Applications of Pattern Recognition Entropy (PRE) and Informatics to Data Analysis

Shiladitya Chatterjee
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Chemistry Commons

BYU ScholarsArchive Citation
Chatterjee, Shiladitya, "Applications of Pattern Recognition Entropy (PRE) and Informatics to Data Analysis" (2019). Theses and Dissertations. 8826.
https://scholarsarchive.byu.edu/etd/8826

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Applications of Pattern Recognition Entropy (PRE) and Informatics to Data Analysis

Shiladitya Chatterjee

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Matthew Richard Linford, Chair
Daniel E. Austin
David V. Dearden
Jaron C. Hansen
Barry M. Lunt

Department of Chemistry and Biochemistry
Brigham Young University

Copyright © 2019 Shiladitya Chatterjee
All Rights Reserved
ABSTRACT

Applications of Pattern Recognition Entropy (PRE) and Informatics to Data Analysis

Shiladitya Chatterjee
Department of Chemistry and Biochemistry, BYU
Doctor of Philosophy

The primary focus of my work is the application of informatics methods to the fields of materials science and analytical chemistry. The statistical analysis of data has become increasingly important in understanding the properties of materials and analytes. Statistical methods like principal component analysis (PCA) and multivariate curve resolution (MCR) are widely used for analysis in chemistry and other fields given their ability to categorize spectra in an unsupervised way. PCA is relatively easy to apply and has appealing mathematical properties. However, the results can be challenging to interpret, even for experienced users. In contrast, MCR results can be more interpretable, because the factors resemble real spectra and do not have negative scores or loadings. Nevertheless, the useful orthogonality properties of the scores and loadings in PCA are sacrificed in doing so. Other statistical analysis methods like cluster analysis and partial least squares regression (PLS-R) present their own challenges. Pattern recognition entropy (PRE) is a novel application of Shannon’s information theory for understanding the underlying complexity in spectra. Unlike PCA and MCR, PRE is a summary statistic that adopts the mathematical quantification of information and applies it for chemometric analysis. PRE values reflect the shape and complexity of spectra. Chapter 1 contains a description of the analytical methods/instruments that provided the data I analyzed by PRE and other informatics tools, including (i) X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) and (ii) liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis (CE), (iii) a discussion of some of the commonly used statistical analysis tools like PCA, MCR, cluster analysis and PLS-R, and (iv) a description of PRE. Chapter 2 describes in much greater detail the theory associated with the statistical tools I used and PRE. Chapter 3 describes the PRE and informatics analysis of depth profiles through thin films by XPS and ToF-SIMS. Chapter 4 introduces the concept of the ‘reordered spectrum’ as an intuitive, visual representation of spectra to address the abstraction associated with PRE result. Total ion current chromatograms (TICCs) generated using LC-MS are often extremely complex and ‘noisy’. Chapter 5 describes the application of PRE as a variable reduction method for producing higher quality TICCs. Chapter 6 discusses the limitations associated with the application of PRE to TICCs and presents a new method using cross-correlation (CC) in conjunction with a PRE analysis. Chapter 7 discusses a new methodology that uses CE and PRE to detect autologous blood doping (ABD). Chapter 8 presents my conclusions of this present work and discusses the scope of future work on PRE. The thesis also contains several appendices. Appendix 1 introduces polyallylamine (PAAm) as a simple, easy-to-apply adhesion promoter for the widely used photoresist SU-8. Appendices 2, 3 and 4 contain articles I wrote that relate to trends in modern XPS instrumentation and 5-8 contain supplemental information relating to Chapters 3, 4, 5, and 7 respectively.

Keywords: informatics, chemometrics, entropy, PRE, XPS, ToF-SIMS, LC-MS, CE, thin films, surface characterization, PCA, MCR
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents, Mr. Tapash Kumar and Mrs. Sumita Chatterjee. The endless mentoring and encouragement of my father and the emotional strength, guidance and sacrifice of my mother have carried me through life to where I am today. I would like to dedicate this dissertation to them.

I would like to thank my paternal (Mr. Anil Kumar Chatterjee and Mrs. Rama Chatterjee) and maternal (Mr. Madhab Chandra Chakraborty and Mrs. Anima Chakraborty) grandparents, because without their blessings and support, I would never have had the chance to be here.

I would like to thank my uncles (Prof. Binay Bhushan Chakrabarti and Mr. Dulal Kumar Chatterjee), because their constant support and counsel has contributed to the man I am today.

I would like to thank my mentor from school, Mr. Ranjan Sarkar, who has been one of the most influential teachers in my life and has left a lasting mark that continues to inspire me to this day.

I would like to thank my undergraduate assistants, George H. Major and Sean C. Chapman, for their work and dedication that made my publications possible. I would also like to thank my lab mates for their help.

Furthermore, I would like to thank my committee members: David V. Dearden, Daniel E. Austin, Jaron C. Hansen and Barry M. Lunt for providing their valuable suggestions during my graduate work.

I am deeply thankful to the Department of Chemistry and Biochemistry of Brigham Young University for providing me the platform for my success. Also, I would like to thank the Roland K. Robins Research Fellowship, the Albert D. and Jennie R. Swenson Graduate Fellowship Award,
the Nicholes-Maw Award, Plant Therapy Inc. and Moxtek Inc. for supporting and funding my research.

Finally, I would like to thank my Ph.D. advisor Dr. Matthew Linford. Dr. Linford has been an amazing mentor the entire time I have been here. He saw my potential and my skills and guided me through my research. Everything I am and everything I have achieved is because he recognized my potential and molded me into an analytical chemist.

Last, but, not the least, I would like to thank The Almighty for blessing me through this journey.
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Page</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>Chapter 1: Overview of XPS, ToF-SIMS, LC-MS, CE, Various Informatics Methods, and PRE</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 X-ray Photoelectron Spectroscopy (XPS)</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Time-of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS)</td>
<td>11</td>
</tr>
<tr>
<td>1.4 Liquid Chromatography-Mass Spectrometry (LC-MS)</td>
<td>14</td>
</tr>
<tr>
<td>1.5 Capillary Electrophoresis (CE)</td>
<td>15</td>
</tr>
<tr>
<td>1.6 Chemometric Methods</td>
<td>17</td>
</tr>
<tr>
<td>1.6.1 Principal Component Analysis (PCA)</td>
<td>18</td>
</tr>
<tr>
<td>1.6.2 Multivariate Curve Resolution (MCR)</td>
<td>19</td>
</tr>
<tr>
<td>1.6.3 Cluster Analysis</td>
<td>20</td>
</tr>
<tr>
<td>1.6.4 Partial Least Squares Regression (PLS-R)</td>
<td>20</td>
</tr>
<tr>
<td>1.7 Pattern Recognition Entropy (PRE)</td>
<td>21</td>
</tr>
<tr>
<td>1.8 References</td>
<td>22</td>
</tr>
<tr>
<td>Chapter 2: Theory of Chemometric Methods and Pattern Recognition Entropy (PRE)</td>
<td>30</td>
</tr>
<tr>
<td>2.1 Principal Component Analysis (PCA)</td>
<td>31</td>
</tr>
<tr>
<td>2.2 Multivariate Curve Resolution (MCR)</td>
<td>34</td>
</tr>
<tr>
<td>2.3 Cluster Analysis</td>
<td>40</td>
</tr>
<tr>
<td>2.4 Partial Least Squares Regression (PLSR)</td>
<td>43</td>
</tr>
<tr>
<td>2.5 Pattern Recognition Entropy, PRE</td>
<td>45</td>
</tr>
<tr>
<td>2.6 References</td>
<td>51</td>
</tr>
<tr>
<td>Chapter 3: A Perspective on Two Chemometrics Tools: PCA and MCR, and Introduction of a New One: Pattern Recognition Entropy (PRE), as Applied to XPS and ToF-SIMS Depth Profiles of Organic and Inorganic Materials</td>
<td>53</td>
</tr>
</tbody>
</table>
5.3.1 LC-MS ........................................................................................................................ 125
5.3.2 Spline Fit to the Data .................................................................................................. 126
5.3.3 Savitzky-Golay Filter ................................................................................................. 127
5.4 Results and Discussion ................................................................................................... 128
5.5 Conclusion ......................................................................................................................... 136
5.6 Acknowledgement ............................................................................................................. 137
5.7 Supporting Information ..................................................................................................... 137
5.8 References ......................................................................................................................... 137

Chapter 6: Using Cross-Correlation with Pattern Recognition Entropy to Obtain Reduced Total Ion Current Chromatograms from Raw Liquid Chromatography-Mass Spectrometry Data ..... 140
6.1 Introduction ....................................................................................................................... 140
6.2 Theory ............................................................................................................................... 142
6.3 Experimental ..................................................................................................................... 144
6.4 Results ............................................................................................................................... 145
6.5 Conclusions ....................................................................................................................... 154
6.6 References ......................................................................................................................... 155

Chapter 7: Informatics Analysis of Capillary Electropherograms of Autologously Doped and Undoped Blood ........................................................................................................................... 157
7.1 Introduction ....................................................................................................................... 157
7.2 Experimental ..................................................................................................................... 161
7.2.1 Sample Preparation and Data Collection ...................................................................... 161
7.2.3 Computations and Data Analysis ................................................................................. 162
7.3 Theory ............................................................................................................................... 163
7.3.1 Pattern Recognition Entropy (PRE) ........................................................................... 163
7.3.2 Euclidean Distance ($d_{Eu}$) ........................................................................................ 164
7.3.3 Peak Fit, Integration (PFI) ......................................................................................... 165
7.3.4 Second Moment (SM) ................................................................................................. 165
7.3.5 Inverse Least Squares (ILS) ....................................................................................... 168
7.3.6 Preprocessing .............................................................................................................. 168
7.4 Results and Discussion ..................................................................................................... 170
7.4.1 Traditional Analyses: Cluster Analysis, PCA, and PLS ............................................. 170
7.4.2 Analysis by Less Traditional Tools: PRE, the Euclidean Distance, Peak Fit - Integration and the Second Moment ........................................................................................................... 176
7.4.3 Inverse Least Squares ........................................................................................................ 180
7.5 Conclusions .......................................................................................................................... 183
7.6 Acknowledgments .................................................................................................................. 184
7.7 Supporting Information ........................................................................................................ 184
7.8 References ............................................................................................................................ 184
Chapter 8: Conclusion of My Present Work ............................................................................... 187
8.1 Future Work .......................................................................................................................... 190
Appendix 1: Polyallylamine as an Adhesion Promoter for SU-8 Photoresist ................. 191
A1.1 Introduction ......................................................................................................................... 191
A1.2 Materials and Methods ...................................................................................................... 196
  A1.2.1 Materials .................................................................................................................... 196
  A1.2.2 Surface Characterization .......................................................................................... 198
A1.3 Results and Discussion ..................................................................................................... 198
  A1.3.1 SU-8 Deposition and Removal on PAAm-Coated Surfaces..................................... 198
  A1.3.2 Reduction in Defects During Spin Coating of SU-8 .................................................. 205
  A1.3.3 Refinement of the SE Model .................................................................................... 206
A1.4 Conclusion .......................................................................................................................... 209
A1.5 Acknowledgments ................................................................................................................. 210
A1.5 References .......................................................................................................................... 210
Appendix 2: Trends in Advanced XPS Instrumentation. 2. Angle-Resolved XPS (AR-XPS) and XPS Surface Mapping ................................................................................................................. 213
A2.1 Discussion .......................................................................................................................... 213
A2.2 References ........................................................................................................................ 223
Appendix 3: Trends in Advanced XPS Instrumentation. 4. A Modern XPS Instrument and How to Manufacture It in Just Four Days ................................................................. 225
A3.1 Introduction ........................................................................................................................ 225
A3.2 Overview of the Thermo Scientific K-Alpha+ System .................................................... 228
A3.3 Analyses with the K-Alpha System .................................................................................... 235
  A3.3.1 Analysis of Contact Lenses ..................................................................................... 235
  A3.3.2 Mapping the Work Function of a Damaged Solar Cell ............................................ 236
A3.4 Manufacturing the K-Alpha ............................................................................................ 238
  A3.4.1 Just in Time Manufacturing ..................................................................................... 238
  A3.4.2 Quality Control ........................................................................................................ 239
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPS</td>
<td>X-ray photoelectron microscopy</td>
</tr>
<tr>
<td>ToF-SIMS</td>
<td>Time-of-flight secondary ion mass spectrometry</td>
</tr>
<tr>
<td>LC-MS</td>
<td>Liquid chromatography-mass spectrometry</td>
</tr>
<tr>
<td>CE</td>
<td>Capillary electrophoresis</td>
</tr>
<tr>
<td>PRE</td>
<td>Pattern recognition entropy</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>MCR</td>
<td>Multivariate curve resolution</td>
</tr>
<tr>
<td>PLS-R</td>
<td>Partial least squares regression</td>
</tr>
<tr>
<td>AES</td>
<td>Auger electron spectroscopy</td>
</tr>
<tr>
<td>LEIS</td>
<td>Low energy ion scattering</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1.1. A schematic representation of an X-ray photoelectron spectrometer (XPS)............ 11
Figure 1.2. Simple schematic representation of a time-of-flight secondary ion mass spectrometer (ToF-SIMS) ................................................................................................................................. 13
Figure 1.3. A block diagram representing the major components in a liquid chromatography-mass spectrometry (LC-MS) system .................................................................................................................. 15
Figure 1.4. A block diagram of the capillary electrophoresis (CE) technique showing key components ............................................................................................................................................. 16
Figure 1.5. The motion of a ball down an incline showing the typical Cartesian coordinate system (axes $x_1$ and $x_2$) and the more computationally convenient, rotated coordinate system (axes $x_1'$ and $x_2'$). .......................................................................................................................... 18
Figure 2.1. A measured ToF-SIMS peak (black) represented as the sum of two pure analyte spectra (blue and red). ............................................................................................................................................... 36
Figure 2.2. A dendrogram illustrating results from a cluster analysis of 5 objects. .................... 42
Figure 2.3. Shannon entropies from coins with a range of probable outcomes. This plot can be interpreted as follows. Left: an unfair coin that never yields a desired outcome ($p = 0$). Middle: a fair coin that has even odds of heads or tails ($p = ½$). Right: an unfair coin that always produces the desired outcome ($p = 1$, right). ........................................................................................................ 47
Figure 2.4. Simulated spectra and their PRE values. ................................................................... 50
Figure 3.1. Three-dimensional graph of 24 positive ion ToF-SIMS spectra from a depth profile through a film of plasma polymerized C3F6 on Si. .......................................................................................... 59
Figure 3.2. Three-dimensional graph of 24 positive ion ToF-SIMS spectra from a depth profile through a film of plasma polymerized PNIPAM on Si .................................................................................. 60
Figure 3.3. Structure of N-isopropylacrylamide. ......................................................................... 61
Figure 3.4. Three-dimensional graph of the 54 Si 2p narrow scans used in this study from an XPS depth profile of SiO2 on Si. ...................................................................................................................... 62
Figure 3.5. Three-dimensional graph of the 54 O 1s narrow scans used in this study from an XPS depth profile of SiO2 on Si. ...................................................................................................................... 63
Figure 3.6 Three-dimensional graph of the 37 Ta 4f narrow scans used in this study from an XPS depth profile of Ta2O5 on Ta .................................................................................................................. 63
Figure 3.7. Three-dimensional graph of the 37 O 1s narrow scans used in this study from an XPS depth profile of Ta2O5 on Ta. .................................................................................................................. 64
Figure 3.8. PCs 1 and 2 of a five PC PCA model of ToF-SIMS spectra from a depth profile through a film of plasma polymerized C3F6 on a silicon substrate. (a) The scores plot as a function of depth index showing the scores on PC 1 and PC 2, and the loadings on (b) PC 1, and (c) PC 2 as a function of the mass (u). .................................................................................................................. 67
Figure 3.9. PCs 3 – 5 of a five PC PCA model of the spectra from a ToF-SIMS depth profile of plasma polymerized C3F6 on a silicon substrate. (a) The scores plots for PCs 3 – 5 as a function of depth index, and the loadings on (b) PC 3, (c) PC 4, and (d) PC 5 as a function of the mass (u). .................................................................................................................. 68
Figure 3.10. MCR components 1 – 3 from a five factor MCR model of a ToF-SIMS depth profile through a film of plasma polymerized C3F6 on Si. (a) Scores on MCR components 1 – 3
as a function of depth index, and the loadings on (b) Component 1, (c) Component 2, and (d) Component 3 as a function of the mass (u). ................................................................. 70

**Figure 3.11.** Components 4 and 5 of a five factor MCR model of ToF-SIMS data from a depth profile through a film of plasma polymerized C$_3$F$_6$ on Si. (a) Scores on MCR components 4 and 5 as a function of depth index, and the loadings on (b) Component 4, and (c) Component 5 as a function of the mass (u). ................................................................................... 71

**Figure 3.12.** Results from a PCA analysis of the ToF-SIMS spectra from a depth profile through a film of plasma polymerized PNIPAM on a Si substrate. (a) and (d) Scores plots as a function of depth index, and the loadings of (b) PC 1, (c) PC 2, (e) PC 3 and (f) PC 4 as a function of the mass (u). ........................................................................................................................................ 74

**Figure 3.13.** Factors 1 – 4 from a six factor MCR analysis of the spectra from a ToF-SIMS depth profile of plasma polymerized PNIPAM on Si. (a) Scores on MCR components (1 – 4) as a function of depth index, and the contributions (loadings) of (b) Component 1, (c) Component 2, (d) Component 3, and (e) Component 4 as a function of mass (u). ........................................................................................................................................ 76

**Figure 3.14.** Factors 5 – 6 from a six factor MCR analysis of the spectra from a ToF-SIMS depth profile of plasma polymerized PNIPAM on Si. (a) Scores on MCR components 5 and 6 as a function of depth index, and the loadings on (b) Component 5 and (c) Component 6 as a function of mass (u). ..................................................................................................................................... 76

**Figure 3.15.** PCA analysis of the spectra from an XPS depth profile through a film of SiO$_2$ on Si. (a) The scores on PC 1 and PC 2 as a function of depth index, and the contributions/loadings of PC 1 and PC 2 as a function of binding energy for (b) the silicon 2p narrow scan and (c) the O 1s narrow scan. (d) The scores on PC 3 and PC 4 as a function of depth index, and the contributions/loadings of PC 3 and PC 4 as a function of binding energy for (e) the silicon 2p narrow scan and (f) the O 1s narrow scan................................................................. 80

**Figure 3.16.** The first four components from a seven-component MCR analysis of XPS data from a depth profile of SiO$_2$ on a Si substrate. (a) Scores on MCR components 1 and 2 as a function of depth index, and the contributions (loadings) of Components 1 and 2 as a function of binding energy corresponding to (b) silicon (Si 2p) and (c) oxygen (O 1s). (d) Scores on MCR components 3 and 4 as a function of depth index, and the contributions (loadings) of Components 3 and 4 as a function of binding energy corresponding to (e) silicon (Si 2p) and (f) oxygen (O 1s). MCR components 5 – 7 are shown in the Appendix 5. ........................................ 81

**Figure 3.17.** PCA analysis of the XPS spectra from a depth profile through a film of Ta$_2$O$_5$ on a Ta substrate. (a) Scores plot for PC 1 and PC 2 as a function of depth index, and loadings on PC 1 and PC 2 as a function of binding energy for (b) tantalum and (c) oxygen. (d) Combined scores plot for PC 3, PC 4 and PC 5 as a function of depth index, and the loadings on PC 3, PC 4, and PC 5 as a function of binding energy for (e) tantalum and (f) oxygen. ................................................................. 85

**Figure 3.18.** Results from a five-factor MCR analysis of an XPS depth profile through Ta$_2$O$_5$ on Ta. The (a) scores on MCR components 1 and 2 as a function of depth index, and the loadings on Components 1 and 2 as a function of binding energy for the (b) Ta 4f and (c) O 1s regions. The (d) scores on MCR components 3, 4, and 5 as a function of depth index, and the loadings on Components 3, 4, and 5 as a function of binding energy for the (e) silicon 2p and (f) oxygen 1s regions. ........................................................................................................ 87
Figure 3.19. (a) PRE values and (b) backward difference PRE values of the mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized C$_3$F$_6$ on Si as a function of depth index. ................................................................. 93

Figure 3.20. (a) PRE values and (b) backward difference PRE values of 19 selected peaks from the mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized PNIPAM on Si as a function of depth index. ...................................................................................................... 94

Figure 3.21. (a) PRE values and (b) backward difference PRE values of the Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. ........................................................................................................ 95

Figure 3.22. (a) PRE values and (b) backward difference PRE values of O 1s narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. ......................................................................................... 96

Figure 3.23. O 1s spectra collected during the XPS depth profiling study of SiO$_2$/Si substrate. The labels in the panels are the depth indices of the spectra. Spectra 50-54 are very similar to spectra 43-49 and so are not included here. ................................................................. 97

Figure 3.25. PRE values of the mock spectra shown in Figure 37 plotted as a function of (a) the amount of random noise (%) and (b) Log (random noise %). ......................................................................................... 99

Figure 3.26. Analysis of the Ta 4f narrow scans from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta. (a) PRE values and (b) backward difference PRE values as a function of depth index. ......................................................................................... 100

Figure 3.27. Analysis of O 1s narrow scans from an XPS depth profile of Ta$_2$O$_5$ on Ta. (a) PRE values and (b) backward difference PRE values as a function of depth index. ................................................................. 101

Figure 3.28. O 1s spectra collected during an XPS depth profile of Ta$_2$O$_5$ on Ta substrate. Spectra 36 and 37 are very similar to spectra 29-25 and so are not included here. .................................................. 102

Figure 4.1. Raw mass chromatograms representing relatively ‘clean’ (a – c) and ‘dirty’ (d – f) m/z channels from an LC-MS experiment. (a) m/z= 1088, (b) m/z= 635, (c) m/z=317, (d) m/z= 1383, (e) m/z= 570, and (f) m/z= 118. These data are found in the PLS_Toolbox from Eigenvector Inc. (Manson, WA) and were originally published by Windig et al. 27 ................................................................ 110

Figure 4.2. (i) Mock, normalized spectra containing 15 data points and (ii) the spectra from (i) in their reordered form. (a) A spectrum containing a ‘spike’ one data channel wide with all other data points having values of zero, and (b) A spectrum consisting of a series of data points with equal values. ......................................................................................... 111

Figure 4.3. The same six spectra in Figure 1 replotted in their reordered state. Data points in the spectra are omitted wherever their values are zero – the spectra go to zero wherever the spectra end. ......................................................................................... 112

Figure 4.4. Areas of nineteen selected peaks/regions from ToF-SIMS spectra obtained from a depth profile through a film of plasma polymerized PNIPAM on Si. ................................................................. 113

Figure 4.5. The same four spectra that are in Figure 4 plotted in their reordered state. Data points in the spectra are omitted wherever their values are zero – the spectra go to zero wherever the spectra end. ......................................................................................... 115

Figure 4.6. (a) PRE values of a series of spectra from a depth profile through a ca. 100 nm film of PNIPAM on silicon. (b) The backward difference PRE values from (a). (c) Spectra 1 – 12 and 22 – 24 from this depth profile in their reordered state. ......................................................................................... 116

Figure 5.1. The total ion current chromatogram (TICC) constituted from all the 1451 mass chromatograms considered in this study ......................................................................................................................... 121
Figure 5.2. Raw mass chromatograms representing relatively ‘clean’ and ‘noisy’ mass chromatograms at (a) m/z= 1088, (b) m/z= 635, (c) m/z=317, (d) m/z= 1383, (e) m/z= 570, and (f) m/z= 118. ....................................................................................................................................... 129

Figure 5.3. (a) Plot of PRE values of individual mass chromatograms vs. corresponding m/z values. (b) Spline fit to the PRE values in (a). (c) Selection of the mass chromatograms that are 1 σ or more away from the spline fit. ............................................................................................. 130

Figure 5.4. Reduced total ion current chromatograms (TICC) obtained by varying the threshold (σ) for the data points in Figure 5.6 relative to their spline fit. (a) 0.0 σ (614), (b) 0.25 σ (320), (c) 0.5 σ (151), (d) 0.75 σ (84), (e) 1.0 σ (40), (f) 1.25 σ (26), (g) 1.5 σ (13), (h) 2.0 σ (3), and (i) 2.5 σ (1). The numbers in parenthesis here are the number of mass chromatograms summed to make the corresponding TICCs. ........................................................................................................ 133

Figure 5.5. Comparison of the TICC obtained via (a) PRE with with 0.5 σ (b) CODA and (c) PRE with smoothing. The CODA result was obtained with the default smoothing window of 5 and MCQ threshold value of 0.89 ........................................................................................................ 134

Figure 5.6. Comparison of mass chromatograms arising from the same components (a) m/z= 627 Da and (b) m/z= 626 Da, and (c) m/z= 1091 Da and (d) m/z= 1089 Da. ................................................ 135

Figure 6.1. A comparison of the TICCs generated by: (a) summing the 1451 mass chromatograms considered in this study, (b) CODA, (c) PRE only (threshold of -0.662), (d) PRE with a Savitzky-Golay smooth (PRE threshold of -0.662 and a smoothing width of seven data points), and (e) PRE-CC (thresholds of -0.662 and 0.275 for PRE and CC, respectively). .... 146

Figure 6.3. A plot representing all of the 1451 mass chromatograms used in our study, where the position of the mass chromatogram (data point) in the plot is given by its PRE value (x-axis value) and CC maximum value (y-axis value), and the quality of the mass chromatogram is denoted by a color: green for highest quality, yellow for moderate quality, and red for poor quality. ....................................................................................................................................... 151

Figure 6.4. TICCs generated by summing the mass chromatograms in Quadrants 1 (a), 2 (b), 3 (c) and 4 (d) in Figure 3, which are defined by the heavy, dashed lines in the plots. .................................................................................................................. 151

Figure 6.5. Change in TICCs at a fixed PRE threshold of -0.662 (0.5 σ) as a function of the CC maximum threshold: (a) 0.0, (b) 0.1, (c) 0.275 and (d) 0.5. ............................................................................. 153

Figure 7.1. Capillary electropherograms of undoped (0 %) and doped (5 and 10 %) blood samples of subjects (a) A, (b) B, and (c) C. Three replicates at each doping level are shown in each panel. (d) Three electropherograms from Subject A at 0, 5, and 10 % doping levels. .... 159

Figure 7.2. Analysis of the main, sharp peak (Labeled ‘1’ and highlighted in green) centered at ca. 1.5 min in an electropherogram of a 5 % doped sample from Subject A using a Shirley background. ........................................................................................................................................ 167

Figure 7.3. Dendrogram from a cluster analysis of the 0, 5, and 10 % electropherograms under consideration in this study. The data were preprocessed using range selection followed by normalization (1-norm) and autoscaling. The dashed, vertical, light blue line indicates a two-cluster model, and the dashed, vertical, black line indicates a three-cluster model. ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0, 5, and 10 % samples, and ‘A’, ‘B’, and ‘C’ represent the three subjects. Replica runs are represented by the number following the ‘A’, ‘B’, or ‘C’....... 172

Figure 7.4. Seven-component, PLS predictions of doping levels from replicate runs for undoped and doped (5 and 10 %) blood samples. Here, a separate seven-component PLS model was
created for each data set with one of its samples left out, and that sample was then predicted by the corresponding model......................................................................................................................................................................................... 176

**Figure 7.5.** Results of the less traditional mathematical/informatics methods used to analyze the doping data. Figure 5a (top left): The average PRE values (heights of bars) with standard deviations (error bars) of the electropherograms of subjects A, B and C for 0, 5, and 10 % doping levels. Figure 5b (top right): The reordered electropherograms from replicate run 1 of subject A at 0, 5, and 10 % doping levels. Figure 5c (middle left): The Euclidean distances between electrophoretic separations of clean and doped (5% and 10%) samples of subjects A, B and C with three replicate runs for each. Figure 5d (middle right): The absolute areas of the broad features (shoulders) to the right of the main peaks from electrophoretic separations of Subjects A, B and C for 0, 5, and 10 % doping levels. Figure 5e: The second moments of the range-selected electropherograms......................................................................................................................................................................................... 180

**Figure 7.6.** Leave-one-out predictions from ILS models based on (a) PRE, \(d_{Eu}\), and PFI, (b) PRE, \(d_{Eu}\), and SM , (c) PRE, PFI, and SM , (d) PRE, \(d_{Eu}\), PFI, and SM summary statistics from replicate runs of A, B and C for undoped (0 %) and doped (5 % and 10 %) blood samples. The \(R^2\) values and data were compared against straight lines \(y = x\) (red lines).......................................................................................................................... 182

**Figure A1.1.** Structure of the average oligomer (four repeat units) in the negative photoresist SU-8. ............................................................................................................................................................................................................................................................................................................... 192

**Figure A1.2.** Structure of the water-soluble adhesion promoter polyallylamine (PAAm). ......... 195

**Figure A1.3.** Possible reaction of the primary amines in PAAm with the epoxy groups in SU-8. ............................................................................................................................................................................................................................................................................................................... 196

**Figure A1.4.** Overview of the coating/fabrication process. ............................................................................................................................................................................................................................................................................................................... 197

**Figure A1.5.** Ellipsometric thicknesses of native oxide coated silicon ‘Before Piranha’, after cleaning, ‘After Piranha’, after cleaning and PAAm deposition, ‘After PAAm’, and after spin coating and development (removal) of SU-8, ‘After SU-8’. ............................................................................................................................................................................................................................................................................................................... 200

**Figure A1.6.** Ellipsometric measurements of plasma-cleaned Si/SiO2, PAAm-coated Si/SiO2, and PAAm-coated Si/SiO2 exposed to SU-8 photoresist for 10, 20, 30, or 40 min prior to underdevelopment............................................................................................................................................................................................................................................................................................................... 202

**Figure A1.7.** Possible reaction of PAAm with the triarylsulfonium salt in SU-8.......................... 202

**Figure A1.8.** Water contact angles of plasma-cleaned Si/SiO2, PAAm-coated Si/SiO2, and PAAm-coated Si/SiO2 exposed to SU-8 photoresist for 10, 20, 30, or 40 min prior to underdevelopment............................................................................................................................................................................................................................................................................................................... 203

**Figure A1.9.** XPS Nitrogen 1s narrow scans of air plasma cleaned surfaces......................... 204

**Figure A1.10.** Surfaces spin coated with SU-8 that were (a) coated with PAAm and that show no defects, and (b) uncoated and that show defects............................................................................................................................................................................................................................................................................................................... 206

**Figure A1.11.** Description of the Spectroscopic Ellipsometry (SE) Models............................ 207

**Figure A2.1.** Illustration of angle resolved XPS. (a) A photoelectron exiting a solid at its maximum take-off angle (90°), with accompanying depth probed in the material of \(\Lambda_e\). (b) A photoelectron exiting a solid at a take-off angle less than 90°, with accompanying depth probed in the material of \(\Lambda_e \sin \theta\). Figure adapted from one in the paper cited here.3........................................ 215

**Figure A2.2.** Relative depth plot for a mixed film of HfO2 and Al2O3 on SiO2 on Si. Figure and caption used/adapted with permission from Thermo Fisher................................................................. 217
Figure A2.3. Thicknesses of carbon, alumina (Al₂O₃), and silica (SiO₂) layers on a silicon substrate versus the number of Al₂O₃ ALD cycles. The thicknesses of the three layers here were determined with Thermo Fisher’s multi-layer thickness calculator Advantage. Figure and caption used/adapted with permission from Thermo Fisher................................. 219

Figure A2.4. Results from a line scan across the diameter of a 200 mm wafer showing the variation in thicknesses of a mixed Al₂O₃/HfO₂ layer and the underlying SiO₂ film. Figure and caption used/adapted with permission from Thermo Fisher...................................................... 220

Figure A2.5. Chemical state maps of the O 1s signal on a 200 mm wafer. The upper map corresponds to oxygen with a higher binding energy (presumably bonded to Al or Si), and the lower map corresponds to oxygen with a lower binding energy (presumably bonded to Hf). Figure and caption used/adapted with permission from Thermo Fisher........................................ 221

Figure A2.6. XPS maps of aluminum (Al 2p signal) (top) and hafnium (Hf 4f signal) (bottom) from a 200 mm wafer. Figure and caption used/adapted with permission from Thermo Fisher.222

Figure A3.1. Cut-away image of the Thermo Scientific K-Alpha® showing the vacuum hardware under the system cladding. Figure and caption used/adapted with permission from Thermo Fisher.................................................. 230

Figure A3.2. Thermo Fisher Scientific XPS instruments, including the K-Alpha® use a focusing monochromater to control X-ray spot size. The incident electron beam from the cathode is focused onto the anode to give the desired X-ray spot size. Figure and caption used/adapted with permission from Thermo Fisher........................................................................................................ 233

Figure A3.3. XPS analysis using the K-Alpha+ instrument of a series of contact lenses. Figure and caption used/adapted with permission from Thermo Fisher.................................................. 236

Figure A3.4. Part (the ends) of the survey spectrum of gold acquired using Al Kα radiation. Figure and caption used/adapted with permission from Thermo Fisher.................................................. 238

Figure A3.5. Production line at Thermo Fisher. Figure and caption used/adapted with permission from Thermo Fisher.................................................. 240

Figure A4.1. Screenshot from the SpecsLab Prodigy software of the page that allows control of the system vacuum components........................................................................................................ 247

Figure A4.2. Screenshot of the Experiment Editor in Prodigy.......................................................... 248

Figure A4.3. Sputter depth profiling interface in the Prodigy software........................................ 250

Figure A4.4. Eight individual survey scans (left) and ten individual C 1s narrow scans (right) of an aqueous solution (Coca Cola®) obtained sequentially by NAP-XPS............................................................. 252

Figure A4.5. Nine individual O 1s narrow scans of an aqueous solution (Coca Cola®) obtained sequentially by NAP-XPS......................................................................................................................... 253

Figure A4.6. Principal component analysis (PCA) of the nine O 1s narrow shown in Figure 5. Top: plot of Q Residuals vs. Hotelling T², middle: plot of scores on PC1 vs. sample number, and bottom: plot of scores on PC2 vs. sample number........................................................................................................................................ 254

Figure A5.1. Scores on Components 5 – 7 from a seven-component MCR analysis of XPS data from a depth profile of SiO₂ on a Si substrate. See Figure 19 in the main text and related discussion.................................................................................................................................. 258

Figure A5.2. Loadings on Components 5 – 7 from a seven-component MCR analysis of XPS data from a depth profile of SiO₂ on a Si substrate corresponding to (a) oxygen (O 1s) and (b) silicon (Si 2p)........................................................................................................................................ 259
Figure A5.3. Oxygen-to-silicon ratio as a function of spectrum number (depth index) during the XPS depth profiling of a film of SiO$_2$ on Si. ................................................................. 260

Figure A5.4. Ta 4f spectra collected during an XPS depth profile study of Ta$_2$O$_5$ on Ta substrate. Spectra 36 and 37 are very similar to spectra 29 – 25 and so are not included here. ................... 261

Figure A5.5. (a) PRE values and (b) backward difference PRE as a function of spectrum number for ToF-SIMS depth profiling of a C$_3$F$_6$ polymer layer on Si. A total of 19 peaks were selected from each spectrum for these analyses. ................................................................. 262

Figure A5.6. (a) PRE values and (b) finite difference PRE values as a function of spectrum number for a ToF-SIMS depth profile of a PNIPAM polymer film on Si. The entire spectra from m/z 0 to 150 were used in this analysis................................. 263

Figure A5.7. Si 2p spectra collected during an XPS depth profiling of SiO$_2$ on Si. Spectra 50-54 are very similar to spectra 43-49 and so are omitted here. ......................................................... 264

Figure A5.8. Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of plasma polymerized C$_3$F$_6$ on Si. ......................................................... 265

Figure A5.9. Scores on PCs 1 and 2 of plasma polymerized C$_3$F$_6$ on Si. ......................................................... 266

Figure A5.10. Loadings on PC 1 and 2 of plasma polymerized C$_3$F$_6$ on Si. ......................................................... 266

Figure A5.11. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets ................. 267

Figure A5.12. Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of plasma polymerized PNIPAM on Si. ......................................................... 269

Figure A5.13. Scores on PCs 1 and 2 of plasma polymerized PNIPAM on Si. ......................................................... 269

Figure A5.14. Loadings on PCs 1 and 2 of plasma polymerized PNIPAM on Si. ......................................................... 270

Figure A5.15. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of plasma polymerized PNIPAM on Si. ......................................................... 270

Figure A5.16. Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of SiO$_2$ on Si. ......................................................... 272

Figure A5.17. Scores on PCs 1 and 2 of SiO$_2$ on Si. ......................................................... 273

Figure A5.18. Loadings on PCs 1 and 2 of SiO$_2$ on Si. ......................................................... 274

Figure A5.19. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of SiO$_2$ on Si. ......................................................... 274

Figure A5.20. Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of Ta$_2$O$_5$ on Ta depth................................................................. 276

Figure A5.21. Scores on PC 1 and 2 of Ta$_2$O$_5$ on Ta................................................................. 277

Figure A5.22. Loadings on PC 1 and 2 of Ta$_2$O$_5$ on Ta................................................................. 278

Figure A5.23. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of Ta2O5 on Ta. ......................................................... 278

Figure A6.1. PRE values of Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. While these results quite closely mirror those we previously published, the number of data points here is different......................................................... 281
Figure A6.2. Representative, normalized Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. ‘Spectrum 1’, ‘Spectrum 3’, ‘Spectrum 18’, and ‘Spectrum 30’ corresponds to the first, third, eighteenth, and thirtieth data points in Figure A6.1........................................ 281

Figure A6.3. The same four Si 2p narrow scans that are in Figure A6.2 plotted in their reordered state ............................................................................................................................................. 282

Figure A6.4. PRE values of XPS Ta 4f narrow scans vs. depth index from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta. ................................................................................................... 283

Figure A6.5. Selected Ta 4f narrow scans from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta. ............................................................................................................................................... 284

Figure A6.6. The same four Ta 4f narrow scans that are in Figure A6.5 plotted in their ordered state. ............................................................................................................................................ 285

Figure A6.7. PRE values of mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized C$_3$F$_6$ on Si as a function of depth index.................................................... 286

Figure A6.8. Representative, normalized mass spectra (actually areas of 19 selected peaks) from a ToF-SIMS depth profile through a film of plasma polymerized C$_3$F$_6$ on Si as a function of depth index................................................................................................................................................. 287

Figure A6.9. The same four spectra in Figure A6.8 plotted in their reordered state. Data points in the spectra are omitted wherever their values are zero. .............................................................. 288

Figure A7.1. A version of Figure 2 from the manuscript with thinner line widths indicating plotting artefacts.......................................................................................................................... 290

Figure A7.2. Comparison of the two curve fitting techniques- Savitzky-Golay (left) with a frame length of 151 vs Spline (right). ................................................................................................... 291

Figure A7.3. (a) and (c): Selection of mass chromatograms above 0.0σ and 0.5σ respectively, (b) and (d) corresponding reduced TICCs........................................................................................................................ 292

Figure A7.4. (a) TICC obtained with a threshold of 0.5σ and (b) individual mass chromatograms constituting this TICC. ................................................................................................................ 293

Figure A8.1. Dendrogram from a cluster analysis of the 0, 5, and 10 % electropherograms under consideration in this study. The data were pre-processed using range-selection followed by normalization (1-norm). ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0, 5, and 10 % samples, and ‘A’, ‘B’, and ‘C’ represent the three subjects. The numbers after ‘A’, ‘B’, or ‘C’ represent replica runs............................................................................................................................................. 294

Figure A8.2. Dendrogram from a cluster analysis of the 0 %, 5 %, and 10 % electropherograms under consideration in this study. ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0 %, 5 %, and 10% samples, and ‘A’, ‘B’, and ‘C’ represent the three subjects. The numbers after ‘A’, ‘B’, or ‘C’ represent replica runs. The data was pre-processed using range-selection followed by autoscaling. ................................................................................................................................................. 295

Figure A8.3. RMSEC (orange) and RMSECV (blue) plots from PCA calculations of the electropherograms from Subjects A, B and C for doped (5 % and 10 %) and undoped samples. ..................................................................................................................................................... 296

Figure A8.4. Scores on PC1 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 1-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A. ................................................................. 297
Figure A8.5. Scores on PC1 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 298

Figure A8.6. Scores on PC2 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 299

Figure A8.7. Scores on PC3 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 300

Figure A8.8. Scores on PC4 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 301

Figure A8.9. Scores on PC5 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 302

Figure A8.10. Scores on PC6 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 303

Figure A8.11. Scores on PC7 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 304

Figure A8.12. Scores on PC8 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 305

Figure A8.13. Scores on PC9 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 306

Figure A8.14. Hotelling T² vs Q residual plot from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 1-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A ................................................................. 307

Figure A8.15. Hotelling T² vs Q residual plot from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped
samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A. ............................................ 308

**Figure A8.16.** RMSEC (orange) and RMSECV (blue) plots from PLS calculations of the electropherograms from Subjects A, B and C for doped (5 % and 10 %) and undoped samples. ..................................................................................................................................................... 309

**Figure A8.17.** Raw PRE values of electrophoretic separations of subjects A, B and C for various doping levels (0 %, 5 % and 10 %). ................................................................................................................................. 310
Chapter 1: Overview of XPS, ToF-SIMS, LC-MS, CE, Various Informatics Methods, and PRE

1.1 Introduction

Material characterization and analysis play a central role in the advancement of essentially all materials, e.g., in the semiconductor industry,1-5 and for nanomaterials,6-7 separation devices,8-13 data storage materials,14-19 hydrophobic coatings,20-24 etc. For many materials, the area of interest is its surface because the surface, based on its physical and chemical properties, interacts directly with its surroundings. For example, catalysis, tribology, wetting, adhesion, corrosion, adsorption, biological signaling and transport, separation science, device failure, and sensing often depend on what is happening in the outermost 0.1 – 1 nm of a material.25-27

Materials science and analytical chemistry have changed dramatically in the last 30 years. In the more distant past of these disciplines, the problems associated with a chemical analysis were solved by using simple techniques like precipitation, weighing, and titration. However, today, with the significant advances in instrumentation and computation capabilities, most problems in analytical chemistry are solved using instruments like LC-MS and CE, and for surfaces XPS and ToF-SIMS, which generate large, even vast, amounts of data.28 The data generated by these instruments needs to be converted into usable information. Moreover, in most surface analyses, a multi-instrument characterization approach is adopted, complicating the problem even further. Different surface analytical techniques provide information at different length scales. XPS29-30 and Auger electron spectroscopy (AES)31 probe ca. 0.5 – 10 nm into surfaces, ToF-SIMS30,32 is
sensitive to ca. 2 – 3 nm into a material, and LEIS$^{33-34}$ characterizes the outermost monolayer of a material. As noted, it is often a combination of analytical techniques that best reveals the composition and nature of a surface.$^{35}$ That is, there is no comprehensive technique for surface analysis that functions like NMR does for small molecule organic chemistry.

Chemometrics, a term coined by Svante Wold and Bruce R. Kowalski in 1971$^{36-38}$, is a discipline associated with the development of quantitative models and the study of sample-variable relationships in chemistry. In other disciplines, the same methods and ideas are referred to by other names, including informatics, bioinformatics (in biology), and data science. The upsurge of data generated by modern analytical instruments with the advancement of the personal computer has ushered in a new age of chemical data acquisition, processing and interpretation. As such, there is a growing need to develop mathematical and statistical techniques to interpret large data structures generated by modern chemical instruments. Over the years, chemometrics has evolved into an independent discipline having a significant role in analytical chemistry and materials analysis. In modern analytical chemistry, one of the most common problems is the multivariate nature of the data, i.e., there exists multiple variables associated with different classes of samples in any analysis. Probably the most common statistical methods used for pattern recognition and variance analysis are principal component analysis (PCA), multivariate curve resolution (MCR), cluster analysis, and partial least squares regression (PLS-R). The theories associated with each of these mathematical will be discussed in Chapter 2.

The Linford research group at BYU has long been engaged in the study and engineering of surface properties for suitable application purposes$^{8-9,39}$. Two of the most widely used surface characterization tools used in the Linford lab are X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (ToF-SIMS)$^{30,40-41}$. Traditionally, for elemental
and compositional analysis, suitable signals originating from the elements of interest are monitored. However, when surface instruments are coupled with a sputter gun for the acquisition of depth profiles, the problem becomes more complex – large numbers of spectra are obtained. Of course this problem is exacerbated when surface imaging is also performed. PCA and MCR can be used to interpret depth profile spectra. PCA is reasonably easy to apply and it has appealing mathematical properties. However, given the orthogonality constraints of PCA, the results can be somewhat abstract, i.e., challenging to interpret, even for experienced users. This is especially true for the higher principal components. On the other hand, MCR results can be very interpretable because the factors generated therein are constrained to resemble real spectra, e.g., they do not have negative scores or loadings. A novel data analyses tool that I have discussed in this dissertation is pattern recognition entropy (PRE). As will be discussed later, PRE is based on Shannon’s groundbreaking work on information theory and is presented as a simple, easy to perform mathematical tool that can be used as a summary statistic for the interpretation of depth profiles. Chapter 3 of this document contains a discussion of the statistical analysis of XPS and ToF-SIMS depth profiles. PCA, MCR, and PRE are applied to four different data sets obtained from: a ToF-SIMS depth profile through ca. 100 nm of plasma polymerized C\textsubscript{3}F\textsubscript{6} on Si, a ToF-SIMS depth profile through ca. 100 nm of plasma polymerized PNIPAM (poly (N-isopropylacrylamide)) on Si, an XPS depth profile through a film of SiO\textsubscript{2} on Si, and an XPS depth profile through a film of Ta\textsubscript{2}O\textsubscript{5} on Ta. PCA, MCR, and PRE reveal the presence of interfaces in the films and often indicate that the first few scans in the depth profiles are different from those that follow. PRE and backward difference PRE provide this information in a straightforward fashion.
However, PRE also has limitations. There is some abstraction and ambiguity associated with it. The concept of the ‘reordered spectrum’\textsuperscript{48} is introduced in Chapter 4 as an intuitive, visual representation of spectra. The shapes of reordered (sorted) spectra correlate with their PRE values and help explain them. These concepts are illustrated with the spectra of liquid chromatography-mass spectrometry (LC-MS), time-of-flight secondary ion mass spectrometry (ToF-SIMS), and X-ray photoelectron spectroscopy (XPS). Reordered spectra have value beyond PRE for comparing and understanding spectra.

The total ion current chromatogram (TICC)\textsuperscript{49} obtained by LC-MS is often extremely complex and ‘noisy’ in appearance,\textsuperscript{50-51} particularly when an electrospray ionization source is used.\textsuperscript{52-54} Accordingly, meaningful qualitative and quantitative information are often best obtained in LC-MS by data mining processes.\textsuperscript{55-60} Here, one or more higher-quality mass chromatograms can be identified/extracted/isolated and combined to form a TICC, wherein much of the background mass noise is eliminated, and quantitative data for chromatographic peaks can be obtained.

During my work with PRE, I discovered that PRE can be employed as a shape recognition tool, i.e., it can effectively differentiate between spectra with many features and relatively featureless spectra. This unique property of PRE, which I explored in my work on noise reduction in TICCs generated by LC-MS, is discussed in Chapter 5.\textsuperscript{61} Reduced TICCs are obtained by first calculating the PRE values of the component mass chromatograms. A plot of PRE value vs. m/z for the mass chromatograms is then generated, and the resulting band of PRE values is fit to a piecewise spline polynomial. The distribution of the differences between the individual PRE values and the spline fit is then used to select good, information-containing mass chromatograms. PRE reduces the number of component mass chromatograms significantly (by an order of
magnitude) and simultaneously preserves most of the chemical information that is collectively in them. Also, it can distinguish between mass chromatograms of chemically similar species. PRE is arguably a less computationally intensive alternative to the widely used component detection algorithm (CODA)\textsuperscript{50, 62} for variable reduction. It produces reduced TICCs of comparable if not higher quality, and it requires only a single user input for variable selection. Moreover, reduced TICCs generated by PRE can be smoothed to further improve their signal-to-noise ratios.

However, while the chromatographic signals in the reduced TICC from PRE were well resolved, some noise remained in the TICC suggesting that the algorithm had selected some false positives, i.e., poor quality mass chromatograms. In Chapter 6, I report an improved version of the PRE algorithm that utilizes a second variable selection filter based on cross-correlation (CC)\textsuperscript{63}. As a check on the ability of PRE and CC to select high quality mass chromatograms, every mass chromatogram in our data set (1451 in total) was individually inspected and rated as either high quality (green), intermediate quality (yellow), or poor quality (red). A color-coded plot of the CC value vs. the PRE value for the mass chromatograms was created, which shows that, as expected, the higher quality mass chromatograms are localized in its upper left quadrant, which corresponds to lower PRE values and higher CC values. In our original paper on this topic, we recommended a threshold of 0.5 $\sigma$ for PRE, which caused the algorithm to select 151 mass chromatograms out of 1451. Of these, 98 were of high quality, 6 were of intermediate quality, and 47 were of poor quality. Using a second threshold for CC, the algorithm kept all the high and intermediate quality mass chromatograms, while removing all 47 of the poor quality ones. The resulting TICC from the PRE-CC algorithm shows less noise compared to the TICC generated from the PRE approach alone. The PRE-CC algorithm is arguably a faster, simpler and more intuitive approach as
compared to the improved version of CODA, the Durbin-Watson based CODA (CODA_DW).62, 64

An autologous blood transfusion (ABT) is the reinjection of blood previously taken from an athlete to increase the oxygen transport capabilities of his/her blood during competition.65-67 This prohibited form of doping is used by endurance athletes to enhance aerobic performance. Despite the World Anti-Doping Agency’s ban on such methods, widespread abuse has taken place, with Lance Armstrong being an infamous example. Autologous blood doping (ABD) can be a challenge to detect because of the similarities between an individual’s doped and undoped blood. Recently, Harrison et al. reported that high-speed capillary electrophoresis may identify ABD.69 First order derivatives of the electropherograms were suggested as a possible tool for the detection of doping. However, given that the results were based on the somewhat subjective analysis of slopes, this method of detection frequently suffered from false negatives. In Chapter 7, I provide a more complete mathematical analysis of the data from the study of Harrison et al. where I provide a contrast between traditional statistical methods and alternative mathematical techniques.70 First, I applied three multivariate statistical analysis tools: cluster analysis, principal component analysis (PCA), and partial least squares (PLS) to develop a calibration and/or obtain definitive groupings of undoped (0%) and 5% and 10% doped blood. Different preprocessing and variable selection methods were considered in these approaches. Unfortunately, due to natural variation in the electropherograms and the limited size of the data set, little or no success was obtained by these efforts. Then, I applied a series of less common mathematical/informatics tools to this problem. These included pattern recognition entropy (PRE), the Euclidean distance between vectors, a peak fitting/integration method and the second moment (SM). Each of these mathematical techniques showed at least some ability to differentiate between the 0, 5, and 10%
doped samples. Inverse least squares (ILS) models involving the combinations of summary statistics were studied to evaluate their prediction capabilities. An ILS calibration based on the summary statistics obtained from PRE, the Euclidean distance, and peak fitting/integration was much more successful than the previously obtained PLS model at predicting levels of blood doping from the corresponding electropherograms. Accordingly, the ILS model could detect the presence of doping (5% and 10%) with 100% accuracy compared to undoped (0%) blood. This methodology may be applicable to other challenging informatics problems like the determination of risk factors for genetically linked diseases, robust pattern finding in peak-like data such as ChIP-seq or other genomic sequencing for understanding the 3D genome.71

Chapter 8 contains my conclusions on PRE and the future work to be done in this area. I believe the shape recognition property of PRE can be used in various applications where the signal-to-noise ratio is crucial to the success of the endeavor as in the automation of data collection in XPS and ToF-SIMS. The change of PRE values with scan number in signal averaging will decrease noise in the spectrum. Other possible applications are discussed in Chapter 8.

Additionally, this dissertation contains several appendices. Appendix 1 discusses my work on the development of an adhesion promoter for the widely used photoresist SU-8.72 SU-8 has emerged as a favorite photoresist for high aspect ratio (HAR) lithography73-75 showing high chemical and mechanical stability and biocompatibility. It is widely used in the microelectronics industry for masking,76 microelectromechanical systems (MEMS),77 and as a structural material for optics.73 In semiconductor fabrication, photoresists are used to create patterns on the silicon wafer for subsequent thin film deposition.78 Unfortunately, its poor adhesion to substrates is a drawback,79-80 with possible solutions being low-viscosity formulations of SU-8,81 surface modification with a low molecular weight adsorbate like hexamethyldisilazane (HMDS)82 or a
commercial adhesion promotion reagent (OmniCoat from MicroChem Inc.). However, HMDS and the commercial reagent require surface dehydration and/or curing, and a modified form of SU-8 is not always desirable. In this work, I demonstrate the use of a water-soluble, amine-containing polymer, polyallylamine (PAAm),\textsuperscript{83-84} which spontaneously adsorbs to silica surfaces, as a simple, easy-to-apply and reactive adhesion promoter for SU-8. Conditions for the use of PAAm are explored, and the resulting materials are characterized by X-ray photoelectron spectroscopy (XPS), spectroscopic ellipsometry (SE), and wetting.

Appendices 2 and 3 discuss recent trends in XPS instrument as part of a collaboration with Thermo Fisher Scientific and Kratos Analytical Ltd. Appendix 2\textsuperscript{85} discusses two less well-known XPS analysis techniques: angle-resolved XPS (AR-XPS) and surface imaging. In AR-XPS, the takeoff angle is varied to provide elemental compositional information from different depths of the sample. Surface imaging can provide information on an extended region of the sample and can be useful in the determination of changes in elemental composition across the region of analysis. In Appendix 3,\textsuperscript{86} I discuss the various aspects of the K-Alpha$^+$ system produced by Thermo Fisher Scientific. It is a relatively low cost and highly automated instrument that is popular in research laboratories around the world. As part of the collaboration, I had the opportunity to visit Thermo Fisher’s factory in East Grinstead, UK and witness its production. I discussed several aspects of this process in the article that I wrote for Vacuum Technology and Coating (VT&C) which helped users better understand the instrument. In Appendix 4,\textsuperscript{87} I discuss some of the advanced software capabilities associated with XPS like remote system operation and automated charge compensation. More specifically, I discussed the Specs Prodigy Software as part of a collaboration with Specs Surface Nano Analysis GmbH, Germany. Finally, Appendices 5-8 contain supporting information for Chapters 3, 4, 5 and 7 respectively.
In what remains of this chapter, I discuss the XPS and ToF-SIMS techniques in some detail. Next, I describe two extensively used analytical separation techniques: LC-MS and CE. Finally, I end this chapter with a short description of each of the chemometric techniques that will be used for various analyses throughout this dissertation.

1.2 X-Ray Photoelectron Spectroscopy (XPS)

XPS is based on Einstein’s photoelectric effect, which describes the emission of electrons from materials when radiation of a certain energy is incident on their surfaces. As such, XPS is used in materials science and the electronics industry and is one of the most frequently used surface analysis technique,\(^{31, 88-91}\) according to the number of publications on it each year. One of the key advantages of XPS is that it generates quantitative elemental information of the outermost ca. 0-10 nm of the material.\(^{92}\) In XPS, X-rays of a fixed energy (wavelength) are used to probe a surface, where the emitted photoelectrons are detected and their kinetic energies are measured. When X-rays of sufficiently high energy are incident on a surface, core electrons from elements comprising the surface are emitted. Elemental analysis is performed given that each of the core electrons has a unique binding energy (the energy required to overcome nuclear attraction) that is dependent on the identity of the element and the core level from which it was ejected. An additional factor that affects the binding energy of a photoelectron is its chemical environment. The fundamental equation of XPS is given below:\(^{93}\)

\[
(1.1) \quad h\nu = BE_F + \Phi_{spec} + KE_{spec}
\]

which can be re-written as:

\[
(1.2) \quad BE_F = h\nu - \Phi_{spec} - KE_{spec}
\]
where, $h\nu$ is the energy of the incident photon, $BE_F$ is the binding energy of a core electron, $KE_{spec}$ is the energy of the photoelectron as measured inside the spectrometer and $\Phi_{spec}$ is the work function of the spectrometer. The typical work function for XPS spectrometers is 4.5 eV.\textsuperscript{94} The two most common X-ray sources used in stand-alone instruments are the Al K$_\alpha$ (1486.6 eV) and Mg K$_\alpha$ (1253.6 eV) lines. Conventional XPS spectra are plotted as counts (number of detected electrons) versus binding energy.

The hemispherical analyzer in an XPS instrument is used to measure the kinetic energy of the photoelectrons. A hemispherical analyzer is comprised of two co-hemispherical plates with a potential difference existing between them. As the electrons pass between the plates, the faster electrons (higher kinetic energy) will take a shorter trajectory and collide with the outer plate. Conversely, the slower electrons (lower kinetic energy) will take a longer path and collide with the inner plate. Only electrons within a certain range of energy which is referred to as ‘pass energy,’ will pass through the analyzer and strike the detector. The resolution of elemental peaks is directly dependent on the absolute magnitude of the pass energy and thus, XPS instruments are typically operated at a fixed low value of pass energy. A scanning retarding field is used before the analyzer to detect electrons across the entire energy range. Other components of XPS include: an electron gun, a suitable target material for creating X-rays, a monochromator for selecting X-rays of a fixed wavelength, optics for focusing the X-ray beams and the photoelectron beams, and a positively biased detector.
1.3 Time-of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS)

XPS is highly surface sensitive and has detection limits of 0.1-1 atom percent. However, it cannot detect hydrogen or helium or distinguish between isotopes. A time-of-flight mass spectrometer is able to detect all elements, distinguish between their isotopes, and provide molecular information. It is also highly sensitive and quantitative. In ToF-SIMS a sample surface is bombarded with high energy ions (in the keV range). Common primary ions for ToF-SIMS include $[\text{Ga}^+, \text{Bi}^{2+}, \text{C}_{60}^+, \text{Au}^+, \text{SF}_5^+]$. Ideally, these ions should not penetrate deeply into a material. Rather, it is best if they deposit their energy in the outermost regions of a sample. Larger ions tend to do this better than smaller ones. As the primary ions collide with the surface atoms,
they transfer their kinetic energy to neighboring surface atoms. During this process, a collision cascade, some of the surface atoms gain the critical energy required for their emission and as a result secondary ions are emitted. This transfer of energy can result in the release of sample ions from the surface, which can then be accelerated and directed into a mass spectrometer. A key advantage of ToF-SIMS is its ability to detect thermally unstable molecular species in the form of secondary ions. The sputtering yield, a measure of the number of secondary ions emitted per impact of every primary ion on the sample surface, is dependent on the chemical environment and often increases drastically in the presence of oxygen for electropositive species. The very basic design of an instrument is shown in Figure 1.2. A pulsed source produces a primary ion beam that is focused by electrostatic lenses onto the sample. Secondary ions emitted from the sample surface are collected using a beam extractor and passed through a reflectron for the elimination of the initial spread in energy. Afterwards, these ions are focused using another set of lenses onto the detector, which usually consists of a combination of photodiodes, scintillators, and photomultipliers.
Figure 1.2. Simple schematic representation of a time-of-flight secondary ion mass spectrometer (ToF-SIMS).
1.4 Liquid Chromatography-Mass Spectrometry (LC-MS)

Chromatography is comprised of a set of separation methods that permit the identification of individual components in a complex mixture of analytes. In chromatography, the sample, dissolved in a ‘mobile phase’, is passed over a ‘stationary phase.’ Due to the chemical interactions of the analytes within the two phases, separation of the analytes is achieved over time. In other words, components that interact strongly with the stationary phase will be slowed down and will be detected later when compared to analytes that interact less with the stationary phase/more with the mobile phase.98-100 Out of various separation methods, liquid chromatography (LC) is the most extensively used, because it is amenable to many molecular species, it is highly sensitive to trace amounts of analytes, and it is able to accurately and quantitatively determine analyte concentrations.101-102 Liquid chromatography-mass spectrometry (LC-MS) couples the separation capability of liquid chromatography (LC) with the detection specificity and versatility of mass spectrometry (MS). The latter provides information-rich mass spectra for each eluting chromatographic peak, beyond that which can be derived from ultraviolet-visible (UV-Vis) absorption detection alone.103-104 In LC-MS, ionization sources like electrospray ionization52 and atmospheric pressure chemical ionization105 are used to convert the liquid phase analytes into gas phase ions for detection in the mass spectrometer. A mass chromatogram or extracted ion chromatogram in LC-MS gives the intensity of an ion at a given m/z value as a function of elution time (no. of scans). The TICC sums all the mass chromatograms in a separation. The terms ‘mass chromatogram,’ ‘extracted ion chromatogram,’ and ‘total ion current chromatogram’ are favored by IUPAC, where the first two of these terms are completely synonymous.49 Figure 1.3 shows the
major components in a LC-MS instrument. Pumping pressures of several hundred atmospheres are often required for achieving reasonable flow rates.

![Block Diagram of LC-MS System](image)

**Figure 1.3.** A block diagram representing the major components in a liquid chromatography-mass spectrometry (LC-MS) system.

### 1.5 Capillary Electrophoresis (CE)

In the early 1990s, capillary electrophoresis (CE) was introduced as an alternative to traditional chromatographic separation methods. By utilizing voltages across a capillary, electrophoresis explores the differential migration rates of charged species in an applied direct current (DC) electric field for separation of analytes. CE is most effective in separation of large
molecules like proteins and peptides. CE is commonly used in the pharmaceutical industry and for DNA separation. CE is faster and cheaper than liquid chromatography. When compared to traditional slab electrophoresis, CE is much faster and provides higher resolution with lower sample volumes (0.1 to 10 nL). The mobility or migration rate of an ion (cm/s) is given as:

\[
\nu = \mu E \tag{1.3}
\]

where \( E \) (V cm\(^{-1}\)) is the electric field strength and \( \mu \) (cm\(^2\) V\(^{-1}\) s\(^{-1}\)) is the electrophoretic mobility. The size and shape of an ion determines its electrophoretic mobility. Neutral species are not separated by CE. Figure 1.4 shows an overview of the CE technique.

**Figure 1.4.** A block diagram of the capillary electrophoresis (CE) technique showing key components.
1.6 Chemometric Methods

I begin this section with an analogy. Figure 1.5 represents a ball rolling down an incline. We know from our basic physics that there is a harder way and an easier way to describe the motion of this ball. The harder way is to use the more standard Cartesian coordinate system represented by the $x_1$ and $x_2$ axes in Figure 1.5. The easier way is to describe the motion of the ball in the rotated coordinate system that is defined by the $x'_1$ and $x'_2$ axes. This latter coordinate system is more ‘natural’ because the path of the ball is parallel to the $x'_1$ axis, and perpendicular to the $x'_2$ axis. Thus, a more complex two-dimensional problem becomes a simpler one-dimensional problem of motion along one axis using a rotated coordinate system. This forms the basis of factor analysis methods. I am not the first to see or use this analogy (I believe it originated with Bonnie Tyler). This form of analysis has several advantages.
Figure 1.5. The motion of a ball down an incline showing the typical Cartesian coordinate system (axes $x_1$ and $x_2$) and the more computationally convenient, rotated coordinate system (axes $x_1'$ and $x_2'$).

In multivariate data analysis, a data matrix contains the information to be analyzed. Here, the rows represent spectra or samples and are referred to as objects and the columns represent wavelengths or properties of the data and are referred to as variables. Multivariate resolution methods like principal component analysis (PCA) and multivariate curve resolution (MCR) aim to project the data matrix into a space with fewer dimensions (line, plane or hyperplane). The decrease in the dimensionality of the projections represents a simplified and meaningful description of the original information in the data matrix.

The primary objectives for the application of chemometric methods are: the classification of spectra/samples, the clustering of spectra/samples and the development of calibration models to predict sample properties. In the analysis of ToF-SIMS data, multivariate methods address the following common questions: (1) How many analytes are present, i.e., what is $K$? (2) What are the pure component spectra, $s_k$, that represent the measured signal? (3) What are the corresponding contributions, $c_k$, of each analyte to the measured signal? (4) Which spectra are most similar and which most different? Question 1 is often addressed using PCA and Questions 2 – 4 can be answered using MCR. These models attempt to enhance the accuracy, sensitivity, and selectivity of data analysis by utilizing signals from multiple variables and multiple measured spectra. In the next few sections, I will provide a short description of each of these methods.

1.6.1 Principal Component Analysis (PCA)
In PCA, a set of data, e.g., spectra, is expressed in a different coordinate system, which is defined by the eigenvectors, a.k.a., principal components or factors, of the data matrix.\textsuperscript{112} The eigenvalues of these eigenvectors provide a quantitative measure of the amount of variance captured by each principal component. PCA can be viewed as plotting spectra as single points in a hyperspace and then rotating the original coordinate system of the data in a way that captures the largest amount of variance possible in the spectra (data points) along new axes as they are sequentially determined. The projections of the data points on the new axes (principal components) are the scores, and the loadings are the contributions of the original axes (variables) to the new axes.

1.6.2 Multivariate Curve Resolution (MCR)

The underlying assumption behind MCR is that Beer’s law is valid for multicomponent systems.\textsuperscript{44,114} Note, while in real systems Beer’s law may not be valid due to chemical and instrumental factors, we assume multicomponent systems as simple additive contributions from individual components. In contrast to PCA, MCR attempts to find the underlying pure component spectra for each analyte based on the fundamental physics and chemistry of the measurement. This is an appealing advantage over PCA, but it also means that the useful orthogonality properties of the scores and loadings of PCA are sacrificed. Beer’s law, a.k.a., the Beer-Lambert law, is a pillar in analytical chemistry and spectroscopy. It is simple to understand when described for a single analyte measured at a single wavelength and offers an intuitive basis for describing systems with multiple analytes measured at multiple wavelengths. As a result, the multicomponent version of
Beer’s law provides a useful metaphor that is familiar to analytical chemists for understanding and interpreting results from chemometrics tools such as classical least squares (CLS) and principal component analysis (PCA). MCR is a powerful class of methodologies based on the CLS model.

### 1.6.3 Cluster Analysis

Cluster analysis relies on the assumption that related spectra/data vectors will be closer in an \( n \)-dimensional space, i.e., similar samples will cluster.\(^{112-113}\) It is primarily an exploratory analysis method that is used for comparing samples based on changes in variables (properties). Spectra/data vectors are aggregated according to the similarity of their features/variables, which will define group memberships at different levels of aggregation. Measurement of distance is a common measure of the similarity between objects: objects with shorter distances of separation are assumed to be more similar. Several distance measures like the Euclidean and Manhattan distance are used. However, cluster analysis is severely dependent on variable scaling, i.e., variables with varying dimensions introduce false correlation.

### 1.6.4 Partial Least Squares Regression (PLS-R)

The fundamental purpose of PLS-R is to find factors (latent variables) that can capture the maximum variation present in a data matrix, \( \mathbf{X} \), for predicting some attribute of the samples, \( \mathbf{Y} \).\(^{115}\) Some of the common examples in chemistry are: \( \mathbf{X} = \) chemical composition and \( \mathbf{Y} = \) measurement of some properties, or \( \mathbf{X} = \) synthesis condition and \( \mathbf{Y} = \) quality parameters, etc. PLS-R is a
mathematical transform that attempts to establish a *cause and effect* relationship using a set of predictor variables (X) and response variables (Y), i.e., it attempts to maximize the covariance between X and Y. Unlike other regression models like multiple linear regression (MLR), PLS-R can analyze noisy data with large numbers of predictor variables. One of the key advantages of PLS-R is its ability to handle collinear predictor variables (X) which enables the analysis and modelling of complex real world data. This is of particular importance given that spectra like chromatographs and spectroscopic measurements have variables (e.g. time points or wavelengths) which are strongly correlated with each other.

### 1.7 Pattern Recognition Entropy (PRE)

PRE is a recent application of Shannon’s information theory that serves as a summary statistic and shape recognition tool for differentiating spectra. \(^{30, 48, 61, 63, 70}\) Shannon’s entropy \((H)^{46-47}\) is a measure of the uncertainty in the system and serves as a quantification of the total information present in a data stream. In both statistical thermodynamics and digital communications, entropy is a measure of the disorder/chaos/number of available states in a system. That is, information is defined as the distribution of the probabilities of a series of events in a message in its context. Based on Shannon’s theory, the information in a signal is quantified and referred to as entropy, or equivalently as information content (IC), which is a measure of the ultimate (data) compression a signal can undergo. Over the past six decades, researchers have exploited the potential and power of his theory, applying it to data storage,\(^{116}\) data compression,\(^{117}\) digital communication,\(^{118}\) and more specifically for MP3s,\(^{119}\) JPEGs,\(^{119}\) mobile phones, optical communications,\(^{120}\) satellite communication, space exploration programs,\(^{120}\) plagiarism
detection,\textsuperscript{121} pattern recognition/detection,\textsuperscript{122} etc. Shannon’s theory has profoundly influenced our world. In fact, his initial paper has been cited nearly 100,000 times. PRE is a modification of Shannon’s entropy where ‘pseudo-probabilities’ in a spectrum/chromatogram are obtained by normalizing the data with the 1-Norm. In PRE, an entire spectrum is treated as a probability distribution to obtain a summary statistic that characterizes it. However, we emphasize that spectra collected in typical science experiments are not probability distributions, at least not in the classical sense. Accordingly, PRE is neither employed in my work to make any statement about probabilities of signals, peaks, spectral features, or noise in spectra. Rather, it is used as a pattern recognition tool because it is sensitive to and can differentiate between spectra with different shapes, where, as will be discussed below, the spectral ‘shape’ is a result of contributions from all the parts of a spectrum – noise, baseline, and signals. Spectra with more features have higher PRE values and vice versa.

1.8 References


(divinylbenzene) microspheres as the core materials. *Journal of separation science* 2013, 36 (24), 3821-3829.


44. Tauler, R., Multivariate curve resolution applied to second order data. *Chemoscmetrics and intelligent laboratory systems* 1995, 30 (1), 133-146.


57. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass


Chapter 2: Theory of Chemometric Methods and Pattern Recognition Entropy (PRE)

The fundamental aim of chemometric methods is relating measurements made on a chemical instrument to the nature of the chemical system that is being studied by the application of mathematical and statistical methods. The starting point of any chemometric analysis is the data which is in the form of a table. Generally, the rows in the table represent individual samples/spectra while the columns represent properties/variables.\(^1\) Table 2.1 is a good example of such a data structure. It contains water contact angle\(^2\) \(x_1\)/ellipsometric thickness\(^3\) \(x_2\) data for a series of surfaces. Over the years, most analytical measurements have transformed into a multivariate nature, i.e., they are multiple measurements relating to a single sample. Under such situation, statistical methods that are classified into the branch of multivariate analysis are primarily useful.

<table>
<thead>
<tr>
<th>Spectrum</th>
<th>(x_1), e.g., value of water contact angle (°)</th>
<th>(x_2), e.g., value of film thickness (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>S2</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>S3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>S4</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>S5</td>
<td>21</td>
<td>11</td>
</tr>
<tr>
<td>S6</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>S7</td>
<td>36</td>
<td>17</td>
</tr>
<tr>
<td>S8</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>S9</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>S10</td>
<td>37</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 2.1. Mock data from surfaces with different water contact angles and thicknesses.
In the first step of any chemometric analysis, the data is reviewed for completeness and false correlations.\textsuperscript{4-5} The collection of all these methods is referred to as \textit{preprocessing}. In case of missing data, it is a good practice to fill the vacancies with the average value of a column/row. \textit{Mean centering} results in the centering of each variable (column) by subtracting the average column value from each of the data points in the column. Note, this operation preserves the spacing between variables and data points, i.e., it doesn’t change the correlation structure of the data. Another common preprocessing method is \textit{range scaling}. This operation is mostly performed when the variables under consideration have widely variable ranges (variances). In range scaling, each variable (column) is scaled by its variance. In many cases, mean centering and range scaling are performed in a single operation called \textit{autoscaling}. \textit{Normalization} (1-Norm) is used in cases where there exists variability in the samples/ objects (rows). A classic example maybe XPS spectra collected from two different spots on the same sample on different days. In this case, while the instrument settings maybe assumed to be constant, there can exist a drift in the instrument parameters during the data collection period. In 1-Norm, each value in each row is scaled such that the sum of the values in each row is unity.

2.1 Principal Component Analysis (PCA)

PCA is ubiquitous in multivariate data analysis, and the statistics of PCA are well defined.\textsuperscript{1, 6-8} PCA is useful in exploratory analysis because it is a variable reduction tool, thus allowing trends
and patterns to be more easily identified. For a data set with each row corresponding to a measured
spectrum, $X$ (size $M \times N$), the PCA decomposition is given as

$$ (2.1) \quad X = TP^T + E $$

where $T$ is an $M \times K$ matrix of scores. Each column in $T$ corresponds to the $k^{th}$ principal component
(PC) for $k = 1, \ldots, K$, and each row $m = 1, \ldots, M$ corresponds to an individual measured spectrum.
Each spectrum is often referred to as an object or individual sample. Thus the scores in PC show
relationships between samples. $K$ is the number of PCs in the model. The columns of $P$ ($N \times K$)
are the corresponding loadings vectors that capture relationships between variables, e.g., the mass
channels in ToF-SIMS. Thus the columns in the scores and loadings matrices are collections of
individual scores and loadings vectors given by

$$ (2.2) \quad T = [t_1 \ t_2 \ \ldots \ t_k] ; \ P = [P_1 \ P_2 \ \ldots \ P_k] $$

The PCs are obtained by maximizing capture of variance. For example, the first PC is obtained by proposing that $t_1 = Xp_1$ and maximizing $t_1^T t_1$ subject to $\|p_1\| = p_1^T p_1 = 1$:

$$ (2.3) \quad \max_{\|p_1\|=1} \{t_1^T t_1\} = \max_{\|p_1\|=1} \{p_1^T X^T X p_1\}. $$

Thus a 1-PC model is

$$ (2.4) \quad X = t_1 p_1^T + E $$

Where $\hat{X} = t_1 p_1^T$ is a model of the data, which can be viewed as the best rank-one estimate of the
original data because it captures the greatest possible variance in the data set. While $p_1$ can be
viewed as the “most common spectrum”, the entries of the scores vector, $t_1^T = [t_{1,1} \ t_{2,1} \ \ldots \ t_{M,1}],$
describe “how much of \( p_1 \)” is in each sample. Typically, \( K \) is larger than one but much smaller than \( \min(M,N) \), i.e., considerable variable reduction takes place, leading to

\[
X = t_1 p_1^T + t_2 p_2^T + \cdots + t_K p_K^T + E = TP^T + E = \hat{X} + E
\]

\( K \) is most often selected to capture systematic variance in the data while leaving noise in the residuals, \( E \). This means that \( \hat{X} = TP^T \) is a “compressed” and “noise-filtered” approximation of the original data, \( X \), where only the \( K \) columns of \( T \) need to be inspected to find trends and patterns in the samples, and the \( K \) columns of \( P \) need to be examined to identify which variables are associated with those trends. The PCs are ordered so that the first PC captures the most variance and each successive PC captures less: \( t_1^T t_1 \geq t_2^T t_2 \geq \cdots \geq t_K^T t_K \). Note that if \( K \) is equal to the mathematical rank of \( X \), \( \text{rank}(X) \leq \min(M,N) \), there is no data compression and the original data set can be re-described perfectly.

It was stated above that the first loadings vector could be viewed as the “most common spectrum.” Although this is a useful metaphor, it is not quite correct and a couple of caveats are required. First, note that the scores and loadings have the useful mathematical property that \( T^T T \) is a diagonal matrix and \( P^T P = I \), where \( I \) is the identity matrix (a diagonal matrix of ones). This means that the loadings are orthogonal such that

\[
\begin{align*}
    p_k^T p_j &= \begin{cases} 1 & \text{for } k = j \\ 0 & \text{for } k \neq j \end{cases} \\
    t_k^T t_j &= \begin{cases} \lambda_k & \text{for } k = j \\ 0 & \text{for } k \neq j \end{cases}
\end{align*}
\]

In Equation (2.6), the eigenvalue \( \lambda \) is introduced where the loadings are eigenvectors with associated eigenvalues \( \lambda \) for \( X^T X p_k = \lambda_k p_k \). However, this property is not necessarily useful for interpreting the scores and loadings in PCA. Because the loadings are orthogonal, they are not pure component spectra. Instead, the loadings are linear combinations of pure component spectra.
This property can make PCA results difficult to interpret, but with some practice it is possible. The good news for XPS and ToF-SIMS is that due to the high selectivity in the measurements, the underlying pure component spectra of the individual analytes don’t overlap as severely as they do for some analytical measurements, e.g., near-infrared spectroscopy. As a result, the first few PCA loadings can, at times, be close approximations to pure component spectra.

A second caveat on interpreting loadings has to do with data preprocessing. For example, often the data are first mean-centered in PCA. This means that the loadings need to be interpreted as directions of deviation away from the multivariate mean. In contrast, no mean centering is a force fit through zero, and the first PC will represent an “average” spectrum (note: with no centering the first PC is not the mean but it may point in the general direction of it). Other types of preprocessing, e.g., variable scaling, often influence how the data are to be interpreted. Knowing the math used to process the data facilitates interpretation of the results. In this work, the data were only processed to the 1-norm, i.e., each row of X (each measured spectrum) was processed such that \( \tilde{x}_{m,n} = \frac{x_{m,n}}{\sum_{n=1}^{N} |x_{m,n}|} \). The tilde, ~, is used to indicate preprocessed data, but it is dropped below to ease interpretation of the subsequent text. That is, in each case the preprocessing is discussed but not indicated explicitly.

2.2 Multivariate Curve Resolution (MCR)

For a single analyte, the dimensionless absorbance at the \( n^{th} \) wavelength, \( a_n \), is given by Beer’s law as
(2.7) $a_n = \varepsilon c l$

where $\varepsilon_n$ is the molar attenuation coefficient (L·mol$^{-1}$·cm$^{-1}$), $c$ is the analyte concentration (mol·L$^{-1}$), and $l$ is the path length through a measured sample (cm). Equation (2) represents a linear relationship between $a_n$ and the analyte concentration, $c$, and also a linear relationship between $a_n$ and the molar attenuation coefficient, $\varepsilon_n$. For constant path length, the definition $s_n = \varepsilon_n l$ can be used to simplify the model so that for $N$ wavelengths ($n = 1, \ldots, N$) the model can be written as

(2.8) $a = sc$

where $a^T = [a_1 \; a_2 \ldots a_k]$ is the measured absorbance spectrum, and $s$ is the analyte spectrum at unit concentration and path length $l$.

We next consider the case where multiple analytes are present. For $K$ analytes, the linear model in Equation (2.8) can be extended to become a linear mixture model given as

(2.9) $a = s_1 c_1 + s_2 c_2 + \cdots + s_K c_K$

where the absorbance is modeled as the sum of the absorbances from each analyte for $k = 1, \ldots, K$ analytes. Equation (2.9) is easily interpreted; the measured absorbance, $a$, is comprised of $K$ individual pure analyte spectra, $s_k$, each with amount $c_k$. For example, Figure 2.1 shows a ToF-SIMS peak measured from 58.00 – 58.25. The measured peak is the sum of two underlying pure analyte spectra, i.e., $K = 2$, that each contribute to the peak.
Figure 2.1. A measured ToF-SIMS peak (black) represented as the sum of two pure analyte spectra (blue and red).

To simplify the notation in Equation (2.9), the columns of spectra can be gathered into a single $N \times K$ matrix $S = [s_1 \ s_2 \ ... \ s_K]$, and the corresponding analyte concentrations can be collected into a single vector $c^T = [c_1 \ c_2 \ ... \ c_K]$. The result is the familiar form of the CLS model given by

$$a = Sc$$

When used in chemometrics and statistical analysis, this model is most often written with an ‘error’ term, $e$, that represents the difference between what is measured ($a$) and the model yielding
\[ a = Sc + e \]

Equation (2.11) has been widely used in a variety of applications because its statistics are well understood and it is easy to interpret. In many cases, for example in ToF-SIMS, the \( c_k \) are not strictly interpretable as concentrations as in Beer’s law but instead represent the contribution of each analyte’s response, \( s_k \), to the overall measured signal given by \( a \). Thus, the CLS model, Equation (2.11), is an enabling tool for analysis and interpretation of data.

To avoid confusion with the absorbance measurements given in Beer’s law, measured XPS and ToF-SIMS spectra are represented by \( N \times 1 \) column vectors, \( x \), and collections of spectra (e.g., depth profile measurements) are arranged in an \( M \times N \) matrix \( X \). The corresponding CLS representation of multiple measurements is written concisely as

\[ X = CS^T + E \]

This equation is no more difficult than Equation (2.11) but it does require some explanation. Each row of \( X \) is a measured spectrum (i.e., it is the transpose of a single measured spectrum given by column vector \( x_m \)), and the rows of the \( M \times K \) matrix \( C \), \( c_m^T \), correspond to contributions for a single point in a depth profile or for a single pixel of an image where \( m = 1, \ldots, M \). Equation (2.12) provides a succinct representation of the data model as a set of pure component spectra, \( S \), and corresponding contributions, \( C \). It also has the same form as the PCA model, but there are some important differences. For example, the contributions and spectra are not generally orthogonal, i.e., \( C^T C \) and \( S^T S \) are not generally diagonal matrices. Additionally, as discussed below, \( C \) and \( S \) are typically restricted to non-negative entries. This is physically reasonable since measured spectra should be non-negative, with the exception of noise or measurement artifacts.
Least-squares estimation is commonly used to minimize $\mathbf{E}$ in Equation (7), i.e., it minimizes the squared Frobenius norm given by $\text{trace}(\mathbf{E}^T\mathbf{E})$. For example, if the pure analyte spectra (columns of $\mathbf{S}$) are known, the estimated contributions, $\hat{\mathbf{C}}$, can be calculated using the least-squares result\(^\text{10}\)

(2.13)  \[ \hat{\mathbf{C}} = \mathbf{X} \mathbf{S} (\mathbf{S}^T \mathbf{S})^{-1} \]

Although many of the mathematical details are beyond the scope of the present discussion, Equation (2.13) is an important result and deserves comment. The calculation here can only work if the matrix inverse $(\mathbf{S}^T \mathbf{S})^{-1}$ exists, i.e., it doesn’t make sense to divide by zero. Mathematically, this means that $\mathbf{S}$ must be of full column rank, but intuitively this means that each of the $\mathbf{s}_k$ must look different from each other, i.e., they must have different shapes.\(^\text{11}\) This means that each $\mathbf{s}_k$ must have different peaks or peak ratios so that the spectral fingerprints are different. To see this, imagine that two spectra look exactly the same. Under this restriction there is no good reason to include more than one identically looking spectrum in Equation (2.9). However, if the spectra are different and the estimate in Equation (2.13) can be made, $\hat{\mathbf{C}}$ corresponds to estimated contributions for each measured spectrum in $\mathbf{X}$. For example, in depth profile measurements, the columns of $\hat{\mathbf{C}}$ represent depth profiles for individual analytes yielding chemical information as a function of depth. Although highly selective mass channels can be used to estimate depth profiles, estimates from Equation (2.13) generally provide better signal-to-noise because it is “averaging” signal from many mass channels. This is a significant advantage of the CLS model.\(^\text{12}\) However, it is often difficult to know $\mathbf{S}$, the pure component spectra, \textit{a priori}. An additional complication here is that spectral libraries can be equivocated by sample matrix effects and instrument-to-instrument differences. In these cases, multivariate curve resolution (MCR) can be used to provide estimates of both $\mathbf{C}$ and $\mathbf{S}$. 
In Equation (2.12), it is also known as end-member extraction and blind source apportionment. There are a number of algorithms for estimating \( C \) and \( S \) in the MCR model, and MCR is used when \( C \) and \( S \) are unknown and both need to be estimated. One of the most common and easy to understand algorithms is the constrained alternating least-squares (ALS) method. This algorithm is initiated with an initial guess of pure analyte spectra, \( S \), given by \( \hat{S}_0 \).

Initial guess \( \hat{S}_0 \)

for \( j = 1: J \)

(a) \[
\hat{C}_j = XS_{j-1}(S_{j-1}^T \hat{S}_{j-1})^{-1}
\]

(b) \[
\hat{S}_j^T = (\hat{C}_j^T \hat{C}_j)^{-1} \hat{C}_j^T X
\]

(c) for \( K = 1: K \)

\[
\hat{s}_{k,j} = \hat{s}_{k,j} / \| \hat{s}_{k,j} \|^{1/2}
\]

end

\[
E_j = X - \hat{C}_j \hat{S}_j^T
\]

\[
e_j^2 = \text{trace}(E_j^T E_j)
\]

end

This algorithm runs over a total of \( J \) iterations but it could be exited sooner if \( e_j^2 \) or \( e_{j-1}^2 - e_j^2 \) are below set tolerances. A few other important points about the ALS algorithm are required. It should be pointed out that comments on the inverse given above for Equation (2.13) also apply
here for Steps (a) and (b). This means that the components in the profiles in \( \hat{C}_j \) and the spectra in \( \hat{S}_j \) must all have different shapes. This can be a problem with some analytical measurements with highly overlapped peaks, but high resolution ToF-SIMS tends to have peaks, or some portion of each analyte’s spectrum, that is unique. As a result, MCR is often fairly easy to use with ToF-SIMS data sets and yields good estimates for contributions and spectra.

Step (c) maintains the estimated spectra to be “unit length” such that \( \hat{s}_k^T \hat{s}_k = 1 \). This convention is necessary when spectral magnitudes are unknown, but it leads to the important realization that there exists a magnitude ambiguity in the MCR decomposition such that for any scalar \( a \neq 0 \), the solutions \( c_k s_k^T = (c_k a)(a^{-1} s_k^T) \) both result in an equivalent model fit. Thus, relative contributions within a profile can be compared but not between factors.\(^{13}\)

As written, the least-squares estimates in Steps (a) and (b) in the algorithm above are unconstrained fits to the data, \( X \), to the estimated spectra and contributions. However, the physics of the measurement system dictates that both contributions and spectra must be non-negative such that \( C \geq 0 \) and \( S \geq 0 \). Therefore, Steps (a) and (b) incorporate non-negativity constraints in the least-squares estimates.\(^{14}\) In fact, the non-negativity constraint is often necessary for ALS to provide an unambiguous decomposition.\(^{15-16}\)

### 2.3 Cluster Analysis

Cluster analysis\(^{4-5}\) is an unsupervised method that aims at combining objects (observations/samples) into groups based on the degree of similarity and difference between them. It is primarily an exploratory analysis tool that is used to understanding classes and compare
observations/samples. Cluster analysis has been widely used for understanding meteorological and climatological phenomena as well as for weather forecasting.\textsuperscript{17-19} Cluster analysis is fundamentally different from discriminant analysis tools in the fact that it does not require any rules for assigning class/group memberships.

The central aspect of cluster analysis is the distance measure, i.e., the shorter the distance between two objects, the more similar they are in their properties/behavior. In general, we define cluster as an aggregation of objects that are separated by small distances as compared to distances between the clusters. A general distance measure is the Minkowski distance (Equation 2.13).

\[
(2.14) \quad d_{ij} = \left[ \sum_{k=1}^{K} |x_{ik} - x_{jk}|^p \right]^{1/p}
\]

where $K$ is the number of variables/properties and $i$ and $j$ are the indices for the objects $i$ and $j$. The most common distance measure is Euclidean distance (Equation 2.14) for $K$-dimensions for which $p=2$. Other distance measures like Karl Pearson distance, Manhattan distance, and Mahalanobis distance are frequently used depending on applications.

Hierarchical clustering is the most common clustering method used. In this method, groups are based on hierarchy where objects are combined by merging one pair from existing groups based on similarities or differences between the objects. In most cases, agglomerative clustering is used where single objects are combined in each step to form clusters of objects. Let us consider an example of a data matrix having $n$ objects. Assuming there is no group structure initially, we start with $n$ groups, each having a single observation. In the next step, two of the groups which are closest to each other in the $K$-dimensional space are combined to form a single group. This process is repeated ($n-1$) times till all the objects have been accounted for. For combining the groups at
each step, similarity measures like Single-linkage, Complete-linkage, Average-linkage, and Centroid clustering are used. In our work, we use Ward’s minimum variance method that merges two groups that minimize the within-group sum of squares error. The sum of squares error is defined as the sum of squared distances between the objects and their respective group centroids which is summed over all groups. A dendrogram is a visual representation of results from a cluster analysis. Let us consider the following example in Figure 2.2. We start at the left corner where in the first step we have 5 separate clusters from the 5 objects. In the first step, objects $x_1$ and $x_2$ are combined with the vertical line connecting the two objects in the dendrogram being proportional to their distance of separation. At the next step, objects $x_3$ and $x_4$ are combined in a similar fashion. The process is repeated until all the objects have been assigned a group membership.

![Figure 2.2. A dendrogram illustrating results from a cluster analysis of 5 objects.](image)
Partial Least Squares Regression (PLSR) is a widely used soft-modeling technique that combines features from principal component analysis and multi component regression. The primary objective of PLSR is prediction of a set of dependent variables $Y$ from a set of predictor variables $X$. One of the key advantages of PLSR is the ability to model multiple response variables simultaneously. PLSR originated in 1975 by Herman Wold for application in the social sciences. The PLSR model is developed from a training set of $X$ and $Y$ matrices and is subsequently used for prediction purposes. A key disadvantage of PLSR is the high dependence on scaling, i.e., scaling can result in emphasis on specific $Y$-variables.

As described by Wold in his paper, the PLSR model attempts to establish a set of “new” latent variables (LVs) which are estimates of the original LVs from the principal component decomposition of the data matrices. In the widely used NIPALS algorithm, scores $U$ and loading $Q$ are calculated for the $Y$ matrix along with scores $T$ and loadings $P$ for the $X$-block. An additional set of vectors called weights $W$ are also calculated. The column (variable) in $Y$ with the greatest variance is chosen for an initial estimate of $w_1$ which is defined as:

\[
(2.16) \quad w_1 = \frac{X^T u_1}{||X^T u_1||}
\]

\[
(2.17) \quad t_1 = Xw_1
\]

For the $y$ data block, the first loading ($q_1$) is given as:

\[
(2.18) \quad q_1 = \frac{y^T t_1}{||y^T t||}
\]
(2.19) $u_1 = Yq_1$

At the end of Equation 2.17, convergence is checked by comparing values of $t_1$ between the current and previous iterations. Once convergence is achieved, the algorithm proceeds to Equation 2.18 and calculates the scores, loadings and weights for the X-block.

(2.20) $p_1 = \frac{X^T t_1}{||t_1^T t||}$

(2.21) $p_{1_{\text{new}}} = \frac{p_{1_{\text{old}}}}{||p_{1_{\text{old}}}||}$

(2.22) $t_{1_{\text{new}}} = t_{1_{\text{new}}} ||p_{1_{\text{old}}}||$

(2.23) $w_{1_{\text{new}}} = w_{1_{\text{new}}} ||p_{1_{\text{old}}}||$

At the end of equation 2.23, the scores and loading for the first LV are calculated and the algorithm proceeds towards calculating the residuals for the X and Y- blocks which are given by E and F matrices. Equations 2.24 and 2.25 describe the E and F matrices after the first iteration.

(2.24) $E_1 = X - t_1 p_1^T$

(2.25) $F_1 = Y - b_1 t_1 q_1^T$

where $b_1$ is the regression coefficient and is defined as:

(2.26) $b_1 = \frac{u_1^T t_1}{u_1^T t_1}$

Next, Equations 2.16 through 2.19 are repeated for each latent variable where $X$ and $Y$ are replaced with $E$ and $F$ after each iteration. Note, Equations 2.16 to 2.26 are adapted from the PLS manual
2.5 Pattern Recognition Entropy, PRE

Information theory was powerfully influenced by Claude Shannon, who developed a mathematical description for communication through the application of the entropy function from statistical thermodynamics. Based on his proposal, the information in a signal is quantified and referred to as entropy, or equivalently as information content (IC), which is a measure of the ultimate (data) compression a signal can undergo. Shannon’s theory profoundly influenced our world, as evidenced by the fact that his initial paper has been cited nearly 100,000 times.

In statistical thermodynamics, the entropy of mixing of an ideal solution is given by:

\[ \Delta S_{\text{mix}} = -R \sum_{k=1}^{c} X_k \ln X_k \]  

where \( R \) is the gas constant, and the \( X_k \) values are the mole fractions of the components to be mixed. Alternatively, the Gibbs formula for entropy is:

\[ \Delta S = -k_B \sum_i p_i \ln(p_i), \]  

where \( k_B \) is Boltzman’s constant, and \( p_i \) is the probability that a microstate will be sampled. Taking his lead from these types of equations, Shannon proposed the following formula to describe the entropy of a communication signal:

\[ H(x_i) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i) \]
where the signal is considered to be a probability distribution with \( p(x_i) \) the probability of observing a signal at a certain channel, and the sum of the probabilities for a signal summing to unity. Equation (2.29) uses the base 2 logarithm, and, accordingly, the final result is given in units of ‘bits’.

A classic example of Shannon’s formula comes from the toss of a fair coin, which, of course, has a probability, \( p \), of \( \frac{1}{2} \) for occurrences of either heads or tails. Inserting these probabilities into Equation (2.29) gives:

\[
(2.30) \quad H_{\text{fair coin}} = -\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{2}\log_2 \frac{1}{2}\right) = 1 \text{ bit}
\]

However, in the case of an unfair coin that would always give either heads or tails, the \( p_i \) values are 1 and 0, which, in Shannon’s formula, give:

\[
(2.31) \quad H_{\text{unfair coin}} = -(0\log_2 0 + 1\log_2 1) = 0 \text{ bit}
\]

Here we have taken: \( 0 \log 0 = 0 \), which is true in the limit of \( p_i \to 0 \). The entropies (information contents) for all such possibilities, ranging from a completely fair coin to an entirely unfair one, as calculated with Equation 2.29, yield the classic plot shown in Figure 2.3. Clearly, the entropy (uncertainty) in the measurement is greatest for the fair coin, and it goes to zero for an entirely unfair coin for which, again, the outcome of the coin toss is certain. Alternatively, we can interpret the entropies of the coin tosses in Equation 2.29 as saying that it requires one bit of information (a zero or a one) to represent the outcome of a fair coin toss (Equation 2.30), while it does not require any bit of information to represent the outcome of the entirely unfair coin (Equation 2.31).
Figure 2.3. Shannon entropies from coins with a range of probable outcomes. This plot can be interpreted as follows. Left: an unfair coin that never yields a desired outcome (p = 0). Middle: a fair coin that has even odds of heads or tails (p = ½). Right: an unfair coin that always produces the desired outcome (p = 1, right).

In this work we take Shannon’s formula and bring it back into science as a tool for data analysis. In particular, we consider a spectrum or a data set to be a probability distribution that can be inserted into Equation 2.29. Because of the ability of this approach to distinguish between shapes of spectra, we have chosen to use PRE (pattern recognition entropy) to refer to the results from Equation 2.29, and not ‘H’, which can be confused with enthalpy. Among other things, PRE seems well suited for analyzing depth profiles in surface and interface analysis. Here, XPS and ToF-SIMS spectra will first be normalized by dividing each energy or m/z value by the sum of all the values in the corresponding spectrum. The resulting ‘probabilities’ are then entered into Equation 2.29. This procedure should also work well with selected peaks (integrated areas) from
spectra. As evident from Equation 2.29, a larger PRE value typically points to a more complex data set, i.e., one that has a larger number of peaks of comparable size, while a spectrum that may be strongly dominated by one peak, such that \( p_i \) approaches 1, will have a lower PRE value. Obviously, Shannon’s use of the base 2 logarithm in Equation 2.29 is not necessary here – other bases should work equally well for comparing spectra. However, for consistency with information theory, we will use the base 2 logarithm here, i.e., the units of all PRE calculations made herein will be in ‘bits’.

To illustrate PRE analysis, Equation 2.29 was applied to the series of mock spectra in Figure 2.4. Figure 2.4a shows the PRE values of three horizontal baselines. Here, these functions (straight lines) have constant values. Accordingly, the ‘probability’ of any point occurring must be the same, i.e., it is \( 1/N \), where \( N \) is the number of data points, and the sum of the probabilities is 1. The calculation for \( N = 1000 \) is provided here in Equation 2.32.

\[
(2.32) \quad \text{PRE} = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = 9.9658 \quad \text{(for } N = 1000) 
\]

The bottom line in Figure 2.4a with \( y = 0 \) is admittedly artificial. If the probabilities are zero everywhere in a distribution, then \( \text{PRE} = 0 \). However, technically speaking, Equation 2.29 should not be applied to this, or any, series of zeros that constitute a spectrum because the sum of these values is not unity – finding the ‘probability’ here would require division by zero. In general, real measurements have noise/uncertainty. And, of course, if necessary, a very small positive offset could be added to the elements of a data set if it includes a spectrum composed entirely of zeros. Figure 3b shows the PRE values for three functions that contain one, two, or three spikes of equal height. Normalization gives the first function a probability of 1 for the spike, and probabilities of zero everywhere else. These values yield \( \text{PRE}_i = 0 \), which is analogous to the situation for the
entirely unfair coin considered above. For the second spectrum, normalization gives two spikes of equal magnitude, i.e., equal probability of \( \frac{1}{2} \), which gives \( \text{PRE}_{ii} = 1 \). This situation is obviously analogous to that of the fair coin toss. For the three spikes of equal probability in Figure 2.4b, we obtain, as expected, a higher PRE value of 1.585 (see Equation 2.33).

\[
(2.34) \quad \text{PRE} = - \sum_{i=1}^{3} \left( \frac{1}{3} \ast \log_2 \frac{1}{3} \right) = 1.585
\]

Thus, increasing the number of features in a spectrum generally increases its PRE value. Figure 2.4c shows two noisy baselines, which differ only by a multiplicative constant. When normalized, they have the same values (probabilities), and therefore the same PRE values. Figure 2.4d shows three functions. The first (i) is a rather narrow peak on a flat baseline. This spectrum has a relatively low PRE value. The second function (ii) is broader, and, as expected, it has a larger PRE value. The third function is the same as a second, except that it has a shoulder at its left side. As expected, it has an even higher PRE value. Figure 2.4e(i) shows a peak with a relatively low amount of noise and its corresponding PRE value. The peak above it is identical to the first, except noisier. Accordingly, this spectrum has a higher PRE value. Finally, a third spectrum (Figure 2.4e(iii)) was produced by taking every third data point from the second spectrum (Figure 2.4e(ii)). As expected, this spectrum shows a lower PRE value. Thus, in general, noisier spectra have higher PRE values, and there is an increase in PRE with an increasing number of data points (we can better understand this statement by considering that more data points generally means more information). These conclusions follow from Equation 2.29. Finally, Figure 2.4f shows two spectra represented by either a solid or a dashed line and their PRE values. As expected, the spectrum with the broader peaks of more comparable size (represented by the dashed line) has a somewhat higher
PRE value. This figure also shows the PRE value of the spectrum obtained by combining the peaks from both spectra into a single spectrum. It is higher than those of the individual spectra.

Figure 2.4. Simulated spectra and their PRE values.

We end this section with an analogy between the PRE value and color. Perhaps it may seem foolish to attempt to represent an entire spectrum with a summary statistic, i.e., a single number. Such a process leads to a significant reduction in information, and, in general, it should be difficult,
if not impossible, to recreate an original spectrum from just its PRE value. Nevertheless, a comparison to our everyday lives may help motivate and justify this approach. The visible spectrum associated with almost any colored object generally contains a series of absorbances and is at least moderately complex. The combination of these features gives an object the color we see/perceive. However, in spite of the fact that our brains reduce a significant amount of spectral information to just one color, the color of an object is still an incredibly useful characteristic of an object – we categorize and identify many objects based on their colors. Similarly, as we will show, PRE values of spectra can often distinguish between them quite well, and in many cases, PRE values are sensitive to relatively small differences between spectra. Of course the PRE value of a spectrum will not be an adequate substitute for it in all situations any more than a single color is a fully satisfactory optical description of a material. Nevertheless, as will be shown below, PRE values can often be quite successful in identifying differences between and characterizing spectra.

2.6 References

10. This result is obtained by minimizing the trace of $E^TE$.
11. More generally, it means that no pure analyte spectrum can be a linear combination of the other pure analyte spectra. The concept of matrix rank is important and central to chemometrics and design of experiments.
12. Sensitivity can also be enhanced using CLS and multivariate weighting strategies such as generalized least squares.
13. Note that for decompositions such as MCR and PCA, the profiles and spectra are generically referred to as “factors.”.
14. Under special considerations these constraints can be relaxed.
24. This analogy between color and the Information Content was originally drawn by Roberto Raso at Kratos, M., UK.
Chapter 3: A Perspective on Two Chemometrics Tools: PCA and MCR, and Introduction of a New One: Pattern Recognition Entropy (PRE), as Applied to XPS and ToF-SIMS Depth Profiles of Organic and Inorganic Materials

3.1 Introduction

The *Surface* of any material governs many of its physical and chemical properties especially interactions with its surroundings. For example, catalysis, tribology, wetting, adhesion, corrosion, adsorption, biological signaling and transport, separation science, device failure, and sensing often depend on what is happening in the outermost 0.1 – 1 nm of a material. Different analytical techniques provide information about surfaces at different length scales. Very often it is a combination of analytical techniques that best reveals the composition and nature of a surface. Each of the surface sensitive techniques (XPS, ToF-SIMS, AES, LEIS) can be coupled with a sputter gun for depth profiling. These depth profiles greatly increase the ‘reach’ of these analytical methods. For example, dopant profiles in silicon have long been determined by ToF-SIMS depth profiles.

In a traditional analysis of an XPS or ToF-SIMS depth profile, one might plot the signal from one element or molecular fragment, or perhaps the ratio of two signals as a function of sputtering time (depth into a material) to determine the compositions of layers. For example, in a
depth profile through a film of SiO$_2$ on Si, one might plot the oxygen signal from the corresponding O 1s XPS narrow scans, the O/Si ratio determined from the O 1s and Si 2p narrow scans, or perhaps the chemically shifted oxide component of the Si 2p narrow scan as a function of sputter time. Certainly, at some level, this is not unreasonable, and useful insights into a material can be gained this way. However, this approach will typically miss useful chemical information about a material because it focuses on a subset of the available data. In the case of ToF-SIMS, simply focusing on a peak or two in the spectra in a depth profile will omit the vast majority of the information collected.

In this work, we discuss the statistical analysis of XPS and ToF-SIMS depth profiles. In particular, we focus on three chemometrics methods that analyze/shed light on whole spectra or large fractions of spectra. The first tool we consider is principal component analysis (PCA), which is probably the most widely used chemometrics/bioinformatics method employed today. The second technique is multivariate curve resolution (MCR). The third data analysis tool discussed herein is novel. It is the pattern recognition entropy (PRE) of a spectrum. This approach has its roots in the groundbreaking work of Shannon, who applied the statistical thermodynamics definition of entropy to signal/communication theory. PRE takes Shannon’s approach in a new direction and is used here to identify transitions in depth profiles. PRE values can be useful in understanding data sets because they are a reflection of the shapes and complexity of spectra. We recommend that PRE be undertaken before PCA and MCR because it is so easy to perform and interpret. The results from PRE can then guide the more complex analyses that follow. Each of the mathematical techniques are applied to four data sets: the XPS depth profile of a film of SiO$_2$ on Si, the XPS depth profile of a film of Ta$_2$O$_5$ on Ta, the ToF-SIMS depth profile of ca. 100 nm of
plasma polymerized C$_3$F$_6$ on Si, and the ToF-SIMS depth profile of ca. 100 nm of plasma polymerized PNIPAM (poly(N-isopropylacrylamide)) on a Si substrate.

Note that researchers have previously used the concept of mutual information for data analysis,\textsuperscript{6-10} which is based on Shannon’s definition of information. However, it should not be confused with the PRE application developed here. Mutual information is a type of metric used to compare variables with regards to a model. It can be used to identify which variables capture the information and/or additional information when measuring or empirically assessing a model. In contrast to this other approach, the Shannon information used herein is a summary statistic that captures the pattern and complexity of a spectrum.

3.2 Experimental

3.2.1 XPS Depth Profiles of Si/SiO$_2$ and Ta/Ta$_2$O$_5$

XPS was performed with a Physical Electronics Quantera Scanning X-ray Microprobe instrument. This system uses a focused monochromatic Al K$_\alpha$ X-ray (1486.7 eV) source for excitation, a spherical section analyzer, and a 32 element multichannel detection system. A 70 W, 200 μm X-ray spot was used for the sputter depth profiles. The X-ray beam is incident normal to the sample and the photoelectron detector is at 45° off-normal. High energy resolution spectra were collected using a pass-energy of 69.0 eV with a step size of 0.125 eV. For the Ag 3d$_{5/2}$ line, these conditions produced a FWHM of 0.91 eV. The binding energy (BE) scale is calibrated using the Cu 2p$_{3/2}$ feature at 932.62 ± 0.05 eV and Au 4f$_{7/2}$ at 83.96 ± 0.05 eV. Low energy electrons at
ca. 1 eV (20 μA), and low energy Ar\(^+\) ions were employed to minimize the variable degrees of charging that were observed. Ar\(^+\) ions (2 kV) were used for the depth profiles with an ion gun incidence angle of 45°, a polar angle of 60°, and an azimuthal angle of 90°. A total of 54 and 37 sputter cycles were performed on the Si/SiO\(_2\) and Ta/Ta\(_2\)O\(_5\) samples, respectively.

### 3.2.2 ToF-SIMS Depth Profiling

ToF-SIMS depth profiles of (i) ca. 100 nm of plasma polymerized C\(_3\)F\(_6\) and (ii) ca. 100 nm of plasma polymerized PNIPAM (poly(N-isopropylacrylamide)), both deposited on 1 cm x 1 cm Si substrates, were used in this study. Prior to polymer deposition, silicon wafers were cleaned by soaking in DI water overnight, which was followed by sonication twice in acetone, methanol, and dichloromethane. Deposition was carried out in a custom built RF plasma system. Deposition of C\(_3\)F\(_6\) was at 150 mTorr at 5 W for 20 min, and deposition of PNIPAM was at 140 mTorr at 5 W for 15 min. We do not have thickness measurements for these samples. The focus of this paper was not the calibration of the depth profile or the sputter rate of the materials, but the application of various analysis methods to the data. However, the settings used in preparing the samples were shown by AFM to produce layers that were ca. 100 nm thick. ToF-SIMS depth profiling was carried out in non-interlaced mode using a ToF-SIMS V instrument (ION-TOF GmbH, Münster, Germany) with 25 kV Bi\(^{3+}\) as the analytical beam and 20 keV C\(_{60}\)^{++} as the sputter beam. Spectra were acquired from a 100 μm x 100 μm area in the center of the sputter crater using a Bi\(^{3+}\) current of 0.09 pA. The total dose per analysis cycle was 1 x 10\(^{11}\) ions/cm\(^2\). Sputtering was done over a 500 μm x 500 μm area with a current of 0.62 nA and a dose of 7.7 x 10\(^{12}\) ions/cm\(^2\) per sputter.
cycle. A total of 24 spectra were collected during each depth profile. An electron flood gun was utilized for charge neutralization. It was not used continuously but was pulsed after each analysis/sputter cycle.

### 3.2.3 Statistical Analyses

PCA and MCR were performed using the PLS Toolbox, version 7.9.3 (Eigenvector Research, Inc., Wenatchee, WA, USA) in MATLAB, version R2015b (Natick, MA). The raw peak areas from the spectra were organized row-wise in data matrices. Although the data were collected using constant instrument settings during the depth profiles, they were normalized using the 1-norm, i.e., each value in each row was scaled such that the sum of the values in each row was unity. This focused the analysis on discriminating between unique spectra. Normalization is a common and important method for preprocessing data sets in chemometrics. It does, however, remove information about the magnitudes of data points from an analysis. Autoscaling, a column operation that scales each channel by its standard deviation and then mean centers it (subtracts the mean of the values in the column from each individual value), was not used. This avoided giving equal weights to the noisy regions of the spectra and the peaks of interest. The data were also not mean centered, which was done to ease comparison of results between PCA and MCR. Not treating the data in this manner is useful for the MCR decomposition. Thus, the preprocessing selected for this analysis was not optimal for PCA, e.g., variables with smaller variances are not as strongly weighted in the analysis as those with larger ones. This approach does, however, allow direct comparison of the PCA and MCR results reported herein, which is the primary reason this was
done. Of course other preprocessing techniques were considered for this analysis, e.g. Poisson scaling for the ToF-SIMS data.\textsuperscript{12-13} But again, to allow a direct comparison between PCA and MCR the data were scaled in the same manner for these methods. PCA and MCR were used together to determine the number of factors to keep in these analyses. This included the use of scree plots in the PCA analysis. See the Appendix 5 for details on the scree plots and also cross validation analyses, which confirmed the statistical validity of our approach. Note that approaches to sample size determination in multivariate analysis have recently been considered by Saccenti and Timmerman.\textsuperscript{14} In addition, we kept in mind the natures of the materials in each depth profile. That is, in the case of MCR, once factors appeared to no longer reflect the expected physics and chemistry of the materials they were assumed to represent interferents, artifacts, etc. and were excluded. Overall, we tried to be conservative with regards to the number of factors kept in each analysis. For the PRE analyses, the data were treated as a probability distribution and arranged column-wise in a matrix. That is, each signal (number of counts) at each m/z (for ToF-SIMS) or binding energy (for XPS) was calculated as the respective signal divided by the sum of all the signals in the respective column (spectrum). The PRE value of each spectrum was then calculated based on these normalized values (vide infra).

3.3 Results and Discussion

3.3.1 The Raw Spectra
Figure 3.1. Three-dimensional graph of 24 positive ion ToF-SIMS spectra from a depth profile through a film of plasma polymerized C3F6 on Si.

It is important not to stray too far from the original data in a chemometrics analysis. That is, where possible it is advisable to examine the original data before any work up, and then to confirm one’s results by referring back to the data again. Accordingly, we show and discuss briefly here the original spectra that are the subject of the analyses described in this paper. Figure 3.1 shows the 24 ToF-SIMS spectra from a depth profile through a film of plasma-polymerized C3F6 on silicon. The monomer for this polymer, C3F6, is obviously unsaturated and fluorinated. At first glance, Figure 3.1 suggests that there is a set of similar spectra from the beginning of the depth profile down to about Spectrum Number 20, and that there is also another group of similar spectra from that point to the end of the depth profile. This latter set of spectra appears to be much simpler than the first set, which is consistent with the first group of spectra corresponding to an organic/polymeric material, where these types of materials often show large numbers of fragment ions by positive ion ToF-SIMS, and the second set corresponding to an inorganic material, which
often yield simpler spectra. Ultimately, PCA, MCR, and PRE were performed on data obtained from the areas of 19 selected peaks from these spectra, i.e., the data matrix used for these analyses had dimensions of 24 x 19. The peaks were selected based on their sizes and chemical relevance. The number of spectra (data points) and integrated areas in this and the other analyses shown in this work are of the same order as in similar chemometrics analyses that have previously been reported by competent chemometrics practitioners.15-17

Figure 3.2. Three-dimensional graph of 24 positive ion ToF-SIMS spectra from a depth profile through a film of plasma polymerized PNIPAM on Si.

Figure 3.2 shows a series of positive ion ToF-SIMS spectra from a film of plasma polymerized PNIPAM, poly(N-isopropylacrylamide), on Si. The structure of the NIPAM monomer is shown in Figure 3.3. Clearly NIPAM belongs to the family of acrylamide monomers.
It contains unsaturation, an amide group (a carbonyl group bonded to a nitrogen atom), and an isopropyl group. As was the case in Figure 3.1, Figure 3.2 appears to initially show a series of similar spectra that are moderately complex, which then transition into a set of simpler similar spectra, i.e., there is evidence for a transition region between these two sets of spectra. It also seems like the first spectrum is different from those that follow it. Ultimately, PCA, MCR, and PRE were performed on data obtained from the areas of 19 selected peaks from these spectra, i.e., the data matrix used for these analyses had dimensions of 24 x 19. The peaks were selected based on their sizes and chemical relevance.

![Structure of N-isopropylacrylamide.](image)

**Figure 3.3.** Structure of \( N \)-isopropylacrylamide.

Figures 3.4 and 3.5 show the Si 2p and O 1s narrow scans, respectively, from an XPS depth profile of a film of SiO\(_2\) on Si. The first sets of scans in Figures 3.4 and 3.5 obviously correspond to the oxide film (the Si 2p signal is shifted to higher binding energy in Figure 3.4 and an O 1s signal is present in Figure 3.5) and the latter sets to the reduced silicon substrate. There appears to
be an interfacial region between these sets of scans. The two narrow scans at the beginning of the depth profile in Figure 3.5 appear to be somewhat different from those that immediately follow them. There appears to be little if any oxygen beyond the interface in the bulk substrate. Ultimately, PCA, MCR, and PRE were performed on data obtained from the areas of 129 energy channels from the O 1s narrow scan and 129 energy channels from the Si 2p narrow scan, which were concatenated (joined together) to form the ‘spectra’ that were used in the PCA and MCR analyses, i.e., a 42 x 258 data matrix was used for these analyses.

Figure 3.4. Three-dimensional graph of the 54 Si 2p narrow scans used in this study from an XPS depth profile of SiO2 on Si.
**Figure 3.5.** Three-dimensional graph of the 54 O 1s narrow scans used in this study from an XPS depth profile of SiO$_2$ on Si.

**Figure 3.6** Three-dimensional graph of the 37 Ta 4f narrow scans used in this study from an XPS depth profile of Ta$_2$O$_5$ on Ta.

Figures 3.6 and 3.7 show the Ta 4f and O 1s narrow scans, respectively, from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta. This set of spectra is moderately complex. The first two
scans in Figure 3.6 contain a doublet at higher binding energy, which suggests they correspond to Ta$_2$O$_5$. They are then followed to about Spectrum Number 24 by a series of broader spectra, the structure of which suggests both an oxidized and a reduced portion to the film. Sputter-induced reduction of the film seems like a logical explanation here. The final spectra consist primarily of simple doublets at lower binding energy, which suggests they correspond to the reduced metal substrate. Figure 3.7 contains the O 1s spectra obtained during this depth profile. There is, again, some indication that the first two spectra are different from those that follow them, and there is a reduction in the intensities of the peaks around the film-substrate interface, after which there appears to be no more oxygen signal. Ultimately, PCA, MCR, and PRE were performed on these O 1s and Ta 4f narrow scans, which were concatenated to form the ‘spectra’ that were used in the PCA and MCR analyses, i.e., the data matrix used for these analyses had dimensions of 37 x 418.

**Figure 3.7.** Three-dimensional graph of the 37 O 1s narrow scans used in this study from an XPS depth profile of Ta$_2$O$_5$ on Ta.
The four sets of spectra in these depth profiles are representative of the types of organic and inorganic materials commonly encountered in XPS and ToF-SIMS analysis, and in materials science in general. Silicon dioxide on silicon is of great technological significance. Tantalum oxide on tantalum is similarly representative of the many metal/oxide combinations found in materials science. The two polymers analyzed by SIMS herein are also important. Fluorinated materials are useful because of their low surface energies (plasma polymerized C$_3$F$_6$ will clearly produce such a material). The acrylamides are a significant class of monomers (the parent compound of PNIPAM is acrylamide).

3.3.2 Plasma Polymerized C$_3$F$_6$ on Si. PCA and MCR of Its ToF-SIMS Depth Profile

Figure 3.8 shows results from a five PC PCA analysis of ToF-SIMS spectra collected during a depth profile through the film of plasma polymerized C$_3$F$_6$ on silicon. Here, PC1 and PC2 capture 69.5 and 29.5% of the variance in the data, respectively, which should be interpreted as variance about zero for 1-normed spectra. PC 1 is attributed to the silicon substrate. This assignment is based on the fact that the scores on PC1 are close to zero and essentially constant until relatively deep into the profile where the substrate should appear (see Figure 3.8a). In addition, the loadings on PC1 (Figure 3.8b) is dominated by a peak at 27.98 amu, which corresponds to $^{28}\text{Si}^+$ (notice that this signal has a mass deficit, which is consistent with its assignment as an inorganic species). This $^{28}\text{Si}^+$ signal is accompanied by corresponding peaks at
$^{29}$Si$^+$ and $^{30}$Si$^+$ in their correct isotopic abundances and masses. The second largest peak in the PC1 loadings plot is at 46.97 amu. Based on its mass and the chemistry of the system it is assigned to $^{28}$SiF$^+$. Perhaps the presence of a strong SiF$^+$ signal in the loadings on PC1 should not come as a surprise because in large measure PC1 appears to account for the transition region between the polymer film and substrate where both Si and F should be present, and Si and F form a strong covalent bond. We assign PC 2 to the C$_3$F$_6$ surface layer. Here, the scores on PC 2 are high for depth indices 1 to 15 after which they decline. Consistent with this designation are the signals at 31.00 (CF$^+$), 69.01 (CF$_3$+), 93.01 (C$_3$F$_3$+), 117.02 (C$_5$F$_3$+), 131.02 (C$_3$F$_5$+) and 141.01(C$_7$F$_3$+) in the loadings on PC 2 (Figure 3.8c). Note also in Figure 3.8 how much more complex the loadings on PC 2 are compared to those in PC 1, which is consistent with them primarily representing an organic and an inorganic material, respectively. The scores on these PCs indicate that Depth Indices (Spectrum Numbers) 16 to approximately 20 correspond to the Si/C$_3$F$_6$ interface region. Here, in Figure 3.8a, we observe a crossing and then inversion of the scores on PC 1 and PC 2. At depth indices beyond 20, the scores on both PCs begin to level out, which is consistent with entry into a new film/layer. Note that some of the scores and loading for PC 2 have negative values. This is a result of the forced orthogonality of the PCs in PCA.
Figure 3.8. PCs 1 and 2 of a five PC PCA model of ToF-SIMS spectra from a depth profile through a film of plasma polymerized C₃F₆ on a silicon substrate. (a) The scores plot as a function of depth index showing the scores on PC 1 and PC 2, and the loadings on (b) PC 1, and (c) PC 2 as a function of the mass (u).

PCs 3 – 5 for this system are shown in Figure 3.9. They clearly represent relatively minor fractions of the total variance that is not captured by PCs 1 and 2. These PCs, and PC 3 and 4 in particular, appear to account for variance at the Si/C₃F₆ interface. In addition, each appears to be doing ‘double duty’, accounting for what appear to be small differences between the outer and inner halves of the films. The suggestion here is that polymer growth is somewhat different when it takes place close to the substrate (perhaps during initiation and the early stages of film growth) compared to when it has been occurring for a while. Indeed, the small step in PC 2 at depth index 8 in Figure 3.8a is also suggestive of a small difference between the upper and lower parts of the film. The scores on PC 3 (Figure 3.9a) are high at the interface and lower going into the Si substrate. The corresponding loadings on this PC (Figure 3.9b) show positive peaks at 31.00 (CF⁺), 46.978 (²⁸SiF⁺), 69.00 (CF₃⁺), and a negative peak from the substrate at 27.98 (²⁸Si⁺). The signs and identities of these signals are consistent with fluorocarbon and mixed fluorine-silicon peaks coming from the top of the interface with increasingly more silicon being present as one passes to the bottom of the interface and into the substrate. The scores and loadings on PC 4 (Figures 3.9a and 3.9c) seem to account for differences between the upper and lower parts of the interface. This loading is dominated by five peaks, one of which is positive: 69.00 (CF₃⁺) and four of which are negative: 46.99 (²⁸SiF⁺), 93.01 (C₃F₅⁺), 117.02 (C₅F₃⁺) and 131.02 (C₃F₅⁺).
Figure 3.9. PCs 3 – 5 of a five PC PCA model of the spectra from a ToF-SIMS depth profile of plasma polymerized C$_3$F$_6$ on a silicon substrate. (a) The scores plots for PCs 3 – 5 as a function of depth index, and the loadings on (b) PC 3, (c) PC 4, and (d) PC 5 as a function of the mass (u).

The MCR analysis of the ToF-SIMS depth profile of the plasma polymerized C$_3$F$_6$ layer on Si is shown in Figures 3.10 and 3.11. As is typically the case in MCR, the scores and loadings here were constrained to be non-negative and so, at least a priori, this analysis appears to be more
interpretable than the corresponding PCA results presented above. In this model (see Figure 3.10a), Component 1 corresponds to the C$_3$F$_6$ layer. It shows significant peaks attributable to the polymer at 31.01(CF$^+$), 69.01 (CF$_3^+$), 93.01 (C$_3$F$_3^+$), 117.02 (C$_5$F$_3^+$), 131.02 (C$_3$F$_5^+$) and 141.01 (C$_7$F$_3^+$). Interestingly, the scores on Component 1 seem to indicate that the composition of the film is changing as it thickens, as was also suggested in the PCA analysis above. Components 2 and 3 appear to correspond to the Si-polymer interface region and Si substrate, respectively. In particular, Factor 2 shows two main signals at 27.98 amu ($^{28}$Si$^+$) and 46.97 amu ($^{28}$SiF$^+$), which, as noted above, is chemically consistent with the expected interface containing both materials. Component 3, which is dominated by the $^{28}$Si$^+$ peak, then appears to account for the bulk substrate. The sputtering of the C$_3$F$_6$ layer is complete by depth index 20, which is in agreement with the PCA analysis.
Figure 3.10. MCR components 1 – 3 from a five factor MCR model of a ToF-SIMS depth profile through a film of plasma polymerized C₃F₆ on Si. (a) Scores on MCR components 1 – 3 as a function of depth index, and the loadings on (b) Component 1, (c) Component 2, and (d) Component 3 as a function of the mass (u).

MCR Components 4 and 5 are shown in Figure 3.11. Factor 4 appears to be associated with the bottom half of the C₃F₆ layer with major peaks at 31.01 (CF⁺), 69.01 (CF₃⁺), 93.01 (C₃F₃⁺),
117.02 (C$_3$F$_3^+$) and 131.02 (C$_3$F$_5^+\text{+}$). The scores on this factor (Figure 3.11a) are further evidence for chemical differences between the upper and lower parts of the C$_3$F$_6$ film. Factor 5 appears to be primarily associated with the base of the C$_3$F$_6$ layer/interface layer with major peaks at 31.00 (CF$^+$) and 69.00 (CF$_3^+$). Nevertheless, all the peaks from the polymer in Component 4 also appear to be present in this factor, but with different ratios that probably reflect chemical differences between the polymer at the interface and polymer in the first half of the film.

**Figure 3.11.** Components 4 and 5 of a five factor MCR model of ToF-SIMS data from a depth profile through a film of plasma polymerized C$_3$F$_6$ on Si. (a) Scores on MCR components 4 and 5 as a function of depth index, and the loadings on (b) Component 4, and (c) Component 5 as a function of the mass (u).

### 3.3.3 Plasma Polymerized PNIPAM on Si. PCA and MCR of Its ToF-SIMS Depth Profile
Figure 3.12 shows the first four PCs from a PCA analysis of the spectra from a ToF-SIMS depth profile through a film of PNIPAM on Si. PC 1 corresponds to the Si substrate with major peaks in its loading (Figure 3.12b) at 27.98 (Si⁺) and 44.98 (SiOH⁺). PC 2 is attributable to the PNIPAM layer with major peaks in its loading (Figure 3.12c) at 27.03 (C₂H₃⁺), 29.05 (C₂H₅⁺), 41.04 (C₃H₅⁺), 43.05 (C₃H₇⁺) and 58.07 (C₃H₈N⁺). All of these peaks from the loadings on PC 2 have mass excesses, which suggests that they are organic in nature and is consistent with the expected organic nature of the film. We also see in this spectrum a series of clusters of peaks separated by ca. 14 amu, i.e., the mass of a methylene unit, which is again consistent with an organic material. The scores on PC1 and PC2 in Figure 3.12a indicate that the PNIPAM layer is present from depth indices 1 to 14, the polymer-substrate interface occurs between depth indices 15 and 20, and the Si substrate is primarily at depth index 21 and beyond. As observed in the previous analysis, the scores and loadings for PC 2 have some negative values, which is again a result of the orthogonality constraints of PCA. The scores on PCs 3 and 4 in Figure 3.12d are complex. While they are certainly associated with the polymer-substrate interface of the material, and they account for this variance in different ways, they also appear to help explain the variance over much of the film and substrate. As was the case for PC 2 in this analysis (Figure 3.12c), the clusters of peaks separated by ca. 14 amu in the loadings on PCs 3 and 4 (Figures 3.12e – f) indicates that they primarily represent an organic material.
Figure 3.12. Results from a PCA analysis of the ToF-SIMS spectra from a depth profile through a film of plasma polymerized PNIPAM on a Si substrate. (a) and (d) Scores plots as a function of depth index, and the loadings of (b) PC 1, (c) PC 2, (e) PC 3 and (f) PC 4 as a function of the mass (u).

An MCR analysis of the data from the PNIPAM layer was also performed, and the scores on the first four factors are shown in Figure 3.13a. The first two of these components are related to the Si substrate. In particular, from depth indices 16 to 25 Factor 1 primarily accounts for the polymer-substrate interface and from depth indices 19 – 25 Factor 2 is more closely associated with the Si substrate. As expected, both Factor 1 and 2 have major peaks at 27.98 (Si+) and 44.98 (SiOH+), while Factor 2 has an additional minor peak at 55.95 (Si2+). Factors 3 and 4 correspond to the PNIPAM layer. This is suggested not only by the positions of the scores on these components (Figure 3.13a), but also by the complexity of the loadings (Figure 3.13d – e). Factor 3 primarily corresponds to the top of the layer (depth indices 1 to 16), and Factor 4 corresponds to the bottom of the PNIPAM layer (depth indices 3 to 18). Thus, this analysis again points to a difference between the upper and lower parts of this film. Factor 4 also contains a greater number of higher mass peaks than Factor 3.

Figure 3.14 shows Factors 5 and 6 from this MCR analysis. They account for quite small amounts of the variance in the data. Factor 5 contributes mostly to the interface. It has strong signals at 27.98 (Si+) and 44.98 (SiOH+), and a series of smaller signals reminiscent of an organic/polymeric material. Factor 6 is strongest at the top of the film. It also appears to account for some of the structure in the films at depth indices 7 to 15, and then to a smaller degree
within the interface. It has major peaks at 43.05 (C$_3$H$_7^+$) and 58.07 (C$_3$H$_8$N$^+$) and a series of signals that suggest that it has an organic nature. Overall, most of the factors in this MCR analysis (Figures 3.13 and 3.14) were readily interpretable, while only the first two PCs in the previous PCA analysis were.
Figure 3.13. Factors 1 – 4 from a six factor MCR analysis of the spectra from a ToF-SIMS depth profile of plasma polymerized PNIPAM on Si. (a) Scores on MCR components (1 – 4) as a function of depth index, and the contributions (loadings) of (b) Component 1, (c) Component 2, (d) Component 3, and (e) Component 4 as a function of mass (u).

Figure 3.14. Factors 5 – 6 from a six factor MCR analysis of the spectra from a ToF-SIMS depth profile of plasma polymerized PNIPAM on Si. (a) Scores on MCR components 5 and 6 as a function of depth index, and the loadings on (b) Component 5 and (c) Component 6 as a function of mass (u).

3.3.4 SiO$_2$ on Si. PCA and MCR of Its XPS Depth Profile

Silicon and silicon dioxide are two of the most important materials in modern technology. They have been extensively characterized by many techniques, including XPS and ToF-SIMS. $^{8, 159}$ Figure 3.15 shows results from a PCA analysis of an XPS depth profile through a film of SiO$_2$
on Si in which Si 2p and O 1s narrow scans were collected at each depth. In this PCA analysis, the Si 2p and O 1s narrow scans were combined (concatenated) into a single spectrum without any weighting of either spectrum. The same approach was taken in the analysis of the Ta 4f and O 1s narrow scans below. The scores on PCs 1 and 2 in Figure 3.15a show clear trends that are consistent with a two-layer structure for the material. Thus, it can be surmised that the SiO$_2$ surface layer extends over depth indices ca. 1 – 15, that there is an interface region from depth indices ca. 16 – 20, and that the Si substrate is present at depth indices ca. 20 – 42. However, there is an obvious abstraction in the scores on these PCs. That is, although there are significant flat regions in these scores plots, neither of them is close to zero where the SiO$_2$ film and/or Si substrate appear to be present, so both PCs appear to be simultaneously describing both key parts of the material – we do not have a simple and desirable one-to-one correspondence between factors and films/materials. Accordingly, it comes then as little surprise that the loadings on these factors are also not as interpretable as one might like. Indeed, both show signals from both key parts of the material, i.e., both factors in Figure 3.15b show peaks due to both oxidized and reduced silicon, and both components in Figure 3.15c show an oxygen peak (one of them is inverted). Adding to this complexity, both factors have a negative component in their loadings (Figures 2.15b – c) – they are unrealistic as spectra. The scores on PCs 3 and 4 in Figure 3.15d indicate that there is something different about the measurements at depth indices 1 and 2, and also at depth indices 17 to 20 compared to the surrounding material. The scores on PC 1 at depth indices 1 and 2 in Figure 3.15a also suggest at least a slight difference between the first two scans and those that follow them. These results are consistent with the fact that (i) there is often a contamination layer, i.e., adventitious carbon, on materials, and (ii) there can sometimes be noticeable chemical differences.
between the surfaces of materials and their corresponding bulks. With regards to this latter point, this silica surface probably has more adsorbed water and a larger number of silanol groups than the bulk SiO$_2$, which will probably be better described as a material connected through siloxane (Si-O-Si) linkages. The complexity of the loadings on Components 3 and 4 between 99 and 103 eV (Figures 3.15e – f) is perhaps suggestive of the presence of suboxides at the Si-SiO$_2$ interface. Overall, the loadings on Components 3 and 4 are difficult to interpret.
**Figure 3.15.** PCA analysis of the spectra from an XPS depth profile through a film of SiO$_2$ on Si. (a) The scores on PC 1 and PC 2 as a function of depth index, and the contributions/loadings of PC 1 and PC 2 as a function of binding energy for (b) the silicon 2p narrow scan and (c) the O 1s narrow scan. (d) The scores on PC 3 and PC 4 as a function of depth index, and the contributions/loadings of PC 3 and PC 4 as a function of binding energy for (e) the silicon 2p narrow scan and (f) the O 1s narrow scan.
Figure 3.16. The first four components from a seven-component MCR analysis of XPS data from a depth profile of SiO$_2$ on a Si substrate. (a) Scores on MCR components 1 and 2 as a function of depth index, and the contributions (loadings) of Components 1 and 2 as a function of binding energy corresponding to (b) silicon (Si 2p) and (c) oxygen (O 1s). (d) Scores on MCR components 3 and 4 as a function of depth index, and the contributions (loadings) of
Components 3 and 4 as a function of binding energy corresponding to (e) silicon (Si 2p) and (f) oxygen (O 1s). MCR components 5 – 7 are shown in the Appendix 5.

A seven-factor MCR analysis of the XPS Si/SiO₂ depth profile was also performed. Factors 1 and 2 in this analysis (Figure 3.16a) clearly correspond to the Si substrate and its SiO₂ over layer, respectively, where Factor 1 extends from depth indices 18 to 42, and Factor 2 extends from depth indices 3 to 17. Unlike the PCA analysis of this data set, the scores on the main factors here show significant regions where they are approximately zero, which makes them considerably more interpretable. This increased interpretability extends to the loadings on these components (Figures 3.16b-c), where Component 1 is primarily described by a peak at ca. 99 eV, which corresponds to bulk (reduced) silicon, and Component 2 is primarily described by a peak at ca. 103 eV, which corresponds to oxidized silicon. Similarly, in the O 1s region of the loadings, Component 1 consists mostly of a horizontal line with a small depression at approximately the O 1s peak energy, while Component 2 corresponds to the O 1s signal. However, while Component 1 appears to describe the silicon substrate in a reasonable fashion, i.e., the scores on Component 1 in Figure 3.16a approximate a step function, the scores on Component 2 in Figure 3.16a from depth indices 3 to 17 suggest some complexity in the film, i.e., they are lower at the top of the film and higher closer to the Si-SiO₂ interface. Complementing Component 2, the scores on Component 3 (Figure 3.16d), which accounts for a smaller fraction of the variation in the data (7.73%) compared to Component 2 (25.01%), indicate that it primarily explains the upper part of the oxide layer. Thus, the scores on Components 2 and 3 again suggest, at least to some degree, a lack of homogeneity in the film, i.e., they indicate there may be a moderate gradient in it. Beyond depth index 20 the scores on
these components are essentially zero. Component 4 largely accounts for the first two data points in the analysis. Both parts of the loadings on Components 3 and 4 (Figure 3.16 e – f) contain silicon and oxygen signals. Finally, we note again the greatly improved intuitive understanding associated with the scores and loadings on Components 1 – 4 in the MCR model (Figure 3.16) compared to the scores and loadings on PCs 1 – 4 in the corresponding PCA model (Figure 3.15).

3.3.5 Ta₂O₅ on Ta. PCA and MCR of Its XPS Depth Profile

Figure 3.17 shows results from a PCA analysis of the spectra from an XPS depth profile of a layer of Ta₂O₅ on Ta. The scores on PC 1 and PC 2 in Figure 3.17a show subtle and pronounced trends, respectively, that are consistent with the expected layered structure of this material. However, as was the case for the PCA of the SiO₂ on Si depth profile that was just discussed, interpretation of these results is challenging because of the presence of negative scores and loadings and the fact that both components appear to be describing both materials. Nevertheless, it can be surmised from Figure 3.17a that the Ta₂O₅ surface layer extends from depth indices 2 or 3 to 17, and that either a different material is present at the very top of the film at depth indices 1 to 2 or 3 or a different form of Ta₂O₅ is there. As noted, the original spectra indicate that sputter-induced reduction of the film is taking place (see Figure 3.6). An interface then appears to be present from depth indices ca. 17 – 20, and the Ta substrate is finally present at depth indices 20 – 27. The loadings on PC 1 and 2 in the Ta 4f (Figure 3.17b) and O 1s (Figure 3.17c) narrow scan regions are at least moderately challenging to interpret. In particular, the loadings on PC 2 in Figure 3.17b show both strong positive and negative peaks, and the loadings on PC1 in Figure 3.17c are
negative. The entire matter is even more complicated with Components 3 – 5. That is, while the scores on these components in Figure 3.17d appear to be fairly consistent with an outermost region of the film that is chemically different from what follows, a film (of Ta₂O₅), a film-substrate interface region, and a substrate layer (the Ta), the loadings on these PCs in Figures 3.17e – f are essentially impossible to interpret – one is denied any intuitive feel for how they are explaining the variance in the data set. Thus, we make no attempt to provide any additional insight into this analysis. We will see below that MCR handles this data set much more effectively.
Figure 3.17. PCA analysis of the XPS spectra from a depth profile through a film of Ta$_2$O$_5$ on a Ta substrate. (a) Scores plot for PC 1 and PC 2 as a function of depth index, and loadings on PC 1 and PC 2 as a function of binding energy for (b) tantalum and (c) oxygen. (d) Combined scores plot for PC 3, PC 4 and PC 5 as a function of depth index, and the loadings on PC 3, PC 4, and PC 5 as a function of binding energy for (e) tantalum and (f) oxygen.

Figure 3.18 shows the scores and loadings from a five-component MCR model of the Ta$_2$O$_5$ on Ta depth profile under consideration in this section. In contrast to the PCA results in Figure 3.17 that were challenging to understand, it is immediately obvious in Figure 3.18 that MCR Component 1 corresponds to the Ta$_2$O$_5$ film and Component 2 corresponds to the Ta substrate. Indeed, the scores on Component 2 resemble a step function, i.e., they consist primarily of two rather flat regions joined by an interface region. These results also suggest that, in contrast to the SiO$_2$ film on Si discussed above (see Figure 3.16), the Ta$_2$O$_5$ layer is fairly homogeneous across its depth. This conclusion is consistent with the scores on PCs 1 and 2 from the PCA analysis in Figure 3.17. As in the PCA analysis, the first part of the depth profile here (depth indices 1 – 2) is not accounted for by Component 1 (the Ta$_2$O$_5$ film), suggesting again that the top of the layer is in some way different from what is below it. Again, we cited sputter-induced reduction of the film as the primary cause for this difference. The loadings on Components 1 and 2 are relatively straightforward to understand and further confirm these general assignments. In particular, Figure 3.18b shows that Component 2 is primarily associated with reduced (metallic) Ta, i.e., as evidenced by the pair of peaks at 21.74 (4f$_{7/2}$) and 23.5 (4f$_{5/2}$) eV, while Component 1 is associated with more oxidized Ta (the signal is chemically shifted to higher binding energy). The loadings on
Components 1 and 2 in the O 1s narrow scan region (Figure 3.18c) are even easier to interpret. Component 1 (the film) has the appearance of an O 1s signal and Component 2 (the substrate) shows no peak – there is a very clean division of the signal here.

The scores and loadings of Components 3 – 5 in this analysis are given in Figure 3.18d – f. Here, Component 3 appears to account for variation in the material over depth indices 3 – 18 (the film), Component 4 accounts for variation in the material over depth indices 4 – 20 (the film and interface), and Component 5 accounts primarily for the first two measurements in the film and part of the interface region. All three of these components show well-defined doublets in Figure 3.18e, where the doublet in Component 4, like the one in Component 2, corresponds to metallic Ta (peaks at 24.38 and 22.62 eV) and those in Components 3 and 5 (peaks at 29.00 and 27.38 eV, and 29.12 and 27.12 eV, respectively) correspond to Ta₂O₅. The scores on these components suggest that there are different forms of Ta₂O₅ in the Ta layer, which is again consistent with the original spectra (Figure 3.6). Once again we see that the factor that appears to represent the metallic part of the material (Component 4) has essentially no oxygen peak (Figure 3.18f), while those attributed to the oxidized material (Components 3 and 5 in Figure 3.18e) have prominent oxygen components (Figure 3.18f).
Figure 3.18. Results from a five-factor MCR analysis of an XPS depth profile through Ta$_2$O$_5$ on Ta. The (a) scores on MCR components 1 and 2 as a function of depth index, and the loadings on Components 1 and 2 as a function of binding energy for the (b) Ta 4f and (c) O 1s regions. The (d) scores on MCR components 3, 4, and 5 as a function of depth index, and the loadings on Components 3, 4, and 5 as a function of binding energy for the (e) silicon 2p and (f) oxygen 1s regions.
3.3.6 Summary of PCA and MCR Analyses of ToF-SIMS and XPS Depth Profiles

PCA and MCR analyses were conducted of ToF-SIMS depth profiles of plasma polymerized C$_3$F$_6$ and PNIPAM on Si, and also of XPS depth profiles of SiO$_2$ on Si and Ta$_2$O$_5$ on Ta. In all cases, PCA identified the general layered structures of the films on the substrates, interface regions between the films and the substrates, and that the first few measurements in the profiles often differed from the subsequent ones. However, the scores on the main PCs in these analyses were sometimes negative, and the corresponding loadings plots also regularly showed negative peaks. The scores and loadings plots corresponding to the higher PCs were sometimes very difficult to interpret. The abstraction, and therefore challenge with interpretability, associated with these analyses is inherent to PCA because of the forced orthogonality of its components. In contrast, the MCR analyses corresponding to the PCA analyses were easier to interpret. Of course this was due in large measure to the fact that its scores and loadings are constrained to be positive, i.e., to have the appearance of real spectra. Ultimately, however, the ‘best’ chemometrics tool to use on a data set may be a combination of them. That is, different data sets often respond somewhat differently to different chemometrics tools – one data set may be better interpreted by PCA while another by MCR, and it is not generally possible to fully understand, a priori, which data set will be better analyzed by one tool or another. In addition, these tools may extract different bits of information from a data set such that the more complete possible analysis will involve the use of multiple chemometrics tools, including PRE (vide infra). Finally, performing more than one analysis provides a check on one’s work – in general one does not expect radically different results from different chemometrics tools.
3.3.7 PRE Analysis of Depth Profiles

As noted above, the PRE value represents a substantial reduction/compression of the information in a spectrum; the PRE value is a summary statistic. In this regard it is similar to our perception of the color of a material, which is, in general, a greatly simplified, but still very useful, representation of the optical properties of an object. Accordingly, it seems unlikely that the PRE approach can compete on a comprehensive chemical basis against techniques like PCA and MCR. Nevertheless, it should be useful in data analysis in the following ways. First, PCA and MCR generally require that all of the spectra be collected and preprocessed prior to any analysis. In many cases, however, it would be advantageous to do at least some mathematical processing of the information during data acquisition. For example, in the case of a long depth profile, it would be advantageous to confirm early in the run that it is progressing reasonably. PRE can do this because the PRE value of a single spectrum can be calculated independently of the other spectra in a depth profile/data set. Second, both PCA and MCR require a fair amount of training to understand and perform, i.e., their theories are based on at least moderately complex linear algebra. In addition, one must learn how to determine the number of factors to keep, and the meaning of such things as ‘scores’, ‘loadings’, ‘Q residuals’, the ‘Hotelling \( T^2 \)’, etc. Of course the underpinnings of Shannon’s formula for entropy are nonobvious. Nevertheless, the general ideas of (i) considering a spectrum to be a simple probability distribution (it isn’t, of course) and (ii) reducing that distribution to a single number, which is a reflection of its shape, are reasonably straightforward concepts. And, as will be shown, the output of both PRE analysis and the corresponding finite difference PRE calculations yield straightforward plots that are much easier
to interpret than many of the PCA scores plots we have encountered. Indeed, a quick, initial PRE analyses can be a useful starting point for subsequent PCA and MCR analyses. Third, PRE calculations are fast. They can be many times faster than MCR analyses involving multiple components, i.e., 5 - 10, which can sometimes take tens of minutes or longer on a personal computer.

In this section we show how PRE values identify interfaces in materials in depth profiles. This approach appears to work because PRE values from different spectra are generally different – PRE recognizes the different patterns represented in different spectra. For example, as one depth profiles through an oxide layer and into its reduced substrate by XPS, one will generally find that the metal peaks from the metal oxide are wider than those from the substrate.21 Thus, a plot of PRE value vs. depth index should indicate the position of an oxide-metal interface, as well as other changes that take place in the material as a function of depth. However, the results from PRE analysis may not be simple step functions because interfaces are rarely perfectly smooth, the act of depth profiling (sputtering) can further roughen/broaden and cause mixing at them, and XPS and ToF-SIMS probe a finite depth into materials. Accordingly, an interface region will often show spectral features from both of the materials that surround it. In these cases, one might expect higher PRE values at and around interfaces. We will encounter this and other scenarios in the examples that follow.

3.3.8 Plasma Polymerized C₃F₆ on Si. PRE of Its ToF-SIMS Depth Profile
Figure 3.19a shows the PRE analysis of the ToF-SIMS spectra from the depth profile through the film of plasma polymerized C$_3$F$_6$ on Si. The results seem fairly easy to interpret. The analysis suggests there is a transition at the beginning of the depth profile (depth indices 1 – 2), a region of nearly constant PRE values that corresponds to the polymer (depth indices 3 – 16), a transition region in which the PRE value rises modestly and then drops towards the value of the substrate (depth indices 17 – 22), and the low PRE values associated with the substrate (depth indices 23 – 24). This lower PRE value of the substrate is expected because of the decreased complexity of its spectrum compared to the more complex spectrum of the polymeric overlayer. The increased PRE value at the polymer-substrate interface (depth indices 18 – 19) suggests that peaks from both entities are in the spectra here, i.e., that there is mixing and/or sampling of both materials in the analysis.

To provide greater insight into PRE analyses, we introduce the use of the finite difference on PRE values. The finite difference is closely related to the derivative – it is essentially a numerical derivative that omits division by the distance between the data points.$^{22}$ In particular, we have found that the backward distance can be helpful in revealing trends in some PRE plots. It is defined as:

\[(3.1)\quad \nabla f(x) = f(x) - f(x - h)\]

or adapted to the task at hand:

\[(3.2)\quad \nabla \text{PRE}_n = \text{PRE}_n - \text{PRE}_{n-1}\]
where \( n \) is the depth index of the PRE value of interest. Thus, the backward distance is the value of the current PRE value minus the value of the previous one. The finite difference, which is defined here as:

\[
(3.3) \quad \Delta \text{PRE}_n = \text{PRE}_{n+1} - \text{PRE}_n
\]

is only moderately different from the backward difference in that it is the next PRE value minus the current one. Similar results were obtained with both types of differences. However, the features in the backward difference plots line up a little better with those in the original PRE data so it is preferred here. As is often the case with derivatives of spectra, the backward difference/finite difference accentuates the changes in the original data. For example, in Figure 3.19b both the initial transition region and interface region of Figure 3.19a are more strongly defined.

To test PRE’s ability to deal with a subset of the total information, analogous calculations to those just described on full spectra were performed on sets of 19 selected peaks from each spectrum. Similar results/shapes of the curves were obtained in each case, except for the initial few points, which showed less pronounced changes/differences (see Appendix 5).
Figure 3.19. (a) PRE values and (b) backward difference PRE values of the mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized C₃F₆ on Si as a function of depth index.
**Figure 3.20.** (a) PRE values and (b) backward difference PRE values of 19 selected peaks from the mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized PNIPAM on Si as a function of depth index.

### 3.3.9 Plasma Polymerized PNIPAM on Si. PRE of Its ToF-SIMS Depth Profile

Figure 3.20 shows a PRE analysis of 19 selected peaks from the mass spectra of a ToF-SIMS depth profile through a film of plasma polymerized PNIPAM on silicon. PRE clearly identifies a reasonably flat region from depth indices 1 – 12, which corresponds to the polymer film, an interface region (depth indices 13 – 21), which shows an initial rise in the PRE value, and the substrate (depth indices 22 – 24), which again shows a lower PRE value because of its inorganic nature. Overall these results are quite similar to those seen in Figure 3.19 for the C₃F₆ film, with the exception that the interface region is more pronounced in Figure 3.20 and the first few points are less well defined. These differences are most likely because only a subset of the data was included in this analysis, i.e., as in the previous example this subset of peaks did not seem to capture the information related to the outer surface of the film. The increase in PRE values at the interface again suggests that the spectra here contain contributions from the spectra of the materials on either side of it and are therefore more complex. Changes in peak intensity are consistent with these chemical changes. For example, over the interface region from depth indices 14 – 17, the intensity of the m/z 44.05 peak, which has a mass excess and should therefore be organic, decreases, while the intensity of the m/z 44.98 peak (SiOH⁺), which has a mass deficit and should be inorganic, increases. The transition region in Figure 3.20a is also clearly defined in the
backward difference PRE plot. Finally, this analysis was repeated with the entire spectra from m/z 0 – 150 (see Appendix 5). These results showed a more pronounced initial transition at depth indices 1 – 2 and less of an increase in the PRE value at the transition region. The corresponding backward difference PRE plot primarily showed a single dip at the transition region of the material.

![Graph](image)

**Figure 3.21.** (a) PRE values and (b) backward difference PRE values of the Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si.

### 3.3.10 SiO$_2$ on Si. PRE of Its XPS Depth Profile

Figure 3.21a shows a plot of the PRE values of the Si 2p narrow scans from an XPS depth profile of SiO$_2$ on Si vs. depth index. This plot is dominated by a pronounced peak that clearly identifies the Si/SiO$_2$ interface. This peak shows a reasonably flat region to its left that corresponds to the SiO$_2$ film and another to its right that corresponds to the Si substrate. These flat regions have
slightly different PRE values. As expected, the first two narrow scans have different PRE values than those that follow. The backward difference of the PRE values in Figure 3.21a (Figure 3.22b) accentuates the differences between the scans from the top of the SiO$_2$ layer and those that follow, and again clearly identifies the Si/SiO$_2$ interface.

**Figure 3.22.** (a) PRE values and (b) backward difference PRE values of O 1s narrow scans from an XPS depth profile through a film of SiO$_2$ on Si.
Figure 3.22 shows the PRE analysis of the O 1s narrow scans from the same SiO$_2$ on Si depth profile considered in Figure 3.21. For reference, most of these narrow scans are shown in Figure 3.23. As before, these spectra come sequentially from the outermost surface of the material, the SiO$_2$ film, the interfacial region, and the bulk Si. PRE analysis shows a small difference between the first two data points and those that follow, which is not obvious in the narrow scans in Figure 3.23. It also yields a steady increase in the PRE values of the next set of points up to the
signal from the substrate. This increase suggests that a change is taking place in the oxygen signals over the course of the depth profile through SiO_2. A plot of the ratio of the O 1s to Si 2p signal vs. depth index (see Appendix 5) shows a small but steady rise up to the Si/SiO_2 interface, which appears to be consistent with the results in Figure 3.22. (The MCR analysis in Figure 3.16 also indicates that the SiO_2 film is not entirely homogeneous.) However, the most abrupt jump in Figure 3.22a occurs at the Si/SiO_2 interface between depth indices 22 and 28. Here, the backward difference analysis in Figure 3.22b is helpful in identifying the transition regions in this material, i.e., it suppresses the initial steady growth in the PRE value because of the modest changes occurring in O1s/Si2p ratio and accentuates the air/SiO_2 and Si/SiO_2 interfaces.

After depth profiling through the film and the Si/SiO_2 interface, the O 1s signal decreases to a constant low level (note the low signal-to-noise O 1s spectra at the higher depth indices in Figure 3.23). The high, constant PRE values for the spectra beyond depth index 27 in Figure 3.23 are consistent with these results, i.e, the increasingly noisy spectra in Figure 3.23 are expected to yield higher PRE values. An arbitrary signal with random noise levels at 0 – 1000% of the height of the original signal (Figure 3.24), and corresponding PRE values for these mock spectra (Figure 3.25) are shown. This modeling shows monotonic increases in the PRE values of the signal+noise spectra up to 100% noise, after which no further increase is observed.
**Figure 3.24.** A mock Gaussian peak with random noise added that ranges from 0% - 1000% of the height of the Gaussian signal.

**Figure 3.25.** PRE values of the mock spectra shown in Figure 37 plotted as a function of (a) the amount of random noise (%) and (b) Log (random noise %).
Figure 3.26. Analysis of the Ta 4f narrow scans from an XPS depth profile through a film of Ta₂O₅ on Ta. (a) PRE values and (b) backward difference PRE values as a function of depth index.

Figure 3.26 shows a PRE analysis of the Ta 4f signal from a depth profile through a film of Ta₂O₅ on Ta. The first two PRE values in this plot, and to a small degree the third one, have noticeably lower values than those of the oxide that follow, which show high PRE values because of their breadth. There is then a decrease in the PRE value at the interface region, which is followed by lower PRE values for the bulk Ta. The changes in the PRE values in Figure 3.26a and the peaks in the backward difference PRE plot in Figure 3.26b are consistent with the presence of two interfaces in this depth profile.
Figure 3.27. Analysis of O 1s narrow scans from an XPS depth profile of Ta$_2$O$_5$ on Ta. (a) PRE values and (b) backward difference PRE values as a function of depth index.
**Figure 3.28.** O 1s spectra collected during an XPS depth profile of Ta\(_2\)O\(_5\) on Ta substrate.

Spectra 36 and 37 are very similar to spectra 29-25 and so are not included here.

Figure 3.27a shows the PRE analysis of the O 1s narrow scans obtained during the Ta\(_2\)O\(_5\) on Ta XPS depth profile. For reference, these O 1s signals are provided in Figure 3.28. Figure 3.27 again suggests two main transitions in the material – one at the air/Ta\(_2\)O\(_5\) interface and the other at the Ta\(_2\)O\(_5\)/Ta interface. As expected from the results in the previous example, the O 1s spectra that contain almost nothing but noise in Figure 3.28 show the highest PRE values in Figure 3.27. Also as observed in the previous example, the PRE values of the spectra in the film change somewhat over the course of the depth profile, suggesting that some change in the material is taking place. That the film may not be entirely homogeneous is also suggested in the MCR analysis in Figure 3.28. The backward difference PRE plot in Figure 3.27 very clearly shows the main transitions in the material.

**3.4 Conclusions**

We have presented PCA and MCR analyses of four different depth profiles: two by ToF-SIMS of polymer films on silicon, and two by XPS of inorganic oxide films on their respective reduced substrates. While both sets of analyses could identify key transitions in the materials in question, and even suggested heterogeneity within the films, the MCR results were consistently more intuitive. In a number of the PCA analyses, the scores and loadings plots for the higher PCs
were not at all easy to understand. Nevertheless, the best approach to these types of analyses will often involve both PCA and MCR. That is, PCA is and will remain an extremely important data analysis tool, but the output of a multivariate analysis will often be more readily conveyed through MCR. Pattern recognition entropy (PRE), which has its roots in Shannon’s information theory, was also introduced. Its advantages were emphasized. It nicely identifies transitions in depth profiles. Interpretation of PRE vs. depth index plots can often be enhanced by considering plots of backward difference PRE values. PRE analysis is proposed to complement and perhaps guide traditional PCA and MCR analyses.

3.5 Acknowledgments and Partial Statement of Attribution

We acknowledge the Department of Chemistry and Biochemistry and College of Physical and Mathematical Sciences at Brigham Young University for their support of this work. We also thank Dan Graham from the Department of Bioengineering, University of Washington, Seattle, for providing the ToF-SIMS data. Collection of the ToF-SIMS depth profiles was supported by NESAC/BIO and NIH grant EB-002027. A portion of the research was performed at EMSL, a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research. Finally, Shiladitya Chatterjee and Bhupinder Singh both made very substantial contributions to this work and should be considered as equal first authors on it.
3.6 Supporting Information Available

Supporting Information is available in Appendix 5.

3.7 References


15. Barnes, C. A.; Brison, J.; Robinson, M.; Graham, D. J.; Castner, D. G.; Ratner, B. D., Identifying individual cell types in heterogeneous cultures using secondary ion mass spectrometry imaging with C60 etching and multivariate analysis. *Analytical chemistry* 2012, 84 (2), 893-900.


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


Chapter 4: Reordered (Sorted) Spectra. A Tool for Understanding Pattern Recognition Entropy (PRE) and Spectra in General

4.1 Introduction

A summary statistic is a number that characterizes a data set. Summary statistics are widely used in analytical chemistry, chemometrics, and statistics because the comparison of a few summary statistics, especially if they capture key characteristics of a data set or spectrum, is often much easier than trying to compare all of the values in a spectrum or data set to those in another. Examples of summary statistics include the mean, standard deviation, median, percentiles, and interquartile range.1-2 The use of these and other less common descriptors abounds.3-7 For example, Antweiler and Taylor studied the estimation of summary statistics of censored data,1 and Aeschbacher and coworkers,4 and also Nunes and Balding,8 presented methods for selecting them in approximate Bayesian computation.

Taking our lead from Claude Shannon’s groundbreaking work,9 we recently introduced Pattern Recognition Entropy (PRE) as a summary statistic for data/spectra.10 Briefly, the PRE of a spectrum is defined as:

\[
\text{PRE}(x_i) = - \sum_{i=1}^{n} p(x_i) \log_2 p(x_i),
\]
where each \( p(x_i) \) value is obtained by dividing the corresponding data point, \( x_i \), in a spectrum by the sum of the \( x_i \) values in the spectrum, i.e., we employ the 1-Norm. These are the ‘probabilities’ for a spectrum. Obviously a logarithm with a different base could be used in Equation 1, but for consistency with Shannon’s approach and the signal processing world he worked in, we use the base 2 logarithm. Accordingly, the units of Equation 1 are ‘bits’. Using Equation 4.1,\(^{10}\) we showed in Chapter 3 that PRE could identify transitions in X-ray photoelectron spectroscopy (XPS)\(^{11-12}\) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) depth profiles.\(^{13-15}\) Of course \( \text{PRE}(x_i) \) can be normalized by division by \( \log n \) such that its value ranges from 0 to 1. To the best of our knowledge, the PRE approach we recently reported\(^{10}\) and the concept of reordered spectra introduced herein are new. They are not the same as previous uses of Shannon’s entropy in spectral analysis. For example, Ferenets and coworkers,\(^{16}\) Viertiö-Oja and coworkers,\(^{17}\) and Ellerkmann and coworkers\(^{18}\) applied Shannon’s entropy formula to the energy spectra of their spectra, where an energy spectrum is obtained by squaring the Fourier transform of the original spectrum (multiplying it by its complex conjugate). The subsequent entropy calculations were then performed over frequency ranges of interest. Yahiatene and coworkers\(^{19}\) calculated local entropy and cross-entropy values pixel-by-pixel that was weighted with higher order statistics to reconstruct super-resolved images from a times series of images with random signal fluctuations. Baldi and coworkers\(^{20}\) on the other hand applied integer entropy codes based on Shannon’s information theory to encode run lengths of large fingerprint vector representations of small molecules. The feature probabilities were reordered and obeyed a power-law distribution. However, the reordering approach here was adopted as a preprocessing method for Elias Gamma codes and is not the same as our approach in which reordering is applied as a visual aid to
understand independent PRE values. Pourhaghighi and coworkers\textsuperscript{21} used information entropy and conditional entropy to calculate orthogonality on multidimensional (2-D) separation systems. A two-dimensional data matrix of peaks from both the dimensions was obtained followed by entropy calculations and finally a figure of merit was designed based on joint and individual entropies of the dimensions. Widjaja and coworkers\textsuperscript{22} applied information theory to convert loading vectors obtained from a singular value decomposition (SVD) of a spectroscopic mixture into pure component spectra. Gonzalez and coworkers\textsuperscript{23} used Shannon’s information entropy to explain the origin of stationary points of the electronic density in a reacting system traced along the reaction path. Our approach is also not the same as the mutual information method that has been applied to spectra.\textsuperscript{24-25} PRE is a whole spectrum analysis tool. In it, an entire spectrum is treated as a probability distribution, where it is normalized and entered into Equation 1 to obtain a summary statistic that characterizes it. However, we emphasize that spectra collected in typical experiments in Analytical Chemistry are \textit{not} probability distributions, at least not in the classical sense. Accordingly, PRE is neither employed here nor in our previous publication to make any statement about probabilities of signals, peaks, spectral features, or noise in spectra. Rather, it is used as a pattern recognition tool because it is sensitive to and can differentiate between spectra with different shapes, where, as will be discussed below, the spectral ‘shape’ is a result of contributions from all of the parts of a spectrum – noise, baseline, and signals.

At least in a comparative sense, the PRE values of spectra can often be estimated. This was the case for the mock spectra presented in the tutorial in our original paper on this subject,\textsuperscript{10} and we have similarly found it to be true for a good fraction of the real spectra we have analyzed. However, some spectra are not so easy to compare. For example, Figure 4.1 shows six mass
chromatograms from a liquid chromatography-mass spectrometry (LC-MS) analysis. Of course, given its simplicity, it is reasonable that Spectrum (a) should have the lowest PRE value. However, explaining why the other spectra have the PRE values they do is not such a simple task.

From Equation 4.1, it is clear that the PRE value of a spectrum consists of the sum of contributions from the individual data points. Because of this commutivity, the same PRE value will be obtained for two spectra with identical \( p_i \) values, whether or not their \( x_i \) values are in the same order. Thus there is an ambiguity in this summary statistic. Of course PRE is not the only summary statistic that contains such an ambiguity/invariance to order.\(^{26}\) These observations suggest an opportunity to view spectra differently and in a way that might make them easier to compare and even interpret, especially in the context of PRE. Here we propose that the values in a normalized spectrum might be reordered (sorted) from highest to lowest value and then replotted. We choose the term ‘reordered’ because the original spectra have a specific and original order. As we will show below, reordered spectra provide a simple and powerful graphical way of viewing and understanding PRE analysis.
Figure 4.1. Raw mass chromatograms representing relatively ‘clean’ (a – c) and ‘dirty’ (d – f) m/z channels from an LC-MS experiment. (a) m/z= 1088, (b) m/z= 635, (c) m/z=317, (d) m/z= 1383, (e) m/z= 570, and (f) m/z= 118. These data are found in the PLS_Toolbox from Eigenvector Inc. (Manson, WA) and were originally published by Windig et al.27
Figure 4.2. (i) Mock, normalized spectra containing 15 data points and (ii) the spectra from (i) in their reordered form. (a) A spectrum containing a ‘spike’ one data channel wide with all other data points having values of zero, and (b) A spectrum consisting of a series of data points with equal values.

4.2 Results and Discussion

Some of the key concepts associated with PRE and reordered spectra are illustrated in a simplified fashion in Figure 4.2. Figure 4.2i shows two mock spectra: one that consists of a series of uniform values (essentially a horizontal line) and another that contains a single spike with all other values in the spectrum equal to zero. A maximum value in the PRE function occurs when each data point in a normalized spectrum has the same non-zero, positive value, e.g., Figure 4.2i,b. In this case, for $n$ data points, $p(x_i) = 1/n$ and Equation 1 reduces to $log_2 n$. In contrast, the spectrum with the spike represents the minimum value that can be obtained by PRE, which is zero. That is, when this spectrum is normalized, the spike is given a value of unity, which yields $log_2(1) = 0$. 
The other points in this spectrum similarly contribute nothing to the PRE value because $p(x_i)log_2p(x_i)$ is zero in the limit of $p(x_i) \to 0$. Figure 2ii shows the reordered (sorted) data in Figure 4.2i. Obviously the spectrum of the set of uniform values is unchanged by this operation. However, when reordered (sorted), the spectrum with the spike consists of a point at high value (one) followed by a series of points at low value (zero). These general ‘shapes’ of reordered spectra: a horizontal line or a spike followed by a series of points at zero intensity, are important for PRE analysis because they represent the extremes of maximum and minimum PRE values, respectively.

![Figure 4.3.](image)

**Figure 4.3.** The same six spectra in Figure 1 replotted in their reordered state. Data points in the spectra are omitted wherever their values are zero – the spectra go to zero wherever the spectra end.
We noted above the challenge associated with estimating the PRE values of Spectra b – f in Figure 4.1. Figure 4.3 shows these spectra in their reordered state. Here it is much easier to rationalize their relative PRE values. The reordered spectra in Figure 4.3 initially consist of a series of overlapping peaks at short reordered time. However, at longer reordered time the spectra with higher values also have higher PRE values – the spectra with higher PRE values are ‘flatter’ and more like the set of identical values (horizontal line) in Figure 4.2ii,b, while the spectra with lower PRE values are more like the spike in Figure 4.2ii,a. Note the log scale in Figure 4.3, which facilitates better visualization of the reordered spectra here than a linear one.

**Figure 4.4.** Areas of nineteen selected peaks/regions from ToF-SIMS spectra obtained from a depth profile through a film of plasma polymerized PNIPAM on Si.
Figures 4.4 and 4.5 further illustrate the concept of reordered spectra. Figure 4 contains a series of selected peaks from a ToF-SIMS depth profile through a thin film of poly(N-isopropylacrylamide) (PNIPAM). These series of peak areas were used in a chemometrics analysis of the data from a depth profile through ca. 100 nm of this material. One of the spectra (Spectrum 24) is rather ‘spike-like’ in appearance and is expected to have quite a low PRE value (it does). In contrast, the other three spectra are more complex, although one of them (Spectrum 15) has a noticeably higher PRE value that the other two. These spectra are reordered in Figure 4.5. We again see here the correlation between spectral shape/structure and PRE value. That is, Spectrum 24, which had the lowest PRE value, is quite ‘spike-like’, Spectrum 15, which had the highest PRE value is ‘flattest’, and the other two spectra lie somewhere in between in both shape and PRE values.

A final example of the usefulness of reordered spectra is shown in Figure 6. Here, Figure 6a shows a plot of the PRE values of the spectra from the ToF-SIMS depth profile of PNIPAM just mentioned. This panel and the panel below it, which shows the backward difference (similar to a numerical derivative) of the PRE values was previously published. In these results, it is implied that Spectra 1 – 12 are different from Spectra 22 – 24. This is confirmed by the corresponding reordered spectra in Figure 6b. Here, Spectra 1 – 12 have higher PRE value, are more ‘horizontal line like’, and form a band, which suggests their similarity. Similarly, Spectra 22 – 24 have lower PRE values, are more ‘spike like’, and form a different band.
Figure 4.5. The same four spectra that are in Figure 4 plotted in their reordered state. Data points in the spectra are omitted wherever their values are zero – the spectra go to zero wherever the spectra end.
Figure 4.6. (a) PRE values of a series of spectra from a depth profile through a ca. 100 nm film of PNIPAM on silicon. (b) The backward difference PRE values from (a). (c) Spectra 1 – 12 and 22 – 24 from this depth profile in their reordered state.

4.3 Conclusions

We have presented reordered spectra as a tool for comparing and understanding spectra, especially in the context of PRE analysis. At both a practical and a theoretical level, reordered spectra correlate with and provide a more intuitive feel for PRE values.
4.4 Supporting Information Available

Supporting Information showing PRE and reordered spectra analyses of various X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) spectra is available in Appendix 6.

4.5 References


26. There may actually be some use for this invariance, i.e., this fact points to a possible method for identifying spectra that contain shifted features but that are otherwise identical; PRE analysis will show a measure of shift invariance.
Chapter 5: Using Pattern Recognition Entropy to Select Mass Chromatograms to Prepare Total Ion Current Chromatograms from Raw LC-MS Data

5.1 Introduction

Liquid chromatography - mass spectrometry (LC-MS) couples the separation capability of liquid chromatography (LC) with the detection specificity and versatility of mass spectrometry (MS). The latter provides information rich mass spectra for each eluting chromatographic peak, beyond that which can be derived from ultraviolet-visible (UV-Vis) absorption detection alone.\textsuperscript{1-2} However, the efficient coupling of LC to MS remains challenging, specifically the efficient transfer of chemical or biochemical species from a condensed phase into the vacuum environment of an MS detector with concomitant ionization. This transition requires the elimination of the majority of the mobile phase, volatilization of the remaining solvent and solute species, and ionization of the analyte, e.g., via electrospray ionization. Another challenge is the high degree of noise that can be present in total ion current chromatograms (TICCs) as a result of chemical and other forms of noise present within each individual mass-to-charge ratio (m/z) signal recorded, i.e., mass chromatogram.\textsuperscript{3-6} (A mass chromatogram or extracted ion chromatogram in LC-MS gives the intensity of an ion at a particular m/z value as a function of elution time (no. of scans). The TICC sums all the mass chromatograms in a
separation. The terms ‘mass chromatogram’, ‘extracted ion chromatogram’, and ‘total ion current chromatogram’ are favored by IUPAC, where the first two of these terms are synonymous. Chemical noise, which can come from solvents, buffers, additives, tubing, and/or other chemical impurities or instrument components, can strongly affect the limits of detection of analytes. This noise can be hard to remove because its pattern can mirror that of the signal. Substantial noise can arise in electrospray ionization or atmospheric pressure chemical ionization (APCI). Mathematically, noise can be of high or low frequency. A prime example of high frequency noise is unwanted transients (spikes). Low frequency noise presents itself as baseline drift, which may be due to gradual elution of strongly retained species from the LC column, other trace contaminants within the mobile phase (particularly during gradient separations).
Figure 5.1. The total ion current chromatogram (TICC) constituted from all the 1451 mass chromatograms considered in this study.

A number of software and hardware approaches have been devised to resolve these problems. Hardware approaches have included the implementation of declustering (desolvation) conditions.\textsuperscript{13,15} These techniques have focused on improving ionization selectivity and prevention of contaminants in ion sources. Often, the complexity of background ions results from inadequate noise reduction.\textsuperscript{9,13} Tandem mass spectrometry (MS/MS) has been proposed as an alternative to improving MS detection specificity. Chemical methods, including the use of neutral reagents like
dimethyl disulfide (DMDS) that selectively react with background ions, have also been developed. \textsuperscript{5, 13} Software techniques provide a different route to obtain high quality TICCs, although they too have their drawbacks. For example, baseline corrections \textsuperscript{16} can suppress extremely small features (trace analytes) in a TICC. Peak picking based on analysis of individual mass chromatograms has also been demonstrated. \textsuperscript{8, 17} The Biller-Biemann method selects mass chromatograms based on their intensities. \textsuperscript{18} However, information extraction by this method is inaccurate and the quality of the resulting TICCs is low. It is not unusual for TICCs to be poorly defined, even while some of the individual mass chromatograms may be of high quality. \textsuperscript{3} Enhancement of the signal-to-noise (S/N) ratio in TICCs has been achieved by matched filtration with experimental noise detection, \textsuperscript{6} the sequential paired covariance, \textsuperscript{19} chemical background noise correction, \textsuperscript{16} and windowed mass selection. \textsuperscript{20} However, in many cases post-acquisition data processing results in distorted MS peak shapes that can compromise mass accuracy. \textsuperscript{20} Figure 5.1 shows the TICC that is the summary of the data considered in this work. It has an irregular baseline and a poor signal-to-noise ratio.

One obvious approach to improving the quality of TICCs is for the operator to manually examine the data and extract the mass chromatograms that appear to contain useful chromatographic information. \textsuperscript{3, 21-22} In general, the mass chromatograms that are thus selected will be devoid of significant noise, contain one or more peaks of reasonable width, and exhibit low, regular baselines. However, manual examination is a tedious and time intensive procedure that requires a trained operator. To more fully automate this process, the component detection algorithm (CODA) \textsuperscript{3, 23} was designed to select mass chromatograms with a minimal amount of high frequency noise and low backgrounds. CODA DW was then developed to account for drifting
baselines. That is, with CODA, a mass chromatogram with a real solute peak is discarded if a drift in the baseline is present. Obviously, this can result in a loss of chemical information. CODA DW on the other hand selects these mass chromatograms.

5.2 Theory of Pattern Recognition Entropy (PRE) Analysis on Mass Chromatograms

Claude Shannon related the statistical thermodynamic concept of entropy to the amount of information in a signal, providing a mathematical description/quantification for it. As stated before, information is defined as the distribution of the probabilities of a series of events in a message in its context. In this study, Shannon’s ideas on the quantification of information/disorder in a data stream, combined with our previous modification of this theory, have been adapted for identifying noisy mass chromatograms from an LC-MS data set. For the PRE calculations performed herein, we simply define the ‘probabilities’ in a mass chromatogram as the values in that mass chromatogram obtained by normalization, i.e., after division of each data point, in a mass chromatogram by the sum of all the data points in that mass chromatogram, where a data point is the intensity value of an individual ion (m/z) at a particular scan number (time). Statistically, this normalization is referred to as the ‘1-Norm’. Accordingly, in this paper, the ‘H’ in Equation 1 is replaced with ‘PRE’. This is done both to avoid any confusion with thermodynamic enthalpy (H), as well as to emphasize the pattern/shape recognition capabilities of this data analysis algorithm.

Obviously, mass chromatograms do not conform to the principles of simple probability. That is, the preprocessing (1-Norm) we perform on our data sets (mass chromatograms) does not
create true probabilities. However, it does allow us to use Shannon’s formula as a pattern recognition tool that can distinguish between/select specific mass chromatograms with different shapes i.e., between those with relatively few features and those with many. As taught in the tutorial in our first paper on this subject, a larger PRE value points to a more complex data set – more data points \( (x_i) \) with significant values. Thus, noisier mass chromatograms generally have higher PRE values. On the other hand, desirable mass chromatograms are more likely to be dominated by one or just a few clean peaks, i.e., peaks with good signal-to-noise ratios, so they will have lower PRE values. Higher quality mass chromatograms can thus be selected based on their lower PRE values and then combined to a form a reduced TICC. Note that the PRE value is a summary statistic, i.e., it reduces an entire mass chromatogram to a single number. For the interested reader, we have created a mock mass chromatogram, and demonstrated its normalization and the calculation of its PRE value in the Appendix 7. We have previously applied PRE analysis to identify transitions in X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) depth profiles. Here, PRE clearly differentiated between spectra containing signal and those with mostly noise. In addition, we have recently introduced the concept of the reordered spectrum to provide a more visual, intuitive representation of PRE analysis.

In this work, PRE values are calculated for all the mass (m/z) chromatograms in an LC-MS data set. A plot of the PRE value for each mass chromatogram vs. its corresponding m/z value is then fitted to a spline curve. The distribution of the distances between the spline fit and the PRE values is next used to identify the mass chromatograms with lowest relative PRE values, i.e., the PRE values below a certain threshold. The selected mass chromatograms are then combined (added) to form a reduced TICC. Significant improvements are observed compared to the original
TICC. Further processing of the reduced TICC is possible, e.g., by smoothing. Compared to CODA, PRE may involve less mathematical processing and user input. Note that van der Greef and coworkers\textsuperscript{28} worked on a similar approach for selecting mass chromatograms. In their algorithm, they first smoothed the mass chromatograms and then baseline corrected them. They next inserted the raw intensity values of the corrected mass chromatograms, or the square of these values, into Shannon’s formula. Finally, they specified a quality factor (IQ = 1-H/H\textsubscript{max}) based on the relative entropy value of a mass chromatograms, H, with respect to the maximum entropy value, H\textsubscript{max}, found in the data set. In our experience, however, pre-processing of data in such a fashion can result in loss of minor features, thus making the algorithm less sensitive to small chromatographic peaks. In addition, the lack of normalization means some of the graphical and accompanying mathematical interpretation of entropy is lost – in general a single spike will no longer have an entropy value of zero. We also believe that the normalization of each mass chromatogram is a better way to compare them to each other as opposed to comparing them to whatever mass chromatogram happens to give the highest entropy value (H\textsubscript{max}). Finally, the van der Greef approach does not account for the shift in average entropy value in the mass chromatograms as a function of m/z value.

5.3 Experimental

5.3.1 LC-MS

The following experimental details were previously reported in the work that generated the data set analyzed herein; note that this data set has previously been used to study the CODA and
CODA DW mass chromatogram selection algorithms.\textsuperscript{3,21-22} The LC-MS chromatograms analyzed herein came from a proprietary surfactant mixture containing at least 15 components that were separated on a Hewlett Packard Hypersil ODS 5 \textmu m column (100 mm x 2.1 mm) using a gradient of 65\% methanol/35\% 0.1 M ammonium acetate in water to 95\% methanol/5\% 0.1 M ammonium acetate in water at a flowrate of 0.3 mL/min on a Hewlett Packard 1090 liquid chromatograph (Analytical Products Group, Waldbronn, Germany). This was connected to a Fisons electrospray interface (electrospray cone voltage of -10 V) and a Fisons Instruments Quattro mass spectrometer (VG BioTech, Altrincham, UK), which was scanned from 50 – 1500 Da every 5 s with a 0.2 s interscan delay. The data from this separation are included in the PLS Toolbox (version 8.1) by Eigenvector Research, Inc. (Manson, WA, USA), which runs under the Matlab computing environment (Version R2015b, Release No. 8.6.0.267246, The Mathworks Inc., 1 Apple Hill Drive, Natick, MA, USA). The computer used for this work was an Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz with 16.0 GB of RAM on a 64-bit Windows 7 Enterprise Edition operating system.

5.3.2 Spline Fit to the Data

The curve of PRE value of the individual mass chromatograms vs. m/z value was fit using a smoothing spline algorithm in MATLAB (the ‘fit’ function).\textsuperscript{29} This curve contained a user inputted smoothing parameter, q, and weights for the individual data points, w_i, where the parameter, q, is defined between 0 and 1, and the weights are unity unless otherwise defined (in our smooth they are all the same: w_i = 1). If q = 1, the smoothing parameter attempts to create a
cubic spline interpolant that will fit every data point. For our data this will create extreme oscillations – this spline will be unreasonable and unphysical. At $q = 0$, the data are fit to a straight line. To find an appropriate value of $q$, its value was decreased from unity by factors of 10, i.e., we considered $q = 1, 0.1, 0.01$, etc. Ultimately, a $q$ value of 0.00001 was selected, which appeared to follow the general trend/mean/variation in the data without undue oscillations. (Note that in the MATLAB spline algorithm the smoothing parameter is actually ‘p’. However, to avoid confusion with the ‘p’ values in Shannon’s formula and the calculations performed herein, we have changed it to ‘q’.)

5.3.3 Savitzky-Golay Filter

Smoothing of the reduced TICC was achieved with a Savitzky-Golay (SG) filter in MATLAB (the ‘sgolayfilt’ function), which smooths the data by approximating it to a polynomial of order that can be specified in MATLAB. The frame length, which is the width of the smoothing window (number of consecutive data points) also needs to be specified. In addition to the spline fit mentioned above, an SG filter was applied to the data points in the plot of PRE value vs m/z value (see Appendix 7).
5.4 Results and Discussion

Figure 5.2a – b shows two mass chromatograms that contain clearly visible chromatographic peaks and relatively low levels of noise. Figure 5.2c shows a mass chromatogram that also appears to contain meaningful chemical information, but with more noise. Figure 5.2e – f shows three mass chromatograms that contain high degrees of noise and appear to be devoid of useful chemical information. As expected, the ‘cleanest’ of these mass chromatograms (Figure 5.2a) has the lowest PRE value (3.7), and what are arguably the noisiest mass chromatograms (Figure 5.2e – f) have the highest PRE values (6.6 and 7.3). The mass chromatograms in Figure 5.2b – d appear to have intermediate complexity and indeed have intermediate PRE values. However, Figure 5.2d looks noisier than Figures 5.2b and c, although it has a lower PRE value. This apparent contradiction is reconciled in two ways. First, upon reordering of the mass chromatograms, which helps visualize their PRE values, Figure 5.2d is indeed seen to have fewer high value data points – this provides an explanation for its lower PRE value. Second, to some degree, the lower PRE value in Figure 5.2d is a plotting artifact. The thicker lines used in this figure are appropriate and even necessary for publication in a journal. However, they overestimate the contributions of a series of spikes in this mass chromatogram. This figure is replotted in Appendix 7 with reduced line widths.
**Figure 5.2.** Raw mass chromatograms representing relatively ‘clean’ and ‘noisy’ mass chromatograms at (a) m/z= 1088, (b) m/z= 635, (c) m/z=317, (d) m/z= 1383, (e) m/z= 570, and (f) m/z= 118.

Figure 5.3a shows the PRE values of all of the 1451 mass chromatograms (from 50 – 1500 Da) considered in this study as a function of m/z value. It is evident in this plot that there is a band of points that cuts across it with a few outliers above it and quite a few below it. The PRE values are higher in magnitude at lower m/z values (50 – 200 Da) and lower at progressively higher m/z values. This implies that the higher mass chromatograms are, in general, less noisy than the lower mass chromatograms. This trend in the mass chromatograms may be a result of solvent ion clusters of mobile phase components, e.g. water playing a major role in the ion currents of lower m/z mass.
fragments. Ultimately, however, the most important and interesting part of this plot may be the presence of outliers with lower PRE values compared to those in the band of values around them. We individually analyzed the mass chromatograms corresponding to these low PRE value outliers and observed that, in general, they did have better S/N ratios and were chemically relevant.

![Figure 5.3](image)

**Figure 5.3.** (a) Plot of PRE values of individual mass chromatograms vs. corresponding m/z values. (b) Spline fit to the PRE values in (a). (c) Selection of the mass chromatograms that are 1σ or more away from the spline fit.

We next looked for a way to rationally select chemically meaningful mass chromatograms with an eye towards automating this process. Of course it was clear here that a single, global
threshold could not be applied to the data because of the variations in the PRE values across Figure 5.3a. And again, what appeared to be most important in this problem was the position of a PRE value relative to the others nearest it. Accordingly, a smoothing spline over all the PRE values was fitted to the data (see the Experimental for details). The original data and spline fit are shown in Figure 5.3b. We next considered the distribution of differences between the PRE values and the spline fit, assuming the distribution to be Gaussian. Note that the assumption of a distribution (Gaussian or otherwise) is not necessary here. All that matters is that a threshold be applied to the PRE values relative to the spline fit to the band. For example, Figure 5.3c shows the PRE values that are one standard deviation, σ, away from the spline fit. A series of thresholds between 0 and 2.5 σ was then used to find the value that would yield the best TICC. (As a comparison to the spline fit, we also looked at a Savitzky-Golay (SG) fit with frame lengths of 51, 101, and 151 data points. The order of the polynomial was 3. To achieve effective smoothing, the frame length was much larger than the polynomial order and comparable to the feature size (about 50 data points). Results with frame lengths of 101 and 151 data points were similar to those obtained with the spline fit in Figure 5.3 (see Appendix 7).)

Figure 5.4 shows the resulting TICCs, which, again, were created by combining all of the mass chromatograms with PRE values below a given threshold. As seen in Figures 5.4a – b, which correspond to σ = 0 and 0.25, respectively, the noise level and background, especially at low scan numbers (time), are fairly high. Arguably the best TICC is obtained at σ = 0.5 (Figure 5.4c). After that, the chromatographic peaks become less distinct relative to each other and peaks are gradually lost. Finally, at σ = 2.5 (Figure 5.4i) only one mass chromatogram remains. It is noisy and clearly a false positive. As can be observed in Figure 5.4, there are changes in peak intensities in the TICC.
as the threshold is increased. This phenomenon is a result of the fact that (i) multiple mass chromatograms may contribute to a single solute peak, (ii) different mass chromatograms will generally contribute to different peaks in the TICC, and (iii) these mass chromatograms have different PRE values. Thus, as the threshold changes, the selected mass chromatograms will change. For example, the mass chromatograms in Figure 5.2(a) and 2.2(c) clearly contain useful chromatographic information, but the differences between their PRE values and the spline fit in Figure 5.3b are different (2.131 and 1.908 along the y-direction in Figure 5.3c, respectively). Accordingly, as the threshold is increased, the mass chromatogram in Figure 5.2c will be lost, followed by the mass chromatogram in Figure 5.2a. This will result in a change in the relative peak intensities in the TICC.
Figure 5.4. Reduced total ion current chromatograms (TICC) obtained by varying the threshold ($\sigma$) for the data points in Figure 5.6 relative to their spline fit. (a) 0.0 $\sigma$ (614), (b) 0.25 $\sigma$ (320), (c) 0.5 $\sigma$ (151), (d) 0.75 $\sigma$ (84), (e) 1.0 $\sigma$ (40), (f) 1.25 $\sigma$ (26), (g) 1.5 $\sigma$ (13), (h) 2.0 $\sigma$ (3), and (i) 2.5 $\sigma$ (1). The numbers in parenthesis here are the number of mass chromatograms summed to make the corresponding TICCs.

In order to test our hypothesis that the mass chromatograms below a given threshold are of higher quality than the unselected mass chromatograms, an inverse selection of the mass chromatograms was performed. Here, TICCs were constituted from all the mass chromatograms above two thresholds (0.0 $\sigma$ and 0.5 $\sigma$). As expected, these TICCs were of very low quality. They contained substantially rising baselines, high levels of noise, and little if any discernible signal (see Appendix 7).

The reduced TICC at the 0.5 $\sigma$ threshold (Figure 5.4c and Figure 5.5a) was constituted from 151 mass chromatograms. All of these 151 mass chromatograms are overlaid in a single plot in the Appendix 7, which shows how the TICC was constructed and particular peaks were formed. For example, the peak just above scan number 150 is a result of two overlapping peaks from two different mass chromatograms. And while the TICC in Figure 5.5a is a little noisier than the one obtained by CODA (Figure 5.5b), the small peaks around the largest peak in Figure 5.5a are arguably better defined. Thus, it appears that the PRE-based approach yields a TICC that is comparable to and in some ways better than the CODA TICC. Practically speaking, one could imagine a slider bar in a program that would allow the user to select the $\sigma$ threshold while watching improvements in the TICC until the best one is found. Additional improvements to the PRE TICC were obtained through a Savitzky-Golay smooth (Figure 5.5c). The smoothing parameter here (7
data points) was chosen based on the feature size, which in this case is the full width half maximum (FWHM) of the chromatographic peaks (generally 6 to 11 data points wide). Our PRE analysis is mathematically and conceptually simpler than the CODA approach. Note, however, that CODA does have a higher degree of variable reduction compared to the PRE approach at the 0.5 σ threshold: 52 vs. 151 channels.

**Figure 5.5.** Comparison of the TICC obtained via (a) PRE with with 0.5 σ (b) CODA and (c) PRE with smoothing. The CODA result was obtained with the default smoothing window of 5 and MCQ threshold value of 0.89.
A closer look at the variables selected by PRE reflects some interesting properties of the algorithm. Indeed, when using the 0.5 $\sigma$ threshold, PRE separated similar peaks originating from the same component. For example, the PRE-based algorithm excluded the noisier mass channel at m/z = 627 Da (Figure 5.6a), while selecting the cleaner one at m/z = 626 Da (Figure 5.6b). Similarly, the noisier m/z = 1091 Da (Figure 5.6c) channel was excluded while the simpler m/z = 1089 Da one (Figure 5.6d) was included. In other words, different isotopes have different abundances and therefore intensities, which means they have different signal-to-noise ratios and different PRE values. Accordingly, these signals can be differentiated from each other by PRE. In general, PRE makes this distinction more readily than CODA or CODA DW.

**Figure 5.6.** Comparison of mass chromatograms arising from the same components (a) m/z= 627 Da and (b) m/z= 626 Da, and (c) m/z= 1091 Da and (d) m/z= 1089 Da.
Finally, there will be other more advanced approaches for creating TICCs by PRE. For example, we have had some success in segmenting the PRE vs. m/z value plot and applying different thresholds to different parts of it. That is, Figure 5.3a might be divided into low, medium, and high m/z regions. Ultimately, however, this approach requires more thought and user involvement than mass chromatogram selection based on a single threshold. A second possibility would be to manually examine the mass chromatograms below one’s σ threshold and to remove any that contain only noise. As shown in Figure 5.4i, at least some of the mass chromatograms selected by the PRE algorithm contribute nothing but noise to the reduced TICC. But again, one must question whether it is worth the extra effort to do this when setting a single threshold for σ and perhaps smoothing the result yield a high quality reduced TICC in a fairly straightforward fashion.

5.5 Conclusion

PRE is a summary statistic and shape recognition tool. We have developed it here into an algorithm for selecting mass chromatograms. PRE analysis successfully selects information rich mass chromatograms leading to the creation of higher quality reduced TICCs. In the example considered herein, PRE selects 151 out of 1451 mass chromatograms with a 0.5 σ threshold. This threshold will probably differ between data sets, separation conditions, instrument parameters, and the nature of the analytes. In general, the unselected (unfavored) mass chromatograms contain significant noise and baseline drift while the selected ones have meaningful and/or higher signal-to-noise ratio chromatographic peaks. Our reduced TICC was comparable in quality to the one
produced by the widely acknowledged CODA algorithm. PRE is easier to understand than CODA. Moreover, PRE requires a single threshold as compared to an MCQ threshold and a smoothing window in CODA. Another advantage of PRE over CODA is its ability to separate isotopic peaks that are similar in noise levels and that arise from the same components. Nevertheless, the algorithm selects some false positives and negatives. Future work on PRE will focus on the elimination of these occurrences.

5.6 Acknowledgement

We thank Willem Windig, the originator of CODA and CODA DW, for his useful comments and suggestions.

5.7 Supporting Information

Supporting Information is available in Appendix 7.

5.8 References


Chapter 6: Using Cross-Correlation with Pattern Recognition Entropy to Obtain Reduced Total Ion Current Chromatograms from Raw Liquid Chromatography-Mass Spectrometry Data

6.1 Introduction

The total ion current chromatograms (TICCs) obtained in liquid chromatography-mass spectrometry (LC-MS) are often limited by high levels of chemical and other electronic noise, making the subsequent extraction of real chromatographic information difficult. The noise in TICCs arises from the noise present in their constituent mass chromatograms, which can have both high frequency (transients and/or spikes) and low frequency (baseline drift) components. Hardware approaches optimizing the transfer of eluents from the liquid chromatograph to the mass spectrometer have been devised to reduce chemical noise. However, limited success has been achieved through these techniques, and LC-MS analysis often relies on post-processing of the TICC to obtain adequate information about analytes. In general, unless noisy mass chromatograms are excluded, poor quality TICCs are obtained.

In an attempt to improve the way mass chromatograms are selected and to create high quality TICCs, we recently demonstrated a variable selection method, described in Chapter 5 based on Pattern Recognition Entropy (PRE) that effectively extracted a high quality TICC from the...
LC-MS analysis of a surfactant mixture. One difference between PRE and other commonly used informatics/chemometrics approaches like principal component analysis (PCA) and multivariate curve resolution (MCR) in that PRE is a summary statistic.\textsuperscript{13-16} Our approach involved the identification of information-containing (high signal-to-noise ratio) mass chromatograms based on their PRE values. Here, information-containing mass chromatograms consisted of well-defined peaks on low backgrounds and showed lower PRE values (a relatively smaller number of higher intensity features). Noisy mass chromatograms had higher PRE values due to a greater number of higher probability events (a relatively larger number of higher intensity features). Thus, in a plot of the PRE values of the mass chromatograms vs. the corresponding m/z values, the desirable mass chromatograms should occupy the lowest part of the band of values. To automate the process of selecting mass chromatograms, an algorithm was designed and code written that was based on the distances between the PRE values of the individual mass chromatograms and a spline that was fitted to all the PRE values of all the mass chromatograms over the entire m/z range. The selected mass chromatograms were combined to form the reduced TICC. The peaks of the resulting TICC were greatly reduced in noise such that the information from the analyte mixture was much more discernable. Indeed, PRE generated a reduced TICC that was comparable, if not of higher quality, than the Component Detection Algorithm (CODA).\textsuperscript{17} Several methods were proposed for producing TICCs with even lower levels of noise. One involved application of a Savitzky-Golay smooth. However, smoothing has inherent disadvantages.\textsuperscript{12} In general, a smooth is a convolution,\textsuperscript{18} which means that the higher frequency components, i.e., detailed information, of a smoothed chromatogram may be lost or suppressed. Our previous work contains additional information on the theoretical treatment of PRE, the normalization of the mass chromatograms, the PRE
calculations, and the formation of reduced TICCs. An approach similar to ours was previously taken by Greef et al.

In this work, we present a pattern recognition entropy-cross-correlation (PRE-CC) algorithm for better selecting LC-MS mass chromatograms. This approach improves upon the previously reported PRE algorithm by further eliminating poor quality mass chromatograms that might otherwise be incorporated into the final TICC. PRE is used here as the initial selection tool, which is followed by a cross-correlation of the PRE-selected mass chromatograms with a boxcar signal that has a width similar to that of the peaks in the chromatogram. More specifically, authentic peaks in our chromatographic data set have widths of 6 – 10 time points. Accordingly, cross-correlation (CC) between a boxcar function of this approximate width and mass chromatograms that contain real peaks should result in a higher maximum value than cross-correlation with mass chromatograms that contain only noise. The maximum CC value that is obtained from a chromatogram can thus be used as an additional figure of merit to eliminate inferior quality mass chromatograms that might otherwise be incorporated into a TICC. This approach is shown here to be successful. At the 0.5 σ threshold recommended for PRE in our previous publication, a CC threshold can be set that eliminates all the false positives selected by PRE, which was ca. one-third of the total number of mass chromatograms selected (47 out of 151), while not removing any of those of high (98) or intermediate (6) quality.

**6.2 Theory**
Pattern Recognition Entropy (PRE), which is based on Claude Shannon’s original work on information theory,\textsuperscript{20-21} is both a pattern recognition tool and a summary statistic. Shannon’s quantification of the information content in a data stream, \( x \), called Shannon’s entropy, \( H \), is given in Equation 1.

\begin{equation}
H (x_i) = - \sum_{i=1}^{n} p(x_i) \ast \log_2 p(x_i)
\end{equation}

where \( p(x_i) \) are the probabilities of each data point, \( x_i \), in the data stream, \( x \).

In PRE, pseudo-probabilities are defined by dividing each data point in a data set (here mass chromatograms) by the sum of all the data points, i.e., the 1-Norm.

Cross-correlation has been widely used in communication theory, serving as an elegant technique for analyzing both periodic and aperiodic signals. The cross-correlation of two spectra, \( A(x) \) and \( B(x) \), taken at equal time intervals, is given in Equation 2.

\begin{equation}
C(\tau) = \sum A(x) \ast B(x + \tau)
\end{equation}

where \( \tau \) is an integer that indicates the shift of signal \( B \) with respect to signal \( A \), and the summation here runs over the intersecting regions of the two signals. Traditionally, cross-correlation has been used for noise filtering. For example, signal \( A(x) \) might contain both noise and signal components, while \( B(x) \) could be a reference signal with little or no noise in it. In most cases, cross-correlation is computed by transforming to and from the frequency (inverse variable) domain via the Fourier transform.\textsuperscript{18}
In this study, to identify ‘good’ mass chromatograms, a reference signal is used that has the approximate full width at half maximum (FWHM) of the peaks in the chromatogram. The maximum value in the resulting cross-correlation should occur where the reference signal and largest chromatographic peak in the mass chromatogram overlap. That is, mass chromatograms with a clearly defined chromatographic peak result in more overlap with the reference signal and a correspondingly higher maximum cross-correlation value. This maximum value can then be used as a figure of merit to assess the quality of the mass chromatograms. No pre-processing of the mass chromatograms was performed prior to these cross-correlations.

6.3 Experimental

The dataset analyzed here was employed in our previous study on mass chromatogram selection by PRE. It was originally used and reported by Windig and Payne. More specifically, the dataset consists of 1451 LC-MS mass chromatograms from 50 – 1500 Da that originated from a separation of a proprietary surfactant mixture with approximately 15 distinct compounds. Details of how this separation was performed are included in the original Windig and Payne paper and in our previous work. This LC-MS dataset is included in the PLS Toolbox (version 8.1) by Eigenvector Research, Inc. (Manson, WA, USA), which functions as a part of the Matlab computing environment (Version R2015b, Release No. 8.6.0.267246, The Mathworks Inc., 1 Apple Hill Drive, Natick, MA, USA). All the computations in this study were coded in MATLAB and run on an Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz computer with 16.0 GB of RAM on a 64-bit Windows 7 Enterprise Edition operating system.
6.4 Results

Figure 6.1 summarizes the results that were obtained in our previous work. It compares the poor-quality TICC that is the sum of all the 1451 mass chromatograms considered in this study (Figure 6.1a) to the results we have obtained via our mass chromatogram selection. Figure 6.1a has an irregular baseline, a high noise level, and few, if any, discernable chromatographic signals. In our previous work, we applied the PRE algorithm to the corresponding mass chromatograms and compared the resulting TICC with the result generated by the widely used CODA algorithm (Figure 6.1b). At a selected threshold (-0.662, i.e., 0.5σ), the PRE algorithm selected 151 mass chromatograms, which were summed to create a greatly improved reduced TICC (Figure 6.1c). Some of the smaller chromatographic peaks in this reduced TICC appeared to be better defined than those generated by the CODA algorithm, i.e., those around the largest peak in this TICC. However, the PRE-produced, reduced TICC contained a somewhat higher noise level. The signal-to-noise level in this TICC could be improved with a Savitzky-Golay smooth (Figure 6.1d).
Figure 6.1. A comparison of the TICCs generated by: (a) summing the 1451 mass chromatograms considered in this study, (b) CODA, (c) PRE only (threshold of -0.662), (d) PRE with a Savitzky-Golay smoth (PRE threshold of -0.662 and a smoothing width of seven data points), and (e) PRE-CC (thresholds of -0.662 and 0.275 for PRE and CC, respectively).
To better understand the source of the noise in the PRE-generated TICC in Figure 6.1c, which may be due to PRE selecting noisy mass chromatograms that are then incorporated into the TICC, each of the 1451 (50-1500 Da) mass chromatograms in the dataset was individually inspected. In this process, the mass chromatograms were classified based on the signals in them and their overall signal-to-noise ratios. Three groups were formed. Chromatograms with quality signals devoid of significant noise were labeled ‘green’, chromatograms with lower quality peaks, lower signal-to-noise ratios, and minor baseline drifts were labeled ‘yellow’, and chromatograms that lacked any meaningful information or that displayed very irregular baselines or an overwhelming amount of noise were labeled ‘red’. All in all, we categorized 128 mass chromatograms as good, 63 as intermediate, and 1260 as poor. Figure 6.2 shows representative good (green), intermediate (yellow), and poor quality (red) mass chromatograms from our data set. Of the 151 mass chromatograms that PRE selected at the 0.5 \( \sigma \) threshold recommended in our previous paper, 98 were of high quality (green), 6 were of intermediate quality (yellow), and 47 were of low quality (red). It was clear that PRE was largely able to filter out the bad mass chromatograms. However, the presence of 47 bad chromatograms among the 151 that were selected is still an issue that contributed to the moderate level of noise in the resulting TICC. It appeared that an additional filter would be needed to remove these poor quality mass chromatograms.
Figure 6.2. Representative mass chromatograms identified as (a) good (green), (b) intermediate
As noted above, cross-correlation is used to confirm or detect the presence of a signal in a waveform/spectrum. It does so by multiplying a reference signal by the signal in question, where the two signals are shifted relative to each other to identify the position of the signal in the waveform. The signal/feature we wished to identify in the mass chromatograms was chromatographic signals. Accordingly, we looked at the full width at half maxima (FWHM) of all the meaningful chromatographic peaks in the 1451 mass chromatograms. That is, we examined every chromatographic peak in every mass chromatogram that contained meaningful information (the mass chromatograms in the green and yellow categories). These peaks were found to have average FWHM values of 7.8 scans. Accordingly, for simplicity, we used a rectangular pulse (boxcar signal) with a width of 8 scans as our reference signal. Obviously, other pulse shapes, e.g., Gaussian pulses, could be contemplated for this task. The figure of merit for our cross-correlation algorithm was the maximum value in the resulting cross-correlation analysis/spectrum (CC maximum).

To understand whether cross-correlation might provide any useful information beyond PRE, i.e., whether it could be used as an additional variable selection tool/filter for mass chromatograms, both PRE values and CC maxima were calculated for each of the 1451 mass chromatograms under consideration in this study. To create a visual representation of this information, the PRE value and CC maximum for each mass chromatogram were considered to form an (x,y) data pair. The resulting plot (Figure 6.3) contains a colored point for each mass chromatogram, where the color (green, yellow, or red) indicates the quality of the corresponding mass chromatogram. It is evident in this plot that the higher quality mass chromatograms fall in
the upper left quadrant of it. The data points here have lower PRE values and higher CC maxima. In our previous work, we recommended a threshold for the PRE value of -0.662 (0.5 standard deviation below the spline fit) that defined a high quality, reduced TICC. This threshold for PRE is given by the heavy, dashed vertical line in Figure 6.3. Figure 6.3 further suggested a threshold for the CC maxima of 0.275 where no ‘red’ mass chromatograms were included in the upper left quadrant. Figure 6.4 shows the reduced TICCs obtained from all the mass chromatograms in the four quadrants defined by the two thresholds (dashed lines) in the plot. It is evident that Quadrant 2 in the plot (Figure 6.4b) yields the highest quality TICC, while the TICCs from the other quadrants are largely devoid of meaningful chromatographic information.
Figure 6.3. A plot representing all of the 1451 mass chromatograms used in our study, where the position of the mass chromatogram (data point) in the plot is given by its PRE value (x-axis value) and CC maximum value (y-axis value), and the quality of the mass chromatogram is denoted by a color: green for highest quality, yellow for moderate quality, and red for poor quality.

Figure 6.4. TICCs generated by summing the mass chromatograms in Quadrants 1 (a), 2 (b), 3 (c) and 4 (d) in Figure 3, which are defined by the heavy, dashed lines in the plots.
Obviously, the thresholds for PRE and CC should be set to capture the maximum number of good (‘green’ and perhaps ‘yellow’) mass chromatograms in the data set. In our previous paper,\textsuperscript{12} we explored the effect of the variation in PRE threshold on the reduced TICC. As the PRE threshold was positioned further and further below the spline fit to the data, the algorithm became more selective and the resulting TICC had a higher signal-to-noise ratio. Ultimately, however, when the threshold was too far below the spline, the analyte peaks were lost in the TICC. An appropriate compromise was found at -0.662 (0.5 \textsigma) below the spline fit. Similarly, we have explored the effect of the CC threshold on the reduced TICC after first applying the PRE filter with the threshold shown in Figure 6.3. Figure 6.5 shows the TICCs that are obtained as the CC threshold is gradually changed. A steady increase in the signal-to-noise ratio of the TICCs is evident as the CC threshold varies from a value of 0.0, i.e., no threshold (Figure 6.5a), to 0.275 (Figure 6.5c). However, when the CC threshold is set too high, e.g., at 0.5 (Figure 6.5d), ‘good’ mass chromatograms with chemically relevant information are excluded such that real chromatographic peaks are lost. A high quality TICC was obtained when the PRE and CC maximum values were set at -0.662 and 0.275, respectively (see Figure 6.5c).
Figure 6.5. Change in TICCs at a fixed PRE threshold of -0.662 (0.5 σ) as a function of the CC maximum threshold: (a) 0.0, (b) 0.1, (c) 0.275 and (d) 0.5.
6.5 Conclusions

As a variable selection and shape recognition tool, PRE was previously used to identify high quality mass chromatograms and subsequently create improved TICCs. However, reduced TICCs generated by PRE contain levels of noise that suggests the selection of false positives by the algorithm. To eliminate false positives (noisy mass chromatograms) and thus improve the accuracy and selection capability of the PRE algorithm, we have introduced a second variable selection tool based on cross-correlation. Of the 151 mass chromatograms originally chosen by PRE, 98 were of high quality, 6 of intermediate quality, and 47 of poor quality. The application of CC after the PRE algorithm with a CC maximum threshold value of 0.275 allows all of the poor quality mass chromatograms selected by PRE to be eliminated and a higher quality TICC to be generated. The new algorithm of PRE followed by CC, which only requires two user inputs (the PRE and CC thresholds), is a simpler alternative to the well-known CODA-DW, where the peaks of the resulting TICC from PRE-CC are arguably better defined.

The applicability of PRE-CC in obtaining reduced high quality TICCs to other analytical chemistry separation techniques like gas chromatography-mass spectrometry (GC-MS) are not understood yet. Since PRE-CC has been demonstrated to be an effective noise reduction method for LC-MS data sets, we suspect that similar excellent results can be replicated with gas chromatography-mass spectrometry, capillary electrophoresis-mass spectrometry etc. However, the noise distribution is different for each separation technique and hence a direct application of PRE-CC as a variable selection tool is expected to be challenging. Our future work will concentrate on finding a unified PRE-CC algorithm that will enable us to extraction high quality TICCs from a wide spectrum of separation techniques.
6.6 References

Chapter 7: Informatics Analysis of Capillary Electropherograms of Autologously Doped and Undoped Blood

7.1 Introduction

Unethical methods for increasing oxygen delivery to skeletal muscle have been in existence for the last four decades despite a ban on such activity by the International Olympic Committee in the mid-1980s. Indeed, according to a World Anti-Doping Agency report, introduction of any quantity of autologous, homologous, or heterologous blood or red blood cells (RBCs) into the circulatory system constitutes doping. Of these doping methods, the detection of autologous blood transfusions (ABTs), i.e., autologous blood doping (ABD), is the most challenging. In an ABT, transfused RBCs are taken from the athlete and stored for reinfusion at a later date. Currently, ABD cannot be directly detected by regular anti-doping tests. Most anti-doping agencies rely on indirect methods, the most common of which consists of maintaining an athlete’s ‘biological passport’. ABD alters the characteristic biomarkers associated with erythropoiesis (red blood cell production). Thus, the observation of the hematological module and the monitoring of specific biomarkers allows for the detection of ABD. However, biological passport based fingerprinting of every athlete’s hematological profile is expensive and time consuming.
Recently, Harrison et al. introduced a fast (ca. 3 min.), direct capillary electrophoresis (CE) based method to detect ABD.1 This approach relies on a decrease in the zeta potential of stored RBCs, which impacts their mobility. The aging of the blood results in significant rheological changes in the RBCs, particularly a decrease in surface area and volume.8-9 Harrison’s work demonstrated the ability of CE to respond to changes in RBC distributions, i.e., ABD resulted in changes to the RBC peak envelope, indicating the presence of aged RBCs. Figure 7.1 shows the raw data from their study, which included undoped (0%) and simulated (5% and 10%) doped samples from three individuals/subjects: A, B, and C. Each electropherogram consists of a sharp peak at earlier time (ca. 1.5 min) followed by a shoulder at longer times (ca. 1.8 – 2.7 min), where the length and height of the shoulder tend to increase with increasing doping levels (see Figure 1d). Harrison et al. presented a first derivative of the data as a mathematical tool for quantifying this difference. Doping was identified by the presence of positive slopes. However, this approach was subjective, where a lack of a clear figure of merit for this approach resulted in false negatives.

The electropherograms in the Harrison study exhibited a substantial amount of variability and complexity, while still showing features that were consistent with doping.1 For example, the initial sharp peak in the electropherograms of the samples from Subjects A and C elutes at 1.5 ± 0.001 s and has a symmetric Gaussian shape (see Figures 7.1a and 7.1c, respectively). However, this initial sharp peak varies in both shape and position in the Subject B samples (see Figure 7.1b). Overall, the electropherograms of the Subject C samples are narrower than those from Subjects A and B. The shoulders following the initial peaks in Subject B have lower absolute intensities than the shoulders on Samples A and C. The raw data suggest that it will be challenging to develop a
universal informatics model that is simultaneously applicable to all three subjects and able to differentiate between 0, 5, and 10% doped samples.

Figure 7.1. Capillary electropherograms of undoped (0 %) and doped (5 and 10 %) blood samples of subjects (a) A, (b) B, and (c) C. Three replicates at each doping level are shown in each panel. (d) Three electropherograms from Subject A at 0, 5, and 10 % doping levels.
In this work, we applied three traditional informatics methods to differentiate between 0, 5, and 10% doping in three subjects. Doping levels of 5% and 10% were chosen because below 5% doping, there are no appreciable physiological effects that increases an athlete’s performance. A single unit of ABD blood transfusion into an athlete (assuming 4 – 5 L of blood for an adult) results in 10% doping. The informatics methods employed in this work included cluster analysis, principal component analysis (PCA), and partial least squares (PLS), which struggled to identify doping due to the limited size of the data set and the large natural variation in the electropherograms that was noted above. For example, cluster analysis achieved separation of the undoped samples from the doped samples at a level of three clusters, but gave meaningless results at a level of two clusters. PCA scores did not show clear clustering of any of the samples, and the PLS calibration showed large error bars after a leave-one-out cross validation of the data. Accordingly, we considered four less traditional methods: pattern recognition entropy (PRE), the Euclidean distance, a peak fitting method (Peak Fit-Integration), and the second moment (SM) to differentiate the electropherograms, all of which showed some success. Combinations of 2, 3 and 4 of the summary statistics generated from these analyses were used in an inverse least squares (ILS) analysis. The resulting ILS calibrations showed solid promise in differentiating between doped and undoped samples and to some extent between different levels of doping. Thus, this approach appears to be able to identify ABD in athletes.
7.2 Experimental

7.2.1 Sample Preparation and Data Collection

Blood samples analyzed in this study were procured from three professional male cyclists and one less active control male subject according to proper ethical practices. Each of the subjects was provided with a written informed consent document, which included details on procedures and biological data handling. Samples (200 µL) were collected by a fingertip lancing process and stored for 41 – 42 days at 4°C before being infused into freshly-drawn blood samples to replicate an autologous blood transfusion. Though the storage of whole blood in the citrate-phosphate-dextrose (CPD) buffer used in this study is potentially not the method of choice used by athletes, it is a reasonable approach to studying autologous blood doping. Indeed, this option was selected because it would not trigger any of the controls set by the Athlete’s Biological Passport (ABP) testing regimen, the most common tool used to detect blood doping. Other storage and transfusion methods, such as cryopreservation of RBCs, could trigger the ABP alarm, as the influx of RBCs without compensating for an increase in total blood volume would push an athlete above the hematocrit limit (%RBCs in total blood volume). Thus, while the doping approach taken here may not have been perfect, it was adequate to simulate what could likely take place. This protocol and study had been approved and funded by the World Anti-Doping Agency. The RBCs contained in the transfused samples were then separated and prepared for the CE separation. The RBCs were isolated via centrifugation and vortex mixing with phosphate-buffered saline (PBS) and 2.5% glutaraldehyde solutions in PBS (gPBS), after which they were given adequate time to stabilize.
The RBCs were further isolated and resuspended in a 45% w/v NaBr solution for the CE separation. A P/ACE™ MDQ capillary electrophoresis system from Beckman Coulter, Inc. (Fullerton, CA) was employed to carry out the subsequent CE analysis using fused capillaries of 365 µm outer diameter and varying internal diameter. Data were acquired every 0.25 s and monitored at 415 nm to identify the RBCs. All separations were performed at a controlled temperature of 25°C for both the sample compartment and the capillary. Further experimental details associated with the sample preparation and data collection were previously reported in the original paper published by Harrison et. al.¹

7.2.3 Computations and Data Analysis

Computer programs used to perform the calculations of pattern recognition entropy (PRE) and the Euclidean distance ($d_{Eu}$) were written in the Matlab computing environment (Version R2015b, Release No. 8.6.0.267246, The Mathworks Inc., 1 Apple Hill Drive, Natick, MA, USA). CasaXPS (Version 2.3.19PR1.0) was used for the peak fitting/area calculations. The computer used for this work was an Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz with 16.0 GB of RAM on a 64-bit Windows 7 Enterprise Edition operating system. Capillary electropherograms were organized row-wise to construct a data matrix. PCA and cluster analysis were performed using the PLS Toolbox, version 7.9.3 from Eigenvector Research, Inc., Wenatchee, WA, USA in the MATLAB programming environment. Cluster analysis was performed on the preprocessed data (preprocessing described below) using Ward’s minimum variance method.
7.3 Theory

7.3.1 Pattern Recognition Entropy (PRE)

PRE is a recent application of Shannon’s Information Theory\textsuperscript{11-13} that serves as a summary statistic and shape recognition tool for differentiating between spectra. Shannon’s entropy (H) of a data stream is defined as:

\begin{equation}
H(x_i) = - \sum_{i=1}^{n} p(x_i) \log_2 p(x_i)
\end{equation}

where the $p(x_i)$ are the probabilities associated with each data point $x_i$. $H$ is a measure of the uncertainty in the system and serves as a quantification of the total information present in a data stream. PRE is a modification of Shannon’s entropy where ‘pseudo-probabilities’ in the electropherograms are obtained by normalizing the data with the 1-Norm. Spectra with more features have higher PRE values (many data points with higher $p(x_i)$ values), and vice versa. PRE has been recently shown to be helpful in analyzing X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) depth profiles.\textsuperscript{18} The ‘reordered spectrum’ is a visual, intuitive tool for better understanding the relationship between normalized spectra and their corresponding PRE values.\textsuperscript{23} PRE has been used to select mass chromatograms to prepare high quality total ion current chromatograms in liquid chromatography-mass spectrometry.\textsuperscript{24-25} Because the CE spectra from doped and undoped blood differ in shape, PRE can be employed to differentiate and identify the samples. As illustrated in Figure 7.1d, the electropherograms of undoped blood tend to be narrower/more ‘spike-like’, i.e., they should have
lower PRE values, with the absence of a wide shoulder arising from an absence of aged RBCs, while the electropherograms of the doped samples tend to be wider/contain more evenly matched values, i.e., they should have higher PRE values.

### 7.3.2 Euclidean Distance ($d_{Eu}$)

The Euclidean distance ($d_{Eu}$) of two vectors\textsuperscript{17-18} in an $n$-dimensional space is the length of the line segment connecting them. For two vectors $u, v \in \mathbb{R}^n$, $d_{Eu}$ is defined as:

$$
(7.2) \quad d_{Eu}(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2 + \ldots + (u_n - v_n)^2}
$$

For example, $d_{Eu}$ for $u (1, 2, 3, 4)$ and $v (2, 3, 4, 5)$ is,

$$
(7.3) \quad d_{Eu}(u, v) = \sqrt{(1 - 2)^2 + (2 - 3)^2 + (3 - 4)^2 + (4 - 5)^2} = 2
$$

An electropherogram, which is a set of intensity values at distinct time points, can be considered a vector in an $n$-dimensional space, where $n$ is the total number of time points at which intensity values are recorded. Accordingly, if two electropherograms are similar, their $d_{Eu}$ value will be closer to zero. On the other hand, $d_{Eu}$ values of less similar electropherograms (or spectra) are expected to be larger.
7.3.3 Peak Fit, Integration (PFI)

As the degree of autologous blood doping (ABD) increases, the shoulder to the right of the main signal in the electropherograms generally becomes longer and higher. A commercial peak fitting software package (CasaXPS – see details above) was used to calculate the areas of the entire signals (main sharp peaks and shoulders) and the areas of just the sharp peaks. The difference between these areas was a measure of the degree of ABD. The background chosen for this purpose was the Shirley background with a five-point average, where an ‘n’ point average in this background defines a ‘2n+1’ window on each side of the region described by the background to establish its starting and ending points. The Shirley background has been widely used in XPS peak fitting,27 A higher window width for the background is preferred when the data contains a higher noise level. Figure 7.2 shows a representative Shirley background under the sharp feature of an ABD electropherogram.

7.3.4 Second Moment (SM)

The second moment (SM), of the electropherograms was calculated using the following formula:

\[(7.4) \quad SM = \sum_{i=1}^{n} y_i t_i^2\]

where, \(y_i\) are the intensity values from the electropherograms and the \(t_i\) are the corresponding time points. Here, the square of the time values enhances the intensity values at increased times. It is
relatively easy to show that the second moment is not shift invariant. To do so, we consider the second moment of a ‘spectrum’ composed of two data points: \((t_i, y_i)\) and \((t_{i+1}, y_{i+1})\):

\[
(7.5) \quad y_i t_i^2 + y_{i+1} t_{i+1}^2
\]

Here, it is assumed that the spacing between the times is \(\Delta t\), such that

\[
(7.6) \quad t_{i+1} = t_i + \Delta t
\]

so that we can write Equation 5 as

\[
(7.7) \quad y_i t_i^2 + y_{i+1}(t_i + \Delta t)^2
\]

which is equivalent to

\[
(7.8) \quad y_i t_i^2 + y_{i+1} t_i^2 + y_{i+1} 2t_i \Delta t + y_{i+1} \Delta t^2
\]

Now, it is imagined that this spectrum is shifted by \(n\) time increments, i.e., by \(n\Delta t\), which converts Equation 7 into:

\[
(7.9) \quad y_i(t_i + n\Delta t)^2 + y_{i+1}(t_i + (n+1)\Delta t)^2
\]

Expanding and simplifying this equation gives:

\[
(7.10) \quad y_i t_i^2 + y_{i+1}(t_i + \Delta t)^2 + y_i [t_i 2n\Delta t + n^2\Delta t^2] + y_{i+1} [t_i 2n\Delta t + 2n\Delta t^2 + n^2\Delta t^2]
\]

If the second moment enjoyed shift invariance, Equations 7.7 and 7.10 would be the same. However, it is clear that if \(n\) is an integer greater than 0, \(\Delta t > 0\), \(y_i \neq 0\), and \(x_i \neq 0\), the third and fourth terms in Equation 7.10 are not zero. Accordingly, we calculated the second moment of our
data set starting from the first data point in the series, and also starting just before the sharp peaks that contain useful information.

**Figure 7.2.** Analysis of the main, sharp peak (Labeled ‘1’ and highlighted in green) centered at ca. 1.5 min in an electropherogram of a 5 % doped sample from Subject A using a Shirley background.
7.3.5 Inverse Least Squares (ILS)

The governing and most simple equation for classical least squares (CLS) is \( A = KC \), where \( A \), \( K \), and \( C \) are matrices containing absorbance spectra, pure component spectra, and concentrations, respectively. As written here, \( K \) organizes the pure component spectra column-wise. CLS models spectra as linear combinations of pure component spectra. Inverse least squares (ILS) is based on a similar equation: \( C = PA \). That is, ILS directly relates measured spectra to concentrations through a matrix \( P \). To develop an ILS calibration, i.e., to solve for \( P \) when \( C \) and \( A \) are known, one must first right-multiply both sides of \( C = PA \) by \( A^T \). The resulting matrix \((AA^T)\) will only have an inverse, i.e., not be rank deficient, if it has at least as many columns as it does rows. That is, ILS requires that there be at least as many samples as there are data points in \( A \). Many spectra, e.g., electropherograms, contain hundreds or thousands of values, and it is not generally feasible to work with hundreds or thousands of specimens (spectra). Hence, a variable reduction technique is necessary for the ILS to function. In this work, we reduced the electropherograms to four numbers: the PRE, \( d_{Eu} \), PFI, and SM values, to develop an ILS model for predicting doping levels.

7.3.6 Preprocessing

Preprocessing plays an important role in many chemometrics analyses. For example, mean centering consists of taking the average of the values of the electropherograms at a given time and then subtracting that average from each individual value at that time. In other words, the average
electropherogram is subtracted from each electropherogram in the data set and the center of the data point cluster (individual electropherograms) is moved to the origin. This is advantageous because otherwise the first principal component (PC 1) in PCA points towards the center of the cloud of data points, i.e., it represents the average spectrum, where this direction may or may not correlate with any chemical trend in the data and PC 1 may have to be discarded. However, with mean centering, the spectral regions (points in time in the electropherograms here) that correspond to greater excursions (spreads) in the data are more heavily weighted in the analysis. Autoscaling overcomes this problem. Autoscaling consists of mean centering the data and then dividing by the corresponding standard deviations, putting the regions of the electropherograms/spectra on equal footing in the analysis. Autoscaling is generally inappropriate for data sets that contain both noisy and signal-containing regions because it gives them equal importance in the analysis. This approach is appropriate for our range-selected data (see below) because the data do not contain regions of significant noise.
7.4 Results and Discussion

The purpose of our work is to find statistical/mathematical tools that differentiate between the electropherograms from doped and undoped blood in the ABD data set in Figure 7.1. Believing it would be important to start with well-accepted tools before considering or introducing others, we first applied three well-known chemometrics methods to the data set: cluster analysis, PCA, and PLS. These traditional methods were inadequate because of the large natural variation in the electropherograms and the limited number of samples (spectra), which made variance analysis difficult. Accordingly, we pursued other possible approaches/algorithms. These included pattern recognition entropy (PRE), which we have recently used multiple times, the Euclidean distance, peak fit-integration (PFI), and the second moment (SM). These results were then combined to develop inverse least squares (ILS) calibrations.

7.4.1 Traditional Analyses: Cluster Analysis, PCA, and PLS

Three different preprocessing methods were applied to the data in the cluster analysis. In the first, a process referred to as ‘range selection’, the data were selected over the range in which they appear to contain meaningful signal(s) (from about 1.3 – 2.7 min., see Figure 7.1). Range selection is a form of scaling in which the data are multiplied by a weighting factor of 0 or 1. The range selected data points were then normalized with the 1-Norm, where this operation consists of division of each data point in an electropherogram by the sum of the data points in that electropherogram, or in the case under current study, each point in the range-selected
electropherogram was divided by the sum of the data points in the range-selected electropherogram. Finally, the data were autoscaled. Replicate runs of each sample introduce correlation into the analysis. However, given the limited sample size and the significant natural variation between the runs, smoothing over this variation by averaging the runs would result in the loss of information. Figure 7.3 shows the dendrogram produced from the cluster analysis of the preprocessed data. The results are mixed. Of the three main clusters in Figure 7.3, which are delineated by the black, vertical, dashed line, the bottom cluster contains seven mostly ‘Clean’ (0%) electropherograms with only two that are not (one 5% and one 10% sample). The middle cluster consists of an even mix of 5% and 10% samples (six of each) and one ‘Clean’ sample, and the top cluster also contains an even mix of 5% and 10% samples (two of each) plus one ‘Clean’ sample. These results suggest that cluster analysis can fairly reasonably separate doped (5% and 10%) from undoped (0% ‘Clean’) samples, but that it cannot distinguish between the two levels of doping considered in this study. However, there is an inconsistency in the clustering here that is revealed in the two-cluster model in the dendrogram (see the light blue vertical dashed line). That is, a priori, one would expect that of the three clusters suggested by the dendrogram, the two that should be most similar, and that should cluster, would contain mostly 5% and 10% samples, i.e., the upper two clusters in the three-cluster model. However, this is not the case. The cluster with the larger number of 5% and 10% samples combines with the bottom ‘Clean’ cluster in the two-cluster model. While this may suggest that the five samples in the top cluster are outliers, it is probably inappropriate to eliminate 5 of our 27 samples in this way.
Figure 7.3. Dendrogram from a cluster analysis of the 0, 5, and 10 % electropherograms under consideration in this study. The data were preprocessed using range selection followed by normalization (1-norm) and autoscaling. The dashed, vertical, light blue line indicates a two-cluster model, and the dashed, vertical, black line indicates a three-cluster model. ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0, 5, and 10 % samples, and ‘A’, ‘B’, and ‘C’ represent the three subjects. Replica runs are represented by the number following the ‘A’, ‘B’, or ‘C’.
The second preprocessing approach taken for our cluster analysis was to repeat the range selection as was done previously and then apply the 1-Norm to the data. These results are shown in Appendix 8. Two main clusters were observed. The top cluster contained 12 samples: 2, 6, and 4 of the 0, 5, and 10% samples, respectively, while the bottom cluster contained 15 samples: 7, 3, and 5 of the 0, 5, and 10% samples, respectively. It is difficult to see any distinct separation of the samples in this analysis. Finally, in a third attempt at cluster analysis, the data were range selected and autoscaled. This approach again produced two distinct clusters (see Appendix 8). The top cluster contained 13 samples: 0, 6, and 7 of the 0, 5, and 10% samples, respectively, while the bottom cluster contained 14 samples: 9, 3, and 2 of the 0, 5, and 10% samples, respectively. That is, with this preprocessing approach, the top cluster only contained doped samples (nearly equal amounts of the 5% and 10% samples), while the bottom cluster contained almost twice as many undoped samples as it did doped samples. While this preprocessing approach is arguably the best of the three methods considered herein, its ability to separate the samples into classes is still arguably weak.

PCA is one of the most commonly used multivariate analysis tools. It is an unsupervised pattern recognition technique, meaning that it requires no prior knowledge of the classes to which objects may belong. PCA has been applied to many different data types from many different types of samples. For example, in our laboratory it has been used to analyze data obtained from the analysis/characterization of alkyl monolayers on silicon, coal samples, mouse livers, nanodiamonds, and chemically treated display glass surfaces. One of the key
limitations of PCA is the large sample size required for analysis of variance and determination of correlation structure. Nevertheless, there are numerous reports containing examples of the successful application of PCA to relatively small data sets.\textsuperscript{27} We performed PCA of our range-selected, normalized, and autoscaled data. To determine the number of PCs to keep, we examined the root mean square error of cross-validation (RMSECV) and root mean square error of calibration (RMSEC) figures of merit against the number of principal components (PCs) (see Appendix 8). The RMSECV here was based on a leave-one-out cross validation. As expected, the RMSEC value decreased monotonically as the number of PCs increased, i.e., an increased number of PCs successively captured more of the variance in the data. The RMSECV value decreased by only a small amount from 1 to 9 components (from 7.513 to 7.251) with only a limited increase in the variance captured. Thus, a one-PC model would appear to be appropriate. However, the resulting scores plot from this one-PC model did not show any reasonable groupings of the samples that corresponded to their degrees of doping (see Appendix 8). A nine-PC model was then considered. It also failed to show any reasonable groupings of the data on any of the nine PCs (see Appendix 8). Hotelling $T^2$ vs. $Q$ residuals plots were then generated for the one- and nine-PC models (see Appendix 8). These plots revealed the distribution of the data both within (Hotelling $T^2$) and outside of the models ($Q$ residuals). In both cases, most of the data points lie within 95% confidence limits. However, in the one-PC model three data points fell far outside these limits, whereas in the nine-PC model, two data points fell slightly outside the limits. Accordingly, a final attempt was made to analyze the data by PCA in which the three outliers in the one-PC model were removed and the model was recreated. Based on the eigenvalues associated with each PC, a three-PC model appeared appropriate for the remaining data. Unfortunately, none of the PCs in
this model showed any reasonable groupings of the data in their scores plots. In summary, multiple attempts with PCA failed to reveal or find any of the expected trends in the electropherograms.

Despite the overall lack of success with PCA, we attempted to use PLS to create a calibration of our data set. For the x-block in this attempt, i.e., the electropherograms, the data were preprocessed by range selection, the 1-Norm, and autoscaling. For the y-block, i.e., the degree of doping, the values were mean centered. Employing leave-one-out cross validation results for this modeling (see Appendix 8), one-, seven-, or eight- components seemed appropriate (these models showed the lowest RMSEC values). Accordingly, we examined the predictions of the one- and seven-component models. (The eight-component model was not considered because its RMSECV value was essentially identical to that from the seven-component model.) The predictions of the seven-component model were of low quality (see Figure 7.4). This model appeared to be able to differentiate between the 0% and 10% doping, but the 5% samples showed strong overlap with both the 0% and 10% samples. The predictions from the one-component model were of a lower quality and were, therefore, useless. We conclude that PLS is fairly unsuccessful in creating the desired calibration between the doping levels and the corresponding electropherograms.
Figure 7.4. Seven-component, PLS predictions of doping levels from replicate runs for undoped and doped (5 and 10 %) blood samples. Here, a separate seven-component PLS model was created for each data set with one of its samples left out, and that sample was then predicted by the corresponding model.

7.4.2 Analysis by Less Traditional Tools: PRE, the Euclidean Distance, Peak Fit - Integration and the Second Moment

Because of the inability of the traditional multivariate approaches (cluster analysis, PCA, and PLS) to model the doping levels of the blood samples, we turned to less traditional
mathematical/statistical analyses. These were pattern recognition entropy (PRE), the Euclidean distance, peak fit-integration, and the second moment.

First, PRE was performed on the electropherograms under consideration in this study. Figure 7.5a shows the average PRE values with standard deviations of the three replicate electropherograms from each subject at each level of doping (see Appendix 8 for the corresponding raw data). The PRE value, which is a summary statistic, is reflective of the shape of the electropherogram, where the presence of additional peaks/shoulders in the electropherogram, which takes place for the 5% and 10% doped samples, results in higher PRE values. As a result, the PRE values gradually increase with doping – PRE is rather effective at responding to the doping levels of all the samples considered in this study.

The reordered spectrum described in Chapter 4 is a visual, intuitive tool for understanding PRE analysis; the absolute magnitude of PRE values are abstract and a graphical way of understanding it can be helpful. A reordered spectrum sorts the values of a spectrum from high to low. For example, three reordered spectra (electropherograms) of undoped, 5% doped, and 10% doped blood from Subject A are shown in Figure 7.5b. The reordered electropherogram corresponding to the undoped sample has the sharpest peak, which is consistent with its lower PRE value, while the reordered electropherogram from the 5% and 10% doped samples have higher numbers of data points with larger values, which is consistent with their higher PRE values.

Second, Figure 7.5c shows the Euclidean distances ($d_{Eu}$) between the electropherograms of the clean and 5% or 10% doped samples, i.e., the distance between the 0% and 5% and also the 0% and 10% samples was calculated for each replicate run. These distances between the electropherograms are expected to progressively increase with increasing doping levels. This is
another way of saying that the vectors corresponding to the doped and undoped electropherograms are expected to be different, and also increase as the degree of doping increases. It is clear from the results in Figure 7.5c that \( d_{Eu} \) always shows a difference between the undoped and doped samples. Furthermore, while, on average, the \( d_{Eu} \) values for the 10% doped samples are greater than those for the 5% samples, there is enough overlap between these results that it would be difficult to differentiate between these two states with this method.

Third, Figure 7.5d shows the ‘Peak Fit - Integration’ (PFI) results obtained by measuring the areas of the shoulders in the electropherograms to the right of the main peaks of the doped and undoped samples. Two things are clear here. First, there is some scatter in the results. Second, the area of the shoulder consistently increases with doping level. The average and standard deviation for each set of measurements are 0.05 ± 0.01, 0.07 ± 0.02, and 0.11 ± 0.04 for the undoped, 5 % doped, and 10 % doped samples, respectively.

The second moment of the electropherograms was calculated in two different ways. In the first case, complete electropherograms were used for SM calculations. This method failed at providing any meaningful difference in the doping levels. In the second, range-selected data were used (Figure 7.5e), and the data were preprocessed by normalization (1-Norm), followed by autoscaling. As was the case with some of the other less traditional methods we employed, this approach showed promise in separating the undoped (0%) and doped (5 and 10%) samples, but not in differentiating between different levels of doping.
Figure 7.5. Results of the less traditional mathematical/informatics methods used to analyze the doping data. Figure 5a (top left): The average PRE values (heights of bars) with standard deviations (error bars) of the electropherograms of subjects A, B and C for 0, 5, and 10 % doping levels. Figure 5b (top right): The reordered electropherograms from replicate run 1 of subject A at 0, 5, and 10 % doping levels. Figure 5c (middle left): The Euclidean distances between electrophoretic separations of clean and doped (5% and 10%) samples of subjects A, B and C with three replicate runs for each. Figure 5d (middle right): The absolute areas of the broad features (shoulders) to the right of the main peaks from electrophoretic separations of Subjects A, B and C for 0, 5, and 10 % doping levels. Figure 5e: The second moments of the range-selected electropherograms.

7.4.3 Inverse Least Squares

Inverse least squares (ILS) is an important method for generating calibrations. In general, ILS uses a relatively small number of variables to create calibrations. For example, an ILS regression can be based on principal component regression (PCR), which uses the PCA of a data set to reduce the number of variables in the data set. Here, we chose to construct ILS calibrations using the PRE, $d_{Eu}$, PFI, and SM summary statistics from the electropherograms. Each of these had demonstrated some ability to differentiate between the samples based on their doping levels. Accordingly, a combination of these metrics would lead to a calibration with greater predictive ability. The coefficient of determination ($R^2$) was used as the figure of merit. $R^2$ is defined as the square of the correlation coefficient $r$ (Equation 7.11) and and is a measure of the percentage
variation in one variable as explained by another variable.29 (We include the formula for ‘R’ here because its definition varies in the scientific literature.)

\[
R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}
\]

(7.11)

ILS models were built based on all possible combinations of the four summary statistics. All of the ILS models based on any two of our summary statistics showed very strong overlap in the predictions of all the three doping levels (0, 5 and 10%), i.e., these models were useless. Accordingly, we considered ILS models based on all possible combinations of three summary statistics (see Figure 7.6). First, Figure 7.6a shows the leave-one-out predictions using the PRE, \(d_{Eu}\) and PFI summary statistics. It can easily differentiate between undoped (0%) and doped (5% or 10%) electropherograms, but not between the two levels of doping. This three summary statistic ILS model was the most successful of the four we considered (\(R^2\) value of 0.994). It seems reasonable that it is based on PRE and \(d_{Eu}\) because these summary statistics appeared to be the most successful in differentiating between the doping levels (see Figure 7.5). Figure 7.6b shows the ILS predictions using PRE, \(d_{Eu}\) and SM. It was less successful as a model (\(R^2\) value of 0.938), which may be explained by SMs greater struggle to differentiate between the samples (see Figure 7.5e). The ILS model based on PRE, PFI and SM in Figure 7.6c was quite poor (it had an \(R^2\) value of 0.574), and the ILS model created using \(d_{Eu}\), PFI and SM, which is not shown, was even worse (it had an \(R^2\) value of 0.328). It is evident from these results that PRE and \(d_{Eu}\) made the largest contributions to the prediction capabilities of the ILS models. As a final attempt, an ILS model based on all four summary statistics was created (see Figure 7.6d). It gave the highest \(R^2\) value (0.997), and like the PRE, \(d_{Eu}\) and PFI-based ILS model, it can clearly differentiate between 0%
(undoped) and 5% or 10% levels of doping. Doping below 5% has little physiological effect. Accordingly, the ILS model shows high accuracy in its ability to differentiate between undoped and ‘meaningfully’ doped blood (see Figures 7.6a and 7.6d).

![Figure 7.6](image)

**Figure 7.6.** Leave-one-out predictions from ILS models based on (a) PRE, $d_{Eu}$, and PFI, (b) PRE, $d_{Eu}$, and SM, (c) PRE, PFI, and SM, (d) PRE, $d_{Eu}$, PFI, and SM summary statistics from replicate runs of A, B and C for undoped (0 %) and doped (5 % and 10 %) blood samples. The $R^2$ values and data were compared against straight lines $y = x$ (red lines).
7.5 Conclusions

The detection of autologous blood doping is critical for banning unscrupulous practices used by athletes to gain an unfair advantage in competition. In this work, we have demonstrated several mathematical techniques that distinguished between doped and undoped blood samples. Capillary electrophoresis was used to separate fresh and stored RBCs, serving as a viable alternative to the widely used and more expensive method of monitoring an athlete’s biological passport. In Harrison’s original work on this topic, a first derivative analysis of slopes was used to detect the presence of doping. However, this method suffered from false negatives, lacking a strong ability to precisely identify doping. In our work, conventional informatics techniques (cluster analysis, PCA, and PLS) had very limited success in distinguishing between electropherograms of samples with different levels of doping. Several preprocessing methods were considered in these analyses. Variance analysis (PCA and PLS) was challenging due to the large natural variation in electropherograms from replicate runs. Four less commonly used summary statistics (PRE, the Euclidean Distance, Peak Fit/Integration, and the Second Moment) were applied to the data. An ILS calibration based on these inputs allowed easy differentiation between undoped and doped samples, and to some degree between the different levels of doping (5 and 10%). We understand that natural variation can exist in the RBCs of athletes due to biological sex, ethnicity, muscle/fat percentage, diet, age, etc. In our study of the three athletes, we observed some minor differences in absolute migration times of the RBCs, but no significant differences between the individuals. In our (Harrison’s) broader studies in this area, he only sees minor differences in absolute migration times of RBCs – he has yet to see any significant differences between individuals. Thus,
the changes induced by transfused cells appears to be a significant, measurably change to the cell population.

**7.6 Acknowledgments**

C. Harrison wishes to offer sincere thanks to the volunteers who participated in this study. Support for the anti-doping study was provided by the World Anti-Doping Agency (reference number 09A23CH).

**7.7 Supporting Information**

Supporting Information is available in Appendix 8.

**7.8 References**


Chapter 8: Conclusion of My Present Work

There is a growing need for more advanced data analysis techniques. As explained in Chapter 1 of this dissertation, the instruments used in the field of analytical chemistry are becoming increasingly complex and increasingly able to generate large sets of data. Traditional data analysis, i.e. monitoring changes in a set of signals over time, is no longer a viable approach, given the vastness of the functional information and time that are sacrificed. The solution to this problem is the mathematical techniques known as summary statistics. Summary statistics are measurements that can summarize the chemical information into a single, useful number.

In Dr. Linford’s lab, we have been applying multivariate statistical analysis techniques like principle component analysis, multivariate curve resolution, cluster analysis, etc. for a wide variety of datasets. In doing so, we have realized that each of these techniques, though extremely capable, versatile, and effective in the extraction of large amounts of chemical information, are problematically complex for the unexperienced user. For example, to understand how PCA works, one must first have a thorough understanding of which preprocessing techniques are appropriate for any given situation to produce the desired information. Further, the use of these multivariate analysis techniques is a complex affair that requires a deep understanding of the theory and intricacies of each algorithm. Often, the results generate by these techniques are difficult to interpret. For example, PCA can generate negative loadings which can be challenging to understand.

There is a need for a simpler data analysis methods that are easier and more direct in their application. In this work, I have introduced the Pattern Recognition Entropy (PRE), a mathematical
algorithm created by adapting Shannon’s Information Theory. As explained in the previous chapters of this dissertation, Claude Shannon’s paper in 1948 uniquely quantified information. Although real spectra cannot be defined by probabilities as in Shannon’s *Entropy*, we described them with pseudo-probabilities for calculation of PRE values. We observed that the PRE values were representative of the shape of the spectrum and could be used to uniquely quantify the complexity of the spectrum, thus serving as a summary statistic.

In Chapter 1, we identified key transitions in several materials through depth profile analysis of data acquired by time-of-flight secondary ion mass spectrometry and X-ray photoelectron spectroscopy. We showed that although PCA and MCR are useful in understanding a thin film and the interactions between the thin film and the substrate, PRE provides the same information faster and more directly. As stated previously, the depth of the information that can be obtained from PCA and MCR is much greater, but PRE demonstrated remarkable usefulness in describing where the interfaces existed, the homogeneity of the thin films, etc.

In Chapter 2, the concept of reordered spectra is presented as a solution for one of the shortcomings of PRE. While PRE distributions are useful in telling a narrative, they can be difficult to interpret, because they are not absolute. Reordered spectra are a visual tool that provides a more intuitive interpretation of PRE values.

Given that PRE can uniquely quantify the complexity of a spectrum and that noisy spectra contain more shapes and complexity, PRE was applied for reducing noise in Total Ion Current Chromatograms generated by Liquid Chromatography-Mass Spectrometry. We calculated PRE values for each mass chromatogram in the given data set. Mass chromatograms with low signal-
to-noise corresponded to low PRE values and were combined to produce a greatly improved and reduced TICC.

In Chapter 6, the follow up work addressing some of the underlying problems of our work in Chapter 5 is presented. Although, PRE was very effective in sorting the high quality mass chromatograms from the low quality ones, a problem persisted: the noise level of the mass chromatograms was not constant throughout the entire mass range (m/z value) of the mass chromatograms channels. As a consequence, a single threshold value for the classification of good and bad mass chromatograms was ineffective. Subsequently, we applied a second algorithm in conjunction with PRE called cross correlation. In tandem, these two algorithms yielded a higher degree of reduction and an improved TICC.

Finally, in our collaboration with Prof. Christopher R. Harrison of San Diego State University, we addressed the problem of autologous blood doping. Autologous blood doping is a major issue in professional endurance sports, because it is difficult to detect. Dr. Harrison invented a method of blood doping detection using capillary electrophoresis and we provided the informatics techniques to analyze the electropherograms and an inverse least square calibration model to confirm the presence of doped blood.
8.1 Future Work

My future work will entail the full exploration of the application of Pattern Recognition Entropy in other fields. For example, PRE can be used for the automation of signal collection. Our goal will be to automate the process, so that the user will not be required to decide the optimal number of scans. The optimization of this process and thereby using the appropriate number of scans saves time and other resources. During the data acquisition process, signals with high noise are common among the initial scans, but over time, the noise level decreases because of signal averaging. We predict that PRE will be useful in quantifying this change in noise level and accordingly suggesting the optimum number of scans.

Additionally, we will apply PRE to hyperspectral images. These images are formed by collecting a spectrum at each pixel. The data sets of hyperspectral images are extremely large, i.e. on the scale of tens of millions to billions of spectra per image. Our goal will be to use PRE as a summary statistic to reduce the information at every pixel by reducing each spectra to a unique number. This will be followed by an analysis to obtain the necessary chemical information. We expect the application of PRE will result more computationally efficient data processing.
Appendix 1: Polyallylamine as an Adhesion Promoter for SU-8 Photoresist

A1.1 Introduction

Resist lithography consists of transferring patterns onto a substrate using an etching process via an irradiation source and a photosensitive polymer 1-3. Here, the solubility of the polymer is either selectively increased or decreased in exposed areas, yielding a positive or negative image. As more advanced techniques and applications for resist lithography have developed, more focus has been given to three-dimensional structures in High Aspect Ratio (HAR) lithography. Combined with low cost replication methods, HAR enables fabrication of microstructures otherwise unobtainable, such as advanced scaffolds for tissue engineering 4-6, miniaturized drug delivery systems 7, and bioinspired surfaces with specific adhesion and haptic properties.

SU-8 (Figure A1.1) has found widespread use in the microelectronics industry 8, as well as in replica molding, as a mask for metallization on organic substrates, and, due to its transparency, as a structural material for optics 9. SU-8 is a negative photoresist, and favored for HAR lithography, due to its high chemical and mechanical stability, wide variety of thicknesses attainable, and biocompatibility 10-12. SU-8 epoxy resin is developed at near-UV wavelengths where the photoresist has very low optical absorption, usually from 365 – 436 nm, which makes HAR photolithography of thick films possible. SU-8 is composed of a Bisphenol A Novolac epoxy
oligomer, up to 10 wt. % photoacid generator, which includes a triarylsulfonium hexafluoroantimonate salt, and an organic solvent. The acid-labile groups in SU-8 produce a photoacid during exposure. Accordingly, irradiation generates a low concentration of a strong acid that catalyzes cross-linking. A post-exposure bake reheats the polymer and regenerates the acid catalyst, inducing further cross-linking. During exposure to radiation, the photoacid generator decomposes into hexafluoroantimonic acid, which protonates the epoxides on the SU-8 oligomer. After heat is applied, the resulting protonated oxonium ions react with neutral epoxides in a series of cross-linking reactions. Because each monomer contains eight reactive glycidyl groups (Fig. 1), extensive cross-linking among the monomers is possible. Thus, two different variables control the degree of cross-linking in the final pattern: exposure dose and thermal treatment.

Figure A1.1. Structure of the average oligomer (four repeat units) in the negative photoresist SU-8.
Unfortunately, SU-8 often shows poor adhesion to hydrophilic inorganic substrates \(^{14-16}\). SU-8 is a hydrophobic organic material (note the methyl and phenyl groups in Figure A1.1), so good wetting of the substrate is necessary to obtain homogeneous and stable coatings. Possible solutions here include the use of low-viscosity SU-8 \(^{17}\), surface modification with a low molecular weight adsorbate like hexamethyldisilazane (HMDS) \(^{18}\), or a commercial adhesion promotion reagent (Omnicoat from Microchem Inc). These commercial primers improve wetting of the inorganic substrate prior to application of SU-8 by either strongly absorbing to it or by reacting to form a thin organic layer with low surface energy. It is crucial to use an adhesion promoter in most applications of SU-8.

HMDS is the most common type of silane used for SU-8 adhesion promotion to oxide surfaces, which include silicon dioxide (SiO\(_2\)) and silicon nitride (Si\(_3\)N\(_4\)) (silicon nitride surfaces become silicon dioxide when exposed to the environment) \(^{19}\). HMDS has the formula (CH\(_3\))\(_3\)SiNHSi(CH\(_3\))\(_3\). Its Si-N groups react readily with polar hydroxyl groups on oxide surfaces, which ultimately tethers trimethylsilyl groups to them. This changes the wetting properties of the oxide surface from hydrophilic to hydrophobic, allowing it to interact more favorably with the hydrophobic SU-8. Nevertheless, these common methods have their drawbacks. HMDS and other commercial reagents require extensive surface dehydration and/or curing because layers of water at the surface of an oxide can react with HMDS and prevent it from creating a permanent hydrophobic surface. Proper application of HMDS is also required to form effective bonds. Indeed, due to its high reactivity, HMDS is often applied using chemical vapor deposition (CVD). Here, after dehydration in a heated vacuum oven, the oxidized substrate is exposed to gaseous HMDS to
deposit a monolayer of -Si(CH$_3$)$_3$ on its surface. Spin-coating is less preferred due to the thicker nature of the coatings, which makes HMDS less effective as an adhesion promoter. Furthermore, if a thick layer of HMDS remains, its vapors will diffuse into resist films that subsequently coat the surface. This vapor may partially crosslink the resist during soft-bake and cause changes in its development rate and even degrade the attainable resolution.

Polyallylamine (Figure A1.2) is a water-soluble polymer that is rich in primary amines. We have studied its deposition in a number of different contexts and on different substrates$^{20-23}$. The silanol groups on piranha solution or plasma cleaned oxide surfaces are acidic and the primary amines in PAAm are basic. Thus, PAAm spontaneously adsorbs from solution onto SiO$_2$ surfaces in a self-limiting fashion$^{24}$. There are similar reports in the literature of the spontaneous adsorption of other amine-containing polymers to oxide surfaces, including polyethylenimine and polylysine$^{25}$. These types of initial layers are often employed as the starting point for polyelectrolyte multilayers$^{26-29}$. 
**Figure A1.2.** Structure of the water-soluble adhesion promoter polyallylamine (PAAm).

Aliphatic polyamines are frequently used as hardeners for epoxy resins. Here we demonstrate PAAm as a novel, easy-to-apply adhesion promoter for SU-8. Part of our motivation for this study is the well-known reaction between amines and epoxides (see Figure A1.3)\(^{30}\), which is one of the most useful reactions in bioconjugate chemistry. Here, conditions for the use of PAAm as an adhesion promoter for SU-8 are explored and the resulting materials are characterized by X-ray photoelectron spectroscopy (XPS)\(^{31}\), spectroscopic ellipsometry (SE), and wetting. Both a simple and more complicated SE models are considered.
Figure A1.3. Possible reaction of the primary amines in PAAm with the epoxy groups in SU-8.

A1.2 Materials and Methods

A1.2.1 Materials

PAAm (20 wt. % in water, $M_w \sim 17,000$) was obtained from Sigma-Aldrich, St. Louis, MO. SU-8 2025; the photoresist, and the SU-8 developer solution were obtained from Microchem Corp., Westborough, MA. These materials were used as received. All substrates employed in this study were native oxide-terminated silicon. Some of these surfaces were cleaned by immersion in freshly prepared piranha solution (3:1 conc. $\text{H}_2\text{SO}_4$: 30% $\text{H}_2\text{O}_2$) for 30 min at 90 °C. Surface
cleaning was also performed with an air plasma in a PDC-32G Plasma Cleaner from Harrick Plasma, Ithaca, NY. All PAAm depositions were performed with 0.1 wt. % solutions of PAAm in water and all immersion times were for 2100 s (35 min). After immersion, surfaces were removed and washed thoroughly with high purity water, and then dried with nitrogen. SU-8 2025 was spun at a final speed of 4000 rpm for 30 s, which yielded a 25 μm thick layer. The wafer was kept on a horizontal surface for 5 min to relax the film, baked at 65 °C for 3 min, and then immediately baked again at 95 °C for 5 min. The wafers were finally allowed to cool at room temperature for 10 min. SU-8 coated wafers were not exposed to light, which allowed us to investigate surface adhesion properties in the absence of polymer crosslinking. The film was developed (immersed) in the SU-8 developer solution for ca. 4 min, rinsed with isopropanol and dried. An overview of this process is presented in Figure A1.4.

![Figure A1.4. Overview of the coating/fabrication process.](image-url)
A1.2.2 Surface Characterization

XPS was performed with an SSX-100 instrument (serviced by Service Physics, Bend, OR) with a monochromatic Al Kα source and a hemispherical analyzer. Survey scans were recorded with a spot size of 800 µm x 800 µm and a resolution of 4 (nominal pass energy of 150 eV). Sessile water contact angles (WCA) were measured with a Ramé-Hart (Netcong, NJ) Contact Angle Goniometer (Model 100-00) fitted with a manual syringe that was filled with high purity (18 MΩ-cm) water. The drop sizes for measuring static water contact angles were ca. 10 µL. Film thicknesses on planar substrates were measured at an incident angle of 75° with an M-2000 spectroscopic ellipsometer (J.A. Woollam, Co. Lincoln, NE), and the data were analyzed using the instrument software (CompleteEASE®) over a wavelength range of ca. 190 – 1700 nm. The optical constants for SiO₂ and Si in the instrument software were used for this modeling.

A1.3 Results and Discussion

A1.3.1 SU-8 Deposition and Removal on PAAm-Coated Surfaces

Figure A1.5 shows the change in film thickness, as measured by SE, which accompanies PAAm deposition onto native oxide-coated silicon that was cleaned with piranha solution. The as-received native oxide layer, ‘Before Piranha’, which would most likely include a small amount of surface contamination, is ca. 2 nm thick. After cleaning, ‘After Piranha’, a small decrease in
thickness is observed, which presumably corresponds to the removal of adventitious hydrocarbons. PAAm is then deposited. It is not expected to go down as a thick layer – for electrostatic reasons the adsorption of polyelectrolytes from water is generally self-limiting. Thus, a small increase in total layer thickness was observed after PAAm deposition. Within 5 – 10 min after PAAm deposition, samples were spin coated with SU-8. Then, after 10 min at room temperature, the SU-8 was developed, dried with nitrogen, and analyzed by SE. Fig. 5 clearly shows that the amount of residual SU-8 on the surface is much higher when PAAm is present. These results points to a favorable interaction between these materials – they are consistent with covalent bonding or significant hydrogen bonding between the materials. Note here that to better probe the PAAm-SU-8 interactions, the surfaces here were underdeveloped. With these shorter development/rinse conditions, there was a visible haze on these surfaces that we attribute to residual SU-8. At longer/full development times, it appears that the SU-8 can be almost entirely stripped from the surface (vide infra).

Figure A1.6 shows results from air plasma cleaned, in contrast to piranha cleaned, PAAm-coated surfaces that were kept in contact with SU-8 for 10, 20, 30, or 40 minutes. In order to probe the interaction of the PAAm with SU-8, these surfaces were either heated to 65 °C or maintained at room temperature. They were then underdeveloped. At the two shorter residence times, 10 and 20 min, both the heated and unheated samples show quite a bit of residual polymer, with the heated samples having more than the unheated. There is a noticeable haze on these surfaces. However, at longer times, 30 and 40 min, the amount of residual SU-8 drops to very low levels – only a few nanometers remain, and there is no longer any visible haze. These results bring up the interesting
possibility that a chemical reaction is not taking place to a significant extent between the SU-8 oligomer and the PAAm. If the reaction depicted in Figure A1.3 were occurring, it would not be reversible, and one might expect more residue on the surfaces with time, not less. However, a different chemical interaction may be taking place between the PAAm and the photoresist. The SU-8 contains rather high concentrations of photoinitiator, which in turn contain triarylsulfonium salts. As indicated in the literature, these salts are expected to be effective electrophiles (arylating reagents), and a possible reaction between them and a primary amine is shown in Figure A1.7 – note that the leaving group here is a neutral molecule, which is a favorable scenario. If this chemistry is operative, it would result in arylated, protonated amines. These would not be reactive with the oxirane rings in SU-8 – protonation/quaternation of an amine destroys its nucleophilicity. Accordingly, different, weaker interactions might be expected between the PAAm and SU-8, including hydrogen bonds and London dispersion interactions. The decrease in residual SU-8 at the 30 and 40 min exposure times might then be attributed to a rearrangement of the PAAm. Rearrangements of amine-containing films on SiO₂ are known. Here, with time, the PAAm may rearrange so that its primary amines can interact with surface silanol groups through acid-base interactions. This process would expose the polyethylene-like backbone of PAAm, creating a hydrophobic surface. The resulting dispersion forces holding these materials together would be weaker than the hydrogen bonds that may have been initially present. All of this may explain why the SU-8 is more effectively removed from the surface with time. Obviously the 10 min results from Figure A1.5 are consistent with the short residence time (10 and 20 min) results in Figure A1.6.
Figure A1.6. Ellipsometric measurements of plasma-cleaned Si/SiO₂, PAAm-coated Si/SiO₂, and PAAm-coated Si/SiO₂ exposed to SU-8 photoresist for 10, 20, 30, or 40 min prior to underdevelopment.

Figure A1.7. Possible reaction of PAAm with the triarylsulphonium salt in SU-8.
Figure A1.8 shows the water contact angles that correspond to the surfaces/materials in Figure A1.6. As expected, the contact angle of the plasma-cleaned surface is very low, that of the PAAm-coated surface is a little higher, and the water contact angles of the SU-8 containing surfaces are quite a bit higher (we noted above the hydrophobicity of the aliphatic and aromatic components of SU-8). While there is some decrease in the contact angle with exposure time, these results suggest that, fundamentally, the surfaces remain coated with at least a thin film of SU-8.

**Figure A1.8.** Water contact angles of plasma-cleaned Si/SiO2, PAAm-coated Si/SiO2, and PAAm-coated Si/SiO2 exposed to SU-8 photoresist for 10, 20, 30, or 40 min prior to underdevelopment.
XPS N 1s narrow scans were used to confirm the PAAm deposition, SU-8 coating, and SU-8 removal. For example, Figure A1.9a shows that, as expected, there is no nitrogen on the plasma cleaned surface, but that noticeable quantities of it are present after PAAm deposition (Figure A1.9b). Here the nitrogen appears to be in two oxidation states, which presumably correspond to protonated and deprotonated amines. There is then no nitrogen observed on the 10 and 20 min surfaces (see above and Figures A1.9c – d), some nitrogen on the 30 min surface, and no nitrogen on the 40 min surface (Figures A1.9e – f). These latter two results appear to depend on where the X-ray spot was focused on the surface. Obviously, the presence of nitrogen on the 30 min surface suggests rather effective removal of the SU-8 from this surface.

**Figure A1.9.** XPS Nitrogen 1s narrow scans of air plasma cleaned surfaces.
A1.3.2 Reduction in Defects During Spin Coating of SU-8

While this is the shortest section in this paper, it is also arguably the most important. Figures 10a – b show the SU-8 surface after spin coating onto both a bare silicon wafer and a PAAm-coated wafer. Spin coating of SU-8 onto bare silicon surfaces always leads to defective films. These defects would be highly detrimental to HAR lithography and multi-layer structures. The type and shape of these defects differed between coatings. No such defects were observed on SU-8 films on the PAAm-coated surfaces.
Figure A1.10. Surfaces spin coated with SU-8 that were (a) coated with PAAm and that show no defects, and (b) uncoated and that show defects.

A1.3.3 Refinement of the SE Model

As shown above, spectroscopic ellipsometry played an important role in characterizing the thin films prepared in this study. However, while the trends and even values obtained by SE in these measurements seemed reasonable, the errors in these measurements, i.e., mean squared errors (MSE), tended to be quite high for the thicker films. In this section we describe attempts to improve our SE modeling of these materials. Figure A1.11 depicts the three models that were considered. Model 1 was the most simple, and it was used to obtain the results reported to this point in this work. It consisted of an SiO$_2$ layer of variable thickness on a silicon substrate. It is well known that the optical constants of thin films have less and less of an effect on ellipsometric film thickness as films becomes thinner $^{34}$. Thus, the optical constants of SiO$_2$ in Model 1 should very effectively model the 1 – 2 nm films created in this study. Models 2 and 3 seemed to more effectively describe the thicker films of residual SU-8 that were prepared. Both of these models contain a Bruggeman Effective Medium Approximation (BEMA) layer. Here, this layer models heterogeneous/porous films of SU-8 using a 50:50 mixture of the optical constants of SiO$_2$ and void. Models 2 and 3 also contain an SiO$_2$ layer that accounts for the native oxide on silicon. Interestingly, when both the BEMA and native oxide film thicknesses are allowed to vary (Model 2), the SiO$_2$ thickness drops below its known value. This model also shows inconsistencies, failing to converge to the same
values when it is repeatedly run. These unphysical results caused us to fix the SiO\textsubscript{2} layer thickness at its measured value of 1.5 nm in Model 3. This model showed both good convergence and MSE values that were consistently lower than those from Model 1. Accordingly, the SE results reported below were obtained with either Model 1 or 3.

**Figure A1.11.** Description of the Spectroscopic Ellipsometry (SE) Models.

Table A1.1 shows the SE results from Model 1 that were reported above for the PAAm-coated, plasma-cleaned surfaces that were exposed to SU-8 for 10, 20, 30, or 40 min prior to development. The MSE values for these measurements, which are also given in the table, are relatively high. The data sets from these samples were reevaluated using Model 3. This approach yielded thicknesses that were approximately double those obtained with Model 1, along with
consistently lower MSE values. It seems reasonable that these thicknesses would be higher by about a factor of two because of the 50% void contribution in the BEMA layer. Perhaps Model 1 can now be seen as giving the thickness of the fully dense, flat film that would exist if all the material in the real films were compacted into a single layer. And while some of the MSE values for Model 3 are still fairly high, the fact that there is a consistent lowering of the MSE values for this model suggests that it is a better approximation of the actual film structures. BEMA layers are often used in SE modeling to account for surface roughness, and the visible scattering (haze) from some of the SU-8 films produced in this study suggested that there was a considerable amount of it. Note that no attempt was made to refit the native oxide and PAAm-on-native-oxide results using Model 3. Model 1 gave low MSE values for these measurements (ca. 1 – 3), and was thus deemed to be sufficient for this purpose. Presumably even better SE results would be obtained from Model 3 if its BEMA layer were based on a 50:50 mixture of the optical constants of SU-8 and void. That is, because of its aromatic rings, the SU-8 is expected to absorb UV light and not be transparent in this wavelength range like the SiO$_2$.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th></th>
<th>Model 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thickness (nm)</td>
<td>MSE</td>
<td>Thickness (nm)</td>
<td>MSE</td>
</tr>
<tr>
<td>10 min (with Heat)</td>
<td>61.1</td>
<td>8.9</td>
<td>128.3</td>
<td>5.4</td>
</tr>
<tr>
<td>10 min (without Heat)</td>
<td>25.9</td>
<td>18.3</td>
<td>54.9</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>Thickness</td>
<td>4.4</td>
<td>108.4</td>
<td>1.7</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>20 min (with Heat)</td>
<td>50.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 min (without Heat)</td>
<td>18.4</td>
<td>9.3</td>
<td>40.8</td>
<td>6.3</td>
</tr>
<tr>
<td>30 min (with Heat)</td>
<td>6.3</td>
<td>3.9</td>
<td>13.8</td>
<td>2.8</td>
</tr>
<tr>
<td>30 min (without Heat)</td>
<td>5.6</td>
<td>2.7</td>
<td>12.7</td>
<td>1.4</td>
</tr>
<tr>
<td>40 min (with Heat)</td>
<td>5.9</td>
<td>6.9</td>
<td>13.1</td>
<td>3.9</td>
</tr>
<tr>
<td>40 min (without Heat)</td>
<td>5.0</td>
<td>4.5</td>
<td>11.9</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table A1.1. Ellipsometric film thicknesses and accompanying MSE values for residual SU-8 on PAAm surfaces after development. Two different ellipsometric models (see Figure 11) were applied to the data. Note that the reported thickness values for Model 1 also contain the native oxide thickness of 1.5 nm.

A1.4 Conclusion

PAAm promotes SU-8 adhesion, giving high quality and smooth coatings of the photoresist. The advantage of PAAm over other adhesion promoters is the simplification of the application process. Pre-cleaning techniques for the silicon surfaces included piranha solution and air plasma. Both the processes showed similar results vis-à-vis PAAm adhesion. The interactions between the PAAm and the SU-8 may include amino-epoxy ring opening reactions, hydrogen bonding, and dispersion forces, where evidence points to an interaction between the primary
amines in PAAm and the SU-8 photoinitiator (triarylsulfonium salts). A BEMA roughness layer improves the quality of the SE modeling. In future studies, solutions of PAAm could be spin coated onto silicon wafers.

A1.5 Acknowledgments

The authors would like to acknowledge Moxtek.Inc for funding and Brigham Young University (BYU) for providing research facilities.

A1.5 References

Appendix 2: Trends in Advanced XPS Instrumentation. 2. Angle-Resolved XPS (AR-XPS) and XPS Surface Mapping

A2.1 Discussion

In this article, we discuss two key aspects of advanced X-ray photoelectron spectroscopy (XPS) instrumentation. In particular, we focus here on the capabilities of the Thermo Scientific Theta Probe XPS spectrometer. As we noted in our last article,\textsuperscript{1} we recently visited Thermo Fisher Scientific as part of a whirlwind trip through England to visit two of the top manufacturers of XPS equipment in the world: Thermo Fisher in East Grinstead and Kratos in Manchester.

In a typical XPS measurement, the angle between the sample surface and the entrance to the X-ray spectrometer, i.e., the take-off angle, is fixed. Figure A2.1 shows an electron exiting a sample at two different take-off angles, $\theta$. In Figure A2.1a the take-off angle is 90°, and in Figure A2.1b it is less than 90°. Of course fixed take-off angles are widely used and extremely important – most XPS measurements are done under such conditions. However, many spectrometers have the capability of moving the components of the instrument such that the take-off angle in a measurement can be varied. This generally involves moving the sample – it’s a lot easier to move a sample than a spectrometer. To understand why this might be important, we recall that the X-rays that illuminate a sample in XPS can penetrate fairly deeply into the material – on the order of a micron. And, of course, photoemission can and will take place throughout this entire depth.
However, the photoelectrons that are thus generated are not nearly as good at traveling through the solid as the X-rays that excited them. These electrons have mean free paths of a few nanometers, which means that they can effectively travel only 5 – 10 nm before they undergo an inelastic collision, i.e., lose energy through a scattering process, which means they either remain trapped in the solid, or escape from it with reduced kinetic energy and contribute to the background relative to peaks in the spectra. Accordingly, if the take-off angle is at its maximum value (90°, see Figure A2.1), the electrons will have left the sample traveling perpendicular to it, and we will be sampling about three mean free paths directly into the material, i.e., $\Lambda_e$. This is the greatest depth we can probe into the sample in this experiment. However, if the take-off angle is between 0° and 90°, we can see from Figure A2.1b that the electrons will now be originating from a shallower depth in the material. And, of course, the smaller we make $\theta$, the less depth we probe into the material. Thus, angle-resolved XPS (AR-XPS), i.e., probing a sample at different take-off angles, provides a mechanism for determining the depth of a particular species in the near surface region of a material. Of course this can be very important in fields such as tribology, electrochemistry, catalysis, and sensing. Finally, note that in spite of its significantly improved surface sensitivity at low take-off angles, AR-XPS is not uniquely/solely sensitive to the outermost monolayer of a material, still probing about 1 nm into the surface at low angles. If one needs this type of information, one should probably consider low energy ion scattering (LEIS).\(^2\)
Figure A2.1. Illustration of angle resolved XPS. (a) A photoelectron exiting a solid at its maximum take-off angle (90°), with accompanying depth probed in the material of $\Lambda_e$. (b) A photoelectron exiting a solid at a take-off angle less than 90°, with accompanying depth probed in the material of $\Lambda_e \sin \theta$. Figure adapted from one in the paper cited here.$^3$

Probably a good way to begin to think about AR-XPS is through relative depth plots. In these plots, one takes the logarithm of the ratio of the peak area for a given species at near grazing emission angle ($\theta$ very small) to that of the peak at near normal emission ($\theta \sim 90°$). Clearly, for an
extremely thin film at a surface, e.g., perhaps a little adventitious (environmental) carbon contamination, the signals from both measurements will be very similar. Accordingly, if these signals are nearly equal, their ratio will be approximately unity, and the resulting logarithm will be about zero. In contrast, if we have a buried layer, its signal at grazing incidence will be very small, while that from the measurement at near-normal emission will be much larger. This ratio will be less than one, and perhaps much less than one, so the resulting logarithm will be a negative value. Thus, the relative depth plot will conveniently show the material that is closer to the surface as higher up in the plot. And while the resulting plots are not quantitative, they do indicate the elements that are nearest the outermost surface of a material versus those that are embedded more deeply below it. As noted by Thermo Fisher: “The relative depth plot has the advantage of being independent of any model and does not require the knowledge of the physical constants for the material. It can show, for example, the change in position of a species due to some form of treatment, such as annealing.”\textsuperscript{4,5}

Figure A2.2 shows a relative depth plot for a film of HfO\textsubscript{2} and Al\textsubscript{2}O\textsubscript{3} on SiO\textsubscript{2} on Si. Here we see the expected adventitious carbon at the top of the film. Below this we find aluminum, hafnium, and oxygen, where the lower binding energy oxygen (‘Low BE’) corresponds to the oxygen bonded to hafnium. That is, there is oxygen bonded to hafnium, aluminum, and silicon in this material. Hafnium has a lower electronegativity than silicon and aluminum, so the oxygen in HfO\textsubscript{2} is expected to exist in a more ionic form than in SiO\textsubscript{2} or Al\textsubscript{2}O\textsubscript{3}. Thus, the oxygen in HfO\textsubscript{2} should have a greater negative charge, which should reduce its binding energy, i.e., it is easier for an electron to exit a more negatively charged atom. Still lower in the plot we see oxygen (‘High BE’) that is bonded to silicon and/or aluminum, oxidized silicon, and finally bulk silicon.
Figure A2.2. Relative depth plot for a mixed film of HfO₂ and Al₂O₃ on SiO₂ on Si. Figure and caption used/adapted with permission from Thermo Fisher.

For more quantitative results, AR-XPS data can be processed by an algorithm in the Thermo Fisher Avantage software to yield the thicknesses of the layers at a surface. Up to three surface layers can be determined in this way. For example, Figure A2.3 shows results from the atomic layer deposition (ALD) growth of Al₂O₃ onto SiO₂ on a silicon substrate. In this deposition, the thickness of the SiO₂ film, which is part of the substrate, stays essentially constant, as it should, while the Al₂O₃ film thickness increases with increasing numbers of ALD cycles. This is also expected. Interestingly, the adventitious carbon layer is thinner on the uncoated SiO₂, but then increases in thickness to an approximately constant level during the first few ALD cycles. An additional application of this quantitative capability is the determination/measurement of film deposition/film thicknesses across a wafer. Figure A2.4 shows the thicknesses of AlHfₓOᵧ (mixture
of Al₂O₃ and HfO₂) and SiO₂ films from a line scan across the diameter of a 200 mm wafer. While the uniformity of these films on this sample are otherwise quite good, there is nevertheless some variation detectable in them that could be significant depending on the application.

These powerful capabilities can be extended to map the surfaces of entire wafers. Results from these types of studies will be of obvious value to process engineers. For example, Figure A2.5 shows a chemical state map of the oxygen 1s signal from an AlHfₓOᵧ film on SiO₂. The lower map shows lower binding energy oxygen, which, as noted above, most likely corresponds to oxygen bonded to Hf, and the upper map shows higher binding energy oxygen, which most likely corresponds to oxygen bonded to Al or Si. The clear result of this analysis is that the right side of this wafer is richer in Hf. XPS maps of the Hf and Al signals from this surface confirm these results (see Figure A2.6). Figures A2.5 and A2.6 also suggest that there are a few local defects on this wafer. Once again, this type of information would be invaluable to anyone interested in optimizing the corresponding depositions and process equipment.
Figure A2.3. Thicknesses of carbon, alumina (Al$_2$O$_3$), and silica (SiO$_2$) layers on a silicon substrate versus the number of Al$_2$O$_3$ ALD cycles. The thicknesses of the three layers here were determined with Thermo Fisher’s multi-layer thickness calculator Advantage. Figure and caption used/adapted with permission from Thermo Fisher.
Figure A2.4. Results from a line scan across the diameter of a 200 mm wafer showing the variation in thicknesses of a mixed Al\textsubscript{2}O\textsubscript{3}/HfO\textsubscript{2} layer and the underlying SiO\textsubscript{2} film. Figure and caption used/adapted with permission from Thermo Fisher.
Figure A2.5. Chemical state maps of the O 1s signal on a 200 mm wafer. The upper map corresponds to oxygen with a higher binding energy (presumably bonded to Al or Si), and the lower map corresponds to oxygen with a lower binding energy (presumably bonded to Hf).

Figure and caption used/adapted with permission from Thermo Fisher.
Angle-resolved XPS and surface mapping are important capabilities of an XPS instrument. In this article we have covered the basics of AR-XPS, discussed relative depth plots, which show AR-XPS results in a qualitative/semi-quantitative fashion, showed results from Thermo Fisher’s Avantage software, which gives quantitative film thicknesses from AR-XPS measurements, and demonstrated how this tool can be applied for line and surface mapping. To conclude this article we include here a series of references on angle-resolved XPS and XPS in general that the reader may find useful.
Acknowledgment. A significant fraction of the content in this article came from an application note published by Thermo Fisher.\textsuperscript{4-5} The authors thank Thermo Fisher for allowing them to do this.

A2.2 References


Appendix 3: Trends in Advanced XPS Instrumentation. 4. A Modern XPS Instrument and How to Manufacture It in Just Four Days

A3.1 Introduction

X-ray photoelectron spectroscopy, XPS, is the most widely used surface analytical technique, where its most useful attribute is that it provides quantitative elemental analysis of all the elements in a material, except hydrogen and helium, in the outer ca. 10 nm of a material. In addition, XPS spectra yield useful information about the chemical (oxidation) states of the elements in a sample, where these spectra can generally be acquired with little or no sample damage. And while there can certainly be some nuance to XPS peak fitting (sometimes a great deal),\textsuperscript{1-5} a reasonable fraction of XPS spectral interpretation is quite straightforward. XPS can be used for chemical imaging, which provides maps of the lateral distributions of surface species, and it can also be used in conjunction with sputter etching of the surface to generate destructive depth profiles. The quantitative, surface sensitive elemental information that XPS provides is not just important for surface scientists; XPS is useful in any discipline where surface composition is expected to play a role. Some typical applications for XPS include the analysis of thin films, polymers, ceramics, organic coatings, chemical sensors, catalysts, and in surface corrosion. Given the wide range of applications for XPS and the wide range of disciplines that make use of this technique, a growing number of XPS instruments are being installed in multi-user facilities. Here,
users from various backgrounds with differing levels of experience operate the instruments. Accordingly, these instruments need to be geared towards automation, high-throughput analysis, and ease-of-use. The Thermo Scientific K-Alpha+ is built with this type of use in mind. In this article we discuss various aspects of this instrument.

This is the fourth installment in our series on ‘Trends in Advanced XPS Instrumentation’. In the first, we described a variety of aspects of the modern instrument, which included a general description of the technique. For example, we showed a schematic of an XPS, discussed the depth sensitivity of the method, provided a picture of a modern instrument (the Kratos AXIS Supra), showed pictures of the sample stage and loading procedure, presented a diagram of the X-ray and electron paths through the instrument, illustrated the electron detection system of a modern XPS, described ‘Snapshot Spectroscopy’, noted that gas cluster ion beams (GCIBs) are being increasingly used for sample sputtering, explained that multiple analytical techniques like ion scattering spectroscopy (ISS), also known as low energy ion scattering (LEIS), are increasingly combined with/added to modern instruments, and discussed automation of the sample analysis, which is becoming increasingly important in newer instruments like the Axis Supra by Kratos Analytical or the K-Alpha+ by Thermo Fisher Scientific. We also showed survey and narrow scans of a sample of native oxide on silicon collected with a modern instrument. One of these narrow scans showed excellent resolution of the Si 2p3/2 and 2p1/2 peaks from reduced/metallurgical silicon at around 99 eV; the width of the Si 2p3/2 component in the corresponding fit was impressive – just 0.46 eV.

As an aside, we have recently published fairly similar spectra of SiO2 on Si in Surface Science Spectra (SSS). SSS is a great place to archive your XPS, Auger, and ToF-SIMS (time-
of-flight secondary ion mass spectrometry) data. You won’t lose it this way – it won’t be misplaced or erased on a hard drive. In addition, if you go to the work of archiving the data in SSS, you will probably find yourself with a more complete set of information about your sample, instrument, and data than you would otherwise have. This more comprehensive view will probably be hard to pull together after your graduate student or post-doc leaves, and may be very useful to you in your future research. Obviously submitting to SSS will help other researchers as well – we all benefit from a well prepared and vetted database. Recently SSS has started to accept submissions of optical constants of materials. If you have the chance, take a look at the submission from our group on the optical constants of Eagle XG® glass9 and from Nik Podraza’s group at the University of Toledo on the optical constants of silicon nitride.10

In our second article,11 we discussed angle-resolved XPS (AR-XPS), which provides a mechanism for probing the thickness and compositional profiles of thin films at surfaces. In AR-XPS, the takeoff angle, which is typically measured from the surface/plane of the sample and not from the sample normal, is varied resulting in analyses at different depths into a thin film or surface. An obvious advantage of this method is that depth information is obtained without sputtering. We showed AR-XPS results of AlHfO and SiO thin films using the Thermo Scientific Theta Probe XPS spectrometer. In each case, AR-XPS successfully analyzed these materials, providing qualitative/semi-quantitative results along with relative depth plots of the elements present in the films.

In our third installment,12 we described XPS imaging techniques in the context of the Axis line of instruments from Kratos Analytical. An advantage of XPS imaging over other surface sensitive imaging techniques like ToF-SIMS and AES (Auger electron spectroscopy) is the non-
destructive nature of the X-ray beams, especially with the advent of modern parallel imaging detectors that significantly reduce X-ray exposure times. Traditional XPS imaging involves a focused X-ray beam that is rastered across a surface to garner elemental/oxidation state information at every point/pixel. However, the spatial resolution here is limited to the spot size of the X-ray beam. Today, modern instruments employ parallel imaging methods that focus photoelectrons with a particular energy onto a two-dimensional detector plane. In combination with an electron analyzer, this plane preserves the spatial information of the electrons.

In this article we discuss various aspects of the K-Alpha+ XPS system from Thermo Fisher Scientific. This is a relatively low cost instrument that allows a high degree of automation and sample throughput. However, in addition to its technical attributes, we also describe its manufacturing. It is noteworthy here that Thermo Fisher has put in place procedures that allow the K-Alpha+ to be manufactured in just four days. Given the complexity of the equipment, this is a remarkable achievement. These improved procedures have also led to the production of a more reproducible instrument.

**A3.2 Overview of the Thermo Scientific K-Alpha+ System**
Figure A3.1 shows a cut-away image of the K-Alpha+ instrument, with some of its key components labeled. In regular use, the instrument is covered with panels. Although there certainly is pedagogical value in exposing all the components of an instrument, and this should probably be done in a more specialized environment with dedicated users, it is probably best to hide most of it in a multi-user facility so that well-intentioned curiosity will not lead to a broken instrument. Sample loading and unloading on the K-Alpha+ is mostly automated – the user simply places samples onto a sample holder, loads the sample holder onto the platen in the load lock, and activates the pumping and sample transfer through the instrument control software. The instrument then performs the necessary pumping and transfers the sample holder to the motion control stage in the analysis chamber. While no system is entirely idiot proof, this one seems pretty close.

The K-Alpha+ has a four-axis motion control system in its main chamber (X, Y, Z, and rotation). The standard sample holder for this instrument has a 60 mm x 60 mm sample area with 9 premarked sample positions and holes for the sample mounting clips. It is possible to mount and analyze all the samples that will fit on the holder in a single run. An optional tilting sample holder is available for performing AR-XPS. Additional sample handling accessories include a bias-able sample holder for performing sample work function measurements and a vacuum transfer module for loading samples that were prepared in an inert environment without exposing them to air.
Figure A3.1. Cut-away image of the Thermo Scientific K-Alpha+ showing the vacuum hardware under the system cladding. Figure and caption used/adapted with permission from Thermo Fisher.

The K-Alpha+ system includes three cameras to aid in automated sample handling. The platter view camera photographs samples on the holder. From this image, sample coordinates can be stored in the instrument software for subsequent analyses. The reflex-view camera uses periscope style optics to provide a magnified plan-view of the samples inside the analysis chamber. This makes it easy to align features of interest with the analysis beam. The height adjustment
camera is used to automatically adjust the height of the sample for optimum count rate. In addition, the K-Alpha+ automatically adjusts charge compensation on insulating samples. Charge compensation here is accomplished using complementary beams of low energy electrons and positive ions. This system works well enough that charge referencing is seldom needed on insulating samples.

The K-Alpha+ employs a hemispherical electrostatic analyzer. This style of analyzer is used on most XPS instruments, and we have discussed its operating principles in a previous article. The high-throughput electron lens and analyzer on this system are designed to provide a high photoelectron count rate at the detector, resulting in shorter analysis times. The detector is a 128 channel, position sensitive resistive anode. One purpose of having multiple data channels at the detector is to provide high energy resolution for the spectra obtained. Another advantage of using position sensitive detectors is that they allow spectra to be obtained over a limited energy range without the need to scan the retarding fields in the hemispherical analyzer. This is known as snapshot spectroscopy. This mode of analysis is especially useful for rapid acquisition of XPS images or depth profiles, and can dramatically reduce acquisition times.

One notable feature of Thermo Fisher’s XPS instruments is that they use a focused monochromator to adjust X-ray spot size (See Figure A3.2). In contrast, the more traditional way of performing small area XPS analysis is to use an aperture to limit the X-ray spot size. While this latter approach allows spectroscopic data to be collected from a reduced area, it also throws away many of the potential photoelectrons, resulting in reduced signals. By focusing the X-ray beam for small area analysis, Thermo Fisher XPS instruments ensure that all the emitted signal possible is obtained from the area of interest, giving relatively high photoelectron count rates and allowing
for faster analysis. One potential drawback of this approach is that some samples are prone to
damage from the focused X-ray beam, though the longer acquisition times of the aperture-defined
approach can also cause this problem. Fortunately, however, the vast majority of samples can be
analyzed long before sample damage interferes with the measurement. The K-Alpha+ monochromator is also designed to operate at low power, 72 W for the largest analysis area
compared to above 300 W for most aperture defined instruments. By careful design of the
monochromator crystal and the electron optics, this low power X-ray source can be used without
compromising signal intensity. The spot size on the K-Alpha+ can be adjusted from 400 µm to 30
µm in 5 µm increments using its microfocused monochromator. This focused X-ray beam is
rastered across the sample when the instrument is operated in imaging mode by moving the stage
underneath the beam. The minimum lateral resolution for XPS images with the K-Alpha+ is around
30 microns. Thermo Fisher have recently introduced a rapid imaging mode on the K-Alpha+
instrument, SnapMap, which continuously moves the stage while acquiring snapshot spectra at set
intervals. This mode offers around a 20x increase in imaging speed compared to traditional
mapping and allows images with a field of view (FOV) of 0.5 x 0.5 mm to 3 x 3 mm to be collected.
For larger areas, up to the full area of the stage platen (60 x 60 mm) can be used in standard
mapping mode. In all cases, full spectra are acquired at each pixel of the image, enabling chemical
state images to be created.
Figure A3.2. Thermo Fisher Scientific XPS instruments, including the K-Alpha⁺ use a focusing monochromater to control X-ray spot size. The incident electron beam from the cathode is focused onto the anode to give the desired X-ray spot size. Figure and caption used/adapted with permission from Thermo Fisher.

In addition to standard spectroscopy and XPS imaging, the K-Alpha⁺ can be equipped with a standard monatomic ion source or Thermo Fisher’s MAGCIS dual mode monatomic and cluster ion source for sputter depth profiling. Cluster ion sources provide significant advantages over their monatomic counterparts. That is, when monatomic sputter sources are used for depth profiling, the ions penetrate deeply (> 10 nm) into the material. Unfortunately, they travel violently, breaking bonds and causing material rearrangement. On sensitive samples, e.g. polymer films, this sample damage makes it difficult to obtain information about the chemical states of the elements in the sample, i.e., carbon, as a function of depth. Cluster sources, on the other hand, spread the kinetic energy of the projectile across many atoms. These atoms do not penetrate as deeply into a sample
and are less prone to break bonds or scramble atomic layers. Thus, the use of cluster ion sputter often enables chemical state information to be obtained from XPS depth profiles of sensitive samples. When operated at very low energies, cluster ion sources can also be used to gently remove/clean adventitious hydrocarbons from sample surfaces prior to analysis. The MAGCIS source offers both modes of operation so monatomic and cluster ions can be produced from the same gun, which gives the operator a wide choice of options when deciding on an analysis strategy for a particular sample.

As might be expected, the K-Alpha+ can be configured with a UV source for ultraviolet photoelectron spectroscopy (UPS), which can provide useful information about a material’s bandgap and valence band. (Although to be fair to the reader, we should point out that UPS is not, in general, nearly as useful as XPS. The spectra tend to be much more difficult to interpret, very often requiring first principle calculations.) The instrument can also be configured to interface with a glove box for preparation and analysis of atmosphere-sensitive samples. The K-Alpha+ uses Thermo Fisher’s Avantage software, which is common to all of their XPS instruments. The instrument is mostly controlled through this software, and it also includes some powerful analysis and automation features. For example, the instrument can perform survey scans of all the samples loaded in it, automatically identify the elements present from the survey scans, and then perform narrow (high resolution) scans for all the elements present. The software also includes peak deconvolution utilities, which are tools for analyzing depth profiles, angle resolved data, and batch processing of similar spectra.

In summary, the K-Alpha+ provides high throughput XPS spectroscopy, imaging, and depth profiling with a high degree of automation in a compact and relatively economical package.
This instrument now constitutes an established platform. A recent literature search showed that since the launch of the K-Alpha line of instruments in 2006, they have been mentioned in more than 6000 publications. (Google Scholar keywords: “K-Alpha” + XPS + ”Thermo”. This approach is not perfect, as you have to manually remove papers that use “K-Alpha” in other contexts, but it ultimately does get you to around 6000+ publications.) Below, we discuss some recent applications of the instrument, and then provide an overview of how it is manufactured.

A3.3 Analyses with the K-Alpha System

Here are two examples of the use of the K-Alpha to solve real world problems.

A3.3.1 Analysis of Contact Lenses

Characterization/quality control of contact lens surfaces is crucial for their performance, where, obviously, contact lenses are in close contact with the eye so understanding/optimizing their surfaces is extremely important. The Thermo Scientific K-Alpha+ XPS has been used to investigate the surface and chemical compositions of contact lenses. In this analysis, the simple turnkey charge compensation of the K-Alpha+ was used along with its depth profiling capabilities to analyze interfacial regions in these devices. Although a cluster ion beam could be used, it was found that a low energy monatomic ion beam (200 eV argon ions) ensured that the polymer chemistry was preserved during depth profiling. In the C 1s narrow scan of the sample, the C=O, C-O/C-N and C-C/C-H chemical states of carbon could be identified. Figure A3.3 shows results
from an XPS analysis of a batch of contact lenses. The analysis reveals the thickness of a proprietary coating on the lenses across the batch. This type of data/methodology can be particularly helpful for understanding problems related to production and also product reliability.

Figure A3.3. XPS analysis using the K-Alpha+ instrument of a series of contact lenses. Figure and caption used/adapted with permission from Thermo Fisher.

A3.3.2 Mapping the Work Function of a Damaged Solar Cell

Work functions can be measured in an XPS instrument by collecting a survey spectrum across the entire spectral range of the instrument – cut-off energies at the two ends of the binding energy scale can be determined in this way. The photoemission in an XPS experiment primarily
consists of two types of electrons: (i) those that have not suffered an inelastic collision and (ii) those that have lost varying amounts of energy in one or more inelastic collisions. Accordingly, the slowest electrons of the spectrum (low kinetic energy or high binding energy) are the electrons that have barely made it out of the sample and therefore have an average kinetic energy of 0 eV. This position is referred to as the “cut-off”. To accurately measure the cut-off, the sample needs to be biased, so that the sample cut-off is shifted away from the spectrometer cut-off. By measuring the Fermi level in this biased condition, the whole spectrum can be shifted back to the correct position, by aligning the Fermi level with the 0 eV binding energy level ($E_b = 0$ eV). Thus, work functions can be calculated by subtracting the higher binding energy cut-off from the photon’s energy, as this cut-off represents the electrons that had just enough energy on arrival at the surface to overcome the work function of the material. Figure A3.4 shows such a representative plot of the survey spectrum of gold using a monochromated source of X-rays (Al Kα, photon energy 1486.6 eV). Thus, gold has a work function of 5.1 eV.
The determination of work function however relies on two key parameters, accurate calibration of the spectrometer and accurate determination of the photon energy. Modern XPS instruments like the K-Alpha™ from Thermo Fisher having internal standards (copper, silver, and gold), which are used by the instrument to automatically calibrate the binding energy scale. Moreover, the photon energy can be accurately determined by measuring the position of an X-ray induced Auger peak on the binding energy scale, typically the Cu L$_{3}$VV signal; the measured deviation from the reference value of 567.9 eV being the difference in the photon energy from 1486.6 eV.$^{14}$

A solar cell based on a thin film of CIGS, Cu(In,Ga)Se$_2$, which had been damaged because of delamination, was used to measure and map the work function of the damaged cell structure. Work function determinations of individual regions were carried out over an area largely dominated by silver, which is used as an electrical contact to the device, returning a value of 4.3 eV. Moreover, principle component analysis (PCA) applied to an area spanning different materials created a work function map of the analyzed region.

A3.4 Manufacturing the K-Alpha

A3.4.1 Just in Time Manufacturing
The production system for the K-Alpha Plus Spectrometer X-ray Photoelectron spectrometer at Thermo Fisher revolves around the business concept of Just in Time (JIT) manufacturing, which is delivery of the right parts in the right amounts at the right times using minimum resources. This paradigm results in a dramatic decrease in inventory and reduces or prevents over production.

### A3.4.2 Quality Control

Quality control in the Thermo Fisher production line involves making intelligent decisions at every stage of the process and ensuring that when a defect is detected, it is reported and root cause analysis is performed to prevent repeats. The goal of the manufacturing system is to run the production line in a continuous fashion. Thus, the K-Alpha+ is the result of avoiding the multiplying effects of problems that go unidentified in earlier production stages. Indeed, the K-Alpha+ system was designed and built around its manufacturing process. The production area is split into four sections:

- Build
- Pre-Test
- Final Test
- Finished Goods

After the build is complete, the system is subjected to a complete vacuum test before the final components are added. A high temperature bake for 72 hours ensures that all internal surfaces are clean prior to a full regime of tests. A rigorous testing regime is required for each component of
the system prior to shipment. K-Alpha$^+$ systems are tested outside their standard operating conditions to ascertain the reliability of the products under extreme conditions. These assessments help monitor/identify batch-to-batch variation in the components. The resulting information is then shared with the suppliers, which helps maintain the quality of the product. A proper feedback loop between the production line, suppliers, and customers is key to the reliability of the instrument. Customer feedback is paramount to the continuous development cycle of the instrument.

Figure A3.5. Production line at Thermo Fisher. Figure and caption used/adapted with permission from Thermo Fisher.
A3.4.3 Competitive Pricing

The K-Alpha+ is one of the most competitively priced XPS systems on the market, which is largely attributed to the waste minimization and efficiency of JIT manufacturing. Team members actively participate in a continuous cycle of optimization, analysis, and waste reduction, which results in cost effective production. An important key to the efficiency of the K-Alpha+ production lies in close monitoring of the production line in the following four areas:

- Setting and maintaining work standards
- Solving daily performance problems
- Participating in the process of continuous improvement
- Efficiently organizing teamwork

The areas on the production floor are defined in such a fashion that the flow of the process is clear and can be analyzed for further optimization. Standard operating procedures are placed at the assembly points. A moving marker line further helps denote the progress at any stage of the production. Using this production process, a K-Alpha+ can be built in just 4 days, and the entire process including extensive testing is complete in just 10 working days, where this testing utilizes automated processes to test overnight and even during weekends.

A3.5 References


Appendix 4: Trends in Advanced XPS Instrumentation. 7. Advanced Software Capabilities

A4.1 Introduction

X-ray Photoelectron Spectroscopy (XPS) is one of the most widely used surface analysis tools – it is garnering more than 10,000 mentions in the literature each year. These citations and references to the technique are an indication that the user base of XPS is steadily growing. XPS has been successfully applied to a wide variety of samples including metals, ceramics, polymers, biological materials, and even liquids. The basic principles of XPS are well understood.\textsuperscript{1-3} XPS is based on the photoelectric effect, i.e., a surface is probed with X-rays that generate photoelectrons that are energy analyzed. The energies of these photoelectrons are characteristic of the elements in the material that is analyzed. XPS is surface sensitive. It only probes the outermost 5 – 10 nm of a material. With the exception of hydrogen and helium, XPS senses all the elements, and it does so quantitatively. Only in the rarest of cases would anyone want to detect/quantify helium at a surface by XPS, but hydrogen is hugely important, so the inability of the technique to detect this element is a significant deficiency. In addition, XPS provides information about oxidation states and chemical environments. For these reasons, XPS is truly an invaluable surface characterization tool. While XPS has existed for years, its instrumentation and accompanying software continue to improve. We have described some of these advances in a series of VT&C articles with the tag line:
“Trends in Advanced XPS Instrumentation”. Like many analytical instruments these days XPS is becoming more automated. This capability is particularly helpful in user facilities with XPS instruments that need to run many samples. The obvious advantage here is that one can set up a run in the evening and let the instrument acquire data at night, and/or let it run during the day to free up the analyst to do other work. However, it should be pointed out that the data analysis associated with XPS can often take considerably longer than the data acquisition itself. Some of our research has focused on various aspects of XPS data analysis. In this article, we focus on the software aspect of XPS. Software packages are an integral part of almost any analytical instrument these days, and purchasing decisions are often based as much on the software of an instrument as its hardware. Increasingly, XPS instruments are entirely controlled through their software – in modern instruments users are not typically found opening and closing valves or manipulating samples manually. Indeed, as the hardware of modern XPS instruments has improved, the software has improved with it, making systems easier for beginners to use and giving greater levels of system control to advanced users. Some features of modern XPS software packages include:

- Remote system operation (What’s better than doing XPS in the lab? Doing it from your couch!)
- Automated data workup and report generation
- Improved data storage formats that allow advanced spectral processing
- Advanced data processing tools including built-in chemometrics (As an aside, one of our most recent articles describes the use of principal component analysis, multivariate curve resolution, and pattern recognition entropy to analyze XPS depth profiles.)
Vacuum system control, monitoring, and logging

Automatic charge compensation and sample height adjustment

Improved user interfaces

This past summer some of us visited SPECS Surface Nanoanalysis GmbH in Berlin, Germany, where we had an opportunity to use their EnviroESCA Near Ambient Pressure XPS (NAP-XPS) system. NAP-XPS greatly expands the range of samples that can be analyzed and experiments that can be performed with XPS – SPECS has analyzed materials like water, coffee beans, and Italian cheese, all of which would be impossible to probe with a conventional, high vacuum XPS system. Obviously, the instrumentation that allows these types of measurements to be made must involve differential pumping and electron optics that can direct photoelectrons to a detector across multiple pumping stages. We saw firsthand how important advanced software is for operating this type of system. Here, we discuss some of the software innovations that are part of this complex instrument.

These capabilities make this instrument and other SPECS instruments relatively easy to use and improve their data collection and workup capabilities. Two areas of focus here will be the software’s ability to save and allow examination of every individual narrow scan, and its ability to interlace during data acquisition, i.e., alternate between taking different narrow scans and survey scans to avoid taking a block of any one type of scan. The former capability is particularly valuable because it gives the user the raw data at quite a fundamental level, and not just a compilation of it, which may be treated in some way by the software before the user gets it. Saving the data at this more basic level is a way of ensuring data safety and compliance. In this contribution we also show a principal component analysis (PCA) of a series of narrow scans that indicate that sample charging/degradation/alteration is not taking place over the course of the experiment. Of course
SPECS is not the only company offering advanced, innovative software with its XPS instruments, and there are also standalone packages dedicated to data analysis, e.g., CasaXPS.

**A4.2 SpecsLab Prodigy Software: The Brain Behind the EnviroESCA Hardware**

Like many other instrument manufacturers, SPECS has a team of software engineers dedicated to improving their instrument software. The software used on most of their XPS systems and related instruments is called Specslab Prodigy, while the EnviroESCA runs on a related software package called Keystone. Through Keystone, the EnviroESCA is fully automated, which includes sample transfer, pump down, venting, gas dosing, and sample heating. Software packages typically control instrument parameters/allow hardware control, establish scan modes, allow data collection to be set up, and are useful for data workup. These capabilities can be particularly useful when working in more advanced modes, e.g., when capturing angle-resolved XPS data or acquiring multidimensional data in XPS image acquisition and/or depth profiling. In all cases, the end goal of XPS software packages is to maximize understanding of the sample with minimum user input and effort.

Prodigy acts as the main control interface for SPECS instruments. Most hardware components are directly controlled through it. This includes the parts of the system responsible for acquiring data, including the X-ray gun and the analyzer, and also the vacuum hardware, e.g., the pumps, valves, and gauges on the system. Figure A4.1 is a screenshot from Prodigy, which shows the vacuum control screen for a custom-built NAP-XPS system. Here, chamber pressures, valve states, pump settings, and mass flow controller (MFC) settings can be viewed and edited.
Figure A4.1. Screenshot from the SpecsLab Prodigy software of the page that allows control of the system vacuum components.

Figure A4.2 shows a screenshot of the Prodigy Experiment Editor, which is used to control data acquisition. Through this interface, the user controls the analyzer, the X-ray system, the
sputter gun, the flood gun, and the motion control system. Experimental templates can be saved here and used to speed up routine analysis tasks.

Figure A4.2. Screenshot of the Experiment Editor in Prodigy.

Having all system components controlled by a software package has enabled XPS instruments to become highly automated; all actions that the system must perform during an experiment can be specified by the user before an experiment begins. Steps in an experiment can be initiated by a timer, or they can be started when a system component reaches a desired state. Setting up these systems for routine analyses is reasonably simple, but the recipes for more advanced experiments can become quite complex. To simplify this process, Prodigy includes
several advanced automation features that simplify certain tasks. These include automation of sputter depth profiling, sample sputter cleaning, flood gun adjustment, and sample height adjustment. In general, automation offers a number of important and tangible benefits. These include improved user safety and reduced risk to the system, e.g., from novice users crashing a pump or contaminating a chamber by throwing valves in the wrong order. In addition, automation prevents dropping of samples into the UHV chamber from manual sample transfer, it avoids overpressure or other dangerous safety conditions, and, in general, it keeps the user physically away from sensitive and dangerous parts of the system, e.g., those at high-voltage. Additional automation commands are included for performing analyses while ramping other system parameters, e.g., performing an XPS analysis at a series of specified sample temperatures while heating a sample holder, or ramping the flowrate of a gas in an NAP-XPS system.

Improvements in system control software has also enabled XPS instruments to be operated remotely. From the perspective of the analyst, the nicest thing about this feature is the ability to monitor the progress of an analysis or to adjust analysis parameters without needing to be next to the instrument. Prodigy additionally includes the capability to remotely control the instrument via TCP/IP with text commands. Thus, Prodigy interfaces with other data collection/laboratory software, e.g., LabVIEW, for situations where XPS is part of a larger production or analysis scheme. Prodigy can also be used to remotely control other devices. This feature can be useful when third-party systems are integrated onto a SPECS instrument. Thus, Prodigy is a versatile platform for controlling custom-configured instruments.

Prodigy is also a powerful platform for XPS data analysis. It handles the standard XPS data types from single spot (spectroscopy) analysis, imaging, sputter depth profiling (see Figure A4.3),
and angle-resolved experiments, and it has features that simplify the processing of large data sets. Templates can again be saved for automatically generating reports. In addition, Prodigy/Keystone have automatic quantification routines for calculating the elemental composition of analyzed surfaces, reporting capabilities including the ability to generate customizable, signed reports (pdfs), and audit trail capabilities.

Figure A4.3. Sputter depth profiling interface in the Prodigy software.

A4.3 Saving/Archiving of Individual Scans
One feature we found particularly interesting in the Prodigy software is that single scans from the measurements are saved so that they can be viewed and manipulated later to understand and enhance data quality. XPS measurements are typically the result of multiple, combined scans. For example, to record a survey spectrum, we might scan from 0 - 1300 eV ten times and record the final summation of all ten spectra. This signal averaging/summing, which improves signal-to-noise ratios, is standard practice in analytical chemistry. Often we are only interested in these types of final results/spectra. However, there are some circumstances in which being able to see the individual scans can be quite helpful. For example, the charging of insulating samples can change dramatically as a function of temperature in heated stage experiments, and it can also vary strongly as a function of pressure in NAP-XPS experiments. These effects result in peaks changing position or drifting with time as an experiment progresses. In an averaged/summed result, the final peaks can be broad or manifest as doublets. However, if the scans can be analyzed individually or in small groups, the spectra can often be grouped/aligned and resummed to produce higher quality data that better reflects the time-evolving signal from the sample. This single scan feature can also be useful for spotting sample damage as a function of X-ray exposure in samples that are prone to it. Figure A4.4 shows eight individual survey scans from an analysis of an aqueous solution (Coca Cola®) as obtained on the SPECS EnviroESCA. The consistency between the survey scans is a good indication that the sample is not charging or changing significantly during data acquisition and that the instrument is also running consistently/repeatably during this time. Survey scans are useful for detecting changes in samples/spectra during data acquisition because they generally have quite good signal-to-noise ratios.
Narrow scans can also be very helpful in detecting sample charging because they will better show changes in sample chemistry than survey scans. However, they generally have lower signal-to-noise ratios than survey scans so best results here will probably be obtained for elements that are abundant and/or have higher atomic sensitivity factors. Figure A4.4 show a series of individual C 1s narrow scans from a sample of Coca Cola®. The concentration of carbon at the surface of this aqueous solution appears to be quite low and the signals are noisy. So while it appears that no significant changes are taking place to this sample as data acquisition proceeds, it would be hard to demonstrate here that any subtle changes are or are not taking place here. In contrast, Figure A4.5 shows a series of individual O 1s narrow scans from the Coca Cola® sample. The narrower peak at higher binding energy corresponds to gas phase water and the broader peak at lower
binding energy corresponds to bulk, liquid water. These spectra have much higher signal-to-noise ratios and their similarity makes a very good case for sample and run-to-run consistency.

Figure A4.5. Nine individual O 1s narrow scans of an aqueous solution (Coca Cola®) obtained sequentially by NAP-XPS.

Principal component analysis (PCA) is a powerful tool for the statistical analysis of spectra. We performed PCA on the O 1s narrow scans in Figure 5 to determine if there was any trend in the samples and/or if any outliers were present – ideally we would not see either. For preprocessing, the data were not normalized – it seemed like this would be unnecessary because the spectra were collected under identical conditions (at least nominally). The spectra were, however, mean centered. Mean centering or autoscaling are two of the most common
preprocessing methods in PCA. Autoscaling seemed inappropriate here because it would give the noisy baseline regions of the spectra the same importance as those containing signal. Figure A4.6 shows the results of this analysis. The lower two plots (the scores on PC1 and PC2 vs. sample number, where ‘PC’ stands for ‘Principal Component’) do not show any strong trend in the data, which suggests that the sample is not changing or changing during data acquisition. The Q Residuals vs. Hotelling T² plot shows that all the spectra (points) are within reasonable limits, which indicates that there are no outliers here.

**Figure A4.6.** Principal component analysis (PCA) of the nine O 1s narrow shown in Figure 5. Top: plot of Q Residuals vs. Hotelling T², middle: plot of scores on PC1 vs. sample number, and bottom: plot of scores on PC2 vs. sample number.
A4.4. Interlacing of Narrow and Survey Scans During Acquisition

Another feature that caught our eye is Prodigy’s looping (interlacing) capability, which refers to Prodigy’s ability to acquire scans in a mixed order in an experiment. In a traditional experiment one will generally take all the scans of a given type at the same time, i.e., all the individual survey scans are collected, followed by all the individual narrow scans of a given type, followed by all the individual narrow scans of another type, etc. In contrast, Prodigy allows one to easily acquire a single survey scan, a few narrow scans of one type, a few narrow scans of another type, another survey scans, a few narrow scans of the first type, etc. This ability to switch repeatedly between scan types can be extremely useful for identifying small changes that may be taking a place in a material due to beam damage or charging that would not otherwise be identifiable if every scan of a given type was acquired at the same time. Obviously this interlacing capability is also a check on instrument stability.

A4.5 Conclusions
We have discussed the features of a modern XPS software package: the Specslab Prodigy system, which is used in most SPECS instruments. Software is a crucial and central part of any modern XPS instrument. Not only does the software monitor and control many essential hardware systems on the instrument, it acts as the user interface and as a tool for data reduction. We have highlighted two specific capabilities of the software – its ability to collect individual narrow scans, and its ability to acquire them in an interlaced fashion.

A4.6 References


Figure A5.1. Scores on Components 5 – 7 from a seven-component MCR analysis of XPS data from a depth profile of SiO$_2$ on a Si substrate. See Figure 19 in the main text and related discussion.
Figure A5.2. Loadings on Components 5 – 7 from a seven-component MCR analysis of XPS data from a depth profile of SiO₂ on a Si substrate corresponding to (a) oxygen (O 1s) and (b) silicon (Si 2p).
Figure A5.3. Oxygen-to-silicon ratio as a function of spectrum number (depth index) during the XPS depth profiling of a film of SiO$_2$ on Si.
Figure A5.4. Ta 4f spectra collected during an XPS depth profile study of Ta₂O₅ on Ta substrate. Spectra 36 and 37 are very similar to spectra 29 – 25 and so are not included here.
Figure A5.5. (a) PRE values and (b) backward difference PRE as a function of spectrum number for ToF-SIMS depth profiling of a C$_3$F$_6$ polymer layer on Si. A total of 19 peaks were selected from each spectrum for these analyses.
Figure A5.6. (a) PRE values and (b) finite difference PRE values as a function of spectrum number for a ToF-SIMS depth profile of a PNIPAM polymer film on Si. The entire spectra from m/z 0 to 150 were used in this analysis.
Figure A5.7. Si 2p spectra collected during an XPS depth profiling of SiO₂ on Si. Spectra 50-54 are very similar to spectra 43-49 and so are omitted here.

PCA of Plasma Polymerized C₃F₆ on Si

Figure descriptions:

Figure A5.8 shows the eigenvalue distribution for the Plasma Polymerized C₃F₆ on Si depth profile data and each includes a knee in the plots indicative of non-random distributions. Auto-scaling is included but not recommended for this data. Figure A5.9 and Figure A5.10 show the scores and loadings on the first two PCs for this data and show clear trends in time/depth (scores), and
variables (loadings) that are consistent with the physics and chemistry of the measurements. Cross-validation results in Figure A5.11, also indicate correlation in the data because the cross-validation error [root mean square error of cross-validation (RMSECV)] decreases as PCs are added. As expected the root mean square error of calibration (RMSEC) decreases as PCs are added.

**Figure A5.8.** Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of plasma polymerized C$_3$F$_6$ on Si.
Figure A5.9. Scores on PCs 1 and 2 of plasma polymerized C$_3$F$_6$ on Si.

Figure A5.10. Loadings on PC 1 and 2 of plasma polymerized C$_3$F$_6$ on Si.
Figure A5.11. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets.

Principal Components Analysis Model:
X-block: 24 by 8984
Included: [1-24] [1-8984]
Included (in axis units): [1-24] [0.029112-149.99]
Preprocessing: Normalize (1-Norm, Area = 1)
Num. PCs: 2
Algorithm: SVD
Percent Variance Captured by PCA Model:

<table>
<thead>
<tr>
<th>Number</th>
<th>Cov(X)</th>
<th>This PC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.15e-02</td>
<td>69.50</td>
<td>69.50</td>
</tr>
<tr>
<td>2</td>
<td>1.76e-02</td>
<td>29.52</td>
<td>99.02</td>
</tr>
</tbody>
</table>
PCA of Plasma Polymerized PNIPAM on Si

Figure descriptions:

Figure A5.12 shows the eigenvalue distribution for the plasma polymerized PNIPAM on Si depth profile data and each includes a knee in the plots indicative of non-random distributions. Auto-scaling is included but not recommended for this data. Figure A5.13 and Figure A5.14 show the scores and loadings on the first two PCs for this data and show clear trends in time/depth (scores), and variables (loadings) that are consistent with the physics and chemistry of the measurements. Cross-validation results in Figure A5.15, also indicate correlation in the data because the cross-validation error [root mean square error of cross-validation (RMSECV)] decreases as PCs are added. As expected the root mean square error of calibration (RMSEC) decreases as PCs are added.
Figure A5.12. Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of plasma polymerized PNIPAM on Si.

Figure A5.13. Scores on PCs 1 and 2 of plasma polymerized PNIPAM on Si.
Figure A5.14. Loadings on PCs 1 and 2 of plasma polymerized PNIPAM on Si.

Figure A5.15. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of plasma polymerized PNIPAM on Si.
Principal Components Analysis Model:

X-block: 25 by 8661

Included: [1-25] [1-8661]

Included (in axis units): [1-25] [0.066049-149.99]

Preprocessing: Normalize (1-Norm, Area = 1)

Num. PCs: 2

Algorithm: SVD

Percent Variance Captured by PCA Model:

<table>
<thead>
<tr>
<th>Number</th>
<th>Cov(X)</th>
<th>This PC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.47e-02</td>
<td>88.51</td>
<td>88.51</td>
</tr>
<tr>
<td>2</td>
<td>4.15e-03</td>
<td>10.57</td>
<td>99.08</td>
</tr>
</tbody>
</table>

PCA of SiO\textsubscript{2} on Si

Figure descriptions:
Figure A5.16 shows the eigenvalue distribution for the depth profile data of SiO$_2$ on Si and each includes a knee in the plots indicative of non-random distributions. Auto-scaling is included but not recommended for this data. Figure A5.17 and Figure A5.18 show the scores and loadings on the first two PCs for this data and show clear trends in time/depth (scores), and variables (loadings) that are consistent with the physics and chemistry of the measurements. Cross-validation results in Figure A5.19, also indicate correlation in the data because the cross-validation error [root mean square error of cross-validation (RMSECV)] decreases as PCs are added. As expected the root mean square error of calibration (RMSEC) decreases as PCs are added.

**Figure A5.16.** Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of SiO2 on Si.
Figure A5.17. Scores on PCs 1 and 2 of SiO$_2$ on Si.
Figure A5.18. Loadings on PCs 1 and 2 of SiO$_2$ on Si.

Figure A5.19. Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of SiO$_2$ on Si.
**Principal Components Analysis Model:**

X-block: 42 by 242

Included: [1-42] [1-242]

Included (in axis units): [1-42] [539-93]

Preprocessing: Normalize (1-Norm, Area = 1)

Num. PCs: 2

Algorithm: SVD

Percent Variance Captured by PCA Model:

<table>
<thead>
<tr>
<th>Number</th>
<th>Cov(X)</th>
<th>This PC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.39e-03</td>
<td>66.45</td>
<td>66.45</td>
</tr>
<tr>
<td>2</td>
<td>4.05e-03</td>
<td>32.06</td>
<td>98.50</td>
</tr>
</tbody>
</table>

**PCA of Ta₂O₅ on Ta**

*Figure descriptions:*
Figure A5.20 shows the eigenvalue distribution for Ta$_2$O$_5$ on Ta depth profile data and each includes a knee in the plots indicative of non-random distributions. Auto-scaling is included but not recommended for this data. Figure A5.21 and Figure A5.22 show the scores and loadings on the first two PCs for this data and show clear trends in time/depth (scores), and variables (loadings) that are consistent with the physics and chemistry of the measurements. Cross-validation results in Figure A5.23, also indicate correlation in the data because the cross-validation error [root mean square error of cross-validation (RMSECV)] decreases as PCs are added. As expected the root mean square error of calibration (RMSEC) decreases as PCs are added.

**Figure A5.20.** Scree plots for three different data pre-processing: autoscale+1-norm, 1-norm, Poission+1-norm of Ta$_2$O$_5$ on Ta depth.
Figure A5.21. Scores on PC 1 and 2 of Ta$_2$O$_5$ on Ta.
**Figure A5.22.** Loadings on PC 1 and 2 of Ta$_2$O$_5$ on Ta.

**Figure A5.23.** Cross-validation and calibration error for two different data splitting. (Left) Random subsets, five subsets, five repeats. (Right) Venetian blinds, five subsets of Ta$_2$O$_5$ on Ta.

*Principal Components Analysis Model:*

X-block: 26 by 370

Included: [1-26] [1-370]

Included (in axis units): [1-26] [540-14]

Preprocessing: Normalize (1-Norm, Area = 1)

Num. PCs: 2

Algorithm: SVD

Percent Variance Captured by PCA Model

<table>
<thead>
<tr>
<th>Number</th>
<th>Cov(X)</th>
<th>This PC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>1</td>
<td>3.67e-03</td>
<td>88.79</td>
<td>88.79</td>
</tr>
<tr>
<td>2</td>
<td>4.15e-04</td>
<td>10.03</td>
<td>98.82</td>
</tr>
</tbody>
</table>
Appendix 6: Supporting Information for Chapter 4

Figure A6.1 shows the PRE values from a series of Si 2p narrow scans that were obtained from an XPS depth profile through a film of SiO_2 on silicon. As we noted previously\(^1\), the first two data points in Figure A6.1 suggest that there is some difference between the very top of this SiO_2 surface and the remainder of the layer. The next set of data points, between Depth Indices 3 and 13, have similar PRE values and correspond to the SiO_2 film. The data points at Depth Indices 22 to 42 also form a set of similar values and correspond to the bulk Si substrate. Finally, the data point at Depth Index 18 at the apex of the sharp peak in Figure A6.1, and also those around it, correspond to the Si-SiO_2 interface. Figure A6.2 shows representative, normalized spectra from each of the four regions identified during PRE analysis of this depth profile. Figure A6.3 shows the spectra in Figure A6.2 in their reordered form. Supporting Information Figures A6.4 – A6.9 similarly show two examples of raw data (spectra), PRE analysis of these spectra, and the corresponding reordered spectra.
**Figure A6.1.** PRE values of Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. While these results quite closely mirror those we previously published, the number of data points here is different.

![Graph of Si 2p narrow scans](image)

**Figure A6.2.** Representative, normalized Si 2p narrow scans from an XPS depth profile through a film of SiO$_2$ on Si. ‘Spectrum 1’, ‘Spectrum 3’, ‘Spectrum 18’, and ‘Spectrum 30’ corresponds to the first, third, eighteenth, and thirtieth data points in Figure A6.1.
Figure A6.3. The same four Si 2p narrow scans that are in Figure A6.2 plotted in their reordered state.
Figure A6.4. PRE values of XPS Ta 4f narrow scans vs. depth index from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta.
Figure A6.5. Selected Ta 4f narrow scans from an XPS depth profile through a film of Ta$_2$O$_5$ on Ta.
**Figure A6.6.** The same four Ta 4f narrow scans that are in Figure A6.5 plotted in their ordered state.
Figure A6.7. PRE values of mass spectra from a ToF-SIMS depth profile through a film of plasma polymerized C₃F₆ on Si as a function of depth index.
Figure A6.8. Representative, normalized mass spectra (actually areas of 19 selected peaks) from a ToF-SIMS depth profile through a film of plasma polymerized C₃F₆ on Si as a function of depth index.
Figure A6.9. The same four spectra in Figure A6.8 plotted in their reordered state. Data points in the spectra are omitted wherever their values are zero.
Appendix 7: Supporting Information for Chapter 5

Mock Normalized Spectrum

<table>
<thead>
<tr>
<th>Data Point No.</th>
<th>Value</th>
<th>Normalized Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>5/346= 0.0144</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>11/346=0.0318</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13/346=0.0376</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>18/346=0.052</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>7/346=0.0202</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>9/346=0.026</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2/346=0.0058</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8/346=0.0231</td>
</tr>
<tr>
<td>9</td>
<td>35</td>
<td>35/346=0.1011</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
<td>42/346=0.1214</td>
</tr>
<tr>
<td>11</td>
<td>111</td>
<td>111/346=0.3208</td>
</tr>
<tr>
<td>12</td>
<td>85</td>
<td>85/346=0.2457</td>
</tr>
<tr>
<td>Total</td>
<td>346</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure A7.1. A version of Figure 2 from the manuscript with thinner line widths indicating plotting artefacts.
Figure A7.2. Comparison of the two curve fitting techniques- Savitzky-Golay (left) with a frame length of 151 vs Spline (right).
Figure A7.3. (a) and (c): Selection of mass chromatograms above 0.0σ and 0.5σ respectively, (b) and (d) corresponding reduced TICCs
Figure A7.4. (a) TICC obtained with a threshold of 0.5σ and (b) individual mass chromatograms constituting this TICC.
Appendix 8: Supporting Information for Chapter 7

Figure A8.1. Dendrogram from a cluster analysis of the 0, 5, and 10 % electropherograms under consideration in this study. The data were pre-processed using range-selection followed by normalization (1-norm). ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0, 5, and 10 % samples, and

**Figure A8.2.** Dendrogram from a cluster analysis of the 0 %, 5 %, and 10 % electropherograms under consideration in this study. ‘Clean’, ‘D5%’, and ‘D10%’ represent the 0 %, 5 %, and 10% samples, and ‘A’, ‘B’, and ‘C’ represent the three subjects. The numbers after ‘A’, ‘B’, or ‘C’ represent replica runs. The data was pre-processed using range-selection followed by autoscaling.
Figure A8.3. RMSEC (orange) and RMSECV (blue) plots from PCA calculations of the electropherograms from Subjects A, B and C for doped (5 % and 10 %) and undoped samples.
Figure A8.4. Scores on PC1 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 1-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.5. Scores on PC1 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.6. Scores on PC2 from replicate runs of Subjects A, B and C for doped (5% and 10%) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.7. Scores on PC3 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.8. Scores on PC4 from replicate runs of Subjects A, B and C for doped (5% and 10%) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.9. Scores on PC5 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.10. Scores on PC6 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.11. Scores on PC7 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.12. Scores on PC8 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.13. Scores on PC9 from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.14. Hotelling $T^2$ vs Q residual plot from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 1-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
Figure A8.15. Hotelling $T^2$ vs Q residual plot from replicate runs of Subjects A, B and C for doped (5 % and 10 %) and undoped samples for a 9-Component PCA Model. Note: The undoped samples are labeled as ‘Clean’ and the replicate runs are labeled with the subject name and run number, i.e., A2 represents the second replicate run of Subject A.
**Figure A8.16.** RMSEC (orange) and RMSECV (blue) plots from PLS calculations of the electropherograms from Subjects A, B and C for doped (5 % and 10 %) and undoped samples.
Figure A8.17. Raw PRE values of electrophoretic separations of subjects A, B and C for various doping levels (0 %, 5 % and 10 %).