Spatial Mass Spectral Data Analysis Using Factor and Correlation Models

Lingli Shen

Thesis submitted for the Degree of Doctor of Philosophy

Department of Automatic Control and Systems Engineering

The University of Sheffield

March 2015

Abstract

ToF-SIMS is a powerful and information rich tool with high resolution and sensitivity compared to conventional mass spectrometers. Recently, its application has been extended to metabolic profiling analysis. However, there are only a few algorithms currently available to handle such output data from metabolite samples. Therefore some novel and innovative algorithms are undoubtedly in need to provide new insights into the application of ToF-SIMS for metabolic profiling analysis. In this thesis, we develop novel multivariate analysis techniques that can be used in processing ToF-SIMS data extracted from metabolite samples.

Firstly, several traditional multivariate analysis methodologies that have previously been suggested for ToF-SIMS data analysis are discussed, including Clustering, Principal Components Analysis (PCA), Maximum Autocorrelation Factor (MAF), and Multivariate Curve Resolution (MCR). In particular, PCA is selected as an example to show the performance of traditional multivariate analysis techniques in dealing with large ToF-SIMS data extracted from metabolite samples. In order to provide more realistic and meaningful interpretation of the results, Non-negative Matrix Factorisation (NMF) is presented. This algorithm is combined with the Bayesian Framework to improve the reliability of the results and the convergence of the algorithm. However, the iterative process involved leads to considerable computational complexity in the estimation procedure.

Another novel algorithm is also proposed which is an optimised MCR algorithm within alternating non-negativity constrained least squares (ANLS) framework. It provides a more simple approximation procedure by implementing a dimensionality reduction based on a basis function decomposition approach. The novel and main feature of the proposed algorithm is that it incorporates a spatially continuous representation of ToF-SIMS data which decouples the computational complexity of the estimation procedure from the image resolution. The proposed algorithm can be used as an efficient tool in processing ToF-SIMS data obtained from metabolite samples.

Acknowledgements

Four years ago I decided to begin the biggest challenge of my life, to study for a PhD program in engineering, I was so passionate at the beginning until bad things kept hitting me since three years ago. But luckily I have made it here to the end, I feel fulfilled no matter what the result will be, I am very fortunate that I have met so many nice people, who have stood by me to face the bad things during this challenging time.

Here I want to give my whole respect to one person, Professor Visakan Kadirkamanathan, who is not only my supervisor on this program, but also a mentor for my life. I made it to the end as one better and stronger person mostly because of his guidance. What he has given to me is not only the professional academic advice, but also the suggestions that has helped me through each barrier in my life. He is the one who has told me never give up when I wanted to, has told me life is not only about gaining the results, has told me that sometimes losing one thing doesn’t mean that you lose your whole life. Thank you, sir, for everything, without you, I could not have this opportunity to become stronger and could never have the braveness to face the challenge.

Dr Seetharaman Vaidyanathan, you are also a marvellous supervisor I have ever had during my study, your patient and careful instruction is one of the most important factors that help me to complete this program. And also you are a very kind and gracious person, your face is always filled with smile, which is one of my best memories during this difficult journey.

And Parham Aram, thank you very much for never ignoring me even when I lost contact with you for ages, and every little help from you is worth a lot to me, also thank you for helping me through that dark time and giving me so many useful advice.

Also I want to thank Andrew Hills, you are one of the nicest person I have ever met, your personality is so great. You and your academic suggestions were very helpful to me at the beginning of this program.

Dazhi, though we haven't been in touch for a year, your kind and sociable personality really impressed me and you are not only a colleague who gave me all the useful academic advice but also a true friend in my life.

And I also want to thank Sean, Xiliang, as well as other colleagues in our group, who are a group of kind people, I appreciate all the help from you guys, I am so lucky to have you as my colleagues and I will always remember your friendship and support.

To my parents, all the relatives in my family and Mr. Wang, without all of you, I could not make the first step as well as the last step of this important part of my life, I love you all so much. A special thank to all my friends for putting up with me. And also thank to myself for not giving up, I adore the whole time of this program, and it will be one of the most valuable memories of my life.

At last I want to dedicate this work to my grandfathers, Guoxiang Shen and Jieren Qian, I know you both have never been away from me, I love you forever.

Table of Content

Abstract I

Acknowledgements III

List of Figures and Tables VIII

Abbreviation List XII

Chapter 1 1

Introduction 1

1.1 Background 1

1.2 Motivation and Purpose 4

1.3 Materials and Methods 6

1.4 Thesis Structure 7

Chapter 2 9

Multivariate Statistical Analysis Methods 9

2.1 Introduction 9

2.2 Background 10

2.3 Multivariate Analysis Techniques 14

2.4 Clustering Analysis 15

2.5 Principal Component Analysis (PCA) 21

2.6 Maximum Autocorrelation Factors (MAF) 24

2.7 Independent Components Analysis (ICA) 26

2.8 Multivariate Curve Resolution (MCR) 29

2.9 Summary 33

Chapter 3 35

Principal Component Analysis 35

3.1 Introduction 35

3.2 Data Description 36

3.3 Principal Component Analysis 39

3.4 Random Sampling 42

3.5 Poisson Scaling 42

3.6 Application Results 43

3.7 Conclusion 62

Chapter 4 64

Non-negative Matrix Factorisation 64

4.1 Introduction 64

4.2 Non-Negative Matrix Factorisation Model 65

4.3 Methodology 67

4.3.1 Multiplicative Update Algorithms 68

4.3.2 Gradient Descent Algorithms 69

4.3.3 Alternating Least Squares Algorithms 70

4.4 Applicable Constraints 71

4.5 Application Results 72

4.6 Conclusion 82

Chapter 5 83

Non-negative Matrix Factorisation under the Bayesian Framework 83

5.1 Introduction 83

5.2 Bayesian Non-Negative Matrix Factorisation 84

5.3 Methodology 85

5.3.1 The Statistical Perspective 85

5.3.2 Factor and Loadings Estimation 87

5.3.3 Model Order Selection 88

5.3.4 B-NMF Iterative Algorithm 90

5.4 Application Results 90

5.5 Conclusion 105

Chapter 6 106

Alternating Non-negative Least Squares 106

6.1 Introduction 106

6.2 Algorithm 108

6.2.1 Model 108

6.2.2 ALS algorithm 110

6.2.3 Model Decomposition 111

6.3 Spatial Frequency Analysis 115

6.4 Application Results 116

6.5 Conclusion 125

Chapter 7 127

Conclusion and Future Work 127

Reference List 132

List of Figures and Tables

Figures:

Figure 1.1 The basic structure of a secondary ions mass spectrometry. The sample is mass analysed using secondary ions mass spectrometer, static SIMS spectra from the surface of samples can be obtained by the end of the spectrometer process. The ions sources can be employed in three ways: surface ionisation, electron ionisation and liquid metal ionisation, with Bi+, Bi3+, Bi3++, Cs+ and C60+ ion sources commonly equipped (Dubey et al.,2011). 3

Figure 2.1 ToF-SIMS working procedure schematic diagram. two secondary ions are accelerated to fly via the field free drift to reach the detector. The light (white) one approaches the detector earlier than the heavy (red) one. 12

Figure 2.2 visualisation of the ToF-SIMS dataset. One ToF-SIMS image of 128×128×100. 13

Figure 2.3 Hierarchical clustering procedure tree. Divisive clustering sets all the objects into one cluster at the beginning and splits them into different clusters step by step while agglomerative approach involves the reverse procedure..16

Figure 2.4 illustration of scaling problem. Two kinds of different classification choices can be made due to the different scale measurements. 18

Figure 2.5Linkage tree of one hierarchical clustering. The 8 objects merge into one cluster at the end of the tree...20

Figure 3.1 Total ion images. Total ion images for three sets replicate measurements of three pure species (C, P, and T) and two mixtures (TC and TPC). 38

Figure 3.2 Mass spectral data. For each sample dataset, one mass spectral plot can be created at every pixel point. 39

Figure 3.3 PCA scree plot. The normalised scree plots for individual components C, T, P, TC and TPC mixtures for three replicates are shown. Each of three replicate samples is plotted in three different colours. Only the first twenty principal components are presented. In each plot dashed lines show 90% cumulative variability, indicating the number of factors required to approximate at least 90% characteristics of the original samples...47

Figure 3.4 Scores and loading plots produced using PCA for C1, C2 and C3 samples. Score images are presented on the left showing the spatial information of each PC and loadings are presented on the right indicating the intensity of PCs...................49

Figure 3.5 Corresponding Score images at each significant peak from the original data for C1, C2 and C3 species samples. (A-C) are the total ion images of C1, C2 and C3 samples. The corresponding scores images at significant peaks are also shown.......51

Figure 3.6 Species C1 score image for peak at m/z = 191.02..52

Figure 3.7 Scores and loading plots produced using PCA for T1 and P1 species samples. 52

Figure 3.8 Scores images and loading plots produced using PCA for TC1 species sample. First PC and second PC are shown in the upper and lower row, respectively. 54

Figure 3.9 Score images and Loading plots produced using PCA for TPC1 mixture samples. 56

Figure 3.10 Sampled datasets from the original dataset. Sample size of 4096, 1024, 256, and 64 (shown as the black points in the images) have been randomly selected from the original dataset, from upper left to bottom right. 58

Figure 3.11 Difference ratio between original dataset and different experiment dataset. Errors between the results of original dataset and sampled dataset for PC1 (blue curve), PC2 (red curve) and PC3 (green curve) are given separately.........................59

Figure 3.12 Loading plots produced using PCA for TPC1 sampled datasets. Three PCs are ordered in descending order according to the degree of importance. 61

Figure 3.13 Scores images and loading plots produced using PCA with random sampling of a sample size of 4096 for TPC1 mixture. Score images are presented on the top while loadings for each of three PCs are shown below. 62

Figure 4.1 Three sets of loadings and scores images from NMF application for each pure component (T, C, and P) samples respectively. Scores images and loading plots are given for each species..75

Figure 4.2 Convergence of experiments with different factors and number of iteration. This diagram shows the Frobenius norm errors between the original data X and the product of the factorisation matrices WH with respect to the changes of different factor numbers and number of iteration, which indicate the speed of convergence of NMF...76

Figure 4.3 Scores images and loading plots from factorisation of TC1 mixture samples using NMF, two factors are utilised....77

Figure 4.4 The Frobenius norm errors of experiments with different iteration numbers. This diagram shows the convergence of the NMF algorithm with different iteration numbers for the mixture TC and TPC respectively..78

Figure 4.5 Scores images and loading plots produced using NMF for TPC1 mixture samples. Three spectral factors are given on the bottom of each pannel. 80

Figure 4.6 Scores images and loading plots produced using sparsity constraint NMF for TPC1 mixture samples. Different regularising parameters were chosen for each experiment. 81

Figure 5.1 NMF model order selection using Chib's method. The plots represent the marginal likelihood for individual component of the three pure chemical samples, T, P, and C. Only the models within 5 factors are presented..91

Figure 5.2 NMF model order selection using Chib's method. The plots represent the marginal likelihood for individual component of the two mixtures samples, TC and TPC. Only the models within 5 factors are presented. 92

Figure 5.3 Convergence rate of B-NMF algorithms for the TPC mixture The B-NMF algorithm converges fairly fast since the cost function is stabilised after only 50 iterations...93

Figure 5.4 Scores images and loading plots produced using B-NMF for T1, C1 and P1

species samples. Scores images are on the left showing the spatial information of each factor while loading plots are on the right indicating the factors. 94

Figure 5.5 Scores images and loadings plots produced using B-NMF for T1, C1 and P1 species samples with additional factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the basis. 97

Figure 5.6 Scores images and loadings plots produced using B-NMF for TC1 mixed species samples with Models of 1, 2 and 3 factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors. 99

Figure 5.7 Scores and loadings images produced using B-NMF for TPC mixed species samples with Models of 2 and 3 factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors. 101

Figure 5.8 Spatial location for each species in TPC1 mixture sample. The three images implies the spatial location for each species in the mixture, T (m/z = 180.06), C (m/z = 191.02) and P (m/z = 164.06) are shown from left to right. 102

Figure 5.9 Scores images and loading plots produced using B-NMF combined random sampling method for TPC mixed species sample. Several sample number have been chosen with the range from 85% to 90% reduction of the raw data. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors. 104

Figure 6.1 Example of a basis decomposition. A 128 pixels by 128 pixels ToF-SIMS image decomposed by a 4×4 grid of basis functions. The basis functions (shown by green curves) are scaled by the weight matrix, A. The centre of each basis function is shown by a yellow dot. The image is mapped onto -1 to 1 with arbitrary units. 116

Figure 6.2 Three sets of loading plots and scores images from ANLS application for each pure component (T, C, and P) samples respectively. Loadings are presented on the top of each panel indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis...............119

Figure 6.3 Three sets of loading plots and scores images using ANLS for TC mixture. Loadings are presented on the top of each panel indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis. 120

Figure 6.4 Three sets of loading plots and scores images using ANLS for TPC mixture. Loadings are presented on the top of each panel indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis. 121

Figure 6.5 loading plots and scores images using ANLS for single species T, P and C replicate samples with one basis. 123

Figure 6.6 Two sets of loading plots and scores images using ANLS for replicate TC mixture samples with two basis...124

Figure 6.7 Two sets of loading plots and scores images using ANLS for replicate TPC

mixture samples. 125

Tables:

Table 2.1 Merging Algorithm 21

Table 3.1 Significant peaks identified from PC loadings for C1, C2 and C3 pure species samples. The peak locations are displayed in an ascending order. 50

Table 3.2 Significant peaks identified from PC loadings produced using PCA for pure T1 and P1 samples. The locations are presented in an ascending order. Total ion images and corresponding score images are given alongside. 53

Table 3.3 Identified m/z values for peak assignment. All of the peaks can be used to identify specific individual chemical compounds while the numbers in red are the given ground truth for each chemicals. They can be used as references for the later identification of different species throughout this thesis. 53

Table 3.4 Significant peaks identified from PC loadings produced using PCA for TC1 mixture samples. 55

Table 3.5 Significant peaks identified from PC loadings produced using PCA for TPC1 mixture species samples. The peak locations are displayed in an ascending order. ..57

Table 4.1 Alternating Least Squares Algorithm for NMF 71

Table 5.1 Gibbs Sampling Procedure 88

Table 5.2 B-NMF Algorithm 90

Table 6.1 ANLS Algorithm Procedure 114

Abbreviation List

ACLS	Alternating Constrained Least Squares
AHCLS	Alternating Hoyer-Constrained Least Squares
ALS	Alternating Least Squares
ANLS	Alternating Non-negative Least Squares
B-NMF	Non-negative Matrix Factorisation under Bayesian Framework
C	Citric Acid
EFA	Evolving Factor Analysis
EPSEM	Equal Probability of Selection Method
ESI	Electrospray Ionisation
ESS	Error Sum of Squares
EVD	Eigenvalue Decomposition
FC-NNLS	Fast combinatorial Non-negativity constrained Least Squares
GC-MS	Gas Chromatography Mass Spectrometry
HMDS	Hexamethyldisilazane
ICA	Independent Component Analysis
ITTFA	Iterative Target Transformation Factor Analysis
LC-MS	Liquid Chromatography Mass Spectrometry
MAF	Maximum Autocorrelation Factor
MALDI	Matrix-assisted Laser Desorption Ionisation
MCMC	Markov Chain Monte-Carlo
MCR	Multivariate Curve Resolution
MCR-ALS	Multivariate Curve Resolution Alternating Least Squares
MLE	Maximum Likelihood Estimation
MS	Mass Spectrometry
MVA	Multivariate analysis
NMF	Non-negative Matrix Factorisation
NMR	Nuclear Magnetic Resonance
P	Phenylalanine
PCA	Principal Components Analysis
PC	Principal Component
SFA	Sub-window Factor Analysis
SIMPLISMA	Simple-to-use Self-modelling Mixture Analysis
SVD	Singular Value Decomposition
T	Tyrosine
ToF-SIMS	Time-of-Flight Secondary Ions Mass Spectrometer
WFA	Window Factor Analysis

Chapter 1

Introduction

1.1 Background

Metabolomics is one of the most profound and significant milestones in the long history of life science research. Since it was developed in the mid-1990s, metabolomics has become a vital part of biological systems and has already penetrated into many important research subjects. While genomics and proteomics strive to explore the activities of life from the aspect of genes and proteins, many of the inter-cellular life activities is actually regulated by metabolites, such as cell signalling, energy transfer, as well as the inter-cellular communication. Metabolites can be considered as a reflection of the environment in the cell, which contains information about the nutritional state, the effects of drug treatment and environmental changes, and the impacts of other external factors (Clarke &

Haselden, 2008). Some researchers believe that, as compared with genomics and proteomics, metabolomics would play an increasingly important role in clinical practice (Schmidt, 2004). It can provide an in-depth examination of the actual impacts from gene expression with less information required.

The term “metabolic profiling” refers to the process of measuring the chemical reactions or dynamic responses of metabolites to external factors (Miura et al., 2009). This terminology was introduced by Horning et al. in the early 1970s when they studied the compounds in human biological samples, which was based on the idea initially developed by Williams et al. (1956) that human biological fluids might carry certain type of patterns or gene expression of genetically caused diseases. Nowadays, metabolic profiling has been widely approved by professionals and academic society, owing to its ability to examine the changes caused by external factors, understand the biological variation, detect genetic diseases in the early stage, and allow more tailored health solutions (Clarke & Haselden, 2008).

The main metabolic profiling tools are nuclear magnetic resonance (NMR) and mass spectrometry (MS) (Beckonert et al., 2007). NMR is a relatively insensitive tool which is particularly suitable for identification of structural information of metabolites (Ibáñez et al., 2013). By detecting the NMR spectra of a series of samples, the pathophysiological state of an organism can be determined with pattern recognition methods. It is also possible to identify the biomarkers in order to provide a predictable platform for the relevant research. By contrast, MS is typically combined with some separation techniques, such as liquid chromatography (LC-MS) and gas chromatography (GC-MS), in order to study specific chemicals or substances of interest (Clarke & Haselden, 2008).

In general, MS related technologies outperform NMR in the sense that it is capable of providing spectra with high sensitivity and resolution (Ibáñez et al., 2013). The most common MS include quadrupole, time-of-flight (ToF) analysers, magnetic sectors, Fourier transform, and quadrupole ion trap, among which ToF-SIMS (time-of-flight secondary ions mass spectrometer) is one of the most powerful surface characterisation techniques that allows spectral analysis and direct chemical state imaging (Choi et al., 2003; Belu, Graham, & Castner, 2003). Similar to many other spectrometers, the main function of ToF-SIMS is to separate or resolve the ions formed in the ionisation source according to their mass-to-charge (m/z) ratios. The *m* denotes the mass number of the molecule since the molecular ion is equal to the molecular weight of the compound, while *z* refers to the charge number of the ion. Tof-SIMS is typically implemented along with some imaging mass spectrometer techniques, such as matrix-assisted laser desorption ionisation (MALDI) and electrospray ionisation (ESI) (Cotter, 2011). With the assistance of ToF-SIMS, researchers can obtain large amount of information about the biomolecules from the mass spectral features of the metabolites samples. The following chart shows the basic structure of a typical secondary ions mass spectrometer (Figure 1.1):

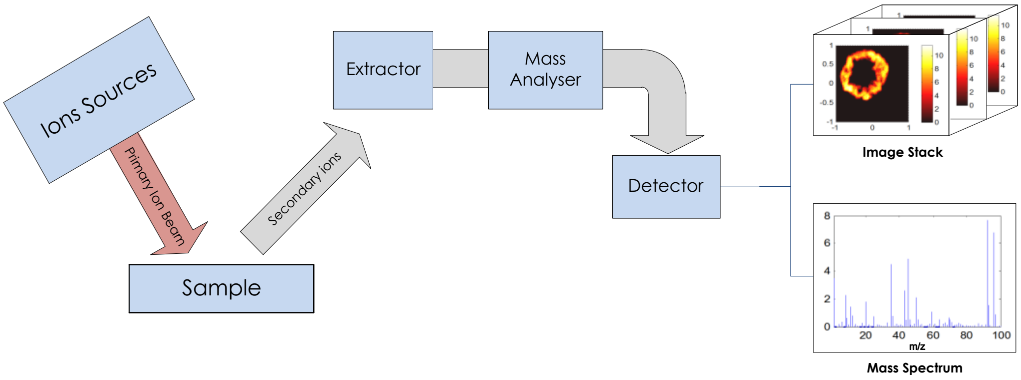


Figure 1.1 The basic structure of a secondary ions mass spectrometry. The sample is mass analysed using secondary ions mass spectrometer, static SIMS spectra from the surface of samples can be obtained by the end of the spectrometer process. The ions sources can be employed in three ways: surface ionisation, electron ionisation and liquid metal ionisation, with Bi+, Bi3+, Bi3++, Cs+ and C60+ ion sources commonly equipped (Dubey et al.,2011).

The flexibility of the ToF-SIMS technique and the high utility of data produced have generated strong interest in its application for biochemical characterisation (Belu, Graham, & Castner, 2003). While ToF-SIMS has been originally utilised in material science, there is a growing research effort on the application in bioscience field, such as analysis of lipid, peptide, tumour spheroids and cancer cell samples (Vickerman & Briggs, 2001; Passarelli & Winograd, 2011; Kotze et al., 2013; Aoyagi et al., 2013).

1.2 Motivation and Purpose

ToF-SIMS is increasingly popular due to its in-situ ion separation methodology. It involves the free flight of the ionised molecules in a field-free drift tube. ToF-SIMS is widely utilised by analysts and researchers because of the following notable features (Belu, Graham, & Castner, 2003):

- Fast parallel detection of all ions and high sensitivity
- High mass range (theoretically unlimited)
- High mass resolution > 10,000
- High mass accuracy (1-10 ppm)
- High transmission and spatial resolution
- Ability to cover all elements, isotopes, as well as molecular species

While the advantages of ToF-SIMS are particularly attractive to metabolomics research and application, the output data can be substantially large due to the high spectral and spatial resolution (Graham, Wagner, & Castner, 2006). It is therefore very difficult to find relevant information or detect specific species, which makes data mining problematic (Sodhi, 2004).

The output data of ToF-SIMS can be represented as a combination of thousands of individual spectrum. One typical ToF-SIMS spectrum contains hundreds or thousands of different intensity peaks, depending on the order, structure, composition, and orientation of the surface species. It is not uncommon that many of the peaks within a given spectrum are somehow interrelated, since they are often derived from the same surface species. As a result, one of the challenges in

ToF-SIMS data analysis is to determine which peaks are interrelated and how they contribute to the chemical differences present on the surface. This is further complicated by the fact that ToF-SIMS dataset typically contains multiple spectra generated from multiple samples, which result in a large and complex data matrix to be analysed.

The large size of the output dataset can cause a number of problems for the interpretation of metabolites. When comparison between two features needs to be made, the high cost of computation caused by a large dataset would hamper the research process and incur considerable costs. Thus ToF-SIMS dataset is usually decomposed into different profiles containing distinct components, which also provide the possibility of template matching with stored templates in a database. Another serious concern for analysing large ToF-SIMS dataset is that it is extremely difficult to separate the original chemical compounds from fragmentation of species resulting in numerous number of peaks, especially when prior knowledge of the components is not available. Therefore, researchers always attempt to explore appropriate and efficient techniques that can be used to address the problems arising from ToF-SIMS data analysis (Tyler, Rayal, & Castner, 2007).

Since metabolic profiling appears to be a new area of application for ToF-SIMS, there are only a few algorithms currently available to handle the output data from metabolite samples. Thanks to prior development of dimensionality reduction and noise removal techniques, several multivariate analysis techniques have been suggested for large and multi-dimensional chemical spectral data processing, such as Principal Components Analysis (PCA), Maximum Autocorrelation Factors (MAF), and Multivariate Curve Resolution (MCR) (Tyler, 2006). However, none of them can efficiently extract information from a large dataset while produce a clear representation and interpretation in the context of metabolic profiling analysis. Therefore development of novel and innovative algorithms are undoubtedly needed to demonstrate the potential of ToF-SIMS for metabolic profiling analysis. In this thesis, novel multivariate analysis techniques for processing ToF-SIMS data extracted from metabolite samples are derived and its application demonstrated.

1.3 Materials and Methods

The data set used throughout this thesis was obtained from the Department of Chemical and Biological Engineering, University of Sheffield. Three metabolites, tyrosine (T), phenylalanine (P) and citric acid (C) (all from Sigma Aldrich, UK) were used in the study. They are spotted on a dish as individual pure species and mixed species, resulting in a total of five separate experiments and each having three replicates. TC mixture contains T and C species in equimolar proportions and TPC mixture comprises T, P, and C species in equimolar proportions. These metabolites were spotted on hexamethyldisilazane (HMDS) (Sigma Aldrich, UK) coated silicon wafers (Compart Technology, UK), prepared as detailed by Salim, Wright, & Vaidyanathan (2012). The images consisted of 128×128 pixels. Each spectrum was calibrated using hydrocarbon fragment peaks. Spectral data up to m/z = 200 was considered for analysis although only the intensities for 100 m/z data points were provided for the image analysis for this work.

The given dataset with known chemical compounds provides us with a controlled environment in which to test the performance of any developed algorithm. The use of the known dataset also provides the ground truth and gives us the ability to interpret whether the results have a valid explanation. This is particularly important when using scale dependent methods such as PCA or MCR since the results obtained will be affected by the assumptions made when pre-processing the data. However, there is no knowledge of the exact spatial localisation of the different species, no quantitative measures exist to test for the complete validation of a given result. The development of the methods and their analysis was carried out using one dataset and tested with the two replicates in order to examine and validate the results.

The underlying properties of the data that any algorithm needs to exploit result in the following requirements for the algorithms to be developed:

1. Dimensionality reduction – Removing redundant information
2. Feature extraction – Identification of discriminatory spectral peaks
3. Factorisation – Separation of spatial and spectral information
4. Sparsity analysis – Exploit redundancy in data (number of components)
5. Spatial correlation analysis – Exploiting spatial correlation

These requirements were the backbone for the development of the methodologies in this thesis.

1.4 Thesis Structure

The remainder of this thesis is organised as follows:

Firstly, a brief background and general working principle of the ToF-SIMS process will be detailed in Chapter 2. An outline of several multivariate analysis methodologies that have previously been suggested for ToF-SIMS data analysis are discussed and their contribution to ToF-SIMS data processing is reviewed.

Chapter 3 presents the implementation of a widely used method, the principal component analysis (PCA) to the ToF-SIMS dataset. This application is an unsupervised analysis procedure aimed at extracting features from large scale dataset while reducing the dimensionality. This implementation results are discussed and is shown to be promising in overcoming the complexity challenges presented by ToF-SIMS data.

Chapter 4 introduces a non-negativity constrained algorithm, namely non-negative matrix factorisation (NMF), which focuses on improving the interpretability of the results. This exploits the fact that ToF-SIMS data are essentially non-negative quantities. Unlike the PCA and other multivariate analysis techniques, this algorithm is capable of providing physically meaningful results and facilitating data mining procedure.

The algorithm in Chapter 4 is extended in Chapter 5 by incorporating a Bayesian framework, and referred to as B-NMF. This method shows its capability in reducing the uncertainty and correlations that exist in the dataset. Moreover, it also provides an appropriate number of components indicative of the number of species in an unknown complex metabolic system.

A novel Alternating Non-negative Least Squares method (ANLS) is presented in Chapter 6. This technique is combined with MCR in order to take advantage of its ability to identify the chemical compounds or species of interest while taking the spatial correlation into account. It provides a simplified approximation of the data by implementing a dimensionality reduction method based on a basis function decomposition approach, significantly reducing the computational demand. This novel algorithm has high potential to be used as a effective tool in processing ToF-SIMS data extracted from metabolite samples.

The conclusions from the findings of the thesis are given in Chapter 7, where we will also discuss possible improvements that can be made to the analysis methods proposed here. It also includes suggestions for future research in metabolic profiling.

Chapter 2

Multivariate Statistical Analysis Methods

2.1 Introduction

Perhaps no other instrument that is more indispensable than mass spectrometer to today's science research. It has also become an essential tool in metabolic profiling analysis (Balmer et al., 2013). Because of the accuracy and high sensitivity provided, a ToF-SIMS can produce a high dimensional data cube, which provides detailed molecular information and high spatial resolution. However, due to the complexity of the species and the fragmentary nature of ToF-SIMS dataset, the resulting data is not always easy to interpret. This poses a serious threat to the usefulness and practical applications of mass spectrometer related techniques. Several multivariate analysis methods have been previously used in addressing this problem. Currently the most popular method is principal component analysis (PCA) with singular value decomposition (SVD) approach, which is a basic as well as one of the earliest decomposition methods (Pearson, 1901; Hotelling, 1933). It identifies discriminatory features by finding the new projection with maximum variances between the components. Maximum autocorrelation factors (MAF) is an alternative to PCA based on maximising the autocorrelation between neighbouring pixels (Switzer & Green, 1984; Larsen, 2002). Another similar method is independent component analysis (ICA), it selects component from one unknown 'blind' mixture with a more rigorous assumption that the components are independent to each other (Linsker, 1992; Bell & Sejnowski, 1995). By contrast, multivariate curve resolution (MCR) is a feature extraction technique that is useful for providing the pure spectra of components in the system (Lawton & Sylvestre, 1971; de Juan & Tauler, 2006). Some other classical statistical methods, like clustering, can provide the benefit of grouping components with similar patterns into subsets.

This chapter will firstly provide a brief background of ToF-SIMS and description of its basic working principle. We will then explain the property a method should have to solve those problems by introducing the general data problem of ToF-SIMS. In addition, we will outline several well-known methodologies that have been extensively used in processing multi-dimensional data, including clustering, PCA, ICA, MAF, and MCR.

2.2 Background

A century has passed since the first prototype of mass spectrometer was originated by the winner of 1906 Nobel Prize in Physics, Sir Joseph John Thomson (Downard, 2012). This special analyser provides mass spectrum of an identical chemical sample, which is a plot of the ion signal as a function of the mass-to-charge ratio.

Mass spectrum can be considered as the fingerprint of the chemicals compounds. It is useful for determining the composition of the molecules, the distribution of chemical species, and chemical structures on the observed surface (Sodhi, 2004). Mass spectrometry can provide numerous possibilities for the analysis of complex systems, especially in the field of chemistry, biology, geology, military, environment and astronomy.

Currently, considerable research effort has concentrated on mass spectrometer and hence the inventions of different kinds of support machines. Different mass analysers vary in features, including the m/z range that can be covered, the mass accuracy, and the achievable resolution. An effective spectrometer will provide a detailed surface characterisation in order to not only identify the temporal and spatial patterns, but also verify the desired changes have been made. These factors require the ability to obtain the distribution, structure and chemical compounds of the surface species (Belu, Graham, & Castner, 2003).

A ToF-SIMS determines the masses of secondary ions by recording their flight time (Choi et al., 2003). It utilises a pulsed ion beam to obtain secondary ions, which are then forced into the ToF analyser by a fixed high voltage (Sodhi, 2004). The extracted secondary ions are subsequently accelerated into the field-free drift tube, and a detector is placed at the finishing point of the flight path in order to monitor the pulses of these secondary ions. To ensure constant ion energy, ToF-SIMS typically incorporates a number of techniques to manage the differences in the initial condition and the energy dispersion of the extracted secondary ions. One ToF analyser working schematic is shown in Figure 2.1.

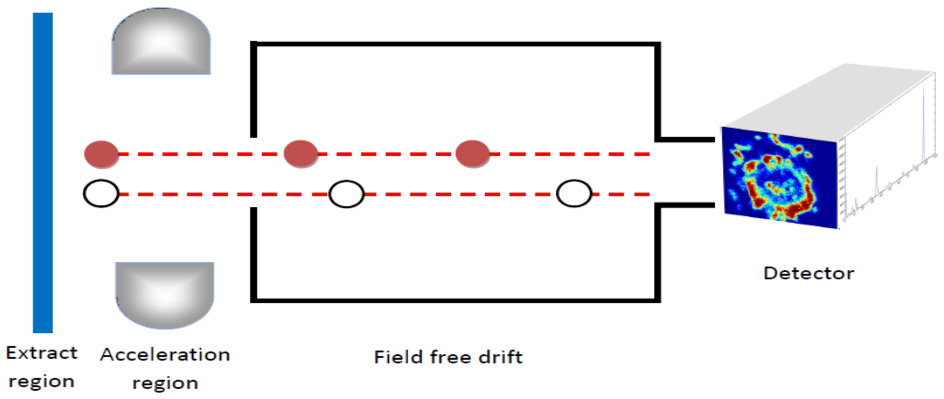


Figure 2.1 ToF-SIMS working procedure schematic diagram. Two secondary ions are accelerated to fly via the field free drift to reach the detector. The light (white) one approaches the detector earlier than the heavy (red) one.

Given the same amount of energy provided during this process, the only difference between secondary ions' flight time is the velocity, which is primarily determined by their masses (Belu, Graham, & Castner, 2003). This relationship is shown as follows:

The velocity of secondary ions in a constant energy state can be simply expressed as:

$$\text{velocity} = \sqrt{\frac{2\times\text{energy}}{\text{mass}}} \qquad (2.1)$$

There is a positive relationship between flight-time and the mass of ion, as it can be demonstrated with a simple algebraic rearrangement:

$$\text{Flight time} = \frac{\text{drift length}}{\text{velocity}} = \text{drift length} \times \sqrt{\frac{\text{mass}}{2\times\text{energy}}} \qquad (2.2)$$

Thus a set of flight times will give a set of mass values that can be plotted as mass spectrum.

A major strength of time-of-flight mass spectrometer is parallel detection of ions, which means that it is possible to capture all the secondary ions of different masses and generate a complete mass spectrum (Boxer, Kraft, & Weber, 2009). Whereas, many other mass spectrometers, such as quadrupole analyser and magnetic sector analyser, are subject to a restricted mass range and lower mass transmission (Reed & Vickerman, 1993). In other words, the mass spectra produced by these mass spectrometers only represent the ions within a given mass range. Thus, parallel ion detection allows ToF-SIMS to handle secondary ions of high masses and have a relatively high sensitivity. Beside excellent sensitivity and high mass range, ToF-SIMS also benefits from high mass and spatial resolution (Belu, Graham, & Castner, 2003). This has made ToF-SIMS a promising instrument for biological analysis applications.

One typical ToF-SIMS mass spectrum may contain hundreds of peaks. The relative intensities of many of these peaks are interrelated since they come from the same surface species. In addition, even for the simplest single component samples, changes in the surface chemistry can affect the relative intensities of the peaks for a given sample system.

A typical ToF-SIMS dataset is illustrated in Figure 2.2 as a microscopic image cube of a sample surface along its mass spectrum (m/z). An image can be created for each mass with the loadings of the corresponding mass scores in every pixel. In our case, the replicate mixture dataset contains several 128×128 image stacks along the mass spectrum up to m/z = 100.

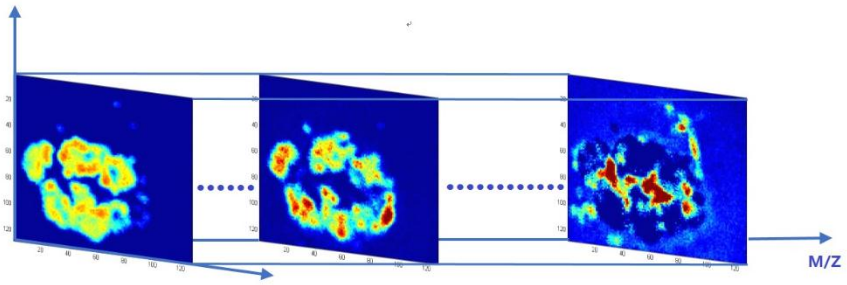


Figure 2.2 visualisation of the ToF-SIMS dataset. One ToF-SIMS image of 128×128×100.

In the remainder of this chapter, we will outline several traditional multivariate techniques which are capable of identifying and extracting the useful information from large dataset, and hence reconstructing a data matrix with lower dimension.

2.3 Multivariate Analysis Techniques

The three-way array data in Figure 2.2 is made of mass spectra throughout the sample surface, in which multivariate analysis (MVA) methods are useful to provide insight to the identification of unknown number and types of the chemicals. This requires feature detection and extraction ability. Two types of defining and learning analysis methods are commonly used: supervised learning and unsupervised learning.

Supervised learning is probably the most straightforward analysis method that aims to seek for one satisfactory model with a set of given inputs and outputs. In our thesis, there is no prior information available, therefore an unsupervised learning would be more appropriate. In unsupervised learning, analysis typically involves detecting patterns and categorising objects purely based on the statistical characteristics.

While supervised learning model is utilised with sets of known inputs and outputs, no examples are given to the model in the unsupervised learning. Instead, patterns are derived directly from the given data, which is the case in our project. Moreover, it is possible to find the hidden structure from the unknown data using unsupervised learning.

Two popular method widely used in unsupervised learning are clustering and factorisation. Clustering is the grouping procedure which classifies similar components according to specific measurements. It is a main task of exploratory data mining as well as a common technique for statistical data analysis. By contrast, factorisation is one “blind” feature identification technique offering dimensionality reduction benefit, it involves many different approaches, such as principle component analysis (PCA), independent component analysis (ICA), non-negative matrix factorisation (NMF), etc. All of these algorithms seek to extract and explain the key patterns of the data.

2.4 Clustering Analysis

Clustering is a statistical procedure for identifying object groups with similar patterns. It became well-known due to its application in psychology for personal trait classification (Cattell, 1943). The objective of clustering analysis is to split a set of objects into distinct groups (classes, clumps, and clusters) based on a chosen criterion (Jain, Murty, & Flynn, 1999).

The process of clustering is similar to classification, as they all deal with finding the relationship inside the dataset. However, a pre-training step for defining the groups’ characters is usually needed for a classifier, it would then learn from the different data group with the ability to classify. This process is a supervised learning as we mentioned previously. Whereas, clustering an unsupervised learning in which the grouping procedure is solely driven by the similarity within the data. It is therefore important to determine how to define the similarity.

Methodology

There are two classes of clustering method, one is called distance-based clustering which uses the distance between each objects as the similarity criterion; another clustering approach is called conceptual clustering which uses the concept in common to all objects as criterion (Jain, Murty, & Flynn, 1999; Michalski & Stepp, 1983). The latter one is much more complicated than the former kind, because the objects are organised according to certain ‘descriptive concept’, which is different from the simple similarity measure.

Various mathematical methods are now widely implemented in the clustering algorithm, among which, Hierarchical method, Partitioning method, Density-based method, Grid-based method, Model-based method are five popular ones. The first two methods are based on the statistic distance of objects. Hierarchical clustering, for instance, is based on the union or the division of the dataset (Johnson, 1967). The procedure can be obtained in two ways: divisive and agglomerative, the principle can be shown as the graph below:

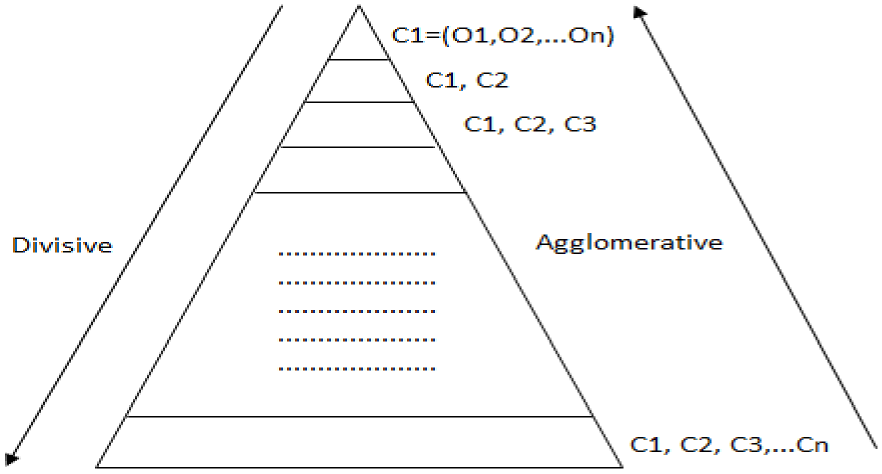


Figure 2.3 Hierarchical clustering procedure tree. Divisive clustering sets all the objects into one cluster at the beginning and splits them into different clusters step by step while agglomerative approach involves the reverse procedure.

Agglomerative procedure is based on the union between the two “nearest” clusters regarding to the distance, whereas, the divisive algorithm is based on the division of each cluster (Jain, Murty, & Flynn, 1999). Because divisive clustering is a global method, in order to gain a global view, it requires other algorithms besides itself, leading to larger amount of computation, which is not practical in many cases.

Distance measure

Distance is the most important factor in many clustering algorithms, as it is one widely approved way to define similarity.

For a mapping $\mathrm{d}: \mathrm{U} \times \mathrm{U} \rightarrow / \mathrm{R}$

It is called a distance function if, for any $\mathrm{x}, \mathrm{y} \in \mathrm{U} : \mathrm{d}(\mathrm{x}, \mathrm{y}) \geq 0$; $\mathrm{d}(\mathrm{x}, \mathrm{x}) = 0$; $\mathrm{d}(\mathrm{x}, \mathrm{y}) = \mathrm{d}(\mathrm{y}, \mathrm{x})$. This distance function is also a metric if: $\mathrm{d}(\mathrm{x}, \mathrm{y}) = 0$ then $\mathrm{x} = \mathrm{y}$; And,

$$\mathrm{d}(\mathrm{x}, \mathrm{y}) \leq \mathrm{d}(\mathrm{x}, \mathrm{z}) + \mathrm{d}(\mathrm{z}, \mathrm{y}) \tag{2.3}$$

The best known distance measurement between two points in a plane, which is the Euclidean metric defined by:

$$\mathrm{d}_2(\mathrm{x}, \mathrm{y}) = \|\mathrm{x} - \mathrm{y}\|_2 = \sqrt{(\mathrm{x} - \mathrm{y})^{\mathrm{T}}(\mathrm{x} - \mathrm{y})} \tag{2.4}$$

The Euclidean metric can be generalised in two ways. The first method is a popular measure called Minkowski metric, which is given by:

$$\mathrm{d}_2(\mathrm{x}, \mathrm{y}) = \|\mathrm{x} - \mathrm{y}\|_{\mathrm{p}} = \sqrt{(\mathrm{x} - \mathrm{y})^{\mathrm{p}}} \tag{2.5}$$

It should be noted that the Euclidean distance is a special case when $\mathrm{p} = 2$, while the Manhattan distance is another special case when $\mathrm{p} = 1$.

The second method of generalisation is obtained by defining:

$$\mathrm{d}_{\mathrm{B}}(\mathrm{x}, \mathrm{y}) = \|\mathrm{x} - \mathrm{y}\|_{\mathrm{B}} = \sqrt{(\mathrm{x} - \mathrm{y})^{\mathrm{T}}\mathrm{B}(\mathrm{x} - \mathrm{y})} \tag{2.6}$$

This equation is related to the famous Mahalanobis distance, however this concept is beyond the scope of our experiment and the Euclidean distance is preferred.

Scaling normalisation

Before the clustering analysis is performed, the relative scaling should be firstly considered, actually, scaling should be considered before many other algorithms.

The importance of the scaling can be illustrated in the following charts:

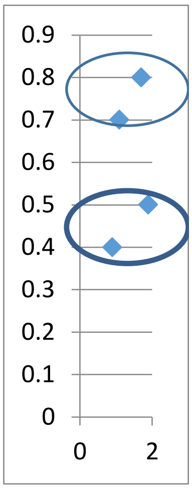


Figure 2.4 illustration of scaling problem. Two kinds of different classification choices can be made due to the different scale measurements.

Figure 2.4 shows a simple example of scaling problem for a 2-dimensional case, in which the axes have the same magnitude but with different scaling, resulting in different visualising positions of the four points, as well as different clusters definitions. In order to solve the problem, normalisation is typically required.

In this thesis, we use a normalisation method described by:

$$x' = \frac{1}{s}(x - \bar{x}) \tag{2.7}$$

Where x' is the normalised new variables, $\bar{x}$ is the mean value of the elements in x, and s is the standard deviation of the vector x.

Agglomerative algorithm

Let $n_k = m$, where n_k is the number of clusters in different clustering level, and m is the number of the objects, or cases need to cluster at the beginning. Therefore there are m clusters containing one object each.

The computation of the distance between clusters can be confusing since the distance between different clusters is not the same as the difference between different objects. This can be demonstrated in the following ways (Jain, Murty, & Flynn, 1999):

1. Single linkage clustering: the distance between two clusters equals to the shortest distance between the elements of each cluster.
2. Complete linkage clustering: the distance between two clusters is the longest distance between the elements of each cluster.
3. Average/weighted average linkage clustering: the distance between two clusters is considered as the (weighted) average of the distances between every element of each cluster.
4. Centroid/weighted centroid linkage clustering: the distance between two clusters is the distance between the (weighted) centres of each cluster.
5. Ward linkage clustering: the distance is defined in terms of the error sum of squares, ESS.

After the distance computation, a merging step would take place. At each iteration, the two clusters with the shortest distance are merged into one cluster. The iterative process would continue until the ideal cluster number is achieved. For example, if you want k clusters, simply cut off the procedure at the $(\mathrm{k}-1)_{\mathrm{th}}$ iteration. The whole processing can be drawn as a linkage tree:

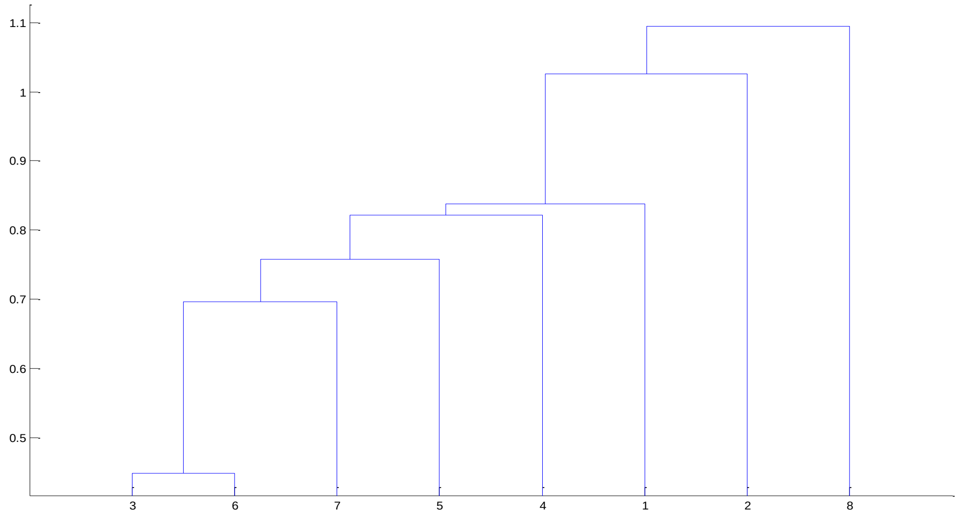


Figure 2.5Linkage tree of one hierarchical clustering. The 8 objects merge into one cluster at the end of the tree.

The horizontal axis in Figure 2.5 represents the labels of the clusters, the vertical axis stands for the distance level at every merging step. It can be seen in Figure 2.3, hierarchical clustering is considered as a bottom-up method and a divisive clustering would be considered as a top-down method, where one (or more) cluster is split into two clusters at every distance level.

Merging Algorithm

Merging steps
Arrange the m objects into a new order that results in a contiguous sequence.
Choose any object to be the first one in the sequence $s(1)$, the first gap (gap is the distance between clusters) is denoted as $G(1) = \infty$.
Select the nearest object as $s(2)$, and the gap between $s(1)$ and $s(2)$ is $G(2) = d(s(1), s(2))$.
From the rest objects, choose the one which is closest to one of $s(1)$, $s(2)$ as $s(3)$. Generalised, choose the $s(k)$ as the closest element to any one of the ready-reordered sequence $s(1), s(2), \dots, s(k-1)$ and the gap is the distance

	between the two elements, $\mathrm{G(k)}$.
	Begin with the disjoint m clusters, and find the gap, which is the maximum of the all gaps (except $\mathrm{G(1)}$), as $\mathrm{G_{max}}$, if the entire gaps before $\mathrm{G_{max}}$ are different, then, merge the elements that has the minimum gap with the related element before into one clusters, there will be one cluster less.
	Delete the rows and columns of the two merged objects, and add new row and column represent the new cluster $\mathrm{s(m+1)}$, update the previous data sequence.
	If all the objects are in one cluster, then the clustering should stop, otherwise, go back to the first step, loop again.

Table 2.1 Merging Algorithm

The computational demand of hierarchical clustering is considerably large though the calculation method is simple, especially the distance matrix. In addition, the algorithm only checks the local distribution at each merging step without checking the global distribution, therefore there is no way to change or revise what has been done. However, hierarchical clustering analysis remains a popular and easily understood method for distinguishing different groups within the data.

2.5 Principal Component Analysis (PCA)

Principal components analysis is claimed to be one of the most valuable contributions from applied linear algebra. It has been used widely in various fields due to its simplicity and outstanding applicability. The aim of PCA is to find the most meaningful basis to reconstruct a complex dataset based on a multi-dimensional orthogonal linear transformation (Hotelling, 1933). It assumes that the variables with the greatest variance are capable of explaining most part of the significant variations in the data (Abdi & Williams, 2010).

In general, the variables in a raw dataset are commonly inter-correlated, leading to unreliable data mining and complicated computation. PCA intends to find the linear combination of the original variables (the principal components) by studying the covariance between the variables. It involves rotation of the covariance matrix into orthogonal factors where variables are no longer spatially correlated (Pearson, 1901).

Our ToF-SIMS dataset in this thesis are two spectral data points which are close to each other on the surface. Due to the inter-correlated nature of the dataset, there might be a large number of superfluous and pleonastic variables, which result in redundant computation and hamper the interpretation of the data. In this case, PCA may be used to remove the correlation in the data while retain the most representative information.

Methodology

The standard PCA algorithm is given by:

$$\mathrm{D} = \mathrm{VX} \tag{2.8}$$

Where D denotes scores matrix of principal components, V and X denote loadings matrix and the original matrix respectively. PCA can be performed using two approaches: eigenvalue decomposition (EVD) and singular vector decomposition (SVD).

1. EVD approach involves the calculation of eigenvalues and eigenvectors in which the eigenvalues refer to the degree of importance of the principal components and the eigenvectors are essentially the principal components. The raw data set is decomposed using EVD into several different 'subsets' with different importance indexes, and the first several important 'subsets' are selected as the principal components, which are believed to contain the most significant properties of the original dataset. One obvious pitfall of EVD approach is that it can only be applied to square matrix, which rarely occurs in reality. The formula of EVD is shown as follows:

Let X be a $\mathrm{n \times n}$ matrix with N linearly independent eigenvectors, $\mathrm{q_i (i = 1, \cdots, n)}$ then we can decompose X as follows:

$$\mathrm{X = E \Lambda E^{-1}} \tag{2.9}$$

Where E is the eigen square matrix made of the X's eigenvectors of $\mathrm{q_i}$ and Λ is the diagonal matrix. The diagonal elements are the corresponding eigenvalues to the eigenvectors.

2. The general principle and formula of SVD are similar to EVD with a more generalised matrix size.

A $\mathrm{m \times n}$ matrix X can be decomposed in the form of:

$$\mathrm{X_{m \times n} = U_{m \times m} \times \Sigma_{m \times n} \times V^T_{n \times n}} \tag{2.10}$$

Where U is an $\mathrm{m \times m}$ orthogonal matrix, Σ is a $\mathrm{m \times n}$ diagonal matrix with non-negative real numbers on the diagonal, and the $\mathrm{n \times n}$ orthogonal matrix $\mathrm{V^T}$ denotes the transpose of V. This factorisation is called a singular value decomposition of X.

The relationship between singular value σ and eigenvalue λ can be illustrated by:

$$\mathrm{(W^T X) v_i = \lambda_i v_i} \tag{2.11}$$

$$\mathrm{\sigma_i = \sqrt{\lambda_i},\ \ u_{i=} \frac{1}{\sigma_i} X v_i} \tag{2.12}$$

Where $\mathrm{v_i}$ denotes the right singular vectors while $\mathrm{u_i}$ is the left singular vectors. The entries of the diagonal matrix Σ are always listed in a descending order for the sake of calculation. In most cases, the first few singular values (principal components) may account for more than 90% of the entries in the data. Therefore the original dataset can be approximated using a far less number of variables, r, without losing the main information of the original dataset.

$$\mathrm{X}_{\mathrm{m}\times\mathrm{n}} \approx \mathrm{U}_{\mathrm{m}\times\mathrm{r}}\Sigma_{\mathrm{r}\times\mathrm{r}}\mathrm{V}^{\mathrm{T}}_{\mathrm{r}\times\mathrm{n}} \qquad (2.13)$$

One drawback of SVD can be illustrated in one O (N^3) calculation, which means that with the expansion of the matrix size, the computation will be complicated by three times, especially with a large number of r.

With the two approaches outlined above, PCA is able to obtain several largest eigenvalues or singular values, which are believed to contain the most significant characteristics of the data, and use them as the transformation matrix.

$$\mathrm{U}^{\mathrm{T}}_{\mathrm{r}\times\mathrm{m}}\mathrm{X}_{\mathrm{m}\times\mathrm{n}} \approx \Sigma_{\mathrm{r}\times\mathrm{r}}\mathrm{V}^{\mathrm{T}}_{\mathrm{r}\times\mathrm{n}} \qquad (2.14)$$

This formula can be generalised to one transformation with the rotation matrix T:

$$\widetilde{\mathrm{X}}_{\mathrm{r}\times\mathrm{n}} = \mathrm{T}_{\mathrm{r}\times\mathrm{m}}\mathrm{X}_{\mathrm{m}\times\mathrm{n}} \qquad (2.15)$$

PCA has been widely used as a dimensionality reduction technique in ToF-SIMS data analysis (Henderson, Fletcher, & Vickerman, 2009). However, Chang (1983) found that the large eigenvalues do not always represent the characteristics of the data; in particular, PCA might not be able to identify the linear combination if all the variables in the data that have the same variance.

2.6 Maximum Autocorrelation Factors (MAF)

Maximum autocorrelation factor (MAF) involves a transformation procedure which takes into consideration of the autocorrelation between neighbouring observations (Larsen, 2002). It was firstly proposed by Switzer and Green in 1984 as an alternative transformation method to PCA. In fact, MAF and PCA are mathematically similar if the covariance matrix is linearly related to the identity matrix (Switzer & Ingebritsen, 1986; Gallagher et al., 2014).

MAF is different from PCA in the way that, instead of the covariance criterion, it employs spatial autocorrelation as the criterion to decorrelate the data. The intuition has been widely accepted due to its sound assumption that noise tends to have a smaller spatial autocorrelation relative to significant components (Storvik, 1993). If noise components in the dataset have larger variance relative to the interesting components, PCA would lead to poor and unreliable representation, as it is unable to recognise whether the linear combination is attributed to the interesting components or noise (Keenan & Smentkowski, 2011). This means that MAF would outperform PCA when the interesting components have lower variance and higher autocorrelation than noise, vice versa (Larsen, 2002).

Methodology

MAF was developed on the basis of PCA. In order to account for autocorrelation between neighbouring observations, MAF employs a shifted matrix that is found by taking the difference between the original data matrix and a spatially shifted duplicate of itself (Tyler, Rayal, & Castner, 2007). The original dataset X can be decomposed by regular PCA method in Equation (2.3), where the matrix V is obtained by an eigenvector rotation of the MAF factor. In order to differentiate from PCA, the MAF transformation can be described by the following linear combinations:

$$\mathrm{S} = \mathrm{A}^{\mathrm{T}}\mathrm{X} \qquad (2.16)$$

Where the MAF factor A is obtained by

$$\mathrm{A} = \mathrm{U}_2^{\mathrm{T}}\Lambda^{-\frac{1}{2}}\mathrm{U}_1 \qquad (2.17)$$

U_1 denotes the eigenvectors while Λ denotes the eigenvalues of the matrix B, where B is the covariance matrix of the original dataset W, which can be specified by the equation below:

$$\mathrm{U}_1\mathrm{B}\mathrm{U}_1^{\mathrm{T}} = \Lambda \qquad (2.18)$$

U_2 is the eigenvectors from the EVD of the shifted matrix, which can be derived from the equation below:

$$U_2X(\Delta)U_2^T = U_2(\frac{1}{2}([\Gamma_W(\Delta)]^T + [\Gamma_Y(\Delta)]))U_2^T \quad (2.19)$$

In this equation, Γ_Y is the spatial correlation, which is defined by Equation (2.20):

$$\Gamma(\Delta) = \mathrm{Cov}\{X_k, X_{k+\Delta}\} \quad (2.20)$$

With the property given by:

$$\Gamma^T(\Delta) = \Gamma(-\Delta) \quad (2.21)$$

Where k denotes the spatial position while Δ is one spatial movement. The matrix derived via the MAF method transforms the variance-covariance matrix to the identity matrix and the shifted matrix for spatial shift of Δ to a diagonal matrix. MAF produces uncorrelated variables with largest autocorrelations using joint diagonalisation of asymmetric covariance matrices.

2.7 Independent Components Analysis (ICA)

Independent Component Analysis (ICA) is also a widely applied tool for identifying components from mixtures and it has been presented in some particular spectral data analyses for the use of identifying the unknown components in the mixture as well as in estimating their concentrations without prior knowledge (Chen & Wang, 2000; Bayliss et al., 1998). ICA was firstly introduced by Herault and Jutten (1986) to address so called “blind source separation” problem based on the assumption that signals originated from different sources in a mixture are mutually independent in distribution (Comon, 1994). ICA is generally considered as an extension of PCA since it also transforms the data into uncorrelated factors. However, ICA employs a more rigorous criterion since statistical independence always leads to uncorrelation, while the converse does not necessarily hold (Hyvärinen & Oja, 2000). In addition, there is no order associated with the components extracted by ICA, whereas PCA assumes that the first principal component has the largest explanatory power to the variation of the data (Langlois, Chartier, & Gosselin, 2010).

There are two major approaches for ICA algorithms, arising from different interpretation of the statistical independence (Haykin, 2009). InfoMax and Maximum Likelihood estimation are algorithms for ICA developed on the basis of information theory which minimises the Shannon mutual information of pairs of variables (Amari, Cichocki, & Yang, 1996; Bell & Sejnowski, 1995; Pham, Garrat, & Jutten, 1992). By contrast, FastICA is an approach based on the intuition that mutually independent distribution can be properly measured by the deviation from normal distribution (non-Gaussianity) (Hyvärinen & Oja, 2000). Therefore, a fundamental limitation of ICA is that the independent components must be non-Gaussian for ICA to be applicable.

Methodology

ICA transform seeks linear combinations that minimise the statistical independence between variables. InfoMax is the approach rooted in the minimisation of mutual information, which utilises entropy as a primary measure of the uncertainty.

InfoMax

Entropy can be considered as the degree of information that the observations of variables provide. Larger entropy is typically related to more random and unpredictable variables (Hyvärinen & Oja, 2000). Conversely, lower entropy means that we have more information about a given system. Entropy can be considered as a measure of non-Gaussianity since a Gaussian variable typically has the greatest entropy among all variables for a given variance. This means that Guassian variables have more “random” distributions. For a discrete random variable X, entropy H is defined as:

$$\mathrm{H}\,(\mathrm{X}) = -\sum_{x} \mathrm{P}(\mathrm{x})\log\mathrm{P}(\mathrm{x}) \tag{2.22}$$

$$\mathrm{H}\,(\mathrm{Y}) = -\sum_{y} \mathrm{P}(\mathrm{y})\log\mathrm{P}(\mathrm{y}) \tag{2.23}$$

$$\mathrm{H}\,(\mathrm{X},\mathrm{Y}) \;=\; -\sum_{x,y}\mathrm{P}(\mathrm{x},\mathrm{y})\mathrm{logP}(\mathrm{x},\mathrm{y}) \qquad (2.24)$$

Where $\mathrm{P}(\mathrm{x})$ is the probability that X is in the state x. Differential entropy is the case when the ordinary concept of entropy is generalised for continuous random variables. The differential entropy H of a random variable x with density $\mathrm{f}\,(\mathrm{x})$ can be described by:

$$\mathrm{H}(\mathrm{x}) \;=\; -\int \mathrm{f}\,(\mathrm{x})\,\log \mathrm{f}\,(\mathrm{x})\mathrm{dx} \qquad (2.25)$$

The mutual information I between m (scalar) random variables, $\mathrm{x_i}$, $\mathrm{i} = 1\ldots\mathrm{m}$ can be defined as follows:

$$\mathrm{I}(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_m) \;=\sum_{\mathrm{i}=1}^{\mathrm{m}}\mathrm{H}(\mathrm{x}_i) - \mathrm{H}(\mathrm{x}) \qquad (2.26)$$

The mutual information can be interpreted as the Kullback-Leibler divergence between the joint density $\mathrm{f}\,(\mathrm{x})$ of random variables (Amari, Cichocki, & Yang, 1996). Therefore, mutual information is a proper measure of independence between random variables as it is non-negative in nature and equal to zero when the variables are statistically independent. By minimising the mutual information, we are able to identify the most statistically independent components. The methods based on mutual information minimisation are preferable in a changing environment (Langlois, Chartier, & Gosselin, 2010).

FastICA

Another approach to measure statistical independence also involves the concept of non-Gaussianity, where negentropy is used a quantitative measure of non-Gaussianity of random variables. Negentropy is a measure of the deviation from normality, which indicates the degree of statistical independence of variables. Negentropy is defined by:

$$\mathrm{J}(\mathrm{x}) \;=\; \mathrm{H}(\mathrm{x}_{\mathrm{Gaussian}}) - \mathrm{H}(\mathrm{x}) \qquad (2.27)$$

Where $\mathrm{x}_{\mathrm{Gaussian}}$ is a Gaussian random variable with the same covariance matrix as that of a non-Gaussian variable, x, and $\mathrm{H(x)}$ denotes the entropy. Similar to the concept of entropy, negentropy is always positive, but equal to zero only when the variable has a Gaussian distribution.

However, the computation of negentropy is complicated and approximation approaches are used. One effective approximation approach is called FastICA, which can be described by:

$$\mathrm{N(V)} = \mathrm{E}\big(\emptyset(\mathrm{V})\big) - \mathrm{E}\big(\emptyset(\mathrm{U})\big)^2 \qquad (2.28)$$

Where V is a non-Gaussian random variable, U is a Gaussian random variable and $\emptyset(\cdot)$ denotes a non-quadratic function. A pre-processing process is required so that all variables are standardised. FastICA offers a computationally inexpensive way to extract independent components with non-Guassian or sub-Guassian distribution (Hyvärinen & Oja, 2000).

2.8 Multivariate Curve Resolution (MCR)

One typical criticism of PCA and other traditional algorithms is that the components extracted are essentially mathematical factors, which may or may not result in meaningful interpretation (Lachenmeier & Kessler, 2008). By contrast, Multivariate Curve Resolution (MCR) is a methodology that not only provides statistically significant results, but also offers practical importance to ToF-SIMS data analysis, especially for chemical and biological data (Wentzell et al., 2006; de Juan, Jaumot, & Tauler, 2014). It is capable of extracting the single properties of the chemical compounds of mixtures (the pure component spectra) and the concentration profiles with incomplete or even no knowledge of the components (de Juan & Tauler, 2006). This means that MCR can be used to process complex dataset or identify unknown chemical compounds.

MCR algorithms can be either non-iterative or iterative. Currently, iterative approaches have gained great popularity due to the ability to process multiset data structures and incorporate known information into the iterative process as constraints (de Juan, Jaumot, & Tauler, 2014). One of the most commonly used iterative MCR algorithms is MCR-ALS which uses alternating least squares (ALS) to solve the optimisation problem at each iteration. We will provide more detailed description of one novel MCR-ALS in Chapter 4.

Although the advantage of MCR is particularly attractive to biological applications, it might produce multiple solutions for the dataset due to intensity and rotational ambiguity (de Juan, Jaumot, & Tauler, 2014). Intensity ambiguity is derived from the indeterminate magnitude of the concentration profiles and pure spectra, leading to different interpretation of identical statistical results (Wise & Kowalski, 1995). However, it is normally easy to be detected and can be mitigated by normalising the concentration profiles or spectra produced, or incorporating known information into the approximation (Tauler, Kowalski, & Fleming, 1993; de Juan, Jaumot, & Tauler, 2014). Analysts are generally more concerned about rotational ambiguity, which is resulted from multiplying or dividing the components by a rotated matrix. Rotational ambiguity can be suppressed by incorporating constraints into the algorithm (Lachenmeier & Kessler, 2008). Common constraints include non-negativity, unimodality, closure, and stoichiometry, among which non-negativity constraint has been used most widely to offer realistic and meaningful results (Tyler, 2006).

Methodology

MCR was initially devised as a tool to study a single second-order data matrix that follows a bilinear structure. It involves a transformation procedure that decomposes the original data matrix into the product of two matrices where each matrix corresponds to an order of the original matrix (Tauler, Kowalski, & Fleming, 1993). The application of MCR has now been extended to multi-dimensional data analysis and more complex systems. However, this requires that the original dataset can be fairly described by a bilinear model. The bilinear model in MCR is described by:

$$\mathrm{X} = \mathrm{CS}^{\mathrm{T}} \tag{2.29}$$

Where X denotes the original dataset that we need to decompose, S^{T} is the pure spectra basis, and C is weighted matrix that indicates the contribution of each basis (concentration profiles). It should be noted that the chemical meaning of the two matrices can be altered to fit the nature of the original dataset. In real world, the original dataset X is always replaced by an estimation matrix $\widehat{\mathrm{X}}$ with error term E. This can be illustrated as:

$$\mathrm{X} = \widehat{\mathrm{X}} + \mathrm{E} = \mathrm{CS}^{\mathrm{T}} + \mathrm{E} \tag{2.30}$$

However, the decomposition of $\widehat{\mathrm{X}}$ can be unreliable without additional information about the concentration profiles. Because of the dynamics of MCR optimisation, various combinations of pure component spectra (S^{T}) and concentration profiles (C) with the similar appearance but different magnitudes may have the identical approximation of the raw data. This is so called intensity ambiguity which can be shown by an example:

$$\widehat{\mathrm{X}} = (\mathrm{Cr})\left(\mathrm{S}^{\mathrm{T}}\frac{1}{\mathrm{r}}\right) + \mathrm{E} = \mathrm{CS}^{\mathrm{T}} + \mathrm{E} \tag{2.31}$$

Where r is a constant number. In addition, if an arbitrary transformation matrix, P, is used in the optimisation problem, multiple possible combinations of C and S^{T} are available to represent the original dataset. This problem is referred to rotational ambiguity and can be shown by:

$$\widehat{\mathrm{X}} = (\mathrm{CP})(\mathrm{P}^{-1}\mathrm{S}^{\mathrm{T}}) = \mathrm{CS}^{\mathrm{T}} \tag{2.32}$$

The MCR algorithms can be realised by several popular methods, including:

1. Evolving Factor Analysis (EFA)

   EFA studies the evolving process of the single values on submatrices that are gradually introduced into the analysis. Therefore, any appearance of a new compound is attributed to the identification of a significant component. EFA can be performed in both top down and bottom up directions of the dataset, where forward EFA and backward EFA investigate the appearance and the disappearance of significant components, respectively (Maeder & Neuhold, 2007).

2. Window/Subwindow Factor Analysis (WFA/SFA)

   WFA is a chemometric tool developed based on EFA with ability to identify the concentration profiles of chemical species by studying the evolutionary process such as chromatography (Malinowski, 1992). It analyses the dataset using "window", which a region along the evolutionary axis where each component lies in. The window size for WFA is important since small window size can lead to indeterminate solutions and large window size may cause the inclusion of new components (Maeder & Neuhold, 2007). In addition, WFA is particularly vulnerable to the noise in the dataset, and a number of improved methods have been developed to address the problem (Chen et al., 2009).

3. Iterative Target Transformation Factor Analysis (ITTFA)

   ITTFA is a method that involves an iterative process to approximate composition profiles and pure component spectra. It is an extended Target Transformation Factor Analysis (TTFA) algorithm, which attempts to identify the components with real chemical meaning (Maeder & Neuhold, 2007). ITTFA generally requires PCA as an initialisation step to provide insight into the number of components and hence the initial estimated concentration profile. A target testing is used to examine whether the projected target and initial target are the same. The resulting data matrix is subsequently reconstructed using the components accepted as a result of target testing. However, the effectiveness of ITTFA is largely affected by the initial target employed (Zhu, Cheng, & Zhao, 2002).

4. Simple-to-use Self-modelling Mixture Analysis (SIMPLISMA)

SIMPLISMA was developed by Willem Windig based on a pure variable approach which resolves spectral mixture data in a user-friendly and time saving manner (Windig & Guilment, 1991). It assumes that there is a so-called pure variable that is significantly contributed by sole one of the pure components in the dataset (Windig et al., 2002). A pure variable can be identified by examining the purity value of variables, which is the ratio of the standard deviation to the mean. The pure variable approach is based on the thought that the intensity at a pure variable can be used as an estimate of a concentration profile when Beer's law is complied with (Windig & Stephenson, 1992). Therefore, the component spectra can be resolved through least-squares regression using the intensity, given that every component in the dataset has minimum of one pure variable. In the situation where the spectral data have many highly overlapping pure components, the pure variable approach based on second derivative spectra can be used to improve resolution of overlapped components (Windig et al., 2002). SIMPLISMA has an advantage of fast resolution since no iterative improvement process is required. Moreover, its interactive process enables the user to refine options at every step. However, the pure variables selected by SIMPLISMA are based on relative purity measure and may not necessarily be the true pure variables (Windig & Stephenson, 1992).

2.9 Summary

In this chapter we have reviewed several traditional unsupervised MVA techniques, which are able to decompose the raw dataset into key components and hence reconstruct the original data with less redundancy. In particular, clustering analysis is devised to categorise variables from a large dataset into distinct subsets. PCA, MAF, and ICA all aim to produce uncorrelated components but using different criteria, namely variance, autocorrelation, and statistical independence. MCR, on the other hand, attempts to provide the extracted components with chemical or physical meaning. All of these MVA techniques could possibly offer the benefit of information extraction to ToF-SIMS data analysis. In fact, the combination of PCA and ToF-SIMS has already been used in many metabolic profiling researches such as the biological molecules in cancer systems (Kotze et al., 2013). However, these MVA methods are also subject to a number of limitations specific to each of them, which may not be compatible to the ToF-SIMS applications in the context of metabolic profiling. In next chapter, we will take PCA as a particular example to demonstrate the application of traditional method to our ToF-SIMS data analysis.

Chapter 3

Principal Component Analysis

3.1 Introduction

ToF-SIMS data are complex even for a simple sample surface, within which the identity and distribution of different species needs to be extracted. Examples are sample composition, molecular orientation, surface order, chemical bonding and sample purity (Graham et al., 2006). The extraction of such information from the ToF-SIMS data is a challenging task. Feature extraction and dimension reduction techniques are of great importance as they can significantly simplify the analysis of complex ToF-SIMS datasets. The application of multivariate analysis techniques has opened new door for the exploration of ToF-SIMS data. In the previous chapter we have mentioned that several MVA techniques can provide promising results in reducing the complexity of ToF-SIMS data analysis. The most popular MVA technique used in this area is perhaps PCA due to its simplicity and easy implementation. The goal of PCA is to extract important information from data and transform this information using a set of orthogonal variables called principal components. In this chapter we will highlight the application of PCA in the analysis of ToF-SIMS dataset. The advantages and limitations of PCA are also discussed by the end of implementation.

3.2 Data Description

As mentioned in Chapter 1, a typical ToF-SIMS spectrum is represented by one three-way dataset, which essentially is the sum of all those secondary ions, including the fragment ions that make the spectrum complex and difficult to interpret. In reality, processing ToF-SIMS spectrum by MVA would require analysing many samples that are simply unavailable.

In this work, data contains measurements of three metabolites, tyrosine (T), phenylalanine (P) and citric acid (C). The chemically pure metabolites were spotted on hexamethyldisilazane (HMDS) and coated silicon wafers. The dataset were exported from the SIMS V instrument (ION-ToF Inc., Germany). Five ToF-SIMS experimental samples were obtained, which contain three individual pure species (T, P and C) and two mixed species (TC and TPC). For each sample, three replicate datasets are available. Each one of those datasets includes images of 128×128 pixels with the spectra up to 200 Da, while only 100 m/z intensities were considered in the analysis of all samples. This is based on the consideration of the computational cost, and all the deprotonated metabolites ions are included. Discriminatory features are first extracted from the estimated scores and loadings by applying the algorithm to the three pure species. Subsequently, the extracted information is used to perform peak assignment in the spectra of TC and TPC mixtures. This way discriminatory information can then be summarised into major peaks, which in turn would have the ability to identify the corresponding species. In order to assess the performance of the algorithm, the extracted spectral information is also utilised in analysing the replicate measurements of each dataset. Figure 3.1 shows the total ion images for each dataset.

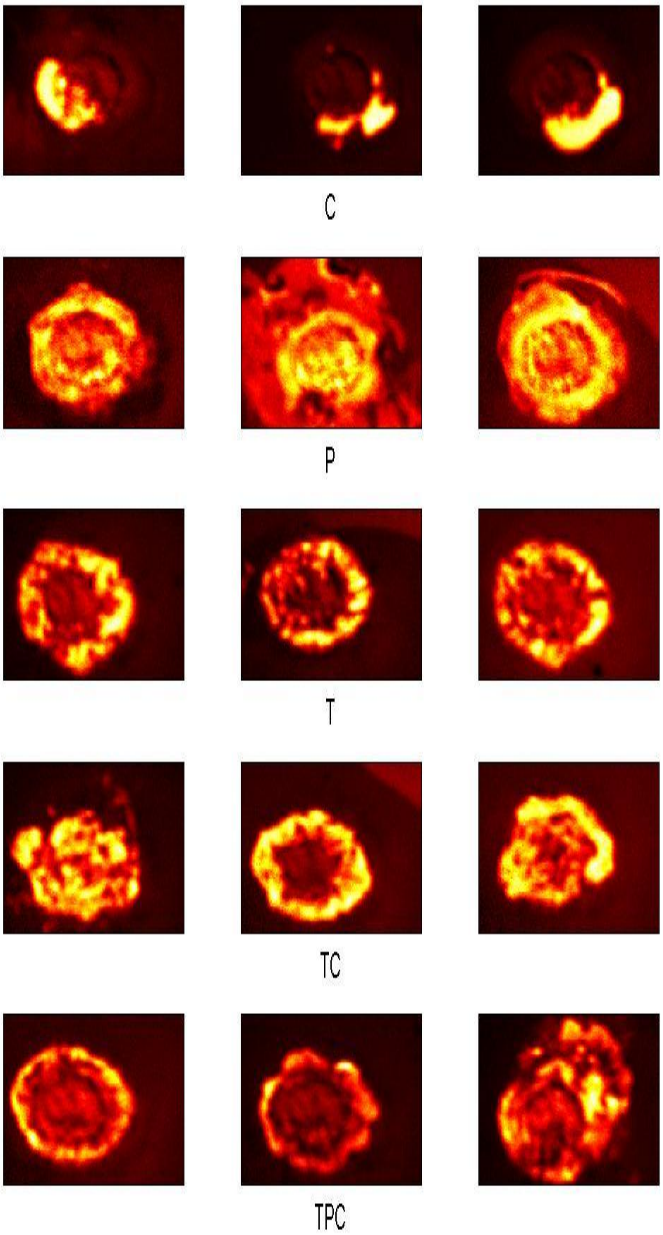


Figure 3.1 Total ion images. Total ion images for three measurements of the pure species (C, P, and T) and the mixtures (TC and TPC).

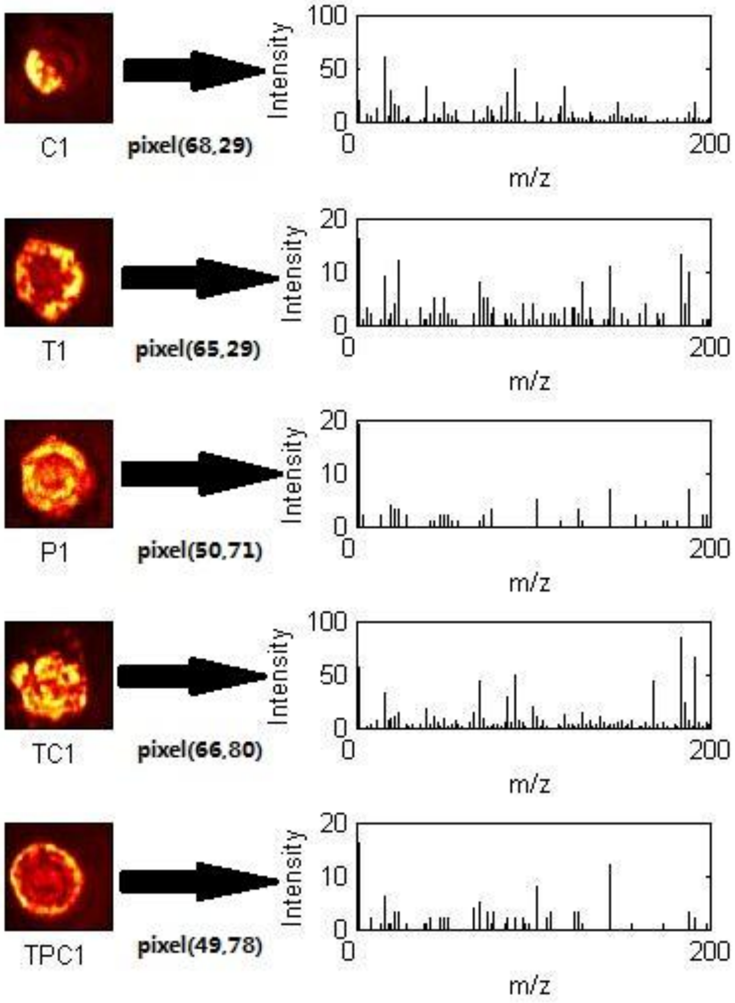


Figure 3.2 Mass spectral data. For each sample dataset, one mass spectral plot can be created at every pixel point.

3.3 Principal Component Analysis

We have already outlined the general theory of PCA in Section 2.5, here we look into its methodology in greater details. PCA attempts to find the linear combination through orthogonal transformation procedure which decorrelates original variables into a number of principal components (Hotelling, 1933). The main principle of the algorithm is presented as follows:

Suppose the original data matrix X of a dimension $m \times n$, with $x_i (i = 1,2,3 \dots m)$ as the row vectors:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_m \end{bmatrix} \tag{3.1}$$

PCA performs transformation by investigating the covariance matrix of the original data, C_x, defined by the following expression:

$$C_x = \frac{1}{n-1} XX^T \tag{3.2}$$

Where C_x is a square symmetric $m \times m$ matrix. The elements of C_x represent the degree of variations among the vectors, which is called correlation. The off-diagonal elements are the covariance of pairs of vectors while the diagonal elements are the variances of vectors themselves. The covariance of the original variables is used to evaluate the level of redundancy or noise in the dataset. Large values can be interpreted as unsound, since the variables tend to be highly interrelated.

The transformation procedure of PCA intends to reduce the covariance of variables to a minimum level that is ideally equal to zero. This means that the original covariance matrix C_x must be somehow transformed into a new covariance matrix C_y, of which off-diagonal entries are all zeros. The transformation procedure used by PCA is to find the linear combinations of the original data matrix X, such that $Y = PX$. Then P can be substituted into the matrix C_y:

$$\begin{aligned} C_y &= \frac{1}{n-1} YY^T \\ &= \frac{1}{n-1} (PX)(XP)^T \\ &= \frac{1}{n-1} PXX^T P^T \\ &= \frac{1}{n-1} P(XX^T)P^T \end{aligned} \tag{3.3}$$

Now we can define a $m \times m$ symmetric matrix, A, such that $A = XX^T$. Equation

(3.3) can hence be rewritten in terms of A:

$$\mathrm{C_y} = \frac{1}{\mathrm{n}-1}\mathrm{PAP^T} \tag{3.4}$$

Therefore, for every symmetric matrix, there is a diagonal matrix $\mathrm{C_y}$ which is comprised of the set of all eigenvalues of $\mathrm{C_x}$ along its main diagonal and zeros elsewhere. The two matrices as defined by the following relationship:

$$\mathrm{A} = \mathrm{EDE^T} \tag{3.5}$$

Where D is the eigenvalue matrix and E is the eigenvectors. Choosing a matrix P that is defined by:

$$\mathrm{P} = \mathrm{E^T} \tag{3.6}$$

And substituting into Equation (3.4) gives:

$$\begin{aligned}\mathrm{C_y} &= \frac{1}{\mathrm{n}-1}\mathrm{PAP^T}\\ &= \frac{1}{\mathrm{n}-1}\mathrm{P(EDE^T)P^T}\\ &= \frac{1}{\mathrm{n}-1}\mathrm{P(P^TDP)P^T}\\ &= \frac{1}{\mathrm{n}-1}\mathrm{(PP^T)D(PP^T)}\\ &= \frac{1}{\mathrm{n}-1}\mathrm{D}\end{aligned} \tag{3.7}$$

As shown above, this results into a new diagonal covariance matrix of all eigenvalues of the original covariance matrix, which eliminates the linear correlation amongst new variables and therefore the redundant information. However, some additional steps are still required as the diagonal elements of the covariance matrix still represent the variances of each variable in the data. Therefore the next step is to transform the matrix such that these variations become more apparent. This essentially means maximising each variance element. PCA selects the variable with the largest variance in Y along with normalised direction in the m-dimensional space P as the first principal component, which is presumed having the greatest explanatory power to the data. This process is repeated until all the directions have been selected once, and subsequently, the vectors in matrix P are ordered in a descending manner from the first principal component to the m^{th} principal component.

3.4 Random Sampling

Although PCA has been widely approved as an effective method to reduce the dimensionality of data, it requires huge amount of computational resources, especially when the original dataset is substantially large. Our experiments with a $128 \times 128 \times 1000$ dataset in the past research using a 64-bit processor computer with 4GB memory could not provide enough resources for the implementation, not to mention the long execution time. This problem is managed by using simple random sampling, which is a basic equal probability of selection method (EPSEM) where each statistical unit of the sample has an equal chance of being selected (Peters & Eachus, 1995). Because the statistical units are randomly selected in a sample, the information provided can be interpreted as an unbiased estimator of the data and used in the application of PCA with much lower computation required.

A simple random sampling can be performed either with replacement or without replacement (Antal & Tillé, 2011). However, random sampling without replacement is generally preferred since sampling the same object more than once would provide no further information (Lohr, 2009). In our thesis, a random sampling without replacement is implemented via MATLAB to ensure representative and unbiased sampling results, which can then be used in the generalisation back to the population (Wong, 1999).

3.5 Poisson Scaling

Data scaling is essential for the effectiveness of MVA in ToF-SIMS data analysis, since the noise in ToF-SIMS data is not uniform as assumed in many conventional

MVA applications (Keenan, Kotula, 2004). In fact, the noise in many ToF-SIMS data by its nature follows a Poisson distribution, where the variance of the data is fairly close to the mean of the data (Henderson, 2013). Consequently, the standard deviation of the data is roughly equal to the square root of the mean. This means that for Poisson-distributed data, the results of MVA could be largely affected by the high intensity and low mass peaks due to higher mean and variance (Lee et al., 2009). Therefore, ToF-SIMS data is generally pre-processed using Poisson Scaling, which accounts for the Poisson noise distribution by dividing each peak by the square root of the peak intensity and by the square root of the mean (Tyler, Rayal, & Castner, 2007; Henderson, 2013). This can be described by:

$$\widetilde{\mathrm{X}} = \mathrm{G}^{-\frac{1}{2}}\mathrm{X}\mathrm{H}^{-\frac{1}{2}} \tag{3.8}$$

Where the scaled data matrix $\widetilde{\mathrm{X}}$ is obtained by dividing the original data X by two scaling matrices G and H, which are the diagonal matrices with the row means and column means of X along the diagonal, respectively. The objective of Poisson scaling is to normalise the non-uniform noise with Poisson distribution, and therefore, the variation in the data purely reflects the chemical concentration and discriminatory pattern (Smentkowski, Ostrowski, & Keenan, 2009; Henderson, 2013).

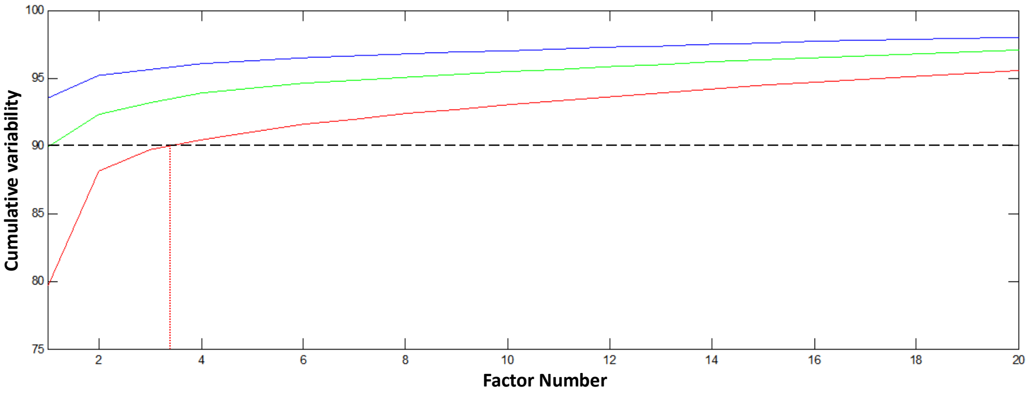
3.6 Application Results

In this section we demonstrate how those previously mentioned MVA algorithms can be applied to ToF-SIMS dataset by using PCA as an illustrative example. The implementation was carried out through the MATLAB using SVD approach. It is important to note that PCA algorithm is a 2-dimensional method (as well as other algorithms discussed in this thesis), while our ToF-SIMS dataset is three dimensional, which is $128 \times 128 \times 100$. However, according to the prior knowledge about the ToF-SIMS images, the first 2 dimensions only reflect the original positions of different components. Since we attempt to find the principal components that are able to represent the original images without 'redundancy', the position of each component is not a major concern because it can be reflected back to the raw image after the analysis. In this case, the original dataset was firstly reshaped into 2-dimensional dataset by combining the first 2 dimensions, which results in a dataset of dimension of 16384×100. Another issue we need to raise is that although only 100 m/z points are provided in the dataset, the results are still presented throughout 200 m/z axis by mapping the points to the actual locations. The same procedure will be utilised before the implementation of other algorithms in this thesis. Poisson scaling procedure also needs to be performed before PCA, the reason is that the PCA is one scaling-dependent algorithm, and the topography of the dataset along with a non-uniform exposure and differential extraction of the secondary ions might cause unclear principal components segregation and incorrect selection.

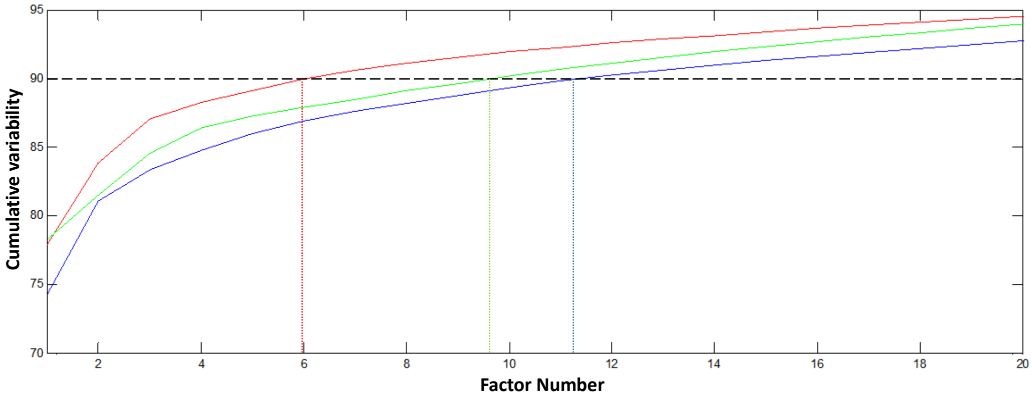
Figure 3.2 shows the cumulative percentage of representation provided by the first 20 principal components for the each of the five samples. The three different colour plots in each image represent the three replicated datasets for each species. There are several approaches to select a proper number of PCs for the PCA application, the most popular ones are (Valle et al., 1999):

1. Akaike information criterion: it provides a measurement of the quality for each model of the dataset by the estimation of the information lost (Akaike,1998).
2. Minimum description length: it gives a good hypothesis of the data by finding a best compression of the original dataset (Grünwald, 2007).
3. Imbedded error function: it is a function of error eigenvalues and can identify the error between models (Brereton, 1992).

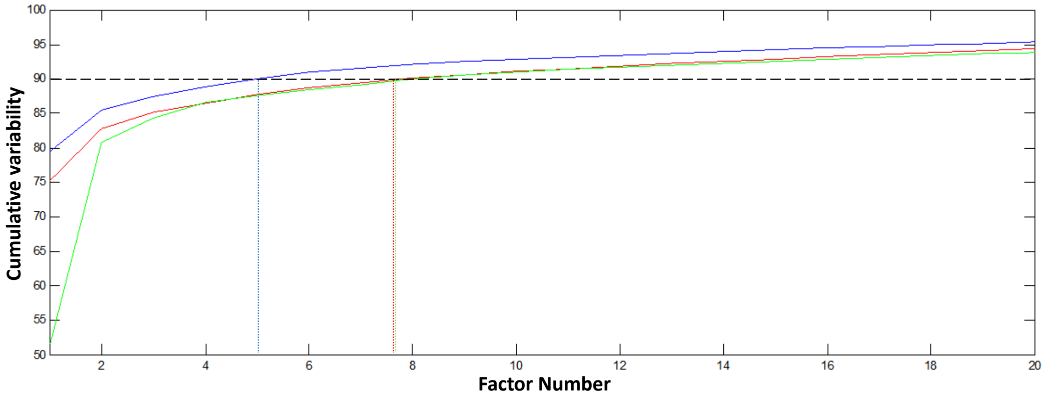
We will also provide one model selection method in the latter chapter to solve this problem, so here in order to reduce the computational cost, we use 90% approximation as the target which is commonly used in the PCA implementation of spectral data. It is clearly shown that for the three pure species, at least 3 principal components (PCs) are required to be able to provide 90% approximation of the raw data, which is the lowest acceptable degree of data representation. In particular, for the single component T and P, it is strongly recommended that 4 or 5 PCs are needed for an informative approximation (Figure 3.2 (T) and Figure 3.2 (P)). This result is different from the expectation based on the data dynamics, since there is only one component in the dataset of T and P species. The possible explanation is that there might be several fragment ions and noise during the ion flying process, the noise are mostly the measurement errors of the ToF-SIMS instrument resulted from the vibration and the support system. The existence of fragment ions and noise significantly hinders the detection of the metabolites signals. It also appears that one PC is sufficient to represent 90% of the data for TC mixture while TC mixture in fact contains two components (T and C) (Figure 3.2 (TC)). The reason is unable to be identified at this point, it might be due to the similarity (similar range of identical peak location and similar fragmentations appearance) between T and C species. This should be interpreted later with the loadings result. Among all the analysis results, only the result from the three component mixture is exactly correct as 3 PCs are required for a reliable representation (Figure 3.2 (TPC)). These results can be considered as reference indicators for the later analyses when choosing the number of chemicals prior to the commencement of the algorithm.


100
95
90
85
80
75
Cumulative variability
2
4
6
8
10
12
14
16
18
20
Factor Number

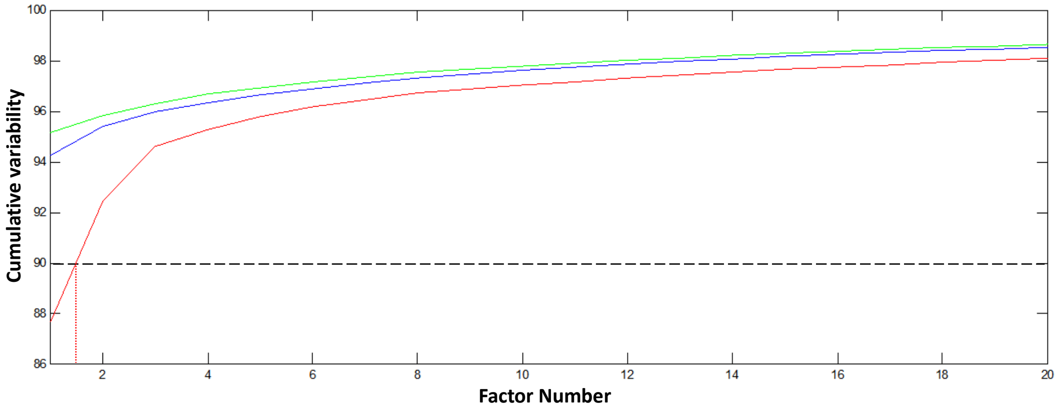
(C)


95
90
85
80
75
70
Cumulative variability
2
4
6
8
10
12
14
16
18
20
Factor Number

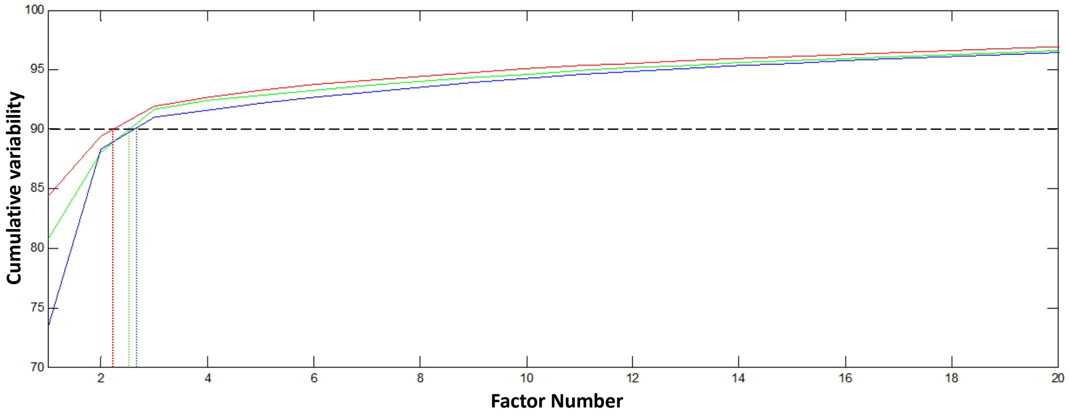
(T)


100
95
90
85
80
75
70
65
60
55
50
Cumulative variability
2
4
6
8
10
12
14
16
18
20
Factor Number

(P)



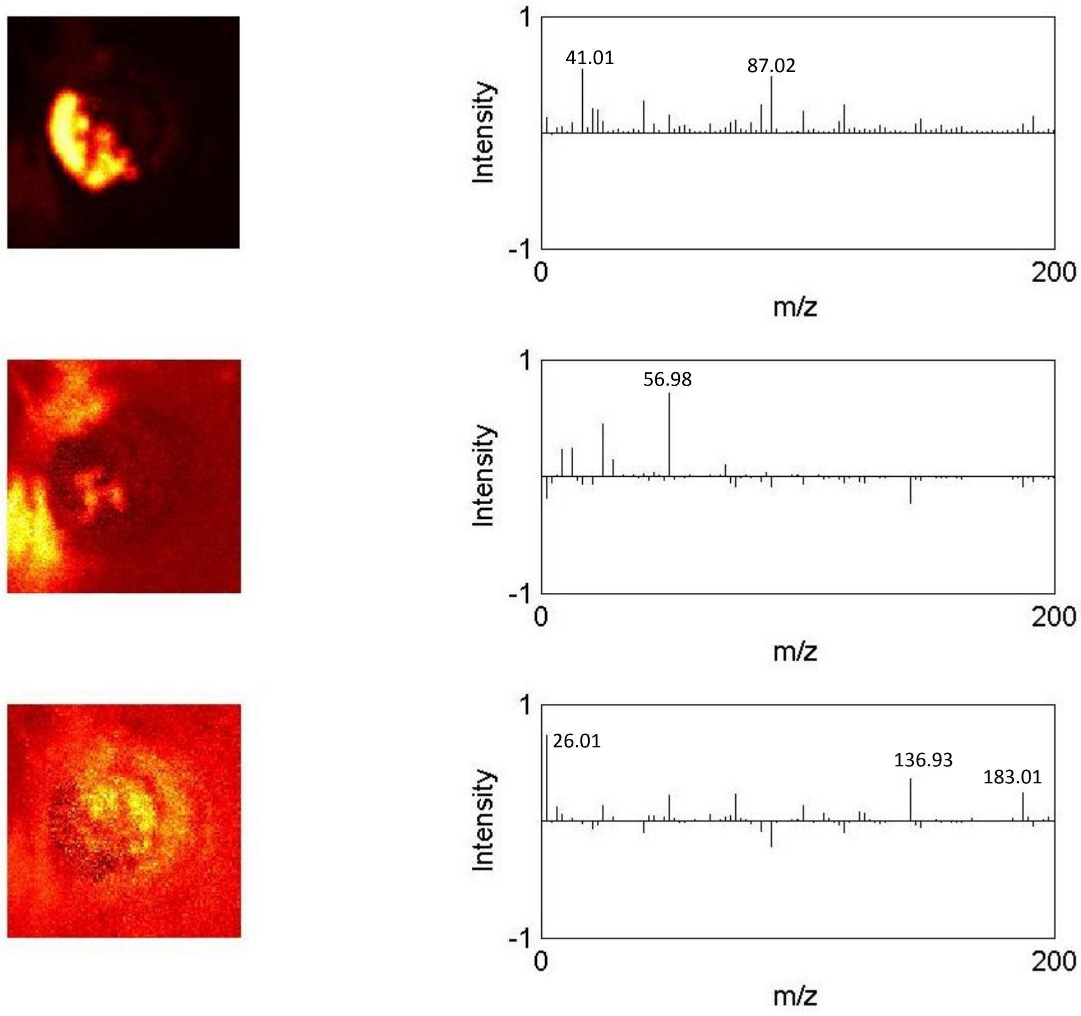
(TC)



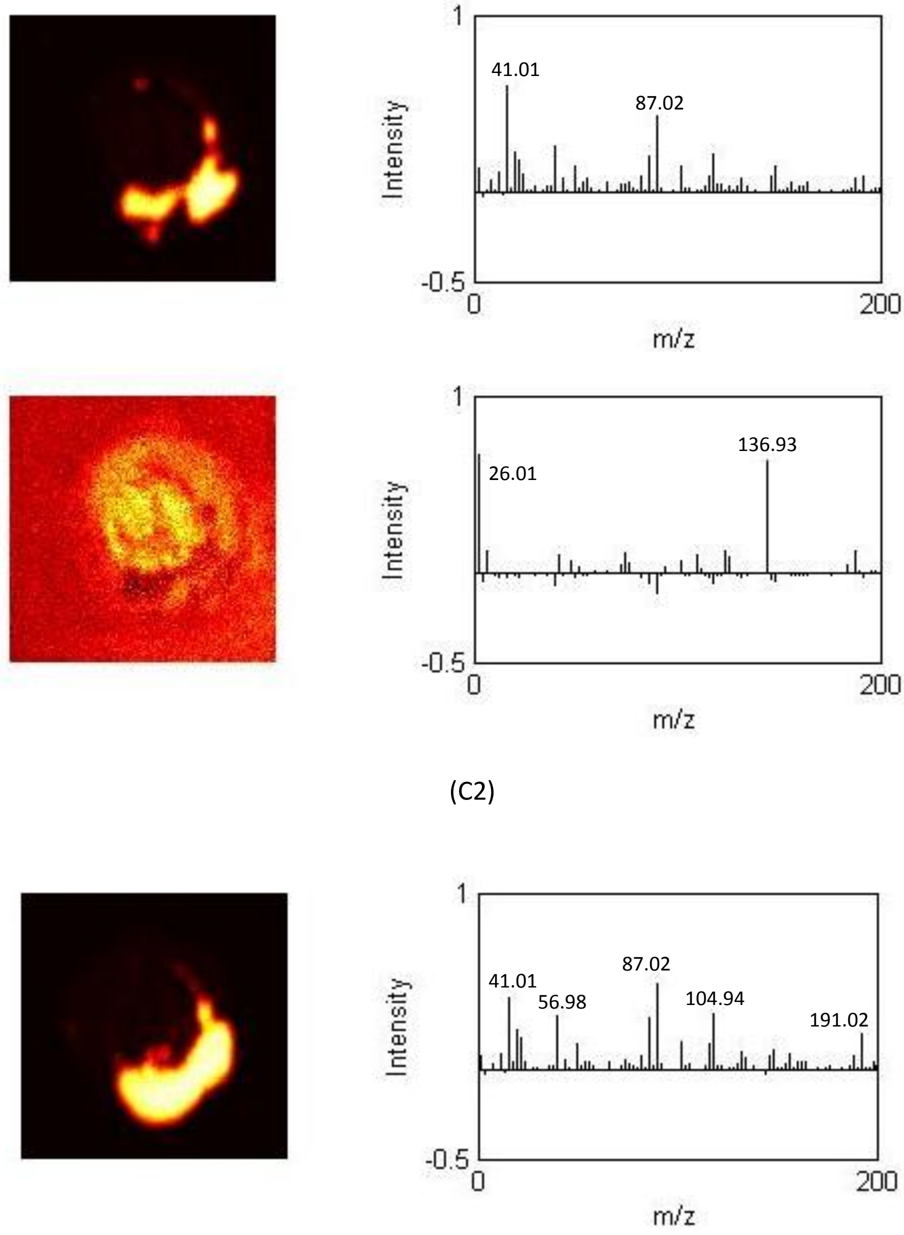
(TPC)

Figure 3.3 PCA scree plot. The normalised scree plots for individual components C, T, P, TC and TPC mixtures for three replicates are shown. Each of three replicate samples is plotted in three different colours. Only the first twenty principal components are presented. In each plot dashed lines show 90% cumulative variability, indicating the number of factors required to approximate at least 90% characteristics of the original samples.

The information derived from the scree plot in Figure 3.2 can be used to evaluate the performance of PCA by applying to three replicate samples of component C. for a better illustration, three PCs, two PCs and one PC are used to approximate sample C1, C2 and C3 respectively. Figure 3.3 illustrates the scores and loading plots of the chosen PCs from C1, C2 and C3 replicate samples. The score images seem very promising for each implementation, while the loadings spectral images plotted alongside are rather unsatisfactory with many negative coefficients that are more difficult to interpret in reality. However, it can still provide information of the original data by transforming the normalised data back to the original one. It is widely used in the applicable research for its simplicity (Biesinger et al., 2002). The Y axis in the loading plots in Figure 3.4 is the signal intensity value of each ions showing at each locations through the m/z axis. The value range is different due to different instrument, which can be calculated by Equation (2.1). In this thesis the value range is from 0 ~ 1 by dividing the real value with the largest intensity and map the loading into the same axis scales from 0 to 1.



(C1)



(C2)

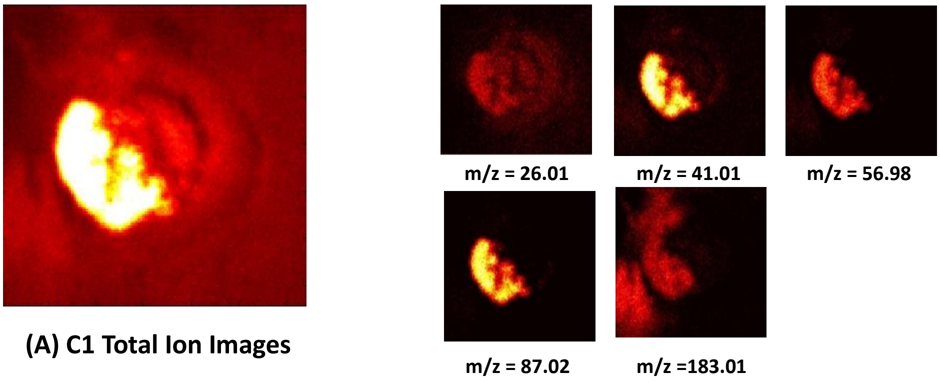
(C3)

Figure 3.4 Scores and loading plots produced using PCA for C1, C2 and C3 samples. Score images are presented on the left showing the spatial information of each PC and loadings are presented on the right indicating the intensity of PCs.

The significant peaks appeared in each PC for three replicate samples are organised in Table 3.1. Intuitively, one specific chemical should have a particular peak location region; therefore three pure samples should contain the same significant peak location as they are all derived from the same pure species. The comparison of the significant peak locations of the three replicate samples indicates that, components can be identified with peaks at m/z = 41.01, m/z = 56.98 and m/z = 87.02 while other spectral locations are most likely due to the chemical fragments and noise. This result is reasonably acceptable as all the identical significant peaks are derived from the first PC without negative values. This can also be identified in Figure 3.3. The corresponding spectra images from the original dataset at each significant peak are illustrated in the Figure 3.4. From these figures the corresponding scores images of peaks at m/z = 26.01, m/z = 136.93 and m/z = 183.01, it can be seen that they all have noisy structures, therefore, these peaks do not have discriminatory information and can be considered noise in the system.

Replicate Sample	Significant Peak Location (m/z)
C1	26.01, 41.01, 56.98, 87.02, 183.01
C2	26.01, 41.01, 56.98, 87.02, 136.93
C3	41.01, 56.98, 87.02, 104.94,191.02

Table 3.1 Significant peaks identified from PC loadings for C1, C2 and C3 pure species samples. The peak locations are displayed in an ascending order.



(A) C1 Total Ion Images

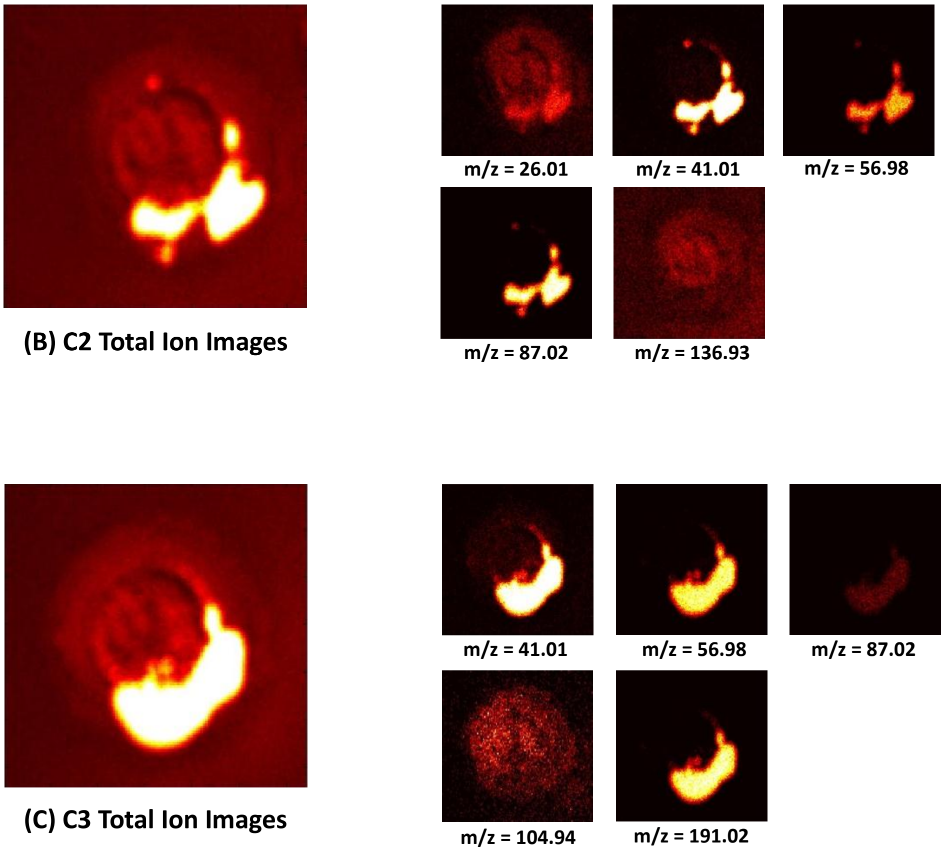
m/z = 26.01

m/z = 41.01

m/z = 56.98

m/z = 87.02

m/z =183.01



(B) C2 Total Ion Images

(C) C3 Total Ion Images

Figure 3.5 Corresponding Score images at each significant peak from the original data for C1, C2 and C3 species samples. (A-C) are the total ion images of C1, C2 and C3 samples. The corresponding scores images at significant peaks are also shown.

From the result above, it can be summarised that, the first PC in each case only disturbed by a small amount of negative values which is useful for the following interpreting work, however, it also can be seen that species C is fragmented highly through the spectrometry process, which leads to the comparing low intensity on the identical peak location m/z = 191.02 (Figure 3.5) and increase the difficulties for the following research. From the spatial aspect, peaks at m/z = 41.01, m/z = 56.98 and m/z = 87.02 are in the area most similar to the identical peak m/z = 191.02, and those three are highly recommended as the fragmentations from species C during the spectrometry ion flying process.

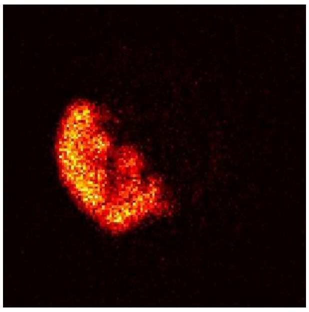


Figure 3.6 Species C1 score image for peak at m/z = 191.02

Similarly, the significant peaks of T1 and P1 samples can be identified using PCA with the chosen PC number m = 1 as we have the prior information that they are both pure species samples. The score images and loading plots for T1 and P1 are shown in Figure 3.6 with the significant peaks and the corresponding score images illustrated in Table 3.2.

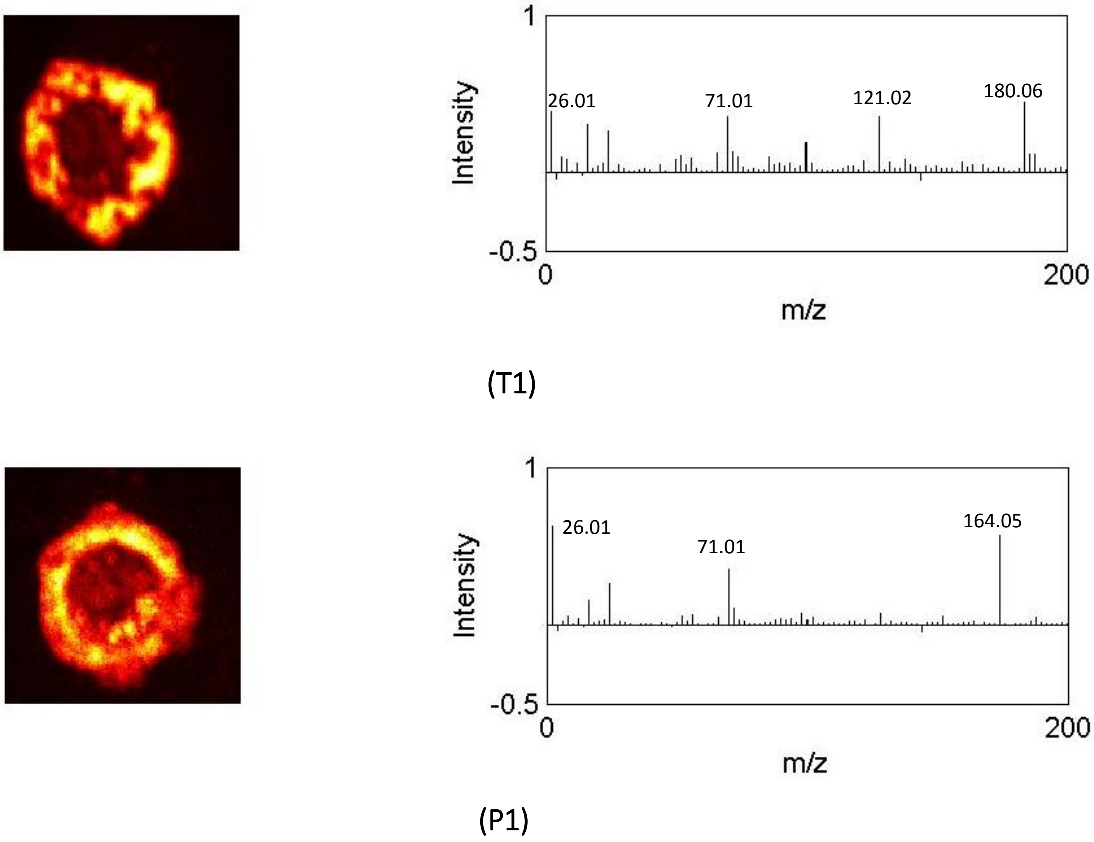


Figure 3.7 Scores and loading plots produced using PCA for T1 and P1 species samples.

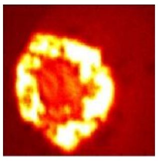
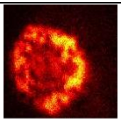
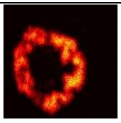
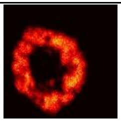
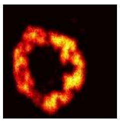
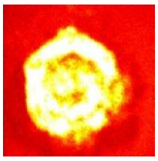
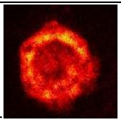
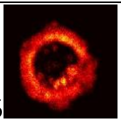
Pure Sample	Significant Peak Location (m/z)
T1 	26.01  71.01  121.02  180.06 
P1 	26.01  164.05 

Table 3.2 Significant peaks identified from PC loadings produced using PCA for pure T1 and P1 samples. The locations are presented in an ascending order. Total ion images and corresponding score images are given alongside.

After applying PCA to the three pure species, the significant peaks are separately summarised in Table 3.3. This information is in turn used to identify individual component by reviewing the peaks specifically attributed to it.

component	Peak location value (m/z)
C	56.98, 87.02, 191.02
T	73.01, 180.06
P	164.05

Table 3.3 Identified m/z values for peak assignment. All of the peaks can be used to identify specific individual chemical compounds while the numbers in red are the given ground truth for each chemicals. They can be used as references for the later identification of different species throughout this thesis.

As we discussed previously, the result of PCA performed on TC mixture suggests that only one PC is sufficient for representing 90% information of the original dataset. However, according to the dynamics of the dataset, we know that TC mixture is a mixed species that consists of two components. Under this circumstance, two PCs are used in order to test the ability of the PCA algorithm given “correct” number of PCs is consistent with the number of chemicals in the mixture. Figure 3.6 represents the scores and loading plots for the two PCs of TC1 sample. For the first PC (shown in the upper row in Figure 3.7), a significant peak at m/z = 191.02 can be attributed to component C when analysing the results on replicate samples for component C. However, the peaks at m/z = 71.01 and m/z = 180.06 can be identified as chemical T as they match the results in Table 3.3. Similarly, peak at m/z = 87.02 is referred to component C. By contrast, only one noise peak at m/z = 26.01 can be found for the second PC as illustrated in the lower row in Figure 3.7. This means, interestingly, that both component T and C can be identified solely using the first PC, which is in agreement with the result obtained from the scree plots where only one PC is needed to reasonably approximate TC mixture. This result shows that the important peaks can be identified, however, the distribution of elements cannot be separated. The information of the two PCs generated is summarised in Table 3.4.

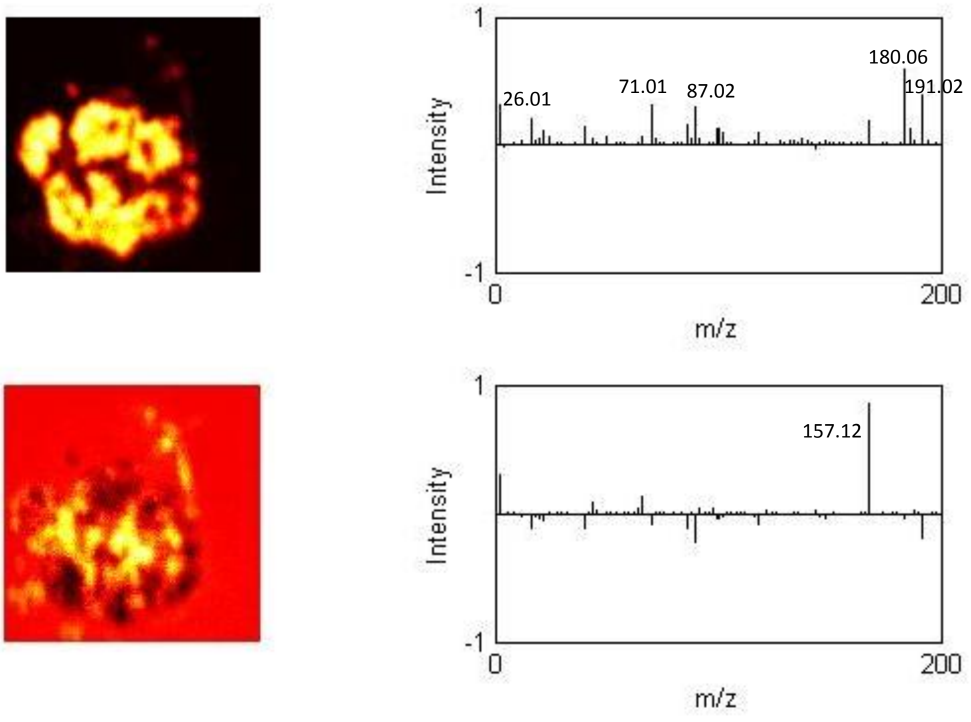


Figure 3.8 Scores images and loading plots produced using PCA for TC1 species sample. First PC and second PC are shown in the upper and lower row, respectively.

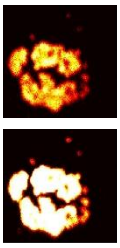
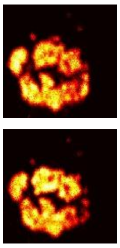
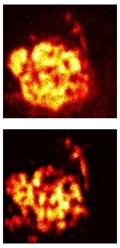
Mixture Sample TC1	Significant Peaks corresponding to Chemical T	Significant Peaks corresponding to Chemical P	Significant Peaks corresponding to Chemical C	Uncertain Peaks
Component 1	71.01 180.06	N/A	87.02 191.02	26.01
Component 2	N/A	N/A	N/A	26.01 157.12
Corresponding Score Images				

Table 3.4 Significant peaks identified from PC loadings produced using PCA for TC1 mixture samples.

From the scree plot of TPC1, three PCs are required to capture the information from the original dataset which corresponds to the correct number of chemicals in the mixture. The result of PCA performing on TPC1 mixture is presented in Figure 3.8, showing the scores images and the loadings for each of the three PCs. It can be seen from the first PC there exist two significant peaks at m/z = 167.024 and m/z = 181.06, which are differentiating peaks of the chemical P and chemical T respectively. These two peaks can be recognised as chemical P and chemical T respectively using the information in Table 3.3. Furthermore, there are also two large peaks appeared at m/z = 191.02 and m/z = 87.02 in the loadings image of the first PC, which can be derived from chemical C. For the second PC (shown in the middle row in Figure 3.8), there is one interesting peak at m/z = 100.02 close to the noise peak at m/z = 26.01. However, we are unable to identify the peak as there is no prior knowledge in relation to chemicals located at that region. As it can be seen in the bottom row in Figure 3.8, one peak at m/z = 87.02 is referred to pure species C for the third PC. The corresponding score images are given in Table 3.5.

The results of PCA application suggests that component C can be identified and distinguished from other chemicals in the third PC while component T and P are still mixed. This is similar to the problem encountered in result of the TC mixture. It should be noted that there is also a strong peak at m/z = 100.019 which only appears in PC2. This could be due to the fragment ion of the process. The performance of PCA on TPC mixture suggests that the detection of different species and segmentation is partly accomplished since still mixed up peaks in one PC.

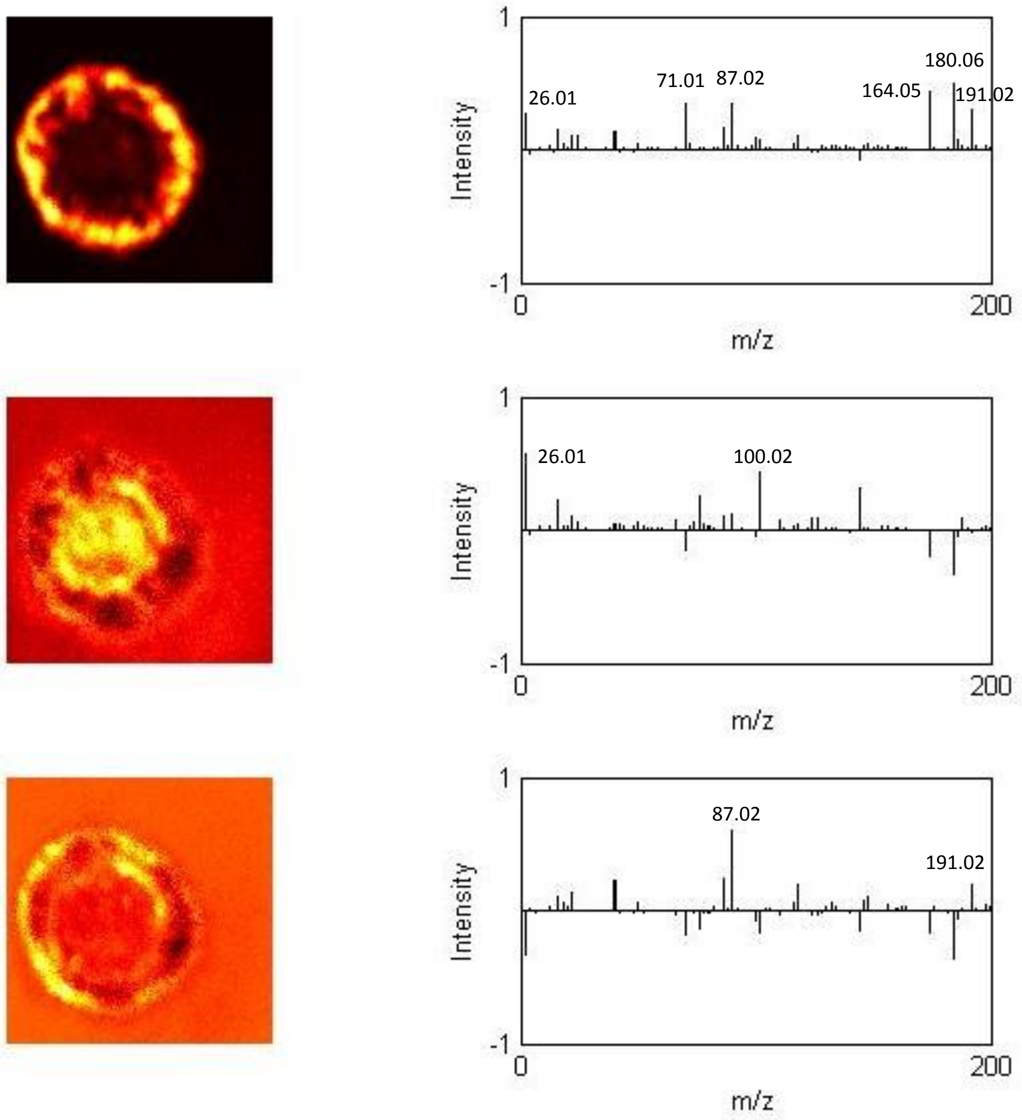


Figure 3.9 Score images and Loading plots produced using PCA for TPC1 mixture samples.

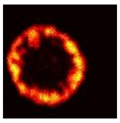
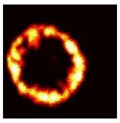
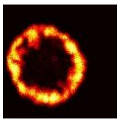
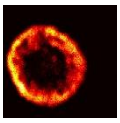
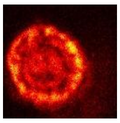
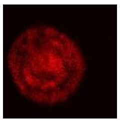
Mixture Sample TPC1	Significant Peaks corresponding to Chemical T	Significant Peaks corresponding to Chemical P	Significant Peaks corresponding to Chemical C	Uncertain Peaks
Component 1	71.01, 180.06	164.05	87.02 191.02	26.01
Component 2	N/A	N/A	N/A	26.01, 100.02
Component 3	N/A	N/A	87.02 191.02	N/A
Corresponding Score Images	 			 

Table 3.5 Significant peaks identified from PC loadings produced using PCA for TPC1 mixture species samples. The peak locations are displayed in an ascending order.

As we mentioned in Section 3.4, random sampling can be used prior to the implementation of PCA in order to reduce the computational demand. The original dataset is randomly sampled, where all spectra at each pixel point have the same probability to be selected into a number of subsamples. These randomly selected spectra constitute a subset of the original dataset with significantly lower data points which therefore reduces the computational complexity of the algorithm.

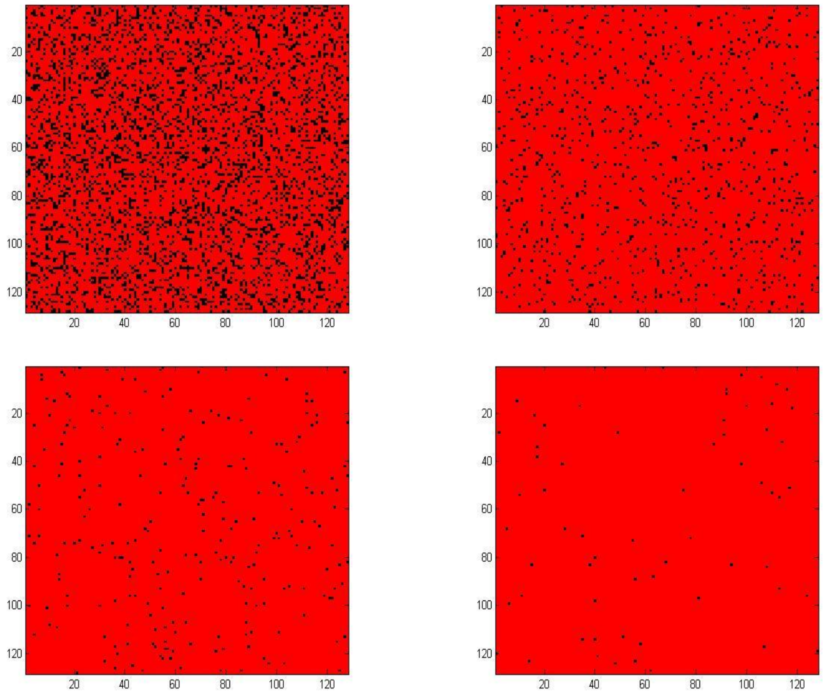


Figure 3.10 Sampled datasets from the original dataset. Sample size of 4096, 1024, 256, and 64 (shown as the black points in the images) have been randomly selected from the original dataset, from upper left to bottom right.

To examine the performance of the algorithm, we randomly sample our replicate dataset for TPC1 mixture with 4 sets of experiments where each set randomly selects 50 trials of samples with the sample sizes of 4,096, 1,024, 256 and 64. The first order error can be considered as a way of comparing the reliability of each sample result. The 'error' of the calculated PCs ranges from 0.004 to 0.4 with sample size changing from 4,096 to 64. The 'error' of the experiment with a sample size of 64 is about 100 times larger than that with a sample size of 4,096. This is intuitive as the PCA results become more reliable when size of sample increases.

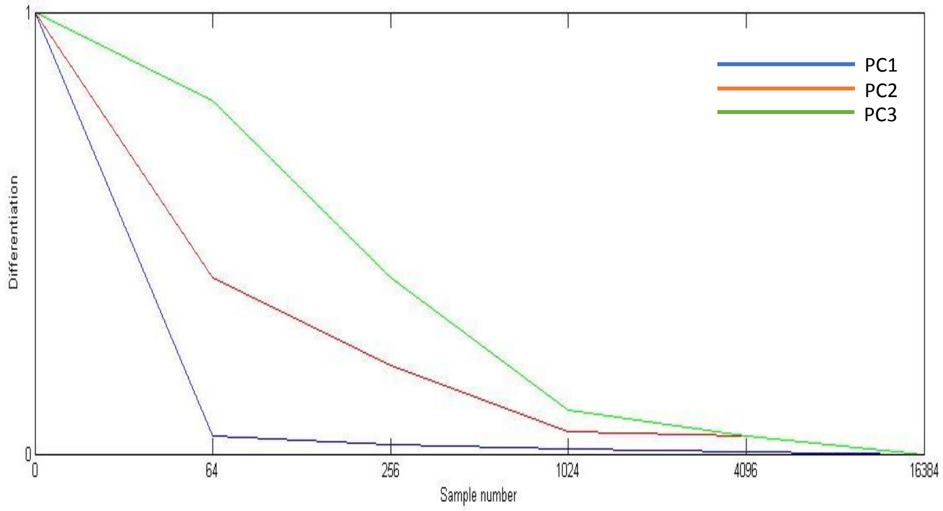
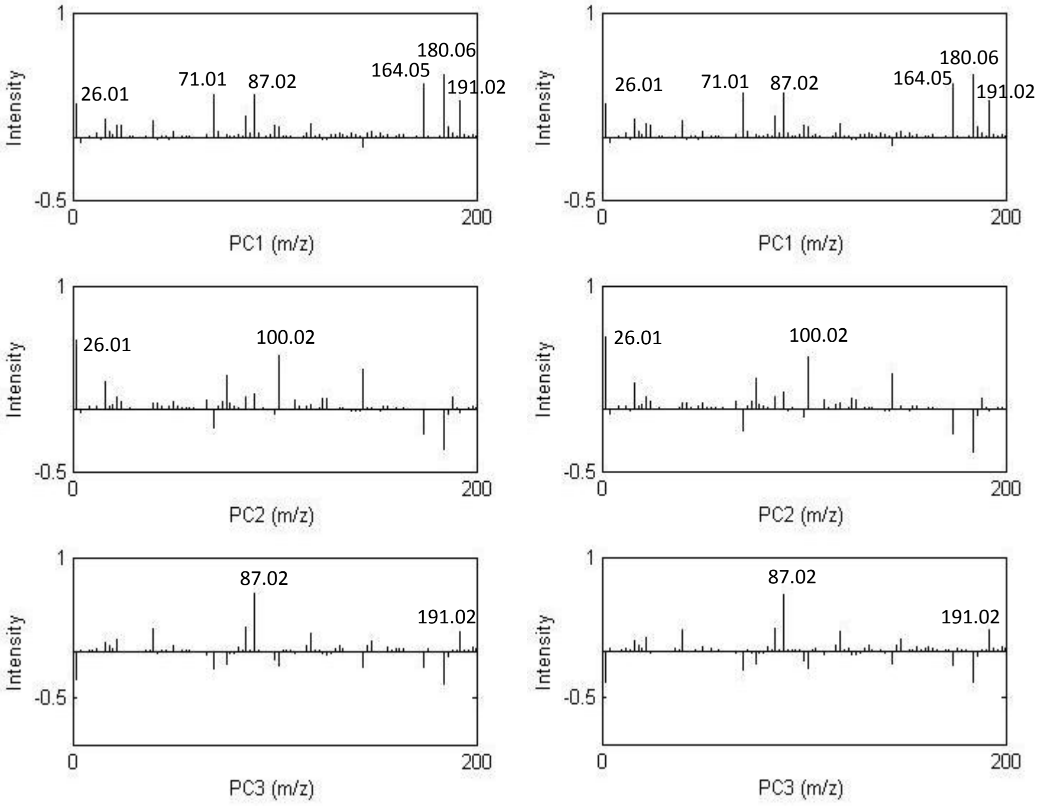


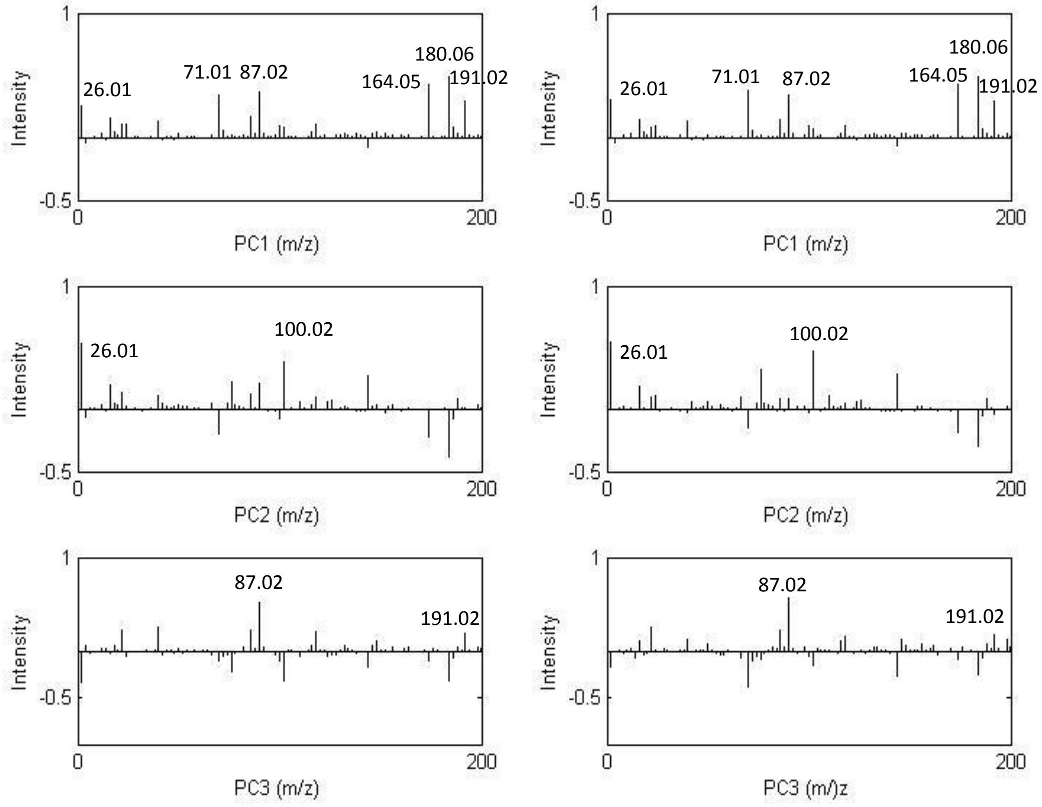
Figure 3.11 Difference ratio between original dataset and different random dataset. Errors between the results of original dataset and sampled dataset for PC1 (blue curve), PC2 (red curve) and PC3 (green curve) are given separately.

The loading plots in Figure 3.11 represent one out of the 50 trials from each sample number experiment. Similar to previous implementation, component T, P and C can be found in all the first PCs for the four different sample sizes with significant peaks at m/z = 181.06, m/z = 164.05 and m/z = 191.02 respectively. Component C also can be identified in all the third PCs with the significant peak at m/z = 87.02, which also appear in the first PCs but with low intensity. The second PCs all contain one noise peak and one uncertain peak. Therefore, the results of the application of PCA to these sampled datasets are approximately the same as the results for the original dataset, but with significant time saving and lower computational demand. This suggests that random sampling can be an effective pre-processing technique for PCA application when the dataset is large and complex.



(4096 samples)

(1024 samples)



(256 samples) (64 samples)

Figure 3.12 Loading plots produced using PCA for TPC1 sampled datasets. Three PCs are ordered in descending order according to the degree of importance.

In order to get a better visualisation, we simply average all the results of the 50 trials from the sample with largest sample size (4,096) and use the three average PCs as the new projections. The projections are in turn used to generate scores for the original data. The results in Figure 3.12 illustrate that the spatial score images are roughly the same with similar PCs. Therefore, the combination of PCA and random sampling offers a relatively simple way to achieve proper solutions.

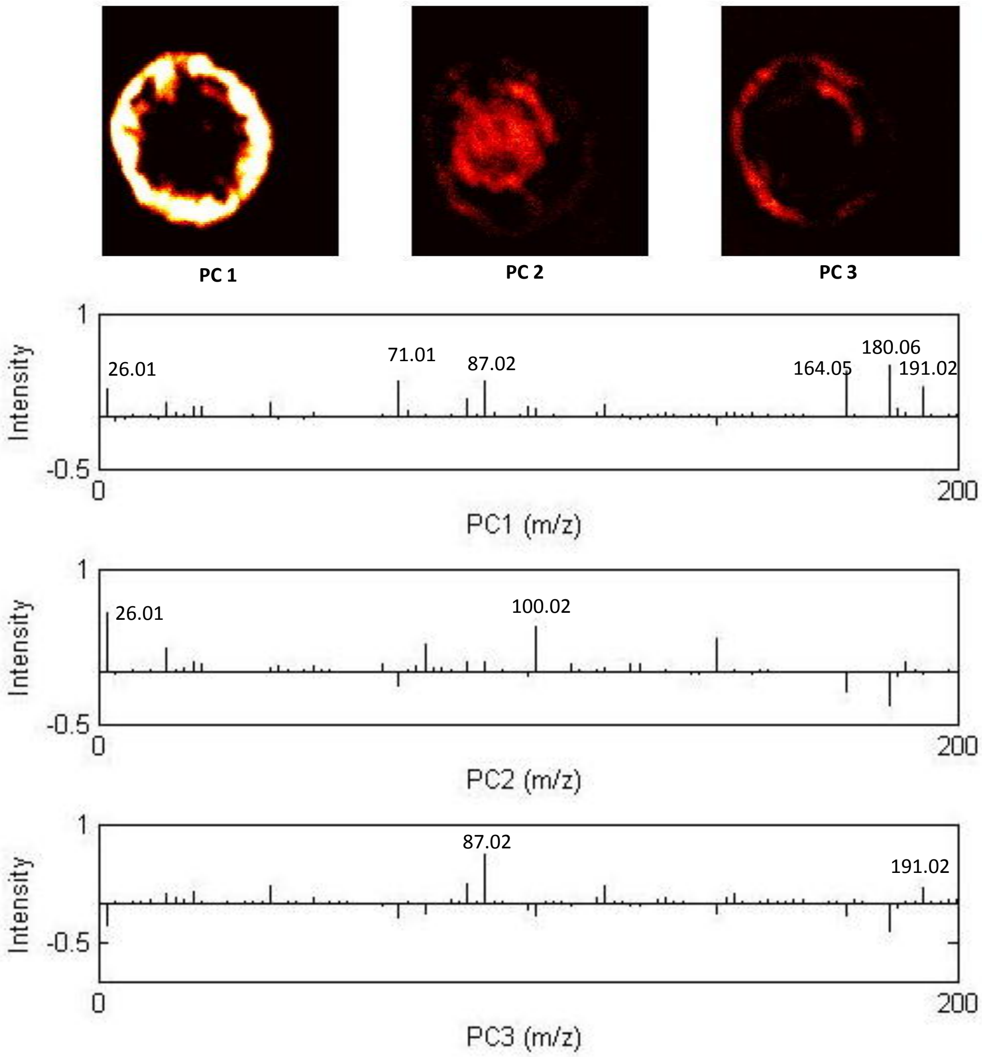


Figure 3.13 Scores images and loading plots produced using PCA with random sampling of a sample size of 4096 for TPC1 mixture. Score images are presented on the top while loadings for each of three PCs are shown below.

3.7 Conclusion

PCA has been used extensively in various research fields for its simplicity and easy implementation, it is a multi-dimension orthogonal linear transformation based on the statistic characteristics. It attempts to compute the most meaningful basis to re-express a complex dataset.

The variables in a “natural” dataset are always linearly interdependent to some high degree, reducing the precision and reliability of the data-mining methods. Highly dependent variables will also make the data collection and the subsequent analysis computation more complicated. However, exploring the data and the related covariance matrices enables the use of some ‘principal components’, a linear combination of the original variables for dimensionality reduction without losing the main information in the dataset.

In this chapter we demonstrated that PCA can be used for information extraction from complex ToF-SIMS datasets generated from mixture of metabolites using a lower number of variables. The random sampling technique was also combined with PCA in order to increase the efficiency of the algorithm. One disadvantage of PCA analysis is the presence of negative values in the computed mass spectra. This is also the case when MAF and MCR methods are used. This can hamper the interpretation of the results since the resulting estimates will be biased even if only small negative peaks exist in the spectra . This is because they do not represent a practically feasible solution. Therefore feasible solutions should be computed through algorithms which incorporate the non-negativity constraint. Such algorithms may reveal hidden information in a more clear and interpretable way. Also, PCA is very sensitive to the pre-processing prior to the analysis and the information with low intensity maybe lost.

Chapter 4

Non-negative Matrix Factorisation

4.1 Introduction

Processing large ToF-SIMS data has created strong demands for dimensionality reduction and noise removal techniques. In this thesis, our primary focus is on the development of novel algorithms with an appropriate transformation method, which can process ToF-SIMS data in an effective manner while being able to be utilised for metabolic profiling analysis in the real world application. The methods that we have discussed in the previous chapter all have one common drawback: negative values may appear during the transformation procedure. For most multivariate techniques, the original data matrix is supposed to be decomposed into a low rank form, meaning that it is likely to end up with negative components. However, negative components have no physical explanation in reality, and hence, affect the interpretability of the results. By imposing non-negativity constraint, the transformation process is organised to be purely additive and the original non-negative structure of the data is maintained.

In this chapter we will introduce the Non-negative Matrix Factorisation (NMF) technique, which can lower the dimensions of the data while provide meaningful results. In addition, it is also capable of producing a sparse representation of the data (Hoyer, 2004). By incorporating NMF with some other extending constraints, the algorithm is capable of ensuring a better visualised and efficient solution and providing an improved convergence property. The results of the application of NMF algorithm to the replicate samples of our ToF-SIMS dataset will be provided at the end of this chapter.

4.2 Non-Negative Matrix Factorisation Model

The basic idea of non-negative matrix factorisation was initially introduced by Paatero and Tapper in 1994 when they proposed an algorithm within alternating non-negative least squares framework. This algorithm was originally referred as “positive matrix factorisation” by Paatero and Tapper (1994) and did not receive much attention from the research society. The concept of NMF was popularised by Lee and Seung in 1999, when they proposed a well-known multiplicative updating algorithm for NMF in their seminal paper (Berry et al., 2007). Compared to Paatero’s method, their algorithm has better performance and is relatively easier to implement (Kim & Park, 2008).

NMF is devised in a way that no negative entries are allowed in the transformation procedure. It is capable of extracting significant features from the data in the form of basis vectors, which in turn, are combined to produce representative patterns.

In NMF, the original non-negative data matrix X of a dimension $\mathrm{n} \times \mathrm{m}$ can be reconstructed by the product of two data matrices $\mathrm{W\ and\ H}$. This can be described by:

$$\mathrm{X} = \mathrm{WH} + \mathrm{E} \tag{4.1}$$

Where $\mathrm{m} \times \mathrm{r}$ matrix W consists of r pure spectral basis vectors, $\mathrm{w_i}$, as its columns and H is a $\mathrm{r} \times \mathrm{n}$ matrix. Matrix E is the residual matrix unexplained by WH. In order to ensure WH is the compression of X, the value of r should satisfy $(\mathrm{n} + \mathrm{m})\mathrm{r} < \mathrm{n} \times \mathrm{m}$. Then each vector $\mathrm{x_i}$ in the original data matrix can be rearranged to the same basis vector $\mathrm{w_i}$ with the corresponding loading vectors $\mathrm{h_i}$. Therefore, the loading vector $\mathrm{h_i}$ can indicate how strongly each basis vector $\mathrm{w_i}$ occurs in relation to the original vector $\mathrm{x_i}$.

Because NMF algorithm intends to find a smaller number of basis which can represent the raw data in a meaningful way, the ambiguity elements E can be removed from the transformation procedure, resulting in a linear approximation of the original data. This can be described by:

$$\mathrm{X} \approx \mathrm{WH} \tag{4.2}$$

This linear representation is an approximation of the original non-negative data matrix X. PCA to some extent can also be considered as a matrix factorisation with no constraints on the negative entries in matrix W and H (Hoyer, 2004). By comparison, NMF involves a reduced rank approximation formed by non-negative factors. This means that the data matrix X is explained by non-subtractive combinations only, which maintain the non-negative structure of the data and produce a combined representation (Berry et al., 2007).

The aim of NMF is to find the best choices of the two non-negative matrices W and H that collectively approximate matrix X, by optimising the minimisation function of the reconstruction error between X and WH (Hoyer, 2004). Paatero and Tapper (1994) solved this problem by implementing non-negativity constrained alternating least squares algorithm for NMF, whereas Lee and Seung (2001) developed a multiplicative updating algorithm which has already been regarded as the standard NMF algorithm.

The minimisation function can not be convex in both W and H, but only one at a time. This means that there is no global optimal solution for NMF and only local optima can possibly be guaranteed. Researchers typically choose the most appropriate local minimum by comparing the local minima generated from different initialisations (Albright et al., 2006). This could be problematic for large dataset such as ToF-SIMS data that we used in this thesis. In addition, the standard NMF algorithm suffers from lack of convergence, because the point that satisfies convergence condition could be a stationary point which does not necessarily result in a local minimum (Berry et al., 2007).

Iterative process is generally used in NMF algorithms and it requires a starting point to initialise the algorithm. At each iteration of the NMF algorithm, the new value of W or H is obtained by updating the current value based on certain functions. An effective initialisation is thus particularly important as it can facilitate the convergence and reduce the processing time.

4.3 Methodology

NMF has gained great popularity due to its property of guaranteed non-negativity, and the emergence of different variations of the general NMF formula (see Equation (4.2)). For example, by multiplying both sides of the equation by a diagonal weighted matrix, the feature redundancy in matrix W can be reduced (Guillamet, Bressan, & Vitria, 2001).

The basic NMF model outlined in Equation (4.1) is originally stated as a minimisation problem described by the Euclidean Distance between X and WH:

$$\mathrm{J(W,H)} = \frac{1}{2}\|\mathrm{X_{m\times n}} - \mathrm{W_{m\times r}H_{r\times n}}\|_\mathrm{F}^2 \quad \text{subject to } \mathrm{W,H} \geq 0 \tag{4.3}$$

Where the product of WH is the matrix factorisation of data matrix X and r is an integer representing the rank of the approximation, given that $\mathrm{r} < \min(\mathrm{m}, \mathrm{n})$.

Several NMF algorithms have been developed to resolve this minimisation problem, including three broad and possibly overlapping methods: multiplicative update algorithms, gradient descent algorithms, and ANLS algorithms (Berry et al., 2007).

4.3.1 Multiplicative Update Algorithms

In the standard NMF algorithm by Lee and Seung (2001), the values of W and H are derived from updating their present values by multiplying a coefficient value, which depends on the approximation function. In most cases, the optimisation function is defined as the Kullback-Leibler divergence (Polani, 2013), which can be expressed by:

$$\mathrm{J_{KL}(X|WH)} = \sum_i \sum_j (\mathrm{x_{ij}} \log \frac{\mathrm{x_{ij}}}{\sum_k \mathrm{w_{ik}h_{kj}}} - \mathrm{x_{kj}} + \sum_k \mathrm{w_{ik}h_{kj}}) \qquad (4.4)$$

NMF can hence be transformed into the optimisation problem given by:

$$\min_{\mathrm{W,H}} \mathrm{J_{KL}(X|WH)} \text{ Subject to } \mathrm{W, H} \geq 0, \sum \mathrm{u_{ij}} = 1, \vee \mathrm{j} \qquad (4.5)$$

The iteration rule can then be described by:

$$\mathrm{H} = \mathrm{H} \frac{(\mathrm{W^T X})}{(\mathrm{W^T WH})} \qquad (4.6)$$

$$\mathrm{W} = \mathrm{W} \frac{(\mathrm{XH^T})}{(\mathrm{WHH^T})} \qquad (4.7)$$

Updating the iteration until the optimisation function in Equation (4.5) is minimised. The optimisation function can also be stated as the Euclidean distance between X and WH as defined in Equation (4.3), which is the standard to measure the similarity between two matrices (Hoyer, 2004). An alternative cost function rooted on the Csiszar's φ-divergence is proposed by Cichocki, Zdunek, and Amari (2006) to solve the problem.

It is important to note the optimisation procedure of multiplicative update algorithms only results in a stationary point, which may not lead to convergence of W and H to a local optimum (Berry et al., 2007). In addition, the convergence procedure of multiplicative update algorithms is considerably slow (Kim, Sra, & Dhillon, 2007). Lin (2007) suggested an optimisation method with bound constraint based on projected gradient technique in attempt to facilitate the convergence of multiplicative update rules.

4.3.2 Gradient Descent Algorithms

Gradient descent based algorithms also involves updating the value of H and W using step wise parameters. In fact, Lee and Seung's multiplicative update algorithm can be regarded as a type of gradient descent method (Chu et al., 2004; Lee & Seung, 2001). The update rules are similar to those in Equation (4.6) and (4.7):

$$\mathrm{H} = \mathrm{H} - \mathrm{s_H}\frac{\partial \mathrm{J}}{\partial \mathrm{H}} \tag{4.8}$$

$$\mathrm{W} = \mathrm{W} - \mathrm{s_W}\frac{\partial \mathrm{J}}{\partial \mathrm{W}} \tag{4.9}$$

Where $\mathrm{s_H}$ and $\mathrm{s_W}$ are the step size parameters. In gradient descent algorithms, non-negativity constraint is simply imposed by setting all negative values in W and H to zero after each update (Hoyer, 2004).

Although gradient descent based algorithms are easy to implement, just like multiplicative update algorithm, they are subject to slow convergence (Berry et al., 2007). Moreover, the application to large dataset can be problematic since gradient descent methods are particularly sensitive to the step size selections (Lee & Seung, 2001). In addition, gradient based methods experiences the phenomenon called zigzagging or jamming, resulting from the convergence to a non-optimal point (Bertsekas, 1982). Kim, Sra, and Dhillon (2007) proposed a modified Newton-type method based on nonnegative least squares that uses a non-diagonal gradient scaling scheme to address the problems associated with gradient descent based methods.

4.3.3 Alternating Least Squares Algorithms

Alternating least squares (ALS) was firstly applied to NMF problems by Paatero and Tapper in 1994. By fixing either W or H, the optimisation problem in Equation (4.3) can be solved using least squares technique in an alternating manner. In particular, ALS algorithms are generally more flexible with the ability to incorporate constraints into the iterative process. However, the original algorithm proposed by Paatero and Tapper was extremely slow as it was not properly fitted into NMF problems (Kim & Park, 2008). A simple and effective ALS algorithm that originally called Alternating Constrained Least Squares (ACLS) solves unconstrained least squares and sets all the negative entries in matrix W or H to zero at each iteration step in attempt to speed up the calculation (Albright et al., 2006):

Initial W as one random dense matrix;
For $\mathrm{i} = \mathrm{i}:\mathrm{k}$ (k is the iteration step number)
Solve H from equation $\mathrm{W^T X} = \mathrm{W^T W H}$
Multiplying both sides of Equation (4.2) by $\mathrm{W^T}$ and setting all negative entries in H to 0
Solve W from equation $\mathrm{H X^T} = \mathrm{H H^T W^T}$
Transporting Equation (4.2) and multiplying both sides by $\mathrm{W^T}$, then setting all negative entries in H to 0
End

Table 4.1 Alternating Least Squares Algorithm for NMF

Although this ALS algorithm offers a fast implementation, it is an inexact method that suffers from lack of convergence (Kim, Sra, & Dhillon, 2007). Compared to the exact ALS algorithms, such as the one used by Paatero and Tapper, ACLS might result in larger approximation errors. Albright et al. (2006) also proposed an advanced algorithm called Alternating Hoyer-Constrained Least Squares (AHCLS) which provides better sparsity than ACLS, however, the convergence to a local minimum is still not guaranteed. Several improvements on alternating non-negativity constrained least squares have been provided to alleviate the convergence problem (Kim, Sra, & Dhillon, 2007; Kim & Park, 2008).

4.4 Applicable Constraints

Owing to the flexibility of NMF, many researchers strive to introduce additional constraints into the algorithm in order to incorporate prior information or other preferred properties. The cost function of NMF is usually extended to include a penalty term which compensates for uncertainties in original data matrix X (Berry et al., 2007). This relationship is given by:

$$\mathrm{J(W,H)} = \|\mathrm{X} - \mathrm{WH}\|_{\mathrm{F}}^{2} + (\alpha, \beta)\mathrm{C(W,H)} \tag{4.10}$$

Where C is the penalty term that accounts for the constraints and α is the regularisation parameters that accounts for the compromise between the estimation error and the required constraints. By setting the regularisation parameter α and β to an appropriate value which is normally very small, the extended optimisation function can be restricted from increasing. The iteration rule is also extended by using partial derivatives of $\mathrm{C(W)}$ and $\mathrm{C(H)}$ with respect to W and H respectively.

Smoothness constraints

One simple smoothness constraint is based on the Tikhonov regularisation (Pauca, Piper, & Plemmons, 2006). It can be written in terms of the penalty term:

$$\mathrm{C(W)} = \|\mathrm{LW}\|_{\mathrm{F}}^{2} \tag{4.11}$$

Where L is the regularisation operator, such as Laplacian operator. The smoothness constraint is used to generalise the noise-contaminated results. It can be applied to matrix H in the similar manner.

Another widely used smoothness constraint is imposed by introducing the Toeplitz matrix T (Chen & Cichocki, 2005). The penalty term can thus be described by:

$$\mathrm{C(W)} = \frac{1}{\mathrm{n}}\|(\mathrm{I}-\mathrm{T})\mathrm{W}\|_{\mathrm{F}}^{2} \tag{4.12}$$

Where n denotes the observation number of the original data matrix X.

Sparsity constraints

One typical problem associated with Paatero and Tapper's ALS algorithm is that there is no sparsity restriction (Albright et al., 2006). Therefore, it is important to impose sparsity constraint to the solutions. There are several ways to derive the measure of sparsity, for instance the Hoyer's measure of data X can be expressed by:

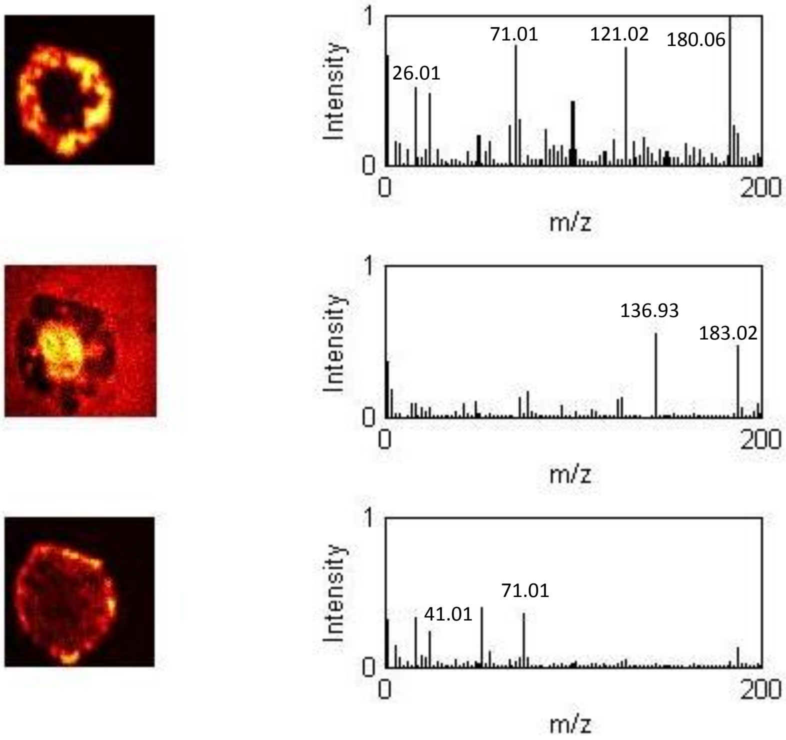
$$\mathrm{S(M)} = \frac{\sqrt{\mathrm{n}}-\|\mathrm{X}\|_1/\|\mathrm{X}\|_2}{\sqrt{\mathrm{n}}-1} \tag{4.13}$$

This matrix can be directly used as the penalty term in form of squaring the sparseness S. In this thesis, we apply a sparsity constrained NMF to our ToF-SIMS replicate samples where the results are provided in Section 4.5.

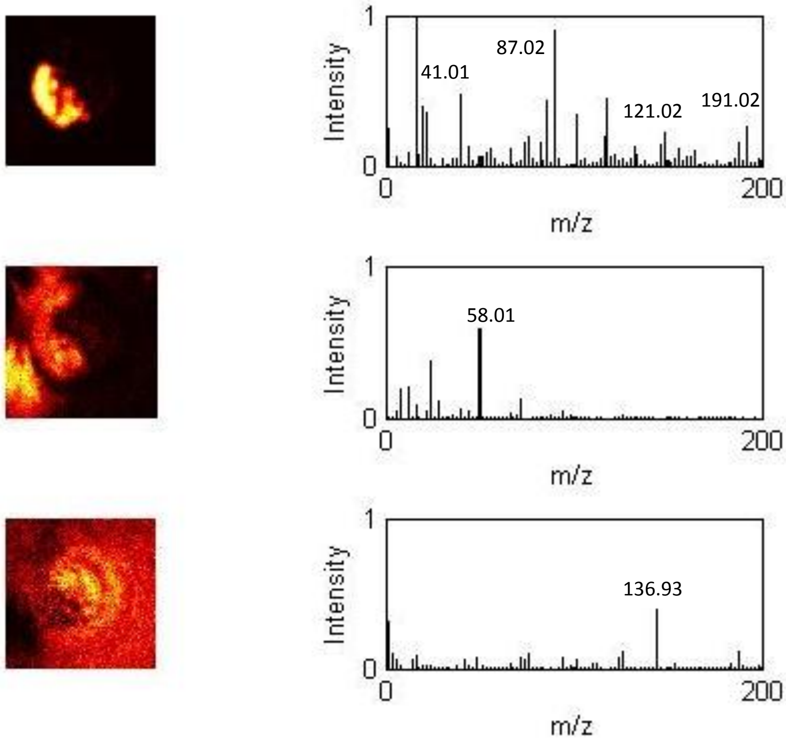
4.5 Application Results

In this section, we provide the results of the application of NMF to our replicate mixture samples. In order to facilitate the computation procedure, the pure component basis number is set according to the prior knowledge we gained from PCA application in Chapter 3. From our results of the implementation of PCA, P and C species all had three principal components even they were pure components. TPC mixture also required three principal components to represent 90% of the original dataset. It is important to note that only one principal component was required for TC mixture, which was insufficient as two pure species were contained in the mixture. Therefore, by incorporating the results of PCA application and the known information about the structure of mixtures, basis number r = 3 was selected for all the NMF applications in this case. Moreover, we used the first 3 principal components from PCA as the initial estimates for W and H in each implementation.

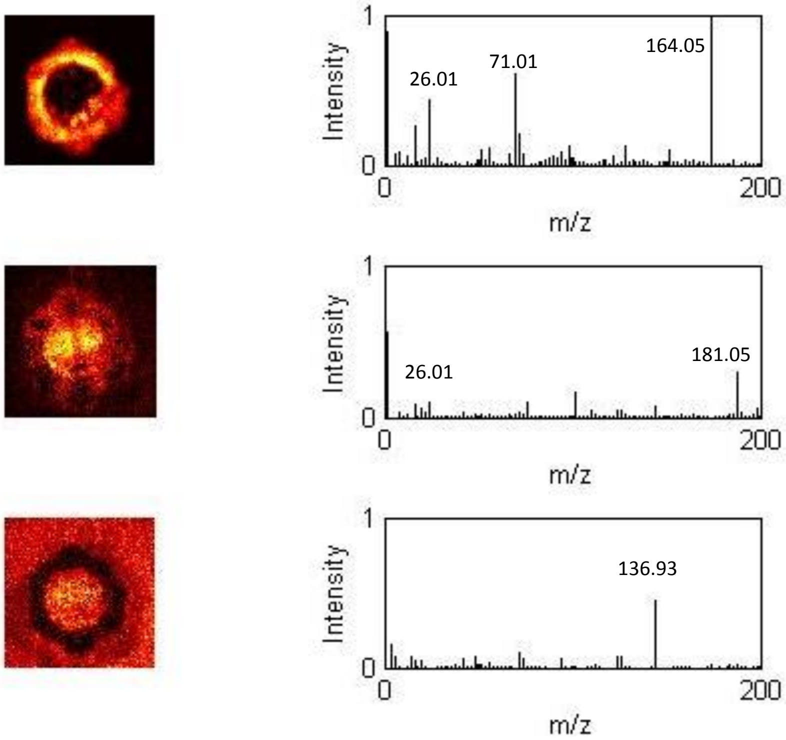
The loadings and scores images produced by NMF are depicted in Figure 4.1. In Figure 4.1 (T), it can be seen that there are salient peaks at m/z = 71.01, 121.02, and 180.06, m/z = 136.93 and 183.02, and m/z = 41.01 and 71.01 within the three spectral basis for the sample T respectively. Significant peaks for sample C and P, can be found at m/z = 41.01, 87.02, 58.01, and 136.93, and m/z =71.01, 136.93, 164.05 and 181.05 separately (Figure 4.1(C) and Figure 4.1(P)). It should be noted that a remarkable peak at m/z = 136.93 appears in the scores images for all three samples (T, C and P). This could possibly be due to the noise in the original ToF-SIMS dataset. In addition, the peak at m/z = 71.01 is present in both sample T and P, which may cause separation problem when the algorithm is applied to TPC mixture.


1
Intensity
0
0
200
m/z
26.01
71.01
121.02
180.06
1
Intensity
0
0
200
m/z
136.93
183.02
1
Intensity
0
0
200
m/z
41.01
71.01

(T)


1
Intensity
0
0
200
m/z
41.01
87.02
121.02
191.02
1
Intensity
0
0
200
m/z
58.01
1
Intensity
0
0
200
m/z
136.93

(C)



(P)

Figure 4.1 Three sets of loadings and scores images from NMF application for each pure component (T, C, and P) samples respectively. Scores images and loading plots are given for each species.

The NMF factorisation performance is demonstrated by the Frobenius norm errors between X and WH in Figure 4.2. The Frobenius norm errors can be defined by:

$$\mathrm{D} = \frac{\|\mathrm{X}-\mathrm{WH}\|_{\mathrm{F}}}{\sqrt{\mathrm{nm}}} \tag{4.14}$$

It is noticeable from Figure 4.2 that the residual is relatively stable after about 1000 iterations. In addition, when the prior knowledge is provided, for example, during one of the experiments the initial parameters are not randomly chosen but set to the result from other experiment (such as PCA), the Frobenius norm errors are stable even for a small iteration number. Therefore, the iteration number $\mathrm{i} = 1000$ is recommended as the time cost is also one of the important concerns especially when the original data is considerably large.

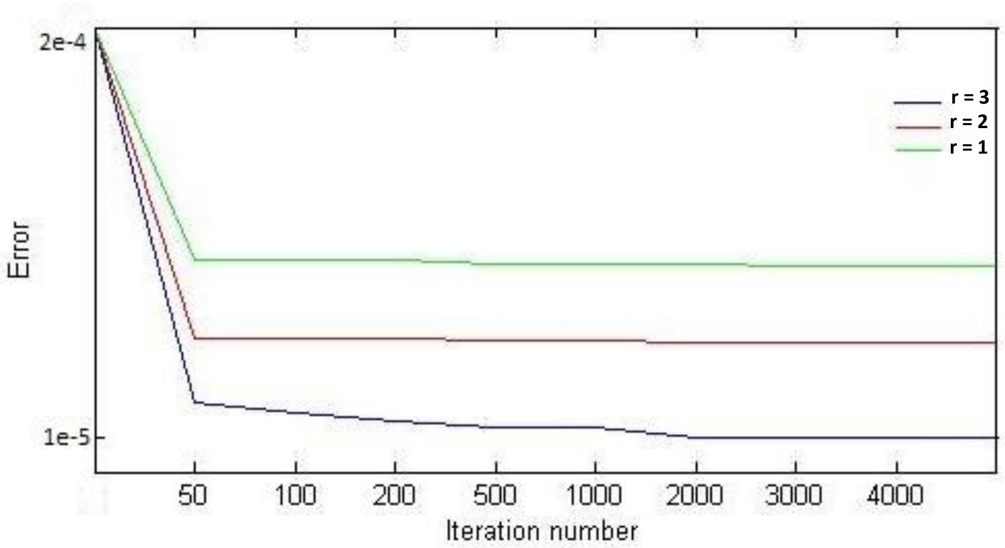
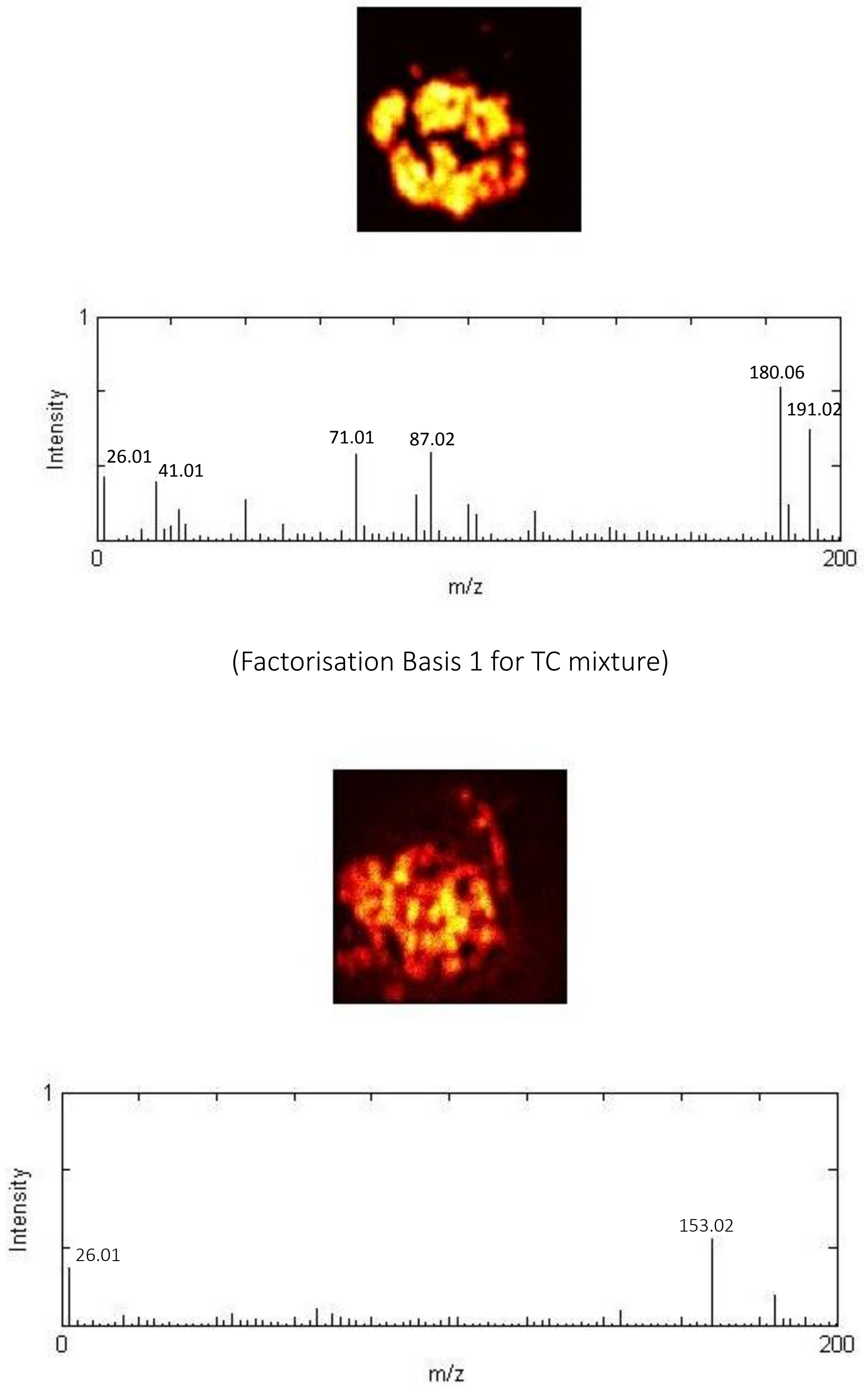


Figure 4.2 Convergence of the algorithm with different factors and number of iteration. This diagram shows the Frobenius norm errors between the original data X and the product of the factorisation matrices WH with respect to the changes of different factor numbers and number of iteration, which indicate the rate of convergence in NMF algorithm.

Figure 4.3 shows the results of NMF application to the mixture sample TC. The first panel in Figure 4.3 represents the first spectral basis within the sample TC. It can be seen that there are two intensive peaks at m/z = 71.01 and 180.06, which can be attributed to the spectrometer noise and component T respectively based on the discriminatory information obtained from Figure 4.1(T). Similarly, for the second spectral basis, two remarkable peaks at m/z = 87.02 and 191.02 should refer to component C with the known information. It should be noted that one distinct peak at m/z = 153.02 appears in the third scores images, which can be caused by the fragments of one component. It could also be due to the principal number r being greater than the actual number of components. Overall, the NMF algorithm is found useful in identifying components in the two component mixture.



(Factorisation Basis 1 for TC mixture)

(Factorisation Basis 2 for TC mixture)

Figure 4.3 Scores images and loading plots from factorisation of TC1 mixture samples using NMF, two factors are utilised.

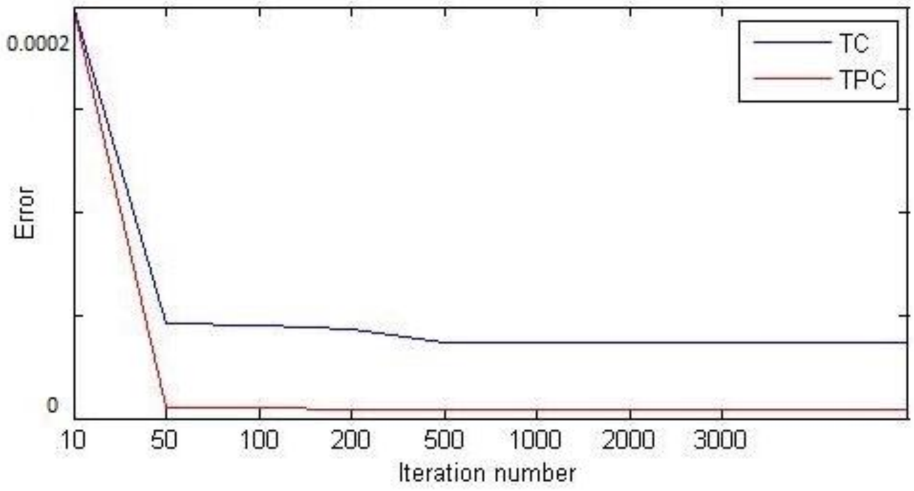
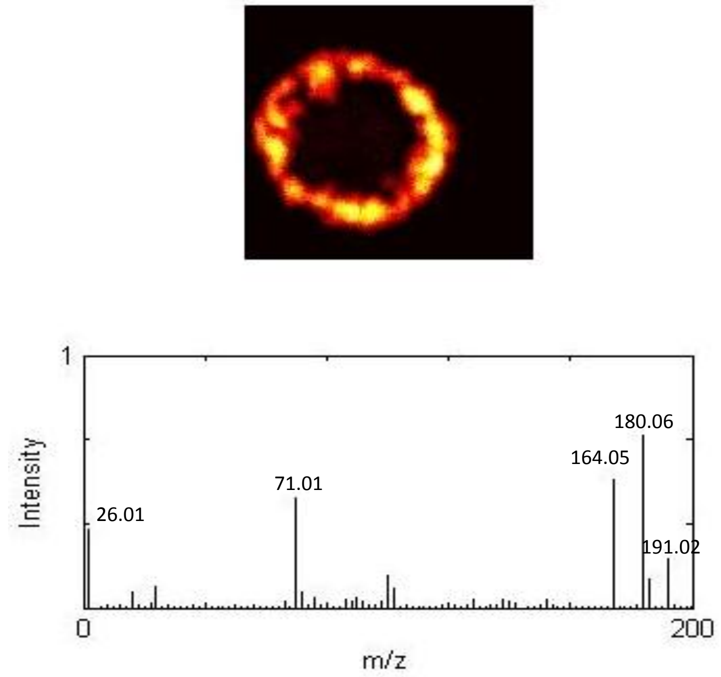


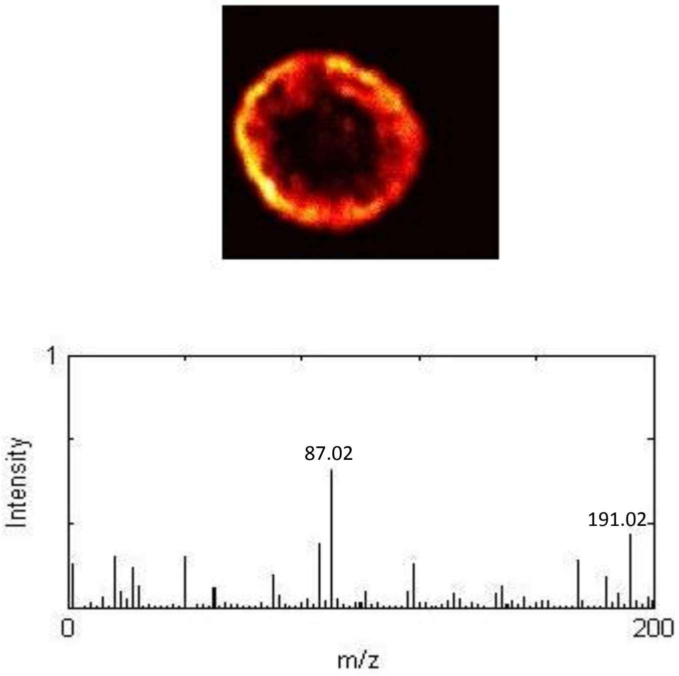
Figure 4.4 The Frobenius norm errors of the algorithm. This diagram shows the convergence of the NMF algorithm for the mixture TC and TPC respectively.

Figure 4.4 illustrated the approximating performance of the factorisation matrices in NMF application to the TC and TPC samples. The Frobenius norm errors are also stabilised when the number of iteration exceeds 1000, hence $\mathrm{i} = 1000$ is chosen as the iteration number for the case in TC and TPC.

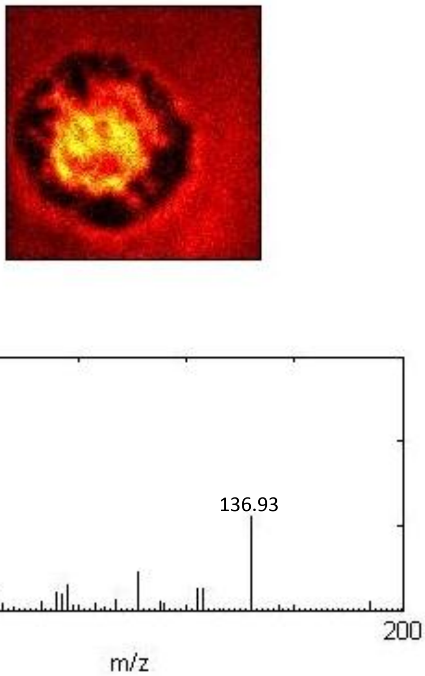
The results of NMF algorithm for replicate TPC mixture sample is shown in Figure 4.5. The first spectral basis has three significant peaks at m/z = 71.01, 164.05 and 180.06, which are inconclusive that whether they are attributed to component T or P, as a common peak is found at m/z = 71.01 and each of them has a discriminatory peak at m/z = 164.05 and 180.06 respectively. The highly intensive peak at the same location may affect the performance of NMF algorithm since there is no available knowledge about the fragmentations and the uncertainty of the dataset cannot guarantee the complete picture the NMF algorithm would present. In the second panel in Figure 4.5, the peaks at m/z = 87.02 and m/z = 191.02 refer to chemical compound C as these two peaks are specific reference for C in PCA. The only peak in the third score image is the noise peak at m/z = 136.93 as discussed previously. It is identified as the third spectral basis because component T and P are mixed into one spectral basis, while this noise peak appears in all the three components with a considerably high intensity.



(Factorisation Basis 1 for TPC mixture)



(Factorisation Basis 2 for TPC mixture)



(Factorisation Basis 3 for TPC mixture)

Figure 4.5 Scores images and loading plots produced using NMF for TPC1 mixture samples. Three spectral factors are given on the bottom of each pannel.

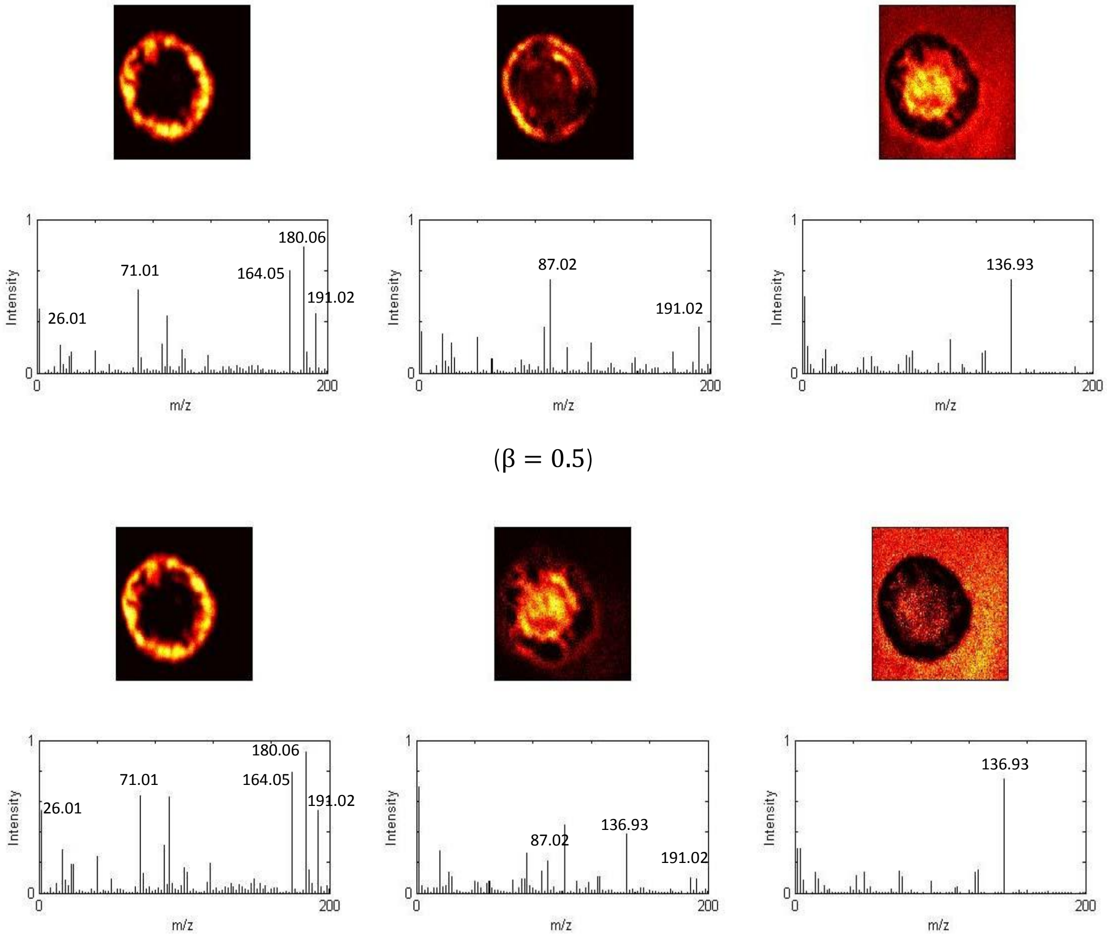
The rest of this section shows the results of sparsity-constrained NMF applying to the TPC mixture samples when the basis number r is set to 3. This implementation aims at finding the solutions of W and H with desired sparseness. The regularised cost function is used in this implementation imposes constraints on both W and H (Pauca, Piper, & Plemmons, 2006):

$$\min\{\|\mathrm{X}-\mathrm{WH}\|_{\mathrm{F}}^{2}+\alpha\|\mathrm{W}\|_{\mathrm{F}}^{2}+\beta\|\mathrm{H}\|_{\mathrm{F}}^{2}\}, \text{subject to } \mathrm{W}, \mathrm{H} \geq 0 \qquad (4.15)$$

Where $\alpha \geq 0$ is the parameter to supress the smoothness of W while $\beta \geq 0$ is the regularisation parameter accounts for the trade-off between the approximation accuracy and the sparseness (Berry et al., 2007). In order to do so, the parameter α is set to the maximum number of X while parameter β can be chosen from 0 to 1; the sparseness can be adjusted using Equation (4.13) by substituting X with the iteration result H which can be derived from Equation (4.15) for each β. The sparseness of H can be simplified as:

$$S(M) = \frac{\sqrt{n}-1/\beta}{\sqrt{n}-1}$$

(4.13)The sparseness becomes more intense when β is larger, thisalso can be validate from Figure 4.6 that the score images are sparser with a larger β and the sparseness is helpful in detecting the specific regions of individual chemical. The results can be compared with the previous results in Figure 4.5 that, because of the inclusion of the dynamics of the mixtures and the spatial problem (this can be seen from the loading images in both experiments, the component T and P are still close to each other), NMF with sparsity constraint is still not effective enough to distinguish nearby components with similarity.



($\beta = 0.5$)

($\beta = 1$)

Figure 4.6 Scores images and loading plots produced using sparsity constraint NMF for TPC1 mixture samples. Different regularising parameters were chosen for each experiment.

4.6 Conclusion

NMF algorithm is a low rank approximate factorisation method which has been extensively used in researches as a dimensionality reduction and segmentation technique. Compared with PCA, it uses a non-subtraction method to avoid the negative loadings, this unbiased algorithm makes the result more reasonable and interpretable for further study. In particular, NMF offers an incomparable feature in terms of retaining non-negativity in the results, and hence, providing physically meaningful interpretation. However, NMF is subject to the limitation that multiple solutions are available due to the removal of ambiguity element. In our thesis, the result may vary with each experiment for a set of randomly selected initial values; in several trials, it also shows that even with the same random initial, the result may be different from each other to some degree, which may also lead to a poor convergence. In order to address the problem, we set the initial value to the PCA result from the previous chapter; this initialisation method provides not only a non-multiple result but also a fast convergence compared with other approaches. Sparsity-constrained NMF is also provided to overcome this problem with improved convergence process, however due to large amount of similar fragments and noise in the dataset, the application is not effective in distinguishing different components (Pauca, Piper, & Plemmons, 2006). Above all, it is still subject to a number of limitations:

- The convergence is only guaranteed to a fixed point which may be a local minimum or saddle point.
- The convergence rate depends on the quality of the initialisation.
- It requires repeated experiments to choose regulation parameters
- A large number of iterations can complicate the computation, leading to a time-consuming estimation procedure.

Chapter 5

Non-negative Matrix Factorisation under the Bayesian Framework

5.1 Introduction

Uncertainties arising from a number of different sources will influence the results obtained from any data analysis method: non-deductible noise occurs in the data collection procedure; correlated variables may result in an overlapping and ambiguous factors. Therefore it is very important to apply a factorisation analysis, which reduces the inexactness of the raw data as well as represents the underlying system with greater accuracy. NMF, one of the simpler methods for factor analysis of non-negative data, is used to accomplish the goal of reducing the number of variables and detecting relationships among the variables. It provides meaningful and physically interpretable solutions in many applications, and is considered as an advanced alternative tool compared with PCA and ICA.

The interpretation of NMF as a low rank matrix approximation is sufficient for the derivation of an inference algorithm, yet this view arguably does not provide the complete picture. The NMF needs to be extended to account for the uncertainties and correlations that exist in the data as well as to robustly identify the number of underlying factors. In this chapter, we describe NMF from a statistical perspective. This view will pave the way for developing extensions that facilitate more realistic and flexible modelling as well as for more sophisticated inference, such as Bayesian model selection. By incorporating NMF into Bayesian framework (B-NMF), the algorithm is capable of ensuring a unique solution for NMF algorithm and providing an improved convergence rate. The results of the application of B-NMF algorithm to the replicate samples of our ToF-SIMS dataset will be provided at the end of this chapter.

5.2 Bayesian Non-Negative Matrix Factorisation

NMF approach as stated in the previous chapter is an approximation of the original non-negative matrix X with the product of two non-negative factorising matrices W and H, where W is the template or sources and H is the expansion coefficients. The algorithm is a process of estimating W and H while minimising the fitting error between raw data X and WH. This can be expressed as:

$$\mathrm{X} = \mathrm{WH} + \mathrm{E} \tag{5.1}$$

Where E is the fitting error and J is the cost function as below.

$$(\mathrm{W}, \mathrm{H}) = \arg\min \mathrm{J}(\mathrm{X} \,\big|\, \mathrm{W}, \mathrm{H}), \ \mathrm{s.t}\, \mathrm{W}, \mathrm{H} \geq 0 \tag{5.2}$$

Where $\mathrm{J} = 0$ when $\mathrm{X} = \mathrm{WH}$, and the minimisation can be iteratively solved by using the multiplicative update rules illustrated in Section 4.3.1. The typical cost functions that should be used in NMF depend on the choice of distance measures, such as squared Euclidean divergence, generalised Kullback-Leibler (KL) divergence and the Itakura-Saito divergence (Févotte & Cemgil, 2009,). These measures have been mentioned in the previous chapter and are always nonnegative and convex for each factor in NMF.

Under appropriate assumptions on the distribution of the original data and factors, this algorithm can be considered as estimating the non-negative factorising matrices W and H through using their maximum likelihood estimates (Schmidt, Winther, & Hansen, 2009). The distance measures in J can be seen as a result of the error E having Gaussian, Poisson, and Gamma error statistics respectively. Therefore, the selection of cost functions is essentially affected by the fitting error, which can be managed by incorporating Bayesian techniques (Févotte & Cemgil, 2009,).

5.3 Methodology

5.3.1 The Statistical Perspective

As we discussed previously, the residual matrix E in Equation (5.1) can be eliminated as the approximation ambiguity. However, this may lead to infinite solutions for the optimisation. This problem can be addressed by introducing prior densities to the iterative process (Schmidt, Winther, & Hansen, 2009). In the Bayesian framework, matrix E can be represented in terms of a likelihood function and the parameters can be expressed in terms of prior densities (Schachtner et al., 2014). By incorporating NMF into the Bayesian framework (B-NMF), prior knowledge about density can be introduced into the factorisation, leading to reliable results and improved convergence.

In order to allow efficient inference in the method, a convenient parametric form is preferred for the prior densities. Bayesian NMF employs the normal likelihood and exponential priors during the Gibbs sampling procedure for their pervasiveness (Schmidt, Winther, & Hansen, 2009). The reconstructed error matrix E is assumed to be distributed as a Gaussian, which can be described by:

$$\mathrm{p}(\mathrm{X}|\mathrm{W},\mathrm{H},\sigma^2) = \prod_{\mathrm{i,j}} \mathrm{N}\left(\mathrm{X}_{\mathrm{i,j}}; (\mathrm{WH})_{\mathrm{i,j}}, \sigma^2\right) = \prod_{\mathrm{i,j}} \frac{\exp(-\frac{1}{2}(\mathrm{X}_{\mathrm{i,j}} - (\mathrm{WH})_{\mathrm{i,j}})^2)}{\sqrt{2\pi}\sigma} \tag{5.3}$$

Where Gaussian density is given by:

$$\mathrm{N}(\mathrm{x}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}} \tag{5.4}$$

Additionally, W and H are assumed as independently and exponentially distributed with scales a and b. The priors can be defined by:

$$\mathrm{p}(\mathrm{W}) = \prod_{\mathrm{i,n}} \mathrm{Exp}\,(\mathrm{W}_{\mathrm{i,n}}; \mathrm{a}_{\mathrm{i,j}}) \tag{5.5}$$

$$\mathrm{p}(\mathrm{H}) = \prod_{\mathrm{i,n}} \mathrm{Exp}\,(\mathrm{H}_{\mathrm{i,n}}; \mathrm{b}_{\mathrm{i,j}}) \tag{5.6}$$

Where $\mathrm{Exp}(\mathrm{x}; \beta) = \beta \mathrm{e}^{(-\beta \mathrm{x})} \mathrm{u}(\mathrm{x})$ is the exponential density with the unit step function $\mathrm{u}(\mathrm{x})$ which guarantees the non-negativity as $\mathrm{u}(\mathrm{x}) = 0$ when $\mathrm{x} < 0$.

Then the inverse gamma density is selected as the prior for the noise variance, with the shape parameter k and scale parameter θ:

$$\mathrm{p}(\sigma^2) = \mathrm{G}^{-1}(\sigma^2, \mathrm{k}, \theta) = \frac{\theta^1}{\Gamma(\mathrm{k})} (\sigma^2)^{-\mathrm{k}-1} \mathrm{e}^{-\frac{\theta}{\sigma^2}} \tag{5.7}$$

The posterior can be derived from the product of the residual likelihood in Equation (5.3) and the priors of W and H and noise variance obtained from Equations (5.5-5.7). The estimation of the factors W and H can be obtained during maximisation of the posterior.

5.3.2 Factor and Loadings Estimation

The estimations of the posterior probability density for both factors are required for the factorisation. The Markov chain Monte-Carlo (MCMC) sampling method is used to estimate the marginal density of the factors, one of which is Gibbs sampling (Smith & Roberts, 1993). MCMC is a broad set of computational algorithms based on the Markov chain convergence theorem; it is widely used in machine learning to solve the integration and optimisation problem (Andrieu et al., 2003). The optimisation can be gained by sampling from a constructed Markov chain with a desired equilibrium distribution. The set of steady chain samples is then used as the optimised distribution. Gibbs sampling is one efficient MCMC method to approximate the marginal density of the variables by obtaining the samples from the specified multivariate distribution (Bishop, 2006). It is applicable when the direct sampling is difficult and it is very adaptable under the Bayesian framework. Gibbs sampling generates a sequence of samples correlated with nearby samples, the sequence of samples can be drawn from the conditional posterior densities of the model parameters, and then the sequence converges to one sample from the joint posterior. The conditional densities of W and H can be considered as the Rectified Normal density (Harva & Kaban, 2006). This is given by:

$$\mathrm{R}(\mathrm{x}) = \Phi\left(-\frac{\mu}{\sigma}\right)\delta(\mathrm{x}) + \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}\right)\mathrm{U}(\mathrm{x}) \tag{5.8}$$

Where Φ is the cumulative distribution function of the standard normal distribution, while U is the unit step function in Equation (5.8). δ is the Dirac delta function given by:

$$\delta(\mathrm{x}) = 0 \text{ when } \mathrm{x} \neq 0, \text{and } \delta(\mathrm{x}) = +\infty \text{ when } \mathrm{x} = 0 \tag{5.9}$$

This rectified Gaussian distribution truncates all the negative entries via the unit step function. Hence the conditional probability of W and H is defined by:

$$\mathrm{p}\left(\mathrm{W}_{\mathrm{i,j}}\middle|\mathrm{x}, \mathrm{W}_{\mathrm{except(i,j)}}, \mathrm{H}, \sigma^2\right) = \prod_{\mathrm{i,j}} \mathrm{R}\left(\mathrm{W}_{\mathrm{i,j}}; \mu_{\mathrm{i,j}}, \sigma^2{}_{\mathrm{i,j}}, \mathrm{w}_{\mathrm{i,j}}\right) \tag{5.10}$$

$$p\left(H_{i,j} \middle| x, H_{except(i,j)}, W, \sigma^2\right) = \prod_{i,j} R\left(H_{i,j}; \mu_{i,j}, {\sigma^2}_{i,j}, h_{i,j}\right) \tag{5.11}$$

Where the conditional probability of the noise variance σ^2 is still set with the Inverse-Gamma density $p(\sigma^2|X, W, H) = G^{-1}(\sigma^2, k_{\sigma^2}, \theta_{\sigma^2})$. Based on the given information, the sampling process is illustrated in the table below:

Sampling Process
Iteration a) For each element in W, draw a sample using rectified Gaussian b) Draw a sample from the inverse-Gamma density for σ^2 c) For each element in H, draw a sample using a rectified Gaussian
Save the sample of W and H

Table 5.1 Gibbs Sampling Procedure

5.3.3 Model Order Selection

Generally, Bayesian probability theory incorporates the prior knowledge to the factorisation problem in order to reduce the uncertainty of the model, hence not only the optimal factorisation parameters but also the factorisation model can be derived (Knuth, 2005). For an unknown dataset, the determination of the model (in NMF, this refers to the number of the factors) is problematic as it cannot be selected directly based on the dataset itself. Normally, NMF is combined with PCA to solve this problem, whereas, under the Bayesian framework, model selection can be performed to determine the number of factors. Model selection requires evaluation of the marginal likelihood $P(X|M)$, which involves an intractable integral over the posterior of the factors W and loadings H. Once the marginal likelihoods for different models are obtained, Bayes factors can be computed to compare and select the more favoured model. The Bayes factor computation is given in Equation (5.12) where models M_0 and M_1 are compared.

$$\mathrm{bf} = \frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_0)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)} \tag{5.12}$$

The factor indicates the comparison ratio between marginalised likelihood of data X under two different models (Sinharay & Stern, 2004). Each of the models is associated with specific hypotheses, where $\mathrm{p}(\mathrm{X}|\mathrm{M})$ is the marginal density under model M. The Bayes factor can also be extended into another form of both posterior and prior odds as below (Kass & Raftery, 1995):

$$\mathrm{bf} = \frac{\mathrm{p}(M_0|\mathrm{x})\mathrm{p}(M_1)}{\mathrm{p}(M_1|\mathrm{x})\mathrm{p}(M_0)} \tag{5.13}$$

The marginal likelihood of model M can be defined by (Bos, 2002):

$$\mathrm{p}(\mathrm{X}|M) = \int \mathrm{p}(\mathrm{X}|\theta, M)\mathrm{p}(\theta|M)\mathrm{d}\theta \tag{5.14}$$

As the integral cannot be calculated analytically in practice, there are several alternative methods available for estimating the Bayes factor, including annealed importance sampling, bridge sampling, path sampling, and Chib's method, all of which can be used with the NMF algorithm (Diciccio et al., 1997; Meng & Wong, 1996; Chib,1995). Among those advanced methods, Chib’s method is the most appropriate choice as it is easily combined it with Gibbs sampling and the computational cost is much lower compared to others.

Chib’s method only uses the posterior sample draws to estimate the marginal likelihood, which suggests estimating the posterior density by:

$$\mathrm{p}(\mathrm{X}|M) = \frac{\mathrm{p}(\mathrm{X}|M,\theta)\mathrm{p}(\theta|M)}{\mathrm{p}(\theta|\mathrm{X},M)} \tag{5.15}$$

The likelihood and prior in the numerator can easily be solved, and the denominator can be estimated from Gibbs sampling output. One efficient way of this is by blocking the parameters in which all θ parameters are partitioned into k blocks, with the dominator rewritten as a product of the k terms and the marginal likelihood can be approximated by k runs during Gibbs sampling (Chib & Jeliazkov, 2001).

5.3.4 B-NMF Iterative Algorithm

The B-NMF algorithm is given in Table 5.2 as an iterative procedure. The parameters W and H in the NMF model can be derived from Gibbs sampling through being set equally to the conditionals at each iteration. In addition, the columns of W and the rows of H can be used as the blocks for the Gibbs sampling (Schmidt, Winther, & Hansen, 2009). The iterative process stops until the convergence to the maximum of joint posterior density obtained.

```
B-NMF Algorithm
Iteration
For $\mathrm{i} = 1\!:\!\mathrm{r}$
    Set $\mathrm{W}_{:,\mathrm{i}}$ to the conditional mode while set negative quantities to zero
End
Updating $\sigma^2$
For $\mathrm{i} = 1\!:\!\mathrm{r}$
    Set $\mathrm{H}_{:,\mathrm{i}}$ to the conditional mode while set negative quantities to zero
End
Save the output of $\mathrm{W}$ and $\mathrm{H}$
```

Table 5.2 B-NMF Algorithm

5.4 Application Results

In this section, we demonstrate the application of Bayesian NMF on our ToF-SIMS replicate samples. In order to facilitate the computation procedure, the factor number can be chosen via Chib's method, by computing the marginal likelihood for the data given factor number (Chib & Albert, 1997). This can be compared to the results with the prior knowledge gained from the PCA application in Chapter 3. Gibbs sampling generates 1000 samples in each block, which is found to be sufficient from several trials. In Figure 5.1, the first model with only one factor has a high potential to represent the original data. The result is more robust than PCA, since Chib's method employs posteriors that are more promising in finding better solutions.

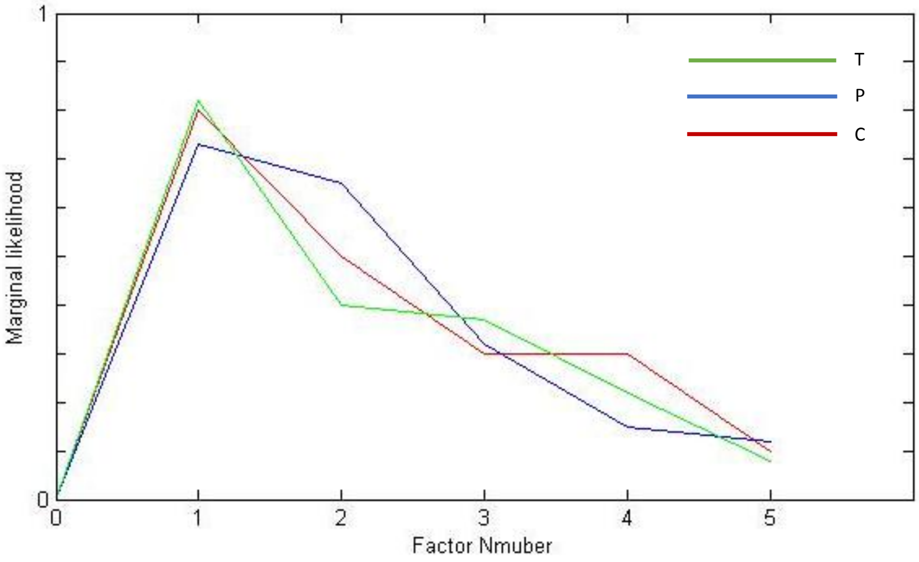


Figure 5.1 NMF model order selection using Chib's method. The plots represent the marginal likelihood for individual component of the three pure chemical samples, T, P, and C. Only the models within 5 factors are presented.

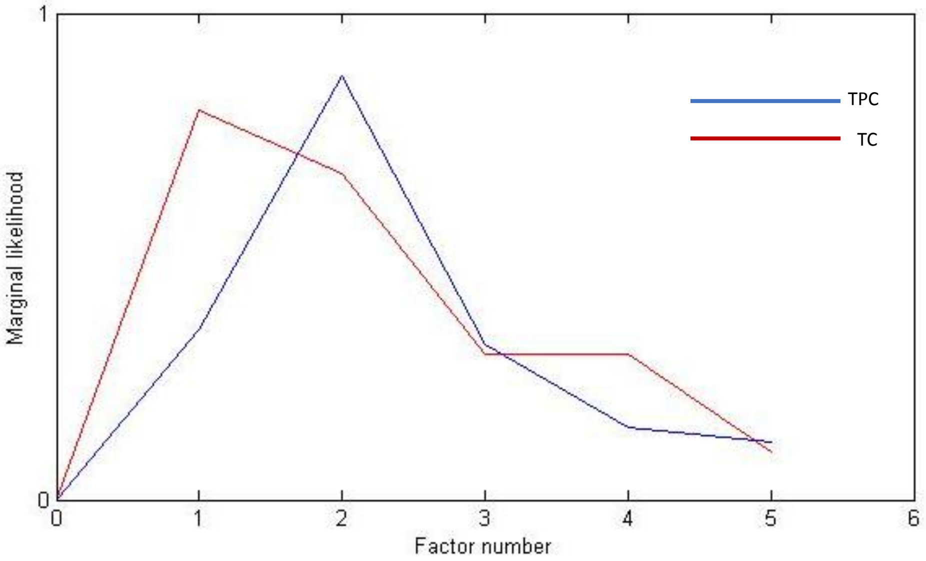


Figure 5.2 NMF model order selection using Chib's method. The plots represent the marginal likelihood for individual component of the two mixtures samples, TC and TPC. Only the models within 5 factors are presented.

Figure 5.2 provides a direct illustration of the factor number chosen using Chib's method for the mixed species sample TC and TPC. It shows that one factor model seems more appropriate for the two species mixture TC and two factor model is preferred for the three species mixture TPC. This result is similar to the results of PCA implementation, which suggest that one PC and three PCs should be chosen for reasonable representation of TC and TPC respectively. This could possibly be due to that the intensity of the identical location peak for species C is considerably low relative to other species and is dominated by other fragmental peaks.

From our results of the implementation of Chib's method, T, P and C species all require one factor to represent the dataset, which is the correct number for the factorisation. However, one factor and two factors are insufficient to represent the original dataset of TC mixture and TPC mixture, as we know that two and three mixed species are contained in the mixture, respectively. By using the results of Chib's method, factor number r is selected to be 1 for three single chemical samples and $\mathrm{r} = 1$ and 2 are selected for the two and three mixed species sample respectively during the application of B-NMF. Furthermore, by incorporating the known information about the structure of mixtures, $\mathrm{r} = 2$ and 3 are also implemented for the two mixtures for the purpose of comparison. The number of iteration is set to 100 as the experiments showed that B-NMF algorithm offers fairly fast convergence rate (Figure 5.3).

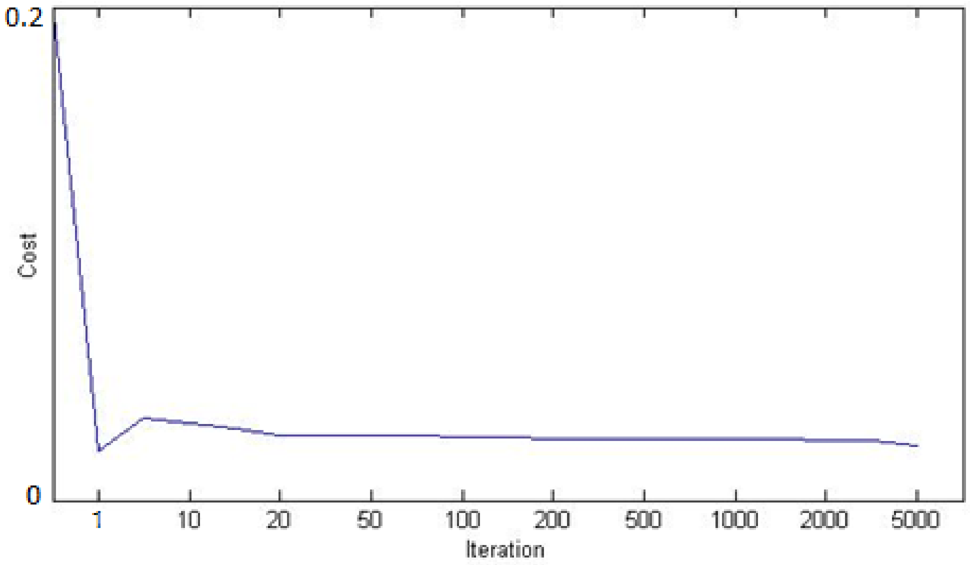
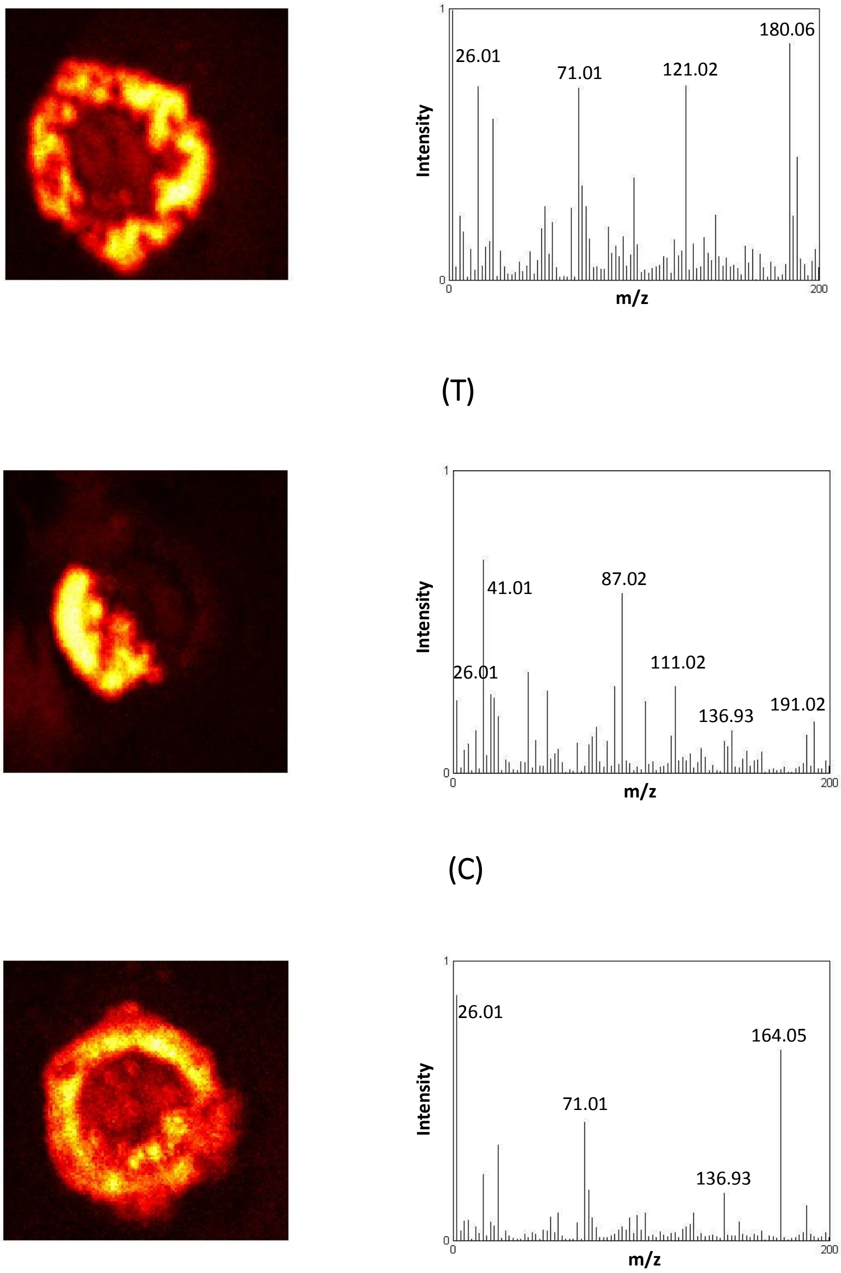


Figure 5.3 Convergence rate of B-NMF algorithm for the TPC mixture The B-NMF algorithm converges fairly fast since the cost function is stabilised after only 50 iterations.



(T)

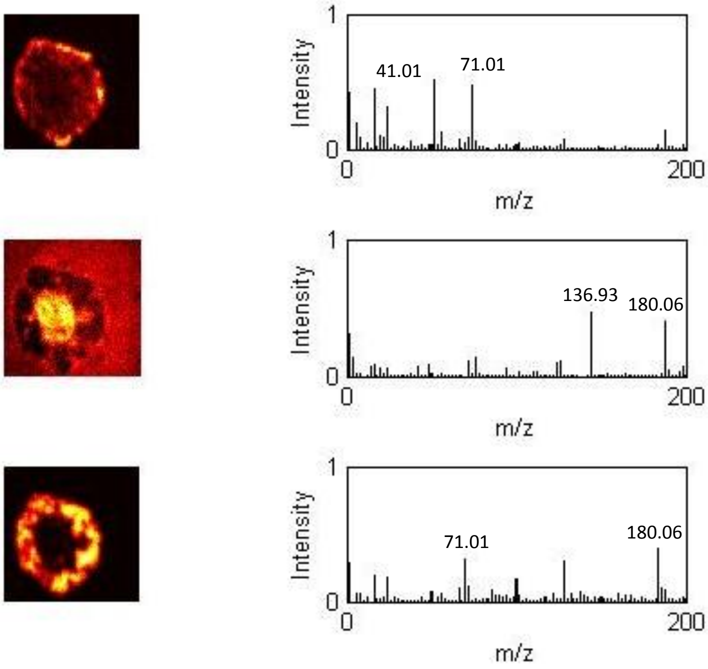
(C)

(P)

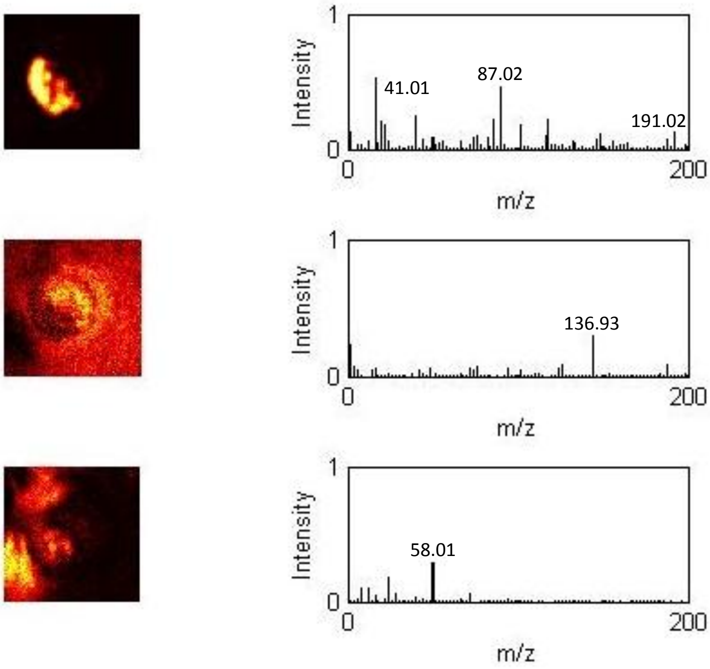
Figure 5.4 Scores images and loading plots produced using B-NMF for T1, C1 and P1 species samples. Scores images are on the left showing the spatial information of each factor while loading plots are on the right indicating the factors.

The loadings and scores images produced by B-NMF are depicted in Figure 5.4. In Figure 5.4 (T), it can be seen that there are salient peaks at m/z = 26.01, 71.01, 121.02 and 180.06. Significant peaks for species C can be found at m/z = 26.01, 41.01, 87.02, 111.02, 136.93 and 191.02 (Figure 5.4(C)). Large peaks for component P appear at m/z = 26.01, 71.01, 136.93 and 164.05 (Figure 5.4(P)). It should be noted that several peaks exist in more than one species. For example, a peak at m/z =26.01 appears in all three species, significant peak at m/z = 71.01 appears in both component T and C, and peak at m/z = 136.93 exists in both component C and P. All these specious peaks can be concluded as noise and fragments of the species from the spectrometry process. In addition, identical peaks for T and C are relatively close to each other, such as peak at m/z = 180.06 and 191.02 respectively. This also increases the difficulty in separating the two species. In addition, it can be seen in Figure 5.4 that identical signals for each metabolite are dominant, but the fragments of each species also have high intensities, which might be the cause for the separation problem.

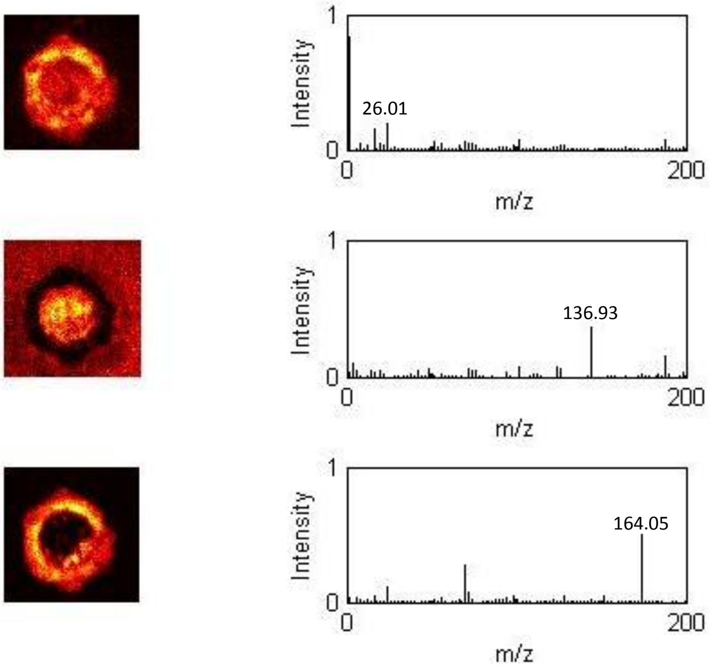
A result for setting the factor number to 3 is also given below to provide some supporting evidences of the correctness of the Chib's method. It can be seen that, for species T with 3 factors, the three basis are split from a single factor derived above with different intensities (Figure 5.5 (T)). They are highly correlated with each other as evidenced by peaks appearing in all factors at the same m/z locations. For species C and P, it can be seen that the peak at m/z=136.93 has been separated from others. However, this cannot provide any further information and is likely to be noise or fragments during the spectrometry process. Therefore, a factor number $\mathrm{r} = 1$ is an appropriate choice for this factorisation.


1
Intensity
0
41.01
71.01
0
200
m/z
1
Intensity
0
136.93
180.06
0
200
m/z
1
Intensity
0
71.01
180.06
0
200
m/z

(T)

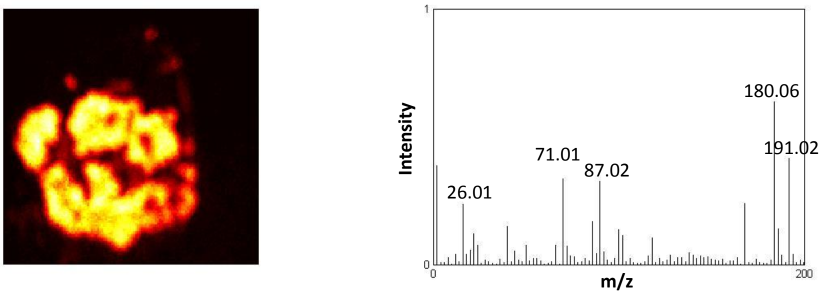

1
Intensity
0
41.01
87.02
191.02
0
200
m/z
1
Intensity
0
136.93
0
200
m/z
1
Intensity
0
58.01
0
200
m/z

(C)

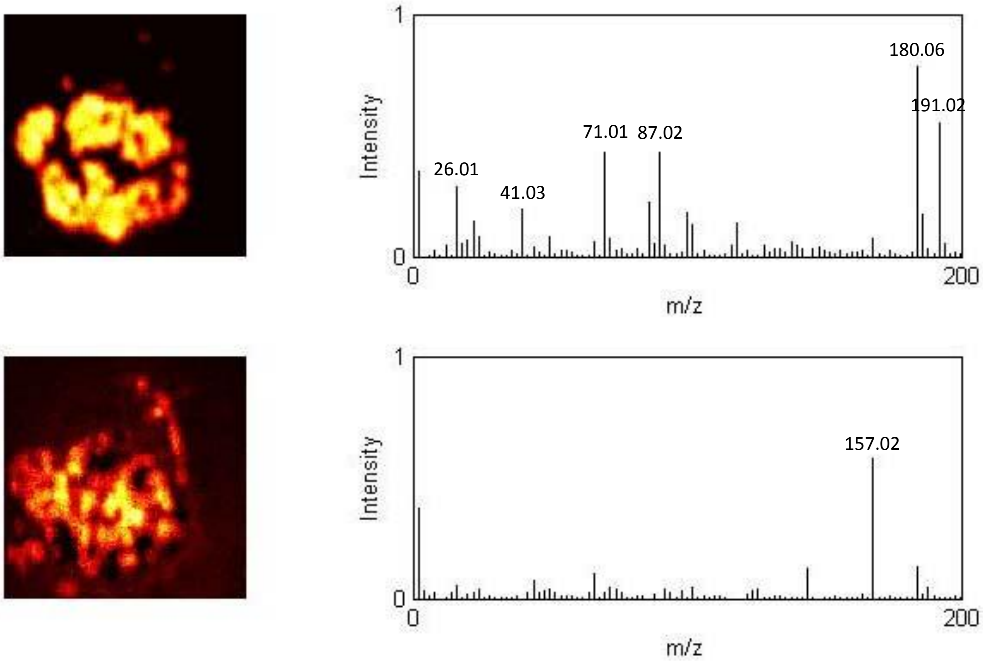


(P)

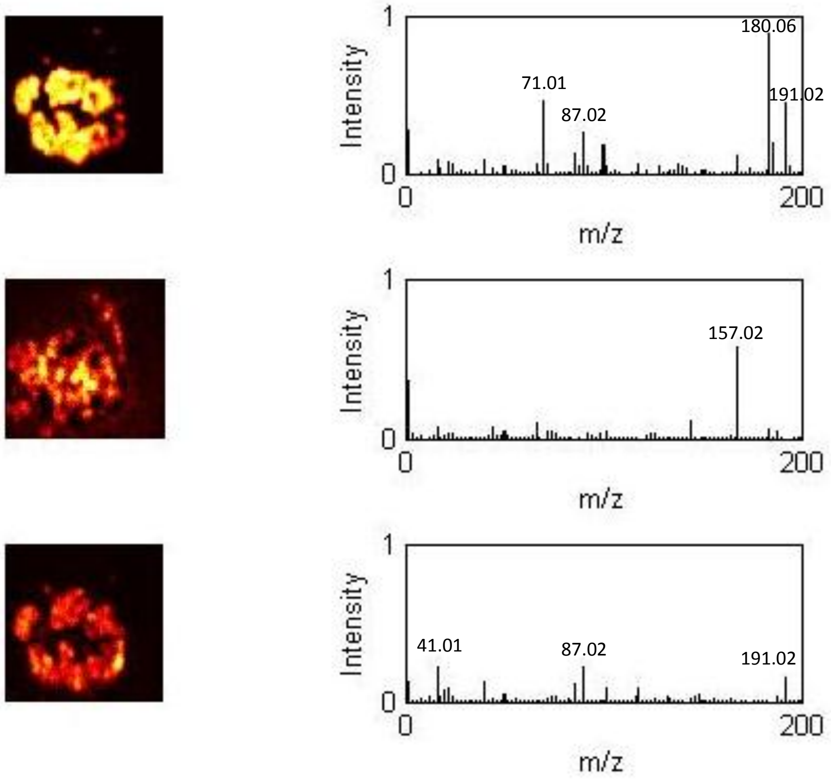
Figure 5.5 Scores images and loadings plots produced using B-NMF for T1, C1 and P1 species samples with additional factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the basis.



(One Factor Model)



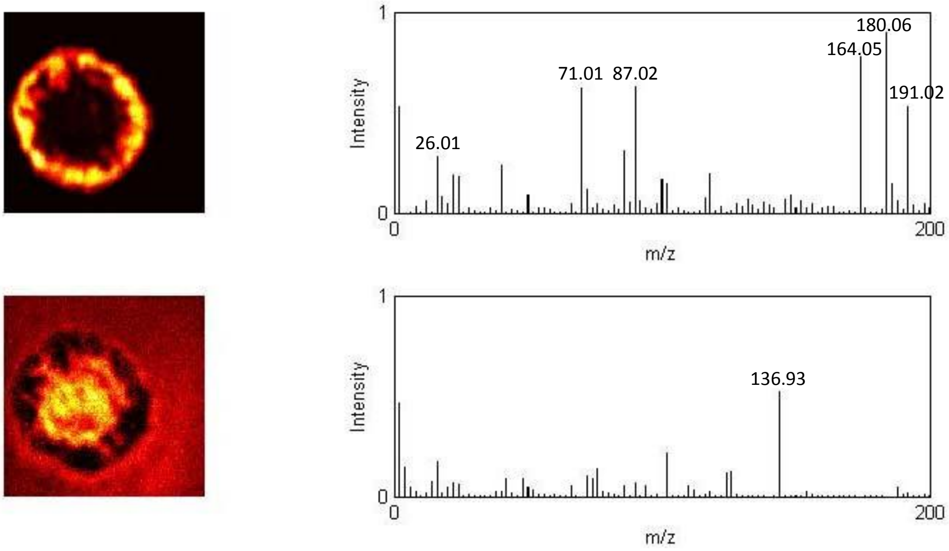
(Two Factors Model)



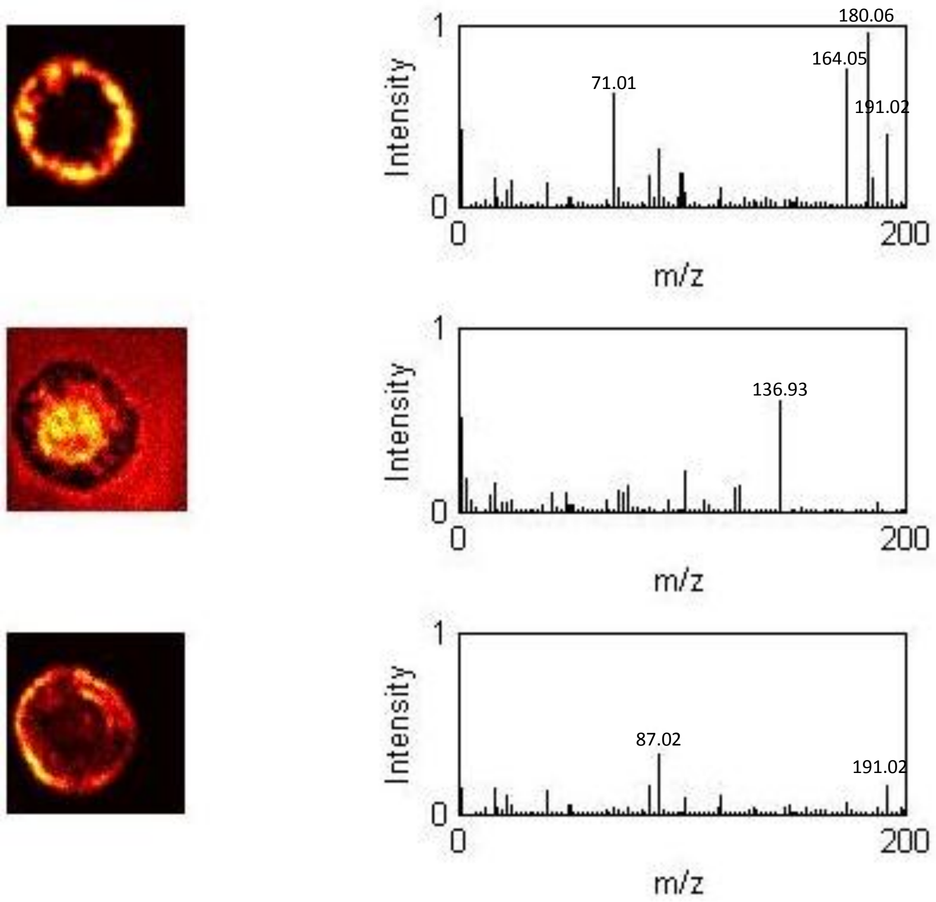
(Three Factors Model)

Figure 5.6 Scores images and loadings plots produced using B-NMF for TC1 mixed species samples with Models of 1, 2 and 3 factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors.

The results of applying B-NMF method to TC mixture sample are shown in Figure 5.6, where the three panels represent the results for different chosen factor numbers. It can be seen that the results of the three models are relatively similar, with slight differences with respect to the first factor basis. Despite of the peaks at m/z = 26.01 and 71.02, which have already been hypothesised as spectrometry process noise, intensive peak at m/z = 180.06 can be attributed to component T based on the discriminatory information we obtained from Figure 5.4(T). Spectra with peaks at m/z = 87.02 and 191.02 can be identified as species C as observed in Figure 5.4(C). Although Chib's method suggests that there is only one factor for TC mixture when the true number of factors ought to be two, we also provide two factors model and three factors model results for comparison purpose. A peak at m/z = 157.02 in factor 2 has been separated from factor 1 of the single factor model, this could be due to the fragment from species C in both two factors and three factors models. The third factor in the three factors model has spectra with low intensities and is highly correlated with the first basis, which suggests that the basis number r is greater than the actual number of components. Overall, the B-NMF algorithm is found useful in identifying components in a two component mixture although the determination of the number of factors needs further improvement.



(Two Factors model)



(Three Factors model)

Figure 5.7 Scores and loadings images produced using B-NMF for TPC mixed species samples with Models of 2 and 3 factor numbers. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors.

The results of the of B-NMF algorithm for TPC mixture sample are depicted in Figure 5.7. When the model factor number $\mathrm{r} = 2$, the first spectral basis has five significant peaks at m/z = 71.01, 87.02, 164.05, 180.06 and 191.02. These peaks are combinations of components T, P or C rather than a single component, because from the prior ground truth, each of them has a discriminatory peak at m/z = 180.06, 164.05 and 191.02 respectively. B-NMF may incorrectly recognise these three components as one since all of them have the same highly intensive noise peak during the data collection and also the factor number may be sufficient to explain the data but insufficient to relate to the ground truth. Moreover, high spatial correlation in the raw data is a serious challenge for species discrimination and spatial separation.

The ion images from the raw data corresponding to the m/z values of T, C and P are presented in Figure 5.8. It can be seen that high values for each species are all concentrated around an annular region, with especially T and P being almost the same shape, making the spatial separation more challenging. In the third panel of the three factors model in Figure 5.7, the peaks at m/z = 87.02 and 191.02 can be identified as species C. Therefore, with a correct factor number, species C can be successfully identified as separate from T and P, while T and P appear more mixed with high spatial correlation.

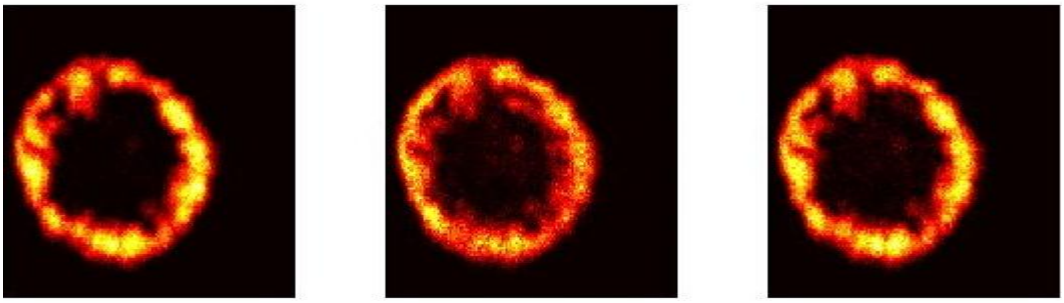
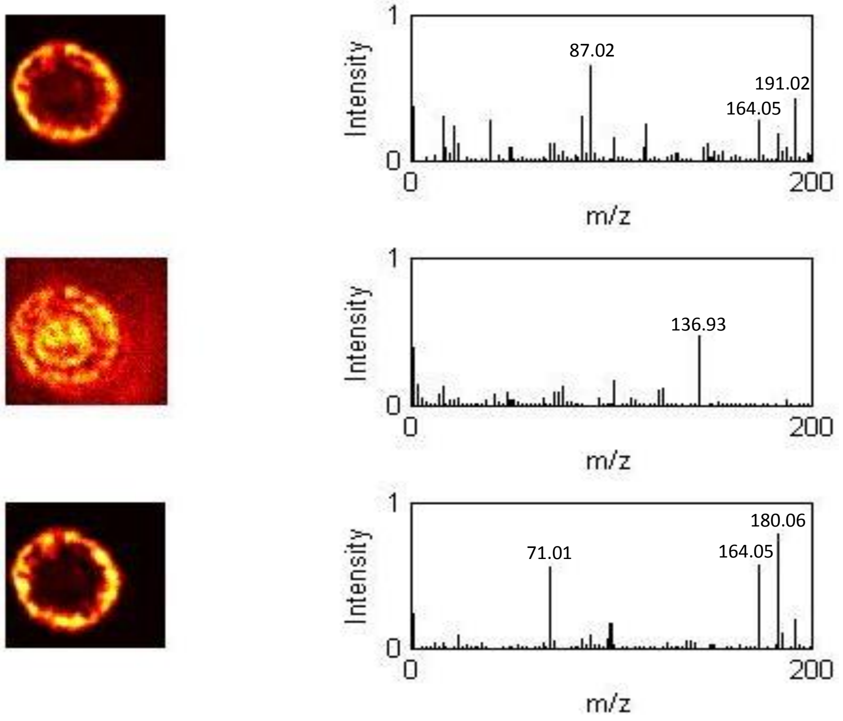
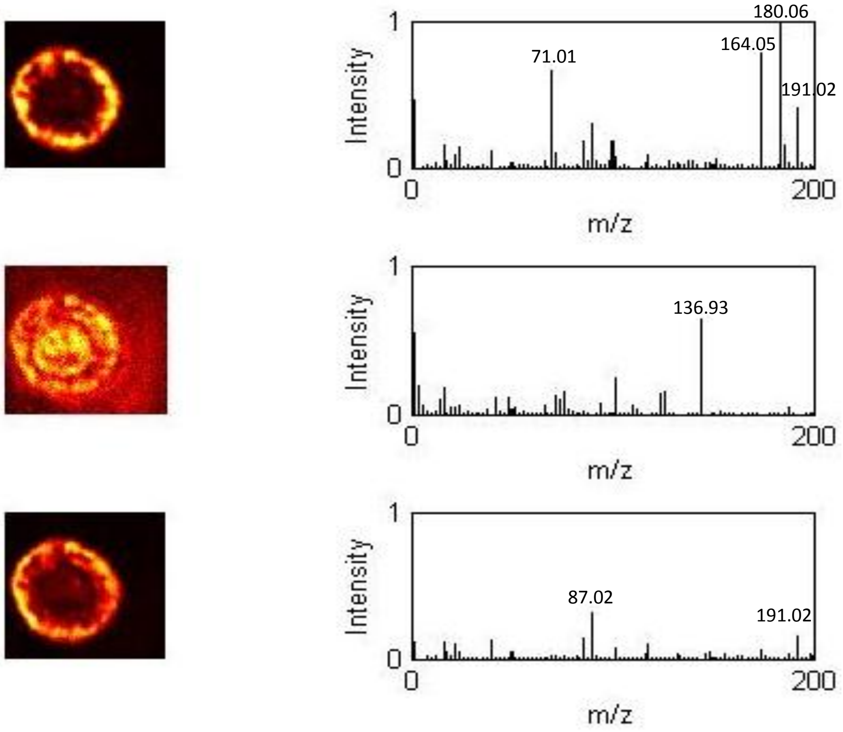


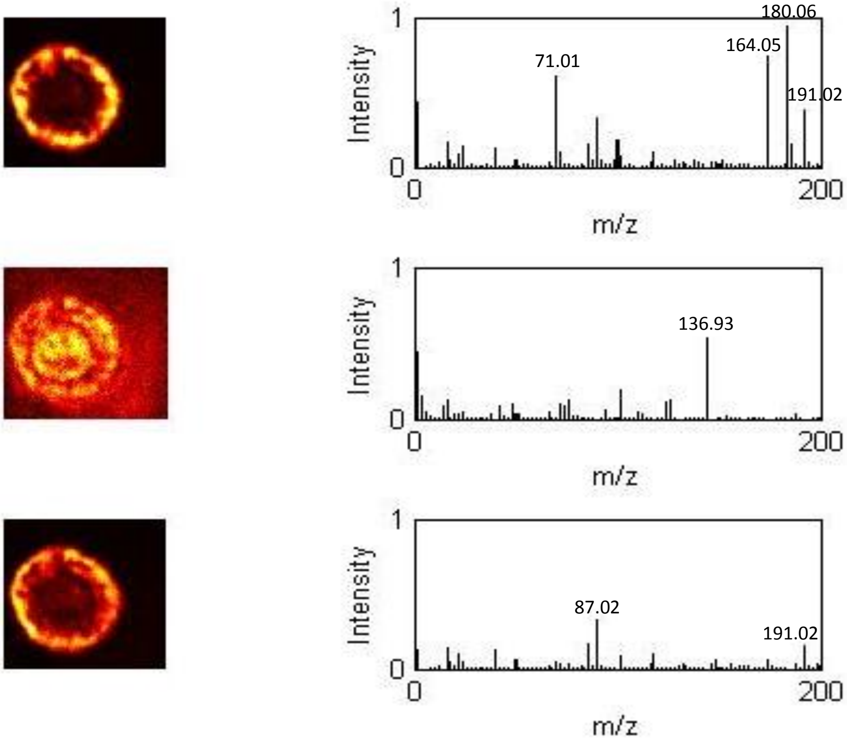
Figure 5.8 Spatial location for each species in TPC1 mixture sample. The three images implies the spatial location for each species in the mixture, T (m/z = 180.06), C (m/z = 191.02) and P (m/z = 164.06) are shown from left to right.



(sample size of 256)



(sample size of 1024)



(sample size of 4096)

Figure 5.9 Scores images and loading plots produced using B-NMF combined random sampling method for TPC mixed species sample. Several sample number have been chosen with the range from 85% to 90% reduction of the raw data. Scores images are on the left showing the spatial information of each factor while loadings are on the right indicating the factors.

Although the convergence can be achieved in a small number of iterations, the B-NMF algorithm still suffers from high computational demand, about 30~50 minutes for one trial with a far smaller data size of $128 \times 128 \times 100$. In order to improve the efficiency, random sampling can be applied to reduce the data size, before the application of the B-NMF algorithm, similar to that presented in Chapter 3. Figure 5.9 indicate that, with a correct factor number $\mathrm{r} = 3$, the B-NMF implementation with 1024 samples from the original datasets is still able to obtain the similar features as the original case. But the factorisation begins to perform poorly while the sample number reduces, which can be given by the high correlated loadings from the 256 samples factorisation. Several trials on the same dataset show that a high percentage reduction of the observations does not affect the

B-NMF algorithm results in this case while improving the efficiency with high speed (Figure 5.9).

5.5 Conclusion

The B-NMF algorithm provides an opportunity to use probability concepts in richly structured data analysis problems where uncertainties are prevalent. It takes the ambiguity into consideration by estimating the maximum likelihood of the parameters which is more advanced than the classical NMF method. The algorithm is computed using MCMC methodology and the resulting samples can also be used directly in model order selection. Order selection offers the possibility to identify the unknown number of factors, an important issue to be addressed in this ToF-SIMS data analysis for metabolic profiling. The result is more credible to the ground truth compared with PCA scree plot. For large datasets, the high computational cost can be resolved by combining B-NMF with a random sampling method. With this combination, both high computational efficiency and fast convergence can be achieved. The B-NMF method chosen here can be applied to many practical problems in bioinformatics. However, it does not take into account any spatial correlation that may exist in the dataset, which may possibly limit its performance as seen in this chapter.

Chapter 6

Alternating Non-negative Least Squares

6.1 Introduction

As we discussed in the previous chapter, NMF offers advantages over traditional multivariate techniques, such as PCA, MAF, and MCR; since the non-negativity is maintained in the results. One simplest method for imposing non-negativity constraint is to overwrite the ordinary (unconstrained) least squares procedure by setting all negative elements in the solution to zero (Berry et al., 2007). However, convergence to optimal minimum is not guaranteed in this method.

An optimised MCR approach, namely multivariate curve resolution alternating least squares (MCR-ALS), is capable of ensuring the non-negativity by imposing constraints on the iterative process, while offering the potential to estimate spectra of pure compounds (Keenan & Kotula, 2005; Wang et al., 2003). MCR-ALS typically requires an initial estimate of either the pure component spectra or the concentration profiles as a starting point for the iterative computation (Tyler, 2006). The initial estimate can be obtained from other algorithms such as PCA and MAF or known information about the data. In addition, many traditional algorithms that based on alternating non-negativity constrained least squares (ANLS), including the earliest NMF method proposed by Paatero and Tapper, appear to be computationally burdensome when applying to large multi-dimensional datasets (Kim & Park, 2008). However, the performance of ANLS can be significantly improved using fast combinatorial non-negativity constrained least squares (FC-NNLS) algorithm, which is designed specifically for multiway data analysis (Van Benthem & Keenan, 2004).

One major problem of applying MCR-ALS algorithm to large ToF-SIMS datasets is that the computation is largely complicated by the high resolution of the output images. As the complexity of scores and loadings estimation is closely related to the number of image pixels in the ToF-SIMS data, higher image resolution of the sample surface would result in greater computational demand and uncertainty. It is therefore of great importance that the number of pixels in TOF-SIMS images is decoupled from the number of unknowns required to be estimated. Another problem for scores and loadings estimation is that ordinary MCR methods do not account for the spatial dependency over the sample surface (Aram et al., 2014). In cases where the spatial distributions of the sample surface are continuous in nature, the characteristics of chemical species at close regions on the surface are relatively more inter-correlated than those at remote regions. Thus, the spatial correlation needs to be taken into consideration in the estimation of scores and loadings.

The framework we proposed in this chapter is MCR that incorporates alternating least squares (ALS) using a basis function decomposition approach. This MCR-ALS approach involves a spatially continuous representation of ToF-SIMS images which describes the spatial correlation across the observed surface. Moreover, by taking advantage of basis function decomposition method, the computational complexity of the estimation procedure can be less affected by the resolution of ToF-SIMS images. In particular, the estimation of individual pixel value is simplified into a set of weights, which subsequently scales the basis functions and leads to considerably lower spatial dimensions. The speed of the ALS algorithm and the reliability of the estimates are improved as a result of less number of unknown factors.

In this chapter, we will firstly outline the MCR method used in ToF-SIMS data analysis. Within the ANLS framework, a model reduction technique that employs a weighted sum of continuous basis functions is used to approximate scores images. The guidelines for basis functions configuration will also be provided. At the end of this chapter we will present the results of the estimation of scores and loadings for ToF-SIMS data analysis using our proposed algorithm. The work in this chapter was conducted involving other researchers in the group. My contribution is in developing the algorithm of the new method and leading the analysis of the dataset in the thesis. The work has been published in the paper (Aram, Shen, Pugh, Vaidyanathan and Kadirkamanathan, 2014).

6.2 Algorithm

6.2.1 Model

As we mentioned in Chapter 2, MCR is a second-order matrix decomposition method which transforms the original data matrix into the product of two smaller data matrices. Our ToF-SIMS dataset can be described as a bilinear model in a way that the spatial and spectral information form a two-way data matrix in the model, i.e. spatial matrix and spectral matrix correspond to each one of the two orders of our ToF-SIMS data matrix. The spatial (scores) data matrix provides information about the distribution of the chemical species and spectral (loadings) data matrix describes the identity of them on the sample surface. The spatial-spectral ToF-SIMS data matrix, as described by the MCR bilinear model, is shown as:

$$\mathrm{Y(f,s) = W(s)B^T(f) + E(f,s)} \tag{6.1}$$

Where f is the mass-to-charge ratio and s is the spatial location in the two dimensional physical surface. There is a transpose operator denoted by superscript T. Each ToF-SIMS image of dimension l by $\mathrm{l'}$pixels is rearranged into a $\mathrm{p \times v}$ data matrix, $\mathrm{Y(\cdot)}$, where $\mathrm{p = l \times l'}$. $\mathrm{W(\cdot)}$ $(\mathrm{p \times m})$ is the scores matrix and $\mathrm{B(\cdot)}$ $(\mathrm{v \times m})$ is the loadings matrix containing m spectral basis vectors. The $\mathrm{p \times v}$ residuals matrix, $\mathrm{E(\cdot)}$, is the error terms that are not explained by the scores and loadings estimation.

At any particular peak, a sum of weighted loadings can be used to represent every element of Y at a given spatial region where the weights are scores at that region. This relationship is given by:

$$\mathrm{Y(f,s) = \sum_{i=1}^{m} w_i(s)\, b_i(f) + E(f,s)} \tag{6.2}$$

Where $\mathrm{w_i}$ denotes the $\mathrm{i_{th}}$ weight at the corresponding region and $\mathrm{b_i}$ denotes the spectral basis vectors.

Here we demonstrate how the proposed algorithm is used to compute the scores estimation $\mathrm{W(\cdot)}$ and the loadings estimations $\mathrm{B(\cdot)}$ in Equation (6.1). The estimation procedure involves a two-step iterative procedure, which in essence is two sequentially performed non-negativity constrained least squares subject to convergence criterion. It is important to note that prior knowledge of the chemical rank or the number of spectral basis vectors is required as a starting point of iterations for MCR-ALS algorithm.

A successful estimation of the scores and loadings matrices should also involve identifying the correct chemical rank. We determine the chemical rank in the algorithm by using the number of principal components as a guide, which is obtained from PCA and the scree test criterion. The effects of measurement noise in the identification of the chemical rank of the system can be mitigated by implementing more sophisticated techniques, such as smoothing methods and subspace comparisons (Jiang, Liang & Ozaki, 2004).

6.2.2 ALS algorithm

A solution to the MCR model described in Equation (6.1) can be obtained by optimising the following minimisation function:

$$\mathrm{J}(\mathrm{W},\mathrm{B}) = \|\mathrm{Y} - \mathrm{W}\mathrm{B}^{\mathrm{T}}\|_{\mathrm{F}}^{2} \tag{6.3}$$

ALS algorithm is often used to handle this optimisation problem (Paatero & Tapper, 1994). At the start of the iterative procedure, an initial estimate of scores is used to compute estimate of the loadings by minimising $\mathrm{J}(\mathrm{B}|\mathrm{W})$ in Equation (6.3). The resulting loadings estimates are then used to update scores estimates i.e. $\mathrm{J}(\mathrm{B}|\mathrm{W})$, which are in turn used in the next iteration. In addition, non-negativity constraint is applied to ALS algorithm in order to provide meaningful and interpretable solutions. The minimisation function in Equation (6.3) is specified by:

$$\mathrm{J}(\mathrm{W},\mathrm{B}) = \|\mathrm{Y} - \mathrm{W}\mathrm{B}^{\mathrm{T}}\|_{\mathrm{F}}^{2} \qquad \mathrm{s.t.}\ \mathrm{W},\mathrm{B} > 0 \tag{6.4}$$

This is the Frobenius norm of the approximation where all the entries of W and B matrices are constrained to be non-negative. The non-negativity constrained least squares problem can be facilitated using FC-NNLS algorithm along with ALS (Van Benthem & Keenan, 2004).

We adopt stopping criteria through monitoring the Frobenius norms of the successive estimates of scores matrices (W) in Equation (6.4):

$$\|\mathrm{W}\|_{\mathrm{F}}^{(\mathrm{k})} - \|\mathrm{W}\|_{\mathrm{F}}^{(\mathrm{k}-1)} < \rho \tag{6.5}$$

Where ρ is a threshold value, and $\|W\|_F$ is given by:

$$\|W\|_F = \sqrt{\sum_{i,j} \left|\omega_{i,j}\right|^2} = \sqrt{\mathrm{tr}(W^T W)} \tag{6.6}$$

6.2.3 Model Decomposition

The spatially continuous nature of the species distribution on the sample surface can hamper the estimation procedure. This problem can be addressed by applying a decomposition method which reconstructs scores images in form of continuous basis functions (Aram et al., 2014). The reason for utilising continuous basis functions is that the spatially continuous locations can contain information about the spatial correlation, leading to more appropriate estimations. The basis decomposition method is given by:

$$w_i(s_p) \approx \sum_{j=1}^{n} \alpha_{ji} \phi_j(s_p) \tag{6.7}$$

Where $\phi(s)$ are known basis functions, α_{ji} are unknown weights, and n is the number of basis functions employed in the decomposition. The basis functions we used are 2 dimensional Gaussian basis functions given by:

$$\phi(s) = \exp(-\frac{(s-\mu_\phi)^T (s-\mu_\phi)}{\sigma_\phi^2}) \tag{6.8}$$

Where σ_ϕ denotes the width of basis functions and μ_ϕ denotes the centre of basis functions. As we will demonstrate later, spatial frequency analysis is used to determine the width and the location of the basis functions in Equation (6.8).

A continuous approximation can be obtained by substituting Equation (6.7) into Equation (6.2):

$$Y(f_v, s_p) = \sum_{i=1}^{m} \left[\sum_{j=1}^{n} \alpha_{ji} \phi_j(s_p)\right] b_i(f_v) + E(f_v, s_p) \tag{6.9}$$

This approximation takes the spatial correlation into account by means of the sum of weighted Gaussian basis functions. We can hence represent Equation (6.9) in a matrix form:

$$\mathrm{Y} = \Phi \mathrm{A}\mathrm{B}^{\mathrm{T}} + \mathrm{E} \tag{6.10}$$

Where A is a $\mathrm{n} \times \mathrm{m}$ matrix describing unknown weights and Φ is a constant $\mathrm{p} \times \mathrm{n}$ matrix given by:

$$\Phi = \begin{bmatrix} \phi_1(s_1) & \phi_2(s_1) & \phi_3(s_1) & \cdots & \phi_n(s_1) \\ \phi_1(s_2) & \phi_2(s_2) & \phi_3(s_2) & \cdots & \phi_n(s_2) \\ \phi_1(s_3) & \phi_2(s_3) & \phi_3(s_3) & \cdots & \phi_n(s_3) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi_1(s_p) & \phi_2(s_p) & \phi_3(s_p) & \cdots & \phi_n(s_p) \end{bmatrix}_{p \times n} \tag{6.11}$$

This representation allows more efficient implementation of ALS algorithm. In general, the applicability of ALS algorithm to large datasets is limited due to the direct link between the complexity of the scores estimation step in the optimisation problem in Equation (6.4) and the images resolution of ToF-SIMS data. We facilitate the scores estimation step by splitting the scores matrix in Equation (6.10) into an unknown weight matrix $\mathrm{A}_{\mathrm{n}\times\mathrm{m}}$ and a constant matrix $\Phi_{\mathrm{p}\times\mathrm{n}}$. These two matrices are approximated using basis functions. This means that instead of estimating the scores matrix directly, we only need to estimate a matrix of weights with much lower dimension. The estimation of loadings matrix can be simplified in the same manner. Therefore, the implementation of basis function decomposition not only improves the convergence property of the algorithm but also reduces the uncertainty in the scores and loadings estimates.

Given the basis function representation, the cost function for the estimation of loadings matrix, B, is specified by:

$$\mathrm{J}(\mathrm{A}, \mathrm{B}) = \left\| \widetilde{\mathrm{Y}} - \mathrm{A}\mathrm{B}^{\mathrm{T}} \right\|_{\mathrm{F}}^{2} \qquad \mathrm{s.t.} \qquad \mathrm{B} > 0 \tag{6.12}$$

Where

$$\widetilde{\mathrm{Y}} = \Phi^{\dagger}\mathrm{Y}, \tag{6.13}$$

$$\Phi^{\dagger} = (\Phi^{\dagger}\Phi)^{-1}\Phi^{\mathrm{T}} \tag{6.14}$$

Where A is constant. The cost function for the estimation of the weight matrix, A, is given by:

$$\mathrm{J(A,B)} = \left\|\widetilde{\mathrm{Y}}^{\mathrm{T}} - \mathrm{BA}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} \qquad \mathrm{s.t.} \qquad \mathrm{A} > 0 \tag{6.15}$$

Where B is constant. The matrix $\widetilde{\mathrm{Y}}$ can be obtained before the algorithm is launched. Here the Frobenius norms of the successive estimates of matrix A can be observed to stop the algorithm as shown in Equation (6.5).

It is important to note that basis decomposition method might lead to smoother estimates of scores images and unclear representation of sharp boundaries and details (Aram et al., 2014). This problem can be solved by introducing an additional step to the estimation algorithm, which involves applying the final estimate of matrix B, generated from the iterative process in Equations (6.12) and (6.15), to a single run of FC-NNLS algorithm in order to optimise the following minimisation function:

$$\mathrm{J(W,B)} = \|\mathrm{Y} - \mathrm{WB}^{\mathrm{T}}\|_{\mathrm{F}}^{2} \qquad \text{subject to} \qquad \mathrm{W} > 0 \tag{6.16}$$

Where B is constant. This arrangement can result in detailed estimates of the scores images.

The complete estimation procedure is presented in the algorithm below. The minimisation of Equations (6.12), (6.15) and (6.16) can be implemented using a fast combinatorial non-negativity constrained least squares (FC-NNLS) provided by Van Benthem and Keenan (2004). Note the overwriting initialisation in Step 2 of the algorithm will not be required as it is already included in FC-NNLS (Gallagher et al, 2004). The complete algorithm is shown in following table:

The ANLS algorithm procedure
1. Decomposition: -determine number of factors m by using PCA, -define basis function centres u using Equation (6.17), -define basis function widths σ_{ϕ} using Equation (6.18), -construct $\widetilde{\mathrm{Y}}$ using Equations (6.11), (6.13) and (6.14),
2. Initialisation: -initialise the weight matrix A_0 as a random dense matrix, -obtain initialisation solution, A_0 and B_0, using the overwriting method,
3. Scores and loadings estimation: -define stopping condition threshold ρ, -set $\mathrm{k} = 1$, while $\lVert \mathrm{A}_\mathrm{k} - \mathrm{A}_{\mathrm{k}-1} \rVert_\mathrm{F} > \rho$, -update the loadings, $\mathrm{B}_{\mathrm{k}-1}$, using FC-NNLS and Equation (6.12), -update the weight matrix, A_k, using FC-NNLS and Equation (6.15), -set $\mathrm{k} = \mathrm{k} + 1$, end while
6. Estimation of high resolution scores matrices: -calculate W from Equation (6.16) using FC-NNLS and the final estimate of B.

Table 6.1 ANLS Algorithm Procedure

6.3 Spatial Frequency Analysis

The basis functions used in the model reduction procedure can be viewed as low-pass filters and limit the spatial bandwidth of the reconstructed scores images to a certain value. The width, σ_{ϕ}, and spacing, Δ_{ϕ}, of these basis functions can be obtained by setting a preferred degree of smoothness in the scores images. Therefore, the spatial cut-off frequency of the reconstructed scores images, ν_{c}, is essentially a design choice. The interested spatial region is divided by Δ_{ϕ} intervals to provide the number of basis functions. The chosen cut-off frequency determines the spacing between basis functions such that Shannon's sampling theorem is satisfied:

$$\Delta_{\phi} \leq \frac{1}{2\rho\nu_{c}} \tag{6.17}$$

Where $\rho \in \mathbb{R} \geq 1$ is an oversampling parameter (Sanner & Slotine, 1992). The spatial cut-off frequency also regulates the width of the basis functions. For an attenuation of $3\,\mathrm{dB}$ at ν_{c}, the width of basis functions can be described by (Freestone et al., 2011):

$$\sigma_{\phi} = \frac{1}{\pi\nu_{c}}\sqrt{\frac{\mathrm{In}2}{2}} \tag{6.18}$$

Where ν_{c} can be set to a high value in order to capture high spatial frequency variations in the scores estimates. However, as the reciprocal role of ν_{c} shown in Equations (6.17) and (6.18), a large number of basis functions with narrow widths can be caused by a high cut-off frequency. This complicates the estimation procedure as it needs to estimate more weights that are associated with the number of basis functions. Thus, there is a trade-off between the accuracy and the computational demands of the estimation procedure.

Taking the dataset used in this work for instance, the decomposition of $128 \times$

128×100 ToF-SIMS data with 4×4 equally spaced grids of basis functions can be performed as shown in Figure 6.1.

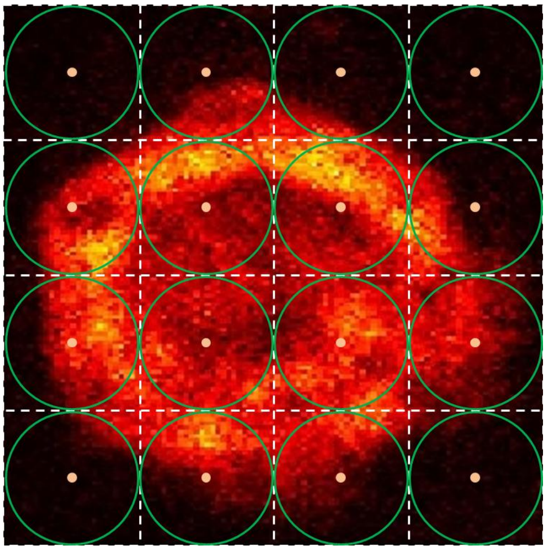


Figure 6.1 Example of a basis decomposition. A 128 pixels by 128 pixels ToF-SIMS image decomposed by a 4×4 grid of basis functions. The basis functions (shown by green circles) are scaled by the weight matrix, A. The centre of each basis function is shown by a yellow dot. The image is mapped onto -1 to 1 with arbitrary units.

The green grids are Gaussian basis functions scaled by the weight matrix A in Equation (6.10) to decompose the observed surface. The yellow dot indicates the centre of each basis function. The image is mapped onto [-1 to 1] with arbitrary units, the estimation of 16384×100 parameters is reduced to the estimation of 16×100 parameters in this particular example.

6.4 Application Results

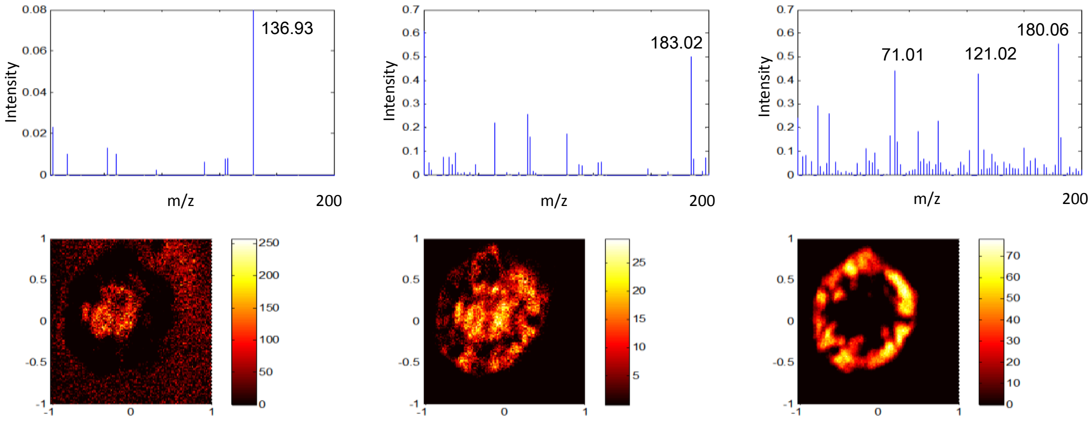
The proposed algorithm was applied to our ToF-SIMS datasets containing three pure species (T, P and C) and two mixed species (TC and TPC). Each one of those datasets includes images of 128×128 pixels with the spectra up to 100 Da.

Significant features were firstly identified from the estimated scores and loadings by applying the proposed algorithm to pure species. Subsequently, peak assignment in the resolved spectra of TC and TPC mixtures was performed using the extracted information. Three major peaks were found using the extracted spectral information in order to describe different datasets. We then assessed the performance of our algorithm by analysing the replicate measurements of each dataset.

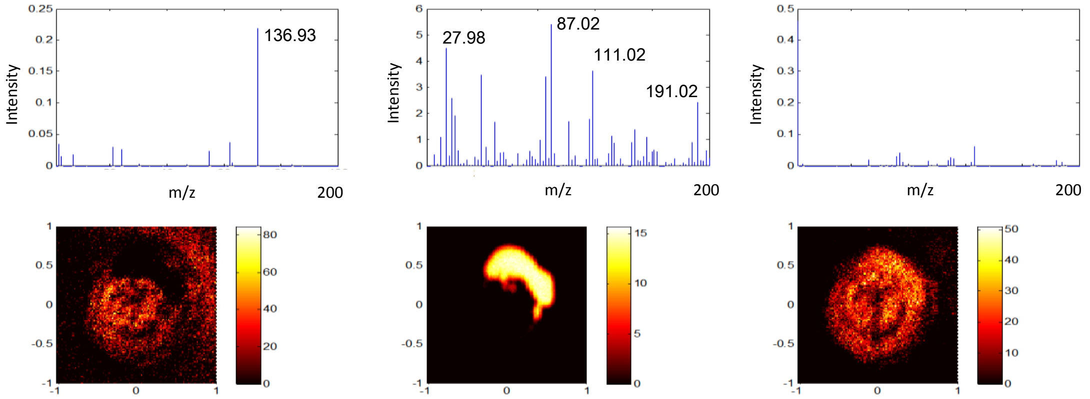
The spatial aspects of the model can be considered arbitrary as ToF-SIMS images were mapped in both x and y directions. The desired cut-off frequency of the reconstructed images was set to $\nu_c = 0.84$; we also set the oversampling parameter of $\rho = 2$ after considering the slow roll-off in the frequency response of Gaussian basis functions. These values were then applied to Equations (6.17) and (6.18), which provided the distance between the centres of adjacent basis functions $\Delta_\phi = 0.3$ and the width of basis functions $\sigma_\phi = 0.22$, leading to a grid of 9×9 equally spaced basis functions in the spatial domain of interest. Under this arrangement, we were able to reduce the number of unknown parameters from $\mathrm{m} \times 16384$ to $\mathrm{m} \times 81$.

In the pre-processing stage, we employed Poisson-scaling and normalising to the total ion counts for the chemical rank analysis and the estimation of scores and loadings, respectively. The initial guess of the chemical rank for each Poisson-scaled dataset was derived from PCA and the scree test. PCA was performed for all the five datasets in Chapter 3. Although the suggested rank of TC mixture was one in the PCA analysis, $\mathrm{m} = 3$ was set for all pure species and mixtures since all the other species have three principal components. The estimation of loadings and the corresponding scores images involved applying the algorithm to normalised ToF-SIMS datasets.

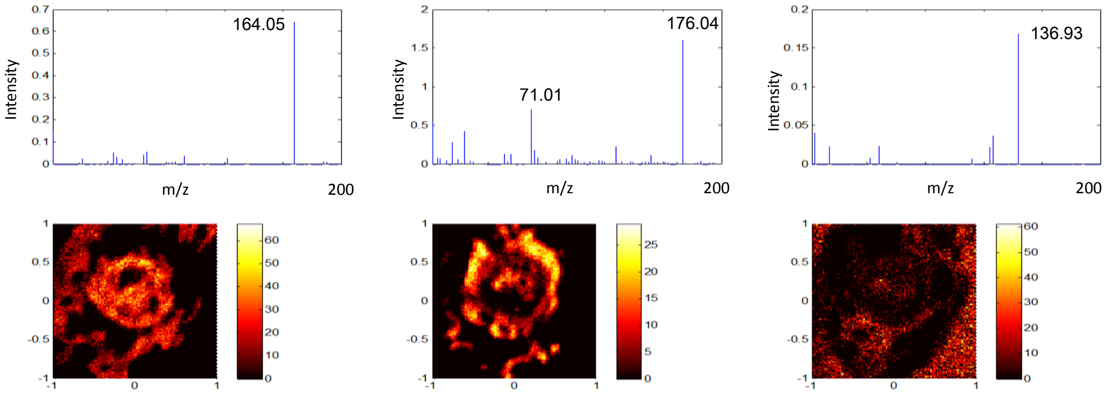
Figure 6.2 shows the results of scores and loadings estimation for T, C and P components. The scores images show the m/z values for dominant peaks, following by figures that represent the corresponding ion images in ToF-SIMS dataset. Large peaks for the first three factors of component T are found at m/z = 136.93, 183.02, 71.01, 121.02, and 180.06, as shown in Figure 6.2(T). Significant peaks for the first three factors of component C and P are illustrated in Figure 6.2(C) and Figure 6.2(P) respectively. The component C has peaks at m/z = 136.93, 27.98, 87.02, 111.02, and 191.02, while peaks at m/z = 164.05, 176.04, 71.01, and 136.93 refer to component P. It should be noted that all the three pure species contain peak at m/z = 136.93. This common peak can be viewed as noise in the system due to lack of discriminatory information. In fact, noisy structures do appear in the corresponding scores images. One possible explanation is that we decomposed the data using higher number factors than required.



(T)



(C)



(P)

Figure 6.2 Three sets of loading plots and scores images from ANLS algorithm for each pure component (T, C, and P) samples respectively. Loadings are presented in top panels indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis.

Figure 6.3 represents the result of the estimation of scores and loadings for TC mixture and the corresponding scores images for each component is illustrated down below. Again, the common peak at m/z = 136.93 is present, which is attributed to the existence of the noise in the system. It is clear that there are two significant peaks at m/z = 87.02 and 191.02 for the first factor in Figure 6.3, we may conclude that it refers to the component C by comparing the peaks in Figure 6.2(C). Similarly, the peaks at m/z = 71.01 and 180.06 for the second factor of TC mixture can be attributed to component T. In addition, an intensive peak is found at m/z = 136.93 through the third factor, it could possibly be due to the noise from the data collection process of ToF-SIMS. In this case, our algorithm performed reasonably well and was found to be effective in separating the distribution of the two pure species (T and C).

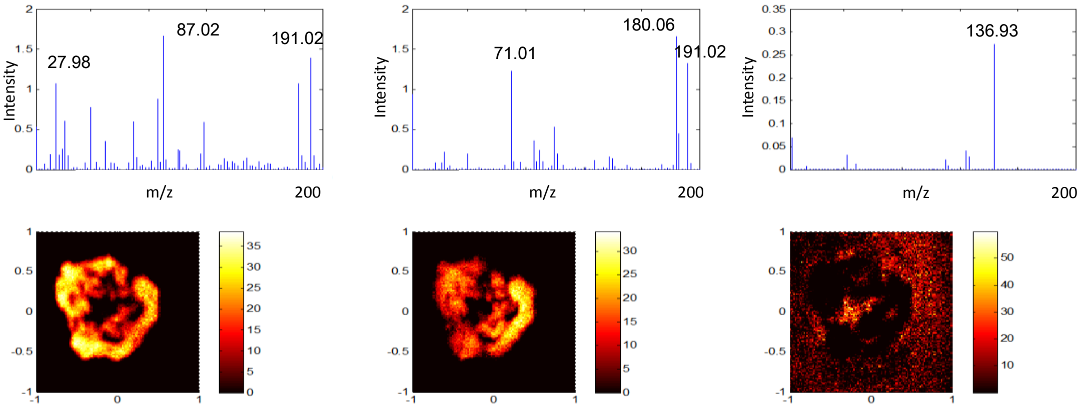


Figure 6.3 Three sets of loading plots and scores images using ANLS for TC mixture. Loadings are presented in the top panels indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis.

Figure 6.4 illustrates the results of the scores and loadings estimation for TPC mixture. As shown in the first loading estimates, the algorithm is capable of identifying the separate distribution of component T with peaks at m/z = 71.01 and m/z = 180.06 in this particular case. P species can be identified in the second graph with one significant peak at m/z = 164.05. Furthermore, peaks at m/z = 41.01, 87.02 and 191.02 in Figure 6.3 suggest that we can identify and segregate component C from the mixture. Despite of the correct number of chemical factors deployed, noise characteristics still exist in the third factor of TPC mixture, which can be identified clearly in the last column graphs of Figure 6.3. This is due to the fragments in the separation of the three pure species.

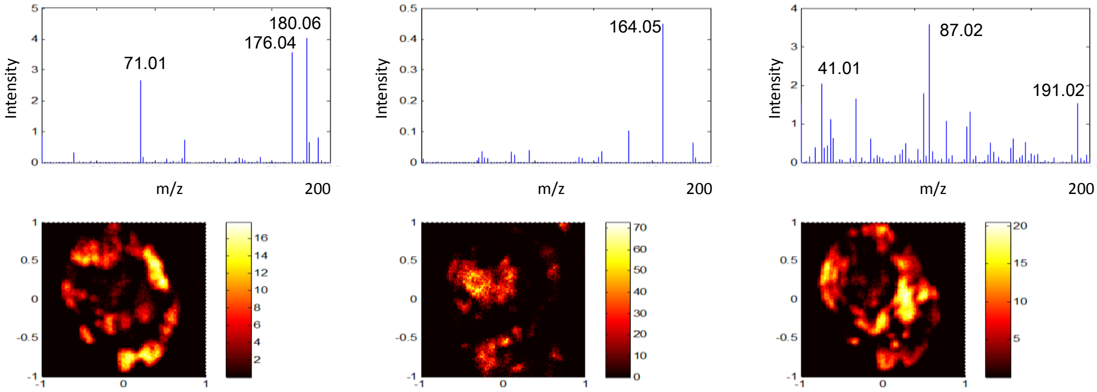
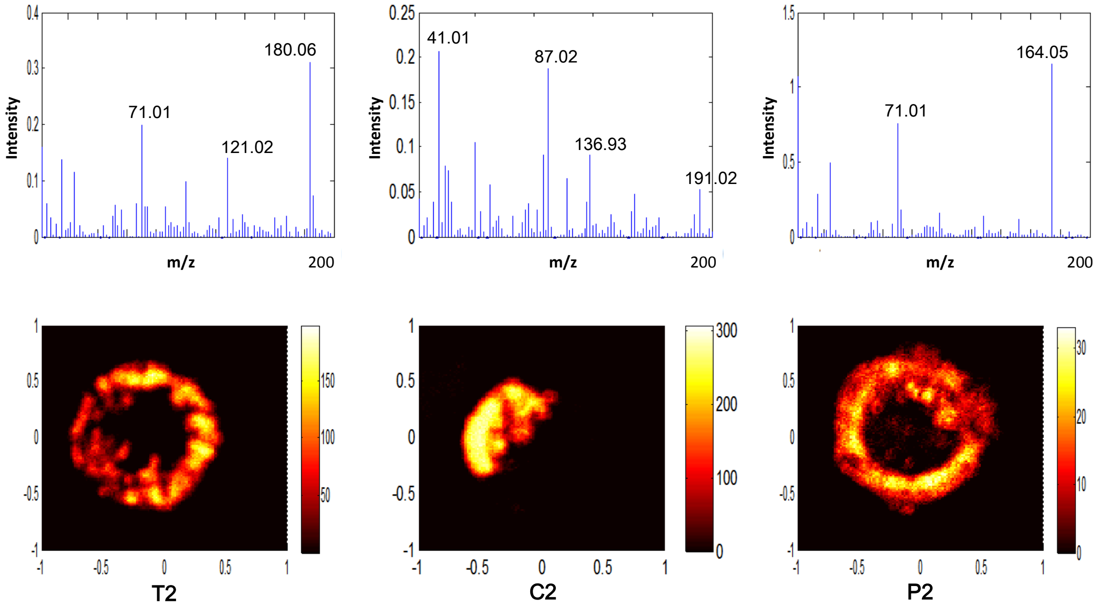


Figure 6.4 Three sets of loading plots and scores images using ANLS for TPC mixture. Loadings are presented in the top panels indicating the intensity of spectral basis while score images are presented on the bottom showing the spatial information of each basis.

The important spectral information required for peak assignment can then be summarised. The results of our application of the proposed algorithm suggest that peaks at m/z = 180.06, 164.05 and 191.02 are essential for identifying component T, P and C respectively, which is in full accordance with the ground truth. Note that the strong peak at m/z = 121.02 for component T and at m/z = 111.02 for component C are attributed to fragments or adducts in the process as they are not present in the analysis of TC and TPC mixtures. In addition, although the peak at m/z = 191.02 for component C in Figure 6.2(C) also appears in scores image of TC and TPC mixtures, the fragment peak of species C at m/z = 87.02 is more significant with greater magnitude.

Similar analysis on replicate measurements of the five species was performed using the extracted spectral information. In particular, m = 1, 2 and 3 was set for pure replicate measurements, TC replicate measurements and TPC replicate measurements respectively. The results of the analysis are summarised below which confirm the capability of our proposed algorithm in identifying different species. We tested the algorithm for a maximum number of 200 iterations, the change in the weights matrix Frobenius norm was reduced to lower than 10^{-5} after as much as 20 iterations. Compare with other multivariate techniques we implemented previously in Chapter 2 and Chapter 3, it is evident that this ANLS algorithm can separate discriminatory components with significantly better execution time.

Figure 6.5 represents the results of each individual pure component with factor number m = 1. Component T, C and P can be recognised easily with their identifiable peaks from left to right respectively. With m = 2 applying to the replicate TC mixture, we can conclude that the mixture is made of component T and C with peaks at m/z = 71.01 and 180.06, and at m/z = 27.98, 87.02, 111.02, and 191.02 (Figure 6.6). For the replicate measurements of TPC mixture, m = 3 was set and pure component T, P and C could be identified with the loading images in Figure 6.5. It should be noted that there is one remarkable peak in the second scores image of Figure 6.5, this might be due to that both component T and P have the same peaks at m/z = 71.01, where the algorithm is incapable of separating the two identical location peaks. Another possible reason is that this peak could be resulted from the fragments or adducts in the estimation process. However, the purposed algorithm performs well in detecting and separating the distribution of the three pure species (T, P and C).



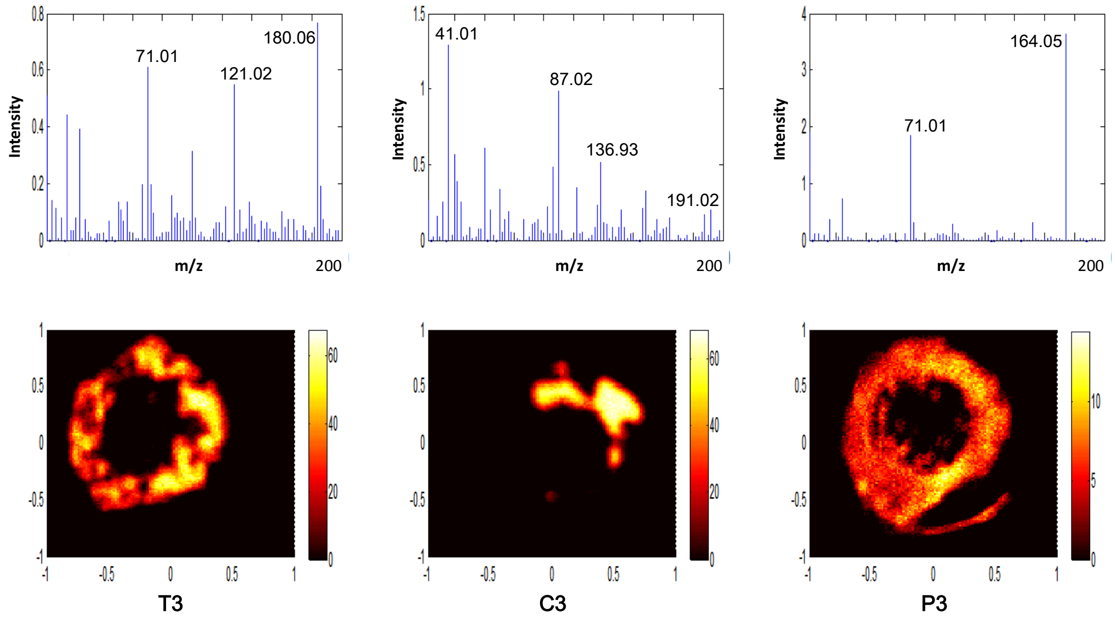
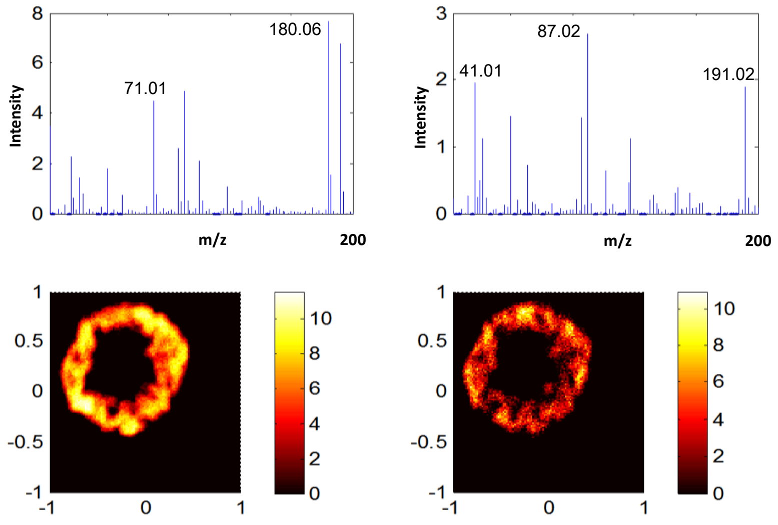
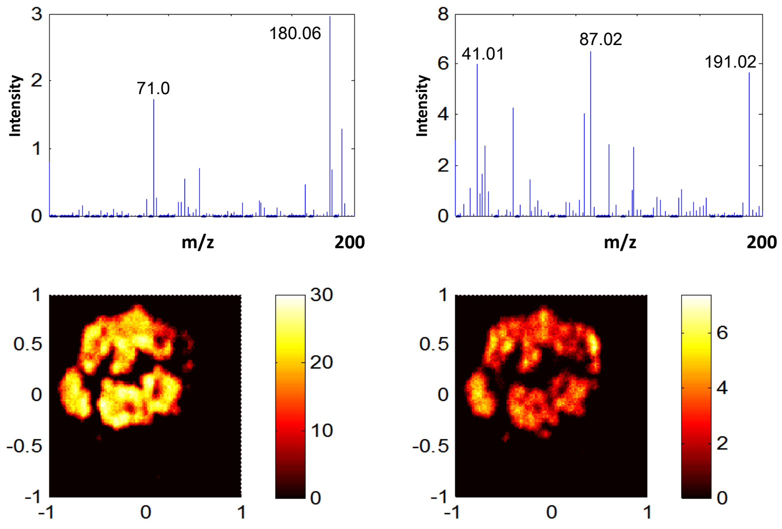


Figure 6.5 loading plots and scores images using ANLS for single species T, P and C replicate samples with one basis.

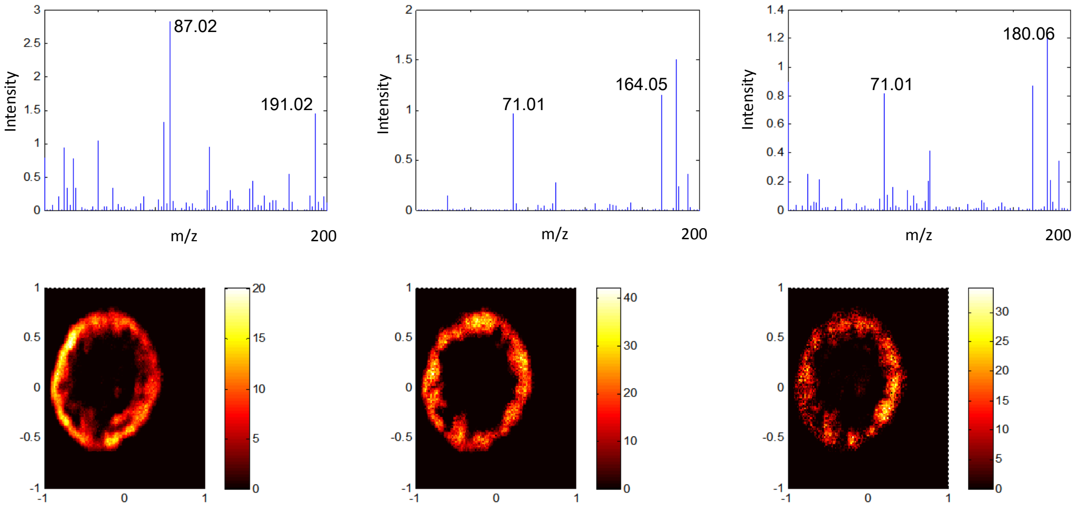


(TC2)

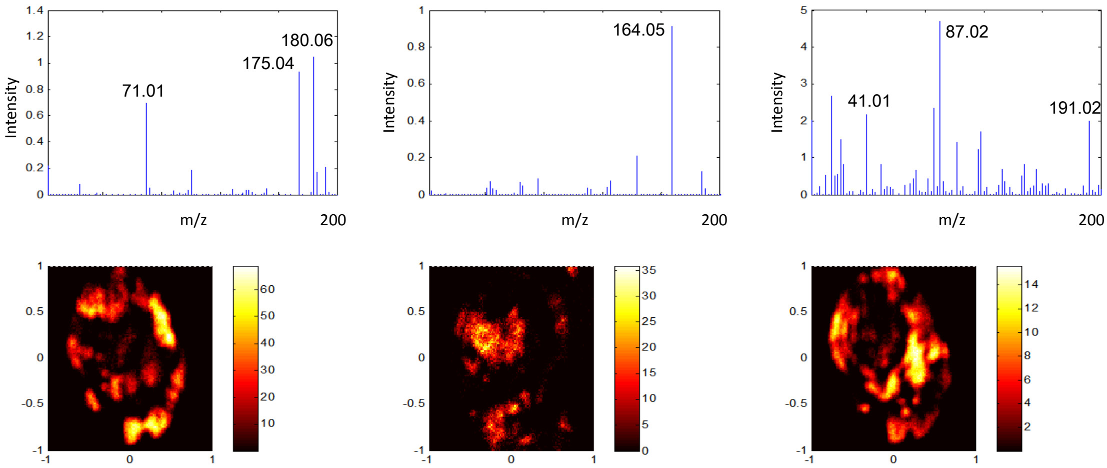


(TC3)

Figure 6.6 Two sets of loading plots and scores images using ANLS for replicate TC mixture samples with two basis.



(TPC2)



(TPC3)

Figure 6.7 Two sets of loading plots and scores images using ANLS for replicate TPC mixture samples.

6.5 Conclusion

Our proposed algorithm for ToF-SIMS data analysis is an MCR method built on the NMF framework. This novel algorithm provides great potential to be used as an efficient tool in processing ToF-SIMS data from metabolite samples. One of the key features of our proposed algorithm is that it incorporates spatially continuous representation of ToF-SIMS dataset by employing a set of continuous basis functions to reconstruct the scores images. This leads to a simplified estimation procedure where only a set of weights are required to approximate the scores images, significantly reducing the spatial dimensions (Aram et al., 2014). Therefore, the algorithm maintains the advantages of ordinary MCR-ALS algorithm, such as identification of pure component spectra and the ability to incorporate known information into the estimation procedure while offering non-negative solutions with reduced computational demand. Furthermore, compared with PCA, the factors in the ANLS are not required to be orthogonal, the calculated solutions resemble the ToF-SIMS data and contribution of chemical components in an effective manner which makes the results more interpretable.

However, it should be noted that, instead of depending on the image resolution, the computational complexity is now related to the spatial frequency of the approximated scores images. This means that by lowering the spatial bandwidth of the reconstructed images, the computation complexity of the estimation procedure can be reduced at the cost of the accuracy.

Chapter 7

Conclusion and Future Work

ToF-SIMS is an advanced chemical analysis platform that is largely new to metabolic profiling (Armitage et al., 2013). Although it is a powerful and information rich tool with high resolution compared to conventional MS, the substantially large and complex output data already becomes a major obstacle to its utility and applicability (Graham, Wagner, & Castner, 2006; Tyler, Rayal, & Castner, 2007). This emphasises the importance of more efficient multivariate algorithms for analysis of such data. The aim of this thesis was to develop and validate novel multivariate analysis techniques for processing ToF-SIMS data extracted from metabolite samples.

In this thesis, we discussed five unsupervised multivariate analysis methods, all of which are capable of decomposing the original complex ToF-SIMS dataset into smaller and simpler matrices while the main features of the data are retained. The traditional multivariate analysis methods, due to their limitations, can be ineffective in processing large and complex ToF-SIMS datasets. In particular, PCA is frequently used with ToF-SIMS to identify chemical compounds in metabolite samples

(Henderson, Fletcher, & Vickerman, 2009; Kotze et al., 2013). Our application of PCA by SVD to ToF-SIMS data extracted from metabolite samples also showed poor performance. Other well-known algorithms, namely, Clustering, MAF, ICA and MCR, were also outlined in the thesis.

The application of PCA to three pure species (T, P, and C) showed that, although they are single components, at least three principal components are required to represent 90% of the original datasets in each individual case. The TPC mixture also had three principal components in the result. It is important to note that only one principal component was required for the TC mixture, which was not appropriate as based on the ground truth of our data, with two mixed species contained in the actual mixture. The scores images of the five species also suggested that negative values are present in the results. Despite the limited performance of the PCA, it is still useful for providing some insights into the identification of chemical compounds in the sense that the principal components obtained can be combined with the prior knowledge of the data to offer a better initialisation for other algorithms.

PCA can be compared to NMF algorithm, which is a reduced rank approximate factorisation maintaining non-negative structure of the data matrix. The major contribution of this method is that it provides a more realistic interpretation of the data. The combination of PCA and known information about the structure of species was used as the starting point for NMF. With enough iterations, NMF is able to extract the factor with identical peaks for each single species. However, the application of the algorithm with mixture samples are not sufficiently satisfactory since the chemical components T, P and C do not manifest as individual factors, in other words, each factor has a spectral pattern that consists of two or more species. The reason may be due to the underlying uncertainties that largely exist in the data. In this thesis, we also introduced NMF with other auxiliary constraints. With sparsity constraint, the NMF results can be more powerful in detecting the intensity peaks as well as the spatial regions for one specific factor.

One of the drawbacks of NMF algorithms is that the iterative procedure leads to considerable computational complexity and hence obstructs convergence. Therefore, instead of using the PCA or the ground truth of the dataset, we used the NMF algorithm under the Bayesian framework, which can address the problem of multiple solutions in NMF while also improving the convergence process. Our application of the algorithm to the ToF-SIMS dataset suggested that it was reasonably effective in identifying components in a two component mixture (TC). However, the B-NMF is unable to identify species which may be attributed to not exploiting the spatial correlation in the data. In particular, components T and P were incorrectly classified into one single factor by the B-NMF algorithm in our results of the TPC mixture. In addition, the score images of the TPC mixture confirmed that these two components are spatially close to each other. As a result, the overlapping part that was represented by the common peaks of component T and P is difficult to distinguish by the B-NMF, leading to an inconclusive identification. Therefore, the B-NMF might not be an ideal method for processing ToF-SIMS data in metabolic profiling analysis, where the species distribution in metabolite samples are continuous in nature.

In this thesis, we proposed an optimised MCR method built on the ANLS procedure. This novel algorithm provides good potential to be used as an efficient tool in the processing of ToF-SIMS data from metabolite samples. One of the key features of the proposed algorithm is that it incorporates spatially continuous representation of ToF-SIMS dataset by employing a set of continuous basis functions to reconstruct the scores images. This leads to simplified estimation procedure where only a set of weights are required to approximate the scores images, significantly reducing the spatial dimension. Therefore, the algorithm maintains the advantages of ordinary MCR-ALS algorithm, such as the identification of pure component spectra and the ability to incorporate known information into the estimation procedure, while offering non-negative solutions with reduced computational demand. This algorithm also requires an initial estimate, which was obtained from the PCA, B-NMF and the ground truth for our specific case. The application of ANLS to our ToF-SIMS dataset suggested that this algorithm is fairly efficient in identifying and separating all the pure components from both TC and TPC mixtures. Furthermore, it is evident that the proposed ANLS variant can separate discriminatory components with a considerable computational speed over the other multivariate techniques analysed in this work. However, it should be noted that, instead of depending on the image resolution, the computational complexity is now related to the spatial frequency of the approximated scores images. This means that by lowering the spatial bandwidth of the reconstructed images, the computational complexity of the estimation procedure can be reduced at the cost of the accuracy.

From the investigation in the previous chapters, comparison can be made by the four different multivariate analysis techniques used on the same metabolic profiling datasets from ToF-SIMS. One conclusion is that all the algorithms can complete unsupervised feature detection and extraction to varying degrees. In the case of the TPC sample where the correct profile number is three, all of the above algorithms when set to identify three factors had different limitations in the results that ensued. The comparison showed that the ANSL outperforms other studied algorithms.

It is important to note that our metabolite samples only contained five species with three pure components. In realistic cases, metabolites information may be more varied and can substantially affect the effectiveness of the algorithms. There are libraries of MS data for different metabolites that can help mitigate the identification problem. The nature of the dataset used in the thesis did not require the imposition of additional conditions and assumed little prior knowledge. However, the inclusion of libraries of metabolites and their spectral patterns are likely to substantially improve the effectiveness of the algorithms. Therefore, in order to complete such a challenge, the available library of ToF-SIMS data must somehow be integrated. Though the present peptide library is limited, as still there are many peptides with unknown ToF-SIMS patterns, factorisation methods should be developed to detect those peptides in the metabolites that have known spectral patterns in the library in addition to identifying unknown peptides. One of the other attributes of the spectral patterns is the fact that the abundance as measured by the height of the spectral peaks includes quantities of metabolites that can be a distraction to the identification process. A classification approach in which prioritisation is given to the location of the important spectral peaks is worth investigating, so that sensitivities to absolute differences in the spectral patterns do not skew the performance of metabolic profiling. To summarise, more advanced analysis methods for the ToF-SIMS data are required to disentangle the complex and high dimensional spatial data meaningfully, to achieve sufficiently accurate metabolic profiling. The promise of ToF-SIMS process crucially depends on this.

Reference List

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (pp. 199-213). Springer New York.Albert, J., & Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *92*(439), 916-925.

Albright, R., Cox, J., Duling, D., Langville, A. N., & Meyer, C. (2006). *Algorithms, initializations, and convergence for the nonnegative matrix factorization* (p. 5). Tech. rep. 919. NCSU Technical Report Math 81706.

Amari, S. I., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, pp. 757-763, Cambridge, MA: MIT Press.

Antal, E., & Tillé, Y. (2011). Simple random sampling with over-replacement. *Journal of Statistical Planning and Inference*, *141*(1), 597-601.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, *50*(1-2), 5-43..

Aoyagi, S., Fletcher, J. S., Sheraz, S., Kawashima, T., Razo, I. B., Henderson, A., Lockyer, N. P., & Vickerman, J. C. (2013). Peptide structural analysis using continuous Ar cluster and C60 ion beams. *Analytical and bioanalytical chemistry*, *405*(21), 6621-6628.

Aram, P., Shen, L., Pugh, J. A., Vaidyanathan, S., & Kadirkamanathan, V. (2014). An efficient ToF-SIMS image analysis with spatial correlation and alternating non–negativity-constrained least squares. *Bioinformatics*, btu734.

Armitage, E. G., Kotze, H. L., Fletcher, J. S., Henderson, A., Williams, K. J., Lockyer, N. P., & Vickerman, J. C. (2013). Time-of-flight SIMS as a novel approach to unlocking the hypoxic properties of cancer. *Surface and Interface Analysis*, *45*(1), 282-285.

Balmer, D., Flors, V., Glauser, G., & Mauch-Mani, B. (2013). Metabolomics of cereals under biotic stress: current knowledge and techniques. *Frontiers in plant science*, *4*.

Bayliss, J. D., Gualtieri, J. A., & Cromp, R. F. (1998, March). Analyzing hyperspectral data with independent component analysis. In *26th AIPR Workshop: Exploiting New Image Sources and Sensors* (pp. 133-143). International Society for Optics and Photonics.Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*, *2*(11), 2692-2703.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7*(6), 1129-1159.

Belu, A. M., Graham, D. J., & Castner, D. G. (2003). Time-of-flight secondary ion mass spectrometry: techniques and applications for the characterization of biomaterial surfaces. *Biomaterials*, *24*(21), 3635-3653.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, *52*(1), 155-173.

Bertsekas, D. P. (1982). Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on control and Optimization*, *20*(2), 221-246.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Biesinger, M. C., Paepegaey, P. Y., McIntyre, N. S., Harbottle, R. R., & Petersen, N. O. (2002). Principal component analysis of TOF-SIMS images of organic monolayers. *Analytical chemistry*, *74*(22), 5711-5716..

Bos, C. S. (2002, January). A comparison of marginal likelihood computation methods. In *Compstat* (pp. 111-116). Physica-Verlag HD

Boxer, S. G., Kraft, M. L., & Weber, P. K. (2009). Advances in imaging secondary ion mass spectrometry for biological samples. *Annual Review of Biophysics*, *38*, 53-74.

Brereton, R. G. (Ed.). (1992). *Multivariate pattern recognition in chemometrics: illustrated by case studies*. Elsevier.Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural computation*, *11*(1), 157-192.

Cattell, R. B. (1943). The description of personality: basic traits resolved into clusters. *The journal of abnormal and social psychology, 38*(4), 476.

Chen, J., & Wang, X. Z. (2001). A new approach to near-infrared spectral data analysis using independent component analysis. *Journal of Chemical Information and Computer Sciences*, *41*(4), 992-1001.Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 267-275.

Chen, Z. P., Jiang, J. H., Li, Y., Shen, H. L., Liag, Y. Z., & Yu, R. Q. (1999). Smoothed window factor analysis. *Analytica chimica acta, 381*(2), 233-246.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270-281.

Choi, B. K., Hercules, D. M., Zhang, T., & Gusev, A. I. (2003). Comparison of quadrupole, time-of-flight, and Fourier transform mass analyzers for LC-MS applications. *Current Trends in Mass Spectrometry*, *18*(5S), 524-531.

Chu, M., Diele, F., Plemmons, R., & Ragni, S. (2004). Optimality, computation, and interpretation of nonnegative matrix factorizations. In *SIAM Journal on Matrix Analysis*.

Cichocki, A., Zdunek, R., & Amari, S. I. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Independent Component Analysis and Blind Signal Separation* (pp. 32-39). Berlin Heidelberg: Springer.

Clarke, C. J., & Haselden, J. N. (2008). Metabolic profiling as a tool for understanding mechanisms of toxicity. *Toxicologic pathology*, *36*(1), 140-147.

Comon, P. (1994). Independent component analysis, a new concept?. *Signal processing*, *36*(3), 287-314.

Cotter, R. J. (2011). Time-of-Flight Mass Spectrometry. In Cole, R. B. (Ed.), *Electrospray and MALDI mass spectrometry: fundamentals, instrumentation, practicalities, and biological applications* (2nd ed., pp. 345-364). John Wiley & Sons.

de Juan, A., & Tauler, R. (2006). Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. *Critical Reviews in Analytical Chemistry*, *36*(3-4), 163-176.

de Juan, A., Jaumot, J., & Tauler, R. (2014). Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods, 6*(14), 4964–4976.

DiCiccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, *92*(439), 903-915.

Downard, K. M. (2012). 1912: a titanic year for mass spectrometry. *Journal of Mass Spectrometry*, *47*(8), 1034-1039.

Dubey, M., Brison, J., Grainger, D. W., & Castner, D. G. (2011). Comparison of Bi1+, Bi3+ and C60+ primary ion sources for ToF‐SIMS imaging of patterned protein samples. *Surface and Interface Analysis*, *43*(1‐2), 261-264.Févotte, C.,

& Cemgil, A. T. (2009, August). Nonnegative matrix factorizations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)* (pp. 1913-1917).

Freestone, D. R., Aram, P., Dewar, M., Scerri, K., Grayden, D. B., & Kadirkamanathan, V. (2011). A data-driven framework for neural field modeling. *NeuroImage*, *56*(3), 1043-1058.

Gallagher, N. B., Shaver, J. M., Bishop, R., Roginski, R. T., & Wise, B. M. (2014). Decompositions using maximum signal factors. *Journal of Chemometrics, 28*(8), 663–671.

Gallagher, N. B., Shaver, J. M., Martin, E. B., Morris, J., Wise, B. M., & Windig, W. (2004). Curve resolution for multivariate images with applications to ToF-SIMS and Raman. *Chemometrics and intelligent laboratory systems*, *73*(1), 105-117.

Gonzalez, E. F., & Zhang, Y. (2005). Accelerating the Lee-Seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*.

Graham, D. J., Wagner, M. S., & Castner, D. G. (2006). Information from complexity: challenges of ToF-SIMS data interpretation. *Applied surface science*, *252*(19), 6860-6868.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.Guillamet, D., Bressan, M., & Vitria, J. (2001). A weighted non-negative matrix factorization for local representations. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. I-942). IEEE.

Harva, M., & Kabán, A. (2007). Variational learning for rectified factor analysis.*Signal Processing*, *87*(3), 509-527.

Haykin, S. (2009). *Neural Networks and Learning Machines*. New Jersey: Pearson

Education, Inc.

Henderson, A. (2013). Multivariate Analysis of SIMS Spectra. In J. C. Vickerman & D. Briggs (Eds.), *ToF-SIMS: Materials Analysis by Mass Spectrometry* (2nd ed., pp. 449-484). Manchester, UK: IM Publications.

Henderson, A., Fletcher, J. S., & Vickerman, J. C. (2009). A comparison of PCA and MAF for ToF-SIMS image interpretation. *Surface and Interface Analysis*, *41*(8), 666-674.

Herault, J., & Jutten, C. (1986, August). Space or time adaptive signal processing by neural network models. In *Neural networks for computing* (Vol. 151, No. 1, pp. 206-211). AIP Publishing.

Horning, E. C., Devaux, P. G., Moffat, A. C., Pfaffenberger, C. D., Sakauchi, N., & Horning, M. G. (1971). Gas phase analytical separation techniques applicable to problems in clinical chemistry. *Clinica Chimica Acta*, *34*(2), 135-144.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, *5*, 1457-1469.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*(4), 411-430.

Ibáñez, C., García-Cañas, V., Valdés, A., & Simó, C. (2013). Novel MS-based approaches and applications in food metabolomics. *TrAC Trends in Analytical Chemistry*, *52*, 100-111.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jiang, J. H., Liang, Y., & Ozaki, Y. (2004). Principles and methodologies in self-modeling curve resolution. *Chemometrics and intelligent laboratory* *systems*, *71*(1), 1-12.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Keenan, M. R., & Kotula, P. G. (2005). Recent Developments in Automated Spectral Image Analysis. *Microscopy and Microanalysis*, *11*(S02), 36-37.

Keenan, M. R., & Smentkowski, V. S. (2011). Simple statistically based alternatives to MAF for ToF-SIMS spectral image analysis. *Surface and Interface Analysis*, *43*(13), 1616-1626.

Kim, D., Sra, S., & Dhillon, I. S. (2007, April). Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem. In *SDM* (Vol. 7, pp. 343-354).

Kim, J., & Park, H. (2008, December). Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 353-362). IEEE.

Knuth, K. H. (2005). Informed source separation: A Bayesian tutorial.

Kotze, H. L., Armitage, E. G., Fletcher, J. S., Henderson, A., Williams, K. J., Lockyer, N. P., & Vickerman, J. C. (2013). ToF-SIMS as a tool for metabolic profiling small biomolecules in cancer systems. *Surface and Interface Analysis*, *45*(1), 277-281.

Lachenmeier, D. W., & Kessler, W. (2008). Multivariate curve resolution of spectrophotometric data for the determination of artificial food colors. *Journal of agricultural and food chemistry*, *56*(14), 5463-5468.

Langlois, D., Chartier, S., & Gosselin, D. (2010). An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative*

Methods for Psychology, *6*(1), 31-38.

Larsen, R. (2002). Decomposition using maximum autocorrelation factors. *Journal of Chemometrics*, *16*(8-10), 427-435.

Lawton, W. H., & Sylvestre, E. A. (1971). Self modeling curve resolution. *Technometrics*, *13*(3), 617-633.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562). MIT Press.

Lee, J. L. S., Gilmore, I. S., Fletcher, I. W., & Seah, M. P. (2009). Multivariate image analysis strategies for ToF-SIMS images with topography. *Surface and Interface Analysis*, *41*(8), 653-665.

Leefmann, T., Heim, C., Kryvenda, A., Siljeström, S., Sjövall, P., & Thiel, V. (2013). Biomarker imaging of single diatom cells in a microbial mat using time-of-flight secondary ion mass spectrometry (ToF-SIMS). *Organic Geochemistry*, *57*, 23-33.

Lin, C. J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation*, *19*(10), 2756-2779.

Linsker, R. (1992). Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, *4*(5), 691-702.Lohr, S. (2009). *Sampling: design and analysis* (2nd ed.). USA: Cengage Learning.

Maeder, M., & Neuhold, Y. M. (2007). *Practical data analysis in chemistry* (Vol. 26). Elsevier.

Malinowski, E. R. (1992). Window factor analysis: theoretical derivation and application to flow injection analysis data. *Journal of chemometrics, 6*(1), 29-40.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*(4), 831-860

Michalski, R. S., & Stepp, R. E. (1983). Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *5*(4), 396-410.

Miura, D., Fujimura, Y., Tachibana, H., & Wariishi, H. (2009). Highly sensitive matrix-assisted laser desorption ionization-mass spectrometry for high-throughput metabolic profiling. *Analytical chemistry*, *82*(2), 498-504.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(1), 111–126.

Passarelli, M. K., & Winograd, N. (2011). Lipid imaging with time-of-flight secondary ion mass spectrometry (ToF-SIMS). *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, *1811*(11), 976-990.

Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, *416*(1), 29-47.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Peters, T. J., & Eachus, J. I. (1995). Achieving equal probability of selection under various random sampling strategies. *Paediatric and perinatal epidemiology*, *9*(2), 219-224.

Pham, D. T., Garrat, P., & Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. *In Proceedings of EUSIPCO*, pp. 771–774.

Polani, D. (2013). Kullback-Leibler Divergence. In Werner, D., Olaf, W., Kwang-Hyun, C., & Hiroki, Y. (Eds.), *Encyclopedia of Systems Biology* (pp. 1087-1088). New York: Springer.

Reed, N. M., & Vickerman, J. C. (1993). The application of static secondary ion mass spectrometry (SIMS) to the surface analysis of polymer materials. In Sabbatini, L., & Zambonin, PG. (Eds.), *Surface characterization of advanced polymers*. Weinheim, Germany: VCH.

Salim, M., Wright, P. C., & Vaidyanathan, S. (2012). A solvation-based screening approach for metabolite arrays. *Analyst*, *137*(10), 2350-2356.

Sanner, R. M., & Slotine, J. J. (1992). Gaussian networks for direct adaptive control. *IEEE Transactions on Neural Networks*, *3*(6), 837-863.

Schachtner, R., Poeppel, G., Tomé, A. M., & Lang, E. W. (2014). A Bayesian approach to the Lee–Seung update rules for NMF. *Pattern Recognition Letters*,*45*, 251-256.

Schachtner, R., Pöppel, G., & Lang, E. W. (2010, June). Bayesian extensions of non-negative matrix factorization. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (pp. 57-62). IEEE.

Schmidt, C. W. (2004). Metabolomics: what's happening downstream of DNA. *Environmental Health Perspectives*, *112*(7), A410.

Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation* (pp. 540-547). Berlin Heidelberg: Springer.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298-321.

Smentkowski, V. S., Ostrowski, S. G., & Keenan, M. R. (2009). A comparison of multivariate statistical analysis protocols for ToF-SIMS spectral images. *Surface and Interface Analysis*, *41*(2), 88-96.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3-23.

Sodhi, R. N. (2004). Time-of-flight secondary ion mass spectrometry (ToF-SIMS):—versatility in chemical and imaging surface analysis. *Analyst*, *129*(6), 483-487.

Storvik, G. (1993). Data reduction by separation of signal and noise components for multivariate spatial images. *Journal of applied statistics*, *20*(1), 127-136.

Switzer, P., & Green, A. A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. *Computer Science and Statistics: The Interface (L. Billard, Ed.)*, 16.

Switzer, P., & Ingebritsen, S. E. (1986). Ordering of time-difference data from multispectral imagery. *Remote Sensing of Environment*, *20*(1), 85-94.

Tauler, R., Kowalski, B., & Fleming, S. (1993). Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical chemistry*, *65*(15), 2040-2047.

Tyler, B. J. (2006). Multivariate statistical image processing for molecular specific imaging in organic and bio-systems. *Applied surface science*, *252*(19), 6875-6882.

Tyler, B. J., Rayal, G., & Castner, D. G. (2007). Multivariate analysis strategies for processing ToF-SIMS images of biomaterials. *Biomaterials*, *28*(15), 2412-2423.

Vaidyanathan, S., Fletcher, J. S., Goodacre, R., Lockyer, N. P., Micklefield, J., & Vickerman, J. C. (2008). Subsurface biomolecular imaging of Streptomyces coelicolor using secondary ion mass spectrometry. *Analytical chemistry*, *80*(6), 1942-1951.

Valle, S., Li, W., & Qin, S. J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research*,*38*(11), 4389-4401.

Van Benthem, M. H., & Keenan, M. R. (2004). Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of chemometrics*, *18*(10), 441-450.

Vickerman, J. C., & Briggs, D. (2001). *ToF-SIMS: surface analysis by mass spectrometry*. IM: Chichester, UK.

Wagner, M. S., & Castner, D. G. (2001). Characterization of adsorbed protein films by time-of-flight secondary ion mass spectrometry with principal component analysis. *Langmuir*, *17*(15), 4649-4660.

Wang, J. H., Hopke, P. K., Hancewicz, T. M., & Zhang, S. L. (2003). Application of modified alternating least squares regression to spectroscopic image analysis. *Analytica Chimica Acta*, *476*(1), 93-109.

Wentzell, P. D., Karakach, T. K., Roy, S., Martinez, M. J., Allen, C. P., & Werner-Washburne, M. (2006). Multivariate curve resolution of time course microarray data. *BMC bioinformatics*, *7*(1), 343.

Williams, R. J., Berry, H. K., CAIN, L., & Rogers, L. L. (1951). Individual metabolic patterns and human disease: an exploratory study utilizing predominantly paper chromatographic methods. *Biochemical Institute Studies (Univ. of Texas Publ.)*, *4*, 7-20.

Windig, W., & Guilment, J. (1991). Interactive self-modeling mixture analysis. *Analytical chemistry, 63*(14), 1425-1432.

Windig, W., & Stephenson, D. A. (1992). Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach.

Analytical Chemistry, 64(22), 2735-2742.

Windig, W., Antalek, B., Lippert, J. L., Batonneau, Y., & Brémard, C. (2002). Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach. *Analytical chemistry, 74*(6), 1371-1379.

Wise, B. M., & Kowalski, B. R. (1995). Process chemometrics. In McLennan, F., & Kowalski, B. R. (Eds.), *Process analytical chemistry* (pp. 259-312). London, UK: Blackie Academic & Professional.

Zhu, Z. L., Cheng, W. Z., & Zhao, Y. (2002). Iterative target transformation factor analysis for the resolution of kinetic–spectral data with an unknown kinetic model. *Chemometrics and intelligent laboratory systems, 64*(2), 157-167.