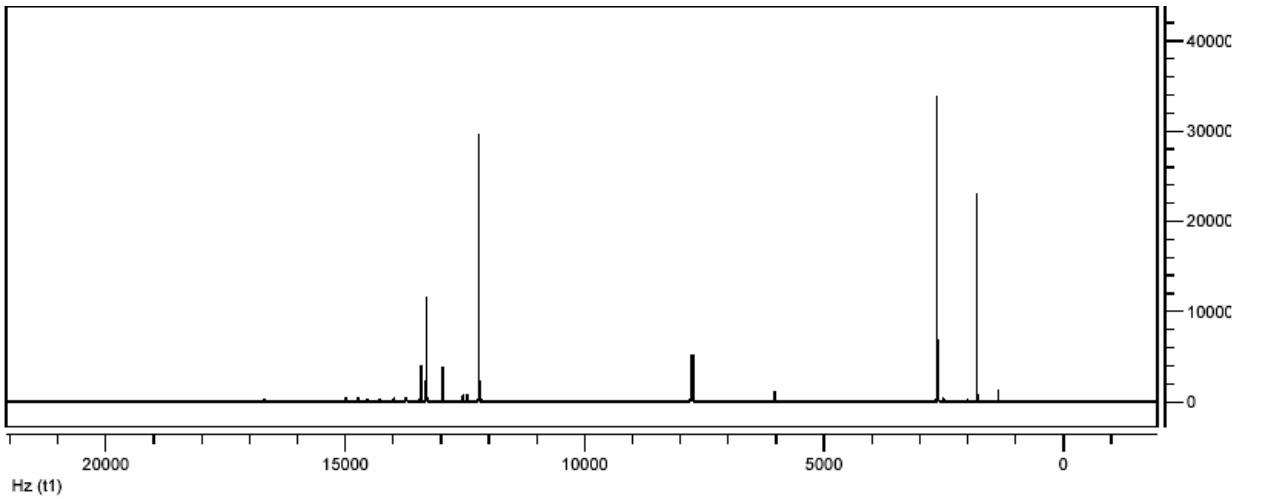


Figure 4.19. One spectrum of the mixture ^{13}C-NMR (in Hz).

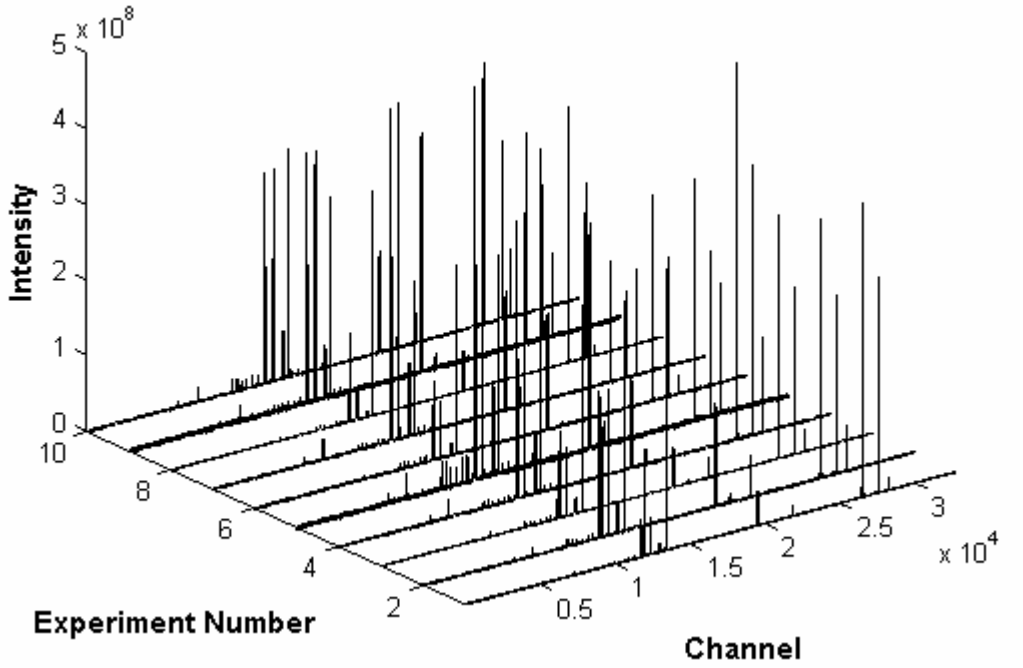


Figure 4.20. Ten original ^{13}C-NMR mixture spectra.

The ^{13}C-NMR data were manipulated according to section 4.5.1.2. Further, BTEM analysis was applied to the pre-treated data. Upon exhaustive targeting of all the interesting features in the V^T , all four ^{13}C-NMR spectra were recovered. The four

recovered ^{13}C-NMR spectra are shown in Figure 4.21. Since all reference spectra were measured by dissolving the analyte in the solvent $CDCl_3$, the reference spectra are actually not the "pure" spectra but the pure spectrum superimposed with the solvent signal (Figure 4.22).

In Figure 4.21, all the three recovered analyte pure spectra are "clean" enough without any "contamination" of solvent. The last spectrum (d) is noisy due to sensitivity issues, and a few signal artifacts arising from the other components present.

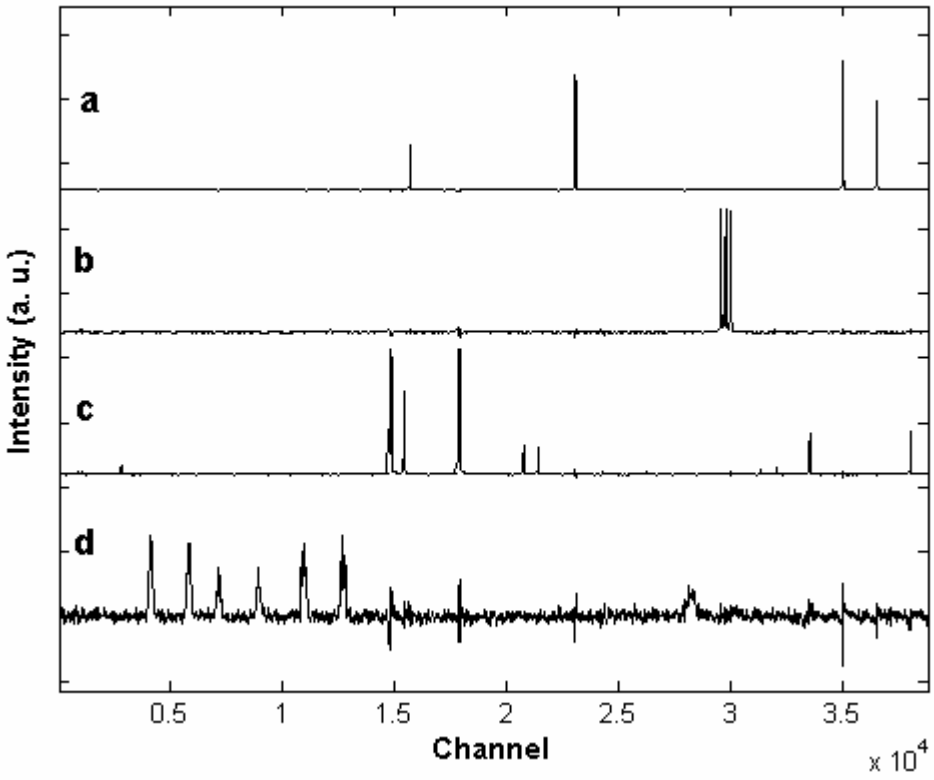


Figure 4.21. The recovered ^{13}C-NMR spectra via BTEM. (a), 2,5-dimethyl-2,4-hexadiene, (b), chloroform-D (c), ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate and (d) tris(pentafluorophenyl)phosphine.

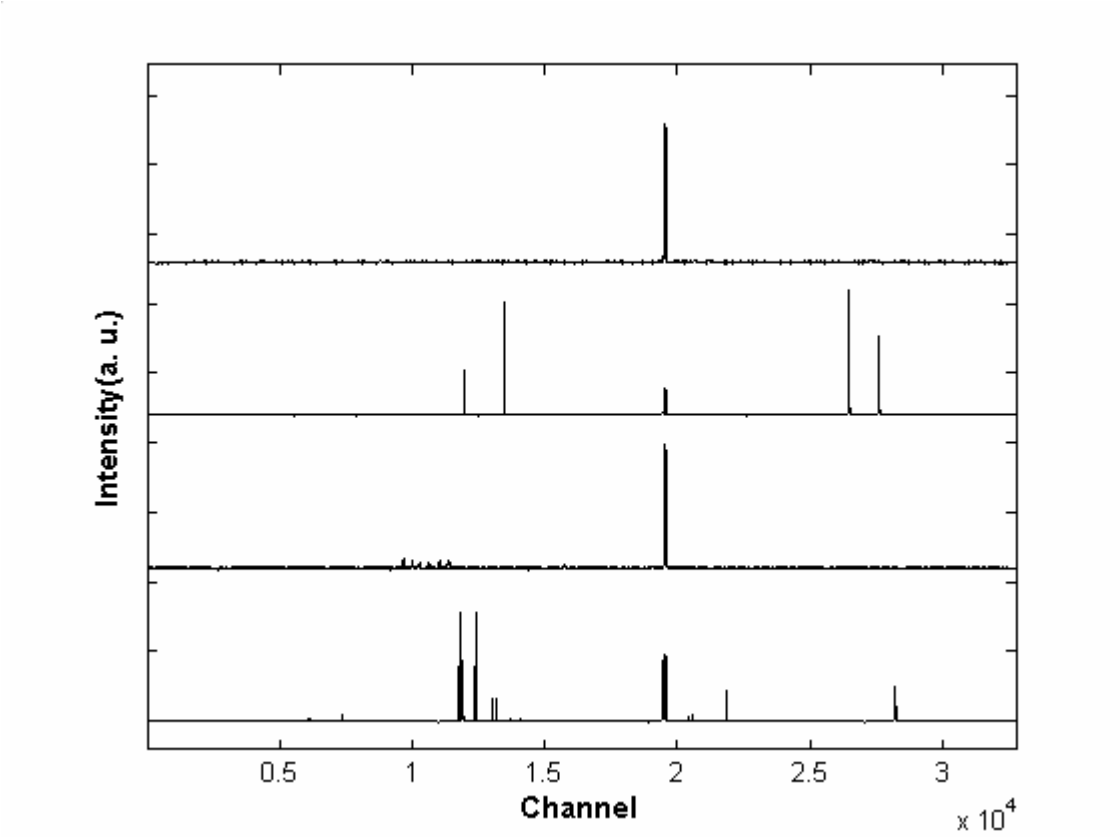


Figure 4.22. The reference ^{13}C-NMR with imbedded solvent signal. (a), chloroform-D (b), 2,5-dimethyl-2,4-hexadiene, (c), tris(pentafluorophenyl)phosphine and (d)ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate.

^{19}F-NMR Data Set

Data arrangement: A multi-component ^{19}F-NMR spectrum is shown in Figure 4.23. In Figure 4.24, all ten mixture ^{19}F-NMR spectra (refer to Table 4.3) are shown. The ^{13}F-NMR data from 10 mixtures were collected in matrix $\boldsymbol{D}$ with size of 10 × 32768, where each spectrum contained 32K (chemical shifts) readings. The ^{19}F-NMR data were manipulated according to section 4.5.1.2. Further, BTEM analysis was applied to the pre-treated data. Upon exhaustive targeting of all the interesting features in the V^T, all the ^{19}F-NMR spectra were recovered. The two recovered ^{19}F-NMR spectra are shown in Figure 4.25.

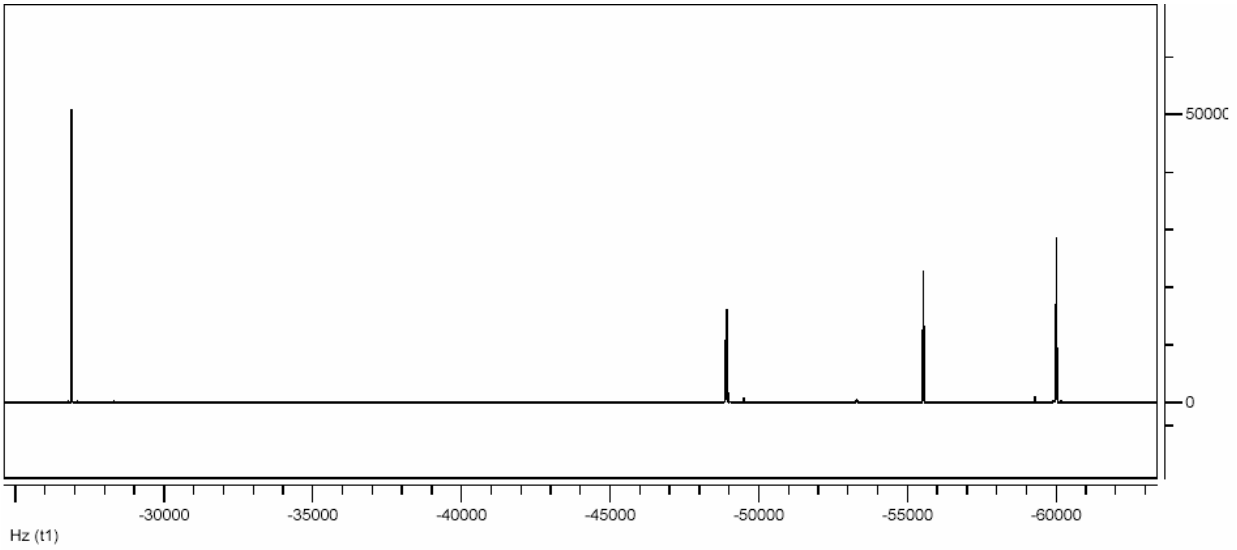


Figure 4.23. One spectrum of the mixture ^{19}F-NMR (in Hz)

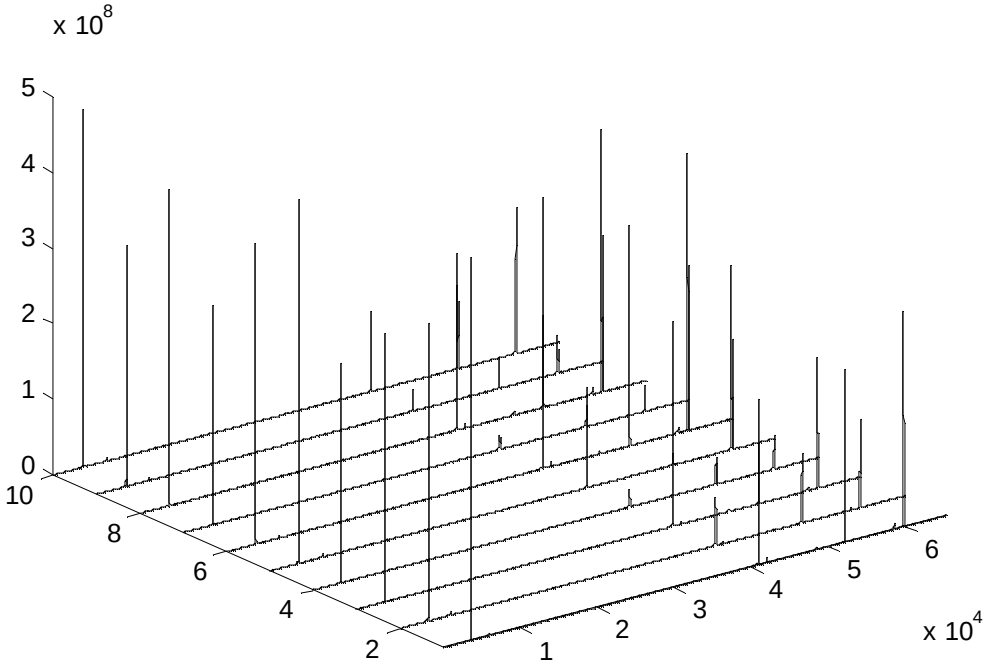


Figure 4.24. Ten original ^{19}F-NMR mixture spectra

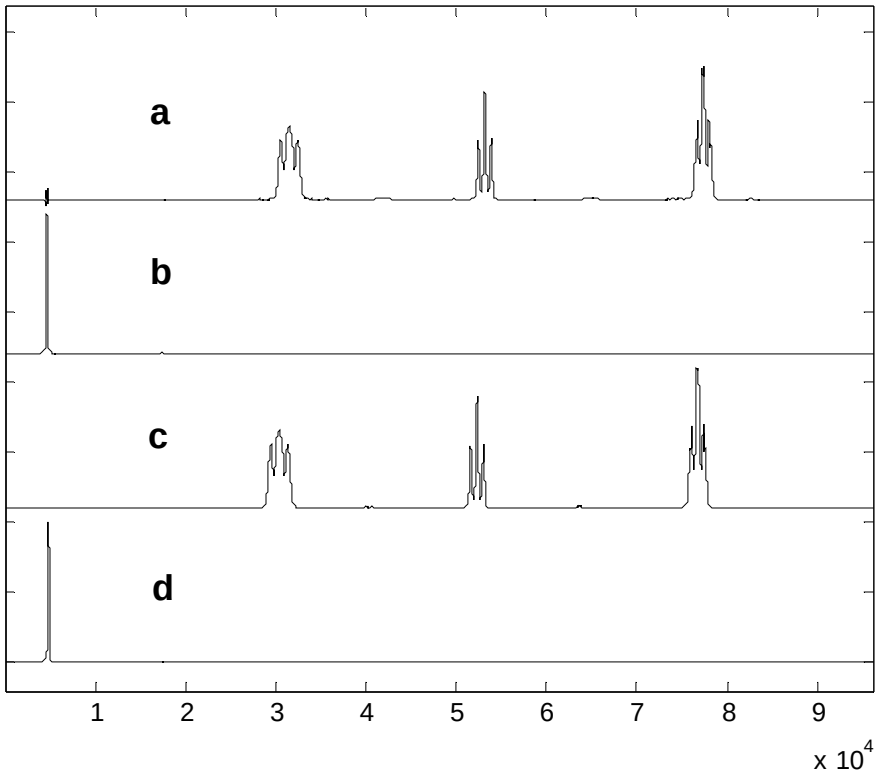


Figure 4.25. The recovered ^{19}F-NMR spectra (a and b) via BTEM and the reference ^{19}F-NMR spectra (c and d). (a) and (c): tris(pentafluorophenyl)phosphine, (b) and (d) ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate.

^{31}P-NMR Data Set

Data arrangement: A multi-component ^{31}P-NMR spectrum is shown in Figure 4.26. In Figure 4.27, all ten mixture ^{31}P-NMR spectra (refer to Table 4.3) are shown. The ^{31}P-NMR data from 10 mixtures were collected in matrix $\boldsymbol{D}$ with size of 10 × 32768, where each spectra contained 32K (chemical shifts) readings. The ^{31}P-NMR data were manipulated according to section 4.5.1.2. Further, BTEM analysis was applied to the pre-treated data. Upon exhaustive targeting of all the interesting features in the V^T, all the ^{31}P-NMR spectra were recovered and the chosen targeted features were also retained. The two recovered ^{31}P-NMR spectra are shown in Figure 4.28.

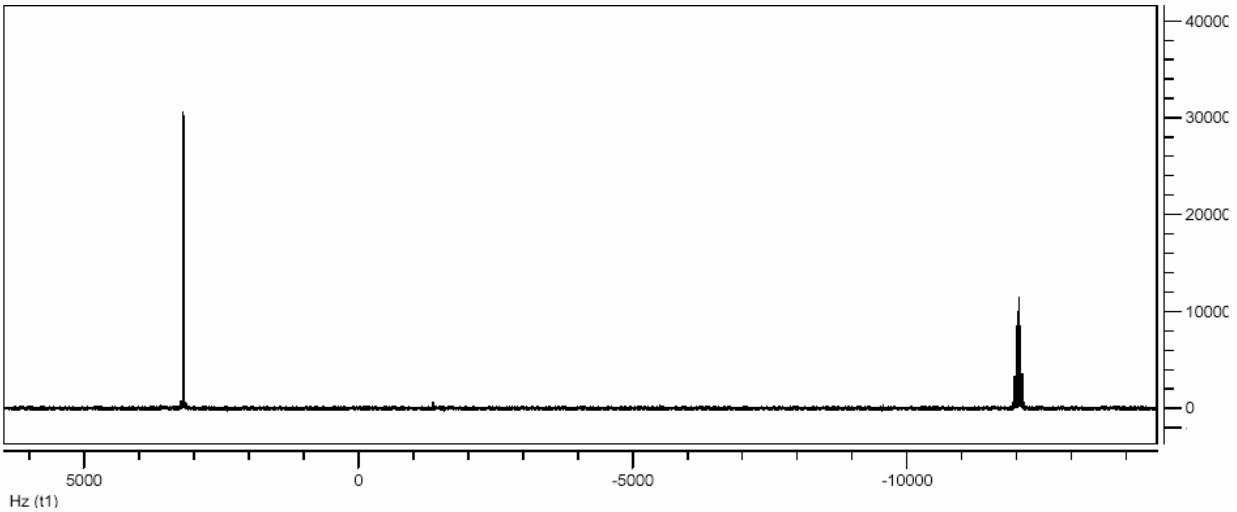


Figure 4.26. One spectrum of the mixture ^{31}P-NMR (in Hz)

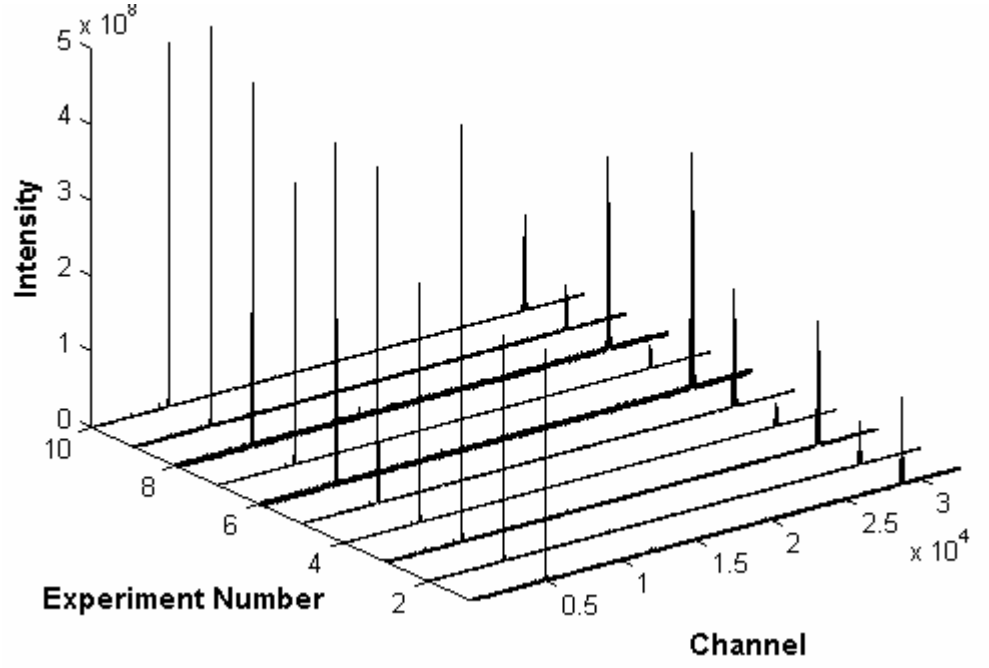


Figure 4.27. Ten original ^{31}P-NMR mixture spectra

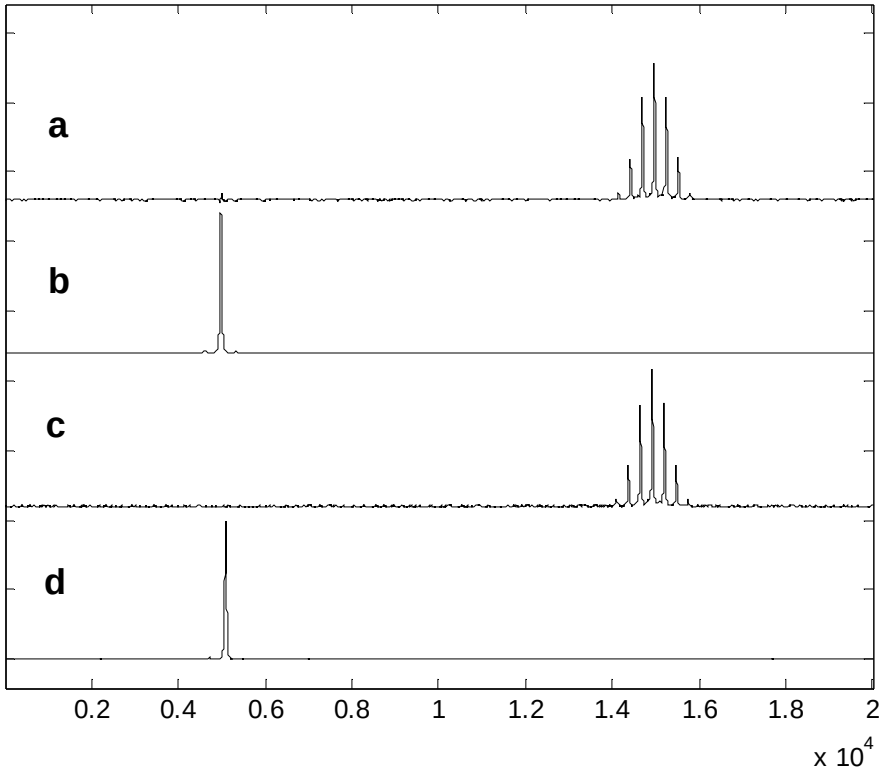


Figure 4.28. The recovered ^{31}P-NMR spectra (a and b) via BTEM and the reference ^{31}P-NMR spectra (c and d). (a) and (c): tris(pentafluorophenyl)phosphine, (b) and (d) ethyl 4,4,4-trifluoro-2-(triphenylphosphoranylidene)acetoacetate

4.5.2. Study of 1D Reaction NMR Data

With successful application of spectral alignment and BTEM on the ^{1}H, ^{13}C, ^{19}F, ^{31}P NMR data individually, the above methodology was further tested in a reaction NMR data set. A cycloaddition reaction was selected in this study.

4.5.2.1. Experimental

Materials: The cycloaddition reaction between 1,3-cyclohexadiene 97% (Aldrich) and dimethyl acetylenedicarboxylate 99% (Aldrich) was studied. Solution 1 was prepared by mixing 0.4 *ml* 1,3-Cyclohexadiene, 0.2 *ml* dimethyl acetylenedicarboxylate and 0.3 *ml* $CDCl_3$. Solution 2 was prepared with mixing 0.2 *ml* 1,3-Cyclohexadiene, 0.4 *ml* Dimethyl

acetylenedicarboxylate and 0.3 *ml* $CDCl_3$. Therefore, these two reactions were carried out with different initial ratios of reagents.

Both reactions were conducted using the same temperature control. For the period 0-15 hours, the reaction temperature was held constant at 295K and for the period 15-22 hours, the reaction temperature was held constant at 310K.

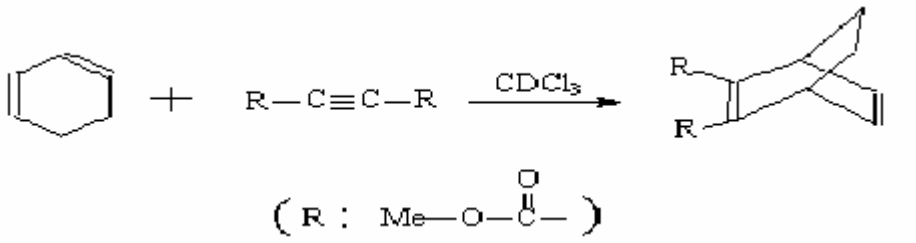


Figure 4.29. The chemical reaction equation for the cycloaddition of 1,3-Cyclohexadiene and Dimethyl acetylenedicarboxylate

Instrumental Aspects: All the data were acquired on a Bruker AVANCE 400 NMR Spectrometer. The interval of each measurement was about 40 mins. ^{13}C-NMR spectra were recorded during the reaction. Moreover, spectra of each pure reagent component were measured as the reference for later comparison with the recovered spectra from BTEM. Reference spectra for (a) Dimethyl acetylenedicarboxylate and (b) 1,3-Cyclohexadiene are shown in Figure 4.30.

Also a time-dependent stack plot of mixture spectra during reaction (Stage I) are shown in Figure 4.31.

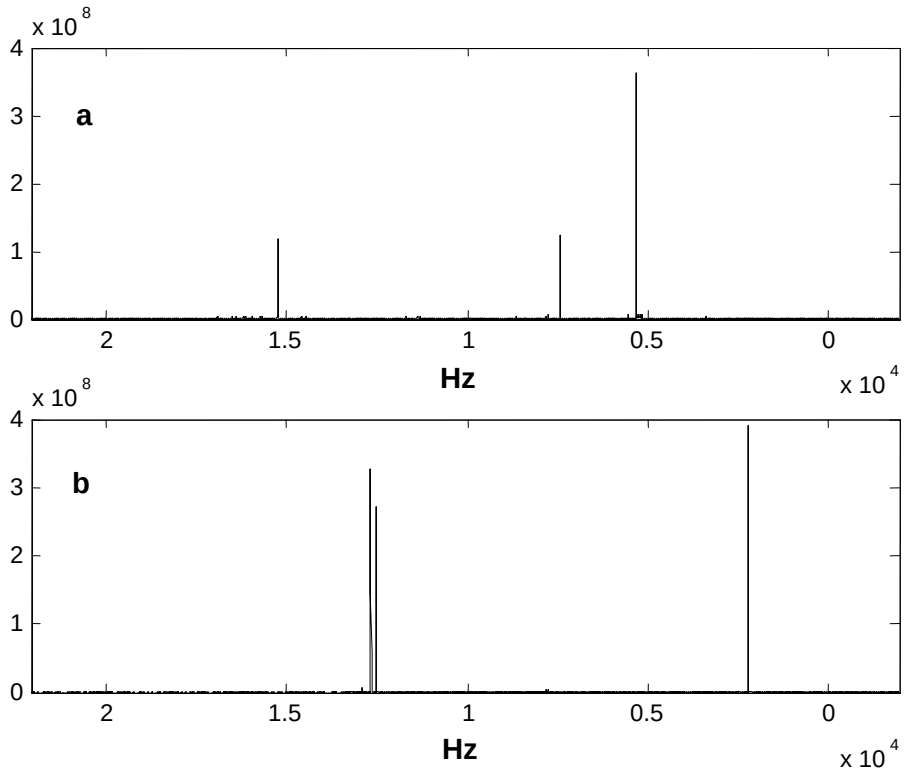


Figure 4.30. Reference experimental ^{13}C-NMR spectra (in Hz) for (a) Dimethyl acetylenedicarboxylate and (b) 1,3-Cyclohexadiene

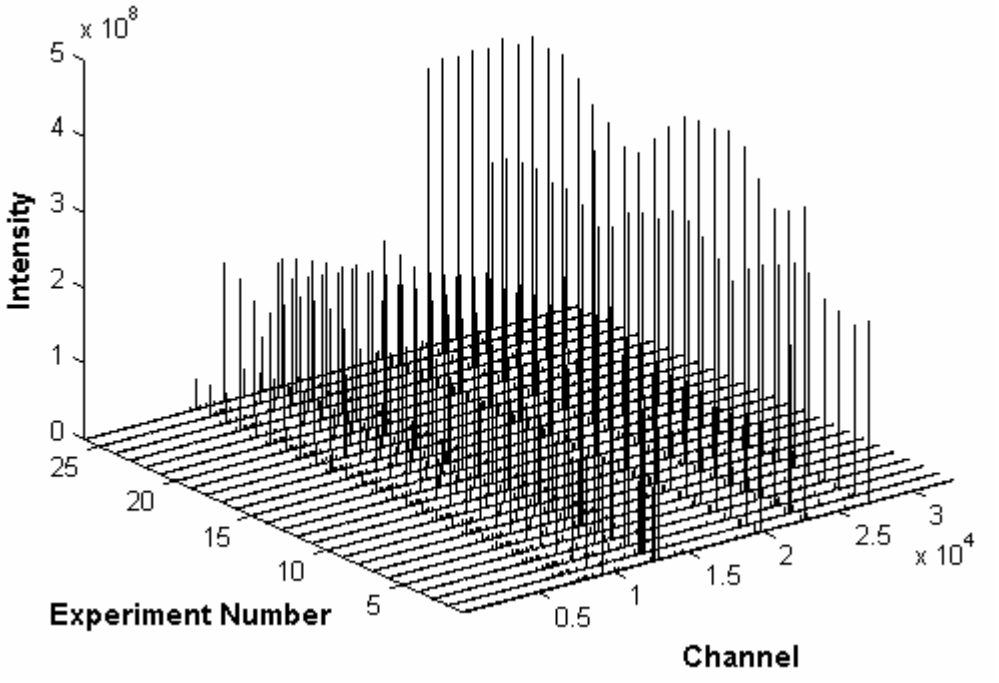


Figure 4.31. A time-dependent stack plot of mixture spectra during reaction (Stage I).

4.5.2.2. Computational Section

Data arrangement: Three sets of ^{13}C-NMR data were collected with 26 measurements from Stage I (Solution 1 at 295K), 12 measurements from Stage II (Solution 2 at 295K), 11 measurements from Stage III (Solution 1 continues at 310K) measurements, respectively. [Note that some instrumental problems occurred and some data was thus discarded] Each ^{13}C-NMR spectra spectrum contained 32K (chemical shifts) readings from -1958.43 Hz to 22080.04 Hz.

As before, non-stationary NMR characteristics were observed in these reaction data sets. The shifts were not constant along the whole range of spectra, and furthermore irregular band shapes also existed. There is no doubt that these nonlinearities should be corrected by pre-processing the data with a re-alignment algorithm and other data processing techniques. An important thing to note is that there were impurities in the reaction. As mentioned in section 3.2.2, the smoothing/filtering procedure will help to minimize the influence of the noise and irrelevant signals which sometime will interfere with the interesting peaks.

Consistent with the algorithm stated in section 4.5.1.2, data were broken down into smaller segments and then aligned individually. In figure 4.32, the top row shows the reconsolidated spectra after segmentation and the bottom row shows the enlarging part from channels 440 to 570 where the drifts of three solvent peaks are prominent. In order to obtain higher quality spectral recovery, the linear interpolation method described in step4, section 4.5.1.2 was applied here to overcome the low resolution data. The re-aligned data were then filtered using the Savitzky-Golay method with the polynomial order and the frame size parameter settings as 3 and 15, respectively.

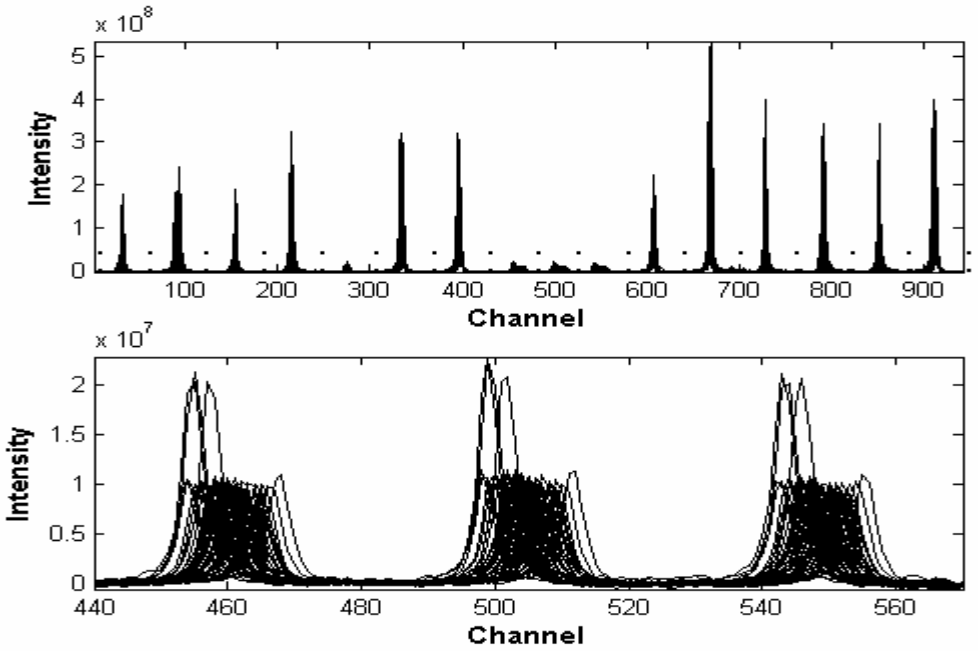


Figure 4.32. The reconsolidated spectra before alignment. Top row: spectra after segmentation; Bottom row: the enlarging part range from 440 to 570 where the shifts of three solvent peaks are prominent.

For each segment, the alignment can be optimized by permitting a constant shift in a controlled manner. Figure 4.33 shows the result after alignment; the top row shows all the spectral results and the bottom row shows the enlarging part ranging from 440 to 570.

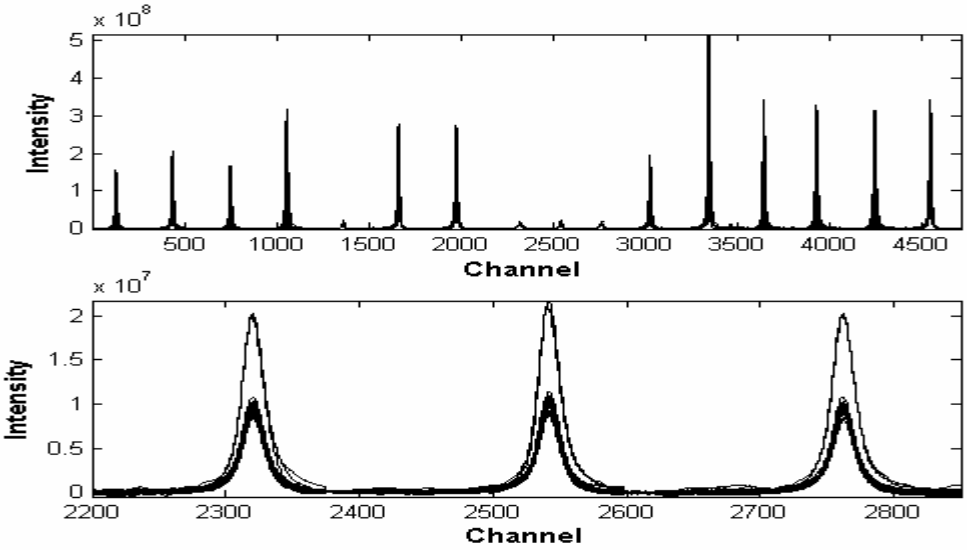


Figure 4.33. The reconsolidated spectra after alignment. Top row: spectra after alignment; Bottom row: the enlargement part from channel 440 to 570 where the shifts are now corrected.

4.5.2.3. Result and Discussion

Recovered Spectra Singular value decomposition was performed on the aligned data. And the right singular vectors in V^T were used for the BTEM analysis. Three major components were obtained and compared with the reference spectra. Two of the reconstructed spectra are close to the reference spectra. These were the reagents. The third species is the product which is easily verified by the presence of six distinguishable ^{13}C resonances.

It is pity that the pure solvent spectrum is not being perfectly recovered in this study. One reason is that the solvent signal is too low compared with the reactants. Another explanation is that the variation of the solvent signal is also too low. A remedy to solve the problem is to introduce more variation of solvent in a better experimental design.

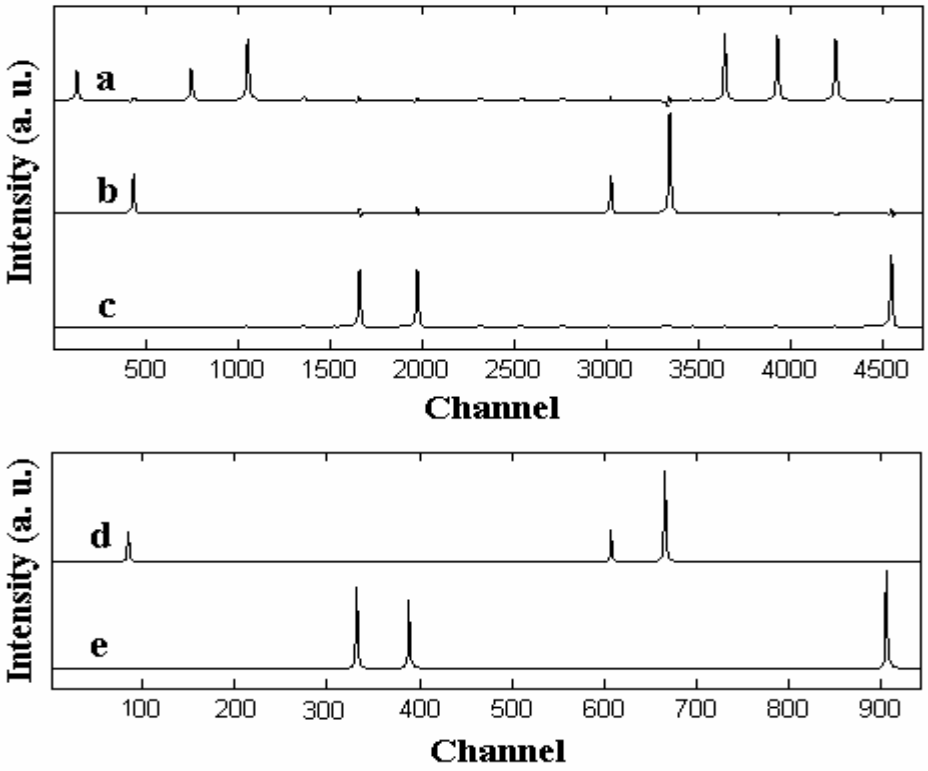


Figure 4.34. The recover spectra (upper figure, a, b, and c) and the reference (bottom figure, d and e). b and d are spectra of Dimethyl acetylenedicarboxylate; c and e are the spectra of 1,3-Cyclohexadiene; a is speculated to be the product spectrum.

Concentration profile. With all the recovered spectra, a first approximation to the profiles of the relative concentrations of the reactants could be obtained by using a least-square fit of the resolved pure component spectra to the mixture spectra. This approach neglected the small contribution to the ^{13}C signal due to chloroform-D.

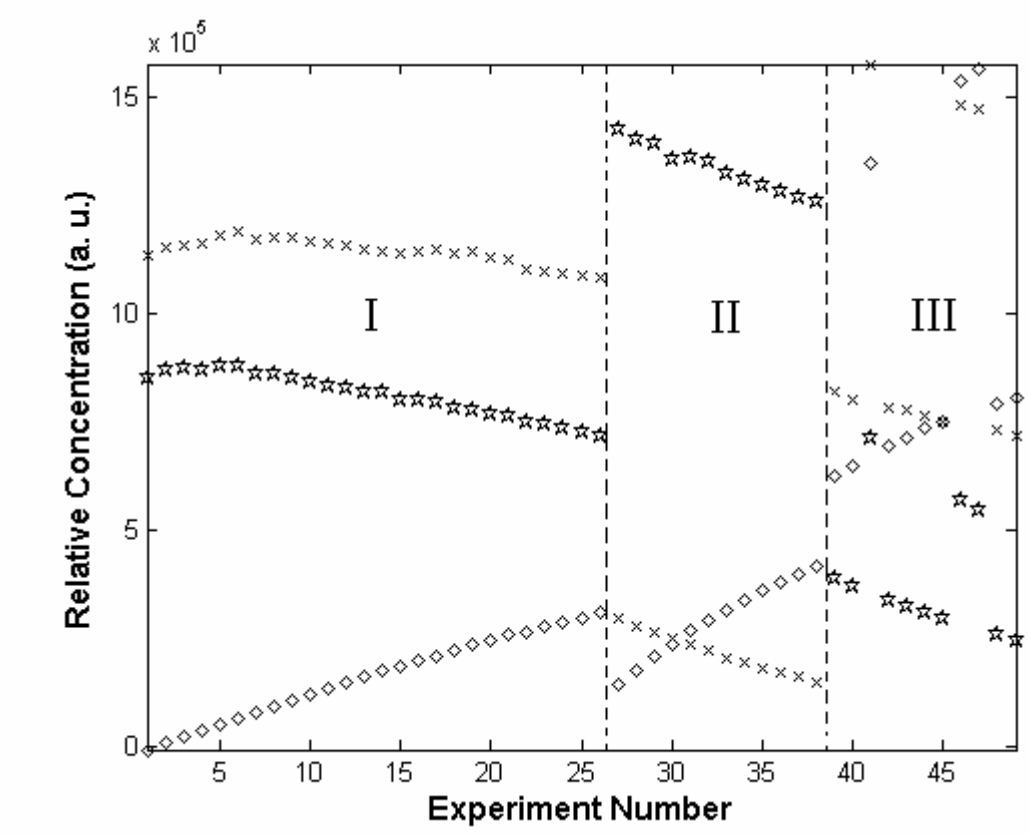


Figure 4.35. The relative concentration profiles for three stage of the reaction before normalization. Cross: Dimethyl acetylenedicarboxylate; Six-point star: 1,3-Cyclohexadiene; Diamond : product.

It is worthy to report here, that the analysis was actually somewhat more complicated. During the first attempt at analysis, inconsistencies in the profiles were obtained for Stage III (Figure 4.35). Some data points in the Stage III were clearly

undergoing unacceptable variations (moving up or down dramatically). It quickly became apparent that the band intensities were varying too much (instability in the spectrometer). Therefore, a spectral normalization procedure[vi] was performed by adopting the middle peak of solvent chloroform-D as internal standard for the normalization. The recalculation of the concentration profiles after the normalization showed reasonable monotonically decreasing or increasing profiles profile throughout duration of the reaction period i.e. Figure 4.36.

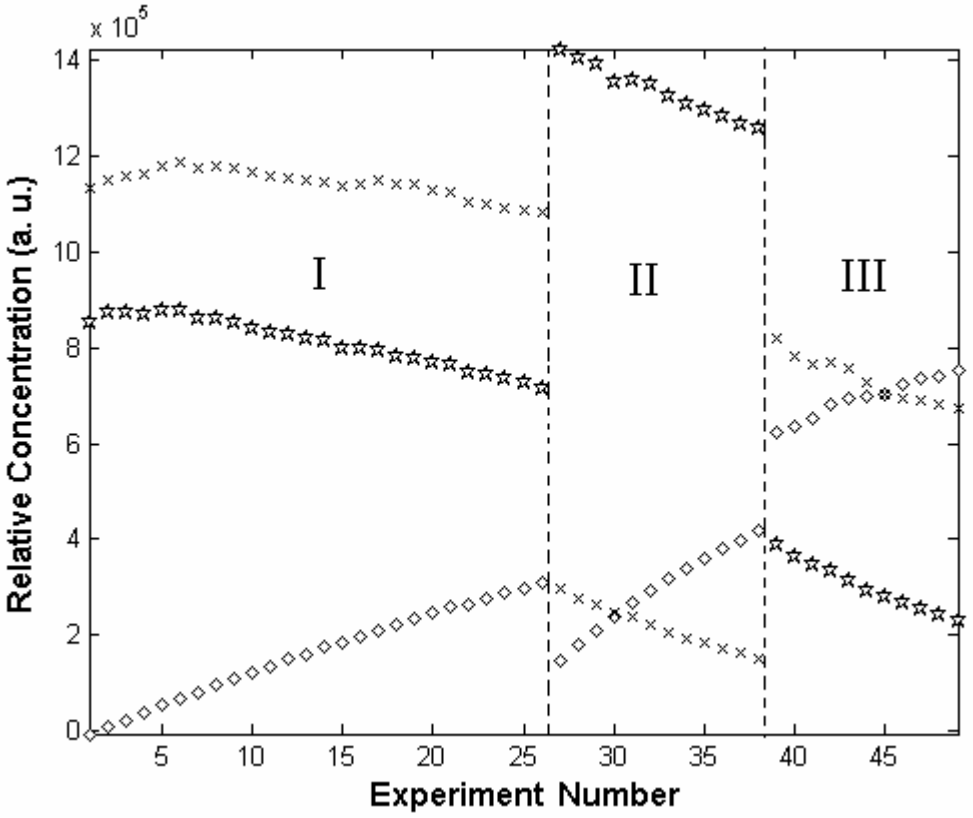


Figure 4.36. The relative concentration profiles for three stage of the reaction after normalization. Cross: Dimethyl acetylenedicarboxylate; Six-point star: 1,3-Cyclohexadiene; Diamond : product.

[vi] The group has been using spectral normalization algorithem to get accurate in-situ concentrations when system pressure, solution density/volume, temperature, spectroscopic pathlength, etc. change during reaction. It is a very reliable method to correct data.

4.5.3. Conclusion

In this section, BTEM showed its usefulness to identify and to recover pure component NMR spectra from a non-reactive mixture system and a reaction system. These experiments show that if (1) there are much more dense data sets (more channels per band), (2) much more spectra per data set[vii] and (3) a proper alignment method like the one introduced here, BTEM can be successfully applied in NMR system identification. The potential for greatly assisting the understanding of the chemistry in poorly understood reactions is obvious.

Further, in order to obtain higher quality spectral recovery, other NMR signal enhancement techniques, just like Cadzow (1988) iterations and Linear Prediction (Tirendi and Martin, 1989) technique may be integrated with the aforementioned techniques to remove undesirable variations in the data.

Also it is important to note that COW and DTW, which were used to correct the position shifts in the spectra in the open literature, were not used here in this chapter. When these procedures were tried with this data, severe distortion was observed. DTW, which works through the signals element by element, has a tendency to destroy the integral of the band/peak, and distortion existed in COW due to the stretch or shrinking of segments as well.

4.6. Summary

In this chapter, BTEM, the self-modeling curve resolution method, which is based on entropy minimization, is introduced after reviews of other curve resolution methods.

[vii] If the number of spectra is very large, then enough information is available about the non-linear characteristics to achieve a very smooth re-construction.

BTEM was successfully implemented to analyze and to reconstruct various cases of multi-component mixture systems including the homogeneous catalytic hydroformylation reaction, and inorganic power X-ray diffraction patterns. Also the numerical experiment of separation mixtures source via entropy minimization method is demonstrated; the usage of the Fourier transform in the acoustics analysis is discussed. This chapter also included discussion of the entropy and the similarities functions that are utilized to improve spectral reconstruction.

Finally, 1D NMR data analysis is the focus of a larger part of this chapter. A five-step procedure was proposed to perform the alignment of the spectra. A non-reactive mixture and a reaction system were tested. Satisfactory results were obtained. Much more work will have to be done to develop routine and automated programs for correcting NMR data before routine BTEM analysis can be performed.

Chapter 5

2D Entropy Minimization Algorithm

There are many types of 2D patterns and spectra in the chemical sciences. Besides multivariate images given by many surface analysis methods (Paul *et al*., 1992; Geladi and Grahn, 1996), there are still quite a lot of different types of 2D spectroscopic data available (discussed in chapter 3, section 3.1). Since there is the pervasive popularity of the 2D format data set in chemical laboratories, the need for 2D data analysis is increasing in demand. The resolution of the blind source separation typically involves 1D and 2D data arrays and is found in a wide variety of disciplines (Cardoso, 1998). It represents a particularly difficult type of inverse problem, where the observables are superpositions/mixtures of source patterns. In the most difficult form of such problems, no *a priori* information is commonly available concerning neither the individual source patterns, nor the number of sources giving rise to the observations.

In this chapter, the 1D entropy-based curve resolution algorithm is generalized and further extended to 2D spectroscopy in order to treat sets of 2D spectroscopic data. A new entropy like function is defined that more fully utilizes the inherent organization found in 2D arrays. The details of matrix-wise 2D Entropy Minimization algorithm as well as the vector-wise 2D Entropy Minimization are shown. Also the 2D and higher dimensional target transformation techniques for testing hypothetical factors are introduced.

5.1. Methodology of 2D Entropy Minimization

In chapter 2, section 2.2.1, various methods proposed for estimating the component spectra from multi-component mixtures with different concentrations have been reviewed. Methods based on self-modeling curve resolution techniques are more attractive since it is remarkable that the estimated basis component spectra can be extracted with no use of other external information. In 1983, Sasaki, *et al.* (1983) proposed a method based on entropy minimization with the merit of providing the unique estimation of the component spectra. Fully revising this algorithm, a new algorithm called Band-Target Entropy Minimization (BTEM) was introduced by Garland's group recently. BTEM is a self-modeling deconvolution technique. The novel entropy minimization method used in BTEM is based on the idea that the real pure components are the simplest patterns involved in the data set and these should have the minimum values of entropy. The pursuit of the minimum entropy thus gives the pattern with greatest simplicity without any *a prior* information. The details of the BTEM methodology were given in chapter 3 and the references therein. The development of the 2D Entropy minimization follows in the next section.

5.2. Overview of Approach

The basic philosophy behind the present 2D entropy minimization curve resolution methodology has 3 primary parts. First, given an arbitrary set of observations (data array), this array should be decomposed into orthogonal components using PCA or SVD etc. It is known that if the number of the observations is smaller than the number of components embedded inside, this leads to an under-determined problem. Fortunately, this is not often the case; collection of copious quantities of data is often an easily met prerequisite for

analysis. Secondly, an appropriate entropy like function must be chosen, and an objective function must be designed to sequentially search for the minima which correspond to the pure patterns. A penalty term is needed to narrow the space and to ensure finding the feasible solution. At the last stage, a reliable global optimizer is utilized to achieve the objective function minima and realize the mixture separation.

5.3. System Representation

First, in order to facilitate discussion of the numerical data analysis for 2D Entropy minimization, it is instructive to have a system model. For 2D absorption spectroscopy, for example, in 2D Nuclear Magnet Resonance spectroscopy which involves spectra with two frequency axes, each element in the matrix is specified by two coordinates corresponding to frequency f_1 and frequency f_2. Let $A_{m\times n}$ denote one measurement of a single multi-component solution with s species where m and n are the number of channels in each spectral direction respectively. It is assumed that there are N_i moles of each species present in the measured sample. Each species possesses a pure component absorptivity $a_{m\times n}$. Since the measured 2D spectrum is the superposition of s pure component absorptivities, $A_{m\times n}$ can be represented as shown in Eq. 5.1 where $\varepsilon_{m\times n}$ denotes the associated instrumental/experimental error.

$$A_{m\times n} = \sum_i N_i a_{m\times n\ i} + \varepsilon_{m\times n} \qquad (5.1)$$

The q spectral measurements give rise to a set of 2D observations. These will be denoted $\underline{A}_{q\times m\times n}$ (an underlined alphabet denotes the 3-way array) where the raw absorbance data is now a 3-way array. It is very important to note that the error term still exists and

consists of instrumental/experimental error *as well as* the system non-linearities arising from the non-stationary signal characteristics.

5.4. Data Decomposition and Model Reduction

In chapter 3 section 3.5, the theoretical basis for generalized decompositions of higher-order tensors and n-arrays has been discussed. The ordering of the decomposition of an n-array can be carried out in more than one way, in other words, with priority given to one or more indices. The numerical realizations of such decompositions have been reported (Kolda, 2001). But it is quite a common situation that the increase of dimension does not change the intrinsic structure of the data, the multidimensional arrays or n-way arrays are still bilinear data which can not be decomposed into independent vectors in each dimension. Therefore, instead of using a full orthogonal tensor decomposition technique (which will probably lead to loss of physical understanding of the problem), a modified tensor decomposition is used. In the following development, a decomposition which retains the natural physical structure of the data set is implemented.

5.4.1. Principle Component Analysis (PCA)

Principle component analysis (also known as the Karhunen-Loève transform) is often used for finding the latent variables in data analysis (see discussion in chapter 3, section 3.3.1). PCA is well known for its use in identifying a few variables which explain all or most of the total variance. The components after the transformation have the highest possible variance, with an important impact on the data analysis and later data reconstruction.

The 2D Entropy Minimization algorithm begins with an appropriate decomposition. In the present case of analyzing a series of 2D data, a 3-way data set is generated. In other words, the set of observations consist of stacks of 2D mixture data. As shown in Eq. 5.2 and Fig 5.1, a 3-way data set can be decomposed into two parts. One is the loading part, the other is the scores part. The 3-way matrix algebra equation is represented as follows:

$$\underline{A} = \sum_{i=1}^{q} F_i \otimes C_i + E \tag{5.2}$$

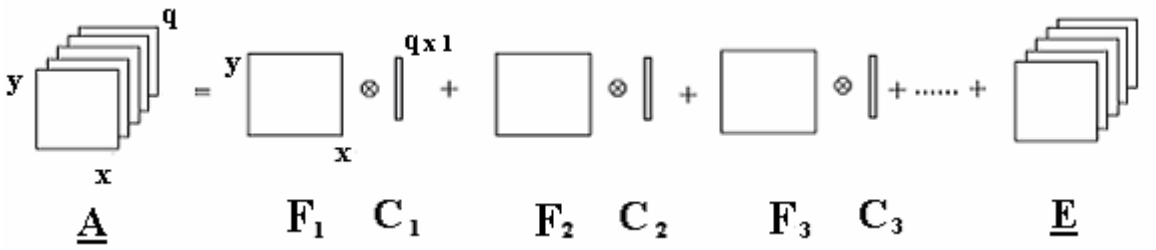


Figure 5.1. A three-way data can be decomposed into a sum of Kronecker products and a residual E.

where, $\underline{A}$ denotes the series of q mixture spectra ($\underline{A}$ is a 3-way array); F_i denotes matrix-formatted component i with size ($x \times y$); C_i denotes the loading for component i (relate to its contribution) which is a vector of length q; $\otimes$ denotes the Kronecker product; and $\underline{E}$ is the residual part. Ideally most of the information in the mixture spectra will be concentrated in the first term of Eq. 5.2, but in real sets of data there are always noise and errors which show up in the residual term, E. Details of the PCA on 3-way array data decomposition can be found in paper by Geladi *et al.* (1989).

5.4.2. Singular Value Decomposition (SVD)

A PCA decomposition can be conveniently calculated using a SVD technique, which possesses a more robust and efficient process (Golub and Kahan, 1965; Golub and Reinsch, 1970). SVD is used in the present contribution. The details of SVD decomposition procedure is outlined here.

In order to achieve appropriate vector-space decomposition of this 3-way array $\underline{A}_{q\times m\times n}$, $\underline{A}_{q\times m\times n}$ must be unfolded into a 2-way array (matrix) $A_{q\times(m\times n)}$ first. This requires that each experimentally measured 2D spectrum is re-ordered by *concatenation/unfolding*. As a result, the number of elements in the 3-way array and the 2-way array are exactly the same. This procedure can be represented in Eq. 5.3, in which a 2-way array is obtained by the transformation of the 3-way array.

$$\underline{A}_{q\times m\times n} \rightarrow A_{q\times(m\times n)} \tag{5.3}$$

Secondly, SVD can be directly applied to $A_{q\times(m\times n)}$. As mentioned in chapter 3, section 3.3.3, three new objects, namely, the left singular matrix $U_{q\times q}$, the diagonal singular values matrix $\Sigma_{q\times(m\times n)}$ and the right singular matrix $V^T{}_{(m\times n)\times(m\times n)}$ are produced.

$$A_{q\times(m\times n)} = U_{q\times q}\,\Sigma_{q\times(m\times n)}\,V^T{}_{(m\times n)\times(m\times n)} \tag{5.4}$$

It is important to note that the number of columns in the matrix $\Sigma_{q\times(m\times n)}$ and the number of rows in the matrix $V^T{}_{(m\times n)\times(m\times n)}$ greatly exceed the number of experimental spectra. The extra columns and rows exist due to the mathematical construct of SVD and are not physical relevant. Accordingly, the last $(m\times n)-q$ columns/rows can be discarded which lead to the truncated form as Eq. 5.5.

$$A_{q\times(m\times n)} = U_{q\times q}\,\Sigma_{q\times q}\,V^{T}{}_{q\times(m\times n)} \tag{5.5}$$

The last stage of the decomposition involves the refolding procedure which involves the undoing of concatenation, or folding on the resulting matrix $V^{T}{}_{q\times(m\times n)}$. This leads to Eq. 5.6 where the *q* physically meaningful right singular vectors have been transformed to *q* physically meaningful right singular matrices which are analogous to F in the PCA decomposition mode.

$$\underline{A}_{q\times m\times n} = U_{q\times q}\,\Sigma_{q\times q}\,\underline{V}^{T}{}_{q\times m\times n} \tag{5.6}$$

If the absorptivities absolutely obey the Lambert-Beer Law, and the bilinear model holds for the system, there would be only *s* degrees of freedom, and subsequently only *s* of the *q* right singular matrices in $\underline{V}^{T}{}_{q\times m\times n}$ in Eq. 5.6 would be physically important from a spectroscopic viewpoint of reconstruction. However, as stated before in chapter 4. section 4.1.2.1, for the system representation, the bilinear model is only locally valid for real data sets. Comparison of the system model, Eq. 5.1, and the decomposition of experimental observations, Eq. 5.6, leads to the conclusion that information on the *s* pure component absorptivities are imbedded, in a nontrivial manner, in the *q* physically meaningful right singular matrices in $\underline{V}^{T}{}_{q\times m\times n}$. If the number of experimentally measured spectra *q* happens to be less than the number of observable components *s*, then the mathematical problem can be considered irrevocably ill-posed and subsequently, there is no possibility for a unique solution to deconvolution.

Here it is necessary to emphasize the importance of SVD with the expression in Eq. 5.6. Consistent with the discussion in section 3.3.3, SVD can be considered a general framework for rank reduction and data compression which indeed provides a crucial initial untangling of the signals. Specifically, SVD untangles the 3-way array observations into *j*

matrices which contain a significant amount of useful physical information relating to the pure 2D spectral patterns. There are *j* essential matrices of meaningful information in the *q* right singular matrices in $V^T_{q\times m\times n}$ (where $s<j<q$) and the remaining right singular matrices are mainly randomly distributed noise. The situation can be expressed by Eq. 5.7 where $\underline{\tilde{A}}_{q\times m\times n}$ is now the expectation for the set of observations (Eq. 5.8). With *q-j* matrices of discarded noise, the potential exists for spectral reconstruction with outstanding signal to noise enhancement.

$$\underline{\tilde{A}}_{q\times m\times n} \leftarrow U_{q\times j}\,\Sigma_{j\times j}\,\underline{V}^T_{j\times m\times n} \tag{5.7}$$

$$\underline{\tilde{A}}_{q\times m\times n} \approx \underline{A}_{q\times m\times n} \tag{5.8}$$

It was demonstrated in 1D spectroscopic data, that circa one thousand spectra can be collected during a reaction and this can be reduced to around 50 right singular vectors which contain most information in the reaction system (Li *et al.*, 2002, 2003). The data reduction greatly facilitates the consequent data analysis procedures.

5.5. The Formulation of 2D Entropy Minimization

The ultimate objective of 2D Entropy Minimization is to obtain accurate estimates of the pure component absorptivities. This is achieved by transforming the abstract V^T information into pure component absorptivity approximations, $\hat{a}_{m\times n}$, one estimate at a time. The computation can be performed on either the right singular vectors in $V^T_{j\times(m\times n)}$ or the right singular matrices in the 3-way array $\underline{V}^T_{j\times m\times n}$. This computational issue requires two different formulations for entropy resulting in two different types of objective functions. Also it is important to note that in the real data set, the real pure component

absorptivities vary somewhat from measurement to measurement, so the approximations of pure component, $\hat{a}_{m\times n}$,is a mean of the observations in some sense.

5.5.1. Vector-Wise 2D Entropy Minimization

For vector-wise 2D-Entropy Minimization algorithm, all the 2D matrix data are re-ordered by *concatenation/unfolding* and represented as vectors. The resultant right singular vectors in ${V^T}_{j\times(m\times n)}$ are used for entropy minimization operation. Thus the 2D pure component recovery is associated with a transformation process from abstract right singular *vectors*, ${V^T}_{j\times(m\times n)}$ by 1D entropy minimization algorithm, into pure component spectral estimates, $\hat{a}$ (vectors) with proper linear combination of the basis vectors. Final 2D pure component spectra should be obtained from re-folding the vector to 2-dimensional matrix.

$$\hat{a}_{m\times n} \leftarrow \hat{a}_{1\times(m\times n)} \leftarrow t_{1\times j}\, {V^T}_{j\times(m\times n)} \tag{5.9}$$

5.5.2. Matrix-Wise 2D Entropy Minimization

The *j* physically meaningful right singular matrices (from 3-way array ${\underline{V}^T}_{j\times m\times n}$) from Eq. 5.6 are used for entropy minimization operation. Thus the 2D pure component recovery is associated with a transformation process from abstract right singular *matrices*. And the proper transformation directly results in the 2D estimate, $\hat{a}_{m\times n}$.

$$\hat{a}_{m\times n} = \hat{a}_{1\times m\times n} \leftarrow t_{1\times j}\, {\underline{V}^T}_{j\times m\times n} \tag{5.10}$$

Some information is in a sense “lost” during unfolding, since the “connectivity or correlation” of the adjacent data elements is erased. Also the unfolding induces period

discontinuities in data which is a small nuisance during entropy evaluation. Therefore, matrix-wise 2D- Entropy Minimization will prove to be the preferred computational route. The higher quality spectral estimates generally result since the spatial correlation is not affected when they are treated as a whole 2D plane.

The optimal determination of the j unknowns in $t_{1\times j}$ is the computationally intensive task. There are two parts involved in solving this problem. The first issue is the repeated evaluation of the entropy of the term $t_{1\times j}\, V^T{}_{j\times(m\times n)}$ or $t_{1\times j}\, \underline{V}^T{}_{j\times m\times n}$. The second issue is the search for the final value of $t_{1\times j}$ such that the global entropy minimum is obtained.

Consistent with the original formulation of Shannon entropy, a definition for the entropy H used in 2D Entropy Minimization can take the form of Eq. 5.11, where the probability distribution p, consistent with Sasaki *et al.*'s original suggestion, is replaced by Eq. 5.12.

$$H = -\sum_{v_1}\sum_{v_2} h_{v_1 v_2} = -\sum_{v_1}\sum_{v_2} p_{v_1 v_2} \ln p_{v_1 v_2} \tag{5.11}$$

$$p_{v_1} = \frac{\left| \frac{d^m \hat{a}_{v_1 v_2}}{dv_1{}^m} \right|}{\sum_v \left| \frac{d^m \hat{a}_{v_1 v_2}}{dv_1{}^m} \right|} \tag{5.12}$$

$$p_{v_1 v_2} = \frac{\left| \frac{d^m p_{v_1}}{dv_2{}^m} \right|}{\sum_v \left| \frac{d^m p_{v_1}}{dv_2{}^m} \right|} \tag{5.13}$$

where the superscript m denotes the degree of spectral differentiation, it can be first derivative, second derivative, or fourth derivative. When m is equal to 1 and the

denominator is dropped as it is just a normalizing factor, Eq. 5.12 will be reduced to $p_{v_1} = \left| \frac{d\ \hat{a}_{v_1 v_2}}{dv_1} \right|$. Substituting the latter expression in Eq. 5.13, $p_{v_1 v_2}$ will turn into the following expression which captures the simultaneous curvature in two directions.

$$p_{v_1 v_2} = \frac{\partial a^2}{\partial v_1 \partial v_2} \tag{5.14}$$

5.5.3. Objective Function Formulation and Optimization

The resolution of a 2D pure component spectrum can be achieved by solving the following minimization problem.

$$\min F_{obj} = H^{2D} + P \tag{5.15}$$

Specifically, the objective function F_{obj} includes the entropy term H^{2D} along with a penalty function P (infra vida). The transformation of right singular matrices into an estimate of a pure spectrum is achieved by Eq. 5.10 but governed by the global optimizer. The global minimization of the objective function is, in principle, achievable using simulated annealing (Corana *et al*., 1987) or a genetic algorithm (Goldberg, 1989) which are both stochastic search techniques.

For the first pure spectral solution, in the objective function given by Eq. 5.15, P is set to zero, and the global entropy minimum is sought without P penalty. In each subsequent search, the penalty function is formulated with the constraints of the P penalty. The minimization solution is admissible only when this subsequent result is dissimilar to all the previous results.

The scheme can be realized by adding a dissimilarity penalty function, P, to the objective function. For example, an admissible function is shown in Eq. 5.16 where $corr2$

denotes the 2D correlation coefficient between any two 2D arrays X and Y (Hogg and Tanis, 1983). The calculation can be implemented with MATLAB function "*corr*2".

$$P = \max\{a \times (e^{\frac{b}{1-corr2}} - 1)\} \tag{5.16}$$

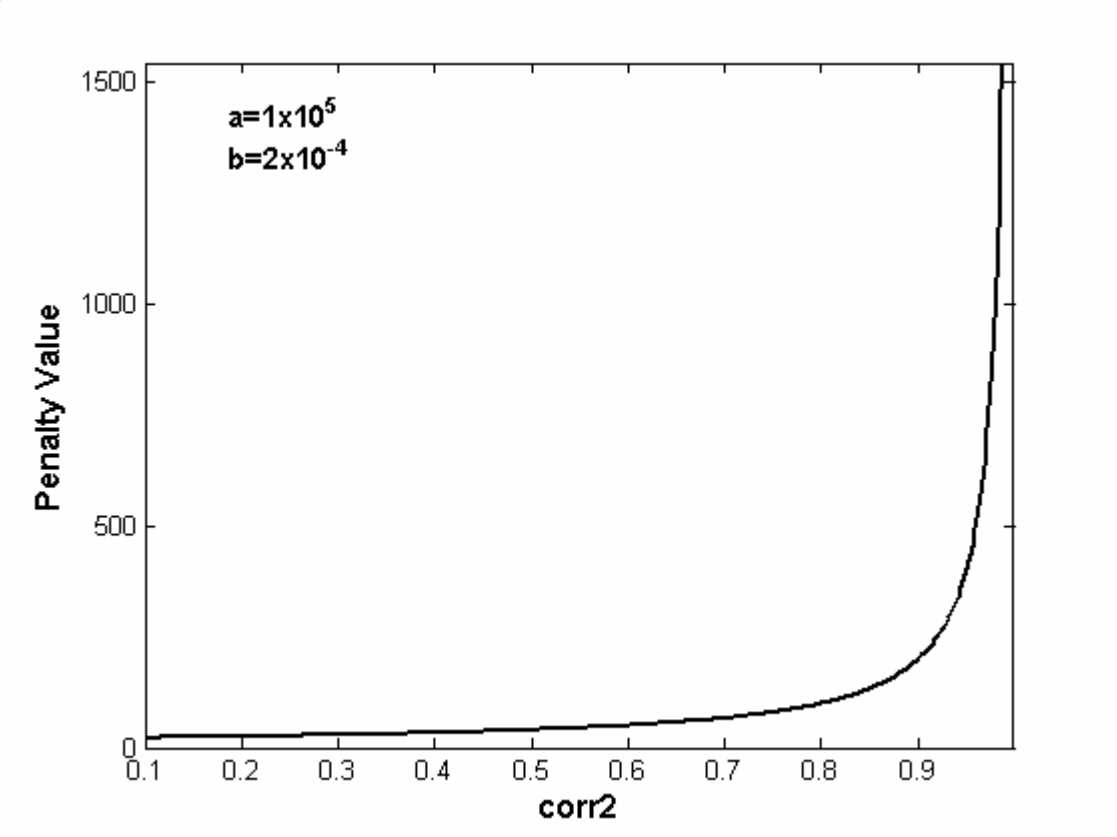


Figure 5.2. The sigmoid penalty function defined by the 2D correlation coefficient between two matrices.

If matrices X and Y are identical, then, the resulting value of $corr2$ will be unity. In practice, X and Y are the estimated 2D result during the current optimization and a previously determined pure spectrum respectively. For more similar matrices X and Y, the bigger the value of $corr2$ and thus the argument on the right hand side of Eq. 5.16. By taking the maximum of the set of values of the 2D correlation coefficients between all

combinations of X and Y, the penalty function prevents any identical reconstruction from occurring in subsequent optimizations.

As a smooth and continuous thresholding function, the sigmoid function is often used in artificial neural networks, where it is designed to introduce nonlinearity in the model and used as a transfer function. As shown in Figure 5.2, the function $corr2$ adopted here is a modified *sigmoid* function which rapidly approaches zero when X and Y become dissimilar and increase quickly when they are alike. This prevents similar 2D patterns from being reconstructed sequentially.

5.5.4. 2D Band Target Entropy Minimization (2D-BTEM)

In a similar spirit, the combination of band-targeting technique with 2D Entropy Minimization was adopted to achieve the recovery of one spectral estimate at a time, but with the targeted feature retained in the reconstruction. 2D-BTEM would have appropriate applications in the spectral reconstruction area where the targeted feature is of particular interest to the analyst.

As a first step, the data observations are decomposed into the basis matrices by SVD using the same procedure stated in section 5.4. In the second step, a careful identification of significant spectral features is required. Through a close examination of the series of right singular matrices, the coordinates of interesting spectral features are recorded. Finally, in the last step, by targeting the selected spectral feature, a 2D pure spectral pattern is obtained with the feature retained.

The objective function adopted here is similar to Eq. 5.15. But a new P penalty is used, which constrains the band-targeting approach to a specific 2D coordinate range. Analogically, other penalty functions, for example, to ensure non-negativities of pure

component spectra and concentration profiles, are allowed if necessary. In the 2D-BTEM algorithm, the use of various penalties in the objective function is often favored if the additional information is available. For example, if the approximation of concentration trajectories is known, so the unimodality or convexity constraint can be imposed.

5.5.5. Variation of the Objective Function

The repeated evaluation of the logarithm term, "*ln*" in Eq. 5.11 requires significant computational time. Previously the expression for *H* which omits the *ln* term, had been tested and the functions H' often yield good quality pure component spectra. In this new formulated objective function, the original $x \ln x$ format entropy term does not exist and it is replaced by *x*. Even though the new form of the entropy has changed, the principle behind the new entropy function is still consistent with the original intention of searching for the simplest patterns in the system. Accordingly, for vector-wise 2D Entropy Minimization or matrix-wise 2D Entropy Minimization, Equation 5.17 and 5.18 can be used respectively.

$$H' = \sum p \tag{5.17}$$

$$H' = \sum_m \sum_n p_{mn} \tag{5.18}$$

Further, if p is replaced with the intensity of the spectrum, then the integral representation of absorbance data in 2 dimensions will be a volume (the volume of the 2D spectrum). The volume minimization often yields good quality pure component spectra in practice.

5.6. Discussion

1. The objective of the proposed methodology is to solve the 2D spectroscopic inverse problem by using an entropy minimization algorithm. This approach includes several rather simple and straightforward steps, namely, (1) defining an appropriate entropy type objective function, (2) defining an appropriate measure to hinder replicate 2D spectrum, and (3) performing a one-result-at-a-time search.

2. The entropy type function and objective function chosen in the present study, obviously, aren't the only choices available. Indeed, there are a multitude of other literature and references that are very important in allowing us to understand entropy and entropy type functions (Frieden, 1975). A number of entropy functions relate to the statistical calculation involving histograms to evaluate the probability value. However, this approach normally does not take into consideration the spatial correlation between the elements in a 2D plane. The spatial correlation implies smoothness and continuity of the spectra in the 2D plane, which is composed of patterns with physical meaning rather that the random variables. The simple derivative function used in the present entropy type function, makes use of nearest neighbor pixel information. By minimizing the element-to-element variations (derivative or local curvature), one can obtain the simplest pure 2D pattern by minimizing the randomness between element values. It is also worthy to note that the objective function is not limited to the sum of the calculation of the entropy along the two canonical and orthogonal directions (x, y). It is also possible to evaluate entropy along diagonal elements in 2D patterns and this may be useful for some specific problems.

3. Instead of trying to solve for all the 2D patterns at once, 2D-BTEM takes a one-at-a-time approach represented by Eq 5.9 and 5.10. The concept of a one-result-at-a-time global search greatly simplifies a number of issues. First numerically, a single spectrum

reconstruction is simpler than an N-objective simultaneous reconstruction. Secondly, constraints to prevent redundant reconstructions in sequential searches are easier to implement than all constraints simultaneously in an N-objective rotation. Third, the user does not need to specify, *a priori*, the dimension or degrees of freedom in the search.

4. The final result involves the expectation for each spectrum $\hat{a}_{m\times n}$, with further corresponding expectation for concentration $\hat{C}_{q\times s}$ given by Eq. 5.19. The implementation of entropy minimization in 2D data analysis would help to achieve the relative concentration profiles which are fairly important in the chemical reaction studies.

$$\hat{C}_{q\times s} = A_{q\times(m\times n)}\left(\hat{a}_{s\times(m\times n)}\right)^{+} \tag{5.19}$$

5.7. 2D Testing of Hypothetical Factors by Target Transformation

Both SVD and PCA are characterized by decomposing the data matrix into bases that represent the data in another space. In practice, one individual (hypothetical) spectrum can be tested to see if it lies in the subspace spanned by the chosen abstract vectors resulted from the factorization of one data set. The essential purpose of target transformation is to form the projection of the target vector onto the subspace spanned by right singular vectors, and then compute the predicted vector using this projection. The identification process involves using a least-square procedure to minimize the deviation between the hypothetical vector and the predicted result. If the hypothetical vector and the calculated result are not significantly different, the hypothetical vector should be regarded as a real factor, otherwise, it would not be a real factor. Target transformation would help to serve as a means of identifying the presence of a hypothetical vector, normally a suspected pure component spectrum. The hypothetical spectrum can be either the real

spectrum from an experimental result, a spectrum in a library, or the theoretical result from a quantum chemistry calculation.

The inherent mathematical relation between abstract basis and real chemical solutions is revealed by target transformation. From a spectral reconstruction viewpoint, the target transformation technique is concerned with the connection between the abstract spectral bases and physically meaningful or experimentally obtained pure component spectra.

In section 5.7.1 that follows a brief review of 1D Target Transformation from the literature. Then in sections 5.7.2, 2D and 3D Target Transformations are developed. As far as I am aware, the results of sections of 5.7.2 are new.

5.7.1. 1D Target Transformation

Developed during the 1960s, the target transformation technique attempts to find linear transformations or oblique rotations of the set of basis vectors to test physical models (Malinowski, 1991). In addition to target transformation, there are many kinds of transformation exist, such as quartimin, quartimax, oblimax, varimax, etc. (Rummel, 1970, Richman, 1986). But target transformation is used to rotate the basis vectors to produce the spectrum that best fits the test spectrum.

Assume that we have a new reaction experiment. After SVD, the spectra from the new experiment give a set of right singular vectors V^T. Linear combinations of these right singular vectors can be made using a transformation vector *t*. Let us call this new estimated spectrum $\hat{a}$. Schematically, this can be shown as Figure 5.3.

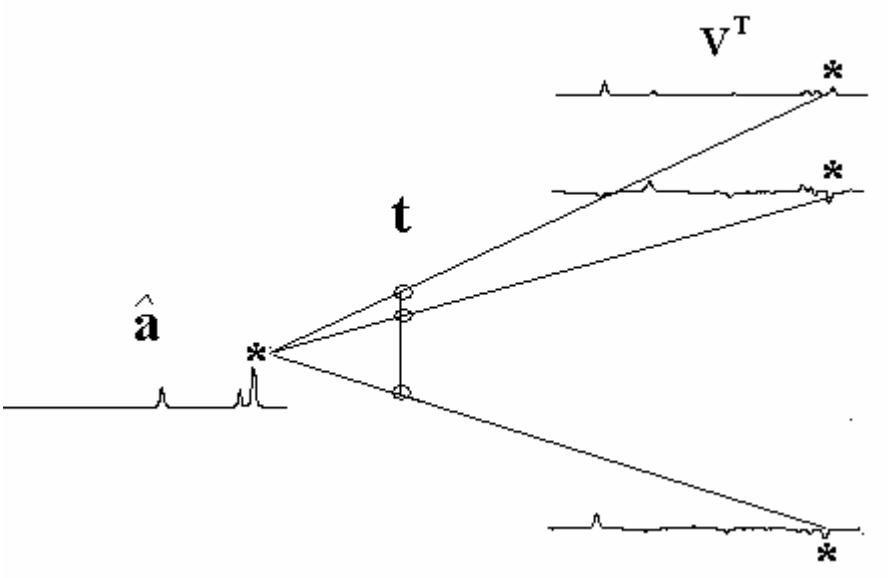


Figure 5.3. A scheme representing a linear combination of right singular vectors which gives an estimated spectrum $\hat{a}$.

If $\hat{a}$ is an estimate, then the linear algebraic expression is given by Eq. 5.21.

$$\begin{bmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \vdots \\ \hat{a}(n) \end{bmatrix}^T = \begin{bmatrix} t(1) \\ t(2) \\ \vdots \\ t(s) \end{bmatrix}^T \times \begin{bmatrix} V^T(1,1) \cdots V^T(1,n) \\ V^T(2,1) \cdots V^T(2,n) \\ \vdots \\ V^T(s,1) \cdots V^T(s,n) \end{bmatrix} \tag{5.21}$$

Now let us introduce a spectrum a from a library which we are interested in testing. The difference Δa between the test spectrum a and the estimated vector $\hat{a}$ can be written as:

$$\Delta a = a - \hat{a} \qquad \text{or} \quad \begin{bmatrix} \Delta a_1 \\ \Delta a_2 \\ \vdots \\ \Delta a_n \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} - \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_n \end{bmatrix} \tag{5.22}$$

The difference between the test spectrum and an estimate should be minimized. This requires the differential equation (Eq. 5.23).

$$\frac{d(\Delta a)^2}{dt} = 0 \tag{5.23}$$

After some mathematical operations, the transformation matrix t can be solved exactly as (5.24) and the spectral estimate $\hat{a}$ is given by (5.25).

$$t = a\ (VV^T)^{-1}V \tag{5.24}$$

$$\hat{a} = t\ V^T \tag{5.25}$$

Finally, one compares the similarity between the test spectrum a and the best estimate $\hat{a}$. This is easily done using e.g. the visual inspection, an inner product, etc. If the estimate $\hat{a}$ is in close agreement with the test spectrum a then there is strong evidence that that species exists in the reaction system.

5.7.2. The Extension to 2D and Higher Dimensions

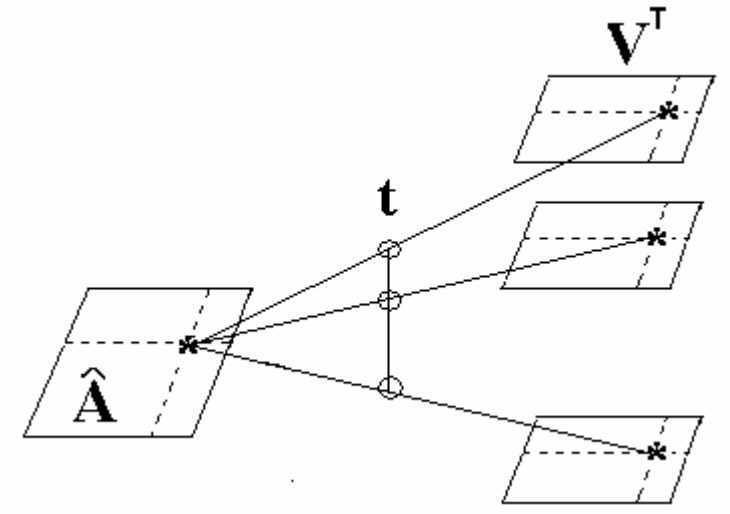


Figure 5.4. A scheme representing a linear combination of right singular matrices which gives an estimated spectrum.

2D target transformation can also be proposed since the 2D decomposition exists. By this 2D projection, the predicted matrix can be computed and compared with the

original 2D spectrum/pattern. 2D target transformation provides an efficient means of identifying possible components contained inside a series of mixture samples. There is also a practical need to check whether one individual 2D spectrum lies in the subspace spanned by the chosen abstract singular matrices (or observations).

After unfolding each 2-dimensional matrix with *m* rows and *n* columns into a vector (with $m \times n$ elements), we can follow the previous procedure easily. When all 2-dimensional matrices are unfolded, a 2-way array is obtained. Sequentially, the V^T vectors are calculated by SVD.

The test spectrum A is also unfolded into a vector a . Sequentially, similar procedures as stated in section 5.7.1 can be used to obtain a estimated vector $\tilde{a}$ by a linear combination of the V^T vectors which is the best approximation to the vector a (obtained from unfolding the test spectrum: matrix A) (Eq. 5.26). Further, the predicted vector $\tilde{a}$ can be unfolded back to the predicted matrix $\hat{A}$ with *m* rows and *n* columns. Finally, one can compare the similarity between the test spectrum A and predicted $\hat{A}$ and decide if the test spectrum is present in the test data set or not.

$$\tilde{a}_{1\times(m\times n)} = t\ V^T_{q\times(m\times n)} \tag{5.26}$$

Further we can extend this method into higher dimensions. For example, in three dimensions, all 3D tensors can be unfolded into *concatenated* vectors in a systematic way. Right singular arrays are obtained afterward. Similar procedures in section 5.7.1 are adopted to predict a estimated vector $\bar{a}$, which is regarded as the best approximation to the vector a (obtained from unfolding the test spectrum: tensor $\underline{A}$)(Eq. 5.27). Further, the estimated vector $\bar{a}$ can be unfolded back to the tensor $\hat{\underline{A}}$ with *m* rows, *n* columns and *o* slices. The illustration is shown in Figure 5.5.

$$\underline{\bar{a}}_{1\times(m\times n\times o)} = t\ V^{T}_{q\times(m\times n\times o)} \tag{5.27}$$

Finally, one can compare the similarity between the test spectrum $\underline{A}$ and predicted $\underline{\hat{A}}$ and decide if the test spectrum is present in the test data set or not.

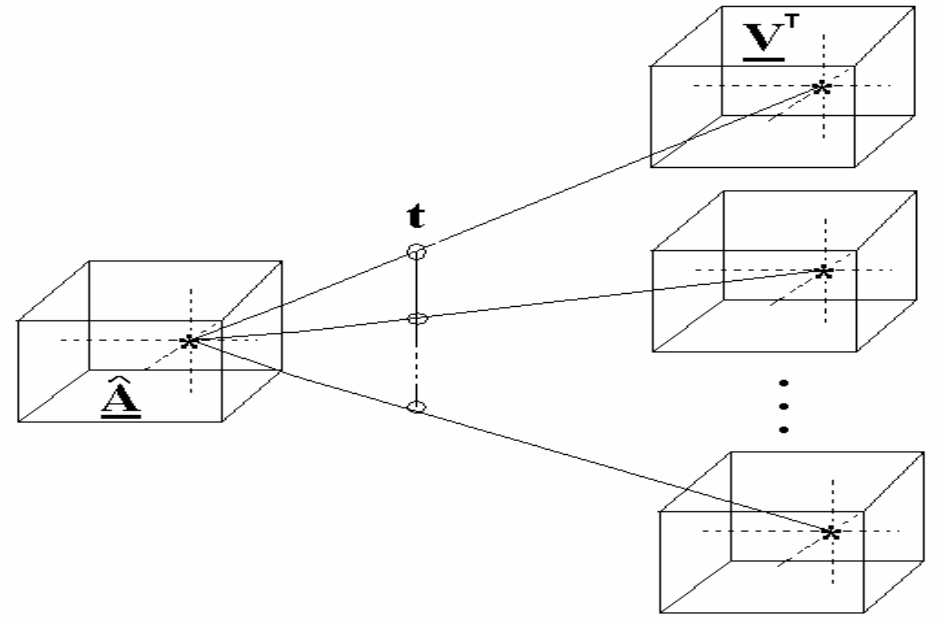


Figure 5.5. A scheme representing a linear combination of right singular array which gives an estimated three-way tensor.

Although quite a few references can be found to 1D target transformation in the literature, the concepts of 2D and 3D target transformations, and their implementation appear to be new.

5.8. Summary

In this chapter, the extension of 1D entropy minimization based curve resolution algorithm is presented. The details of system representation of the 2D spectroscopic data set, the singular value decomposition and model reduction of 2D spectroscopic data are illustrated. The procedure about objective function formulation and entropy type function

selection are brought into discussion. 2D-BTEM is a combination of band-targeting with Entropy Minimization methodology, both in 2D aspects. The band-targeting makes sure that each feature is retained in the final reconstruction.

Also, in this chapter, the concept of 1D target transformation is extended to 2D and higher dimensional domains. Target transformation can be used as a tool to test whether a hypothetical spectrum can be regarded as a real spectrum which is embedded inside the data set. The extension makes the technique available for higher dimensional data analysis.

Chapter 6

2D Entropy Minimization Algorithm —Application to Simulated Data and Image Signal Processing

In this chapter, the 2D Entropy Minimization algorithm and 2D Band-Target Entropy Minimization developed in chapter 5 are applied to simulated data as well as image data. This chapter is divided into three major sections. The first section describes the separation of a set of matrix mixture data generated from three texturally different matrices. Further, the second section presents the application of Entropy Minimization algorithm to simulated 2D mixture spectra. In the third section, image reconstruction analysis is also explored; image mixture separation is demonstrated in detail.

The first problem encountered in 2D data is the visualization of matrix-formatted data, in other words, the way to represent such a spectrum on paper. Generally speaking, there are two major techniques used to visualize the matrix, namely, the mesh routine and contours routine. Both are implemented as the build-in functions in MATLAB. The function “mesh”, draws a wire frame mesh with colour, and colour is defined by surface height. Another function “contour”, in the same way as a topographical map, makes a plot in which the intensity of the peaks is represented by contour lines drawn at proper intervals. It is easier for a user to find the feature, if we plot the matrix in a 3D view with “mesh” function. On the other hand, if there are a multitude of features in the plot, the smaller ones behind the front peaks may not be seen clearly, the contour plot would avoid this problem with a bird's eye view.

6.1. The Use of Entropy Minimization for Matrix Mixture Separation

In this section, a generalize source separation problem is formulated. A set of mixtures originating from three texturally different matrices were simulated in order to explore and test the performance of the 2D Entropy Minimization method.

6.1.1. Data Simulation

The numerical experiment began with the simulation data of matrix mixtures. The first matrix was a sparse matrix with only several entries that were non-zero (Sparse Matrix). The positions and magnitudes were arbitrarily chosen on the interval [0 1]. The second one (Tri-diagonal Matrix) was a banded matrix whose entries were all zeros except the ones on the diagonal and adjacent. The positions were fixed by the fact that it must have a tri-diagonal structure, and the magnitudes were chosen arbitrarily. The third one (Random Matrix) was a random matrix produced with the MATLAB function "rand" [i]with elements between [0,100]. These three original pure matrices are shown in Figure 6.1. The values of the matrix-wise entropies for the original matrices were 14.826(Random Matrix), 12.411 (Tri-diagonal Matrix) and 2.079 (Sparse Matrix), respectively.

A random mixing matrix with non-negative entries was generated. This mixing matrix is shown in Eq. 6.1 and the mixture matrices are shown in Figure 6.2 with the mesh plot. Note that these mixtures are very similar, because of the large magnitudes involved in the original Random Matrix.

$$A = \begin{pmatrix} 0.6435 & 0.7266 & 0.2679 \\ 0.3200 & 0.4120 & 0.4399 \\ 0.9601 & 0.7446 & 0.9334 \end{pmatrix} \tag{6.1}$$

[i] If not specific further, the simulated non-negative random matrix are all produced with MATLAB function "rand".

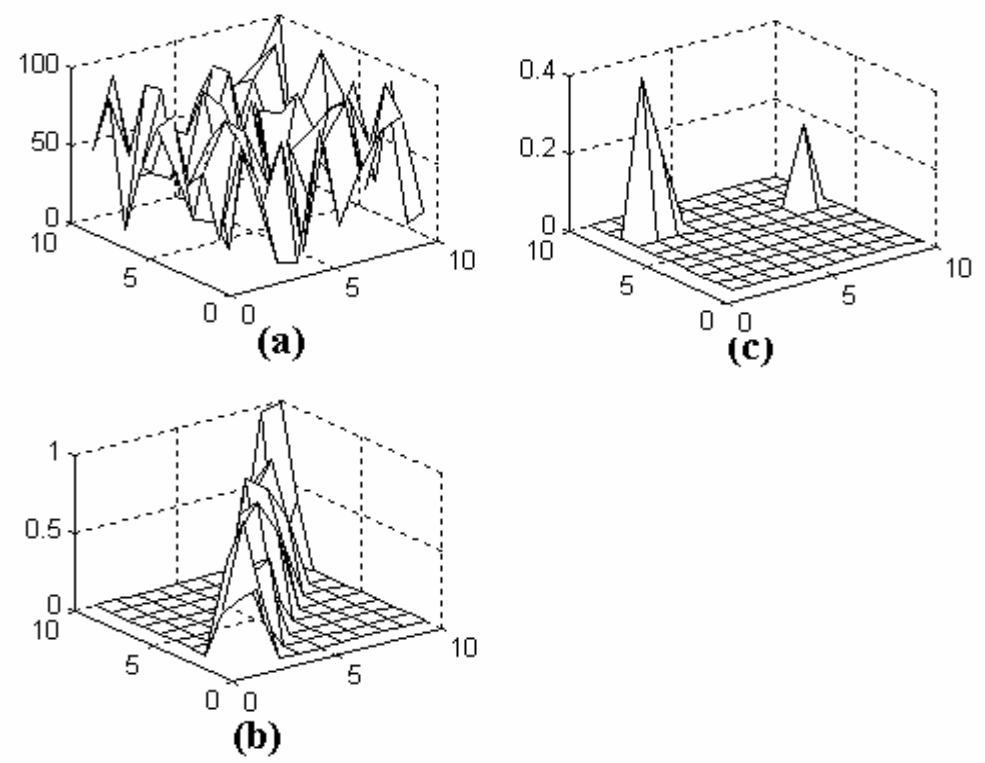


Figure 6.1. The mesh plot of pure matrices: (a) Random Matrix, (b) Tri-diagonal Matrix and (c) Sparse Matrix.

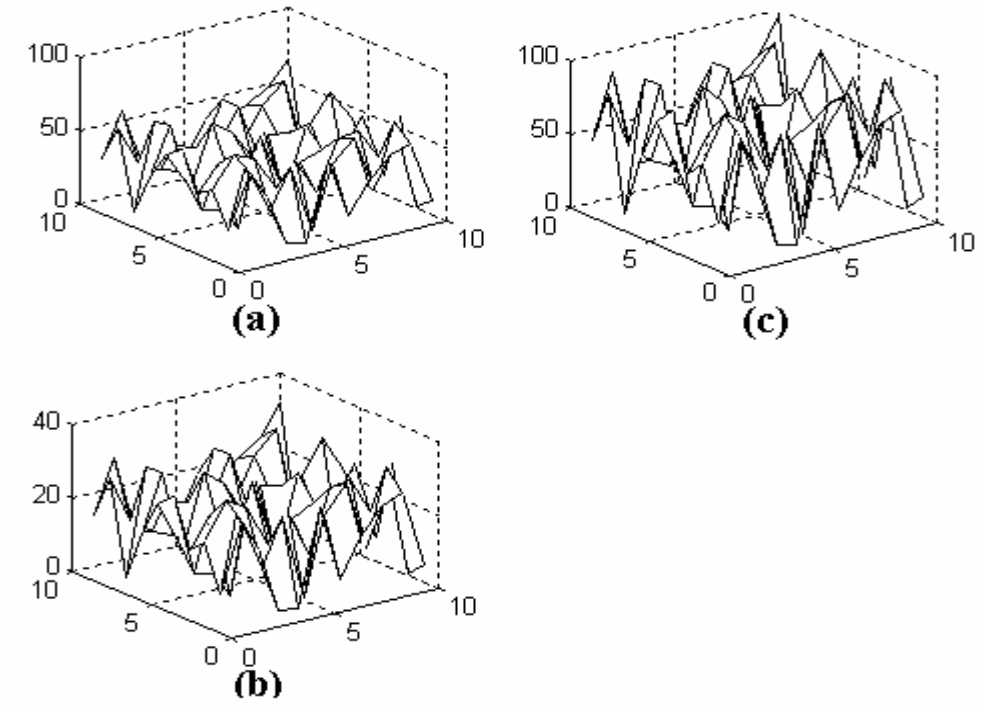


Figure 6.2. The mesh plot of the mixture matrices (a) ,(b) and (c).

6.1.2. Result

The three mixture matrices were analyzed with the 2D Entropy Minimization algorithm. First, all these three matrices were decomposed into right singular matrices using SVD procedure described in chapter 5, section 5.4.2. Secondly, the right singular matrices were transformed into the pure source matrices with 2D Entropy Minimization algorithm in a systematic way. Thirdly, 2D-BTEM also was applied to these right singular matrices to obtain the pure matrices.

6.1.2.1. Result of 2D Entropy Minimization

The mesh plots of right singular matrices are shown in Figure 6.3. The set of right singular vectors were then subjected to a global search for minimum entropy. This first estimated matrix result was the “Sparse Matrix”. After that, the second result, which should not be identical with the first one, will be sought using the search strategy adopted in the 2D Entropy Minimization with the dissimilarity penalty. In the present adaptation to matrix formatted data, and in order to take advantage of the natural 2D structure, we sequentially take the first derivative in two directions (Eq. 5.14). In other words, the entropy like function H used in matrix-wise 2D entropy minimization estimates the smoothness of the 2D data in 2 dimensions.

The mesh plots of resulting set of 3 recovered matrices via 2D Entropy Minimization with dissimilarity constraints, one with a global entropy minimum and two others with local entropy minima are shown in Figure 6.4. The values of the matrix-wise entropies for these recovered matrices were 8.331 (Sparse Matrix), 12.411 (Tri-diagonal Matrix) and 14.285 (Random Matrix), respectively.

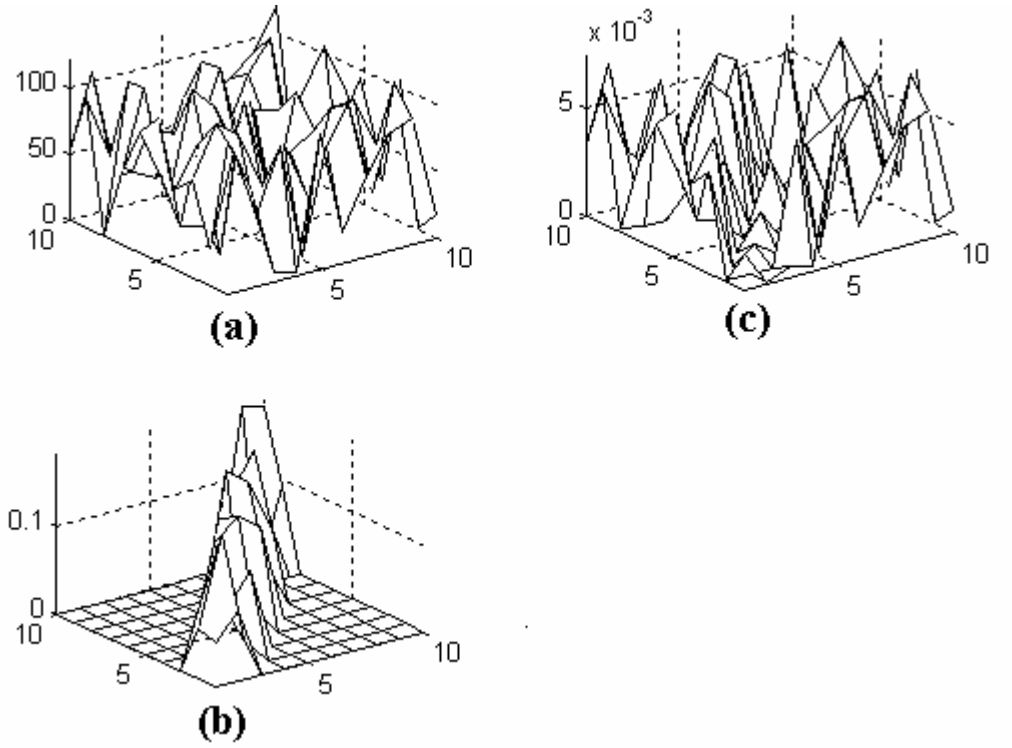


Figure 6.3. The mesh plot of 1st (a), 2nd (b) and 3rd (c) right singular matrices obtained from the mixture matrices via SVD decomposition procedure.

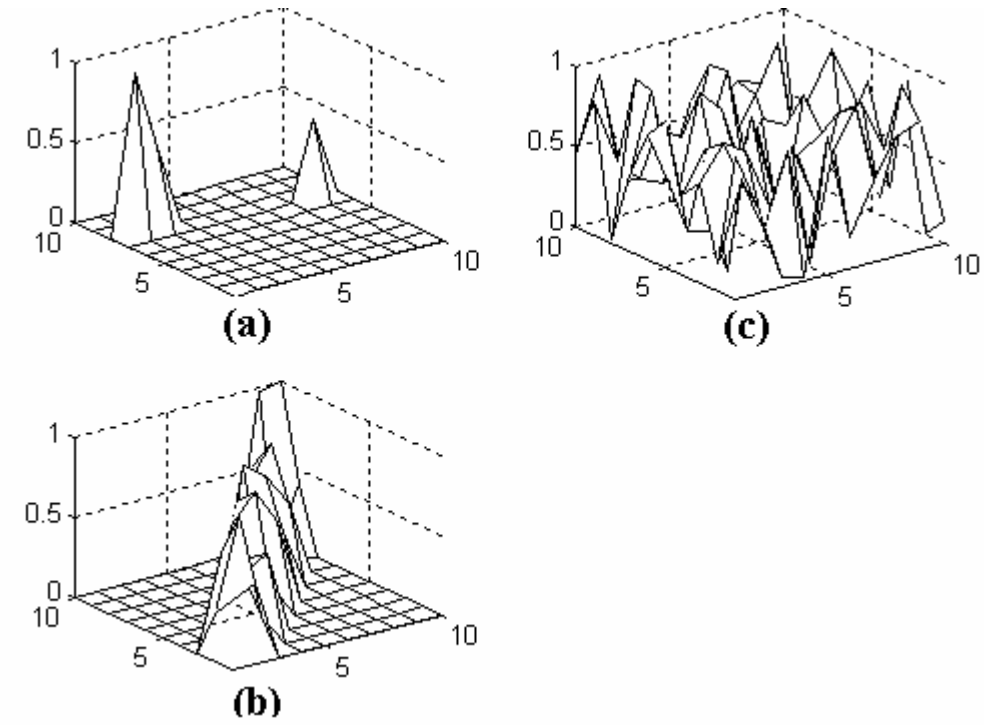


Figure 6.4. The mesh plot of recovered matrices (a), (b) and (c).

In order to examine the quality of the reconstruction, the recovered matrices were compared with the original matrices by calculations of the 2D correlation coefficient (*corr*2). The result is shown in Table 6.1.

Table 6.1 Comparisons between the recovered results and references

Simulated matrices	**Entropy value of Original matrix**	**Entropy value of recovered result**	***corr*2**
Random Matrix	14.826	14.285	1.00
Tri-diagonal Matrix	12.411	12.411	1.00
Sparse Matrix	2.079	8.331	0.986

There is a noticeable difference between the entropy of the original Sparse Matrix (2.0794) and the recovered one (8.3305). The computation error can be shown to arise from small non-zeros entries. In the original Sparse Matrix, all of the entries are zeros except three points, with the definition of entropy of matrix, the entropy is extremely low. But the result of computation inevitably produces very small values instead of zero entries, which will result in a large accumulation of entropy. A numerical experiment was carried out to calculate the entropy of the projected result from right singular matrices with the pure Sparse Matrix as the hypothetical matrix by target transformation technique (chapter 5, section 5.7). The matrix-wise entropy of the final projected result is 8.7258 with summation of fitting errors, 2.519888e-027. A similar calculation of the projected result from original mixture matrices with the pure Sparse Matrix as the test matrix produces the result with entropy of 8.6872 and error 5.825685e-019. There seems to be some sort of sensitivity issues involved. Although very accurate entries are obtained for the individual elements of the Sparse Matrix, the entropies remain somewhat different.

6.1.2.2. Result of 2D Band-Target Entropy Minimization

One of the purposes of the simulation of three matrices with discernably different features is to test the possibility of utilizing these features to facilitate the recovery. As mentioned before, 2D-BTEM is designed to recover patterns with the selected features retained in the result.

From inspection of the right singular matrices, it is apparent that the 2nd right singular matrix possesses a unique and prominent feature with a diagonal ridge cross the plane. Actually it is quite similar to the first pure matrix component in Figure 6.1, even though we don't have any knowledge about this before the recovery is accomplished. We can easily select this feature, and the first estimated result can be obtained by targeting the region with [1 to 3; 1 to 3] or [6 to 8; 8 to 10][ii]. This result is shown Figure 6.5a. It is not so obvious to find a practical second feature in the series of right singular matrices (This is not often the case; explanation follows in discussion part). A small trick is needed to pick out another feature of interest. The third right singular matrix (c) in Figure 6.3 is subtracted from the first recovered result (Tri-diagonal Matrix). The latent pattern appears when the proper subtraction is performed, at the same time; the feature of Tri-diagonal Matrix fades. A trial and error method can be used here. Figure 6.5b shows the result of subtraction (for better view, the figure was flip upside down by plotting its negative transformation as shown in Figure 6.5c), it is clear that another interesting feature with a pyramid at the left corner is found. Further band-targeting is now possible. The new estimate with a feature located around [7 to 9; 1 to 3] was obtained via 2D-BTEM and shown in Figure 6.5d.

[ii] Consistent with the contour plot of the matrix in MATLAB, in this thesis, the format of a targeted region is represented as [y-coordinate range; x-coordinate range]

Obviously without any feature, the Random Matrix can not be targeted out via 2D-BTEM at present.

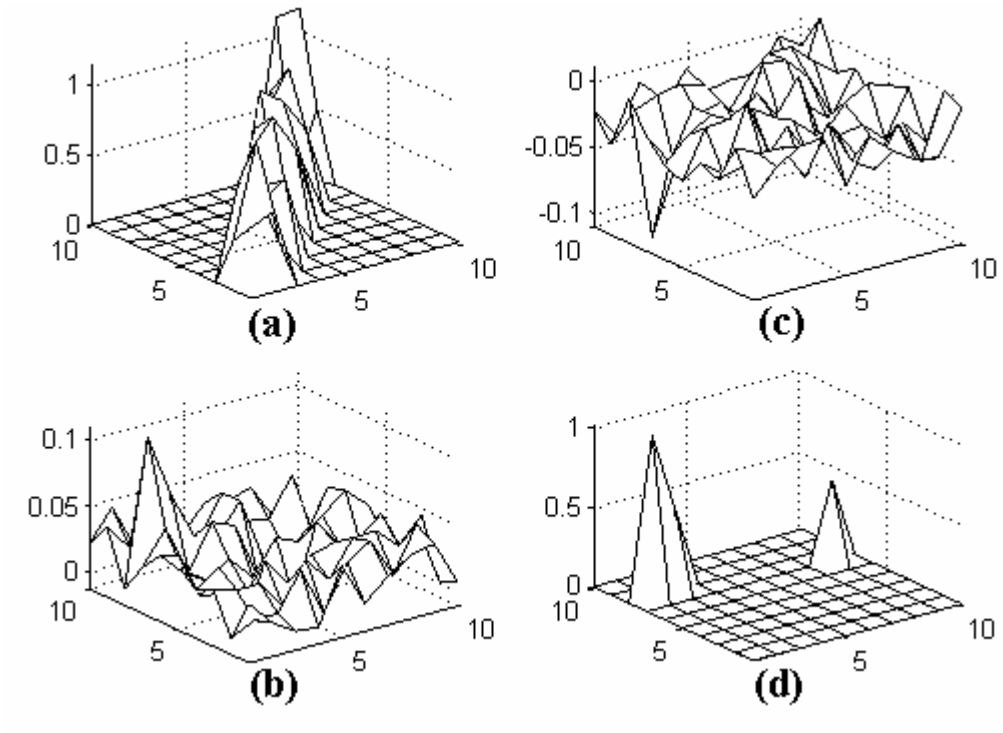


Figure 6.5. The mesh plots of the first 2D-BTEM result (a); the latent pattern found by subtracting the third right singular matrix by Tri-diagonal Matrix (b); matrix b with sign change (c); the second estimate obtained via 2D-BTEM (d).

6.1.3. Discussion

Mixing matrices In order to compare the quality of the reconstructions, the original and recovered final mixing matrices are compared. A mixing matrix can be recovered by least-square fit to the original mixture data with the estimated results from 2D Entropy Minimization. The final estimated mixing matrix is shown in Eq. 6.2.

$$A_{estimated} = \begin{pmatrix} -2.4668 & 11.926 & 63.062 \\ -1.1042 & 5.9814 & 31.363 \\ -3.4671 & 17.457 & 94.090 \end{pmatrix} \tag{6.2}$$

For easy visual comparison, each column was normalization by making the second row unity. One issue that should be pointed out is that the order of the estimated result used in the calculation is not same as the corresponding ones in mixing so rows have to be shuffled. So the final estimated mixture matrix is given in Eq. 6.3 and rescaled original mixing matrix is given in Eq. 6.4.

$$A_{est\_shifted} = \begin{pmatrix} 2.0107 & 1.9938 & 2.2341 \\ 1.0000 & 1.0000 & 1.0000 \\ 3.0000 & 2.9186 & 3.1399 \end{pmatrix} \tag{6.3}$$

$$A_{mix} = \begin{pmatrix} 2.0107 & 1.7639 & 0.6091 \\ 1.0000 & 1.0000 & 1.0000 \\ 3.0000 & 1.8074 & 2.1217 \end{pmatrix} \tag{6.4}$$

It is clear that the first column of the original and the calculated mixing matrices are very close. Also the middle columns of both matrices are similar, but the last column differs. The reason lies in the extremely low contribution of Sparse Matrix to the mixture (corresponding to the third column) compared to Random Matrix (corresponding to the first column). The maximum value of the entries in Sparse Matrix is 0.4 and most of the entries are zeros while the maximum value of Random Matrix is 99 and all entries are non-zero. The contribution of Random Matrix component (average value 52.8) is about 7000 times than the Sparse Matrix (average value 0.008) that would probably make the least-square regression to fail to predict precisely the correct weighting due to the ill-conditioning.

Also this is the main reason why it is difficult to find out the second feature in the series of right singular matrices.

2D BTEM and 2D entropy Minimization Both of these two algorithms were tested in this example. It is clear in the example, that the 2D entropy Minimization can be used for matrix separation. Each result is extracted one by one and at each step, previous results are used to prevent the trivial result from repeating. Also using 2D-BTEM, we can extract the specific results for feature that can be detected and targeted in the right singular matrices one-at-one-time.

A little consideration has been given to applied problems that can be represented by different weightings of matrices and which might benefit from inverse solutions of this type. There is one area where the present development might have an application. In some areas of mathematical physics, there are problems which are represented by operators or sets of operators (i.e., in optics, scattering theory etc). Perhaps there are problems where one obtains a complex observation which arises from a few sources simultaneously. In such problems, if one could vary the contributions, it might be possible to recover the operators that are giving rise to the observations.

6.2. The Use of Entropy Minimization for the Solution of Simulated Five-Component Spectral Mixture Data

As shown in the section 6.1, the mixture matrices separation was approached by 2D Entropy Minimization method which included a sparse matrix, tri-diagonal matrix and a random structure matrix. In this section, simulations of absorptive spectra were performed in order to illustrating and testing the performance of the entropy minimization method.

6.2.1. Simulation

6.2.1.1. Numerical Simulation with 2D Pearson VII Model

A simple 5-component system was simulated. It is known that several methods have been devised so far to accomplish the peak modeling by means of Gaussian, Lorentzian, Voigt, Pearson VII and other models. Pearson VII is one of useful band shape functions. As shown in Eq. 6.5,

$$a(v) = \frac{K}{\left[1+\left(\frac{2(v-v^0)\sqrt{2^{-M}-1}}{W}-\right)^2\right]^M} \tag{6.5}$$

where v denotes wavenumber; v^0 denotes the peak center position; M denotes the Pearson width parameter and K denotes the amplitude. When M approaches 1, Pearson VII resembles a Lorentzian model and it would become Gaussian model when M approaches infinity.

Obviously, Pearson VII model is more flexible than the Guassian and Lorentzian model. For 2-D Pearson VII spectra modeling, the individual 2D peak can be formulated as:

$$Peak_i(v_x, v_y) = \frac{K_x}{\left[1+\left(\frac{2(v_x-v_x^0)\sqrt{2^{-M_x}-1}}{W_x}-\right)^2\right]^{M_x}} \frac{K_y}{\left[1+\left(\frac{2(v_y-v_y^0)\sqrt{2^{-M_y}-1}}{W_y}-\right)^2\right]^{M_y}} \tag{6.6}$$

A 2D spectrum can be modeled by superimposing several 2D peaks located in different positions. Especially if M_x equal to M_y, W_x equal to W_y, it would be a symmetric peak with exact same shape in x direction and y direction.

6.2.1.2. Numerical Simulation of 2D Spectra

In this study, five 2D spectra, each with size of 100 by 100, were simulated according Eq. 6.6. The positions and parameters were arbitrarily chosen. Fifteen mixture spectra were simulated from these five initial "pure" spectra with a random mixing matrix (15 by 5) with "rand" function. Also fifteen sets of noise matrices (normally distributed random numbers with mean zero, variance 9 and standard deviation 3, simulated with MATLAB function "randn") with the same array size were added in order to mimic the real world data.

As shown in Figure 6.6, the first five are the pure component spectra without noise. A close examination of one of the 15 mixture spectra with added noise (shown in the right bottom of Figure 6.6) reveals circa 20 peaks crowding one spectrum with heavy overlapping.

Note that a new representation method is used – one that resembles NMR. Each figure is a composite which has 3 sub-figures, where the large sub-figure is the 2D contour map, the left sub-figure is the 1D projection along left-right direction, and the bottom sub-figure is the projection along up-down direction.

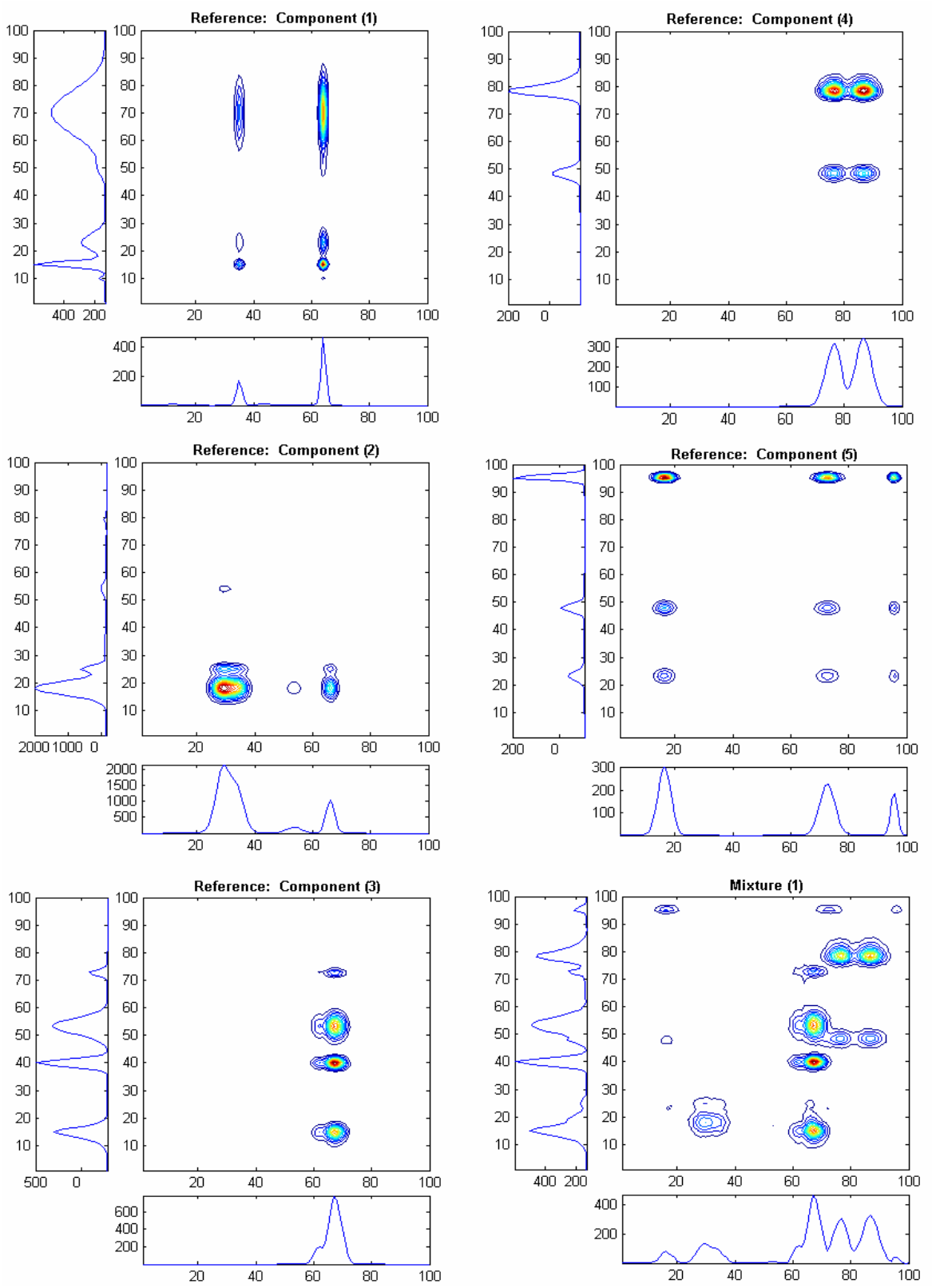


Figure 6.6. The contour plot of the five pure simulated 2D spectra (component 1-5) and one mixture spectrum with added noise (bottom-right).

6.2.2. Result and Discussion

Fifteen right singular matrices $\underline{V}^{T}_{15\times100\times100}$ were obtained with the SVD procedure developed in this thesis. The resulting right singular matrices are shown in Figure 6.7. Physically meaningful spectral features were observed in only the first five right singular matrices. The sixth matrix is essentially featureless, which holds true for the matrices seven-fifteen as well. The variances of the sixth to 15th right singular matrix were calculated and they are 9.5450, 9.4115, 9.2021, 9.0262, 9.0130, 8.9045, 8.7248, 8.6423, 8.6034 and 8.3909. Their average value of the variance (8.9464) is very close to variance of the noise (~9.0) which was added during simulation procedure. It can be interpreted that the first five right singular matrices already can explain all the "physical" variance (except the noise) in the fifteen sets of data. Additionally, the original fifteen spectra are reduced into five right singular matrices which contain almost all useful information. The remaining ten are nothing but noise. Therefore it is reasonable to discard all the other ten right singular matrices.

From visual inspection of the right singular matrices, several spectral features are identified and marked with arrows in Figure 6.7.

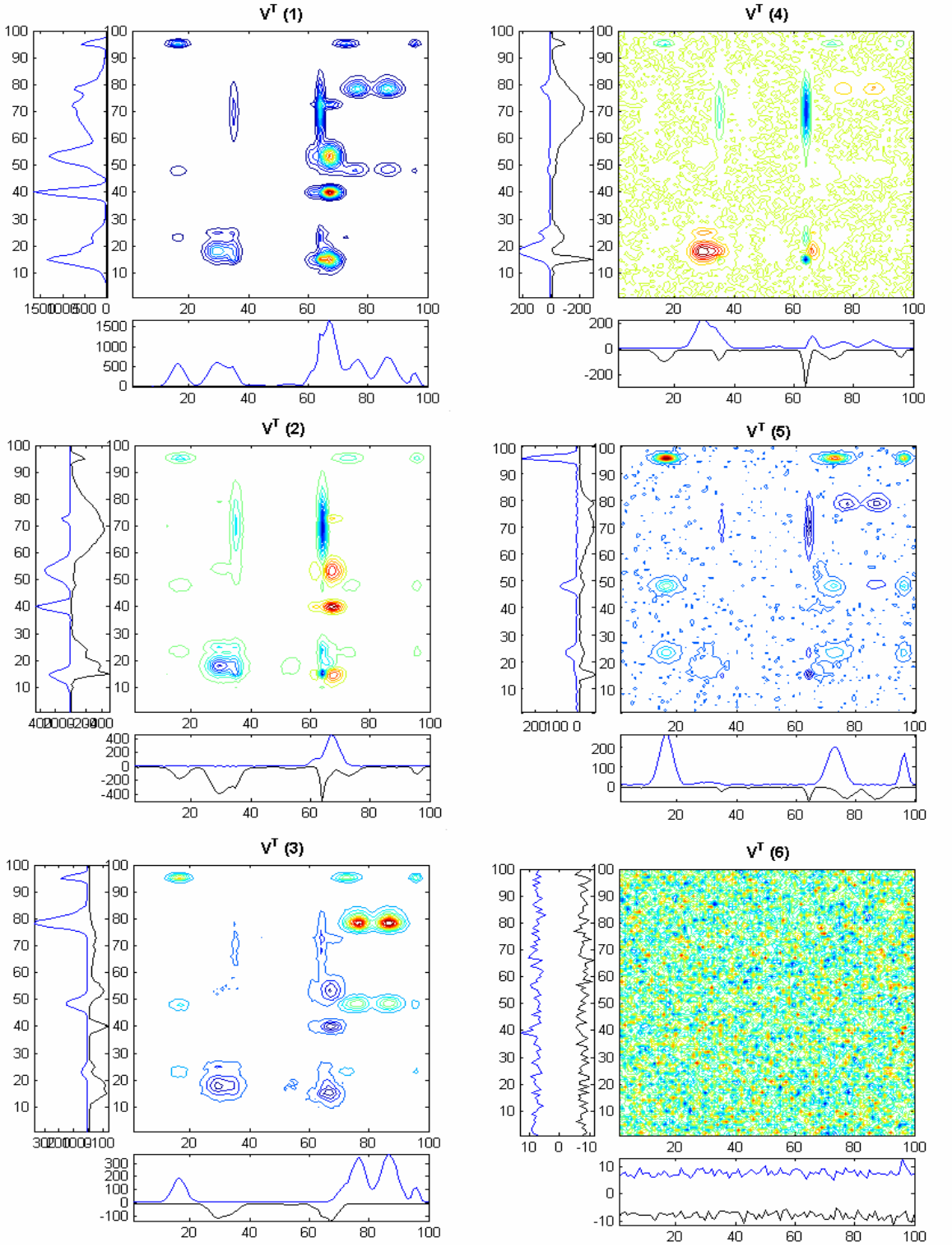


Figure 6.7. The resultant right singular matrices(1st to 6th). Several spectral features are marked with arrows. Note that yet another representation is now introduced where the left and bottom 1D projection possess two lines for positive and negative contributions.

2D-BTEM was performed using $\underline{V}^{T}_{5\times100\times100}$ by targeting observable features in the right singular matrices one at one time.

Two types of objective functions were implemented. In the first objective function, only a volume term is used, but in second objective function, the derivate cost is added. Reconstruction results are close, and the former results are presented in Figure 6.8 with individually targeted region 1 ([10 to 30; 25 to 35]), region 2 ([10 to 30; 60 to 70]) , region 3 ([50 to 60; 60 to 70]), region 4 ([62 to 85; 30 to 40]), region 5 ([70 to 90; 80 to 90]) and region 6 ([90 to 100; 70 to 75]). It was found that the result from targeting region 4 was nearly the same as targeting region 2 since these two targeted features belong to the same pattern (they are correlated). Exhaustive searches proved that only five separable and irreducible 2D spectral patterns are present.

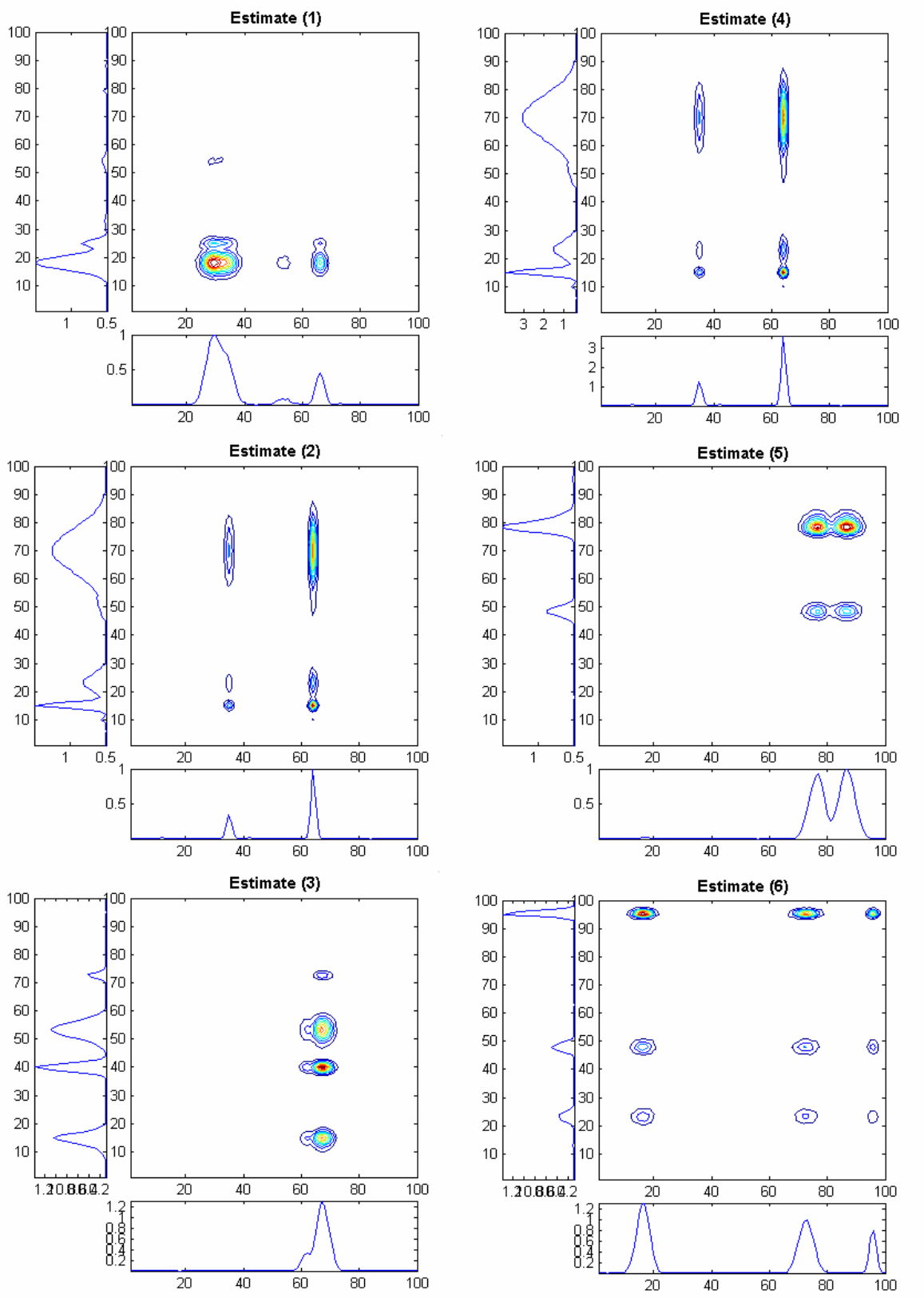


Figure 6.8. The resolved spectra via 2D-BTEM by targeting the feature peaks shown in right singular matrices

The recovered pure 2D spectra can be used to solve the concentration profile. For the purpose of comparing the quality of the reconstructions, the relative concentrations were obtained with the 2D-BTEM results and compared with the original mixing matrix. As shown in Figure 6.9, the normalized concentration profiles of both the real and estimated results are very close. This is a solid evidence for the high quality performance of the recovery method.

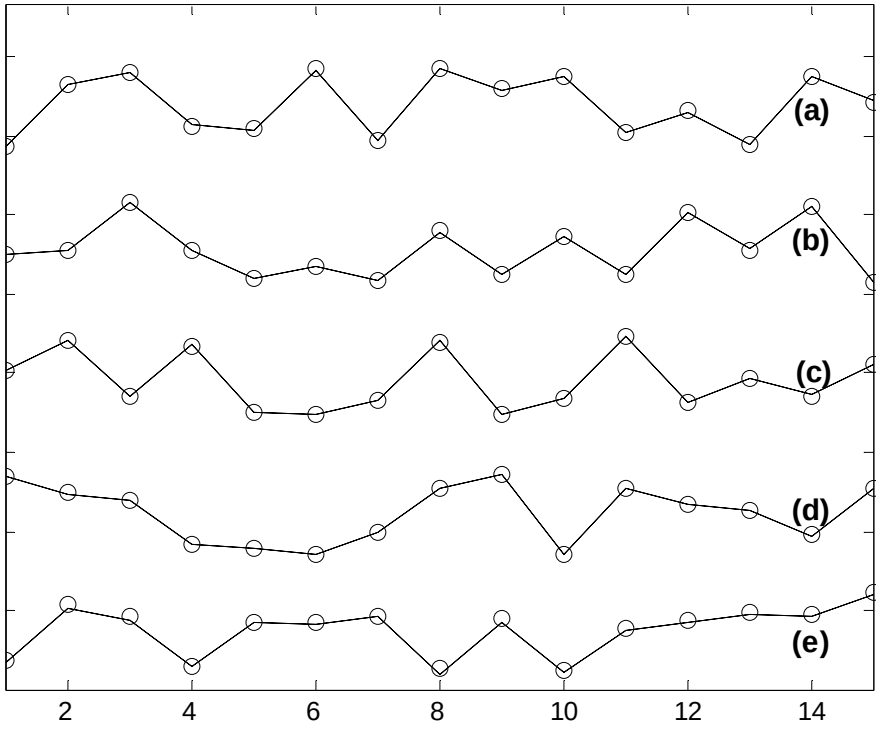


Figure 6.9. L^2 normed concentration of five components (a, b, c, d and e corresponding to the reference components 1-5) associated with the 15 simulated mixtures. Circles: original mixing loading. Solid line: estimated loading.

6.2.3. Summary

It is shown in this simulated example that entropy based spectral reconstruction method can be implemented to analyze the linear mixing multi-component mixture systems. This encouraged the application to real systems.

6.3. The Application of Entropy Minimization for Blind Source Separation Problems in Image Analysis

6.3.1. Introduction

Among the various types of 2D data arrays, images represent the most typical 2D data arrays, and images are now commonplace in a number of diverse fields such as biological research, medical diagnosis, metallurgy, remote sensing etc. The problem of exacting the sources in the blind system represents a particularly difficult type of inverse problem(Sabatier, 1978), where the only information is the mixture patterns. In the general sense, and also in the most challenging form, *a priori* information concerning either the individual source patterns or the number of sources giving rise to the observations are unavailable. A variety of approaches have been proposed to solve the associated blind source separation problem in the electrical engineering and image processing literature. These include high order statistics (Cardoso, 1993)**,** mutual information maximization approach (Infomax) (Bell and Sejnowski, 1995), and nongaussianity approach (FastICA) (Hyvärinen, 1999), etc. It should be noted that a lot of work has been done in this area.

As already apparent, entropy can be used as a measure of information. It has been used somewhat in image processing, but in a very different way than this thesis uses it. Therefore, as an important informatics character, entropy is closely associated with pattern

recognition. In this section, for the first time, this 2D entropy minimization algorithm is applied to three different types of image blind source separation problems. The first case involves 3 texturally dissimilar black and white images (photographs). The second involves 3 geometrically similar color images (photographs). The third case involves image enhancement for an underdetermined problem. In the first 2 cases, outstanding blind source recovery is achieved and in the last case significant image enhancement is observed. These results appeared in *Pattern Recognition* (Guo and Garland, 2006). A reprint can be found in Appendix E.

6.3.2. Results

6.3.2.1. Analysis of Texturally Different Images

Three texturally dissimilar images with size of 128×128 pixels, consisting of a building, a fabric and a tile, were downloaded from the MIT VisTex public database[iii]. The present reference images have been used by other researchers as a test case for the development of other blind source separation approaches (Hashimoto, 2002). Although the recovered images are reasonably good, the final recovered images possess quite a few visible inconsistencies, and an error rate of 6.2% is stated. One purpose of this study is to compare the performance of Entropy minimization with other methods.

These color images are stored as a sequence of truecolor RGB triplets, i.e. separate red, green and blue layers (in other words, there are 3 matrices associated with each color image). Without loss of generality, only the "Red" layer was used as the pure images in this study. Three images are displayed in black and white mode in the top row of Figure 6.10. Using the same entropy function described in Section 6.1.2.1, the

[iii] available from http://vismod.media.mit.edu/pub/VisTex/VisTex.tar.gz

calculated values of the matrix-wise entropies for these pure images were 533.27 (Building), 570.50 (Fabric) and 531.59 (Tile).

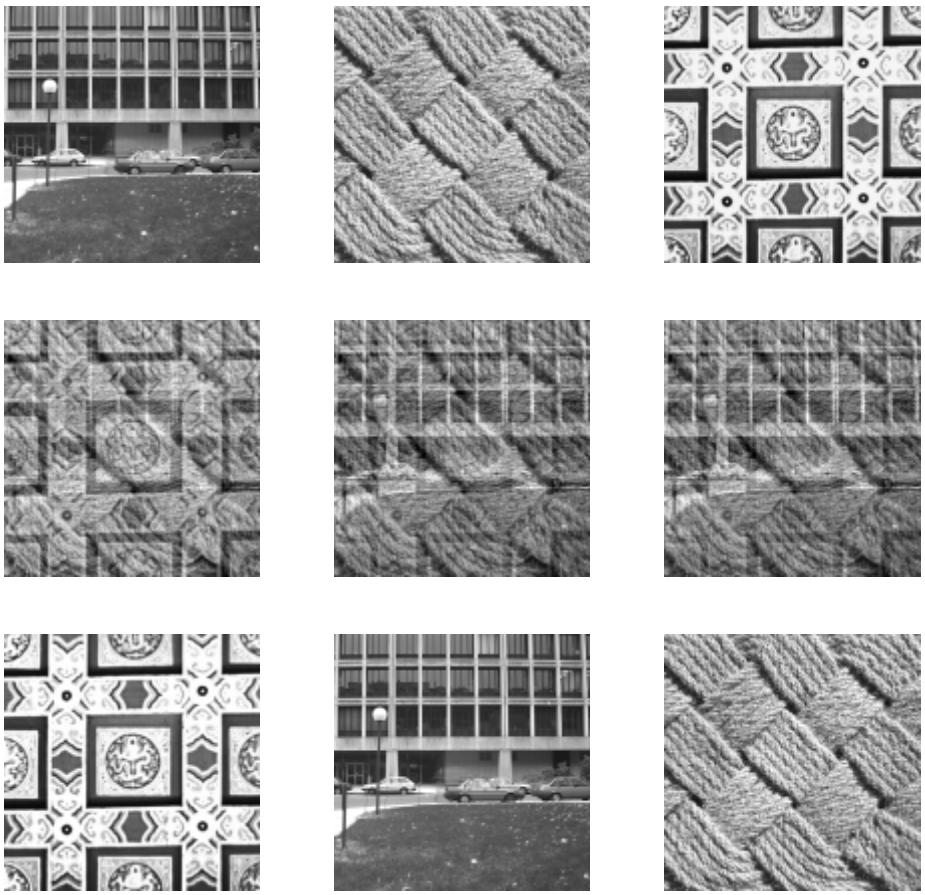


Figure 6.10. Top row: original images from MIT database, the "Red" lay images were used as the pure images and displayed in black and white mode. Middle row: mixture images. Bottom row: recovered images.

A random mixing matrix with non-negative entries was generated and is shown in Eq. 6.7.

$$A = \begin{pmatrix} 0.3202 & 0.7200 & 0.5479 \\ 0.8207 & 0.8876 & 0.3240 \\ 0.6186 & 0.6395 & 0.2308 \end{pmatrix} \tag{6.7}$$

The matrix-wise entropies for the mixture images shown in middle row of Figure 6.10 were 569.3511, 570.3221 and 570.2690, respectively. According to the 2D entropy minimization algorithm, SVD was first applied to the unfolded mixture images (3 × (128×128)) and right singular matrices were obtained. Three physically meaningful right singular matrices were used for matrix-wise 2D- Entropy Minimization. The first recovered result was the building image. Further non-similar images were sought using the search approach outlined in section 5.5.3. The penalty function adopted here was a modified *sigmoid* function (Eq. 5.16) with parameters a=1e5 and b=2e-4.

The resulting set of 3 recovered images is shown in the bottom row of Figure 6.10. The values of the matrix-wise entropies for these images were 531.55 (Tile), 533.27 (Building) and 570.52 (Fabric), respectively.

In Figure 6.10, it is obvious that the recovered images are very similar to the original source images in the top row. A least-square fit was performed to project the recovered images back onto the mixture data and the resulting mixing matrix was obtained. Further, comparison of the quality of the reconstructions is achieved by column permutations to re-order the spectra, and column re-normalization. The original and calculated mixing matrices are given in Eq. 6.8 and 6.9 and they are very similar.

$$A_{original} = \begin{pmatrix} 0.3901 & 0.8111 & 1.0000 \\ 1.0000 & 1.0000 & 0.5912 \\ 0.7537 & 0.7205 & 0.4213 \end{pmatrix} \tag{6.8}$$

$$A_{calculated} = \begin{pmatrix} 0.4017 & 0.8129 & 1.0000 \\ 1.0000 & 1.0000 & 0.5963 \\ 0.7524 & 0.7205 & 0.4250 \end{pmatrix} \tag{6.9}$$

6.3.2.2. Analysis of Geometrically Similar Images

After the success of the separation of the texturally dissimilar images, a more complex data set is needed to test the proposed image recovery approach. Three geometrically similar images, consisting of buildings in Singapore, were used in this study[iv]. A tall skyscraper is centered in these images with both sides of the image bordered by adjacent buildings. Original image files were in truecolor (RGB) with separate red, green and blue overlays, and each one was represented by 256×192×3 array. These color images are shown in Figure 6.11. Their matrix-wise entropies, which were obtained by summing over the 3 RGB layers, were 3411.0, 3364.4 and 3386.9, respectively.

Figure 6.11. Original images in color. PWC Building (left), Republic Building (center), CapitaLand Building (right).

A random mixing matrix with non-negative entries was generated and is shown in Eq. 6.10.

$$A = \begin{pmatrix} 0.6449 & 0.3420 & 0.5341 \\ 0.8180 & 0.2897 & 0.7271 \\ 0.6602 & 0.3412 & 0.3093 \end{pmatrix} \tag{6.10}$$

[iv] These images are archived and down loadable at http://www.chee.nus.edu.sg/research/chbe_freeware.html

The 3 corresponding super-imposed images are shown in Figure 6.12. The values of the matrix-wise entropies for these images are obtained by summing over the 3 RGB layers, and these were 3590.9, 3576.6 and 3583.7, respectively.

Figure 6.12. Mixture image obtained from mixing matrix A defined in Eq. 6.10.

Each image was unfolded three times according to the 3 RGB layers. After unfolding each image, SVD was then applied separately to each of the RGB data sets, and hence 3 times to 3 × (256×192) matrices. It is worth noting that there are two different approaches to the solution of whole image separation in three lays. First, we can start with mixtures in any color layer - red, green or blue. With the recovered images via entropy minimization, the mixing matrix is sequentially obtained by mapping back the reconstruction image to the mixture. The unmixing/separation matrix, equivalently, the inverse of the mixing matrix, can be utilized to separate the other two mixture layers. Alternately, the entropy minimization is executed over all the three layers. The final reconstruction result is composed of each recovered pure layer in red, green, blue order. The latter strategy was adopted here in order to test the performance of the algorithm.

To start, the red set of right singular matrices was taken, and a similar procedure was used as described in Section 6.3.2.1. This resulted in a set of 3 red images and the same general approach was then used for the green and blue data sets independently. The red, green and blue data sets for each image were then consolidated to generate the three reconstructed truecolor images each with 256×192×3 pixels shown in Figure 6.13. The values of the matrix-wise entropies for the images in Figure 6.13 were 3386.9, 3364.4 and 3411.1, respectively.

Figure 6.13. Reconstructed images in color.

Table 6.2 depicts the entropies of the different layers for the different images. For completeness, details about the sensitivity to the global search during the reconstruction procedure are included. It shows that the Republic building possesses the lowest entropies of each layer as well as in the sum among the three images. Although the CapitaLand Building image has the 2nd lowest entropy in all three RGB layers, its image was always recovered first, even though a number of genetic algorithm or simulated annealing searches were conducted with random initial values. This suggests the search space was very complex in this problem.

Table 6.2. The entropies of different layers for different building photos

Image Name	Matrix-wise entropies value for three images			
	Red layer	Green layer	Blue layer	Sum of 3 layers
PWC Building	1134.8	1131.9	1144.3	3411.0
Republic Building	1129.3	1114.5	1120.5	3364.4
CapitaLand Building	1133.7	1121.5	1131.7	3386.9

The reconstructed images in Figure 6.13 are obviously very similar to the original images in Figure 6.11. The red images were used to reconstruct the mixing matrix A_{red} for this set of data. The same general approach was then used for the green and blue data sets independently, resulting in mixing matrices A_{green} and A_{blue}. The resulting mixing matrices are also very similar to the original mixing matrix.

6.3.2.3. The Underdetermined Problem and 2D-BTEM Method

It is known that if the number of experimentally measured images happens to be less than the number of observable components, then the mathematical problem can be considered irrevocably ill-posed and subsequently, it is impossible for a unique solution to problem. In this section, band-targeting entropy minimization method is employed in the simulated undetermined problem to help extract useful information.

A new image, consisting of the capital letters NUS, was created as a watermark with matrix-wise entropy of 174.4975 (Figure 6.14a). This watermark was imbedded at a 10% level into each of the three mixture images used in Section 6.3.2.2. An example of a mixture image with a 10% watermark is shown in Figure 6.14b. And it is noticed that the

watermark is not really discernable at this level of imprinting. The 3 mixture images with imbedded watermarks were separated into red, green, blue and SVD performed.

In order to enhance the watermark intensity with 2D-BTEM, *a priori* information about that the lower left region near pixels x=45-46 and y=40-43 (located in the upper part of the letter U) is needed. This coordinate region with an interest feature needs to be retained and enhanced after image recovery. With the constraint of retaining the selected spectral feature, 2D-BTEM was implemented for the pursuit of enhancing the image pattern. The resulting recovered image is shown in Figure 6.14c. The watermark is now more prominent, but it cannot be fully recovered due to the underdetermined nature of the problem.

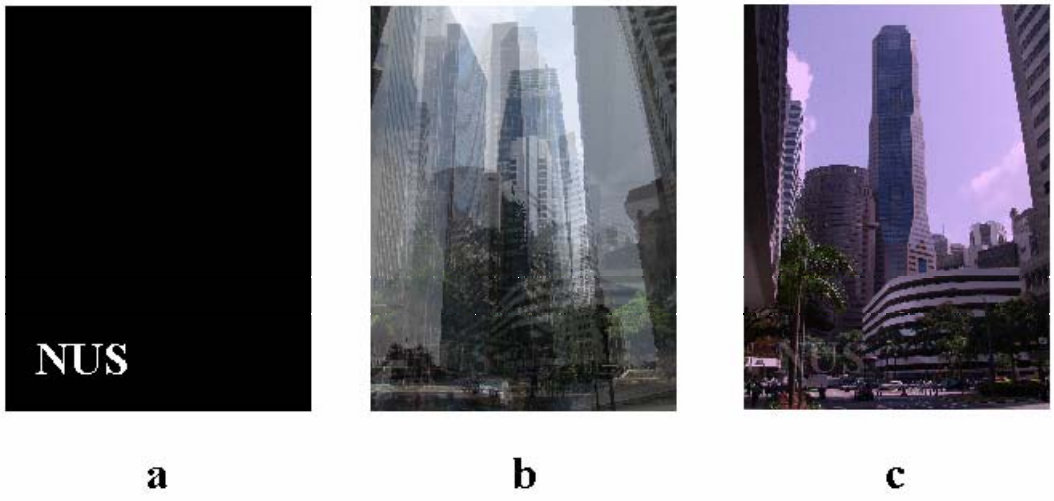


Figure 6.14. A simulated watermark (a), an example of a mixture image with a 10% watermark (b) and the resultant recovered image (c).

6.3.3. Discussion

The results of the determined problem show that with the direct use of the entropy function, blind source separation of mixtures is possible. The utility of this approach was

further shown using the associated underdetermined problem, where image enhancement was achieved.

Many forms of entropy are used in the image processing science. The entropy type function and objective function chosen in the present contribution are not, by any means, the only choices available. Indeed, there are a multitude of entropy (Frieden, 1975; Kapur, 1993), and entropy type functions that could be used or at least tried on various classes of images. Several different definitions of entropy appear in the literature of image processing. Starck and Murtagh (1999) had a discussion about entropy functions used in image processing, including Burg entropy (Burg, 1978), Freiden entropy (Frieden, 1975), Gull and Skilling entropy (Gull and Skilling, 1984). Even though entropy is extensively used in image processing, the approach and application here appears to be quite original and new.

In this analysis, the two-direction derivative function used in the present entropy type function would make use of nearest neighbor pixel information. The minimization of the randomness between pixel values would be obtained by minimizing the pixel-to-pixel variations (the smoothness of local curvature). Also the least randomness would result in the retention of a structured pattern here.

6.4. Summary

In this chapter, 2D entropy minimization, which involves a one-spectrum-at-a-time global search approach, was successfully applied to a set of matrix mixture data set generated from three different texture matrices. Further simulated mixture spectra were also briefly explored.

In the image reconstruction analysis, the texturally different and the geometrically similar images were tested sequentially. The mixture spectra were decomposed using SVD and then global stochastic optimization was used to find the first irreducible image pattern. Further images were then subsequently reconstructed, by imposing a 2D correlation coefficient for dissimilarity to prevent repeated images, until all images were exhaustively enumerated. In another test, the watermark was enhanced after targeting the region for 2D-BTEM. All these results support that 2D entropy minimization algorithm would have the potential for a wide variety of applications, including image and spectroscopic analysis.

Chapter 7

2D BTEM: Application to Real Experimental Systems

This chapter emphasizes the development and the application of numerical algorithms for pure component reconstruction of 2D NMR spectroscopy and 2D fluorescence data. This chapter is divided into two major sections. The first section describes the application of the proposed methodology to the 2D NMR systems. In this mixture system, we would like to test the 2D-BTEM algorithm on a real multi-component mixture spectroscopic data. After that we present the application of 2D-BTEM to 2D NMR reaction data. In the second part, 2D fluorescence data is analyzed. A simulation data set and an experimental data set are treated in sequence. At the end of this chapter, other types of 2D spectroscopic data are discussed.

7.1. Application of 2D Band-Target Entropy Minimization Method (2D-BTEM) to 2D NMR Data

The NMR experimental results used in this section were obtained in collaboration with Dr Anette Wiesmat (ICES), Peter Sprenger (Bruker Biospin) at ICES in Singapore and Peter Sprenger at Bruker Biospin AG in Zurich, Switzerland.

7.1.1. Introduction

NMR spectroscopy is an absorption spectroscopy involving the absorption of radio frequency electromagnetic waves. And it is one of the most important and most

informative analytical techniques available for characterizing the structures of compounds. It is now widely used in various studies such as physics, chemistry, biology, material science, etc.

In catalysis study, either heterogeneous catalysis or homogeneous catalysis, NMR spectroscopy always is regarded as a very important and fascinating tool for understanding the catalysis. Most importantly, NMR spectroscopy can also be used to obtain both qualitative and quantitative information of the chemical species in the complex reaction system. NMR spectroscopic data will be used to reveal the concentrations of catalytic intermediates as well as the reagents and products. It will facilitate the acquisition of kinetic data and establishment of the mechanism.

7.1.2. *In situ* NMR Spectroscopy used in Catalysis

In many technical processes, there is a need to study complex multi-component mixtures and gain insights into their behaviour during the physical and chemical processes. Once the physicochemical behaviour of such mixtures is understood, predictive models for their properties can be developed.

In situ spectroscopy is a general methodology; it allows on-line investigation of the reaction in real time and under real operation conditions without interruption to the reaction. With this merit, monitoring chemical reactions (Sarazin *et al.*, 1996) especially, complex catalytic reactions (also including the biocatalysts, biotransformation) (Weber and Brecker, 2000) by using non-invasive NMR spectroscopy is of growing interest (Maiwald *et al.*, 2003; Keifer, 1999; Iggo *et al.*, 1998; Cobb *et al.*, 1996; Kim *et al.*, 1999; Maiwald *et al.*, 2004). *In situ* NMR methodology plays an important role in the characterization of organometallic complexes in catalytic reactions. Firstly, observing

intermediates and reactive transients during chemical reaction is essential, as these may not be observable once the reaction conditions are changed. Further elucidation of reaction mechanisms and catalytic cycles provides rational development and improvement of the reaction system.

7.1.3. Two-Dimensional NMR Spectroscopy

In the last three decades, NMR spectroscopy has experienced a dramatic development in sophistication of instrumentation as well as great progress in signal processing. The notion of this new revolutionary technique of 2D NMR was proposed by Jeener (1971) in 1971 and later demonstrated by Ernst *et al.* leading to a tremendous increase in the capability of NMR and the subsequent explosion in experimental techniques for higher dimensions (Gunther, 1983; Duddeck and Dietrich, 1992; Farrar, 1987). There are now a large number of experimental 2D techniques including COSY (Aue *et al.*, 1976; Piantini *et al.*, 1982), HSQC(Bodenhausen and Ruben, 1980), HMQC (Bax *et al.*, 1983), HMBC (Bax and Summers, 1986), TOCSY (Braunschweiler and Ernst, 1983), NOSEY (Jeener *et al.*, 1979; Kumar *et al.*, 1980), ROESY (Bothner-By *et al.*, 1984), 2D-INADEQUATE (Bax *et al.*, 1980; 1981) etc. have been widely used for the analysis of structurally complex molecules, including the structural determination of bio-molecules such as proteins, peptides and nucleic acids.

The use of 2D NMR techniques potentially enables us to obtain, in a single experiment, hundreds or thousands of structural constraints, which can ultimately lead to a high-resolution structure of the molecule. 2D NMR is routinely used in an ever increasing array of specialized experiments designed to aid in spectral assignment and structural characterization of macromolecules.

Two dimensional NMR spectra show signal intensities as the function of two frequencies (f_2 and f_1). The f_2 axis typically depicts ^{1}H frequencies, while the f_1 axis depicts either ^{1}H in homonuclear spectra, or another nucleus (e.g. ^{13}C or ^{15}N) in heteronuclear spectra. The primary advantage of 2D NMR over 1D NMR spectroscopy is the resolution enhancement and simultaneous acquisition of certain correlation information between the spins which are usually inaccessible from the corresponding conventional 1D NMR experiment. Also 2D NMR methods effectively resolve crowded regions in the spectrum by mapping the spectral information onto two frequency axes rather than the conventional 1D (chemical shift) plot of the spectrum.

7.1.3.1. Homonuclear Correlation Spectroscopy

Homonuclear correlation experiment is designed to record the correlation of protons and homonuclei. In the typical H-H COSY experiment, both of the coordinates are the chemical shifts of proton nucleus. The COSY spectrum shows how the protons are coupled with other protons as indicated by the coordinates of peak's position. Other homonuclear correlation techniques include, 2D TOCSY(Total correlation Spectroscopy), 2D INADEQUATE (Incredible Natural Abundance DoublE QUAntum Transfer Experiment, which is useful for determining the signals arise from neighbouring nuclei), 2D NOESY(Nuclear Overhauser Effect spectroscopy, which is useful for giving information about interactions between protons that are close in space rather than those are connected by a short through-bond) and 2D ROESY (Rotational Nuclear Overhauser Effect spectroscopy) .

7.1.3.2. Heteronuclear Correlation Spectroscopy

Apart from protons, chemical compounds normally contain other magnetically active nuclei, such as ^{15}N, ^{13}C and other important elements. The heteronuclear correlation experiment is similar to the homonuclear experiment with the exception that it concerns two different nuclei. Also the second indirectly detected dimension contains chemical shift information about the heteronucleus. The use of these hetero nuclei allows some new features in NMR which facilitates structure determination especially of complex molecules, for example, bio-molecule or organometallic molecule. For the structure elucidation, if the proton nuclei have already been assigned, the additional carbon-13 proton correlation spectrum will facilitate the assignment of all the protons related to the carbons. In the techniques of Heteronuclear Multiple Quantum Correlation (HMQC) and the Heteronuclear Single Quantum Correlation (HSQC), the resultant spectrum is shown with the ^{1}H and ^{13}C axes being plotted against each other. Heteronuclear Multiple Quantum Correlation (HMBC) enables assignments of signals in cases where ^{13}C and ^{1}H nuclei are coupled through two or more bonds.

7.1.4. Application of SMCR in NMR

Most effort in NMR data analysis focus on the FID processing and NMR spectra enhancement, for example, linear prediction, DFT (discrete Fourier transform), MEM (maximum entropy method) and pattern recognition. As mentioned in the literature review in section 2.2.3, few SMCR chemometric techniques have been transferred to applications in NMR. The unique characteristics of NMR signals prevent these chemometric techniques from being used in most applications.

7.1.5. 2D BTEM: Application to Mixture System

In this section, we will present the application of the algorithm to 2D NMR measurements of solutions containing three solutes. Two types of 2D NMR experiments were implemented in this mixture system. They are (1) COSY (H-H Correlation Spectroscopy): two dimensional NMR experiment that reveals both direct coupling (but not indirect coupling) between protons within a spin system and (2) HSQC (^{1}H, ^{13}C Correlation Spectroscopy): two dimensional NMR experiment where the f_2 and f_1 axes depict the frequencies of directly bonded ^{1}H and ^{13}C nuclei, respectively and each peak represents an individual, coupled ^{13}C-H group.

7.1.5.1. Experimental Section

Sample Preparation and Measurements

The samples for NMR were prepared by dissolving varying amounts of 1,5 chloro-1-pentyne (Aldrich), 4-nitrobenzaldehyde (Aldrich) and 3-methyl-2-butenal (Aldrich) and topping with $CDCl_3$ to achieve a total volume of 500ul.

Seven solutions for NMR measurements were prepared. The reference samples were prepared by dissolving 20 µl 1,5 chloro-1-pentyne, 20 mg 4-nitrobeanzaldehyde, 20 µl 3-methyl-2-butenal in $CDCl_3$ and topping up to achieve a total volume of 500ul. Therefore three reference solutions were obtained separately. It is worth noting that the use of a constant liquid phase volume in all sample preparations was crucial in the quantitative aspects of this study.

Instrumental Aspects

All the spectroscopy data were acquired at 298 K on a Bruker Avance 400 WB NMR spectrometer equipped with a 5mm $^{1}H/^{31}P/^{13}C/^{15}N$ QNP probe with *z* gradient. All the 2D NMR spectra were acquired at 400.13 MHz (^{1}H) and 100.62 MHz (^{13}C) with standard Bruker-supplied pulse sequences. Two spectra of each solution were measured. The spectral parameters were as follows: the ^{1}H spectral width was 5208 (COSY) and 4807 Hz (HSQC) and for the ^{13}C dimension 20120 Hz, number of scans per increment 2, number of t_1 increments 128, each with 1K acquisition points, and repetition time 1.5s. The 2D spectra were processed as 1K-1K complex matrices with unshifted sine weighting functions in both two dimensions. The final data set for 2D-BTEM was a 3-array of dimension $\underline{A}_{14\times1024\times1024}$.

All these experiments were carried out by Dr. Anette Wiesmat and Peter Sprenger at Institute of Chemical and Engineering Sciences (ICES), Singapore.

7.1.5.2. Computation Section

Data Pre-treatment

The contour plot of the 2D HSQC NMR spectrum of one mixture solution is presented in Figure 7.1 and the complex spectral features of this mixture system are clearly presented. It is clear that for computations involving very large arrays, i.e. matrix multiplications, and the computation of the SVD, difficulties can be encountered even on a high end workstations with considerable RAM. Moreover, during the implementation of

BTEM algorithm, there would be thousands of iterations needed during the stochastic optimization/searching technique.

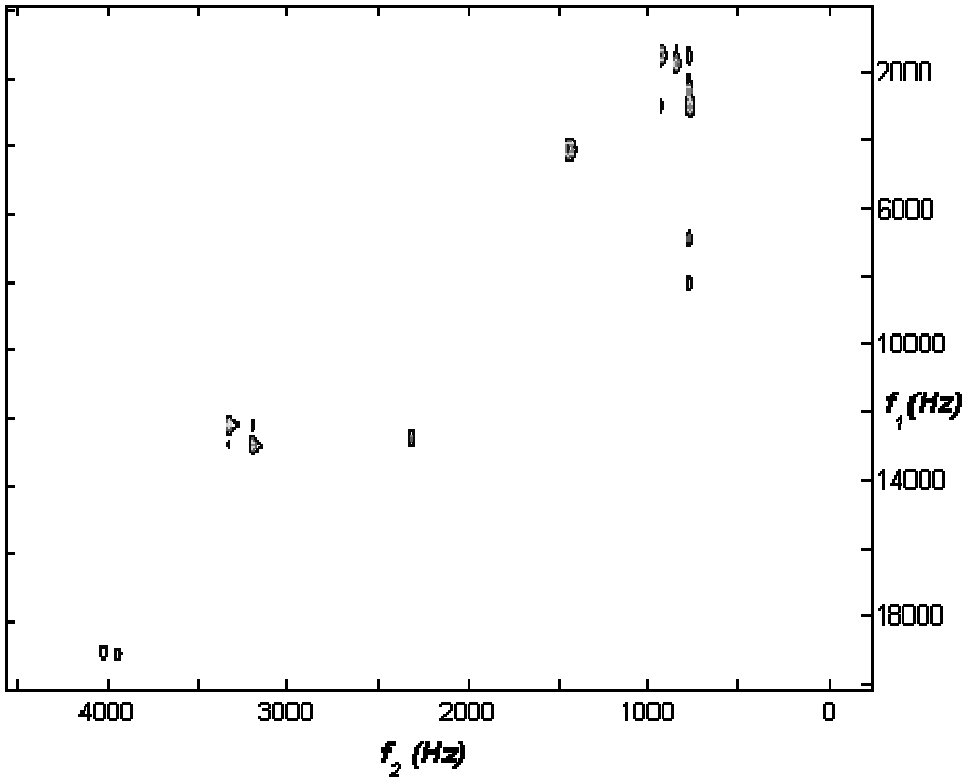


Figure 7.1. The contour plot of the 2D HSQC NMR spectrum of one mixture solution.

As we see in Figure 7.1, there are only several peaks and most of the regions are blank where no significant signal exists. In other words, about 90 percent of the original 2D spectral data does not contain useful physical information. Therefore, in order to decrease the computational burden, only 4 rectangular regions (6 small pieces) containing the real physical spectral features (peaks) were taken. The small rectangular regions were assembled into a new consolidated data array $A_{14\times(539\times107)}$.

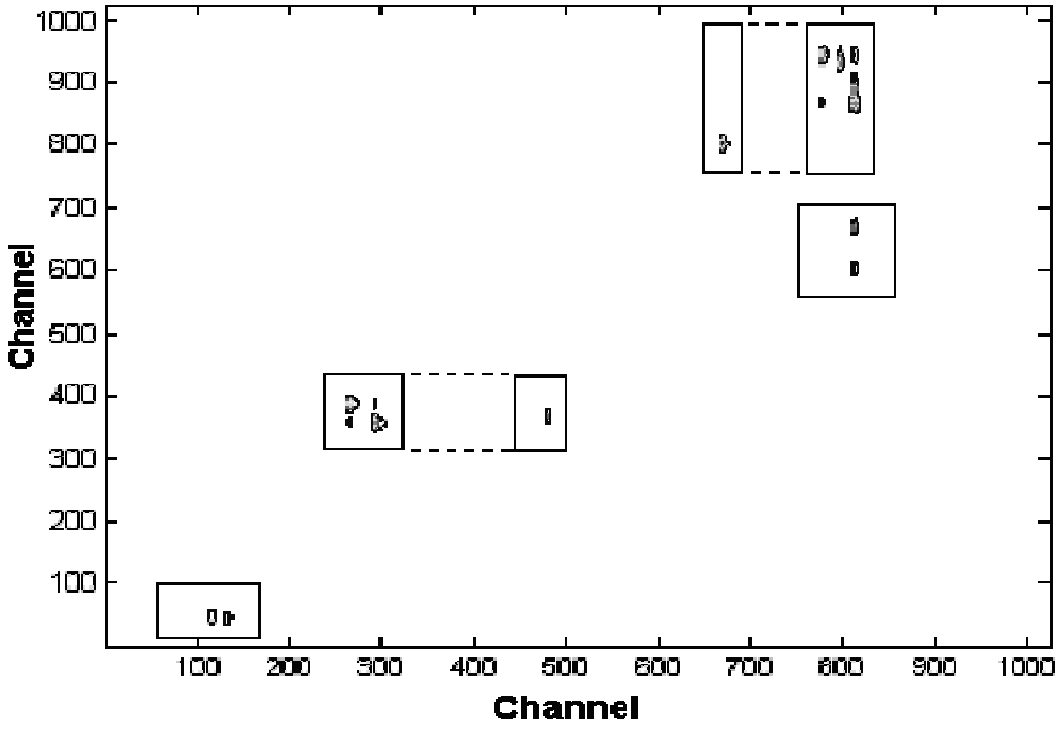


Figure 7.2. Only 4 rectangular regions (6 small pieces) containing the real physical spectral features (peaks) were used in subsequent analysis. (x and y coordinates are shown in channels).

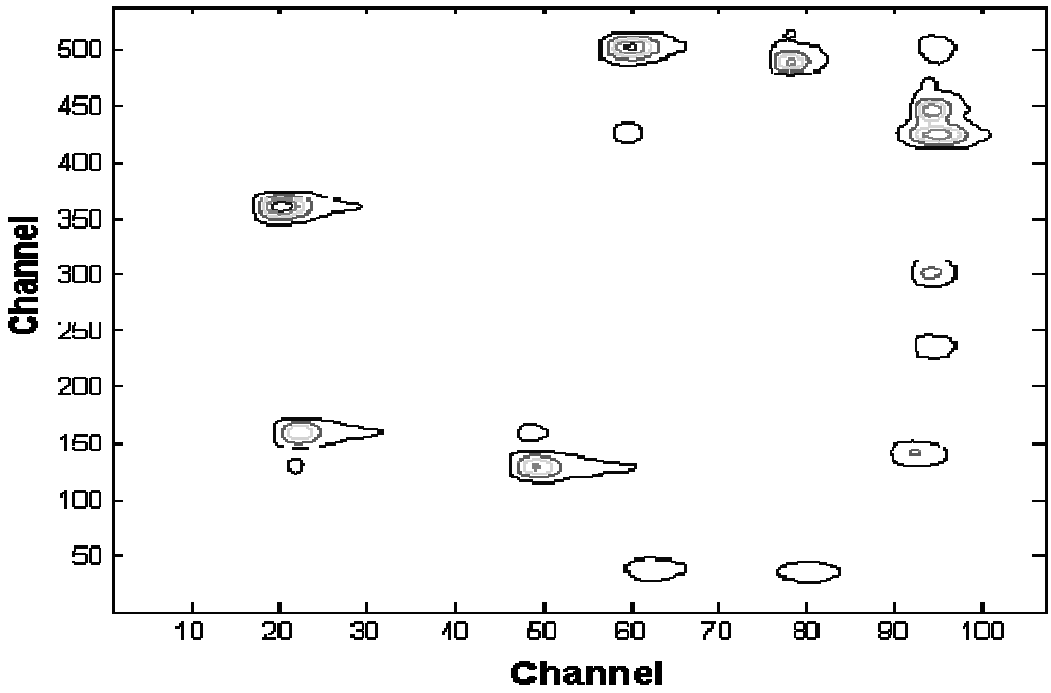


Figure 7.3. The contour plot of one consolidated data set resulting from the small rectangular regions (shown in channels).

As shown in Figure 7.3, all the significant signals are now concentrated together in the small matrix (539×107). Compared to the original 1024 by 1024 data, only 5.5% elements in the original data will be used during the following computational procedure. It is clear that using the pre-processed small patches instead of the raw data would facilitate the extensive mathematical computation.

2D Wiener Filtering

An adaptive 2D Wiener filter is commonly used to filter degraded images in image processing and was used in the present contribution to filter the experimental 2D NMR data. The function WIENER2 is available in MATLAB(1995) which performs 2D adaptive noise-removal filtering. Y = WIENER2 (X, [a b]) filters the matrix X using pixel-wise adaptive Wiener filtering, using neighbourhoods of size a-by-b to estimate the local image mean and standard deviation. In the present study, the parameters *a* and *b* were set to 10 and 10.

7.1.5.3. Result

Singular Value Decomposition

SVD was performed on the matrix $A_{14\times(539\times107)}$ to obtain the 14 right singular vectors in $V^T_{14\times(539\times107)}$. The vector-format right singular vectors are shown in Figure 7.4. Different from the normal 1D spectrum, the signals in the 2D plane are sliced and connected, producing lots of peaks extending in one line. After that, concatenation was undone. The resulting right singular matrices are shown in Figure 7.6. Physically meaningful spectral features were observed in only the first seven matrices. The 8th matrix

is essentially featureless. This holds true for the 9th-14th matrices as well. Consequently, the 3-way array $V^T_{14\times(539\times107)}$ was reduced to $V^T_{7\times539\times107}$ where *j* is set to seven.

2D-BTEM

Both vector-wise and matrix-wise 2D-BTEM were applied in HSQC data. First, vector-wise 2D-BTEM was performed by targeting interesting features range from channel 25900 to 26100(*a*), 41980 to 42020(*b*) and 32293 to 32313(*c*) individually since these features/signals repeat in the series of V^T. These ranges are labelled as *a, b* and *c* in Figure 7.4 respectively. Three estimated spectra resulted from the vector-wise algorithm are shown in Figure 7.5.

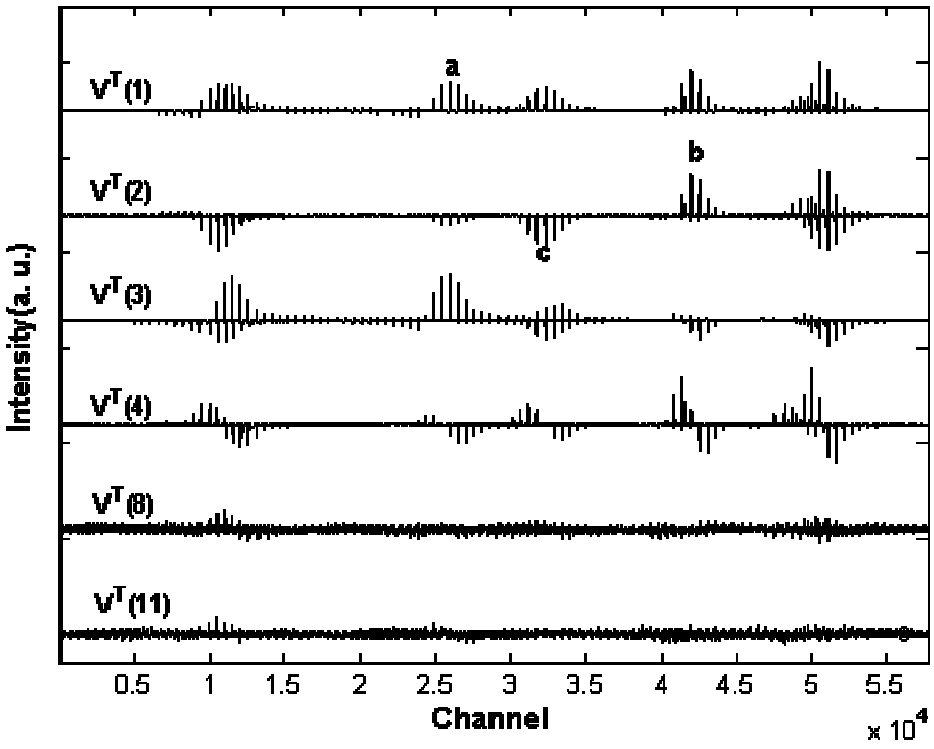


Figure 7.4. The vector-formatted right singular vectors resulted from HSQC data $A_{14\times(539\times107)}$, Only 1st-4th, 8th and 11th V^T are shown here. Label *a b* and *c* indicate the interesting features.

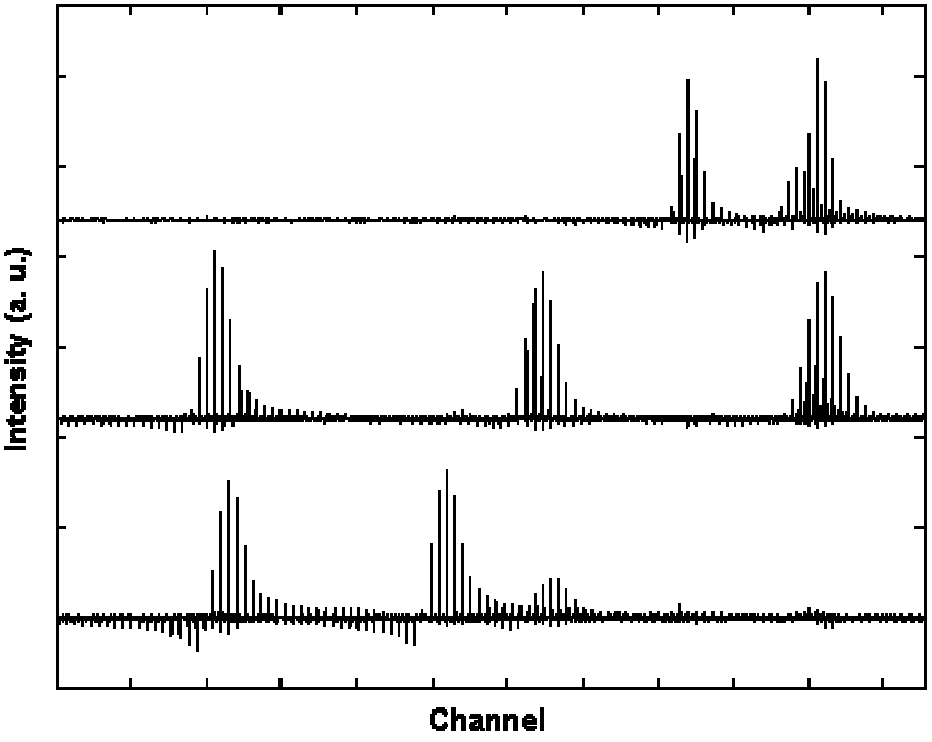


Figure 7.5. The recovered 1D patterns resulting from the vector-wise algorithm.

Matrix-wise 2D-BTEM was implemented with $V^{T}_{7\times539\times107}$ by targeting observable features in the seven right singular matrices. They are region 1([355 to 365; 18 to 20]), region 2([442 to 448; 94 to 96]) and region 3([125 to 135; 47 to 50]). Exhaustive searches produced only three 2D spectral patterns. These estimates are shown in Figure 7.6(*a*, *b* and *c*).

After all these three patterns were extracted from the mixture source, the next step was to put these patches back to their original position. In other words, these patterns were then imbedded into matrices with 1024×1024 channels. After imbedding, the resultant 2D HSQC estimated pure component spectra and the reference spectra are shown in Figure 7.7. The spectral estimates appear quite good when compared to authentic experimental references.

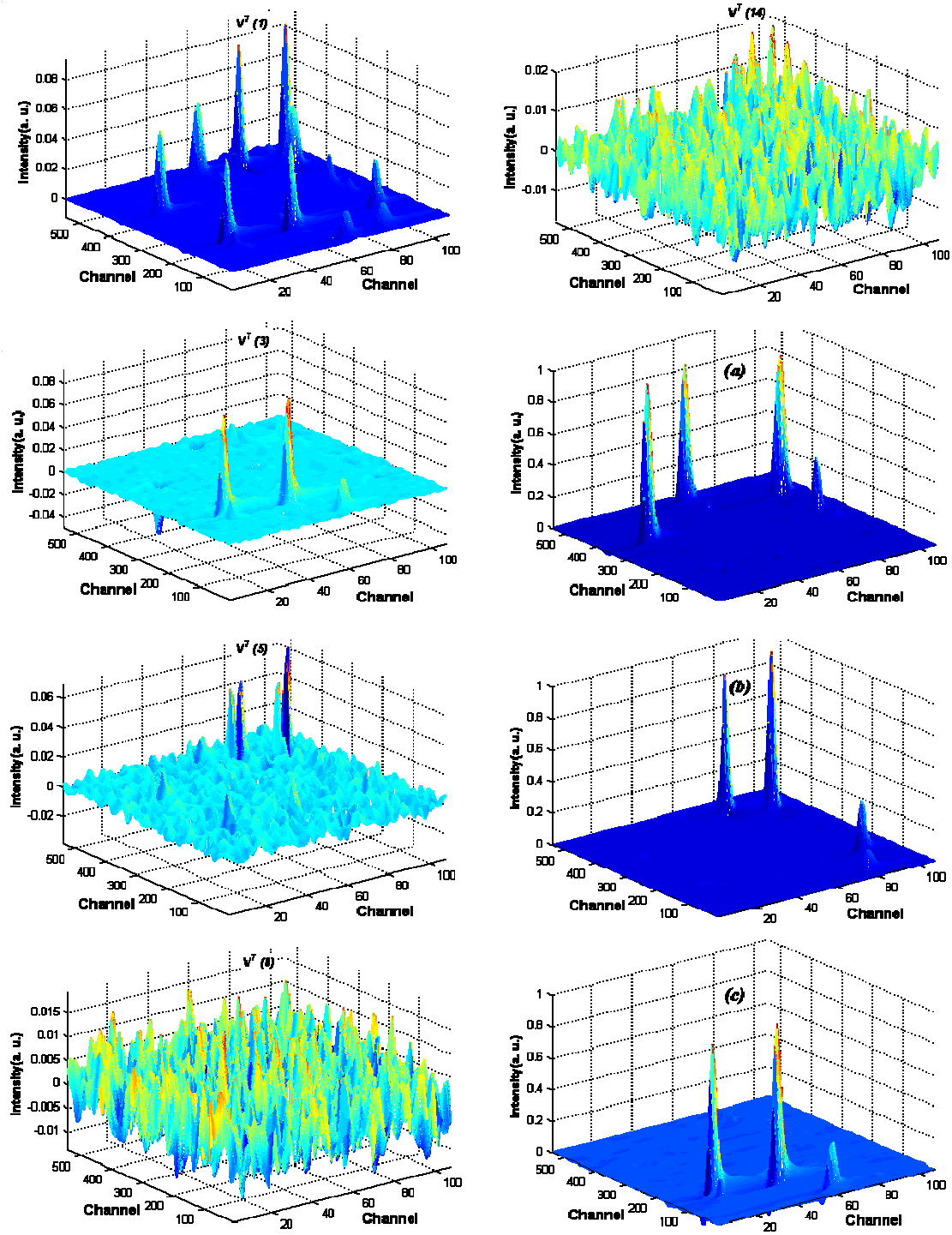


Figure 7.6. The resulting right singular matrices (1st , 3rd, 5th, 8th and 14th are shown only) and the exhaustive search results with three patterns(a, b and c). A negative part in the signal is observable in c which is related to the phase problem[i].

[i] The phase problem is a notorious problem in NMR studies. Discussions are continuing with Peter Sprenger at Bruker Biospin and with researchers at the ICES on ways to nullify this problem, possibly by a combination of hardware / pulse considerations and post processing of data.

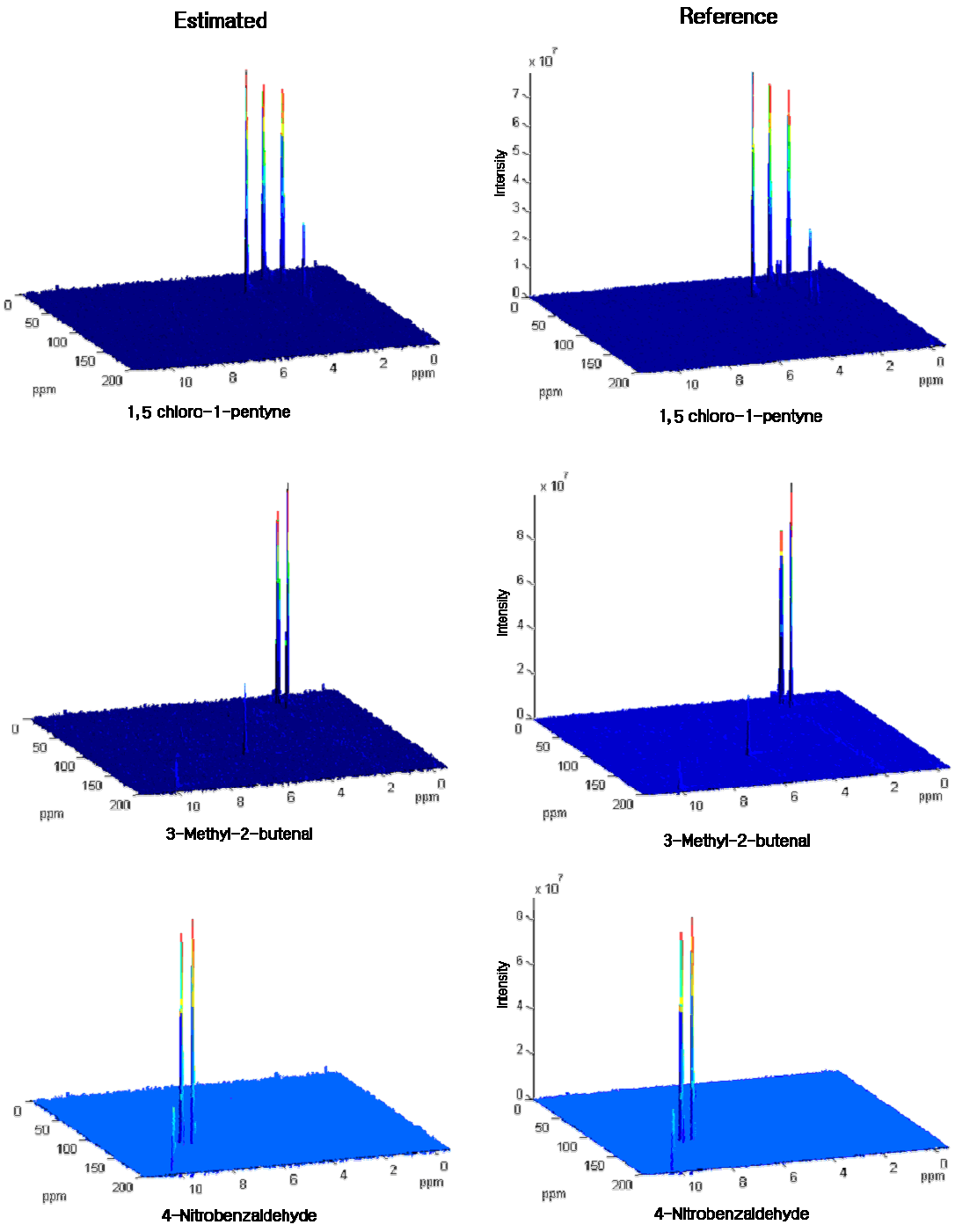


Figure 7.7. The estimated HSQC spectra and reference spectra.

Moreover, the dual problem for relative concentrations can be solved. Figure 7.8 shows the relative concentrations as determined by a least-square fit with the reference spectra versus the relative concentrations as determined by a least-square fit with the estimated pure component spectra. It is noticed that the relative concentrations calculated for samples 1-7 are almost the same as 8-14 since samples 8-14 represent the replicate measurements (there were seven physical samples, each sample measured twice). The calculated concentration profiles are very similar.

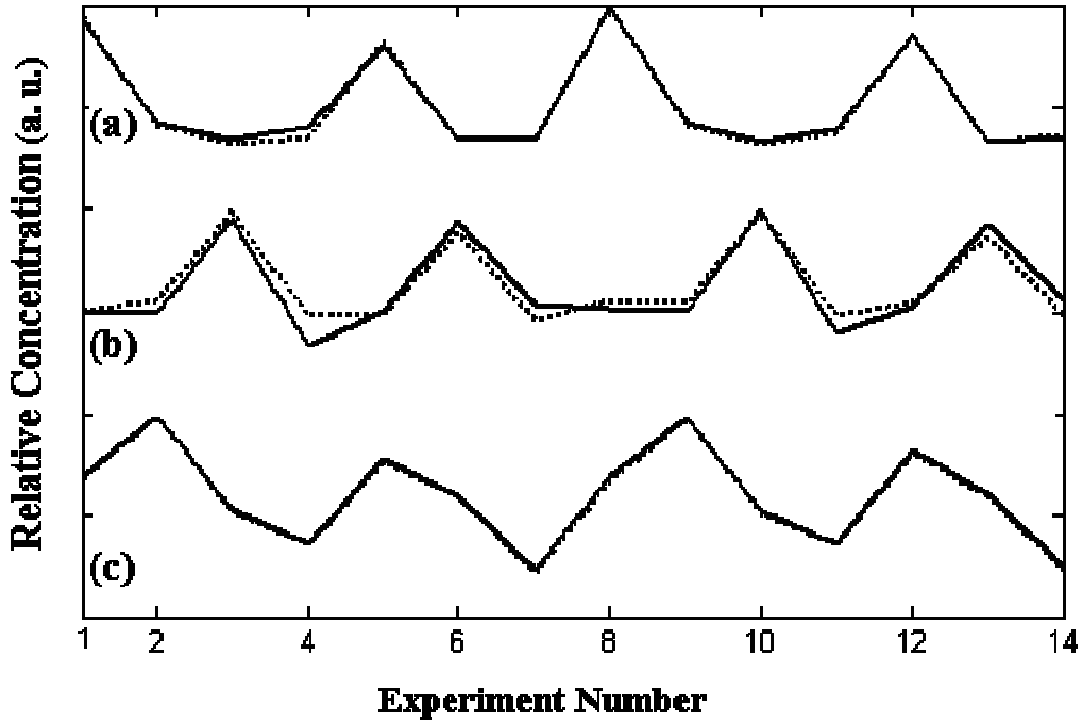


Figure 7.8. The relative concentrations for HSQC experiments as determined by a least squares fit with the reference spectra (solid line) versus estimated pure spectra (dotted line). Top row for 1,5 chloro-1-pentyne(a), middle row for 3-methyl-2-butenal (b) and bottom for 4-nitrobenzaldehyde(c).

The 2D COSY measurements were analyzed in a similar manner. Only twelve measurements were available (two measurements were outliers). The contour plot of the 2D COSY NMR spectrum of one mixture solution is shown in Figure 7.9. Only 8

rectangular regions containing the real physical spectral features (peaks) were taken. The small rectangular regions were assembled into a new consolidated data array $A_{12\times(468\times150)}$.

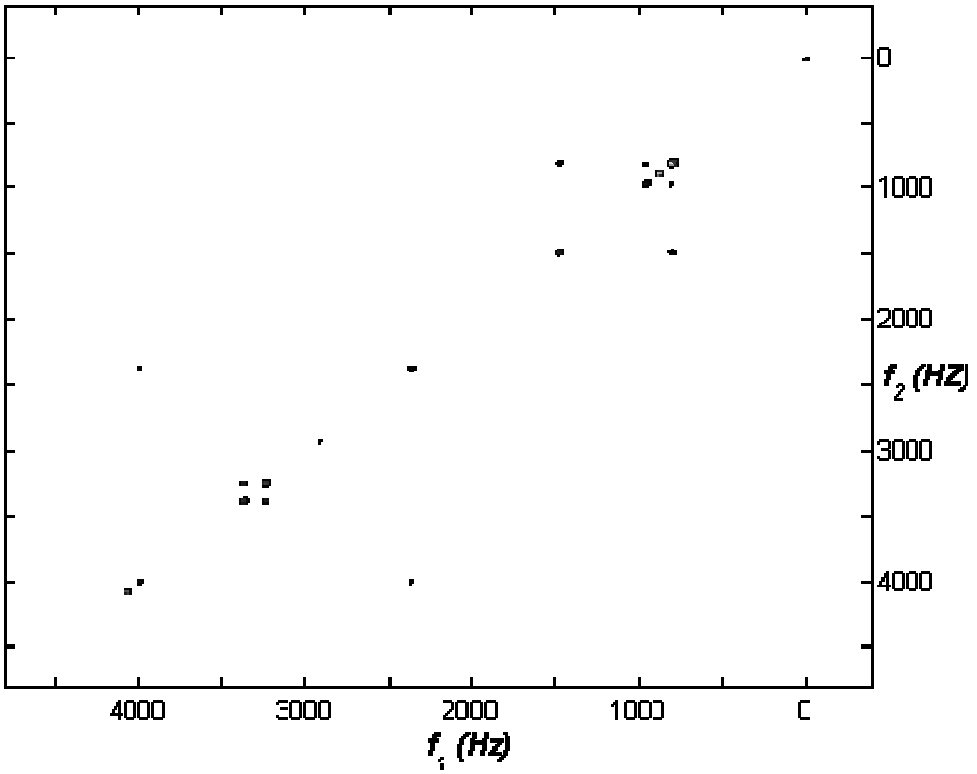


Figure 7.9. The contour plot of the 2D COSY NMR spectrum of one mixture solution (shown in Hz).

SVD was performed on the matrix $A_{12\times(468\times150)}$ to obtain the twelve right singular vectors in $V^T_{12\times(468\times150)}$. Concatenation was undone. Physically meaningful spectral features were observed in only the first seven matrices. Consequently, the 3-array $V^T_{12\times468\times150}$ was reduced to $V^T_{7\times468\times150}$ where j is set to seven. Both vector-wise and matrix-wise methods were applied to this set of data. Exhaustive searches produced only three 2D spectral patterns. These patterns were then imbedded into matrices with 1024×1024 channels. The resultant 2D COSY estimated pure component spectra and the reference spectra are

shown in Figure 7.10. The spectral estimates appear quite good when compared to authentic experimental references.

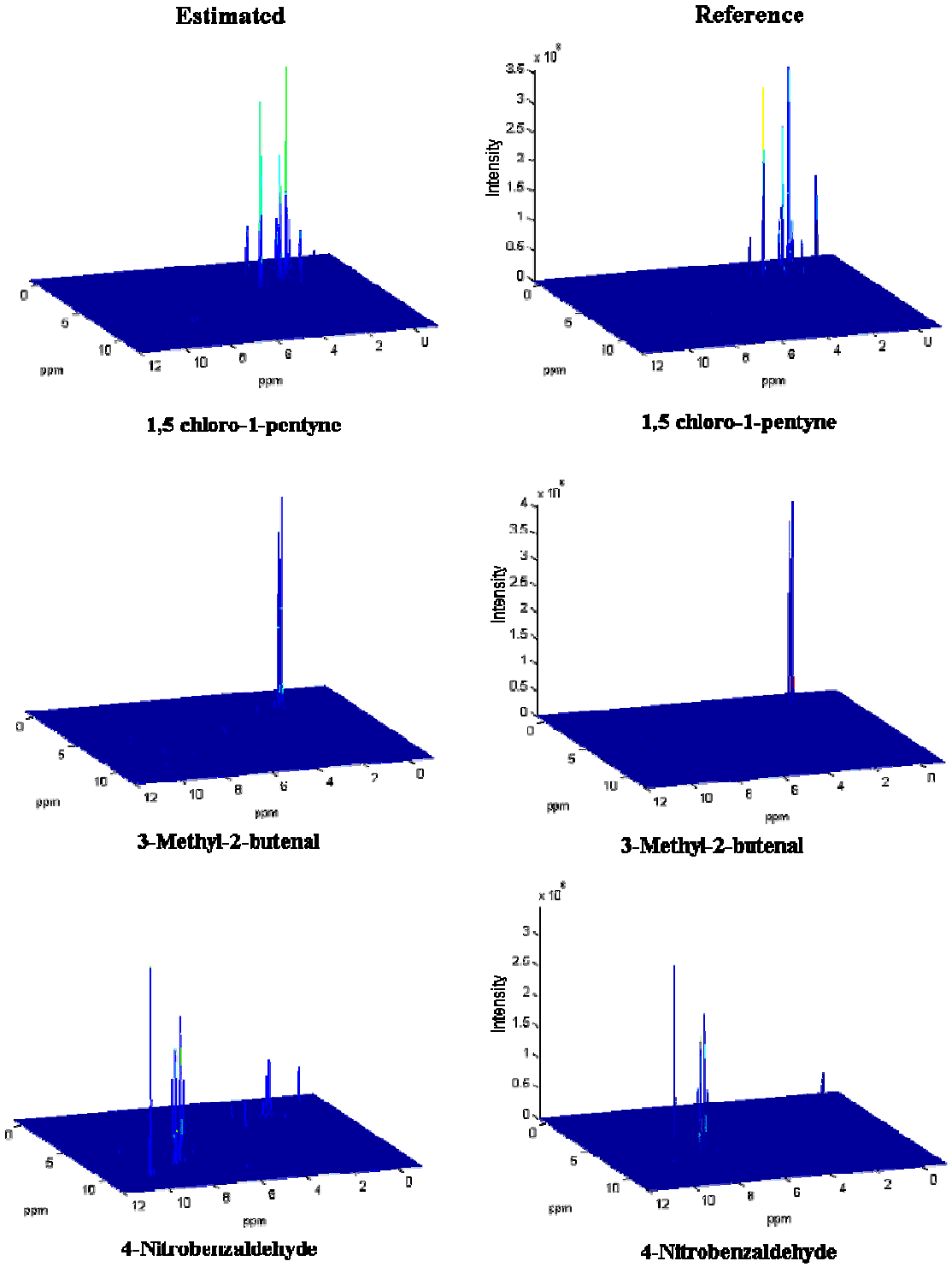


Figure 7.10. The estimated 2D COSY spectra and reference spectra.

The dual problem for relative concentrations was also solved. Figure 7.11 shows the relative concentrations as determined by a least-square fit with the reference spectra versus the relative concentrations as determined by a least-square fit with the estimated pure component absorptivities. The calculated concentration profiles are good for the last two components but only fair for the 1st component. This may be due to the somewhat higher non-stationary characteristics of COSY versus HSQC. Again, the relative concentrations calculated for samples 1-6 are almost the same as the relative concentrations of samples 7-12 since samples 7-12 represent the replicate measurements (there were 6 physical samples, each sample measured twice).

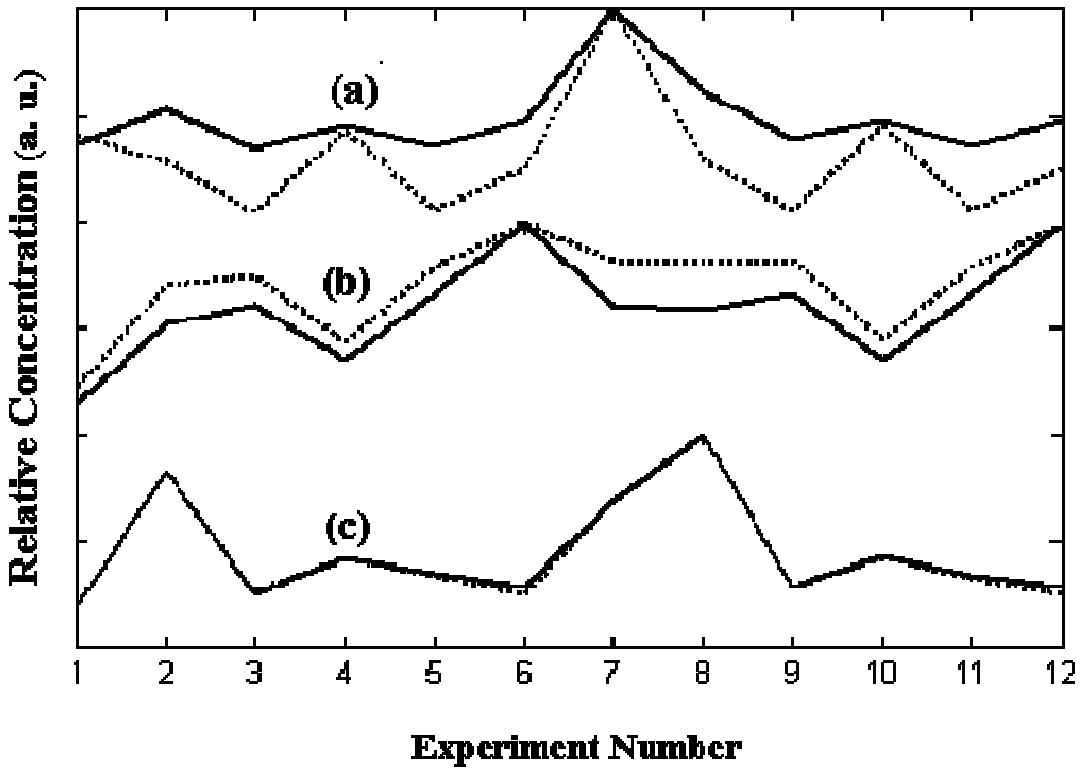


Figure 7.11. The relative concentrations for COSY experiments as determined by a least squares fit with the reference spectra (solid line) versus estimated pure spectra (dotted line). Top row for 1,5 chloro-1-pentyne(a), middle row for 3-methyl-2-butenal (b) and bottom for 4-nitrobenzaldehyde(c).

7.1.5.4. Discussion

1. Both vector-wise and matrix-wise 2D-BTEM methods can be used in this data set. But matrix-wise 2D-BTEM proved to be the preferred computational route, since some information is in a sense "lost" during concatenation.

Another practical reason is the difficulty to find out the features in the V^T series. The resulted 1D spectrum from concatenating will contain huge amount of peaks/feature. And it will not be realistic to 1) identify the feature appearing repeatedly in V^T and 2) to exhaustive search all the features in the V^T domain. However, it is natural for us to study the V^T series in matrix format and select the specific regions. In matrix format, the number of features/peaks (in 2D shape) is numerable in contrast with the number in the 1D form. Matrix-wise 2D-BTEM will produce higher quality spectral estimates generally.

2. Through the comparison of the result of COSY and HSQC, we found that the result of HSQC is much better than COSY. This may be due to the somewhat higher non-stationary characteristics of COSY versus HSQC. The non-stationary characteristics may due to the change of peak position, the change of peak shape, noise and nonlinearity introduced by some of the NMR data processing procedures.

In the following section, we will discuss how to deal with 2D NMR spectroscopy corrupted by the nonlinearity which is induced by the non-stability of peak positions during measurements.

3. Since the stacks of the 2D NMR data do not belong to the category of tri-linear data, the direction application of three-way decomposition, such as PARAFAC, would not be suitable for this data set.

7.1.5.5. Conclusion

An advanced entropy minimization based algorithm for 2D NMR spectroscopic data reconstruction has been proposed and verified on real experimental spectroscopic data. The quality of the recovered spectra is found to be quite satisfactory when compared to references obtained from pure component measurements. Also the calculated relative loadings of each component based on the recovered spectra are consistent with the loadings calculated from reference spectra. These results have implications for general chemical identification of inseparable multi-component mixtures system. Full details of the study of 2D NMR data can be found in Journal: *Analytical Chemistry* (Guo *et al*., 2005). A reprint is provided in Appendix F.

7.1.6. 2D BTEM: Application to Reaction System

As mentioned before, as a non-invasive analytical method, NMR provides information about constituents from the data collected during the reaction. Therefore in this section, a reactive multicomponent mixture system was investigated by an HSQC experiment. 2D-BTEM was also applied to this data set to verify the methodology.

7.1.6.1. Experimental Section

Cycloaddition is a kind of reaction in which two or more unsaturated molecules react with the formation of a cyclic adduct. In this experiment, the same reaction system described in chapter 4, section 4.5.2 was used. The two reagents were 1,3-cyclohexadiene 97% (Aldrich) and dimethyl acetylenedicarboxylate 99% (Aldrich). Two reactions were carried out with different initial ratios of reagents. The temperature for both reactions was 310K. The 2D HSQC NMR spectra were taken circa every 40 mins.

All the data were acquired on a Bruker Avance 400 NMR spectrometer. 2D HSQC were collected. The 2D spectra were processed as 1K-1K complex matrices in both dimensions. The final data set with twelve spectra were collected for 2D-BTEM.

7.1.6.2. Computation Section

Data Pre-treatment

The same strategy used in section 7.1.5.2 was implemented for reaction data, only one rectangular region containing the physical spectral features (peaks) were taken in order to lighten the computational burden. The rectangular regions from twelve samples were assembled into a new data array $\underline{A}_{12\times401\times301}$. In Figure 7.13, the mesh plot and the contour plot of one reaction mixture spectrum are shown.

A close examination of all these spectra found that all the signals were not stationary during the reaction. This characteristic can be shown by examination of the tracks of peak centres in mixture spectra during the reaction.

From Table 7.1, it is easy to observe that, during the reaction, for example, the *Y* coordinate of the center of peak3 was moving from channel 76 to 79. It means during the reaction the signals were not stationary and the peaks were moving around. This shift definitely introduced an unavoidable non-linearity to the system and destroyed the bilinear model. Therefore, it was a prerequisite to eliminate or reduce the nonlinearities before further consequent analysis.

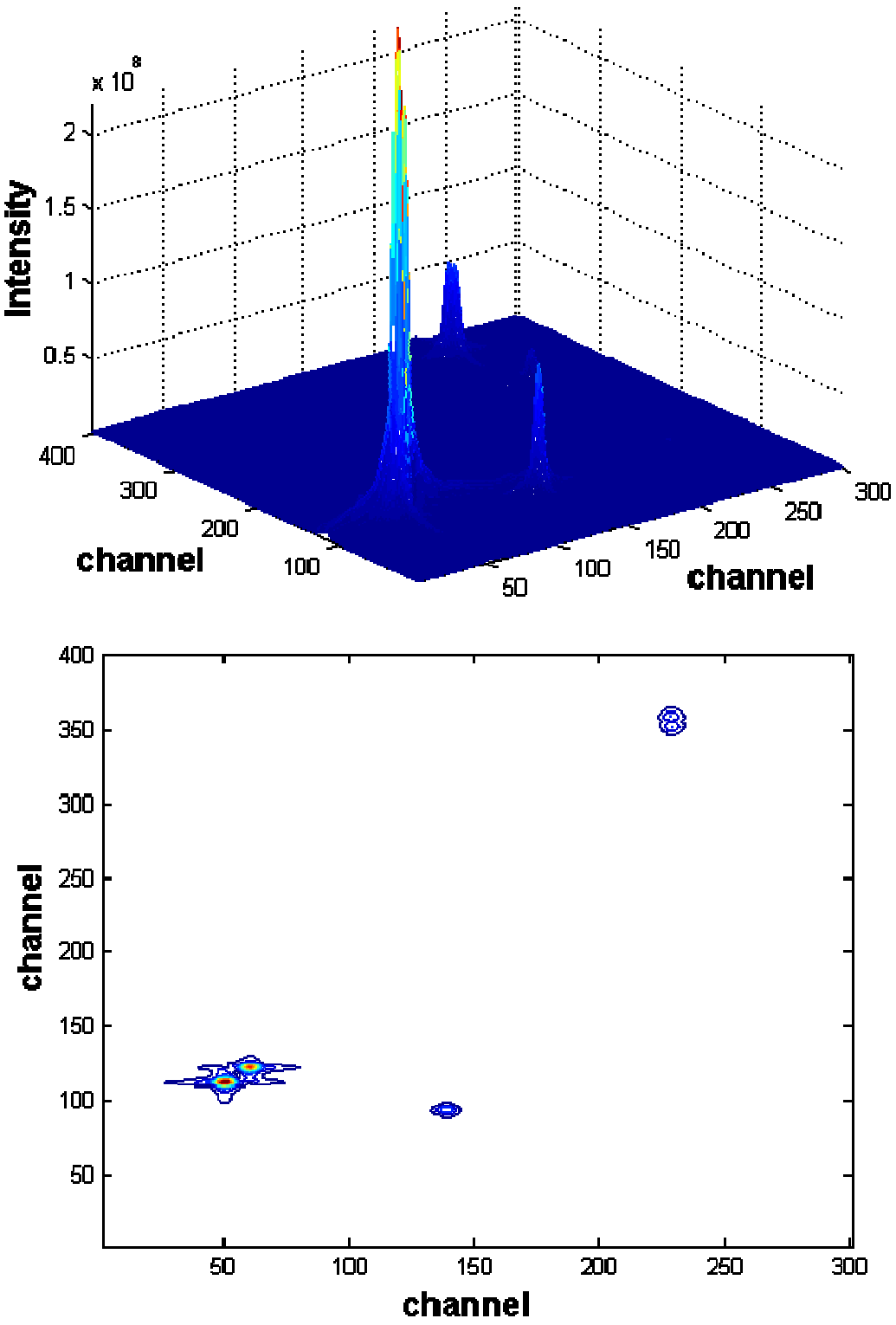


Figure 7.13. The mesh (top) and contour (bottom) plot of one reaction mixture spectrum.

Table 7.1. The coordinates of peak centres for the 6 peaks in 12 spectra

Spec. no.	X-coordinate(abscissa)						Y-coordinate(ordinate)					
	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Peak 6	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Peak 6
1	50	59	137	227	228	242	97	104	75	342	333	267
2	50	59	137	227	228	242	98	105	76	343	334	267
3	50	59	137	227	228	242	98	105	76	343	334	268
4	50	59	137	227	228	242	99	105	77	343	334	268
5	50	59	137	227	228	242	99	106	77	344	335	268
6	50	59	138	227	228	242	99	106	77	344	335	269
7	50	59	138	227	228	242	100	107	77	344	335	269
8	50	59	138	227	228	242	100	107	78	345	336	269
9	50	59	138	227	228	242	100	107	78	345	336	269
10	50	59	138	227	228	242	101	107	78	345	336	270
11	50	59	138	227	228	242	101	108	78	346	336	270
12	50	59	138	227	228	242	101	108	79	346	337	270

A good 2D alignment method may help to realign all these peaks systematically. But it is apparent that the alignment with constant move is impractical since some peaks heavily overlap in 2D NMR spectra.

A practical approach is to find a suitable model for the 2D NMR spectra. With the 2D peak fitting we can separate overlapping signals and do further alignment.

2D Peak Fitting

The first challenge to do the 2D peak fitting is to determine all the peak centers. It is well known that the residual method (Peakfit v4.0 User's Manual) and derivative method (Kauppinen *et al.*, 1981) can be used to find out the peak center in 1D spectroscopy. Unfortunately it is not so easy to implement the above methods in 2D spectroscopy. In this study, a simple and easy strategy was used to find out the vertex

position of each peak in the 2D coordinates. This consists of projecting along the x and y coordinates to help distinguish overlapping peaks.

The determination of peak center is often a preliminary requirement for us to investigate the shift problem in this series of 2D NMR spectra. Only after obtaining all peak center information can we approach subsequent analysis such as curve fitting. The curve fitting procedure would be very useful for the removal of nonlinear and noisy components from spectra. Curve fitting methods have been widely used to determine the area and parameters of absorption bands and to separate overlapping bands in a composite contour. 1D curve fitting is a quite common technique used in spectral analysis. 1D spectroscopic curve fitting has been investigated in detail (Chen and Garland, 2003; Vickers *et al.*, 2001). But to our knowledge little work has been performed on 2D curve fitting problems. In a similar spirit to 1D curve fitting, the least-square method is applied to construct the objective function for minimizing spectral difference as shown in Eq. 7.2.

$$\textbf{Min} \quad f = (A_{measurement} - A_{model})^2 \tag{7.2}$$

It is known that in an ideal case, each transition in an NMR spectrum should be represented by a Lorenzian shape since Lorentzian functions are theoretically related to the Fourier transformation of a decaying exponential signal (Carazza, 1976). Therefore, in accordance with the characteristic of NMR signals, a 2D Lorenzition model was selected as a model for 2D NMR. For 1D spectroscopy, the peak shape function of a Lorentzian which is centered about the frequency v^0 can be formulated as

$$a(v) = \frac{K}{1 + \left(\frac{2 \times (v - v^0)}{W} \right)^2} \tag{7.3}$$

where K denotes its amplitude, v^0 denotes the peak center parameter and W is the width parameter.

In the 2D case, a 2D spectrum can be modelled by superimposing several 2D peaks located in different positions. And the expression of a 2D spectrum with n peaks is formulated in a more complex form (Eq. 7.4)

$$A(v_x, v_y) = \sum_i^n \frac{K_{xi} \times K_{yi}}{\left(1+\left(\frac{2(v_x - v_{xi}^0)}{w_{xi}}\right)^2\right) \times \left(1+\left(\frac{2(v - v_{yi}^0)}{w_{yi}}\right)^2\right)} \tag{7.4}$$

For completeness, the Pearson VII 2D model (Ord, 1972) is noted here. (discussed in section 6.2.1.1.) Pearson VII model is one of the most useful band shape functions for combination models. Obviously, Pearson VII model is more flexible than Guassian and Lorentzian model.

A 2D spectrum can be modelled as a superposition of several 2D peaks located in different positions. After that, an efficient method is needed to solve this fitting procedure. Several parameters in the model are needed to be optimized. In this study, the 2D Lorenzian model was selected, since less parameter is needed in comparison to the Pearson VII model.

The conventional gradient-based optimization methods and the stochastic optimization methods including GA and SA methods are all available to obtain the final optimal band parameters.

Singular Value Decomposition

After the 2D curve-fitting procedure, a new data array $\underline{A}_{12\times401\times301}$ was obtained. SVD was performed on the matrix $\underline{A}_{12\times401\times301}$ to obtain the 12 right singular vectors in $V^{T}_{12\times(401\times301)}$. Concatenation was undone. Physically meaningful spectral features were observed in only the first 5 matrices. The 6th matrix was essentially featureless. This holds true for the others as well.

7.1.6.3. Result

From the right singular matrices, it is easy to find the interesting features in the data set. By targeting these interesting features, three 2D spectral patterns were recovered with the 2D-BTEM algorithm. The estimated 2D spectra are shown in Figure 7.14. It is found that in Figure 7.14, (*a*) and (*b*) are very similar with reagent (*d*) and reagent (*e*). And the estimated (*c*) is the cycloaddition product.

The dual problem for relative concentrations was also solved. Figure 7.17 shows the relative concentrations as determined by a least-square fit with the reference spectra versus the relative concentrations as determined by a least-square fit with the estimated 2D spectra. The profiles of the reagents monotonically decreased along the experiment. The concentration profile of product is shown as an increasing profile throughout duration of the reaction period.

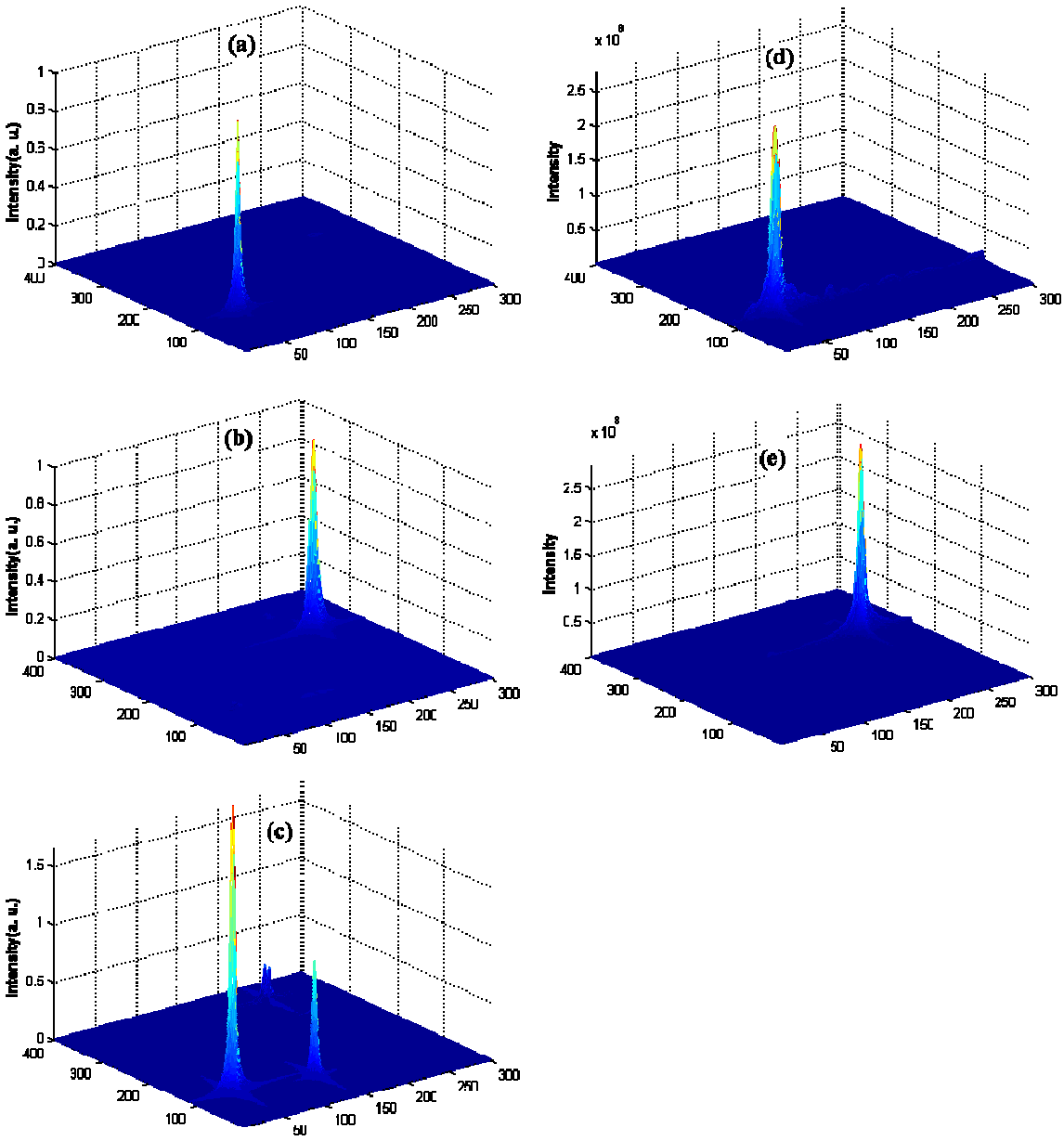


Figure 7.14. Estimated spectra (*a*, *b* and *c*) and the reference (*d* and *e*).

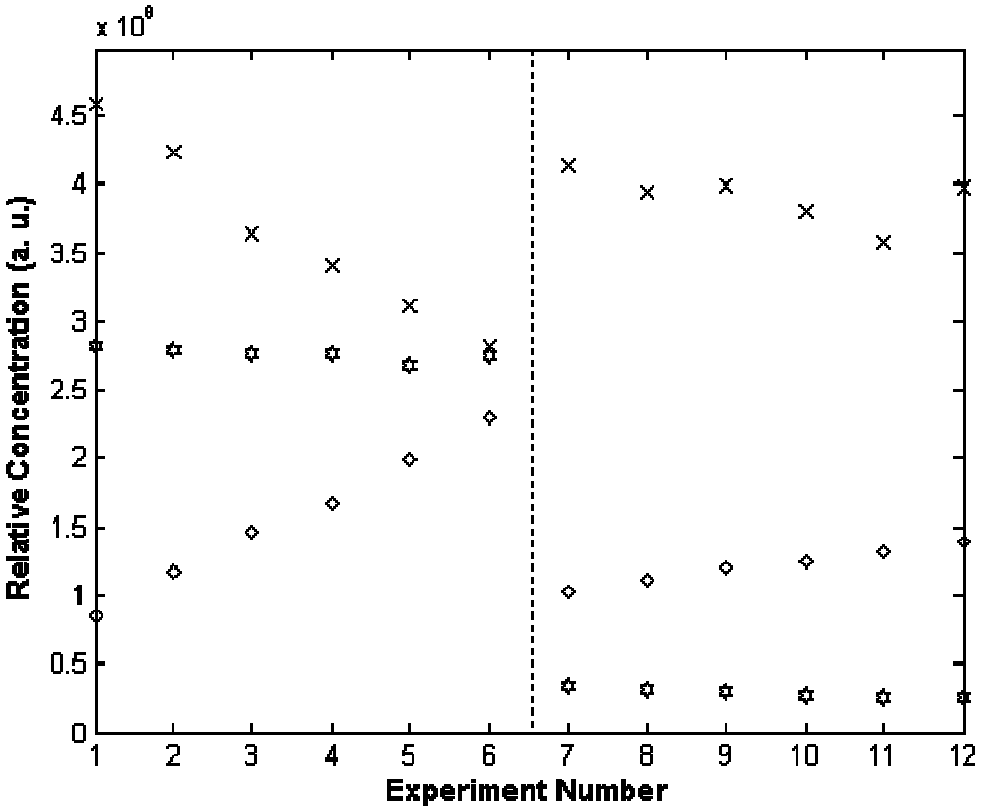


Figure 7.17. The relative concentration profiles. Cross: dimethyl acetylenedicarboxylate; Six-point star: 1,3-cyclohexadiene; Diamond : product.

7.1.6.4. Discussion

In a real chemical system, the signals are contaminated by noise. And the peak shape would not be symmetric. Especially in NMR there are some phase problems which create both positive and negative parts of the spectra. It is apparent that it is a straightforward idea to model the entire 2D resonance signals and align them in a systematic way. Further work is needed to alleviate the phase problem possibly by a combination of hardware/pulse considerations and post processing of data.

Also it should be noted that in some cases, the baseplane (two-dimensional baseline) would affect the curve fitting process, so it is also important to implement the baseline correction if necessary.

7.2. Application of 2D Band-Target Entropy Minimization (2D-BTEM) to Fluorescence Data

7.2.1. Introduction

In a modern chemical laboratory, it is common that there is a lot of multi-channel detector data or multiplexed spectroscopic data which are different from conventional one-dimensional data. In a simple 2D fluorescence measurement, the excitation/emission matrix (EEM) is regarded as a function of both the excitation and emission wavelengths. Multidimensional fluorescence measurements circumvent the deficiency of the conventional single-wavelength measurement in evaluating multicomponent samples; there the latter always results in broad, structureless and severely overlapping spectra (Townshend, 1995). Wide applications of multidimensional fluorescence in chemistry (Nahorniak *et al.*, 2005) and environmental science (Muroski *et al.*, 1996; Booksh *et al.*, 1996) (especially for the identification and quantification of polycyclic aromatic hydrocarbons (PAH))(JiJi *et al.*, 1999; Hart *et al.*, 2002), clinical analysis (Olivieri *et al.*, 2004), drug science (da Silva *et al.*, 2002) and food analysis (Bro, 1998) have made use of the sufficient information contained in EEM to differentiate species, detect minute perturbations in mixture samples and monitor the chemical and bio-chemical processes.

The resulting matrix-formatted data obtained from one measurement generally possesses more chemical characteristics and chemical information than a 1D measurement. As mentioned in section 2.2.2, there are two major categories to consider in the mathematical sense: bilinear data and non-bilinear data.

The matrix-formatted data of 2D fluorescence in a dilute solution is the prototype for ideal bilinear data. Other types of data with “bilinear” structure are LC-MS, GC-MS

and HPLC-DAD data (Frenich *et al.*, 2000). Mathematically, an ideal EEM of a single fluorophore (matrix Q) can be represented as the outer production of the excitation spectrum (vector x) and emission spectrum (vector y) (Eq. 7.10)

$$Q = xy^T \tag{7.10}$$

In the case of the mixture with a multiple fluorescent species, it is assumed that the measurement is additive over constituents and the response of a mixture sample is the sum of the individual constituent response. The Eq. 7.10 is generalized into Eq. 7.11 where I , the EEM measurement of the mixture sample, is the sum of s "pure" responses from individual species.

$$Q = \sum_{i}^{s} x_i y_i^T \tag{7.11a}$$

or

$$Q = XY^T \tag{7.11b}$$

where X, Y are matrices whose columns x_i and y_i are vectors representing the excitation and emission spectrum of each pure component, respectively. Further, with simplification, a collection of fluorescence EEMs for mixture solutions with differing composition can be represented by the expression:

$$\underline{Q} = XCY^T \tag{7.12}$$

where $\underline{Q}$ is the stacks of the EEM measurements (a three-way array), C is a matrix, whose element represents the relative concentrations of the constituent in samples. By introducing the dimension of concentration, the collections of data with bilinear form as Eq. 7.10 produce a trilinear data set.

For data with bilinear structure, especially fluorescence EEM data, quite a lot of chemometric methods are available. A detailed review of RAFA and GRAM methods can be found in chapter 2, section 2.2.2. Besides RAFA and GRAM, Direct Trilinear Decomposition (DTLD) also has variety of applications in fluorescence data which focuses on the trilinear property of fluorescence EEM data. (Leurgans and Ross, 1992; Frenich *et al.*, 2003). In 1990, Neal *et al.* (1990) proposed a constrained nonlinear optimization method to solve the mixture spectra by imposing the constraints of nonnegative, diagonal and degree of the overlapping.

A collection of matrix-formatted data will lead to a three-way structure, by introducing a concentration axis, for example, collecting a series of samples along the time direction for a reaction system. Parallel Factor (PARAFAC) analysis was a data analysis tool originating from Psychometrics. Inspired by the early work of Cattell (1944), Harshman (1970) and Carroll and Chang (1970) independently developed an easily interpreted model for fitting an n-linear model to an n-way array. There are a number of applications of PARAFAC on fluorescence data (Thygesen *et al.*, 2004; Stedmon and Markager, 2005; Christensen[a] *et al.*, 2005; Nahorniak *et al.*, 2005; Christensen[b] *et al.*, 2005). Actually, the mathematical model of PARAFAC will be regarded as the three dimensional extension of GRAM, and DTLD algorithm.

In this section, we will present the applications of 2D-BTEM to real experimental 2D fluorescence EEM data including a simulation data set and an experimental data set carried out at our laboratory. Also the comparison of the result of 2D-BTEM with the result of the PARAFAC method is depicted.

7.2.2. Simulation Data

Spectra of amino acids which contained in the data set named Claus.mat was downloaded form website[ii] .The samples were generated and measured by Claus A. Andersson (Bro, 1997). The original data set consisted of five simple laboratory-made amino acid samples in phosphate buffered water solution. These samples contained different amounts of the amino acids: tryptophan, tyrosine and phenylalanine. All the data were acquired on a PE LS50B spectrofluorometer. Technical parameter: excitation 250-300 nm, emission 250-450 nm, 1 nm intervals , excitation slit-width of 2.5 nm, emission slit-width of 10 nm and a scan-speed of 1500 nm/s. Seven simulated data set were produced through multiplying the original data set with a random matrix, also complicated by adding white noise. The size of the data array is $7 \times 61 \times 201$.

7.2.2.1. Singular Value Decomposition

After SVD decomposition, a series of right singular matrices are obtained. In Figure 7.18 several right singular matrices and one mixture sample are shown. From Figure 7.18 we can directly observe that only the first several right singular matrices are relevant, since from the fifth right singular matrix, most of matrices represent noise only.

7.2.2.2. Result

Accordingly, the series of right singular matrices were truncated and transformed appropriately by the criteria of entropy minimization due to 2D BTEM.

Since these mixture spectra were moderately overlapping, a first derivative entropy measurement was used in this case. The result shows that all the excitation and emission

[ii] http://www.models.kvl.dk/research/data/Amino_Acid_fluo/index.asp

matrices of the pure components in this multi-component system were successfully recovered. These three estimated pure spectra are close matches for the pure spectra of tryptophan, tyrosine and phenylalanine.

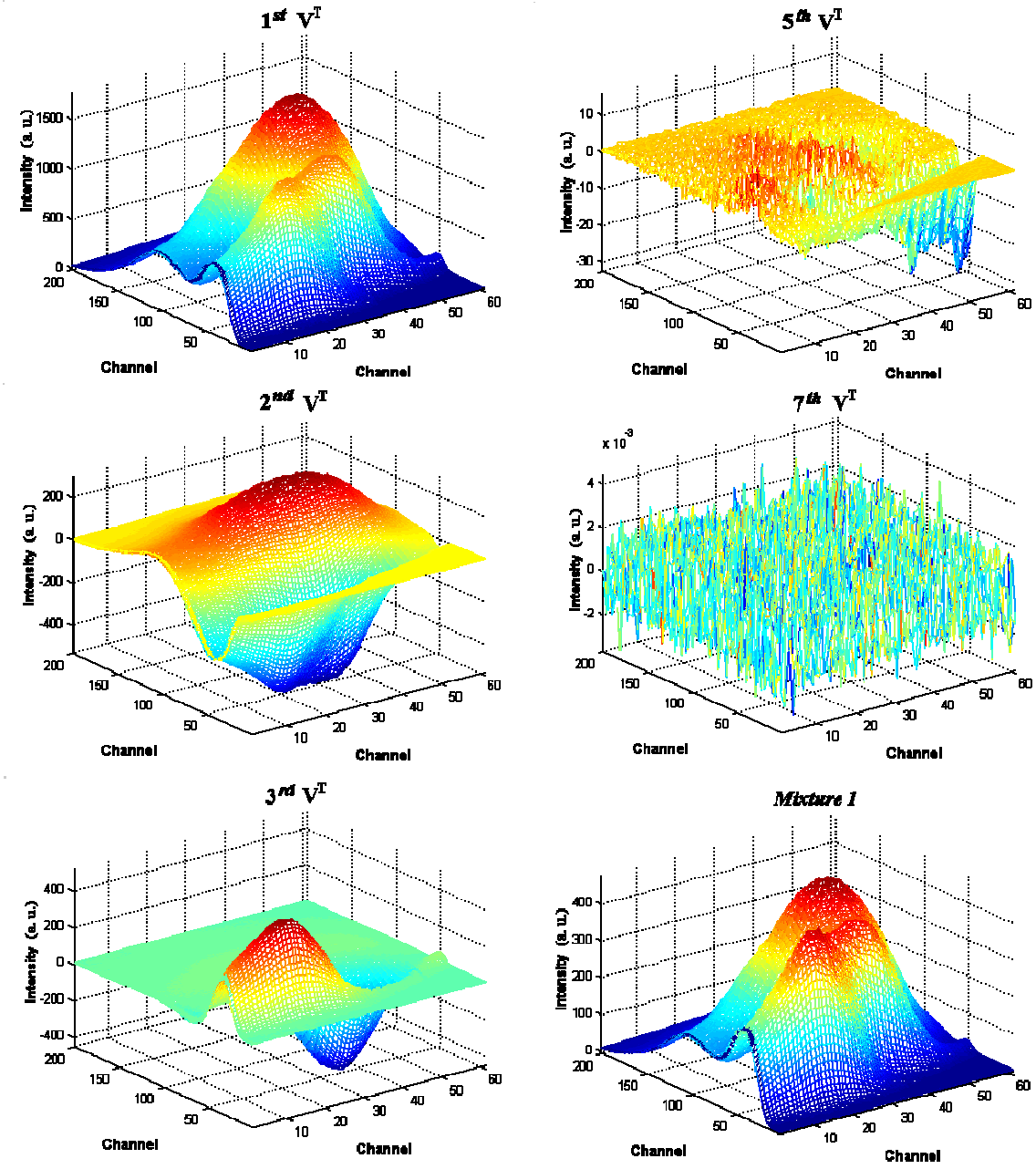


Figure 7.18. Mesh plot of some right singular matrices (1st, 2nd, 3rd, 5th and 7th) resulting from SVD procedure and one simulated mixture data set which consists of 3 amino acids. (shown in channels)

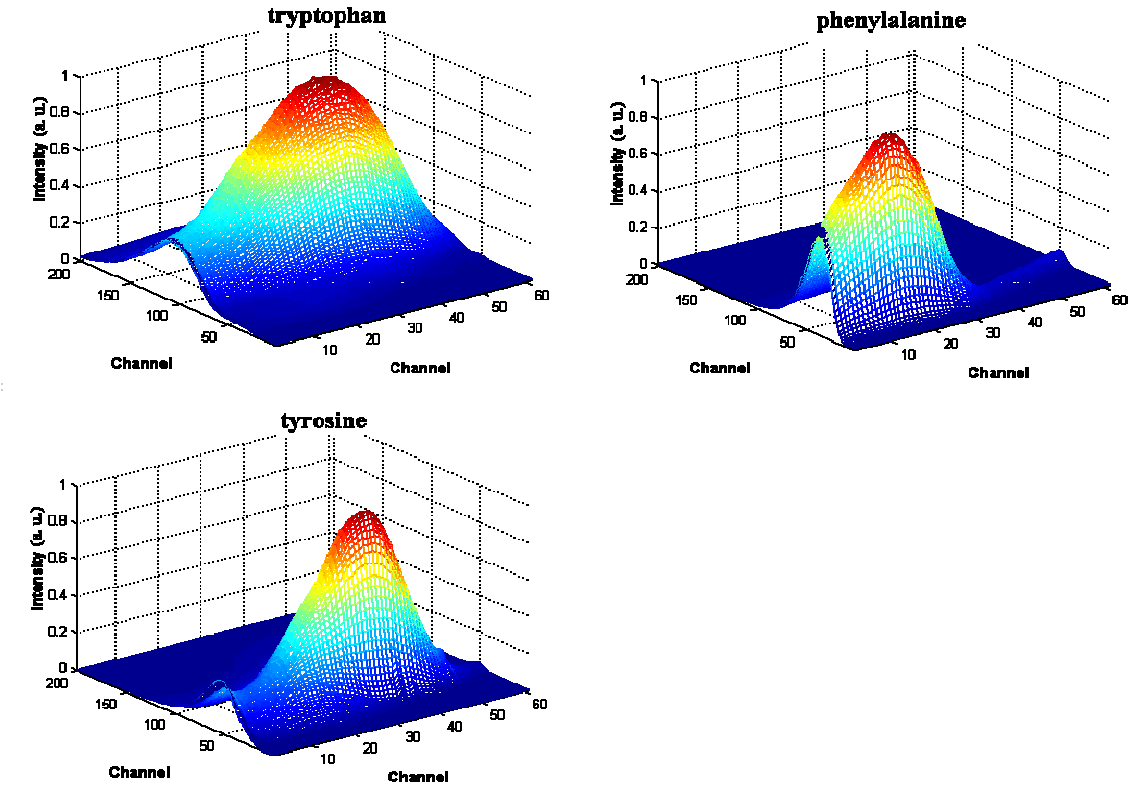


Figure 7.19. Mesh plots of the estimated pure spectra of the pure components extracted by 2D BTEM. (shown in channels)

7.2.3. Experimental Data

7.2.3.1. Experiment Section

Sample Preparation

First, phosphate buffer was prepared (0.2mol/L, pH7.4): using 9ml NaH_2PO4 (0.2mol/L), and mixing with 81ml $Na_2HPO4{\cdot}12H_2O$ (0.2mol/L), then adjusting to pH 7.4. This solution was then stored at room temperature.

Three amino acids: L-tryptophan (99%, Acros Organics), L-tyrosine (99+%, Aldrich) and L-phenylalanine (98.5+%, Acros Organics) were dissolved in deionized water. Dilute solution should be prepared, since it is important that the concentrations

must not be too high. Therefore, the stock solutions: tyroptophan 0.68 $\mu g / ml$, tyrosine 2.4 $\mu g / ml$ and phenylalanine 8.30 $\mu g / ml$ were prepared. Seven mixture samples were prepared according the Table 7.2.

Table 7.2. The mixing table for preparation of mixture samples with the stock solutions

Sample	Mixing Table (ml)					
	tryptophan	tyrosine	phenylalanine	phosphate buffer	deion. water	total vol.
1	0.4	0.9	0.7	1.0	12	15
2	0.1	0.7	0.4	1.0	12.8	15
3	0.2	0.2	0.8	1.0	12.8	15
4	0.3	0.4	0.1	1.0	13.2	15
5	0.5	0.5	0.5	1.0	12.5	15
6	0.7	0.6	0.9	1.0	11.8	15
7	0.9	0.3	0.6	1.0	12.2	15

Apparatus

Fluorescence spectra were obtained with a Perkin-Elmer LS50B luminescence spectrofluorometer. Technical parameter: excitation 200-350 nm, emission 200-450 nm, 0.5 nm intervals , excitation slit-width of 10 nm, emission slit-width of 10 nm and a scan-speed of 1500 nm/min. The spectrometer was interfaced to a computer. Seven mixture samples were measured as well as the reference solutions.

7.2.3.2. Data Pretreatment

Removing Rayleigh Scattering Regions

It is known that in fluorescence, there are interferences (Townshend, 1995). Solvents, containing carbon-hydrogen and oxygen-hydrogen bonds, may produce Raman

Scattering bands. Also several background signals and noise including Raman scattering from the cuvette windows, Rayleigh scattering from solvent(s) exist. All these interferences would cause serious problems in quantitative analysis of fluorescence EEM data. Rayleigh scattering as an elastic scattering of exciting light also causes interference. Details of Rayleigh scattering and Raman scattering and the background signals can be found in references(Ingle *et al.*, 1988; Lakowicz, 1999). Another possible serious interference may come from the contamination of the cell with luminescent impurities(JiJi and Booksh, 2000).

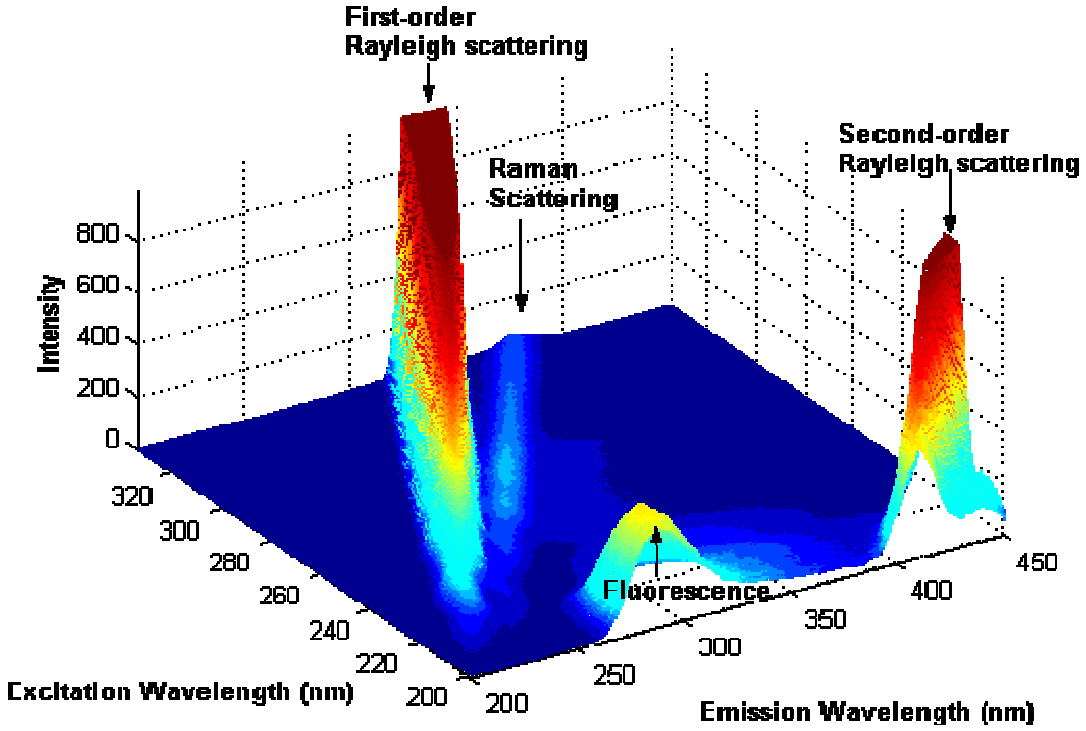


Figure 7.20. The mesh plot of the pure phenylalanine sample. The 1st order , 2nd order Rayleigh scattering and Raman scattering are critical background signals.

Since the Rayleigh scattering is so strong, it has to be removed. The reason to discard the Rayleigh scattering part is that it does not fit in the linear model and would destroy the bilinear model. Also the high intensity of Rayleigh scattering sometimes

exceeds the limit of measurement. JiJi and Booksh(2000) discussed the different ways of reducing and removing Rayleigh and Raman scattering, and they summarized four weighting strategies, namely, hard positive, soft positive, hard negative, and soft negative weighting, which either can be used to enhance the interesting signal region or eliminate the nonlinear parts. Also Rinnan *et al.* (2005) modeled the first order Rayleigh scattering by rotating and shifting the spectra.

In this study, we simply cut the Rayleigh scattering regions and replace them with zeros values. Reference spectra after removing Rayleigh scattering regions are shown in Figure 7.21.

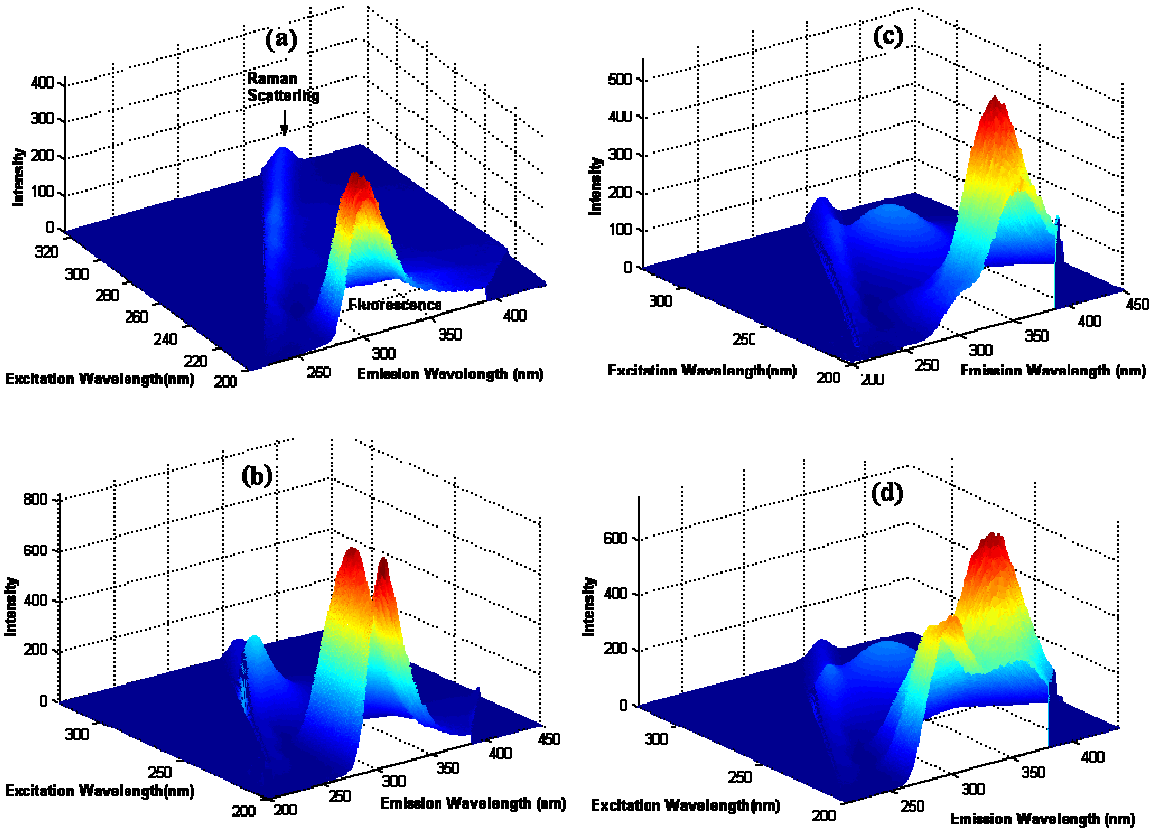


Figure 7.21. Reference spectra of phenylalanine (a), tyrosine (b), tryptophan(c) and a mixture example (d). It is shown that the fluorescence signals are prominent after removing some background signals.

It is important to emphasis that in Figure 7.20 the first-order Rayleigh scattering and second-order Rayleigh scattering should not be fitted by the linear model. Rather these elements must be set to zero values in order not to bias the bilinear model in this analysis. Also the original data matrix is truncated into 250×450 array (Emission wavelength range from 213.5nm to 438nm, Excitation wavelength range from 200nm to 324.5nm) since the most interesting fluorescence features are already included. As shown in Figure 7.21a, the fluorescence signal is prominent after removing some background signals in comparison to Figure 7.20. The only undesired part left is the Raman scattering which overlaps with some desired signal in Figure 7.21.

7.2.3.3. 2D BTEM

All the 7 measurement set of mixture matrix data were tested with the 2D-BTEM algorithm. First all matrices were decomposed into right singular matrices using SVD. Secondly, close examinations of all the right singular matrices indicate many interesting features. Thirdly, 2D-BTEM was applied to target the features one by one by manipulating these right singular matrices to obtain the pure source matrix in a systematic way.

The first several right singular matrices are shown in Figure 7.22. The first one approximates the average of all mixture spectra. But from the 2^{nd} one, it is easier to identify the interested features shown in the mesh plots. As indicated, the region, A, B, C, and D are the interested features which we would like to retain after entropy minimization transformation. After the 5^{th} right singular matrices, no physically meaningful spectral features were observed, but always some noisy parts show up at the region E indicated in the Figure 7.22 which consists of heteroscedastic noise, i.e. the variance of noise changes with the intensity of the signal. Therefore, we can conclude that only the first several right

singular matrices are relevant, since from the fifth one, right singular matrices represent noises only. Accordingly, the series of right singular matrices are truncated and transformed by the criteria of entropy minimization according to 2D BTEM.

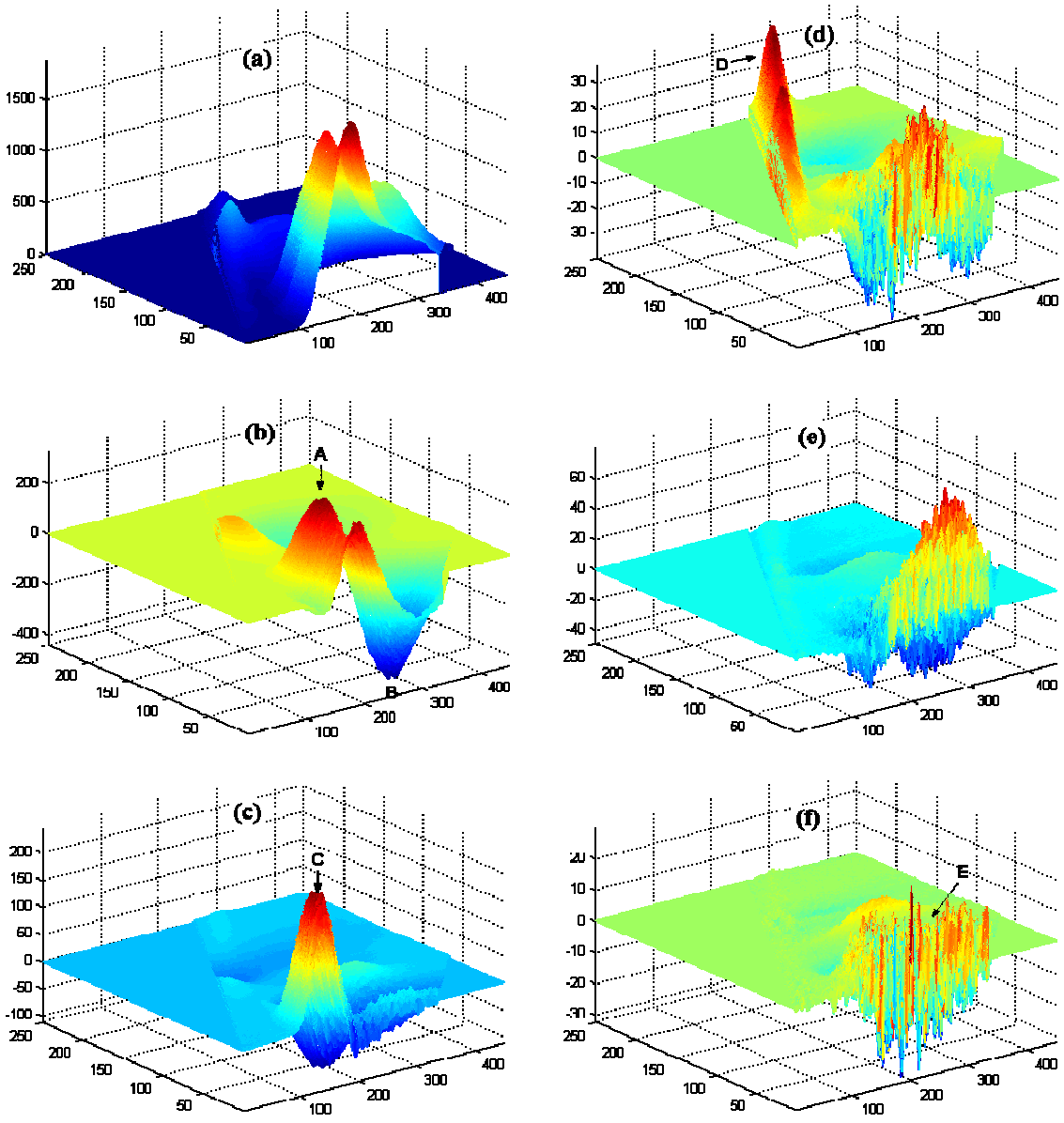


Figure 7.22. The mesh plots of the 1st (a), 2nd (b), 3rd (c), 4th (d), 5th (e) and 7th (f) right singular matrices. The x and y coordinates are now data channels and z is the arbitrary magnitude.

7.2.3.4. Result and Discussion

The fluorescence EEM spectra of the mixtures of 3 amino acids are highly overlapping. But they can be directly recovered by 2D-BTEM method without separation. The resultant estimated spectra by individually targeting region A, B, C are shown in Figure 7.23. The spectra are fairly good compared to the experimental references in Figure 7.24. It is important to note that in the experimental reference, each spectrum still consist of the ridges which originates from Raman scattering, but most of the interference signals are eliminated in the estimated results from the 2D-BTEM algorithm. By band-targeting the region D, we obtained the approximation of the Raman scattering component, which is shown as (d) in Figure 7.23. It is clear that the Raman scattering ridge is the dominant contribution. The other feature comes from the heteroscedastic noise which is difficult to separate.

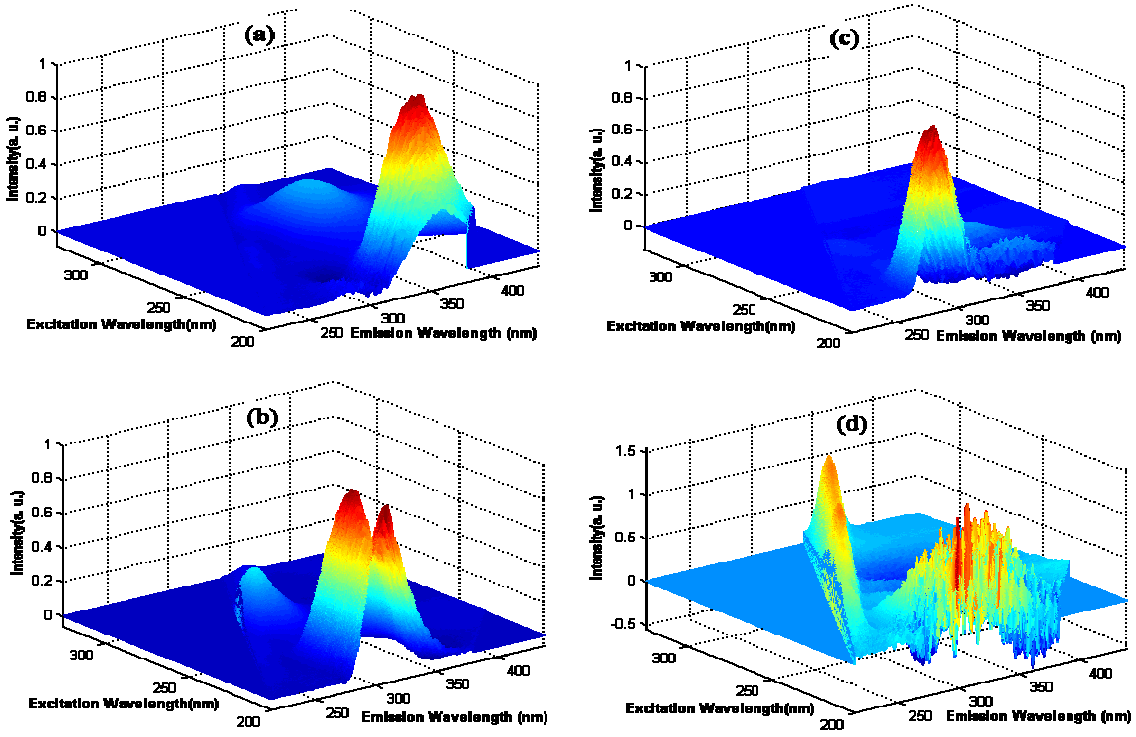


Figure 7.23. The four estimated components obtained by 2D-BTEM: tryptophan (a), tyrosine (b), phenylalanine (c) and Raman scattering (d).

Furthermore, the dual problem for the curve resolution can be solved after obtaining all the estimated spectra. The relative concentration profiles determined by a least-square fit with the estimated pure spectra versus the authentic experimental concentrations are tabulated in Table 7.3.

Table 7.3. The comparison of reference and recovered concentrations with three components and four components

Sample No.	**Refer. Conc.** $(10^{-7}\, g/ml)$			**Recovered Conc. (3)**			**Recovered Conc. (4)**			
	Tyro.	Phen.	Tryp.	Tyro.	Phen.	Tryp.	Tyro.	Phen.	Tryp.	Raman
1	1.44	3.87	0.18	944.92	365.88	364.06	937.33	359.61	359.38	12.49
2	1.12	2.21	0.05	745.76	249.20	182.10	729.98	236.16	172.37	25.97
3	0.32	4.43	0.09	354.37	434.15	256.12	330.36	414.31	241.31	39.53
4	0.64	0.55	0.14	479.37	98.40	302.37	458.59	81.21	289.55	34.23
5	0.80	2.77	0.23	605.17	297.01	447.17	592.01	286.12	439.05	21.68
6	0.96	4.98	0.32	718.50	481.67	613.94	711.92	476.23	609.88	10.83
7	0.48	3.32	0.41	480.41	336.16	786.20	466.91	325.00	777.87	22.23

Since the estimated spectra are normalized according to the 2D-BTEM algorithm, the real magnitudes of the pure spectra are left unknown. That is so-called scale ambiguity in the curve resolution problem, even though, the comparison can be represented after normalization. In Figure 7.24, the relative concentration profiles from a least-square fit and the experimental concentration are all normalized by their L^2 norm respectively. Figure 7.24 indicates that the calculated concentration profiles are quite similar with the real experimental concentration profiles. It is worth noting that in the Table 7.3, two sets of recovered concentrations are tabulated, one resulted from least-square fit with three estimated amino acid's spectra, and the other resulted from four components including the estimated Raman scattering component. Although the number of components used for the calculation is different, the concentration results are quite close which can not be discerned in the plot (Figure 7.24). The reason may lie in the fact that Raman scattering is

a quite distinct component which has little overlapping with other major components, and most of its systematic variation can not be explained and replaced with other components.

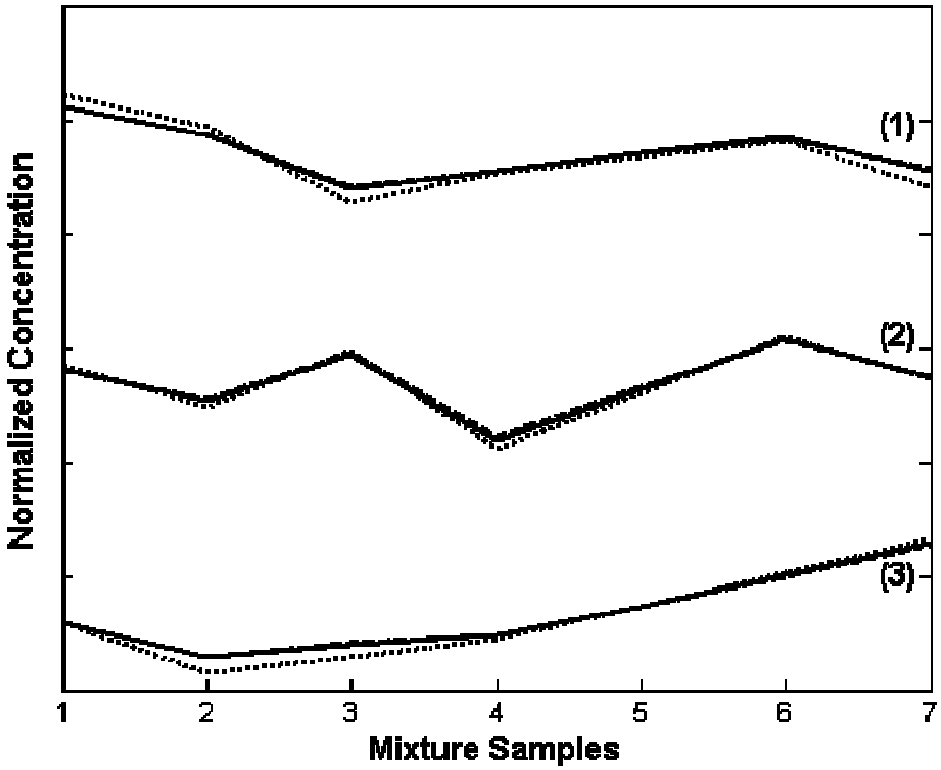


Figure 7.24. L^2 Normalized concentrations associated the seven mixtures. Dotted line represents the experimental concentration. Solid line represents the least-square fit result with three estimated spectra from 2D-BTEM. Dashed line represents the least-square fit result with four estimated spectra from 2D-BTEM. (1) tyrosine, (2) phenylalanine and (3) tryptophan.

Figure 7.25a shows the residual of one mixture spectrum by extracting these three recovered components. It is clear that the residual is close to the fourth component (d) shown in Figure 7.23 which is regarded as Raman Scattering. If four components are used in the least-square fit, the new residual is equivalent to the nonlinearity which cannot be accounted for inside the bilinear model (Figure 7.25b). This idea is further supported by calculating the signal recovery of the reconstruction. 94.29% of the signal is recovered

with three components and it would reach 97.68 % if the Raman Scattering component is included. So even though there is non-linearity imbedded in the data, the estimated pure components obtained from 2D-BTEM account for 97.68 % of the total integrated signals in the seven mixtures. The high degree of the recovery shows that the most of the important components in the system have been extracted.

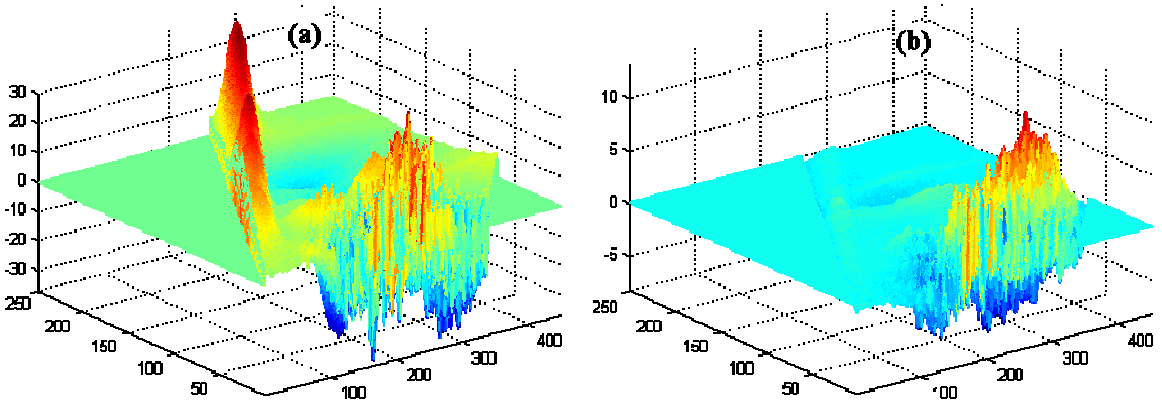


Figure 7.25. The residual of one mixture spectrum extracted by the reconstruction spectra with three recovered components (a) and with four recovered components (b).

7.2.3.5. Comparison with the PARAFAC (Trilinear Model)

As mentioned in introduction part, PARAFAC is a multi-way method which employs the trilinear model. For comparison, a MATLAB multi-way toolbox (Andersson and Bro, 2000) is used to implement the PARAFAC algorithm. It is difficult to determine the proper number of components, so an arbitrary guess, three and four are set to the PARAFAC modeling. A three-component PARAFAC solution is shown in left column of Figure 7.26. Also the four-component solution is shown in right column of Figure 7.26.

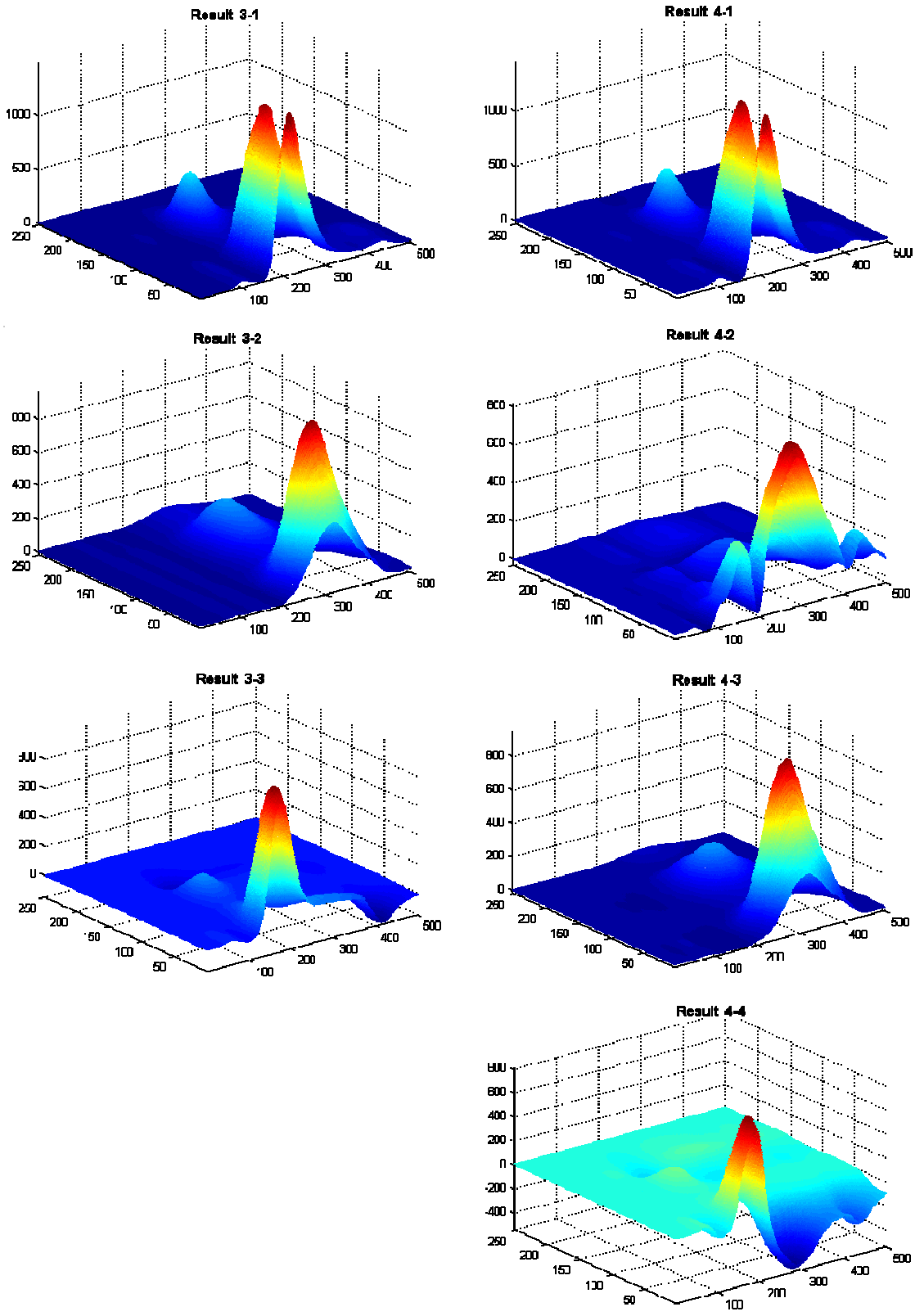


Figure 7.26. Result from PARAFA model with three components (left) and four components (right).

Theoretically, the PARAFAC solution should be identical to the real solution if the trilinear model is valid. In the PARAFAC, if the trilinear model is valid then the calculated core consistency from the test data should be close to 100% (Bro and Kiers, 2003). In this study, the core consistency was 91.7662% for three components, but only 48.8129 % for four components. These results counter the "appropriateness" of the trilineary model for this data set. The explained variation stated in the program was 98.9379 % (for three components) and 99.1523% (for four components) respectively – but as we see, the pure components are not necessarily correct.

Figure 7.27 shows one of the residuals of the mixture spectrum resulting from subtracting the three major components resulting from PARAFAC model. It is clear that the residual is not close to the Raman Scattering and still some fluorescence signals exist.

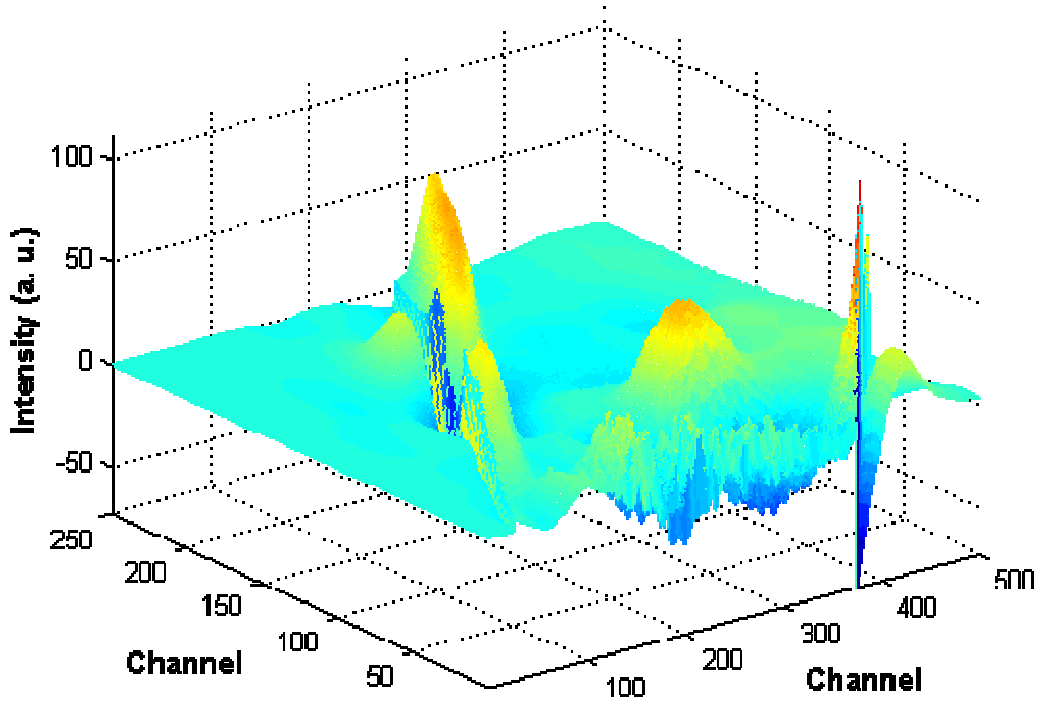


Figure 7.27. The mesh plot of one residual of a mixture spectrum after subtracting the three major components resulting from PARAFAC model.

7.2.3.6. Discussion

The Rayleigh scattering region with its large magnitude possesses a large part of signal; therefore it plays a very important part in the modeling. The Rayleigh and Raman scattering will deteriorate the mathematical modeling of the fluorescent spectroscopic data. Since the Rayleigh scattering cannot be regarded as a stationary component in the bilinear model, thus removing the Rayleigh scattering part will enormously reduce the nonlinearity of the data. Even though the Raman scattering still exist as it overlaps with the fluorescence signal, the removal of the Rayleigh scattering effects significantly facilitate the bilinear modeling and 2D-BTEM algorithm.

In real world, the theoretical mulitlinearity of the fluorescence data are not totally valid, it is apparent that there is nonlinearity which prevents the algorithm from performing properly as it is heavily based on assumption of trilinearity. This nonlinearity may originate from the troublesome background signals, and noise. Also there are still two issues related to the trilinear modeling, the robustness of the algorithm when applied to noisy data and the prediction of correct number of the components.

7.2.3.7. Conclusions

The application of 2D-BTEM on the fluorescence EEM data was successful where 2D-BTEM has been proved to be a very powerful tool to deconvolute the 2D fluorescence spectroscopy. 2D-BTEM has been applied to simulated and experimental spectra. The quality of recovered spectra is good compared with experimental references. These results have implications for a large variety of intrinsically inseparable multi-component mixtures system encountered in science research.

7.3. Other Types of 2D Spectroscopic Data

Two-dimensional (2D) NMR methods have been subject of great interest in recent years because these methods allow measurement of the intra- and intermolecular interactions that are central to structure-function relationships in chemical and biochemical systems. There are various types of 2D vibrational spectroscopy analogous to 2D NMR spectroscopy(Wright, 2002), namely, 2D-IR (Zhao and Wright, 2000; Zhao *et al.*, 2000), 2D-Raman(Tanimura and Mukamel, 1993; Keunok *et al.*, 2003; Wright, 2002).

MS-MS is another type of important 2D spectroscopic data. Mass Spectrometry is an important identification tool used in industry and academia for both routine and research purpose. It provides accurate molecular weight of the charged ions created from the chemical molecules of interest and this information is used to identify of studied chemicals. Tandem mass spectrometers (MS-MS) are instruments that have two analyzers or more analyzers in a tandem arrangement. It can be used for structural and sequencing studies of complex molecules. During the parent ion scanning, the first analyzer allows the transmission of all samples mass. The specific fragments ions would be further generated by bombarding the sample ions which would be observed by the second analyzer. Tandem mass spectrometer is extremely useful to give most unambiguous information than the conventional mass spectroscopy which makes it a powerful tool in organic synthesis and biomolecules analysis.

Electron paramagnetic resonance (EPR) technique is also often called ESR (electron spin resonance) or EMR (electron magnetic resonance) (Wertz and Bolton, 1986). Similar to NMR spectroscopy in the presence of a static magnetic field, instead of the absorption of radio frequency electromagnetic waves, EPR is an absorption spectroscopy involving the absorption of microwave irradiation and detects species that

have unpaired electrons. With a wide range of applications in chemistry, physics, biology, and materials science, medical science, EPR (Pluschau and Dinse, 1994; Ren *et al.*, 2004; Fauth *et al.*, 1991; Pluschau and Dinse, 1995; Kababya *et al.*, 1996) are used to probe the "static" structure of solid and liquid systems, also the dynamic processes. The 2D EPR is natural 2D spectroscopic pattern.

7.4. Summary

The focus of this chapter was applying the 2D-BTEM in real experimental 2D NMR (HSQC and COSY) data. Also successful application in 2D fluorescence data verifies its general applicability. In all these analysis, the resolved spectra and concentration profiles are satisfactory when compared with references obtained from pure component measurements.

Chapter 8

Three-Dimensional Entropy Minimization Algorithm

In this chapter, the entropy minimization method is further extended to three-dimensional version. The potential application of 3D entropy minimization in multidimensional Nuclear Magnetic Resonance spectroscopy is illustrated.

8.1. Multidimensional Nuclear Magnetic Resonance Spectroscopy

NMR is the most powerful tool for determining the structures of the molecules. The 1D conventional high resolution NMR spectroscopy can be easily interpreted for small molecules. However, using 1D NMR for more complex molecule, the ^{1}H spectra will consist of many resonances. The assignment of resonance is proved to be difficult for two reasons. First, it is still quite challenging to accurately predict the chemical shift of each hydrogen atoms. Second, when the number of resonances in the spectrum increases, and heavy overlapping happens, the assignment and the successive analysis of component will be complicated. Resolving the overlapping problem by spreading the information along two orthogonal dimensions instead of one axis, 2D NMR is widely used as it provides an extraordinary increase of resolution, thus making the analysis of the complex spectra possible.

Further, a 3D NMR experiment can be constructed from a 2D NMR by introducing another evolution time and a second mixing period, equivalently, adding the third dimension in the resultant data. The first type of 3D experiment can be constructed by combining two kinds of 2D experiments, such as HSQC-TOCSY, HSQC-NOESY, and TOCSY-NOESY. It consists of correlating the various nuclei either through scalar

coupling (COSY, TOCSY, HMQC, HSQC) or through space (NOESY), spreading this overlapping along the third chemical shift axis by combining two "classical" 2D experiments. The addition of the third (or further, the fourth) dimension would increase the resolution of resonance and reduce the overlap due to the smaller chance of signal overlaps which are dispersed in an enlarged space.

Second kind of the 3D experiments is called "triple resonance" (Bax *et al.*, 1990) which is a heteronuclear NMR experiment involving 3 (or more) nucleis. Typically, ^{1}H, ^{13}C, ^{15}N and ^{31}P are correlated. The experiments are frequently performed on doubly labelled (^{13}C, ^{15}N) proteins (Oschkinat *et al.*, 1988). It is worthy to note that with the aid of the multidimensional NMR technique, four dimensional and even higher dimensional NMR is possible. A representation for a 3D NMR spectrum is shown in Figure 8.1.

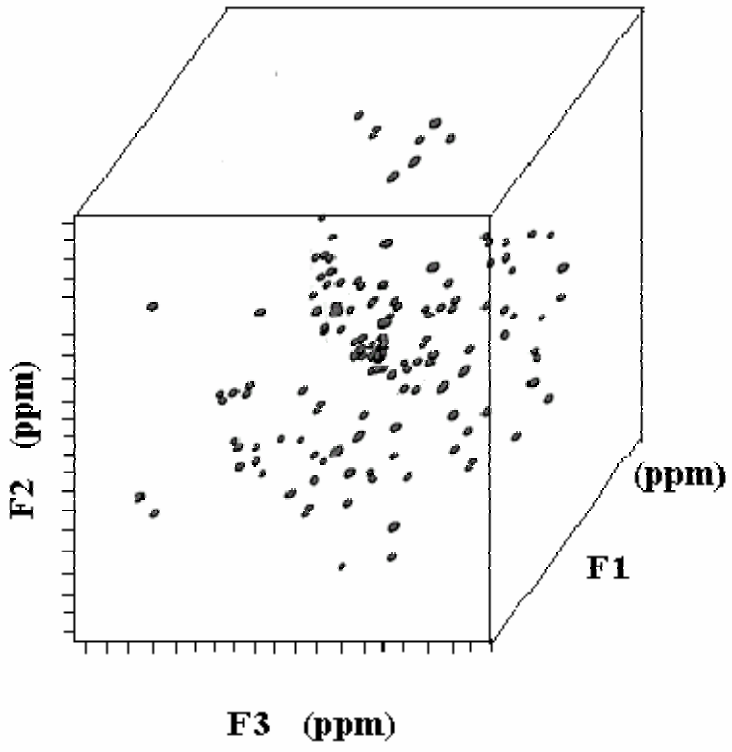


Figure 8.1. A representation for a 3D NMR spectrum (not actual data).

Since the technical consideration relate to the NMR spectrometer, the time needed to collect the higher dimensional NMR data set normally is quite long, often many days. In practice, the acquisition of FIDs (free induction decay) are truncated, which often introduce the distortion and the noise that stretches across the entire spectra. Considerable efforts have been invested in the multidimensional NMR data analysis, Linear Prediction method (Hoch and Stern, 1996; Stern *et al.*, 2002) and Maximum Entropy methods (Zhu, 1996), filter diagonalization method (Mandelshtam, 2001), etc.(Lin and Hwang, 1993) which are widely documented in literature. But to our knowledge, there is still no chemometric techniques methods address to the data analysis of the mixture system.

8.2. Visualization of 3D Data

It is known that high dimensional data visualization is very important in data analysis since it gives a direct view of data. We already discussed the display of a 2D data with 3D mesh plot to depict their positions and corresponding intensities. But it is hard to directly display data with more than three dimensions. So it is impossible to directly display a 3D data and their intensity in a 3-dimension coordinate. Therefore the users should understand that the display of data with more than three dimensions has to be transformed in some way before they can be rendered.

In this section, three-dimension contour plot is employed which is represented as an iso-surface graph, that is, a surface in a three-dimensional space where the intensity values of each point are the same. An iso-surface is essentially the intersection of a surface cutting through a volume of data. For example, as shown in Figure 8.2a, the iso-surface graph suggests that the points on the surface of ball region have same value. But still there is not an immediate way to show the 3D data and their intensities without anther

additional artificial dimension, for example, color index. In the other words, if data are in the form with coordination of *x,y,z* and intensity, one possible way is that we can represent 3D data on a plane (using *x,y* and *z* coordinates) and use colours to represent their intensity.

8.3. Overview of the 3D Entropy Minimization Approach

The basic philosophy behind the present methodology is analogous to 2D entropy minimization method. The approach consists of three primary stages.

First, the series of 3D tensors are systematically arranged into a 2D array which is indexed by sample-number of the spectra in one direction, and by unfolding each 3D tensor into one vector in another direction. This 2D array is decomposed into orthogonal components using PCA (Principal Component Analysis) or SVD (Singular Value Decomposition) etc.

Secondly, an objective function is created using appropriate entropy like function with penalty terms which aim to pursuit the simple pattern corresponding to the minima of entropy function.

Third, a reliable global optimizer such as simulated annealing or genetic algorithm is employed to achieve the objective function minima. The sequential evaluations of these minima correspond to the pure pattern present in the observed set of 3D mixtures. All of the following simulations and image reconstructions were conducted using in-house code developed in MATLAB.

8.4. Numerical Simulations

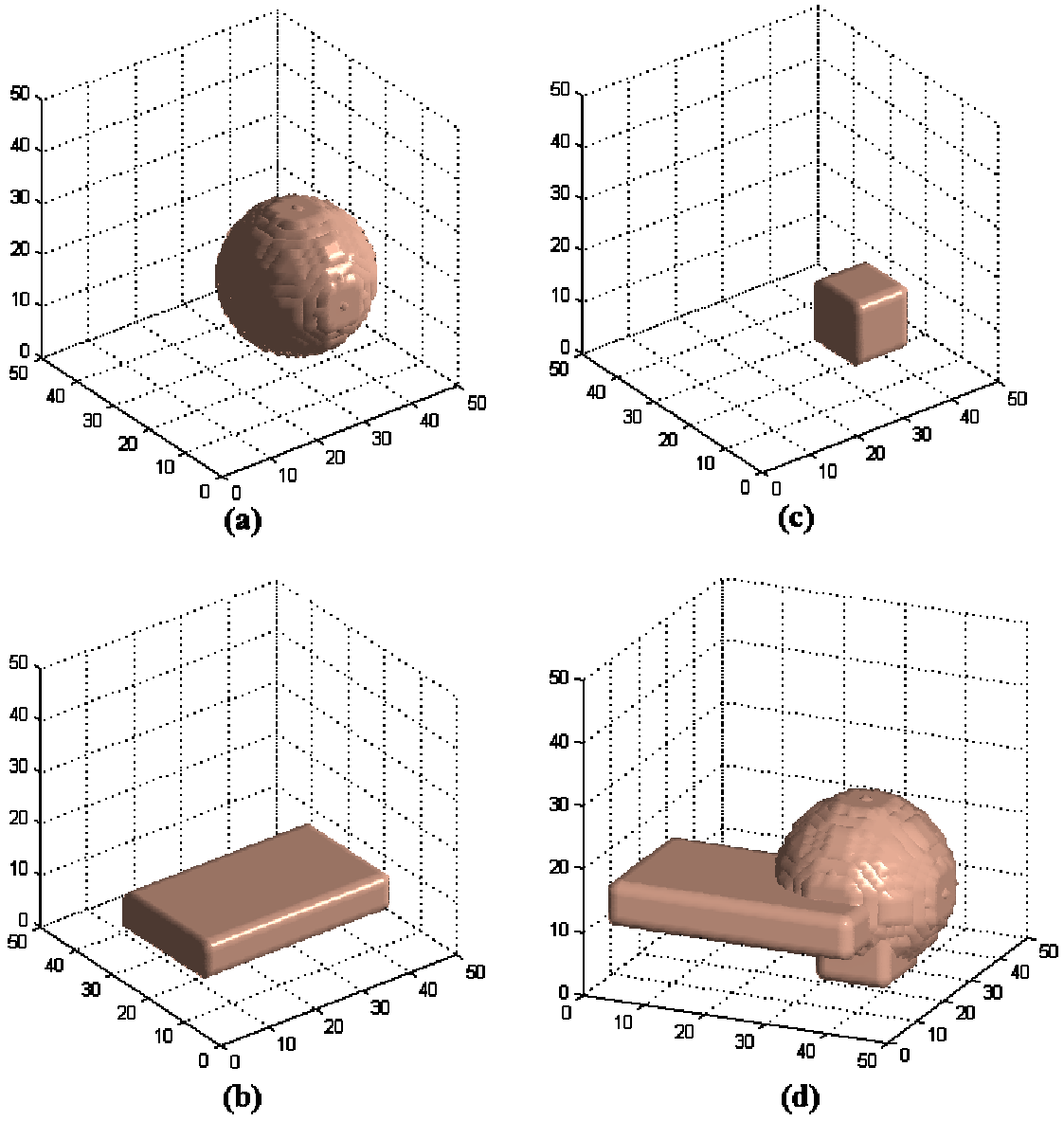


Figure 8.2. The three simulated objects, namely, a ball (a), a rectangle (b) and a cube(c) are shown. At the right corner, an example of the superimposition of these three objects is shown as (d).

In this section, two sets of 3D simulations are formulated in order to test the 3D Entropy Minimization mixture separation algorithm.

Simulation 1:

Three simulated objectives composed of one ball, one rectangle and one cube were used (shown in Figure 8.2). In a $50\times50\times50$ space, the elements inside specific region defined by each objective were filling with unity value data and leaving other elements' value with zeros. Mixing procedure is carried out with an arbitrary mixing matrix (5by3) which gives the different weighing factors. The superimposition of three different tensors with various contributions creates a mixture tensor. Five mixture tensors were fabricated in this simulation and one is shown in Figure 8.2d.

Simulation 2:

In the second simulation, in a $50\times50\times50$ space, the elements inside specific region defined by each objective were filling with non-negative random value data(simulated with MATLAB function "rand") and leaving other elements' value with zeros. Again, five mixture tensors were fabricated in this simulation.

8.5. Result

8.5.1. Simulation 1

Five 3D mixture tensors were systematically arranged into a 2D array, and the 2D matrix was decomposed by SVD. It is known that the middle diagonal matrix Σ contains the *singular values* of data and the square of its diagonal elements represent the amount of information corresponding to each principal component. In simulation 1, the five singular values were 29933, 8064.2, 951.25, 2.2759e-012 and 1.2963e-012 individually. In other words, the first 3 right singular arrays already explain all the variance (except from the noise) in the 5 sets of data, and by the way, it proves that only three components were

embedded inside mixtures in this ideal linear mixing system. The left 2 right singular arrays were nothing but the noise originated from computational errors (which are validated by their extremely small singular values). Therefore it is reasonable only the first three right singular arrays were used.

Due to the aforementioned reasons, the resulting right singular arrays can not be shown in normal way. Additional colour index is needed to visualize the data. In Figure 8.3, the three physically meaningful right singular arrays are shown in iso-surface plot.

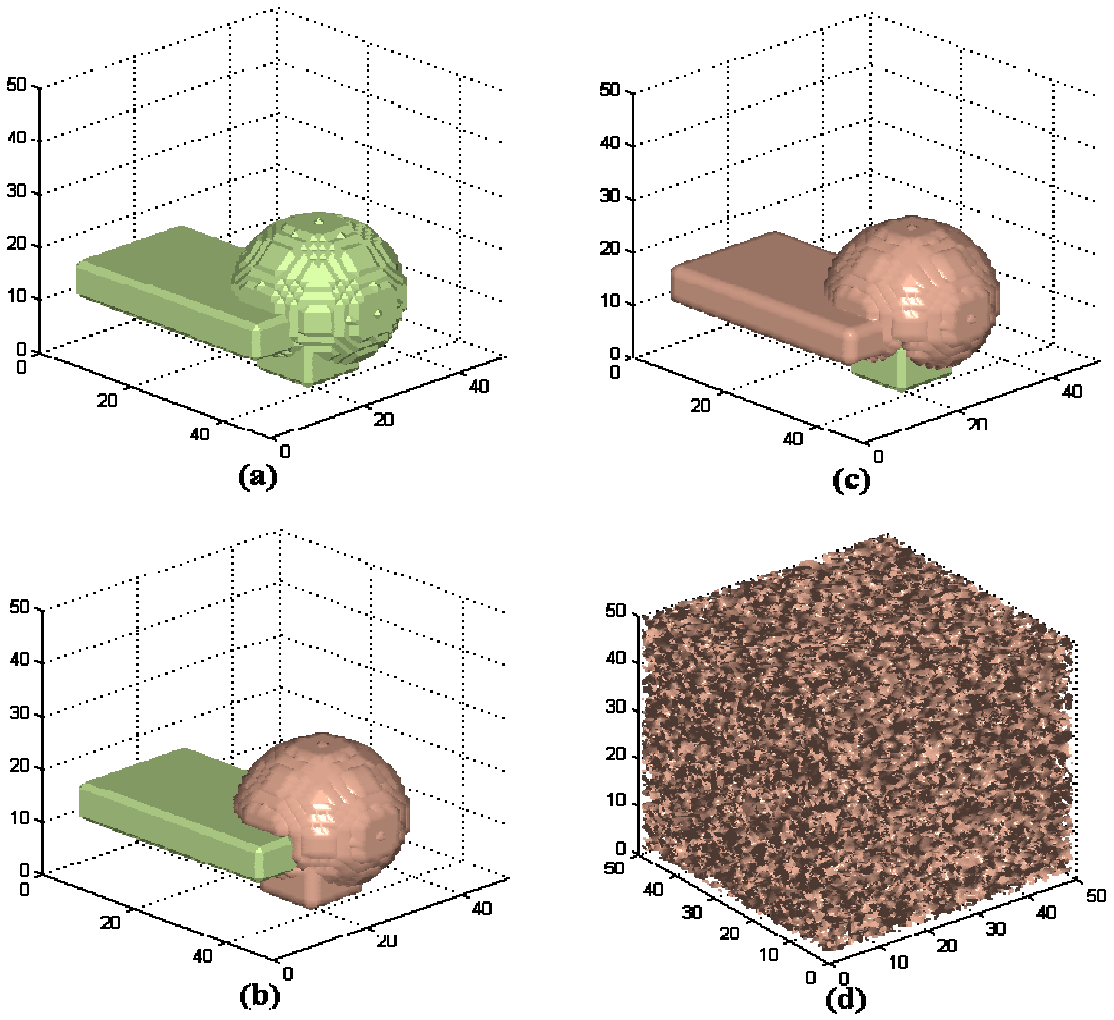


Figure 8.3. The first four resulting right singular arrays, 1st (a) 2nd (b) 3rd (c) and 4th (d). The greenish part suggests the elements in that region are negative meanwhile the elements are positive in the brownish region.

For simplicity, only two colours are used here, the greenish part suggests the elements in that region are negative meanwhile the elements are positive in the brownish region. The first right singular array shows all the values are negative. In the second right singular array, the rectangle object is negative and other parts are positive. Meanwhile, the cubic object is negative in the third right singular array. It is reasonable to conclude that these three objects are independent and possess different variations.

It is shown in Figure 8.3 that the fourth right singular array is filled with homoscedastic noise. In order to investigate the data structure of the fourth right singular array, histogram plot with the elements in the fourth right singular array was created.

The histogram graphically summarizes the distribution of the data. First, with unfolding, the 3D data set was transformed into a one dimension data. All elements in the data set were grouped according to their numeric values. The histogram's x-axis reflects the range of values and its y-axis shows the counts of elements that fall within the groups.

In Figure 8.4, the histogram of the fourth right singular array (a) and fifth right singular array (b) are shown. The histograms show that all these arrays were filled with computation errors which have random distribution values range from -5e-3 to 5e-3. Therefore, it proves that the aforementioned statement that only the first three right singular arrays contain information and they will be used in the sequential reconstruction procedure.

As discussed, the visual inspections imply that, three patterns are independent. It is natural to perform 2D-Entropy Minimization using three right singular arrays by targeting the small region inside each observable feature once at one time. Exhaustive searches produced only three 3D patterns which are similar to the original three objects in Figure

8.2. They were obtained by individually targeting region 1([19 to 21; 28 to 30; 8 to 10]), region 2([7 to 8; 2 to 5; 11 to 13]) and region 3([23 to 25; 34 to 35; 24 to 26]).

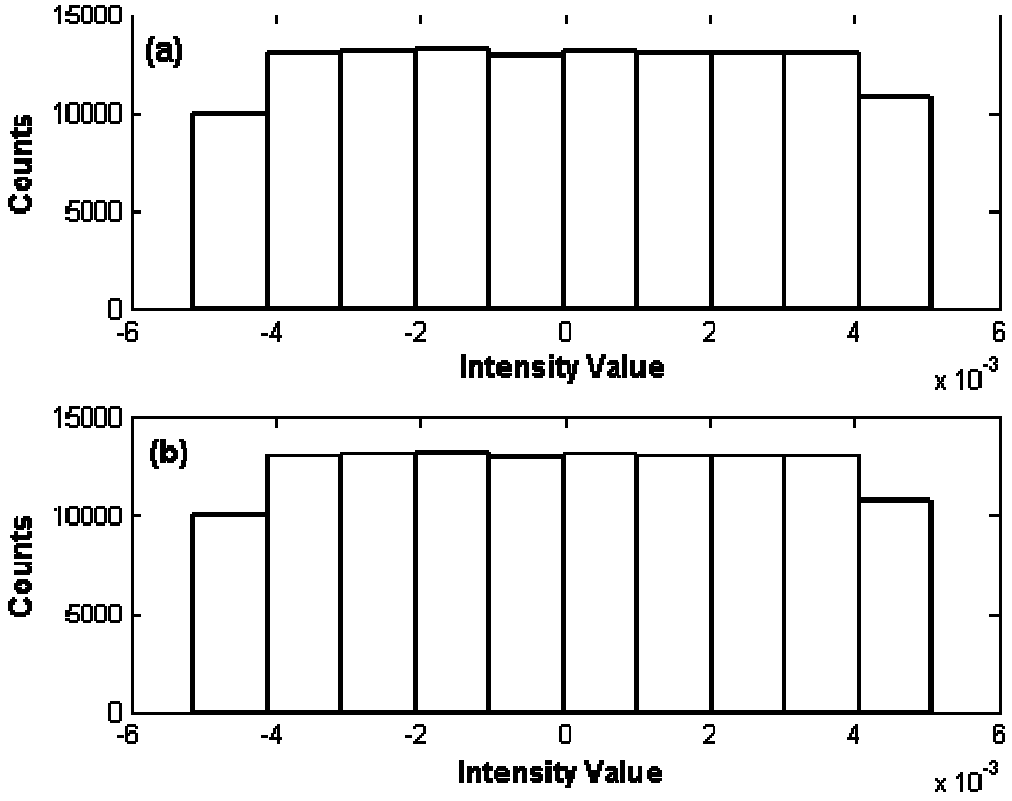


Figure 8.4. The histogram of the fourth right singular array (a) and fifth right singular array (b).

Two types of objective functions were implemented. In the first objective function, only the term concerning summation of the intensity (equivalently the volume term for 2D pattern) was used, but in second objective function, the derivate cost was added. Reconstruction results are similar.

In order to check the quality of the reconstructed patterns, a least-square procedure was performed to project the recovered 3D tensors back onto the mixture 3D data and the resulting mixing matrix was obtained. It is clear that the entries of the original mixing matrix (Eq. 8.1) and this new mixing matrix (Eq 8.2) differ very slightly.

$$A_{original} = \begin{pmatrix} 1.0 & 32.0 & 3.0 \\ 33.0 & 3.0 & 15. \\ 133.0 & 93.0 & 105. \\ 53.0 & 193. & 30.5 \\ 153.0 & 120. & 75. \end{pmatrix} \quad (8.1)$$

$$A_{calculated} = \begin{pmatrix} 1.0000 & 32.0000 & 3.0002 \\ 33.0000 & 3.0000 & 15.0000 \\ 133.0000 & 93.0000 & 105.0006 \\ 53.0000 & 193.0000 & 30.5012 \\ 153.0000 & 120.0000 & 75.0008 \end{pmatrix} \quad (8.2)$$

8.5.2. Simulation 2

In simulation 2, similar steps were taken as depicted in section 8.5.1. After SVD process, the singular value shown in descending order were 16815, 4796.7, 582.52, 1.8774e-012 and 3.6866e-013 individually. In other wards, the first 3 right singular arrays already can explain all the variance (except from the noise) in the 5 sets of data, and by the way, it proves that only three components are embedded inside the mixture (not shown here). Therefore it is reasonable only the first three right singular arrays should be used. By individually targeting regions inside the three objects, the pure 3D patterns were obtained. Exhaustive searches proved only three 3D patterns exist (not shown here). They are all highly consistent with the original pure 3D spectra.

In the same spirits as simulation 1, the normalized unmixing matrix based on the recovered objectives from Simulation 2 were obtain via least-square method (shown in Eq. 8.10) And the normalized original mixing matrix is shown Eq. 8.11. Except for the column wise order is different; they are identical with present display.

$$A_{calculated} = \begin{pmatrix} 0.0286 & 0.0065 & 0.1658 \\ 0.1429 & 0.2157 & 0.0155 \\ 1.0000 & 0.8693 & 0.4819 \\ 0.2905 & 0.3464 & 1.0000 \\ 0.7143 & 1.0000 & 0.6218 \end{pmatrix} \quad (8.10)$$

$$A_{mixing} = \begin{pmatrix} 0.0286 & 0.1658 & 0.0065 \\ 0.1429 & 0.0155 & 0.2157 \\ 1.0000 & 0.4819 & 0.8693 \\ 0.2905 & 1.0000 & 0.3464 \\ 0.7143 & 0.6218 & 1.0000 \end{pmatrix} \quad (8.11)$$

In the study, both simulations showed that the extension of entropy minimization algorithm to its three-dimensional mode was very successful. The numerical results were good and very accurate pure component spectral reconstructions were obtained.

8.6. Summary

In this section, the 3D entropy minimization algorithm was initiated. A methodology was suggested and successfully tested on a simulated 3D mixture data. The purpose of developing a high dimensional data deconvolution algorithm is the exploration of large, complex, multi-dimensional scientific data, especially the emerging NMR data. By applying and extending ideas from entropy minimization and pattern recognition, we are developing a new generation of computational tools and techniques that are being used to improve the way in which scientists extract useful information from their data, especially the high dimensional NMR mixture spectra.

Chapter 9

Conclusions and Future Work

9.1. Conclusions

In the present studies, Minimum-Entropy based pure component spectral reconstruction method has been successfully applied to real data, such as, NMR spectroscopic data, powder X-ray diffraction data where very accurate pure component spectral reconstructions are obtained from mixture data. More importantly, 2D Band-Target Entropy Minimization algorithm and associated techniques were introduced, these methodology development was validated using synthetic spectroscopic data, and were further modified and tailored and applied to various real 2D experimental data including 2D NMR data and 2D fluorescence data.

Initial work began with the extension of entropy minimization algorithm to 1D NMR spectra which possess unique characteristic in comparison with other spectra, such as FTIR, Raman spectroscopy. Also four sets of data from different types of homogeneous catalytic hydroformylation were investigated. After some modification, BTEM was successfully applied for the first time, to sets of acoustic data.

The second focus was the extension of entropy minimization algorithm to its multidimensional mode, and eventually applied to the identification of multidimensional data. 2D-BTEM has been successfully implemented to real experimental cases as it is shown in chapter 6 and 7, which including the simulated, non-reactive and reactive mixture systems. There are two clear benefits of this approach. First, it will certainly extend the scope of problems that can be treated in the chemical sciences. In other words,

the extension brings enormous opportunities of applying the entropy minimization method to various types of data set in chemical engineering fields including various matrix-formatted data and tensor-formatted data. The second major benefit is that we can now more confident to apply the multidimensional spectroscopy in the system identification of complex chemical reaction via on-line monitoring and *in-situ* measurements.

9.2. Future Work

Firstly, the simulated 3D pattern recognition in a superimposed system has been explored in current studies. However it is very necessary to verify the methodology with real data, including various kinds of 3D NMR data, either non-reactive or reactive mixture data. Normally the collection of 3D NMR data is hindered by the long duration of the measurement, often days. Very recently a 10-100 times advance in data acquisition time was made by Kupce and Freeman (2004); this just might allow us to make many 3D measurements in the future.

In the long term, the future work will focus on the system identification of multidimensional NMR data (since NMR data is the most directly available and evident multidimensional type data). It is believed that there is so much more we can do for the real experimental data, even though certainly many efforts should be invested in improving the algorithm before its application to the real 3D data set. This domain is stimulating and challenging.

Also using chemometric techniques to explore the large-scale spectroscopic data is still a field very much in its infancy. In order to extend chemometric technique to large-scale data, several barriers must be overcome. The large data manipulation including data storage and management is a critical issue that must be addressed first, prior to the

application of the chemometric algorithms. Also the visualization of the large scale data, the computing resource problem and the complex optimization related to the large data are all challenging.

Secondly, even though, in chapter 4 and chapter 7, 1D and 2D NMR data are treated by a various pre-treatment procedure and the results are acceptable. But in the real data, the reconstruction algorithm is not sufficient to, the non-linearites generated from severe spectral band moving and changing band shape. Also due to the serious non-linearities, the recovery of minor components having weak signals is sometime quite difficult. Further developments which focus on effectively solving the problem of spectral non-linearities and retrieving weak signal in NMR spectra are needed. Much effort will have to be invested in developing routine and automated programs for correcting the on-line NMR spectroscopic data. These pre-treated methods are critical to the subsequent performance of chemometric methods including BTEM analysis.

REFERENCES

Alam, T.M. and M.K. Alam. Chemometric Analysis of NMR Spectroscopy Data: A Review, Annu. Rep. NMR. Spectro., *54*, pp.41-80. 2005.

Albert, K., M. Krucker, T. Glaser, A. Schefer, A. Lienau and D. Zeeb. Hyphenated techniques, Analytical and Bioanalytical Chemistry, *372* (1), pp.25-26. 2002.

Alsberg, B.K., A.M. Woodward and D.B. Kell. An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach, Chemom. Intell. Lab. Syst., *37* (2), pp. 215-239.1997.

Alter, O., P.O. Brown and D. Botstein. Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling, Proc. Natl. Acad. Sci. U.S.A., *97,* pp.10101-10106. 2000.

Andersson, C.A. and R. Bro. The N-way Toolbox for MATLAB, Chemom. Intell. Lab. Syst., *52* (1), pp.1-4. 2000.

Antalek, B. and W. Windig. Generalized Rank Annihilation Method Applied to a Single Multicomponent Pulsed Gradient Spin Echo NMR Data Set. J. Am. Chem. Soc., *118*, pp. 10331-10332. 1996.

Antalek, B. Using Pulsed Gradient Spin Echo NMR for Chemical Mixture Analysis: How to obtain Optimum Results, Concepts Magn. Reson., *14*, pp.225-258. 2002.

Armstrong, G.S., K.E. Cano, V.A. Mandelshtam, A.J. Shaka and B. Bendiak. Rapid 3D NMR using the Filter Diagonalization Method: Application to Oligosaccharides Derivatized with C-13-Labeled Acetyl Groups, J. Magn. Reson., *170* (1), pp.156-163. 2004.

Aue, W.P., E. Bartholdi, and R.R. Ernst. Two-Dimensionalspectroscopy. Application to Nuclear Magnetic Resonance, J. Chem. Phys., *64*(5), pp.2229-2246. 1976.

Banerjee, S. and D.Y. Li. Interpreting Multicomponent Infrared Spectra by Derivative Minimization. Appl. Spectrosc., *45*, pp. 1047-1049. 1991.

Bax, A., R. Freeman and S.P. Kempsell. Natural Abundance ^{13}C–^{13}C Coupling Observed via Double Quantum Coherence, J. Am. Chem. Soc., *102*, pp.4849-4851. 1980.

Bax, A., R. Freeman and T.A. Frenkiel. An NMR Technique for Tracing out the Carbon Skeleton of an Organic Molecule, J. Am. Chem. Soc., *103*, pp.2102-2104. 1981.

Bax, A. , R.H. Griffey and B.L. Hawkins. Sensitivity-Enhanced Correlation of ^{15}N and ^{1}H Chemical Shifts in Natural Abundance Samples via Multiquantum Coherence, J. Am. Chem. Soc., *105*, pp.7188-7190. 1983.

Bax, A. and M.F. Summers. ^{1}H and ^{13}C Assignments from Sensitivity-Enhanced Detection of Heteronuclear Multi-Bond Connectivity by 1D Multiquantum NMR, J. Am. Chem. Soc., *108*, pp.2093-2094. 1986.

Bax, A., G.M. Clore and A.M. Gronenborn. Proton–Proton Correlation via Isotropic Mixing of Carbon-13 Magnetization, a new Three-Dimensional Approach for Assigning Proton and Carbon-13 Spectra of Carbon-13-Enrichedproteins, J. Magn. Reson., *88*, pp.425–431. 1990.

Bell, A. and T. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution, Neural Comput., *7*(6), pp.1004–1034. 1995.

Berry, M.W. Large-Scale Sparse Singular Value Computations, Int. J. Supercomp. Appl., *6*, pp.13-49. 1992.

Berry, M.W., T. Do, G.W. Obrien, V. Krishna, and S. Varadhan. SVDPACKC: Version 1.0 User's Guide. Knoxville: University of Tennessee. 1993.

Bijlsma, S., D.J. Louwerse, and A.K. Smilde. Rapid Estimation of Rate Constants of Batch Processes Using On-Line SW-NIR. AICHE J. *44*, pp. 2713-2723. 1998.

Bijlsma, S. and A.K. Smilde. Application of Curve Resolution Based Methods to Kinetic Data, Anal. Chim. Acta, *396*, pp. 231-240. 1999.

Bijlsma, S. Estimating Rate Constants of Chemical Reactions Using Spectroscopy. Ph.D. Thesis, University of Amsterdam. 2000.

Blackburn, J.A. Computer Program for Multicomponent Spectrum Analysis Using Least-Squares Method, Anal.Chem., *37*(8), pp. 1000-10003. 1965.

Blanco, M. and D. Valdes. Influence of Temperature on the Predictive Ability of Near Infrared Spectroscopy Models, J. Near Infrared Spectrosc., *12* (2), pp.121-126. 2004.

Bodenhausen, G. and D. Ruben. Natural Abundance ^{15}N NMR by Enhanced Heteronuclear Spectroscopy, Chem. Phys. Lett., *69*, pp.185-188. 1980.

Booksh, K.S., A.R. Muroski and M.L. Myrick. Single Measurement Excitation/Emission Matrix Spectrofluorometer for Determination of Hydrocarbons in Ocean Water .2. Calibration and Quantitation of Naphthalene and Styrene, Anal. Chem., *68* (20), pp.3539-3544. 1996.

Borgen, O.S. and B.R. Kowalski. An Extension of the Multivariate Component-Resolution Method to Three Components, Anal.Chim. Acta, *174*, pp. 1-26. 1985.

Borgen, O.S., N. Davidsen, M.Y. Zhu and O. Oeyen. The Multivariate N-Component Resolution Problem with Minimum Assumptions, Mikrochim. Acta, *2*, pp. 63-73. 1987.

Bothner-By, A. A., R.L. Stephens, J.M. Lee, C.D. Warren and R.W. Jeanloz. Structure Determination of a Tetrasaccharide: Transient Nuclear Overhauser Effects in the Rotating Frame, J. Am. Chem. Soc., *106*, pp.811-813. 1984.

Braunschweiler, L. and R.R. Ernst. Coherence Transfer by Isotropic Mixing: Application to Proton Correlation Spectroscopy, J. Magn. Reson., *53*, pp.521-528. 1983.

Brekke, T. and O.M. Kvalheim. Analysis of Nuclear Magnetic Resonance Spectra of Mixtures Using Multivariate Techniques. In Signal Treatment and Signal Analysis in NMR, Vol. 20, ed. by D.N. Rutledge, pp. 422-451. New York : Elsevier Science. 1996.

Brereton, R.G. Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems. New York: Ellis Horwood Ltd., 1990.

Brixner, T., J. Stenger, H.M. Vaswani, M. Cho, R.E. Blankenship and G.R. Fleming. Two-Dimensional Spectroscopy of Electronic Couplings in Photosynthesis, Nature, *434* (7033), pp. 625-628. 2005.

Bro, R., J.J. Workman, P.R. Mobley and B.R. Kowalski. Review of Chemometrics Applied to Spectroscopy: 1985-95. 3. Multi-Way Analysis, Appl. Spectrosc. Rev., *32* (3), pp.237-261. 1997.

Bro, R. PARAFAC: Tutorial and Applications, Chemom. Intell. Lab. Syst., *38*, pp.149-171. 1997.

Bro, R. Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications. Ph.D. Thesis, University of Amsterdam (NL) and Royal Veterinary and Agricultural University (DK). 1998.

Bro, R. and H.A.L. Kiers. A New Efficient Method for Determining the Number of Components in PARAFAC Models, J. Chemom., *17*(5), pp.274-286. 2003.

Bronkhorst, A.W. The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions , ACUSTICA, *86* (1), pp.117-128. 2000.

Brooks, S.P. and B.J.T. Morgan. Optimization Using Simulated Annealing, statistician, *44*, pp. 241-257.1995.

Brown, S.D., T. Barke and R. Larivee. Chemometrics, Anal. Chem., *60*, pp. 252R-273R. 1988.

Brown, S.D. Chemometrics, Anal. Chem., *62*, pp. 84R-101R. 1990.

Brown, S.D., R.S. Bear and T.B. Blank. Chemometrics, Anal. Chem. ,*64*, pp. 22R-49R. 1992.

Brown, S.D., T.B. Blank, S.T. Sum and Lois G. Weyer. Chemometrics, Anal. Chem. *66*, pp. 315R-359R. 1994.

Brown, S.D., S.T. Sum, F. Despagne and B.K. Lavine. Chemometrics. Anal. Chem. *68*, pp. 21R-61R. 1996.

Bu, D.S. and C.W. Brown. Self-Modeling Mixture Analysis by Interactive Principal Component Analysis, Appl. Spectrosc. *54*, pp. 1214- 1221. 2000.

Burg, J.P. Annual Meeting International Society Exploratory Geophysics, In Modern Spectral Analysis, ed by D.G. Childers, pp. 34–41. New York: IEEE Press. 1978.

Bylund,Dan., R. Danielsson, G. Malmquiest and K.E. Markides. Chromatographic Alignment by Warping and Dynamic Programming as A Pre-Processing Tool for PARAFAC Modelling of Liquid Chromatography-Mass Spectrometry Data, J. Chromatogr. A, *961*, pp. 237-244. 2002.

Cadzow, J. Signal Enhancement, A Composite Property Mapping Algorithm, IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-*36*, pp. 49-66. 1988.

Carazza, B. On the Lorentzian Shape and the Information Provided by an Experimental Plot, J. Phys. A: Math. Gen., 9, pp.1069-1072.1976.

Cardoso, J.F. and A. Souloumiac. Blind Beamforming for Non-Gaussian Signals, IEE. Proc-F, *140*(6), pp.362–370. 1993.

Cardoso, J.F. Informax and Maximum Likelihood in Blind Source Separation, IEEE Signal Processing Letters, *4*(4), pp. 112-114. 1997.

Cardoso, J.F. Blind Signal Separation: Statistical Principles, Proc. of the IEEE, *86* (10), pp.2009-2025.1998.

Carroll, J.D. and J. Chang. Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of 'Eckart-Young' Decomposition. Psychometrika, *35*, pp. 283-319.1970.

Cattell, R.B. Parallel Proportional Profiles and Other Principles for Determining the Choice of Factors by Rotation, Psychometrika, *9*, pp.267–283. 1944.

Cattell, R.B. The Scree Test for the Number of Factors, Multivar. Behav. Res., *1*, pp.245-276. 1966.

Chen, J. and X.Z. Wang. A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis , J. Chem. Inf. Comp. Sci., *41*(4), pp. 992-1001. 2001.

Chen, L. and M. Garland. Computationally Efficient Curve-Fitting Procedure for Large Two-Dimensional Experimental Infrared Spectroscopic Arrays Using the Pearson VII Model, Appl. Spectrosc., *57* (3), pp.331-337. 2003.

Chen, L., W. Chew and M. Garland. Spectral Pattern Recognition of *in Situ* FT-IR Spectroscopic Reaction Data Using Minimization of Entropy and Spectral Similarity (MESS): Application to the Homogeneous Rhodium Catalyzed Hydroformylation of Isoprene. Appl. Spectrosc., *57*, pp. 491-498. 2003.

Chen, Z.P., Y.Z. Liang, J.H. Jiang, Y. Li, J.Y. Qian and R.Q. Yu. Determination of the Number of Components in Mixtures Using a New Approach Incorporating Chemical Information, J. Chemom., *13* (1),pp. 15-30.1999.

Chen, Z.P., Z. Liu, Y.Z. Cao and R.Q. Yu. Efficient Way to Estimate the Optimum Number of Factors for Trilinear Decomposition, Anal.Chim. Acta, *444* (2), pp. 295-307. 2001.

Chew, W., E. Widjaja, and M. Garland. Band-Target Entropy Minimization (BTEM): An Advanced Method for Recovering Unknown Pure Component Spectra. Application to the FTIR Spectra of Unstable Organometallic Mixtures. Organom., *21*(9), pp.1982-1990. 2002.

Chew, W. Exploratory Chemometric Studies of Unmodified Rhodium Catalyzed Conjugated Diene Hydroformylations via *in situ* FTIR Spectroscopy. Ph.D Thesis, National University of Singapore. 2003.

Christensen[a], J., E.M. Becker and C.S. Frederiksen. Fluorescence Spectroscopy and PARAFAC in the Analysis of Yogurt, Chemom. Intell. Lab. Syst., *75* (2), pp.201-208. 2005.

Christensen[b], J.H., A.B. Hansen, J. Mortensen and O. Andersen. Characterization and Matching of Oil Samples Using Fluorescence Spectroscopy and Parallel Factor Analysis, Anal. Chem., *77* (7), pp.2210-2217. 2005.

Cobb, J.B.C., A. Bennett, G.C. Chinchen, L. Davies, B.T. Heaton and J.A. Iggo. The Characterisation of Distinct Adsorption Sites for Hydrogen on Copper in Copper/Alumina Catalysts by *In Situ* ^{1}H NMR Spectroscopy, J. Catal., *164*, pp.268-275. 1996.

Conway, A.R.A., N. Cowan and M.F. Bunting. The Cocktail Party Phenomenon Revisited: The Importance of Working Memory Capacity, Psychonomic Bulletin & Review, *8* (2), pp.331-335. 2001.

Corana, A., M. Marchesi, C. Martini, and S. Ridella. Minimizing Multimodal Functions of Continuous Variables with the "Simulated Annealing" Algorithm, ACM Trans. Math. Softw. 13, pp. 262-280. 1987.

Cornils, B. and W.A. Herrmann. Applied Homogeneous Catalysis with Organometallic Compounds: a Comprehensive Handbook, Vol.1. Weinheim: Wiley-VCH. 1996.

Cornwell, T.J. and K.F. Evans. A Simple Maximum Entropy Deconvolution Algorithm, Astron. Astrophys., *143*, pp.77-83. 1985.

Craddock, J.M. and C.R. Flood. Eigenvectors for Representing the 500 mb Geopotential Surface over the Northern Hemisphere, Quart. J. Roy. Meteor. Soc., *95*, 576–593. 1969.

Croux, C., and G. Haesbroeck. Principal Components Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies, Biometrika, *87*, pp.603-618. 2000.

Csihony, S., A. Bodor, J. Rohonczy, L.T. Horvath. *In situ* IR and NMR Spectroscopic Investigation of the Formation and Structure of Protonated Diacetylketene Tetrachloroaluminate, J. Chem. Soc., Perkin Transactions 1, pp.2861-2865. 2002.

Da Silva, J.C.G.E., J.M.M. Leitao, F.S. Costa and J.L.A. Ribeiro. Detection of Verapamil Drug by Fluorescence and Trilinear Decompositim Techniques, Anal. Chim. Acta, *453* (1), pp.105-115. 2002.

Davis, J.E., A. Shepard, N. Stanford and L.B. Rogers. Principal-Component Analysis Applied to Combined Gas Chromatographic-Mass Spectrometric Data, Anal. Chem., *46*, pp. 821-825. 1974.

De Juan, A. and R. Tauler. Comparison of Three-Way Resolution Methods for Non-Trilinear Chemical Data Sets, J. Chemom., *15*, pp. 749-772. 2001.

De Juan, A. and R. Tauler. Chemometrics Applied to Unravel Multicomponent Processes and Mixtures: Revisiting Latest Trends in Multivariate Resolution, Anal. Chim. Acta, *500*(1-2), pp.195-210. 2003.

Delwiche, S.R., K.H. Norris and R.E. Pitt. Temperature Sensitivity of Near-Infrared Scattering Transmittance Spectra of Water-Adsorbed Starch and Cellulose, Appl. Spectrosc., *46* (5), pp.782-789. 1992.

Deprettere, F. SVD and Signal Processing: Algorithms, Analysis and Applications. Amsterdam: Elsevier Science Publishers. 1988.

Dijksterhuis, G.b. and W.J. Heiser. The Role of Permutation Tests Inexploratory Multivariate Data Analysis, Food Qual. Pref., *6*, pp.263-270. 1995.

Duddeck, H. and W. Dietrich. Structure Elucidation by Modern NMR: A Workbook. New York: Springer-Verlag. 1992.

Dyrby, M., D. Baunsgaard, R. Bro and S.B. Engelsen. Multiway Chemometric Analysis of the Metabolic Response to Toxins Monitored by NMR, Chemom. Intell. Lab. Syst., *76*(1) , pp.79-89. 2005.

Eastment, H.T. and W.J. Krzanowski. Cross-Validatory Choice of the Number of Components from a Pricnipal Component Analysis, Technometrics, *24*, pp.73-77. 1982.

Einax, J.W. Chemometrics in Analytical Chemistry, Anal. Bioanal. Chem., *380*(3), pp.368-369. 2004.

Ernst, R.R., G. Bodenhausen and A. Wokaun. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford: Oxford University Press. 1987.

Esteves da Silva, J.C.G., J.M.M. Leitão, F.S. Costa and J.L.A. Ribeiro. Detection of Verapamil Drug by Fluorescence and Trilinear Decomposition Techniques, Anal. Chim. Acta, *453*, pp.105-115, 2002.

Faber, K. and B.R. Kowalski. Critical Evaluation of two F-Tests for Selecting the Number of Factors in Abstract Factor Analysis, Anal. Chim. Acta, *337*, pp.57-71. 1997.

Faber, N.M., J. Ferre and R. Boque. Iteratively Reweighted Generalized Rank Annihilation Method 1. Improved Handling of Prediction Bias, Chemom. Intell. Lab. Syst., *55* (1-2), pp. 67-90, 2001a.

Faber, N.M., R. Boque and J. Ferre. Iteratively Reweighted Generalized Rank Annihilation Method 2. Least Squares Property and Variance Expressions Chemom. Intell. Lab. Syst., *55* (1-2), pp. 91-100, 2001b.

Farmer, S.A. An Investigation into the Results of Principal Component Analysis of Data Derived from Random Numbers, Statistician, *20*, pp.63-72. 1971.

Farrar, T.C. An Introduction to Pulse NMR Spectroscopy. Chicago: Farragut Press. 1987.

Fauth, J.M., S. Kababya and D. Goldfarb. Application of 2D-FT-EPR Spectroscopy to Study Slow Intramolecular Chemical-Exchange, J. Magn. Reson., *92*(1), pp.203-207. 1991.

Feng, J.H. and M. Garland. Unmodified Homogeneous Rhodium-Catalyzed Hydroformylation of Styrene. The Detailed Kinetics of the Regioselective Synthesis, Organomet., *18*, pp. 417-427. 1999.

Forland, G.M., F.O. Libnau, O.M. Kvalheim and H. Hoiland. Self-Association of Medium-Chain Alcohols in *n*-Decane Solutions, Appl. Spectrosc., *50*, pp. 1264-1272. 1996.

Forshed, J., I. Schuppe-Koistinen and S. P. Jacobsson. Peak Alignment of NMR Signals by Means of a Genetic Algorithm, Anal. Chim. Acta, *487*, pp.189–199. 2003.

Frenich, A.G., M.M. Galera, J.L.M. Vidal, D.L. Massart, J.R. Torres-Lapasio, K. De Braekeleer, J.H. Wang and P.K. Hopke. Resolution of Multicomponent Peaks by Orthogonal Projection Approach, Positive Matrix Factorization and Alternating Least Squares, Anal. Chim. Acta, *411* (1-2), pp.145-155. 2000.

Frenich, A.G., D.P. Zamora and M.M. Galera. Vidal JLM Application of GRAM and TLD to the Resolution and Quantitation of Real Complex Multicomponent Mixtures by Fluorescence Spectroscopy, Anal. Bioanal. Chem., *375* (7), pp.974-980. 2003.

Frieden, B. Image Enhancement and Restoration, in Picture Processing and Digital Filtering. Topics in Applied Physics, Vol. 6, ed R. by T. S. Huang, pp. 177-248. New York: Springer-Verlag, Berlin Heidelberg. 1975.

Gaetz, M., H. Weinberg, E. Rzempoluck and K.J. Jantzen. Neural Network Classifications and Correlation Analysis of EEG and MEG Activity Accompanying Spontaneous Reversals of the Necker Cube, Cognitive Brain Research, *6* (4), pp.335-346. 1998.

Garland, M. and G. Bor. Infrared Spectroscopic Studies on Metal Carbonyl Compounds. 24. Observation of the Infrared Spectrum of an Acylrhodium Tetracarbonyl during the Hydroformylation of Olefins with Rhodium-Containing Catalyst Precursors. Inorg. Chem., *28*, pp.410-413. 1989.

Garland, M. and P. Pino. Kinetics of the Formation and Hydrogenolysis of Acylrhodium Tetracarbonyl, Organom., *10*, pp. 1693-1704. 1991.

Garland, M. Horvath, I.T., G. Bor and P. Pino. Thermodynamic Parameters for the Formation of Cobalt-Rhodiumm Heptacarbonyl and Cobalt-Rhodium Octacarbonyl, Organom., *10*, pp.559-567. 1991.

Garland, M., E. Visser, P. Terwiesch, D.W.T. Rippin. On the Number of Observable Species, Observable Reactions and Observable Fluxes in Chemometric Studies and the Role of Multichannel Integration, Anal. Chim. Acta, *351*, pp.337-358. 1997.

Ge, N.H. and R.M. Hochstrasser. Femtosecond Two-Dimensional Infrared Spectroscopy: IR-COSY and THIRSTY, Phys. Chem. Comm., *3*, pp.17-26. 2002.

Geladi, P., H. Isaksson, L. Lindqvist, S. Wold and K. Esbensen. Principal Component Analysis of Multivariate Images, Chem. Int. lab. Sys., *5*, pp. 209-220. 1989.

Geladi, P. and H. Grahn. Multivariate Image Analysis, pp.38, Table 2.3. New York: Wiley, Chichester.1996**.**

Gemperline, P.J. A Prior Estimate of the Elution Profiles of the Pure Component in Overlapped Liquid Chromatography Peaks Using Target Factor Analysis, J. Chem. Inf. Comput. Sci., *24*, pp. 206-212. 1984.

Gemperline, P.J. Target Transformation Factor Analysis with Linear Inequality Constraints Applied to Spectroscopic-Chromatographic Data, Anal. Chem., *58*, pp. 2656-2663. 1986.

Gemperline, P.J. Mixture Analysis using Factor Analysis I: Calibration and Quantitation, J. Chemom., *3*, pp. 549-568. 1989.

Gentle, J.E. Singular Value Factorization. In Numerical Linear Algebra for Applications in Statistics. pp. 102-103. Berlin: Springer-Verlag. 1998.

Gilmore, C.J, G. Bricogne and C. Bannister. A Multisolution Method of Phase Determination by Combined Maximization of Entropy and Likelihood. 2: Application to Small Molecules, Acta Cryst. A, *46,* pp. 297–308. 1990.

Gilmore, C.J., K. Shankland and G. Bricogne. Applications of the Maximum Entropy Method to Powder Diffraction and Electron Crystallography, Proc. Roy. Soc., *442* (A), pp.97-111. 1993.

Goldberg, D.E., B. Korb and K. Deb. Messy Genetic Algorithms: Motivation, Analysis, and First Results, Complex Systems, *3*(5), pp. 493—530. 1989.

Goldberg, D.E. Genetic Algorithms in Search, Optimization, and Machine Learning. MA: Addison-Wesley. 1989.

Golub, G. and W. Kahan. Calculating the Singular Values and Pseudo-Inverse of a Matrix, SIAM J. Numer. Anal. B, *2,* pp.205-221.1965.

Golub, G. H. and C. Reinsch. Singular Value Decomposition and Least Squares Solutions, Num. Math., *14,* pp.403-420. 1970.

Golub, G. H. and C. F. Van Loan. The Singular Value Decomposition and Unitary Matrices. In Matrix Computations, pp. 70-73. MD: Johns Hopkins University Press, Baltimore. 1996.

Gull, S. F., and J. Skilling. Maximum-Entropy Method in Image-Processing, IEE Proc., *131*(F), pp.646-659. 1984.

Gunther, H. Modern Pulse Methods in High-Resolution NMR Spectroscopy, Angew. Chem.. Int. Ed. Engl., *22,* pp.350-380.1983.

Guo, L.F, F. Kooli and M. Garland. A General Method for the Recovery of Pure Powder XRD Patterns from Complex Mixtures Using No *a priori* Information. Application of Band-Target Entropy Minimization (BTEM) to Materials Characterization of Inorganic Mixtures, Anal. Chim. Acta, *517*(1-2), pp.229-236. 2004.

Guo, L.F., A. Wiesmath, P. Sprenger, M. Garland. Development of 2D Band-Target Entropy Minimization and Application to the Deconvolution of Multicomponent 2D Nuclear Magnetic Resonance Spectra, Anal. Chem., *77*, pp. 1655-1662. 2005.

Guo, LF.and M. Garland. The use of entropy minimization for the solution of blind source separation problems in image analysis. Pattern Recognition, *39*, pp.1066-1073. 2006.

Hadamard, J. Sur les probl´emes aux d´eriv´ees partielles et leur signification physique, Bull. Univ. Princeton, 13. pp.49-52. 1902.

Hamilton, J.C. and P.J. Gemperline. Mixture Analysis using Factor Analysis II: Self-modeling Curve Resolution, J. Chemom, *4*, pp. 1-13. 1990.
Harshman, R. Foundations of the PARAFAC Procedure: Models and Conditions for an Explanatory Multi-Modal Factor Analysis, UCLA Working Papers in Phonetics, *16,* pp.1-84. 1970.

Harshman, R.A. and M.E. Lundy. Uniqueness Proof for a Family of Models Sharing Features of Tucker's Three-Mode Factor Analysis and PARAFAC/CANDECOMP, Psychometrika, *61,* pp.133-154. 1996.

Hart, S.J., G.J. Hall, J.E. Kenny. A Laser-Induced Fluorescence Dual-Fiber Optic Array Detector Applied to the Rapid HPLC Separation of Polycyclic Aromatic Hydrocarbons, Anal. Bioanal. Chem., *372,* pp.205–215. 2002.

Hashimoto, W. Separation of independent components from data mixed by several mixing matrices, Signal. Process., *82*(12), pp.1949-1961. 2002.

Haykin, S. Cocktail party phenomenon: What is it, and how do we solve it? In European Summer School on ICA, Berlin, Germany, 2003.

Henrion, R. On Global, Local and Stationary Solutions in Three-Way Data Analysis, J Chemom., *14*(3), pp. 261-274. 2000.

Ho, C.N., Christian, G.D. and Davidson, E.R. Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer, Anal. Chem., *50*, pp. 1108 – 1113, 1978.

Ho, C.N., Christian, G.D. and Davidson, E.R. Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons, Anal. Chem., *52*, pp. 1071 – 1079, 1980.

Ho, C.N., Christian, G.D. and Davidson, E.R. Simultaneous Multicomponent Rank Annihilation and Applications to Multicomponent Fluorescent Data Acquired by the Video Fluorometer, Anal.Chem., *53*(1) , pp. 92-98. 1981.

Hoch, J.C. and S.S. Alan. NMR Data Processing.New York: Wiley-Liss. 1996.

Hogg, R.V. and E.A. Tanis. Probability and Statistical Inference. New York: Maxwell Macmillan International. 1993.

Holland J.H., Adaptation in Nature and Artificial Systems. The University of Michigan Press. 1975. Reprinted by MIT Press. 1992.

Holmes, E. and H. Antti. Chemometric Contributions to the Evolution of Metabonomics: Mathematical Solutions to Characterising and Interpreting Complex Biological NMR Spectra, Analyst, *127,* pp.1549-1557. 2002.

Hopke, P.K, P. Paatero, H. Jia, R.T. Ross and R.A. Harshman. Three-Way (PARAFAC) Factor Analysis: Examination and Comparison of Alternative Computational Methods as Applied to Ill-Conditioned Data, Chemom. Intell. Lab. Syst., *43* (1-2), pp. 25-42. 1998.

Hörchner, U. and J.H. Kalivas. Further Investigation on a Comparative Study of Simulated Annealing and Genetic Algorithm for Wavelength Selection, Anal. Chim. Acta, *311,* pp.1-13. 1995.

Horst, P. Sixty Years with Latent Variables and still more to come, Chemom. Intell. Lab. Syst., *14*, pp.5-21. 1992.

Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components, J. Educ. Psychol., *24*, pp. 417-441,498-520.1933.

Huo, R., R. Wehrens, J. Van Duyhoven and L.M.C. Buydens, Assessment of techniques for DOSY NMR data processing, Anal. Chim. Acta., *490*, pp. 231–251.2003.

Huo, R., R. Wehrens and L.M.C. Buydens, Improved DOSY NMR data processing by data enhancement and combination of multivariate curve resolution with non-linear least square fitting, J. Magn. Reson., *169*, pp. 257–269. 2004.

Hyvärinen, A. Fast and Robust Fixed-Point Algorithms for Iindependent Component Analysis, IEEE Trans. Neural Networks, *10*(3), pp.626–634. 1999.

Iggo, J.A., D. Shirley and N. C. Tong. A High Pressure NMR Flow Cell for the *in situ* Study of Homogeneous Catalysis, New J. Chem. Commun., *22*, pp.1043 -1045.1998.

Ingle, J.R., D. James and R. Stanley. Spectrochemical Analysis. pp.438-464, Prentice Hall: Englewood Cliffs, N.J. 1988.

Jackson, J.E. A User's Guide to Principal Components. New York: John Wiley. 1991.

Jaynes, E.T. Information Theory and Statistical Mechanics, Phys. Rev., *106*, pp.620-630. 1957.

Jeener, J. Ampere Int. Summer School 11. Basko Polje, Yugoslavia. 1971.

Jeener, J., B.H. Meier, P. Bachmann and R. R. Ernst. Investigation of Exchange Process by Two-Dimensional NMR Spectroscopy, J. Chem. Phys., *71*, pp.4546-4553. 1979.

Jenkins, R. and R. L. Snyder. Introduction to X-ray Powder Diffractometry, pp.205-230. New York: Wiley. 1996.

Jetter, K., U. Depczynski, K. Molt and A. Niemoller. Principles and Applications of Wavelet Transformation of Chemometrics, Anal. Chim. Acta, *420* (2), pp.169-180. 2000.

Jiang, H.J., Y.Z. Liang and Y.Ozaki. Principles and Methodologies in Self-Modeling Curve Resolution, Chemom. Intell. Lab. Syst., *71*, pp.1– 12. 2004.

JiJi, R.D., G.A. Cooper and K.S. Booksh, Excitation-Emission Matrix Fluorescence Based Determination of Carbamate Pesticides and Polycyclic Aromatic Hydrocarbons, Anal.Chim. Acta, *397*, pp.61–72. 1999.

JiJi, R.D., K.S. Booksh. Mitigation of Rayleigh and Raman Spectral Interferences in Multiway Calibration of Excitation-Emission Matrix Fluorescence Spectra, Anal. Chem. *72* (4), pp.718-725. 2000.

Joaquim J., M. Vives, R. Gargallo and R. Tauler. Multivariate Resolution of NMR Labile Signals by Means of Hard and Soft-Modelling Methods, Anal. Chim. Acta, *490* (1-2), pp. 253-264. 2003.

Jolliffe, I.T. Principal Component Analysis. New York: Springer. 1986.

Jones, J.A. and P.J. Hore. The Maximum-Entropy Method and Fourier Transformation Compared, J. Magn. Reson., *92* (2), pp.276-292. 1991.

Jutten, C. and J. Hérault. Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture, Sign. Proc., *24*, pp. 1-20. 1991.

Kababya, S., I. Bilkis and D. Goldfarb. Dynamic Processes of 1,1'-dihydroxy-2,2',6,6'-tetra-tert-butylbiphenyl Radical Cation in Sulfuric Acid as Studied by Two-Dimensional FT-EPR Spectroscopy, J. Am. Chem. Soc., 118 (40), pp.9680-9690. 1996.

Kaiser, G. A Friendly Guide to Wavelets, pp. 44-45. Boston: Birkhauser. 1994.

Kapur, J.N. Maximum-Entropy Models in Science and Engineering. pp.3-9 & pp.513-523, New Delhi: Wiley Eastern Ltd. 1993.

Karlsmore, J., A. Koggersbol, N. Jensen and S. B. Jorgensen. A Two Stage Procedure for Control Structure Analysis and Design, Comput. Chem. Eng., *18*, pp. S465–S470. 1994.

Karjalainen, E.J. The Spectrum Reconstruction Problem-Use of Alternating Regression for Unexpected Spectral Components in Two-Dimensional Spectroscopies, Chemom. Intell. Lab. Syst., *7* (1-2), pp. 31-38. 1989.

Karjalainen, E.J. Isolation of Pure Spectra in GC/MS by Mathematical Chromatography, Entrophy Consideration. In Computer Enhanced Analytical Spectroscopy, Vol 2, Ed by H. Meuzelaar, pp.49-70. New York : Plenum Press.1990.

Kauppinen, J.K., D.J. Moffatt, H.H. Mantsch, and D.G. Cameron, Fourier Transforms in the Computation of Self-Deconvoluted and First-Order Derivative Spectra of Overlapped Band Contours, Anal. Chem., *53*, pp.1454.1981.

Kawata, S., K. Sasaki, S. Minami and H. Komeda. Advanced Algorithm for Determing Component Spectra Based on Principal Component Analysis, Appl. Spectrosc., *39*, pp. 610-614. 1985.

Kawata, S., K. Sasaki and S. Minami. Component Analysis of Spatial and Spectral Patterns in Multispectral Images. I. Basis. J. Opt. Soc. Am., *4*, pp.2101-2106. 1987.

Kawata, S., K. Sasaki and S. Minami. Component Analysis of Spatial and Spectral Patterns in Multispectral Images. II. Entropy Minimization. J. Opt. Soc. Am., *6*, pp. 73-79. 1989.

Keifer, P.A. NMR tools for biotechnology, Current Opinion in Biotechnology, *10*(1), pp.34-41. 1999.

Keller, H.R., D.L. Massart, Y.Z. Liang and O.M. Kvlheim. A Comparison of the Heuristic Evolving Latent Projections and Evolving Factor-Analysis Methods for Peak Purity Control in Liquid-Chromatography Aphy with Photodiode Array Detection, Anal. Chem., *267* (1), pp.63-71. 1992.

Keunok, H.Y., H.Y. Yang, J.M. Rhee, M.H. Lee, and S.C. Yu. Two-Dimensional Raman Correlation Spectroscopy Study of the. Pathway for the Thermal Imidization of Poly(amic acid), Bull. Korean Chem. Soc., *24*(3), pp.357. 2003.

Kiers, H.A.L. Three-Way SIMPLIMAX for Oblique Rotation of the Three-Mode Factor Analysis Core to Simple Structure, Comp. Stat. Data Anal., *28*(3), pp. 307-324. 1998.

Kiers, H.A.L., J.M.F. Ten Berge and R. Bro. PARAFAC2 - Part I. A Direct Fitting Algorithm for the PARAFAC2 Model, J. Chemom., *13*(3-4), pp.275-294. 1999.

Kim, B. M. Development of a New Multivariate Receptor Model and its Application to Los Angeles Airborne Particle Data. Ph.D. Dissertation, University of Southern California, Los Angeles, CA. 1989.

Kim, I.S., K.D. Barrow and P.L. Rogers. Application of Nuclear Magnetic Resonance Spectroscopy to Analysis of Ethanol Fermentation Kinetics in Yeasts and Bacteria, Biotechnol. Lett., *21* (10), pp.839-848. 1999.

Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi. Optimization by Simulated Annealing, Science, *220*, pp. 671–680. 1983.

Koehl, A. Linear Prediction Spectral Analysis of NMR Data, Progress in NMR spectroscopy, *34*, pp.257-299, 1999.

Kolda, T. G. Orthogonal Tensor Decompositions SIAM, J. Matrix Anal. A , *23* (1), pp. 243-255. 2001.

Kroonenberg, P. M. and J. de Leeuw. Principal Component Analysis of Three-Mode Data by Means of Alternating Least Squares Algorithms, Psychometrika, *45*, pp.69-97. 1980.

Kumar, A., R. R. Ernst and K. Wüthrich. A Two-Dimensional Nuclear Overhauser Enhancement (2D NOE) Experiment for the Elucidation of Complete Proton-Proton Cross-Relaxation Networks in Biological Macromolecules, Biochem. Biophys. Res. Commun., *95*, pp.1-6. 1980.

Kupce, E. and R. Freeman. The Radon Transform: A New Scheme for Fast Multidimensional NMR, Concepts Magn. Reson. Part A, *22*A (1), PP.4-11. 2004.

Kvalheim, O.M. and Y.Z. Liang. Heuristic Evolving Latent Projections: Resolving Two-Way Multicomponent Data: 1. Selectivity, Latent-Projective Graph, Datascope, Local Rank and Unique Resolution, Anal. Chem., *64* (8), pp. 936–946. 1992.

Ladroue, C., A.R. Tate, F.A., Howe and J.R. Griffiths. Unsupervised Feature Extraction of in Vivo Magnetic Resonance Spectra of Brain Tambours Using Independent Component Analysis. In Intelligent Data Engineering and Automated Learning - IDEAL 2002, Third International Conference, August 2002, Manchester, UK, pp.441-446. Proceedings. Lecture Notes in Computer Science 2412, Springer 2002.

Lakowicz, J.R. Principles of fluorescence spectroscopy. New York: Kluwer Academic/Plenum. 1999.

Larsen , R.M. August 1998, URL: http://soi.stanford.edu/~rmunk/PROPACK/

Laue, E.D., M.R. Mayger, J. Skilling and J. Staunton. Reconstruction of Phase-Sensitive Two-Dimensional NMR-Spectra by Maximum Entropy, J. Magn. Reson., *68*, pp.14-29. 1986.

Lavine, B.K. Chemometrics, Anal. Chem., *70*, pp. 209R-228R. 1998.

Lavine B.K. and A.J. Moores. Genetic Algorithms in Analytical Chemistry, Anal. Lett., *32* (3), pp. 433-445. 1999.

Lavine, B.K. Chemometrics, Anal. Chem., *72*, pp. 91R-97R. 2000.

Lavine, B.K. Chemometrics, Anal. Chem., *74*, pp. 2763-2770. 2002

Lavine, B.K. and JJ. Workman. Chemometrics, Anal. Chem., *76*(12), pp. 3365-3371. 2004.

Lawton, W.H. and E.A. Sylvestre. Self Modeling Curve Resolution, Technometrics, *13*, pp. 617-633, 1971.

Lazzaroni, R., R. Settambolo, A. Caiazzo and M.A. Bennett. Rhodium-Catalyzed Hydroformylation of 4-vinylpyridine: 4-Ethylpyridine Formation via an Unusual Cleavage of the Rh-C Bond by the Enolic Form of the Oxo Product. Organometallics., *21*(12), pp. 2454-2459. 2002.

Leardi, R. Genetic Algorithms in Chemometrics and Chemistry: a Review, J. Chemom., *15* (7), pp. 559-569. 2001.

Lee, D.D. and H.S. Seung. Learing the Parts of Objects by Nonnegative Matrix Factorization, Nature, *401*, pp. 788-791. 1999.

Leger, M.N. and P.D. Wentzell. Dynamic Monte Carlo Self-Modelling Curve Resolution Method for Multicomponent Mixtures, Chemom. Intell. Lab. Syst., *62*, pp. 171-188. 2002.

Lehoucq, R.B., D.C. Sorensen and C. Yang. ARPACK Users's Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. Philadelphia: SIAM. 1998.

Lenz, E.M., J. Bright, R. Knight, I.D. Wilson and H. Major. Cyclosporin A-Induced Changes in Endogenous Meta-Bolites in Rat Urine: a Metabonomic Investigation Using High Field H-1 NMR Spectroscopy, HPLC-TOF/MS and Chemometrics, J. Pharmaceut. Biomed., *35* (3), pp. 599-608 .2004.

Lepre, C.A., J.M. Moore and J. W. Peng. Theory and Applications of NMR-Based Screening in Pharmaceutical Research, Chem. Rev., *104* (8), pp. 3641-3675. 2004.

Leurgans, S. and R.T. ROSS. Multilinear Models: Applications in Spectroscopy(with discussion) , Statist. Sci., 7(3), pp. 289-319. 1992.

Li, C. Z., E. Widjaja, W. Chew, and M. Garland. Rhodium Tetracarbonyl Hydride: The Elusive Metal Carbonyl Hydride. Angew. Chem. Int. Ed., *41*, 20, pp. 3785-3789. 2002.

Li, C. Z., E. Widjaja and M. Garland. The $Rh_4(CO)_{12}$-Catalyzed Hydroformylation of 3,3-Dimethylbut-1-ene Promoted with $HMn(CO)_5$. Bimetallic Catalytic Binuclear Elimination as an Origin for Synergism in Homogeneous Catalysis, J. Am. Chem. Soc., *125*, pp.5540-5548. 2003.

Li, C.Z., L.F. Guo and M. Garland. Homogeneous Hydroformylation of Ethylene Catalyzed by $Rh_4(CO)_{12}$. The Application of BTEM to Identify a New Class of Rhodium Carbonyl Spectra: $RCORh(CO)_3(-C_2H_4)$. Organometallics , *23*(9), pp.2201-2204. 2004a.

Li, C. Z., L. F. Guo and M. Garland, Identification of Rhodium-Rhenium Nonacarbonyl $RhRe(CO)_9$: Spectroscopic and Thermodynamic Aspects, Organometallics, *23*(22), pp.5275-5279. 2004b.

Liang, Y.Z., O.M. Kvlheim, H.R. Keller, D. L. Massart, P. Kiechle and F. Erni. Heuristic Evolving Latent Projections-Resolving 2-Way Multicom ponent Data .1. , Detection and Resolution of Minor Constituents, Anal. Chem., *64* (8), pp. 946-953. 1992.

Liang, Y.Z. and O.M. Kvalheim. Resolution of Two-Way Data: Theoretical Background and Practical Problem-Solving - Part 1: Theoretical Background and Methodology, Fresenius J. Anal. Chem., *370* (6), pp.694-704. 2001.

Libnau, F.O., A.A. Christy and O.M. Kvalheim. Determination of the Equilibrium Constant and Resolution of the HOD Spectrum by Alternating Least-Squares and Infrared Analysis, Appl. Spectrosc., *49*, pp. 1431-7. 1995.

Lin, Y.Y and L.P. Hwang. NMR Signal Enhancement Based on Matrix Property Mappings, J. Magn. Reson. A, *103*, pp. 109-114, 1993.

Liu, G. Kinetics and Mechanism Study of Homogeneous Hydroformylation With Unmodified Rhodium Carbonyls. M. Eng Thesis, National University of Singapore. 1999.

Liu, G. and M. Garland. The Competitive and Non-competitive Hydroformylation of Conjugated Dienes. Starting with Tertrarhodium Dodecacarbonyl, An In-Situ High-Pressure Infrared Spectroscopic Study, J. Organomet. Chem., *608*, pp. 76-85. 2000.

Liu GW, C. Z. Li, L. F. Guo and M. Garland. Experimental evidence for a significant homometallic catalytic binuclear elimination reaction: Linear-quadratic kinetics in the rhodium catalyzed hydroformylation of cyclooctene, J. Catal., *237*, pp. 67-78. 2006.

Lorber, A. Quantifying Chemical Composition from Two-Dimensional Data Arrays, Anal.Chim. Acta, *164*, pp. 293-297. 1984.

Lorber, A. Features of Quantifying Chemical Composition from Two-Dimensional data Arrays by the Rank Annihilation Factor Analysis Method, Anal. Chem., *57*, pp. 2395-2397.1985.

Lucasius, C.B., M.L.M. Beckers and G. Kateman. Genetic Algorithms in Wavelength Selection: a Comparative Study, Anal. Chim. Acta, *286*, pp.135-153.1994.

Maeder, M. Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks, Anal. Chem., *59*, pp.527-530. 1987.

Maiwald, M., H.H. Fischer, Y.K. Kim and H. Hasse. Quantitative on-line High-Resolution NMR Spectroscopy in Process Engineering Applications, Anal. Bioanal. Chem., *375* (8), pp.1111-1115. 2003.

Maiwald, M., H.H. Fischer and Y. K. Kim et al. Quantitative High-Resolution on-line NMR Spectroscopy in Reaction and Process Monitoring, J. Magn. Reson., *166* (2), pp.135-146. 2004.

Makeig, S., T.P. Jung, A.J. Bell, D. Ghahremani and T.J. Sejnowski. Blind Separation of Auditory Event-Related Brain Responses into Independent Components, Proceedings of the National Academy of Sciences of the United States of America, *94* (20), pp.10979-10984. 1997.

Malinowski, F.R. Determination of the Number of Factors and the Experimental Error in a Data Matrix, Anal. Chem., *49*, pp.612-617. 1977.

Malinowski, E.R. Obtaining the Key Set of Typical Vectors by Factor Analysis and Subsequent Isolation of Component Spectra, Anal. Chim. Acta, *134*, pp. 129-137. 1982.

Malinowski, E.R. Theory of the Distribution of Error Eigenvalues resulting from Principal Component Analysis with Applications to Spectroscopic Data. J. Chemom., *1*(1), pp. 33-40. 1987.

Malinowski, E.R. Factor Analysis in Chemistry. pp.10-22, New York: John Wiley. 1991.

Malinowski, E.R. Window Factor-Analysis - Theoretical Derivation and Application to Flow-Injection Analysis Data, J. Chemom., *6*, pp. 29-40. 1992.

Malinowski, E.R. Factor Analysis in Chemistry, 3rd ed., New York: Wiley, 2002.

Mandelshtam, V.A . FDM: the Filter Diagonalization Method for Data Processing in NMR Experiments, Prog. Nucl. Magn. Reson. Spectrosc., *38* (2), pp.159-196. 2001.

Matlab, MathWorks Inc. MatLab Reference Guide, 1995.

Metropolis, N., M. Rosenbluth, A. Teller and E. Teller. Equation-of-State Calculations by Fast Computing Machines, J. Chem. Phys., *21*, pp. 1087-1092. 1953.

Meyer, Y. Wavelets: Algorithms and Applications, Society for Industrial and Applied Mathematics, Philadelphia, pp. 13-31, 101-105. 1993.

Millican D.W. and L.B. Mcgown. Fluorescence Lifetime Resolution of Spectra in the Frequency-Domain using Multiway Analysis, Anal.Chem., *62* (20), pp. 2242-2247. 1990.

Mobley, P.R., B.R. Kowalski, J. J. Workman and R. Bro. Review of Chemometrics applied to Spectroscopy: 1985-95.2., Appl. Spectrosc. Rev., *31* (4), pp. 347-368. 1996.

Muroski, A.R., K.S. Booksh and M.L. Myrick. Single-Measurement Excitation/Emission Matrix Spectrofluorometer for Determination of Hydrocarbons in Ocean Water. 1. Instrumentation and Background Correction, Anal. Chem., *68*, pp. 3534-3538. 1996.

Myers, C.S. and L.R. Rabiner. A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition, Bell. Sys. Tech. J., *60*(7), pp.1389-1409. 1981.

Nahorniak, M.L., G.A. Cooper, Y.C. Kim and K.S. Booksh. Three and Four-Way Parallel Factor (PARAFAC) Analysis of Photochemically Induced Excitation-Emission Kinetic Fluorescence Spectra, Analyst, *130* (1), pp.85-93. 2005.

Neal, S.L., E.R. Davidson and I.M. Warner. Resolution of Severely Overlapped Spectra from Matrix-Formatted Spectral Data Using Constrained Nonlinear Optimization, Anal. Chem., *62*, pp.658-664. 1990.

Neuhold, Y.M. and M. Maeder. Hard-Modelled Trilinear Decomposition (HTD) for an Enhanced Kinetic Multicomponent Analysis, J. Chemom., *16* (5), pp.218-227. 2002.

Nielsen, N.PV, J.M. Carstensen and J. Smedsgaard. Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis using Correlation Optimised Warping, J. Chromatogr. A, *805* (1-2), pp.17-35. 1998.

Nodland, E., F.O. Libnau, O.M. Kvalheim, H.J. Luinge and P. Klaeboe, Influence and Correction of Peak Shift and Band Broadening Observed by Rank Analysis on Vibrational Bands from Variable-Temperature Measurements, Vib. Spectrosc.,*10*, pp. 105. 1996.

Ohta, N. Estimating Absorption Bands of Component Dyes by Means of Principal Component Analysis, Anal. Chem., *45*, pp. 553-557. 1973.

Olivieri, A.C., J.A. Arancibia, A.M. de la Pena, I. Duran-Meras and A. E. Mansilla. Second-Order Advantage Achieved with Four-Way Fluorescence Excitation-Emission-Kinetic Data Processed by Parallel Factor Analysis and Trilinear Least-Squares. Determination of Methotrexate and Leucovorin in Human Urine, Anal. Chem., *76* (19), pp.5657-5666. 2004.

Ong, L.R., E. Widjaja, R. Stanforth and M. Garland. Fourier Transform Raman Spectral Reconstruction of Inorganic Lead Mixtures using a Novel Band-Target Entropy Minimization (BTEM) Method, J. Raman Spectrosc., *34* (4), pp.282-289. 2003.

Ord, J.K. Families of Frequency Distributions. London: Charles Griffin and Co.1972.

Oschkinat, H., C. Griesinger, P.J. Kraulis, O.W. Sorensen, R.R. Ernst, A.M. Gronenborn and G.M. Clore. Three-dimensional NMR spectroscopy of a protein in solution, Nature, *332*(6162), pp.374-376. 1988.

Paatero, P. and U. Tapper. Positive matrix factorization: A Nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values, Environmetr., *5* ,pp.111-126. 1994.

Paatero, P. Least Squares Formulation of Robust Non-Negative Factor Analysis, Chemom. Intell. Lab. Syst., *37* (1), pp. 23-35. 1997.

Patonay, G., G. Nelson and I. M. Warner. Recent Advances in Multidimensional Fluorescence Spectrometry, Prog. Anal. Spectrosc., *10* (6), pp.561-571, 1987.

Paul, G., B. Ewert, E. Kim and G. Hans. Image Analysis in Cemistry I. Properties of Images, Greylevel Operations, the Multivariate Image, Trends Anal. Chem., *11*(1), pp. 41-53.1992.

Peakfit v4.0, User's Manual, Jandel Scientific, 259 Kerner Blvd. San Rafael, CA, 1995.

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space, Phil. Mag., *2* (6), pp.559-572. 1901.

Pedersen, H.T., R. Bro and S.B. Engelsen. Towards Rapid and Unique Curve Resolution of Low-Field NMR Relaxation Data: Trilinear SLICING versus Two-Dimensional Curve Fitting, J. Magn. Reson., *157* (1), pp.141-155. 2002.

Piantini, U., O.W.Sorensen and R.R.Ernst. Multiple Quantum Filters for Elucidating NMR Coupling Networks, J. Am. Chem. Soc., *104*(24), pp.6800-6801. 1982.

Plumbley, M.D., S.A. Abdallah , J.P. Bello, M.E. Davies, G. Monti and M.B. Sandler. Automatic Music Transcription and Audio Source Separation, Cybernetics and Systems, *33* (6), pp.603-627. 2002.

Pluschau, M. and K.P. Dinse. 2D EPR Study of a Photoinduced Proton Abstraction in the System Anthraquinone and 4-Methyl-2,6-di-tert-butylphenol in 2-Propanol, J. Magn. Reson., Ser A, *109* (2), pp.181-191. 1994.

Pluschau, M., K.P. Dinse. 2D EPR data Dnalysis using Linear Prediction Singular Value Decomposition and Zero-Filling, Appl. Magn. Reson., *9*(2), pp.299-304. 1995.

Ramos, L.S., E. Sanchez and B.R. Kowalski. Generalized Rank Annihilation Method. 2. Analysis of Bimodal Chromatographic Data, J. Chromatogr., *385,* pp. 165-180. 1987.

Rasmussen, G.T., T.L. Isenhour, S.R. Lowry and G.L. Ritter. Principal Component Analysis of the Infrared Spectra of Mixtures, Anal. Chim. Acta, *103,* pp. 213-221. 1978.

Ren, J.Y., C.Q. Chang, P.C.W. Fung, J.G. Shen and F.H.Y. Chan. Free Radical EPR Spectroscopy Analysis Using Blind Source Separation , J. Magn. Reson., *166* (1), pp. 82-91. 2004.

Renée, D., R.D. JiJi and K.S. Booksh. Mitigation of Rayleigh and Raman Spectral Interferences in Multiway Calibration of Excitation-Emission Matrix Fluorescence Spectra, Anal. Chem., *72*(4), pp.718-725. 2000.

Richman, M.B. Rotation of principal components, J. Climatol., 6, 293-335. 1986.

Rinnan, A., K.S. Booksh and R. Bro. First Order Rayleigh Scatter as a Separate Component in the Decomposition of Fluorescence Landscapes, Anal. Chim. Acta, *537* (1-2), pp.349-358. 2005.

Ritter, G.L., S.R. Lowry, T.L. Isenhour and C.L. Wilkins. Factor Analysis of The Mass Spectra of Mixtures, Anal. Chem., *48*, pp. 591-595.1976.

Romo, T.D., J.B. Clarage, D.C. Sorensen and G.N. Phillips. Jr. Automatic Identification of Discrete Substates in Proteins: Singular Value Decomposition Analysis of Time-Averaged Crystallographic Refinements, Proteins, *22,* pp.311-321. 1995.

Rummel, R.J. Applied Factor Analysis, Evanston, Illinois: Northwestern Universtiy Press. 1970.

Sabatier, P.C. Introduction to applied inverse problem. In Applied inverse problems : lectures presented at the RCP 264 in Montpellier, ed by P.C. Sabatier, pp.1-27. NewYork: Springer-Verlag, Berlin/Heidelberg. 1978.

Sanchez, E. and B.R. Kowalski. Generalized Rank Annihilation Factor Analysis, Anal. Chem., *58* , pp. 496-499. 1986.

Sanchez, E., L.S. Ramos and B.R. Kowalski. Generalized Rank Annihilation Method. 1. Application to Liquid-Chromatography Diode-Array Ultraviolet Detection Data, J. Chromatogr., *385,* pp.151-164 ,1987.

Sanchez, E. and B.R. Kowalski. Tensorial Resolution: a Direct Trilinear Decomposition, J. Chemom., *4,* pp.29-45. 1990.

Sanchez, F.C., M.S. Khots, D.L. Massart and J.O. Debeer. Algorithm for the Assessment of Peak Purity in Liquid-Chromatography with Photodiode-array Detection, Anal. Chim. Acta, *285* (1-2), pp. 181-192. 1994.

Sanchez, F.C., B. van den Bogaert, S.C. Rutan, and D.L. Massart. Multivariate Peak Purity Approaches. Chemom. Intell. Lab. Syst. *34*, pp. 139-171. 1996a.

Sanchez, F.C., J. Toft, B. van den Bogaert, and D.L. Massart. Orthogonal Projection Approach Applied to Peak Purity Assessment. Anal. Chem. *68*, pp. 79-85. 1996b.

Sarazin, C., F. Ergan, J.P. Seguin, G. Goethals, M.D. Legoy and J.N. Barbotin. NMR Online Monitoring of Esterification Catalyzed by Cutinase, Biotechnol. Bioeng., *51*(6), pp.636-644. 1996.

Sasaki, K., S. Kawata and S. Minami. Constrained Nonlinear Method for Estimating Component Spectra from Multicomponent Mixture, Appl. Optics., *22,* pp. 3599-3603. 1983.

Sasaki, K., S. Kawata and S. Minami. Estimation of Component Spectral Curves from Unknown Mixture Spectra, Appl. Optics. , *23,* pp. 1955-1959. 1984.

Saurina, J., S. Hernandez-Cassou, R. Tauler and A. Izquierdo-Ridorsa. Multivariate Resolution of Rank-Deficient Spectrophotometric Data from First-Order Kinetic Decomposition Reactions, J. Chemom., *12* (3), pp.183-203. 1998.

Savitzky, A. and M. J. E. Golay. Smoothing and Differentiating of Data by Simplified Least Squares Procedures, Anal. Chem., *36,* pp.1627-1639. 1964.

Schmidt, M., S. Rajagopal, R. Zhong and K. Moffat. Application of Singular Value Decomposition to the Analysis of Time-Resolved Macromolecular X-Ray Data, Biophys. J., *84*, pp.2112-2129. 2003.

Schulze, D. and P. Stilbs. Analysis of Multicomponent FT-PGSE Experiments by Multivariate Statistical Methods Applied to the Complete Bandshapes, J. Magn. Res. *105*(A), pp.54-58. 1993.

Shaffer, R.E. and G.W. Small. Learning Optimization from Nature. Genetic Algorithms and Simulated Annealing, Anal. Chem. , *69*(7), pp. 236A-242A. 1997.

Shannon, C.E. A Mathematical Theory of Communication, Bell Syst. Tech. J., *27*, pp.379-423, 623-656. 1948.

Shao, X.G., G.Q. Wang, S.F. Wang and Q.D. Su, Extraction of Mass Spectra and Chromatographic Profiles from Overlapping GC/MS Signal with Background, Anal. Chem., *76*(17), pp. 5143-5148. 2004.

Siarry, P., G. Berthiau, F. Durbin and J. Haussy. Enhanced Simulated Annealing for Globally Minimizing Functions of Many Continuous Variables, ACM. Trans. Math. Softw., *23*, pp. 209–228. 1997.

Sibisi, S. Two-Dimensional Reconstructions from One-Dimensional Data by Maximum Entropy, Nature, *301*, pp.134–136. 1983.

Sibisi, S., J. Skilling, R.G. Brereton, E.D. Laue and J. Staunton. Maximum Entropy Signal Processing in Practical NMR Spectroscopy, Nature, *311*, pp.446–447. 1984.

Silva, S. M., R.P.J.Bronger, Z. Freixa, J. Dupont and van Leeuwen, Piet W. N. M. High Pressure Infrared and Nuclear Magnetic Resonance Studies of the Rhodium-Sulfoxantphos Catalysed Hydroformylation of 1-octene in Ionic Liquids, New J. Chem., *27*(9), pp.1294-1296. 2003.

Simonetti, A.W., G. Postma, R. Huo, F. Szabo and L.M.C. Buydens. Application of Independent Component Analysis to ^{1}H MR Spectroscopic Imaging Exams of Brain Tambours, Anal. Chim. Acta, *544*, pp.36-46. 2005.

Sin, S.Y., E. Widjaja, L.E. Yu and M. Garland. Application of FT-Raman and FTIR Measurements using a Novel Spectral Reconstruction Algorithm , J. Raman Spectrosc., *34* (10), pp.795-805. 2003.

Singh, A. Outliers and Robust Procedures in some Chemometic Applications, Chemom. Intell. Lab. Syst., *33*, pp.75-100. 1996.

Skilling, J. and R.K. Bryan. Maximum Entropy Image Reconstruction: General Algorithm, Mon. Notices R. Astron. Soc., *211*, pp.111-124. 1984.

Skilling, J. (ed). Maximum Entropy and Bayesian Methods. Norwell, MA: Kluwer Academic. pp. 45-52. 1989.

Smilde, A. K. Three-way analyses: Problems and Prospects, Chemom. Intell. Lab. Syst, *15,* pp.143-157. 1992.

Starck, J.L. and F. Murtagh. Multiscale Entropy Filtering, Signal Processing, *76*(2), pp.147-165. 1999.

Stedmon, C.A. and S. Markager. Resolving the Variability in Dissolved Organic Matter Fluorescence in a Temperate Estuary and its Catchment using PARAFAC Analysis, Limnology and Oceanography, *50* (2), pp.686-697. 2005.

Steinbock, O., B. Neumann, B. Cage, J. Saltiel, S.C. Muller and N.S. Dalal. A Demonstration of Principal Component Analysis for EPR Spectroscopy: Identification of Pure Compounds from Complex Spectra, Anal. Chem., *69* , pp. 3708-3713. 1997.

Stern, A.S., K.B. Li and J.C. Hoch. Modern Spectrum Analysis in Multidimensional NMR Spectroscopy: Comparison of Linear-Prediction Extrapolation and Maximum-Entropy Reconstruction, J. Am. Chem. Soc., *124* (9), pp.1982-1993. 2002.

Stilbs, P., K. Paulsen and P.C. Griffiths. Global Least-Squares Analysis of Large, Correlated Spectral Data Sets: Application to Component-Resolved FT-PGSE NMR Spectroscopy, J. Phys. Chem. ,*100* (20), pp. 8180-8189. 1996.

Stogbauer, H., A. Kraskov, S.A. Astakhov and P. Grassberger, Least-Dependent-Component Analysis Based on Mutual Information, Phys. Rev. E., *70* (6), Art. No. 066123, 2004.

Stoyanova, R. and T.R. Brown. NMR spectral quantitation by principal component analysis - III. A generalized procedure for determination of lineshape variations, J. Magn. Reson.,*154,* pp.163-175. 2002.

Stoyanova, R., A.W. Nicholls, J.K. Nicholson, J.C. Lindon and T.R. Brown. Automatic alignment of individual peaks in large high-resolution spectral data sets, J. Magn. Reson., *170*, pp. 329–335. 2004

Susithra, L. In-Situ Kinetics Studies Of The Homogeneous Catalytic Hydroformylation With Unmodified Cobalt Carbonyls. M. Eng Thesis, National University of Singapore. 1999.

Swierenga, H., A.P. de Weijier, R.J. van Wijk and L.M.C. Buydens, Strategy for Constructing Robust Multivariate Calibration Models, Chemom. Intell. Lab. Syst., *49*, pp.1-17. 1999.

Tanimura, Y. and S. Mukamel. Two-Dimensional Femtosecond Vibrational Spectroscopy of Liquids. J. Chem. Phys., 99(12), pp.9496-9511. 1993.

Tauler, R., E.Casassas, and A. Izquierdo-Ridorsa. Self-Modeling Curve Resolution Applied to Spectroscopic Titration Data using Factor Analysis. Anal. Chim. Acta, *248*, pp. 447-458. 1991.

Tauler, R., I. Marqués and E. Casassas. Multivariate Curve Resolution Applied to Three-way Trilinear Data: Study of a Spectrofluorimetric Acid-base Titration of Salicylic Acid at Three Excitation Wavelengths, J. Chemom., *12*, pp. 55-75. 1998.

Tauler, R. Calculation of Maximum and Minimum Band Boundaries of Feasible Solutions for Species Profiles Obtained by Multivariate Curve Resolution, J. Chemom., *15*, pp. 627-646. 2001.

Thygesen, L.G. and S.O. Lundqvistt. NIR Measurement of Moisture Content in Wood under Unstable Temperature Conditions. Part 1. Thermal Effects in Near Infrared Spectra of Wood, J. Near Infrared Spectrosc., *8* (3), pp.183-189. 2000.

Thygesen, L.G., A. Rinnan, S. Barsberg and J.K.S. Maller. Stabilizing the PARAFAC Decomposition of Fluorescence Spectra by Insertion of Zeros Outside the Data Area. Chemom. Intell. Lab. Syst., *71* (2), pp.97-106. 2004.

Tirendi, C.F. and J.F. Martin. Quantitative Analysis of NMR Spectra by Linear Prediction and Total Least Squares J. Magn. Reson., *85* (1), pp. 162-169. 1989.

Tomasi, G., F. van den Berg and C. Andersson. Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data, J. Chemom., *18* (5), pp.231-241. 2004.

Tomita, F. and S. Tsuji. Extraction of Multiple Regions by Smoothing in Selected Neighborhoods, IEEE. Trans. Syst. Man. Cybern., *7*, pp. 107–109. 1977.

Townshend, A. (editor in chief). Encyclopedia of Analytical Science. Vol 3, pp.1373-1401, London / San Diego: Academic Press. 1995.

Tucker, L.R. Some Mathematical Notes on Three-Mode Factor Analysis, Psychometrika, *31*, pp. 279.1966.

Vandeginste, B.G.M., W. Derks, and G. Kateman. Multicomponent Sel-Modelling Curve Resolution in High-Performance Liquid Chromatography by Iterative Target Tranformation Factor Analysis, Anal. Chim. Acta, *173*, pp. 253-264. 1985.

Van Leeuwen, P.W.N.M. and C. Claver. Rhodium Catalyszed Hydroformylation. pp.1-15, Boston: Kluwer Academic. 2000.

Van Smaalen, S., L. Palatinus and M. Schneider. The Maximum-Entropy Method in Superspace , Acta Cryst. A, *59*, pp. 459-469. 2003.

Vetterli, M. and C. Herley. Wavelets and Filter Banks: Theory and Design, IEEE Trans. Sig. Proc., *40,* pp. 2207-2232. 1992.

Vickers, T.J., R.E. Wambles and C.K. Mann. Curve Fitting and Linearity: Data Processing in Raman Spectroscopy, Appl. Spectrosc., *55* (4), pp.389-393. 2001.

Vigario, R.N. Extraction of Ocular Artefacts from EEG using Independent Component Analysis, Elecroencephalography and Clinical Neurophysiology, *103*(3), pp.395-404. 1997.

Vives, M., R. Tauler, V. Moreno and R. Gargallo. Study of the Interaction of a Cis-dichloroaminopyrrolidine Pt(II) Complex and the Polynucleotide Poly(I)–Poly(C) Acid by means of ^{1}H-NMR and Multivariate Curve Resolution, Anal.Chim. Acta, *446,* pp. 439–450.2001.

Vogels, J.T.W.E., A.C. Tas, F. van den Berg and J. van der Greef. Partial linear fit: A New NMR Spectroscopy Preprocessing Tool for Pattern Recognition Applications, Chemom. Intell. Lab. Syst, *21*, pp.249–258. 1993.

Volkov, V.V. Separation of Additive Mixture Spectra by a Self-Modeling Method, Appl. Spectrosc., *50,* pp.320-326. 1996.

Wall, M.E., P.A. Dyck and T.S. Brettin. SVDMAN -- Singular Value Decomposition Analysis of Microarray Data, Bioinformatics, *17*, pp.566-68. 2001.

Walczak, B. Outliers Detections in Bilinear Carlibration, Chemom. Intell. Lab. Syst., *29*, pp. 63-73. 1995.

Wallace, R.M. Analysis of Absorption Spectra of Multicomponent Systems, J. Phys. Chem., *64*, pp.899. 1960.

Wasim, M. and R.G. Brereton. Determination of the Number of Significant Components in Liquid Chromatography Nuclear Magnetic Resonance Spectroscopy, Chemom. Intell. Lab. Syst., *72* (2), pp.133-151. 2004.

Watanabe, S. Pattern-Recognition as a Quest for Minimum Entropy, Pattern Recognition, *13*, pp. 381-387. 1981.

Weber, H. and L. Brecker. Online NMR for Monitoring Biocatalysed Reactions, Current Opinion in Biotechnology, *11* (6), pp.572-578 . 2000.

Wells, W.M. Efficient Synthesis of Gaussian Filters by Cascaded Uniform Filters, IEEE Trans. Patt. Anal. Mach. Intell., *8*, pp. 234–239. 1986.

Wertz, J.E. and J.R. Bolton. Electron Spin Resonance: Elementary Theory and Practical Applications. New York: Chapman and Hall .1986.

Wider, G., S. Macura, A. Kumar, R.R. Ernst and K. Wüthrich. Homonuclear Two-Dimensional ^{1}H NMR of Proteins: Experimental Procedures, J. Magn. Reson., *56*, pp.207-234. 1984.

Widjaja, E. Development of Band-Target Entropy Minimization (BTEM) and Associated Software Tools. Ph.D Thesis, National University of Singapore. 2002.

Widjaja, E. and M. Garland. Pure Component Spectral Reconstruction from Mixture Data using SVD, Global Entropy Minimization and Simulated Annealing. Numerical Investigations of Admissible Objective Functions Using a Synthetic 7-Species Data Set. J. Comput. Chem. *23*, pp. 911-919. 2002.

Widjaja, E., C.Z. Li, and M. Garland. Semi-Batch Homogeneous Catalytic In-Situ Spectroscopic Data. FTIR Spectral Reconstructions Using Band-Target Entropy Minimisation (BTEM) without Spectral Preconditioning. Organometallics. *21*, pp. 1991-1997. 2002.

Widjaja, E., C.Z. Li, W. Chew and M. Garland. Band-Target Entropy Minimization. A Robust Algorithm for Pure Component Spectral Recovery. Application to Complex Randomized Mixture of six Component, Anal. Chem., *75* (17), pp.4499-4507. 2003.

Widjaja E. and M. Garland. Entropy Minimization and Spectral Dissimilarity Curve Resolution Technique Applied to Nuclear Magnetic Resonance Data Sets, J. Magn. Reson., *173* (1), pp.175-182. 2005.

Windig, W. Mixture Analysis of Spectral Data by Multivariate Methods, Chemom. Intell. Lab. Syst., *4*, pp.201-213. 1988.

Windig, W., J.L. Lippert, M.J. Robbins, K.R. Kresinske, J.P. Twist and A.P. Snyder.Interactive Self-Modeling Multivariate Analysis, Chemom. Intell. Lab. Syst., *9*, pp.7-30. 1990.

Windig, W. and J. Guilment. Interactive Self-Modeling Mixture Analysis, Anal. Chem., *63*, pp. 1425-1432. 1991.

Windig, W. and D.A. Stephenson. Self-Modeling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data using the Simplisma Approach, Anal. Chem., *64*, pp. 2735-2742. 1992.

Windig, W. Spectral Data Files for Self-Modeling Curve Resolution with Examples using the Simplisma Approach. Chemom. Intell. Lab. Syst. *36*, pp. 3-16. 1997.

Windig, W. and B. Antalek. Direct Exponential Curve Resolution Algorithm (DECRA): A Novel Application of the Generalized Rank Annihilation Method for a Single Spectral Mixture Data Set with Exponentially Decaying Contribution Profiles. Chemom. Intell. Lab. Syst., *37*, 241-254. 1997.

Windig, W., B. Antalek, L.J. Sorriero, S. Bijlsma, D.J. Louwerse, and A.K. Smilde. Applications and New Developments of the Direct Exponential Curve Resolution Algorithm (DECRA). Examples of Spectra and Magnetic Resonance Images, J. Chemom. *13*, pp. 95-110. 1999a.

Windig, W., J. Hornak and B. Antalek. Multivariate Image Analysis of Magnetic Resonance Images with the Direct Exponential Curve Resolution Algorithm (DECRA) Part 1: Algorithm and Model Study, J. Magn. Reson., *132*, pp.298-306, 1999b.

Windig, W., J. Hornak and B. Antalek. Multivariate Image Analysis of Magnetic Resonance Images with the Direct Exponential Curve Resolution Algorithm (DECRA) Part 2: Application to Human Brain Images, J. Magn. Reson., *132*, pp.307-315. 1999c.

Witjes, H., W.J. Melssen, H.J.A. in 't Zandt, M. van der Graaf, A. Heerschap and L.M.C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets. J. Magn. Reson., *144*, pp. 35–44. 2000.

Wold, S. Cross-Validatory Estimation of the Number of Components in Factorand Principal Components Analysis, Technometrics, *20*, pp.397-405. 1978.

Wold, S., K.E sbensenk and P. Geladi. Principal Component Analysis, Chemom. Intell. Lab. Syst., *2* (1-3), pp.37-52. 1987.

Wold, S. Chemometrics: What do we mean with it, and what do we want from it? Chemom. Intell. Lab. Syst., *30*(1), pp. 1-196 . 1995.

Workman, J.J., P.R. Mobley, B. R. Kowalski and R. Bro. Review of Chemometrics Applied to Spectroscopy: 1985-95 .1., Appl. Spectrosc. Rev., *31* (1-2), pp.73-124. 1996.

Wright, J.C. Coherent Multidimensional Vibrational Spectroscopy, Int. Rev. Phys. Chem., *21* (2), pp.185-255. 2002.

Xie, Y.L., P.K. Hopke, and P. Paatero. Positive Matrix Factorization Applied to A Curve Resolution Problem. J. Chemom. *12*, pp. 357-364. 1998.

Yeung, M.K., J. Tegner and J.J. Collins. Reverse Engineering Gene Networks using Singular Value Decomposition and Robust Regression, Proc. Natl. Acad. Sci. U.S.A., *99*, pp.6163-68. 2002.

Zhang, H.J., M. Garland, Y.Z. Zeng and P.Wu. Weighted two-Band Target Entropy Minimization for the Reconstruction of Pure Component Mass Spectra: Simulation Studies and the Application to Real Systems, J. Am. Soc. Mass. Spectrom., *14* (11), pp. 1295-1305. 2003.

Zhao, W., K. M. Murdoch, D.M. Besemann, N.J. Condon, K.A. Meyer and J.C. Wright. Nonlinear Two-Dimensional Vibrational Spectroscopy, Appl. Spectrosc., *54*(7), pp.1000-1004. 2000.

Zhao, W. and J.C. Wright. Doubly Vibrationally Enhanced Four Wave Mixing: The Optical Analog to 2D NMR, Phys. Rev. Lett, *84* (7), pp.1411-1414. 2000.

Zheng, P., P.B. Harrington, A. Craig and R. Fleming. Variable Alignment of High Resolution Data by Cluster Analysis, Anal. Chim. Acta, *310*, pp. 485-492. 1995.

Zhu, G. Application of a Three-Dimensional Maximum-Entropy Method to Processing Sections of Three-Dimensional NMR Spectra. J. Magn. Reson., Ser B, *113* (3), pp.248-251. 1996.

Zimanyi, L., A. Kulcsar, J.K. Lanyi, D.F. Sears and J. Saltiel. Singular Value Decomposition with Self-Modeling Applied to Determine Bacteriorhodopsin Intermediate Spectra: Analysis of Simulated Data ,Proc. Natl. Acad. Sci. U.S.A. , *96*(8) , pp.4408-4413. 1999.

Zimanyi, L. Analysis of the Bacteriorhodopsin Photocycle by Singular Value Decomposition with Self-Modeling: A Critical Evaluation Using Realistic Simulated Data , J. Phys. Chem. B., *108*(13), pp. 4199-4209. 2004.

APPENDICES

Appendix A: Reprinted from Journal of Catalysis, Liu GW, C. Z. Li, L. F. Guo and M. Garland. Experimental evidence for a significant homometallic catalytic binuclear elimination reaction: Linear-quadratic kinetics in the rhodium catalyzed hydroformylation of cyclooctene, *237*, pp. 67-78, Copyright (2006), with permission from Elsevier.

Appendix B: Reproduced with permission from [Li, C.Z., L.F. Guo and M. Garland. Homogeneous Hydroformylation of Ethylene Catalyzed by $Rh_4(CO)_{12}$. The Application of BTEM to Identify a New Class of Rhodium Carbonyl Spectra: $RCORh(CO)_3(-C_2H_4)$. Organomet., *23*(9), pp.2201-2204], Copyright (2004) American Chemical Society.

Appendix C: Reproduced with permission from [Li, C. Z., L.F. Guo and M. Garland, Identification of Rhodium-Rhenium Nonacarbonyl $RhRe(CO)_9$: Spectroscopic and Thermodynamic Aspects, Organomet., *23*(22), pp.5275-5279], Copyright (2004) American Chemical Society.

Appendix D: Reprinted from Analytica Chimica Acta, Guo, L.F, F. Kooli and M. Garland. A General Method for the Recovery of Pure Powder XRD Patterns from Complex Mixtures Using No *a priori* Information. Application of Band-Target Entropy Minimization (BTEM) to Materials Characterization of Inorganic Mixtures, *517*(1-2), pp.229-236., Copyright (2004), with permission from Elsevier.

Appendix E: Reprinted from Pattern Recognition, Guo, LF. and M. Garland. The use of entropy minimization for the solution of blind source separation problems in image analysis. *39*, pp.1066-1073, Copyright (2006), with permission from Elsevier.

Appendix F: Reproduced with permission from [Guo, L.F., A. Wiesmath, P. Sprenger, M. Garland. Development of 2D Band-Target Entropy Minimization and Application to the Deconvolution of Multicomponent 2D Nuclear Magnetic Resonance Spectra, Anal. Chem., *77*, pp. 1655-1662], Copyright (2005) American Chemical Society.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Catalysis 237 (2006) 67–78
ELSEVIER

Appendix A

JOURNAL OF CATALYSIS
www.elsevier.com/locate/jcat

Experimental evidence for a significant homometallic catalytic binuclear elimination reaction: Linear-quadratic kinetics in the rhodium catalyzed hydroformylation of cyclooctene

Guowei Liu, Chuanzhao Li, Liangfeng Guo, Marc Garland *

Department of Chemical and Biomolecular Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore 119260

Received 14 July 2005; revised 29 September 2005; accepted 29 September 2005
Available online 23 November 2005

Abstract

The hydroformylation of cyclooctene was studied using $Rh_4(CO)_{12}$ as precursor in *n*-hexane solvent in the temperature range 293–308 K and P_T of 4.0–8.0 MPa, using quantitative in situ infrared spectroscopy. During the course of reaction, the degradation of $Rh_4(CO)_{12}$ to the intermediate $RCORh(CO)_4$ was observed, with accompanying formation of cyclooctane carboxaldehyde. The limited conversion of $Rh_4(CO)_{12}$ to $RCORh(CO)_4$ was shown to be equilibrium-controlled. Some ketone was also formed. Spectral deconvolution was performed with band-target entropy minimization (BTEM). The reaction kinetics for product formation, in terms of the observable organometallics, were rate = $k_1[RCORh(CO)_4][CO]^{-1}[H_2] + k_2[RCORh(CO)_4][Rh_4(CO)_{12}]^{0.25}[H_2]^{0.5}[CO]^{Y}$. This can be rewritten as rate = $k_1[RCORh(CO)_4][CO]^{-1} \times [H_2] + k_2[RCORh(CO)_4][HRh(CO)_4][CO]^{X}$. The hydride $HRh(CO)_4$ could be identified by BTEM but not accurately quantified. This unusual linear-quadratic expression in rhodium species represents the kinetic form for a simultaneous interconnected unicycle catalytic mechanism and a homometallic catalytic binuclear elimination reaction (CBER). The second term accounted for ca. 40% of the observed product formation at the mean reaction conditions used in this study. The implications and opportunities presented by homometallic CBER are discussed. In particular, a form of modified homometallic CBER is proposed that will permit greater utilization of the nonlinear kinetics.

Keywords: Catalytic binuclear elimination reaction; Linear-quadratic kinetics; Rhodium-catalyzed hydroformylation; In situ quantitative FTIR spectroscopy

1. Introduction

One of the most widely studied transition metal homogeneous catalyzed reactions [1,2], the hydroformylation reaction provides a versatile route for the synthesis of a vast array of bulk and specialty chemicals [3]. The primary transition metals used are cobalt [4], rhodium [5], and platinum [6].

Because of the versatility of this reaction, a very wide range of unfunctionalized and functionalized alkenes have been used as substrates [7]. In the case of unfunctionalized rhodium-catalyzed hydroformylation, comparisons of activity (reaction rates) have been studied in detail for various classes of unmodified alkene substrates [8–13], and significant variation has been found. In situ studies of the unmodified rhodium-catalyzed hydroformylation of unfunctionalized alkenes has revealed that although the rates are substantially different between classes, the turnover frequencies (TOFs), based on the instantaneous concentrations of the observable acyl intermediate are for the most part very similar [14–17]. This indicates a wide variability in the conversion of precursor to intermediates. In addition, a fair number of anomalies in activity or selectivity patterns have been observed; for example, (1) in the case of ethylene, the corresponding ketone is frequently formed [18,19]; (2) in the case of cyclohexene, unusually high CO pressures are needed to form the acyl complex; (3) for some cycloalkenes, there is very low equilibrium-controlled precursor conversion [16]; and (4) in the case of methylene cyclopropane, ring cleavage occurs [20].

The observation of anomalous substrate behaviors, even within a homologous series of substrates, is not restricted to the hydroformylation reaction alone. Indeed, other well-studied

* Corresponding author. Fax: +65 6 779 1936.
E-mail address: chemvg@nus.edu.sg (M. Garland).

0021-9517/$ – see front matter
doi:10.1016/j.jcat.2005.09.033

catalytic reactions, such as hydrogenation, have numerous documented examples [21]. The primary issue to emphasize is that further study of the origins of these anomalous observations is infrequently pursued. The eventual clarification of the underlying mechanisms behind the anomalous observations of activity and selectivity may provide rich scientific opportunities.

The study of binuclear elimination, and even catalytic binuclear elimination, has been gaining interest over the past few decades. These studies have been driven in part by the need to explain anomalous activity and selectivity patterns. Presently, there are ca. 14 well-defined homometallic stoichiometric binuclear elimination reactions (Table 1). Of course, the existence of homometallic stoichiometric binuclear elimination reactions raises the issue of related catalytic possibilities. Such a mechanism would provide the basis for a catalytic binuclear elimination reaction (CBER).

The cobalt-mediated hydroformylation reaction represents the best example of both stoichiometric and catalytic investigations. The reaction of acyl cobalt tetracarbonyls with cobalt tetracarbonyl hydride was first reported by Heck and Breslow [25], then extensively investigated by the Vesprem group [26–28]. Both groups suggested that a catalytic analog to stoichiometric CBER may be present in addition to a unicyclic topology during cobalt-catalyzed hydroformylations. The possible presence of CBER was further suggested by in situ spectroscopy of the cobalt-catalyzed hydroformylation [41]. In 1983, Mirbach conducted an in situ infrared IR study of the homogeneous cobalt-catalyzed hydroformylation of 1-octene and found that perhaps 4% of the product formation arose from a homometallic CBER [42].

Based on studies of the elimination mechanism of osmium alkyls and osmium hydrides, Norton suggested that binuclear elimination is probably much more common than has been realized, because of the extraordinary ability of metal hydrides to fill vacant coordination sites on other metals [30]. He also argued that binuclear elimination is probably involved to some extent in the cobalt-catalyzed oxo reaction. In his later work on the relative nucleophilicity of metal hydrides, Norton further indicated that binuclear elimination could be the terminal step in catalytic hydroformylation [37].

In addition, in the studies of a surface-tethered silica-supported rhodium hydroformylation of styrene, Collman et al. [43] found a nonlinear rate dependence of the hydroformylation of styrene with the surface concentration of the rhodium catalyst species. Accordingly, these authors discussed the possibility of dinuclear reductive elimination as the site–site interaction [43].

An in situ spectroscopic attempt to identify a homometallic CBER in the unmodified homogeneous rhodium-catalyzed hydroformylation reaction was conducted in 1999 [44]. In that study, cyclohexene was used as a substrate, because the known equilibrium-controlled precursor conversion would result in only ca. 10% conversion of $Rh_4(CO)_{12}$ at 6.0 MPa CO partial pressure. Thus the catalytic system should contain considerable quantities of $HRh(CO)_4$ in equilibrium with $Rh_4(CO)_{12}$ in addition to the $RCORh(CO)_4$ present, and hence the probability of binuclear elimination between $HRh(CO)_4$ and $RCORh(CO)_4$ should be nonnegligible. However, no statistical evidence of a homometallic CBER could be found. It is worth mentioning that no spectroscopic evidence for the presence of observable quantities of $HRh(CO)_4$ could be obtained in that study, most likely due to the lack of appropriate signal processing tools at that time.

The development of band-target entropy minimization (BTEM) [45–50] and associated algebraic tools [51] has made it possible to conduct more detailed in situ studies. Recently we obtained strong evidence for *bimetallic* CBER kinetics in the rhodium-catalyzed hydroformylation of 3,3-dimethyl-but-1-ene and cyclopentene promoted with $HMn(CO)_5$ [52,53]. This prompted renewed interest in anomalous homometallic rhodium hydroformylation. Because the rhodium-catalyzed hydroformylation of cyclooctene is also known to exhibit equilibrium-controlled precursor conversion, we have reinvestigated the rhodium-catalyzed hydroformylation of cyclooctene in an attempt to better understand the kinetics [54]. The present paper reports this effort and our conclusion that a significant homometallic CBER is present.

2. Experimental

2.1. General information

All solution preparations and transfers were carried out under purified argon (99.9995%, Saxol; Singapore) atmosphere using standard Schlenk techniques [55]. The argon was further purified before use by passing it through a deoxy and zeolite column. Purified carbon monoxide (research grade, 99.97%, Saxol; Singapore) and purified hydrogen (99.9995%, Saxol; Singapore) were also further purified through deoxy and zeolite columns before being used in the hydroformylation experiments. Purified nitrogen (99.9995%, Saxol; Singapore) was used to purge the Perkin–Elmer 2000 Fourier transform infrared (FTIR) spectrometer system.

$Rh_4(CO)_{12}$ (98%) was purchased from Strem Chemicals and was used without further purification. The cyclooctene (99.9%, Chemsampco) was dehydrated with CaH_2 before use and stored under argon in a refrigerator. After dehydrating, no other species could be detected by GC (HP6890; HP-FFAP polyethylene glycol TPA capillary column, 100 °C; flame ionization detector, 250 °C).The puriss-quality *n*-hexane (99.6%; Fluka) was distilled from sodium-potassium alloy under argon for ca. 5 h to remove trace water and oxygen.

2.2. Apparatus

In situ kinetic studies were performed in a 1.5-L stainless-steel (SS316) autoclave (P_{max} = 22.5 MPa; Buchi–Uster), connected to a high-pressure IR flow cell. The system is the same as that used in earlier studies [14–17], and a schematic diagram of the experimental setup has been provided previously [17].

2.3. In situ spectroscopic and kinetic studies

A total of 16 kinetic experiments in 5 sets were performed. In each set, one experimental parameter was systematically

Table 1
Presently known homometallic stoichiometric binuclear elimination reactions

No	Reaction	Reference
1	$2HCo(CO)_4 \rightarrow Co_2(CO)_8 + H_2$	[22,23]
2	$2HMn(CO)_5 \rightarrow Mn_2(CO)_{10} + H_2$	[24]
3	$RCOCo(CO)_4 + HCo(CO)_4 \rightarrow Co_2(CO)_8 + RCHO$	[25–27]
4	$RCo(CO)_4 + HCo(CO)_4 \rightarrow Co_2(CO)_8 + RH$	[28]
5	$2Os(CO)_4H_2 \rightarrow H_2Os_2(CO)_8 + H_2$	[29–31]
6	$2Os(CO)_4(H)CH_3 \rightarrow HOs(CO)_4Os(CO)_4CH_3 + CH_4$	[29–31]
7	$Os(CO)_4H_2 + Os(CO)_4(CH_3)_2 \rightarrow HOs(CO)_4Os(CO)_4CH_3 + CH_4$	[29–31]
8	$Os(CO)_4H_2 + Os(CO)_4(H)CH_3 \rightarrow H_2Os_2(CO)_8 + CH_4$	[29–31]
9	$RMn(CO)_5 + HMn(CO)_5 \rightarrow Mn_2(CO)_9 + RCHO$	[32–34]
10	$2HRh(CO)_2(PPh_3)_2 \rightarrow Rh_2(CO)_2(PPh_3)_2 + H_2 + 2CO$	[35]
11	$2HIr(CO)(PPh_3)_3 \rightarrow Ir_2(CO)_2(PPh_3)_4 + H_2 + 2PPh_3$	[36]
12	$EtRe(CO)_5 + HRe(CO)_5 \rightarrow EtCHO + Re_2(CO)_9$	[37]
13	$EtHFe(CO)_4 + CH_3CH_2C(=O)FePPh_3(CO)_3 \rightarrow CH_3CH_2CHO + EtFe_2PPh_3(CO)_7$	[38]
14	$(\eta^5\text{-}C_5H_5)Mo(CO)_3H + (\eta^5\text{-}C_5H_5)Mo(CO)_3R \rightarrow (\eta^5\text{-}C_5H_5)_2Mo_2(CO)_4 + RCHO + CO$	[39,40]

Table 2
Experimental design for the $Rh_4(CO)_{12}$ catalyzed hydroformylation of cyclooctene

Experiment	CO (MPa)	H_2 (MPa)	Cyclooctene (mL)	$Rh_4(CO)_{12}$ (mg)	Temperature (K)
Standard	4.0	2.0	10	100.1	298.0
CO	2.0	2.0	10	99.9	298.0
variation	3.0	2.0	10	99.9	298.0
	5.0	2.0	10	103.2	298.0
H_2 variation	4.0	1.0	10	105.2	298.0
	4.0	3.0	10	104.0	298.0
	4.0	4.0	10	98.2	298.0
Cyclooctene	4.0	2.0	5	100.7	298.0
variation	4.0	2.0	15	100.9	298.0
	4.0	2.0	20	98.8	298.0
$Rh_4(CO)_{12}$	4.0	2.0	10	50.5	298.0
variation	4.0	2.0	10	198.7	298.0
	4.0	2.0	10	255.3	298.0
Temperature	4.0	2.0	10	100.2	293.0
variation	4.0	2.0	10	100.1	303.0
	4.0	2.0	10	101.6	308.0

varied while the remaining variables were kept essentially constant. The detailed experimental design for this study is given in Table 2. The experimental design of the experiments involved 300 mL of solvent and the following intervals: temperature, 293–308 K; P_{H_2}, 1.0–4.0 MPa; P_{CO}, 2.0–5.0 MPa; initial alkene, 5–20 mL; and initial $Rh_4(CO)_{12}$, 50.5–255.3 mg.

All of the experiments were performed in a similar manner. A typical procedure for the standard experiment was as follows. First, background spectra of the IR sample chamber were recorded. Then 150 mL of *n*-hexane was transferred under argon to the autoclave. Under 0.2 MPa CO pressure, IR spectra of the *n*-hexane in the high-pressure cell were recorded. The total system pressure was raised to 4.0 MPa CO, and the stirrer and high-pressure membrane pump were started. After equilibration, IR spectra of the CO/*n*-hexane solution in the high-pressure cell were recorded. A solution of 10 mL of cyclooctene dissolved in 50 mL of *n*-hexane was prepared, transferred to the high-pressure reservoir under argon, pressurized with CO, and then added to the autoclave. After equilibration, IR spectra of the cyclooctene CO/*n*-hexane solution in the high-pressure cell were recorded. A solution of ca. 100 mg of $Rh_4(CO)_{12}$ dissolved in 50 mL of *n*-hexane was prepared, transferred to the high-pressure reservoir under argon, pressurized with CO, and then added to the autoclave. After equilibration, IR spectra of the $Rh_4(CO)_{12}$/cyclooctene/CO/*n*-hexane solution in the high-pressure cell were recorded. After this, 2.0 MPa of hydrogen was added to initiate the synthesis.

The in situ spectra were obtained every 15 min during each 6-h experiment in the range of 1000–2500 cm^{-1} with a resolution of 4 cm^{-1}. A total of 320 spectra were obtained for further spectroscopic and kinetic analyzes.

Two relevant reviews on in situ IR spectroscopic studies of general catalytic systems [56] and in situ IR spectroscopic studies of the hydroformylation reaction [57] have recently appeared.

2.4. Transport considerations

The primary transport issues to consider in a homogeneous catalytic reaction being monitored by in situ spectroscopy are (1) the mixing times in the continuous-stirred tank reactor (CSTR) and recycling system, (2) the rate of gas–liquid mass transfer compared with the rate of reaction, and (3) the extent of reaction occurring outside the CSTR and inside the recycle loop (i.e., the composition difference between outlet and inlet). A review of these issues and the associated calculations have been provided previously [58]. The mixing time in the CSTR and recycling loop was on the order of a few minutes. The experimentally measured overall mass transfer coefficients, K_La, for hydrogen and carbon monoxide into *n*-hexane at 200 rpm were approximately 0.1 and 0.06 s^{-1}, respectively. Taking into account the gas solubility at the mean reaction conditions, the maximum rates of mass transfer were ca. 3×10^{-3} mol fraction/s. Because the maximum observed rate of hydroformylation in this study was ca. 3.5×10^{-7} mol fraction/s, all hydroformylation experiments exhibited product formation rates belonging to the category H of Hatta classifications—infinitely slow reaction compared with gas–liquid mass transfer. Saturation at

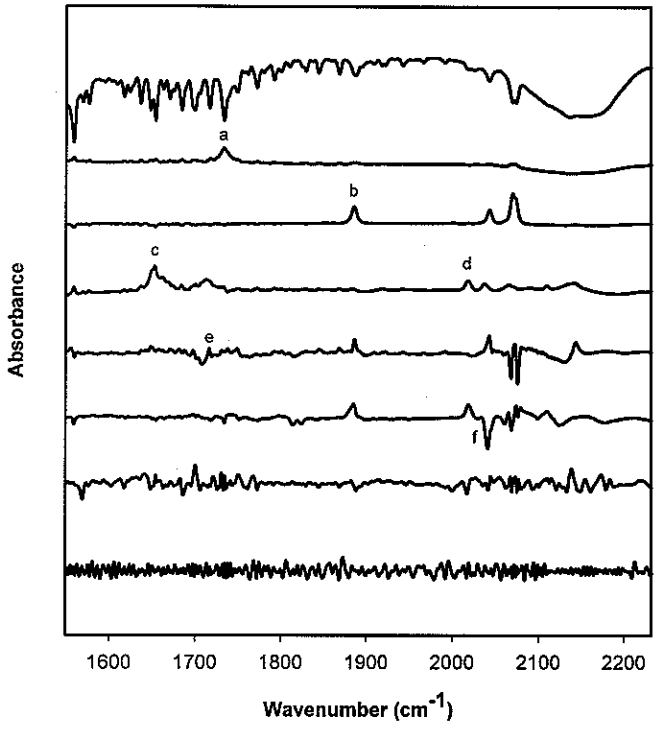


Fig. 1. A singular value decomposition of the in situ spectroscopic data showing the 1st, 2nd, 4th, 7th, 12th, 19th, 50th significant vectors and the 320th vector. The labeled extrema are those which were used to recover the pure organometallic component spectra as well as alkene and aldehyde by BTEM. The reaction conditions are P_{CO} = 2.0–5.0 MPa, P_{H_2} = 1.0–4.0 MPa, cyclooctene = 5–20 mL, $Rh_4(CO)_{12}$ = 50.5–255.3 mg in 300 mL *n*-hexane at 298 K.

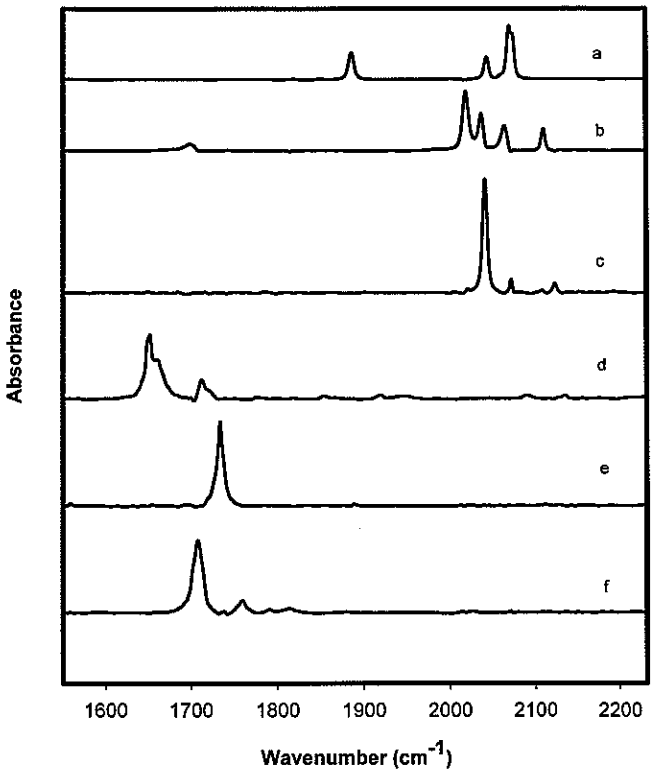


Fig. 2. The recovered pure component spectra of the organic and organometallic species using BTEM: (a) $Rh_4(CO)_{12}$, (b) $RCORh(CO)_4$, (c) $HRh(CO)_4$, (d) cyclooctene, (e) cyclooctane carboxaldehyde, and (f) ketone. The reaction conditions are P_{CO} = 2.0–5.0 MPa, P_{H_2} = 1.0–4.0 MPa, cyclooctene = 5–20 mL, $Rh_4(CO)_{12}$ = 50.5–255.3 mg in 300 mL *n*-hexane at 298 K.

the beginning of a run was achieved on the order of a few minutes. The mixing issues together with the initial gas–liquid mass transfer issue indicate that transport effects influence the first few minutes of reaction. The residence time in the recycle loop was on the order 2 min. The concentrations of the reagents dissolved CO, H_2, and cyclooctene varied by <1% along the recycling loop during this period. Accordingly, the concentrations of these components as well as the organometallics at the spectrometer are only differentially removed from the concentrations within the CSTR, and thus the measurements are truly in situ.

2.5. Computations

The newly developed algorithms of BTEM for spectral deconvolution and algebraic system identification for catalytic reaction modeling were used to analyze the in situ IR spectra [45–51].

3. Results

3.1. Spectroscopic aspects

The 320 in situ FTIR spectra in this experimental study were analyzed over the spectral interval 1550–2230 cm^{-1} with data intervals of 0.2 cm^{-1}. Fig. 1 shows notable vectors from the singular value decomposition of the spectroscopic matrix $A_{320\times3401}$. These vectors contain the chemically important spectral features for the pure component spectra. Fig. 2 shows the correspondingly recovered pure component spectra obtained from the BTEM analysis (solvent hexane, atmospheric moisture, and CO_2 and dissolved CO are omitted).

Fig. 2 shows that the reconstructed pure component spectra are consistent with those obtained in the numerous previous in situ FTIR spectroscopic studies of the unmodified rhodium-catalyzed hydroformylation of alkenes [14–17,44–51]. These studies have shown that the rhodium precursors are transformed under reaction conditions to generate observable quantities of a mononuclear acyl complex, namely $RCORh(CO)_4$, with characteristic features at 1698, 2020, 2039, 2065, and 2111 cm^{-1}. To date, approximately 20 acyl rhodium tetracarbonyl complexes have been observed in situ [20], but evidence for observable quantities of other intermediates has been exceptionally difficult to obtain. The most obvious coordinately saturated 18e-mononuclear species expected under hydroformylation conditions are the hydride $HRh(CO)_4$ [59,60] and the alkyl $RRh(CO)_4$. Spectroscopic evidence of the existence of the former under syngas has developed over the years, and outstanding spectra of both $HRh(CO)_4$ [2002.8(vw), 2041.6(vs), 2071.8(m), 2123.6(vw) cm^{-1}] and $DRh(CO)_4$ were obtained recently [48]. However, analyzes of spectra from active unmodified hydroformylations have not conclusively indicated observable quan-

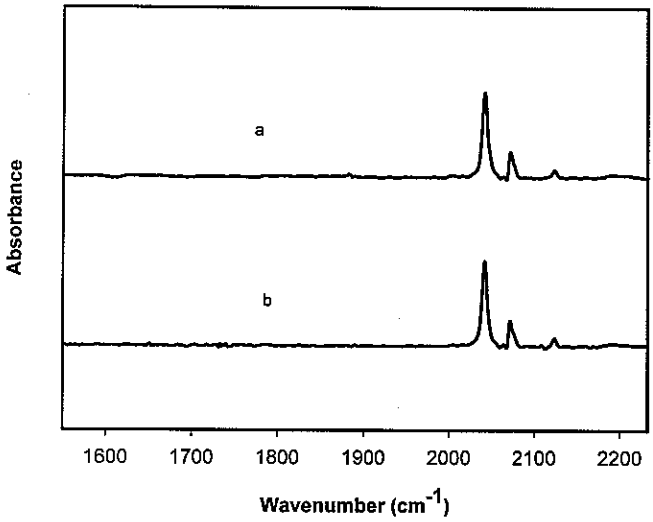


Fig. 3. The target and projected pure component spectra of $HRh(CO)_4$: (a) target vector and (b) projected vector. The reaction conditions are P_{CO} = 2.0–5.0 MPa, P_{H_2} = 1.0–4.0 MPa, Cyclooctene = 5–20 mL, $Rh_4(CO)_{12}$ = 50.5–255.3 mg in 300 mL *n*-hexane at 298 K.

tities of $HRh(CO)_4$ until now. The spectrum of $HRh(CO)_4$ was recoverable in the present study.

The resolved spectrum is a little distorted, although the primary bands at ca. 2042, 2071, and 2123 cm^{-1} could be recovered. Some random signals associated with measurement noise are observed. In addition, target factor analysis (TFA) [61] was performed to provide independent confirmation of $HRh(CO)_4$. A pure component spectrum of $HRh(CO)_4$ obtained from a noncatalytic experiment in this laboratory was used as the target vector and was projected onto 50 vectors of the V^T matrix. Fig. 3 shows the target and projected pure component spectra. These results indeed confirm the presence of the component at very low concentration. The projected vector has a good signal-to-noise ratio of >25:1.

In most investigations of hydroformylation reactions, the corresponding aldehyde is by far the predominant overall organic product, particularly when rhodium is used. However, sometimes ketone and even polyketone formation is nonnegligible [18,19]. In the present experiments, ketone formation occurred as confirmed by the spectral feature at 1720 cm^{-1}.

Table 3 gives the contributions of each component to the spectral absorbance, as well as the total percentage of signal recovery. The contribution of moisture to the total signal in this 1999 experimental dataset was a very significant 27.49%. Our later studies show significantly less moisture in the FTIR spectra. Such a large contribution of moisture, with its sharp and nonstationary bands, puts considerable restrictions on the spectral recovery and total percentage of signal that can be modeled. Despite this difficulty, however, ca. 99% of the signal could be modeled. The total signal associated with the organometallics was only ca. 2.8%.

In the interest of completeness, it should be mentioned that the presence of trace amounts of conjugated 1,3-cyclooctadiene in the cyclooctene substrate cannot be excluded. (Conjugated dienes are common impurities in cycloalkenes.) As shown previously [62], the presence of conjugated dienes in rhodium-catalyzed hydroformylations often results in a new characteristic metal–carbonyl vibration at ca. 1992 cm^{-1}. A weak local extremum at 1992 cm^{-1} was observed in the V^T vectors of this study, but pure component spectral recovery was unsuccessful.

Table 3
Percentage of integrated absorbance for each component compared to the total original experimental data

Component	Integrated intensity of each component (%)
Moisture	27.49
n-hexane	33.65
Dissolved CO	31.58
$Rh_4(CO)_{12}$	2.26
$C_8H_{15}CHO$	2.09
Cyclooctene	0.89
$RCORh(CO)_4$	0.50
Ketone	0.44
$HRh(CO)_4$	0.08
$Rh_6(CO)_{16}$	[a]
Total	98.98

[a] Negligible contribution. Pure component spectrum not recoverable but presence verified by TTFA.

3.2. *Representative experiments and kinetics*

The 16 experiments provided 5 distinct subsets, which were then analyzed for precursor conversion and organic product formation. In what follows, to save space, we summarize the results of the equilibrium conversion and present only the rhodium series in their entirety. The remaining results can be found in the supporting information.

3.2.1. *The rhodium variation experiments*

The set of experiments associated with the variations in rhodium were all conducted with the initial conditions of 4.0 MPa CO, 2.0 MPa hydrogen, and 10 mL cyclooctene in 300 mL *n*-hexane at 298 K. The initial amounts of rhodium precursor $Rh_4(CO)_{12}$ were 52.6, 102.1, 199.8, and 253.3 mg.

3.2.1.1. Precursor conversion Fig. 4 shows the mole fractions of the catalyst precursor and the one and only quantifiable mononuclear intermediate, $RCORh(CO)_4$, as a function of time. The time series show very little scatter in the mole fraction data. As seen in this figure, the precursor and $RCORh(CO)_4$ rapidly achieved a more-or-less steady-state situation in ca. 30 min. The percent conversion for these four experiments was ca. 20.9, 12.0, 7.6, and 6.3%. Fig. 4 indicates that the formation of $RCORh(CO)_4$ and disappearance of $Rh_4(CO)_{12}$ are consistent with a good mass balance on rhodium. A small increase in the concentration of $Rh_4(CO)_{12}$ and a small decrease in the concentration of $RCORh(CO)_4$ occurred over time due to consumption of the substrate.

3.2.1.2. Aldehyde Fig. 5 shows the mole fractions of the aldehyde product as a function of time. Again, the time series show very little scatter in the mole fraction data. A small induction period occurred in the first 30 min of each series. The rates

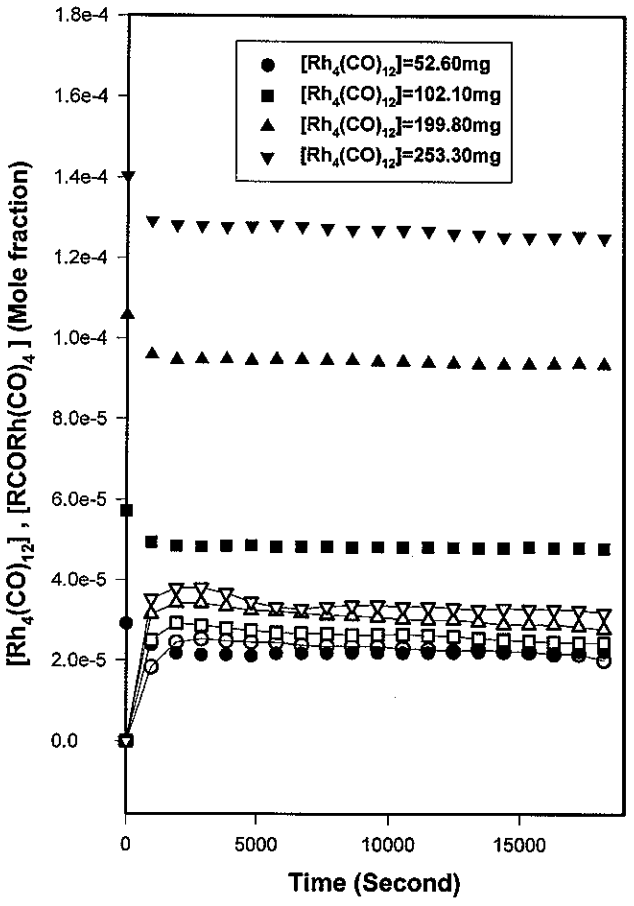


Fig. 4. Time series data for the primary quantifiable organorhodium species present during the rhodium variation set of experiments. The data points with line represent $RCORh(CO)_4$, while those without line represent $Rh_4(CO)_{12}$. The remaining reaction conditions are held constant at $V_{alkene} = 10$ mL, $P_{CO} = 4.0$ MPa, $P_{H_2} = 2.0$ MPa, at 298 K.

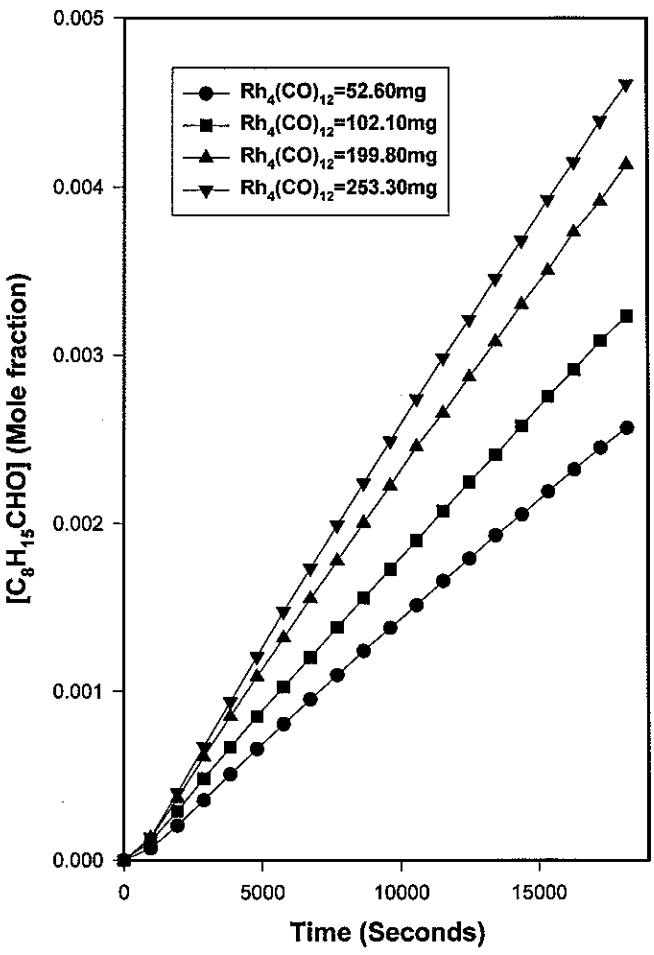


Fig. 5. Time series data for the organic product $C_8H_{15}CHO$ present during the rhodium variation set of experiments. The remaining reaction conditions are held constant at $V_{alkene} = 10$ mL, $P_{CO} = 4.0$ MPa, $P_{H_2} = 2.0$ MPa, at 298 K.

of reaction for these four experiments were ca. 1.4×10^{-7}, 1.8×10^{-7}, 2.3×10^{-7}, and 2.6×10^{-7} mol fraction/s.

3.2.1.3. Turnover frequency Fig. 6 shows the TOFs for aldehyde formation based on the instantaneous mole fractions of $RCORh(CO)_4$ in each experiment as a function of $Rh_4(CO)_{12}$ loading and as a function of time. These time series data demonstrate increased scatter due to the fact that two independent experimental observations are needed for each data point. Regression of the data for each run provided the values TOF (52.60 mg) = $(6.20 \pm 0.16) \times 10^{-3}$ s^{-1}, TOF (102.10 mg) = $(6.85 \pm 0.10) \times 10^{-3}$ s^{-1}, TOF (199.80 mg) = $(7.50 \pm 0.09) \times 10^{-3}$ s^{-1}, and TOF (253.30 mg) = $(7.93 \pm 0.57) \times 10^{-3}$ s^{-1}, where the errors were listed as twice the standard deviation (i.e., 95% confidence limit). Clearly, TOF is not a constant; it increases with increased loading of $Rh_4(CO)_{12}$.

Previous studies of unmodified rhodium-catalyzed hydroformylations of 3,3-dimethyl-but-1-ene have repeatedly shown that TOF is a function of temperature, hydrogen pressure, and carbon monoxide pressure [15,52]. The present results clearly show that TOF is also dependent on rhodium loading. This is not expected for a unicycle catalytic mechanism, which has rates linear in the total concentration of intermediates. The observed nonlinear aspects of TOF indicate the possible contribution of a CBER to the final product formation.

3.3. Analysis of precursor conversion

The precursor $Rh_4(CO)_{12}$ should be equilibrated with the hydride species $HRh(CO)_4$, and this species should be equilibrated with its coordinately unsaturated hydride, $HRh(CO)_3$. As derived previously [16], a steady-state assumption can be imposed on the concentration of $HRh(CO)_3$ in terms of its disappearance due to alkene coordination and its formation due to hydrogenolysis of the acyl complex. This results in an expression for *equilibrium-controlled* precursor formation in unmodified rhodium-catalyzed hydroformylations of the form in Eq. (1), where the subscript denotes steady-state concentrations. In the hydroformylation of cyclohexene, the experimentally determined relationship was that of Eq. (2) [16], where $[Rh_4(CO)_{12}]$, [CO], $[H_2]$, and [R′] represent the mole fractions of precursors, dissolved gases, and substrate used:

$$[RCORh(CO)_4]_{SS} = \Phi[Rh_4(CO)_{12}]_{SS}^{0.25}[CO][H_2]^{-0.5}[R']^1, \quad (1)$$

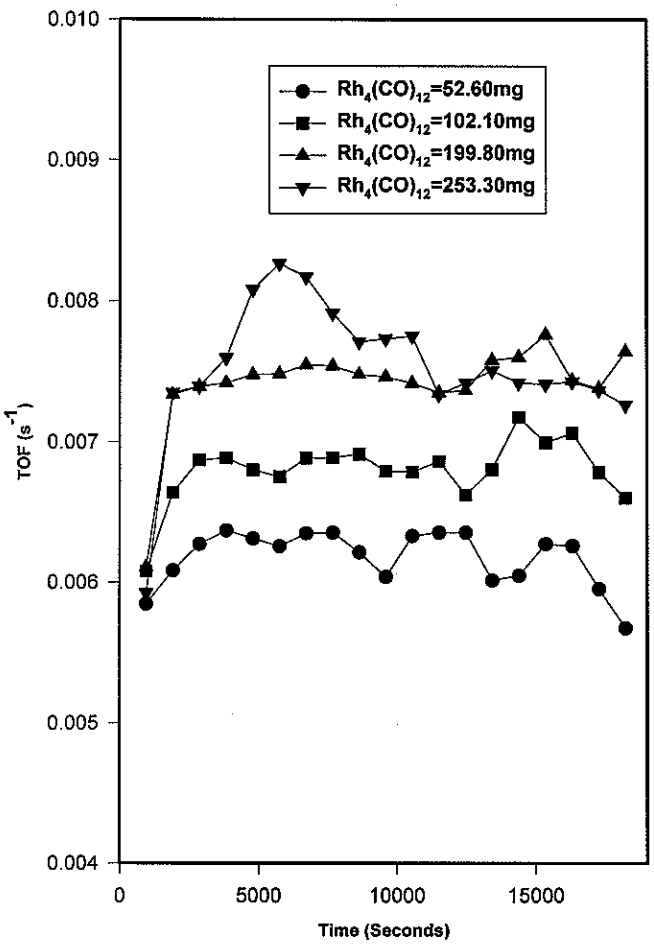


Fig. 6. The effects of $Rh_4(CO)_{12}$ loadings on the turnover frequencies (TOF) for aldehyde formation. The remaining reaction conditions were held constant at $V_{alkene} = 10$ mL, $P_{CO} = 4.0$ MPa, $P_{H_2} = 2.0$ MPa, at 298 K.

$$[RCORh(CO)_4]_{SS} = \Phi[Rh_4(CO)_{12}]_{SS}^{0.3}[CO]^{1.1}[H_2]^{-0.3}[R']^{0.9}. \quad (2)$$

In the present experimental study for the hydroformylation of cyclooctene, a multilinear regression was performed for the pseudo-steady-state concentration of the acyl rhodium tetracarbonyl in terms of the remaining reactants in the system (with all experimental runs used). The obtained relationship is given in Eq. (3). The values of the exponents have rather acceptable statistical bounds, so the data and regression appear to be rather good. The exponents are quite similar to those previously obtained experimentally in the case of cyclohexene and consistent with the hypothesis of equilibrium-controlled precursor conversion. In particular, the nuclearity difference between precursor and intermediate is clearly shown by the exponent of 0.23 for $Rh_4(CO)_{12}$:

$$[RCORh(CO)_4]_{SS} = \Phi[Rh_4(CO)_{12}]_{SS}^{0.23\pm0.1}[CO]^{1.4\pm0.2} \times [H_2]^{-0.3\pm0.2}[R']^{0.7\pm0.3}. \quad (3)$$

3.4. Analysis of catalytic kinetics

TOF analysis of the experimental data associated with the effect of initial $Rh_4(CO)_{12}$ (Fig. 6) suggests that a possible CBER exists in this system. In principle, the bimolecular reaction of $RCORh(CO)_3$ with molecular hydrogen is not the only mechanism available for the hydrogenolysis of $RCORh(CO)_4$. The possibility also exists that some aldehyde formation occurs via a bimolecular elimination reaction, similar to that found for the unmodified cobalt system under stoichiometric conditions [25], which is shown by

$$HCo(CO)_4 + RCOCo(CO)_{3,4} \rightarrow RCHO + Co_2(CO)_{7,8}. \quad (4)$$

Under the experimental condition, there may exist several active species that could be involved in product formation, including $RCORh(CO)_3$, $RCORh(CO)_4$, $HRh(CO)_3$, $HRh(CO)_4$. In principle, both hydride rhodium species can react with both acyl rhodium species to yield aldehydes,

$$HRh(CO)_{3,4} + RCORh(CO)_{3,4} \rightarrow RCHO + Rh_2(CO)_{6,7,8}. \quad (5)$$

We now state the working hypothesis that the experimental system consists of a simultaneous unicyclic reaction mechanism and a homometallic CBER. Accordingly, the rate of hydroformylation should take the general form

$$\text{rate} = k_1[RCORh(CO)_4][CO]^{-1}[H_2] + k_2[RCORh(CO)_4][HRh(CO)_4][CO]^X. \quad (6)$$

Because it is difficult to quantify the $HRh(CO)_4$ in the present study due to its weak signal, the following equation is used to replace the concentration of saturated $HRh(CO)_4$ with $[Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[CO]^1$:

$$Rh_4(CO)_{12} + 2H_2 + 4CO \leftrightarrow 4HRh(CO)_4. \quad (7)$$

Thus the new expression for the rate of hydroformylation in terms of observable organometallics becomes

$$\text{rate} = k_1[RCORh(CO)_4][CO]^{-1}[H_2] + k_2[RCORh(CO)_4][Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[CO]^Y. \quad (8)$$

Dividing through by the acyl rhodium tetracarbonyl concentration provides the following working expression for the apparent TOF of aldehyde formation:

$$TOF_{total} = k_1[CO]^{-1}[H_2] + k_2[Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[CO]^Y. \quad (9)$$

Again, the first term presents the classic unicyclic hydroformylation mechanism, and the second term is the contribution from the CBER. All of the experimental data at 298 K were used to regress the foregoing equation to provide the values

$$k_1 = (1.600 \pm 0.109) \times 10^{-2}\ s^{-1},$$
$$k_2 = (5.603 \pm 1.143) \times 10^{-2}\ s^{-1},$$

and

$$y = -(6.275 \pm 0.732) \times 10^{-1}.$$

To further evaluate the contribution of a CBER, we prepared a plot of the TOFs versus $[Rh_4(CO)_{12}]^{1/4}$ from the $Rh_4(CO)_{12}$ series of experiments (Fig. 7). Regression of all of the rhodium series data provided the following result:

$$TOF = 0.00360 \pm 0.00030\ s^{-1} + 0.03806 \pm 0.00328\ s^{-1} \times [Rh_4(CO)_{12}]^{0.25}, \quad (10)$$

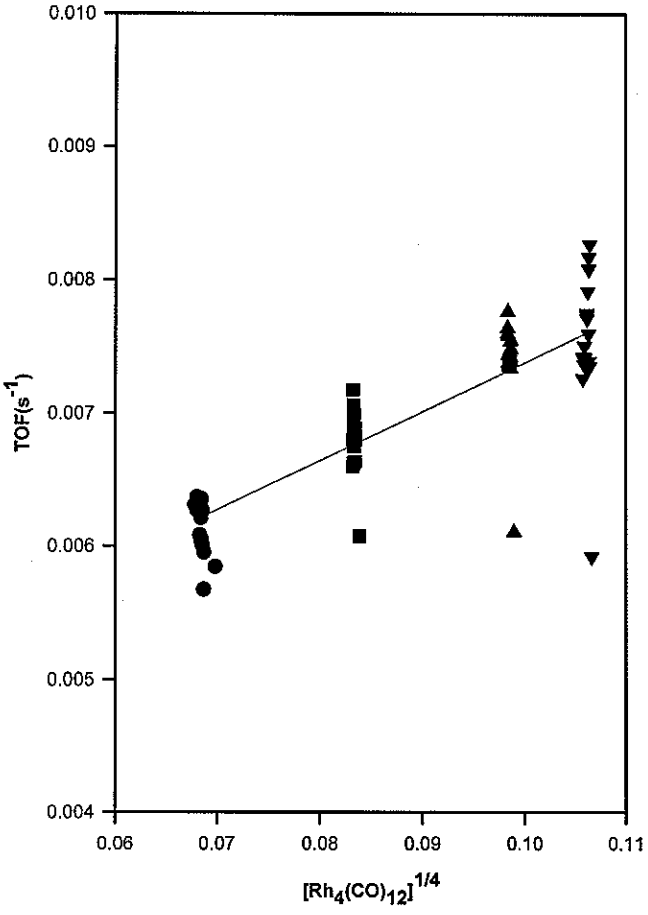


Fig. 7. The dependence of the instantaneous values of TOF on the instantaneous values of observable $Rh_4(CO)_{12}$ for the set of experiments with varying nominal rhodium concentration. The remaining reaction conditions are held constant at $V_{alkene} = 10$ mL, $P_{CO} = 4.0$ MPa, $P_{H_2} = 2.0$ MPa at 298 K.

where the errors are presented as twice the standard deviation. This indicates that the contributions of catalytic binuclear elimination are 36.7, 42.7, 47.7, and 48.6% at initial $Rh_4(CO)_{12}$ loadings of 52.60, 102.10, 199.80, and 253.30 mg, respectively. The lowest points in Fig. 7 represent the first measurements in each run and, as such, they represent outliers.

4. Discussion

4.1. Spectroscopic considerations

The in situ spectroscopic measurements and analyzes result in relatively good pure component spectra and concentration profiles for the organic substrate cyclooctene, the primary organic product cyclooctane carboxaldehyde, the organic ketone side product, the organometallic precursor $Rh_4(CO)_{12}$, and the primary organometallic intermediate $RCORh(CO)_4$. In addition, using BTEM and TFA made it possible to confirm the presence of observable quantities of $HRh(CO)_4$ during catalysis, but the signals were too weak to obtain accurate concentration profiles as a function of time. The unusually high atmospheric moisture content in these spectra was a significant factor contributing to limitations in both pure component spectral reconstructions and the concentration profiles.

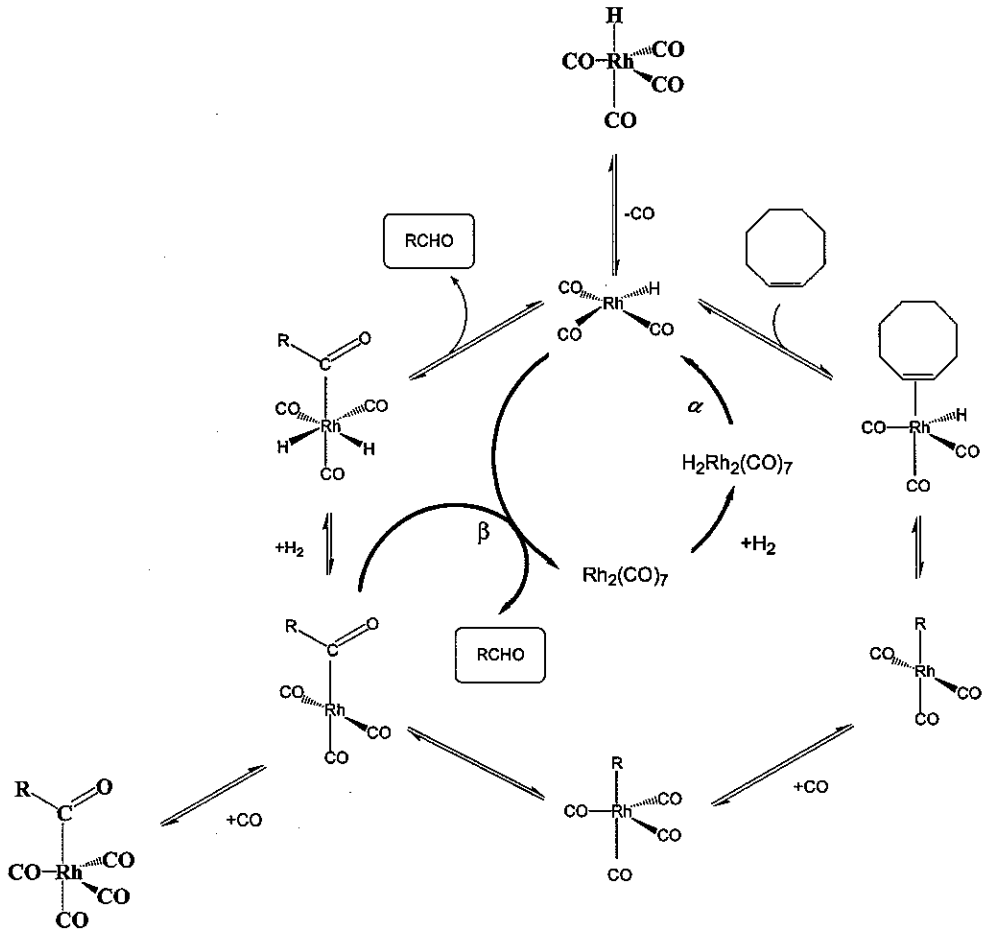
4.2. Equilibrated $HRh(CO)_4$

The existence of a rapidly obtainable equilibrium between observable quantities of a rhodium carbonyl complex $[Rh_4(CO)_{12}]$ and observable quantities of $HRh(CO)_4$ at room temperature under noncatalytic experiments and relatively moderate syngas pressures has been firmly established [48]. The catalytic case is more complex. Marko et al. [63] observed that the rate of hydroformylation of cyclohexene starting with $Rh_6(CO)_{16}$ is proportional to the 1/6th power of the initial carbonyl concentration, that is, $[Rh_6(CO)_{16}]^{1/6}$. These data suggest an equilibrium between the hexanuclear rhodium carbonyl complex precursor and a mononuclear rhodium hydride during catalysis. Our work starting with $Rh_4(CO)_{12}$ as precursor and cycloalkenes as substrates showed that equilibrium-controlled precursor conversion occurs among the observable tetranuclear precursor, an unobservable mononuclear hydride, and the observable mononuclear rhodium acyl carbonyl [16]. The rate of cyclohexene hydroformylation starting with $Rh_4(CO)_{12}$ was proportional to the 1/4th power of the initial carbonyl concentration, that is, $[Rh_4(CO)_{12}]^{1/4}$. This was then shown to be a more general phenomenon, because equilibrium precursor conversion was also seen with other cyclic alkenes, including cyclooctene [20]. Rhodium carbonyl hydride was not observable in either of the aforementioned spectroscopic studies.

In the present study, the development of BTEM made it possible to detect $HRh(CO)_4$. Although equilibrium-controlled precursor conversion is seen in this system, and although $HRh(CO)_4$ must be in equilibrium exchange with the precursor, it was not possible to experimentally quantify the equilibrium. Again this was due to the weak signals involved and, more importantly, the unusually high atmospheric moisture content in the spectra and the resulting difficulties in obtaining good concentration profiles. Accordingly, it was necessary to model the equilibrium concentrations of $HRh(CO)_4$ in terms of the more accurate measurements of the instantaneous concentrations of $Rh_4(CO)_{12}$.

4.3. Kinetics and evidence for homometallic CBER

The classic unicyclic unmodified rhodium catalytic cycle must exist during the hydroformylation of cyclooctene as it does during the hydroformylations of other alkenes that have been modeled successfully [14–16]. Accordingly, the observable rate of hydroformylation must have a linear term in the total concentration of rhodium intermediates. Indeed, this is experimentally verified by the term $k_1[RCORh(CO)_4][CO]^{-1}[H_2]$. However, a second catalytic mechanism is operating simultaneously, arising from the attack of $HRh(CO)_4$ on the primary observable rhodium intermediate $RCORh(CO)_4$ [or, more appropriately, the coordinately unsaturated species $RCORh(CO)_3$]. This generates a second term in the rate expression involving the product $[RCORh(CO)_4][HRh(CO)_4]$. This term is quadratic



Scheme 1. Proposed reaction mechanism for the simultaneous interconnected unicyclic and homometallic CBER hydroformylation reactions. The species $HRh(CO)_4$ and $RCORh(CO)_4$, represented in bold type, are the observable organometallics under reaction conditions.

in rhodium. This second term accounts for the contribution of the homometallic CBER.

The topology of this interconnected catalytic rhodium system, consisting of both a unicyclic mechanism and homometallic CBER involving rhodium alone, and giving rise to linear-quadratic kinetics, is shown in Scheme 1.

It is possible, perhaps even probable, that the binuclear mechanism exists in many (or most) other hydroformylations of alkenes, including very reactive α-olefins. However, the experimentally determined kinetics probably will not indicate the presence of a statistically significant quadratic term in most of these systems due to the ultra-low steady-state concentrations of $HRh(CO)_4$ and hence low probability for CBER. In this study with cyclooctene, the quadratic term is statistically verifiable. It arises in large part due to the relatively high concentrations of $HRh(CO)_4$ present throughout the entire reaction period due to equilibrated conversion with the large pool of $Rh_4(CO)_{12}$ available.

4.4. Clarification of equilibrium-controlled precursor conversion

Equilibrium-controlled precursor conversion was first experimentally observed and successfully modeled for the case of $Rh_4(CO)_{12}$ as precursor in the hydroformylation of cyclohexene [16]. The hydroformylation kinetics are consistent only with a unicyclic catalytic reaction topology. Under this working assumption, a pseudo-steady-state hypothesis was posited for the concentration of $HRh(CO)_3$, and Eq. (3) was derived. The crucial equation for the pseudo-steady-state hypothesis is

$$\begin{aligned} &d[HRh(CO)_3]/dt \approx 0 \\ &= k_i[RCORh(CO)_4][CO]^{-1}[H_2] \\ &\quad - k_{ii}[Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[\text{alkene}]. \end{aligned} \tag{11}$$

In the present contribution, the system has simultaneously one unicylic and one homo-bimetallic CBER reaction topology. Under this circumstance, there exists an extra term in the pseudo steady state hypothesis, namely that arising from the CBER contribution. The resulting equation is

$$\begin{aligned} &d[HRh(CO)_3]/dt \approx 0 \\ &= k_i[RCORh(CO)_4][CO]^{-1}[H_2] \\ &\quad - k_{ii}[Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[\text{alkene}] \\ &\quad + k_{iii}[RCORh(CO)_4][Rh_4(CO)_{12}]^{1/4}[H_2]^{1/2}[CO]^{-1}. \end{aligned} \tag{12}$$

Appendix A

Table 4
Comparison of unicyclic TOF for previous rhodium hydroformylations

Substrate	Metals used	TOF_{uni} (min^{-1})	Reaction topologies present	Reference
3,3-Dimethyl-but-1-ene	Rh	0.11[a]	Unicyclic	[15]
Cyclohexene	Rh	0.135[b]	Unicyclic	[16]
Styrene	Rh	0.056, 0.076[c]	Unicyclic	[17]
3,3-Dimethyl-but-1-ene	Rh, Mn	0.079[d]	Unicyclic and hetero-bimetallic CBER	[43]
Cyclopentene	Rh, Mn	0.12[e]	Unicyclic and hetero-bimetallic CBER	[44]
Cyclooctene	Rh	0.21[f]	Unicyclic	[20]

[a] At 293 K, 2.0 MPa CO, 2.0 MPà H_2 in *n*-hexane.
[b] At 293 K, 6.0 MPa CO, 2.0 MPa H_2 in *n*-hexane.
[c] Minor and major region-selective cycles at 298 K, 5.0 MPa CO, 0.5 MPa H_2 in *n*-hexane.
[d] At 298 K, 2.0 MPa CO, 1.0 MPa H_2 in *n*-hexane.
[e] At 289.7 K, 2.0 MPa CO, 2.0 MPa H_2 in *n*-hexane.
[f] At 298 K, 4.0 MPa CO, 2.0 MPa H_2 in *n*-hexane.

Therefore, in the case of a mixed unicyclic and CBER topology, the corrected expression for equilibrium precursor control is given by Eq. (13). Comparison with Eq. (1), which is strictly valid only for a unicyclic topology, shows that the anticipated observable orders for CO and alkene should remain the same, but that the orders of $Rh_4(CO)_{12}$ and H_2 shift. As the CBER hydroformylation contribution becomes very large compared with the unicyclic hydroformylation contribution, Eq. (13) becomes proportional to just the product [alkene][CO]. In the present set of experiments, the contribution of the CBER is at most only ca. 40% of the total aldehyde formation:

$$[\mathrm{RCORh(CO)_4}]_{SS} = k_{ii}[\mathrm{Rh_4(CO)_{12}}]_{SS}^{1/4}[\mathrm{H_2}]^{-1/2}[\text{alkene}][\mathrm{CO}] \Big/ \left(k_i + k_{iii}[\mathrm{Rh_4(CO)_{12}}]_{SS}^{1/4}[\mathrm{H_2}]^{-1/2}\right). \quad (13)$$

4.5. Comparison of unicyclic TOF to previous rhodium hydroformylations

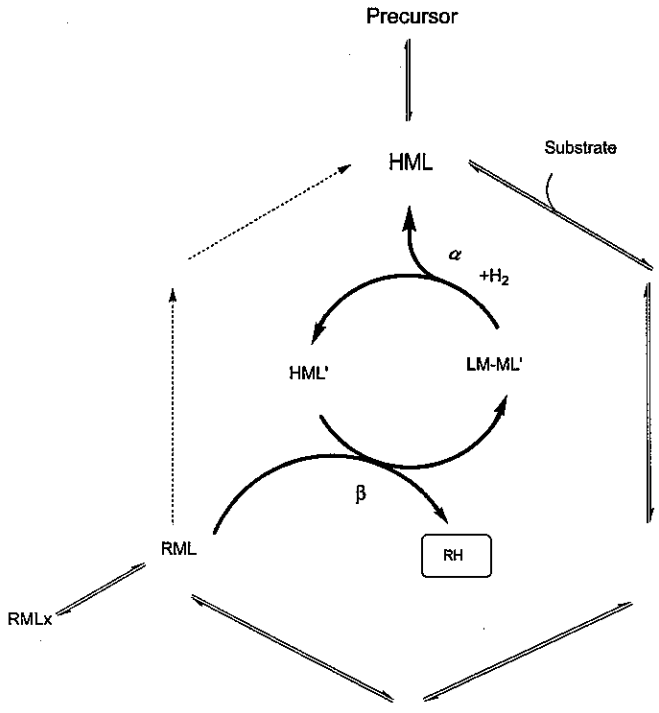
The detailed kinetics of a few unmodified rhodium hydroformylation reactions (i.e., involving full rate expressions based on observable intermediates) have been determined using simultaneous in situ spectroscopic measurements. Table 4 lists the unicyclic TOFs determined from these studies. As Table 4 shows, the unicycle TOFs vary over a narrow range, regardless of the substrate used. In all of these studies, the functional form of the unicyclic TOF was the same, namely $[CO]^{-1}[H_2][\text{alkene}]^0$. In a previous comparative study with rhodium at ca. 6×10^{-5} mol fraction, involving homologous series of alkenes (i.e., cycloalkene, terminal linear alkenes, internal terminal alkenes, and methylene cycloalkanes), it was observed that all of the observed TOFs were ca. 0.04–0.20 min^{-1} measured at 293 K, 2.0 MPa CO, and 2.0 MPa H_2 in *n*-hexane. In addition, cyclooctene exhibited the highest TOF, at ca. 0.20 min^{-1} [20]. The present contribution is consistent with an elevated TOF for cyclooctene.

The issue of parametric sensitivity must be raised. First, regression of the data given in Fig. 7 assumes an exponent equal to 0.25 for $Rh_4(CO)_{12}$ so that good bounds are obtained for the rate coefficients. But this might impose a small bias, because the value of the unicyclic TOF is highly dependent on the exact value of the exponent. If the regression were performed once again with an exponent equal to 0.2, then we would obtain a TOF_{uni} equal to 0.0027 ± 0.0003 s^{-1} (i.e., 0.16 min^{-1}), which is more consistent with other systems. This reemphasizes the foregoing argument about parametric sensitivity. Finally, an extensive comparison between the TOF in this study and that obtained previously [20] is unnecessary, because the latter value of TOF was only a rough first approximation (determined from a single experimental run).

4.6. Other types of systems exhibiting nonlinear or quadratic terms

It is important to mention that a few other homogeneous catalytic systems are known to exhibit rates of reaction that are nonlinear or even quadratic in metal loading. The most studied systems involve ring-opening reactions of epoxides. For example, the asymmetric ring opening of epoxides with trimethyl azide catalyzed by $(Salen)Cr^{III}$ complexes exhibited rates that are second order in chromium [64]. The proposed catalytic mechanism is in fact a monometallic CBER, because it involves mononuclear species and dinuclear species, in which the key step is the intermolecular reaction of two chromium complexes with Salen, epoxide and azide ligands. As far as we know, no detailed kinetic and mechanistic follow-up study has been published to date.

Detailed kinetic and mechanistic studies of related ring-opening reactions of epoxides using zirconium complexes and zinc complexes have appeared, however. Zirconium complexes of the *C*3-symmetric ligand (+)-(*S*,*S*,*S*)-triisopropanolamine as precursors have been shown to promote the stereoselective reaction of cyclohexene oxide and trimethyl azide, with a reaction order of ca. 0.5 in total zirconium observed [65]. The mechanism of this reaction is believed to involve a catalytic cycle with exclusively dimeric zirconium intermediates. In the case of zinc-diiminate-catalyzed copolymerization of cyclohexene oxide and CO_2, a reaction order of 1.0–1.8 in zinc was observed [66]. The mechanism of this reaction is believed to involve a catalytic cycle with exclusively dimeric zinc intermediates, but the possible existence of a simultaneous catalytic cycle with exclusively mononuclear zinc intermediates was also considered.



Scheme 2. Proposed reaction topology for homometallic CBER with purely quadratic kinetics.

4.7. Metal utilization and synthetic efficiency

In the present contribution, the quadratic effects to the overall rate of product formation increases only slowly as the nominal metal concentration is increased. This is due to the fact the precursor is tetranuclear and its conversion is equilibrium-controlled. Such a situation is particularly disadvantageous. Systems with mononuclear precursors controlled by equilibrium conversion will show significantly more pronounced rate increases as a function of total metal loading.

Increasing the nominal metal loadings in small-volume syntheses provides a means of using the metal complexes in a far more efficient manner. Practical reaction engineering limits to such endeavors would soon be encountered as loading is increased, including (i) the solubility limits of the precursor/intermediates under reaction conditions, (ii) possible gas–liquid mass transfer limitations, and (iii) the limited heat transfer capacity of the reactor configuration. As the nominal metal loading increased, the contribution of a unicyclic mechanism would decrease.

4.8. Generating a homometallic CBER with purely quadratic kinetics

It is interesting to consider the possibilities and prerequisites for a homometallic CBER that would exhibit purely quadratic kinetics in organometallic species. Restricting ourselves to catalytic reactions involving hydride species, we can image a situation similar to that shown in Scheme 1. As shown in Scheme 2, the metal, M, is modified with a ligand, L, and this modified hydride, HML, can react with substrate. However, further molecular hydrogen activation on RML proves difficult. At the same time, another hydride species, HML′, with ligand L′, is also present in solution, but this species does not readily undergo reaction with substrate. The homometallic CBER would dominate the catalytic kinetics, because step β dominates, and a higher-order utilization of the metal would be achieved. (It is implicitly understood that the hydrogen activation [step α] must be efficient.) To avoid ligand exchange, L, L′, or both may be (or should be) nondissociating.

A qualifying statement needs to be made regarding the generalized kinetics of Schemes 1 and 2. The kinetic polynomial governing a homometallic CBER would by itself have a linear-quadratic form. The quadratic term controls the kinetics when the mononuclear species are predominant and the dinuclear species are minor, and thus the rate-determining step is bimolecular elimination. At very high metal loadings and low rates α, a shift in species distribution is possible to predominantly dinuclear species with mononuclear species as minor species. In this limit, hydrogen activation is rate-controlling, and the linear kinetic term dominates.

5. Conclusion

The present study has identified a catalytic system that exhibits linear-quadratic kinetics. These linear quadratic kinetics arise from a simultaneous unicyclic mechanism and a homometallic CBER. The homometallic CBER contributes a significant amount (40%) of the reaction product. Identifying the mechanistic reasons for system activity required detailed and quantitative in situ spectroscopic measurements. The implications of CBER include a rationale for nonlinear kinetic effects that differs from those in many previously proposed mechanisms. This study has also provided further insight into equilibrium-controlled precursor conversion.

Acknowledgments

Financial support for this experimental research was provided by the Academic Research Fund of the National University of Singapore (NUS). Research scholarships for GL and LG were provided by the Graduate School of Engineering (NUS). CL thanks Singapore Millennium Foundation (SMF) for a postdoctoral fellowship. The authors thank Ferenc Ungvary, Jack Norton, and Istvan Kovacs for the valuable discussions.

Supporting information

Plots of $RCORh(CO)_4$ versus time, RCHO versus time, and TOF versus time for the experimental series involving CO variation, hydrogen variation, cyclooctene, and temperature are available free of charge at DOI:10.1016/j.jcat.2005.09.033.

References

[1] C.D. Frohning, C.W. Kohlpaintner, in: B. Cornils, W.A. Herrmann (Eds.), Applied Homogeneous Catalysis with Organometallic Compounds: A Comprehensive Handbook, vol. 1, Wiley–VCH, Weinheim, 1996, chap. 2.

[2] P.C.J. Kamer, J.N.H. Peek, P.W.N.M. van Leeuwen, in: B. Heaton (Ed.), Mechanism in Homogeneous Catalysis, Wiley–VCH, Weinheim, 2005, p. 231.
[3] J. Falbe, New Syntheses with Carbon Monoxide, Springer, New York, 1980.
[4] H. Adkins, G. Kresk, J. Am. Chem. Soc. 70 (1948) 383.
[5] G. Schiller, Ger. Pat. 953,605, 1956.
[6] E.L. Jenner, R.V. Lindsey, US Patent 2,876,254, 1959.
[7] P. Pino, F. Piacenti, M. Bianchi, in: L. Wender, P. Pino (Eds.), Org. Synth. Met. Carbonyls, vol. 2, Wiley, New York, 1977, pp. 233–296.
[8] J. Palagyi, G. Palyi, L. Marko, J. Organomet. Chem. 14 (1968) 238.
[9] G. Csontos, B. Heil, L. Marko, Ann. N.Y. Acad. Sci. 239 (1974) 47.
[10] A. Sisak, F. Ungvary, L. Marko, Organometallics 2 (1983) 1244.
[11] L. Marko, F. Ungvary, J. Organomet. Chem. 432 (1992) 1.
[12] F. Ungvary, Coord. Chem. Rev. 170 (1998) 245.
[13] F. Ungvary, Coord. Chem. Rev. 218 (2001) 1.
[14] M. Garland, G. Bor, Inorg. Chem. 28 (1989) 410.
[15] M. Garland, P. Pino, Organometallics 10 (1990) 1693.
[16] C. Fyhr, M. Garland, Organometallics 12 (1993) 1753.
[17] J. Feng, M. Garland, Organometallics 18 (1999) 417.
[18] G. Consiglio, B. Studer, F. Oldani, P. Pino, J. Mol. Catal. 58 (1990) L9.
[19] C. Li, L. Guo, M. Garland, Organometallics 23 (2004) 2201.
[20] G. Liu, R. Volken, M. Garland, Organometallics 17 (1999) 3429.
[21] J.A. Osborn, F. Ardine, J. Young, G. Wilkinson, J. Chem. Soc. Am. 12 (1966) 1711.
[22] K.H. von Brandes, H.B. Jonassen, Z. Anorg. Allg. Chem. 343 (1966) 215.
[23] F. Ungvary, L. Marko, J. Organomet. Chem. 20 (1969) 205.
[24] B.H. Byers, T.L. Brown, J. Am. Chem. Soc. 99 (1977) 2528.
[25] D.S. Breslow, R.F. Heck, Chem. Ind. (London) (1960) 467.
[26] F. Ungvary, L. Marko, Organometallics 2 (1983) 1608.
[27] I. Kovacs, F. Ungvary, L. Marko, Organometallics 5 (1986) 209–215.
[28] C. Hoff, F. Ungvary, R.B. King, L. Marko, J. Am. Chem. Soc. 107 (1985) 666.
[29] J. Norton, W. Carter, J. Kelland, S. Okrasinski, Adv. Chem. Ser. 167 (1978) 170.
[30] J. Norton, Acc. Chem. Res. 12 (1979) 139.
[31] R.T. Edidin, K. Hennessy, A. Moody, S. Okrasinski, J. Norton, New J. Chem. 12 (1988) 475.
[32] M.J. Nappa, R. Santi, S. Diefenbach, J. Halpern, J. Am. Chem. Soc. 104 (1982) 619.
[33] M.J. Nappa, R. Santi, J. Halpern, Organometallics 4 (1985) 34.
[34] I. Kovacs, C.D. Hoff, F. Ungvary, L. Marko, Organometallics 4 (1985) 1347.
[35] C.K. Brown, D. Georglou, G. Wilkinson, J. Chem. Soc. A (1971) 3120–3127.
[36] J. Schwartz, J. Cannon, J. Am. Chem. Soc. 96 (1974) 4721.
[37] B. Martin, D.K. Warner, J. Norton, J. Am. Chem. Soc. 108 (1986) 33.
[38] J.C. Barborak, K. Cann, Organometallics 1 (1982) 1726.
[39] W.D. Jones, R.G. Bergmann, J. Am. Chem. Soc. 79 (1979) 5447.
[40] W.D. Jones, J. Huggins, R.G. Bergman, J. Am. Chem. Soc. 103 (1981) 4415.
[41] N.H. Alemdaroglu, J.L.M. Penninger, E. Oltay, Monatsh. Chem. 107 (1976) 1153.
[42] M.F. Mirbach, J. Organomet. Chem. 265 (1984) 205.
[43] J.P. Collman, J. Belmont, J. Brauman, J. Am. Chem. Soc. 105 (1983) 7288.
[44] J. Feng, M. Garland, Organometallics 18 (1999) 1542.
[45] M. Garland, in: B. Heaton (Ed.), Mechanism in Homogeneous Catalysis, Wiley–VCH, Weinheim, 2005, p. 151.
[46] W. Chew, E. Widjaja, M. Garland, Organometallics 21 (2002) 1882.
[47] E. Widjaja, C. Li, M. Garland, Organometallics 21 (2002) 1991.
[48] C. Li, E. Widjaja, W. Chew, M. Garland, Angew. Chem. Int. Ed. 20 (2002) 3785.
[49] C. Li, E. Widjaja, M. Garland, J. Catal. 213 (2003) 126.
[50] E. Widjaja, C. Li, W. Chew, M. Garland, Anal. Chem. 75 (2003) 4499–4507.
[51] E. Widjaja, C. Li, M. Garland, J. Catal. 223 (2004) 278.
[52] C. Li, E. Widjaja, M. Garland, J. Am. Chem. Soc. 18 (2003) 5540.
[53] C. Li, E. Widjaja, M. Garland, Organometallics 23 (2004) 4131.
[54] G. Liu, Kinetics and mechanism study of homogeneous hydroformylation, M. Eng. Thesis, National University of Singapore, 1999.
[55] D.F. Shriver, M.A. Drezdzon, The Manipulation of Air-Sensitive Compounds, Wiley, New York, 1986.
[56] A. Haynes, in: B. Heaton (Ed.), Mechanism in Homogeneous Catalysis, Wiley–VCH, Weinheim, 2005, p. 107.
[57] L. Damoense, M. Datt, M. Green, C. Steenkamp, Coord. Chem. Rev. 248 (2004) 2393.
[58] M. Garland, in: I.T. Horvath (Ed.), Transport Effects in Homogeneous Catalysis, Encyclopedia of Catalysis, Wiley, New York, 2002.
[59] R. Whyman, In Situ Spectroscopic Studies in Homogeneous Catalysis, Adv. Chem. Ser., Homogeneous Transition Met. Catal. React. 230 (1992) 19–31.
[60] J.L. Vidal, W.E. Walker, Inorg. Chem. 20 (1981) 249.
[61] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 1991.
[62] G. Liu, M. Garland, J. Organomet. Chem. 608 (2000) 76.
[63] G. Csontos, B. Heil, L. Marko, Ann. NY Acad. Sci. 239 (1974) 47.
[64] K.B. Hansen, J.L. Leighton, E.N. Jacobsen, J. Am. Chem. Soc. 118 (1996) 10924.
[65] B.W. McCleland, W.A. Nugent, M.G. Finn, J. Org. Chem. 63 (1998) 6656.
[66] D.R. Moore, M. Cheng, E.B. Lobkovsky, G.W. Coates, J. Am. Chem. Soc. 125 (2003) 11911.

Appendix B

Homogeneous Hydroformylation of Ethylene Catalyzed by $Rh_4(CO)_{12}$. The Application of BTEM to Identify a New Class of Rhodium Carbonyl Spectra: $RCORh(CO)_3(\pi\text{-}C_2H_4)$

Chuanzhao Li, Liangfeng Guo, and Marc Garland*

Department of Chemical and Environmental Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore 119260

Received March 17, 2003

Summary: The homogeneous rhodium-catalyzed hydroformylation of ethylene was studied, starting with $Rh_4(CO)_{12}$ in n-hexane solvent. The organometallic intermediates and the organic products were measured under isobaric and isothermal conditions using in situ high-pressure infrared spectroscopy. The newly developed algorithm of band-target entropy minimization (BTEM) was applied for the spectral reconstruction to the semibatch data. The classic acyl complex $RCORh(CO)_4$ was observed, as well as a new metal carbonyl spectrum. Aldehyde and ketone were observed as organic products. The new organometallic spectrum possesses an acyl band and three strong terminal carbonyl vibrations. The new spectrum was most pronounced at very high ethylene concentrations. Although some very minor vibrations are also apparent, the spectrum is consistent with a trigonal bipyramid structure $RCORh(CO)_3L$ ($L = C_2H_4$). Given the presence of the ketone formed and the small vibrational intensity at ca. 1725 cm^{-1} in the new spectrum, the presence of some observable $RCOCH_2CH_2Rh(CO)_3L$ ($L = C_2H_4$) cannot be excluded.

Introduction

An extraordinary wide range of alkenes have been hydroformylated.[1] Although linear (terminal and internal) and cyclic alkenes have been the focus of most studies, more unusual structures such as methylene cycloalkanes and various highly functionalized alkenes have also been studied. In most of these investigations, hydroformylation to the corresponding aldehyde is by far the predominant overall organic transformation, particularly when rhodium is used. Competitive hydrogenation to alkanes is normally an extremely minor side reaction.[2] The simplest alkene, ethylene, is one of the few noticeable exceptions to the general rule.

Ethylene is known to hydroformylate rapidly in the presence of unmodified rhodium complexes; however, ketone and even polyketone formation is usually nonnegligible.[3] Ethylene appears to be rather unusual in its ability to insert into an acyl–rhodium bond, thereby leading to ketone/polyketone formation. Such repeated alternating insertions of ethylene and CO have been observed with cobalt and ruthenium carbonyl complexes as catalyst precursors as well,[3,4] and the most productive system presently known is palladium catalyzed.[5]

In situ FTIR studies of the hydroformylation of ethylene starting with unmodified rhodium carbonyl species have been reported by at least two groups.[6] In situ FTIR studies have been seriously hindered by the fact that the spectrum of dissolved ethylene changes dramatically as partial pressure is increased. A new polymer matrix method was recently applied to rhodium-catalyzed ethylene hydroformylation. A new species was observed and assigned the stoichiometry $RCORh(CO)_3(\pi\text{-}C_2H_4)$.[7]

The acyl rhodium tetracarbonyl $RCORh(CO)_4$ has been observed during active hydroformylations of a large number of alkene substrates.[8] This information has been used repeatedly to develop rate expressions for aldehyde formation.[9] With the exception of two substrates, notably methylene cyclopropane and ethylene, only the rhodium-containing precursors, $Rh_6(CO)_{16}$ and $RCORh(CO)_4$, could be identified during syntheses. Methylene cyclopropane hydroformylation leads to complex in situ spectra, and ethylene hydroformylations exhibit a spectral pattern containing at least one new rhodium carbonyl species.

In the present contribution, a semibatch in situ study of ethylene hydroformylation is performed. The new spectral processing program BTEM[10] is used to recover the pure component spectra of all observable species and their concentrations. BTEM is able to overcome the problematic ethylene spectral features and to recover a

* To whom correspondence should be sent. E-mail: chemvg@nus.edu.sg. Fax: 65-6779-1936.

(1) (a) Marko, L.; Ungvary, F. *J. Organomet. Chem.* **1992**, *432* (1–3), 1. (b) Ungvary, F. *Coord. Chem. Rev.* **1998**, *170*, 245. (c) Ungvary, F. *Coord. Chem. Rev.* **2001**, *218*, 1.

(2) Pino, P.; Piacenti, P.; Bianchi, M. In *Organic Synthesis Via Metal Cabonyls*; Wender, I., Pino, P., Eds.; John Wiley & Sons: New York, 1977; Vol 2.

(3) Consiglio, G.; Studer, B.; Oldani, F.; Pino, P. *J. Mol. Catal.* **1990**, *58*, L9.

(4) (a) McClure, J. D. Ger. Offenlegungsscrift No. 2,046,060, 1971. (b) Cooper, J. L. US Pat. No. 4,602,116, 1986. (c) Koelliker, R. Ph.D. Thesis #8704, ETH Zurich, 1988.

(5) Drent, E.; Van Broekhoven, J. A. M.; Doyle, M. J. *J. Organomet. Chem.* **1991**, *417*, 235.

(6) (a) King, R. B.; King, A. D.; Iqbal, M. Z. *J. Am. Chem. Soc.* **1979**, *101*, 4893. (b) Liu, G.; Garland, M. *J. Organomet. Chem.* **2000**, *613*, 124. (c) Li, C.; Guo, L.; Garland, M. 13th International Symposium on Homogeneous Catalysis, Paper 206, 2002.

(7) Zhang, J.; Poliakoff, M.; George, M. W. *Organometallics* **2003**, *22*, 1612.

(8) (a) Garland, M.; Bor, G. *Inorg. Chem.* **1989**, *28*, 410. (b) Garland, M.; Pino, P. *Organometallics* **1991**, *10*, 1693. (c) Liu, G.; Garland, M. *Organometallics* **1999**, *18*, 3457.

(9) (a) Garland, M.; Pino, P. *Organometallics* **1990**, *9* (6), 1943. (b) Garland, M. *Organometallics* **1993**, *12*, 535. (c) Fyhr, C.; Garland, M. *Organometallics* **1993**, *12* (5), 1753. (d) Feng, J.; Garland, M. *Organometallics* **1999**, *18*, 417.

10.1021/om030202e CCC: $27.50
Publication on Web 03/31/2004

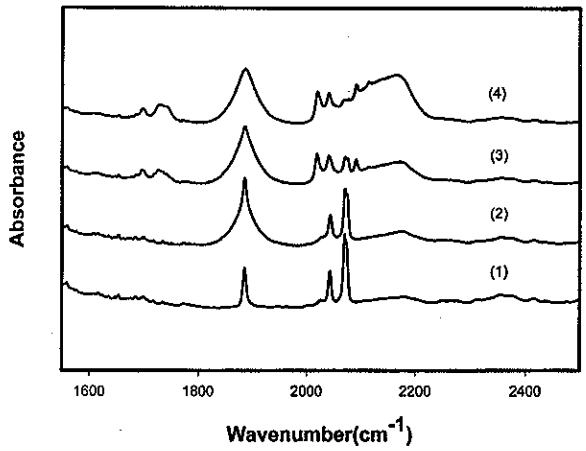


Figure 1. Four in situ high-pressure FTIR spectra of the active hydroformylation semibatch reaction. The initial conditions were 298.0 mg of $Rh_4(CO)_{12}$ and 200 mL of hexane at 20.5 °C. Spectrum 1: 5 bar ethylene, 4.5 bar CO, 3.5 bar H_2. Spectrum 2: 15 bar ethylene, 4.5 bar CO, 3.5 bar H_2. Spectrum 3: 30 bar ethylene, 4.5 bar CO, 3.5 bar H_2. Spectrum 4: 30 bar ethylene, 15 bar CO, 15 bar H_2.

relatively clean pure component spectrum of the new class of rhodium carbonyl complexes $RCORh(CO)_3L$ (L = alkene) under real reaction conditions. The present results help to clarify the preliminary experimental results obtained by this group,[6b] provide further detail about the BTEM analysis of this system,[6c] and are consistent with the polymer matrix results recently reported by Poliakoff et al.[7]

Results and Discussion

Reaction Spectra and Spectral Distortions. Two semibatch ethylene hydroformylations were performed at ca. 9 and 20 °C starting with $Rh_4(CO)_{12}$ as catalyst precursor. The maximum partial pressure of ethylene used was 50 bar. The problematic spectral features of ethylene are shown in Figure 1 where four in situ spectra at different reaction conditions are presented. At low partial pressures, i.e., 5 bar, ethylene has a reasonably narrow bandwidth for the band at 1885 cm^{-1} (possibly overtone from C=C−H bending mode). However, as indicated by the figure, the width at half-height for ethylene can increase to ca. 80 cm^{-1} at 30 bar ethylene partial pressure.

BTEM. The first semibatch experiment contained 6 steps/perturbation in reactants or solvent, and the second contained 8 steps. The total number of spectra taken was 366. Eleven pure component spectra were recovered by BTEM from the raw experimental spectra. These include the atmospheric moisture and CO_2, the solvent hexane, and dissolved CO, as well as three organic solutes and four organometallics.

The three organic pure component spectra were ethylene (1), propanal (2), and ketone/polyketone (3). Their pure component spectra are shown in Figure 2. The maximum for the recovered ethylene spectrum is at 1885.8 cm^{-1}. It was obtained using 50 right singular

(10) (a) Chew, W.; Widjaja, E.; Garland, M. *Organometallics* **2002**, *21*, 1882. (b) Widjaja, E.; Li, C.; Garland, M. *Organometallics* **2002**, *21*, 1991. (c) Li, C.; Widjaja, E.; Chew, W.; Garland, M. *Angew. Chem., Int. Ed.* **2002**, *41* (20), 3785. (d) Widjaja, E.; Li, C.; Chew, W.; Garland, M. *Anal. Chem.* **2003**, *75*, 4499.

Appendix B

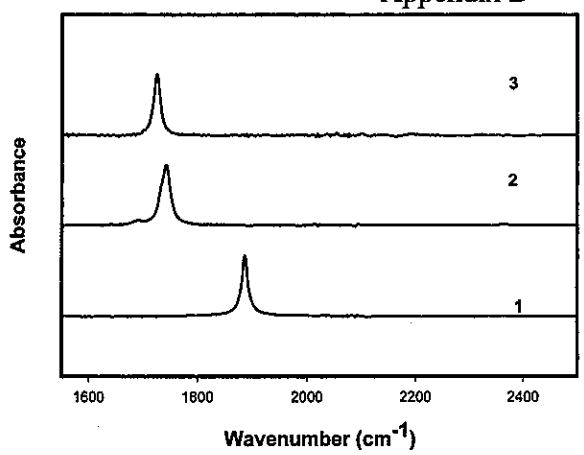


Figure 2. BTEM spectra of recovered organic species. Spectrum 1: ethylene. Spectrum 2: propanal. Spectrum 3: ketone/polyketone.

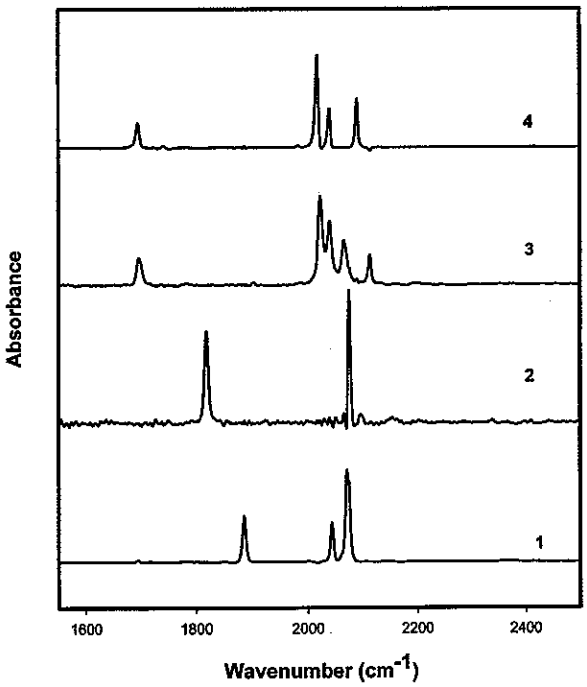


Figure 3. BTEM spectra of recovered organometallic species. Spectrum 1: $Rh_4(CO)_{12}$. Spectrum 2: $Rh_6(CO)_{16}$. Spectrum 3: $RCORh(CO)_4$. Spectrum 4: new species.

vectors in the optimization and targeting the region 1883−1887 cm^{-1}. This maximum is coincident with the bridging carbonyl of $Rh_4(CO)_{12}$. The maximum absorbance of aldehyde occurs at 1742.4 cm^{-1} with a shoulder at 1693.2 cm^{-1}. A total of 25 vectors were used, and the region targeted was 1741−1745 cm^{-1}. The maximum absorbance of the ketone/polyketone is 1725.2 cm^{-1} with a shoulder at 1701 cm^{-1}. A total of 25 vectors were used, and the region targeted was 1723−1727 cm^{-1}. The BTEM reconstructions of the organic species show very few spectral artifacts.

The four pure component organometallic spectra recovered included the three expected spectra $Rh_4(CO)_{12}$ (1), $Rh_6(CO)_{16}$ (2), and $RCORh(CO)_4$ (3) as well as one new spectrum (4). The four organometallic spectra are shown in Figure 3. The primary vibrational features in the $Rh_4(CO)_{12}$ spectrum are 1885.4, 2044.2, 2071.2, and

2074.4 cm^{-1}. A total of 25 vectors were used, and the region targeted was 2069–2073 cm^{-1}. The maxima in the $Rh_6(CO)_{16}$ spectrum are 1818.4 and 2075.4 cm^{-1}, obtained using 100 vectors, and the region targeted was 1816–1820 cm^{-1}. The maxima in the acyl rhodium tetracarbonyl $RCORh(CO)_4$ (R = $-CH_2CH_3$) spectrum are 1696.6, 2023.2, 2040.6, 2066.6, and 2113.6. cm^{-1}, obtained using 50 vectors, and the region targeted was 2112–2115 cm^{-1}. The maxima in the new complex are 1695.2, 2017.6, 2040.2, and 2090.4 cm^{-1}, obtained using 50 vectors, and the region targeted was 2017–2019 cm^{-1}.

The BTEM reconstructions of the first three complexes are very consistent with our previous studies, and the spectra show few artifacts. The relatively high levels of noise in the reconstructions of $Rh_6(CO)_{16}$ and $RCORh(CO)_4$ are due to the extremely low concentration of the former due to its solubility and the unusually low conversion to acyl in the case of ethylene hydroformylation. Higher concentrations of $RCORh(CO)_4$ have been observed using $Rh_4(CO)_{12}$ and most other alkenes.

$RCORh(CO)_3(\pi\text{-}C_2H_4)$ The new spectrum shows a distinct artifact at 2114.4 cm^{-1} due to residual mixed-in signals from $RCORh(CO)_4$, but the signal-to-noise ratio is quite good. Three minor absorptions are apparent at 1724.2, 1741.4, and 1886.6 cm^{-1}. The last two are certainly due to residual mixed-in signals from propanal and $Rh_4(CO)_{12}$ and/or ethylene.

It would appear that the band at 1695.2 cm^{-1}, which is almost exactly coincident with the acyl band of $RCORh(CO)_4$, must in fact belong to a C=O in a RCO group. This together with the simplicity of the spectrum suggests it is mononuclear. This implies that the new species is an $RCORh(CO)_xL_y$ ($x < 4$, $y \geq 1$). Furthermore, given the low C_s symmetry of such a species, the number of terminal vibrations will equal the number of terminal CO ligands present. Therefore the stoichiometry reduces to $RCORh(CO)_3L$.

To test this hypothesis a little further, we integrated the areas under the terminal CO regions and the acyl bands. The ratio R is 4/2.7. This indicates again that $x = 3$.

$$R = \frac{\int_{2010}^{2120} \hat{a}_{RCORh(CO)_4} / \int_{1680}^{1710} \hat{a}_{RCORh(CO)_4}}{\int_{2010}^{2120} \hat{a}_{RCORh(CO)_3L} / \int_{1680}^{1710} \hat{a}_{RCORh(CO)_3L}} \quad (1)$$

The most obvious ligand L is π-coordinated ethylene. Indeed, the formation of $RCORh(CO)_3(\pi\text{-}C_2H_4)$ would be promoted by higher ethylene concentrations and lower CO concentrations. Furthermore, the insertion of the ethylene ligand into the acyl rhodium bond in $RCORh(CO)_3(\pi\text{-}C_2H_4)$, followed by hydrogen activation, provides a mechanistic reason for the observed ketone/polyketone in the system.

Due to the presence of organic ketone in the system and due to the very minor vibration at 1724.2 cm^{-1} in the reconstructed spectrum of $RCORh(CO)_3(\pi\text{-}C_2H_4)$, the presence of a minute but observable level of $RCOCH_2CH_2Rh(CO)_3(\pi\text{-}C_2H_4)$ cannot be excluded.

Mole Fractions and Further Stoichiometric Considerations. Since the "pure" component spectrum of ethylene is so variable, determination of the moles of all observable components is not straightforward. Indeed, even the pure component spectrum of hexane is variable under extreme dilution by ethylene. Instead, we solved a reduced problem where only part of the spectral data was used. The original matrix of spectral data $A_{366\times4751}$ was partitioned. Two spectral windows, 1600–1740 and 2000–2100 cm^{-1}, were taken and a new composite matrix formed $A_{366\times1202}$. The hexane absorbance at 1138 cm^{-1} was used to convert the spectra to a renormalized form, $A^{Dml}{}_{366\times1202}$. Information on the moles of rhodium put into the reactor was used as a constraint, and thereby were obtained the moles of the three observable organometallic species $Rh_4(CO)_{12}$ and $RCORh(CO)_4$ and the moles of rhodium in the new organometallic species $RCORh(CO)_3(\pi\text{-}C_2H_4)$. Absorptivities from laboratory reference spectra of aldehyde and ketone were used to calculate approximate moles of the organic products. The profiles of moles of species versus spectrum number for the three organometallics and two organics from one of the two sets of data with 200 data points (the 293K semibatch run) are shown in Figure 4. The discontinuous quality of the data arises from the various perturbations in the reaction condition imposed on the system. The primary features to note in Figure 4 are (a) the decline in precursor $Rh_4(CO)_{12}$ and associated increase in intermediates $RCORh(CO)_4$ and $RCORh(CO)_3(\pi\text{-}C_2H_4)$ and (b) the increases in aldehyde and ketone/polyketone.

The mole data suggest that the approximate maximum ratio of ketone/polyketone to aldehyde in this system was a very significant 0.45. GC–MS was used to confirm the presence of diethyl ketone. Higher molecular weight ketone oligiomers could not be identified by the GC–MS analysis.

Appendix B

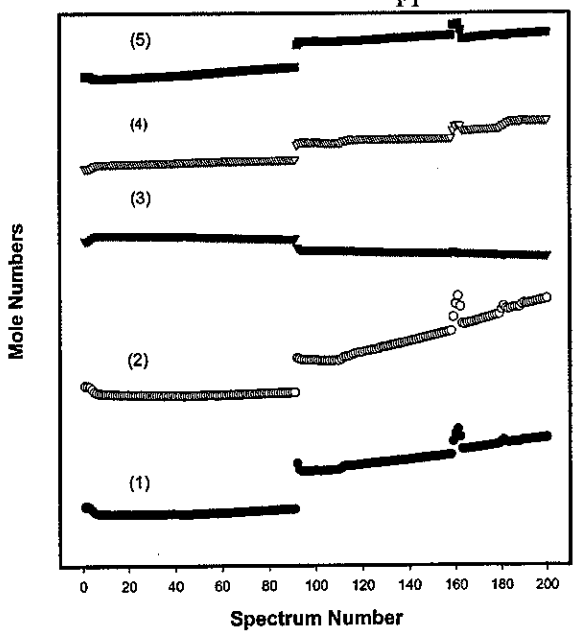


Figure 4. Moles of species versus spectrum number for the three primary organometallic species and two organic products. Curve 1: propanal. Curve 2: ketone/polyketone. Curve 3: $Rh_4(CO)_{12}$. Curve 4: $RCORh(CO)_4$. Curve 5: $RCORh(CO)_4$. The 200 data sets correspond to the first 2 days of the first semibatch experiment. The initial conditions were 298.0 mg of $Rh_4(CO)_{12}$ and 200 mL of hexane at 20.5 °C, 30 bar ethylene, 4.5 bar CO, 3.5 bar H_2. Scale on y-axis is arbitrary due to widely different concentrations.

Experimental Section

General Information. All solution preparations were carried out under argon (99.9995%, Soxal, Singapore) using standard Schlenk techniques.[11] The argon was further purified prior to use by passage through deoxy and zeolite columns. All reactions were carried out under carbon monoxide (99.97%, Soxal, Singapore) and hydrogen (99.999%, Soxal, Singapore) after further purification through deoxy and zeolite columns.

The precious metal complex $Rh_4(CO)_{12}$, with stated purity of 98% min., was obtained from Strem Chemicals (Newport, MA) and was used without further purification, although trace quantities of the high-nuclearity cluster $Rh_6(CO)_{16}$ are virtually always present. The *n*-hexane solvent (stated purity >99.6%, Fluka AG) was refluxed over sodium potassium alloy under argon. Ethylene (99.9%, Soxal, Singapore) was used as obtained.

Equipment. Two semibatch experiments were conducted in *n*-hexane as solvent in a 1.5 L high-pressure batch reactor system with in situ FTIR capability. Details of the equipment can be found elsewhere.[8d]

GC–MS was run on an HP6890 (HP-1 methyl siloxane capillary column, 100°C) connected to a HP 5973.

In Situ Spectroscopic Studies. Two single semibatch experiments were carried out in the following manner. First 150 mL of *n*-hexane was transferred under argon to the autoclave. The total system pressure was raised to 0.4 MPa CO, and the stirrer and high-pressure membrane pump were started. A solution of 298–374 mg of $Rh_4(CO)_{12}$ dissolved in 50 mL of *n*-hexane was prepared, transferred to the high-pressure reservoir under argon, pressured with CO, and then added to the autoclave. Ethylene (3.0–4.0 MPa) was then added to the autoclave. Hydrogen (0.5 MPa) was then added to initiate the synthesis. Spectra were recorded every 5 min at 0.2 cm^{-1} intervals in the range 1000–2500 cm^{-1}.

In the following steps, the partial pressures of carbon monoxide (0.4–2.4 MPa), hydrogen (0.5–3.0 MPa), and ethylene (3.0–5.0 MPa), the solvent *n*-hexane (200–250 mL), and the reaction temperature (282–293 K) were repeatedly changed. A total of 6 steps were performed in one experiment and 8 steps in another experiment. In each step, ca. 20–40 spectra were taken, and a total of 366 spectra were obtained for the spectroscopic analysis. Mass transfer considerations were taken into account,[12] and in the present configuration, the rate of reaction was very slow compared to transport.

Computational Section

A newly developed algorithm of band-target entropy minimization (BTEM) was applied for the spectral reconstruction to analyze the spectra. The consolidated experimental spectra were first subjected to a data decomposition approach using singular value decomposition (SVD). The orthogonal basis vectors that spanned the subspace of observations were transformed into pure component spectral estimates using the BTEM algorithm.[10] Details of the mathematical procedures used are described in ref 10.

Acknowledgment. Financial support for this experimental research was provided by the Academic Research Fund of the National University of Singapore (NUS) under R-279-000-089-112. Research scholarships for C.L. and L.G. were provided by the Graduate School of Engineering (NUS). In addition, the authors wish to thank Prof. I. T. Horvath and Dr. Andrea Bodor at Department of Technology and Environmental Chemistry, Eotvos University, for attempting high-pressure ^{13}C NMR experiments of this system. Organic product was observed but no intermediates. Mixing and mass transfer difficulties are suspected, consistent with discussions concerning NMR tubes found in ref 12.

OM030202E

(11) Shriver, D. F.; Drezdzon, M. A. *The Manipulation of Air-Sensitive Compounds*; Wiley: New York, 1986.

(12) Garland, M. Transport Effects in Homogeneous Catalysis. In *Encyclopedia of Catalysis*; IT Horvath, Ed.; Wiley: New York, 2002; Vol. 6, p 550.

Organometallics **2004**, *23*, 5275–5279

Identification of Rhodium–Rhenium Nonacarbonyl $RhRe(CO)_9$. Spectroscopic and Thermodynamic Aspects

Chuanzhao Li, Liangfeng Guo, and Marc Garland*

Department of Chemical and Biomolecular Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore 119260

Received June 3, 2004

The hydrido metal carbonyl $HRe(CO)_5$ reacts rapidly with the cluster $Rh_4(CO)_{12}$ at room temperature in *n*-hexane solvent under CO and H_2 to give the coordinately saturated dinuclear carbonyl complex $RhRe(CO)_9$. At room temperature and low partial pressures of CO and H_2 (P_t < 2.2 MPa), an observable equilibrium is established between the reactants: $4HRe(CO)_5 + Rh_4(CO)_{12} + 4CO \rightarrow 4RhRe(CO)_9 + 2H_2$. This observation implies the facile activation of molecular hydrogen on $RhRe(CO)_9$ at mild conditions and in the presence of CO. A pure component spectrum of $RhRe(CO)_9$ was obtained by BTEM analysis from the in situ FTIR spectroscopic measurements of the equilibrated solutions. The new species has absorbance maxima at 1985.6(s), 2012.2(w), 2026.6(vs, br), 2075(s), and 2127.2(w) cm^{-1}, indicative of local trigonal bipyramidal geometry on the $-Rh(CO)_4$ moiety and square bipyramidal geometry on the $-Re(CO)_5$ moiety. Equilibrium measurements on the temperature interval T=289.7–308.2 K and the partial pressure intervals 0.2 MPa < P_{CO} < 2.2 MPa and 0.05 MPa < P_{H2} < 2.0 MPa were performed, providing an enthalpy of reaction $\Delta_r H = -116 \pm 29$ kJ/mol and entropy of reaction $\Delta_r S = -312 \pm 99$ J/(mol K). Attempts to isolate $RhRe(CO)_9$ by crystallization at −78 °C were unsuccessful.

Introduction

Metal carbonyl compounds have played a rather special historical role in the development of organometallic chemistry since they represent many of the first identified and isolated metal-organics and were used in many of the first detailed metal-organic mechanistic studies.[1] Mononuclear carbonyls of many of the transition metal elements have been synthesized. Additionally, many homometallic dinuclear and polynuclear carbonyl clusters have been synthesized.[2] Metal carbonyls are frequently used for stoichiometric as well as catalytic metal-mediated organic syntheses, to introduce organo-carbonyl functionalities.[3] In turn, carbonyl-containing functional groups are widely considered to be the backbone of many further synthetic strategies.[4]

Considerable interest in simple hetero-bimetallic metal carbonyls exists. Much of this interest arises from (a) outstanding structural questions, (b) the desire to further understand basic reaction mechanisms particularly where activation of small molecules occurs, and (c) the need to explain the observation of synergism in many bimetallic metal-mediated organic syntheses. Although a large variety of bimetallic polynuclear carbonyls have been synthesized and characterized, the same cannot be said of the most simple mixed-metal organometallics, namely, unsubstituted hetero-bimetallic dinuclear carbonyls. An extensive list of known unsubstituted hetero-bimetallic dinuclear carbonyls is provided in Table 1.

The species listed in Table 1 have a number of interesting attributes, but particular noteworthy are the nonisolatibility and reactivity under mild conditions of some of these species. For example, $CoRh(CO)_7$ is nonisolatable,[12] it readily reacts with molecular hydrogen at subambient temperatures,[14] and it initiates fairly

* To whom correspondence should be addressed. Phone: +65-6-874-6617. Fax: +65-6-779-1936. E-mail: chemvg@nus.edu.sg.

(1) Basolo, F.; Ralph, G. P. *Mechanisms of inorganic reactions; a study of metal complexes in solution*; Wiley: New York, 1958.

(2) (a) Housecroft, C. E. *Metal–metal bonded carbonyl dimers and clusters*; Oxford University Press: New York, 1996. (b) Dyson, P. J.; McIndoe, J. S. *Transition metal carbonyl cluster chemistry*; Gordon and Breach Science Publishers: The Netherlands, 2000.

(3) Wender, I.; Pino, P. *Organic syntheses via metal carbonyl*; Wiley: New York, 1977; Vol. 1 and Vol. 2.

(4) Seebach, D.; Weidmann, B.; Widler, L. In *Modern Synthetic Methods*; Scheffold, R., Ed.; Otto Salle Verlag: Frankfurt, 1983; p 324.

(5) (a) Joshi, K. K.; Pauson, P. L. *Z. Naturforsch.* **1962**, *17b* (8), 565. (b) Bor, G.; Sbrignadello, G. *J. Chem. Soc., Dalton Trans.* **1974**, (4), 440–8. (c) Beck, K.; Alexander, J.; Krause Bauer, J. A.; Nauss, J. L. *Inorg. Chim. Acta* **1999**, *288* (2), 159–173.

(6) Addison, S. J.; Connor, J. A.; Kinkaid, J. A. *J. Organomet. Chem.* **1998**, *554* (2), 123–127.

(7) Kovacs, I.; Hoff, C.; Ungvary, F.; Marko, L. *Organometallics* **1985**, *4*, 1347–1350.

(8) (a) Kruck, T.; Hoefler, M. *Ber.* **1964**, *97* (8), 2289–300. (b) Kruck, T.; Hoefler, M. *Angew. Chem.* **1964**, *76* (18), 786.

(9) Sbrignadello, G.; Tomat, G.; Magon, L.; Bor, G. *Inorg. Nucl. Chem. Lett.* **1973**, *9* (10), 1073–1077.

(10) (a) Nesmeyanov, A. N.; Anisimov, K. N.; Kolobova, N. E.; Kolomnikov, I. S. *Ser. Khim.* **1963**, 194. (b) Sbrignadello, G.; Batiston, G.; Bor, G. *Inorg. Chim. Acta* **1975**, *14* (1), 69–78. (c) Knox, S. A. R.; Hoxmeier, R. J.; Kaesz, H. D. *Inorg. Chem.* **1971**, *10* (11), 2636–2637. (d) Michels, G. D.; Svec, H. J. *Inorg. Chem.* **1981**, *20* (10), 3445–3447.

(11) Martin, B.; Warner. D. K.; Norton, J. R. *J. Am. Chem. Soc.* **1986**, *108* (1), 33–39.

(12) (a) Spinder, F.; Bor, G.; Dietler, U. K.; Pino, P. *J. Organomet. Chem.* **1981**, *213*, 303. (b) Garland, M.; Horvath, I. T.; Bor, G.; Pino, P. *Organometallics* **1986**, *10*, 559–567. (c) Horvath, I. T.; Bor, G.; Garland, M.; Pino, P. *Organometallics* **1986**, *5* (7), 1441–1445.

(13) Garland, M.; Horvath, I. T.; Bor, G.; Pino, P. *Organometallics* **1991**, *10*, 559–567.

(14) (a) Horvath, I.; Garland, M.; Bor, G.; Pino, P. *J. Organomet. Chem.* **1988**, *358*, C17–C22. (b) Garland, M.; Pino, P. *Organometallics* **1990**, *9*, 1943–1949.

10.1021/om0496030 CCC: $27.50
Publication on Web 09/25/2004

Table 1. List of Known Unsubstituted Hetero-bimetallic Dinuclear Carbonyls and Relevant References to Their Characterization

species	refs for syntheses and character-ization	refs for thermo-dynamics	refs for mechanistic studies	refs for catalysis
$CoMn(CO)_9$	5	6	7	
$CoRe(CO)_9$	5b, 8	6		
$CoTc(CO)_9$	5b, 9			
$ReMn(CO)_{10}$	10	6	11	
$MnTc(CO)_{10}$	10b,10d			
$ReTc(CO)_{10}$	10b, 10d			
$CoRh(CO)_7$	12	13	14	14b, 15, 18
$CoRhCO)_8$	12b	13	14	14, 15
$RhRe(CO)_9$	this work	this work	this work	

rapid catalytic hydroformylation at ambient or even subambient temperatures at a total CO/H_2 pressure of 0.1 MPa.[15] These conditions are similar to those for the Wilkinson catalyst $HCORh(PPh_3)_3$.[16] The synergism observed in catalytic hydroformylation with $CoRh(CO)_7$ as precursor has been attributed to (a) facile hydrogen activation and subsequent generation of $HRh(CO)_3$/$HRh(CO)_4$ for the active rhodium cycle at low temperatures[17] and (b) cluster catalysis at high temperatures.[18] As noted previously, low temperature activation of molecular hydrogen is often very difficult to achieve particularly in the presence of carbon monoxide.[14,19]

Further, in the context of synergism, it can be mentioned that the existence of highly reactive hetero-bimetallic dinuclear carbonyls $MnRh(CO)_9$/$MnRh(CO)_8$ has been postulated in recent low-temperature hydroformylations where $HMn(CO)_5$ and $RCORh(CO)_4$ are simultaneously observed under catalytic reaction conditions. A new catalytic reaction topology, called bimetallic catalytic binuclear elimination, has been invoked to explain the kinetics. The active catalytic system consists of three sets of intermediates: mononuclear manganese, mononuclear rhodium, and hetero-bimetallic dinuclear species.[20]

In the present contribution the synthesis and characterization of the new hetero-bimetallic dinuclear carbonyl $RhRe(CO)_9$ is presented.

Experimental Section

General Information. All solution preparations and transfers were carried out under purified argon (99.9995%, Saxol, Singapore) atmosphere using standard Schlenk techniques.[21] The argon was further purified before use by passing it through a deoxy and zeolite column. Purified carbon monoxide (research grade, 99.97%, Saxol, Singapore) and purified hydrogen (99.9995%, Saxol, Singapore) were also further purified through deoxy and zeolite columns before they were used. Purified nitrogen (99.9995%, Saxol, Singapore) was used to purge the Perkin-Elmer FT-IR spectrometer system.

$Rh_4(CO)_{12}$ (98%) and $Re_2(CO)_{10}$ (99%) were purchased from Strem Chemicals (Newport, MA) and was used as obtained. $HRe(CO)_5$ was custom synthesized by Strem Chemicals.

Apparatus. In situ kinetic studies were performed in a 1.5 L stainless steel (SS316) autoclave (P_{max} = 22.5 MPa, Buchi-Uster, Swizerland), which is connected with a high-pressure infrared cell. The system is the same as that used previously.[22]

In Situ Spectroscopic Studies. An experimental design was planned, which would utilize the semibatch approach and algorithm.[23] All the experiments were performed in a similar manner. For example, first, single-beam background spectra of the IR sample chamber were recorded. Then 150 mL of *n*-hexane was transferred under argon to the autoclave. Under 0.2 MPa CO pressure, infrared spectra of the *n*-hexane in the high-pressure cell were recorded. The total system pressure was raised to the set CO pressure, and the stirrer and high-pressure membrane pump were started. After equilibration, infrared spectra of the CO/*n*-hexane solution in the high-pressure cell were recorded. A solution of $Rh_4(CO)_{12}$ dissolved in 50 mL of *n*-hexane was prepared, transferred to the high-pressure reservoir under argon, pressured with CO, and then added to the autoclave. After equilibration, infrared spectra of the $Rh_4(CO)_{12}$/CO/*n*-hexane/solution in the high-pressure cell were recorded. A solution of $HRe(CO)_5$ dissolved in 50 mL of *n*-hexane was prepared, transferred to the high-pressure reservoir under argon, pressured with CO, and then added to the autoclave. In the following steps, the semibatch experiment methodology was performed, namely, the CO/H_2 pressures were varied in each semibatch step. Once the system reached equilibrium, the CO/H_2 pressures were varied to conduct the next semibatch step. The detailed experimental design is shown in Table 2.

The experimental design of the experiments involved 250 mL of solvent and the intervals 289.7−308.2 K, P_{H2} = 0−2.0

Appendix C

Table 2. Experimental Design

expt	step	P_{H2}, MPa	P_{CO}, MPa	$Rh_4(CO)_{12}$, mg	$HRe(CO)_5$, μL	hexane, mL	T, K
1	1	0.11	0.31	88.68	58	250	289.7
	2	0.21	0.31	88.68	58	250	289.7
	3	0.31	0.31	88.68	58	250	289.7
	4	0.51	0.31	88.68	58	250	289.7
2	1	0	2.0	77.87	70.0	250	298.3
	2	1.0	2.0	77.87	70.0	250	298.3
	3	2.0	2.0	77.87	70.0	250	298.3
3	1	0.05	0.2	83.86	70.0	250	298.3
	2	0.10	0.2	83.86	70.0	250	298.3
	3	0.15	0.2	83.86	70.0	250	298.3
	4	0.40	0.2	83.86	70.0	250	298.3
4	1	0.05	0.4	168.88	55.0	250	298.3
	2	0.10	0.4	168.88	55.0	250	298.3
	3	0.15	0.4	168.88	55.0	250	298.3
	4	0.20	0.4	168.88	55.0	250	298.3
	5	0.40	0.4	168.88	55.0	250	298.3
5	1	0.11	0.4	160.75	42.0	250	308.2
	2	0.21	0.4	160.75	42.0	250	308.2
	3	0.42	0.4	160.75	42.0	250	308.2
	4	0.80	0.4	160.75	42.0	250	308.2
	5	0.8	0.8	160.75	42.0	250	308.2
6	1	0.0	1.6	124.9	64.5	250	308.2
	2	1.0	1.6	124.9	64.5	250	308.2
	3	1.58	1.6	124.9	64.5	250	308.2
	4	2.0	1.6	124.9	64.5	250	308.2
	5	2.0	2.2	124.9	64.5	250	308.2

(15) Garland, M. Low-temperature homogeneous catalytic hydroformylation with mixed cobalt and rhodium carbonyls as catalyst precursors, Ph.D. Thesis, ETH No 8585, 1988.

(16) (a) Osborn, J. A.; Jardine, F. H.; Young, J. F.; Wilkinson, G. J. Am. Chem. Soc. **1966**, *12*, 1711−1732. (b) Evans, D.; Yagupsky, G.; Wilkinson, G. *J. Am. Chem. Soc.* **1968**, *11*, 2660−2665. (c)Yagupsky, G.; Brown, C. K.; Wilkinson, G. *J. Am. Chem. Soc.* **1970**, *9*, 1392−1401.

(17) Garland, M. *Organometallics* **1993**, 12, 535−543.

(18) Ojima, I.; Li, Z. Catalysis of Rh, Rh−Co, and Ir−Co multinuclear complexes and its applications to organic syntheses. In *Catalysis by Di- and Polynuclear Metal Cluster Complexes*; Adams, R. D., Cotton, F. A., Eds.; Wiley-VCH: New York, 1998; pp 307−343.

(19) (a) Halpern, J. *Adv. Catal.* **1959**, *11*, 301. (b) Pino, P.; Oldani, F.; Consiglio, G. *J. Organomet. Chem.* **1983**, 250491.

(20) (a) Li, C.; Widjaaj, E.; Garland, M. *J. Am. Chem. Soc.* **2003**, *125* (18), 5540−5548. (b) Li, C.; Widjaaj, E.; Garland, M. *Organometallics* **2004**, *23* (17), 413−418.

(21) Shriver, D. F.; Drezdzon, M. A. *The Manipulation of Air-Sensitive Compounds*; Wiley: New York, 1986.

(22) Feng, J.; Garland, M. *Organometallics* **1999**, *18* (3), 417−427.

(23) Widjaja, E.; Li, C.; Garland, M. *Organometallics* **2002**, 21, 1991−1997.

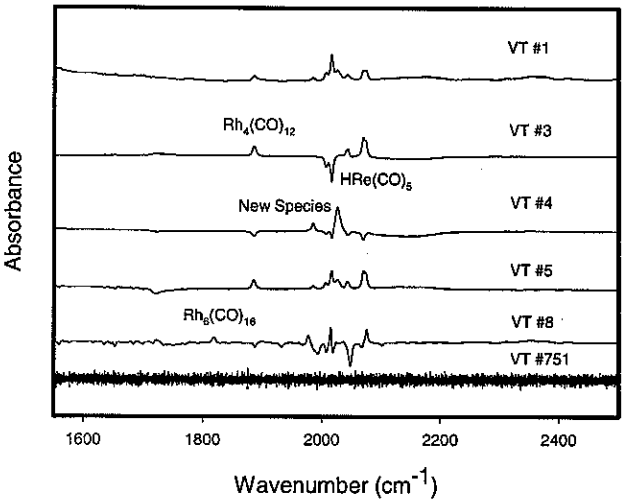


Figure 1. Singular value decomposition of the in situ spectroscopic data showing the first, third, fourth, and fifth significant vectors and the 751th vector. The marked extrema are those that were used to recover the organometallic pure component spectra by BTEM.

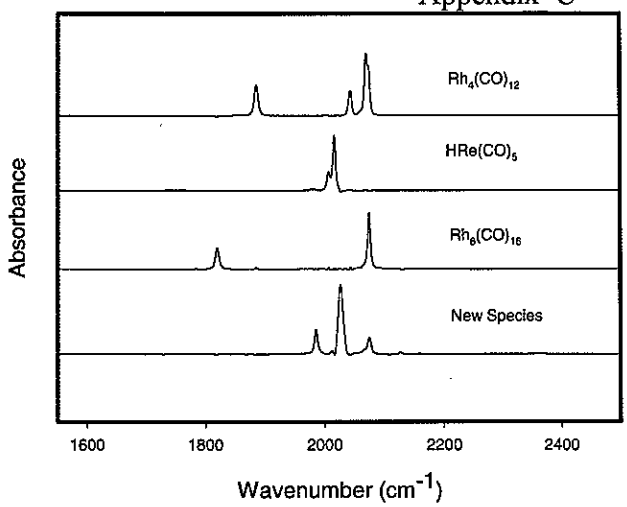


Figure 2. Recovered pure component spectra of the organometallic species using BTEM.

MPa, P_{CO} = 0.2–2.2 MPa, initial $Rh_4(CO)_{12}$ = 77.87–168.88 mg, and initial $HRe(CO)_5$ = 42.0–64.5 μL.

The in situ spectra were taken every 10 min in the range 1000–2500 cm^{-1} with a resolution of 4 cm^{-1}. Since equilibrium was normally established in hours, most of these spectra were recorded under nonequilibrium conditions. A total of 751 spectra were obtained for further spectroscopic analyses.

Attempt at Isolation/Crystallization. $HRe(CO)_5$ (40 μL) was reacted with 80.5 mg of $Rh_4(CO)_{12}$ at room temperature in 200 mL of *n*-hexane solvent under 2.0 MPa CO and in the absence of added H_2 for 6 h. The bright orange reaction mixture was then released to a 250 mL Schlenk tube and cooled with dry ice for 24 h under CO protection. Some orange solids were obtained after removing solvent/dissolved CO, and these were analyzed with FTIR. The spectra obtained are distinctly different than the solution spectrum of $RhRe(CO)_9$.

Results and Discussions

Spectral Analyses. BTEM was used to analyze the spectral data and obtain pure component spectra.[24] A total of 751 spectra were obtained in the experiments. Singular value decomposition (SVD) was employed first to decompose the experimental absorbance data matrix $\mathbf{A}_{751\times4751}$, without any preconditioning,[25] to give the orthonormal matrixes $\mathbf{U}_{751\times751}$ and $\mathbf{V}^{T}_{4751\times4751}$.

The significant spectral extrema in the first few right singular vectors ($\mathbf{V}^{T}_{751\times4751}$) were inspected and were used as targets in the BTEM algorithm. These noteworthy right singular vectors are presented in Figure 1. Four significant bands, as indicated by the numbers in Figure 1, are highlighted. These bands are prominent spectral features associated with the organometallics present and were targeted by BTEM.

The targeting of the highlighted bands in Figure 1 resulted in four pure component organometallic spectra. The reconstructed pure component spectra via BTEM are presented in Figure 2. These spectra were $Rh_4(CO)_{12}$, $HRe(CO)_5$, $Rh_6(CO)_{16}$, and a new species. The new species has peaks at 1985.6(s), 2012.2(w), 2026.6(vs, br), 2075(s), and 2127.2(w) cm^{-1}. Exhaustive use of BTEM on the remaining features in Figure 1 did not result in the recovery of any additional organometallic spectra.

Since the IR spectrum of the new species has no features indicating bridging C–O groups, a higher nuclearity cluster involving two or more Rh atoms is unlikely. Indeed, all known mixed Rh–M clusters having two or more Rh atoms have bridging carbonyls. Good examples include $Co_2Rh_2(CO)_{12}$, $CoRh_3(CO)_{12}$, $Rh_2Ir_2(CO)_{12}$, and $Rh_3Ir(CO)_{12}$.[14,26] Consequently, a low-nuclearity mixed rhodium and rhenium carbonyl is suspected. The simplest example would be a dinuclear complex with a single metal–metal bond and coordinatively saturated moieties $-Rh(CO)_4$ and $-Re(CO)_5$.

It is known that the local symmetry of the $-Rh(CO)_4$ group is C_{3v} and the local symmetry of the $Re(CO)_5$ group is C_{4v}.[27] Figure 3 presents the IR spectra of $BrRe(CO)_5$ and $HRh(CO)_4$[16b,28] obtained in our lab for comparison with the new species. As shown in Figure 3, $BrRe(CO)_5$ has three vibrations at 1986(s), 2016.4(w), and 2046(vs) cm^{-1} and $HRh(CO)_4$ has three metal–carbonyl vibrations at 2041.6(vs), 2071.8(m), and 2123.6(w) cm^{-1}. Assuming that the two " local symmetries"[29] rule holds for a species $(CO)_4Rh–Re(CO)_5$, we would expect six bands, three ($2A_1$+1E) for the $Rh(CO)_4$ part of the molecule and another three ($2A_1$+1E) for the $Re(CO)_5$ part of the molecule. Accordingly, two very strong bands (E) should be found in the spectrum. The intensity ratio of the two very strong bands (E) should be ca.

(24) (a) Chew, W.; Widjaja, E.; Garland, M. *Organometallics* **2002**, 21, 1982–1990. (b) Li, C.; Widjaja, E.; Chew, W.; Garland, M. *Angew. Chem.* Int. Ed. **2002**, 20, 3785–3789. (c) Widjaja, E.; Li, C.; Garland, M. *Proceeding of the International Conference on Scientific & Engineering Computation (IC-SEC)*, Recent Advances in Computational Sciences and Engineering; Lee, H. P., Kumar, K., Eds.; Imperial College Press: London, 2002; pp 36–40. (d) Widjaja, E.; Li, C.; Chew, W.; Garland, M. *Anal. Chem.* **2003**, *75*, 4499–4507. (e) Widjaja, E.; Li, C.; Garland, M. *J. Catal.* **2004**, *223* (2), 278–289.

(25) Chen, Li; Garland, M. *Appl. Spectrosc.* **2002**, *56* (11), 1422–1428.

(26) Martinengo, S.; Chini, P.; Albano, V. G.; Gariati, F. *J. Organomet. Chem.* **1973**, *59*, 379–394.

(27) Braterman, P. S. *Metal Carbonyl Spectra*; Academic Press: New York, 1973; pp 155–177.

(28) (a)Whyman, R. In *In situ spectroscopic studies in homogeneous catalysis;* Advances in Chemistry Series 230; 1992; pp 19–31. (b) Vidal, J. L.; Walker, W. E. *Inorg. Chem.* **1981**, *20*, 249–54.

(29) Cotton, F.; Liehr, A. D.; Wilkinson, G. *J. Inorg. Nucl. Chem.* **1956**, *2*, 141–148.

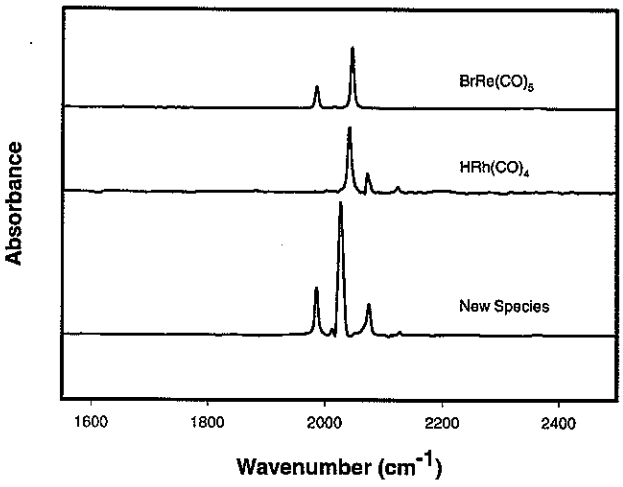
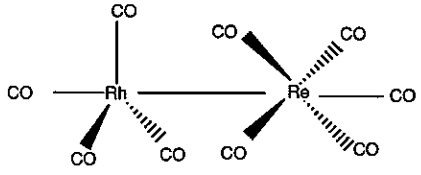


Figure 3. Spectral comparison of the new species with the spectra of $BrRe(CO)_5$ and $HRh(CO)_4$ obtained in our lab.

Scheme 1. Proposed Structure of $RhRe(CO)_9$



3/4, and they should represent ca. 7/9 of the total CO stretching intensities.

Indeed, the new spectrum appears to meet the above assumptions. The two bands 2075(s) and 2127.2(w) cm^{-1} are quite close to the $2A_1$ belonging to the $Rh(CO)_4$ moiety, while the other two bands 1985.6(s) and 2012.2(w) cm^{-1} are quite close to $2A_1$ belonging to the $Re(CO)_5$ moiety. Furthermore, close inspection of the spectrum indicates that the peak at 2026.6 cm^{-1} is very strong and broad (the half-width is ca. 10 cm^{-1} instead of the usual 5 cm^{-1} anticipated) and is accordingly a combination of the two required strong bands (E). Integrated intensities at this position account for ca. 74% of the total CO stretching intensities.

As mentioned in regard to Table 1, the cobalt–rhenium nonacarbonyl is known. Bor et al.[5] reported that $ReCo(CO)_9$ has six peaks at 1971.7(w, br), 1990(s), 2007(w), 2032.8(vs), 2059.8(s), and 2133.9(w) cm^{-1}. Except for the weak broad peak at 1971.7 cm^{-1}, all the other peaks are quite similar to the present new complex. In their assignment, a "free rotational model" was employed to explain the peak at 1971.7 cm^{-1}. Bor et al. also reported that the presence of the isotopic satellite (1956 cm^{-1}) from natural ^{13}C. An ^{13}C isotopic satellite for the lowest vibration at 1944.8 cm^{-1} can be found in the new complex $RhRe(CO)_9$. The proposed structure for $RhRe(CO)_9$ is shown in Scheme 1. The attempts to isolate $RhRe(CO)_9$ by crystallization were unsuccessful. (See Experimental Section for more details.)

Concentration Profiles. Figure 4 shows the time-dependent moles of $Rh_4(CO)_{12}$, $HRe(CO)_5$, and $RhRe(CO)_9$ respectively for experiment 6 (Table 2). As all the peaks of the $HRe(CO)_5$ are overlapping with other species, its calculation is somewhat sensitive.

Appendix C

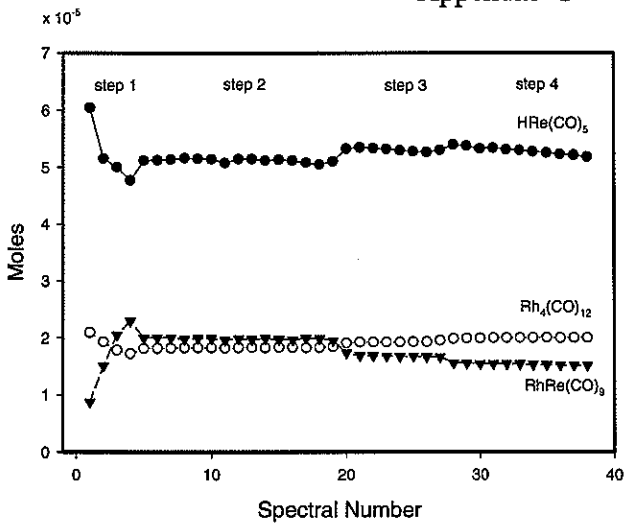


Figure 4. Representative example of the concentration profiles of the organometallic species. This experiment was performed with 250 mL of *n*-hexane at 308.2 K with 0–2.2 MPa CO and 0–2.0 MPa H_2. The initial conditions were 124.9 mg of $Rh_4(CO)_{12}$ and 64.5 μL of $HRe(CO)_5$.

As shown in Figure 4, step 1, with only $Rh_4(CO)_{12}$/$HRe(CO)_5$/CO in the system, the moles of both $Rh_4(CO)_{12}$ and $HRe(CO)_5$ have a very rapid decline, while the moles of $RhRe(CO)_9$ increased rapidly. This means that with little or no molecular hydrogen in the system, the reaction $Rh_4(CO)_{12} + 4HRe(CO)_5 + 4CO \leftrightarrow 4RhRe(CO)_9 + 2H_2$ goes very fast to the right-hand side. In step 2, when molecular hydrogen was added, it can be seen that the $RhRe(CO)_9$ moles decreased, and accordingly the moles of both $Rh_4(CO)_{12}$ and $HRe(CO)_5$ increased until they reached a new equilibrium state. This is clear evidence for the rapid and reversible molecular hydrogen activation by the heterometallic dinuclear carbonyl $RhRe(CO)_9$. In the following steps, with more hydrogen introduced in the reactor, similar trends can be observed in Figure 4.

Thermodynamics. Six series of experiments were conducted at three temperatures to determine the equilibrium constants for the reaction.

$$Rh_4(CO)_{12} + 4HRe(CO)_5 + 4CO \leftrightarrow 4RhRe(CO)_9 + 2H_2 \quad (2)$$

$$K_{eq} = [RhRe(CO)_9]^4[H_2]^2/[Rh_4(CO)_{12}][HRe(CO)_5]^4[CO]^4 \quad (3)$$

In the above equations, all the concentrations were measured in mole fractions. As shown in Table 2, the equilibrium experiments were performed under different partial pressures of CO and H_2 with different initial loadings of $Rh_4(CO)_{12}$ and $HRe(CO)_5$.

The regression analyses yield the equilibrium constants at different temperatures. The equilibrium constants were $K_{eq}(289.7\ K) = 94\ 113$ ($R^2 = 0.88$), $K_{eq}(298.3\ K) = 16\ 715$ ($R^2 = 0.99$), and $K_{eq}(308.2\ K) = 3278$ ($R^2 = 0.89$). The R^2 values indicate that two of the three regressions could be better. The difficulties responsible for the two low R^2 values include (1) the known difficulties in measuring the equilibria of very reactive organometallics,[13] i.e., $CoRh(CO)_7$, (2) the deviation in

the CO and H_2 Henry constants in the binary gas vapor–liquid equilibria, and (3) the extremely large values of the exponents in both the numerator and denominator of the equilibrium constant. The latter sixth- and ninth-order dependencies are most surely a significant source of error. Nevertheless, the equilibrium data are consistent with the overall reaction stoichiometry.

The equilibrium thermodynamics was regressed to provide the thermodynamic parameters $\Delta_r H = -116 \pm 29$ kJ/mol and $\Delta_r S = -312 \pm 99$ J/(mol K). The negative value of the enthalpy $\Delta_r H$ is consistent with the observation of higher conversion to the hetero-bimetallic dinuclear complex $RhRe(CO)_9$ at lower temperatures. In addition, the negative sign of the entropy of the reaction $\Delta_r S$ is in agreement with the stoichiometric equation for this reaction; namely, nine molecule reagents give only six molecules as products. Due to the very low experimental ΔP_T used in this study, a three-parameter thermodynamic expression involving $\Delta_r V$ was not used to analyze the data.

Discussion

It was somewhat surprising not to find $Re_2(CO)_{10}$ in the reaction of $HRe(CO)_5/Rh_4(CO)_{12}$ under CO/H_2. Indeed, the $HRe(CO)_5$ might be expected to undergo bimolecular elimination to form $Re_2(CO)_{10}$, but this reaction is apparently negligible under the present conditions.

The formation of $RhRe(CO)_9$ was also attempted without the use of $HRe(CO)_5$. One semibatch experiment was performed with a reactant combination comprising $Re_2(CO)_{10}$ and $Rh_4(CO)_{12}$ in *n*-hexane under 2.0–3.0 MPa CO and 2.0 MPa H_2, at 298–308 K. Spectral analysis with both BTEM and target factor analysis (TFA)[24e] did not provide any evidence for the presence of $ReRh(CO)_9$.

The intrinsic reaction kinetics for the formation of $RhRe(CO)_9$ and/or the activation of molecular hydrogen on $RhRe(CO)_9$ were too rapid to be captured with the present experimental setup. Indeed, characteristic times for gas–liquid mass transfer as well as liquid-phase mixing were on the order of a few minutes.[30] Therefore, it is quite possible, indeed probable, that transport-controlled rather than intrinsic kinetics was observed. Nevertheless, it appears likely that an intermediate such as an unobservable $H_2RhRe(CO)_8$ and/or $H_2RhRe(CO)_7$ is formed and subsequently undergoes fragmentation to give $HRe(CO)_5$ and a very unstable hydridorhodium species such as $HRh(CO)_4$[24b,28] and/or $HRh(CO)_3$,[31] rapidly yielding $Rh_4(CO)_{12}$.

The relatively infrequent identification of a new unsubstituted hetero-bimetallic dinuclear carbonyl probably justifies a small discussion of potential reactivity patterns. Foremost among these reactivity patterns would be the chemistry of the carbonyl ligands. The rapid hydrogen activation on the coordinately saturated $RhRe(CO)_9$ is a certain indication of facile CO dissociation. In this regard, the rapid dissociation of CO and subsequent hydrogen activation is similar to the case of $CoRh(CO)_8$: both occur at room temperature under significant partial pressures of CO. Furthermore, the issue of facile CO dissociation prompts questions about fluxionality. It is well known that the CO ligands on $Rh_4(CO)_{12}$,[32] but more importantly $CoRh(CO)_7/CoRh(CO)_8$, are fluxional on the NMR time scale. This issue is surely one that should be addressed in the future.

The CO dissociation also prompts questions about possible facile substitution reactions. Substitution with 1 equiv of phosphine almost surely occurs, and substitution on the rhodium is the probable position, as is the case for $CoRh(CO)_7 + PEt_3 \rightarrow CoRh(CO)_6PEt_3 + CO$.[33] Analogously, substitutions with NR_3, PPN, etc., can be expected. More complex addition reactions involving unsaturated organic molecules should not be ruled out either. It is known that $CoRh(CO)_7$ readily adds 1 equiv of alkyne ($C_6F_5C{\equiv}CC_6F_5$, $PhC{\equiv}CPh$) to make $CoRh(CO)_6(\mu\text{-}\eta^2\text{-alkyne})$.[34] A related reaction with $RhRe(CO)_9$ may occur.

The known facile CO dissociation and rapid activation of molecular hydrogen and the presumed reactivity toward addition reactions suggest the strong potential for $RhRe(CO)_9$ as a catalyst precursor in a variety of catalytic syntheses, but particularly carbonylations. Future work in this area is planned.

Acknowledgment. Financial support for this experimental research was provided by the Academic Research Fund of the National University of Singapore (NUS). Research scholarships for C.L. and L.G. were provided by the Graduate School of Engineering (NUS). C.L. also thanks Singapore Millenium Foundation (SMF) for a SMF postdoctoral fellowship.

OM0496030

(30) Garland, M. Transport Effects in Homogeneous Catalysis. In *Encyclopedia of Catalysis*; Horvath, I. T., Ed.; Wiley: New York, 2002.

(31) Chini, P.; Martinengo, S. *Inorg. Chim. Acta* **1969**, *3*, 21.

(32) Evans, J.; Johnson, B. F. G.; Lewis, J.; Norton, J.R.; Cotton, F. A. *J. Chem. Soc., Chem. Commun.* **1973**, 807.

(33) Horvath, I. T. *Organometallics* **1986**, *5* (11), 2333–2340.

(34) (a) Horvath, I. T.; Zsolnai, L.; Huttner, G. *Organometallics* **1986**, *5* (1), 180–182. (b) Horvath, I. T. *Polyhedron* **1988**, *7* (22–23), 2345–2349.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Analytica Chimica Acta 517 (2004) 229–236
Appendix D
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

A general method for the recovery of pure powder XRD patterns from complex mixtures using no a priori information Application of band-target entropy minimization (BTEM) to materials characterization of inorganic mixtures

Liangfeng Guo[a,b], Fethi Kooli[b], Marc Garland[a,b,*]

[a] *Department of Chemical and Biomolecular Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore 117576, Singapore*
[b] *Institute of Chemical and Engineering Sciences, Singapore Block 28, Ayer Rajah Crescent #02-08, Singapore 139959, Singapore*

Received 10 January 2004; received in revised form 4 May 2004; accepted 4 May 2004

Abstract

The recovery of pure component spectra from multi-component mixtures is one of the most common analytical problems in the chemical sciences. In cases where separation of the unknown components is not possible, the problem is often intractable. In materials science research, X-ray diffraction (XRD) and particularly X-ray powder diffraction (XRPD) are perhaps the primary characterization tools. Recently, we introduced band-target entropy minimization (BTEM), an essentially model-free deconvolution technique, applicable to sets of unknown mixture samples and initially applied to liquid-phase characterization (Chew et al., Organometallics, 2002, 21, 1982–1990). In the present study, a set of 12 unknown inorganic powder mixtures were prepared and the XRPD patterns measured. BTEM was then applied. The analysis provided the right prediction that five components were present. Outstanding pure component XRD patterns were obtained for all five components ($Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$, and PbO) present. These results have implications for a large variety of intrinsically inseparable multi-component mixtures encountered in material science research. These include un-reactive as well as reactive systems, and ex situ as well as in situ studies, involving organic, inorganic and even metallic/alloy components. Initial tests suggest that BTEM may be well suited for recovering the trace component diffraction patterns present and hence greatly aiding material characterization.

Keywords: XRD patterns; BTEM; Deconvolution

1. Introduction

The recovery of pure component spectra from multi-component mixtures is a ubiquitous problem in the chemical sciences. In those cases where the multi-component mixtures can be physically separated, the analysis is often greatly simplified. However, in a wide variety of problems and situations, physical separation is either very difficult and/or intrinsically impossible. In the latter cases, the recovery of the pure component spectra from the complex multi-component mixture is extraordinarily difficult. This problem is particularly severe in materials science research where constituents, either compositionally or structurally different, are often intimately and irreversibly associated.

Especially for liquid-phase mixtures, considerable effort has been invested by researchers to devise methods for achieving spectral recovery for more-or-less general cases, namely, where no a priori information is known, neither about the chemistry involved nor the number of components present. Until recently, SIMPLISMA [1,2], IPCA [3], and OPA-ALS [4] were the state-of-the-art chemometric algorithms for solving the problems, but numerous difficulties persist. In the interest of completeness, it should be noted that the recovery of pure component spectra from unknown mixture spectra belongs to an important class of mathematical problems known as ill-posed inverse problems [5].

Recently, we introduced a new and very sophisticated algorithm called band-target entropy minimization (BTEM).

* Corresponding author. Tel.: +65-6874-6617; fax: +65-6779-1936.
E-mail address: chemvg@nus.edu.sg (M. Garland).

0003-2670/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2004.05.006

In BTEM, a set of spectra of unknown composition are collected, and then analyzed. The analysis begins with the use of singular value decomposition (SVD) to provide an initial untangling of the signals present. This is then followed by identification of interesting features present in the vector-space decomposition. Each of these interesting features is then targeted to reconstruct the entire whole-function (spectrum) which is associated with that feature. Concepts from information entropy theory as well as global search routines are brought together to identify these spectra. The exhaustive set of all pure component spectra constitute all the underlying patterns in the data.

Although initially developed for liquid-phase systems, BTEM has now begun to be applied to solid-state problems. BTEM, using FTIR [6–8], Raman [9], NMR [10] and MS [11] spectroscopies has been successfully applied to problems ranging from chemical syntheses to environmental science. The most powerful feature of BTEM is that trace constituents at sub-ppm levels, and constituting less than 0.1% of the total signal of the data set, can be recovered with a signal to noise ratio approaching 50:1 or better [12]. It is worthwhile to repeat that [1] the spectra are recovered using no a priori information what-so-ever (no recourse to chemical information, libraries, previous experience, etc.) [2], the pattern recognition is model-free, so any type of line shapes with any degree of overlap can be recovered [3], the exhaustive search then enumerates the number of components (observable species) present and [4] the algorithm can deal with moderately non-stationary signals (non-linearities) in the data.

Many of the generic problems associated with pure component signal recovery are also found in many types of XRPD studies. These include the assumptions that [1] the components present in the multi-component samples are unknown [2], the number of components present are unknown, and [3] non-linearities in the data are present. In XRPD studies, non-linearities can arise from a variety of sources. The most common source is the packing/orientation of the powder particles [13]. Thus, even a pure component can give visibly different patterns when exact replicates are prepared. Another complication of most conventional XRPD measurements is that the signal-to-noise ratio is typically lower than more wide-spread and common analytical measurements such as UV–Vis, FTIR and Raman. The latter frequently exhibit S/N of better than 100 using scan times of a few minutes. A wide variety of difficult experimental XRPD problems, ranging from the phases present in alloys [14] to the active constituents of heterogeneous catalysts under reaction conditions [15–17], to semiconductor materials/junctions [18] and to polymorphism in pharmaceutical products [19,20] could benefit from a significantly improved methodology.

It is useful to repeat that the general setting for such a methodology would require the measurement of a number of states of the system, in order to enumerate the number of constituents present and their pure component patterns (i.e. different samples of an alloy system containing more than one phase, in situ measurements of an inorganic catalyst at different reaction conditions, or multiple samples of a crystallization at different times). Multiple replicate measurements of more-or-less identical samples are inappropriate. Accordingly, the problem statement is very ill-posed. Multiple observations of a system in numerous states with significant inter-sample variation provide a much more appropriate approach to the problem.

In the present contribution, BTEM is successfully applied to a set of XRPD pattern consisting of inorganic compounds. Twelve different multi-component mixtures were prepared and their XRPD patterns measured. The matrix of pattern data was then analyzed independently, given no prior knowledge of the number nor composition of the constituents present. The correct number of components was found and good XRPD pure patterns were obtained. The recovered pure components are almost identical to reference diffraction patterns of the pure components. A considerable reduction in noise was achieved during the reconstructions (infra vide).

2. Experimental

2.1. Materials, sample preparation and pure XRPD

The five inorganic compounds used in this study were $Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$, and PbO. Lead oxide and zirconium oxide were purchased from Aldrich and used without further purification. The remaining lead compounds were synthesized from lead nitrate and the appropriate bases [9]. The compounds used in this study were chosen since a complementary Raman BTEM study exists [9].

The powder X-ray diffraction patterns of the pure reference inorganic compounds were collected at room temperature using a Bruker D8 Advance diffractometer, with Cu Kα radiation. The diffractometer was operated at 40 kV and 40 mA. The samples in their powder form were scanned in the diffraction angle range 2.5–75° 2θ in steps of 0.01° for 2s/step. Fig. 1 shows the pure reference XRPD patterns obtained from the five inorganic compounds.

2.2. Mixture sample preparation and multi-component mixture XRPD

Twelve different multi-component mixtures were prepared by grinding the five compounds in an agate mortar for a few minutes, at room temperature. In order to vary the composition of the mixtures, different weight percentages of the inorganic compounds were used. The weight percentages of each prepared sample are shown in Table 1.

The 12 different mixtures were measured over a period of circa 1 month when free instrument time was available. This issue will be important for discussion involving diffraction pattern re-alignment.

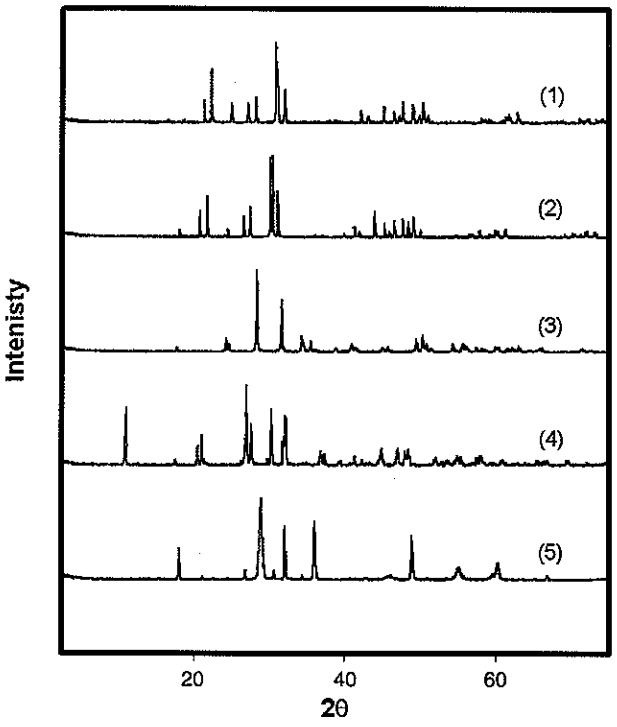


Fig. 1. Reference experimental XRPD patterns (Cu Kα) for: (1) $Pb_3(PO_4)_2$, (2) $Pb_3(PO_4)_3Cl$, (3) ZrO_2, (4) $Pb(OH)_2$ and (5) PbO.

3. Computational section

A newly developed algorithm of band-target entropy minimization for spectral reconstruction was applied to analyze the X-ray powder diffraction data. The consolidated experimental data was first subjected to (i) re-alignment and then (ii) filtering as a data pre-treatment procedure. Then singular value decomposition (SVD) was performed on the pre-treated data matrix. The orthogonal basis vectors that spanned the subspace of observations were transformed into pure component diffraction estimates using the band-target entropy minimization algorithm.

Table 1
Composition of the 12 synthetic mixtures consisting of five components (values are given in wt.%)

Sample	Composition (wt.%)				
	$Pb_3(PO_4)_2$	$Pb_3(PO_4)_3Cl$	ZrO_2	$Pb(OH)_2$	PbO
Mix1	4.07	10.02	48.01	16.08	21.82
Mix2	15.03	9.77	34.66	22.09	18.44
Mix3	30.48	39.36	15.21	10.30	4.64
Mix4	43.96	5.49	9.78	5.97	34.81
Mix5	46.74	16.53	4.09	3.56	29.07
Mix6	21.58	28.77	0.82	39.96	8.87
Mix7	1.02	1.34	24.60	29.10	43.94
Mix8	64.35	21.73	0.00	2.18	11.74
Mix9	20.03	20.03	20.12	20.03	19.79
Mix10	10.00	10.02	35.01	9.98	34.99
Mix11	35.80	34.09	10.23	10.23	9.66
Mix12	10.03	34.90	10.01	10.06	35.00

3.1. Alignment

It is known that shifts in X-ray diffraction data occur due to calibration/alignment experimental error [21]. As mentioned previously, the present data set was acquired over a relatively long period of time, so shifts in data were expected. Indeed, after careful examination of the 12 diffraction mixture patterns, shifts in some characteristic peaks were clearly discernable. This non-stationary characteristic of the data was corrected by pre-processing the data with a re-alignment algorithm. The original raw diffraction data was denoted by the matrix $A^{\text{raw}}_{12\times7251}$. No additional non-stationary characteristics were taken into account.

The Euclidean inner product of two vectors is often used as a measure of similarity. Accordingly, we chose one representative diffraction pattern from our 12 mixtures—this was our fixed "reference" vector. The choice of this representative pattern is not particularly important, indeed, it can even be arbitrary. The important issue is that the reference pattern has at least some features common with all other vectors/patterns. The remaining 11 mixture patterns can be compared one-by-one with the "reference" vector. Mathematically, the more similar the two patterns are, the larger their inner product value would be. The alignment can be optimized by shifting the second vector until a maximum in the inner product is achieved. At this point the "reference" pattern and shifted vector are aligned—that is to say, there is a maximum correspondence between their diffraction maxima. Mathematically, this can be expressed as

$$\max G = \langle v_{\text{ref}}, v_{\text{shifted}} \rangle \quad (1)$$

This procedure was used on all 11 mixture patterns. The final re-aligned vectors and the reference vector were then adjusted for length (note that the overall number of channels will be reduced slightly). After consolidation, the diffraction data can be represented as the matrix $A^{\text{align}}_{12\times7052}$.

3.2. Filtering

The Savitzky–Golay (SG) filtering method [22], also called a digital smoothing polynomial filter, is a widely used filtering technique in spectroscopic data treatment. For normal filtering techniques such as the moving average filter, when one applies a higher filter order, the smoother the results will become, but this is at the cost of distortion of the characteristic features. Ideally, a good SG filter would preserve the primary features such as peak heights and widths while removing a considerable amount of noise.

We chose to implement a SG filtering of the diffraction data using the "golayfilt" function in MATLAB [23]. As known, higher orders of the SG polynomial filter are more suitable at preserving the feature heights and width, but sufficient smoothing is still achieved for most purposes. Minimal distortion is more important than maximal smoothing in the present problem. Accordingly, the polynomial order and the frame size parameter in the SG filter operation were

set as 7 and 41, respectively. The resulting filter data was denoted as $A^{\text{filter}}_{12\times7052}$.

3.3. BTEM

For a wide variety of problems in spectroscopy, the set of observations for a multi-component system, can be modeled (at least locally) as a bilinear product involving the individual constituent characteristics and error:

$$A_{k\times v} = C_{k\times s}a_{s\times v} + \varepsilon_{k\times v} \quad (2)$$

where A is an absorbance/intensity matrix, $C_{k\times s}$ a concentration matrix, and $a_{s\times v}$ a pure component intensity pattern matrix, k the number of mixture patterns, s the number of observable components, v the number of data channels and $\varepsilon_{k\times v}$ the experimental noise and model non-linearities [24]. For a well-posed problem k, the number of mixture patterns must be significantly larger than s, the number of observable components.

The really difficult inverse problem, and the one which concerns us here, is the determination of the intensity pattern matrix $a_{s\times v}$ given absolutely no a priori information. In other words, we want to determine of all patterns $a_{s\times v}$ given A alone. Implicitly, this also necessitates the determination of the number of component involved. This inverse problem is ill-posed.

Mathematically, there will be many solutions for the above inverse problem. To break this ambiguity (ill-posedness) several constraints have been proposed to minimize the set of admissible solutions for the component patterns; these range from rather straightforward constraints such as non-negativity of concentrations and patterns, to more elaborate (and potentially less general) unimodality of concentrations and signals, signal shape assumptions, and kinetic characteristic assumptions. But for a totally blind system identification procedure, it is important to make sure that the search is not over-constrained. Ideally, this would imply that no assumptions what-so-ever are imposed.

Band-target entropy minimization is a self-modeling deconvolution technique based on Shannon information entropy criterion [25]. The novel entropy minimization method used in BTEM is based on the idea that the real pure components are the simplest patterns involved in the data set. Such simple patterns should have the minimum values of entropy in the system of observations. In the case of XRPD patterns, we know that each point in the patterns must be non-negative and this is the only constraint imposed.

The first procedure in BTEM is performing singular value decomposition on the set of spectroscopic data $A_{k\times v}$. The matrix of right singular vectors, V^{T} is the primary object of interest:

$$A_{k\times v} = U_{k\times k}\Sigma_{k\times v}V^{\mathrm{T}}_{v\times v} \quad (3)$$

The experimentalist then identifies local features in the first few several right singular vectors which are of interest. These features are "targeted" one-at-a-time by BTEM. This means that BTEM ensures that each targeted feature is retained in the final reconstruction. Not only is the feature retained, but BTEM recovers the entire function associated with the feature.

To achieve this result, a global optimization must be performed to determine the elements of a transformation matrix T such that minimum entropy and non-negativity are ensured. Thus the primary numerical manipulation is associated with a transformation process from abstract right singular vectors, V^{T} into pure component spectral estimates, $\hat{a}$, where z vectors from V^{T} are used:

$$\hat{a}_{1\times v} = T_{1\times z}V^{\mathrm{T}}_{z\times v} \quad (4)$$

The transformation is governed by the optimization of an objective function corresponding to the final estimate of the pure component spectral estimate, $\hat{a}$. There are two terms in the objective function, one in entropy (H) and one in a penalty (P).

$$F_{\text{obj}} = H + P \quad (5)$$

where H, the entropy value is defined by Eq. (6) and P a penalty in pattern non-negativity. H uses a function h_v which in its classic form is a discrete probability distribution function. Most often h_v involves the derivative amplitude of the estimated pure spectrum in a L^1 norm (maximum intensity equal to 1):

$$H = -\sum_v h_v \ln h_v \quad (6)$$

The final estimate of the pure component spectra, $\hat{a}_{s\times v}$, corresponds to the global minimum of the proposed objective functions. More details about the BTEM concept, developments and applications can be found from [6,10].

4. Results

4.1. XRPD data for 12 mixtures

Seven typical X-ray powder diffraction data sets (samples 1–6 and 12) are shown in Fig. 2. Comparison of this data with the pure component XRPD (from Fig. 1) confirms that there is considerable overlap of diffraction features.

4.2. Re-alignment, filtering and singular value decomposition of 12 mixtures

Due to shifts present in the data acquired over the 1-month measurement period, the raw experimental diffraction data was first re-aligned according to procedure in Section 3.1. The re-aligned data was then filtered using the Savitzky–Golay method as described in Section 3.2. The final pre-treated data was consolidated into a single matrix of dimension $A_{12\times7052}$. Singular value decomposition was performed on the pre-treated data matrix, $A_{12\times7052}$ yielding

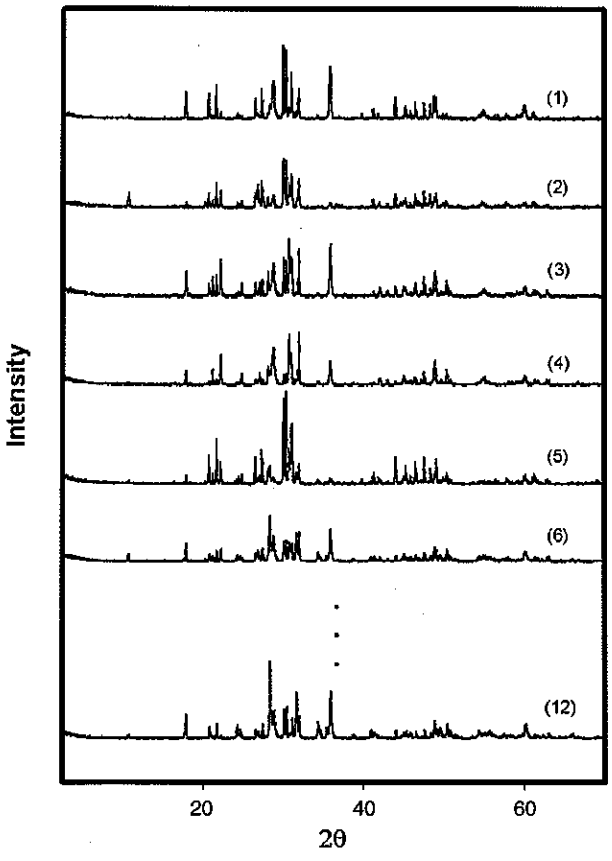


Fig. 2. Experimental X-ray powder diffraction patterns of the mixtures 1–6 and 12.

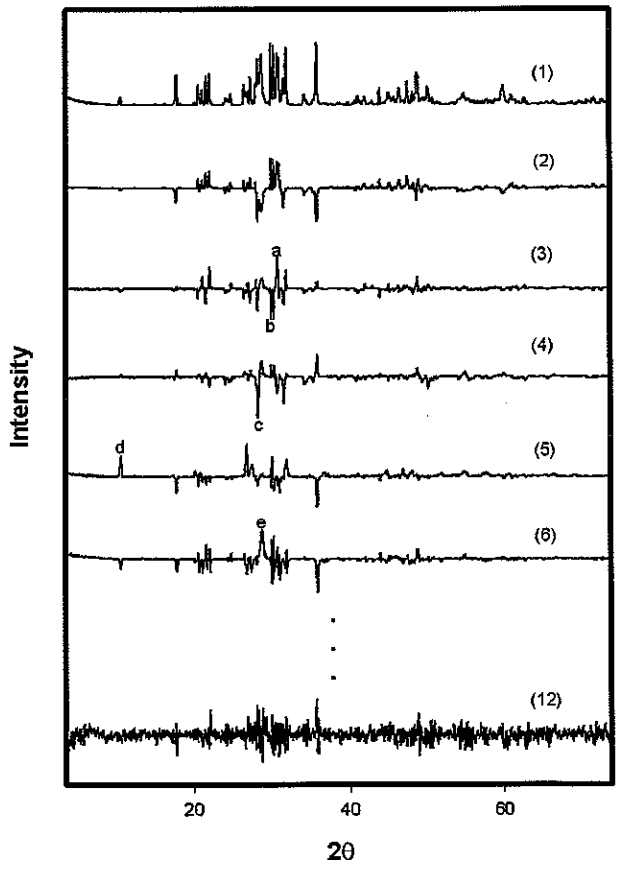


Fig. 3. Plot of the first six as well as the 12th right singular vectors of V^T obtained from singular value decomposition of the experimental mixture diffraction data. First vector is more-or-less an average pattern for the data set. Subsequent vectors account for decreasing contributions to the variance of the data set. The last vector contains primarily heteroscedastic noise. Letters a–e indicate diffraction peaks subsequently targeted by BTEM.

the orthonormal matrices $U_{12\times12}$ and $V^T_{7052\times7052}$, and the diagonal singular value matrix $\Sigma_{12\times7052}$.

The right singular vectors in V^T are the needed input for the BTEM analysis. The first six vectors as well as the 12th vector are shown in Fig. 3. These vectors are ordered in descending contribution to the total variance of the measurements. Thus the first vector possesses the largest contribution to the observations, while the 12th possesses the smallest contribution to the observations. In fact, the 12th vector does not appear to have any prominent physically meaningful features. Instead, the 12th vector seems to be primarily heteroscedastic noise (noise localized at positions of significant signal intensity). Many diffraction extrema in the V^T vectors were targeted for use in BTEM. Five of these targeted bands, labeled a–e are identified in Fig. 3.

4.3. BTEM of 12 mixtures

Clear physically meaningful and prominent diffractions were apparent in only the first 10 of the 12 right singular vectors. Accordingly, only the first 10 vectors from V^T were used in BTEM. The labeled diffractions a–e were used, one-at-a-time, in the BTEM algorithm. Upon completion of each optimized recovery, an entire individual diffraction pattern was recovered, where the chosen targeted feature was retained. The five recovered diffraction patterns are shown in Fig. 4.

The diffraction maxima appearing in the recovered patterns account for all the prominent diffractions observed in the mixtures. Therefore, there do not seem to be any other components present. Many other diffractions in addition to those labeled a–e were also targeted by BTEM. The recovered patterns all belonged to the set represented in Fig. 4. Inspection of the five recovered diffraction patterns, and comparison to the references in Fig. 1 shows that five recovered patterns do, in fact, correspond to the compounds $Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$, and PbO.

The BTEM pure component pattern estimates can be used to solve the dual problem for relative concentrations (see Eq. (2)). Since the BTEM results are actually *normalized*, the exact magnitudes of the pure component patterns are not known. Therefore, only *normalized* concentrations are particularly instructive. Accordingly, the real concentrations from Table 1 were re-normalized and compared to estimates

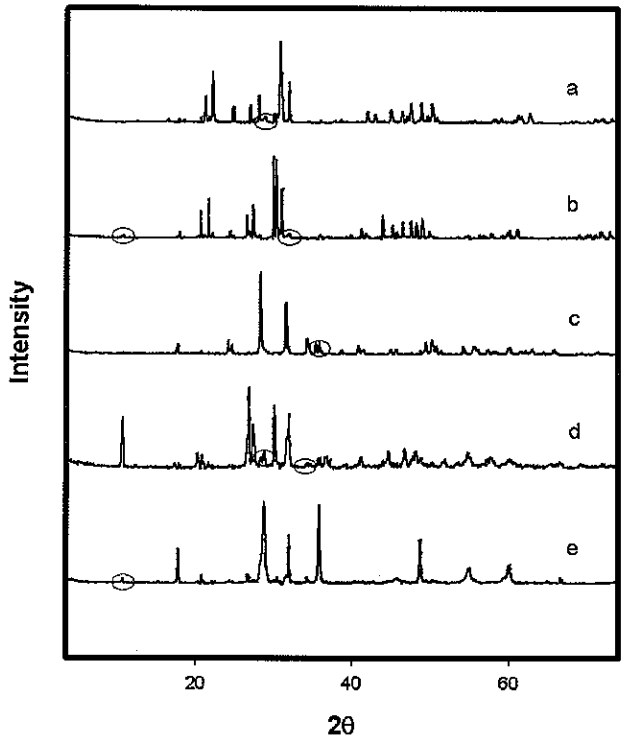


Fig. 4. Pure component diffraction patterns recovered using BTEM using the targeted features a–e in Fig. 3. (a) $Pb_3(PO_4)_2$, (b) $Pb_3(PO_4)_3Cl$, (c) ZrO_2, (d) $Pb(OH)_2$ and (e) PbO. Circled regions indicate small artifacts remaining after optimized pattern reconstruction. These are features which are not seen in reference experimental pure component patterns.

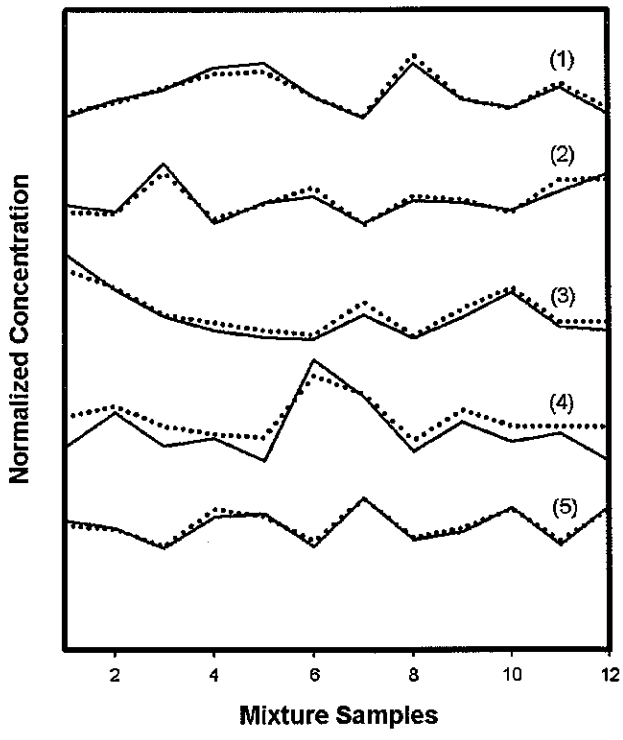


Fig. 5. L2 normed concentrations associated with the 12 mixtures. Dotted line represents experimental powder loadings from Table 1. Solid line represents powder loadings predicted by BTEM. (1) $Pb_3(PO_4)_2$, (2) $Pb_3(PO_4)_3Cl$, (3) ZrO_2, (4) $Pb(OH)_2$ and (5) PbO.

obtained from BTEM and Eq. (2). These results are shown in Fig. 5. Although proper scaling information on the absolute magnitudes of the pure patterns is not known (no calibration was performed), the profiles of the normed concentrations are consistent.

The experimental reference pure component diffraction patterns can be compared to the recovered patterns in a meaningful way. The inner product provides a measure of similarity. If it is zero, the two vectors are orthogonal. If it is unity, the vectors are identical. Comparison of the reference and the recovered patterns for $Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$ and PbO provided values of the inner product of 0.9848, 0.9621, 0.9839, 0.9462, 0.9829, respectively (average equal to 0.9720). These values reconfirm the considerable accuracy of the recovered patterns.

The actual prepared concentrations (C_{real}) were calculated using the normed reference XRD patterns and the estimated concentrations (C_{est}) were calculated using the normed BTEM recovered XRD patterns by solving the least squares problem equation 2. The errors between the actual prepared concentrations (C_{real}) and the estimated concentrations (C_{est}) were 1.9, 1.1, 17.0, 48.5 and 3.6% for species 1–5, respectively (using Eq. (7)). The errors are quite good for species 1, 2 and 5. The non-negligible errors for ZrO_2 and $Pb(OH)_2$ can be readily traced to their weak signals. Indeed, ZrO_2 and $Pb(OH)_2$ contribute only 12.2 and 5.0% of the total integrated intensity obtained from all the 12 measured samples. Accordingly, the noise level in the BTEM estimates (Fig. 4) are greater for these species, and at the same time, accurate estimates of their concentrations using a least squares fit is more difficult due to the low signal intensity. A larger set of samples and measurements would have greatly improved the results:

$$\text{error}\,(\%) = \frac{\sum_{12} |C_{real} - C_{est}|}{\sum_{12} C_{real}} \times 100 \quad (7)$$

4.4. Comparison with SIMPLISMA

A number of "pure component" spectral reconstruction algorithms can be found in the chemometrics literature. Foremost among these algorithms are SIMPLISMA, IPCA and OPA-ALS (see Section 1). All of these algorithms require some sort of a priori estimate of the number of components present—which may severely restrict their applicability in many situations. Nevertheless, we have applied the most common algorithm SIMPLISMA to the present data set including the a priori knowledge that only five components are present. The result is shown in Fig. 6.

As seen from Fig. 6, SIMPLISMA provided a rather good first approximation to the pure component XRD patterns. However, a number of artifacts are present in

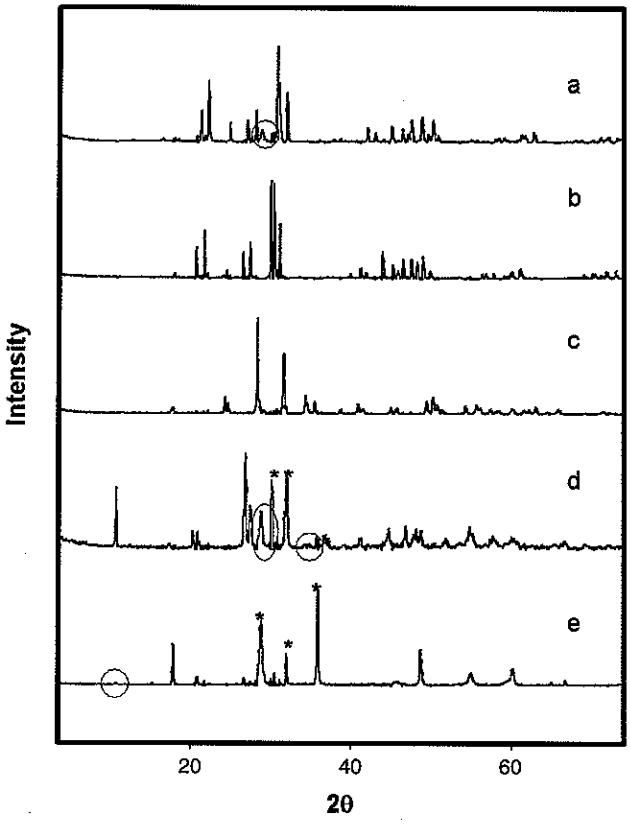


Fig. 6. Pure component diffraction patterns recovered using SIMPLISMA using 5 degrees of freedom. (a) $Pb_3(PO_4)_2$, (b) $Pb_3(PO_4)_3Cl$, (c) ZrO_2, (d) $Pb(OH)_2$ and (e) PbO. Circled regions indicate inconsistencies with reference diffraction patterns. Asterisks (*) indicate sets of significant diffraction intensity whose ratio differ significantly from the reference diffraction patterns.

the reconstructed XRD patterns. These include regions of diffraction intensity which should not be present, as well as incorrect ratios between the intensity of 2 or more diffraction bands. Again, the experimental reference pure component diffraction patterns can be compared to the recovered patterns in a meaningful way. Comparison of the reference and the recovered patterns for $Pb_3(PO_4)_2$, $Pb_3(PO_4)_3Cl$, ZrO_2, $Pb(OH)_2$ and PbO provided values of the inner product of 0.9826, 0.9667, 0.9910, 0.9279, 0.9400, respectively (average equal to 0.9616). These values are in general reasonably good, but not quite as good as the BTEM results.

5. Discussion

The results presented in Section 4 clearly show that recovery of pure component diffraction patterns from mixtures data alone is possible. Diffraction data from an unknown set of mixtures can be measured, decomposed to untangle the signals, and then searched feature-by-feature to obtain a complete set of pure component diffraction patterns. No recourse to libraries nor prior knowledge of the components present is needed what-so-ever. The search for simple underlying patterns is reliable and reproducible.

In the course of analysis, it was found desirable to perform automatic re-alignment of the mixture diffraction patterns. The problem of alignment loss in X-ray diffraction studies is well known. This is often seen as a constant shift in 2θ diffraction angles. There exist a whole host of underlying reasons that can produce such a shift. Although such shifts are typically of very little consequence for the experimentalist who wants to confirm the presence of a phase (qualitative work) etc., such shifts are extremely detrimental in pure pattern recovery by BTEM. The first analysis of the data set without re-alignment was unsatisfactory. Only after re-alignment were superior pure diffraction patterns recovered.

It was also found that filtering of diffraction data was desirable with this relatively small set of 12 mixture patterns. Filtering resulted in smoother right singular vectors after SVD and this in turn is useful during the entropy minimization procedure. Extremely noisy right singular vectors lead to poor reconstructed patterns. It should be noted, however, that need for filtering is probably by-passed all together if a significantly larger number of experimental observations can be made. A large number of observations lead to smoother primary right singular vectors and a larger number of secondary right singular vectors that can be discarded before BTEM. Indeed, this conclusion is consistent with our previous results using hundreds of FTIR spectra in other studies [8,12]—more observations make the pattern recovery more accurate. In this regard, it can be noted again that the recovered pure component diffraction patterns shown in Fig. 4 show very minor signal artifacts. Mixed-in signals, such as those seen in the circled regions of Fig. 4, typically arise because there is not enough information in the primary vectors in order to achieve the level of pattern approximation really desired. More data leads to better primary vectors and consequently to really efficient entropy minimization. Finally, as more measurements are made, minute signals can be recovered. For example, pure component patterns with less than 0.1% of the total integrated signal of the data set have been recovered [12].

This discussion can most appropriately be concluded by referring back to the difficult inverse problems mentioned in Section 1. In particular, we focus on in situ heterogeneous surface catalyzed reactions. It is a well accepted fact that heterogeneous catalysts restructure under working conditions, and new phases arise. Given the temperatures typically used, these phase changes can be rather rapid. An elegant application of the present method would be the in situ X-ray analysis of a heterogeneous working catalyst where the diffraction patterns can be recorded in rapid succession, for example, by using a high intensity X-ray source (synchrotron). In such a situation, the catalyst could be observed under a wide variety of operational conditions (partial pressures of gases, temperature, total pressure, etc.). Since 100s if not 1000s of X-ray diffraction patterns could be measured in a reasonable amount of time, subsequent BTEM analysis could be anticipated to recover the patterns of very minor phases present.

In this regard, strong metal support interaction (SMSI) is particularly intriguing. BTEM analysis as applied to XRPD material science problems with unknown constituents can be viewed as a material systems identification procedure.

6. Conclusions

An easy to implement and general solution to the inverse problem for pure component pattern reconstruction from X-ray diffraction data has been presented. Although tested on a simple laboratory system, the procedure appears applicable to a host of difficult in situ material science problems with unknown constituents. The procedure relies on entropy minimization and global optimization as the primary numerical tools. The results from the pattern searches for this particular problem are in excellent agreement with references. The BTEM method relies on no a priori information what-so-ever. The matrix of mixture diffraction data is the only input.

Acknowledgements

The lead inorganic samples used in this study were provided by Professor Robert Stanforth of the Department of Chemical and Environmental Engineering at the National University of Singapore.

References

[1] W. Windig, J. Guilment, Anal. Chem. 63 (1991) 1425.
[2] W. Windig, Chemom. Intell. Lab. Syst. 36 (1997) 3.
[3] D.S. Bu, C.W. Brown, Appl. Spectrosc. 54 (2000) 1214.
[4] F.C. Sanchez, J. Toft, B. Van den Bogaert, D.L. Massart, Anal. Chem. 68 (1996) 79.
[5] P.C. Sabatier, Applied Inverse Problems, Springer-Verlag, Berlin, 1978.
[6] W. Chew, E. Widaja, M. Garland, Organometallics 21 (2002) 1982.
[7] C. Li, E. Widjaja, M. Garland, J. Catal. 213 (2) (2003) 126.
[8] C. Li, E. Widjaja, M. Garland, J. Am. Chem. Soc. 125 (2003) 5540.
[9] L.R. Ong, E. Widjaja, R. Stanforth, M. Garland, J. Raman Spectrosc. 34 (4) (2003) 282.
[10] E. Widjaja, Ph.D. Thesis, National University of Singapore, 2002, Chapter 7, pp. 257.
[11] H.J. Zhang, M. Garland, Y.Z. Zeng, P.J. Wu, Am. Soc. Mass Spectrum 14 (2003) 1295.
[12] C. Li, E. Widjaja, W. Chew, M. Garland, Angew. Chem. I.E. 20 (2002) 3785.
[13] S. Hagopian-Babikian, R.F. Hamilton, S.S. Iyengar, R. Jenkins, J. Renault, General introduction, in: V.E. Buhrke, Ron. Jenkins, D.K. Smith, A Practical Guide for the Preparation of Specimens for X-ray Fluorescence and X-ray Diffraction Analysis, Wiley, pp. 1–34.
[14] M. Ikeda, S.Y. Komatsu, Y. Nakamura, Mater. Trans. 43 (12) (2002) 2984.
[15] L. Marosi, C.O. Arean, J. Catal. 213 (2) (2003) 235.
[16] J.A. Rodriguez, J.Y. Kim, J.C. Hanson, M. Perez, A.I. Frenkel, Catal. Lett. 85 (3–4) (2003) 247.
[17] S. Velu, K. Suzuki, C.S. Gopinath, J. Phys. Chem. B 106 (49) (2002) 12737.
[18] X.H. Wu, Y.D. Wang, Y.F. Li, Z.L. Zhou, Mater. Chem. Phys. 77 (2) (2003) 588.
[19] G.L. Perlovich, L.K. Hansen, A. Bauer-Brandl, J. Pharm. Sci. 91 (4) (2002) 1036.
[20] J.M. Llacer, V. Gallardo, R. Delgado, J. Parraga, D. Martin, M.A. Ruiz, Drug. Dev. Ind. Pharm. 27 (9) (2001) 899.
[21] R. Jenkins, R.L. Snyder, Introduction to X-ray Powder Diffractometry, Wiley, New York, Chapter 8, pp. 205.
[22] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (8) (1964) 1627.
[23] MathWorks Inc. MatLab Reference Guide, 1995.
[24] M. Garland, E. Visser, P. Terwiesch, D.W.T. Rippin, Anal. Chim. Acta 351 (1–3) (1997) 337.
[25] C.E. Shannon, Bell Syst. Tech. J. 3 (1948) 379.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 39 (2006) 1066–1073

Appendix E

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

The use of entropy minimization for the solution of blind source separation problems in image analysis

Liangfeng Guo, Marc Garland*

Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576, Singapore

Received 28 June 2005

Abstract

Entropy minimization is closely associated with pattern recognition. The present contribution uses a direct minimization of an entropy like function to solve the blind source separation problem for image reconstruction. The mixture patterns are decomposed using SVD and then global stochastic optimization is used to find the first irreducible image pattern. Further images are then subsequently reconstructed, by imposing a 2D correlation coefficient for dissimilarity to prevent repeated images, until all images are exhaustively enumerated. Three test cases are used, including (1) a set of three black and white texturally different photographs (2) a set of three RGB geometrically similar photographs and (3) an underdetermined problem involving an imbedded watermark. Cases 1 and 2 are easily solved with outstanding image quality. Both searches are conducted in an unsupervised manner—no a priori information is used. In Case 3, the watermark is enhanced after targeting the region for entropy minimization. The present results have a wide variety of applications, including image and spectroscopic analysis.

Keywords: Information entropy; Inverse problems; Image processing; Blind source separation; Entropy minimization

1. Introduction

Blind source separation problems typically involve 1D and 2D data arrays and are found in a wide variety of disciplines [1]. They represent a particularly difficult type of inverse problem [2], where the observables are superpositions/mixtures of source patterns. In the most difficult form of such problems, no a priori information is commonly available concerning the individual source patterns nor the number of sources giving rise to the observations.

Images represent the most typical 2D data arrays analyzed. A variety of approaches have been taken to solve the associated blind source separation problem in the electrical engineering and image processing literature. These include high order statistics [3], mutual information maximization approach [4] (Infomax), and nongaussianity approach (FastICA) [5], etc.

Although information entropy is widely used to characterize signals [6], and it is recognized that entropy is closely associated with pattern recognition [7], its direct implementation has until recently found rather limited use in the solution of blind source separation problems. In the chemical sciences, particularly for spectroscopic problems, information entropy has become increasingly used with considerable success. In such problems, spectra of inseparable chemical mixtures are frequently encountered, and the need to identify pure component spectra arises. Entropy minimization as a tool for recovering 1D spectra was apparently first used by Sasaki and coworkers [8]. Since then, the approach has been refined considerably, resulting in an algorithm called band-target entropy minimization (BTEM), and applied to 1D spectroscopic arrays, namely, infrared spectroscopy [9,10], Raman spectroscopy [11,12], mass spectroscopy [13], NMR [14] and X-ray diffraction [15]. These studies have shown that excellent estimates of pure source spectra can be recovered even for signals constituting less than 1% of the total observable signal, and this approach has helped to solve

* Corresponding author. Tel.: +65 6 874 6617; fax: +65 6 779 1936.
E-mail address: chemvg@nus.edu.sg (M. Garland).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.09.006

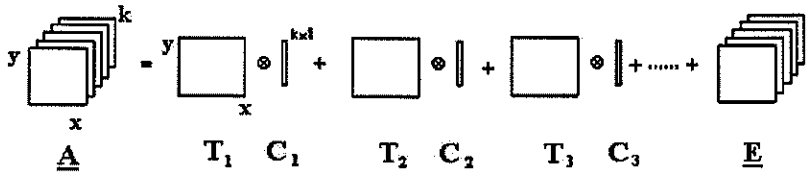


Fig. 1. A multi-way data can be decomposed into a sum of Kronecker products and a residual $\underline{E}$.

a number of previously intractable problems where more than a dozen components are simultaneously present [16–18]. Since increased signal entropy is associated with randomness and disorder, the search for entropy minimization has become synonymous with the search for simplicity [19].

Very recently, this entropy minimization approach has been generalized and further extended, in order to treat sets of 2D spectroscopic data arrays [20]. A new entropy like function was defined that more fully utilizes the inherent organization found in 2D arrays. This approach was then successfully applied to 2D nuclear magnetic resonance (NMR) spectra, specifically 2D COSY and HSQC data. The 2D source patterns were recovered with a very high degree of accuracy, without the use of any a priori information whatso-ever.

In the present contribution we apply for the first time this 2D entropy minimization approach to three different image blind source separation problems. The first case involves three texturally dissimilar black and white images (photographs). The second involves three geometrically similar color images (photographs). The third case involves image enhancement for an underdetermined problem. In the first two cases, outstanding blind source recover is achieved and in the last case significant image enhancement is observed.

2. Methods of investigation

2.1. Overview of approach

The basis philosophy behind the present methodology has three primary parts. First, given an arbitrary set of observations (data array), which may even lead to an overdetermined problem, this array should be decomposed into orthogonal components using PCA (principal component analysis) or SVD (singular value decomposition) etc. Secondly, an objective function should be created using an appropriate entropy like function with a penalty term which hinders the repeated recovery of images which are too similar in appearance. Third, use a reliable global optimizer such as simulated annealing (SA) to achieve the objective function minima. The sequential evaluations of these minima correspond to the pure source images present in the observed set of mixed images. All of the following simulations and image reconstructions were conducted using in-house code developed in MATLAB [21].

2.2. Decomposition of image set

Decomposition of sets of multivariate images is often implemented using PCA (also known as the Karhunen–Loève transform [22]). PCA is often used for finding the latent variables in a data set and thus determine the underlying multivariate information. PCA allows the exploration/analysis of patterns in the data set. The first description of PCA was apparently made by Cauchy in 1829 [23]. It was later developed further by Pearson [24] and Hotelling [25]. For rather thorough reviews of the issues see the books by Joliffe [26] and Jackson [27].

The PCA transform will produce a series of the components with the highest possible variance. In the present case of image analysis, three-way data sets are involved. In other words, the set of observations consist of stacks of 2D *mixture* data. As shown in Eq. (1) and Fig. 1, we can decompose a three-way data set into three parts.

$$\underline{A} = \sum_{j=1}^{n} T_j \otimes C_j + E. \tag{1}$$

The three-way matrix algebra equation Eq. (1) has the following terms; $\underline{A}$ denotes the series of k mixture images ($\underline{A}$ is a three-way array), T_j denotes matrix-formatted component j with size $(x \times y)$; C_j denotes the loading for component j (relate to its contribution) which is a vector of length k; $\otimes$ denotes the Kronecker product; and $\underline{E}$ is the residual part. Ideally most of the information in the mixture images will be concentrated in the first term of Eq. (1), but in real sets of data there is always noise which shows up in the residual $\underline{E}$. Details of the PCA on three-way array data decomposition can be found in Geladi and coworkers' paper [28]. A PCA decomposition can be conveniently calculated using SVD technique, which possesses a more robust and efficient process [29,30]. In SVD, the unfolded observation matrix A (two-way arrary) can be decomposed into three parts: $A = U\Sigma V^T$, where the rows of V^T contain the elements of the *right singular vectors*. These *right singular vectors* can be folded into the right singular matrices which are analogous to T in the PCA decomposition mode. SVD is used in the present contribution.

2.3. Reconstruction based on entropy minimization

2.3.1. Goal

The series of T_j matrices can always be recombined, as a linear combination, to produce a new image I. In most cases, linear combinations of the T_j matrices will yield another superposition of source images (like the original k mixture images). In a few particular cases, (particular linear combinations) the new image I will in fact be one of the original source images. The primary difficulty is to choose an algorithm that permits the estimation of the linear combinations with weights s_j which result in the pure source images without the use of any a priori information.

$$I = \sum T_j s_j \tag{2}$$

2.3.2. Entropy function

Entropy minimization could provide a proper setting for constraining an unsupervised search of the image space. Assuming that a proper entropy like function can be formulated, then the pure source images would have the lowest entropies (global and local minima) and their mixtures would have higher entropy values. Stated in other terms, the more complex mixture images can be simplified to irreducible source images.

At this juncture, it is possible to consider two alternative representations for the images I, namely their 1D vector representation obtained by concatenation of each row and their native 2D matrix format. The former case of 1D vector representation does not take full advantage of the information available, whereas the native 2D matrix format provides the opportunity to use the image structure to its fullest. Accordingly, only the latter case will be treated.

A working approximation for the entropy H used in matrix-based 2D entropy minimization can take the form of Eq. (3), where $\hat{I}_{xy}$ is the estimate of the new image I_{xy} at each updated iteration. Here, consistent with Sasaki and coworkers' original suggestion for 1D data, the probability distribution p is replaced by a derivative of the data [31]. In the present adaptation to matrix formatted data, and in order to take advantage of the natural image structure, we take the second derivative (Eq. (4)). In other words, the entropy like function H used in matrix-based entropy minimization estimates the smoothness of the images in 2D.

$$H^{2D} = -\sum_x \sum_y h_{xy} = -\sum_x \sum_y p_{xy} \ln p_{xy}, \tag{3}$$

$$p_{xy} = \frac{\left| d^2 \hat{I}_{xy} / \mathrm{d}x\, \mathrm{d}y \right|}{\sum \left| d^2 \hat{I}_{xy} / \mathrm{d}x\, \mathrm{d}y \right|}. \tag{4}$$

2.4. Objective function formulation and optimization

The resolution of the pure source images can be achieved by solving the following minimization problem.

$$\min F_{obj} = H^{2D} + P. \tag{5}$$

Specifically, the objective function F_{obj} includes the entropy term H^{2D} along with a penalty function P (infra vida). The transformation of T into an estimate of a pure source image I is achieved by Eq. (2) but governed by the optimization algorithm used. For a well-behaved function with simple response surface, gradient-based solvers are rather efficient and suitable algorithms. But in high-dimensional problems where the gradients are not easy available, random search methods are favored. The global minimization of the objective function is, in principle, achievable using SA or a genetic algorithm (GA) which are both stochastic search techniques [32,33].

In the first optimization to recover the first pure source image, P is set to zero, and the global entropy minimum is sought. In each subsequent search, the penalty function is formulated such that only images dissimilar to all previous source image results are admissible.

This effect can be achieved by adding a dissimilarity penalty function, P, to the image reconstruction objective function. Such an admissible function is shown in Eq. (6) where *corr*2 denotes the 2D correlation coefficient between any two 2D arrays A and B [34].

$$P = \max\left\{ p \times \left(e^{q/(1-corr\,2)} - 1 \right) \right\}. \tag{6}$$

If A and B are identical, then the value of *corr*2 will be unity. In the practice, the images A and B used are the estimated image during the current optimization and a previously determined source image(s), respectively. For more similar images A and B, the bigger the values of *corr*2 and thus the argument on the right-hand side of Eq. (6). By taking the maximum of the set of values of the 2D correlation coefficient between all combinations of A and B, the penalty function prevents any identical reconstruction from occurring in subsequent optimizations. The penalty function adopted here is a modified *sigmoid* function with parameters $q = 0.0002$ and $p = 100000$ which rapidly increases when A and B become similar. This allows distinct and texturally dissimilar source images to be reconstructed.

3. Results

3.1. Analysis of texturally different images

Three texturally different images, consisting of a building, a fabric and a tile, were downloaded from the MIT Vistex public database (available from http://vismod.media.mit.edu/pub/VisTex/VisTex.tar.gz). All the images were 128 × 128 pixels. Original files contain truecolor (RGB) images, with each image file composed of three layers. Without loss of generality, the "Red" color layer data were used as the pure image. These images in black and white are shown in Fig. 2. The values of the matrix-wise entropies for the images in Fig. 2 were 533.27 (building), 570.50 (fabric) and 531.59 (tile).

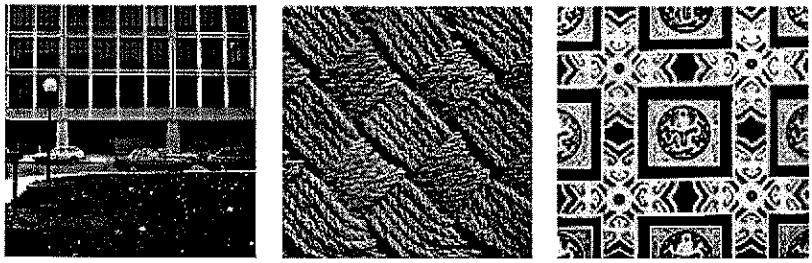


Fig. 2. Original black and white images from MIT database.

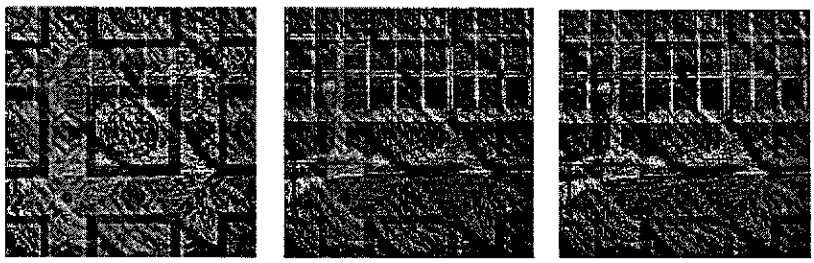


Fig. 3. Mixture images obtained from mixing matrix A.

Fig. 4. Recovered images.

A random mixing matrix with non-negative entries was generated. This mixing matrix is shown in Eq. (7).

$$A = \begin{pmatrix} 0.3202 & 0.7200 & 0.5479 \\ 0.8207 & 0.8876 & 0.3240 \\ 0.6186 & 0.6395 & 0.2308 \end{pmatrix}. \quad (7)$$

The three corresponding images are shown in Fig. 3. The values of the matrix-wise entropies for the superimposed images in Fig. 3 were 569.3511, 570.3221 and 570.2690, respectively.

After unfolding each image, SVD was applied to the $3 \times (128 \times 128)$ matrices. This set of right singular matrices was then taken, and a global search using SA method was used to find the image with minimum entropy. This image was the building. At this point, further non-similar images were sought using the search approach outlines in the section *objective function formulation and optimization.* The resulting set of three images, one with a global entropy minimum and two others with local entropy minima are shown in Fig. 4. The values of the matrix-wise entropies for the images in Fig. 4 were 531.55 (tile), 533.27 (building) and 570.52 (fabric), respectively.

The reconstructed images in Fig. 4 are obviously very similar to the original images in Fig. 2. If the recovered images are mapped back onto the mixture data, the resulting mixing matrix can be obtained.

In order to compare the quality of the reconstructions, the original and recovered final mixing matrices are compared. After column permutations to achieve correspondence in ordering of the images used, and after column re-normalization, the mixture matrices are given in Eqs. (8) and (9). It is clear that the original and calculated mixing matrices are very similar.

$$A_{original} = \begin{pmatrix} 0.3901 & 0.8111 & 1.0000 \\ 1.0000 & 1.0000 & 0.5912 \\ 0.7537 & 0.7205 & 0.4213 \end{pmatrix}, \quad (8)$$

$$A_{calculated} = \begin{pmatrix} 0.4017 & 0.8129 & 1.0000 \\ 1.0000 & 1.0000 & 0.5963 \\ 0.7524 & 0.7205 & 0.4250 \end{pmatrix}. \quad (9)$$

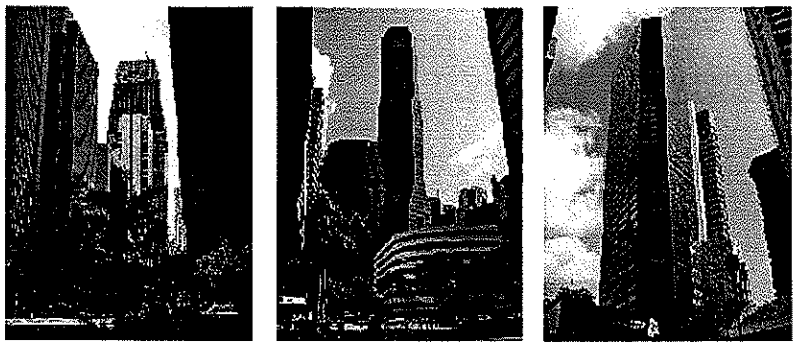


Fig. 5. Original images in color. PWC Building (left), Republic Building (center), CapitaLand Building (right).

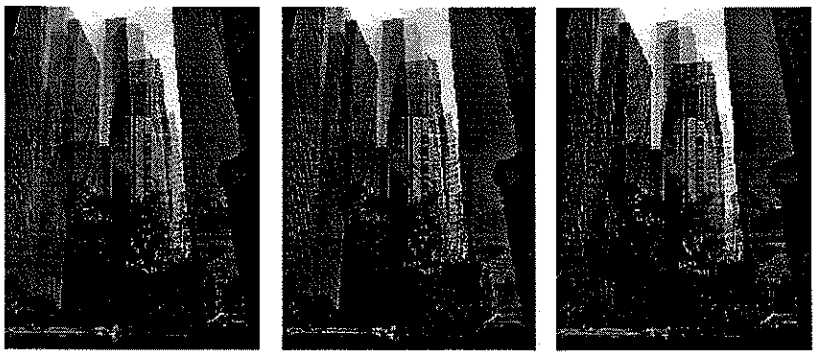


Fig. 6. Mixture images obtained from mixing matrix A defined in Eq. (10).

For completeness, it should be mentioned that the present reference images were not randomly chosen. Indeed, the same images have been used by other researchers as a test case for the development of other blind source separation approaches. Hashimoto applied a variation of independent component analysis (ICA), employing moment analysis, to these same images [35]. Although the recovered images are reasonably good, the final recovered images possess quite a few visible inconsistencies, and an error rate of 6.2% is stated.

3.2. *Analysis of geometrically similar images*

Three geometrically similar images, consisting of buildings in Singapore, were used as a more difficult test of the proposed image recovery approach. The images chosen are all geometrically similar because a tall skyscraper is centered in the image, both sides of the image are bordered by adjacent buildings, and blue sky is the background. Original images files were in truecolor (RGB), and contained $256 \times 192 \times 3$ pixels. The original files were used in their entirety. These color images are shown in Fig. 5. The values of the matrix-wise entropies for the images in Fig. 5, where the entropies are summed over the three RGB layers were 3411.0, 3364.4 and 3386.9, respectively [36].

A random mixing matrix with non-negative entries was generated. This mixing matrix is shown in Eq. (10).

$$A = \begin{pmatrix} 0.6449 & 0.3420 & 0.5341 \\ 0.8180 & 0.2897 & 0.7271 \\ 0.6602 & 0.3412 & 0.3093 \end{pmatrix}. \tag{10}$$

The three corresponding super-imposed images are shown in Fig. 6. The values of the matrix-wise entropies for the superimposed images in Fig. 6, where the entropies are summed over the three RGB layers, were 3590.9, 3576.6 and 3583.7, respectively.

Each image was unfolded three times according to the three RGB layers. After unfolding each image, SVD was then applied separately to each of the RGB data sets, and hence three times to $3 \times (256 \times 192)$ matrices. To start, the red set of right singular matrices were taken, and a global search using SA was used to find the image with minimum entropy. At this point, further non-similar images were sought using the search approach outlined in section 3.1. This resulted in a set of three red images, one with a global entropy minimum and two others with local entropy minima. The same general approach was then used for the green and blue data sets independently. The red, green and blue data sets for each image were then consolidated to generate the three reconstructed truecolor images each with $256 \times 192 \times 3$

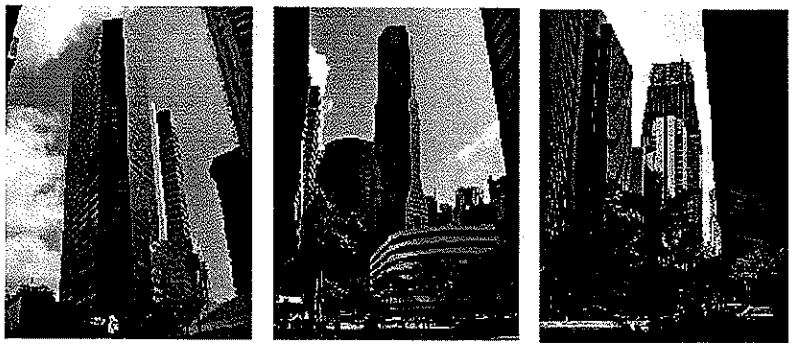


Fig. 7. Reconstructed images in color.

Table 1
The entropies of different layers for different Singapore building photos

Building name	Matrix-wise entropies for the images			
	Red layer	Green layer	Blue layer	Sum
PWC building	1134.8	1131.9	1144.3	3411.0
Republic building	1129.3	1114.5	1120.5	3364.4
CapitaLand building	1133.7	1121.5	1131.7	3386.9

pixels shown in Fig. 7. The values of the matrix-wise entropies for the images in Fig. 7 were 3386.9, 3364.4 and 3411.1, respectively.

The entropies of the different layers for the different buildings are shown in Table 1. The Republic Building had the lowest entropies in each layer as well as in the sum. Sensitivity to the global search was observed with the present images. Although the CapitaLand Building image has the second lowest entropy in all three RGB layers, its image was always recovered first, even though a number of GA or SA searches were conducted. This suggests that in this problem, the search space was very complex. Indeed, if the initial starting point for the search was made in the vicinity of the Republic Building image, the global entropy minimum was easily obtained.

The reconstructed images in Fig. 7 are obviously very similar to the original images in Fig. 5. The red images were used to reconstruct the mixing matrix A_{red} for this set of data. The same general approach was then used for the green and blue data sets independently, resulting in mixing matrices A_{green} and A_{blue}. The resulting mixing matrices are also very similar. The accuracy of the recovered images can be further appreciated by examining a 5×5 pixel blue layer at the interface between a building and the sky. For this purpose, the right roof corner of the PWC Building is used at pixel locations $x = 116$–120, and $y = 214$–218. Fig. 8 provides an expanded view of the image region under consideration in RGB and the corresponding values of the original and recovered pixels in the blue layer.

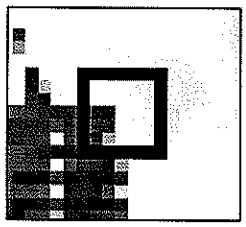
3.3. The underdetermined problem

The three original geometrically similar building images from Singapore were taken again. A new image, consisting of the capital letters NUS, was created as a watermark with matrix-wise entropy of 174.4975 (Fig. 9a). This watermark was imbedded at a 10% level into each of the three mixture images. An example of a mixture image with a 10% watermark is shown in Fig 9b. It is important to note that the watermark is not really discernable at this level of imprinting. The three mixture images with imbedded watermarks were separated into red, green, blue and SVD performed.

At this point, the previously mentioned BTEM algorithm (see introduction) which has been used for spectroscopic pattern recognition was applied to the underdetermined problem. In order to simplify the problem statement and search space, we permit ourselves to know a priori that there is some sort of interesting feature in the lower left region near pixels $x = 45$–46 and $y = 40$–43 (this located in the upper part of the letter U) which needs to be retained and enhanced after image recovery. BTEM was then applied to the right singular matrices. The resulting recovered image is shown in Fig 9c. The watermark is now more prominent, but it cannot be fully recovered due to the underdetermined nature of the problem. In fact, the background is not black, but instead, the Republic Building which has the lowest entropy of the three building images is recovered. The matrix-wise entropy of the Fig. 9c is 3367.3 which is slightly higher than the entropy of the Republic Building alone. The end result, and the one which is important in the present context, is that there is enhancement of the watermark.

4. Discussion

The results of the determined problem indicated that the blind source separation problem can be solved, with outstanding image reconstructions, using the methodology proposed. Also, the underlying aspects of this approach are rather simple and straightforward, namely, (1) define an appropriate entropy type objective function, (2) define an



(a)

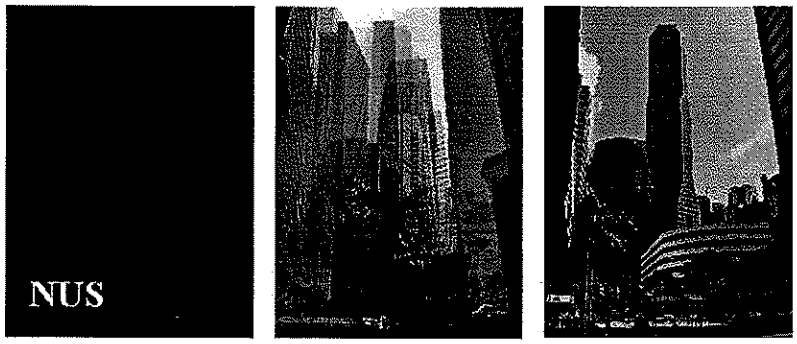
253	255	254	254	252
255	255	255	254	253
148	142	254	253	252
154	170	255	254	254
140	207	255	255	252

(b)

252.9991	254.9991	253.9991	253.9991	251.9991
254.9991	254.9991	254.9991	253.9991	252.9991
147.9991	141.9991	253.9991	252.9991	251.9991
153.9991	169.9991	254.9991	253.9991	253.9991
139.9991	206.9991	254.9991	254.9991	251.9991

(c)

Fig. 8. Right roof corner of PWC Building in RGB (a) and corresponding blue layer values in the original (b) and recovered (c) images for this 5×5 pixel region.



(a) (b) (c)

Fig. 9. Watermark (a). An example of a mixture image with a 10% watermark (b). The resulting recovered image (c).

appropriate measure to hinder replicate images, and (3) perform a one-image-at-a-time global search. The utility of this approach was further shown using the associated underdetermined problem, where image enhancement was achieved.

The entropy type function and objective function chosen in the present contribution are not, by any means, the only choices available. Indeed, there are a multitude of entropy [19,37], and entropy type functions that could be used or at least tried on various classes of images. In this contribution, the simple derivative function used in the present entropy type function, makes use of nearest neighbor pixel information. By minimizing the pixel-to-pixel variations (derivative or local curvature), one is minimizing the randomness between pixel values, and hence strongly enforcing the retention of a structured variation (a pattern).

Although the derivative based entropy like function chosen in the present contribution worked well, one somewhat unexpected finding was obtained. The random mixture images in Fig. 3 possessed entropy values less than one of the pure images (fabric). This appears to be an inherent possibility in dense image sets evaluated with a derivative based function—superpositions of pure patterns may be less "random" than an original pure image. This issue needs to be raised since this situation was not observed to arise in sparse spectroscopic NMR data sets [20]. Superpositions of sparse data sets appear to lead to mixture images whose entropies are generally greater than any of the pure images present.

The concept of a one-image-at-a-time global search greatly simplifies a number of issues. First numerically, a single image reconstruction is simpler than an N-image simultaneous reconstruction. Secondly, constraints to prevent redundant image reconstructions in sequential searches are easier to implement than all constraints simultaneously in an N-image rotation. Third, the user does not need to specify, a priori, the member of pure images to be recovered. Instead, images are recovered until all image data have been accounted. In the present approach a non-negativity constraint was unnecessary with the image data.

5. Conclusion

In the present contribution, a rather straightforward solution using an entropy like formalism, is proposed for the image-mixture blind source separation problem. The methodology, which involves a one-image-at-a-time global search approach, was successfully applied to a set of texturally dissimilar images and geometrically similar images with outstanding image reconstruction quality. Further extensions, such as imbedded image enhancement was also briefly explored.

References

[1] J.F. Cardoso, Blind signal separation: statistical principles, Proc. IEEE. 86 (10) (1998) 2009–2025.
[2] P.C. Sabatier, Introduction to applied inverse problem, in: P.C. Sabatier (Ed.), Applied Inverse Problems: Lectures Presented at the RCP 264 in Montpellier, Springer, Berlin, Heidelberg, New York, 1978, pp. 1–27.
[3] J.F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, IEE. Proc-F 140 (6) (1993) 362–370.
[4] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1004–1034.
[5] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Networks 10 (3) (1999) 626–634.
[6] T. Cover, J. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[7] S. Watanabe, Pattern recognition as a quest for minimum entropy, Pattern Recognition 13 (1981) 381–387.
[8] K. Sasaki, S. Kawata, S. Minami, Estimation of component spectral curves from unknown mixture spectra, Appl. Opt. 23 (1984) 1955–1959.
[9] W. Chew, E. Widjaja, M. Garland, 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 (2002) 1982–1990.
[10] E. Widjaja, C. Li, M. Garland, Semi-batch homogeneous catalytic in-situ spectroscopic data, FTIR spectral reconstructions using band-target entropy minimisation (BTEM) without spectral preconditioning, Organometallics 21 (2002) 1991–1997.
[11] L.R. Ong, E. Widjaja, R. Stanforth, M. Garland, Fourier transform Raman spectral reconstruction of inorganic lead mixtures using a novel band-target entropy minimization (BTEM) method, J. Raman Spectrosc. 34 (4) (2003) 282–289.
[12] S.Y. Sin, E. Widjaja, L.E. Yu, M. Garland, Application of FT-Raman and FTIR measurements using a novel spectral reconstruction algorithm, J. Raman Spectrosc. 34 (10) (2003) 795–805.
[13] H.Z. Zhang, M. Garland, Y.Z. Zeng, P. Wu, Weighted two-band target entropy minimization for the reconstruction of pure component mass spectra: simulation studies and the application to real systems, J. Am. Soc Mass Spectrum. 14 (2003) 1295–1305.
[14] E. Widjaja, M. Garland, Entropy minimization and spectral dissimilarity curve resolution technique applied to nuclear magnetic resonance data sets, J. Magn. Reson. 173 (2005) 175–182.
[15] L.F. Guo, F. Kooli, M. Garland, A general method for the recovery of pure powder XRD patterns from complex mixtures using no a priori information - Application of band-target entropy minimization (BTEM) to materials characterization of inorganic mixtures, Anal. Chim. Acta 517 (1–2) (2004) 229–236.
[16] C.Z. Li, E. Widjaja, W. Chew, M. Garland, Rhodium tetracarbonyl hydride: the elusive metal carbonyl hydride, Angew. Chem. I. E. 20 (2002) 3785–3789.
[17] C.Z. Li, E. Widjaja, M. Garland, The Rh-4(CO)(12)-catalyzed hydroformylation of 3,3-dimethylbut-1-ene promoted with HMn(CO)(5). Bimetallic catalytic binuclear elimination as an origin for synergism in homogeneous catalysis, J. Am. Chem. Soc. 125 (18) (2003) 5540–5548.
[18] A.D. Allian, M. Garland, Spectral resolution of fluxional organometallics. The observation and FTIR characterization of all-terminal [Rh4(CO)12], Dalton Trans. 11 (2005) 1957–1965.
[19] J.N. Kanpur, in: Maximum-Entropy Models in Science and Engineering, Wiley Eastern Ltd, New Delhi, 1993, p. 3.
[20] L.F. Guo, A. Wiesmath, P. Sprenger, M. Garland, Development of 2D band-target entropy minimization and application to the deconvolution of multicomponent 2D nuclear magnetic resonance spectra, Anal. Chem. 77 (6) (2005) 1655–1662.
[21] MATLAB 6.5, The Mathworks Inc., USA.
[22] C.W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, Oxford, 1968.
[23] A.L. Cauchy, Oeuvres IX (2) (1829) 172–175.
[24] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.
[25] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psy. 24 (1933) 417–441, 498–520.
[26] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[27] J. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[28] P. Geladi, H. Isaksson, L. Lindqvist, S. Wold, K. Esbensen, Principal component analysis of multivariate images, Chem. Int. Lab. Sys. 5 (1989) 209–220.
[29] G. Golub, W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, SIAM. J. Numer. Anal. (B) v2 (1965) 205–221.
[30] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420.
[31] Y. Zeng, M. Garland, An improved algorithm for estimating pure component spectra in exploratory chemometric studies based on entropy minimization, Anal. Chim. Acta 359 (3) (1998) 303–310.
[32] A. Corana, M. Marchesi, C. Martin, S. Ridella, Minimizing multimodal functions of continuous-variables with the "simulated annealing" algorithm, ACM. Trans. Math. Software 13 (1987) 262–280.
[33] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[34] R.V. Hogg, E.A. Tanis, Probability and Statistical Inference, fourth ed., Maxwell Macmillan International, New York, 1993.
[35] W. Hashimoto, Separation of independent components from data mixed by several mixing matrices, Signal. Process. 82 (12) (2002) 1949–1961.
[36] These images are archived and downloadable at < http://www.chee.nus.edu.sg/research/chbe_freeware.html >
[37] B.R. Frieden, Image Enhancement and Restoration, in: T.S. Huang (Ed.), Topics in Applied Physics, vol. 6, Springer, New York, 1975, pp. 179–246.

About the author—LIANGFENG GUO received his BS degree in chemical engineering from Tianjin University in 1994, and subsequently worked in chemical process development. He has been a Ph.D. candidate since 2001 at National University of Singapore where his research is focused on pattern recognition and spectroscopic inverse problems.

About the author—MARC GARLAND received his Ph.D. in chemical engineering at the Eidgenossische Technische Hochschule Zurich. He worked in central research at the pharmaceutical firm CIBA GEIGY AG and was lecturer at ETH-Z before joining NUS. His research focuses on system identification problems in reaction engineering.

Anal. Chem. **2005**, *77*, 1655–1662

Development of 2D Band-Target Entropy Minimization and Application to the Deconvolution of Multicomponent 2D Nuclear Magnetic Resonance Spectra

Liangfeng Guo,[†] Anette Wiesmath,[‡] Peter Sprenger,[§] and Marc Garland*[,†,‡]

Department of Chemical and Biomolecular Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore, 117576, Institute of Chemical and Engineering Sciences, Singapore, 1 Pesek Road, Jurong Island, Singapore, 627833, and Bruker Biospin AG, Branch Office Thailand, 41 Soi Lertpanya, Lertpanya Building, 1407, Bangkok 10440, Thailand

Spectral reconstruction from multicomponent spectroscopic data is the frequent primary goal in chemical system identification and exploratory chemometric studies. Various methods and techniques have been reported in the literature. However, few algorithms/methods have been devised for spectral recovery without the use of any a priori information. In the present studies, a higher dimensional entropy minimization method based on the BTEM algorithm (Widjaja, E.; Li, C.; Garland, M. *Organometallics* 2002, *21*, 1991–1997.) and related techniques were extended to large-scale arrays, namely, 2D NMR spectroscopy. The performance of this novel method had been successfully verified on various real experimental mixture spectra from a series of randomized 2D NMR mixtures (COSY NMR and HSQC NMR). With the new algorithm and raw multicomponent NMR alone, it was possible to reconstruct the pure spectroscopic patterns and calculate the relative concentration of each species without recourse to any libraries or any other a priori information. The potential advantages of this novel algorithm and its implications for general chemical system identification of unknown mixtures are discussed.

In the last thirty years, NMR spectroscopy has experienced a dramatic development in both the sophistication of the instrumentation and the signal processing. 2D NMR was proposed by Jeener[1] in 1971 and later demonstrated by Ernst et al., leading to a tremendous increase in the capability of NMR and the subsequent explosion in experimental techniques for two dimensions. By the introduction of 2D experiments, the complexity of heavily overlapping 1D spectra could be significantly simplified. Moreover, unique structural information based on the correlation of the frequencies is obtained. There are now a large number of experimental 2D techniques including COSY,[2,3] HSQC,[4] HMBC,[5] etc.[6] 2D techniques have been widely used for the analysis of structurally complex molecules, including the structural determination of biomolecules such as proteins, peptides, and nucleic acids.

The analysis of multicomponent solutions, particularly those containing unknown constituents, is a general problem in spectroscopy, and it presents numerous analytical difficulties. Considerable effort has been focused on the one-dimensional FT-IR, Raman, and NMR data using a variety of chemometric techniques such as SIMPLISMA,[7] IPCA,[8] and OPA-ALS.[9] In addition, second-order blind identification has been used for both 1D and 2D NMR.[10] However, progress on really complex blind source problems[11] has been severely limited due to the frequent need for some form of a priori system information such as an estimate of the number of species present or the frequent intractability of systems containing more than about three components. Progress in the analysis of multicomponent solutions is particularly important for complex syntheses and other reactive systems where separation of the labile constituents is not a realistic option.

Recently, a new technique, band-target entropy minimization (BTEM), was developed and applied to 1D spectra of multicomponent systems, using FT-IR,[12,13] Raman,[14] XRD,[15] and MS.[16]

* To whom correspondence should be addressed. Tel: +65 6874 6617. Fax: +65 6779 1936. E-mail: chemvg@nus.edu.sg.

† National University of Singapore.

‡ Institute of Chemical and Engineering Sciences, Singapore.

§ Bruker Biospin AG.

(1) Jeener, J. Ampere Int. Summer School 11. Basko Polje, Yugoslavia, 1971.
(2) Aue, W. P.; Bartholdi, E.; Ernst, R. R. J. *Chem. Phys.* **1976**, *64* (5), 2229–2246. Nagayama, K.; Kumar, A.; Wuethrich, K. *J. Magn. Reson.* **1980**, *40* (2), 321–334.
(3) Piantini, U.; Sorensen, O. W.; Ernst, R. R. *J. Am. Chem. Soc.* **1982**, *104* (24), 6800–6801.
(4) Bodenhausen, G.; Ruben, D. *Chem. Phys. Lett.* **1980**, *69*, 185–188.
(5) Bax, A.; Summers, M. F. *J. Am. Chem. Soc.* **1986**, *108*, 2093–2094.
(6) Duddeck, H.; Dietrich, W. *Structure Elucidation by Modern NMR: A Workbook*, 2nd. ed.; Springer-Verlag: New York, 1992.
(7) Windig, W. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 3–16.
(8) Bu, D. S.; Brown, C. W. *Appl. Spectrosc.* **2000**, *54*, 1214–1221.
(9) Sanchez, F. C.; Toft, J.; Van den Bogaert, B.; Massart, D. L. *Anal. Chem.* **1996**, *68*, 79–85.
(10) Nuzillard, D.; Bourg, S.; Nuzillard, J. M. *J. Magn. Reson.* **1998**, *133* (2), 358–363.
(11) Hua, KB. *Circ. Syst. Signal Pr.* **2002**, *21* (1), 91–108.
(12) Chew, W.; Widjaja, E.; Garland, M. *Organometallics* **2002**, *21*, 1982–1990.
(13) Widjaja, E.; Li, C.; Garland, M. *Organometallics* **2002**, *21*, 1991–1997.
(14) Ong, L. R.; Widjaja, E.; Stanforth, R.; Garland, M. *J. Raman Spectrosc.* **2003**, *34* (4), 282–289.
(15) Guo, L. F.; Kooli, F.; Garland, M. *Anal. Chim. Acta* **2004**, *517*, 229–236.
(16) Zhang, H. J.; Garland, M.; Zeng, Y. Z.; Wu, P. *J. Am. Soc Mass Spectrom.* **2003**, *14*, 1295–1305.

10.1021/ac0491814 CCC: $30.25
Published on Web 02/17/2005

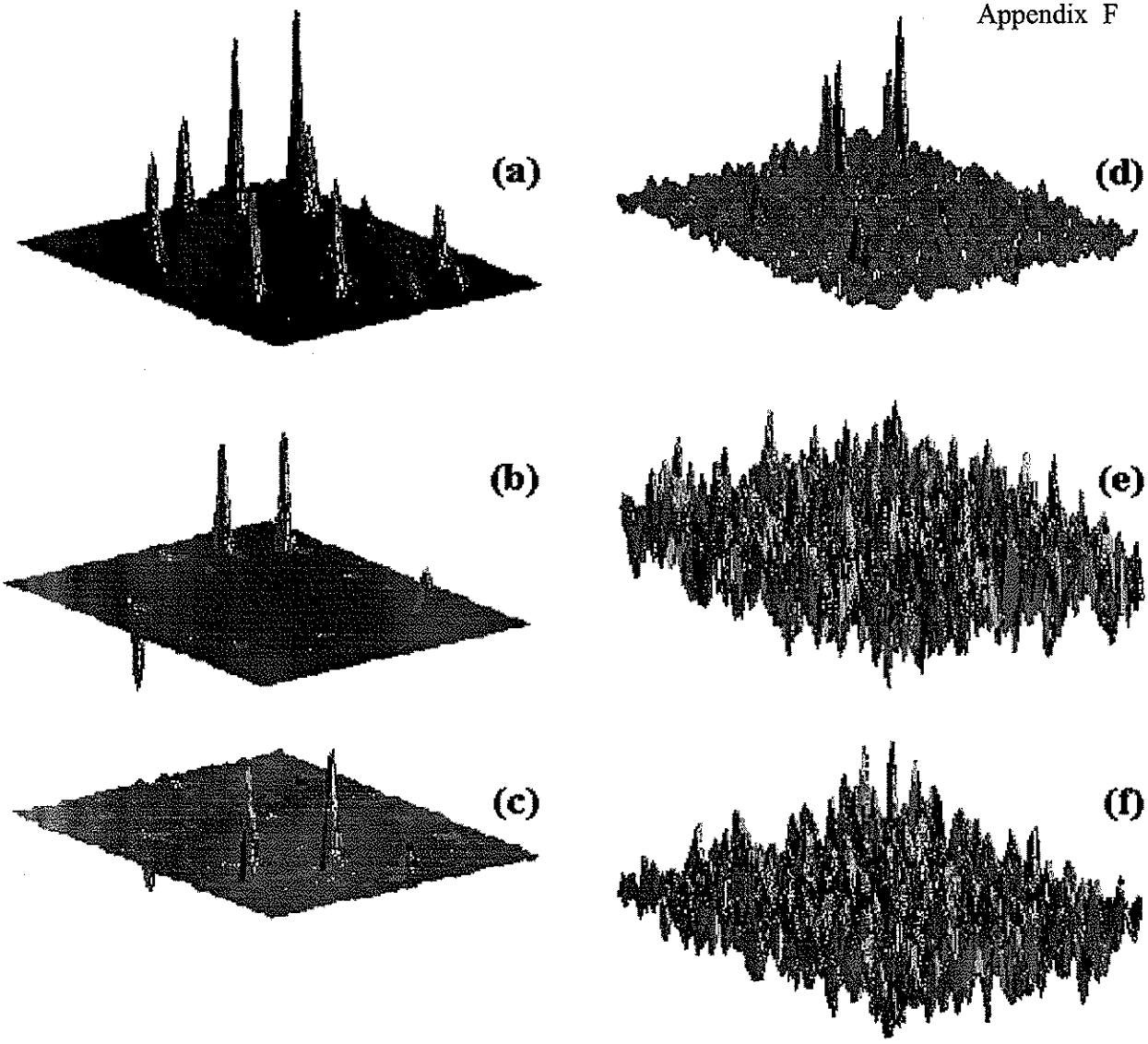


Figure 1. The 1st (a), 2nd (b), 3rd (c), 5th (d), 8th (e), and 14th (f) right singular matrices (V^T).

BTEM uses no libraries or a priori knowledge and works with nonreactive as well as reactive systems. Systems with 10 or more components have been successfully analyzed,[17] and the spectra of trace components constituting less than 0.1% of the total signal can be recovered with significant signal-to-noise enhancement.[18] The primary utility of BTEM comes from (1) the fact that no a priori estimate of the number of species present is needed, (2) considerable noise reduction can be obtained, and (3) its goal-oriented approach; the user targets a single spectral feature of interest, and the algorithm returns the full-range deconvoluted spectrum.

In this contribution, we present for the first time the mathematical constructs of 2D-BTEM and apply the algorithm to COSY and HSQC 2D NMR measurements of solutions containing three solutes. The extension of 1D-BTEM to 2D-BTEM requires a reformulation of the vector space decomposition as well as the definition of entropy. Details of the unfolding procedure for the three-array data set as well as the entropy minimization are presented. Although only a few measurements were made, the deconvoluted 2D spectra are in good to very good agreement with the pure component reference spectra. The goal-oriented approach of targeting spectral features is retained upon extension of 1D-BTEM to 2D-BTEM. The applicability of this methodological approach to even more complex multicomponent mixtures is readily apparent.

EXPERIMENTAL SECTION

Sample Preparation. The samples for NMR were prepared by dissolving varying amounts of 1,5-dichloro-1-pentyne (Aldrich), 4-nitrobenzaldehyde (Aldrich), and 3-methyl-2-butenal (Aldrich) and topping with $CDCl_3$ to achieve a total volume of 500 μL. Seven solutions for NMR measurements were prepared, and each solution contained all three solutes. The approximate concentrations of solutes varied over the range 1.0–2.0 wt %. Consequently, the variation in solute concentrations from sample to sample can be considered quite low. Two spectra of each solution were

(17) Li, C.; Widjaja, E.; Garland, M. *J. Am. Chem. Soc.* **2003**, *125*, 5540–5548.
(18) Li, C.; Widjaja, E.; Chew, W.; Garland, M. *Ange. Chem., Int. Ed.* **2002**, *20*, 3785–3789.

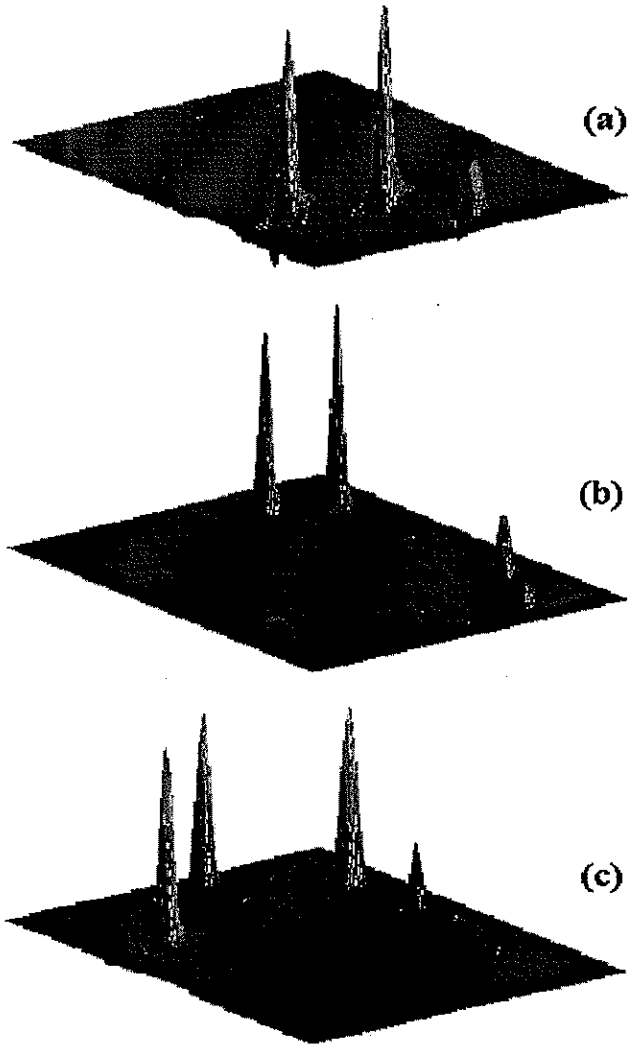


Figure 2. Exhaustive searches with 2D-BTEM producing three 2D spectral patterns. These correspond to 4-nitrobenzaldehyde (a), 3-methyl-2-butenal (b), and 1,5-dichloro-1-pentyne (c).

measured. The use of a constant liquid-phase volume in all sample preparations was crucial in the quantitative aspects of this study.

Instrumental Aspects. All the data were acquired at 298 K on a Bruker Avance 400 WB NMR spectrometer equipped with a 5-mm $^{1}H/^{31}P/^{13}C/^{15}N$ QNP probe with z gradient. All the 2D NMR spectra were acquired at 400.13 (^{1}H) and 100.62 MHz (^{13}C) with standard Bruker-supplied pulse sequences. The spectral parameters are as follows: the ^{1}H spectral width was 5208 (COSY) and 4807 Hz (HSQC) and for the ^{13}C dimension 20 120 Hz, number of scans per increment 2, number of t1 increments 128, each with 1K acquisition points, and repetition time 1.5 s. The 2D spectra were processed as 1K−1K complex matrices with unshifted sine weighting functions in both dimensions. The final data set for 2D-BTEM was a three-array of dimension 14 × 1024 × 1024.

COMPUTATIONS

Pattern Recognition and Information Entropy. As previously mentioned, various chemometric methods for estimating pure component spectra from multicomponent mixtures have been proposed. Methods based on self-modeling curve resolution,[19] such as SIMPLISMA[7] and IPCA,[8] have attracted considerable interest since pure component spectra can be extracted without the use of any reference library information. In 1984, Sasaki et al. proposed a new approach based on entropy minimization.[20] The concept of signal entropy was first introduced by Shannon.[21] Watanabe has mentioned the association between entropy minimization and pattern recognition.[22] It is now recognized that minimization of entropy is closely associated with the *principle of simplicity*.[23]

The development of band-target entropy minimization was made possible after the implementation of a number of additional considerations including the concept of band-targeting, signal enhancement and signal noise reduction,[12] and experimental design.[13,18] Although entropy expressions similar to those used by Sasaki et al. were originally used, simplified spectral measures have been used recently to expedite computation, particularly for the analysis of very large data sets. In the present contribution, the objective function used is the modified form. However, the name "entropy minimization" has been retained due the original problem formulation and the fact that the goal remains a search for spectral simplicity.

System Representation. It is well known that there exist many complications associated with quantitative NMR measurements.[24,25] These complications are associated in part with signal-to-noise issues, but more importantly with issues associated with spin relaxation and the time scales for data acquisition. Nevertheless, it is instructive to have a system model in order to facilitate discussion of the numerical data analysis.

NMR is an absorption spectroscopy.[26] Accordingly, let $A_{v\times v}$ denote a 2D spectrum where v is the number of channels in each spectral direction. Let $A_{v\times v}(T,P,x)$ be the measurement of a single multicomponent solution with s species, where the solution has a state specified by the temperature T, pressure P, and species mole fractions x. It is assumed that there are N_i moles of each species present in the measured volume of sample. Each species possesses a pure component absorptivity $a_{(v\times v)i}(T,P,x)$. Since the measured 2D spectrum is the superposition of s pure component absorptivities, $A_{v\times v}(T,P,x)$ can be represented as shown in eq 1 where $\epsilon_{v\times v}$ denotes the associated instrumental/experimental error.

$$A_{v\times v}(T,P,x) = \sum N_i a_{(v\times v)i}(T,P,x) + \epsilon_{v\times v} \qquad (1)$$

In the entire experimental study, k measured spectra will be acquired. The experimental design requires variations in the composition of samples. Accordingly, each mole fraction x must be varied, and the set of mole fractions will span a region of composition space. Consequently, the pure component absorp-

(19) Lawton, W. H.; Sylvestra, E. A. *Technometrics* **1971**, *13*, 617−633.
(20) Sasaki, K.; Kawata, S.; Minami, S. *Appl. Opt.* **1984**, *23*, 1955−1959.
(21) Shannon, C. E. *Bell Syst. Technol. J.* **1948**, *3*, 379−423.
(22) Watanabe, S. *Pattern Recognit.* **1981**, *13*, 381−387.
(23) Kanpur, J. N. *Maximum-Entropy Models in Science and Engineering*; Wiley Eastern Ltd.: New Delhi, 1993; p 3.
(24) Kasler, F. *Quantitative Analysis by NMR Spectroscopy*; Academic: New York, 1973.
(25) Tirendi, C. F.; Martin,J. F. *J. Magn. Reson.* **1989**, *85*, 162−169.
(26) Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Clarendon Press: Oxford, 1987.

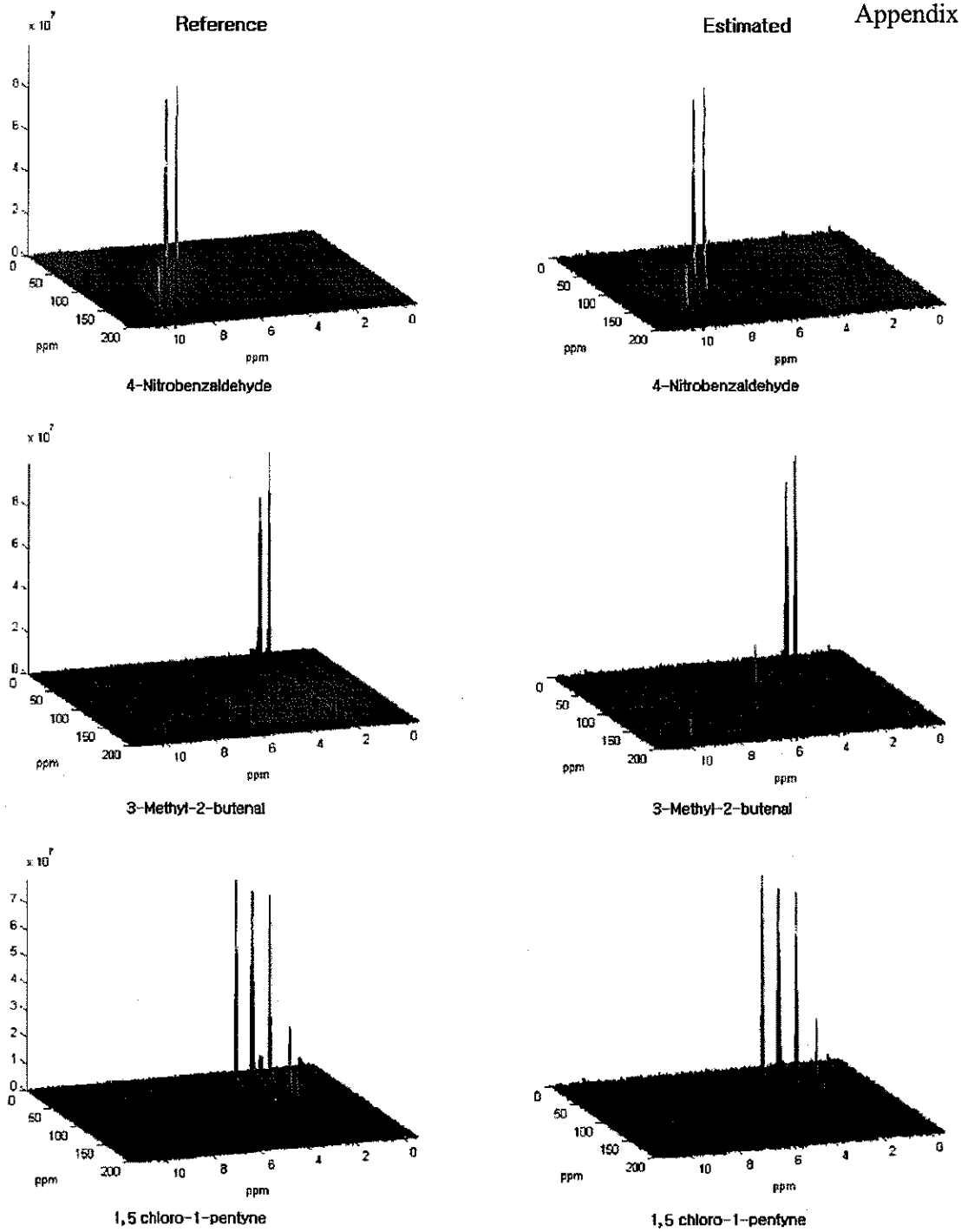


Figure 3. Reference HSQC spectra for 4-nitrobenzaldehyde, 3-methyl-2-butenal, and 1,5-dichloro-1-pentyne and comparison to the estimated pure 2D HSQC NMR spectra by 2D-BTEM.

tivities of any particular species will be assumed to vary somewhat from sample to sample. Indeed, it is very well known that changes in chemical shifts are induced by composition changes. Moreover, there is no guarantee that all samples can be measured at exactly the same temperature and pressure (in some chemometric studies, these parameters may be intentionally varied as well in the experimental design). In other words, the pure component absorptivities are nonstationary. Their position, intensity, and band shapes are not constant over the set of all k measurements. The mean absorptivites for the study will be denoted $a_{(v\times v)i}(\overline{T,P,x})$.

The k spectral measurements give rise to a set of 2D NMR observations. These will be denoted $A_{k\times v\times v}(T,P,x)$ where the raw absorbance data is now a three-array. The experimental data $A_{k\times v\times v}(\overline{T,P,x})$ can be represented in terms of the mean absorptivites $a_{(v\times v)i}(\overline{T,P,x})$by eq 2. It is very important to note that the error $\epsilon'_{k\times v\times v}$ now consists of instrumental/experimental error *as well as* the system nonlinearities arising from the

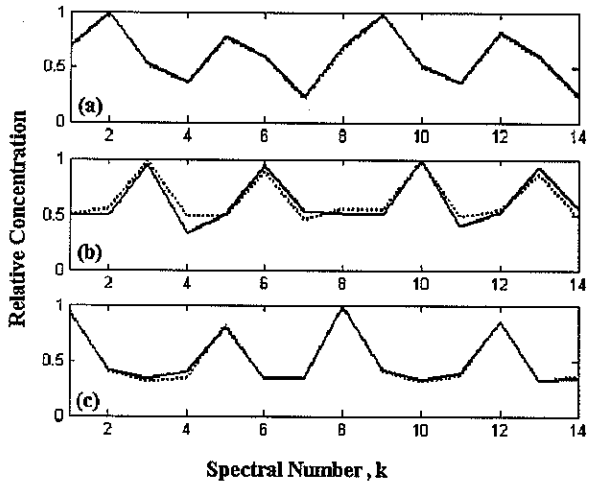


Figure 4. Relative concentrations for HSQC experiments as determined by a least-squares fit with the reference spectra (solid line) versus estimated pure spectra (dotted line). Top curve for 4-nitrobenzaldehyde (a), middle for 3-methyl-2-butenal (b), and bottom for 1,5-dichloro-1-pentyne(c).

nonstationary signal characteristics.

$$A_{k\times v\times v}(T,P,x) = N_{k\times s}a_{s\times v\times v}(\overline{T,P,x}) + \epsilon'_{k\times v\times v} \qquad (2)$$

Singular Value Decomposition and Model Reduction. The theoretical basis for generalized decompositions of higher-order tensors and *n*-arrays has been addressed in a few research articles. Foremost among these are the Tucker3[27] and PARAFAC[28] decomposition. The ordering of the decomposition of an *n*-array can be carried out in more than one way, in other words, with priority given to one or more indices. The numerical realizations of such decompositions have been reported.[29] In the following development, we implement a decomposition that retains the natural physical structure of the data set.

An appropriate vector-space decomposition of the experimental three-array $A_{k\times v\times v}(T,P,x)$, which leads to abstract representations of the absorptivities, is an important computational step. To achieve this the three-array, $A_{k\times v\times v}(T,P,x)$ must first be unfolded to a two-array (matrix) $A_{k\times(v\times v)}(T,P,x)$. This requires that each experimentally measured 2D spectrum is reordered by *concatenation* of each row. The three-array and two-array have exactly the same number of elements.

$$A_{k\times v\times v}(T,P,x) \rightarrow A_{k\times(v\times v)}(T,P,x) \qquad (3)$$

Singular value decomposition (SVD) is a generalization of eigenvector analysis, and it is the preferred computational method to study the structure of rectangular matrices.[30] Well-documented programs/functions for carrying out SVD are available for the primary scientific and engineering languages including the ELLS and MatLab libraries. SVD can be directly applied to $A_{k\times(v\times v)}(T,P,x)$. Typically, one would formulate the problem such that $k < (v \times v)$. This condition will normally be satisfied since the experimentalist performs a few individual measurements and the typical number of channels on current commercial NMR instruments are between (1024 × 1024) and (16384 × 16384). SVD results in three new objects, namely, the left singular matrix $U_{k\times(v\times v)}$, the singular values $\Sigma_{(v\times v)\times(v\times v)}$, and the right singular matrix $V^T_{(v\times v)\times(v\times v)}$. Regardless of the type of problem studied, the matrix $\Sigma_{(v\times v)\times(v\times v)}$ is always diagonal. In the present case of spectroscopic data analysis, the matrices $U_{k\times(v\times v)}$ and $V^T_{(v\times v)\times(v\times v)}$ are dense.

$$A_{k\times(v\times v)}(T,P,x) = U_{k\times(v\times v)}\Sigma_{(v\times v)\times(v\times v)}V^T_{(v\times v)\times(v\times v)} \qquad (4)$$

It should be pointed out that the number of rows in the matrices $\Sigma_{(v\times v)\times(v\times v)}$ and $V^T_{(v\times v)\times(v\times v)}$ greatly exceed the number of experimental spectra k. The extra rows exist due to the mathematical constructs of SVD. Accordingly, the last $(v \times v) - k$ rows are not physically relevant and can be discarded. This leads to the truncated expression eq 5.

$$A_{k\times(v\times v)}(T,P,x) = U_{k\times k}\Sigma_{k\times k}V^T_{k\times(v\times v)} \qquad (5)$$

It is also possible to undo concatenation. This leads to eq 6 where the k physically meaningful right singular vectors have been transformed to k physically meaningful right singular matrices.

$$A_{k\times v\times v}(T,P,x) = U_{k\times k}\Sigma_{k\times k}V^T_{k\times v\times v} \qquad (6)$$

If the absorptivities were absolutely stationary signals, then the system is linear, there would be only s degrees of freedom, and subsequently only s of the k right singular matrices in $V^T_{k\times v\times v}$ in eq 6 would be physically important from a spectroscopic viewpoint. However, as stated before for the system representation, the physical system is nonlinear. Comparison of the system model eq 2 and the decomposition of experimental observations eq 6 leads to the conclusion that information on the s pure component absorptivities are imbedded, in a nontrivial manner, in the k physically meaningful right singular matrices in $V^T_{k\times v\times v}$. If the number of experimentally measured spectra k happens to be less than the number of observable components s, then the mathematical problem can be considered irrevocably ill-posed, and subsequently, there is no possibility for a unique solution to deconvolution.

Consider a mildly nonlinear system. In the present 2D NMR case, this would be a system where the chemical shifts, band shapes, and intensities of the absorptivities change only slightly in the experimentally accessible region of the measurements ($T_{min} - T_{max}$, $P_{min} - P_{max}$, $x_{min} - x_{max}$). Now, imagine that the number of experimentally measured spectra k are only slightly larger than s. A significant amount of useful physical information can be expected to be present in most if not all the k right singular matrices in $V^T_{k\times v\times v}$. Next, imagine that the number of experimentally measured spectra k are much larger than s. A significant amount of useful physical information can be expected to be present in only j of the k right singular matrices in $V^T_{k\times v\times v}$ where

(27) Tucker, L. R. In *Problems in Measuring Change*; Harris, C. W., Ed.; University of Wisconsin Press: Madison WI, 1963; pp122–137.
(28) Carroll, J. D.; Chang, J. *Psychometrika* **1970**, *35*, 283–319.
(29) Kolda, T. G. *SIAM J. Matrix Anal. A* **2001**, *23* (1), 243–255.
(30) Golub, G. H.; van Loan, C. F. *Matrix Computations*; Johns Hopkins University Press: Baltimore, 1996.

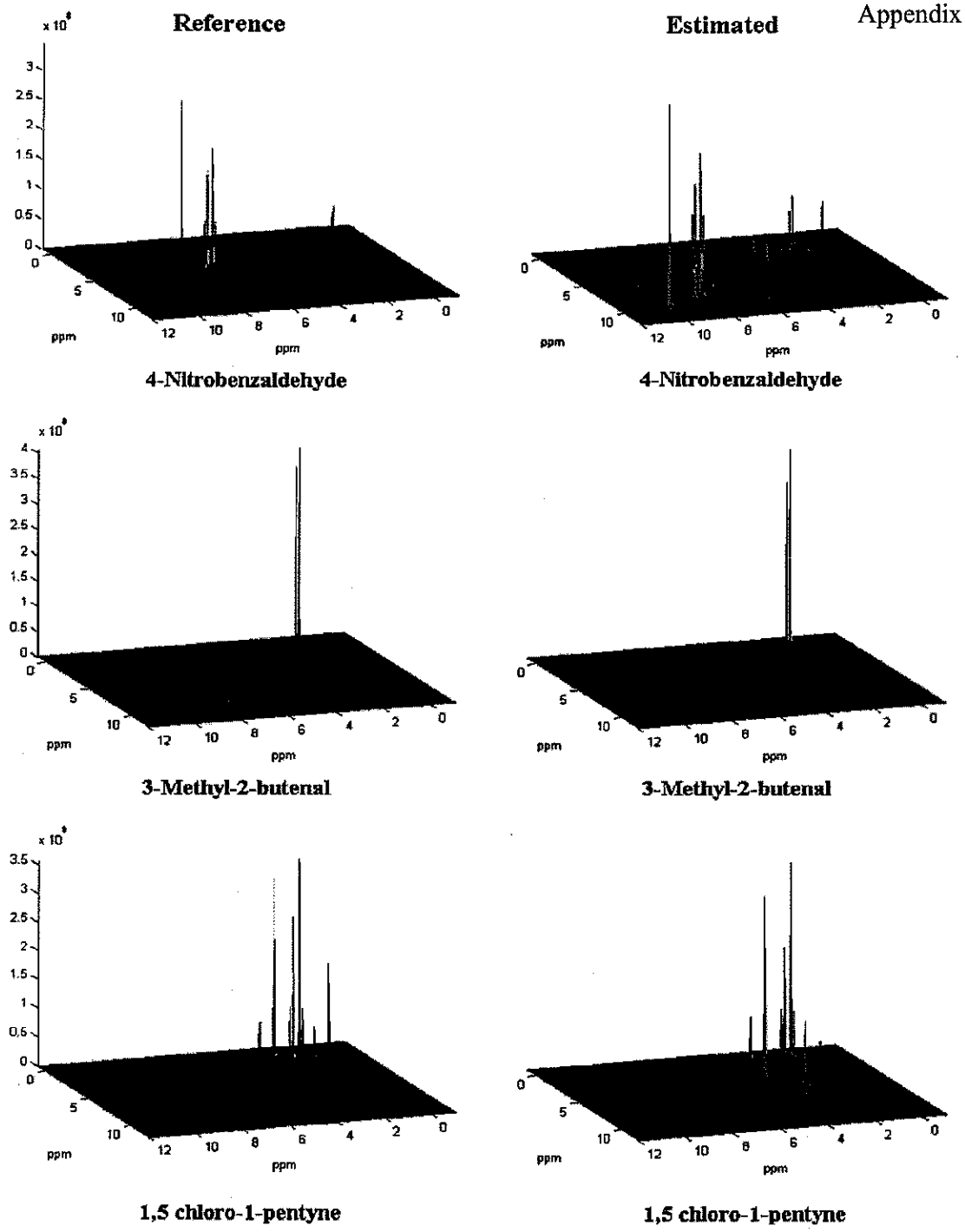


Figure 5. Reference and the estimated pure 2D COSY NMR spectra using 2D-BTEM.

$s < j < k$. The remaining right singular matrices are mainly randomly distributed noise. This important observation can be expressed by eq 7, where $\bar{A}_{k\times v\times v}(T,P,x)$ is now the expectation for the set of observations.

$$\bar{A}_{k\times v\times v}(T,P,x) \leftarrow U_{k\times j}\Sigma_{j\times j}V^{T}_{j\times v\times v} \tag{7}$$

$$\bar{A}_{k\times v\times v}(T,P,x) \approx A_{k\times v\times v}(T,P,x) \tag{8}$$

The importance of singular value decomposition as expressed by eq 7 cannot be underestimated. First, SVD provides a crucial initial untangling of the signals. Specifically, SVD untangles the observations into j matrices, which contain the abstract information concerning the pure component absorptivities. Second, when the system is mildly nonlinear and k is much larger than s, it is possible to discard $k-j$ matrices of information, which consist almost entirely of noise and have little usable physical spectroscopic information. With $k-j$ matrices of discarded noise, the potential exists for spectral reconstruction with outstanding signal-to-noise enhancement.

Formulation of 2D-BTEM. The ultimate objective of 2D-BTEM is to obtain accurate estimates of the mean pure component absorptivities. This is achieved by transforming the abstract V^T information into pure component absorptivity approximations $\hat{a}_{v\times v}(\overline{T,P,x})$, one estimate at a time. Since the real pure component absorptivities vary somewhat from measurement to measurement, the approximations are in some sense a mean of the observations. The computation can be performed on either the right singular vectors in $V^T_{j\times(v\times v)}$ or the right singular matrices in the three-array $V^T_{j\times v\times v}$. This computational issue requires two different formulations for entropy resulting in two different types of objective functions. It is convenient to refer to the two formulations as vectorwise 2D-BTEM and matrixwise 2D-BTEM. Since some information is in a sense "lost" during concatenation, matrixwise 2D-BTEM will prove to be the preferred computational route since higher quality spectral estimates generally result.

The approximated pure component 2D absorptivities $\hat{a}_{v\times v}(\overline{T,P,x})$ will be represented by eq 9 for vectorwise 2D-BTEM and by eq 10 for matrixwise 2D-BTEM. The vector $T_{1\times j}$ maps the corresponding matrix $V^T_{j\times(v\times v)}$ or three-array $V^T_{j\times v\times v}$ into exactly one spectral estimate. In the case of vectorwise 2D-BTEM, the initial mapping produces a vector that is refolded to obtain $\hat{a}_{v\times v}(\overline{T,P,x})$. In the case of matrixwise 2D-BTEM, the transformation directly results in the 2D estimate $\hat{a}_{v\times v}(\overline{T,P,x})$.

$$\hat{a}_{v\times v}(\overline{T,P,x}) \leftarrow\leftarrow T_{1\times j}V^T_{j\times(v\times v)} \tag{9}$$

$$\hat{a}_{v\times v}(\overline{T,P,x}) \leftarrow T_{1\times j}V^T_{j\times v\times v} \tag{10}$$

The optimal determination of the j unknowns in the vector $T_{1\times j}$ is the computationally intensive task. There are two parts. The first issue is the repeated evaluation of the entropy of the term $T_{1\times j}V^T_{j\times(v\times v)}$ or $T_{1\times j}V^T_{j\times v\times v}$. The second issue is the search for the final value of $T_{1\times j}$ such that the global entropy minimum is obtained.

Consistent with the original formulation of Shannon entropy, a definition for the entropy H used in vectorwise 2D-BTEM can take the form of eq 11, where the probability distribution p is consistent with the original suggestion of Sasaki et al. that a second derivative ($m = 2$) of the spectral data can be used eq 12.

$$H = \sum_v h_v = -\sum_v p_v \ln p_v \tag{11}$$

$$p_v = \frac{\left|\dfrac{\mathrm{d}^m \hat{a}_v}{\mathrm{d}v^m}\right|}{\sum_v \left|\dfrac{\mathrm{d}^m \hat{a}_v}{\mathrm{d}v^m}\right|} \tag{12}$$

In a similar spirit, we can now formulate a definition for the entropy H used in matrixwise 2D-BTEM eq 13, where the term p takes into consideration the smoothness of the spectra in two dimensions, eqs 14 and 15.

$$H = -\sum_{v_1}\sum_{v_2} h_{v_1v_2} = -\sum_{v_1}\sum_{v_2} p_{v_1v_2} \ln p_{v_1v_2} \tag{13}$$

$$p_{v_1} = \frac{\left|\dfrac{\mathrm{d}^m \hat{a}_{v_1v_2}}{\mathrm{d}v_1^m}\right|}{\sum_v \left|\dfrac{\mathrm{d}^m \hat{a}_{v_1v_2}}{\mathrm{d}v_1^m}\right|} \tag{14}$$

$$p_{v_1v_2} = \frac{\left|\dfrac{\mathrm{d}^m p_{v_1}}{\mathrm{d}v_2^m}\right|}{\sum_v \left|\dfrac{\mathrm{d}^m p_{v_1}}{\mathrm{d}v_2^m}\right|} \tag{15}$$

The repeated evaluation of the log terms can require significant computational time. We have previous tested expressions for H that omit the log term, and the functions H' often yield good-quality pure component spectra. Accordingly, for vectorwise and matrixwise the following two expressions are used.

$$H' = \sum_v p_v \tag{16}$$

$$H' = \sum_{v_1}\sum_{v_2} p_{v_1v_2} \tag{17}$$

The above-mentioned functions H' are needed in the global optimization. Specifically, they are included in the objective function F_{obj} along with a penalty function P and a term for peak integration (eq 18). Further details concerning P and PeakInt can be found refs 12 and 13. The transformation of V^T (eqs 9 and 10) is governed by the optimization of an objective function corresponding to the final estimate of the pure component spectral approximation, $\hat{a}$. The use of PeakInt in the objective function of 2D-BTEM is often favored. The minimization of the integrated estimate is consistent with spectral simplicity. The global minimization of the objective function is achieved using simulated annealing, which is a stochastic search technique.[31]

$$\min F_{obj} = H + P + \text{PeakInt} \tag{18}$$

2D Wiener Filtering. An adaptive 2D Wiener filter is commonly used to filter degraded images in image processing and was used in the present contribution to filter the experimental 2D NMR data. The Function WIENER2 is available in Matlab,[32] which performs 2D adaptive noise-removal filtering. Y = WIENER2 (X, [a b]) filters the matrix X using pixelwise adaptive Wiener filtering, using neighborhoods of size a-by-b to estimate the local

(31) Corana, A.; Marchesi, M.; Martin, C.; Ridella, S. *ACM. Trans. Math. Software* **1987**, *13*, 262–280.
(32) Matlab, MathWorks Inc. MatLab Reference Guide, 1995.

image mean and standard deviation. In the present study, the parameters a and b were set to 10 and 10.

RESULTS AND DISCUSSION

HSQC Data. The 14 2D HSQC measurements form a three-array $A_{14\times1024\times1024}$. The SVD on an array of this size can be problematic, even on a high-end workstation with considerable RAM.[33] To decreasing the computational burden, only four rectangular regions containing the real physical spectral features (peaks) were taken. About 90% of the original 2D spectral data did not contain useful physical information. The small rectangular regions were assembled into a new concatenated data array $A_{14\times(539\times107)}$.

SVD was performed on the matrix $A_{14\times(539\times107)}$ to obtain the 14 meaningful right singular vectors in $V^T_{14\times(539\times107)}$. Concatenation was undone. The resulting right singular matrices are shown in Figure 1. Physically meaningful spectral features were observed in only the first seven matrices. The eighth matrix is essentially featureless. This holds true for the matrices 9−14 as well. Consequently, the three-array $V^T_{14\times539\times107}$ was reduced to $V^T_{7\times539\times107}$ where j is set to 7.

Both vectorwise and matrixwise 2D-BTEM were performed using $V^T_{7\times539\times107}$ by targeting observable features in the seven visualized right singular matrices. Exhaustive searches produced only three 2D spectral patterns. The estimates are shown in Figure 2.

These patterns were then imbedded into matrices with 1024 × 1024 channels. The results of the matrixwise 2D-BTEM after imbedding, to yield the 2D HSQC estimated pure component absorptivities $\hat{a}_{1024\times1024}$, are shown in Figure 3. The spectral estimates appear quite good when compared to authentic experimental references.

Moreover, the dual problem for relative concentrations can be solved. Figure 4 shows the relative concentrations as determined by a least-squares fit with the reference spectra versus the relative concentrations as determined by a least-squares fit with the estimated pure component absorptivities. The relative concentrations calculated for samples 1−7 are almost the same as the relative concentrations of samples 8−14 since samples 8−14 represent the replicate measurements (there were 7 physical samples, each sample measured twice). The calculated concentration profiles are very similar. Figure 4 indicates another important aspect, namely, the least-squares fitting of an entire matrix for concentration determination is generally better than relying on the intensity of just one point (peak height).

COSY Data. The 2D COSY measurements were analyzed in a similar manner. As only 12 samples were available, the dimension of the 3-array was 12 × 1024 × 1024. To decreasing the computational burden, only eight rectangular regions containing the real physical spectral features (peaks) were taken. About 90% of the original 2D spectral data did not contain useful physical information. The small rectangular regions were assembled into a new concatenated data array $A_{12\times(468\times150)}$.

SVD was performed on the matrix $A_{12\times(468\times150)}$ to obtain the 12 meaningful right singular vectors in $V^T_{12\times(468\times150)}$. Concatenation was undone. Physically meaningful spectral features were observed in only the first seven matrices. The eighth matrix was essentially featureless. This holds true for the matrices 9−12 as well. Consequently, the three-array $V^T_{12\times468\times150}$ was reduced to $V^T_{7\times468\times150}$ where j is set to 7.

Both vectorwise and matrixwise 2D-BTEM was performed using $V^T_{7\times468\times150}$ by targeting observable features in the seven visualized right singular matrices. Exhaustive searches produced only three 2D spectral patterns. These patterns were then imbedded into matrices with 1024 × 1024 channels. The results of the matrixwise 2D-BTEM after imbedding, to yield the 2D COSY estimated pure component absorptivities $\hat{a}_{1024\times1024}$ are shown in Figure 5. The spectral estimates appear quite good when compared to authentic experimental references.

The dual problem for relative concentrations was also solved. Figure 6 shows the relative concentrations as determined by a least-squares fit with the reference spectra versus the relative concentrations as determined by a least-squares fit with the estimated pure component absorptivities. The calculated concentration profiles are good for the first two components but only only fair for the third component. This many be due to the somewhat higher nonstationary characteristics of COSY versus HSQC. Again, the relative concentrations calculated for samples 1−6 are almost the same as the relative concentrations of samples 7−12 since samples 7−12 represent the replicate measurements (there were 6 physical samples, each sample measured twice).

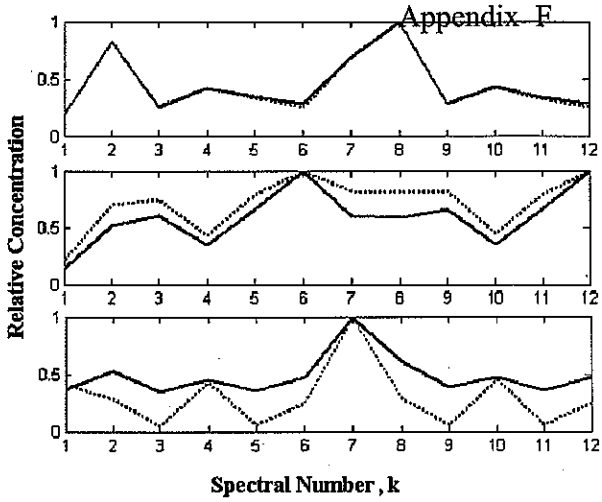


Figure 6. Relative concentrations for COSY experiments as determined by a least-squares fit with the reference spectra (solid line) versus estimated pure spectra (dotted line). Top curve for 4-nitrobenzaldehyde (a), middle for 3-methyl-,2-butenal (b) and bottom for 1,5-dichloro-1-pentyne (c).

CONCLUSION

An advanced entropy minimization based algorithm 2D-BTEM has been proposed and verified with real experimental data from multicomponent COSY and HSQC 2D nuclear magnetic resonance spectroscopy. The quality of recovered spectra is good when compared with authentic experimental references obtained from pure component measurement.

Received for review June 4, 2004. Accepted January 5, 2005.

AC0491814

(33) All calculations were run on a NT workstation with 2-GB RAM and two Xeon processors running MatLab 6.5.

LIST OF PUBLICATIONS

Li, C.Z., L.F. Guo and M. Garland. Homogeneous Hydroformylation of Ethylene Catalyzed by $Rh_4(CO)_{12}$. The Application of BTEM to Identify a New Class of Rhodium Carbonyl Spectra: $RCORh(CO)_3(-C_2H_4)$. Organomet., *23*(9), pp.2201-2204. 2004.

Guo, L.F, F. Kooli and M. Garland. A General Method for the Recovery of Pure Powder XRD Patterns from Complex Mixtures Using No *a priori* Information. Application of Band-Target Entropy Minimization (BTEM) to Materials Characterization of Inorganic Mixtures, Anal. Chim. Acta, *517*(1-2), pp.229-236. 2004.

Li, C. Z., L.F. Guo and M. Garland, Identification of Rhodium-Rhenium Nonacarbonyl $RhRe(CO)_9$: Spectroscopic and Thermodynamic Aspects, Organomet., *23*(22), pp.5275-5279. 2004.

Guo, L.F., A. Wiesmath, P. Sprenger, M. Garland. Development of 2D Band-Target Entropy Minimization and Application to the Deconvolution of Multicomponent 2D Nuclear Magnetic Resonance Spectra, Anal. Chem., *77*, pp. 1655-1662. 2005.

Tjahjono, M., L.F. Guo and M. Garland. The development of a response surface model for the determination of infinite dilution partial molar volumes and excess volumes from dilute multi-component data alone. Implications for the characterization of non-isolatable solutes in complex homogeneous reactive systems. Chem. Eng. Sci. *60* (12), pp.3239-3249. 2005.

Liu GW, C. Z. Li, L. F. Guo and M. Garland. Experimental evidence for a significant homometallic catalytic binuclear elimination reaction: Linear-quadratic kinetics in the rhodium catalyzed hydroformylation of cyclooctene, J. Catal., *237*, pp. 67-78. 2006.

Guo, LF.and M. Garland. The use of entropy minimization for the solution of blind source separation problems in image analysis. Pattern Recognition, *39*, pp.1066-1073. 2006.

Guo, LF and M. Garland. Application of 2D Band-Target Entropy Minimization (2D-BTEM) to Fluorescence data. (In preparation)

Guo, LF and M. Garland. Multinuclear 1D NMR spectroscopic data analysis using Band-Targeting Entropy Minimization (BTEM) method. (In preparation)

Guo, LF and M. Garland. Three-Dimensional Singular Value Decomposition and Pattern Recognition via entropy minimization. (In preparation)