THE DEVELOPMENT OF MULTIVARIATE ANALYSIS METHODOLOGIES FOR COMPLEX TOF-SIMS DATASETS: APPLICATIONS TO MATERIALS SCIENCE

Gustavo Ferraz Trindade BSc, MSc

A thesis submitted for the degree of Doctor of Philosophy

Department of Mechanical Engineering Sciences
Faculty of Engineering and Physical Sciences
University of Surrey

April 2018

© Gustavo Ferraz Trindade
ABSTRACT

Secondary ion mass spectrometry (SIMS) is a technique that has evolved to be one of the most powerful techniques for the analysis of organic samples. Modern instruments are capable of obtaining three-dimensional information with high spatial resolution of a material with information as rich as a full mass spectrum at every voxel of the 3D structure, thus generating very large and complex datasets. Multivariate analysis (MVA) methods are used within the SIMS community, however, the absence of MVA in the software packages of instrument manufacturers together with constant increase in data and data analysis complexity demands practical data analysis solutions that are accessible to scientists of diverse backgrounds. This thesis aims to expand the applicability of three major MVA methods to complex SIMS datasets: Principal component analysis (PCA), non-negative matrix factorisation (NMF) and $k$-means clustering. This is achieved by establishing and validating existing and novel methodologies for the processing of large and complex datasets. Furthermore, it presents the development of a software that encompasses these methodologies and provide accessible and flexible analysis and data visualisation tools. Finally, it presents the application of the software to a series of experiments carried out at The Surface Analysis Laboratory of the University of Surrey in which data processing enabled deeper interpretation of the results and helped to achieve insights towards scientific and industrial problem solving.
I dedicate this thesis to my loving wife and my caring parents.

Helen, obrigado por ter embarcado nesta jornada e cruzado o atlântico comigo.

Adriana e Ailton, obrigado por serem a base de tudo o que eu construí.

Amos vocês.
In loving memory of Marry

(*1996 - †2015)
DECLARATION

This dissertation is the result of my own work and it has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signed: Gustavo Ferraz Trindade

Date: Sunday, 29 April 2018

Gustavo Ferraz Trindade BSc, MSc

Guildford, UK.
ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude and appreciation to my supervisor, Professor John Watts, for his vital guidance and constant input on this project and all his teachings on topics ranging from surface analysis and materials science to life in Britain and real ales. John, you are an inspiration to me as an academic and as a leader. Thank you!

I would also like to extend my gratitude to my co-supervisors Dr. Marie-Laure Abel; for her appreciated intellectual input to this thesis, her numerous pieces of advice and for looking after me inside and outside the academic sphere; and to Dr. Mark Baker for all his advice.

My sincere thankfulness also goes to Professor James Castle for all his wisdom kindly shared and his guidance on the production and submission of SSS papers; to Dr. Steve Hinder for all his valuable teachings and assistance with the analytical instruments in the laboratory; and to Mrs. Tracy Johanson for all her essential assistance with travel arrangements and bureaucratic matters.

I also kindly acknowledge all post-graduate colleagues of the MES department for insightful discussions spanning the most varied topics, which made the PhD experience even more enjoyable. A special vote of thanks goes to my colleagues on SIRG research group, especially Dr. Taraneh Bozorgzad-Moghim, Dr. Jorge Banuls-Ciscar, Dr. Rene Tshulu, Mr. Kristof Marcoen, Mr. Simon Bacon, Dr. Chris Mallinson,
ACKNOWLEDGEMENTS

Mr. Mark Minta, Dr. Sabrina Tardio and Mr. Min Jang, without whose support, encouragement and feedback, simsMVA would never have existed. Thank you!

My appreciation also goes to Dr. Jose Ferreira, with whom I had an extremely fruitful two years-long collaboration at Surrey and for teaching me how to be more audacious.

I have had many productive collaborations throughout these years, and for that I wish to extend my gratitude to Dr. Lucio Rosa, Dr. Elis Rosa and Dr. Carla dos Santos for their efforts in the Brazilian coffee experiments, which joyfully connected my research to my most loved delicacy from my home-country; to Dr. David Williams, for plasma treating the polypropylene samples; to Dr. Tiago Silva, for his advice, constant encouragement and the collaboration on MVA of PIXE imaging data; and to Dr. Sebastiaan Van-Nuffel, for the fruitful discussions and collaboration on the PCA-subsampling project.

Finally, I wish to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Science without Borders programme for funding this project; and the UK Surface Analysis Forum, the Worshipful Company of Armourers and Brasiers and the Institute of Physics for enabling parts of this work to be presented at international conferences.

Gustavo Ferraz Trindade, February 2018
# CONTENTS

## PART I  CONCEPTS

1  **INTRODUCTION**  3

1.1  **CONTEXT OF THIS THESIS**  3

1.2  **HISTORICAL DEVELOPMENT OF MVA IN THE SIMS COMMUNITY**  4

1.3  **THE NECESSITY OF MVA SOFTWARE**  7

1.4  **DEVELOPMENT TIMELINE OF SIMS MVA**  8

1.5  **AIMS AND OBJECTIVES**  11

1.6  **THESIS STRUCTURE**  11

2  **SECONDARY ION MASS SPECTROMETRY**  13

2.1  **INTRODUCTION**  13

2.2  **FUNDAMENTALS OF SECONDARY ION FORMATION**  13

2.3  **THE STATIC LIMIT OF SIMS**  17

2.4  **SECONDARY ION ANALYSIS AND DETECTION**  18

2.4.1  **Time-of-flight analysis**  20

2.5  **THE TOF.SIMS 5 INSTRUMENT**  22

2.5.1  **Primary ion beam column**  24

2.5.2  **Dual-source column**  27

2.5.3  **Analyser**  28

2.5.4  **Electron flood gun for the analysis of insulating materials**  30

2.5.5  **Imaging modes of analysis**  32

2.5.6  **Stage raster for large area imaging**  35

2.5.7  **Depth profiling modes**  36

2.6  **SAMPLE REQUIREMENTS AND PREPARATION**  39

3  **MULTIVARIATE DATA ANALYSIS METHODS**  43
THE DEVELOPMENT OF MULTIVARIATE ANALYSIS METHODOLOGIES FOR COMPLEX TOF-SIMS DATASETS:
APPLICATIONS TO MATERIALS SCIENCE

3.1 INTRODUCTION 43

3.2 PRINCIPAL COMPONENT ANALYSIS 44
  3.2.1 Algebraic solution of PCA 44
  3.2.2 Algorithms for PCA 52

3.3 NON-NEGATIVE MATRIX FACTORISATION 53
  3.3.1 Alternating least squares NMF 55
  3.3.2 Multiplicative update NMF 57
  3.3.3 Initialisation of matrices and choice of number of components 59

3.4 K-MEANS CLUSTERING 59

3.5 FINAL REMARKS 62

4 MVA METHODOLOGIES FOR TOF-SIMS DATA 63

4.1 INTRODUCTION 63

4.2 DATA PRE-PROCESSING 64
  4.2.1 Selection of variables 64
  4.2.2 Data unfolding 66
  4.2.3 Mean-centring 69
  4.2.4 Normalisation 70
  4.2.5 Scaling 71
  4.2.6 Matrix augmentation 73
  4.2.7 Pixel/voxel subsampling 74

4.3 VISUALISATION OF MVA RESULTS OF TOF-SIMS DATA 75
  4.3.1 Use of a priori information and minimisation of the number of plots 77
  4.3.2 Show scores and loadings together: the biplot 80
  4.3.3 Clustering algorithm to separate regions in a scores plot: Voronoi cells. 81
  4.3.4 Web plots 82
  4.3.5 Visualisation of intensity maps: right choice of colour maps 84
  4.3.6 Normalisation for contrast enhancement on intensity maps 88
  4.3.7 Connecting loadings and scores with matching colour maps 91
4.3.8 Multi-colour overlays for NMF results 91
4.3.9 Visualisation of 3D MVA results 93
4.3.10 3D overlays of NMF results 98

4.4 TIME-RESOLVED DATASETS 101
4.4.1 Damage in a standard polystyrene sample 101
4.4.2 Diffusion in stainless steel 102

4.5 FINAL REMARKS 105

5 THE SIMSMVA SOFTWARE 107
5.1 INTRODUCTION 107
5.2 MAIN FEATURES OF SIMSMVA 107
5.2.1 Spectra mode 110
5.2.2 Depth Profile mode 113
5.2.3 Images mode 116
5.2.4 3D mode 120
5.2.5 Exporting data from the SurfaceLab software 123
5.2.6 Algorithms and typical running times 126
5.2.7 System requirements 129
5.2.8 External media 129

PART II VALIDATION 131

6 PCA & NMF OF DATA OF STANDARD DICARBOXYLIC ACIDS 133
6.1 INTRODUCTION 133
6.2 LINEAR SATURATED DICARBOXYLIC ACIDS 134
6.3 EXPERIMENTAL CONDITIONS 135
6.4 TOF-SIMS RESULTS 137
6.5 MVA 141
6.5.1 PCA results 143
CONTENTS

9.1 INTRODUCTION 195

9.2 ToF-SIMS OF POLYPROPYLENE 196
  9.2.1 Energetic surface treatments 197

9.3 EXPERIMENTAL 199
  9.3.1 Polypropylene samples 199
  9.3.2 Plasma composition and treatment 199
  9.3.3 ToF-SIMS analyses 199

9.4 DATA HANDLING AND PROCESSING 200
  9.4.1 Spectral alignment 201

9.5 NMF RESULTS 204
  9.5.1 Effect of treatment distance 208
  9.5.2 Effects of plasma mixing with atmosphere 209
  9.5.3 Behaviour of additive 210

9.6 CONCLUSION & FUTURE WORK 214

10 ANALYSIS OF WOOD GROWTH REGIONS 215

10.1 INTRODUCTION 215
  10.1.1 Structure of wood 216
  10.1.2 ToF-SIMS of wood 219

10.2 METHODOLOGY 220
  10.2.1 Sample preparation 220
  10.2.2 ToF-SIMS analyses 221
  10.2.3 Data processing 222

10.3 RESULTS 224

10.4 DISCUSSION 228

10.5 CONCLUSIONS AND FUTURE WORK 231

11 ANALYSIS OF BREWED COFFEE 233

11.1 INTRODUCTION 233
11.2 Mass spectrometry of coffee

11.3 Methodology
   11.3.1 Samples provenance and preparation
   11.3.2 ToF-SIMS analysis
   11.3.3 Spectra calibration and peak picking
   11.3.4 Multivariate data analysis

11.4 Results & Discussion
   11.4.1 Influence of sample preparation methods on data quality
   11.4.2 Differentiation of preparation methods
   11.4.3 Classification of coffee samples by their origin

11.5 Conclusion & Future Work

12 Analysis of Coatings Formed by Polyester Resin Blends

12.1 Introduction

12.2 Characterisation of Polymer Blends

12.3 Experimental
   12.3.1 Sample preparation
   12.3.2 ToF-SIMS analyses

12.4 Data Analysis
   12.4.1 Peak picking and exporting
   12.4.2 Loading data into analysis software
   12.4.3 Multivariate analysis

12.5 Results

12.6 Discussion
   12.6.1 Effect of curing
   12.6.2 Immiscibility of polyesters
   12.6.3 A step further: k-means clustering

12.7 Conclusion & Future Work

x
CONTENTS

PART IV DISCUSSION

13 THE IMPACT OF THE METHODOLOGIES

13.1 PRACTICAL IMPACT ON MATERIALS SCIENCE RESEARCH

13.2 FUNDAMENTAL IMPACT ON METHOD DEVELOPMENT

13.3 RECOMMENDATIONS FOR ACHIEVING MEANINGFUL RESULTS
    13.3.1 Initial steps in simsMVA

14 CONCLUSIONS & PERSPECTIVES

14.1 CONCLUSIONS

14.2 THE SUCCESS OF SIMS MVA

14.3 FUTURE PERSPECTIVES

REFERENCES

APPENDIX 1: PUBLISHED WORK AND CONFERENCE PRESENTATIONS

APPENDIX 2: FURTHER APPLICATIONS TO TOF-SIMS DATA

A2.1 LITHIUM-BASED CORROSION PROTECTIVE LAYER

A2.2 ANALYSIS OF POLYURETHANE COATINGS

A2.3 ANALYSIS OF PARCHMENT

A2.4 DRUG DETECTION ON FINGERPRINTS

A2.5 PLASMA TREATED STAINLESS STEEL

A2.6 TOF-SIMS/MVA DATABASE OF COMMERCIAL RESINS

A2.7 STUDY OF DEGRADATION OF AIRCRAFT TOPCOATS

A2.8 EFFECTS ON ION BOMBDAMENT OF POLYMERS

APPENDIX 3: FURTHER APPLICATIONS TO OTHER ANALYTICAL TECHNIQUES

A3.1 FTIR OF POLYESTERS

A3.2 XPS DEPTH PROFILES
### A3.3  PIXE OF COFFEE

Page 323

### A3.4  LARGE AREA PIXE MAPPING

Page 324
PART I

CONCEPTS

The following Chapters present the conceptual development necessary for the work presented in this thesis. These include an introduction to the thesis, the principles and experimental aspects of ToF-SIMS, a mathematical description of the multivariate methods employed in this thesis and the methodologies for the specific application of multivariate analysis to ToF-SIMS datasets.
1 INTRODUCTION

1.1 Context of this thesis

The Surface and Interface Reactions Group (SIRG), led by Professor John F. Watts, has most of its research carried out at The Surface Analysis Laboratory of the University of Surrey. Throughout the years, SIRG has been established as a global reference on the surface characterisation of industrial samples mainly by means of time-of-flight secondary ion mass spectrometry (ToF-SIMS) and X-ray photoelectron spectroscopy (XPS), techniques in which members of the group have developed elevated levels of expertise using state-of-art instruments. Such environment makes the ideal set-up for the development of a doctorate project based on practical data analysis methodologies, since the candidate is guaranteed to receive training that will enable the production of excellent quality data and will be exposed to challenging samples from various sectors of industry and academy.

This thesis deals with the development of methodologies for the application of three major multivariate data analysis (MVA) methods to ToF-SIMS datasets: principal component analysis (PCA), non-negative matrix factorisation (NMF) and $k$-means
clustering. This introductory Chapter intends to contextualise the reader with regards to MVA in the SIMS field up to the current day and also to present the structure of this thesis. The fundamentals of the technique and the MVA methods will be thoroughly discussed throughout the remaining Chapters.

1.2 Historical development of MVA in the SIMS community

The first publication that results from a search for the keywords “ToF-SIMS” and “PCA” in the web of science database dates back from 1996 and is a piece of work that was in fact presented at the 9th International Conference on Quantitative Surface Analysis held at the University of Surrey and chaired by Castle, Seah and Watts. In their paper, Vanden Eynde and Bertrand [1] proposed a method for quantification of SIMS spectra of polystyrene that explored the use of PCA to take into account all information available in each spectrum. In the following years, a series of works from the University of Washington had major impact in the community and showed how powerful MVA can be [2–4]. These did not deal with spectra of standard polymers (that had their fragmentation patterns and SIMS spectra very well established at the time) but with rather complex biological samples such as cells and proteins. At the same time, Bertrand’s group followed up on the PCA of polymers and used MVA for the identification of small molecules, such as additives, on the surface of model polymer samples [5, 6].

Seven years after the first application of PCA to SIMS, another important breakthrough was the first application of MVA to ToF-SIMS imaging datasets of
polymers reported by Tyler [7]. Although polymer samples are simpler than proteins, the analysis of images represented an increase of dataset complexity, which led to more mathematically rigorous studies of the specific application of MVA to ToF-SIMS [8] and widened the possibilities with applications of methods such as multivariate curve resolution (MCR), imported from the very well established chemometrics community [9, 10].

By 2008, a significant number of researchers were already using MVA with review papers describing its use by the group at University of Washington [11, 12]. However, researchers around the globe were using various terminology and often MVA methods were being applied as black boxes without concern for the peculiarities of ToF-SIMS data. This demanded standardisation of ToF-SIMS data pre-processing steps and appropriate terminology for MVA, which were both achieved by a series of works by Lee and co-workers at the National Physical Laboratory (NPL) [13–18]. In the following year, the journal “Surface and Interface Analysis” had two special issues edited by Gilmore and Wagner dedicated to MVA with 12 out of the 15 papers dealing with ToF-SIMS data [19–21]. Such an initiative of creating a special issue seems to have given confidence to researchers using ToF-SIMS to apply MVA to their datasets, as the following years saw a significant increase in number of publications with applications in many diverse fields and publications featuring in high impact factor journals [22–24]. However, in terms of method development, the next breakthrough happened very recently with novel approaches for the analysis of extremely large datasets. Van Nuffel et. al used voxel subsampling for the first time to perform PCA of large 3D datasets [25, 26] and Cumpson et al. expanded the approach by combining subsampling with a
random vectors-based algorithm which enabled the calculation of principal components of very large datasets using graphic processing units (GPUs) [27, 28]. These two contributions led to a work on the use of non-negative matrix factorisation (NMF) of large ToF-SIMS datasets described in this thesis [29]. Another important recent contribution was done by Tuccito et al. with the application of wavelet compression-based PCA to large datasets [30]. These recent contributions are setting the scene for future developments in a scenario where every SIMS dataset will be a large dataset: the recently launched orbiSIMS spectrometer [31] generates datasets thousands of times larger than any available conventional ToF-SIMS instrument.

Apart from the popular MVA methods within the SIMS community, the rapid growth of the machine learning research field has been inspiring researchers to apply novel and less memory/computationally demanding methods such as the application of Fourier transform analysis [32] and deep learning neural networks [33] to ToF-SIMS spectra. The potential success of these alternative approaches means that a new cycle may begin and events similar to those described here may occur until the new methods are accepted and established in the ToF-SIMS community. Figure 1.1 shows the number of publications per year in the web of science database containing the keywords ToF-SIMS and MVA, PCA, MCR, NMF or multivariate analysis.
Figure 1.1: Publications per year in the web of science database containing the keywords: ToF-SIMS and MVA, PCA, MCR, NMF or multivariate analysis.

1.3 The necessity of MVA software

The absence or limitation of MVA in the software packages of instrument manufacturers has always been a problem and the increase in complexity of both data and data analysis methods described in the previous Section created a niche inside the community, where data analysis solutions that would not be achievable with the available software being extremely dependant on scientists highly skilled in mathematics or computer programming. ToF-SIMS is a field that is, every year, more interdisciplinary with researchers with a variety of backgrounds, so this problem should be addressed in order
to enable MVA of data of very complex systems in, for example, biological or industrial contexts. In terms of software, the most popular spectrometer manufacturers (IONTOF, Ionoptika, Physical Electronics) do not provide a complete set of MVA tools in their analysis software, which make researchers go for independently developed alternatives. The three most widely employed software for MVA within the SIMS community are the PLS.MIA toolbox by Eigenvector Research [34], the NBtoolbox, developed by Graham [35] and the MCR-ALS toolbox developed by Jaumot, Gargallo, de Juan and Tauler [36]. Graham made an important contribution with a MATLAB-based graphical user interface (GUI) which was the first one dedicated exclusively for the analysis of SIMS data [35] but its development seems to be focused on the analysis of biological samples and some of the factorisation routines present in the GUI are dependant on another independently developed toolbox from the chemometrics community [36]. Moreover, many of the processing methods and visualisation tools used and described throughout this thesis were not available in any software and this was the initial motivation for the author to create his own set of routines in MATLAB. Over the course of the work, these routines evolved into a software package called simsMVA and many of its features were developed based on analysis challenges faced by the author and the research group at the University of Surrey.

1.4 Development timeline of simsMVA

Chapter 5 of this thesis contains a description of simsMVA as it currently stands. However, it is important to appreciate that the current version is a product of constant
development throughout the years and several of the features were added in order to solve analytical challenges mostly of industrial samples typically analysed in The Surface Analysis Laboratory of the University of Surrey. Figure 1.2 shows a development timeline of simsMVA with the major milestones indicated. Several of the features were developed to deal with the problems presented in the Chapters of Part III of this thesis, in addition, there were also features that were added as a result of collaborations between the author of this thesis and other researchers within his own group and from different institutions.

Initially, only a few functions were created and used within the MATLAB environment to analyse multiple ToF-SIMS spectra. After one year, functions for the analysis of imaging datasets had been created and those were grouped in a graphical user interface named *pcaGFT* and presented at the PhD confirmation viva of the author. The positive feedback from the examiners brought the confidence for further development and by June 2016 pcaGFT was being used by several members of the research group at the University of Surrey. This was of extreme importance as the different applications demanded the development of specific tools such as a mode for the analysis of 3D datasets or the incorporation of inorganic/organic mass filtering of peak lists (the research related to these developments is described in Appendix 2). In 2017 the GUI was completely redesigned and pcaGFT was renamed to *simsMVA* and, shortly after, it was presented to external audiences at three conferences with a particular focus on the international SIMS conference in Krakow. Since then, many researchers have been requesting copies and giving valuable feedback in return, a fact that keeps improving simsMVA.
The Development of Multivariate Analysis Methodologies for Complex TOFSIMS Datasets: Applications to Materials Science

Figure 1.2: SimsMVA development timeline.
1.5 Aims and objectives

The aim of this thesis is to expand the applicability of MVA to large and complex ToF-SIMS datasets and this was achieved by the creation and validation of novel methodologies that incorporate a set of flexible and versatile MATLAB routines. The additional objectives related to this aim were the development of a software encompassing such methodologies; the application of the software to materials science research problems with experiments carried out by the author and the further application/validation of the software by releasing it to external users. Moreover, in order to appreciate the challenges related to materials science and ToF-SIMS, one must understand the technique at a high level. Therefore, another objective was the achievement of excellence in the experimental aspects and understanding of ToF-SIMS.

1.6 Thesis structure

This thesis is structured in four mains Parts: I - Concepts, II - Validation, III - Applications and IV - Discussion. Part I contains the current Chapter and four other Chapters that present the conceptual development necessary for this thesis. Chapters 2 and 3 are respectively descriptions of the ToF-SIMS technique and of the mathematics and algorithms of the multivariate analysis methods of established. Chapter 4 deals with the application of multivariate analysis specifically to ToF-SIMS data. It contains a mix between recommendations from literature and novel methodologies developed by the author. Chapter 5 describes the main functionalities of simsMVA, which encompasses all methodologies present in Chapter 4. Part II contains three Chapters that all have the
purpose of validating the methodologies that are part of simsMVA. This is achieved via
the description of acquisition and processing of SIMS data of standard or model samples
(Chapters 6 and 7) or the creation and systematic tests of synthetic datasets (Chapter 8).
Part II is therefore both conceptual and experimental. Part III encompasses 4
experimental Chapters containing applications of all methods described in Parts I and II
to experiments carried out by the author at the University of Surrey (Chapters 9 to 12).
Part IV contains a general discussion with regards to the impact of the methodologies
presented in this thesis (Chapter 13) followed by conclusions (Chapter 14). This thesis
also contains three Appendices. Appendix 1 presents a list of all publications that are
related to the research of the thesis, Appendices 2 and 3 contain a series of further
applications of the simsMVA software to ToF-SIMS and other analytical techniques
datasets in research carried out by collaborators and independent users.
2 Secondary Ion Mass Spectrometry

2.1 Introduction
Secondary ion mass spectrometry (SIMS) is based on the detection of ionised molecules, molecular fragments or atoms generated as a consequence of the bombardment of a primary ion beam on the surface of the sample under analysis. The technique has its roots on the characterisation of materials in the semiconductors industry and evolved to be one of the most powerful techniques for the analysis of organic and inorganic samples [37, 38].

2.2 Fundamentals of secondary ion formation
The principal phenomenon related to the formation of secondary ions is the sputtering process which can be described as a cascade of collisions of the incident particles in the
material being analysed (a general view of the process is represented in Figure 2.1). Ions from the source collide with the surface of the sample generating atomic motion by direct collisions between the primary ions and the atoms of the material or non-direct collisions between atoms in motion (which leads to a collective molecular motion regime for polymers [39]). As typical primary ion beam energies are of the order of keV, the collisions energies are much higher than the binding energies of the atoms in the material, resulting in an extensive fragmentation around the collision region and producing essentially only the emission of atomic particles. As the collision cascades spread they become less energetic, resulting in less fragmentation and generating the emission of molecular fragments (see schematic in Figure 2.1). The fragments from the first three monolayers of the sample will have sufficient energy to overcome the surface binding energy and leave the sample. Such fragments are ejected as atoms and molecules. The description of the sputtering phenomenon and ion formation is still the object of much research, with several models that consider both theoretical aspects and molecular dynamics simulation results [39–45].
The sputtering yield ($Y$) determines the rate of material desorption after incidence of primary ions. It is defined as the ratio between the number of atoms removed and the number of incident ions. However, only a small fraction of the atoms and fragments removed are ionised (from 0.0001% to 10%) [46]. The basic SIMS equation relates the number of ionised species of mass $m$ to their sputtering yield $Y_m$:

$$I_m = I_p Y_m \alpha \theta_m \eta$$

where $I_m$ is the current of ions with mass $m$, $I_p$ is the primary ion current, $\alpha$ is the ionisation probability, $\theta_m$ is the concentration of $m$ in the surface and $\eta$ is the transmission of the analyser employed.

**Figure 2.1: Schematic of the sputtering process upon primary ion beam collision with a surface.**
Once an atom or molecule is sputtered, there are several ways that they can be ionised [47]. Among these are i) direct ionisation via collision with material, resulting in positively charged odd electron ions; ii) heterogeneous bond scission that separates charge and forms ions of both negative and positive polarities; iii) cationisation of neutral species by incorporation of protons or association to metals, resulting in positively charged even electron ions; iv) neutral species that become electronically excited and lose an electron later and v) neutral species that capture or collide with secondary electrons released from inelastic collisions. An odd electron ion is an ion containing unpaired electrons in its ground state while an even electron ion contains no unpaired electrons in its ground state. Figure 2.2 shows a positive SIMS spectrum of a very simple molecule (malonic acid). This spectrum is part of a series of measurements carried out by the author of this thesis and are presented in more detail in Chapter 6. Both the sputtering yield and the ionisation probability will depend on parameters such as the primary ion beam energy, incidence angle and the type of molecule. Furthermore, the electronegativity of species present in the surface of the material under analysis will influence the ionisation probability of different fragments of a same molecule. This is known as matrix effect and is the main problem when it comes to quantification from SIMS data.
Figure 2.2: Positive SIMS spectrum of malonic acid.

2.3 The static limit of SIMS

Every time an energetic (keV) primary ion strikes the surface of a sample, it causes a damage region of around 10 nm² around the impact point [48]. This means that $10^{13}$ strikes/cm² will have affected all atoms in the surface. This number is the so called static limit and it is the value of which the primary ion beam dose must stay below in order to achieve SIMS characterisation of a true, undamaged surface. The observation that low-dose SIMS could be used as a surface-sensitive technique was first made by Benninghoven [49] and the static limit was first systematically tested and determined for organic materials by Briggs using a primary ion beam of monoatomic argon atoms.
and polystyrene samples [50]. In the recent years, the adoption of cluster primary ion beams (such as Bi\(^+_n\), Au\(^+_n\), C\(^+_n\) or Ar\(^+_n\)) has become widespread and it is expected that the impact of cluster ions causes much less damage. When a cluster impacts a surface, its atoms are separated and each one only retains a fraction of the initial ion energy, resulting in much lower depth of penetration and much higher sputtering yields [42, 51].

2.4 Secondary ion analysis and detection

There are a number of different methods for the analysis and detection of ions generated in the SIMS technique. All kinds of analysers are based on the detection of the mass/charge ratio of the ions and spectrometers based on it include magnetic sector field, quadrupole, ion-trap and time-of-flight analysers, the latter being the most popular for the characterisation of organic materials and the type of detection used for the results presented in this thesis. The high-resolution capabilities of modern instruments allow the detection of multiple ions with the same “nominal mass”. Such differentiation is physically possible due to the difference in mass between a proton and a neutron and the imbalance of such particles in the nuclei of different atoms. Figure 2.3 shows a zoomed in view in the region of nominal mass 57 u of a SIMS spectrum of a pre-treated metallic substrate measured by the author of this thesis. Figure 2.4 shows the difference between atomic and nominal masses for a range of nominal masses in the periodic table. The main constituents of typical organic fragments (C and H) have a positive difference while the most common metals (with masses between 20 and 200) have a negative difference. This enables, in most cases, prompt identification of peaks for inorganic and
organic fragments by just observing their relative position to the molecule or atom nominal mass in a spectrum.

Figure 2.3: Seven different ion species detected at the same nominal mass.

Figure 2.4: Difference between atomic and nominal masses for a range of masses in the periodic table
2.4.1 Time-of-flight analysis

A time-of-flight (ToF) analyser has unique characteristics that make it the ideal detection system for SIMS. Secondary ions are separated by their mass/charge ratio \((m/z)\) and their masses are determined by the time they take to travel a distance \(L\) in a flight tube free of electromagnetic fields after being accelerated by an extractor to a common energy \(E\) (Figure 2.5 (A)). The relation between \(E\) and the flight time, \(t\), is straightforward:

\[
E = zU = \frac{mv^2}{2} = \frac{mL^2}{2t^2}
\]

where \(z\) is the ion charge state, \(U\) is the extractor potential and \(v\) is the velocity of the secondary ion. Therefore, lighter ions travel at higher speeds and reach the detector before heavier ions in a time \(t\):

\[
t = L\left[\frac{m}{2E}\right]^{\frac{1}{2}} = L\left[\left(\frac{m}{z}\right)\frac{1}{2U}\right]^{\frac{1}{2}}
\]

In ToF-SIMS, the primary ions are generated in very short pulses and, for each pulse, all the secondary ions that can reach the detector within the pulse length (typically 100 ns for masses up to 600 u) are detected, meaning that a secondary ions spectrum is produced with the minimum amount of primary ions.

There are several factors that influence on the ToF resolution (or width of a secondary ion packet) [52] such as angular spread of ion trajectories or effective depth of the detector but the major source of resolution degradation is the energy spread \((\Delta E)\) of ions of the same mass (initial relative positions of ions also degrades resolution but this is regarded as a sample effect rather than instrumental and will be discussed later).
From the previous equation, it can be seen that resolution $R$ is, in first order, determined only by the energy spread:

$$ R \equiv \frac{t}{2\Delta t} = \frac{m}{\Delta m} = \frac{E}{\Delta E} $$

In order to overcome such problem, scientists came up with the idea of an ion mirror. Considering the scheme in Figure 2.5 (B) if immediately after being generated, ions of mass $m$ are influenced by a potential that varies quadratically with the distance travelled, the time it will take to stop and reaccelerate an ion until it leaves the field is [53]:

$$ t = 2\sqrt{\frac{m}{2}} \int_{0}^{x_{\text{max}}} \frac{dx}{\sqrt{E_0 - zU(x)}} $$

Where $E_0$ is the initial kinetic energy, $U = ax^2$ and $x_{\text{max}}$ is the distance required to stop the ion ($\sqrt{E_0/a}$). Substituting the values and solving the integral:

$$ t = \sqrt{\frac{E_0}{a}} \int_{0}^{\sqrt{\frac{E_0}{a}}} \frac{dx}{\sqrt{E_0 - zax^2}} = \frac{\pi}{\sqrt{2za^2}} = \left[ \frac{m}{z} \right] \left[ \frac{2a}{2a} \right]^{1/2} $$

Which shows that, under a quadratic field, the detection time does not depend on the initial kinetic energy and thus any mass resolution degradation will come from minor sources. However, experimental designs of such kind of ion mirror have the disadvantage of reducing the transmission of ions [54] and manufacturers adopted designs known as single stage and double stage reflectrons with the ion path consisting of two field-free regions and one region (reflectron) where ions are subject to at least one linear increasing potential $U = ax$ step (Figures 2.5 (C) and 2.5 (D)). Such designs
guarantee at least first order focusing of the initial energy spread of secondary ions of the same mass [54, 55]. Among other design strategies for the improvement of resolution are delayed extraction and post-acceleration voltages.

![Diagram of ToF analyser designs](image)

**Figure 2.5**: Different designs for a ToF analyser: linear (A), quadratic field ion mirror (B), single stage reflectron (C), double stage reflectron (D). For each design diagram: top: ions path and grid positioning; bottom: potential curves.

Figures reproduced from [54] with permission.

### 2.5 The TOF.SIMS 5 instrument

The TOF.SIMS 5 instrument installed at the University of Surrey was commercialised by the company IONTOF GmbH (Muenster, Germany) in 2005 and is still their current
SIMS product line. The system supports three different primary ion sources the main one being a bismuth liquid metal ion gun (LMIG), that can produce monoatomic or cluster Bi\textsuperscript{+\textsubscript{n}} beams, and a dual-source column with sources capable of producing carbon cluster ions of C\textsubscript{60}\textsuperscript{+} atoms and Cs\textsuperscript{+} ions. The C\textsubscript{60}\textsuperscript{+} and Cs\textsuperscript{+} beams are typically used for etching in depth profiling modes. The instrument also has an electron gun (flood gun) for charge compensation. The two sources are positioned at 45° to the surface of the sample and in between there is a ToF analyser with a single stage reflectron and a multi-channel plate detector capable of up to 20 kV post acceleration voltage, resulting in a time resolution of nanoseconds for typical pulsed acquisition. The analysis chamber is maintained at ultra-high vacuum at pressure of around 10\textsuperscript{−9} mbar with a turbo-molecular pump and a titanium sublimation cryogenic pump. Figure 2.6 shows a photograph of the equipment installed at University of Surrey with its main components indicated. The following sections will detail the main components of the equipment.
Figure 2.6: Left: Photography of the TOF.SIMS 5 equipment installed at University of Surrey with main components indicated. Right: Schematic of the analysis chamber with ion beam columns and mass analyser (taken from Iontof’s training material).

2.5.1 Primary ion beam column

Figure 2.8 shows all the components present in the primary ion beam column. The column contains a liquid metal ion gun (LMIG) with a bismuth tip with design schematic shown in Figure 2.7 (A)). The potential of the emitter defines the ion beam energy and the emission of ions starts after the field strength at the needle tip exceeds a threshold value. The field strength is set by the extractor voltage and can be fine adjusted by the suppressor voltage. Both voltages can be changed by the user or in an automated manner in order to maintain a constant emission current. The alignment of the beam is done with three deflection systems (X/Y) to pass the ions through three apertures. It is possible to
measure the current through apertures 1 and 2 to find the optimal focusing parameters and the focusing through the blanking aperture can be optimised by measuring the current hitting the target (with a Faraday cup or even a conducting sample). For all analysis modes, the beam must be pulsed and this is done at the pre-chopper and chopper which is also where the beam mass filtering is done by blocking ions of different masses that will have different velocities for the same beam energy (the highest current beam species are Bi$^+$ and Bi$_3^+$). The maximum energy of the beam is 30 keV for single charged species.

Figure 2.7: Principles of the three types of ion source present in the TOF.SIMS 5 equipment: liquid metal ion gun (A); electron impact (B); surface ionisation (thermal sputtering) (C). Adapted from the instrument manual.
Figure 2.8: Schematic with all components present in the primary ion beam column of the TOF.SIMS 5 equipment. Adapted from the instrument manual.
2.5.2 Dual-source column

The dual-source ion column contains an electron impact source capable of producing C_{60}^{+} and a thermal sputtering ion source capable of producing Cs^{+}. A 90 degrees deflection magnet is used for beam selection and the focusing, bunching and pulsing work in a similar way to the primary ion beam column. These sources are used more commonly as etching guns in depth profiling modes (see next sections) although they can also be used as primary ion guns for specific applications. Figure 2.9 shows a schematic with the main components present in the dual-source column. Typical energies for the C_{60}^{+} and Cs^{+} beams are respectively 1 keV and 3 keV.

Figure 2.9: Schematic with main components present in the dual-source column of the TOF.SIMS 5 equipment.
2.5.3 Analyser

The TOF.SIMS 5 instrument is equipped with a mass analyser that contains a single stage reflectron. Figure 2.10 shows a schematic with its main components. Positively or negatively charged secondary ions are extracted depending on the extraction bias and are subsequently focused in the flight tube. The reflectron is gridless and is set in a way that the optimal energy focusing occurs at the microchannel plate detector. Before reaching the detector, secondary ions experience a post-acceleration potential in order to increase the signal conversion rate in the scintillator.

![Schematic of TOF.SIMS 5 mass analyser](image)

Figure 2.10: Schematic with main components of the mass analyser of the TOF.SIMS 5 instrument. Taken from the instrument manual.

There are several parameters that can be changed in order to optimise the mass resolution of the analyser but the most commonly adjusted by a user are the distance of the extractor
tip to the surface of the sample and the reflectron voltage. The optimal distance of the tip is 1.5 mm and this has to be adjusted depending on features of the sample under analysis and the optimal reflectron voltage is 20 V for positive ions and –20 V for negative ions. Figure 2.11 (A) shows a series of $^{28}\text{Si}^+$ peaks measured from a silicon wafer sample to illustrate the mass resolution degradation for reflectron voltages deviating from the optimal. These are part of a series of measurements carried out by the author of this thesis for the optimisation of acquisition parameters. Another parameter that is changed depending on the material analysed is the cycle time, which is inversely proportional to the repetition rate of the primary ion gun. Lower repetition rates will increase the cycle time and enable the detection of longer time-of-flights, effectively increasing the range of masses analysed. The only compromise is the decrease in count rate. Figure 2.11 (B) shows $^{28}\text{Si}^+$ peaks acquired with different cycle times and Figure 2.11 (C) shows the same peaks normalised by their respective spectra total ion counts to demonstrate that there is no degradation on mass resolution.
2.5.4 Electron flood gun for the analysis of insulating materials

The majority of organic materials are electrically insulating and therefore they will accumulate charge on their surface during a SIMS analysis. Such an effect can reduce or completely suppress the secondary ions reaching the analyser. The sample charging occurs during the sputtering process after the bombardment of the surface by positive charged primary ions and the simultaneous loss of secondary electrons. The sputtered region will exhibit positive charging unless the material has enough electrical conductivity to carry the electrons to the region in question. In order to neutralise the charge build up during SIMS analysis the surface of the sample is flooded with electrons with energy sufficiently low (up to 21 eV) to not cause any damage between the pulses of primary ions. The TOF.SIMS 5 equipment is equipped with a flood gun that, in

Figure 2.11: $^{28}$Si$^+$ peak acquired with different conditions of the analyser. (A): varying reflectron voltage. (B): varying cycle time. (C): varying cycle time and normalised by total ion counts.
operation, provides in a pulsing scheme with the primary ion beam and the extraction voltage of the analyser as shown in Figure 2.12.

![Diagram of pulsing scheme](image)

**Figure 2.12: Pulsing scheme when flood gun is turned on in the TOF.SIMS 5.**

Another parameter that must be changed when the flood gun is turned on is the reflectron voltage. As the sample surface potential shifts from ground, this must be corrected by shifting the reflector voltage as shown in Figure 2.13.
2.5.5 Imaging modes of analysis

The TOF.SIMS 5 instrument is capable of raster scanning the primary ion beam with a maximum area of 500 x 500 µm². For all analysis modes, a full mass spectrum is recorded per pixel in the raster pattern, which enables post-reconstruction of datasets entirely or partially. When there is no interest on spatial distribution of secondary ion fragments, a single spectrum is acquired per measurement. However, in order to minimise damage, this is not done in a “point analysis” fashion, instead, the beam is

Figure 2.13: Adjustment of the reflector voltage (VR) for insulating samples.
raster scanned over smaller areas and the spectra of all pixels are collapsed into one final spectrum representative of the surface under analysis. The three most used modes of operation of the primary ion gun are the **high current (bunched) mode**, the **burst alignment mode** and the **burst mode**. The high current mode requires two beam cross-overs in the deflection systems in order to achieve very short pulse lengths. As a consequence, it can provide good mass resolution \( \frac{m}{\Delta m} \approx 9,000 \) at 28 u with high beam current (~30 nA DC on target) but reduced spatial resolution (2 - 10 µm spot size). With only one beam cross-over, the burst alignment mode results in lower current (~400 pA DC on target) but enables much better beam focusing and therefore higher spatial resolution (~300 nm spot size) but with unit mass resolution. Figure 2.14 shows the mass spectra at nominal mass 1 u and total ion counts maps of 500 x 500 µm² areas of an adhesive sample acquired using both high current and burst alignment mode.
Figure 2.14: Mass spectra at nominal mass 1 u (bottom) and total ion counts maps of 500 x 500 µm² areas (top) of an adhesive sample acquired using both high current and burst alignment modes.

In order to achieve both good mass and spatial resolution (at the cost of a lower beam current), the burst mode chops the burst alignment mode beam in very short pulses using an RF source, which results in repeated signals within each unit mass. Figure 2.15 shows a comparison between the high current and burst modes for the same 500 x 500 µm² area of a polypropylene sample contaminated with potassium-rich particles. The image for the burst mode is reconstructed by adding together the areas of all peaks representing the same ion. Such repeated pattern in the spectra of burst mode may be problematic
when several peaks are present in the same nominal mass (as in the example shown in Figure 2.3). In all imaging modes, care must be taken in order to select the number of pixels that will match the beam spot size so that a full monolayer will be hit after a full scan and no overlapping (with possible damage) will occur.

![Figure 2.15: Mass spectra at nominal mass 39 u and integrated counts maps for the K⁺ peak of 500 x 500 µm² areas (top) of a polypropylene surface acquired using both high current and burst modes.](image)

**2.5.6 Stage raster for large area imaging**

In order to overcome the 500 x 500 µm² beam raster limit for ToF-SIMS imaging, the TOF.SIMS 5 equipment has an automated stage raster mode that acquires secondary
ions of a set of regions and stitches the maps into one dataset. Every separate region rastered by the primary ion beam is called a “patch”. The main parameters that can be controlled by the user are the total area of analysis, the pixel density in pixels/mm\(^2\), the patch side length (or maximum beam rastering range of each patch) and the number of scans per patch. Such automation can be problematic when the conditions of the sample change across its surface (topography, conductivity) and the parameters optimisation of the initial patch are not valid for the others. This will be discussed further in the applications chapter of this thesis.

2.5.7 Depth profiling modes
Depth profiles are obtained in the TOF.SIMS 5 in a dual-beam fashion where a primary ion beam (typically Bi\(^+\) or Bi\(_3\)\(^+\)) and an etching (or sputter) beam are optimised independently. The two beams operate intermittently to provide in depth information of secondary ions. Typically, the Cs\(^+\) ion source is used for hard materials such as metals and the C\(_{60}\)\(^+\) is used to profile organic materials. Modern instruments will also have argon cluster sources that reportedly give less damage to organic materials and result in profiles with better depth resolution [56]. There are two possible modes of operation for depth profiles: **interlaced mode** and **non-interlaced mode**. In the interlaced mode, mostly used for conductive samples, the sputter beam operates between extraction pulses in a quasi-simultaneous fashion with the primary ion beam. In the non-interlaced mode, ideal for insulating samples, the primary ion beam and analyser operate separately from the sputter gun, with intervals in the order of seconds that give time for surface relaxation and electron flooding. Figure 2.16 shows the timing sequences for both interlaced and
non-interlaced modes. For both modes, the primary ion gun raster scans a much smaller region than the sputter gun in order to avoid crater wall effects, however, if one wants to acquire the raw data containing both lateral and in-depth information (3D analysis), appropriate raster size and parameters of the primary ion beam must be set up accordingly. Furthermore, 3D analysis will only be valid for the non-interlaced mode with sequential acquisition of images of the bottom of the crater formed after the sputtering cycle.
Figure 2.16: Comparison between interlaced and non-interlaced modes of depth profiling.
2.6 Sample requirements and preparation

ToF-SIMS is an ultra-high vacuum technique, therefore samples must in principle be in solid and low-vapour pressure states. The TOF.SIMS 5 equipment contains three kinds of sample holder. The **back mounted** sample holder (Figure 2.17 (A) and (B)) contains a series of holes with maximum size of 1 cm x 1 cm and also contains a Faraday cup for current measurement. The samples are inserted from its back side and held with clipped pieces of metal. In contrast to the limiting size and shape of samples, the biggest advantage of the back mounted sample holder is that the optimal height and the X-Y coordinates of the stage are automatically set up. The **top mounted** sample holder (Figure 2.17 (C)) allows any solid sample to be mounted on it (with maximum size of 10 cm x 8 cm x 2 cm). The samples can be held with metal clips or vacuum-compatible double-sided tape. The **Heating/Cooling** sample holder (Figure 2.17 (D)) allows the temperature on its surface to be ramped via software and also has a copper finger that can be used to maintain samples at low temperatures inside both the prep chamber and the analysis chamber of the TOF.SIMS 5.

ToF-SIMS is a highly surface sensitive technique with very low detection limits. Therefore care must be taken in order to avoid sample contamination. All sample holders must be handled using silicone free gloves and samples must be manipulated using clean tweezers. Storage conditions can vary but, for most samples analysed for this thesis, keeping specimens wrapped in aluminium foil showed to be the best practice to minimise surface contamination.
As it was mentioned in Section 2.3, the initial position of the secondary ions will affect mass resolution. In the TOF.SIMS 5 equipment, the effective primary ions pulse width depends on the impact point within the raster region. As a consequence of this, ions generated at regions further from the centre of the analysed area will have longer detected time-of-flight, resulting in slightly misaligned spectra for different pixels. The software of the equipment contains a built-in function that corrects for this effect. However, secondary ions generated at different heights will also suffer delays in their detected time-of-flight, which means that sample topography will highly influence mass
resolution. As described in Section 2.5.4, charge build-up at the surface of the sample will shift the focal plane of the analyser, so mixed samples that might suffer from differential charging effects will also yield mass spectra with degraded resolution. Moreover, some samples will also present topographic field effects that instead of delaying time-of-flight, will suppress any detection of ions coming from specific regions, as shown in Figure 2.18. Among strategies to minimise topographic field effects are delayed extraction and change of flood gun parameters such as electron current and energy [13, 14].

![Figure 2.18: Topographic field effects during ToF-SIMS analysis of an insulating sample. Image reproduced from the literature with permission [13].](image-url)
3 MULTIVARIATE DATA ANALYSIS METHODS

3.1 Introduction

Multivariate data analysis (MVA) encompasses a set of methods for the statistical analysis of data sets containing more than one variable. These methods are particularly useful when groups of variables are correlated, which is the case of ToF-SIMS data where every spectrum will have thousands of peaks representing secondary ion fragments that are often generated from the same parent molecule, thus their peak intensities are expected to correlate. Such peak intensities are regarded as the variables of a ToF-SIMS dataset and their peculiarities will be discussed in the next Chapter. In order to effectively apply MVA to ToF-SIMS data, the fundamentals of the methods must be understood. This Chapter describes the mathematical basis of the three MVA techniques used throughout this thesis: principal component analysis, non-negative
matrix factorisation and $k$-means clustering. Matrices will be represented in **bold** and a superscript $^T$ means the transpose of a matrix.

### 3.2 Principal component analysis

Principal component analysis (PCA) is the best starting point to explore a multivariate dataset. The method accounts for covariance amongst variables and one of its biggest advantages is that it has an exact solution. In contrast, its main limitations are the constraint of orthogonality between components and the fact that it only captures linear variations within the dataset, which may lead to incorrect interpretations. For ToF-SIMS data, non-linearity comes mainly as consequence of detection dead-time and the *matrix effects* described in Chapter 2 [57]. There are methods that expand PCA to non-orthogonal components, such as independent component analysis [58], and accounts for data non-linearity, such as kernel PCA [59, 60] but these are beyond the scope of this introduction.

The following detailed description includes a ToF-SIMS practical example with data acquired for this thesis and the mathematical steps follow the ones shown in [61].

#### 3.2.1 Algebraic solution of PCA

Consider a mass spectrum with only two peaks. As an example, consider the intensities $I_1$ and $I_2$ of two characteristic peaks from polydimethylsiloxane (PDMS) at masses 73.04 u (SiC$_3$H$_9$) and 147.06 u (Si$_2$C$_5$H$_{15}$O), we can define a 2-dimensional *data vector* $\mathbf{X}$ for the sample:
\[ \mathbf{X} = \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} \]

Expanding \( \mathbf{X} \) to more than one spectrum, each sample will be a 2-dimensional vector:

\[ \mathbf{X}_1 = \begin{pmatrix} I_{11} \\ I_{21} \end{pmatrix}; \quad \mathbf{X}_2 = \begin{pmatrix} I_{12} \\ I_{22} \end{pmatrix}; \quad \mathbf{X}_3 = \begin{pmatrix} I_{13} \\ I_{23} \end{pmatrix}; \ldots \]

Figure 3.1 shows the representation of normalised peak intensities of both peaks for 97 arbitrary spectra plotted against each other. The normalisation was done in order to minimise differences due to roughness and surface charging.

**Figure 3.1:** Representation of normalised peak intensities of peaks at masses 73 and 147 u for 97 arbitrary spectra plotted against each other.

In the 2-dimensional \( I_1 \times I_2 \) space represented in Figure 3.1, every data point (or vector) can be expressed as a linear combination of the orthonormal basis \( \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \). Extending the set to \( m \) peaks and \( n \) spectra, we can arrange the original dataset in an \( m \times n \) matrix \( \mathbf{M} \):
The Development of Multivariate Analysis Methodologies for Complex ToF-SIMS Datasets: Applications to Materials Science

\[ M = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1n} \\ I_{21} & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ I_{m1} & \cdots & \cdots & I_{mn} \end{bmatrix} \]

The first PCA assumption is the linearity of the data. The method asks whether there is another basis, which is a linear combination of the original basis, that best re-express the data set in terms of its covariance. The linear transformation of the original set is \( S = PM \). Where the rows of \( P \) are a set of new basis vectors for expressing the columns of \( M \).

Writing \( S \):

\[ S = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ p_{m1} & \cdots & \cdots & p_{mm} \end{bmatrix} \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1n} \\ I_{21} & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ I_{m1} & \cdots & \cdots & I_{mn} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{m} p_{1k}I_{k1} & \cdots & \sum_{k=1}^{m} p_{1k}I_{kn} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ \sum_{k=1}^{m} p_{mk}I_{k1} & \cdots & \sum_{k=1}^{m} p_{mk}I_{kn} \end{bmatrix} \]

The Equation shows that every \( j^{th} \) column of \( S \) is a vector projected on the \( j^{th} \) dimension of the new basis. With this linearity assumption (that every data point can be expressed as a linear combination of an orthonormal basis), now it is necessary to find the best way to re-express the data set by finding a new basis with the principal components.

Recalling the example with two peaks shown in Figure 3.1, it is possible to see that the peaks at 73 u and 147 u have a linear relationship. If we include two more peaks, 55 u (\( \text{C}_4\text{H}_7 \)) and 69 u (\( \text{C}_5\text{H}_9 \)) in the examples, we can investigate which pairs are dependant by plotting their intensities against each other as shown in Figure 3.2.
Figure 3.2: Normalised intensities of peaks at masses 55, 69, 73 and 147 u plotted against each other for 97 arbitrary samples (each data point represents the peak areas for a different spectrum) in a symmetric matrix where the diagonal shows the histograms of intensities for each mass.

The 55 and 69 u peaks intensities do not show a linear relationship with the intensities for the 73 and 147 u but are linear between them, meaning that the intensity of one depends on the intensity of the other, so it is possible to infer the presence of the second by the counts of the first. This means they are redundant or correlated. To generalize the problem for m peaks and quantify redundancy, we can define the variance of this data set. For example, the variances \( \sigma \) of the peaks at 73 u and 147 u individually are defined as:
The development of multivariate analysis methodologies for complex ToF-SIMS datasets: Applications to materials science

\[
\sigma^2_{73} = \frac{1}{n} \sum_k (I_{73,k})^2; \quad \sigma^2_{147} = \frac{1}{n} \sum_k (I_{147,k})^2
\]

Therefore, a straightforward generalisation of the covariance between a pair of two peak intensities is:

\[
\sigma^2_{I_1, I_2} = \frac{1}{n} \sum_i I_{1,i} I_{2,i}
\]

Considering the data set \( \mathbf{M} \), its covariance matrix \( \mathbf{C}_M \) is defined as:

\[
\mathbf{C}_M = \frac{1}{n} \mathbf{M} \mathbf{M}^T
\]

where each \( ij \)th element is:

\[
c_{ij} = \frac{1}{n} \sum_{k=1}^{n} I_{1,k} I_{j,k}
\]

Since for any matrix \( \mathbf{A} \), \( \mathbf{A} \mathbf{A}^T \) is symmetric [23], \( \mathbf{C}_M \) is a symmetric square \( m \times m \) matrix and captures, between every possible pair of peak intensities, signal/noise in the diagonal terms (variances) and redundancy in the off-diagonal terms (covariances). A possible practice for ToF-SIMS data is to use standardized (auto-scaled) variables \( Z(I) \):

\[
Z_i = \frac{I_i - \mu_i}{\sigma_i}
\]

where \( \mu \) is the mean value of \( I_i \) for all \( n \) spectra. The covariance matrix calculated from standardized variables is also called correlation matrix and its values only vary in a real scale from -1 (anti-correlated) to 1 (correlated) with all diagonal values being 1. Figure 3.3 represents the correlation matrix of the 4 peaks dataset shown in Figure 3.2.
Figure 3.3: Correlation matrix of the 4 peaks dataset shown in Figure 2.3. The colour scales mean the calculated values of correlation. NB: the values for correlation are within -1 and 1.

As expected, the correlation matrix shows that there are two pairs of correlated peaks (55 & 69 and 73 & 147) and no correlation at all (values around zero) for cross combinations. The covariance matrix has much higher values for the peaks with lower masses because their intensities are much higher in comparison to the background noise.

Remembering the transformation $S = PM$, to minimise redundancy and maximise signal, the covariance matrix of $S$ ($C_S$) should have zero values for off-diagonal terms (which means $S$ is de-correlated) and each successive dimension in $S$ should be rank-ordered by variance. As the PCA assumption is that $P$ is an orthonormal matrix, the goal is: Find some orthonormal matrix $S$ such that $C_S = (SS^T)/n$ is a diagonal matrix. The rows of $P$ are the principal components of $M$. Rewriting $S$: 

\[ C_S = SS^T/n \]
\[
C_S = \frac{1}{n} SS^T = \frac{1}{n} (PM)(PM)^T = \frac{1}{n} PMM^T P = P \left[ \frac{1}{n} MM^T \right] P^T
\]

\[
C_S = PC_M P^T
\]

Which was achieved using the property \((AB)^T = B^T A^T\). For the next step, two theorems are needed [61]:

**Theorem 1.** The inverse matrix of an orthonormal matrix \(A\) is its transpose matrix \((A^{-1} = A^T)\).

**Theorem 2.** For a symmetric matrix \(A\), \(A = EDE^T\), where \(D\) is a diagonal matrix and \(E\) is a matrix of eigenvectors of \(A\) arranged in columns.

If we choose \(P\) to be a matrix where each row is an eigenvector of \(C_M\):

\[
C_S = PC_M P^T = P(P^TDP)^T = PP^{-1}DP = PP^{-1} = D
\]

Which means the choice of \(P\) diagonalises \(C_S\).

In summary, the principal components of a dataset are a new basis defined by the eigenvectors of the covariance matrix of the data set and are rank-ordered by the variance of the transformed data set. If the matrix \(M\) was instead arranged as \(n\) spectra \(x m\) peaks, the same conclusions could be reached by using the transformation \(S = MP\).

Matrix \(M\) can then be expressed:

\[
M = SP^T
\]

Matrix \(S\) is called *Scores* and its columns are interpreted as the projections of each data point in the new basis. Matrix \(P\) is called *Loadings* and its columns are the projections
of the old axis onto the new ones. A simple visualization of this rotation can be seen for the two peaks case in Figure 3.4.

![Figure 3.4: Simple visualization of PCA for the two peaks case. Since the new components are always orthogonal, PCA is a rotation of axis.](image)

Another algebraic solution is derived from singular value decomposition (SVD) and can be seen in detail in [61]. The SVD is a factorisation of $M$ of the form:

$$M = U\Sigma P^T$$

Where $\Sigma$ is a diagonal matrix of positive numbers called the singular values of $M$; the columns of $U$ and $P$ are orthogonal unit vectors called left and right singular vectors of $M$. The covariance matrix of $M$, $C_M$, can be written in terms of the SVD:

$$n \cdot C_M = M^TM = P\Sigma U^T U\Sigma^T = P\Sigma^T \Sigma P^T$$

Which is on the same form as the eigenvector decomposition of $C_M$ given in Theorem 2.
3.2.2 Algorithms for PCA

In order to account for the true variance of a dataset, matrix $M$ must be mean-centred before the calculation of principal components. Mean-centring consists of subtracting each column of $M$ by their average so that the values are distributed around zero. Figure 3.5 shows the problem with non mean-centred data for a simple case with two variables. If the data is not mean-centred, the first principal component will represent the average of the dataset and the subsequent components, due to the orthogonality constraint, will not be the directions of highest variance. Mean-centring is part of a series of steps called data pre-processing. The relevant pre-processing steps for ToF-SIMS data and their consequences will be presented in Chapter 4.

Figure 3.5: The problem with non mean-centred data for PCA in a 2-dimensional case. If the data is not mean-centred (A), the first principal component will represent the mean point of the dataset. When the data is mean centred (B), the principal components show the true directions of variance of the data.
In summary, a basic algorithm for obtaining the principal components of \( \mathbf{M} \) is as follows:

1. Mean-centre \( \mathbf{M} \)
2. Calculate the covariance matrix \( \mathbf{C}_\mathbf{M} = \mathbf{M}\mathbf{M}^T/n \)
3. Solve \( \mathbf{C}_\mathbf{M}\mathbf{V} = \lambda\mathbf{V} \) to obtain eigenvectors \( (\mathbf{V}^i) \) and eigenvalues \( (\lambda^i) \) of \( \mathbf{C}_\mathbf{M} \)
4. Construct loadings matrix \( (\mathbf{P}) \) with eigenvectors \( \mathbf{V}^i \) in each column ranked by corresponding eigenvalues \( \lambda^i \)
5. Calculate scores matrix \( \mathbf{S} = \mathbf{M}\mathbf{P} \)

For very large datasets, that are typical of ToF-SIMS, step 2 is computationally very demanding. For this reason, alternative algorithms for iterative approximate calculations of the principal components have been developed \([62–64]\). Some of the algorithms will approximate the values of the covariance matrix, other will skip its calculation by, for example, computing the singular value decomposition of matrix \( \mathbf{M} \). These are implemented in most statistics packages of modern matrix-based programming languages such as MATLAB, R or Python.

### 3.3 Non-negative matrix factorisation

The previous Section showed that the original data matrix can be expressed as a product between its principal components scores and loadings matrices. Non-negative matrix factorisation (NMF) is a similar method to PCA but rather than the orthogonality constraint between components, all values on both matrices must be non-negative. This is of great usefulness for analytical chemistry data where negative values will most of the times make no physical/chemical sense, which makes difficult the interpretation of
results. The problem formulation for NMF, thus, is [65]: For a non-negative data matrix $M$ sized $n \times m$, find two non-negative matrices $W$ (sized $n \times k$) and $H$ (sized $m \times k$, $k \ll m$) that factorise $M$ as well as possible:

$$M = \sum_{i=1}^{k} w_k h_k^T + E = WH^T + E$$

Where $w_k$ and $h_k$ are the columns of matrices $W$ and $H$ and matrix $E$ represents the approximation error. This bi-linearity is often called in analytical chemistry an unmixing model where the original data matrix is interpreted to be a linear combination of “pure” components spectra (matrix $H$) at different concentrations (matrix $W$). Therefore, $W$ and $H$ will be much smaller matrices as there are usually fewer pure components than there are variables in the original dataset. Figure 3.6 shows a visual representation of the NMF problem for ToF-SIMS data.

Figure 3.6: NMF interpretation for ToF-SIMS data.
There are more formulations for an NMF problem (such as symmetric, semi-orthogonal, multi-factor, etc [65]) but these are not as popular in the ToF-SIMS community and are outside the scope of this thesis.

The fact that NMF is an approximation means that there is no exact solution and iterative approximate solutions must be employed. Furthermore, the available algorithms will find, for most problems, only local minima. The non-uniqueness of an NMF solution can be easily shown if one considers any non-negative, invertible matrix D so that for given solutions W and H, WD and HD are also solutions:

\[ M \approx WH^T \]

\[ (WD)(HD)^T = WDD^TH^T = WH^T \approx M \]

In order to achieve the best possible NMF solution, one must use appropriate data pre-processing, addition of constraints and algorithms [65, 66]. Aspects of data pre-processing and constraints will be discussed in Chapter 4. Next sub-sections aim to describe the mathematical concepts of the two classes of NMF algorithms used throughout this thesis.

### 3.3.1 Alternating least squares NMF

Given an approximate matrix factorisation of the form \( X = aB + e \), the approximation error e can be expressed as:

\[ e = X - aB \]
So, the “sum of squares” of the elements in \( e \) (as a function of matrix \( B \)) can be written as [67]:

\[
S(B) = \sum e_i^2 = e^T e = (X - aB)^T(X - aB) = X^T X - X^T aB - B^T a^T X + B^T a^T a B
\]

Therefore, in order to minimise \( S \):

\[
\frac{\partial S}{\partial B} = -2a^T X + 2a^T a B = 0,
\]

\[
a^T a B = a^T X,
\]

\[
B = (a^T a)^{-1} a^T X
\]

This is known as the “least squares estimation” of \( B \). An alternating least squares (ALS) NMF, as the name suggests, is an algorithm that consists of iterations that will have a least squares solution for matrix \( H \) followed by another least squares solution for matrix \( W \). The fact that either matrix \( H \) or \( W \) are fixed allows a simple least squares solution of the problem. A basic ALS NMF algorithm is described as follows [66, 68]:

1. Initialise \( W \)
2. \( H = M^T W (W^T W)^{-1} \)
3. Set all negative elements of \( H \) to zero
4. \( W = M H (H^T H)^{-1} \)
5. Set all negative elements of \( W \) to zero
6. Repeat 2-5 until convergence
The non-negativity of the matrices is implemented in steps 3 and 5. Depending on the application, more constraints can be added in those steps such as unimodality, closure and low rank constraints [10, 69].

3.3.2 Multiplicative update NMF

The multiplicative update (MU) algorithms for NMF were first introduced by Lee and Seung [70, 71] and many of the modern approaches are modifications of one of their original algorithms that seeks to minimise the Euclidean distance between $M$ and $WH^T$, defined as:

$$
\|M - WH^T\|^2 = \sum_{ij} [M_{ij} - (WH^T)_{ij}]^2,
$$

that clearly is minimum when $M = WH^T$. The original MU algorithm is as follows:

1. Initialise $W$ and $H^T$ with respectively $n \times k$ and $k \times m$ random positive numbers
2. $H^T = H^T .* (W^T M) ./ (W^T W H^T)$
3. $W = W .* (M H) ./ (W H^T H)$
4. Repeat 2 and 3 until convergence

The symbols “.*” and “./” are the MATLAB notation for term by term multiplication and division, respectively. Those multiplicative updates guarantee that $H$ and $W$ will remain non-negative and constantly decrease the Euclidean distance after every iteration [66, 71].

With a Bayesian approach, NMF can be regarded as a maximum likelihood estimation problem [65, 72, 73]. If the measured data is expected to have a Gaussian
probability distribution, the estimation consists on minimising the Euclidean distance
and the already introduced algorithm suffices. However, if the data is, for example,
expected to have a Poisson probability distribution, a different objective function must
be minimised, which is called the Kullback-Leibler divergence $D$:

$$ D(M|WH^T) = \sum_{ij} \left( M_{ij} \log \frac{M_{ij}}{(WH^T)_{ij}} - M_{ij} + (WH^T)_{ij} \right) $$

The divergence is also minimum when $M = WH^T$ and the MU algorithm proposed by
Lee and Seung that minimises it is as follows [71]:

1. Initialise $W$ and $H$ with respectively $m \times k$ and $n \times k$ random positive numbers

2. For each element $H_{aj}^T$ and $W_{ia}$:
   
   a. $H_{aj}^T = H_{aj}^T \frac{\sum_i W_{ia} M_{ij} / (WH^T)_{ij}}{\sum_k W_{ka}}$
   
   b. $W_{ia} = W_{ia} \frac{\sum_j H_{aj}^T M_{ij} / (WH^T)_{ij}}{\sum_l H_{al}^T}$

3. Repeat 2a and 2b until convergence

As discussed before, the variables of a ToF-SIMS dataset are the spectral channels and,
in a micro channel plate detector, these will be ruled by counting statistics and the counts
per channel can be approximated by the Poisson distribution, which makes the latter MU
algorithm a better approach for ToF-SIMS data. However, this is the least popular
algorithm within the community and the potential implications of not using it will be
discussed later in this thesis.
3.3.3 Initialisation of matrices and choice of number of components

The three NMF algorithms described in this Chapter suggest the initialisation of matrices \( W \) and \( H \) with random numbers, however, more specific initialisation can aid NMF by avoiding undesired local minima of the approximate solutions. Two common initialisation methods are: i) selection of random samples of dataset \( M \) (row-wise for \( H \) and column-wise for \( W \)). ii) performing PCA prior to NMF in order to identify the “purest” rows or columns of \( M \) and use them to initialise matrices \( W \) and \( H \).

With regards to choosing the optimal number of components, one approach that works well with ToF-SIMS data is the calculation of principal components prior to NMF to estimate how many “pure” compounds may be present in a dataset. This is usually used in conjunction with knowledge about the sample under analysis. A more robust method is the calculation of cophenetic coefficients which is also known as contrast analysis [74] and consists of calculating NMF components several times for various increasing number of components and analysing stability by checking metrics such as standard deviation within the obtained results per number of components. When there is a significant change in stability, it means that the optimal number of components has been reached. The problem of such an approach is that it may be very time-consuming for very large datasets which are typical of ToF-SIMS.

3.4 \( k \)-means clustering

The \( k \)-means algorithm is an iterative method that partitions multidimensional data into clusters accordingly to their means so that the clusters contain points with the minimum
distance to a number of centres \((k)\) determined beforehand [75]. Each sample will have \(m\) coordinates which are the variables of the dataset. The basic steps of the algorithm are as follows:

1. Initialise \(k\) centres.

2. **Assignment:** A specific distance measure between every data point and the \(k\) centres is calculated and each data point is attributed to their closest cluster centre.

3. **Update:** Attribute new values to the clusters centres by calculating the mean of the data contained in each cluster.

4. Repeat 2 and 3 until convergence in assignment.

The initialisation of the \(k\)-centres can be done in several different ways from selecting random coordinates to more sophisticated approaches such as running the algorithm in a much smaller subsample of the dataset until convergence.

Figure 3.7 shows an example of the algorithmic steps for a dataset with 2 dimensions and 4 clusters. In the case of ToF-SIMS datasets the dimensions will be the hundreds or thousands of mass peaks, which can be a problem as the distances between different points become indistinguishable when a large number of coordinates is used, leading to instability and lack of convergence. Such effect is known as the *curse of dimensionality* [76] and one way to avoid it is to apply \(k\)-means clustering *after* NMF or PCA in a space with reduced dimensions.
Figure 3.7: Algorithmic steps of $k$-means clustering for an example with 2 dimensions and 4 clusters.

The most commonly used distance measure is the Euclidean distance, however, there are more distance measures that can be used for different datasets. One particularly effective measure is the correlation distance: for a pair of $n$-dimensional points $P$ and $Q$
with respective coordinates $q_1, \ldots, n$ and $p_1 \ldots n$, the correlation distance between $P$ and $Q$, $d(P, Q)$ is defined as:

$$d(P, Q) = 1 - \frac{(P - \overline{P})(Q - \overline{Q})}{\sqrt{(P - \overline{P})^2 \sqrt{(Q - \overline{Q})^2}}} = 1 - \frac{\sum_{i=1}^{n}(p_i - \overline{p})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2} \sqrt{\sum_{i=1}^{n}(q_i - \overline{q})^2}}$$

Where $\overline{P}$ and $\overline{Q}$ are the component-wise mean of each point.

### 3.5 Final remarks

The three MVA methods presented in this Chapter are nowadays categorised as *unsupervised machine learning* as they do not require any training datasets or controlled input from the analyser as opposed to *supervised learning* methods that have recently started to become more popular within the SIMS community (particularly classification models [33])

Some comments were already made in this Chapter with regards to the specific application of those methods to ToF-SIMS data, but there is still a number of peculiarities and procedures that must be adopted in order to achieve practical application and interpretation of MVA of ToF-SIMS. These will be discussed in the next Chapter.
4 MVA

METHODOLOGIES FOR TOF-SIMS DATA

4.1 Introduction

Chapter 3 of this thesis described the mathematical steps necessary to perform multivariate analysis of a data matrix. However, there are several ways to obtain a data matrix from a set of ToF-SIMS spectra, and a series of steps must be taken to obtain appropriate datasets that will return MVA results of the required quality. Furthermore, because of the complexity of ToF-SIMS data, post-MVA processing such as data visualisation methods must be carried out to achieve effective interpretation of results. The following Sections describe all pre-processing and post-processing steps that are specific of ToF-SIMS.
4.2 Data pre-processing

Before the application of MVA to ToF-SIMS data, there is a series of steps that must be taken in order to make the dataset suitable for the desired MVA method. These steps are known as “data pre-processing” and this Section aims to describe them.

4.2.1 Selection of variables

The variables of a ToF-SIMS dataset are the chemical information provided by the spectra, which is mostly present in the peaks at different spectral channels. Disregarding possible overlaps between neighbour peaks, all channels comprising a specific peak will represent the same secondary ion and therefore be somewhat redundant. For this reason, a data matrix is usually constructed by integrating the areas under peaks of each measurement and assigning to it either the peak centre channel, the mass related to its centre or the chemical structure of the ion. This practice is known as “peak-picking” and is the most common way of creating a data matrix out of ToF-SIMS measurements. The main advantages of peak picking are the elimination of channel redundancy, the increase of signal-to-noise ratio for every variable and the fact that it keeps high mass resolution assignments. To achieve precise peak-picking, the analyst must assess every spectral channel, determine the right mass span for each peak and make assignments. This practice demands a lot of time and alternatives such as automated peak search algorithms or mass binning were developed [77, 78]. Peak search algorithms will depend upon search parameters such as peak width or minimum signal-to-noise ratio. It is a much faster approach but it will not have ion assignments and is subject to artefacts such as
noise spikes being regarded as peaks. Channel binning is simply done by integrating all counts within each nominal mass and assigning the values as variables. A more sophisticated channel binning approach separates counts on the right and left-hand sides of the nominal mass, effectively separating inorganic from organic fragments in most cases (as illustrated in Chapter 2). This third approach relies on precise channel-to-mass calibration and may split peaks in half when they are of very high masses or organic fragments with many oxygen or nitrogen atoms are present, as these have exact mass higher than nominal mass and will shift a molecular fragment mass towards its nominal mass. Figure 4.1 shows a comparison among the most common ways of selecting variables.

![Figure 4.1: Ways of selecting variables for a SIMS dataset. (A): peak-picking; (B): inorganic/organic nominal mass binning; (C): nominal mass binning.](image-url)
All these procedures can be regarded as data compression, which means information is lost after peak selection. There has recently been a trend within the SIMS community towards processing uncompressed datasets in order to remove any bias added by peak search routines. The work developed for this thesis contributed to the field by performing NMF of uncompressed SIMS data for the first time, topic of which will be discussed later, in Chapter 7.

4.2.2 Data unfolding

The most complex dataset one can generate with ToF-SIMS is a 3D dataset. As described in Chapter 2, with the use of a dual beam set up it is possible to have lateral and depth information arranged in a 3D hyperspectral cuboid where every voxel will contain a full mass spectrum. All MVA methods used throughout this thesis are based on having a 2-dimensional data matrix with variables in one dimension and observations (or samples) in another, therefore, in order to perform matrix factorisation methods such as PCA and NMF on those datasets, the data must first be *unfolded* in a manner that enables the final results to be folded back without loss of spatial information. This is done by stacking the spectra (or rows of peak areas) following a specific order so that their original spatial information can be restored after any data manipulation. Figure 4.2 shows a flowchart describing how the unfolding process is done for a 3D dataset in order to create a matrix $M$ that will undergo MVA following the methods described in Chapter 3. In case of imaging datasets there are no further levels and only level 1 is unfolded and processed whereas in case of depth profiling data, the lateral information is collapsed into one data point per level. For spectra analysis, every measurement is regarded as one
sample and technical repeats are acquired often to increase statistics. There are more MVA methods (such as tensor factorisation [65] or PARAFAC [79]) that do not require data unfolding but these are outside of the scope of this thesis.
Figure 4.2: Data unfolding and construction of matrix M for a 3D dataset.
4.2.3 Mean-centring

As introduced in Chapter 3, matrix \( M \) must be *mean-centred* so that PCA accounts for the true variance of the dataset. Mean-centring consists of subtracting each column of \( M \) by their average so that the values are distributed around zero. Figure 3.5 shows the mean centred process for a matrix with two variables but a ToF-SIMS dataset will have many more variables and a good way to visualise the effect of mean-centring is to plot the total “spectrum” of all samples before and after centring, as shown in Figure 4.3. Mean-centring is not done prior to NMF as it would create negative values.

![Figure 4.3: Different pre-processing steps for a ToF-SIMS dataset.](image-url)
4.2.4 Normalisation

For a ToF-SIMS dataset, the values of the elements of matrix $M$ will always need some sort of correction. In most cases, the first correction that must be applied is data normalisation. Normalisation consists in dividing all elements of each row of $M$ by a common value. In other words, it is a correction that is done on the individual spectra of the dataset. The values used for normalisation can be a specific variable (intensity of a specific peak of the spectrum) or a combination of a number of variables, such as the total ion intensity of a spectrum or the total intensity of all variables present in a dataset (after variables selection). Typically, the main reason for normalisation is to account for differences in acquisition conditions of individual spectra. These conditions can be, for example, different acquisition times, different primary ion beam current or secondary ion yield hindered by charging and topography effects. Topography effects are usually seen in imaging datasets, where some regions rastered by the primary ion beam will have different heights and may be out of the optimal focal plane of the secondary ion analyser, which will result in differences in intensity for all variables. Such differences are not related to surface chemistry and, unless corrected, will be picked up by MVA methods. Figures 4.4 and 4.5 shows the example of a ToF-SIMS imaging dataset of a particle containing potassium. If the data is not normalised, it looks like there is less potassium in some regions of the particle, which is not the reality. This is an extreme case where some regions of the particle are almost completely out of the depth of field of the analyser, which results in very noisy spectra for the pixels covering such regions and a normalised map that shows edge artefacts.
4.2.5 Scaling

The second correction that must be done to a ToF-SIMS dataset is scaling. The major cause for the necessity of data scaling is the fact that variables will have different amounts of noise and error (heteroscedasticity). As briefly discussed in Chapter 2, the statistics of secondary ion detection in ToF-SIMS can be described, at each spectral channel, by the Poisson probability distribution \( P(N) \), that predicts the probability of recording \( N \) counts in a single measurement of duration \( t \):

\[
P(N) = \frac{e^{-t} t^N}{N!}
\]
\[ P(N) = \frac{\mu^N e^{-\mu}}{N!} \]

Where \( \mu \) is the mean value. The standard deviation of a Poisson variable can be estimated as simply the square root of the variable itself, thus a straightforward way of “normalising” the error across all variables is by dividing each value of data matrix \( M \) by its square root:

\[ Ms = \frac{M}{M^{\frac{1}{2}}} \]

Where “./” denotes MATLAB element-wise division. However, typical ToF-SIMS imaging datasets will have very low counting rates that reflect in large relative uncertainties and make the square root estimation inaccurate. For this reason, Keenan and Kotula [8] proposed a scaling approach that takes into account the relationship among all variables by using the row-wise and column-wise means of the data matrix. A scaled dataset is then written as:

\[ Ms = (G^{-1/2})M(H^{-1/2}) \]

Where \( G \) is a diagonal matrix with the unfolded mean image (row-wise mean) along its diagonal and \( H \) is a diagonal matrix with the mean spectrum along its diagonal (column-wise mean). This method of scaling is nowadays widely used in the SIMS community and is known as “Poisson scaling”.

On the other hand, when a dataset has too many counts, detection dead-time effects will occur and lead to non-linearities between peaks [57]. In this case, similar statistical approaches were applied and there are now established models for appropriate
pixel-wise dead-time correction [80–82]. Modern instruments such as the TOF.SIMS 5 will have this correction by default in any data exported from the analysis software.

Another method that is more commonly used in other fields (but sometimes applied to ToF-SIMS data) is known as “auto-scaling” and consists of mean-centring the data and dividing each element by its column-wise standard deviation. In other words, auto-scaling converts each value to its distance to the mean in units of standard deviation. Figure 4.3 shows the average spectrum of an example dataset before and after mean-centring, Poisson scaling and auto-scaling.

All pre-processed datasets shown in Figure 4.3 can produce meaningful MVA results when careful interpretation is aligned with knowledge of the experiment and the sample. Sometimes the “true” direction of variance of a dataset is not the most relevant and centring the data around a different point can be more useful [83]. Other times the list of variables may have very similar signal-to-noise ratio and Poisson scaling the data may only enhance noise.

### 4.2.6 Matrix augmentation

One way to go about dealing with several similar large imaging ToF-SIMS datasets is by means of the application of MVA on a matrix that contains the intensities of a set of peaks for every pixel of each image arranged in rows. This is known in the literature as “row-wise matrix augmentation” [84] and has been demonstrated with ToF-SIMS data [85]. “Column-wise augmentation” is also possible when more than one independent detector acquire data for the same set of samples (or pixels) [86, 87]. Another practice
for MVA is the insertion of rows containing measurements of well-known standard samples that may be part of the samples under investigation. The Chapters in Part III of this thesis will establish this approach to be useful as it sets directions of high variance within the dataset and ultimately aids the identification of such pure samples in data of mixed systems.

### 4.2.7 Pixel/voxel subsampling

Imaging or 3D datasets generated by modern ToF-SIMS instruments can be extremely large (millions of pixels x millions of variables), which makes it impossible to perform the required matrix manipulations and operations presented in Chapter 3 using conventional computers due to memory and time limitations. One approach to overcome this problem is the use of *training sets* by means of pixel/voxel subsampling [25–27, 29]. Figure 4.6 shows the basic steps of the subsampling method for the NMF case: Matrix $H_s$ is calculated by factorising a reduced set of rows ($A_s$) of matrix $A$. Matrix $W$ is then calculated using the pseudo-inverse of $H_s$ and the original $A$. The exact same steps are taken for PCA analysis and for $k$-means, the iterative process can be done in the subsampled dataset, with all pixels only being considered for a last iteration.

Subsampling works for ToF-SIMS data because, for every impact point of primary ions, there is a fundamental volume (that usually spans more than one voxel) where all generated secondary ions will be highly correlated. However, potential problems are the discrepancy of random selection of voxels, especially on datasets containing compounds at very low concentrations. This has been addressed by Cumpson et. al [27] with the use
of Sobol sequences and a direct application of this approach for NMF followed by a few systematic tests are discussed in Chapters 7 and 8 of this thesis.

Figure 4.6: Sub sampling scheme. Matrix Hs is calculated by factorising a reduced set of rows of matrix A. Matrix W is then calculated using the pseudo-inverse of Hs and the original A.

4.3 Visualisation of MVA results of ToF-SIMS data

MVA intends to simplify complex data but its results are often difficult to visualise. Both PCA and NMF methods will always yield two main results matrices, which are known as Scores and Loadings when they are PCA outputs and Intensities and Characteristic spectra when NMF outputs. This Section will present several options to explore MVA results data by means of visualisation always with a final goal of having one final figure containing the most information in a non-cluttered way.

As presented in Chapter 4, there are a considerable number of publications that deal with MVA of non-spatially structured SIMS data and it is fair to say that there is a
very well-established manner to visualising these results: Scatter plots for the scores matrix and discrete sequence plots (e.g. bar plots) for the loadings matrix. Therefore, these are going to be the starting point of the visualisation, rather than looking at numbers in the matrices.

To illustrate different levels of complexity of visualisation of non-spatially structured data, standard carboxylic acids are used. Two to three ToF-SIMS spectra were acquired for each of the following acids with structural formula present in Figure 4.7: adipic, pimelic, suberic and tri-carboxylic. The experimental details of the first three acids (in addition to four others) and results are discussed more fully in Chapter 6. Peak areas of ion fragments were obtained for each spectrum. These areas were normalised by total ion counts, Poisson scaled and mean centred and matrix $M$ was constructed with the peak areas of each spectrum organised in columns.

Figure 4.7: Structural formula of standard carboxylic acids. (a): Adipic acid. (b): Pimelic acid. (c): Suberic acid and (d): Tri-carboxylic acid.
Each column of a scores matrix will be the coordinates of the data points in the principal components space and each line of a loadings matrix will be the contributions of the original variables to the principal components. Figure 4.8 shows both scores and loadings values for the first three principal components plotted separately.

Figure 4.8: Scores and loadings of the carboxylic acids dataset plotted separately for first three principal components.

4.3.1 Use of a priori information and minimisation of the number of plots

Figure 4.8 is already much simpler than having to plot all spectra of all samples and look for differences, however, there are more and better ways to visually explore these results. For the scores matrix, in most cases it is useful to input prior information regarding the measurements, which in this example is the fact that there are only three kinds of sample
with 2 to 3 technical repeats. This means that the scores of each kind of sample can be grouped and plotted using varied symbols and colours (Figure 4.9 (A)). Furthermore, as the principal components can be seen as a rotation of the original variables (peak areas) coordinates system, it makes sense to investigate “2D planes” or “3D hyperplanes” of this new space by plotting the scores of specific principal components against each other (Figures 4.9 (B) and (C)).

**Figure 4.9: 2D and 3D scatter plot of scores of principal components.**

Sometimes more than 3 principal components will be necessary to describe a dataset. A very compact way of visualising scatter plots of all relevant components against each other is a “plot matrix” like the one presented in Figure 4.10 (5 components were chosen arbitrarily in this case). Every element of a “plot matrix” is a scatter plot between two variables represented by the rows and columns. If the rows and columns represent the same variables (which is the case of Figure 4.10), the plot matrix will be symmetric and the plots along the diagonal can be replaced by histograms of their respective variables.
Figure 4.10: Plot matrix of scores 5 principal components of the carboxylic acids dataset.
4.3.2 Show scores and loadings together: the biplot

A biplot is a projection of the original variables axes onto a plane of the principal components space. In other words, it is a single plot that shows both loadings and scores of 2 or 3 principal components and even though it is a very standard way for the visualisation of PCA results [88], it is not commonly used within the mass spectrometry community because of the often large number of variables that make the plot look polluted and non-informative. However, a threshold can be applied for the visualisation of only the most significant variables that would still aid interpretation of the results. A biplot for PCA results of ToF-SIMS data is more effective if both scores and loadings are normalised by the maximum of their absolute value, which creates a common scale ranging from -1 to 1.

The contributions of the original variables are not exclusive of specific principal components, so a biplot helps to find out the most characteristic peaks of each group of samples (or each carboxylic acid in case of Figure 4.11).

Figure 4.11: 2D and 3D scatter plot of a PCA biplot.
4.3.3 Clustering algorithm to separate regions in a scores plot: Voronoi cells.

Voronoi cells are delimited regions based on distances to specific points in a plane or cube [89]. Those points can be for example, centres of clusters in the principal components space. For the acids dataset, they can be determined using a $k$-means clustering algorithm applied to the scores of the first three principal components. Once the centres are determined, all points in the visible space PC1 x PC2 x PC3 are tested to check which centre is the closest and coloured accordingly. This is a useful visualisation tool especially for the 3D scatter plots that can be difficult to see in depth.

Figure 4.12 (A) shows the Voronoi cells obtained with two categories clustering: it helps to visualise the differences between dicarboxylic and tricarboxylic acids. Figure 4.12 (B) shows the same space now classified with 4 clusters: 1 category for each acid. This kind of visualisation is sometimes more useful for live presentations (with the possibility of rotating the axes) or meetings for discussion of results or even for the analyst to investigate their data.
Figure 4.12: 2D and 3D scatter plots of principal components scores with voronoi cells created based on $k$-means clustering. Top: 4 clusters. Bottom: 2 clusters.

4.3.4 Web plots

All plots presented in this Section would work as well for outputs of NMF. One very simple and useful plot that can be done for non-negative components intensities that allows the presentation of more than 3 components in one single plot, is a web (or
radar) plot of the averages of scores by sample groups. Figure 4.13 shows a web plot with the averaged intensities of NMF results of the acids dataset (5 components).

![Web plot with NMF results with 5 components for the acids dataset.](image)

**Figure 4.13: Web plot with NMF results with 5 components for the acids dataset.**

As expected, 5 components were too many in this case and components 2 and 4 present the same information, which makes one of the components redundant. Moreover, each component was exclusive for just one acid due to the simplicity of this dataset. In more complex cases there will be different relative amounts of each component and the web plot will still be useful to visualise results of more than 3 components. Figure 4.14 shows results from ToF-SIMS data of coffee from 4 different farms in Brazil, more details of the experiment are presented in Chapter 11 of this thesis.
Figure 4.14: NMF results with 6 components for the coffee dataset.

There is one assumption that one must have in mind when producing web plots: the pre-assigned groups are indeed groups in terms of their component intensities and computing their average is practical.

4.3.5 Visualisation of intensity maps: right choice of colour maps

With the appropriate unfolding procedures described in Section 4.2.2, it is possible to fold back the PCA scores, NMF intensities (matrix $W$) or $k$-means clusters so that the results can be visualised with their spatial distribution. Figure 4.15 shows the folding/unfolding process for a 3D dataset. For 2D datasets the procedure is the same except there are no further levels than the first one.
Figure 4.15: Flowchart describing the unfolding/folding process of a 3D dataset

Intensity maps are the most usual way of visualising a 2D folded matrix and therefore will be the starting point of imaging data visualisation. Image functions such as MATLAB’s `imagesc` will map every value of the matrix into colour scales that are usually referred as “colourmaps”. The right choice of colourmaps will strongly depend on the research field they are used and the information that is wished to be conveyed. A current useful definition for colourmaps is based on the Lab system, where $L$ represents lightness and $a$ & $b$ represent green-red and blue-yellow colour components. Such representation is designed to mimic human perception and all colourmaps can be grouped in three different categories: sequential, diverging and qualitative [90, 91].
Sequential maps will have a constant increase of $L$ between a pair of colours. Diverging maps will be centred around $L = 0$ or $L = 100$ and decrease or increase to two distinct colours. Qualitative (or discrete) colormaps will not have a constant increase or decrease in lightness but will rather be made of very distinct colours. The most popular colormap across all science fields is the classic “rainbow” map, also known by its name in MATLAB: “jet”. However, recent research argues that rainbow maps can be misleading because of their non-linearity and suggests the use of alternative maps [92].

For PCA scores, diverging colormaps are usually the best choice since values will be distributed around zero. Figure 4.16 shows results of a large area ToF-SIMS imaging dataset from a cross-sectional taper of an organic coating on a metallic substrate. The full details and results involving this dataset and other tapers are discussed in Chapter 14 of this thesis. Only principal components 1 and 4 are shown. The maps were created using three different colormaps: a popular rainbow colormap (MATLAB’s jet) and two diverging colormaps to the same pair of colours with the difference being the central point (white or black). The map for PC 1 is quite simple and all colormaps are equally effective. However, for PC 4 the contrast is much clearer when diverging colormaps are used.
Figure 4.16: Folded scores of principal components 1 and 4 for the coating dataset using three different colour maps.
4.3.6 Normalisation for contrast enhancement on intensity maps

Most of the time that PCA is used for the analysis of imaging datasets, the goal, at least initially, is to assess the data in an exploratory and qualitative manner. Thus, in order to enhance the contrast in the intensity maps, one can manipulate the data by re-normalising the Scores matrix. Figure 4.17 shows a comparison of different normalisation methods for the scores of the same PC 1 and PC 4 presented in Figure 4.15 (now only with the colourmap diverging from black). The third column of plots are scatter plots of PC 4 x PC 1 for all pixels (samples in this case). The two most effective normalisation approaches are: normalise by maximum absolute value of a given pair of PC scores or auto-scale the scores matrix as described in Section 4.2.5.
Figure 4.17: Different scores normalisation. Images on the left-hand-side are folded scores maps for PCs 1 and 4. Plots on the right-hand side are scatter plot of scores of PC 1 against PC 4 for all pixels.
If there are still contrast issues for the visualisation of a given principal component scores, another approach that can be taken is shown in Figure 4.18: split positive from negative scores (and loadings) and show them as separate intensity maps (a sequential colourmap is then more appropriate).

Figure 4.18: Separation of positive from negative values. Top: map of scores of PC 4 with Bottom: loadings of PC 4.
4.3.7 Connecting loadings and scores with matching colour maps

Most times MVA results will be assessed by looking at scores and loadings at the same time. For this reason, connecting the visualisation of both matrices can aid interpretation and improve data presentation. Figure 4.19 shows the same map for the scores of PC 4 with a different diverging colourmap together with its loadings plotted with matching colours depending on whether the values are positive or negative.

Figure 4.19: Loadings and scores connected by having the same colour schemes.

4.3.8 Multi-colour overlays for NMF results

A very effective way of presenting NMF results is by means of multi-colour overlays of the components intensity maps. When the overlay is composed of up to three
components, each component, per pixel, can be attributed to separate channels in a red-green-blue (RGB) additive colour model. To overlay more than three maps, each component can be attributed to a colour and their relative intensities are mapped into the transparency/opacity of each colour per pixel, as shown in Figure 4.20. Qualitative colourmaps are the most appropriate for overlays as the more distinct the colours are, the better contrast will be achieved.

Figure 4.20: Map overlay of 5 NMF components

The same plot matrix presented in Figure 4.10 can be constructed for imaging datasets with a set of components scores plotted against each other, as shown in Figure 4.21
where every point in the plot matrix represents a pixel. To connect the plot matrix to the spatial information of each point (or pixel), \( k \)-means clustering can be performed on the scores matrix so that each pixel/point can be coloured accordingly to their cluster, which makes qualitative colourmaps an appropriate choice also for this case.

Figure 4.21: \( k \)-means clustering applied to PCA results of a ToF-SIMS imaging dataset. Left: spatial distribution of clusters. Right: plot matrix of scores of all pixels.

4.3.9 Visualisation of 3D MVA results

Once PCA scores, NMF intensities or \( k \)-means clusters are folded back to their spatial arrangement, the MVA results of a 3D dataset can be regarded as a stack of 2D images that can be visualised individually. In that case, all visualisation approaches described in the previous Sections apply, however, there are also ways of visualising 3D volumes
that may help with data interpretation. In order to exemplify 3D visualisation strategies, a non-interlaced mode ToF-SIMS data stack of a talc-filled automotive grade polypropylene will be used. The sample has a similar formulation to the material analysed in Chapter 9 of this thesis with the addition of talc particles as a filler. The 3D ToF-SIMS data were acquired using the dual-beam depth profiling set-up of the TOF.SIMS 5 with a 25 keV Bi$_3^+$ primary ion beam delivering 0.18 pA of current and raster scanned over a 400 × 400 μm$^2$ area at the centre of the etch crater formed using a 1 keV C$_{60}^+$ beam raster scanned over an area of 600 × 600 μm$^2$. The analysis was performed in the ‘non-interlaced’ mode with an electron flood to neutralise charge build up. The sputter time and pause time per level were set respectively as 0.5 s and 1 s.

A starting point for the visualisation of 3D MVA results can be volumetric slices, which are cross-sectional views of the data and can be presented individually or in a combination of orthogonal slices, as shown in Figure 4.22 (A). From this point, there is a series of visual optimisation steps. For the specific sample of the example, particles are present at the surface of the material and since depth profile data is acquired and stored in a “layer-by-layer” fashion, this creates a visual artefact that gives the wrong impression about the depth distribution of such particles. The correction for such artefact is called “z-correction” and has been reported in the literature mostly for 3D depth profiling data of cells [25, 93, 94]. At the most basic level, the z-correction disregards any possible differences in etch rates of different materials present in the sample, which means that it is simply a shift of specific voxels in the z-direction. To determine which voxels to shift, all cross-sectional views of a specific secondary ion map (or PCA/NMF component) must be assessed. Figure 4.23 shows the steps for the z-correction of the
polymer dataset: a threshold is applied to every cross-section of a specific component in order to identify the presence or absence of particles, then the pixels that fall within this threshold are shifted up. Figure 4.22 (B) shows the same slices of Figure 4.22 (A) but after z-correction.

Figure 4.22 Slices view of PC 1 scores of a 3D dataset (A): before z-correction; (B): after z-correction.
Figure 4.23: Steps for z-correction for one XZ plane. (a): cross-sectional view of intensities of a component or variable. (b): application of threshold to identify positions of particles. (c) edge of particles set number of levels each voxel will be shifted.

Volume slices can be confusing to visualise and interpret after z-correction, so different approaches, such as the use of scatter plots, are more effective. A 3D scatter plot can be seen as an extension of the 2D intensity maps as it will have one data point for each voxel and the colour of a point is proportional to the intensity of the component shown. Figure 4.24 shows a scatter plot of the same principal component before and after z-correction.
Typically, 3D ToF-SIMS depth profiles will probe only from a few tens to hundreds of nanometres in depth, which is much smaller than the maximum beam raster size of 500 x 500 µm². This makes it most times impossible to present data with the correct aspect ratio and an appropriate choice must be made depending on which features the analyst wants to highlight. Figure 4.25 shows the same scatter plot with different aspect ratios. A more expanded z-scale may show hidden information in the filler particles, where a more compressed (and closer to reality) scale will work better as a description of the dataset as a whole.
4.3.10 3D overlays of NMF results

3D NMF results can be shown as multi-colour overlays in a similar way that is done to imaging datasets: each component is attributed to a colour and every point in a scatter point (or pixel in slice views) has its components intensities mapped into their respective colours transparency/opacity. Moreover, 3D overlays can be rendered into isosurfaces to produce a final, visually appealing, representation of MVA results. An isosurface is a surface that represents points of a constant value or fall within the same threshold. Figure 4.26 shows NMF intensities obtained for 3 components of the polymer dataset with a comparison amongst individual components of a scatter plot, a 3D overlay and rendered isosurfaces. Care must be taken with the use of isosurfaces as they are created based on relative intensities thresholds of each component. Figure 4.27 shows the different isosurfaces views that can be obtained by applying different thresholds to the same dataset.
Figure 4.26: Combination of 3D scatter plots of NMF components into an overlay scatter plot that can be rendered as an overlay of isosurfaces.

Figure 4.27: Isosurface overlay obtained upon application of different thresholds.
Sometimes isosurfaces will cover each other and if one wants to show the distribution of NMF components in the inner volume of the data, cross-sectional views should be used. There is, however, another possibility which is the use of transparency. Figure 4.28 shows an XZ plane view of an isosurfaces overlay of only the red and green components with transparency applied to both surfaces. It shows that the particles seem to be covered by the green component that fades at certain depth.

Figure 4.28: Top: summary of MVA results of 3D polymer dataset. Bottom: Side view of isosurfaces overlay of two components with transparency applied.
4.4 Time-resolved Datasets

Several important phenomena in materials science are time-dependant and the way they are observed with analytical techniques such as ToF-SIMS is by periodic acquisition of data of the sample of interest. The resulting datasets can be processed with MVA in the same manner as any other, more standard, kind of measurement with every point in time being a sample (or observation). The following datasets were acquired to illustrate how such processing can be done by using two time-resolved datasets as examples.

4.4.1 Damage in a standard polystyrene sample

ToF-SIMS of a standard polystyrene sample sheet was acquired repeatedly over time using a pulsed 0.3 pA Bi\textsubscript{3}\textsuperscript{+} beam and with each spectrum acquired for 10 seconds. The time scale was then converted into dose scale. Typically, ratios of characteristic secondary ions are tracked in order to obtain damage cross-sections of a material. However, if NMF is applied to a dataset constructed with all peaks, trends can be observed. Figure 4.29 shows the NMF results with the components intensities plotted against the primary ion beam dose. As the sample was not previously cleaned, component 3 (blue) represents hydrocarbon contamination that is removed as more primary ions sputter the surface. Component 1 (red) contains the polystyrene characteristic peaks and component 2 (green) had peaks that are typical of an ion beam damaged polymer material [95]. The maximum value for the polystyrene characteristic peaks occurs for a dose of 1.7 x 10\textsuperscript{13} ions/cm\textsuperscript{2} and the crossover point between undamaged and damaged polystyrene is at 2.5 x 10\textsuperscript{13}. Both these values are slightly
higher than the SIMS static limit introduced in Chapter 2 but this is expected since the primary ion beam used as a cluster beam of Bi$_3^+$. 

![Graph of overlay vs. dose/um cm$^{-2}$](image)

**Figure 4.29:** NMF results for damage study of polystyrene.

### 4.4.2 Diffusion in stainless steel

In order to illustrate the processing of a more complex time-resolved dataset, the following experiment was carried out: a coupon of a stainless steel was marked with a scratch and mounted on a special holder of the TOF.SIMS 5 instrument that is capable of controlling the sample temperature. A direct current (DC) beam of Bi$_3^+$ was rastered over a 200 x 200 µm$^2$ area used to remove any contamination and oxide layers. The sample then was heated up to 300 ºC and 500 x 500 µm$^2$ maps of positively charged secondary ions were acquired periodically on the same area, creating a dataset with the same structure as a 3D depth profiling measurement, as shown in Figure 5.30. The maps
were acquired with 5 scans each using a pulsed Bi$_3^+$ beam in the high current bunched mode, which kept the total dose under the SIMS static limit.

Figure 4.30: Experiment to create a 3D structured time-resolved dataset.

A peak list was then created using the most intense peaks of the total spectrum and the data was unfolded for NMF analysis using 4 components. The intensity maps (matrix $W$) per “level” are essentially “frames” that can be shown as an overlay in a video form as a time-lapse of the diffusion of atoms characteristic of each component. A video with all frames has been published online at www.youtube.com/watch?v=Oq6wlwwJQYw and Figure 4.31 shows a profile view with the total intensity of each component and snapshots of overlays at four time frames. It also shows the monitored values of sample temperature and pressure in analysis chamber.
Figure 4.31: NMF results of time resolved experiment of a varying surface temperature of stainless steel.
Component 1 is dominated by Al\(^+\) related ions, component 2 by Mn\(^+\), component 3 by Cr\(^+\) and component 4 by Si\(^+\). Initially, this experiment was created only to illustrate MVA analysis of a time-resolved dataset, however, the NMF results brought up insights of different diffusion rates of minor compounds present in the steel sample, such as aluminium, silicon and magnesium. A more thorough study is intended as future work.

### 4.5 Final remarks

Pre-processing steps and appropriate data visualisation are essential for achieving meaningful and reliable MVA results of complex ToF-SIMS dataset. Next Chapter present a software suite developed by the author of this thesis that encompasses all methodologies described in this Chapter.
5 The simsMVA Software

5.1 Introduction

As described in previous Chapters, many of the processing methods and visualisation tools used and described in this thesis were not available in any software and this was the initial motivation for the author to create his own set of routines in MATLAB. Over the course of the years, those routines evolved into a software called simsMVA and as Chapter 1 presented, many of the features were developed based on analysis challenges faced by the research group at the University of Surrey. The following Sections aim to present and describe simsMVA as it stands, with the main focus on its usability.

5.2 Main features of simsMVA

simsMVA is developed on MATLAB and, alongside in-house developed functions, makes use of the Statistics & Machine Learning and the Image Processing toolboxes. The current version is optimised to import files generated and exported by IONTOF’s
TOF.SIMS 5 spectrometer but any matrix can also be loaded from MATLAB’s workspace. The main window contains a menu bar with three options: “New tab”, “Theme” and “Help”. The “New Tab” menu contains four different options: “Spectra”, “Images”, “Profiles” and “3D”. Every time one of these options is selected, a new tab is created and the user is prompted to give it a name. The “Theme” menu contains a set of different colour schemes that can be applied to the GUI. A detailed description of each feature present in each kind of tab is given in Table 5.1. Every time a dataset of any kind goes through multivariate analysis (PCA, NMF, k-means or PLS), an “MVA results” tab is created. The next Sections will describe the main features of each mode of simsMVA with some example datasets acquired at The Surface Analysis Laboratory of the University of Surrey.

Table 5.1: Detailed description of main buttons present on each kind of tab.

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Spectra</th>
<th>Images</th>
<th>Profiles</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>New list (Load)</td>
<td>Opens a file selection dialog that allows the user to select matrices from Matlab’s workspace or one or more .txt (Spectra and Profiles) or .BIF6 (Images and 3D) files exported from the SurfaceLab software. Files for different set of samples must have the same peak list.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Add extra</td>
<td>Opens a file dialog that allows the user to select one or more .txt files exported from the SurfaceLab software that will be added to the already loaded files.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set groups</td>
<td>Pops up a window with all sample names. The user can then give the same name to technical repeats and select which samples to be considered for analysis. One the groups are set, the scatter plot on the bottom left figure will be updated with matching colours for samples within the same group.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Description</td>
<td>Available</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>Performs principal components analysis and creates an “MVA results” tab with the PCA results.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMF</td>
<td>Pops up a window (NMF menu) that allows the user to select input parameters for non-negative matrix factorisation (Algorithm, number of components, number of iterations). The Run button creates an “MVA results” tab with the NMF results.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS</td>
<td>Under construction</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k-means</td>
<td>Pops up a window (k-means menu) that allows the user to select input parameters of the Matlab kmeans function (Distance, number of clusters, number of iterations and number of replicates). The Run button creates an “MVA results” tab with the k-means clustering results.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Save project</td>
<td>Opens a save file dialog that allows the user to save the project as a .mat file.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load project</td>
<td>Opens a load file dialog that allows the user to open any saved projects.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line scan</td>
<td>Changes the mouse cursor to a cross. The user then has to click on two or more points on the right hand side map and press enter to display (in the figure at the bottom left hand side) the ion intensity across the line connecting the selected points.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overlay</td>
<td>Pops up a window that enables the overlay of different ion maps.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D view</td>
<td>Pops up a window that shows a 3D view of the intensities of a selected mass peak. The sliders on the left hand side allow the user to slice through the data cuboid.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2.1 Spectra mode

Figure 5.1 shows a screenshot of a Spectra mode tab loaded with data of different wood growth regions. In addition to the wood, samples of reference cellulose and organosolv lignin were also measured. More details about the experiment and results are described in Chapter 10 of this thesis. Several different regions were analysed for each sample at both early and late growth regions and a high spectral resolution peak list with areas of forty-nine characteristic fragments of lignin and cellulose was created. A Spectra mode tab contains a table with all loaded data: peak areas of different samples and their associated masses and labels. The first column contains tick boxes that allow the user to unselect specific variables prior to multivariate analysis. The plots on the right-hand side show the average peak list for all samples. The upper plot will always show the original data and the bottom plot will show the pre-processed data. The panel on the bottom left will contain a list of all variables where the user can select which peak areas to plot against each other in a Matrix. The main data pre-processing steps employed were normalisation by total counts, Poisson scaling and mean centring.
Figure 5.1: Screenshot of a Spectra mode tab loaded with ToF-SIMS data of wood, lignin and cellulose.

If the user clicks the “PCA” button for example, a new “MVA results” tab is created. The tab will contain the loadings on the top panel, the scores on the bottom right panel and the principal components captured variance on the bottom left panel. Figure 5.30 shows a screenshot of a PCA results tab of the wood dataset. The Loadings panel has a slider that controls a threshold of which variables to be named for each principal component. It is also possible to split positive and negative loadings and switch between mass number and labels (usually chemical assignments). The PC 1 loadings shown in Figure 5.2 have lignin characteristic peaks on the positive side and cellulose characteristic peaks on the negative side.
The development of multivariate analysis methodologies for complex ToF-SIMS datasets: Applications to materials science

Figure 5.2: Screenshot of an MVA results tab for Spectra data.

The Scores panel has three drop down menus to select up to three principal components scores to be shown as 2D and 3D scatter plots. The colours of the symbols correspond to previously assigned groups in the Spectra mode tab. There are a few different functions for data visualisation such as averaging groups, radar plots or creation of a custom scale to plot the Scores against. For the wood data set, it is possible to observe that PC 1 separates lignin from cellulose+wood and PC2 separates lignin + cellulose from wood. Two useful functions to visualise these groups are the use of biplots combined with the creation of Voronoi cells, as shown in Figure 5.3.
Figure 5.3: Screenshot of a Scores panel with a combination of biplot and Voronoi cells created on the scatter plots.

For the creation of Voronoi cells, a $k$-means clustering algorithm was applied to the scores of the first three principal components. Once the cluster centres are determined, all points in the visible space PC1 x PC2 x PC3 are tested to check which centre is the closest and coloured accordingly. With the Voronoi cells and biplot on the left-hand side of Figure 5.3 (scatter plot of scores of PC 1 against PC2) it is clear that there are three groups comprising of the pure lignin samples, the pure cellulose samples and the wood samples. The biplot shows the most characteristic peaks of each group.

5.2.2 Depth Profile mode

Figure 5.4 shows a screenshot of a Depth Profile mode tab loaded with data of a metallic layered sample. The profiles were acquired using the dual-beam depth profiling mode of the TOF.SIMS 5 instrument with a 25 keV Bi$_3^+$ primary ion beam delivering 0.18 pA of current and raster scanned over a $50 \times 50 \, \mu m^2$ area at the centre of the etch crater.
formed using a 3 keV Cs\(^+\) beam raster scanned over an area of 400 × 400 μm\(^2\). The depth profiling analysis was performed in the ‘interlaced’ mode, where the sources can operate in a simultaneous and continuous fashion. A simsMVA Depth Profile mode tab contains a table on the right-hand side that allows the selection of which peak intensity profile to plot on the large set of axes on the left-hand side. The small axes on the right-hand side show the total spectrum of all levels. It is possible to select specific ranges of masses and levels to be processed. This is useful for example in the presence of artefacts on the first or last few levels or when there is implantation of ions of the sputter beam. The slider on the top left applies a moving average filter to the profiles of all ions and there are other functions such as normalisation and data pre-processing.

**Figure 5.4:** Screenshot of a Depth Profiles mode tab loaded with ToF-SIMS data of a layered metallic sample.
If the user clicks the "NMF" button, a small window pops up with a number of options for the factorisation. The available options are "algorithm", "number of components", "number of iterations", "number of repeats", "live visual output", "calculate lack-of-fit" and "use sparse matrices". Once the factorisation is done, similarly to PCA, a new "MVA results" tab is created. The tab will contain the components spectra on the top panel, the components intensities on the bottom right panel and the error per iteration on the bottom left panel. Figure 5.5 shows a screenshot of a Components Intensities panel for the metallic layered sample. NMF was performed using 6 components, a multiplicative update-based algorithm and 500 iterations. The "Overview" button pops up a window containing an overlay of the profiles of all components together with their characteristic spectra (Figure 5.6).

**Figure 5.5:** Screenshot of a Components intensities tab with NMF results of depth profile data of a layered metallic sample.
Figure 5.6: Screenshot of an overlay window with NMF results of depth profile data.

5.2.3 Images mode

Figure 5.8 shows a screenshot of an Images mode tab loaded with data of the taper-section of an organic coating presented earlier in this Chapter and described in detail in Chapter 12 of this thesis. The area covering the whole taper is larger than the raster range of the primary ion beam of the equipment, therefore in order to analyse the whole region, several patches must be acquired. The TOF.SIMS 5 offers an automated way of analysing large areas by rastering the stage, however, when the sample is formed of regions with very different chemistry and conductivity, such automated modes will only
be optimised for one end of the analysed area and the solution for getting good quality data is the separate acquisition of several 500 x 500 μm² patches. The images mode of simsMVA offers a tool for *stitching* several hyperspectral patches and transform them into one dataset, as shown in Figure 5.7. The total size of the dataset present in Figure 5.8 is 384 pixels x 256 pixels x 699 peaks.

**Figure 5.7: Dataset “stitching” process in simsMVA.**

An Images mode tab will have at the top left-hand side the intensity maps for a selected ion that can be normalised by the total ion intensity or any other ion map. The two plots on the right-hand side will show the intensity distribution of the peak list (after pre-processing) outside (top) and inside (bottom) of the region of interest determined by a red resizable polygon on the left-hand side map. The user can choose to process only the region within the polygon or perform subsampling of the data using low discrepancy which, as described in section 4.2.7, have been shown to generate, in much less time, results as good as if the whole dataset was processed [25, 29, 96]. The image at the bottom left will contain an overlay of different ion maps selected by the user.
Figure 5.8: Screenshot of an Images mode tab loaded with ToF-SIMS data of a cross-sectional taper of an organic coating.

Figure 5.9 shows a PCA results tab for the coating dataset. Prior to PCA, the data was normalised by total ion counts, Poisson scaled and mean centred. The Loadings panel looks the same for all previously described modes of analysis apart from the option to use different colour maps for the plots. A good practice is to match the colour maps for the loadings and scores plots. This is useful for PCA results where the Scores are shown using a diverging colour map. In both Loadings and Scores panels, positive values are scaled to red and negative to blue, while values closer to zero will be coloured black. The scatter plot on the Scores panel represent the scores of two chosen principal
components for all pixels. The “Brush” button enables the user to select pixels on the scatter plot and those will be highlighted on the intensity map on the left-hand side.

Figure 5.9: Screenshot of a MVA results tab with PCA results of a cross-sectional taper of an organic coating.

The PCA results of the taper section showed that there are two phases on the bulk of the organic coating, with a different composition on the top surface. It also showed that the metal substrate is covered with a thin chromium layer. If instead of PCA, NMF with 5 components is done, a useful feature is the “Overlay” button, that creates a window that
enables to overlay all components in a single map, according to a discrete colourmap, as shown in Figure 5.10. For this dataset specifically, the NMF results show clearly the top surface (red), polymer 1 (orange), polymer 2 (purple), chromium layer (green) and metal substrate (blue).

![Figure 5.10: Screenshot of a Components overview window.](image)

### 5.2.4 3D mode

Figure 5.11 shows a screenshot of a 3D mode tab loaded with data of an automotive grade polypropylene sample that was introduced in Chapter 4. More details of the samples and effects of plasma and flame treatment are discussed in Chapter 9 of this thesis. A simsMVA 3D mode tab will have a map on the top left side that will show the intensity maps for the selected ion at the selected level. Similar to an Images mode tab, the two plots on the right hand side will show the intensity distribution of the peak list.
(after pre-processing) outside (top) and inside (bottom) of the region of interest determined by a red polygon on the left hand side map. The axes on the bottom left will have a 3D map of intensities with a slice on the currently selected level.

Figure 5.11: Screenshot of a 3D mode tab loaded with ToF-SIMS data of an automotive grade polypropylene sample.

Figure 5.12 shows a screenshot of a 3D NMF Components Intensities panel for the polypropylene sample. NMF was performed subsampling 1% of the voxels, using 3 components, a multiplicative update-based algorithm and 500 iterations. The “Profile view” button collapses all pixels of the NMF components of each level to one data point per level and plots the results on the right-hand side axes.
The development of multivariate analysis methodologies for complex ToF-SIMS datasets: Applications to materials science

Figure 5.12: Screenshot of a Components Intensities panel for 3D ToF-SIMS data of an automotive grade polypropylene sample.

As described in Section 4.3.9, sometimes the correction of the z-coordinates of the voxels will aid visualisation and interpretation of 3D MVA results [93, 94]. When the user clicks the “3D view” or “3D overlay” buttons, they are prompted with an option to perform z-correction using the XZ or the YZ planes of a specific component. Figure 5.13 shows both “3D View” (for one component) and “3D overlay” (for all three components) of the NMF results of the polypropylene dataset after z-correction. There are three possible styles of 3D visualisation (“Scatter”, “Slices” or “Rendered Isosurfaces”) and parameters such as colour map, transparency and aspect ratio can be controlled.
Figure 5.13: 3D visualisation of the NMF components intensities of the polypropylene dataset. Left: Screenshot of a 3D view window. Right: Screenshot of a 3D overlay window.

5.2.5 Exporting data from the SurfaceLab software

The current version of simsMVA is optimised to import files generated and exported by the TOF.SIMS 5 spectrometer. Figures 5.14 to 5.16 are intended as a guideline on how to export files using SurfaceLab software provided by IONTOF alongside the instrument. A pre-requisite for generating new datasets from the same measurement is that it was recorded in an ITM raw data file format.

In SurfaceLab’s Spectra program, once the spectra of all samples are calibrated, the user can then apply a peak list to one of them and follow the steps shown in Figure
5.14: 1) On the left-hand side menu, select all samples desired to be on the peak list. 2) Click on “statistics” button, a window with the peak list will pop up. 3) On the top icons bar, unselect all extra statistics metrics. 4) Click “options” and change “description” to “Mass”. 5) Click on save button and save peak list as a .txt file.

Figure 5.14: Guideline of how to export a peak list from SurfaceLab’s Spectra program.

In SurfaceLab’s Profiles program, the user has to reconstruct the data from an .ITM file using a desired peak list and then follow the steps shown in Figure 5.15: 1) On the top menu bar, click “File->Export”, an options window will pop up. 2) Select “Data Point” and “Sputter Time” for X-Axis and one of the other options for Y-axis, press OK and save the profile data as a .txt file.
Figure 5.15: Guideline of how to export a peak list from SurfaceLab's Profiles program

In SurfaceLab’s Images program, the user also has to reconstruct the data from an .ITM file using a desired peak list and then follow the steps shown in Figure 5.16: 1) On the top menu bar, click “File->Export”, an options window will pop up. 2) On the left-hand side menu, select the desired mass range. 3) Select “Export Summed Image” for Images data or “Export Scan Resolved Images” for 3D data. 4) In “Exported Data Format” Choose “Binary (BIF6)” and press the “Export” button to create the BIF6 files that can then be loaded into simsMVA.
Figure 5.16: Guideline of how to export a peak list from SurfaceLab’s Profiles program

5.2.6 Algorithms and typical running times

For very large and sparse datasets, the app switches the memory allocation of the data matrix automatically to MATLAB sparse prior to MVA. For PCA and k-means, the statistics and machine learning toolbox functions are used (svd, svds and kmeans). For NMF, an in-house developed function that is able to handle sparse matrices is used. The use of sparse matrices has the advantage of saving memory but it increases the computational times for MVA. Figure 5.17 shows a plot of dataset size x computational
time for PCA, NMF and $k$-means clustering for the imaging ToF-SIMS dataset presented in Section 5.2.3. The size of the bubbles is proportional to the number of zero elements in the data matrix. The different dataset sizes were obtained by selecting different mass ranges (some fragments will have lower yield and therefore less counts per pixel) and different degrees of pixel binning. The data used has integrated peaks data so most of the zero elements of the original raw dataset are already excluded. When using raw datasets, depending on memory size, using sparse allocation may be the only viable option [29].

![Graph](image)

**Figure 5.17:** Computational time for PCA, NMF and $k$-means clustering for the imaging ToF-SIMS dataset presented in Section 5.2.3.

From the plots in Figure 5.17, it can be observed that PCA computational time is not influenced by matrix sparsity (function `pca` for non-sparse and `svds` for sparse data allocation) whereas the NMF computational time is quicker for sparse matrices when the total matrix size is less than $1 \times 10^5$, becoming slower for higher values.

For NMF, apart from computational time, an important parameter is the *error*, which is defined as:
Neighbour pixels in ToF SIMS data will be extremely correlated and therefore there must be a subsampling limit where the achieved result is as good as if one did the analysis using all pixels. Assuming that memory is not an issue, it is also important to take into account the gain in computational time and error for larger subsamples. A maximum value of the “quality parameter” will represent the ideal subsample size:

\[ \text{quality parameter} = (\text{error} \times \text{time})^{-1} \]

Fig. 5.18 shows the error and quality parameter for NMF (3 components) of the polymer blend dataset in function of subsample size and after different number of iterations.

**Figure 5.18:** Error and quality parameter for NMF of one patch of the polymer blend dataset. The different colours represent different numbers of iterations.

The results show that after around 3% subsample size, the results do not get much better as the subsample size and calculation time increase, resulting in a drop of the “quality
parameter”. Using 3% subsample sizes as standard can be very useful since it speeds up
the calculation time significantly.

5.2.7 System requirements
simsMVA is available both as an application inside MATLAB and as a standalone
version. The MATLAB version runs in any operational system and requires an
installation of MATLAB 2015b or newer with the Statistics and Machine Learning and
the Image Processing toolboxes. The standalone version runs on 64 bit Windows PCs
and does not need a MATLAB installation. For large area imaging or 3D datasets a
minimum of 8 GB of RAM is required.

5.2.8 External media
Some of the features of simsMVA are better shown in video form, therefore a website
and a YouTube channel have been created and videos are often published. The web
addresses for the page and for the YouTube channel are: http://www.mvatoools.com and
https://www.youtube.com/channel/UC5CGYrTPrJkJhGOxglMonA respectively.
PART II

VALIDATION

The next three Chapters contain studies of datasets that were used solely to systematically test the proposed MVA methodologies with no particular application of scientific or industrial interest on the samples analysed but conducted for the purpose of methodology validation. Such validation is important to be carried out in well-known systems and synthetic datasets before the application to real-life, industrial problems and although as a secondary purpose, the experiments of this part also yielded results of scientific importance.
6 PCA & NMF OF DATA OF STANDARD DICARBOXYLIC ACIDS

6.1 Introduction

This Chapter presents results of experiments that were carried out by the author of this thesis as part of a collaboration with visiting fellow Dr. Jose Ferreira from University of Sao Paulo (Brazil). The main research project deals with dicarboxylic acids-based pre-treatments of metallic substrates for corrosion protection. In order to achieve a better understanding of the interactions between acids and metal substrates, it was decided to acquire surface analysis data of a set of standard dicarboxylic acids. This resulted in publications of X-ray photoelectron spectroscopy (XPS) and ToF-SIMS data in the archive journal Surface Science Spectra [97]. Here, the ToF-SIMS results are presented and the dataset is used for validation of the PCA and NMF functions of simsMVA.
6.2 Linear saturated dicarboxylic acids

Linear saturated dicarboxylic acids are a class of organic chemical compounds with two carboxyl functional groups (-COOH) at the extremities of their aliphatic chains. This class of organic acids can be represented by the general molecular formula HOOC-(CH₂)ₙ-COOH. The most common values for n with their respective acid names are present in Table 6.1. The general chemical behaviour and reactivity of these compounds are similar to monocarboxylic acids, and they are all widely used in the production of copolymers, such as polyamides and polyesters [98–100]. The easy conversion of carboxyl groups to esters has industrial importance since many esters are used as taste and odour enhancers. Carboxylic acids are also used as catalysts, replacing ecologically unfavourable organic halides [101]. Over the last three decades, interest in such acids has increased, specifically regarding their application to improve the corrosion resistance of metallic substrates such as zinc, copper, iron and aluminium [102–107]. Research has also shown that carboxylic acids can be used as additives for the electro-synthesis of polymeric protective coatings. Such coatings promote passivation of different metallic substrates, allowing the oxidation of the carboxylic acid monomers without concomitant reactions [108–110]. More recently, carboxylic acids have been used to generate hydrophobic surfaces on various metallic substrates (Fe, Al, Cu, Mg, Zn, Ti, etc.) forming self-assembled layers by adsorption, via carboxyl groups, to the positively charged metal surfaces [111–115].
CHAPTER 6: PCA & NMF OF DATA OF STANDARD DICARBOXYLIC ACIDS

Table 6.1: The most common dicarboxylic acids with general formula HOOC-(CH2)n-COOH.

<table>
<thead>
<tr>
<th>n</th>
<th>Preferred IUPAC name</th>
<th>Common name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Ethanedioic acid</td>
<td>Oxalic acid</td>
</tr>
<tr>
<td>1</td>
<td>Propanedioic acid</td>
<td>Malonic acid</td>
</tr>
<tr>
<td>2</td>
<td>Butanedioic acid</td>
<td>Succinic acid</td>
</tr>
<tr>
<td>3</td>
<td>Pentanedioic acid</td>
<td>Glutaric acid</td>
</tr>
<tr>
<td>4</td>
<td>Hexanedioic acid</td>
<td>Adipic acid</td>
</tr>
<tr>
<td>5</td>
<td>Heptanedioic acid</td>
<td>Pimelic acid</td>
</tr>
<tr>
<td>6</td>
<td>Octanedioic acid</td>
<td>Suberic acid</td>
</tr>
<tr>
<td>7</td>
<td>Nonanedioic acid</td>
<td>Azelaic acid</td>
</tr>
<tr>
<td>8</td>
<td>Decanedioic acid</td>
<td>Sebacic acid</td>
</tr>
<tr>
<td>9</td>
<td>Undecanedioic acid</td>
<td>Basilic acid</td>
</tr>
<tr>
<td>10</td>
<td>Dodecanedioic acid</td>
<td>Dodecanedioic acid</td>
</tr>
<tr>
<td>14</td>
<td>Hexadecanedioic acid</td>
<td>Thapsic acid</td>
</tr>
</tbody>
</table>

6.3 Experimental conditions

Seven different dicarboxylic acids were analysed: ethanedioic (oxalic), propanedioic (malonic), butanedioic (succinic), pentanedioic (glutaric), hexanedioic (adipic), heptanedioic (pimelic) and octanedioic (suberic), which are representative of acids with both odd and even numbers of carbon atoms (C2-C8 chains). The samples were acquired in pure, homogeneous powder form from Sigma-Aldrich. Each powder was sprinkled into a small hole (1 mm diameter) of a metal stub and compacted using a clean spatula. No adhesive was used. Immediately after mounting, each sample was moved to the TOF.SIMS 5 loadlock chamber and kept there for 12 hours in order to achieve optimal working pressure and check for stability. Figure 6.1 shows, for the succinic acid sample,
the pressure in both loadlock chamber before analysis and main chamber during SIMS analysis in function of time. All acids presented similar behaviour with the exception being oxalic acid that, due to its extreme instability, was frozen at temperatures of around -5 °C using a copper cold finger in a special sample holder that is capable of monitoring the temperature of the sample. The finger was frozen prior to pumping the chamber in order to minimise water adsorption onto the surface of the sample. The sudden drops at around 03:26 and 08:37 occurred as a consequence of the automated activation of a backing pump.

Figure 6.1: Monitored pressure (for succinic acid sample) in loadlock chamber before analysis and main chamber during SIMS analysis.

The ToF-SIMS analyses were carried out using a 25 keV Bi$_3^+$ beam operated in the high current bunched mode delivering 0.2 pA and rastered over regions of 50 x 50 μm$^2$. Such
choice of smaller raster areas was made because of the high surface roughness and consequent difficulty of finding suitable regions that would yield good quality secondary ion spectra.

### 6.4 ToF-SIMS results

Figure 6.2 shows the general structure of all six acids and the fragmentation pattern of the main observed peaks in their ToF-SIMS spectra. The fragmentation mechanism follows the one described by Spool [116]. The ions ([M+H]+ and [M-OH]+ for positives and [M-H]- and [M-COOH]- for negatives) also occur for the respective oligomers of all acids (2M, 3M, 4M, 5M, etc.).

![Diagram](image)

**(A)**

**Figure 6.2: General structure of all seven measured acids and the fragmentation pattern of the main observed positive (A) and negative (B) peaks in their ToF-SIMS spectra.**
Figures 6.3 and 6.4 show a comparison overview of the positive and negative spectra obtained for all acids. The results show agreement to the fragmentation pattern presented in Figure 6.2. More details of the positive and negative spectra of each acid will be discussed in the following Sections.

Figure 6.3: Comparison of the positive spectra obtained for all acids.
Figure 6.4: Comparison of the negative spectra obtained for all acids.

The lower mass range (0 to 70 u) of the positive ion spectra is also very characteristic of these acids. The larger the molecule is, the more peaks are observed (due to further fragmentation of the carbon chains) and the more the spectra look like a typical linear saturated hydrocarbon polymer spectrum. Figure 6.5 shows a comparison of all the positive spectra in the 0 to 70 u range together with a polypropylene spectrum.

The acids are prone to undergo degradation when in vacuum or exposed to X-rays [19-25] with their stability increasing with the increase of chain length. This means that the longer the carbon chain is, the more difficult it is to lose a carboxyl. Figure 6.6 shows the ratios of the negative ions [M-COOH]$^-$ / [M-H]$^-$ for all acids analysed. It can
be seen that the ratios decrease exponentially with the increase of the chain length, in agreement with their stability. The value for oxalic acid falls off the trend, having a lower value than its successor. A reason for this may be related to the fact that the sample was analysed in a frozen state.

![Figure 6.5: Comparison overview of the positive spectra for all measured acids together with a standard polypropylene spectrum in the 0 to 70 u range.](image-url)
6.5 MVA

As illustrated in the previous section, the spectra of these dicarboxylic acids are relatively simple and the peaks that distinguish them from each other are easy to identify. Thus, they make up a good dataset for testing the basic MVA routines of simsMVA. After calibration using characteristic peaks, a peak-list containing the main positively charged ions was created (Table 6.2) and applied to a series of repeat measurements to create a data matrix. The peak areas were normalised by total ion counts and Poisson scaled and both PCA and NMF were applied to the resulting dataset.
Table 6.2: Positive and negative ions peak list used for MVA of ToF-SIMS data of dicarboxylic acids.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Mass</th>
<th>Structure</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHO⁺</td>
<td>29.001</td>
<td>C₂HO⁻</td>
<td>41.010</td>
</tr>
<tr>
<td>C₃H₅⁺</td>
<td>41.04</td>
<td>CHO₂⁻</td>
<td>45.007</td>
</tr>
<tr>
<td>C₂H₅O⁺</td>
<td>43.02</td>
<td>C₂H₃O₂⁻</td>
<td>59.029</td>
</tr>
<tr>
<td>C₃H₇⁺</td>
<td>43.057</td>
<td>C₃H₃O₂⁻</td>
<td>71.033</td>
</tr>
<tr>
<td>CHO₂⁺</td>
<td>45.001</td>
<td>C₂H₃O₂⁻</td>
<td>73.012</td>
</tr>
<tr>
<td>C₃H₃O⁺</td>
<td>55.022</td>
<td>C₄H₇O₂⁻</td>
<td>87.033</td>
</tr>
<tr>
<td>C₄H₇⁺</td>
<td>55.059</td>
<td>C₂HO₄⁻</td>
<td>89.023</td>
</tr>
<tr>
<td>C₅H₉⁺</td>
<td>69.08</td>
<td>C₅H₅O₂⁻</td>
<td>101.009</td>
</tr>
<tr>
<td>C₃H₅O₃⁺</td>
<td>87.016</td>
<td>C₅H₃O₄⁻</td>
<td>103.025</td>
</tr>
<tr>
<td>C₂H₅O₄⁺</td>
<td>91.029</td>
<td>C₅H₁₁O₂⁻</td>
<td>115.031</td>
</tr>
<tr>
<td>C₄H₅O₃⁺</td>
<td>101.033</td>
<td>C₄H₅O₄⁻</td>
<td>117.008</td>
</tr>
<tr>
<td>C₃H₅O₄⁺</td>
<td>105.036</td>
<td>C₅H₁₃O₂⁻</td>
<td>129.023</td>
</tr>
<tr>
<td>C₄H₃O₅⁺</td>
<td>115.055</td>
<td>C₅H₇O₄⁻</td>
<td>131.038</td>
</tr>
<tr>
<td>C₄H₅O₄⁺</td>
<td>119.029</td>
<td>C₆H₉O₄⁻</td>
<td>145.030</td>
</tr>
<tr>
<td>C₆H₉O₃⁺</td>
<td>129.066</td>
<td>C₇H₁₁O₄⁻</td>
<td>159.054</td>
</tr>
<tr>
<td>C₅H₉O₄⁺</td>
<td>133.067</td>
<td>C₆H₁₃O₄⁻</td>
<td>173.081</td>
</tr>
<tr>
<td>C₇H₁₁O₃⁺</td>
<td>143.08</td>
<td>C₆H₇O₄⁻</td>
<td>179.048</td>
</tr>
<tr>
<td>C₆H₁₁O₄⁺</td>
<td>147.058</td>
<td>C₆H₇O₅⁻</td>
<td>207.043</td>
</tr>
<tr>
<td>C₈H₁₃O₃⁺</td>
<td>157.094</td>
<td>C₈H₁₁O₈⁻</td>
<td>235.065</td>
</tr>
<tr>
<td>C₇H₁₃O₄⁺</td>
<td>161.089</td>
<td>C₁₀H₁₅O₈⁻</td>
<td>263.088</td>
</tr>
<tr>
<td>C₈H₁₅O₄⁺</td>
<td>175.104</td>
<td>C₈H₁₀O₁₂⁻</td>
<td>269.068</td>
</tr>
<tr>
<td>C₄H₅O₅⁺</td>
<td>181.041</td>
<td>C₁₂H₁₉O₈⁻</td>
<td>291.138</td>
</tr>
<tr>
<td>C₆H₉O₈⁺</td>
<td>209.027</td>
<td>C₉H₁₁O₁₂⁻</td>
<td>311.094</td>
</tr>
<tr>
<td>C₈H₁₃O₈⁺</td>
<td>237.064</td>
<td>C₁₄H₂₃O₈⁻</td>
<td>319.151</td>
</tr>
<tr>
<td>C₁₀H₁₇O₈⁺</td>
<td>265.124</td>
<td>C₁₆H₂₇O₈⁻</td>
<td>347.156</td>
</tr>
<tr>
<td>C₆H₇O₁₂⁺</td>
<td>271.06</td>
<td>C₁₂H₁₇O₁₂⁻</td>
<td>353.127</td>
</tr>
<tr>
<td>C₁₂H₂₁O₈⁺</td>
<td>293.136</td>
<td>C₈H₁₇O₁₆⁻</td>
<td>359.093</td>
</tr>
<tr>
<td>C₉H₁₃O₁₂⁺</td>
<td>313.075</td>
<td>C₁₅H₂₃O₁₂⁻</td>
<td>395.128</td>
</tr>
</tbody>
</table>


<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{C}<em>{14}\text{H}</em>{25}\text{O}_{8}^+$</td>
<td>321.153</td>
<td>$\text{C}<em>{12}\text{H}</em>{15}\text{O}_{16}^-$</td>
<td>415.114</td>
</tr>
<tr>
<td>$\text{C}<em>{16}\text{H}</em>{29}\text{O}_{8}^+$</td>
<td>349.178</td>
<td>$\text{C}<em>{18}\text{H}</em>{29}\text{O}_{12}^-$</td>
<td>437.181</td>
</tr>
<tr>
<td>$\text{C}<em>{12}\text{H}</em>{19}\text{O}_{12}^+$</td>
<td>355.086</td>
<td>$\text{C}<em>{10}\text{H}</em>{9}\text{O}_{2}^0$</td>
<td>449.111</td>
</tr>
<tr>
<td>$\text{C}<em>{15}\text{H}</em>{25}\text{O}_{12}^+$</td>
<td>397.183</td>
<td>$\text{C}<em>{16}\text{H}</em>{23}\text{O}_{16}^-$</td>
<td>471.142</td>
</tr>
<tr>
<td>$\text{C}<em>{12}\text{H}</em>{17}\text{O}_{16}^+$</td>
<td>417.072</td>
<td>$\text{C}<em>{21}\text{H}</em>{35}\text{O}_{12}^-$</td>
<td>479.178</td>
</tr>
<tr>
<td>$\text{C}<em>{18}\text{H}</em>{31}\text{O}_{12}^+$</td>
<td>439.213</td>
<td>$\text{C}<em>{15}\text{H}</em>{19}\text{O}_{2}^0$</td>
<td>519.188</td>
</tr>
<tr>
<td>$\text{C}<em>{10}\text{H}</em>{11}\text{O}_{2}^0$</td>
<td>451.099</td>
<td>$\text{C}<em>{24}\text{H}</em>{41}\text{O}_{12}^-$</td>
<td>521.151</td>
</tr>
<tr>
<td>$\text{C}<em>{16}\text{H}</em>{25}\text{O}_{16}^+$</td>
<td>473.101</td>
<td>$\text{C}<em>{20}\text{H}</em>{31}\text{O}_{16}^-$</td>
<td>527.173</td>
</tr>
<tr>
<td>$\text{C}<em>{21}\text{H}</em>{37}\text{O}_{12}^+$</td>
<td>481.194</td>
<td>$\text{C}<em>{12}\text{H}</em>{11}\text{O}_{24}^-$</td>
<td>539.126</td>
</tr>
<tr>
<td>$\text{C}<em>{15}\text{H}</em>{21}\text{O}_{20}^+$</td>
<td>521.068</td>
<td>$\text{C}<em>{20}\text{H}</em>{25}\text{O}_{12}^-$</td>
<td>589.178</td>
</tr>
<tr>
<td>$\text{C}<em>{24}\text{H}</em>{43}\text{O}_{12}^+$</td>
<td>523.236</td>
<td>$\text{C}<em>{18}\text{H}</em>{23}\text{O}_{24}^-$</td>
<td>623.077</td>
</tr>
<tr>
<td>$\text{C}<em>{20}\text{H}</em>{33}\text{O}_{16}^+$</td>
<td>529.208</td>
<td>$\text{C}<em>{32}\text{H}</em>{55}\text{O}_{16}^-$</td>
<td>695.410</td>
</tr>
<tr>
<td>$\text{C}<em>{12}\text{H}</em>{12}\text{O}_{24}^+$</td>
<td>541.093</td>
<td>$\text{C}<em>{24}\text{H}</em>{35}\text{O}_{24}^-$</td>
<td>707.282</td>
</tr>
<tr>
<td>$\text{C}<em>{24}\text{H}</em>{41}\text{O}_{16}^+$</td>
<td>585.255</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{C}<em>{20}\text{H}</em>{31}\text{O}_{20}^+$</td>
<td>591.123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{C}<em>{18}\text{H}</em>{25}\text{O}_{24}^+$</td>
<td>625.092</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{C}<em>{28}\text{H}</em>{49}\text{O}_{16}^+$</td>
<td>641.255</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{C}<em>{25}\text{H}</em>{41}\text{O}_{20}^+$</td>
<td>661.286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{C}<em>{24}\text{H}</em>{37}\text{O}_{24}^+$</td>
<td>709.166</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 6.5.1 PCA results

Figure 6.7 shows the variance captured per principal component for the positive and negative peak lists. In both cases, 5 components are enough to explain more than 95% of the dataset variance.
Figures 6.8 and 6.9 show the scores and loadings of the first six principal components obtained for the acids dataset. The first principal component (PC 1) seems to be related to the carbon chain length of the acids structure as the scores gradually change from positive to negative with increase of chain length. The loadings of PC 1 reflect what was observed in the spectra: acids with longer chains will generate more hydrocarbon ion fragments which are typical of linear chain polymers. The loadings of subsequent components will have different relative intensities for the various characteristic peaks of the acids and a combination of their scores is enough to separate them all as it is shown in the two biplots presented in Figure 6.10. As expected from the variance analysis presented in Figure 6.7, very similar results are obtained if the negatives peak list is used.
Figure 6.8: Scores (left) and loadings (right) for principal components (PC) 1 to 3 of the acids dataset.
The Development of Multivariate Analysis Methodologies for Complex ToF-SIMS Datasets: Applications to Materials Science

Figure 6.9: Scores (left) and loadings (right) for principal components (PC) 4 to 6 of the acids dataset.
6.5.2 NMF results

NMF was done using 7 components and a multiplicative update algorithm. Convergence was achieved after 500 iterations. Figure 6.11 and 6.12 show web plots (generated as described in Chapter 5) of the intensities of the 7 NMF components and their respective characteristic positive and negative spectra.
Figure 6.11: Relative intensities of the 7 NMF components obtained using a positive peak list and their respective characteristic spectra obtained for the acids dataset.
Figure 6.12: Relative intensities of the 7 NMF components obtained using the negative peak list and their respective characteristic spectra obtained for the acids dataset.

Each component was exclusive of one acid with minimum overlapping of their intensities. This consequently reflected on their characteristic spectra that contained the respective peaks of their attributed acids. This is obviously an expected result but nevertheless it shows that NMF can indeed identify pure components in a ToF-SIMS dataset.
6.5.3 Low mass range

As discussed in Section 6.4, the low mass range of the positive ion spectra are very characteristic of the acids but their differences are less obvious with most peaks appearing for all samples. In order to check whether MVA would pick up such differences, a survey peak list containing all positive polarity spectra peaks with masses from 10 to 70 u was created and both PCA and NMF was applied to the resulting dataset. Figure 6.13 shows the web plot generated from the NMF results using the same number of components as before. The components intensities are not exclusive for each acid but nevertheless each one will have its own “fingerprint” and this is shown in their different patterns in their web plots.
The PCA results were also enough to achieve the separation presented in Figure 6.10, however, in contrast to the results of the dataset using all characteristic peaks, the second principal component (PC 2) scores showed the same trend presented in Figure 6.6: an initial value for oxalic acid, that raises for malonic and decreases with the increase of acids carbon chain length. This may mean that the multivariate structure present in the loadings of PC 2 is related to relative intensities of peaks that are representative of the stability of the acids. The scores and loadings of PC 2 are shown in Figure 6.14.
6.6 Final remarks

The ToF-SIMS dataset created with standard dicarboxylic acid powders allowed the validation of the MVA routines of simsMVA in a practical, experimental case.
Moreover, without any bias, the results confirmed the interpretation that was made prior to MVA.

Apart from being an ideal set of data for testing MVA routines, these spectra are of great value for a database of standard materials and the publications in the Surface Science Spectra journal are the only available data for these materials. Even though the spectra are relatively simple, this was a challenging experiment to be carried out given the difficulties in sample preparation and acquisition of good quality spectra. One should also note that the XPS spectra of the same acids were acquired and published in the same journal [97] following an even more careful experimental methodology since the materials have shown to undergo significant X-ray degradation.
7 NMF OF UNCOMPRESSED DATA

7.1 Introduction

In Chapter 5, it was briefly mentioned that the variables of a dataset can be all spectral channels instead of peak areas. This Chapter aims to explore further on that topic and presents a new methodology for performing NMF of uncompressed data.

7.2 Trends in the analysis of uncompressed SIMS data

There has recently been a new trend within the SIMS community towards multivariate analysis of raw, unbinned datasets, as opposed to the well-known practice of peak picking, which can be biased by the eye of the analyst or by automated peak search routines. Cumpson et. al. [28] have recently proposed a method for computing the singular value decomposition (SVD) of large datasets using a random vectors algorithm,
enabling PCA of such huge datasets to be done in hours using conventional computers or even minutes using GPU arrays [27, 28]. Furthermore, high-resolution mass spectrometry raw data matrices are extremely sparse with typically less than 0.1% non-zero elements and methods that take advantage of such sparsity can speed up the processing even more. Those methods include low discrepancy subsampling [25, 27], pre-processing data compression [117, 118] and wavelet-based compression [30]. PCA is a powerful tool but its outcome may be difficult to interpret, especially when it comes to hyperspectral images. Non-negative matrix factorisation (NMF) provides results that can make more sense in terms of surface chemistry, since the non-negative factors can be directly related to mass spectra and also to distribution maps [119]. No attempt has been reported previously to this work on the non-negative matrix factorisation of large uncompressed ToF-SIMS datasets. The next Sections intend to describe and explore approaches on how to perform NMF on ToF-SIMS data retaining both spatial and spectral resolution. Pre-processing challenges are also discussed and a typical experimental result is used as an example.

7.3 Experimental

A large area (1cm²) ToF-SIMS dataset of a chemically contaminated fingerprint on a clean silicon wafer was chosen to illustrate NMF analysis of large datasets. Extensive work has been done at Surrey on the application of mass spectrometry to fingerprint characterisation over different substrates [120–122]. The most recent developments seek to identify metabolized drugs present in latent fingerprints and since the concentration
of those substances can be very low, multivariate methods using raw datasets can aid
the identification of hidden spectral and image features. ToF-SIMS analysis of
negatively charged secondary ions was carried out using the TOF.SIMS 5 system at
University of Surrey using a 25 keV Bi$_3^+$ ion beam operated in the high current bunched
mode delivering 0.3 pA, pulsed at 10 kHz and images were acquired using the stage
raster mode. The total 1 x 1 cm$^2$ area contains 2000 x 2000 pixels which comprise of 20
x 20 patches of 500 x 500 µm$^2$ areas and sized 100 x 100 pixels each. The whole area
was scanned 10 times. When the primary ions beam is bunched, the spectral resolution
is maximum (m/Δm ~ 9,000 at $^{29}$Si) but there is an increase in beam diameter (typically
1-2 µm for Bi$_3^+$ at 25 keV) and thus a decrease in spatial resolution, resulting in
hyperspectral images with optimal pixel density of approximately 0.06 pixels/µm$^2$.
Figure 7.1 shows a total ion intensity map of the entire analysed region. The fingerprint
donor was a volunteer who provided it recognising that the data would not be stored in
a fingerprint database or used for identification purposes.

Figure 7.1: Total ion intensity map and total pixels spectrum of a chemically
contaminated fingerprint.
### 7.4 Loading raw data

The IONTOF SurfaceLab® software exports the data into a general raw data (.GRD) binary file format. The file contains information of every secondary ion detected (scan, x-coordinate, y-coordinate and detected time-of-flight). The best way found to deal with such large raw datasets was to convert it directly into pre-allocated sparse arrays in MATLAB. The resulting data is arranged in a sparse matrix sized \( N \times M \) containing the \( N \) spectra of every single pixel, with \( M \) spectral channels each. The procedure adopted also allows the programme to perform pixel and spectral channels binning on the go.

The completely unbinned dataset discussed in this Chapter has 1,300,000 spectral channels within a mass range from 9 to 500 u.

As described in Chapter 2, in the TOF.SIMS 5 equipment, the effective primary ions pulse width depends on the impact point within the raster region. As a consequence of this, ions generated at regions further from the centre of the analysed area will have longer detected time-of-flight, resulting in slightly misaligned spectra for different pixels. The SurfaceLab software contains a built-in function that corrects for this effect. Secondary ions generated at different heights will also suffer delays in their detected time-of-flight. The correction for secondary ions time-of-flight delays and consequently the alignment of peaks are essential for multivariate analysis of unbinned data, since each spectral channel is regarded as a variable.
7.5 NMF algorithms

The algorithm used to achieve non-negative factorisation of all datasets presented in this Chapter was the one with multiplicative update rules (MU), introduced in Chapter 3.

The computer used was a conventional Dell Optiplex 9020 PC with a core i5 processor and 32 GB of RAM. Even with such high amount of memory, loading the full data was only possible using sparse allocation (the data matrices analysed would have between 0.1% and 0.5% of non-zero elements) and even so it was not possible to perform the required multiplications presented in the algorithm using conventional matrix multiplication in MATLAB. The approach chosen to overcome this problem was to use a training set by means of pixel subsampling using low discrepancy Sobol sequences, as suggested by Cumpson et. al [27] and illustrated in Chapter 5 of this thesis.

As initial conditions, matrix $H$ was always started with spectra of random samples of the original dataset and the matrix $W$ started with random numbers. The number of factors was chosen as three mainly to illustrate the issues with spectral channel alignment discussed in next session. The original data is sized 4,000,000 pixels x 1,300,000 channels and a subset of 15,000 pixels was used. The time per iteration was 36 seconds, making a total time of 5 hours after 500 iterations. Because of memory limitations it was not possible to reconstruct the original matrix and calculate the error of the factorisation. However, one way to estimate whether the subsampled percentage (15,000 pixels ~ 0.35%) was good enough is by using a reduced range of masses (rather than a binned dataset in order to maintain the same sparsity of the full dataset) and compare $H$ matrices calculated using all pixels to matrices calculated using different
sizes subsamples of the presented results against matrices calculated using smaller subset sizes with a difference defined by:

$$D (s \%) = \sqrt{\sum_{i} \sum_{j} (H_{ij}^{100 \%} - H_{ij}^{s \%})^2}$$

Figure 7.2 shows the difference calculated for several different subsets after 100 iterations and the time it took to achieve the solution. It can be seen that from 0.0375\% subset size, the results for matrix $H$ do not change much, however, the time per iteration increases significantly.

Figure 7.2: Difference $D$ (black) of matrix $H$ to 100 \% subsampling calculated for several different subsets and time/100 iterations (blue).
7.6 NMF results

Figure 7.3 presents NMF results of the fingerprint dataset factorised with three components. Matrices $H$ are plotted as they output from NMF and matrices $W$ are normalised per pixel to make the red-green-blue overlay easier to interpret.

![Figure 7.3: NMF results with three components for the fingerprint dataset. Field of view: 1cm x 1cm.](image)

The NMF results separated the substrate from the fingerprint, as shown in the intensity maps (matrix $W$) on the left hand side of Figure 7.3. The third component contains all
peaks of the silicon oxide substrate (OH, Si, and so forth). The first and second components spectra look very similar at first sight and contain the characteristic peaks (SO$_2^-$, SO$_2^-$, CH$_3$(CH$_2$)$_{11}$SO$_4^-$, etc) of sodium dodecyl sulphate (CH$_3$(CH$_2$)$_{11}$SO$_4^-$Na), which is a surfactant present on the personal care product used to chemically contaminate the deposited fingerprint.

In spite of the fact that the dataset has very few counts per pixel, NMF was successfully achieved. This shows the significant advantage of performing multivariate analysis of noisy, very large datasets. Figure 7.4 shows an unbinned spectrum of a single pixel. A pixel by pixel assessment of the spectral view will not contain relevant information but the whole data would still have latent structure and be able to undergo factorisation without binning.

![Figure 7.4: Spectrum of one pixel of the uncompressed dataset. There is not enough statistics to infer any chemical information.](image)

Since negative secondary ions analysed were considered, the silicon and silicon oxide peaks have very low intensity and even so NMF managed to separate them perfectly.
from the fingerprint signal, as shown in Figure 7.5. This reinforces the advantage of using unbinned datasets when it comes to finding hidden features.

![Graph showing spectral view of Si- and SiO-](image)

**Figure 7.5: High resolution spectral view of Si- and SiO- for all components of matrix H. The colour of each component is the same as for Figure 7.3**

### 7.6.1 Spectra alignment issues

As mentioned before, components 1 and 2 contain all characteristic fragments of the chemically contaminated fingerprint and their spectra look very similar. Closer inspection of both the spectra and their distribution show that there is a systematic misalignment of all peaks in components 1 and 2, with always the first component distributed on the upper part of a patch and the second on the lower part. Figure 7.6 shows a zoomed in view of the NMF results presented in Figure 7.3. The spectra of all three components are shown at nominal mass 80 u (SO$_3^-$) and all components distribution in six patches of analysis (sized 500 x 500 um$^2$ each).
The fact that NMF finds two misaligned components means that the original dataset had misaligned spectra at different pixels and this is what matrix \( W \) is capturing. Components 1 and 2 are chemically the same and only differ in terms of their calibration. The reason for this observed misalignment is uncertain. It can either be due to a non-perfect primary ion time-of-flight correction (as discussed in Chapter 2) or to a roughness artefact. In any case, in order to correct this and enable chemical, rather than...
physical/instrumental, characterisation, the solution would be to perform the alignment of all 4 million (or at least all subsampled) spectra before NMF. This huge number of spectra to be aligned makes manual calibration not feasible and therefore the alignment methods sought were the ones that do not require pre-assignment of peaks. Within those, the most popular alignment algorithms are based on correlation matrixes [123], but the alignment of a single spectrum (with no compression) can take hours. An alternative algorithm proposed by Wong et. al [124, 125] based on recursive Fourier transform correlation (RAFFT) is shown to be as effective, taking several minutes per spectrum. The alignment can be parallelized in multi core machines but it can still take a significant amount of time. Moreover, even if the alignment could be made in a reasonable time, there is another limiting factor which arises when one attempts to analyse unbinned spectra of single pixels from typical ToF-SIMS datasets with very low statistics. As shown in Figure 7.4, the unbinned spectrum of a single pixel will have, in total, less than 20 counts and since the mentioned alignment algorithms depend purely on the data and not on instrumental or calibration parameters, aligning peaks which are not very well defined becomes impossible. If a full dataset is factorised using NMF, the alignment can be done within the components of Matrix $H$ and then transformed back to the original dataset, as shown in Figure 7.7:

**Figure 7.7**: Scheme for peak alignment through reconstruction of factorised data.
To understand the effect of the proposed correction, a synthetic dataset was used. Figure 7.8 shows NMF results with 3 components before and after alignment of a dataset with two Gaussian peaks mixed with different proportions over 2000 x 2000 pixels (The 3D structure is to illustrate the potential misinterpretation due to alignment issues: the green region, instead of having a different “chemical” signal, has in fact the same constitution as the red region.). In this case, due to the simplicity of the dataset, the alignment was done by shifting the elements of matrix $H$ by the correct amount. Originally, the region shown as green is found to contain a third component which is in fact a misalignment artefact. After alignment, the three components represent only two regions, with the green and red being mixed equally. Thus, the approach for alignment via data reconstruction succeeded for a synthetic dataset.

**Figure 7.8:** NMF results with 3 components before and after alignment of a synthetic dataset. The left-hand side plots show the three factorised components (Matrix $H$). The right-hand side images show the intensity maps of the factorised components (Matrix $W$).
In order to test the NMF-alignment method on real data, the fingerprint dataset was compressed down to 200 x 200 pixels, 200,000 channels and only a small number of mass channels was used, resulting in a dataset sized 40,000 x 170. Figure 7.9 shows NMF results (after 100 iterations) with 2 and 3 components for this subset before and after alignment using both NMF-alignment (alignment of matrix $H$) and conventional RAFFT (analysis of spectra for all pixels). Matrix $H$ plot is only shown at nominal mass 80 u. A first observation is that a factorisation using two components is enough to separate the fingerprint from the substrate even for the non-aligned data. However, in more complex datasets or even in fingerprint datasets where the chemicals are not homogeneously distributed, a third component with extra information might be suppressed by the alignment issue. Differently from the simulated dataset, factorising the non-aligned data into three components and performing the alignment was not enough to remove the gradient observed in Figures 7.3 and 7.6. A factorisation and alignment of 8 components was enough to create the results presented in the third column of Figure 7.9. This is still more advantageous than having to align the spectra of 40,000 pixels, and in situations where there is not enough statistics per pixel, NMF-based alignment is the only option. The RAFFT-Aligned (fourth column of Figure 7.9) data also did not show the gradients in matrix $W$ but matrix $H$ presented an artefact on the left-hand side of the peak at 80 u.
Figure 7.9: NMF results (after 100 iterations) with 2 and 3 components a subset sized 40,000 x 170 before and after alignment using both NMF-alignment (alignment of matrix H) and conventional RAFFT (Analysis of spectra for all pixels). Matrix H plots only show nominal mass 80 u.

7.7 MapReduce as an alternative

The proposed alignment procedure worked well with smaller datasets but it is not feasible for large datasets that need subsampling, since the reconstructed product $WH^T$ will not fit in the memory. Furthermore, the datasets presented here are relatively small if compared to 3D ToF-SIMS depth profiles, especially ones generated by state-of-art instruments such as the new orbi-SIMS spectrometer [126]. Compression pre-processing
steps have been shown to be of great value and the development of more sophisticated
compression routines may be the way to go. However, at the current stage, there will
always be a compromise with spatial or spectral resolution that specific applications
could not afford. For this reason, alternative methods to process non-compressed full
mass spectrometry datasets must be developed. Cumpson et. al [28] have presented an
algorithm to perform singular value decomposition of huge matrices which reads data
from the hard disk and can be done in conventional computers. For the non-negative
factorisation, here it is proposed the use of the MapReduce framework. MapReduce is a
programming model introduced in 2004 by Dean and Ghemawat [127] and it is still
intensively used in several big data applications, where the matrix sizes are of several
orders of magnitude larger than the ones presented in the current work. It is meant to be
carried out by parallel and distributed computing. Every MapReduce program will have
a Map function and a Reduce function. The Map function reads segments of data and
sort them in key-value pairs (or stuples). The data is then shuffled so that every worker
node receives values assigned to the same key and performs the Reduce function to all
values.

There are a few algorithms for NMF and MapReduce in the literature [128–130]
and it was decided to implement in MATLAB the one presented by Liu et. al [130]. The
algorithm will contain 5 Map/Reduce steps to update each of matrices \( W \) and \( H \) and its
biggest advantages are that it partitions the data along the long dimension of the matrices
(see Figure 7.10), taking full advantage of the data matrix high sparsity and enabling it
to be applied in distributed systems. A General scheme on how the matrix \( H \) is updated
using MapReduce and Lee & Seung’s algorithm is shown in Figure 7.10, where key-
value pairs are represented in the form <key, value>. The partitioned lines of \( W \) and the partitioned columns of \( H \) are represented respectively as \( w_i \) and \( h_j \). The multiplicative update rule for matrix \( H \) can be rewritten as \( H = H \times (X/Y) \). The first two sets of MapReduce operations are used to calculate matrix \( X = W^T A \). Key-value pairs of \(<j, A_{ij}, w_i^T>\) are obtained only for non-zero \( A_{ij} \) elements of matrix \( A \). Those pairs are then added together for all values of \( i \), resulting in matrix \( X \) partitioned along its short dimension with key-value pairs of \(<j, x_j>\). The third and fourth sets of MapReduce operations are used to calculate matrix \( Y = W^T WH \). First, the square matrix \( C = W^T W \) is calculated by adding together the products \( w_i^T w_i \) for every \( i \) and mapping it to a common key in the form \(<0, C>\). Secondly, every partition \( h_j \) is multiplied by \( C \), resulting in matrix \( Y \) also partitioned along its short dimension with key-value pairs of \(<j, y_j>\). The fifth MapReduce step finally updates matrix \( H \) by multiplying every partition \( h_j \) by the division of their respective \( x_j/y_j \). More details of the algorithm can be found in the literature [130].
\[ H \leftarrow H.* \frac{X}{Y} = H.* \frac{W^T A}{W^T W H} \]

Figure 7.10: MapReduce Scheme for updating H.

Mathworks added MapReduce to MATLAB only in version 2014b and thus it still lacks example applications, making the learning process stiffer. Our implementations have so far only been tested on a single machine (parallelized over its processors) and the time per iteration using the most recent implementation is 10 times faster than the first one, which shows that there is still room for improvement. Figure 7.11 shows NMF (2 components) results of a compressed dataset in a limited mass span (matrix A sized 2,500 x 1,200) of the contaminated fingerprint using MapReduce NMF and standard
NMF. Both approaches were able to identify the fingerprint features and the resulting matrices $H$ show the same chemical information.

![Map/Reduce NMF vs NMF comparison](image)

**Figure 7.11:** Comparison of NMF (2 components) between MapReduce NMF and standard NMF. The dataset is compressed and a limited mass span was used (matrix $A$ sized 2,500 x 1,200).

A few tests were done with different subsets with different sparsity. High sparsity datasets were obtained by selecting different mass ranges of the uncompressed fingerprint data. The low sparsity datasets also represent different mass ranges but of a pre-compressed fingerprint dataset (both in terms of pixels and spectral channels).
Therefore, in this case, high sparsity data will contain a much shorter mass range than a low sparsity dataset with the same number of elements. Nevertheless, the plot in Figure 7.12 shows that the MapReduce algorithm implemented does indeed take advantage of data sparsity, with a sparse dataset with approximately $10^{11}$ elements having the same time per iteration of a non-sparse dataset 1000 times smaller. The circle diameters in Figure 7.12 are proportional to the number of zero elements in the data matrix.

![Figure 7.12: Time per iteration of MapReduce NMF of different subsets of the fingerprint dataset. The low sparsity datasets have a higher degree of compression than the high sparsity ones. The circle diameters are proportional to the number of zero elements in the data matrix.](image)

All MapReduce results presented were obtained on a single machine with 4 workers. In order to make full use of the advantages of the methodology and process the large
uncompressed datasets in reasonable time, it is necessary to deploy the routines in a distributed computing fashion, which is intended as future work. The first tests using more workers are presented in Figure 7.13 with a comparison of MapReduce-NMF time per iteration in function of number of workers in a single machine (Dell precision 7910 with Intel® Xeon® Processor E5-2687W-v3 and 64 GB of memory) for a compressed (with low sparsity) fingerprint dataset sized 1600 x 200,000. It can be seen that the time per iteration scales exponentially with the number of workers.

![Figure 7.13: Time per MapReduce NMF iteration in function of number of workers in a single machine.](image)

7.8 Conclusions

This Chapter has presented an effective way of performing NMF of ToF-SIMS large datasets by means of sparse pre-allocation and subsampling. NMF results of unbinned datasets may enable insights on the surface chemistry of hidden features and noisy
signals from low concentration compounds that can be overlooked if analyst-biased peak-selection techniques are employed. MapReduce in distributed frameworks can be an alternative to process next generation extremely large datasets. The next steps of the work will be the application of the methodology to more complex samples and the implementation of MapReduce using a high-performance computing cluster and MATLAB’s distributed computer server license at University of Surrey, which may enable the processing of entire uncompressed datasets without sub-sampling.

The peak-shifting effect is deleterious for MVA and must be corrected, however, as well as the total ion intensity in each pixel, the delay in time-of-flight also carries topographical information about the sample under analysis, which means the SIMS data itself can be used obtain qualitative topography maps, as demonstrated in a recent publication [131].

The subsampling methodology described here is now fully implemented in simsMVA and can be used for any imaging or 3D dataset loaded in the programme. The peak alignment routines still need more work to achieve a higher level of automation and consequently implementation in simsMVA, with its biggest challenges lying on processing time and variability across different datasets. Chapter 9 of this thesis will present an application for this methodology for the analysis of industrial samples and there an additional step towards peak alignment automation is made by incorporating a compression step that excludes from the dataset all spectral channels without any counts.

The publication of this work and subsequent discussions with colleagues at conferences sparked questions about the applicability of both subsampling and standard
MU NMF to low statistics data. This led to collaborations where it is intended to systematically test those two approaches in order to find out the limits of their validity on applications. This is topic of discussion of the next Chapter.
8 The Limits for the Application of PCA and NMF

8.1 Comparison of NMF algorithms for low statistics datasets

Chapter 3 described three algorithms that can be used in order to achieve an approximate solution for the NMF problem. In the SIMS community, most of the published work used either algorithms with alternating least squares or multiplicative update rules with the Euclidean norm as objective function that aims to maximise the likelihood of two normal distributions. The detection of secondary ions in ToF-SIMS is ruled by counting statistics and follow a Poisson distribution and pre-processing steps such as Poisson scaling (described in Chapter 4) aim to compensate for differences in noise due to Poisson statistics, however, the signal itself will still be Poissonian in nature. For a high number of counts per channel the central limit theorem [132] allows the approximation
to normal distributions but that is not always the case especially for high resolution imaging or 3D datasets where the number of counts per pixel is very low. This Section presents a systematic comparison amongst the three different NMF algorithms applied to synthetic datasets containing varied levels of counts per pixel. The reflection on possible limitations of the standard NMF algorithm arose only in the last year of the development of this thesis, therefore, these are preliminary attempts on such investigation.

8.1.1 Synthetic datasets

The synthetic datasets consisted of 50 x 50 pixels data cuboids with every pixel containing a spectrum with 100 channels. Pixels of distinct positions (concentric regions of different radii) had exclusive counts for one of three different peaks across the channels, as shown in Figure 8.1. The peaks span in number of channels were defined by a gaussian shape and the counts in each channel were created using a Poisson distribution-based number generator. Prior to NMF, all datasets were Poisson scaled and normalised by total counts following the methodologies described in Chapter 4. The NMF algorithms used are the ones described in Chapter 3: alternating least squares (ALS); normal multiplicative update (normal MU) and Poisson multiplicative update (Poisson MU). For each dataset, each algorithm run for 1000 iterations and 3 components were used. Matrix $W$ was initialised with random numbers and matrix $H$ initialised with random spectra from the dataset
8.1.2 Results & Discussion

Figure 8.2 shows the NMF results, using the three algorithms described in Chapter 3, for a series of synthetic datasets with varying total counts. Each column of the grid in Figure 8.2 represents a dataset and the numbers at the top of represent their counts per pixel. The first row contains the total spectra of each dataset and the subsequent rows alternate between an overlay of NMF components intensities (matrix $W$) and a plot of NMF characteristic spectra (matrix $H$). All overlays were done using false colour images in the RGB space.
Figure 8.2: NMF results of synthetic data. The numbers indicate counts/pixel.
A result is considered “successful” when it manages to separate the three regions into separate components and the spectrum of each component contains only the peak related to its respective region.

For the lowest statistic datasets (up to 7 counts per pixel), both the normal MU and ALS algorithms fail to reveal the correct circles structure and their respective peaks in the NMF results, with only the Poisson MU algorithm being successful. From 8 counts per pixel, the ALS algorithm becomes viable but the normal MU still fails. All algorithms start to succeed only from datasets with 10 counts per pixel.

Even though the ALS algorithm is more effective than normal MU for some datasets, it has repeatedly produced (not all runs are shown in Figure 8.2) results with mixed peaks in the components, which are reflected in mixed colour distributions across the circles in the components overlays (such as the ALS results for 25 counts per pixel). Since the datasets are synthetic and the circles are known to have exclusively only one of the three peaks, this result, that may look reasonable, is completely wrong and goes against the analytical chemistry view of NMF where there is the idea of un-mixing the signal of the dataset. Such problem is overcome in the chemometrics community by adding “unimodality” constraints to the algorithm, which would allow only one peak per component [68]. This works for some analytical chemistry techniques and also would work for the synthetic datasets proposed here, however, in the case of ToF-SIMS datasets this would not be viable as there are always several peaks per component. This means that care must be taken when using the ALS algorithm for NMF of ToF-SIMS data.
8.1.3 Final remarks

The systematic tests of NMF using synthetic datasets with Poisson variables showed that when the counts per variable are low, the Poisson MU algorithm is more suitable to find the real structure of the dataset. This is of relevance for ToF-SIMS imaging datasets, especially if one seeks to process uncompressed data, that would not benefit from integrating spectral channels into one variable (peak-picking).

For higher counts per variable per pixel, all three algorithms seem to succeed, however, care must be taken when using the ALS as it repeatedly resulted in mixed components, which can lead to wrong interpretations of data of real samples.

8.2 The Limits of Subsampled PCA

These are initial results of systematic tests using synthetic datasets to find when the pixel/voxel subsampling methodology described in Chapter 4 fails to detect a minor compound in a 3D dataset. PCA was chosen because the non-uniqueness of NMF would add uncontrolled sources of variation at this initial stage. These preliminary results are part of a collaboration with Dr. Sebastiaan Van Nuffel from the French National Centre for Scientific Research who provided the synthetic datasets.

8.2.1 Synthetic datasets

The datasets are sized $128 \times 128 \times 128 = 2,097,152$ voxels, which is a typical size of ToF-SIMS 3D depth profiling data. There is a total of 72 datasets and they all contain one major compound with one minor compound randomly dispersed with varying voxel
concentrations (9 levels between 10 ppm and 1% of the total number of voxels). The two compounds have 6 mass peaks in common and respectively 47 and 28 distinct mass peaks, giving a peak list of 81 variables. Various levels of Poisson noise were added. In order to avoid any bias, the exact compounds concentrations for each dataset were not disclosed. Figure 8.3 contains a 3D view of PCA results of one of the datasets showing the distribution of the two compounds and the variables related to each of them.

![Figure 8.3: 3D view of PCA results of one of the synthetic datasets.](image)

**8.2.2 Proposed workflow with detailed results**

All datasets are Poisson scaled. The scaled datasets are processed using MATLAB’s “pca” function of the Statistics & Machine Learning toolbox. A total of 56 different subset size percentage was considered and each case was repeated 10 times for subsample sizes under 1% and 2 times for the others, resulting in 1370 analyses per dataset. During the analyses, four **metrics** were recorded:
i) “Number of components” (NC): Total number of principal components that add up to 95% accumulated variance. Figure 8.4 shows a comparison of accumulated variance across principal components when processing 100% and 0.001% of the voxels.

![Figure 8.4: Accumulated variance across principal components when processing 100% (left) and 0.001% (right) of the voxels of a synthetic dataset.]

ii) “Correlation of PCs” (CORR): Sum of squares of correlation matrix in principal components space minus diagonal elements. Figure 8.5 shows correlations matrices for the first 4 principal components obtained by sampling 100% and 0.001% of the voxels. As described in Chapter 3, PCA decorrelates the dataset under analysis, which means CORR must be zero when 100% of the voxels are processed.
Figure 8.5: Correlation matrix for first 4 PCs. (A): 100%. (B): 0.001%
iii) “TIME”: Total time taken to obtain Scores and Loadings matrices.

iv) “Concentration of minor compound” (CONC): Percentage of voxels attributed to the minor chemical component. Two approaches are proposed for the identification:

a. $k$-means clustering of Scores matrix using two clusters: this is a somewhat robust but very time-consuming approach because of the size of the dataset and the need for lots of iterations to achieve convergence. Perhaps this method can be more useful in more complex systems (with more than two compounds for example). One alternative is to perform the clustering using only the first principal component, which is still time-consuming.

b. Separate voxels with negative and positive scores in the first principal component: this carries the assumption that the first principal component will always separate the major from minor component and their respective voxels will always have positive and negative scores. In order to guarantee this will happen most times, the Scores of first principal component were transformed prior to classification (using MATLAB notation):

$$S(:,1) - 0.5 \times (\text{min}(S(:,1)) - 1)$$
8.2.3 Overall results of first run

Figure 8.6 present the metrics recorded from the first runs of calculations. Every point represents an average across repeats and the error bar represents their standard deviation. In this case, approach a (k-means clustering) was taken to estimate CONC.

**Figure 8.6:** Four metrics recorded for subsampled PCA of multiple synthetic datasets.

Regardless of the dataset being processed, the metrics NC and TIME show very similar values for different subsample sizes. This result is expected for TIME as all datasets have the same total number of voxels and variables. The reason why NC also showed consistent values may be related to the fact that those datasets are synthetic with always
only two compounds. For all datasets, NC reaches a plateau at around 0.05 % subsample size which may be related to the subsampling limit in order to still retain all information.

**CORR** shows maximum value at around the same sample size for all datasets but the right-hand side tails are different. One hypothesis is that this is due to different levels of noise. As the number of components yielded may change for the different sample sizes, for future analysis **CORR** will be normalised by total number of components.

Moreover, the values of **CORR** and **NC** were found to be connected according to the equation:

\[ NC(s) \sim \log \left( \sum_{\text{SUBSIZES}}^s \text{CORR}(i) \right) \]

Which means that the log of the cumulative sum of **CORR** is proportional to **NC** at a given subsample size. Figure 8.7 shows plots of the cumulative sum of **CORR** against subsample sizes and values of **NC**. The implications of this relationship are still not clear but they may be useful in the future to predict information of different datasets.
Figure 8.7: For all datasets: Left: Cumulative sum of CORR in against subsample sizes. Right: NC against cumulative sum of CORR.

CONC is not very clear when plotted against subsample sizes. Figure 8.8 shows CONC in function of datasets and each point is a different subsampling size.

Figure 8.8: CONC values for various datasets. Each point represents a subsample size.
There are clearly many more over-estimations than correct ones but this is not due to subsampling. For both approaches a and b, the estimation failed in several cases (including cases with high concentration of minor compound and high subsampling percentage). This means that another approach must be taken in order to have a more reliable estimation. In spite of these estimation errors, analysis of individual datasets allowed one to see cases where voxels were successfully and unsuccessfully classified for low subsample sizes, as shown in Figure 8.9.
Figure 8.9: PC 1 scores of datasets with successful (A) and unsuccessful (B) PCA identification of the two compounds by subsampling both 100 % and 0.001 % of voxels.
8.2.4 Final remarks

An automated routine was successfully developed for the application of PCA to several synthetic datasets and recording of important metrics. A different approach for the identification of minor compound voxels in the principal components space must be taken and more metrics are intended to be recorded such as the cosine similarity between loadings obtained from subsampled dataset and the ones obtained from processing the full dataset [26]. Once all steps are optimised, more complex synthetic datasets can be created with features such as background noise and voxels containing levels of mixture of the two compounds.
PART III

APPLICATIONS

The following Chapters contain applications of all MVA methodologies described so far to scientific and industrial case studies of samples typically analysed at The Surface Analysis Laboratory of the University of Surrey. As it is described throughout the Chapters, those pieces of research have in fact been the main driving forces for development and implementation of specific functionalities of simsMVA.
CHAPTER 9: ANALYSIS OF ATMOSPHERIC PLASMA TREATED POLYPROPYLENE

9 ANALYSIS OF ATMOSPHERIC PLASMA TREATED POLYPROPYLENE

9.1 Introduction

This Chapter presents a combination of large area ToF-SIMS imaging with NMF in order to characterise a cold-atmospheric plasma treatment of the surface of an automotive grade polypropylene. All data analysis steps are a direct application of the methodology introduced in Chapter 7 for the analysis of uncompressed data. The results enabled the investigation of the spatial distribution of the treatment for different standoff distances as well as the effects of plasma plume mixing with atmosphere.
9.2 ToF-SIMS of polypropylene

There is now an extremely large body of data relating to the ToF-SIMS analysis of homopolymers, co-polymers and more complex formulations [37]. Such results have been obtained from reasonably pure samples of the candidate polymers and thus the development of fragmentation patterns and assignment of fragment ions has become a straightforward undertaking [37, 133]. Also well-known within the SIMS community is the prevalence of small molecules, added as processing aids to such polymers [134] which have a tendency to migrate to the surface possibly as part of processing by the usual shaping methods such as extrusion or injection moulding. Similar phenomena occur during the processing of a range of automotive grade polypropylene (PP) formulations. Such materials have ethylene propylene diene monomer (EPDM) or other polymers such as polystyrene or polycarbonate incorporated in their formulation to improve impact resistance [135]. The copolymer formulations are then filled with particles such as carbon black and/or talc and number of processing aids and other additives are used to improve behaviour and performance [135, 136].

Extensive research has been done on automotive grade polypropylene using SIMS, mainly focusing on the polymer formulation itself [137, 138]. Galuska [137] has established calibration curves to determine contents of ethylene propylene based on ratios between characteristic SIMS fragments from the surface of different pure PP polymer blends. Tomasetti et al [138] have applied such calibration to standard injection moulded samples. Recent work from the group at University of Surrey [139] identified two additives by ToF-SIMS in such formulations.
9.2.1 Energetic surface treatments

Energetic surface treatments of polymers have been extensively used to improve adhesion, biocompatibility and other surface-active processes. The most commonly used treatment methods are flame, plasma, corona discharge and more recently cold-atmospheric plasma [140–148]. Atmospheric plasma jets became popular as a result of their relatively low cost and in-situ implementation possibilities. They are also more flexible since a vacuum system is not required and the low temperature of the plume makes it less detrimental to polymer surfaces [149].

ToF-SIMS has been shown to be an appropriate technique for the characterisation of plasma surface-treated polyolefins, with several pieces of work describing the characteristic secondary ion fragments that a treated surface will yield [142, 150–156]. Figure 9.1 shows the main oxidation mechanism that occurs in a polypropylene chain with two examples of characteristic SIMS fragments. Studies have also reported that energetic surface treatments will drive the surface segregation of additives and the presence of such molecules at the surface of a polymer can have a deleterious effect on the desired application performance [142].

ToF-SIMS characterisation of the treated surfaces is often aided by MVA methods such as PCA to determine the main characteristic peaks after treatment or even the depth of oxidation using a dual beam depth profiling set up [150, 153]. The reactivity zones of a plasma treated polymer surface have been investigated with X-ray photoelectron spectroscopy (XPS), which is also widely used for such characterisation [142, 151, 157]. No similar investigation has been carried out using ToF-SIMS. The
spatial distribution of the treatment is much larger than the typical raster size of a ToF-SIMS primary ion beam, but modern instruments provide stage-raster modes that enable imaging analysis of such large areas. The current challenges lie on the acquisition of automated large area datasets of strong insulators and the post-processing of the large amount of data generated. The work described in this Chapter presents large area ToF-SIMS imaging of the surface of an automotive grade polypropylene after cold-atmospheric plasma treatment. The main challenges regarding pre-processing will be presented as well as NMF results, which have been made possible to achieve after the development of methodologies presented in Chapter 7 of this thesis for the processing of large area ToF-SIMS datasets.

Figure 9.1: main oxidation mechanism that occurs in a polypropylene chain with two examples of characteristic positive SIMS fragments.
9.3 Experimental

9.3.1 Polypropylene samples
The polymer is a carbon black filled automotive grade polypropylene (PP) provided by Sabic (Redditch, UK). These materials are used mainly in dashboards and fan shrouds and undergo plasma/flame treatment prior to paint application. The samples were supplied as 5 mm thick injection moulded plaques that were cut into small coupons (approximately 4 cm x 4 cm x 5 mm) prior to treatment.

9.3.2 Plasma composition and treatment
The plasma torch used in this work is a Plasma Tact system from Ad-Tec Europe with a modification that allows the use of argon or helium as a primary gas and has an additional input to allow doping of the inlet gas with other gases. A detailed description of the plasma system is published elsewhere [158]. For this work a mixture of argon and air were chosen as the input gases. Two samples were treated at different stand-off distances (4 mm and 8 mm). For both samples, the inlet gas flow rate and the torch velocity were set respectively as 8 L min\(^{-1}\) and 150 mm min\(^{-1}\).

9.3.3 ToF-SIMS analyses
ToF-SIMS analysis of positively charged secondary ions was carried out using the TOF.SIMS 5 system at the University of Surrey using 25 keV Bi\(_3^+\) ion beam operated in the high current bunched mode delivering 0.3 pA with 100 µs cycle time, resulting in
a mass range between 0 and 694 u. Secondary ion maps were acquired using the stage raster mode. One patch represents the maximum beam raster area of 500 x 500 µm². The total 2 x 2 cm² analysed area contains 2000 x 2000 pixels which comprise of 40 x 40 patches sized 50 x 50 pixels each. The whole area was scanned once with one shot per pixel, ensuring static conditions. A low energy, 20 eV, electron flood was used to neutralise charge build-up on the sample surface. For both samples, ToF-SIMS characterisation was performed immediately after surface treatment.

9.4 Data handling and processing

In order to apply the methodology proposed in Chapter 7 to a more challenging, industrial problem, the MVA method chosen to process the data was NMF using the MU algorithm. The computer used was a Dell Optiplex 9020 PC with a core i5 processor and 32 GB of RAM. Datasets were exported from SurfaceLab software (IONTOF GmbH) to General Raw Data (GRD) format and imported into MATLAB 2016b. After pre-processing, the data was loaded into simSMVA to undergo multivariate analysis. The pre-processing steps employed were peak alignment, normalisation by total ion counts and Poisson scaling following recommendations described in Chapter 4. NMF was done in subsets of the whole datasets following the methodology described in Chapter 7. The number of components used in the factorisation (four for both samples) was determined by variance investigation of principal component analysis of the same subsets processed with NMF.


9.4.1 Spectral alignment

To acquire large area ToF-SIMS maps in an automated and quick manner, it is necessary to give up some adjustments such as height and reflector voltage correction at every patch measured. For this reason, the sample being analysed must be fairly homogeneous both in terms of charging and topography. The polypropylene coupons are strong insulators and their surface were not perfectly flat within such a large area of analysis. This led to different detected time-of-flight for ions of the same mass coming from different regions, resulting, ultimately, in misaligned spectra at different pixels. The correction for secondary ion time-of-flight delays and consequently the alignment of peaks are essential for MVA of uncompressed data, since each spectral channel is regarded as a variable. There are four main approaches to overcome the misalignment problem, two of them related to the experimental set up and the other two related to data analysis:

1) Better sample preparation by cutting thinner slices of the samples surface to reduce the differential charging effect.

2) Mount samples over thick glass (e.g. a microscope slide) to ensure homogeneous charging.

3) Integrate the area under specific channels ranges and regard them as variables (peak-picking).

4) Align the spectra of all pixels to a reference pixel spectrum.

In order to perform the numerical alignment of the spectra of all pixels the data must have enough counts per pixel that result in well-defined peaks. For this paper, the main
concerns for the characterisation of the samples were on the *large area* and *high spectral resolution* aspects of the data, therefore, in order to achieve enough counts per pixel it was decided to compromise on the least significant aspect of the dataset which is the spatial resolution. The original 2,000 x 2,000 pixels x 2,000,000 channels dataset did not have sufficient statistics per pixel but the data shown in this work was compressed down to 50 x 50 pixels x 2,000,000 channels, which enabled the spectra alignment to be done using a well-known Fourier transform-based alignment routine [124]. An additional compression step was done prior to alignment: masses outside the range of integer nominal mass +/- 0.2 u were excluded from the dataset as counts in those regions are zero or only due to noise. Figure 9.2 shows a general schematic of the compression steps.

![Figure 9.2: Data compression steps. Data of neighbour pixels are joined together (pixel binning) and spectral channels without counts are excluded.](image)

Figure 9.3 illustrates the loss of chemical information and the impossibility of performing multivariate analysis of misaligned data. It shows a comparison of NMF results with 3 components (without merits to the chemistry involved yet) for two datasets: one acquired with a sample mounted directly on the sample holder and a second
one where the sample was mounted on a thick glass slide. The maps are RGB overlays of the folded W matrices obtained from NMF. The use of thinner slices of the samples’ surface did improve the differential charging but it increased surface roughness, therefore not improving much the quality of the data. Mounting the samples on glass was beneficial to data quality and facilitated the acquisition but it was not enough to suppress the alignment issue. Peak picking may have worked for a different dataset were all peaks are well defined but for this specific experiment, the relevant peaks of the pristine and oxidised material will often have the same nominal mass and thus be very close together in the spectra (for example the three peaks at nominal mass 69 u - C₅H₅⁺, C₄H₅O⁺ and C₃HO₂⁺) and impossible to be resolved unless the spectra across pixels are aligned. Mounting samples on glass (Figure 9.3c) improves the quality of data. However, NMF still picks up small shifts (due to differential charging or surface roughness).
9.5 NMF results

Figure 9.4 shows NMF results with four components for the samples treated with the plasma torch at 4 mm (first row of plots) and 8 mm (second row of plots) from the surface. The two datasets were not processed together in a row-wise matrix augmentation fashion [87, 159] because of two main reasons that have the potential to confuse automated alignment routines i) differences in mass resolution between datasets
and ii) very different calibration from one dataset to the other. Such an approach could have been done with peak-picked or spectral channels-binned datasets (similar to the stainless-steel data presented in Chapter 13), however, this would have not kept the high-resolution feature in the spectra of NMF components which are essential for the characterisation of the levels of oxidation of the surface. The images on the left-hand side of Figure 9.4 are an overlay of all components intensities (or columns of matrix $W$). One colour was attributed to each component and the opacity level of each colour at each pixel is proportional to the intensities of their respective component (or values in matrix $W$). The plots on the right-hand side of Figure 4 are the component spectra (or rows of matrix $H$). Component 2 (dark blue) has all characteristic peaks of polypropylene and represents the untreated surface. The other three components (yellow, light blue and red) contain roughly the same set of peaks but at different relative intensities. The peaks are a combination of polypropylene characteristic peaks and oxygen containing peaks. The most relevant peaks of all components are present in Table 9.1. It is important to note that the fragments containing one or two oxygen atoms will often have the same nominal mass of characteristic peaks of the pristine material (e.g. 55 u and 69 u). This makes the peak-alignment pre-processing step crucial for the correct characterisation of the treatment. Furthermore, as observed before [142, 150, 155], fragments with two oxygen atoms are representative of higher degrees of oxidation of the surface.
8 mm standoff distance

4 mm standoff distance

Figure 9.4: Left: overlay of all components intensities (field-of-view: 2 cm x 2 cm).

Right: characteristic spectra of components. The inset images are the individual maps of each component.

Both samples presented very similar results even though their datasets have undergone individual factorisations. The practice of processing datasets from similar samples as a single matrix has proven to help interpretation before [85, 160], however, the datasets from the two samples had very different channel to mass calibrations (due to charging effects) and different peak shapes,
Table 9.1: Most relevant ToF-SIMS peaks found for non-treated (hydrocarbons) and treated (oxygen-containing) polypropylene surface.

<table>
<thead>
<tr>
<th>FRAGMENT</th>
<th>MASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHO⁺</td>
<td>29.001</td>
</tr>
<tr>
<td>C₂H₅⁺</td>
<td>29.038</td>
</tr>
<tr>
<td>CH₃O⁺</td>
<td>31.017</td>
</tr>
<tr>
<td>CH₂O⁺</td>
<td>33.034</td>
</tr>
<tr>
<td>C₂H₂O⁺</td>
<td>42.008</td>
</tr>
<tr>
<td>C₃H₆⁺</td>
<td>42.044</td>
</tr>
<tr>
<td>C₂H₃O⁺</td>
<td>43.0189</td>
</tr>
<tr>
<td>C₃H₇⁺</td>
<td>43.0559</td>
</tr>
<tr>
<td>CHO₂⁺</td>
<td>44.9958</td>
</tr>
<tr>
<td>C₂H₅O⁺</td>
<td>45.0329</td>
</tr>
<tr>
<td>C₃H₃O⁺</td>
<td>55.0228</td>
</tr>
<tr>
<td>C₄H₇⁺</td>
<td>55.0583</td>
</tr>
<tr>
<td>C₃H₅O⁺</td>
<td>57.0373</td>
</tr>
<tr>
<td>C₄H₉⁺</td>
<td>57.0768</td>
</tr>
<tr>
<td>C₂H₃O₂⁺</td>
<td>59.014</td>
</tr>
<tr>
<td>C₅H₇O⁺</td>
<td>59.0511</td>
</tr>
<tr>
<td>C₃HO₂⁺</td>
<td>68.9968</td>
</tr>
<tr>
<td>C₄H₅O⁺</td>
<td>69.035</td>
</tr>
<tr>
<td>C₃H₉⁺</td>
<td>69.0754</td>
</tr>
</tbody>
</table>
9.5.1 Effect of treatment distance

Both samples treated at 4 mm and 8 mm stand-off distances have similar NMF results, with an outer treated region less oxidised. The difference between the two samples concerns the central highly oxidised region. The sample treated at 8 mm has a homogeneous core and the treated at 4 mm has highly oxidised halo with a less oxidised core. A reason for this is probably that the closer distance presents a less oxidising part of the flame to the surface as shown in the schematic of Figure 9.5. This can be a problem in industrial applications were the plasma tip is rastered across a large surface area and a homogeneous treatment is sought.

Figure 9.5: Schematic for the effect of standoff distance to surface oxidised regions.
9.5.2 Effects of plasma mixing with atmosphere

As a consequence of the high intensities of the oxygen-containing species and additives at the surface of the treated samples, lower intensity fragments containing nitrogen atoms were not grouped exclusively by any NMF component. However, different reactivity zones around an atmospheric plasma jet have been investigated using XPS [149] with an indication that nitrogen species will have a different surface distribution. Figure 9.6 shows, for the sample treated at 8 mm standoff distance, the intensity maps of the ion fragments NO$_2^+$, CH$_4$NO$^+$ and C$_2$H$_8$N$^+$. It is important to note that those three peaks are all at the same nominal mass (46 u) and therefore their intensity distribution across the sample surface would not be possible without the peak alignment step.

![Intensity maps of ion fragments NO$_2^+$, CH$_4$NO$^+$ and C$_2$H$_8$N$^+$.](image)

**Figure 9.6:** Top: intensity maps of the ion fragments NO$_2^+$, CH$_4$NO$^+$ and C$_2$H$_8$N$^+$. Bottom: total spectrum at nominal mass 46 u.
The maps for CH$_4$NO$^+$ and C$_2$H$_8$N$^+$ show high intensity in the core area, similarly to the highly oxidised region identified by NMF in Figure 11.3. The presence of carbon and hydrogen atoms on those fragments indicates that nitrogen groups from atmosphere reacted with the activated surface during and even after treatment. The peak identified as NO$_2^+$ (45.992 u) could also be attributed to CH$_2$O$_2^+$ (46.005 u) or $^{13}$CHO$_2^+$ (46.001 u), which are characteristic of treated PP surfaces. However, its map shows a halo-shaped distribution similarly to what has been observed for nitrogen distribution by XPS on both polymer and steel plasma-treated samples [149, 158]. As the fragment does not contain carbon or hydrogen atoms, the halo distribution may be exclusively from nitrogen incorporation as a consequence of air flow around the plasma jet.

**9.5.3 Behaviour of additive**

The most abundant additive detected at both samples surfaces was the antioxidant Irganox 1010 with chemical structure shown in Figure 9.7.
Figure 9.7: Chemical structure of Irganox 1010 antioxidant additive.

The most characteristic secondary ion fragments of Irganox 1010 are $\text{C}_{15}\text{H}_{23}\text{O}^+$ (219 u) and $\text{C}_{17}\text{H}_{25}\text{O}_2^+$ (259 u). Figure 9.8 shows the spatial distribution of the peak at 219 u for both the 4mm standoff and 8mm standoff treatment distances. As expected, the additive surface segregates to more oxidised regions and its spatial distribution correlates with the distribution of component 4 of the NMF results. In fact, the peaks at 219 u and 259 u are present in the characteristic spectra of component 1 with higher intensity than in the other 3 components.
As future work, it is intended to investigate whether the different regions will present different levels of oxidation of the oxidised additive in a similar manner that the polymer oxidises. The first reaction of the molecules after surface segregation is expected to follow a chain-breaking donor mechanism where the polymer radical takes a H atom from the inhibitor (AH) which is transformed into the radical (A°) which can interact with another peroxo radical [143, 161]. The second proposed step is that all CH₃ ends of the additive will also oxidised following the same mechanism described in Figure 9.1. All steps are illustrated in Figure 9.9.

Figure 9.8: Spatial distribution of additive Irganox 1010 tracked with area under the peak at 219 u. Top: 8 mm standoff distance. Bottom: 4 mm standoff distance.
Figure 9.9: Oxidation of the additive Irganox 1010 after surface segregation in a polypropylene matrix.
9.6 Conclusion & Future work

The proposed data analysis methodology has successfully simplified a very large and complex ToF-SIMS dataset of the surface of an atmospheric plasma treated polypropylene coupons. The results kept the high spectral resolution, which was essential for the characterisation of the oxidised and reactivity zones of the plasma at two different plasma torch standoff distances. As future work, it is aimed to further explore the very rich NMF results, focusing on the behaviour of additives and other small molecules and their spatial distribution after treatment.

This was a direct application of the methodology proposed in Chapter 7 and is an excellent example of cases were the only option for effective MVA is the processing of the uncompressed spectra.
10 Analysis of Wood Growth Regions

10.1 Introduction

This Chapter presents a methodology to characterise the surface of Scots Pine wood \((Pinus sylvestris \text{ L.})\) samples cut at different geometries by performing NMF of ToF-SIMS data. With the aid of two standard samples of lignin organosolv and cellulose, the NMF results reduced the data to two compounds which are representative of the surface coverage of lignin and cellulose. Differences between the relative amounts of lignin and cellulose in early and late wood were in agreement to the anatomy of wood, making the proposed methodology of multivariate analysis aided by data of standard samples a potential way to differentiate species and potentially age of the wood in terms of their natural components.
10.1.1 **Structure of wood**

The major chemical components of wood are cellulose, hemicellulose, lignin and extractives. Cellulose is the main component present in plants cell walls and thus the most abundant organic polymer on earth [162]. It is mostly responsible for the high strength of wood and is also the basic building block of textiles and paper. Cellulose is a natural polymer made up of repeating units of beta linked D-glucose joined by glycosidic bonds with the formula \((\text{C}_6\text{H}_{10}\text{O}_5)_n\). It is thus a polysaccharide consisting of hundreds to thousands of the D-glucose molecules. In plants, several cellulose chains are arranged in parallel to form cellulose microfibrils. These microfibrils range from 2 to 20 nm in diameter and can be 100 to 40000 nm [162]. Cellulose microfibrils are subsequently joined together to form cellulose fibres that provide the structural support in a plant cell wall, as shown in Figure 10.1.

![Cellulose microfibrils and fibrils in a plant cell wall](image)

**Figure 10.1:** The arrangement of cellulose microfibrils and fibrils in a plant cell wall as the chemical diagram of a single cellulose chain. Figure taken from [163].

Hemicellulose is any heteropolysaccharide present in the plant cell wall alongside cellulose. It can be made up of a wide variety of sugar monomers such as glucose,
xylose, galactose, mannose, arabinose and rhamnose. As a consequence of the difference in monomers, hemicellulose is a random and amorphous material that is relatively weak compared to cellulose. Two of the most predominant hemicelluloses in softwoods are xylan and glucomannan the structures of which are shown in Figure 10.2.

![Figure 10.2 - Two possible molecular structures of Xylan and Glucomannan.](image)

Lignin is a complex polymer of coniferyl, syringyl and coumaryl. It fills the space between the cellulose and hemicellulose components in a plant cell wall, essentially acting as the matrix in a composite material. Lignin covalently bonds with hemicellulose and therefore forms crosslinks with the plant polysaccharides that make up the hemicellulose. These cross-links contribute to the mechanical strength of the plant cell wall and consequently the wood [165]. The molecular structure of lignin is shown in Figure 10.3.
The various properties of wood make it one of the most adaptable construction materials available. Wood is strong and relatively light and flexible compared to other construction materials such as steel. The natural structure of wood defines it as an anisotropic material, so that each direction has different physical and chemical properties. Wood has a wide range of applications including the production of paper by the pulp and paper industry which is one of the largest industrial sectors in the world. Another important application is the manufacture of oriented strand board (OSB) which is used in the construction industry for walls, floors and roofs. The understanding of

Figure 10.3 - Molecular Structure of Lignin

(https://en.wikipedia.org/wiki/Lignin)
interfacial properties between wood and coatings or adhesives is crucial and a first step is the investigation of the surface of wood prior to any treatment and ToF-SIMS is a suitable technique for this purpose.

10.1.2 ToF-SIMS of wood

Since the advent of the applicability of ToF-SIMS analysis to organic materials, a number of characteristic ion fragment assignments for wood have been identified for lignin, cellulose and extractives. Saito et al. [166] and Goacher et al. [167] have established an extensive list of characteristic fragments generated by soft wood samples which, together with the work from Fardim and Duran [168], provide a detailed database for ToF-SIMS spectra of wood.

Multivariate analysis can be a powerful tool to aid such investigation because of the similarity of the spectra from different regions and the high level of complexity in ToF-SIMS data of wood samples. Goacher et al. have applied PCA and NMF to ToF-SIMS spectra and images of wood materials for the spatial identification of lignin and cellulose contents, within wood cells, in wood cross-sections [167] and in the study of enzymatic activity on wood substrates [169]. No work, so far, has used multivariate analysis to characterise the surface of radially and tangentially cut wood, of which the superficial and interfacial properties are important in the coating and bonding of wood.

The work described in this Chapter aims to assess the ToF-SIMS data of the surface of Scots Pine wood samples cut with different geometries in order to differentiate growth regions in terms of their chemical composition. Such assessment
was achieved through NMF, making use of well-defined databases of characteristic secondary ions.

10.2 Methodology

10.2.1 Sample preparation

Samples of Scots Pine wood approximately 70 µm thick were cut using microtomy performed on a Microm HM355S motorised rotary microtome (Optech Scientific Instruments, Thame, UK) equipped with a standard specimen clamp and a tungsten carbide knife (cleaned with acetone) to provide radial and tangential cuts. In total, 4 different sections of each cut were obtained from the same ring in one block of wood. The cuts exposed both early and late growth rings for the radial cut samples as shown in Figure 10.4. No prior cleaning or extraction was performed to the wood samples. Samples of cellulose (Sigma Aldrich) and organosolv lignin (Lignol Innovations now acquired by Fibria Innovations) were prepared from reference powders in the following manner: Cellulose was compressed into a dimpled stainless-steel sample holder and lignin was dissolved with acetone and deposited onto a piece of silicon wafer (previously sonicated in acetone for 15 minutes) to form a thin film. Organosolv lignin was employed as it was readily available from a separate experiment, and of known provenance. It is well known, however, that there are differences between this standard material and milled wood lignin.
ToF-SIMS analyses

The ToF-SIMS analysis was carried out using the TOF.SIMS 5 system at the University of Surrey using a 25 keV Bi\textsuperscript{3+} ion beam operated in the high current bunched mode delivering 0.3 pA, pulsed at 10 kHz and rastered over areas of 75 x 75 µm\textsuperscript{2}. Several different regions were analysed for each sample at both early and late growth regions. Positive ions spectra were acquired for 100 seconds each, keeping the total primary ion dose well below the static limit of 10\textsuperscript{13} ions/cm\textsuperscript{2}/analysis. All the samples were mounted on a glass plate using double-sided adhesive tape as shown in Figure 10.4 to ensure
homogeneous charging and a pulsed low energy electron flood gun was used to eliminate charge building up on the samples surfaces during the analyses.

10.2.3 Data processing

All spectra were calibrated using one unique mass list containing lignin and cellulose characteristic peaks. The broad range of masses in the calibration list provided enough calibration points to achieve accurate peak assignments. A high spectral resolution peak list with areas of forty-nine fragments (see Table 10.1) was created based on the work by Goacher et al [167] and using the SurfaceLab software (ION-TOF GmbH). Peaks at m/z = 73, 147 and 281 u were excluded from the list because of possible overlap with peaks of the common surface contaminant polydimethylsiloxane (PDMS). The main data pre-processing steps employed were normalisation by total counts, Poisson scaling and mean centring. In order to improve the statistics, the initial data matrix of 35 measurements x 49 peaks was resampled to a larger matrix sized 1000 x 49. The resampling was made by bootstrapping the median of each of the five groups of observations (pure cellulose, pure lignin, radial early, radial late and tangential) to new sets of 200 observations each. By bootstrapping the median of the multivariate distribution of the data points, it was possible to better infer any statistical property of such distribution with no prior assumptions (except the pre-defined groups from technical replicates). Also, the larger number of data points will aid multivariate analysis (especially for methods that account for different kinds of correlations). All pre-processing and multivariate analysis were performed on MATLAB R2015a using an early version of the simsMVA software. PCA was performed with the peak areas
(variables) for each of the spectra (observations). NMF was performed on the same data matrix using the multiplicative algorithm described in Chapter 3.

Table 10.1: List of peaks for secondary ion fragments characteristic of wood

<table>
<thead>
<tr>
<th>Centre Mass (u)</th>
<th>Assignment</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.022</td>
<td>CH$_3^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>19.018</td>
<td>H$_3$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>22.990</td>
<td>Na$^+$</td>
<td>Sodium</td>
</tr>
<tr>
<td>27.022</td>
<td>C$_2$H$_5^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>28.027</td>
<td>C$_2$H$_4^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>29.039</td>
<td>C$_2$H$_5^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>31.017</td>
<td>CH$_3$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>44.022</td>
<td>C$_{13}$CH$_3$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>44.056</td>
<td>C$_2$H$_3$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>45.035</td>
<td>C$_2$H$_5$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>47.012</td>
<td>CH$_3$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>51.020</td>
<td>C$_4$H$_5^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>59.013</td>
<td>C$_2$H$_3$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>60.021</td>
<td>C$_2$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>61.031</td>
<td>C$_2$H$_3$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>63.021</td>
<td>C$_3$H$_5^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>65.038</td>
<td>C$_3$H$_5^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>67.058</td>
<td>C$_3$H$_7^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>71.013</td>
<td>C$_3$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>73.031</td>
<td>C$_3$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>77.035</td>
<td>C$_6$H$_5^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>79.052</td>
<td>C$_6$H$_7^+$</td>
<td>Lignin</td>
</tr>
<tr>
<td>81.030</td>
<td>C$_5$H$_5$O$^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>81.073</td>
<td>C$_6$H$_9^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>83.013</td>
<td>C$_4$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>85.035</td>
<td>C$_4$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>87.048</td>
<td>C$_4$H$_5$O$_2^+$</td>
<td>Polysaccharide</td>
</tr>
</tbody>
</table>
### 10.3 Results

PCA was performed in the full dataset to obtain initial conditions to feed the NMF analysis. The loadings of the first principal component (PC1 - 94% of total explained variance) agreed with the classification presented by Goacher et al [167] as shown in

<table>
<thead>
<tr>
<th>Mass</th>
<th>Formula</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.055</td>
<td>C₇H₇⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>93.079</td>
<td>C₇H₉⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>95.097</td>
<td>C₇H₁₁⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>97.038</td>
<td>C₅H₅O₂⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>99.047</td>
<td>C₅H₅O₂⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>101.037</td>
<td>C₄H₅O₃⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>105.075</td>
<td>C₈H₉⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>107.093</td>
<td>C₇H₇O⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>109.031</td>
<td>C₆H₅O₂⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>113.028</td>
<td>C₃H₅O₃⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>115.047</td>
<td>C₃H₅O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>127.046</td>
<td>C₆H₅O₃⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>128.051</td>
<td>C₆H₅O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>131.042</td>
<td>C₅H₅O₄⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>133.044</td>
<td>C₃H₉O₄⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>137.059</td>
<td>C₈H₉O₂⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>145.093</td>
<td>C₆H₉O₄⁺</td>
<td>Polysaccharide</td>
</tr>
<tr>
<td>147.120</td>
<td>C₆H₁₁O₄⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>151.051</td>
<td>C₈H₇O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>152.047</td>
<td>C₈H₈O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>153.053</td>
<td>C₈H₉O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>165.059</td>
<td>C₉H₉O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>167.071</td>
<td>C₉H₁₁O₃⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>181.068</td>
<td>C₉H₉O₄⁺</td>
<td>Lignin</td>
</tr>
<tr>
<td>189.04600</td>
<td>C₁₁H₉O₃⁺</td>
<td>Lignin</td>
</tr>
</tbody>
</table>
Figure 10.5, with positive direction loadings containing lignin peaks and negative direction loadings containing polysaccharides (cellulose and hemicellulose) peaks.

**Figure 10.5**: Loadings of PCA results for principal component 1 (PC1). Values normalised to maximum in each direction (positive and negative).

The positive loadings of PC1 were separated from the negative loadings, making two non-negative matrices that were used as initial conditions for the NMF analysis of which model considered two pure compounds. The final results for the peaks comprising the pure compounds after 500 NMF iterations are shown in Figure 10.6.
Figure 10.6: Peaks comprising pure compounds obtained after 500 iterations of NMF. Top: compound 1. Bottom: compound 2. Values normalised to maximum.

Neither of the two final compounds presented exclusively peaks for lignin or cellulose. Although this might indicate a local minimum of the NMF, such overlap is expected as there will be background counts of all peaks in both lignin and cellulose. Regardless of the reason, standard lignin samples presented large amounts of compound 1 and little of compound 2 whereas standard cellulose samples presented the opposite behaviour, enabling “relative quantification” in terms of both compounds. Figure 10.7 (a) shows the scores or relative “concentrations” for all samples analysed (values are normalised to maximum).
CHAPTER 10: ANALYSIS OF WOOD GROWTH REGIONS

Figure 10.7: (a) Relative “concentrations” for all samples of both compounds obtained by NMF. (b) Average of each group of spectra for the same samples. The error bars represent the standard deviation of the mean of measurements in a set.

It is important to note that all data points can be represented with only a slight deviation from a straight line with no additional constraints applied to the problem, meaning that the more of compound 1, the less of compound 2. Considering the almost pure relative concentration values for lignin (> 80 % for component 1) and cellulose (> 90 % for component 2), it is acceptable to name compound 1 “lignin” and compound 2 “cellulose”. Values for both radial early and late presented larger scatter in comparison to the tangential ones. This might indicate that some measurements were performed at the interface of early and late regions.
10.4 Discussion

Even though the spectra of the different regions of the wood samples are very similar, the multivariate analysis made it possible to differentiate very clearly between them. Two main observations can be made for their relative concentrations: firstly, all wood samples have relatively more cellulose than lignin, which agrees with the fact that cellulose and hemicellulose comprise up to 75 % of wood. Also for the radially cut samples, late growth rings have relatively more cellulose than early growth rings. To better illustrate such observations, Figure 10.7 (b) shows the average of each group of spectra for the same samples and the standard materials.

Scots Pine wood consists of cellulose (40 %), lignin (25 – 30 %), hemicellulose (28.5 %) and extractives (3.5 %) [170]. The basic units of wood are micro fibrils which make up the cells of the tree, which have different growth rates depending on the season of the year. Cells grow faster in spring and forms what it is known as *early wood* while in the summer the growth is slower, forming *late wood*. The early and late wood together form an annual ring. When a block of wood is cut, radially or tangentially, the resulting sample will always present exposed cell walls at the surface. It is known that the fibres in the cell walls are cellulose rich and the middle lamella that binds the cells together are more concentrated of lignin. Figure 10.8 shows a schematic for an ideal cut of a wood sample with views for a cross section (left) and a plan view of the region analysed by ToF-SIMS (right).
Figure 10.8: A scheme for an ideal cut of early (upper) and late (lower) wood samples with views of a cross section (left) and the plan view (right) of the region analysed by ToF-SIMS. Not to scale.

Even for non-ideal cuts, it is reasonable to consider that the area analysed will yield an average behaviour of the scheme presented in Figure 10.8, which shows that there is indeed more cellulose being probed than lignin in any case. The differences between the values for early and late growth rings observed in Figure 10.7 can be explained by the different growth rates of early and late cells that cause late wood to have smaller cells and thicker cellulose-rich cell walls than early wood (2–4 μm and 4–8 μm for early and
late cells respectively [171]). As it is shown in Figure 10.8, late wood will have more cells and the average cellulose coverage is higher than the value for early wood within the same analysis area. Therefore, the higher amounts of cellulose observed for late growth regions in the NMF results are also in agreement with the general anatomy of wood and the difference between radial and tangential can be explained by the higher asymmetry of late wood cells. It is important to mention that the scheme of Figure 10.8 is ideal and does not consider the possibility of the ToF-SIMS signal for the characteristic peaks of lignin and cellulose being hindered by other compounds such as extractives that can segregate to the surface in different amounts for each kind of growth ring. Goacher et al. [172] have reported that wood extractives can yield secondary ion fragments similar to those characteristic of lignin, allowing potential mass interferences. This would also explain the fact that the relative concentrations obtained for lignin and cellulose do not add up to 1 in Figure 10.7. One experiment to be carried out, to improve and validate the methodology presented in this work, is to perform the removal of extractives of the wood using specific solvents and repeat the ToF-SIMS analysis. The comparison of the results with XPS analyses of the same set of samples will also aid the validation of the methodology. Following a methodology proposed by Johansson et. al. [173], Banuls-Ciscar et. al. [174] have analysed wood strands by XPS and reported that the differences between early and late growth rings in the C1s peak are reduced after extraction.
10.5 Conclusions and future work

This work used an early version of simsMVA and was the first work that the methodology of inputting data of standard materials was employed. Therefore, the necessity of understanding the chemistry of wood drove the development of the NMF routines that are present in the current version of the software.

The agreement between the NMF results and the anatomy of wood shows how powerful the methodology can be to quantify relative amounts of lignin and cellulose in a wood sample and differentiate between early and late growth rings. The next step of this work is to expand the analysis to growth ring of more wood species and increase the list of characteristic peaks by including ToF-SIMS fragments of common extractives and consider the presence of other compounds such as hemicellulose by performing NMF with more factors or “pure compounds”. The results will confirm whether or not extractives are present at a greater concentration in one of the growth regions.

This work also branched into other pieces of research carried out at University of Surrey: it held the basis for the development of a PhD thesis on the analysis of interfacial properties between wood and polyurethane-based adhesives [175] and also branched into a Master’s project on the differentiation of hardwood species using ToF-SIMS and MVA [176].
11 Analysis of Brewed Coffee

11.1 Introduction
This Chapter describes results of a collaboration between researchers of the University of Surrey and Federal University of Rio Grande, in Brazil. The group in Brazil has a long tradition of characterisation of food and beverage samples using ion beam analysis techniques such as particle induced X-ray emission (PIXE) [177, 178]. However, for the characterisation of coffee, a more chemically specific technique was necessary, hence the collaboration with the author.

11.2 Mass spectrometry of coffee
Coffee is the third most widely consumed beverage in the world, exceeded only by water and tea [179]. This complex beverage, which was reported by Avicenna in his classic book The Canon of Medicine [180] in the 11th century, arrived in Europe via Italy and spread to the American continent [179] by France during the 18th century. Since then
its production has been growing around the world. According to the International Coffee Organization, from 1990 to 2016, the total production of all exporting countries grew 65% [181], while the export grew by 43%, which also shows growth in the internal consumption.

Considering this growth, the complexity of the beverage (that has about 900 volatile substances [182]) and the central role played by the coffee in the lifestyle and the world economy, different types of coffee certifications were created to guarantee the attributes of the product [183] and at the same time improve farmers’ profits [184]. As a consequence of the relevance of coffee taste, there are a number of papers that focus on the use of analytical chemistry techniques mainly to detect the compounds responsible for different flavours in coffee [183]. Other papers use coffee samples from different countries to determine their geographical origin [185, 186], once the characteristics of the coffee are in part a combination of environmental and genetic factors. Other papers focus on detecting specific compounds such as chlorogenic acids CGA [187], because of its capability of reducing blood pressure [188], or even influence the quality of the coffee that is drunk [185, 189, 190].

The majority of the studies have employed some sort of mass spectroscopy, such as gas chromatography mass spectrometry (GC-MS) or proton transfer reaction mass spectrometry, to analyse the volatile compounds, which involves some sample preparation and is optimised for one specific range of compounds. Other techniques such as easy ambient sonic-spray ionization mass spectrometry (EASI-MS) and
electrospray ionization mass spectrometry (ESI-MS) [191] are often used to analyse the non-volatile compounds.

Here it is proposed a complete workflow for ToF-SIMS characterisation of coffee, from sample preparation, through spectra acquisition, to data analysis. Samples were produced before and after brewing in water following different procedures that were tested systematically and the MVA methodologies developed for this thesis were applied to characterise samples based on brewing methods and their provenance.

11.3 Methodology

This section contains a detailed explanation of the methodology that has been developed in order to achieve good results in the characterisation of coffee. It consists of three major parts: sample preparation, ToF-SIMS data acquisition and data analysis. Before having the definitive sample preparation methods described here, a few attempts were made until an optimal and systematic set of procedures was established. Data from natural and biological samples have great variability inherently, therefore a thorough ToF-SIMS data acquisition was done in order to avoid more sources of error. Data with spectral resolution, $m/\Delta m$, lower than 5,000 (@ 29 Da) was discarded. Finally, an effective characterisation methodology that depends on multivariate analysis needs appropriate data analysis software and simsMVA was used for this purpose.
### 11.3.1 Samples provenance and preparation

The as-received samples are ground roast coffee from four different farms of Sao Paulo state in Brazil. The granules are stored in plastic bags therefore contamination is expected. For all four kinds of coffee, four different sets of samples were prepared following the systematic procedures described as follows:

i) **Compact ground coffee:** a sample of each ground coffee was compressed into a small hole (2 mm diameter, 5 mm deep) of a stainless-steel stub using a spatula. The stub and spatula were previously washed with IPA and acetone.

ii – iv) **Dilution followed by thin film deposition:** these procedures consisted of four main steps. First the ground coffee was diluted in deionized water with 200 mg of coffee for 8 mL of water. The water temperature varied in order to mimic the resulting solution of different brewing methods: ii) *cold brew:* water at room temperature and brewing for 24 hours; iii) *hot brew:* water at 92 °C and brewing for 10 minutes; iv) *wrong brew:* water at room temperature and brewing for 10 minutes. Individual beakers and pipettes were used for each type of coffee to avoid cross contamination. After brewing, two drops of the solution were dropped onto clean pieces of silicon wafer. The samples were then covered with aluminium foil and taken to an oven at 100 °C for 20 minutes. Amongst other possible solvents, water was the obvious choice regardless of all the difficulties with its evaporation. Diluting the coffee in water will result in samples that contain what is consumed when drinking brewed coffee, which makes such analysis much more relevant. Figure 11.1 shows 3 photographs of the ToF-SIMS top mounted sample holder with the solution drops before and after evaporation.
11.3.2 ToF-SIMS analysis

The ToF-SIMS analysis was carried out using the TOF.SIMS 5 system at the University of Surrey using 25 keV Bi$_3^+$ ion beam operated in the high current bunched mode delivering 0.3 pA, pulsed and rastered over areas of 50 x 50 µm$^2$. Two to four different regions were analysed for each ground coffee and film sample. Positive secondary ion spectra were acquired for 30 seconds each, keeping the total primary ion dose below the static limit of $10^{13}$ ions/cm$^2$/analysis. For the film samples, the analysed areas were always in the centre of the drop to avoid inhomogeneous regions. For the ground coffee
samples, a pulsed low energy (21 eV) electron flood gun was used to eliminate charge building up on the samples surfaces during the analyses.

11.3.3 Spectra calibration and peak picking

All spectra were calibrated using one unique mass list with a large range of masses that provided enough calibration points to achieve accurate peak assignments. A high spectral resolution peak list was created using The SurfaceLab software (ION-TOF GmbH) following the recommendations described in Chapter 4. As a first approach, the list contained peak areas of 86 ion fragments which were selected based on molecules that are reported to relate to brewed coffee flavour [182]. In some cases, these peaks are not the most intense of the spectrum, but nevertheless they still presented multivariate variance that enabled the characterisation of the coffee samples, as shown in the Results & Discussion section. Spectra of the silicon wafer substrate and of a drop of pure water were also acquired and the peaks that showed to be dominant in those samples were excluded from the peak list as they would hinder the characterisation of coffee by adding non-relevant directions of variance. Table 11.1 contains all the ions considered for multivariate analysis.

Table 11.1: ToF-SIMS ions considered for multivariate analysis. Their flavour description is based on reference [182].

<table>
<thead>
<tr>
<th>Mass</th>
<th>Assignment</th>
<th>Flavour</th>
</tr>
</thead>
<tbody>
<tr>
<td>46.996</td>
<td>CH$_3$S+</td>
<td>Cooked potato</td>
</tr>
<tr>
<td>47.980</td>
<td>CH$_4$S+</td>
<td>Cooked potato</td>
</tr>
<tr>
<td>48.992</td>
<td>CH$_5$S+</td>
<td>Cooked potato</td>
</tr>
<tr>
<td>Mass</td>
<td>Formula</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>57.035</td>
<td>C₃H₅O⁺</td>
<td>Roasted/fruity</td>
</tr>
<tr>
<td>58.041</td>
<td>C₃H₆O⁺</td>
<td>Roasted/fruity</td>
</tr>
<tr>
<td>59.051</td>
<td>C₃H₇O⁺</td>
<td>Roasted/fruity</td>
</tr>
<tr>
<td>71.053</td>
<td>C₄H₇O⁺</td>
<td>Malty/fruity</td>
</tr>
<tr>
<td>72.049</td>
<td>C₄H₈O⁺</td>
<td>Malty/fruity</td>
</tr>
<tr>
<td>85.033</td>
<td>C₄H₈O₂⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>85.072</td>
<td>C₃H₆O⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>86.037</td>
<td>C₄H₉O⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>86.064</td>
<td>C₃H₁₀O⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>87.051</td>
<td>C₄H₁₀O₂⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>87.080</td>
<td>C₃H₁₁O⁺⁺</td>
<td>Buttery</td>
</tr>
<tr>
<td>99.051</td>
<td>C₄H₇O₂⁺</td>
<td>Oily-buttery</td>
</tr>
<tr>
<td>99.088</td>
<td>C₄H₁₁O⁺⁺</td>
<td>Butter rancid</td>
</tr>
<tr>
<td>100.048</td>
<td>C₅H₈O₂⁺</td>
<td>Oily-buttery</td>
</tr>
<tr>
<td>100.082</td>
<td>C₅H₁₂O⁺</td>
<td>Butter rancid</td>
</tr>
<tr>
<td>100.996</td>
<td>C₅H₁₃O⁺</td>
<td>Smoke/Roasted</td>
</tr>
<tr>
<td>101.034</td>
<td>C₅H₁₄S⁺</td>
<td>Smoke/Roasted</td>
</tr>
<tr>
<td>101.112</td>
<td>C₅H₁₅O⁺</td>
<td>Butter rancid</td>
</tr>
<tr>
<td>102.045</td>
<td>C₅H₁₁O₂⁺</td>
<td>Smoke/Roasted</td>
</tr>
<tr>
<td>103.051</td>
<td>C₅H₁₁S⁺</td>
<td>Smoke/Roasted</td>
</tr>
<tr>
<td>104.060</td>
<td>C₅H₁₂S⁺</td>
<td>Smoke/Roasted</td>
</tr>
<tr>
<td>105.038</td>
<td>C₅H₁₃O⁺</td>
<td>Cooked potato</td>
</tr>
<tr>
<td>107.052</td>
<td>C₅H₁₃N₂⁺⁺</td>
<td>Hazelnut/Roasted/Peanuts</td>
</tr>
<tr>
<td>108.056</td>
<td>C₅H₁₄N₂⁺⁺</td>
<td>Hazelnut/Roasted/Peanuts</td>
</tr>
<tr>
<td>109.068</td>
<td>C₅H₁₅N₂⁺⁺</td>
<td>Hazelnut/Roasted/Peanuts</td>
</tr>
<tr>
<td>111.050</td>
<td>C₅H₁₆O₂⁺⁺</td>
<td>Sweet/caramel</td>
</tr>
<tr>
<td>112.051</td>
<td>C₅H₁₆O₂⁺⁺</td>
<td>Sweet/caramel</td>
</tr>
<tr>
<td>113.002</td>
<td>C₅H₁₂O⁺</td>
<td>Smoke/roasted</td>
</tr>
<tr>
<td>113.069</td>
<td>C₅H₁₃O₂⁺</td>
<td>Sweet/caramel</td>
</tr>
<tr>
<td>114.015</td>
<td>C₅H₁₄O⁺</td>
<td>Smoke/roasted</td>
</tr>
<tr>
<td>115.023</td>
<td>C₅H₁₅O⁺</td>
<td>Smoke/roasted</td>
</tr>
<tr>
<td>117.035</td>
<td>C₅H₁₆O⁺</td>
<td>Meat</td>
</tr>
<tr>
<td>119.057</td>
<td>C₅H₁₂SO⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>120.052</td>
<td>C₅H₁₂SO⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>121.070</td>
<td>C₅H₁₃SO⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>122.065</td>
<td>C₅H₁₀N₂⁺⁺</td>
<td>Peanuts/Roasted</td>
</tr>
<tr>
<td>123.048</td>
<td>C₅H₁₄O₂⁺⁺</td>
<td>Phenolic/Roasted</td>
</tr>
<tr>
<td>Mass</td>
<td>Molecular Formula</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>123.089</td>
<td>C₇H₁₁N₂⁺</td>
<td>Peanuts/Roasted</td>
</tr>
<tr>
<td>124.046</td>
<td>C₇H₉O₂⁺</td>
<td>Phenolic/Roasted</td>
</tr>
<tr>
<td>125.067</td>
<td>C₇H₉O₂⁺</td>
<td>Phenolic/Roasted</td>
</tr>
<tr>
<td>125.107</td>
<td>C₈H₁₃O⁺</td>
<td>Mushroom</td>
</tr>
<tr>
<td>126.098</td>
<td>C₈H₁₃O⁺</td>
<td>Mushroom</td>
</tr>
<tr>
<td>127.048</td>
<td>C₈H₁₅O⁺</td>
<td>Caramel/Sweet</td>
</tr>
<tr>
<td>127.119</td>
<td>C₈H₁₅O⁺</td>
<td>Mushroom</td>
</tr>
<tr>
<td>128.029</td>
<td>C₈H₁₅O⁺</td>
<td>Caramel/Sweet</td>
</tr>
<tr>
<td>128.058</td>
<td>C₈H₁₅O⁺</td>
<td>Spicy</td>
</tr>
<tr>
<td>129.033</td>
<td>C₈H₁₅O⁺</td>
<td>Caramel/Sweet</td>
</tr>
<tr>
<td>129.066</td>
<td>C₈H₁₅O⁺</td>
<td>Spicy</td>
</tr>
<tr>
<td>130.068</td>
<td>C₈H₁₉N⁺</td>
<td>Coconut</td>
</tr>
<tr>
<td>131.050</td>
<td>C₈H₁₉N⁺</td>
<td>Coconut</td>
</tr>
<tr>
<td>131.086</td>
<td>C₈H₁₉N⁺</td>
<td>Coconut</td>
</tr>
<tr>
<td>133.063</td>
<td>C₉H₁₁N⁺</td>
<td>Coconut</td>
</tr>
<tr>
<td>133.105</td>
<td>C₉H₁₁N⁺</td>
<td>Coconut</td>
</tr>
<tr>
<td>134.069</td>
<td>C₈H₁₉N₂⁺</td>
<td>Earth/Hazelnut/Roasted</td>
</tr>
<tr>
<td>135.085</td>
<td>C₉H₁₁N₂⁺</td>
<td>Earth/Hazelnut/Roasted</td>
</tr>
<tr>
<td>140.118</td>
<td>C₉H₁₆O⁺</td>
<td></td>
</tr>
<tr>
<td>141.064</td>
<td>C₉H₁₅O⁺</td>
<td>Spicy/Sweet/Caramel</td>
</tr>
<tr>
<td>142.068</td>
<td>C₉H₁₆O⁺</td>
<td>Spicy/Sweet/Caramel</td>
</tr>
<tr>
<td>143.084</td>
<td>C₉H₁₁O₂⁺</td>
<td>Spicy/Sweet/Caramel</td>
</tr>
<tr>
<td>147.079</td>
<td>C₉H₁₁SO₂⁺</td>
<td>Cat/Green/cassis</td>
</tr>
<tr>
<td>148.086</td>
<td>C₉H₁₂SO₂⁺</td>
<td>Cat/Green/cassis</td>
</tr>
<tr>
<td>149.069</td>
<td>C₉H₁₃SO₂⁺</td>
<td>Cat/Green/cassis</td>
</tr>
<tr>
<td>149.123</td>
<td>C₉H₁₃N₂⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>150.070</td>
<td>C₉H₁₃O₂⁺</td>
<td>Cravo</td>
</tr>
<tr>
<td>150.114</td>
<td>C₉H₁₃N₂⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>151.078</td>
<td>C₉H₁₁O₃⁺</td>
<td>Cravo</td>
</tr>
<tr>
<td>151.118</td>
<td>C₉H₁₃N₂⁺</td>
<td>Hazelnut/Roasted</td>
</tr>
<tr>
<td>152.055</td>
<td>C₉H₁₃O₃⁺</td>
<td></td>
</tr>
<tr>
<td>153.054</td>
<td>C₉H₁₃O₃⁺</td>
<td></td>
</tr>
<tr>
<td>155.041</td>
<td>C₉H₁₁O₃⁺</td>
<td>Caramel</td>
</tr>
<tr>
<td>155.082</td>
<td>C₉H₁₁O₃⁺</td>
<td>Caramel</td>
</tr>
<tr>
<td>156.046</td>
<td>C₉H₁₃O₃⁺</td>
<td>Caramel</td>
</tr>
<tr>
<td>156.084</td>
<td>C₉H₁₃O₃⁺</td>
<td>Caramel</td>
</tr>
<tr>
<td>157.066</td>
<td>C₉H₁₃O₃⁺</td>
<td>Caramel</td>
</tr>
</tbody>
</table>
### 11.3.4 Multivariate data analysis

The peak areas for all samples were exported into a .txt file and loaded into *simsMVA* using the *Spectra mode* of analysis. Once the data is loaded, the main window shows a table with the areas of all peaks (in rows) for each sample (in columns). It is possible to change the labels of the peaks and also assign IDs to peaks that may have something in common (such as be daughter fragments of the same molecule). For the coffee samples, the IDs of the ions were established based on the flavour properties of their parent molecule, according to [182]. Another useful feature is the easy pre-assignment of groups to the data points. All repeats of a single sample can be grouped based on names defined by the user and these are going to be plotted with the same symbol and colour for all subsequent analysis. This can be very useful when it comes to separate the data by sample preparation method or coffee origin (see Figures 11.3 and 11.4). The *simsMVA* software also enables quick pre-processing of ToF-SIMS data prior to multivariate analysis. The data pre-processing steps employed were normalisation by total counts, Poisson scaling and mean centring.
11.4 Results & Discussion

The influence of sample preparation methods showed greater variance across the data and therefore hindered the differentiation of the samples based on their farms. For this reason, different sets of data were processed: the first one containing all measurements and the other ones containing subsets of data for each sample preparation method.

11.4.1 Influence of sample preparation methods on data quality

Figure 11.2 shows overview spectra of all four kinds of coffee analysed and each sample preparation method with mass ranges 0 – 700 u for ground coffee and 0 – 1000 u for brewed coffee. Such difference in mass range is because an increase of the instrument cycle time for the brewed coffee samples, that showed information in the high mass range. As mentioned previously, for the acquisition of the spectra of ground coffee, the samples were thick enough to act as insulators, which required an electron flood gun to be used for charge compensation. This and the high roughness of the samples contributed to degradation of the spectral resolution if data were acquired in certain regions, which required a careful search of suitable regions and made the data acquisition process for these samples more time consuming. All spectra of the brewed samples are extremely similar at first glance, yielding always the same peaks regardless of the kind or brewing method. The differences among the spectra rely on relative intensities of their peaks, which can be better assessed if multivariate analysis is employed.
Figure 11.2: Spectra of all four kinds of coffee and sample preparation method.
11.4.2 Differentiation of preparation methods

Figure 11.3 shows a scatter plot of the scores of principal components (PC) 1 and 2 alongside their loadings plots of the dataset containing all samples. The symbols are categorised by sample preparation method. There are two categories for the room temperature brewed samples because those were prepared and measured on different days and the fact that both showed very similar scores establishes that the method is reproducible. To aid visualisation, a $k$-means clustering algorithm was applied to the scores of the first 3 principal components and the resulting groups are shown as the background colours of the plot of Figure 11.4 (a): PC 1 separates the ground samples from the dried film samples while PC 2 separate the hot and cold brewed samples from the room temperature brewed and granules samples and therefore can be considered as a scale for the “level of brewing”. Properly brewed samples presented high positive scores of PC 2 and the positive loadings of PC 2 contain the peaks related to “roasted” flavours, which are usually the most notable taste in “over brewed” coffee.
11.4.3 Classification of coffee samples by their origin

Figures 11.4 and 11.5 show scatter plots of the scores of the PCs that better separated the samples in terms of their origins (farms 1 to 4). Next to each scatter plot there is also a web plot that allows the visualisation of the averaged scores of each group for 5 principal components (the shaded area represents negative scores). The web plot is
useful to show that each group has its own “fingerprint” in terms of their principal components scores, which can be a first step towards building a classification model. Amongst the four sample preparation methods, the results for the “hot brew” samples showed the best discrimination of the farms. More data needs to be acquired in order to confirm and quantify such discrimination.

Granules

![Granules Diagram]

Hot brew

![Hot brew Diagram]

**Figure 11.4:** See caption of Figure 11.5.
Cold brew

Wrong brew

Figure 11.5: For each sample preparation method: left - scatter plot of the scores of the PCs that better separated the samples in terms of their origins. Right - web plot of averaged scores of each group for 5 PCs (the shaded area represents negative scores).
11.5 Conclusion & Future work

This Chapter reported the successful development of a methodology for the characterisation of coffee samples using ToF-SIMS which contributed towards the understanding of the brewing process and now enables more experiments to be carried out. One important feature of the simsMVA software was developed specifically for the analysis of the coffee data: the use of $k$-means clustering to visualise groups of samples in the scores scatter plots. This shows the importance of using different applications to develop MVA methodology and, moreover, gave a go-ahead sign for the development of multivariate classification to be incorporated into simsMVA. This was part of a collaboration with University of Surrey researcher Dr. Lucio Rosa in which the next step will use more data from more samples to increase accuracy. Initial tests using a $k$-nearest neighbours classifier have been shown to be promising.
12 Analysis of Coatings Formed by Polyester Resin Blends

12.1 Introduction

This Chapter presents an MVA-based data analysis extension to a well-established experimental methodology for the assessment of organic coatings using ToF-SIMS. In addition to microtome-based sample preparation steps of tapers for the analysis through sample thickness, standard samples in cured and uncured conditions are introduced and analysed in the same manner as the specimens under investigation. Such an approach yields results that can be analysed using a multivariate analysis procedure that enables the simultaneous processing of spatially and chemically related datasets.
12.2 Characterisation of polymer blends

There are many areas of scientific and technological endeavour where it is advantageous to use blends or mixtures of polymeric materials. In some instances blends are formed as a result of complete miscibility of the two or more polymeric phases, but in many cases this is not the case and phase separation of the component materials occurs following the mixing procedures [192]. If the polymers are easily identifiable (such as PMMA and PVC, for example) segregation can be identified by standard chemical analysis methods (such as FT-IRS), microbeam methods (such as energy dispersive X-Ray analysis in a scanning electron microscope) or surface specific chemical analysis methods such as X-ray photoelectron spectroscopy (XPS) or ToF-SIMS [193–198]. ToF-SIMS, of course, has much to commend it as an analytical technique for this purpose as it provides surface mass spectra with molecular specificity at an extremely good detection limit. However, ToF-SIMS becomes difficult to apply with a high degree of certainty should the polymers comprising the blend be extremely similar in chemical composition but still possess a tendency to phase segregate. The work described in this Chapter deals with such a case. A paint has been formulated with two fully pigmented, but extremely similar, commercial polyester resins, both of which are crosslinked with the same curing agent (hexamethoxy methyl melamine – HMMM). The ToF-SIMS spectra of each of the commercial resins are extremely similar and there are no peaks of sufficient intensity that are uniquely diagnostic of either material to provide an unambiguous identification of each using imaging. In order to resolve the problem, ToF-SIMS data has been processed using MVA methods, in particular NMF, in an effort to identify phase separation in the cured coating.
The coatings investigated in this work are approximately 20 µm in thickness and rather than prepared as cross-sections of the paints on their substrates, a novel taper-sectioning method has been employed. The group at the University of Surrey introduced a methodology over a decade ago and which is now a very well established way of investigating buried polymer/polymer and polymer/metal interfaces in organic coatings [199, 200]; the exposure of cross-sections by creating tapers using ultra-low-angle microtomy (ULAM). In contrast to the more usual ToF-SIMS approach of dual-beam depth profiling, one of the advantages of using ULAM is the massive extension of the possible depths of analysis that can be achieved. ToF-SIMS analysis of ULAM exposed cross-sections are usually done in imaging mode [201, 202] and in some cases the areas comprising all interfaces of interest will be larger than the maximum primary ion beam raster size, which demands the use of large area methodologies such as stage rastering. Moreover, in order to truly evaluate interfacial properties, one must prepare and measure more than one taper for each coating. All this results in several very large ToF-SIMS datasets that are potentially very similar.

This Chapter proposes a methodology for row-wise matrix augmentation (introduced in Chapter 4 and referred to here as “map stitching”) in combination with standards samples for the study of four large area datasets of exposed interfaces in a mixed organic system. The organic systems under consideration are blends of two fully formulated and pigmented paint systems based upon commercial polyester resins.
12.3 Experimental

12.3.1 Sample preparation

The polyester resin films were prepared by industrial partner Becker Industrial Coil Coatings. A resin mixture containing the commercial polyester resins and HMMM cross-linking agent, together with usual pigmentation and additives was stirred together and applied onto chromium (VI) pre-treated Al and Zn substrates using a spiral bar coater. The coated panels were then cured at 232°C for 45 seconds (Zn substrate) and 237°C for 35 seconds (Al substrate) in a hot air oven. The final coating thicknesses of the films were within the range 20 – 22 µm. The formulation description of all the specimens are detailed in Table 12.1.

Ultra-low angle microtomy (ULAM) was employed to expose the interfacial regions “buried” around 20-25 µm below the true top surface of the coating, making it available for subsequent analysis with ToF-SIMS. Specimens of the coatings were cut using an industrial guillotine from the ~20 cm x 10 cm Zn and Al panels to ~1 cm² square specimens. Following a well-established methodology described by Hinder et al [199, 200], the squares were then mictrotomed at an angle of 2° using a Microme HM355S motorized rotary microtome (Optech Scientific Instruments, Thame, UK) fitted with a tungsten carbide knife. For the analysis of these specific coatings, the angle chosen was bigger than the usual so that both top surface/primer and primer/substrate interfaces could be revealed in a single taper (see Figure 12.1). The resulting sample is tilted by 2° in the vacuum chamber before ToF-SIMS analysis to ensure the surface is
normal to the extraction optics of the time-of-flight mass spectrometer, making it suitable for ToF-SIMS analysis. The two polyester resins that compose the blend yield extremely similar secondary ions spectra, therefore, in order to aid the spatial differentiation of the two incompatible resins forming the blend, standard pure samples of both polyester resins (in both the cured and uncured conditions) were prepared on zinc substrates, following a similar application procedure used for the mixed polyester coating films. The use of standard samples to achieve relative quantification of ToF-SIMS data through multivariate analysis has been shown to be effective [203, 204].

Figure 12.1: Schematic of sample preparation using ULAM. Two main interfaces are exposed on the obtained tapers.
12.3.2 ToF-SIMS analyses

ToF-SIMS analyses were carried out on the TOF.SIMS 5 equipment at the University of Surrey using a 25 keV Bi$_3^+$ primary ion beam delivering 0.3 pA in the high current bunched mode with 100 µs cycle time. Positively charged secondary ions were analysed. For the pure standard samples, the beam was rastered over 500 x 500 µm$^2$ regions producing secondary ions images of 128 x 128 pixels. This is consistent with the spot size of the Bi$_3^+$ beam in the high-current bunched mode (ca. 3.5 µm) which is used to achieve high spectral resolution, at a resolution, m/Δm on $^{29}$Si$^+$ peak, of ca. 8000. The whole area comprising both interfaces was too irregular (both in terms of charging and topography) to be analysed using an automated stage raster. Instead, six patches of 500 x 500 µm$^2$ (128 x 128 pixels) each were measured with charge compensation optimisation prior to each measurement, resulting in a total analysed area of 1 x 1.5 mm$^2$ per sample. A low energy, 20 eV, electron flood was used to neutralise charge build-up on the sample surface. In order to ensure homogeneous charging, all samples were mounted on glass slides. All datasets were stitched together subsequently as described in next Section.
12.4 Data analysis

12.4.1 Peak picking and exporting
The channel to mass scales of the spectra of each standard sample and all large area patches were calibrated using the same calibration points. A peak list containing 669 peaks was then created using the SurfaceLab software (IONTOF GmbH) from automated peak search routines on the pure spectra of the two polyester resins. Data cubes containing the maps of the areas of all selected peaks were then exported for each 500 x 500 µm² scan (see schematic in Figure 12.2). The export format chosen was the binary file BIF6.

12.4.2 Loading data into analysis software
The BIF6 files were loaded into simsMVA using the stitch function which creates a single matrix containing all (stitched) patches and the standard sample images, for all the samples described in Table 12.1. The first row of Figure 12.2 shows a schematic with the stitching procedure and the resulting map for the peak at 149 u (C₈H₅O₃⁺). simsMVA facilitates this methodology which has been used before [85]. Furthermore, having the maps stitched enables the entire dataset to be processed as a single matrix, which is a detail of great importance for the spatial differentiation of the two resins as will be shown in the Results Section. The entire dataset of 24 patches plus 4 standard samples has a total of 393,216 pixels x 669 peaks.
Figure 12.2: Schematic of the data stitching procedure followed by multivariate analysis. The letters correspond to samples described in Table 12.1
12.4.3 Multivariate analysis

Once loaded into simsMVA, the images could be further processed. Prior to any MVA, two pre-processing steps were performed: normalisation of all maps intensities by total counts per pixel and Poisson scaling of the peak intensities. Those steps are easily done using the drop-down menus of the main window of the app. After a series of tests, the chosen multivariate analysis workflow was:

i) Selection of a subset of pixels using low discrepancy subsampling as proposed in [96]. The computer used was a Dell Optiplex 9020 PC with a core i5 processor and 32 GB of RAM. Such amount of memory could easily handle the 393,216 x 669 elements of the dataset but the use of training sets enables the analysis to be done in any conventional PC and it has shown to be effective for multivariate analysis of ToF-SIMS imaging data, in which neighbour pixels are highly correlated [25, 29, 96]. Furthermore, with reduced subsets the total processing time is conveniently reduced to a few seconds for the dataset presented in this paper. The percentage of pixels to be subsampled can be set using one of the drop-down menus of simsMVA.

ii) PCA of the subsampled data in order to set the number of components for NMF. The number of components for NMF was decided based on the variance per principal component, a priori knowledge of the samples and maps of PCA scores.

iii) NMF of the subsampled dataset. For the application of NMF to ToF-SIMS images, the subsampled data is unfolded into a Matrix $A_s$ containing samples (pixels) in rows and variables (spectral channels) in columns. The second and
third rows of Figure 12.2 show a schematic of the dataset unfolding and factorisation. The algorithm used to achieve non-negative factorisation was the multiplicative update rule described in Chapter 3.

iv) Calculation of the full matrix \( W \) using the original unfolded dataset \( A \) and the pseudo inverse of matrix \( H_s \): \( W = A \times H_s / (H_s^T \times H_s) \).

12.5 Results

Figure 12.3 presents NMF results of the stitched dataset. Optimal factorisation was achieved after 500 iterations. The image on the left-hand side of Figure 12.3 is an overlay of all components (matrix \( W \)) with one colour attributed to each component. The labels A-D and S₁-S₄ correspond to the different samples as described in Figure 12.2 and Table 12.1. The intensity of each component determines the transparency of its colour across the pixels. The plots on the right-hand side of Figure 12.3 are the characteristic spectra of each component (matrix \( H \)). The results demonstrate the ability of NMF analysis in the identification and visualisation of the interfacial layers across the bulk surfaces of both the pigmented and the un-pigmented coating specimens under comparison. In the overlay maps, it is possible to observe that component 1 (in red) is very intense at the cured top surface of the coatings across the samples set. This is consistent with the images representative of the cured standard surfaces \( S_1 \) and \( S_3 \). The spatial distributions shown by the intensity maps of components 3 and 6 (green and pink, respectively) discriminates the immiscible polyester resins 1 and 2 across the bulk of all four coating specimens. They are also identified accordingly at the uncured standard
surfaces $S_2$ and $S_4$. This enables prompt phase identification and also brings insights into the interfacial chemistry across the bulk. Component 2 (in yellow) is attributed to the inorganic pigments dispersed in the resin layers of coatings C and D, where it can be seen that the pigmentation seems to have agglomerated at the interfacial region between polyesters 1 and 2. Component 4 corresponds to the Al substrates of coatings A & D. Component 5 corresponds to the Zn substrates of Coatings B and C.

Figure 12.3: Phase identification using NMF analysis. The image on the left-hand side is an overlay of all components (matrix $W$) with one colour attributed to each component. The plots on the right-hand side are the characteristic spectra of each component (matrix $H$). The Labels $S_1$ to $S_4$ and A-D refer to table 12.1.
12.6 Discussion

12.6.1 Effect of curing
In the NMF results, the surface of the cured reference samples matches the top surface of all coatings while the uncured standard samples are the ones that enable differentiation of the two polyesters across the layers within the ULAM tapers. A straightforward conclusion for this would be that the cross-linked polyesters become indistinguishable by ToF-SIMS and the layers buried underneath it are not completely cured hence why they show similarities to the uncured standard samples. However, one cannot discard the possibility of something else being on top of all surfaces that undergo curing (such as, for example, oven contamination or the surface segregation of minor components). ToF-SIMS is highly surface sensitive and in this case, all that is being measured at the top surfaces is whatever is covering them, leaving only the uncured standard samples with distinguishable features. One way to find out whether this is only a surface effect of cured resins could be to cut the top surface of the standard samples and investigate their bulk with ToF-SIMS.

12.6.2 Immiscibility of polyesters
It is clearly shown in the results that the two immiscible polyester resins are well separated across the bulk of the four specimens and MVA enabled the identification of interfacial phases without prior knowledge of the system. Two different morphologies were identified and differentiated by NMF analysis across the bulk of the un-pigmented
coating specimens (A and B). In both coatings A and B, polyester 2 is concentrated at the top coat interface. However, coating B presents discrete globular structures throughout the bulk towards the Zn interface, while in coating A polyester 2 forms a less concentrated, irregular web-like shape. The pigmented coatings (C and D) presented similar behaviour but polyester 2 is now concentrated at both interfaces, with smaller globules formed across the bulk. There are two factors that may have influenced those observed differences, the first being the presence of pigments that seem to follow the distribution of polyester 2 (even around the globules). The second possible factor is the curing process: regardless of the pigments, the phase distribution of polyesters 1 and 2 seems to represent “snapshots” of different stages of their separation (or mixing), with coating A showing an initial stage where both resins are still disputing space and coating D showing a final layers structure. This may mean that the curing process needs more control if a more reproducible manufacture procedure is desired. In addition, it is clear that the substrate type (aluminium or hot dipped galvanised steel - HDGS) has a secondary role to play in the phase separation of the polyester blend. Comparing the unpigmented blends (A and B of Figure 12.3), the extent of phase separation is not the same, the dispersed phase (polyester 2) having a significantly larger size in the case of the HDGS substrate. Turning now to the pigmented system a similar occurrence is revealed, in this case of the HDGS substrate a fine dispersion of polyester 2 is evident in the middle (polyester 1) layer in addition to the two adjacent layers of polyester 2 either side of the inner layer. In the case of the aluminium substrate only the three-layer “sandwich” structure is observed with little, if any, phase separation of polyester 2 in the well define layer of polyester 1. The reason for this is unclear but it is thought to be
unlikely the specific interaction between substrate and resin, which would occur over very short distances. More likely is that the difference in thermal mass of the two substrates is responsible. Zinc and low carbon steel have values of thermal conductivity significantly below that of aluminium (about one half and one quarter respectively) thus the coatings deposited on aluminium benefit from a higher input of thermal energy to aid phase segregation thus the fine phase distribution on these substrates is better developed than the coarser distributions on the HDGS substrate. As phase separation develops so, eventually, a three-layer structure develops.

12.6.3 A step further: k-means clustering

NMF results allow a complete description of a very complex dataset with a much reduced number of variables. However, if one wants to quickly classify samples with the aid of standards, or in the case of this work, identify spatially where each polyester is, a final step of the multivariate analysis can be the application of a k-means clustering algorithm to the W matrix of the NMF results. For the purpose of comparison, k-means was also applied to the PCA scores matrix. The k-means algorithm is described in Chapter 3 and was applied using the correlation distance \( d(P,Q) \) defined for two points, \( P \) and \( Q \) as:

\[
d(P,Q) = 1 - \frac{(P - \bar{P})(Q - \bar{Q})}{\sqrt{(P - \bar{P})^2\sqrt{(Q - \bar{Q})^2}}} = 1 - \frac{\sum_{i=1}^{n}(p_i - \bar{P})(q_i - \bar{Q})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{P})^2} \sqrt{\sum_{i=1}^{n}(q_i - \bar{Q})^2}}
\]

Where \( \bar{P} \) and \( \bar{Q} \) are the component-wise mean of each point. It is important to apply k-means clustering after NMF or PCA in a space with reduced dimension because the
distances between different points become indistinguishable when many coordinates are used, this is known as the *curse of dimensionality*. Figure 12.4 and 12.5 present $k$-means results with 4 clusters. Figure 12.4 shows the images with the categories attributed to each pixel and Figure 12.5 shows a matrix with scatter plots of all components intensities against each other (every point represents a pixel) for both NMF+$k$-means and PCA+$k$-means analysis. One colour is attributed to each category (or cluster). The standard samples at the bottom row present homogeneous squares which are used for the spatial identification of the resin components in the tapers. In the NMF+$k$-means resulting image, the first and second category separate the two polyesters, the third category matches the surface of both cured standard samples to the surface of all coatings, which indicates as well as the NMF results that a curing effect common to every sample which undergoes curing is being detected. The fourth category is shared by the two different metal substrates (aluminium and zinc). The PCA+$k$-means could not identify the resins as well as the previous approach, with the first and second categories separating the resins only in terms of the presence of pigmentation. This result persisted even when the analysis was done with more clusters or more principal components, which shows that the use of NMF+$k$-means was more effective for the purpose of this work.
In comparison to the NMF results of Figure 12.3, the k-means clustering results present a more definitive classification and the row of standard samples at the bottom work as labels, since they are homogeneous and end up having all pixels with the same cluster-classification. The downside of using k-means are the loss of spectral information with the degree of purity achieved by NMF results. However, spectra can still be reconstructed by adding together the original data of all pixels in a single cluster.
Figure 12.5: Plot matrix containing scatter plots of components intensities against each other. (A): NMF components. (B): PCA components. Only components used for k-means clustering are shown.
12.7 Conclusion & Future work

The proposed methodology successfully identified the spatial distribution of two very similar polyester resins across tapers of 4 different coatings. As future work, it is intended to investigate the effects of pigmentation and the details of the chemistry that happens along the coatings production, especially the cross-linking process.

This methodology expands to imaging datasets the usage of data of standard materials presented in Chapter 10 and was the main reason for the development of the “stitching” functions in simsMVA, which is in fact now being used for different research projects that are described in Appendices 2 and 3.
PART IV

DISCUSSION

The previous Chapters presented the successful development of MVA methodologies for complex ToF-SIMS datasets and a series of applications to materials science. The next two Chapters discuss the impact of such methodologies and their potential implications for future work.
13 The Impact of the Methodologies

13.1 Practical impact on materials science research

The practical impact of the methodologies developed for this thesis are thoroughly illustrated in Part III (Chapters 9 to 12) with the analysis of diverse datasets from a variety of samples. These are products of experimental research carried out by the author in The Surface Analysis Laboratory at the University of Surrey.

The uncompressed data processing methodology employed in Chapter 9 was a direct application of the methodology proposed in Chapter 7 and is an excellent example of cases where the only option for achieving meaningful results is by performing MVA of uncompressed spectra. The results have successfully simplified a very large and complex ToF-SIMS dataset of atmospheric plasma treated polypropylene samples, with results retaining the high spectral resolution, which was essential for the characterisation of the oxidised and reactivity zones of the plasma. Furthermore, the exact same methodology for large area imaging data, in combination with the map stitching
approach taken in Chapter 12, was used to characterise the removal of hydrocarbon species from the surface of atmospheric plasma treated stainless steel samples. Appendix 2.5 briefly presents NMF results of two stitched large area imaging datasets of samples treated with an atmospheric plasma with and without incorporation of deuterated water.

The “map stitching” methodology and the inclusion of measurements of standard samples has, in fact, been applied to many datasets other than the polyester-based coatings presented in Chapter 12 and of which results are briefly described in Appendices 2.1 and 2.2. These include the identification of minerals developed at the surface of a lithium-based corrosion protective layer by stitching large area ToF-SIMS imaging datasets of corroded samples together with two standard minerals and the comparison of 3D depth profiling for the identification of buried metal-organic interfaces using two different etching ion beams (C$_{60}$ and Ar$_{1200}$). Another application is part of a project that seeks to detect metabolized drugs on fingerprints using ToF-SIMS (Appendix 2.4). For the effective characterisation of the fingerprints, a combination the approaches presented in Chapters 7, 10 and 12 of this thesis had to be employed. These were the alignment of peaks across spectra of different pixels in a stitched large area imaging dataset followed by NMF without compression, aided by measurements of standard samples of the drugs potentially present in the fingerprints.

With regards to the analysis of spectra datasets, the careful selection of peak lists has proven to be effective for the analysis of very complex samples that yield numerous secondary ion fragments. This is illustrated in the analysis of wood in Chapter
10 and in the analysis of corrosion products in Appendix 2.1, where, in both cases, NMF was used to quantify relative amounts of mixed components in a set of complex samples by using a reduced list of variables containing only characteristic fragments of the “pure” materials. Reduced peak lists were also used for the characterisation of flavour molecules of coffee in Chapter 11 and the identification of products and reactants in an ozone-treated polyurethane coating, in Appendix 2.7.

Most of these results were published in peer-reviewed journals, and presented at international conferences (see Appendix 1), which unquestionably establishes their utility. However, in the context of this thesis, these applications, together with most applications present in Appendices 2 and 3 had vital importance for the development of all the methodologies and the simsMVA software, since most of the analysis and visualisation functions were developed with the aim to solve the problems faced when analysing these samples. This shows that the development of simsMVA was not a solution in search of a problem but instead a very versatile tool with key features created with the purpose of solving specific problems. A clear advantage of simsMVA over the usual “black box” approach of MVA (previously applied by The Surface and Interface Reactions Group at the University of Surrey), is the responsive manner in which requests and suggestions for modifications of the software can be carried out very promptly.

13.2 Fundamental impact on method development

Apart from having successfully expanded the applicability of MVA methods to complex ToF-SIMS datasets, this thesis also contributed more fundamentally on method
development with the work in Chapter 7 that presented, for the first time, NMF of uncompressed ToF-SIMS data and all challenges related to it. The publication of these results (see Appendix 1) triggered discussions that ended up in collaborations, the results of which results are presented in Chapter 8, carried out with the aim of understanding the limits of the applicability of conventional NMF algorithms and pixel subsampling in the context of large ToF-SIMS datasets. The relevance of this is reflected in the number of citations that the research has received and also in the direct application of it which is presented in Chapter 9 for the analysis of plasma treated polypropylene samples. In fact, the first attempt on the analysis of uncompressed datasets was on a treated polypropylene dataset that had a very complex surface chemistry, which hindered the progress on experimental issues such as charge compensation and peak alignment in large area datasets. This led to the author taking a step back and first investigating a simpler sample, which was the contaminated fingerprint deposited on a silicon wafer substrate, presented in Chapter 7. simsMVA has also been successfully applied to other analytical techniques such as FTIRS, XPS and PIXE as indicated in Appendix 3.

13.3 Recommendations for achieving meaningful results
simsMVA is able to analyse datasets using all the methodologies described in this thesis, however, reliable and meaningful results are only achieved if the user inputs suitable data and take the appropriate analysis steps. These steps can be summarised as a series
of recommendations starting from data acquisition and ending on post processing and visualisation of MVA results.

1) In order to effectively take the most information out of a complex ToF-SIMS dataset, one must, prior to any data processing, acquire good quality data. This is only achieved with careful configuration of acquisition parameters that best suit a specific sample under analysis. For the TOF.SIMS 5 at the University of Surrey, these include adjusting the reflectron voltage; fine-tuning flood gun parameters to have effective charge compensation in insulating samples; choosing the appropriate mode of operation of the primary ion gun (with the compromise between lateral and mass resolution in mind); using appropriate primary ion beam raster mode (a random raster pattern is more effective for insulating samples as the induced charge build-up will not be concentrated in a specific region, facilitating charge compensation by the electron flood gun); selection of appropriate number of pixels that will guarantee that the primary beam will not probe the same spot twice on a sample surface; maximising count rate by focusing the primary beam and using appropriate settings for the secondary ion extracting, focusing and post-acceleration; ensuring static conditions by monitoring primary ion beam current and acquisition time.

2) For a flat, homogeneous and conductive sample, such analysis conditions will not deviate much from the standard and some quantitative criteria can be used to assess data quality. The first two criteria used at The Surface Analysis Laboratory are, for
a positive polarity total area spectrum, the width, in nanoseconds, of the $H^+$ peak, that should be under 1 ns; and the $m/\Delta m$ mass resolution at a peak on nominal mass 29, that should be over 7,000. Further criteria include checking homogeneity in a total secondary ion map, total area spectrum background and presence of peaks saturating the detector. However, as shown in this thesis, some samples can be rather complex and in such case the analyst must make sure that the source of data quality degradation is not from instrumental parameters but instead inevitable sample aspects such as topography, differential charging or non-homogeneity. These sample aspects are, most of the time, a source of greater variance than the actual surface chemistry, therefore, they will be picked up by MVA methods and skew the identification of differences in chemistry within a dataset, potentially outputting misleading results such as the NMF components obtained before peak alignment presented in Chapters 7 and 9.

3) With regards to the creation of peak lists, sometimes the spectra of different samples will significantly differ in terms of their spectral resolution. These could be pixels in areas containing different materials in an imaging dataset, levels of different layers in a depth profile or even completely different samples in a spectra dataset. In all cases, one must ensure that the integrated mass range used to create the variables spans both the high and low-resolution peaks.

4) Finally, a practice that has been shown and illustrated multiple times throughout this thesis, is the inclusion of data of standard materials in situations where the analyst
seeks to achieve relative quantification. Data of standard materials are likely to have many correlated variables and at high intensity, this sets the main directions of variance within the entire dataset, which results in PCA or NMF components that are representative of quantities of such “pure” materials.

13.3.1 Initial steps in simsMVA

If a dataset is carefully generated following the steps described in this Section, simsMVA is more prospective to provide meaningful results. As described in Chapter 5, there are many ways to explore a dataset using the software, however, some practices are recommended for any data under analysis.

1) As it is usual to acquire repeats of the same sample when analysing spectra, the first thing to be considered after loading a dataset in the Spectra mode of simsMVA is the assignment of “groups”. This is done by clicking on the “Set Groups” button, which enables the user to rename each observation so that the ones given the same name will be plotted using the same symbol and colour. The advantages of such approach were illustrated in Chapter 4.

2) Imaging and 3D datasets are usually very large and the calculation of their MVA results can take long. Therefore, even if the computer has enough memory to process all pixels/voxels, it is recommended to first perform MVA in a subsampled dataset in order to check whether the selected peak list or parameters such as number of
NMF components are appropriate for the dataset. This practice saves a considerable amount of time and enables the user to test various pre-processing and input parameters before settling for a definitive MVA approach.

3) In depth profiling datasets it is very common to have the first few levels presenting very different chemistry from the remaining layer structure. This is sometimes a very thin top layer indeed but most of times it is just an adventitious contamination layer with light molecules such as polydimethylsiloxane or atmospheric hydrocarbons. Such layers may hinder MVA as they will have high intensity in the initial levels and take up PCA or NMF components. For this reason, it is recommended to perform the intended MVA method both including and excluding those layers. This can be done by clicking on the “Set Range” button that enables the user to select an interval of levels to be considered for MVA.

4) Once meaningful MVA is achieved, simsMVA can be used to further explore the results by means of the post-processing and visualisation methods presented in Chapters 4 and 5.
14 Conclusions & Perspectives

14.1 Conclusions

The work described in this thesis has successfully expanded the applicability of MVA to large and complex ToF-SIMS datasets. Such expansion was achieved by first reaching an excellent level in the experimental practice of ToF-SIMS, which enabled the production of good quality results from challenging samples. The complexity of these results required the creation and validation of novel methodologies that were encompassed in simsMVA, fulfilling the need for a versatile software that can be applied and modified to solve various materials science and industrial problems.

Apart from the software and all output from this and related research, another important outcome of this thesis is the legacy that will be left with the research group at the University of Surrey. All methodologies were established to a point that is now part of the routine analysis procedure carried out by all group members using ToF-SIMS.
14.2 The success of simsMVA

From its early days of development, simsMVA has always been used by research group members other than the author of this thesis. Essentially with colleagues acting as beta testers. The feedback received from colleagues throughout the years helped to increase, even more, the flexibility of the software which is mainly what attracted the attention of new external users. Thus, there are two factors that measure the success of simsMVA. The first is research outputs of the author of this thesis himself or through collaborations with other researchers. Such publications are listed in Table 14.1 and details of all collaborative research that used simsMVA are presented in Appendices 2 and 3. The second factor is the number of users that requested the software via website and are doing independent research using simsMVA, following its initial release in September 2017 at the International SIMS conference in Krakow. There is a total of 25 active users and these are researchers working in all parts of the world as shown in Figure 14.1.

Table 14.1: Research output that used simsMVA.

<table>
<thead>
<tr>
<th>Research output</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master of Engineering (MEng) dissertations</td>
<td>3</td>
</tr>
<tr>
<td>Doctor of Philosophy (PhD) theses</td>
<td>6</td>
</tr>
<tr>
<td>Research Papers</td>
<td>9</td>
</tr>
<tr>
<td>Posters</td>
<td>5</td>
</tr>
<tr>
<td>Conference presentations</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>35</strong></td>
</tr>
</tbody>
</table>

278
Figure 14.1: Countries with simsMVA users indicated in red.
14.3 Future perspectives

Some of the methodologies employed in this thesis have already inspired other researchers who used similar approaches to analyse their data (some of it is presented in Appendices 2 and 3). This creates the expectations that more pieces of research will follow the same path, including any future work by the author, which will ultimately result in the improvement of the methodologies.

The release of simsMVA to colleagues and external users created a feedback channel that constantly improves the applicability of simsMVA, which will continue to be developed further.

Another branch of this work that is expected to be further developed in the following years is the analysis of very large (often uncompressed) datasets. Most of the research has so far dealt with standard datasets (such as the one in Chapter 7) or had some sort of data compression (such as the results in Chapter 9), but the recent release of the OrbiSIMS shows how the next generation of high-end spectrometers will look, with extremely large datasets being generated. This means that very complex datasets will be generated and this will demand methods that handle very large datasets in more practical ways.

With regards to standardisation, NPL has done an excellent job on nomenclature and guidelines but perhaps it is time to consider the creation of standard samples for MVA testing of software such as simsMVA or any other, including in-house developed routines. This would bring even more confidence to researchers in both industry and academia and reliability to published results.
REFERENCES


The Development of Multivariate Analysis Methodologies for Complex ToF-SIMS Datasets: Applications to Materials Science


THE DEVELOPMENT OF MULTIVARIATE ANALYSIS METHODOLOGIES FOR COMPLEX TOF-SIMS DATASETS: APPLICATIONS TO MATERIALS SCIENCE


[53] D.J. Griffiths, C. Inglefield, Introduction to Electrodynamics, Am. J. Phys. 73


THE DEVELOPMENT OF MULTIVARIATE ANALYSIS METHODOLOGIES FOR COMPLEX TOF-SIMS DATASETS: APPLICATIONS TO MATERIALS SCIENCE


REFERENCES


THE DEVELOPMENT OF MULTIVARIATE ANALYSIS METHODOLOGIES FOR COMPLEX TOF-SIMS DATASETS: APPLICATIONS TO MATERIALS SCIENCE

doi:10.1021/ac3005346.


[183] Ö. Özdestan, S.M. van Ruth, M. Alewijn, A. Koot, A. Romano, L. Cappellin, F.


[193] T. Kono, E. Iwase, Y. Kanamori, TOF-SIMS analysis of polystyrene/polybutadiene blend using chemical derivatization and multivariate


[207] M.J. Bailey, S. Francese, Rapid detection of cocaine, benzoylecgonine and
methylecgonine in fingerprints using surface mass spectrometry, 140 (2015).
APPENDIX 1: PUBLISHED WORK AND CONFERENCE PRESENTATIONS

The following lists contain all research output that is somehow related to this thesis. All presented posters and published research papers are attached in the annex volume 2 of this thesis and some of the slides of the oral presentations can be found online on the website: https://www.slideshare.net/GustavoFerrazTrindad

PUBLICATIONS

1. **Trindade, G. F.; Bañuls-Ciscar, J.; Ezeh, C. K.; Abel, M.; Watts, J. F.** Characterisation of wood growth regions by multivariate analysis of ToF-SIMS data. Surface and Interface Analysis (Print), V.48, P.N/A - N/A, 2015.


**PRESENTATIONS AT CONFERENCES**


- **Trindade, G.F.;** Banuls-Ciscar, J.; Ezeh, C.; Abel, M.; Watts, J. F.
  *Characterisation of Wood Growth Regions by Multivariate Analysis of ToF-SIMS Data*,
  ECASIA 16, Granada, Spain (2015).

- **Trindade, G. F.;** Abel, M.; Watts, J. F.
  *Processing Large ToF-SIMS Datasets for the Study of Surface Segregation of Polymer Additives*,
  CAC 2016, Barcelona, Spain (2016).

- **Trindade, G. F.;** Abel, M.; Watts, J. F.
  *Processing Large ToF-SIMS Datasets*,
  SIMS Europe 2016, Muenster, Germany (2016).

- (INVITED) **Trindade, G.F.**
  *MVA methodologies for surface analysis data*,
  UKSAF summer meeting, Teignmouth, UK (2017).

- **Trindade, G.F.;** Abel, M.; Watts, J. F.
  *simsMVA: A Matlab tool for multivariate analysis of ToF-SIMS datasets*,
  SIMS 21, Krakow, Poland (2017).

- **Trindade, G.F.;** Williams D. F.; Abel, M.; Watts, J. F.
  *Analysis of atmospheric plasma treated polypropylene by large area ToF-SIMS imaging and NMF*,
  ECASIA 17, Montpellier, France (2017).

- **Trindade, G. F.;** Rosa, L. F. S.; Abel, Stori, E. M.; Santos, C. E. L.; Watts, J. F.
  *Characterisation of coffee by ToF-SIMS*,
  ECASIA 17, Montpellier, France (2017).
APPENDIX 2: FURTHER APPLICATIONS TO TOF-SIMS DATA

The next two Appendices present summaries of a series of studies that were mostly not led by the author of this thesis but made use of the MVA methodologies described in earlier Chapters. Appendix 2 contains studies that applied the methodologies to ToF-SIMS data and Appendix 3 the application onto different analytical techniques. Some of the studies resulted in publications of which the author of this thesis is a co-author.

A2.1 Lithium-based corrosion protective layer

As part of a collaboration with Kristof Marcoen from Vrije Universiteit in Brussels, Belgium, simsMVA was used for the analysis of ToF-SIMS data obtained from a Lithium-based corrosion protective layer after various different times exposed to corrosive solutions [204]. The analysis methodology followed exactly the same one proposed in Chapter 10 of this thesis for the analysis of wood samples. NMF was applied to data from corroded samples and measurements of pure materials were added in order to achieve relative quantification. A summary of the results is shown in Figure A2.1.
Figure A2.1: NMF results from ToF-SIMS data of Li-based coatings exposed to corrosive solutions for different lengths of time. Figures from reference [204].

The second part of the collaboration extended the analysis to a set of large area imaging datasets and used the stitching functions of simsMVA to analyse the same series of measurements and investigate the spatial distribution of the pure components identified in the first part. This was again based on a methodology proposed in this thesis (Chapter 12), which is the NMF analysis of stitched imaging datasets aided by images of pure samples. Initial results are shown in Figure A2.2.
Figure A2.2: Spatial distribution of 2 NMF components obtained from data of Li-based coatings exposed to corrosive solutions. The last two columns are measurements of standard materials.

A2.2 Analysis of polyurethane coatings

A recent PhD project by Dr. Jorge Banuls-Ciscar at the University of Surrey was based on the understanding of interactions between polyurethanes, wood and steel. Part of the project involved 3D dual beam depth profiling ToF-SIMS characterisation of polyurethane coatings deposited onto metallic substrates and the datasets generated were of great importance for the development of the 3D mode of simsMVA. NMF analysis
aided the identification of organic-metal interfacial SIMS fragments that were used for further understanding of adhesion phenomena [175]. Figure A2.3 shows the NMF intensities of 3 components found in one of the 3D datasets.

![Figure A2.3: Intensities of NMF components obtained from a 3D ToF-SIMS dataset of a polyurethane coating deposited onto a metallic substrate. Figure from reference [175].](image)

The same project also used the “stitching” tool of simsMVA in order to analyse, as a single matrix, data from depth profiling of a coating obtained with two different etching beams: \( \text{C}_{60}^+ \) and \( \text{Ar}_{1200}^+ \). Results (Figure A2.4) showed that apart from different etch rates, interfacial chemistry could be successfully identified using either of the etching beams [175].
Appendix 2: Further Applications to ToF-SIMS Data

Figure A2.4: simsMVA datasets stitching procedure applied to data of two different 3D ToF-SIMS depth profiles [175].

Similar investigation has been done as part of the PhD project of Dr. Sabrina Tardio where several ToF-SIMS spectra of thin films of polyurethane monomers of different thicknesses deposited onto metallic substrates were measured and simsMVA/PCA was
applied to the resulting dataset in order to identify secondary ions characteristic of the
organic-metal interface [205].

A2.3 Analysis of parchment

Several samples of parchment were analysed with the TOF.SIMS 5 instrument as part
of a MSc project of Vladmir Vilde carried out at the University of Surrey. The
application of the simsMVA routines helped to identify characteristic secondary ions
from parchments made from different animal sources (goat flesh, goat skin, sheep flesh
and sheep skin) [206]. Figure A2.5 shows a scores plot of PC 1 x PC 2 for a dataset with
samples of different parchments. More details can be found in reference [206].

![Figure A2.5: Principal components scores scatter plot (PC 1 vs. PC 2) for a ToF-
SIMS dataset of parchment samples. Figure from [206].](image)
A2.4 Drug detection on fingerprints

The group of forensic chemistry at the University of Surrey has a project that seeks to detect metabolized drugs on fingerprints using several techniques including ToF-SIMS [120, 122, 207]. In collaboration with PhD student Min Jang, several very complex large area images were acquired and initial tests were done using the maps stitching functions and adding measurements of reference drugs, which is the methodology proposed in Chapter 12. This is a great challenge as sometimes the drugs will be present at very low concentration levels and initial tests showed to be promising. Figure A2.6 shows a screenshot of simsMVA with PCA results of a matrix-augmented dataset containing four large area mappings of fingerprints.

Figure A2.6: Screenshot of simsMVA with PCA results of four large area ToF-SIMS datasets of fingerprints.
A2.5 Plasma treated stainless steel

The EngD project of Dr. David Williams dealt with atmospheric plasma treatment of stainless steel and part of the work involved large area ToF-SIMS imaging of treated surfaces using humid plasma. [160]. The ToF-SIMS experiments were in fact carried out by the author of this thesis and served as a basis for the subsequent experiments on plasma-treated polypropylene samples which are presented in Chapter 9 of this thesis. Two samples were prepared by treating their surface with humid plasma containing both water (H₂O) and heavy water (D₂O) and a matrix-augmented dataset was created (following methodology described in Chapters 4 and 12) and processed with NMF, the results are shown in Figures A2.7 and A2.8. The approach of processing both datasets as a single matrix was shown to be advantageous as 4 components showed exactly the same distribution for both samples and 1 component seems to be present only when the surface is treated with heavy water.
Figure A2.7: Overlay of NMF results obtained by processing both humid plasma-treated steel datasets as a single matrix. Top: D$_2$O; Bottom: H$_2$O.
Figure A2.8: 5 NMF components obtained by processing both humid plasma-treated steel datasets as a single matrix. Top: D₂O; Bottom: H₂O. Field of view of each dataset: 2 x 2 cm².

**A2.6 ToF-SIMS/MVA database of commercial resins**

The PhD project of Dr. Rene Tshulu at the University of Surrey involved the characterisation of many commercial polyester resins. simsMVA was used to perform both PCA and NMF on ToF-SIMS datasets which enabled the classification of such resins based on their polymer repeat units [208]. Figure A2.9 shows the PCA results for a set of standard resins. This database was then used for the understanding of more complex formulations such as the ones presented in Chapter 12 of this thesis.
Figure A2.9: PCA scores obtained from a dataset of ToF-SIMS of standard resins [208].

A2.7 Study of degradation of aircraft topcoats

The PhD project of Dr. Taraneh Bozorgzad Moguim at the University of Surrey investigated the degradation of aircraft topcoats by exposing samples to UV-ozone environments for different periods of time [209]. Early versions of simsMVA have been used to analyse ToF-SIMS data of these samples and generated important feedback for the development of additional features. One example is related to the pigmentation present in the coatings which yielded very intense secondary ion signals from inorganic atoms, skewing the dataset and making PCA only accounts for variance due to pigmentation signal. This resulted in the creation of the “mass filter” tool in the spectra mode of simsMVA. The tool makes the removal of inorganic (or organic) peaks automated and practical by excluding peaks based on a mass value threshold (typically values under the nominal mass will represent inorganic peaks). After removal of inorganic peaks, the first principal component of the resulting dataset separated organic
reactants from products, which enabled important insights about the degradation mechanism of the coatings. Figure A2.10 shows the PCA results after removal of inorganic masses from the dataset.

Figure A2.10 PCA results of ToF-SIMS data of an aircraft topcoat exposed to UV/ozone for different periods of time [209].
A2.8 Effects on ion bombardment of polymers

One of the aims of the EngD project of Dr. Robin Simpson at the University of Surrey was to understand the effects caused by bombardment of organic materials with both monoatomic and cluster argon ion beams. simsMVA was applied to a dataset created from a series of ToF-SIMS measurements of a standard PMMA material before and after etching with both types of beams. Although it is often said in the literature that cluster bombardment does not impinge damage to the material, PCA of the data showed that there is difference on the spectra of the surface after etching (Figure A2.11) [210].

![Figure A2.11: PCA scores of ToF-SIMS data of PMMA samples bombarded with monoatomic and cluster Argon beams][210].
APPENDIX 3: FURTHER APPLICATIONS TO OTHER ANALYTICAL TECHNIQUES

A3.1 FTIR of polyesters

As a follow-up investigation to the results presented in Chapter 12, the industrial collaborator provided the group at University of Surrey with Fourier-transform infrared (FTIR) spectra of standard polyester resins. The data was loaded into simsMVA and PCA/NMF in conjunction to classification models managed to achieve relative concentrations in of pure compounds in mixtures following a similar procedure to the one presented in Chapter 12 (Figures A3.1 and A3.2).

![Figure A3.1: PCA scores of FTIR dataset of standard polyester resins.](image-url)
A3.2 XPS depth profiles

XPS depth profiling datasets were acquired for the same samples of polyurethane samples mentioned in Appendix 2 using cluster argon beams at the University of Surrey. NMF was applied to C 1s scans using all spectral channels as variables and each depth level as samples. Unlike ToF-SIMS NMF results, for XPS, matrix $H$ is not visualised with plots of “pure components” but instead the column products $w_i h_i$ for each component $i$ are overlaid together with the original signal (Figure A3.3). Such approach was used as an “unbiased peak fitting estimator” for each level that, for the polyurethane datasets, enabled the identification of peak characteristic of the organic-metal interface, as shown in the results of Figure A3.4.
Figure A3.3: Schematic for visualisation of NMF results of XPS depth profiling data.
Figure A3.4: NMF results of dataset created from C 1s XPS depth profile of polyurethane thin film onto metal substrate [175].

Another project in collaboration with Dr. Jose Ferreira from Federal University of Bahia in Brazil is related to time-resolved datasets with XPS data of the dicarboxylic acids presented in Chapter 6. Samples of the acid were subject to continuous X-ray irradiation and C 1s XPS spectra were recorded periodically in order to investigate the effects of X-ray induced degradation of the acids. The next step is to apply NMF to the resulting datasets in a similar manner to that described for time-resolved datasets in Chapter 4 of this thesis.
A3.3 PIXE of coffee

The four kinds of coffee presented and analysed in Chapter 11 were also analysed by particle induced X-ray emission (PIXE) at Federal University of Rio Grande do Sul in Brazil. The raw spectra were loaded in simsva so that each spectral channel is regarded as a variable. Initial results showed that PCA can potentially be used to classify the samples based on their PIXE spectra (Figure A3.5). Next step it the application of NMF followed by correction factors that will enable peak fitting and quantification of NMF components.

Figure A3.5: PCA results of PIXE data from four coffee samples. Left: Scores scatter plot of PC 2 vs. PC 3. Right: Loadings of PC 2 and PC 3.
A3.4 Large area PIXE mapping

In collaboration with Dr. Tiago Silva and Prof. Marcia Rizzuto from the University of Sao Paulo in Brazil, MVA was applied to large area PIXE scans of a decorative tile from a 19th century building [74]. The novelty in the approach for this data was the use of column-wise matrix augmentation, which joined data from three different X-ray detectors as a single set of variables. The NMF results shown in Figure A3.6 enabled the identification of characteristic PIXE spectra for the different pigments used to paint the tile.

Figure A3.6: NMF results of large area PIXE imaging data of a decorative tile.

The dotted lines in the characteristic spectra separate data obtained from different X-ray detectors. Figure reproduced from [74].