Weighted Two-Band Target Entropy Minimization for the Reconstruction of Pure Component Mass Spectra: Simulation Studies and the Application to Real System

ZHANG HUAJUN

THE NATIONAL UNIVERSITY OF SINGAPORE

2003

Weighted Two-Band Target Entropy Minimization for the Reconstruction of Pure Component Mass Spectra: Simulation Studies and the Application to Real Systems

ZHANG HUAJUN

(B. Eng. Zhejiang Univ. PRC)

(M. Sc. Zhejiang Univ. PRC)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPT OF CHEMICAL & ENVIRONMENTAL ENGINEERING

THE NATIONAL UNIVERSITY OF SINGAPORE

2003

ACKNOWLEDGEMENT

First of all**,** I'd like to express my appreciation to Institute of High Performance Computing Singapore (IHPC) which provided research scholarship to help me to pursue the higher degree studies.

I want to express my deepest appreciation to my supervisors, Dr. Marc Garland and Dr. Wu Ping, both of them led me to the interesting area of chemometrics, for their professional guidance, inspiring discussions, great encouragement and continuous supervision.

I would like to thank Ms. Zeng Yingzi and Mr. Yang Shuowang (IHPC) for their great help. Besides, I want to thank Mr. Chew Wee, Mr. Effendi Widjaja and Dr. Chen Li. They gave me many useful suggestions and much guidance in my research work.

TABLE OF CONTENTS

SUMMARY

How to get pure components spectra from mixture spectra by mathematical methods is a great challenge to chemical researchers. In the past two decades, efforts were dedicated to extracting pure component spectra from mixture spectra. Many of these studies have to use reference spectra or database to find the pure spectra. During the past few years, Dr. Garland and his group have developed a series of algorithms based on the Entropy Minimization and singular value decomposition (SVD) to get the pure spectra from mixture spectra mathematically. By using these algorithms, pure spectra can be reconstructed without relying on prior information. These algorithms are also very useful in studying chemical reactions.

These algorithms have been successfully used in reconstructing pure FTIR, UV, RAMAN spectra in non-reactive or reactive systems. The latest algorithm, Band Target Entropy Minimization (BTEM), has been successfully applied to studying mechanisms of chemical reactions and to find the spectra of unstable components. Currently, these algorithms are expected to handle other spectroscopic data such as Mass Spectra, XRD, and NMR. Due to the specific features of each type of spectrum, effort has to be invested to modify the algorithms for the use in a particular type of spectroscopic data.

Mass spectroscopy is widely used nowadays in chemical analysis. Different from other kinds of continuous spectroscopies like FTIR, the mass spectroscopy is discrete. Another aspect of mass spectroscopy is that the pattern for every pure component spectrum is not fixed. It varies from time to time and from machine to machine.

The original entropy minimization function usually uses 1st, 2nd and 4th order derivatives, which would cause problem when non-differentiable discrete data like mass spectra used. In this present work, discrete spectra, MS, are studied by using entropy minimization algorithm to reconstruct pure component spectra from mixture spectra.

In this thesis, first, a special objective function for discrete spectra is developed to reconstruct pure component spectra from mixture spectra. Peak heights instead of their derivatives are used in the objective function. The algorithm is computationally efficient as fewer mathematical operations are needed to evaluate the objective function.

Secondly, the effect of noise on the system is studied. A method using weighted information $\boldsymbol{V}^T$ vectors is proposed to reduce noise effect in the system. This algorithm weights $\boldsymbol{V}^T$ vectors according to their contribution to the total variance of the observations. Therefore, it makes the system much less sensitive to the noise present in the system.

Thirdly, a two-peak/band targeting method (tBTEM) is used to deal with overlapped peaks. It uses two peaks into targeting rather than one peak as in BTEM. The two-peak targeting method can rearrange the objective function values of different pure spectra, which could find out all pure spectra with their highest peak from mixture. In addition, it can be extended to deal with highly overlapped systems by using multi-peak in targeting. Due to the use of two peaks, it is also less sensitive to the noise because big signal to noise ratio is used.

Fourthly, an exhaustive search method is proposed to find all components in a synthesized mixture with 10 components. In this method, several criteria have been used in discarding bad, duplicated and combined estimated spectra.

Fifthly, a method called fast multi-start simplex method (FMSS) is carried out to accelerate the optimization speed. FMSS uses a multi-start method to find global minimum and to discard non-promising start points in advance by using different stop criteria at different searching steps. The method dramatically reduces the optimization time compared with the multi-start simplex method. Compared with the popular optimization method, simulated annealing, it is also much faster in reconstructing pure mass spectra.

Finally, the new set of algorithms for reconstructing discrete spectra from mixture is tested successfully on a real MS data set which contains four components. Some strategies to lower the non-stationary effect are discussed.

NOMENCLATURE

BTEM and tBTEM Methods

$\boldsymbol{A}$	Experimental mixture data matrix
$\hat{\boldsymbol{a}}$	Normalized estimated spectrum/spectra
$\hat{\boldsymbol{a}}^{est}$	Un-normalized estimated spectrum
$\boldsymbol{c}$	Concentration of pure component (s)
$\hat{\boldsymbol{c}}$	Estimated concentration
F	Penalty function term
G	Objective function Value of BTEM or tBTEM
h	Shannon entropy measure
P	Penalty function of BTEM or tBTEM
$\boldsymbol{P}$	matrix containing 10 reference pure component spectra
$\boldsymbol{S}$	Square diagonal singular value matrix after truncation from S
$\boldsymbol{T}$	Transform matrix/vector
$\boldsymbol{U}$	Left singular matrix
$\boldsymbol{V}^{T}$	Transposed right singular matrix
X Y and **Z**	Estimated spectra which are linearly dependent

Greek Letters

S	Diagonal singular value matrix
a and b	Linear coefficients of estimated spectra in exhaustive search
e	Error matrix
g	Coefficients of penalty function

Subscripts

a	Estimated spectrum/spectra
c	Estimated concentrations
j	Number of right singular vectors used in optimization

k	Number of experiments
s	Number of observable species in mixture
u	Number of channels of spectrum /spectra

Fast Multi-Start Simplex Method (FMSS)

r	Reduction coefficient
R	Number of total searching rounds
n	Number of retained searching points

Greek Letters

e	Stop criterion

Subscripts

p	Number of searching points
i	The ongoing searching round

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

In chemical sciences, one constant challenge is to identify pure components from mixtures and/or evolving reactive systems. Usually, one attempts to purify/separate the components before analysis. To date, many advanced combined analytical instruments, e.g. GC-MS, HPLC-MS and HPLC-NMR, have been used to perform separations, obtain pure component spectra and finally present positive identification of the pure components. Even aided with these state-of-the-art analytical instruments, there are many types of situations where separation is difficult (i.e. a mixture with tens of components in it) or even impossible (i.e. reactive systems with transient or labile species). Moreover, it is difficult to apply traditional methods, separation-then-analysis, to changing systems such as reactions. Therefore, a lot of effort has been put in reconstructing pure component spectra from mixture data without chemical separation i.e. interpreting mixture spectra by mathematical methods to get pure spectra.

Today, there are many mathematical methods of interpreting mixtures. These methods could be classified into two groups. One group uses reference spectra or spectra databases to extract pure component spectra from mixtures. The other group reconstructs pure component spectra without using references. The later one is more useful in finding unknown component spectra since there is no reference available. Dr. Garland and his group developed a series of algorithms (Zeng and Garland, 1998; Pan *et al.*, 2000; Widjaja and Garland 2002; Chew *et al.*, 2002) to reconstruct pure component spectra from mixtures without prior information. Their algorithms used entropy minimization and singular value decomposition (SVD). Besides the abilities in reconstructing pure components spectra, these algorithms are very useful in studying transition metal homogeneous chemical syntheses (Widjaja *et al*., 2002; Li, *et al*., 2002, 2003a and 2003b)

Till now, many achievements of entropy minimization methods have been made in interpreting continuously differentiable spectroscopic data, such as FTIR, RAMAN, and UV-VIS. The latest algorithm is band-target entropy minimization (BTEM) (Chew *et al*., 2002). Currently, these algorithms are expected to apply to other spectroscopic data such as Mass Spectra, XRD, and NMR. Due to the specific features of each type of spectrum, effort has to be invested to modify the algorithms for their use in a particular type of spectroscopic data.

MS is a popular analytical method in many fields such as organic chemistry, food analysis, drug analysis and biological analysis. There are many kind of mass spectrometry such as electrospray ionization mass spectrometry (ESI MS), membrane introduction mass spectrometry (MIMS) *etc*. MS combined with other devices such as GC-MS and HPLC-MS are powerful tools in modern analytical chemistry, they can perform good separation and detection at the same time. But when the number of components in mixture increases, it would be difficult to separate all the components in the mixture at the same time. This kind of problem always happens in food and medicine analyses (*e.g.* Chinese medicines).

Compared with IR and other spectra, mass spectrum has specific features; it is well known for its un-differentiable nature. BTEM method demonstrates its ability in dealing with continuous spectra. Since BTEM method reconstructs pure spectra with

1st, 2nd and 4th order derivatives; it seems that BTEM is not suitable for discontinuous data. For discrete data systems, a different kind of objective function should be used.

In BTEM method, an optimization method named simulated annealing (SA) is used in finding global minima. Widjaja and Garland (2000) successfully used SA as global optimization method to reconstruct spectra on large-scale problems. However, the optimization time of SA in reconstructing pure spectra is relatively long. It would be better to find a faster optimization method.

In this thesis, effort was devoted to modify BTEM and make entropy minimization method applicable to discrete spectra on both simulated systems and real systems. Effort also was dedicated to finding a faster optimization method in reconstructing spectra.

In this present work, effort mainly focuses on mathematical aspects which include:

1) Developing an objective function for discrete spectra which uses peak heights instead of derivatives. The new objective function needs less computational operations than BTEM.
2) Weighting the abstract $\boldsymbol{V^T}$ vectors to get estimated spectra to reduce the effect of noise *i.e.* make the algorithm less sensitive to the number of $\boldsymbol{V^T}$ vectors used in the system. This is very useful in treating data obtained from real systems.
3) Using a two-peak targeting strategy (tBTEM) to deal with strongly overlapping peaks. With this method, all pure component spectra can be reconstructed with their highest peaks. This method also makes the system less sensitive to noise because of higher signal to noise ratio.

4) An exhaustive search method is developed to find all pure components spectra automatically.

5) A global optimization method, fast multi-start simplex method (FMSS), is proposed. Compared with SA, it is much faster in reconstructing pure mass spectra. FMSS uses different kinds of stop criteria in different steps to speed up its searching speed by discarding some non-promising points in advance. Besides its fast speed, FMSS is a totally parallel method.

These new mathematical methods are tested on synthesized and real mixture mass spectra data sets. When these new methods are applied to real mass spectra, some experimental strategies are used to lower the non-stationary effect.

In this thesis, these contributions are organised as follows: Chapter 2 is a literature review of chemometrics development, entropy minimization methods, chemometrics methods in MS, and optimization. In chapter 3, the special aspects of mass spectra and reasons of modifying BTEM are discussed, and mathematical modifications are described in detail. The methodology of fast multi-start simplex optimization method (FMSS) is also discussed and presented in detail in this chapter. In chapter 4, these new mathematical methods are tested by synthesized mixture data; their advantages are discussed and shown by different examples. In chapter 5, these new algorithms are tested on real mass spectra data. Special features of real mass spectra and experimental strategies which are used to lower non-stationary effect of MS are discussed. Conclusions and discuss of future works are presented in chapter 6. Finally, references are attached at the end of the thesis.

Chapter 2：Literature Review

In chemical engineering, especially in chemical reaction engineering, studies of mechanism and kinetic of chemical reactions play important roles in the development of chemical theories. Based on the understanding of the processes of chemical reactions, it would be feasible to design, optimize and control chemical reactions. Furthermore, in chemical industries, the high yield and reproducibility of the targeting components are highly desired. Kinetic studies could provide a basis to achieve these goals. Today, mechanism and kinetic studies are still very active and challenging areas (Garland *et al.* 1997) although they are very old sciences.

In terms of the mechanism and kinetic of a chemical reaction, there are some very important questions that should be answered. They are:

(1) How many observable species present in the reaction system?

(2) How many observable reactions take place in the system?

(3) How the unknown observable species look like?

In many cases, the initial reactants are known. In most cases, the resultants are known too. In terms of reactions, additional information concerning the species and reactions presenting in the system may not always be available especially when transient, reactive or labile species are presented in the systems.

Computer aided analysis of ***in-situ*** spectroscopic measurements could help to solve these problems and allow modelling complex processes since it does not interfere with systems and would get the total information of systems. However, it is usually difficult to interpret of spectroscopic measurements obtained from reactive systems. These difficulties come not only from the complexity of the large amount of data but also from the absence of reference data if newly observed but still unidentifiable transient species presented in system. Also, the presence of random experimental error as well as noise in the spectra would make the interpretation difficult.

Effective interpreting chemical data which are gotten from various analytical instruments is very important. Mathematical methods which could analyze the data and get exact and meaningful chemical information could greatly help us to achieve this aim. Computer aided analytical methods have been received considerable attention in the past decades. A research area named “chemometrics” has developed fast. Chemometrics is a science that combines mathematics and statistics with chemistry to handle, interpret even predict chemical data. Powerful chemometric methods have opened new vistas and provided useful solutions to many complex chemical problems. (Kowalski, 1980; Frank and Kowalski, 1982; Delaney, 1984; Brereton, 1987; Brown *et al*., 1988, 1992, 1994 and 1996; Brown, 1990; Lavine, 1998 and 2000).

Among many chemometric methods, one of the most important techniques is factor analysis (Malinowski, 1991 and 1999), especially the principal component analysis

(PCA) and singular value decomposition (SVD). These techniques demonstrated their abilities in extracting real factors associated with the number of components and reactions from a large number of experimental data which are linearly combined by different components. Among many applications of PCA and SVD, few of them focus on on-line chemical reaction (Furusjoe *et al*., 1998; Bijlsma and Smilde, 1999; Bijlsma *et al*, 1998). Recently, there is a kind of method which combines SVD and entropy minimization in finding pure spectra and dealing with reactive chemical reactions. This kind of method demonstrates as a powerful tool in studying mechanism and kinetic of chemical reactions and finding elusive components (Li *et al*., 2002). The latest algorithm of this kind of method is called band-target entropy minimization (BTEM) (Chew *et al*., 2002).

As mentioned in chapter one, the emphasis of present work is to develop the numerical techniques to apply BTEM to discrete spectra. Based on this reason, the outline of chapter 2 is as follows. Section 2.1 discusses the theoretical basis of SVD. Section 2.2 mainly focuses on entropy minimization methods. Section 2.3 discusses the methods reconstructing pure spectra with/without prior information and entropy minimization method on reconstructing pure spectra with SVD method. In section 2.4, the topic is mainly on chemometric methods on discrete data such as mass spectrometry. Optimization methods are discussed in section 2.5.

2.1 Singular Value Decomposition (SVD)

When a large dimension chemical matrix needs to be studied by chemometric methods, techniques to lower the matrix of data to the lowest dimension are needed. The mathematical methods for determining the number of real factors such as species and reactions are called eigenanalysis, these methods yielding eigenvalues and associated eigenvectors of a matrix. The four most commonly used methods are SVD (Shrager, 1984 and 1986), the power method (POWER), the Jacobi method, and non-linear iterative partial least-squares (NIPALS) (Winter, 1992).

Singular value decomposition is a very powerful technique in dealing with sets of equations or matrices that are either singular or numerically very close to singular. SVD allows one to diagnose the problems in a given matrix. SVD is the preferred algorithm and the most stable (Lawson and Hamson, 1972; Shrager 1986) under the widest range of applications. SVD can distinguish eigenvectors which have minute difference. For large matrices involving thousands of scalars, the use of SVD is preferable (Shrager 1986). With these advantages, SVD is more and more popular.

When performing singular value decomposition on a spectroscopic data matrix $\boldsymbol{A}$, the matrix $\boldsymbol{A}$ is expressed as equation (2.1) (Shrager 1986; Scheick, 1997).

$$\boldsymbol{A}_{\boldsymbol{k}\times u} = \boldsymbol{U}_{\boldsymbol{k}\times\boldsymbol{k}} \times S_{\boldsymbol{k}\times u} \times \boldsymbol{V}^{\boldsymbol{T}}_{u\times u} \tag{2.1}$$

Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal singular vector matrices that satisfy $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix. Matrix S is a diagonal matrix whose diagonal elements, called singular values, are equal to the square roots of the respective eigenvalues. The matrix containing the singular values S contains a square diagonal matrix $\boldsymbol{S}_{k\times k}$ in the first k columns and a zero matrix $\boldsymbol{O}_{k\times(v\text{-}k)}$. If the data matrix is square, the SVD problem reduces to the classic eigenvalue problem (the j^{th} singular value is related to the eigenvalue l_j as the square root). The singular values in S are arranged in decreasing magnitude, representing the decreasing contribution of each corresponding vector in $\boldsymbol{V}^T$ to the total variance of the signals, i.e., the first few vectors contain a significant amount of meaningful information while the latter vectors contain considerable noise.

In real chemical processes, let k denotes the number of spectra taken and u denotes the total number of channels (commonly $u >> k$), an $\boldsymbol{A}_{k\times u}$ data matrix can be obtained when all spectra are collected. It is assumed that both k and υ are greater than s, the number of components in the data set. The set of spectroscopic measurements is related to pure component spectra for the s observable species denoted as $\boldsymbol{a}_{s\times n}$, the relative concentration of the s species and the error are denoted as $\boldsymbol{c}_{k\times s}$ and $e_{k\times n}$, respectively. It has been mentioned in considerable detail elsewhere (Garland *et al.* 1997) that $\boldsymbol{a}_{s\times n}$ can usually be considered as constant, and the error matrix $e_{k\times n}$ represents random experimental error, instrumental error and non-linearity in the absorptivities.

$$\boldsymbol{A}_{k\times u} = \boldsymbol{c}_{k\times s}\boldsymbol{a}_{s\times u} + e_{k\times u} \tag{2.2}$$

When performing SVD on a set of experimental spectra $\boldsymbol{A}$ which is obtained sequentially with time, a set of singulars value or eigenvalues are gotten. For an ideal case without any kind of noise, the real factors should equate to the number of non-zero singular values. For real system, the number of non-zero singular values is always greater than that of real factors. The significant singular values correspond to real factors while the remaining represents the noise in the system. To determine the number of significant factors, visual checking or statistical testing (Malinowski, 1988) should be incorporated. Filtering techniques (Smit, 1992a and 1992b) are always preformed before SVD to enhance the quality of data to get correct factors.

In fact, $\boldsymbol{V}^T$ is an abstract matrix associated with the pure component spectra matrix $\boldsymbol{a}_{\mathrm{s\times v}}$ (Malinowski, 1991). If there are no nonlinearities in the system (and this is seldom the real case), then the first $\boldsymbol{s}$ $\boldsymbol{V}^T$ vectors contain all the information associated with the absorptivities of pure components, $\hat{\boldsymbol{a}}$, and the two matrices are related by a square transformation matrix $\boldsymbol{T}$ as shown in equation (2.3). Similarly, the estimated concentration matrix $\hat{\mathbf{c}}$ is related to $\boldsymbol{U}$ and the rotation matrix $\boldsymbol{T}$ in equation (2.4). It is important to point out that in most real physically meaningful situations, equations (2.3) and (2.4) are at best crude approximations. First, the spectroscopic problem is not linear, and secondly, one does not know the number of species $\boldsymbol{s}$ present in advance.

$$\hat{\boldsymbol{a}}_{s\times u} = \mathbf{T}_{s\times s} \times \mathbf{V}^T_{s\times u} \tag{2.3}$$

$$\hat{\boldsymbol{c}}_{\boldsymbol{k\times s}} = \boldsymbol{U}_{\boldsymbol{k\times s}} S_{\boldsymbol{s\times s}} \boldsymbol{T}^{-1}_{\boldsymbol{s\times s}} \tag{2.4}$$

2.2 Entropy minimization method in chemometrics

Entropy minimization is known to be a powerful pattern recognition tool (Watanabe, 1981) and is associated with the principle of simplicity. Sasaki *et al* (1983) have shown that it is possible to transform the eigenvectors from the second moment of the spectroscopic data (i.e. basis vectors similar to the right singular vectors $\boldsymbol{V^T}$), into a set of vectors which approximate the shape of the pure component absorptivities. The procedure is based on solving a Shannon's entropy minimization problem (Kanpur, 1993). The Sasaki's algorithm used equation (2.5) as its objective function, where the entropy function h_{sv} is given by the normalized second derivatives of the estimated absorptivities (equation 2.6).

$$\underset{w.r.t.T_{s\times s}}{min}(G) = -\sum_{1}^{S}\sum_{1}^{u} h_{su}\, ln(h_{su}) \tag{2.5}$$

$$h_{sv} = \frac{\left|\hat{a}''_{su}\right|}{\sum_{u=1}^{L}\left|\hat{a}''_{su}\right|} \tag{2.6}$$

Sasaki used equation (2.3) to get estimated pure spectra $\boldsymbol{\hat{a}}_{s\times u}$ by using a square matrix $\boldsymbol{T}_{s\times s}$. Equation (2.3) reconstructs out all potential pure spectra in one optimization. This method is called "square *s*×*s* problem". For an unknown system, the number of species *s* is unknown. When Sasaki's entropy minimization method is applied to get pure spectra, an arbitrary number of species, *s*, should be used, *i.e.* the arbitrary number should exactly equal to the real number of components in system.

In Sasaki's method, two constraints should be imposed to ensure the non-negativity of estimated pure component spectra and concentrations; it is achieved by introducing a penalty function into the objective function, i.e. equation (2.7). Note that the pure component spectra and concentrations matrices are related to the rotation matrix $\boldsymbol{T}_{s\times s}$ in equations (2.3), (2.4) and (2.6), respectively.

$$\min_{w.r.t.T_{s\times s}} (G) = -\sum_{1}^{S}\sum_{1}^{u} h_{su}\, ln(h_{su}\,) + P(\hat{a},\hat{c}) \qquad (2.7)$$

where P is the penalty function and defined by equation (2.8), F_1 and F_2 are the functions associated with the pure component spectra and concentration matrices respectively; γ_a and γ_c are penalty factors that are empirical constants. The detail about these equations, refer to Sasaki's paper (1983)

$$P(\hat{\boldsymbol{a}},\hat{\mathbf{c}}) = g_{\,a}F_1(\hat{\boldsymbol{a}}) + g_{\,c}F_2(\hat{\mathbf{c}}) \qquad (2.8)$$

Following Sasaki's method, Zeng and Garland (1998) suggested the use of a 4th order derivative (for high quality differentiable data) within an entropy type functional and reformulated an appropriate objective function. Such entropy function is anticipated to produce final approximations for pure component spectra which are smoother and more symmetric, and possess fewer spectral artefacts arising from other components in the multi-component solution.

Pan *et al* (2000) proposed an algorithm using weighted spectral regions either on the entire spectrum or part-of-the-spectrum. This method is potentially very useful for problems where signal variance differs greatly from one region of the spectrum to another, and where one spectral window may contain very highly overlapping features.

The above efforts dealt with systems possessing very few (e. g. two) components, due to the limitations of the model used and the optimization algorithms used. Widjaja and Garland (2002) successfully extended the entropy minimization algorithm to a synthetic seven-component system by using Corana's simulated annealing (SA) (Corana, 1987) as the optimization method, so that large-scale entropy minimization problems with multiple observable species could be solved. In above-mentioned algorithms, the rotation square matrix $\boldsymbol{T}$ with dimensions of $s \times s$ has to be solved at the same time by an optimization method. This "square problem" may encounter computational difficulties when s increases, band positions shift and band shapes change.

As mentioned previously, entropy minimization methods have problem in deciding the number of components s in system. For a real system, the observable number of species would be determined by the significant number of eigenvectors. There are several statistical criteria for determining the significant number of eigenvectors (Carey, 1975) in entropy minimization methods, Malinowshi F-test method (Malinowshi, 1990 and 1999) was used to determine the number of observable species in system.

Other than reconstructing pure spectra from mixture, entropy minimization method can be used in many fields. Chen *et al* (2002) used Shannon's entropy minimization method in automatic phase correction of ^{1}H NMR spectra. The results of automatic phase correction are found to be comparable to, or perhaps better than, manual phase correction. Chen and Garland (2002) applied entropy minimization method to precondition ***in-situ*** FTIR spectra. By their method, background spectra such as H_20 and CO are deduced from experimental spectra that makes kinetic research easier.

2.3 Band-target entropy minimization (BTEM)

Based on entropy minimization method, Chew *et al* (2002) developed a band-target entropy minimization method (BTEM). Different from reconstructing all the pure spectra at the same time, BTEM reconstructs one pure spectrum every time. For square problem entropy minimization method, when a certain number of $\boldsymbol{V^T}$ vectors used in optimization, the same number of pure spectra will be reconstructed even the assumptive number of components *s* is wrong. In BTEM, whatever the number of $\boldsymbol{V^T}$ vectors used in the optimization, the result is always one spectrum. It is less important to know the exact number of observable species in BTEM. A rough number of $\boldsymbol{V^T}$ vectors, say $j > s$, are taken from $\boldsymbol{V^T}$ matrix, these *j* vectors are then transformed, one-spectrum-at-a-time, into an estimate pure component spectra. In this way, the "square problem" of solving $s \times s$ unknowns is avoided and instead, each with $1 \times j$ unknowns is solved. The utility of BTEM in solving the "blind" spectral reconstructing problem (e.g. given no prior information) arises because (1) no assumption concerning the number of observable species is necessary (2) spectral quality after reconstruction is greatly improved because non-linearity can be taken into account by targeting a small band of peak and (3) the algorithm is goal oriented – one targets an observed feature in $\boldsymbol{V}^T$ and then recovers the associated entire function (pure component spectrum).

The BTEM algorithm is initiated by targeting a feature in the matrix $\boldsymbol{V}^T = [\boldsymbol{V}_{V\times k}, \boldsymbol{V}_{v\times(v\text{-}k)}]^T$, where only the first k vectors have physical meaning (k experimental measurements were made). The algorithm retains this interesting spectral feature and forces a reconstruction of the associated entire pure component spectrum. In terms of mathematical aspects, the main difference between the original square problem entropy minimization and BTEM is the way to reconstruct pure spectra. The estimated pure spectrum in every reconstruction is shown in equation (2.9)

$$\boldsymbol{a}_{\boldsymbol{1}\times n}{}^{\mathrm{est}} = \boldsymbol{T}_{1\times \mathrm{j}} \times \boldsymbol{V}_{\mathrm{j}\times n}^{\mathrm{T}} \tag{2.9}$$

The objective function of BTEM is similar to that of square problem entropy minimization. Compare with “square problem” entropy minimization method, BTEM uses a column $\boldsymbol{T}_{1\times j}$ to replace the rotate square matrix $\boldsymbol{T}_{s\times s}$ as in equation (2.3) and reconstructs spectra one by one.

The BTEM algorithm has been successfully applied to many real systems such as FTIR (Chew, 2002), RAMAN (Ong, 2001; Sin, 2002) and *in-situ* reactions (Widjaja *et al*, 2002) and has shown its considerable usefulness in finding unknown components and dealing with the pure component spectra of unstable species (Li *et al.*, 2002). It should be noted that these applications involved only continuous spectra like FTIR and RAMAN.

2.4 Pure spectra reconstruction methods in Mass spectrometry

Different from those continuous spectra such as IR and RAMAN, mass spectra are discrete and non-differentiable. Different kinds of chemometric methods are used to deal with MS. Sharaf and Kowalski (1982) reported applications of Lawton and Sylevstr (Lawton and Sylevstr, 1971) self- modelling curve resolution methods to resolve overlapping GC-MS peaks of a binary mixture. Chen and Huang (1981) presented a method of spectral estimation for three components. Their method required regions of unique spectral response for each of the components. Ritter *et al*. (1976) used factor analysis of mass spectra to identify the number of components in mixture and predicted other unknown mixture. In their method, they discarded some particular *m/z* positions from the data matrix and identified components that have unique mass positions. Visual checking is needed in their method to delete some *m/z* positions. The peak discarding method would be useful in finding major components but would be dangerous in dealing with minor components.

Feng and Liang (2000) proposed an approach to retrieve components' mass spectra. The procedure first checked the weighted reference to determine the presence of the reference spectra in a mixture and then used a non-negative least squares regression to find the contributions of the components in the mixture. In Gong's newly published papers, several methods are used to retrieve pure mass spectra from mixture especially to study Chinese traditional medicines (Gong *et al*, 2001a and 2001b). In these methods, reference database should be used to search possible pure components in the mixture just like many other library search methods (Tong and Cheng, 1999) in MS.

Phalp *et al* (1995) reported a modified Simple-to-use interactive self-modelling mixture analysis (Windig, 1992) (SIMPLISMA) approach (TSIMPLISMA) that used the concept of "representative-spectra" of MS. Instead of using the pure-variables in the SIMPLISMA, TSIMPLISMA defines and evaluates the purity of the spectra. The spectrum with the highest purity value represents the "representative" spectrum for a component and the contributions from this component are removed from the data set. The procedure proceeds sequentially for the remaining components. The incorporation of the expert knowledge also contributes to the usefulness of TSIMPLISMA. Windig *et al.* (2002) combined conventional SIMPLISMA (for pure variables of wide peaks) with second-derivative spectra data (for pure variables of narrow peaks, overlapping with the wide peaks) get pure spectra. In summary, the identification of pure component mass spectra from mixture spectra usually requires some sort of information concerning the reference spectra or expert knowledge. Although SIMPLISMA method does not need reference, when a mixture presented, an arbitrary number of pure components in system need to be specified, also their estimated results always have negative MS peaks.

2.5 Optimization

Optimization methods are used in many fields of science, engineering and business. They are also frequently used in chemometrics field.

Optimization methods can be classified into two groups in terms of their optimization problems. One group is constrained problem, the other is unconstrained problem. Commonly, solving a constrained problem is not only time-consuming but also much more difficult than its counterpart – the unconstrained problem. In practice, unconstrained optimization is usually preferred. A constrained optimization is easy to be dealt as an unconstrained optimization method by converting its constraints to penalty functions.

There are many methods dealing with unconstrained problems. Each method has its advantages and disadvantages in different kinds of problems. In general, computational methods to deal with unconstrained problems use iteration methods. These optimization methods may fall into two categories: direct methods and indirect methods. Direct method such as simplex (Spendley *et al*., 1962) and random search method (Dixon and James, 1980) use only the value of objective function. The indirect methods such as steepest descent method and Newton's method (Edgar *et al*., 2001) use objective functions derivatives in finding their search directions whenever their derivative formulas are implicit or explicit. Most of the time, the objective functions are differentially implicit, therefore numerical differential methods are needed which would result into computational inefficient and make problems more difficult. Compared with indirect methods, direct methods generally are less efficient but much robust.

All optimization methods also can be classified as local minima optimization methods or global optimization methods. Local optimization methods would only find the nearest local extrema depending on their initial points. Global optimization methods can find global minima wherever the initial points are. Global optimization methods can be sub-classified as two groups: exact methods and heuristic methods. Exact methods can find the global optimization points and can prove that they have found. Branch-and-bound methods (BB), methods based on interval arithmetic (Kearfott, 1996) and some multistart procedures (Rinnooy and Timmer, 1987; Locatelli and Schoen, 1999) belong to the category of exact methods. Heuristic methods can not prove that they have found global minima although they would find often. Simulated annealing (SA), genetic algorithm (GA) and scatter search belong to heuristic method.

BTEM used simulated annealing (SA) method as its global optimization method. SA method is a popular method today. Original SA method (Kirkpatrick *et al*, 1983) is a single point stochastic search technique where in each iteration, a neighbour point, whenever higher or lower, is created from a current solution by an acceptance probability. SA is a naturally sequential algorithm and difficult to be a parallelize algorithm (chen *et al.*, 1998). There are many efforts on extending SA into parallel algorithm (Onbasoglu and Ozdamar, 2001; Chu *et al*., 1999). Chen *et al.* (1998) reported a hybrid parallel SA method with GA to take both methods' merits. Although SA is a global optimization method, its optimization result sometimes is affected by the starting points.

Before BTEM method, Sasaki, Zeng and Pan used Nelder and Mead's simplex method (Nelder and Mead, 1965) as their optimization method in a low dimension square problem entropy minimization. Simplex method is an efficient and robust method compared with other direct search method. Although simplex method is not a global method, it is not easy to be trapped on a local minimum and has the ability of following the gross behaviour of the test functions despite many local minima. Simplex method is fit for solving problems with small number of variables. For large-scale problem, it is not reliable. Huang *et al* (1998) extended the simplex method to a global optimization method to synthesized and real magnetoencephalography problems by using many random starting points to find many local minima. By comparing these minima, the lowest value is treated as the global minimum. This kind of method is a heuristic method for it can not guarantee to find a global minimum. But if enough start points given, the global minimum would be found.

Chapter 3: Development of tBTEM

In this chapter, emphasis is placed on finding the right kind of objective function form and fast optimization method for reconstructing pure component spectra from discrete data. Band-target entropy minimization (BTEM) is mainly used for dealing with continuous data such as FTIR, UV and RAMAN. BTEM's objective function uses 1st, 2nd and 4th derivatives and does not fit for discrete data. Since the non-differentiable nature of discrete data, it would be very important to find a suitable objective function for discrete data. On the other hand, the optimization method used in BTEM is simulated annealing (SA). Although SA's performance in BTEM is very good, its optimization speed is relatively slow especially when the number of data channel υ increases.

3.1 BTEM

BTEM algorithm reconstructs one pure component spectrum at every reconstruction by targeting a single peak (or a small range of interval) every time. The detailed steps are shown below.

1. Perform singular value decomposition (SVD) on a spectroscopic data array $\boldsymbol{A}_{k\times v}$.

$$\boldsymbol{A}_{\boldsymbol{k}\times u} = \boldsymbol{U}_{\boldsymbol{k}\times\boldsymbol{k}} \times S_{\boldsymbol{k}\times u} \times \boldsymbol{V}^{\boldsymbol{T}}_{u\times u} \tag{3.1}$$

   where k is the number of experiments taken, υ is the number of spectroscopic data channels ($k < \upsilon$). Since the $\boldsymbol{V^T}$ vectors after k^{th} vector has no physical meaning in matrix $\boldsymbol{V}^{\boldsymbol{T}}_{u\times u}$ and it can be truncated to$\boldsymbol{V}^{\boldsymbol{T}}_{\boldsymbol{k}\times u}$ (Garland *et al.*, 1997; Golub and Van Loan, 1996, Malinowski, 1991).

2. Inspect the right-singular matrix $\boldsymbol{V^T}_{k\times u}$ to identify last k - j vectors that appear to represent only noise and discard these k - j vectors. (Usually, every $\boldsymbol{V^T}$ vector is plotted in a figure and checked. The abscissa and y-axis of the figure represent channel number and intensity respectively.) This leaves the truncated right singular matrix $\boldsymbol{V^T}_{j\times u}$ (j is the number of $\boldsymbol{V^T}$ vectors which will be used in following targeting).

3. Identify an interesting local extremum in the first j singular vectors by checking every $\boldsymbol{V^T}$ vector. This local extremum will be targeted by BTEM. Usually, a small interval of wave numbers around the extremum is chosen to deal with peak shift in real system.

4. A set of random numbers is used as an initial guess for the vector $\boldsymbol{T}_{1\times j}$ by optimization method. $\boldsymbol{T}$ is always updated by optimization method until the right value is found. Usually simulated annealing (SA) is used as the optimization method. Then the spectrum $\boldsymbol{a}^{\text{est}}$ is estimated by equation (3.2).

$$\boldsymbol{a}_{1\times u}{}^{\text{est}} = \boldsymbol{T}_{1\times j} \times \boldsymbol{V}_{j\times u}^{\text{T}} \tag{3.2}$$

5. Normalize the estimated spectrum $\boldsymbol{a}^{\text{est}}$ by the maximum peak within the targeted region. Let the normalized spectrum be denoted as $\hat{\boldsymbol{a}}$. If only a peak instead of a small region used, the denominator of equation (3.3) is the value of this peak.

$$\hat{\boldsymbol{a}}_{1\times u} = \frac{\boldsymbol{a}_{1\times u}{}^{est}}{max(a')} \tag{3.3}$$

where a' is the targeted region within the estimated spectrum.

6. Formulate the objective function in terms of the normalized $\hat{\boldsymbol{a}}$

$$min(G) = -\sum_{\upsilon} h_\upsilon ln h_\upsilon + P(\hat{\boldsymbol{a}}_{1\times\upsilon}, \hat{\boldsymbol{c}}_{k\times 1}) \tag{3.4}$$

where the entropy function is defined by equation (3.5) and the first order derivative is used. Usually, 1st, 2nd and 4th order derivatives have been used in the entropy function for differentiable spectra, but this varies from case to case. The penalty function ***P*** is defined by equation (3.6) which contains two terms, non-negativity of estimated spectrum and non-negativity of estimated concentrations. These two penalty functions are defined in equations (3.7) and (3.8) and their coefficients are defined in equations (3.9) and (3.10). The estimated concentration column is defined in equation (3.11).

$$h_\upsilon = \left|\frac{d\hat{a}_\upsilon}{d\upsilon}\right| \Big/ \sum \left|\frac{d\hat{a}_\upsilon}{d\upsilon}\right| \tag{3.5}$$

$$P(\hat{\boldsymbol{a}}_{1'\upsilon}, \hat{\boldsymbol{c}}_{k'1}) = \gamma_a F_1(\hat{\boldsymbol{a}}_{1'\upsilon}) + \gamma_c F_2(\hat{\boldsymbol{c}}_{k'1}) \tag{3.6}$$

$$F_a(\hat{a}_{1\times\upsilon}) = \sum_\upsilon (\hat{a}_\upsilon)^2 \quad \forall \ \hat{a}_\upsilon < 0 \tag{3.7}$$

$$F_c(\hat{c}_{k\times 1}) = \sum_k (\hat{c}_k)^2 \quad \forall \ \hat{c}_k < 0 \tag{3.8}$$

$$\gamma_a = \begin{cases} 0 & F_1(\hat{\mathbf{a}}_{1\times u}) < 10^{-3} \\ 10 & 10^{-3} \le F_1(\hat{\mathbf{a}}_{1\times u}) < 10^{-2} \\ 10^4 & F_1(\hat{\mathbf{a}}_{1\times u}) \ge 10^{-2} \end{cases} \tag{3.9}$$

$$\gamma_c = 10^3 \quad \forall \ F_2(\hat{C}_{k\times 1}) \tag{3.10}$$

$$\hat{\boldsymbol{c}}_{k'1} = \boldsymbol{A}_{k'u} \ ' \ \hat{\boldsymbol{a}}^T_{u'1} \ ' \ (\hat{\boldsymbol{a}}_{u'1} \ ' \ \hat{\boldsymbol{a}}^T_{u'1})^{-1} \tag{3.11}$$

7. Check the objective function value *G* against a stopping criterion. If the stopping criterion is not met, then generate another ***T*** by the optimization method. Repeat steps 4 to 7 until one ***T*** associated with a pure component spectrum is reconstructed.

8. Repeat steps 3 to 7 by targeting another extremum to find the remaining pure component spectra.

Practically, to find a pure component spectrum, first an interested peak/band is chosen from $\boldsymbol{V}^{\boldsymbol{T}}$ vectors. An estimated spectrum will come out after targeting this peak/band. If the targeted peak/band is not the highest peak/band within the estimated spectrum, another round of targeting should be performed by using the highest peak/band's position within the estimated spectrum. Commonly, the second result would be better. If the new peak in the second estimated spectrum is the highest one, it means a pure component spectrum was found. The reason for using the highest peak to targeting is its greatest signal to noise ratio. Although many pure component spectra can be found by their highest peaks, sometimes if more than two components' highest peaks locate at the same position, only one pure component spectrum can be reconstructed by targeting at their common highest peak. The others pure spectra should be reconstructed by targeting different peaks.

3.2 Modification on objective function for discrete spectra

The original entropy functional equation (3.4) causes a problem with non-differentiable discrete data like mass spectra. Therefore, it was suggested that the peak heights instead of their derivatives should be used to formulate the objective function. It was also found that the use of normalized peak heights of estimated spectrum instead of the expression of $h_u \ln h_u$ has good performance in reconstructing spectra. The objective function for discrete spectra required less computational time in evaluating the objective function. Specially, the following objective function is proposed.

$$min(G) = \sum_u \hat{\boldsymbol{a}}_{1\times u} + P(\hat{\boldsymbol{a}}_{1\times u}, \hat{\boldsymbol{c}}_{k\times 1}) \qquad (3.12)$$

where the penalty functions are defined similar to that of the BTEM algorithm. They are expressed in equations (3.13) and (3.14).

$$P(\hat{\boldsymbol{a}}_{1'u},\hat{\boldsymbol{c}}_{k'1}) = g_{a}F_{a}(\hat{\boldsymbol{a}}_{1'u}) + g_{c}F_{c}(\hat{\boldsymbol{c}}_{k'1}) \tag{3.13}$$

$$F_a(\hat{a}_{1\times\upsilon}) = \sum_{\upsilon}(\hat{a}_{\upsilon})^2 \quad \forall \;\; \hat{a}_{\upsilon} < 0$$

$$F_c(\hat{\boldsymbol{c}}_{k\times 1}) = \sum_{k}(\hat{c}_{k})^2 \quad \forall \;\; \hat{c}_{k} < 0 \tag{3.14}$$

$$\gamma_a = 10^4; \quad \gamma_c = 10^3;$$

Besides the changes of the objective function, the number of the coefficients of penalty function for non-negativity of estimated spectrum is changed from three to one. In general, the objective function for discrete spectra is simpler compared to that of BTEM.

3.3 Weighted V^T

As mentioned in the introduction section of SVD, every $\boldsymbol{V^T}$ vector is a unit vector because $\boldsymbol{VV^T} = \boldsymbol{I}$. As shown in equation (3.2), every $\boldsymbol{V^T}$ vector plays the same important role in BTEM algorithm. The BTEM algorithm can not tell the differences between vectors mainly containing useful information and vectors mainly containing noise. Furthermore, as we know, the vectors in the $\boldsymbol{V^T}$ matrix are ordered according to their contribution to the total variance of the observations. Therefore, the first few vectors associate with real chemically important signals in the system and the rest associate primarily with the instrumental and experimental noise. This means even in an nth $\boldsymbol{V^T}$ vector (n >> the number of components), it would have useful information in it together with a large amount of noise. In the original BTEM algorithm, in order to fully use the information, a very large number of vectors (>> the number of components) are used in optimization. As mentioned, because the vectors associated with both real signals and noise play equal roles in the BTEM algorithm, adverse effects on the spectral reconstruction may be introduced when too many $\boldsymbol{V^T}$ vectors are used in optimization.

BTEM avoids the problem in deciding or guessing the exact number of species s presented in system, as that occurs in square problem entropy minimization. BTEM still has difficulty in deciding how many $\boldsymbol{V^T}$ vectors $\boldsymbol{j}$ should be used in optimization. To fully use the information and not to introduce much noise in system, there should be an optimum value of the number of $\boldsymbol{V^T}$ vectors which should be used in system. For less noise systems, it would be a minor problem; but for noise system, it would be very important.

Obviously, it would be advantageous to lower the effect of noise while retaining the useful information. Instead of using the $\boldsymbol{V^T}$ vectors directly, we multiply the $\boldsymbol{V^T}$ matrix by a set of weights, namely, the diagonal matrix $\boldsymbol{S}$ that is readily available from SVD (see section 2.1). Thus, the significance of the vectors associated with the real signals is increased and the effect of noise is reduced. In particular, equation (3.2) is modified as below.

$$\boldsymbol{a}^{est}{}_{1'u} = \boldsymbol{T}_{1'j}\ '\ (\boldsymbol{S}_{j'j}\ '\ \boldsymbol{V}^{T}_{j'u}) \qquad (3.15)$$

where $\boldsymbol{S}$ is the square diagonal matrix containing singular values (see section 2.1). By scaling the right singular vectors, the importance of the noise vectors is reduced during optimization. Accordingly, it is less likely that the optimization method becomes trapped in a local minimum. In other words, the optimization is less sensitive to the choice of the number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors.

3.4 Two-peak/band target method

In Step 5 of the BTEM algorithm, the estimated spectrum is normalized by using the maximum value within the targeted band. Since the targeted peak plays a central role in retrieving the entire associated spectrum, using a higher peak is advantageous due to the larger signal to noise ratio and the result is often less affected by the noise presented. For systems having moderately overlapping spectra, the BTEM algorithm is usually successful in estimating all the pure spectra with their highest peaks. However, if strongly overlapping spectra present at the highest peak position / channel, only the spectrum with the smallest objective function value would be recovered. To deal with this challenge, a two-peak/band targeting strategy uses a modification of equation (3.3) as shown below, where a′ and a″ are the two targeted peaks/bands within the estimated spectrum.

$$\hat{\boldsymbol{a}}_{1'u} = \frac{\boldsymbol{a}_{1'u}{}^{est}}{max(a') + max(a'')} \tag{3.16}$$

In two-peak/band targeting method, it would be better that the two highest peaks within a spectrum are used in equation (3.16). Since the chances that two pure component spectra have their highest peaks locating at the same positions / channels are small, permutation for different pairs of the highest peaks provides an exhaustive search for pure component spectra. Consequently, it is possible to recover most if not all the pure component spectra that are overlapping. If extremely overlapping pure component spectra are suspected in the system, ***e.g.*** more than two pure spectra whose highest peaks locate at the same location, or one spectrum is almost totally overlapped by another spectrum; a multiple targeted peak strategy may be employed to handle such a complex system. In these cases, the normalization in equation (3.16) will be replaced by equation (3.17).

$$\hat{\boldsymbol{a}}_{1\times u} = \boldsymbol{a}_{1\times u}{}^{est} \Big/ \sum_{i=1}^{p} max(a_i\ ); \qquad \boldsymbol{p} > 2 \tag{3.17}$$

When one peak is targeted, every pure component in the system will have an objective function value. Since SA optimization method only finds out the value of global minimum, the spectrum whose objective function value corresponding to global minimum would be reconstructed out. The other spectra with higher objective function values locating at different local minima would not be found. It is found that each local minimum would indicate a pure spectrum in one-peak-targeting method, there is no local minima indicating mixture spectra which are combined by two pure spectra. When more than one peak is used in targeting, the searching spaces are more complex than that of single-peak-targeting method. There are many local minima which indicate mixture spectra. For example, in two-peak targeting method, there are many local minima indicating spectra which are linearly combined by two pure spectra. In three-peak targeting-method, there are many local minima whose spectra are combined by two or three pure spectra. Therefore, when multi-peak-targeting method is used, it is not always good to engage many peaks (***e.g.*** >4 peaks) in targeting. Using many peaks in targeting would make the searching space more complex. It would make reconstruction more difficult.

3.5 Overall tBTEM algorithms

With all those modifications in the previous sections, the overall tBTEM algorithm can be expressed in the following steps.

1. Perform singular value decomposition (SVD) on spectroscopic data array $\boldsymbol{A}_{k\times u}$ as equation (3.1), where k is the number of experiments taken, υ is the number of spectroscopic data channels ($k < \upsilon$). After truncating off physically meaningless parts of right singular matrix $\boldsymbol{V}^{\boldsymbol{T}}_{u\times u}$ and zero part of diagonal matrix $\boldsymbol{S}_{\boldsymbol{k}\times u}$, they become to $\boldsymbol{V}^{\boldsymbol{T}}_{k\times u}$ and $\boldsymbol{S}_{k\times k}$ (Garland *et al.*, 1997; Golub and Van Loan, 1996, Malinowski, 1991).
2. Inspect every right-singular vector in matrix $\boldsymbol{V}^{\boldsymbol{T}}_{k\times u}$ to identify last k - j vectors that appear to represent only noise. Discard these k - j vectors. This leaves the truncated right singular matrix $\boldsymbol{V}^{\boldsymbol{T}}_{j\times u}$.
3. Identify two interesting local extrema in the first j singular vectors to be targeted by tBTEM. Usually, only these two extrema themselves will be chosen in discrete spectra.
4. A set of random numbers is used as an initial guess for the test vector $\boldsymbol{T}_{1´j}$ by optimization method. $\boldsymbol{T}$ is always updated by optimization method until the right spectrum is found. Then the spectrum $\boldsymbol{a}^{\mathrm{est}}$ is estimated by equation (3.18)

$$\boldsymbol{a}^{est}{}_{1´u} = \boldsymbol{T}_{1´j}\ ´\ (\boldsymbol{S}_{j´j}\ ´\ \boldsymbol{V}^{T}_{j´u}) \tag{3.18}$$

5. Normalize the estimated spectrum $\boldsymbol{a}^{\mathrm{est}}$ by targeting two peaks/bands with equation (3.19). Let the normalized spectrum be denoted as $\hat{\boldsymbol{a}}$.

$$\hat{\mathbf{a}}_{1\times u} = \frac{\mathbf{a}_{1\times u}^{est}}{max(a') + max(a'')} \tag{3.19}$$

where a' and a'' are the targeting peaks within the estimated spectrum.

6. Formulate the objective function by equation (3.20) in terms of the normalized $\hat{\boldsymbol{a}}$

$$min(G) = \sum_u \hat{\boldsymbol{a}}_{1\times u} + P(\hat{\boldsymbol{a}}_{1\times u}, \hat{\boldsymbol{c}}_{k\times 1}) \tag{3.20}$$

$$P(\hat{\boldsymbol{a}}_{1\times u}, \hat{\boldsymbol{c}}_{k\times 1}) = g_a F_a(\hat{\boldsymbol{a}}_{1\times u}) + g_c F_c(\hat{\boldsymbol{c}}_{k\times 1}) \tag{3.21}$$

$$F_a(\hat{\boldsymbol{a}}_{1\times u}) = \sum_v (\hat{a}_u)^2 \quad \forall\ \hat{a}_u < 0$$

$$F_c(\hat{\mathbf{c}}_{k\times 1}) = \sum_k (\hat{c}_k)^2 \quad \forall\ \hat{c}_k < 0 \tag{3.22}$$

$$g_a = 10^4; \quad g_b = 10^3;$$

The penalty function $\boldsymbol{P}$ is defined in equation (3.21) to guarantee non-negativity in the reconstructed spectrum and concentrations (equation (3.22)). The estimated concentrations are defined in equation (3.23).

$$\hat{\boldsymbol{c}}_{k\times 1} = \boldsymbol{A}_{k\times u} \times \hat{\boldsymbol{a}}_{u\times 1}^{T} \times (\hat{\boldsymbol{a}}_{u\times 1} \times \hat{\boldsymbol{a}}_{u\times 1}^{T})^{-1} \tag{3.23}$$

7. Check the objective function value against a stopping criterion. If the stopping criterion is not met then generate another $\boldsymbol{T}$ by optimization method. Repeat steps 4 to 7 until one pure component spectrum is reconstructed.

8. Repeat steps 3 to 7 by targeting other extrema to find the remaining pure component spectra.

Other than FTIR and NMR spectra, mass spectra has no (or tiny) peak shift. Commonly, for tBTEM, peaks themselves only would be used in targeting. tBTEM could easily be changed to a single-peak targeting method by choosing two same peaks.

3.6 Fast multi-start simplex method (FMSS)

In addition to different kinds of entropy functions needed to reconstruct continuous or discrete spectra, optimization methods play very important role in BTEM to find results. The original square problem entropy minimization uses simplex method; but it fails to deal with more than 3 components in a mixture. Although simplex is a robust method and not easily to be trapped in local minima, it is difficult to find a global minimum in higher dimension problems. After Widjaja and Garland (2002) used simulated annealing method as a global minimization method, square problem entropy minimization and BTEM can deal with more than 10 components. For continuously differentiable data such as FTIR and RAMAN, with u = 2500-10000 channels of data, and j = 25-100, typical workstation CPU time for a single spectral reconstruction is 6-12 hours (dual Intel Xeon 500 MHz CPUs, 2GB RAM, Win NT 4.0 workstation).

Multi-start methods attempt to find a global minimum by starting the searches from many different starting points. Many papers focus on this kind of method such as multilevel single linkage method (MLSL), Multiple Local searches with clustering (LC) (Törn, 1978), and controlled random searching (CRS) (Price, 1978). In these methods, none of them embeds local minima searching methods.

Other than these multi-start methods which do not include local minima searching methods, there is another kind of multi-start method which embeds some local minima searching methods. The latter methods start searching from many random starting points by using local minima searching methods. Since every starting point reaches a local minimum, if many starting points are used, the lowest one would be the global minimum. However, the latter methods can not guarantee to find global minima. This kind of multi-start method has one advantage that it can find many different local minima other than global minimum. Multi-start methods combined with local minima searching methods should be classified as heuristic searching methods.

Simplex methods are local minima searching methods. Spendly, Hex and Himsworth (1962) first formulated a sequential simplex method in 1962. It used an equilateral polyhedron in search. Nelder and Mead (1965) refined the Sequential Simplex by permitting the geometric figures to expand and contract continuously during the search to improve its searching efficiency. The Nelder-Mead’s Simplex algorithm is an elegant method for function minimization. Although it is not a global optimization method, it is able to crawl out of some local minima to find better minima. Compared with gradient methods such as Powell's method, Nelder-Mead’s simplex method is generally less efficient but more robust. It neither uses line minimizations nor builds an implicit model of the derivative structure of the function. These aspects of the Nelder-Mead simplex method make it quite popular.

Huang *et al* (1998) successfully used multi-start simplex method to deal with simulated and real magnetoencephalography problems. They combined Nelder-Mead’s simplex method with a great number of initial points, after every starting point reaches its end points with the same stopping criterion, all objective function values were compared and the lowest one was assumed to be the global minimum.

Based on the idea of multi-start methods mentioned above, a global optimization algorithm named fast multi-start simplex method (FMSS) is proposed. It dramatically speeds up the searching speed by using different stopping criteria in different searching steps in order to discard points which are not promising in the future. In BTEM algorithm, FMSS is successfully used to reconstruct pure component mass spectra on both synthesized and real data sets. Compared with multi-downhill simplex method, it dramatically reduces the computational time. Also it is much faster than SA in reconstructing pure mass spectra.

The following procedure specifies the FMSS algorithm. First, FMSS starts from many starting points and converges every starting point at a coarse stopping criterion. Then it retains some points that have the smallest objective function values (the number of the retained points is smaller than the number of the starting points). Second, from these retained points, a new round of searching, which uses a finer stopping criterion, is performed and a smaller number of points are retained again. After several rounds of searching by refined stopping criteria, the final objective function values are compared and the lowest one is assumed to be the global minimum. The detailed steps are shown below.

1. Choose a number of starting points n_0, reduction coefficient of number of searching points $0 < r_p < 1$, coarse stopping criterion ε_0, stopping criterion reduction coefficient r_ε, and the number of searching rounds R.
2. 1st searching round: start searching from every starting point by Nelder-Mead simplex method using the stopping criterion ε_0, compare these different minima and retain the $n_1 = n_0 \times r_p$ temporary terminal points with smallest objective function values. If n_1 is a fraction, then round it to its nearest bigger integer. These temporary terminal points will be used as starting points in the next round.
3. 2nd to R^{th} searching round: start searching from every retained point from the previous round with a decreasing stopping criterion $e_i = e_0 \times (r_e)^{(i-1)}$ (i: the ongoing searching round number, $1 \le i \le R$). Retain $n_i = n_{i-1} \times r_p$ temporary terminal points according to their objective function values.
4. Compare the n_R objective function values, which are obtained from R^{th} searching round, to get the lowest one. The obtained smallest objective function value is assumed as the global minimum.

Although a fixed reduction coefficient r_p is used in the above steps, a changeable reduction coefficient may be used in different situations. For example, when the number of initial points is huge, in the first few searching rounds, one may use a small reduction coefficient r_p so that to reduce the number of points fast. After that, a bigger coefficient could be used in the following searching rounds.

The FMSS method obviously is a heuristic method. As a common property of heuristic methods, it can not guarantee to find a global minimum, but if enough random starting points are given, it would find the global minimum.

Chapter 4: tBTEM Tests on Simulation System

In this chapter, tBTEM and FMSS algorithms are tested on synthesized mixture data sets.

4.1 Tests on tBTEM algorithm

4.1.1 Simulation method for synthesized data set

Ten real pure component spectra are used to synthesize mixture data set for testing tBTEM. These ten components are ethanol, hexane, toluene, acetone, acetonitrile, cyclohexene, acetic acid, $(CH_3)_2CHOH$, CH_2Cl_2 and $CH_3CH_2COCH_3$. The reasons of choosing these ten organic components are:

1) All of them have low molecular weights (< 100).

2) There are many same alkyl groups such as CH_3- and CH_3CH_2- and functional groups such as CH_3CO-.

In MS detection method, these components would have many same charged fragments in electric-impact ionization (EI) method such as $CH_3^{\bullet}$ and $CH_3CH_2^{\bullet}$, therefore there are many overlapped channels in mixture spectra.

Every pure component mass spectrum is obtained from GC-MS (GC: Hewlett-Packard 6890, MS: Hewlett-Packard 5973) with its pure sample. Each spectrum spans m/z = 10 – 100 with 1 m/z interval. The ionization method is EI. Their pure spectra are shown in Figure 4.1.

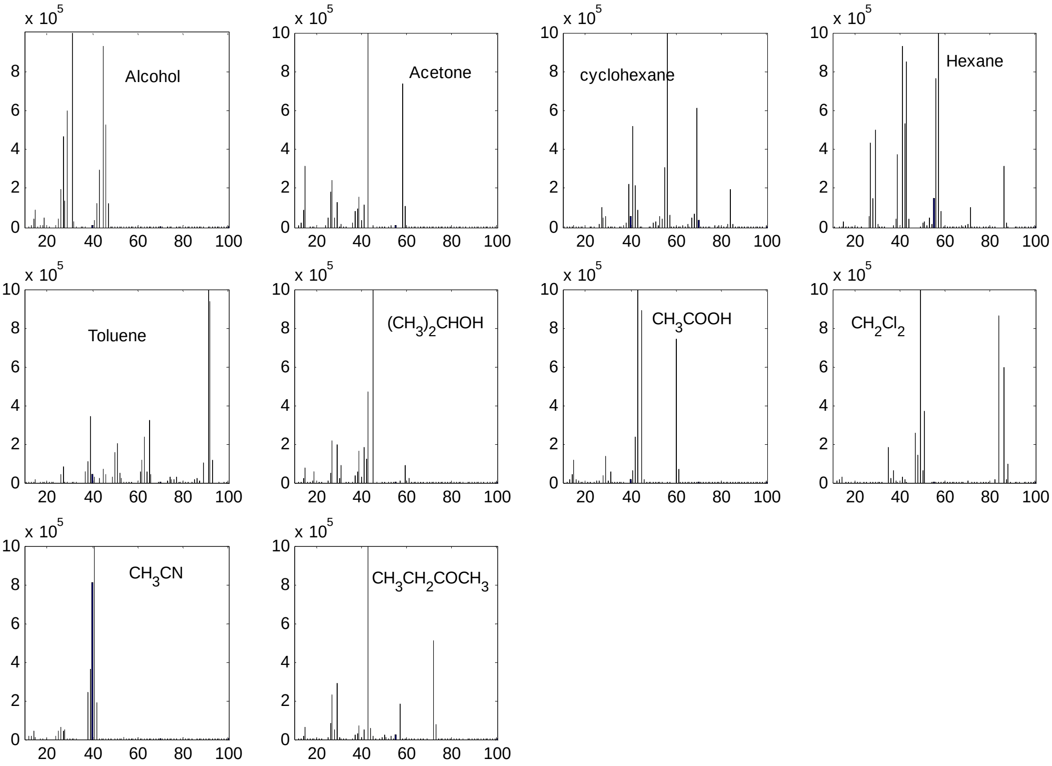

Figure 4.1: Pure mass spectra of 10 organic components

Fifty synthesized mixture data were generated from these 10 pure experimental spectra and minor white noise was added to the spectra. This data set was simulated with randomly generated non-negative concentrations for each mixture (arbitrary units, range from 0 to 1). Accordingly, a concentration matrix, random generated from 0 to 1, for the 50 mixtures $\boldsymbol{C}_{50\times10}$ was obtained. In order to make a meaningful simulation, all pure component spectra were first scaled to a maximum peak height of 10^6 (arbitrary unit). The mass spectra of the mixtures were simulated with equation (4.1), where $\boldsymbol{P}$ denotes the pure spectra of the ten components and e the noise matrix that was randomly generated with a level of 0 to 10^2. In the present study, the total channels for each spectrum was set to 91. The 2-D array set of 50 simulated mass spectra of the mixtures is shown in Figure 4.2.

$$\mathbf{A}_{50\times91} = \mathbf{C}_{50\times10} \times \mathbf{P}_{10\times91} + \varepsilon_{50\times91} \tag{4.1}$$

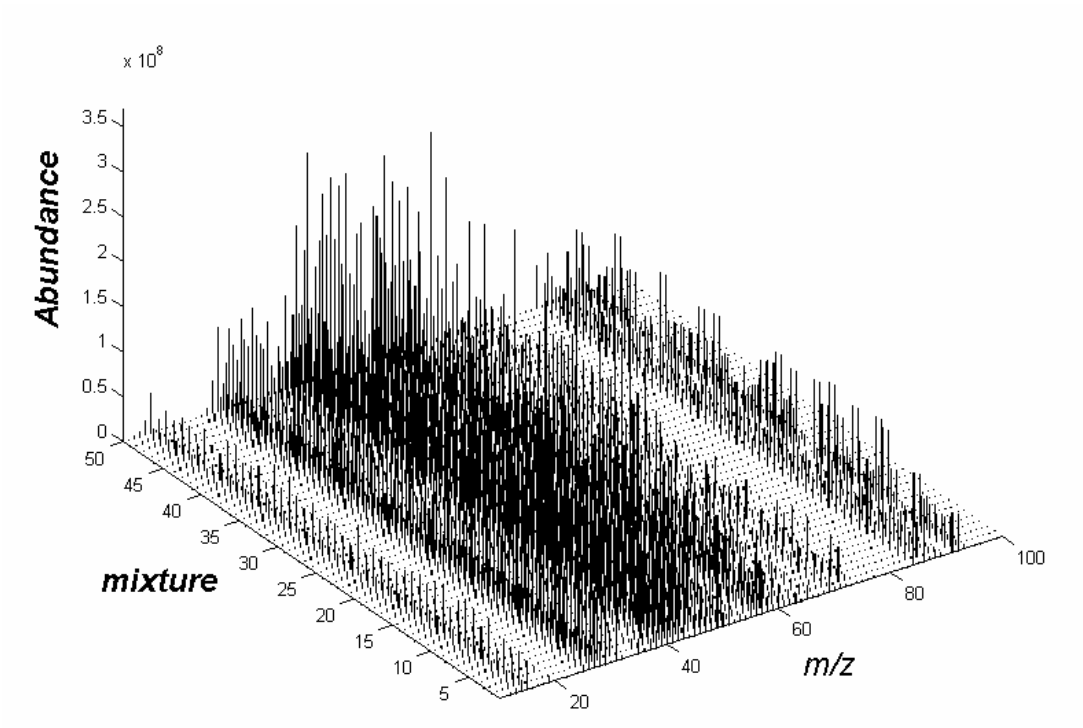

Figure 4.2: 2-D simulated fifty mixture mass spectra

4.1.2 Parameters of SA optimization method

SA parameters: starting temperature T_0 = 10, step variation N_s = 20, temperature reduction coefficient r_T = 0.85, and N_T = max [(100, 5×N)/N_s], where N is the number of decision variables to be optimized.

4.1.3 Results

4.1.3.1 Effect of noise presented in tBTEM system

In this section, studies focus on how the noise in system affects the estimated results. tBTEM objective function used in this test is not weighted. Since there are 10 components present in synthesized mixture, the choices of the number of $\boldsymbol{V^T}$ vectors j are from 10. Table 4.1 shows the estimated results at different number of $\boldsymbol{V^T}$ vectors used in optimization. All these estimated results are gotten by targeting at m/z 43 and 58. To evaluate the similarity between the estimated spectra and reference spectrum (CH_3COOH), an inner product (IP) is used. The inner product provides a measurement of the degree of similarity between two spectra. When the two compared spectra are normalized to unit vectors, the IP value should fall in the range of 0 to 1. If two spectra are identical, the IP equals 1. On the other hand, if the IP is 0, the two spectra are orthogonal.

In Table 4.1, every objective function value is the lowest one among five repeated tests under the same set of parameters, and its corresponding estimated spectrum is used to get IP value. The reference pure spectrum, CH_3COOH, has an objective function of 2.0323 while targeting at m/z 43 and 58. The effect of noise in system is shown in Figure 4.3 and Figure 4.4.

Table 4.1: Noise effect on optimization

Number of V^T Vectors used	**Objective Function Value**	**IP Value**	**Number of V^T Vectors used**	**Objective Function Value**	**IP Value**
10	1.9813	0.9985	23	2.0396	0.9395
15	1.9791	0.9982	24	2.0528	0.9188
17	1.9792	0.9981	25	2.0030	0.9215
18	1.9796	0.9977	26	2.0260	0.9309
19	1.9785	0.9980	27	2.0055	0.9069
20	1.9781	0.9980	28	1.9991	0.9134
21	1.9831	0.9951	2 9	1.9799	0.9388
22	2.0642	0.9143	30	1.9770	0.8961

From Figure 4.3, the objective function values slowly decrease at first. When the number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors increases to 21, the objective function value begins to increase and there is a jump at 22. After that the objective function values decrease. From Figure 4.4, the trend of IP values at different number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors is similar to that of objective function values. At the beginning, the estimated results are good, when the number of $\boldsymbol{V}^{\boldsymbol{T}}$

vectors reaches 21, the estimated spectra become worse. In other words, if the number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors used in BTEM is far more than the number of species in mixture, the estimated result will be affected by noise. In this synthesized data set, the noise level is quite low. So the critical number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors (22) is far bigger than the real number of species (10) in system. While in a real system, the noise level is far bigger than that of this synthesized case, the effect of noise in real case would be more prominent.

It would be better to have an example of noise effect of real system; but for mass spectra, problem is encountered in choosing reference pure spectrum. As we know, pattern for a pure component spectrum is not fixed. At different machines, there would have different kinds of patterns. Even in one machine and in one injection, the patterns would change at different retention time due to non-stationary effect. It would be a problem to choose the reference spectrum. So it would not be feasible to compare the results between the estimated spectra and reference spectrum.

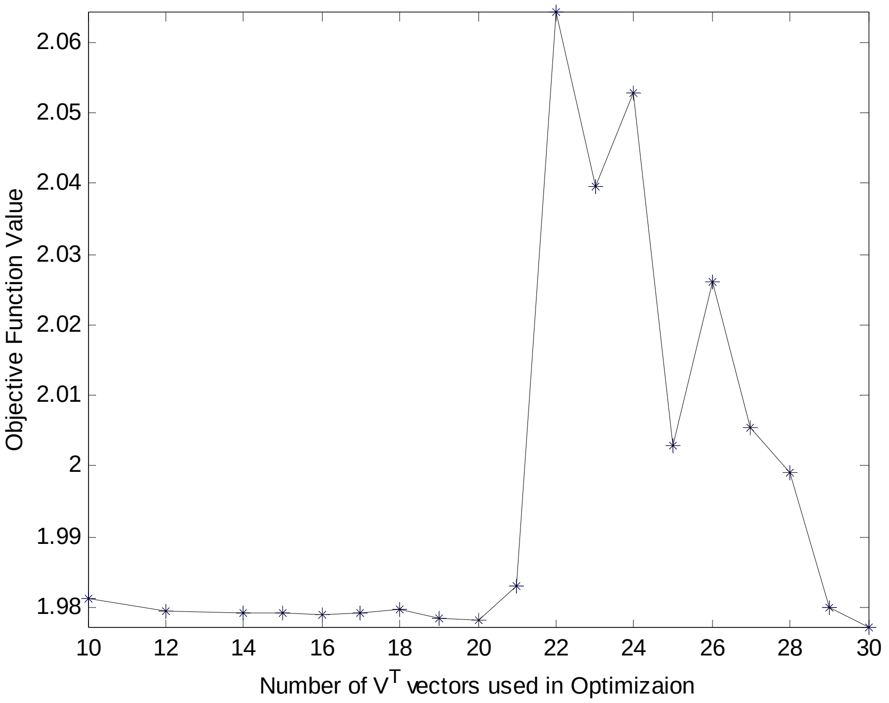

Figure 4.3: Changing of objective function values

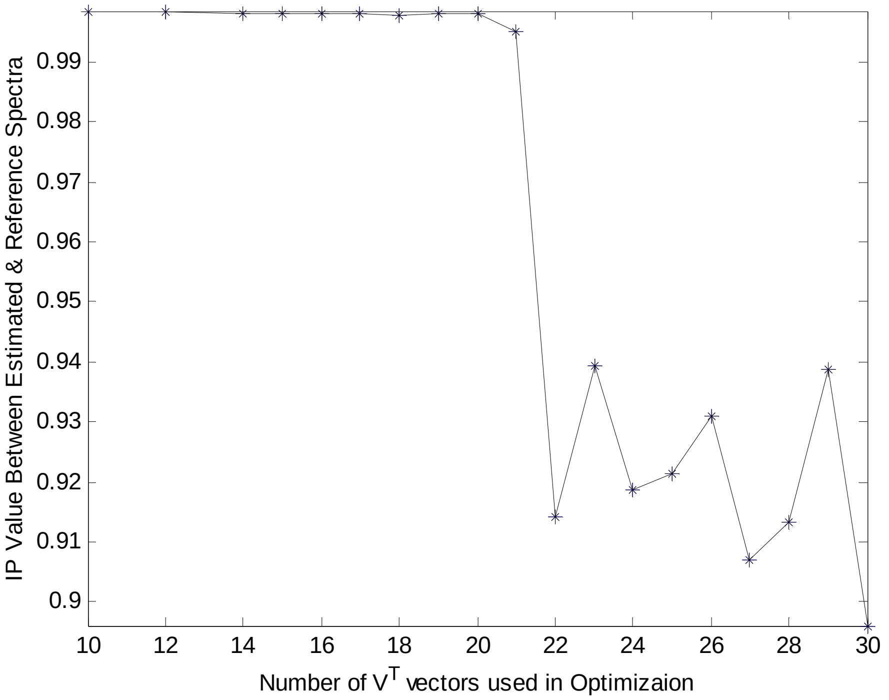

Figure 4.4: Changing of IP values

4.1.3.2 Performance of tBTEM algorithm for discrete data

In this section, the performance of tBTEM is tested. Following tests use the first 10 $\boldsymbol{V^T}$ vectors and target only one peak at every reconstruction (*i.e.* two targeted peaks are same). It should be noted that the number of $\boldsymbol{V^T}$ vectors used is equal to the number of the pure components in the system. The results are shown in Table 4.2 where the targeted peaks and the values of the objective function are given. For comparison purpose, the objective function values of all the real pure spectra are also calculated and listed. It is evident that both the reference and reconstructed objective function values are very similar, where the latter is usually slightly smaller.

Table 4.2: Comparison of estimated spectra with real spectra

Component	Real spectra Objective Function value	Estimated spectra		
		Peak targeted at (m/z)	Objective Function value	Inner Product
Ethanol	4.7648	31	4.7601	1.00000
Hexane	6.6033	57	6.5028	0.99983
Toluene	4.4888	91	4.4761	0.99996
Acetone	4.7731	58	4.645	0.99737
Acetonitrile	2.9309	41	2.9305	1.00000
Cyclohexene	3.9546	56	3.8604	0.99776
$(CH_3)_2CHOH$	2.9658	45	2.8805	0.99886
Acetic acid	4.7739	60	4.7714	0.99996
CH_2Cl_2	3.8334	49	3.8312	1.00000
$CH_3CH_2COCH_3$	2.9028	43	2.8536	0.99972

The performance was further examined by taking the inner product of the estimated and corresponding real spectra. Obviously, all the IP values are very close to 1 which is further confirmation of the excellent reconstructions of all the component spectra.

4.1.3.3 Comparison of weighted and un-weighted V^T vectors

On the basis of the results of section 4.1.3.2, the number of $\boldsymbol{V^T}$ vectors used was increased to 50, the number of mixtures synthesized, without introducing any weights.

It is found that the reconstructed pure component spectra are in general not acceptable. For example, the estimated cyclohexane spectrum (Figure 4.5c) does not resemble the real one (Figure 4.5a). In contrast, the result (Figure 4.5b) was improved greatly when the 50 $\boldsymbol{V}^T$ vectors were weighted by the diagonal matrix obtained from SVD according to equation (3.15). This test indicates that weighted $\boldsymbol{V}^T$ vectors make the algorithm less sensitive to the choice of the number of $\boldsymbol{V}^T$ vectors used. Indeed, although a lot of noise vectors are incorporated into the optimization due to the large number of vectors used, the performance of the algorithm is very good. The weighting is very helpful in spectral reconstruction for real experimental systems because the exact number of components in a system is often unknown *a priori*.

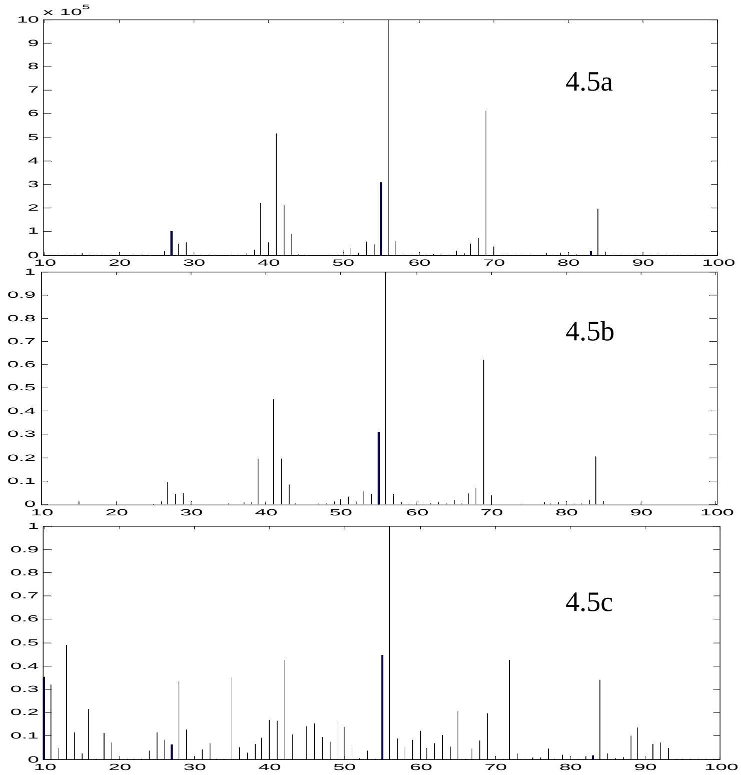

Figure 4.5: The real and estimated cyclohexane spectra.

4.5a: real reference spectrum; 4.5b: estimated spectrum with 50 weighted $\boldsymbol{V}^T$ vectors;
4.5c: estimated spectrum with 50 un-weighted $\boldsymbol{V}^T$ vectors.

Obviously, the objective function value of the spectrum in Fig 4.5c is larger than that of the spectrum in Fig 4.5b. This type of result occurs often, namely, the un-weighted algorithm produces an incorrect pure component spectrum while the weighted algorithm produces an excellent reconstruction. The weighting / scaling of $\boldsymbol{V^T}$ clearly improves the results due to a better condition number. It should be further noted that if the weighting is not performed, a crude estimate of the number of species has to be known *a priori* (see section 4.1.3). Otherwise, the optimization would likely stuck in a local minimum.

4.1.3.4 Comparison of tBTEM and BTEM

Although it is not absolutely necessary, it is common to target the highest peak of a pure component spectrum using the original BTEM because of its large S/N ratio. To target the highest peak, the following steps can be performed. First, choose a peak of interest in the $\boldsymbol{V^T}$ vectors to initialize targeting, and get an estimated spectrum. If in the reconstructed spectrum, the targeted peak is not the highest, then the highest peak is targeted next, and the reconstruction is repeated. The result is often improved compared to the first one.

Difficulties may be encountered, particularly with BTEM, if the highest peaks of two or more spectra all locate at the same channel (m/z). This is because both the global and some of the local minima (obtained during the optimization) correspond to the real pure component spectra. In the BTEM algorithm, only the global minimum is kept and the local minima are discarded. As a result, only one spectrum will be recovered while the others will not be found. For instance, in the present system, $CH_3CH_2COCH_3$,

CH_3COCH_3 (Acetone) and CH_3COOH (Acetic Acid) have their respective highest peaks locating at m/z 43 which indicates the presence of the fragment of $CH_3CO^\bullet$. When the peak at m/z 43 was chosen, the algorithm was successful in reconstructing the spectrum of $CH_3CH_2COCH_3$ with the minimum objective function value equal to 2.9028. In contrast, although the local minima objective function values of 3.5342 and 3.5388 corresponded to the solutions of CH_3COCH_3 and CH_3COOH, they were discarded during the iteration of the algorithm. Therefore, the algorithm, using only a single targeted peak could not find the spectra of CH_3COCH_3 and CH_3COOH using the common highest peak.

As mentioned in the previous case shown in Table 4.2, the single peak targeting algorithm was able to identify the pure component spectra for $CH_3CH_2COCH_3$, CH_3COCH_3 and CH_3COOH. Note that in this case the targeted peak for the reconstruction of $CH_3CH_2COCH_3$ was m/z 43, the highest peak, while the reconstructions of CH_3COCH_3 and CH_3COOH used m/z 58 and 60 which are not the highest peaks for these two spectra.

The tBTEM algorithm has the intrinsic ability to deal with this difficulty. By choosing two big peaks (usually the highest two peaks) of a spectrum, tBTEM obtains the minimum objective function corresponding to the solution of one of the spectra. Then in another run, another pair of highest peaks is used and a new solution corresponding to another spectrum is obtained. For instance in Table 4.3, using the two highest peaks of m/z 43 and 58 resulted in the minimum objective function 2.0306 which led to the reconstruction of spectrum of CH_3COCH_3. Similarly, the choice of the two highest peaks of m/z 43 and 60 obtained the minimum objective function 2.0323 which recovered the spectrum of CH_3COOH.

Table 4.3: Comparison of objective function values by using different targeted peaks

Target at (m/z)	**$CH_3CH_2COCH_3$**	**CH_3COCH_3**	**CH_3COOH**
43 only	***2.9028***	3.5342	3.5388
43 and 58	2.8837	***2.0306***	3.5388
43 and 60	2.9025	3.5156	***2.0323***

Most importantly, in comparison with BTEM algorithm, the tBTEM is less sensitive to the number of $\boldsymbol{V}^T$ vectors used. For example, when an arbitrary number of 35 $\boldsymbol{V}^T$ vectors were used, the resultant spectra of hexane obtained from the BTEM (targeted at m/z 57) and tBTEM (targeted at m/z 57 and 86) are shown in Fig 4.6c and Fig 4.6b respectively. Clearly, the resulting spectrum from tBTEM resembles the real one very closely (Fig 4.6a) while a higher level of noise is present in the spectrum from BTEM. The inner product of real and the estimated spectrum from tBTEM and BTEM are 0.99366 and 0.95635 respectively, which again shows that tBTEM outperforms the single targeted peak algorithm.

The improvement of the performance of tBTEM is probably due to the use of more spectra information, *i.e.* two peaks instead of only a single peak. The enriched spectra information may facilitate the reconstruction of pure component spectra from highly overlapped mixture spectra.

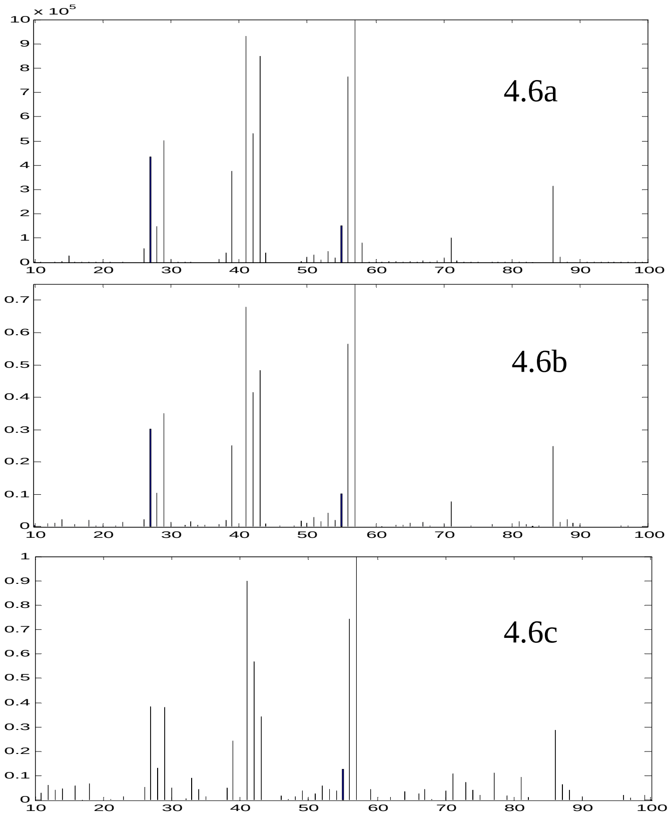

Fig 4.6: Estimated hexane spectra by using one-peak or two-peaks for targeting

4.6a: Real spectrum; 4.6b: targeted at M/z 57 and 86; 4.6c: targeted at M/z 57 only

4.2 Exhaustive search using tBTEM

For either the BTEM or the tBTEM algorithms, each pure spectrum is recovered from a separate run. Therefore, many runs are needed in order to recover all pure spectra, and these runs are directed by the user's choice of the targeted bands (peaks). Clearly, a problem arises when many channels of data must be considered. Recovery of all potential pure component spectra may not practical if the search is left to the user's judgement alone. Accordingly, an automated exhaustive search is desirable.

The exhaustive search starts with the full permutation of the targeted peaks to generate a super-set of estimated spectra. It is clear that such a super-set will contain not only real pure component spectra but also non-real spectra which arise from the superposition (additivity) of other pure component spectra. In addition, a few heuristic rules are needed to disregard the non-real spectra and eventually to identify the real component spectra. In this section, in order to describe the exhaustive search, we will use the simulated data with tBTEM as discussed in the previous sections.

4.2.1 Full permutation of targeted peak pairs

Let *m* denote the number of channels in the spectroscopic data. Then a full permutation of any pair of peaks (channels) from the entire set of channel leads to *n* combinations. In other words, the number of targeting is equal to the full permutation of *m* and the same number of estimated spectra is obtained.

In most real physical problems, a full permutation over the set of *m* channels is not necessary. Indeed, many channels will have little or no real physical information. Accordingly, let *m'* denote the number of channels considered after the data (the first few $\boldsymbol{V}^{\boldsymbol{T}}$ vectors) are filtered. This filtering can be automated, thereby eliminating channels whose values are less than a critical threshold value, or manually whereby the user inspects the data set and identifies a sub-set of interest. The latter maybe of more efficient (since a smaller set *m'* is identified) but it may require experience / some expert knowledge. Then, the realistic number of runs *n'* can be expressed as equation 4.2.

$$n' = C_2^{m'} = m' \times (m'-1)/2 \tag{4.2}$$

Following the idea further, in order to develop the automated routine, the targeted peaks are chosen from the first $\boldsymbol{V^T}$ vector using the threshold method. The set of targeted peaks chosen in this way seems reasonable as the first $\boldsymbol{V^T}$ vector represents an average of total mixture spectra. A rather stringent threshold can be set if desired, thus maximizing the admissible channels to be searched. A useful threshold can be easily defined as the ratio of a peak to the maximum value in the first $\boldsymbol{V^T}$. For the simulated system considered here, if the threshold is set to 0.05 (case 1) then *m'* is equal to 33; while if the threshold is 0.1 (case 2) then *m'* becomes 26. As a result, the total numbers of the estimated spectra *n'* for cases 1 and 2 are 528 and 325 respectively.

In the following sub-sections, a few heuristic rules are developed to filter/eliminate the superposition spectra and duplicated spectra from the total super-sets of 528 or 325 estimated spectra and eventually extract all the real component spectra. Since a large number of peaks were used for targeting, the robustness of the developed algorithm was also examined.

4.2.2 Initial filtering / rejection of undesirable estimated spectra

A real spectrum can always be reconstructed by tBTEM using an exhaustive search. Also, the reconstruction is expected to possess a high degree of accuracy if the highest peak is targeted because the highest peak has larger signal to noise ratio. The rule therefore rejects those estimated spectra whose highest peak is not one of the 2 targeted peaks. The collection of the candidate spectra is thus reduced. For cases 1 and 2, the numbers of the candidates are reduced to 67 and 58 respectively.

4.2.3 Reject the duplicate spectra for a real component

For real pure component spectra having a number of fragments, an exhaustive search using tBTEM will always result in multiple estimates of a real spectrum. In order to detect and reject these duplicate estimates, the inner product of every two estimated spectra can be calculated after normalizing to unit vectors. Due to the presence of noise, a threshold of the inner product can be set ***e.g.*** a typical value may be 0.95. If the resulting inner product is larger than 0.95, one of the two spectra under investigation is regarded as the duplicate and has to be rejected from the collection of potential candidates. The estimated spectral with the smaller objective function is kept while the other is discarded, because the former indicates that the two higher peaks are used for targeting and a better signal/noise ratio resulted. In contrast, if the inner product is less than 0.95, both spectra are regarded as promising and thus kept for further examination. In doing so, the numbers of the candidate spectra in the simulated system decrease to 24 and 22 for cases 1 and 2 respectively.

4.2.4 Reject spectra that have linear relationship with others

As mentioned in section 3.4, when two peaks are used in targeting, there are many local minima at where spectra are corresponding to mixture spectra that are linearly combined by two pure spectra.

This rule rejects spectral estimates that are linear combinations of other spectra. Assume that a spectrum ***X*** is linearly combined by two other spectra ***Y*** and ***Z***, then the following relationship holds, where *a* and *b* are positive coefficients, and ***err*** represents the vector of errors which should approach zero. A full permutation of combinations for ***X*, *Y*** and ***Z*** are searched to reveal linear dependent combinations, within some pre-specified tolerance for the term ***err***.

$$\boldsymbol{X} = \alpha \times \boldsymbol{Y} + \beta \times \boldsymbol{Z} + \boldsymbol{err} \tag{4.3}$$

A common way to estimate the coefficients and error term is to perform least square regression. It is noted that, however, the least square procedure always generates results for the coefficients and errors, no matter whether there exists the linear relationship among the ***X***, ***Y*** and ***Z***. In other words, using least square alone may not guarantee the correct solution.

The physical meaning of the method is to check if a channel's value of ***X*** is almost solely contributed by ***Y***. Then an estimated coefficient *a* can easily be obtained as $a \approx \boldsymbol{X}_I / \boldsymbol{Y}_I$ at this channel. Therefore get the estimated value of *b* and the relationship among ***X***, ***Y*** and ***Z***. If ***X***, ***Y*** and ***Z*** are linearly dependent, then a least squares method would be used to find the combined spectrum.

4.2.4.1 Find a channel having the highest ratio of spectra *Y* and *Z*

Sequentially choose three spectra ***X***, ***Y*** and ***Z*** from the resulting collection of Section 4.2.3, and then compare the peak values of the two spectra ***Y*** and ***Z*** at each channel ***i*** and find a channel ***I*** where the maximum ratio is presented. Let $\boldsymbol{R}_I$ denotes the maximum ratio, we have

$$\boldsymbol{R_I} = \boldsymbol{max(}\left|\boldsymbol{Z_i / Y_i}\right|\boldsymbol{,}\left|\boldsymbol{Y_i / Z_i}\right|\boldsymbol{)} \qquad \boldsymbol{i} = 1,2,...,91 \tag{4.4}$$

The presence of noise may result in misleading highest ratio $\boldsymbol{R}_{\mathrm{I}}$ since many channels of mass spectra are zeros. Thus for all the estimated spectra that are normalized by their respective highest peak, the constraints of equations (4.5) and (4.6) are imposed. This indicates that at channel *I*, the maximum peak value of $\boldsymbol{Y_I}$ or $\boldsymbol{Z_I}$ has to be larger than 0.01. The ratio $\boldsymbol{R_I}$ is larger than 200.

$$\boldsymbol{max(Y_I, Z_I)} > 0.01 \tag{4.5}$$

$$\boldsymbol{R_I} > 200 \tag{4.6}$$

Assuming that the peak value $\boldsymbol{Y_I}$ is larger than $\boldsymbol{Z_I}$ at channel *I*, we proceed to the following steps to estimate the coefficients.

4.2.4.2 Determine the coefficients of the linear relationship

Rewriting equation (4.3) at any channel ***i***, we have the following relationship.

$$\boldsymbol{X_i} = \alpha \times \boldsymbol{Y_i} + \beta \times \boldsymbol{Z_i} + \boldsymbol{err_i} \tag{4.7}$$

Dividing equation (4.7) by $\boldsymbol{Y_I}$ gives equation (4.8).

$$\boldsymbol{X_i / Y_I} = \alpha + \beta \times \boldsymbol{Z_i / Y_I} + \boldsymbol{err_i / Y_I} \tag{4.8}$$

Recall that if $\boldsymbol{i} = \boldsymbol{I}$, $\boldsymbol{Y_I} > \boldsymbol{Z_I}$ then $\boldsymbol{Z_I/Y_I} < 0.005$ (from equation 4.6). Since the estimated spectra ***X***, ***Y*** and ***Z*** are all normalized by their highest peaks, the coefficients *a* and *b* would be less than 1. Therefore at channel *I*, the last two terms of equation (4.8) can be cancelled out. Consequently, the coefficient *a* can be estimated by equation (4.9) as below.

$$\boldsymbol{\alpha} \approx \boldsymbol{X_I / Y_I} \tag{4.9}$$

Substituting equation (4.9) into equation (4.3), we obtain equation (4.10).

$$\boldsymbol{X} = \boldsymbol{X_I} / \boldsymbol{Y_I} \times \boldsymbol{Y} + \boldsymbol{\beta} \times \boldsymbol{Z} + \boldsymbol{err} \tag{4.10}$$

Dividing equation (4.10) by $\boldsymbol{X_I}$ and rearranging the equation result in equation (4.11).

$$\boldsymbol{X} / \boldsymbol{X_I} - \boldsymbol{Y} / \boldsymbol{Y_I} = (\boldsymbol{\beta} / \boldsymbol{X_I}) \times \boldsymbol{Z} + \boldsymbol{err} / \boldsymbol{X_I} \tag{4.11}$$

Let vector $\boldsymbol{W}$ represent the left-hand side of equation (4.11),

$$\boldsymbol{W} = \boldsymbol{X} / \boldsymbol{X_I} - \boldsymbol{Y} / \boldsymbol{Y_I} \tag{4.12}$$

In order to determine whether there is a linear relationship between ***W*** and ***Z***, the inner product (IP) of the two unified vectors is evaluated in the same manner as described previously. In particular, a threshold is set as 0.98. If the inner product is larger than the threshold, the spectra ***X***, ***Y*** and ***Z*** are regarded as linearly related and the algorithm will proceed to the step described in Section 4.2.4.3. Otherwise, the three spectra are mutually independent and another set of ***X***, ***Y*** and ***Z*** are chosen and the steps starting from Section 4.3.4.1 are repeated. It is noted that in this process, we do not intend to get accurate value of a.

4.2.4.3 Determine the relationship between the linear related spectra

Once the spectra ***X***, ***Y*** and ***Z*** are found to be linearly related, the least squares algorithm can be used on equation (4.3) to determine the coefficients a and b. The relationship of the three spectra can be identified by checking the signs of the coefficients. For instance, if both a and b are positive, ***X*** is the linear combination of ***Y*** and ***Z.*** Similar, if a is positive and b is negative, ***Y*** is the linear combination of ***X*** and ***Z***, and so on. For the spectrum that is the linear combination of the other two, it will be removed from the collection of the candidate component spectra.

After checking the full permutation of estimated spectra ***X***, ***Y*** and ***Z*,** the set of remaining spectra is reduced in size, and these retained spectra are regarded as the pure component spectral estimates for the system. For the present simulated system, the final numbers of the remaining spectra for both case 1 (*n*' = 528) and case 2 (*n*' = 325) are reduced exactly to 10. These 10 resultant estimates do in fact correspond to the 10 reference component used in the simulation.

4.2.5 Computational Considerations

The spectral reconstruction algorithm tBTEM has to be executed hundreds of times during the exhaustive targeting. This represents the primary computational demand in the exhaustive search. Indeed, a typical run time for each run of tBTEM was on the order of a few minutes, and therefore, the few hundred executions required for case 1 and case 2 required about 10 hours. These calculations were performed on a 2GB RAM, Intel Xeon 500 MHz, and Windows NT 4.0 workstation. After the generation of the superset of spectral estimates, the heuristic rules are applied. This latter sorting and reduction of the set of spectral estimates requires little time compared to the numerous global searches required for the simulated annealing driven tBTEM.

4.3 Performance of fast multi-start simplex method

As we know, BTEM and tBTEM are algorithms to reconstruct pure spectra. They must use global optimization methods in reconstructing pure spectra. Although simulated annealing (SA) is successfully used in BTEM and tBTEM algorithms, its optimization speed is slow. In this section, SA is replaced by fast multi-start simplex method (FMSS) as optimization method in tBTEM. The FMSS results show that FMSS is much faster than SA in reconstructing pure mass spectra.

4.3.1 Simulation method for synthesized data

The simulated method is similar to that discussed in section 4.1.1 except that the number of mixtures increases to 60 and noise level is increased. The noise matrix e is randomly generated at a level of 0 to 10^3 instead of 0 to 10^2. The simulated method is shown in equation (4.13) and the 2-D array set of mixture is shown in Figure (4.7)

$$\boldsymbol{A}_{60´91} = \boldsymbol{C}_{60´10}\ ´\ \boldsymbol{P}_{10´91} + \varepsilon_{60´91} \qquad (4.13).$$

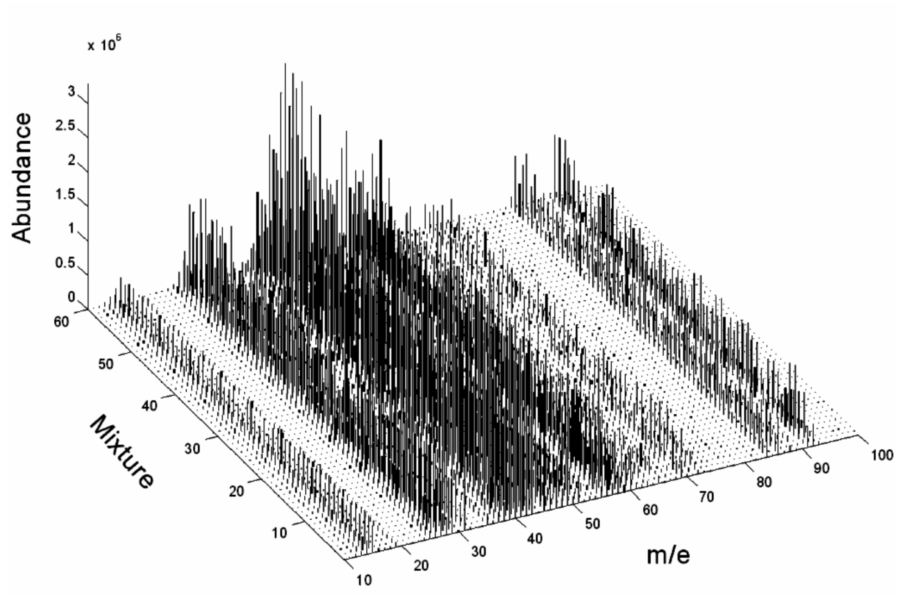

Figure 4.7: 2-D simulated sixty mixture mass spectra

4.3.2 Testing parameters

After several tests, the parameters of FMSS for reconstructing pure component mass spectra are fixed as:

(1) Points reduction coefficient $\boldsymbol{r_p}$ = 0.2

(2) Initial stopping criterion ε_0 = 0.1

(3) Stopping criterion reduction coefficient $\boldsymbol{r_\varepsilon}$ = 0.1

(4) The number of search rounds $\boldsymbol{R}$ = 3.

In FMSS method, the number of variables refers to the number of $\boldsymbol{V^T}$ vectors used in optimization; the stopping criterion refers to the stopping criterion of objective function value. The termination criteria for scalars of $\boldsymbol{T}$ is fixed at 10, which is a very coarse tolerance. Other parameters of Nelder-Mead simplex method are the default settings of software MATLAB 5.3 (the optimization function used is “fminsearch”). The scalars of initial testing $\boldsymbol{T}$ are randomly given with uniform distribution from -5 to 5. The number of starting points $\boldsymbol{n_0}$ is 30 unless otherwise mentioned.

The speed of FMSS method is compared with Corana’s SA in reconstructing pure component spectra in tBTEM algorithm. In Corana’s SA, r_T and N_T determine the simulated temperature’s (T) decrease speed and the number of evaluations preformed respectively. Larger value of r_T and N_T means larger number of evaluations needed ***i.e.*** longer optimization time. The parameters of SA are the same as those used by Widjaja and Garland (2002), that is, starting temperature T_0 = 10, step variation N_s = 20, temperature reduction coefficient r_T = 0.85, and N_T = max [(100, 5×N)/N_s], where N is the number of decision variables to be optimized.

The computational work was performed on a workstation with dual Xeon PIII 500 MHz CPU and 2048M ram by using commercial software MATLAB 5.3. Since MATLAB 5.3 does not support parallel method, therefore every time only one of the two CPUs is used.

4.3.3 Validity tests of FMSS

The validity of the FMSS optimization method is tested by using tBTEM with different number of optimization variables. When a certain number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors, j, is used in optimization, therefore the number of decision variables in the vector $\boldsymbol{T}$ is also j (see equation 3.15). In every test, when a certain number of variables and a certain number of starting points are chosen, 30 repeated reconstructions using the same parameters are performed in order to test the robustness of the algorithm. The inner product (IP) of two unit vectors is used to check the similarity between the reference spectrum and estimated spectra. In these tests, the reference spectrum is acetone and the estimated spectra are all targeted at the peaks m/z 43 and 58. Table 4.4 shows the IP values between the reference and the estimated spectra.

Table 4.4: Validity tests of different number of variables

Number of variables	IP results								
	30 starting points			25 starting points			20 starting points		
	Max	Min	Avg.	Max	Min	Avg.	Max	Min	Avg.
10	0.9970	0.9968	0.9969	0.9987	0.9791	0.9959	0.9981	0.9688	0.9943
35	0.9970	0.9960	0.9967	0.9976	0.9728	0.9960	0.9970	0.9464	0.9921
60	0.9960	0.9853	0.9931	0.9970	0.9621	0.9944	0.9981	0.9443	0.9889

The results show that the performance of the FMSS depends on the number of starting points. When 20 and 25 starting points are used, some estimated spectra are quit well; some of the estimated spectra are poor. It shows that FMSS is not reliable when only 20 or 25 initial points are used in one reconstruction. However, when 30 starting points are used, all the estimated spectra are quite well even when 60 variables are used in the optimization. Comparison of the worst estimated spectrum with the reference spectrum under 30 initial points is shown in Figure 4.8, which shows that even the worst one is acceptable.

Although it seems not safe that using only 30 starting points for an optimization problem with 60 variables, the results in Table 4.5 indicate that 30 initial points are enough. The reason is probably due to the nature of simplex method. Although simplex is not a global optimization method, it is able to follow the gross behavior of the test functions despite many local minima, and is not easy to be trapped in shallow local minima.

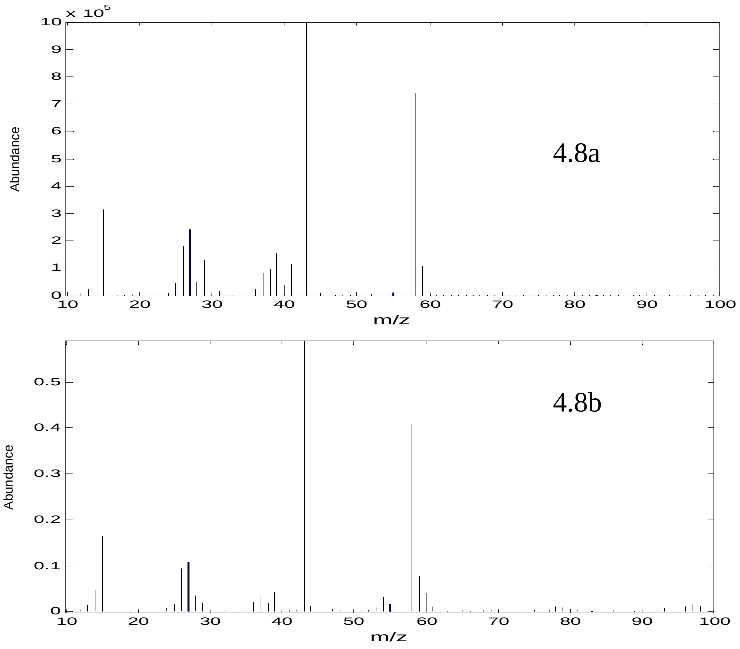

Figure 4.8: The worst estimate (4.8b) and the reference spectrum (4.8a)

4.3.4 Computational efficiency

SA is a popularly used global optimization method; it has been successfully used in BTEM and tBTEM. In this section, optimization speed of FMSS is compared with SA and multi-start simplex (MSS) methods. Several tests under different number of variables are carried out to compare the optimization speeds of three methods. Table 4.5 shows the computational times of different optimization methods under different number of variables. Every method is tested 30 times under a certain number of variables. The stopping criteria of the objective function values for MSS and SA methods are set to 10^{-3}. The stopping criterion of FMSS is converged to 10^{-3} step by step as in section 4.3.2.

Table 4.5: The computational time of FMSS, MSS and SA

Number of variables	**Computation Time (s)**								
	Fast Multi-start Simplex Method			**Multi-start Simplex Method**			**Simulated Annealing (SA)**		
	Min	Max	Avg.	Min	Max	Avg.	Min	Max	Avg.
10	14.8	56.9	34.3	110.8	130.5	117.3	98.2	151.2	109.7
35	59.8	319.9	201.8	927.0	1009.2	962.3	445.6	556.5	460.4
60	132.0	865.6	460.4	2815.3	2929.2	2882.6	937.2	1072.0	955.7

Among these three optimization methods, the variation of FMSS optimization speed is the biggest. The other two methods are relatively constant. The average optimization time of FMSS is the smallest among these three methods at j = 10, 35 and 60. Comparing the longest time of FMSS to the shortest time of MSS and SA at different j, even the slowest speeds of FMSS are always faster than the fastest speeds of MSS and SA. The average optimization speed ratios of FMSS to SA at j = 10, 35 and 60 are 31.27%, 43.83% and 48.17% respectively.

It is demonstrated that 30 initial points for FMSS method are enough for reconstruct pure mass spectra; the optimization speeds of FMSS at different number of initial points are also studied. These tests would help us to understand at what number of initial points where FMSS would be slower than SA. Table 4.6 shows the results of average optimization time of SA and FMSS under different kinds of parameters. SA is tested by 7 different numbers of variables (*j* from 10 to 60). Besides using 7 different numbers of variables, FMSS also is tested under different number of starting points (30, 50 and 70). Every result in Table 4.6 is a mean of 30 repeated tests.

Table 4.6: Comparison of optimization time of FMSS and SA at different kinds of parameters

Number of starting Points	**Optimization time at different number of variables (s)**						
	10	**18**	**26**	**35**	**43**	**51**	**60**
30	34.3	74.4	147.3	201.8	316.8	410. 8	460.4
50	62.5	130.2	257.5	398.3	472.8	629.5	801.2
70	77.8	185.2	344.3	545.7	655.5	831.2	1070.2
SA results	109.7	203.1	334.5	460.4	628.3	793.5	955.7

The variation of optimization times of FMSS and SA under different numbers of parameters are shown in Figure 4.9. The line 2 is the optimization times of SA; the lines 1, 3 and 4 are those of FMSS. The trends of these 4 lines indicate that in our synthesized system: 1) when 30 and 50 initial points used in optimization, FMSS would always faster than SA at different number of variables. 2) When 70 initial points used in optimization, the relative optimization speed of FMSS would slow down compared with SA when the number of variables increases. It seems that around 60 points, the speed of FMSS would be equal to the speed of SA.

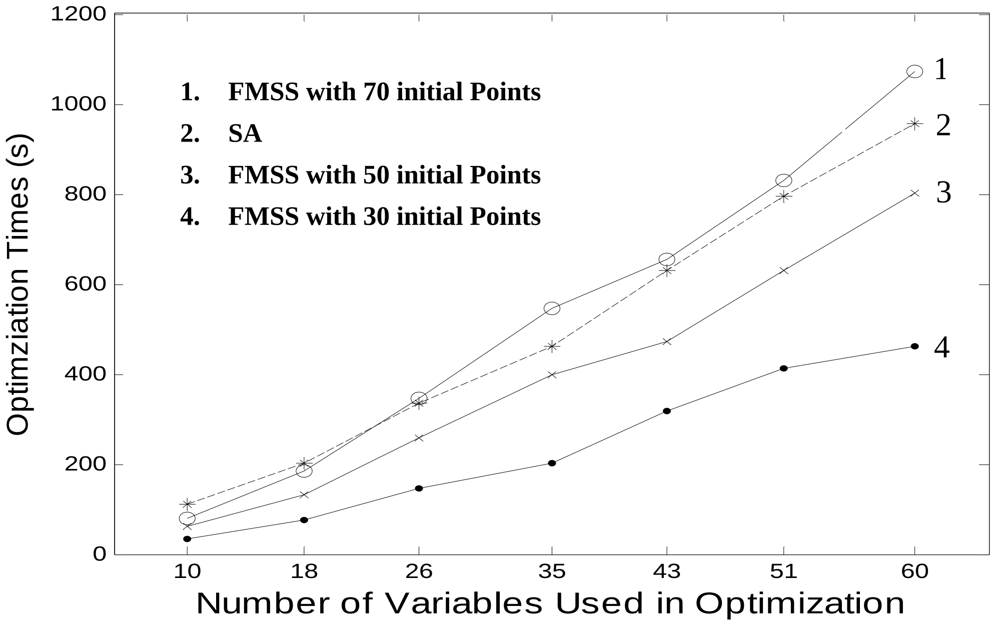

Figure 4.9: FMSS *vs.* SA at different number of variables and starting points

4.4 Summary

In this chapter, tBTEM and FMSS methods are tested on synthesized mixture. The advantages of modifications of tBTEM are explained by examples.

By reformulating the objective function with the peaks heights instead of their derivatives, the algorithm is able to reconstruct pure component spectra from mixture mass spectra. Furthermore, the algorithm is more computationally efficient as fewer mathematical operations are needed for the evaluation of the objective function.

Weighting of the abstract $\boldsymbol{V}^T$ vectors reduces the adverse effect of noise and the sensitivity of choosing the number of $\boldsymbol{V}^T$ vectors. As a result, the algorithm is more robust. Weighting also allows a larger number of $\boldsymbol{V}^T$ vectors to be used thereby increasing the amount of recoverable information even with the presence of more noise.

A significant improvement in the reconstruction of highly overlapped spectra has been achieved by using the tBTEM algorithm where two peaks are used in targeting. In principle, the salient idea in tBTEM might be extended to deal with really highly overlapped spectra by considering multiple (>2) targeted peaks.

Since the tBTEM algorithm is based on the idea that each component spectrum is reconstructed one-at-a-time by choosing of targeted peaks, an exhaustive search method provides a strategy for generating all possible pure component spectra. The searches execute tBTEM many times and then reduce the super-set of estimates to the sub-set of only real component spectra.

A global optimization method named fast multi-start simplex method (FMSS) is developed. It dramatically reduces the optimization time compared with multi-start simplex method and is much faster than SA. FMSS method optimizes its speed by not wasting time in searching useless points. Compared with SA, which is difficult to be paralleled, FMSS has advantage for its parallel nature. FMSS method is a totally parallel method; it can be changed to any number of parallel jobs without changing the algorithm. Following the idea of FMSS, many local searching methods can be used to find global optimums with faster speeds. Furthermore, FMSS can also find many local minima other than global minimum. It would be possible to use FMSS to find many if not all components by one targeting using BTEM or tBTEM.

Chapter 5: tBTEM on Real System

In the chapter 4, tBTEM was successfully applied to synthesized discrete spectra. In this chapter, tBTEM is applied on a real system by using FMSS optimization method.

5.1 Challenge of real mass spectra

Commonly, when one uses a GC-MS system to find pure component spectra in mixture, first, he/she will try to separate mixture into pure components by using suitable GC columns and parameters (mobile phase, flow speed and column temperature *etc*.). The separated mixture would go into MSD to get pure component spectra. Although it is possible to separate a mixture into all pure components before MS detection, much effort should be dedicated to find suitable columns and to adjust parameters. If a mixture has many components such as food samples and Chinese traditional medicines, it would spend much time and much money in separating these kinds of mixtures into pure components. As it is shown in the previous chapter, tBTEM is very powerful in dealing with a MS mixture with 10 components on synthesized data set. If tBTEM can be used in real system, it would have many advantages.

Problems encountered when applying tBTEM to real systems due to the nature of mass spectra. For a pure component mass spectrum, there is no fixed pattern for it. Mass spectra of the same component would have different patterns from different machines and from different retention time even in one injection. For example, the two highest peaks of acetone in Figure 4.1 would change their order or ratio at different spectra. We call this phenomenon as non-stationary effect. As mentioned before, the basic rule applying BTEM or tBTEM to a system is that the mixture spectra should be a linear mixture of different pure component spectra. Recall the equation (2.1), ideally, every pure component spectrum in matrix $\boldsymbol{\alpha}_{k\times s}$ should have fixed pattern, and there should has no error in system. In terms of equation (2.1), the non-stationary aspects would make the mixture spectra not a linear combination of different pure spectra, ***i.e.*** non-stationary effect would cause the system have big “non-linearity” in terms of our algorithms. We know that MS has bigger signal to noise ratio in terms of a single mass spectrum. When a set of mixture spectra is under investigation, because of the nature of our algorithms, the non-stationary effect would make MS has “small S/N” especially the small peaks in MS.

Other than the non-stationary effect of pure spectra, there is another major difference between synthetic mixtures discussed in chapter 4 and real mixture mass spectra analyzed in this section. The most significant difference is that fragments from the various components injected as a mixture will undergo a host of complex recombination and new signals will appear. The resulting spectral reconstructions from the major components may possess some contributions from these new signals, and these may even exist in channels where zero intensity is expected. This can lead to considerable complications when further identification is required.

Although BTEM is known to be able to deal rather well with non-stationary signals from FTIR and RAMAN data, such non-linearity can be tolerated because of the relatively good S/N ratio typical of these spectroscopes. Therefore, in MS data where S/N ratio is typically lower, non-stationary pure component spectra may be more difficult to recover.

5.2 Experimental setup and data collection

A GC-MS spectrometer (GC: Hewlett-Packard 6890, MS: Hewlett-Packard 5973) with a 5 meter long 100% methyl stationary phase capillary column were used together with a helium mobile phase and a 1μL syringe. It should be noted that this column has low separation ability and therefore is usually used for retention gaps and guard columns in GC but not for separation purposes. The use of a 1μL syringe injection into the GC-MS system lowers the level of fragment recombination because of the lower overall amount of solutes/solvent. Each mixture sample was injected slowly and smoothly, resulting in a somewhat flat total-ion-count (TIC) peak at the detector. Consequently, little separation of components occurred.

Five samples were prepared from four low molecular weight compounds, i.e., ethanol, hexane, toluene and acetone. Every sample is randomly mixed by 4 compounds. Each 1μL mixture sample was injected over a period of a few seconds (around 5 seconds). Some variation in the composition of the components occurred at the detector during the detection period. The range of the exported data was from m/z 10 to m/z 100 with the interval of 1 m/z. As a result, a series of circa 400 MS spectra data were acquired from the five sample injections. Only a small number of these spectra were used in the analysis. The manually selected mass spectra were collected in the more-or-less flat regions of the TIC peaks. This yielded 16 mixture mass spectra (as shown in Fig. 5.1) used in the subsequent reconstructions. It is noted that there is no mixture containing only one component.

The FMSS parameters are the same as the parameters used in section 4.4.1 with 30 initial points.

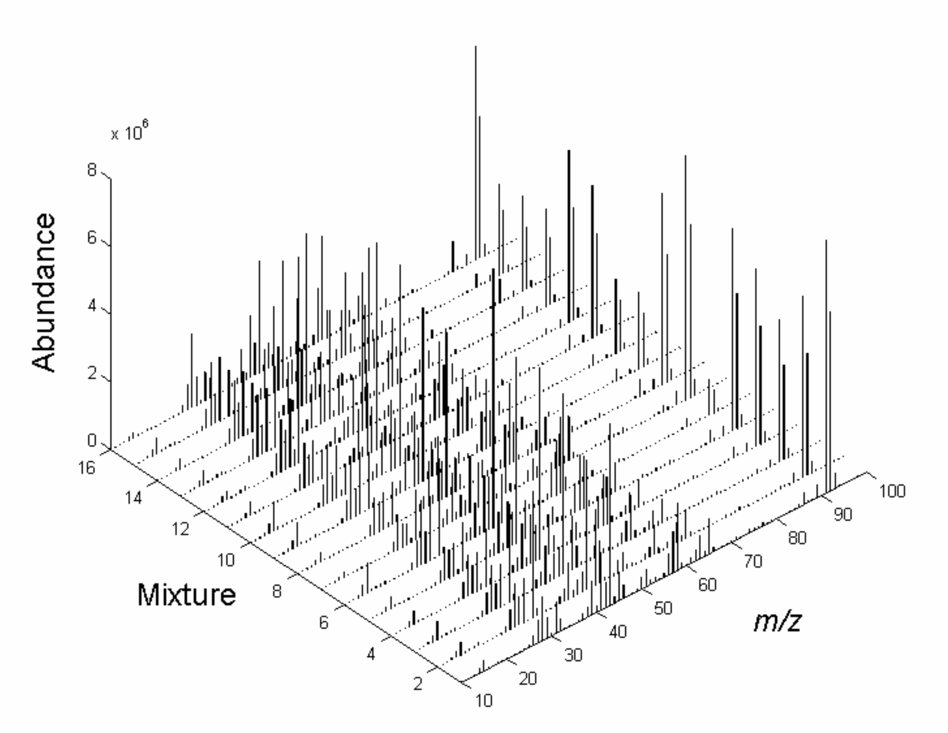

Figure 5.1: 2-D sixteen mass spectra of the real mixtures

5.3 Results of spectral reconstruction

In the spectral reconstructions performed with tBTEM, all 16 $\boldsymbol{V}^{\boldsymbol{T}}$ vectors were weighed by the singular value matrix, *i.e.* the diagonal matrix $\boldsymbol{S}$ from SVD. As shown in Table 5.1, each pure component spectrum was reconstructed by targeting two specified peaks. The four pure component spectra were successfully recovered as shown in Figure 5.2.

Table 5.1: Peaks targeted for reconstructing pure component spectra in the real system

Component	**ethanol**	**acetone**	**hexane**	**toluene**
Targeted peaks (m/z)	31 and 45	43 and 58	57 and 86	91 and 92

The characteristic fragmentation patterns for the ethanol, acetone, hexane and toluene are readily apparent in the spectral reconstructions – they are rather good. Closer comparison of these reconstructions with the "references" provided in Figure (5.2) indicates an interesting artifact. The reconstructions are in some ways "cleaner" *i.e.* the small peaks of spectra are not reconstructed well. Although the primary fragments are still very prominent, the intensities of some of the other fragments are reduced. This is the case for all four component spectral reconstructions. This reduction in the intensity of the other channels is probably related to the non-stationary quality of the mass spectra, since the high non-linearity of real mass spectra set, the small peaks would have poor signal to noise ratio in terms of tBTEM algorithm *i.e.* the small peaks are all merged by high level of noise which results in "smooth" reconstruction.

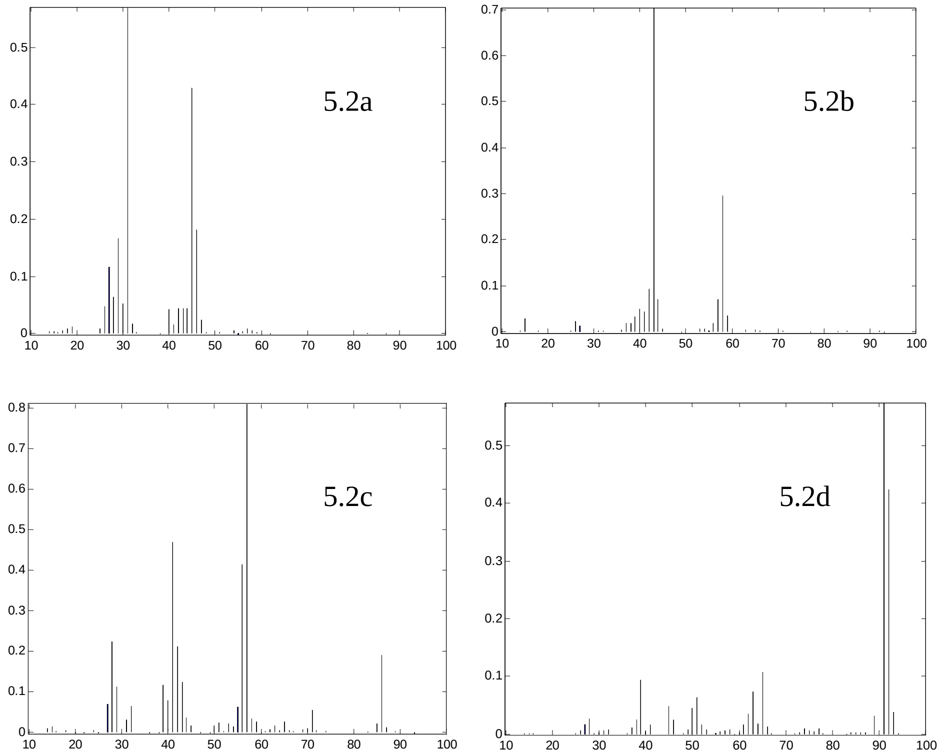

Figure 5.2: Reconstructed spectra from real mixture spectra

5.2a: ethanol; 5.2b: acetone; 5.2c: hexane. 5.2d: toluene.

Again, it is important to emphasize the utility of using weighted $\boldsymbol{V}^{\boldsymbol{T}}$ vectors. Ideally, if the experimentalist had prior knowledge of the system, the number of the $\boldsymbol{V}^{\boldsymbol{T}}$ vectors used would be equal to the number of pure components in the system in order to get the best results. However, the experimentalist probably wants to explore an unknown system, and therefore information of the number is unknown *a priori*. Another reason is that the information in every $\boldsymbol{V}^{\boldsymbol{T}}$ vector reduces smoothly, which makes difficult to decide how many $\boldsymbol{V}^{\boldsymbol{T}}$ vectors should be used in optimization. Consequently, more $\boldsymbol{V}^{\boldsymbol{T}}$ vectors are used than really needed, but much noise is introduced at the same time. Figures 5.3a and 5.3b are the reconstructed ethanol spectra using un-weighted four and six $\boldsymbol{V}^{\boldsymbol{T}}$ vectors respectively. As can be seen from these two figures, spectral features belonging to toluene at m/z 91 and 92 appear. This is more prominent in the six $\boldsymbol{V}^{\boldsymbol{T}}$

case (Figure 5.3b) than the four $\boldsymbol{V^T}$ case (Figure 5.3a). Clearly, the weighted 16 $\boldsymbol{V^T}$ vector solution presented in Figure 5.2 is much better.

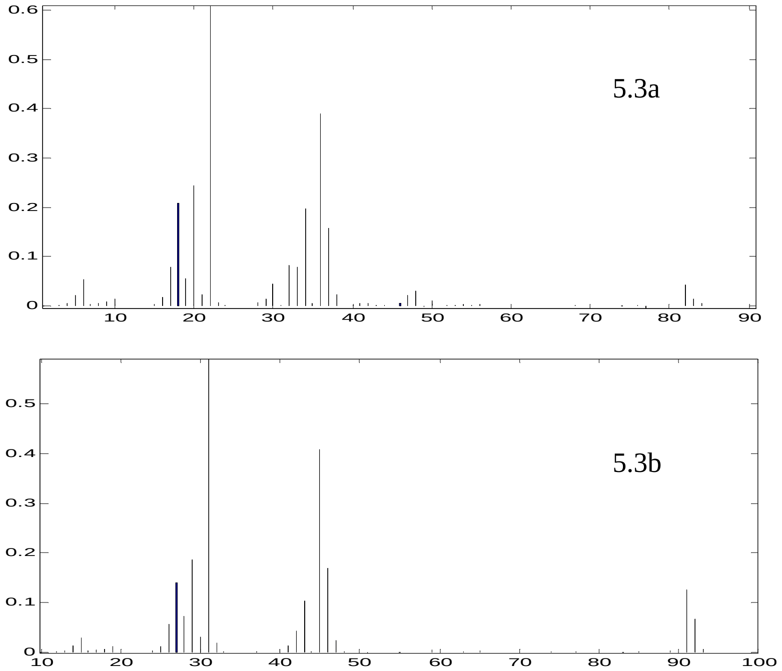

Figure 5.3: Results of estimated ethanol spectra with different number of $\mathbf{V^T}$ vectors

5.3a: using 4 un-weighted $\boldsymbol{V^T}$ vectors; 5.3b: using 6 un-weighted $\boldsymbol{V^T}$ vectors.

5.4 Conclusion and discussion

For real systems, difficulties can be encountered and these are related to the presence of non-stationary signals, and the presence of fragmentation recombination in electron-impact ionization which would occur in high pressure MSD. With the present experimental configuration, two difficulties arose: namely a slight order simplification of the pure component spectra after tBTEM, and difficulties related to the exhaustive search. The difficulties observed here do not negate the possibility that analysis from other types of experimental configurations / methods may be easier. Good examples may include (1) ultra-high vacuum studies of desorbed species in surface science studies, (2) chemical vapor deposition at low system pressures, (3) time-of-flight data from reactive systems, and (4) electrospray ionization Mass Spectrometry (ESI MS). In the last case, the tBTEM may work rather well as fewer fragments are present and the spectral patterns are perhaps more stationary. A successful application to ESI MS may greatly facilitate spectral reconstruction in the fields of biochemistry and organometallics where complex and poorly separable mixture are common.

One may argue that since a good separation can be achieved if GC is used, the use of the GC column alone may not be necessary. As mentioned in section 5.2, a poor-performance GC column was used to induce a series of different experimental measurements from 1 sample for testing tBTEM. This idea might actually be of some utility when sample size is very small / limited. One does not have to find a really good column for the separation. Accordingly, one could use tBTEM plus a poor-performance GC column before MS to obtain pure component information. It would be possible to get good reconstructions from using any “poor” separation of sample on any arbitrarily “poor” column.

In summary, the performance of the tBTEM algorithm was examined on experimental mixture spectra. The recovered spectral estimates are quite acceptable. These tests suggest that tBTEM has considerable potential for many real mass spectroscopy applications.

Chapter 6: Conclusions and Future Work

In this thesis, studies mainly focus on mathematical aspects: 1) modify band-target entropy minimization method (BTEM) to apply to discrete data such as mass spectra. 2) Propose a global optimization method to speed up optimization speed.

A special objective function formula was developed to reconstruct pure mass spectra from mixture data set. Compared with original BTEM method, this new objective function of tBTEM has no derivative and is suitable for discrete data. Moreover it also has advantage in optimization speed since fewer evaluations are needed.

It is found that the noise in system will affect the performance of BTEM. If the number of $\boldsymbol{V}^{\boldsymbol{T}}$ vectors used in optimization is much larger than the number of species in mixture, the result would be bad even for a system with less noise. Accordingly, a method that uses weighted $\boldsymbol{V}^{\boldsymbol{T}}$ vectors is proposed to lower the effect of noise.

In real MSD method, for different pure components, there are many charged fragments which have same m/z values. Therefore, it would have many overlapped peaks in MS mixture. Furthermore, the non-stationary effect in MS is much bigger that in IR and RAMAN, therefore, targeting at the highest peaks which have the biggest S/N ratio is meaningful in MS systems. When more than 2 spectra whose highest peaks locate at the same place, It would have problem for BTEM to find all of them by using their highest peaks. Hence, a two-peak targeting method is used to reconstruct all pure spectra by using their highest peaks. By targeting at different pairs of peaks, the objective function values would be rearranged, therefore all spectra which overlapped at their highest peak could be found. This method also has advantage on dealing with noise due to the use of more spectra information.

By targeting at different pairs of peaks from a full permutation of selected peaks, tBTEM also enables to target all pure spectra out from mixture automatically. Following this idea, an exhaustive searching method successfully extracts all pure spectra from a huge collection of estimated spectra.

A global optimization method, simulated annealing, is used in BTEM. When dealing with a data set which contains more than 2500 channels, the optimization time is about 10 hours. When targeting at a series of peaks using SA, the overall optimization time of BTEM would be very long. Therefore, a new global optimization method named fast multi-start simplex method (FMSS) is developed to accelerate the searching speed. Although the original simplex is a local optimization method, it is extended as a (pseudo) global optimization method by using multiple starting points. The original multi-start simplex can find a global minimum, but its optimization speed is very slow due to a large number of unpromising points in evaluation. FMSS method discards lots of unpromising points in advance at different steps by different stop criteria. It dramatically reduces the optimization time. Compared with popular simulated annealing method, FMSS is faster than SA more than 50%.

With all these new mathematical methods, a real MS data is tested. Since the non-stationary effect is the nature of MS and its magnitude is high. Some experimental strategies were employed in dealing with non-stationary effect. First, a small volume syringe (1μL) was used in injection over long period of time which lowers the flow rate of mixture and results in less overall concentration in MS detection. Second, a smaller non-stationary effect is found in flat area. Therefore, spectra were picked from the flatter area. Based on these experimental strategies, a four components mixture was successfully studied and all four pure component spectra were reconstructed.

The present work focuses on mathematical aspects and does limited application to real system. Future works may apply tBTEM to real systems. On the other hand, although discrete spectra cannot be treated as continuous spectra, continuous spectra could be treated as discrete spectra. To apply tBTEM to continuous data to take advantage of its fast optimization speed would be one meaningful application.

References

Bijlsma, S. *et al*. Rapid Estimation of Rate Constants Using On-line SW-NIR and Trilinear Models, *Anal. Chim. Acta,* 376, pp. 339-355. 1998.

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

Brereton, R.G. Chemometrics in Analytical Chemistry: A Review, *Anal. Chem.*, 112, pp. 1635-1657. 1987.

Brown S.D. *et al*. Chemometrics, *Anal. Chem.*, 60, pp. 252R-273R. 1988.

Brown S. D. Chemometrics, *Anal. Chem.*, 62, pp. 84R-101R. 1990

Brown, S. D. *et al*. Chemometrics, *Anal. Chem.,* 64, pp. 22R-49R. 1992.

Brown, S. D. *et al*. Chemometrics *Anal. Chem.*, 66, pp. 315R-359R. 1994.

Brown, S. D. *et al*. Chemometrics, *Anal. Chem.*, 68, pp. 21R-61R. 1996.

Delaney, M. F. Chemometrics, *Anal. Chem.*, 56, pp. 261R-277R. 1984.

Carey R. N. *et al*. Principal Component Analysis: Alternative to Referee Methods in Method Comparison Studies, *Anal. Chem.*, 47, pp. 1824-1829. 1975

Chen, H. *et al*. Parallel Genetic Simulated Annealing: A massively Parallel SIMD Algorithm, *IEEE T. Parall. Distr.*, 9, pp. 126–136. 1998.

Chen, J. H. and Huang L. P. Reconstruction of Mass Spectra of Components of Unknown Mixtures Based on Factor Analysis, *Anal. Chim. Acta,* 133, pp. 271-281. 1981.

Chen, L. and Garland, M. Use of Entropy Minimization for the Preconditioning of Large SD Spectroscopic Data Arrays: Application to *in situ* FT-IR Studies from the Unmodified Homogeneous Rhodium Catalyzed Hydroformylation Reaction, *Appl. Spectrosc.*, 56, pp. 1422-1428. 2002.

Chen, L. *et al*. An Efficient Algorithm for Automatic Phase Correction of NMR Spectra Based on Entropy Minimization, *J. Maga. Reson.*, 158, pp. 164-168. 2002.

Chew, W. *et al*. 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, pp. 1982-1990. 2002.

Chu, K. W. *et al*. Parallel Simulated Annealing by Mixing of States, *J. Comput. Phys.*, 148, pp. 646-662. 1999.

Corana, A. *et al.* Minimizing Multimodal Functions of Continuous Variables with the "Simulated Annealing" Algorithm, *ACM Transactions on mathematical software*, 13, pp. 262-282. 1987.

Dixon, L. C. W. and James, L. On Stochastic Variable Metric Methods. In Analysis and Optimization of Stochastic Systems. ed. by Jacobs Q. L. R. *et al*. London: Academic Press. 1980.

Edgar, T. F. *et al*. Optimization of Chemical Processes (2nd edition). pp. 190-210. Singapore: McGraw-Hill Int. Ed. 2001.

Frank, I. E. and Kowalski, B. R. Chemometrics. *Anal. Chem.,* 54, pp. 232R-243R. 1982.

Feng, G. and Liang, Y. Z. A novel Approach to the Retrieval of the Mass Spectrum of a Mixture, *Analytical Sciences,* 16, pp. 603-607. 2000.

Furusjoe, E. *et al.* Evaluation Techniques for Two-way Data from *in-situ* Fourier Transform Mid-infrared Reaction Monitoring in Aqueous Solution. *Anal. Chem.*, 70, pp. 1726-1734. 1998.

Garland, M. *et al.* On the Number of Observable Species, Observable Reactions and Observable Fluxes in Chemometric Studies and the Role of Multichannel Integration, *Anal. Chim. Acta*, 350, pp. 337-358. 1997.

Golub, G. H. and Van Loan, C. F. Matrix Computations. pp. 70, Baltimore, MD: The Johns Hopkins University Press. 1996.

Gong, F. *et al*. Determination of Volatile Components in Peptic Powder by Gas Chromatography-mass Spectrometry and Chemometric Resolution, *J. Chromatogr. A.*, 909, pp. 237-247. 2001a.

Gong, F. *et al*. Gas Chromatography-mass Spectrometry and Chemometric Resolution Applied to the Determination of Essential Oils in Cortex Cinnamomi, *J. Chromatog. A*, 905, pp. 193-205. 2001b.

Huang, M. *et al*. Multi-start Downhill Simplex Method for Spatio-temporal Source Localization in Magnetoencephalography, *Enectroencephalography and clinical neurophysiology*, 108, pp. 32-44. 1998.

Kanpur, J. N. Maximum Entropy Models in Science and Engineering. pp. 3, New Delhi: Wiley Eastern Ltd. 1993.

Kearfott, D. S. Rigorous Global Search: Continuous Problem. Norway, MA: Kluwer Academic Publishers. 1996.

Kirkpatrick, S. *et al*. Optimization by Simulated Annealing, *Science*, 220, pp, 671-680. 1983.

Kowalski, B.R. Chemometrics, *Anal. Chem.*, 52, pp. 112R-122R. 1980.

Lavine, B. K. Chemometrics, *Anal. Chem.*, 70, pp. 209R-228R. 1998.

Lavine, B. K. Chemometrics, *Anal. Chem.*, 72, pp. 91R-97R. 2000.

Lavine, B. K. and Workman J. Chemometrics, *Anal. Chem.,* 74, pp. 2763-2769. 2002.

Lawson, C. L. and Hamson R. J. Solving Least Square Problems. Englewood Cliffs, NJ: Prentice-Hall. 1974.

Lawton, W. H. and Sylvestre, E. A. Self Modeling Curve Resolution, *Technometrics*, 13, pp. 617-633. 1971.

Li, C. *et al*. Rhodium Tetracarbonyl Hydride: the Elusive Metal Carbonyl ydride, *Angew. Chem. Int. Ed.*, 41, pp. 3785-3789. 2002.

Li, C. *et al*. 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 the Homogenous Catalysis, *J. Am. Chem. Soc.*, 125, pp. 5540-5548. 2003a.

Li, C. *et al*. Spectral Reconstruction of In-situ FTIR Spectroscopic Reaction Data Using Band-target Entropy Minimization (BTEM): Application to the Homogeneous Rhodium Catalyzed Hydroformylation of 3,3-Dimethylbut-1-ene Using $Rh_4(CO)_{12}$, *J. Catalysis*, 213, pp. 126-134. 2003b.

Locatelli, M. and Schoen, F. Random Linkage: A Family of Acceptance /rejection Algorithms for Global optimization. *Math Prog.*, 85, pp. 379-396. 1999.

Malinowski E. R. Statistical F-tests for abstract factor analysis and target testing, *J. Chemometrics*, 3, pp. 49-60. 1988.

Malinowski, E. R. Factor Analysis in Chemistry. pp. 62, New York, NY: Wiley. 1991.

Malinowski, E. R. Abstract Factor Analysis of Data with Multiple Sources of Error and a Modified Faber-Kowalski F-test, *J. Chemometrics*, 13, pp. 68-81. 1999.

Nelder, J. A. and Mead, R. A simplex Method for Function Minimization, *Comput. J.*, 7, pp. 308-313. 1965.

Onbasoglu, E and Ozdamar, L. Parallel Simulated Annealing Algorithms in Global Optimization, *J. Global Optim.*, 19, pp. 27-50. 2001.

Ong, L.R. Extension of Pure Component Spectra Reconstruction in Exploratory Chemometrics to Solid Samples for XPS and FT-Raman, B.Eng. Thesis, National University of Singapore. 2001.

Pan, Y. *et al*. Pure Component Reconstructions Using Entropy Minimizations and Variance-weighted Piecewise-continuous Spectral Regions: Application to the Unstable Experimental System $Co_2(CO)_8/Co_4(CO)_{12}$, *J. Chemometrics*, 14, pp. 63-77. 2000.

Phalp, J. M. *et al*. The Resolution of Mixtures Using Data From Automated Probe Mass Spectrometry, *Anal. Chim. Acta*, 318, pp. 43-53. 1995.

Price, W. L. A Controlled Random Search Procedure for Global Optimization. In Towards Global Optimization 2, ed by Dixon, L. C. W. and Szegö, G. P., pp. 71-84. Amsterdam: North-Holland Pub. Co. 1978.

Rinnooy, Kan, A. H. G and Timmer, G. T. Stochastic Global Optimization Methods, Part 2; Multi Level Method, *Math Prog*, 39, pp. 57-78. 1987.

Ritter, G. L. *et al*. Factor analysis of the mass spectra of mixtures, *Anal. Chem.,* 48, pp. 591-595. 1976.

Sasaki, K. *et al*. Constrained Nonlinear Method for Estimating Component Spectra from Multicomponent Mixtures, *Appl. Optics.,* 22, pp. 3599-3603. 1983.

Scheick, J. T. Linear Algebra with Applications (International Edition). pp. 373-377, Singapore: McGraw-Hill. 1997.

Sharaf, M. A. and Kowalski, B. R. Quantitative Resolution of Fused Chromatographic Peaks in Gas Chromatography/Mass Spectrometry, *Anal. Chem.*, 54, pp. 1291-1296. 1982.

Shrager, R. I. Optical spectra from chemical titration: and analysis by SVD, *SIAM Journal on algebraic and Discrete methods*, 5, pp. 351-358. 1984.

Shrager, R. I. Chemical transitions measured by spectra and resolved using singular value decomposition, *Chemometrics Intell. Lab. Syst.* 1, pp. 59-70. 1986.

Sin, S. Y. Application of FT-Raman Spectroscopy Measurements. B. Eng. Thesis, National University of Singapore. 2002.

Smit, H.C. Specification and Estimation of Noisy Analytical Signals. Part I: Characterization, Time Invariant Filtering and Signal Approximation. In Chemometrics Tutorials II, ed by Brereton R.G. *et al.*, pp. 39-51. Amsterdam-London-New York-Tokyo: Elsevier. 1992.

Smit, H.C. Specification and Estimation of Noisy Analytical Signals. Part II: Curve Fitting, Optimum Filtering and Uncertainty Determination. In Chemometrics Tutorials II, ed by Brereton R. G. *et al*. pp. 53-65. Amsterdam-London-New York-Tokyo: Elsevier. 1992.

Spendley, W. *et al*. Sequential Application of Simplex Designs in Optimization and Evolutionary Operations, *Technometrics 4*, pp. 441-461. 1962

Tong, C. S. and Cheng, K. C. Mass Spectral Search Method Using the Neural Networks Approach, *Chemometrics and Intelligent Laboratory systems*, 49, pp. 135-150. 1999.

Törn, A. A search-clustering Approach to Global Optimization. In Towards Global Optimization 2, ed by Dixon, L. C. W. and Szegö, G. P., pp. 49-62. Amsterdam: North-Holland Pub. Co. 1978.

Wan, *et al*. Comparing Similar spectra: From Similarity Index to Spectral Contrast Angle, *J. Am. Soc. Mass. Spectrom.,* 13, pp. 85-88. 2002.

Watanabe, S. Pattern Recognition as a Quest for Minimum Entropy, *Pattern Recog*., 13, pp. 381-387. 1981.

Widjaja, E. and Garland, M. 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.

Windig, W. *et al*. Combined Use of Conventional and Second-Derivative Data in the SIMPLISMA Self-modeling Mixture Analysis Approach, *Anal. Chem.*, 74, pp. 1371-1379. 2002.

Widjaja, E. *et al*. Semi-Batch Homogeneous Catalytic In-Situ Spectroscopic Data. FTIR Spectral Reconstructions Using Band-Target Entropy Minimization (BTEM) without Spectral Preconditioning, *Organometallics*, 21, pp. 1991-1997. 2002.

Windig, W. A Simple-to-use Interactive Self-modelling Mixture Analysis, *Comput.-Enhanced Anal. Spectros*., 3, pp. 95-126. 1992.

Winter, D. J. Matrix Algebra. ed. by Robert W. Pirtle. New York: Macmillan Publishing Co., 1992.

Zeng, Y. and Garland, M. An Improved Algorithm for Estimating Pure Component Spectra in Exploratory Chemometric Studies Based on Entropy Minimization, *Anal. Chim. Acta,* 359, pp. 303-310. 1998.

List of publications

1. Zhang Huajun *et al*. Synthesis of calix[4]arene derivatives and studies on their UV spectrum, *Journal of Zhejiang University (Science Edition)*, 27, pp. 279-282. 2000.
2. Zhang Huajun *et al*. Synthesis of calix[4]arene derivatives and studies on their extracting power, *Journal of Zhejiang University (Science Edition)*, 28, pp. 64-66. 2001.
3. Zhang Huajun *et al*. On the Development of Weighted Two-band Target Entropy Minimization for the Reconstruction of Pure Component Mass Spectra. in Recent Advances in Computational Science & engineering. ed. by pp. 49-53. London: Imperial College Press. 2002.
4. Zhang Huajun *et al*. Pure component mass spectrum reconstruction with two-band target entropy minimization: application to mass spectral data. International Conference of ACS, USA, 2003.
5. Zhang Huajun *et al*. Weighted Two-Band Target Entropy Minimization for the Reconstruction of Pure Component Mass Spectra: Simulation Studies and the Application in Real Systems, *J of the American Society of Mass Spectrometry*, 14, pp. 1295-1305. 2003.