Practical guide on chemometrics/informatics in x-ray photoelectron spectroscopy (XPS). I. Introduction to methods useful for large or complex datasets

T. Avval, N. Gallagher, D. Morgan, P. Bargiela, N. Fairley, V. Fernandez, M. Linford

Practical Guide on Chemometrics/Informatics in X-ray Photoelectron Spectroscopy (XPS), Part 1: Introduction to Methods Useful for Large or Complex Data Sets

Running title: Practical guide on chemometrics/informatics in XPS. Part 1. Introduction.

Running Authors: Avval et al.

Tahereh G. Avval,[1] Neal Gallagher,[2] David Morgan,[3,4] Pascal Bargiela,[5] Neal Fairley,[6] Vincent Fernandez,[7] Matthew R. Linford[1,a)]

[1]Department of Chemistry and Biochemistry, Brigham Young University, C100 BNSN, Provo, Utah 84602, USA
[2]Eigenvector Research, Inc., Manson, Washington 98831, United States
[3]Max Planck- Cardiff Centre on the Fundamentals of Heterogeneous Catalysis FUNCAT, Cardiff Catalysis Institute, School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff, CF10 3AT, UK
[4]HarwellXPS – EPSRC National Facility for Photoelectron spectroscopy, RCaH, Didcot, Oxon. OX11 0FA, UK
[5]The Institute for Research on Catalysis and the Environment of Lyon (IRCELYON), 2 Avenue Albert Einstein, 69626 Villeurbanne, France
[6]Casa Software Ltd., Bay House, Teignmouth, UK
[7]Nantes Université, CNRS, Institut des Matériaux de Nantes Jean Rouxel, IMN, F-44000 Nantes, France

[a)] Electronic mail: mrlinford@chem.byu.edu

Chemometrics/informatics, and data analysis in general, are increasingly important topics in X-ray photoelectron spectroscopy (XPS) because of the large amount of information (data/spectra) that are often collected in degradation, depth profiling, *operando,* and imaging studies. In this guide, we discuss vital, theoretical aspects and considerations for chemometrics/informatics analyses of X-ray photoelectron spectroscopy (XPS) data with a focus on exploratory data analysis (EDA) tools that can be used to probe XPS data sets.

These tools include a summary statistic (pattern recognition entropy, PRE), principal component analysis (PCA), multivariate curve resolution (MCR), and cluster analysis. The use of these tools is explained through the following steps: A. Gather/use all the available information about one's samples, B. Examine (plot) the raw data, C. Developing a general strategy for the chemometrics/informatics analysis, D. Preprocess the data, E. Where to start a chemometrics/informatics analysis, including identifying outliers or unexpected features in data sets, F. Determine the number of abstract factors to keep in a model, G. Return to the original data after a chemometrics/informatics analysis to confirm findings, H. Perform multivariate curve resolution (MCR), I. Peak fit the MCR factors, J. Identify intermediates in MCR analyses, K. Perform cluster analysis, and L. How to start doing chemometrics/informatics in one's work. This guide has a companion paper that illustrates these steps/principles by applying them to two fairly large XPS data sets. In these papers, special emphasis is placed on MCR. Indeed, in this paper and its companion, we believe that, for the first time, it is suggested and shown that (i) MCR components/factors can be peak fit as though they were XPS narrow scans, and (ii) MCR can reveal intermediates in the degradation of a material. The other chemometrics/informatics methods are also useful in demonstrating the presence of outliers, a break (irregularity) in one of the data sets, and the general trajectory/evolution of the data sets. Cluster analysis generated a series of average spectra that describe the evolution of one of the data sets.

I. INTRODUCTION

Chemometrics/informatics methods have been used for years to analyze large and complex data sets. However, in spite of previous work in this area,[1-5] this capability has been overlooked by much of the X-ray photoelectron spectroscopy (XPS) community. Indeed, multivariate/chemometrics methods may not have been significantly adopted and employed by XPS practitioners because of the general unfamiliarity of many scientists with these methods. Chemometrics/informatics methods are particularly relevant to XPS because of the trend to collect increasingly large data sets in degradation, depth profiling, *operando*, and imaging studies. That is, the XPS community needs tools to deal with the very large amounts of information that are generated in these ways. The first extensive use of chemometrics /informatics algorithms was in a degradation study of a PVC/polymethymethacrylate (PMMA) blend using tools that included principal component analysis (PCA), multivariate curve resolution (MCR), and image classification.[6, 7] Chemometrics/informatics methods reduce the dimensionality of large and complex data sets and may extract hidden features in data. These techniques can be used in combination with conventional XPS peak fitting. Fundamentally, multivariate chemometrics/informatics methods work in XPS data analysis because of the high degree of correlation between the spectra in many data sets. Chemometrics/informatics can also guide experimental design to maximize the interpretability of experimental results. The time-of-flight secondary ion mass spectrometry (ToF-SIMS) community appears to have recognized the importance of chemometrics methods to a somewhat greater extent than the XPS community.[1, 7-13]

XPS is the most widely used and important method for chemically analyzing surfaces.[14-17] In XPS, a beam of X-rays, which is directed onto a surface, generates photoelectrons via the photoelectric effect. The kinetic energies of these photoelectrons are measured, converted into binding energies, and used to identify the elements present at sample surfaces. Relatively small 'chemical shifts' in the resulting peak positions (typically 1 – 4 eV, but sometimes as large as 10 eV) reveal the chemical (oxidation) states of the elements.[18] While the X-rays used in XPS can penetrate ca. 1 micron into a material, the photoelectrons they generate can only escape in an unattenuated fashion from the upper ca. 5 – 10 nm of it (or deeper with hard X-ray sources). Accordingly, XPS is a surface sensitive spectroscopy. Furthermore, while sample damage is often minimal in XPS, e.g., for many inorganic materials, it does occur in some cases. Because XPS peak widths and chemical shifts are of similar magnitudes, peak fitting is often necessary in XPS data analysis. For quite a few years, XPS experts have expressed concern over the quality of some of the XPS peak fitting in the scientific literature. In response to this issue, which is part of the larger problem of reproducibility in science,[19, 20] a group of experts has recently produced a series of guides that cover multiple aspects of XPS.[17, 21-30] These guides follow many efforts by XPS experts to educate the broader community, including through ISO and ASTM standards. This particular guide is part of a second series of guides that covers additional topics related to XPS, and also other surface analytical techniques.

This paper is a guide for analyzing large XPS data sets using chemometrics/informatics methods. It provides instructions for applying multiple exploratory data analysis (EDA) methods to XPS spectra, which include a summary

statistic (pattern recognition entropy, PRE), principal component analysis (PCA), multivariate curve resolution (MCR), and cluster analysis. In particular, this guide is developed below around twelve key points/sections that include A. Gathering/using all the available information about one's samples, B. Examining (plotting) the raw data, including looking for outliers and other irregularities in the data, C. Developing a general strategy for the chemometrics/informatics analysis, D. Preprocessing the data, E. Knowing where to start a chemometrics/informatics analysis, including identifying outliers or unexpected features in the data, F. Determining the number of abstract factors to keep in a model, G. Returning to the original data after the informatics analysis to confirm findings, H. Performing MCR, I. Peak fitting the MCR factors, J. Identifying intermediates in MCR analyses, K. Performing cluster analysis, and L. Knowing how/where to start doing chemometrics/informatics in one's technical work. These topics have been organized into a flowchart (Figure 1). In this approach, we recommend that PCA and perhaps a summary statistic first be performed on data sets. As suggested by the dashed lines in the flowchart, we believe that chemometrics/informatics analyses should always point one back to the original data. By returning to the original data, chemometrics/informatics predictions can be confirmed, and the original data may be better understood, dissected, and reconsidered so that more correct and refined chemometrics/informatics analyses can then be undertaken.

Certainly, even with a flowchart to help, the large amount of information in this work may seem overwhelming. That is, someone new to chemometrics/informatics may wish to apply these methods in their work, but be put off by all of the new vocabulary, concepts, and techniques in this paper and the subsequent one. Does one really have to

master all these concepts and methods to be able to do chemometrics/informatics, or is there an easier way? We think there is an easier way. Of course, we believe that (i) all the methods described in this work are important, where each has strengths that let it solve certain problems better or more conveniently than the others, and (ii) there is value in probing data sets with different statistical/mathematic tools because the results from these methods can reinforce each other. Nevertheless, in our opinion, those who wish to most quickly benefit from chemometrics/informatics in their XPS analyses should focus on MCR, first reading (and following) Sections A and B of the Results and Discussion below and then skipping to the sections on MCR. The other sections of this document and the information in the sequel to this paper can then be referred to as needed. In our opinion, not only do the most exciting and important results in this study come from MCR, MCR is easier to apply than PCA and its results are generally more intuitive. For example, spectra taken under identical conditions do not, in general, need to be preprocessed prior to MCR. In contrast, some form of preprocessing is required before most PCA analyses, and it is not always clear what that best preprocessing approach is. MCR factors are also much easier to interpret than PCA loadings because they generally look like (and very often represent) real spectra. In addition, while PCA is often used to estimate the number of factors that are needed in an MCR analysis, one can do this with MCR itself by (i) looking at the amount of variance captured by the different MCR factors (in a good model, the number of factors that are kept will generally account for most of the variance in the data set), (ii) examining the abstract factors to see where they no longer show meaningful structure, (iii) examining trajectories of the scores to see where they become overly noisy, (iv) perhaps reconstructing spectra from one's data set

with MCR factors as is done in Figures 13 – 15 in the companion to this paper for PCA, and (v) using what one knows about one's sample to determine the appropriate number of MCR factors to keep/expect. While our view may not be shared by all chemometricians/data scientists, we believe that MCR is the most powerful tool discussed in this work for analyzing many large XPS data, and that if one were to learn and apply only one of these techniques, it should be MCR. However, to be effective in this space in the long run, one should become familiar with at least PCA (and the preprocessing methods associated with it), and, in time, with other chemometrics/informatics methods as well.

The following are additional recommendations/caveats associated with this paper. First, as has been noted, this work has a companion guide that applies the principles and tools discussed in this work to two rather large XPS data sets.[31] We believe it will be helpful to those who wish to see the practical application of EDA methods to XPS data. Second, because the chemometrics/informatics methods employed in this study have been reviewed and discussed many times in the literature,[4, 5, 32-37] we do not provide a theory section for this paper. Third, while we have tried to focus on chemometrics/informatics methods and accompanying preprocessing methods that are proven and effective, there are other methods that may be considered for XPS data analysis – we have not covered all possible EDA methods in this guide. Fourth, a glossary listing the terms in this paper is provided at the end of this work.

We now attempt to answer a question that some readers may have about doing chemometrics/informatics, which is: What can I actually get out of

chemometrics/informatics analyses? Some of these benefits (and also limitations) of these tools are as follows:

1. *Better deal with large amounts of information*. For decades, a trend in analytical instrumentation has been for it to produce more information, spectra, and data. There is every reason to expect that this trend will continue. However, it can be difficult to fully understand and explore large amounts of information by, for example, comparing individual spectra to each other. Chemometrics/informatics methods provide a rational, accepted way of sifting through large amounts of information.
2. *Find outliers in data sets*. As shown in the companion guide to this one,[31] chemometrics/informatics methods are good a finding outliers in data sets. Of course, outliers are easy to spot in small data sets, but may be more difficult to identify in larger ones.
3. *Find irregularities in data sets*. As shown in the companion guide to this one,[31] chemometrics/informatics methods are good at identifying irregularities in data sets.
4. *Group/identify similar samples*. It can be tedious to group/organize the spectra in large data sets in a spectrum-by-spectrum fashion, i.e., manually. In contrast, chemometrics/informatics methods are generally very good at grouping similar spectra/data sets together – identifying which samples are more similar and which are more different in a data set. These grouping are not always obvious in the raw data. In a recent study,[5] we showed that chemometrics/informatics methods group similar Fourier transform infrared (FTIR) spectroscopy and XPS spectra.

5. *Better analyze hyperspectral images*. Hyperspectral images are increasingly collected by analytical techniques like XPS, ToF-SIMS, low energy ion scattering (LEIS), FTIR, and Raman. These data sets can be enormous, containing hundreds, thousands, or even tens of thousands of spectra. Chemometrics/informatics EDA methods can effectively probe these images, identifying regions that are chemically similar and different.[38]
6. *Better combine data from multiple analytical techniques*. Because there is no single surface or material analytical technique that can fully characterize a material, surfaces and materials are often characterized with multiple analytical methods.[39] Indeed, chemometrics/informatics methods provide a natural way of combining and comparing all this information. For example, chemometrics/informatics methods allow spectra from multiple methods to be joined together (concatenated) and then analyzed. In addition, it is common for different analytical methods to yield different bits of information about samples. For example, spectroscopic ellipsometry may yield a film's thickness and refractive index at a particular wavelength, XPS might give the ratio of two signals from two different elements, atomic force microscopy (AFM) might provide the film roughness and/or a step height value, low energy ion scattering (LEIS) may provide the surface concentration of a particular element, and contact angle goniometry (wetting) may provide the degree of surface hydrophobicity or hydrophilicity in the form of a water contact angle. Chemometrics/informatics methods provide a natural way of comparing all of this data. Here, the different pieces of analytical data from a sample can be made into a vector, and the EDA

methods discussed in this guide (or others) can then be used to analyze these vectors, where this analysis will reveal which samples are both most similar and most different. This general approach is also mentioned below in the section on preprocessing (autoscaling).

7. *Confirm what is expected*. Chemometrics/informatics methods allow one to either confirm that there are only minimal differences between samples, or that significant differences exist between them.
8. *Simplify the data.* Chemometrics/informatics methods can simplify data sets. For example, years ago, some of us analyzed some ToF-SIMS hyperspectral images with cluster analysis. Patterns in the data were more easily revealed in the images that resulted from the data being forced into two, three, or four clusters.[40]
9. *Identify the underlying chemistry, e.g., intermediates, in data sets*. In the companion guide to this one,[31] we show for the first time that a chemometrics/informatics analysis (MCR) can reveal intermediates in the degradation of a material. We believe these results constitute a 'killer app' for chemometrics/informatics analyses of XPS data sets.
10. *Explore data at a deeper level*. We have almost always found surprises in the large data sets we have analyzed using chememetrics/informatics methods.
11. *Not every data set needs a chemometrics/informatics analysis*. Chemometrics/informatics analyses do not need to be applied to every data set. Small numbers of spectra usually do not need chemometrics/informatics analyses. Even for larger data sets, a simple analysis may be all that is required to obtain the information that is desired about a material. For example, one may only need to

know whether a certain element is present (or not present) at certain locations at a surface. A simple, false-color map of the intensity of a peak from this element from a hyperspectral XPS image may answer this question.

This paper is an introduction to the concepts and terminology associated with chemometrics/informatics as they may be applied to XPS data analysis. It should be useful to those who are unfamiliar with these concepts; those who may be unfamiliar with chemometrics/informatics may wish to read this paper before looking at the companion paper to this one. However, these papers can also be read together – the sections in the two papers, as numbered, refer to the same concepts, where the first paper is more conceptual and the second shows the implementation of the concepts. Of course, the best way to become familiar with chemometrics/informatics methods is to use them. Accordingly, the data analyzed in part 2 has been made available for readers to download and analyze on their own, allowing them to cross check their analyses with those in the companion to this work.

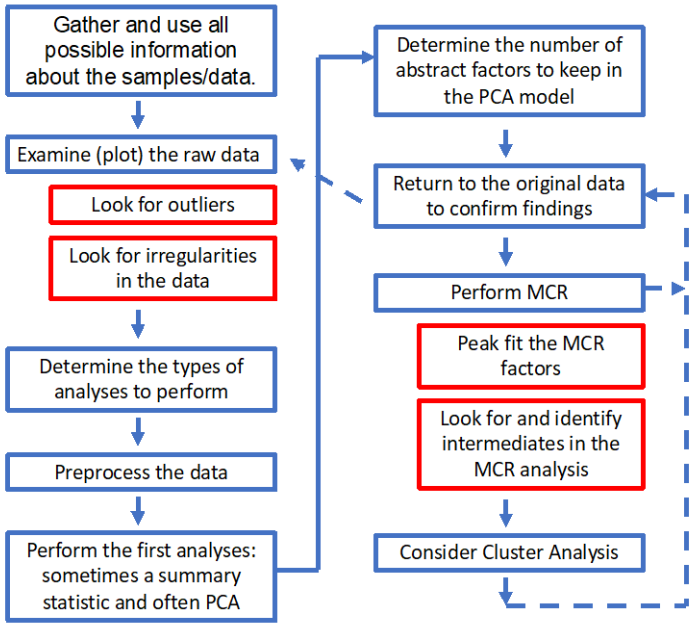

Figure 1. Flowchart of the topics covered in this work (blue boxes). The red boxes indicate important subtopics.

II. Results and Discussion

This paper is a conceptual guide to performing chemometrics/informatics analyses of XPS data sets. The following subsections cover concepts and/or steps that should be followed in performing these analyses: A. Gathering/using all the available information about one's samples, B. Examining (plotting) the raw data, C. Developing a general strategy for the chemometrics/informatics analysis, D. Preprocessing the data, E. Knowing where to start a chemometrics/informatics analysis, including identifying outliers or unexpected features in data sets, F. Determining the number of abstract factors to keep in a model, G. Returning to the original data after a chemometrics/informatics analysis to confirm findings, H. Performing multivariate curve resolution (MCR), I. Peak fitting the MCR factors, J. Identifying intermediates in MCR analyses, K. Performing

cluster analysis, and L. Knowing how to start doing chemometrics/informatics in one's work. The same subsections are used in the companion paper to this one, where, again, this paper focuses on the more foundational aspects of the concepts in question and the other paper shows their implementation. Emphasis in both papers is placed on MCR as a powerful, intuitive, factor-based method.

A. Gather/use all the information you have about your samples

All of the information that is available about a sample should be considered in a chemometrics/informatics analysis of it; chemometrics/informatics analyses should be performed holistically. The following information is often available about materials that are analyzed – it should be considered in a chemometrics/informatics analysis.

- One often has quite a bit of chemical information about the samples and materials one is analyzing. For example, one generally knows which substrate was used in a deposition and the chemical and elemental natures of the materials that were deposited on it. Thus, the chemical structures and elemental natures of the polymers, molecules, targets, precursors, etc. used to make one's surfaces and materials should be looked up and examined, where one should pay special attention to the oxidation states of the elements in these materials.[18] One should not propose a composition or chemical identify for a material unless the necessary elements are both present in appropriate quantities/concentrations and in their appropriate oxidation states.
- It is common for surfaces to be characterized by multiple analytical techniques. Some of us have drawn the analogy between 'the fable of the blind men and the

elephant' and surface analysis[39] – a high-level understanding of surfaces and materials often only comes as a result of employing multiple analytical techniques. This other information should be considered in a chemometrics/informatics analysis. For example, ToF-SIMS will often give significant clues about a sample's elemental and molecular composition,[41] low energy ion scattering (LEIS) reveals the atomic compositions of the outermost atomic layers of materials,[42] contact angle goniometry (wetting) quantifies a material's hydrophobicity or hydrophobicity, spectroscopic ellipsometry (SE) can yield film thicknesses, roughnesses, and film and substrate optical constants,[43] various atomic force microscopy (AFM) modes, e.g., phase images, provide direct or indirect chemical information about surfaces.[44, 45]

- Additional information in XPS analyses themselves is often overlooked. For example, were unexpected elements identified in a survey scan? Do the valence band regions or the Auger peaks in the survey spectra provide additional clues about the chemical nature of the material? Are the baselines (their rises or drops) consistent with the structure proposed for a material, e.g., buried materials usually show significant increases in their baselines on the high binding energy sides of their peaks.

B. Examine (plot) the raw data

There may be a temptation for those new to chemometrics/informatics to begin an analysis by entering data into a software package and immediately performing advanced analyses/calculations on it. The results of these efforts may be complicated-looking

graphs. This approach may be fine if one understands both one's data and the chemometrics/informatics methods one applies. However, for EDA, it is generally better to begin by examining the raw data itself. Sometimes, simply plotting and inspecting a data set may provide enough information about it to answer the questions at hand – chemometrics/informatics analyses are not needed for every data set. For example, in industry, one may simply be looking for an answer to a question, like whether an impurity is present at a surface or whether a sample is similar to previously made ones, i.e., whether it meets spec – in these cases, probing a data set to its limits, e.g., by multiple chemometrics/informatics methods, may be a waste of resources. However, the larger or more complicated a data set is, the more likely it is that advanced statistical methods will be necessary to fully understand it. We have almost always found that chemometrics/informatics analyses of large data sets yield some surprises.

One may begin a chemometrics/informatics investigation of a data set by plotting the data on top of itself in an overlay plot. Overlay plots provide a preliminary, big picture, view of a data set. They reveal whether the sample/spectra are changing – if the overlaid spectra look like a single spectrum (to within the noise on the data), then no changes are occurring and no additional analysis may be needed of the data set! However, considerable changes in these spectra may be observed. Such changes in XPS spectra may be an indication of sample degradation or charging. They often suggest that further examination of the data set is appropriate through a chemometrics/informatics analysis. However, overlay plots do not naturally reveal the order in which spectra are collected, and an overlay plot may be confusing or hide features in a data set when large numbers of spectra are overlaid/plotted together (see, for example, Figure 2a). Waterfall

plots show spectra in a side-by-side, sequential fashion. For example, the waterfall plot in Figure 2b of the data in Figure 2a shows the changes in the spectra as a function of scan number. Because of the more three-dimensional nature of waterfall plots, it can be advantageous to view them from different angles. Figure 2c shows the 'low binding energy' view of the data in Figure 2b. In this case, both the high and low binding energy views are useful for understanding the changes taking place in the data – changes in the data are taking place from both perspectives. We emphasize again that not every data set needs a thorough chemometrics/informatics analysis. In some cases, overlay or waterfall plots may provide a sufficient amount of insight into a data set to conclude an analysis.

While the spectra in the overlay plot in Figure 2a are colored, there are so many of them that it is difficult to see any trends in them, although the arrow on the plot helps the reader understand how the spectra are changing. The C 1s spectra in the tartaric acid data set in Figure 3 are colored more effectively. That is, the color scale on the left of the plot shows that the violet-colored spectra were collected first and the green-colored spectra last. Accordingly, it is clearer in the presentation of the data that a low-energy, reduced carbon peak around 285 eV grows in during data collection.

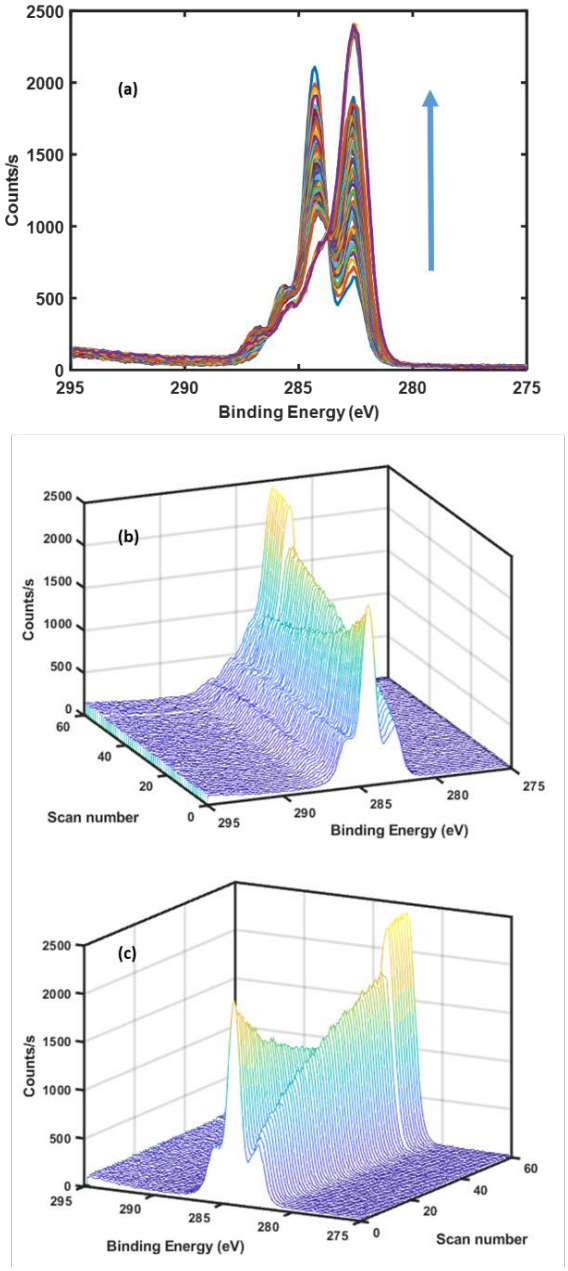

Figure 2. (a) Overlay plot, and (b – c) waterfall plots of 60 C 1s narrow scans from an XPS analysis of cellulose, where (b) and (c) show different views of the same data. The arrow in (a) indicates the order in which the data were collected.

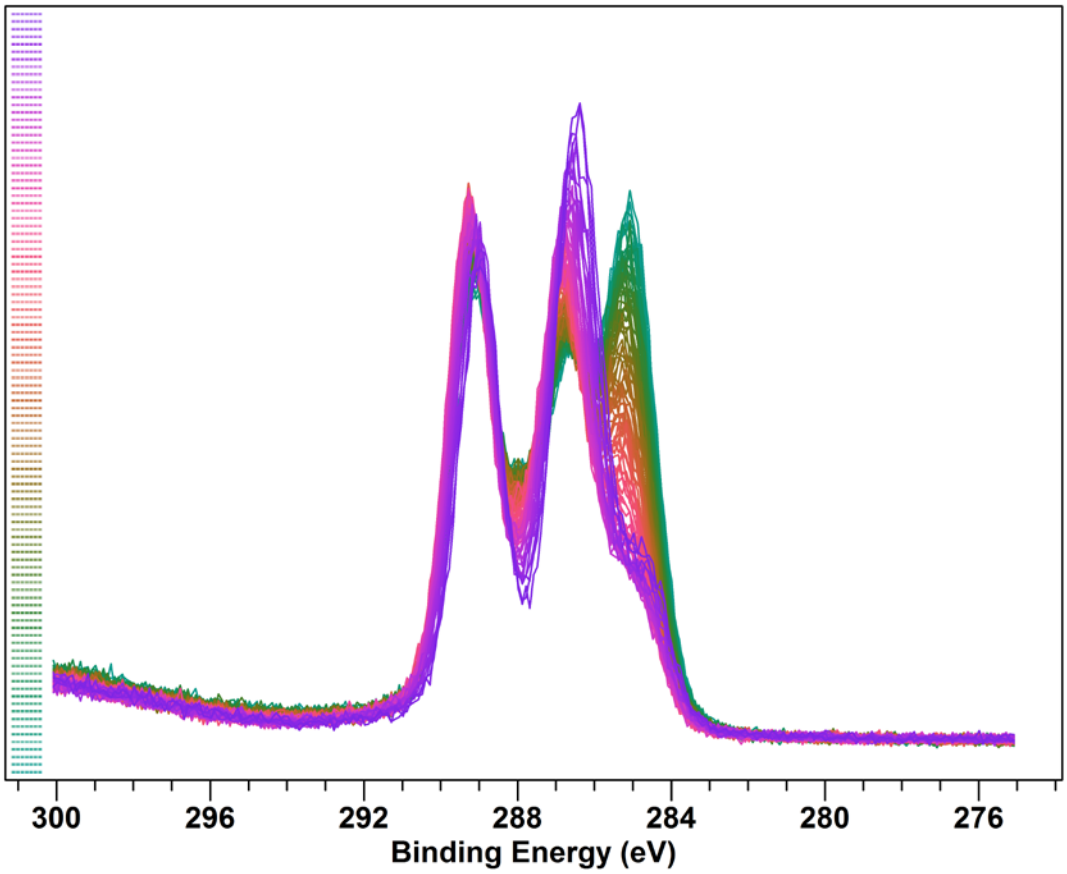

Figure 3. Overlay plot of the C 1s spectra from the tartaric acid data set, where the order in which the spectra were collected is indicated by the colors of the spectra. The scale on the left of the plot indicates that the violet-colored spectra were collected first and the green-colored spectra were collected last.

Unlike Auger spectra in Auger electron spectroscopy (AES), which are often shown as derivative spectra because of the high background in this technique (differentiation removes a constant signal/offset in a spectrum), photoemission spectra are not usually differentiated. However, one can consider this possibility. Figure 4 shows the first and last C 1s and O 1s narrow scans from the cellulose data set shown in Figure

2 and their derivatives. The derivatives of a the Gaussian-like O 1s peaks essentially produces two new peaks – a lobe that is positive (above the x-axis) and another that is negative (below the x-axis) (see Figure 4d). That is, differentiation essentially doubles the number of 'curves' in the signal – it adds complexity that can be used to identify and distinguish between spectra. For example, the derivatives of the C 1s spectra in Figure 4a (in Figure 4b) are more complex than the original spectra. For a Gaussian-like signal, the point at which the middle of its derivative spectrum crosses the x-axis corresponds to the maximum in the original peak. Indeed, the relatively small difference between the peak positions of the two O 1s spectra in Figure 4c is more easily seen in the x-axis crossing points of their derivative spectra in Figure 4d. However, differentiation (especially with finite differences) can increase the noise in a spectrum. Savitzky-Golay (SG) smoothing and differentiation filters,[46-50] which act via numerical convolution, reduce this problem. (An SG filter was used to create Figure 4.) Another possibility for differentiating spectra is to first fit the data with a high-order polynomial and to then differentiate the polynomial. This approach, which has been used to calculate the so-called D parameter of carbon Auger peaks,[51] yields noise-free derivative spectra.

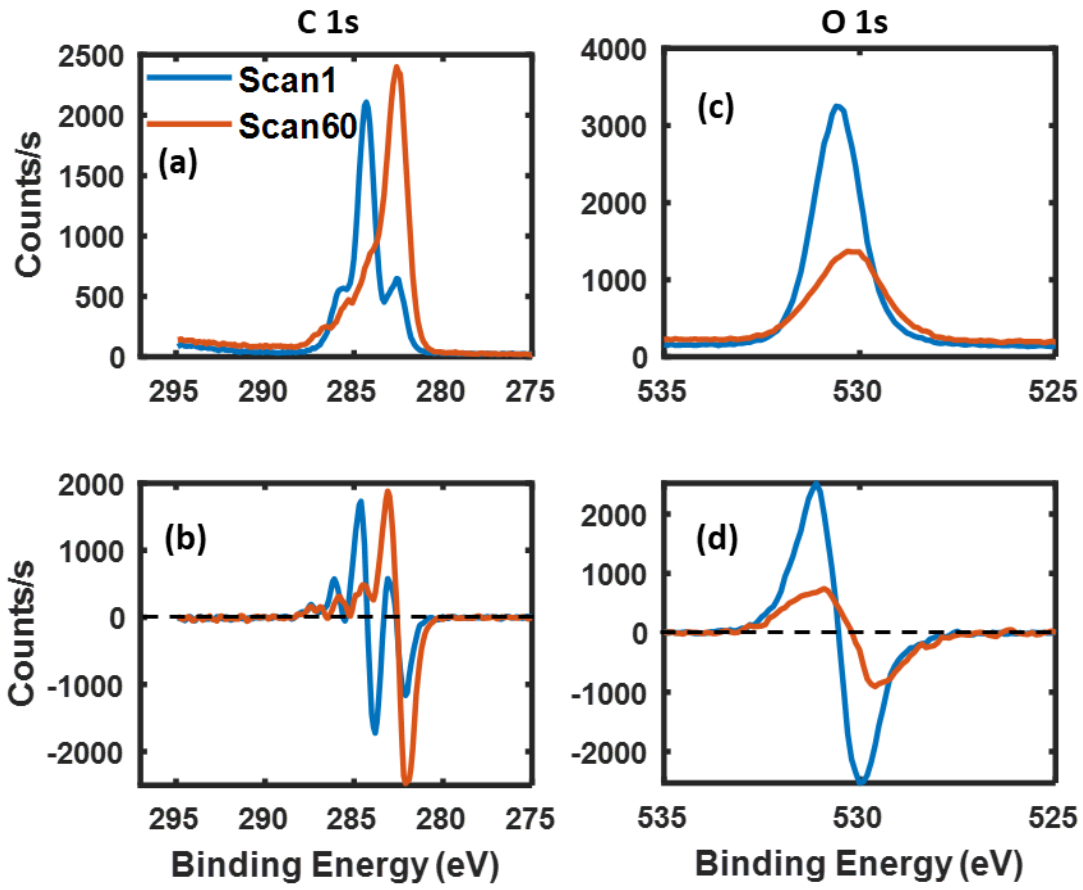

Figure 4. The first and the last (60th) undifferentiated (a and c) and differentiated/derivative (b and d) C 1s and O 1s narrow scans of the cellulose data set.

C. Develop a general strategy for the chemometrics/informatics analysis

It can be challenging for a beginner in chemometrics/informatics to know which analyses/tools to apply to a data set. Accordingly, if an analyst is unsure how to proceed, we recommend the approach in the flowchart in Figure 1. Of course, there are other chemometrics/informatics tools and protocols that the analyst may learn or develop. However, one new to this area may wish to follow the approach outlined in Figure 1 because it is based on three well-established, proven techniques (PCA, MCR, and cluster

analysis). We have also found summary statistics to be helpful in the initial evaluation of our data.

D. Preprocess the data (when necessary)

Some sort of preprocessing is applied before most chemometrics/informatics analyses. ‘Data preprocessing’, or just ‘preprocessing’, refers to any mathematical treatment of a data set prior to a chemometrics/informatics analysis. The objective of data preprocessing is to suppress signal that is not of interest and bring signal that is of interest to the forefront. Appropriate preprocessing depends on the objectives of the analysis and how the signal manifests itself. Good practice in data preprocessing uses knowledge of the science of the measurements and mathematics of the preprocessing techniques to enhance the sum-of-squares (SSQ) of the signal of interest, e.g., data might be preprocessed to remove artifacts and/or to improve visualization. That is, the chemometrics/informatics tools used to analyze multivariate measurements, such as XPS depth profiles, are often focused on maximizing the capture of SSQ or minimizing model residual SSQ. For example, principal components analysis (PCA) is designed to find factors (loadings) that maximize the capture of SSQ while partial least squares regression is used to find a factor (regression vector) that minimizes the estimation error for a predictand. The results of these analyses can often be improved and clarified by data preprocessing. Some of the more common ways to preprocess XPS and ToF-SIMS spectra are now described.[52]

1. No preprocessing

No preprocessing is often employed when spectra are collected under essentially identical conditions, e.g., in a depth profile, image, *operando*, or damage study. No preprocessing may also be useful when determining the number of principal components (PCs) with which to represent a data set (see Section G below), and when using an abstract factor-based approach to denoise/smooth spectra. However, because PCA is essentially a rotation of a coordinate system, where the spectra act as single points in a hyperspace, when no preprocessing is performed, the first PC (PC1) can account for a disproportionately large amount of the variance in the data set, and the chemical variation in the data set may not be well correlated with the scores (projections) on this PC (and even higher PCs). That is, when no preprocessing is applied, PC1 simply points in the direction of the 'cloud' of data points, but does not usually correlate strongly with the chemical information in the data set.

2. Normalization

Spectra are often normalized using the so-called '1-norm', which consists of dividing each data point in a spectrum by the sum of the data points in that spectrum. This form of pre-processing makes sense for spectra that only have positive values. Normalization is often used to account for different data acquisition times or conditions. For example, two high-quality (low noise) XPS spectra taken with different acquisition times may contain the same information, but have significantly different numbers of counts. Normalization would put them on equal footing/reveal their equivalence. Other less common forms of normalization (at least for XPS spectra) may also be considered.

3. Baselining

An important part of XPS peak fitting is defining the baseline/background of the narrow scan under analysis, where the area fit and used for quantitation in XPS is almost always taken as the that between the peak envelope and a background. The three most common backgrounds used in XPS data analysis are the linear, Shirley, and Tougaard backgrounds.[22] Baselining refers to removing a baseline signal from a spectrum. In general, baselines should not be removed from peak-fit XPS spectra that are shown in the literature – it is important to show the original data. However, baselining of XPS spectra may be appropriate in some chemometrics analyses.

4. Charge correction/peak shifting

XPS spectra and peak positions may be shifted to account for sample charging. Appendix 1 of the companion paper to this one[31] discusses this issue for the tartartic acid data set.

5. Variable selection

In variable selection, one focuses on certain peaks/signals that are of interest. Accordingly, variable selection can be used to remove either regions of noise or measurement artifacts from spectra. For example, it is common in ToF-SIMS data analysis to select and integrate the peaks in a set of spectra while ignoring the noise/baseline between them. These ToF-SIMS peak areas are then used to represent the entire spectra, i.e., they may be preprocessed and analyzed by one of the chemometrics/informatics tools discussed herein. XPS often uses a similar approach of selecting and integrating a subset of the peaks from a set of spectra (either the narrow scans or regions in a survey scan). In some cases, these peaks are ratioed.

6. Mean centering

Mean centering is often employed before PCA. In mean centering, the mean of all the values at each x-axis value in the spectra, e.g., binding energy for XPS, is subtracted from each value. In essence, mean centering moves the center of the data points (spectra) to the origin. Mean centering often allows one to see the variation between spectra more easily. For example, when the data are not mean centered in PCA, PC1 points in the average direction of the spectra. In these cases, PC1 often accounts for a very large fraction of the variance in the data set, where it may or may not correlate with meaningful chemical changes in the data set. In contrast, PC1 points in the direction of greatest change/variance in the data set when it is mean centered. Disadvantages of mean centering and autoscaling (see below) are that they (i) remove information from a data set, and (ii) add complexity to the data in the sense that mean centered spectra generally have negative peaks that are more difficult to interpret than the peaks in traditional spectra. As an example of mean centering, Figure 5 shows three C 1s and O 1s spectra from the beginning, middle, and end of the cellulose data set before (Figure 5a,d) and after mean centering (Figure 5b,e). In each case, the middle spectrum, which is more or less the average of the other two spectra, has values close to zero after mean centering.

7. Autoscaling

Autoscaling consists of mean centering each column of data points in a data set and then dividing each data point by the standard deviation of all the data points in the column. Autoscaling gives all the variables in a data set equal statistical weight (the same variance of unity). Autoscaling is useful when data from different techniques are combined in an analysis. For example, one might compare surfaces/thin films based on their advancing water contact angles, ellipsometric film thicknesses, and selected XPS,

ToF-SIMS, and/or LEIS peak areas, or, alternatively, some measure of the peaks like their equivalent widths.[53, 54] Autoscaling allows the information from these different techniques to be more fairly combined. Otherwise, a chemometrics analysis that minimizes SSQ will favor the variables with large averages. Autoscaling is not generally recommended for XPS narrow scans because it puts the signal and noise on equal footing. For example, Figure 5c,f show autoscaled versions of the raw spectra in Figure 5a,d. While one can identify the positions of the original peaks, at least some of the noise at the baseline has been enlarged to the point that these spectra/data sets would be of limited value in a chemometrics/informatics analysis. Autoscaling is similarly problematic in ToF-SIMS when applied to whole spectra. In summary, autoscaling is generally more appropriate for data sets that consist of integrated peak areas and specific measurements, while mean centering is generally more appropriate when complete XPS or ToF-SIMS spectra are analyzed.[4]

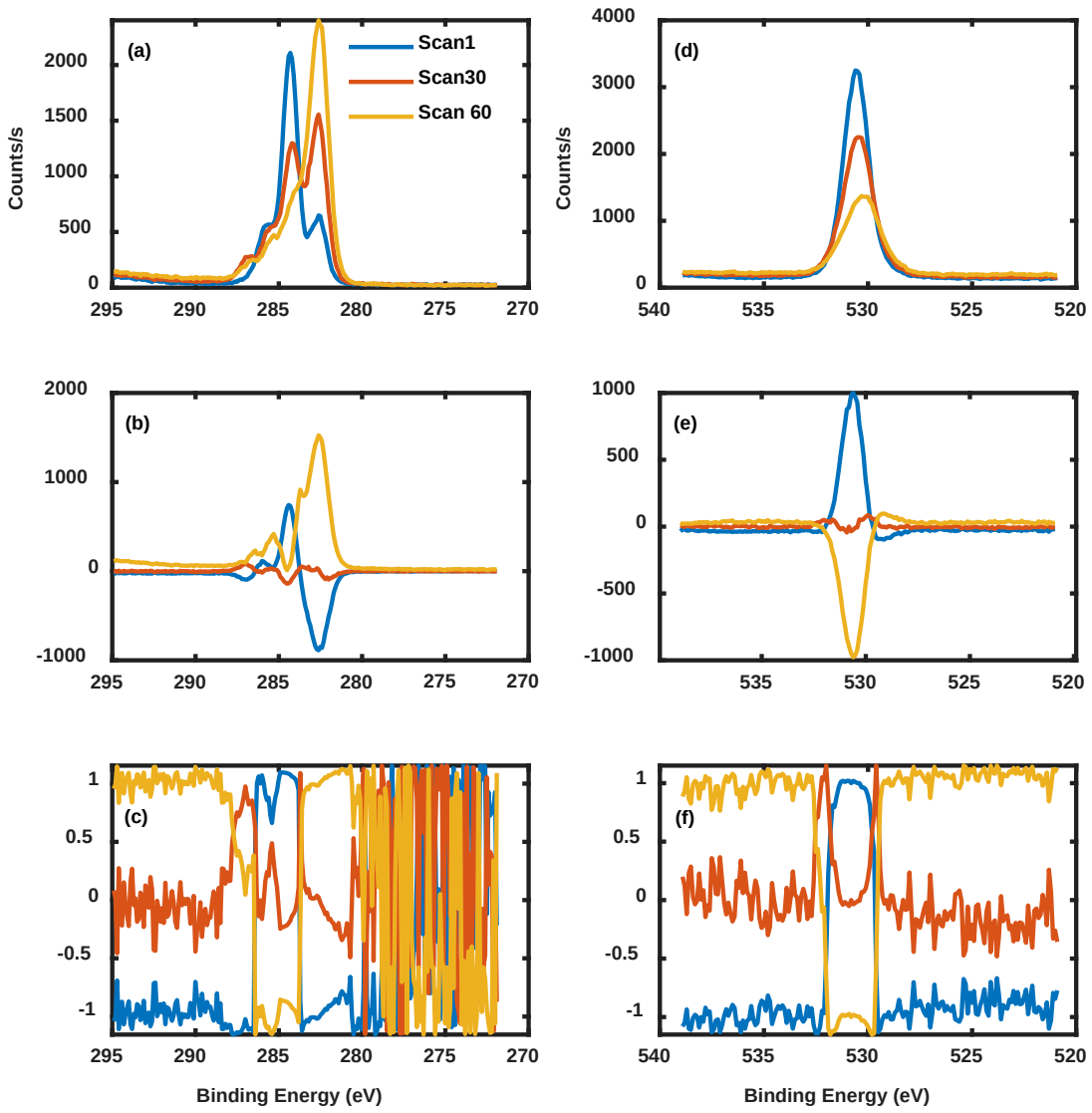

Figure 5. Raw (a) C 1s and (d) O 1s spectra (scans 1, 30, and 60) from the cellulose data set. (b, e) These spectra after mean centering. (c, f) These spectra after autoscaling.

8. Poisson scaling

Poisson scaling consists of dividing each instance of a variable by the square root of the mean of that variable. It is often appropriate for pulse-counted data. As shown in Figure 6, Poisson scaling can accentuate smaller features in a spectrum. Poisson scaling, autoscaling, and variance scaling are forms of variable weighting, which can be used to

suppress large peaks that have a large SSQ and allow smaller peaks to have greater influence on a model. Care must be taken here to avoid dividing by zero (or by a very small number) so as to not add SSQ due to noise. To avoid this, a small offset is often added to the values before division.

9. Concatenation

Concatenation consists of joining together/combining multiple spectra. While most data analysis in surface analysis and analytical chemistry does not employ concatenated data, the concatenation of spectra can be very helpful in chemometrics/informatics analyses. Concatenation ensures that the variation in coupled spectra is simultaneously considered, i.e., that a group of related narrow scans are considered as a unit. However, one must be cautious in concatenating XPS spectra. Because different narrow scans come from photoelectrons with different kinetic energies, they sample different depths in a material. Accordingly, it is generally better to reserve concatenation for spectra with similar binding energies. The C 1s and O 1s narrow scans are relatively close in binding energy.

10. Spectral differentiation

Spectral differentiation is another form of data preprocessing (see discussion above). While we mention it here for completeness, we do not strongly recommend it for XPS spectra.

11. Smoothing

XPS data may be smoothed, which can be useful for revealing the structure of noisy spectra. However, as discussed below, we generally discourage the smoothing of XPS spectra.

12. Applying multiple preprocessing methods

It is not uncommon for multiple forms of preprocessing to be applied to data sets. For example, spectra might be normalized, concatenated, and finally mean centered prior to a chemometrics/informatics analysis.

13. Preprocessing of spectra acquired under the same conditions

We end this section with some recommendations for preprocessing XPS data acquired either under the same conditions, which is generally the case for the large data sets we have been discussing here, or under different experimental conditions or with different instruments. XPS data acquired under the same experimental conditions often do not need to be preprocessed in any way prior to PRE, MCR, and cluster analysis. However, it is usually advisable to mean center them prior to PCA. Data obtained from different XPS instruments and/or taken under different experimental conditions might be normalized prior to PRE, MCR and cluster analysis, and preprocessed by normalization and mean centering prior to PCA.

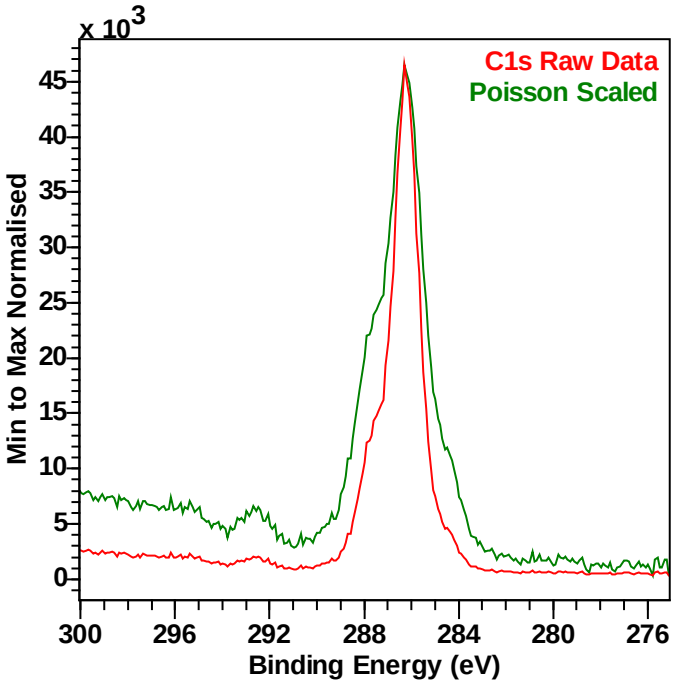

Figure 6. Raw C 1s spectrum of sucrose (red/bottom), and the same spectrum preprocessed by Poisson scaling (green/top). See Section D8 for details on Poisson scaling.

E. Where to start an informatics analysis, including identifying outliers or unexpected features in data sets

A summary statistic is a single number that characterizes a spectrum. Summary statistic analyses are quite easy to perform and can be helpful in identifying trends in data/spectra. We recommend that a summary statistic analysis be performed early in a data analysis. Common examples of summary statistics include the mean, standard deviation, and range. Some of us recently showed the usefulness of the PRE summary statistic (a form of Shannon's entropy) in understanding series of XPS spectra, and also for probing spectra from other techniques, including ToF-SIMS.[4, 55 5, 35, 56, 57] PRE often clusters and reveals trends in data; its results are often similar to PCA scores plots. (In the companion to this paper, Figure 6 shows PRE results and Figures 7 – 9 show PCA scores

plots.) We are early proponents/adopters of summary statistics – while we believe that summary statistics will be used to a greater extent in data analysis, widespread adoption of summary statistics in chemometrics/informatics has not yet occurred.

As single numbers, summary statistics are limited in the amount of information they can provide about a spectrum. Accordingly, we next recommend that a whole-spectrum analysis be performed. The most common, and probably most important, of these EDA methods is PCA. In essence, in PCA, the spectra are represented as single points in a hyperspace, and the coordinate system for these spectra is rotated to capture the largest amount of variance possible in the data set. For XPS spectra, each original axis corresponds to a different binding energy at which data were collected. The new axes (principal components, PCs) created by the rotation of the coordinate system are orthogonal to each other and arranged to account for decreasing amounts of variation in a data set. The projections of the data points (spectra) on the new axes (PCs) are called scores. The loadings are the contributions of the original axes to the new ones (PCs). It is not always clear how many PCs (abstract factors) one should keep in the PCA of a data set. This issue, which is shared by MCR, is discussed in more detail below. However, even though most of the variance in a data set may be captured by a few PCs, the higher PCs sometimes contain useful, and even important, information about a data set.

PCA scores plots are used to represent the relationships, i.e., similarities and differences, between spectra. Figure 7 is an example of a scores plot obtained from the PCA of the concatenated C 1s and O 1s narrow scans from the cellulose data set. It shows the scores (projections) of data points (spectra) on the new axes (PCs). The relationships between data points (spectra) in scores plots are often revealed by the presence of

outliers, trajectories of data points, groupings of data points, and other irregularities. Each of these possibilities is discussed below. These features may be present in XPS spectra for the following reasons:

- Some XPS practitioners begin their depth profiles with the sputter source turned off, which may result in the first few scans of a depth profile being different from those that follow.
- The first few spectra in a series of spectra may be different from those that follow if the electronics of an instrument are not warmed up.
- Desorption of powder materials may occur during their analysis, which may cause gradual or abrupt changes in the samples.
- Samples may gradually charge as data is collected either through sample charging or material damage.[58, 59]

1. Outliers

All of the chemometrics/informatics tools described in this work can identify outliers in data sets. An outlier is a spectrum or data point (or other form of information) that is inconsistent with the rest of the data and perhaps even incorrect. The presence of outliers may complicate or confound data interpretation. Of course, strong justification must be present before outliers can be removed from data sets. For example, one may be justified in removing a spectrum from a data set if the equipment that produced it was malfunctioning or the person who prepared the sample failed to follow the appropriate protocol for its preparation. For transparency, the authors of a study should report if one or more outliers has been removed from a data set. Outlier identification is an early step in a chemometrics/informatics analysis. Outliers often appear in scores plots as ‘points’

that are in some way different/separated from those around them. The point by itself on the right side of Figure 7 may be an outlier. In general, one should return to the original data to confirm such an interpretation.

2. Trajectories

Figure 7 shows a series of continuously varying data points (spectra) that form a horseshoe-shaped trajectory. Such trajectories are observed in many PCA scores plots. They usually indicate that there is a steady change taking place in a series of spectra. The spectra in Figure 7 change because the material that produced them is steadily degrading.

3. Groupings of data points (spectra)

Sometimes scores plots show groupings of data points (spectra). These are usually interpreted to mean that the spectra in these groups are similar to each other. Such interpretations should be confirmed in the original data.

4. Other irregularities

An example of an irregularity in a PCA scores plot is the rather large break on the left side of the trajectory in Figure 7. This break is present because there was a break in data collection during which time another analysis was performed. After this analysis, acquisition of this data set resumed.

Finally, additional information may be added to PCA scores plots. Most scores plots do not have colored data points, as in Figure 7. However, the elapsed time of the analysis has been added to the scores plot in Figure 7 as the color of the data points. This approach allows additional information/another dimension to be rather easily added to a graph.

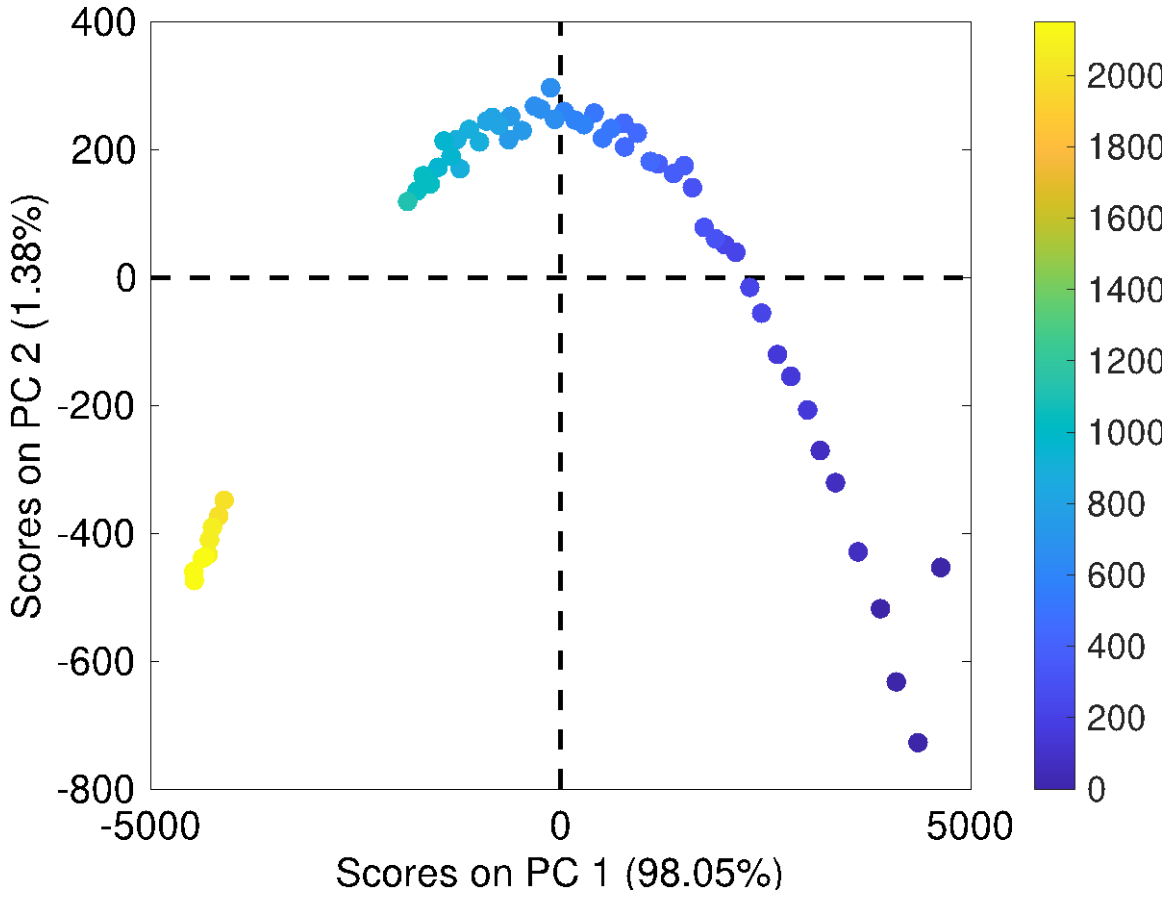

Figure 7. Two-dimensional PCA scores plots of the concatenated C 1s and O 1s narrow scans of the cellulose data set with the elapsed time shown as the color of the data points. In PCA, the spectra are, in essence, plotted as single points in a hyperspace, where the axes of this coordinate system are rotated to align with the greatest amount of variance in the data. The 'scores' of the data points (spectra) are the projections of the data points (spectra) on the new (rotated) axes, where these new axes as called 'principal components' or 'PCs'.

F. Determine the number of abstract factors to keep in a model

One of the challenges associated with PCA and MCR is determining the 'right' number of abstract factors to keep, i.e., the number that appropriately describes/captures the relevant variance in a data set. If too few abstract factors are kept, important variation/information in the data set will be omitted. If too many are kept, unnecessary noise will be included in the model. (Although overfitting, i.e., having too many abstract

factors in a model, is not always a problem in EDA, it can be quite problematic for models used for process monitoring and control.) While there is no simple formula or approach for determining the appropriate number of abstract factors to keep, there are accepted tools that can be used for this purpose. In this section, we describe three methods for determining the number of abstract factors that describe a data set: scree plots, cross-validation, and reconstructing the spectra from abstract factors. Section J describes yet another approach to this problem, which is to create models with increasing numbers of abstract factors and to then evaluate these models based on their chemical and statistical reasonableness. It is important to be familiar with, and to use, multiple methods for finding the best number of abstract factors that describe a data set. Indeed, in a study of different methods used to determine the number of abstract factors that describe data sets, Jackson[36] concluded that 'no one method is good for all data sets'.

Scree plots are plots of the variance captured by the series of ranked PCs, where, in general, the amount of variance captured per PC decreases as higher and higher PCs are considered. (For examples of scree plots, see Figures 10 and 11 in the companion paper to this one.) Scree plots may appear in different ways. For example, they may show the cumulative variance captured by the PCs or the log of the eigenvalues (the eigenvalues are proportional to the amount of variance captured by a PC). In a scree plot, e.g., where the log of the eigenvalues is plotted vs. the PC number, one often looks for a discontinuity or 'knee' in the plot where the slope of the results (as viewed from right to left) changes. This point is often taken as the number of PCs that describe the data set. Scree plots have pros and cons. That is, while they do not always provide clear guidance regarding the number of PCs to keep, they are usually easy to produce. Also, as noted

above, it is not uncommon for a small number of PCs to capture most of the variance in data set, but higher PCs to still contain meaningful information/variance (not just noise). Indeed, in some cases, the variance/variation in a data set captured by one or more of the higher PCs is the most useful information in an analysis.

Cross-validation methods are also commonly used to determine the number of PCs to keep in PCA. Cross-validation is a procedure in which subsets of the data are left out during the chemometrics/informatics modeling and then used to test/estimate the prediction error of the model. That is, a model built with a fraction of the data is applied to the data that has been left out. One then evaluates the ability of this model to describe the left-out data. Typically, each possible subset of the data is left out and modeled, but other cross-validation methods, such as with random subsets, are also used.

A more graphical approach for finding the number of abstract factors that describe a data set is to first perform PCA on the data and then to reconstruct the spectra from increasing numbers of PCs. Both the reconstructed spectra and the loadings (abstract factors) are then examined. In general, the number of abstract factors to keep is determined by the point at which the reconstructed spectra no longer change substantially when they are reconstructed with additional abstract factors, and also when the loadings cease to show chemically meaningful structure. (This approach is illustrated in Figures 12 – 14 of the companion paper to this one.) In general, one should *not* preprocess the data (spectra) before doing this, or only minimally preprocess it. If one were to mean center or autoscale spectra prior to this approach, one would create spectra with both positive and negative peaks – it is generally easier to recognize when spectra have been adequately

reconstructed when they look like real spectra. The companion guide to this one[31] shows the reconstructions of spectra from increasing numbers of abstract factors.

This paragraph contains five additional comments/observations about the reconstruction of data from abstract factors and abstract factors in general. First, this approach should be applied to different spectra in a data set because a given spectrum may be fortuitously described by the first few abstract factors, i.e., the number of abstract factors needed to reconstruct a given spectrum may not be representative of the number of abstract factors needed to reconstruct all the spectra. Second, abstract factors (PCs) in PCA may be hard to interpret because of negative peaks in them, which are a consequence of the forced orthogonality of the factors. MCR, which is discussed next, generally yields abstract factors that look like real spectra and are, therefore, easier to interpret. Third, if XPS spectra are corrected/shifted, e.g., to the maximum of one of their signals, fewer abstract factors will generally be needed to describe them. Fourth, preprocessing can influence the number of abstract factors that describe a data set. For example, mean centering a data set removes the average from it. Accordingly, one might expect that it would take one abstract factor (PC) less to describe such a data set. Fifth, the number of abstract factors that describe a data set modeled by PCA is about the same as the number needed to describe the data set modeled by MCR (these numbers of abstract factors usually agree to plus or minus one).

Finally, there are other reasons for reconstructing data from abstract factors. For example, one can denoise/smooth a spectrum by reconstructing it from a limited number of abstract factors. (Other approaches for smoothing spectra include Savitzky-Golay smooths,[48] Fourier analysis,[46, 60] and high-order polynomial smooths. In the Fourier

approach, the higher frequency components of a Fourier transform are excluded before the data are back transformed. This approach also allows specific frequency components to be excluded/removed from the spectra.) However, we discourage people from smoothing their XPS spectra prior to fitting/analyzing it because this form of preprocessing can distort the data. Nevertheless, series of spectra with high noise levels may benefit from denoising because high levels of noise obscure the underlying structure of the data, which may become more apparent when it is denoised. That is, smoothing may help determine the number of fit components/synthetic peaks to include in a fit, where an advantage of using a tool like PCA to reconstruct/smooth spectra is that it can receive input from a large number of spectra, which can allow it to better discriminate between signal and noise. However, if at all possible, one should endeavor to collect high quality data. One should not attempt to 'fix' poor quality/noisy results with chemometrics/informatics. Subtle features in data sets are best revealed (and believed by others) when the data are of high quality. As an aside, a related error, which is sometimes seen in the scientific literature, is for only the fit components or the sum of the fit components of a fit, and not the original data, to be shown in a figure – in some sense, the data are presented as completely smoothed/entirely noise free. Unfortunately, such a representation of the results provides the reader with no information about the quality of either the original data or the fit to it. The original XPS data should always be shown with any fit of it.

G. Return to the original data after a chemometrics/informatics analysis to confirm findings

It is easy to perform many different analyses on data sets with modern chemometrics/informatics software. However, the predictions and findings from these analyses should always be confirmed in the originally data; one should never stray too far from the original data in a chemometrics/informatics analysis. The second, companion paper to this one provides an example of confirming predictions made in chemometrics/informatics analyses in the original data (see Figure 15 of the companion paper).

H. Perform Multivariate Curve Resolution (MCR)

MCR has become popular among chemometricians as it offers various advantages over PCA. Indeed, because of the non-negativity constraints that are usually applied in MCR, MCR loadings/components have the appearance of real spectra, making them easier to interpret, while PCA loadings often have negative peaks. In addition, because MCR loadings are not forced to be orthogonal to each other, as they are in PCA, MCR loadings often reveal the true, underlying spectra of data sets. (Figure 16 of the companion paper to this one shows an MCR analysis of the cellulose data set.) MCR is often performed on unpreprocessed, or minimally preprocessed, data – preprocessing by mean centering or autoscaling would create negative peaks, which, again, are not allowed in the typical implementation of MCR. Of course, as in all chemometrics/informatics analyses, preprocessing may strongly affect MCR results. XPS narrow scans may be better analyzed by MCR when they are concatenated. For example, a more easily interpreted narrow scan, e.g., the C 1s, may be paired/concatenated with a less easily interpreted one, e.g., the O 1s. Variations/trends in the more easily interpreted narrow

scan may then be used to understand the less easily interpreted narrow scan. Differentiated spectra concatenated with undifferentiated spectra may be considered in MCR if its non-negativity constraints are relaxed.

In addition to the methods mentioned in Section G, one can determine the number of abstract factors that describe a data set in MCR by creating models from increasing numbers of abstract factors and then evaluating the chemical and statistical reasonableness of these models. Two aspects of the statistical reasonableness of a model are the amount of noise in its scores and loadings, and also the amount of structure (information that appears to be chemically meaningful) in its loadings. Higher abstract factors, i.e., those accounting for less variance in a data set, generally contain more noise than the earlier abstract factors. Accordingly, reconstructed spectra typically become noisier as more abstract factors are used to create them. Thus, one option for determining the number of abstract factors to keep in an MCR model is to increase the number of abstract factors in the model until (i) the abstract factors cease to have meaningful structure, (ii) the scores, e.g., trajectories in scores plots, are overly noisy, and/or (ii) the models become chemically unreasonable. Obviously, there is some subjectivity in this process.

The reader may ask why more than one chemometrics/informatics method is described in this guide. Isn't one method good enough? It is often advantageous to probe data sets with multiple chemometrics/informatics methods because the analyses can confirm and complement each other. That is, the results of a chemometrics/informatics analysis are more believable when different methods with different underlying mathematical bases yield the same results. For example, a significant anomaly suggested

in one chemometrics/informatics analysis ought to be present in other analyses. In addition, the groupings of spectra, e.g., in scores plots, suggested by the different chemometrics/informatics methods should usually be the same, or at least very similar. As a final example of using multiple chemometrics/informatics analyses to analyze data sets, note that PCA is often applied to data sets before MCR to get a sense for the number of abstract factors to keep in the subsequent MCR analysis.

I. Peak fit the MCR factors

We believe that, in this work, we are the first to observe that chemical information can be extracted from MCR factors of XPS data sets by peak fitting them. (See Figure 17 of the companion paper to this one for an example of peak fitting MCR factors.) Such fits should help reveal the chemical changes taking place in a material. We believe the C 1s narrow scan is a particularly good candidate for this type of peak fitting because:

(i) C 1s spectra are often fairly simple (they exhibit neither spin-orbit splitting (the signal originates from a 1s orbital) nor multiplet splitting (as do some metals in some oxidation states)), although conjugated organic materials often show shake-up signals,

(ii) They can often be fit with symmetric (not asymmetric) peaks,

(iii) When there is no peak tailing/asymmetry, the peak width for a given chemical state of carbon is usually fairly narrow,

(iv) The baselines below them are often relatively straightforward (many organic materials, e.g., many common polymers, are insulators),

(v) The range of chemical shifts for carbon is large (in contrast, oxygen shows much less chemical shifting, which often makes it harder to peak fit in a meaningful fashion).

However, one should be aware that MCR components may contain artifacts. For example, there may be small peaks or distortions that are not present in the spectra, e.g., the MCR components may not be well fit in places with typical Gaussian/Lorentzian/Voigt signals. Nevertheless, MCR is an extremely powerful tool for understanding series of spectra. Indeed, the possibility of artifacts, which are usually small, underscores the importance of utilizing all the information available in an analysis, i.e., from both the raw data and (ideally) multiple informatics analyses of it. That is, an artifact in one chemometrics/informatics analysis will probably not be present in the results of a different one.

J. Identify intermediates in MCR analyses

To the best of our knowledge, the companion paper to this one is the first time that intermediates have been observed by MCR in the degradation of a material during XPS (see Figure 19 in the companion paper to this one). That is, some of the MCR factors describe the data at intermediate stages of the analysis. These results showcase the power of MCR to reveal the underlying chemistry of data sets. We are not aware of another technique or approach that can extract this type of information from an analysis.

K. Perform cluster analysis

Cluster analysis is another widely used EDA method. Cluster analysis groups similar samples/spectra according to their distances in a multidimensional/multivariate space. Different measures of distance may be applied to the points (spectra) in a cluster analysis, where two of the most common of these distance measures are the Euclidean and Manhattan (city block) distances. The relationships between the spectra in a cluster analysis are often shown as a dendrogram, where a dendrogram shows spectra that are closer in a multidimensional space cluster as clustered together. For example, Figure 8 shows a dendrogram with five selected/color coded clusters. Figure 22 in the companion paper to this one shows the average spectra in each of these clusters. It was obtained from a cluster analysis of the C 1s spectra in the tartaric acid data set. it shows a significant amount of clustering of the spectra. Indeed, five groupings of similar spectra are identified in this cluster analysis, where, again, the spectra in these clusters are closer to each other in a multivariate space than they are to the spectra in the other clusters. Cluster analysis has advantages and disadvantages. It is relatively easy to apply and conceptually simpler than some other informatics methods. Indeed, it can be a good starting point for an EDA. It can also be useful for identifying outliers, which might appear as a cluster that only contains one sample (or a few samples), where these clusters are significantly different from the others. However, cluster analysis does not generally provide as much insight or information as MCR or PCA. Cluster analysis itself could lead to additional multivariate analyses and/or XPS peak fitting. For example, one might perform MCR or PCA on the spectra in a specific cluster. In addition, the average XPS spectrum of the data points (spectra) in a cluster might be peak fit.

L. How to start doing chemometrics/informatics

The reader who wishes to learn more about chemometrics/informatics and incorporate it into their technical work may wish to consider the following:

1. *Read and study the subject*. There are many useful guides, papers, and books[61] on chemometrics/informatics.

2. *Attend a short course*. Short courses on chemometrics/informatics are given by various companies and professional organizations. They can provide quite a bit of information about the subject in just a few days. Short courses can create useful connections to experts (the instructors), who may be willing to consult or collaborate on future projects. In our opinion, the best short courses have both theoretical and 'hands on' components, i.e., they both teach the theory and provide students with sample data sets that they fit/work up themselves with a software package.

3. *Collaborate with an expert*. Collaborating with an expert can be an excellent way to gain more knowledge about a field and avoid making mistakes in it.

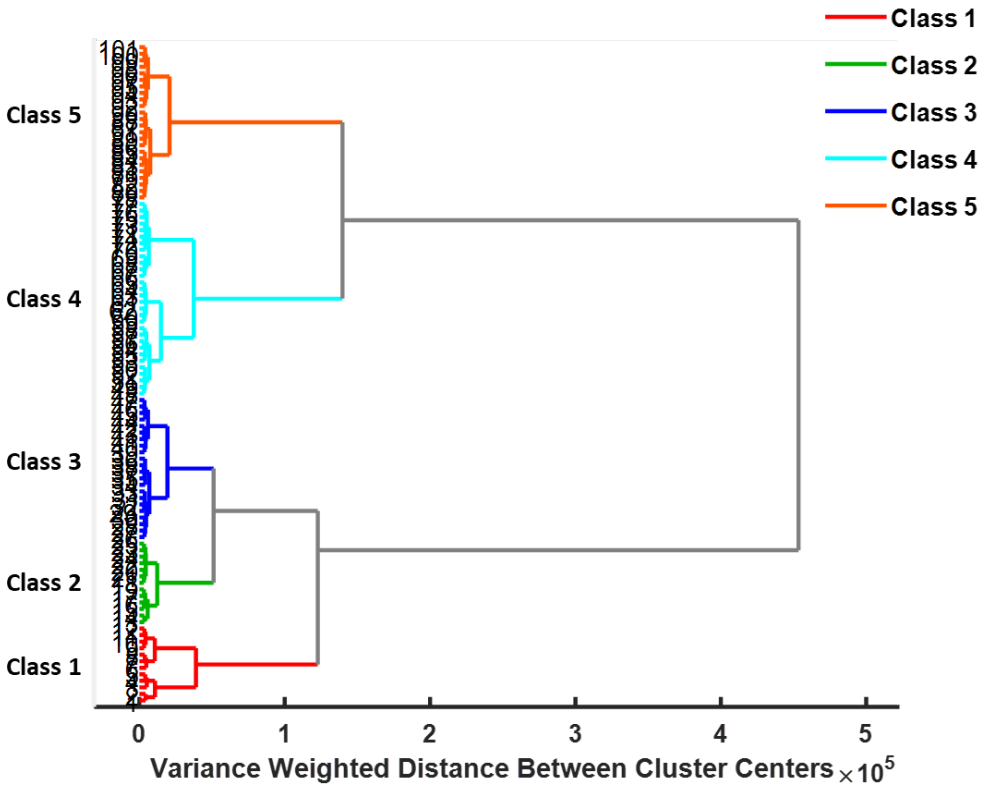

Figure 8. Example of a dendrogram. The numbers on the left side of the figure correspond to spectra in a data set. The user selects the number of clusters in a cluster analysis. That is, imagine moving a vertical line back and forth across the dendrogram. The number of horizontal lines crossed by the vertical line is the number of clusters selected, e.g., if the line were at 3 x 10^5 on the x-axis, two clusters would have been selected. Five clusters (classes) have been selected here, which are color coded and numbered. The distance between the left edge of the plot and the point where the clusters meet is proportional to the distance between them.

III. SUMMARY AND CONCLUSIONS

XPS is the most common and important technique for chemically analyzing surfaces. The current trend to collect increasingly large data sets in degradation, *operando*, depth profiling, and imagining studies should make chemometrics/informatics techniques increasingly relevant in XPS. Multiple chemometrics/informatics methods, including summary statistics, PCA, MCR, and cluster analysis can be used to analyze complex XPS data sets. It is often advantageous to apply more than one chemometrics/informatics tool to a data set. In this guide, we discuss considerations

associated with the use of chemometrics/informatics EDA tools in analyzing XPS data sets. We have *not* mentioned or discussed all the possible EDA methods that might be applied to them. There are many. Rather, the purpose of this article has been to (i) give the reader an introduction to some of the more common EDA methods that might be used in XPS spectral analyses and (ii) introduce the reader to some of the thinking that should accompany these analyses. The companion to this guide[31] provides examples of applications of chemometrics/informatics EDA methods to two real XPS data sets, explaining additional considerations and principles associated with these analyses.

We now repeat and comment on the steps we recommend for performing a chemometrics/informatics analysis of XPS data sets.

A) *Gather and consider whatever information is available about one's material*. This information should include what is known about the sample chemistry and the preparation of the sample. One should also consider information from other characterization techniques, and additional information from XPS.

B) *Examine and plot the raw data in different ways*. We recommend both so-called overlay and waterfall plots. A less conventional method of plotting the data may be with first (and even second) derivatives.

C) *Develop a general strategy for a chemometrics/informatics analysis*. Those newer to chemometrics/informatics may wish to follow the suggestions in this paper and its companion (see flowchart in Figure 1). As one becomes more experienced in chemometrics/informatics, it will become easier to determine the types of analyses that are best suited for a given data set.

D) *Preprocess the data*. Preprocessing plays an important role in many chemometrics/informatics analyses; data preprocessing can sometimes strongly affect these analyses. A partial list of the possible data preprocessing methods is included in this work, which include no preprocessing at all, normalization, baselining, variable selection, mean centering, autoscaling, Poisson scaling, concatenation of data sets, spectral differentiation, and smoothing. It is common for more than one preprocessing method to be applied to data.

E) *Know where to start a chemometrics/informatics analysis*. We recommend that PCA and perhaps a summary statistic first be applied to data. In this initial analysis, one should look for outliers and/or other anomalies in the data. However, as noted in the Introduction, readers may wish to start with MCR – we strongly recommend MCR.

F) *Determine the number of abstract factors to keep in PCA and MCR analyses*. In general, there is no simple method for determining the number of abstract factors to keep in a PCA or MCR analysis, although this process can be somewhat easier for MCR than for PCA. Various methods, including scree plots, cross-validation, reconstructing the data from increasing numbers of abstract factors, and evaluating models based on their chemical reasonableness should be employed/considered. Sometimes, higher abstract factors may contain useful information about a data set, even when a strong majority of the variance in a data set is captured by a few of the earlier factors.

G) *Return to the original data after a chemometrics/informatics analysis*. The original data should confirm what is found in a chemometrics/informatics

analysis; one should never be too far from the original data in chemometrics/informatics analyses. Indeed, one may need to return to the original data multiple times. As more information is revealed about a data set, additional chemometrics/analyses may then need to be performed on it. The process of performing a high quality chemometrics/informatics analysis is often iterative.

H) *Perform MCR*. MCR is a powerful chemometrics/informatics method. It does not force abstract factors/loadings to be orthogonal, as in PCA. Accordingly, produces abstract factors that look like real spectra so MCR results are often easier to interpret than those from PCA. Because of MCR's (typical) non-negativity constraints, mean centering and autoscaling are not appropriate preprocessing methods for MCR.

I) *Peak fit MCR factors*. Because MCR spectra look like 'real' spectra, MCR factors of XPS data may be peak fit, which can increase the amount of information that can be derived from this technique. We believe the companion work to this one shows the first time the evolution of an XPS data set has been revealed in this manner.[31] The protocols for peak fitting MCR factors may be based on fits to the raw XPS data. Concatenation of data can be useful in MCR (and PCA) analyses – by linking two or more spectra to become a single spectrum, the ratios of the peaks in each spectrum are 'locked', which can lead to more meaningful results.

J) *Identify intermediates in MCR analyses*. To the best of our knowledge, the companion paper to this one shows the first time that an MCR analysis of an XPS data set has revealed intermediates in the evolution (degradation) of a material. Intermediates are common in complex chemical reactions – we do not believe it

should come as a surprise that they have been identified here. A criterion used in this analysis was the level of noise on the scores and loadings. Excessively noisy loadings suggest that noise (not signal) is being incorporated into them, i.e., that the maximum number of abstract factors/loadings that should be kept in a model may have been exceeded. Excessively noisy trajectories for the scores similarly suggest that useful information is no longer being added to the analysis.

K) *Perform cluster analysis*. Cluster analysis is a widely used EDA method. Cluster analysis results are summarized in dendrograms. Dendrograms can allow outliers to be easily spotted. In the companion paper to this one,[31] we calculate the average XPS spectrum of each cluster in a cluster analysis. The resulting series of average spectra allows us to follow the evolution of a material as it degrades.

L) *Know how/where to start doing chemometrics/informatics in one's technical work*. Those wishing to use chemometrics/informatics methods will probably want to read other tutorial papers and books on this subject and may also wish to attend a short course or take a class on this subject. 'Hands-on' short courses and classes that provide the students with sample data sets that they fit/work up themselves can be particularly helpful. Those interested in incorporating chemometrics into their research may also wish to hire a consultant (if they are in industry) or find an academic collaborator (if they are at a university).

IV. GLOSSARY

Abstract factor. In the case of PCA, the abstract factors are (i) the new, rotated axes (PCA loadings), (ii) linear combinations of the original axes, and (iii) orthogonal to each

other. In the case of MCR, the abstract factors (i) often represent chemical states, and (ii) generally are not orthogonal to each other.

Autoscaling. A chemometrics preprocessing method that consists of mean centering a set of numbers and then dividing these results by the standard deviation of the numbers.

Chemometrics. The use of statistical (often multivariate) methods to analyze chemical data and spectra.

Classical least squares (CLS). A chemometrics/informatics tool based on the equation **S** = **Pc**+**e**, where **S** refers to measured spectra, **P** to pure-component spectra, **c** to the concentrations of the pure-component spectra in the spectra, and **e** to the errors.

Cluster analysis. A chemometrics/informatics method that compares, differentiates, and clusters spectra based on the distances between them in a multivariate set of measurements.

Concatenation. The uniting/combining of two or more spectra to make a single spectrum. For example, the concatenation of the vectors [1 5 3] and [2 6 9 0] is [1 5 3 2 6 9 0].

Cross-validation. A process by which a fraction of a data set is modeled (analyzed with a chemometrics/informatics tool) and the remainder of the data is then represented with the model. The quality of the model is judged based on its ability to represent the left-out data.

Dendrogram. Dendrograms are plots used to represent the results from cluster analyses. Typically, the spectra are listed in a column on the left side of a dendrogram. The lengths of the horizontal lines in a dendrogram correspond to the similarities between spectra. One can select the number of clusters to keep in a cluster analysis by drawing a vertical line through the dendrogram – the number of horizontal lines that the vertical line cuts through is the number of clusters kept in the analysis.

Depth profiling. In general, the processing of repeatedly removing a small layer of material from a surface with an ion beam, and then probing it with an analytical method like XPS.

Exploratory data analysis (EDA). The use of chemometrics/informatics methods to explore the structure of data sets where no prior knowledge, information, or categorization of the data sets is included in the analysis.

Hyperspace. A space with more than three dimensions. Chemometrics/informatics methods often treat spectra as single points (vectors) in a hyperspace (multivariate space).

Hyperspectral image. An image obtained by analyzing a surface at a pattern of points on it, collecting a spectrum at each point/pixel. The resulting data ‘cubes’ (parallelepipeds) have two spatial dimensions and a third dimension containing individual spectra.

Informatics. The use of statistical and multivariate methods to analyze data and spectra.

Loadings plot. In PCA, a plot showing the contributions of the original axes to a given rotated axis (loading). That is, loadings plots reveal the degree to which the original axes/variables, e.g., energies or wavelengths, contribute to a loading (new axis created by the rotation of the original coordinate system). Loadings provide chemical information in PCA. That is, a loading (and its underlying chemistry) contribute significantly to spectra (data points) that have high scores on that loading, and vice versa.

Mean centering. A preprocessing method that consists of subtracting the mean of a set of numbers from each number. When a series of XPS spectra are mean centered, this process is repeated at each binding energy. Mean centering centers a set of data points (spectra) about the origin.

Multivariate curve resolution (MCR). A chemometrics/informatics tool that finds underlying components/factors that describe spectra. It is based on solving the governing equation of CLS twice and then iterating to find the components/factors, where an initial guess is made for the factors/components. In general, MCR is applied with non-negativity constraints such that the new components/factors only contain positive values.

Operando. This term usually refers to the analysis of materials that are studied under their working conditions.

Outlier. A piece of data or measured spectrum that is significantly different from the general trends in a data set. Outliers describe unusual signal that may be anomalous or incorrect.

Overlay plot. A plot in which spectra are plotted on top of each other. They can be useful for assessing whether changes are taking place in spectra. Overlay plots can be difficult to interpret when a large number of spectra are plotted on top of each other.

Pattern recognition entropy (PRE). An EDA technique that is based on Shannon's concept of entropy. PRE is a summary statistic.

Peak fitting. The process of fitting XPS narrow scans with a baseline and one or more synthetic (mathematical) peaks that represent chemical states.

Preprocessing. A mathematical treatment or modification of raw data that is used to prepare it for a subsequent chemometrics/informatics analysis. Preprocessing is often used to remove or suppress signal that is not of interest, e.g., baselines or clutter.

Principal component (PC). Factors estimated in principal components analysis. PCs are factors that define an orthogonal coordinate system that maximize capture of variance (or sum-of squares) in a data set. Each PC has associated scores and loadings.

Principal component analysis (PCA). One of the most important EDA methods. In essence, in PCA, spectra are considered to be single points in a multivariate space. PCA is designed to capture the maximum sum-of-squares in a data set with the fewest factors; the coordinates of the original space are rotated to capture the greatest possible amount of variance in the spectra. The loadings are the contributions of the old axes to the new ones, and the projections of the data points (spectra) onto the new axes (PCs) are called 'scores'. In PCA, the new axes (PCs) are orthogonal to each other.

Scores plot. A plot of the projections (scores) of the spectra (data points) on the new axes (principal components/abstract factors). Two- and three-dimensional scores plots are plots of the scores of the spectra (data points) on two or three principal components, respectively. Scores plots often reveal various features of data sets, including groupings of spectra, outliers, anomalies in data sets, and trajectories in data sets.

Scree plot. A plot of the variance captured per abstract factor, often represented as the log of the variance captured per abstract factor (the eigenvalues of the cross-product of the data matrix are proportional to variance captured by each PCA . Scree plots may also be plotted as the cumulative variance captured by the factors. A discontinuity or 'knee' in a scree plot often indicates the number of abstract factors that describe a data set.

Summary statistic. A single number that is used to characterize a data set or spectrum. The mean, median, mode, standard deviation, variance, and PRE are examples of summary statistics.

Three-dimensional scores plot. See 'Scores plot'.

Time-of-flight secondary ion mass spectrometry (ToF-SIMS). A mass spectrometric surface analytical technique that probes surfaces with primary ions and detects the resulting secondary ions.

Two-dimensional scores plot. See 'Scores plot'.

Waterfall plot. Waterfall plots show spectra in a side-by-side, sequential fashion. In general, the sequence of the spectra will be the same as the order in which they were acquired.

X-ray photoelectron spectroscopy (XPS). A surface sensitive analytical technique based on the photoelectric effect in which surfaces are irradiated with X-rays and the resulting photoelectrons are detected.

CONFLICT OF INTEREST

Some of the authors work for organizations that sell the chemometrics software (NG,

PLS_Toolbox) and XPS data analysis software (NF, CasaXPS) used in this study.

ACKNOWLEDGMENTS

Because of the similarity and close connection between this guide and its companion,[31] some of the text and figures in these papers are either the same or very similar. This reuse of this material is done with permission from the publisher.

DATA AVAILABILITY

The cellulose, tartaric acid, and sucrose data sets used in this work are provided in the Supplemental Materials of the companion guide to this document[31] as VAMAS files.

REFERENCES

1. K. Artyushkova and J. E. Fulghum, Journal of Electron Spectroscopy and Related Phenomena **121** (1), 33-55 (2001).
2. S. Pylypenko, K. Artyushkova and J. E. Fulghum, Applied Surface Science **256** (10), 3204-3210 (2010).
3. M. P. Felicissimo, J. L. S. Peixoto, R. Tomasi, A. Azioune, J.-J. Pireaux, L. Houssiau and U. P. Rodrigues Filho ¶, Philosophical Magazine **84** (32), 3483-3496 (2004).
4. S. Chatterjee, B. Singh, A. Diwan, Z. R. Lee, M. H. Engelhard, J. Terry, H. D. Tolley, N. B. Gallagher and M. R. Linford, Applied Surface Science **433**, 994-1017 (2018).
5. T. G. Avval, B. Moeini, V. Carver, N. Fairley, E. F. Smith, J. Baltrusaitis, V. Fernandez, B. J. Tyler, N. Gallagher and M. R. Linford, Journal of Chemical Information and Modeling **61** (9), 4173-4189 (2021).
6. H. Ahn, D. W. Oblas and J. E. Whitten, Macromolecules **37** (9), 3381-3387 (2004).
7. K. Artyushkova and J. E. Fulghum, Surface and Interface Analysis **31** (5), 352-361 (2001).
8. C. Bittencourt, M. P. Felicissimo, J.-J. Pireaux and L. Houssiau, Journal of Agricultural and Food Chemistry **53** (16), 6195-6200 (2005).
9. B. J. Tyler, Applied Surface Science **252** (19), 6875-6882 (2006).
10. R. E. Peterson and B. J. Tyler, Applied Surface Science **203-204**, 751-756 (2003).
11. L. Yang, Y.-Y. Lua, G. Jiang, B. J. Tyler and M. R. Linford, Analytical Chemistry **77** (14), 4654-4661 (2005).
12. D. J. Graham and D. G. Castner, Biointerphases **7** (1), 49 (2012).

13. M. S. Wagner, D. J. Graham and D. G. Castner, Applied Surface Science **252** (19), 6575-6581 (2006).
14. P. Van der Heide, *X-ray photoelectron spectroscopy: an introduction to principles and practices*. (John Wiley & Sons, 2011).
15. S. Hofmann, *Auger-and X-ray photoelectron spectroscopy in materials science: a user-oriented guide*. (Springer Science & Business Media, 2012).
16. F. A. Stevie and C. L. Donley, Journal of Vacuum Science & Technology A **38** (6), 063204 (2020).
17. D. R. Baer, K. Artyushkova, C. Richard Brundle, J. E. Castle, M. H. Engelhard, K. J. Gaskell, J. T. Grant, R. T. Haasch, M. R. Linford, C. J. Powell, A. G. Shard, P. M. A. Sherwood and V. S. Smentkowski, Journal of Vacuum Science & Technology A **37** (3), 031401 (2019).
18. V. Gupta, H. Ganegoda, M. H. Engelhard, J. Terry and M. R. Linford, Journal of Chemical Education **91** (2), 232-238 (2014).
19. D. R. Baer and I. S. Gilmore, Journal of Vacuum Science & Technology A **36** (6), 068502 (2018).
20. E. National Academies of Sciences and Medicine, (2019).
21. D. R. Baer, Journal of Vacuum Science & Technology A **38** (3), 031201 (2020).
22. S. Tougaard, Journal of Vacuum Science & Technology A **39** (1), 011201 (2020).
23. C. J. Powell, Journal of Vacuum Science & Technology A **38** (2), 023209 (2020).
24. A. G. Shard, Journal of Vacuum Science & Technology A **38** (4), 041201 (2020).
25. T. R. Gengenbach, G. H. Major, M. R. Linford and C. D. Easton, Journal of Vacuum Science & Technology A **39** (1), 013204 (2021).
26. M. J. Sweetman, S. M. Hickey, D. A. Brooks, J. D. Hayball and S. E. Plush, Advanced Functional Materials **29** (14), 1808740 (2019).
27. D. R. Baer, G. E. McGuire, K. Artyushkova, C. D. Easton, M. H. Engelhard and A. G. Shard, Journal of Vacuum Science & Technology A **39** (2), 021601 (2021).
28. J. V. Macpherson, Physical Chemistry Chemical Physics **17** (5), 2935-2949 (2015).
29. P. A. Navrátil, B. Westing, G. P. Johnson, A. Athalye, J. Carreno and F. Rojas, presented at the Advances in Visual Computing, Berlin, Heidelberg, 2009 (unpublished).
30. J. Wolstenholme, Journal of Vacuum Science & Technology A **38** (4), 043206 (2020).
31. Tahereh G. Avval, Hyrum Haack, Neal Gallagher, David Morgan, Pascal Bargiela, Neal Fairley, Vincent Fernandez and M. R. Linford, Journal of Vacuum Science & Technology A **Submitted** (2022).
32. R. Bro and A. K. Smilde, Analytical Methods **6** (9), 2812-2831 (2014).
33. A. de Juan and R. Tauler, Critical Reviews in Analytical Chemistry **36** (3-4), 163-176 (2006).
34. N. B. Gallagher, J. M. Shaver, E. B. Martin, J. Morris, B. M. Wise and W. Windig, Chemometrics and intelligent laboratory systems **73** (1), 105-117 (2004).
35. S. Chatterjee and M. R. Linford, Bulletin of the Chemical Society of Japan **91** (5), 824-828 (2018).
36. J. Jackson, New York (1991).
37. B. M. Wise and N. B. Gallagher, Journal of Process Control **6** (6), 329-348 (1996).

38. F. Zhang, R. J. Gates, V. S. Smentkowski, S. Natarajan, B. K. Gale, R. K. Watt, M. C. Asplund and M. R. Linford, Journal of the American Chemical Society **129** (30), 9252-9253 (2007).
39. D. S. Jensen, S. S. Kanyal, N. Madaan, J. M. Hancock, A. E. Dadson, M. A. Vail, R. Vanfleet, V. Shutthanandan, Z. Zhu, M. H. Engelhard and M. R. Linford, Surface and Interface Analysis **45** (8), 1273-1282 (2013).
40. L. Pei, G. Jiang, R. C. Davis, J. M. Shaver, V. S. Smentkowski, M. C. Asplund and M. R. Linford, Applied Surface Science **253** (12), 5375-5386 (2007).
41. A. M. Spool, *The Practice of TOF-SIMS: Time of Flight Secondary Ion Mass Spectrometry*. (Momentum Press, 2016).
42. C. V. Cushman, P. Brüner, J. Zakel, G. H. Major, B. M. Lunt, N. J. Smith, T. Grehl and M. R. Linford, Analytical Methods **8** (17), 3419-3439 (2016).
43. H. G. Tompkins and J. N. Hilfiker, *Spectroscopic ellipsometry: practical application to thin film characterization*. (Momentum Press, 2015).
44. B. Voigtländer, *Atomic force microscopy*. (Springer, 2019).
45. G. Haugstad, *Atomic force microscopy: understanding basic modes and advanced applications*. (John Wiley & Sons, 2012).
46. P. Kraniauskas, *Transforms in Signals and Systems (Modern Applications of Mathematics)*. (Addison-Wesley Longman, Incorporated, 1992).
47. R. W. Schafer, IEEE Signal Processing Magazine **28** (4), 111-117 (2011).
48. W. H. Press and S. A. Teukolsky, Computers in Physics **4** (6), 669-672 (1990).
49. J. Luo, K. Ying and J. Bai, Signal Processing **85** (7), 1429-1434 (2005).
50. A. Savitzky and M. J. E. Golay, Analytical Chemistry **36** (8), 1627-1639 (1964).
51. N. Fairley, (https://youtu.be/uQYVBlouJs0, 2017).
52. O. E. de Noord, Chemometrics and Intelligent Laboratory Systems **23** (1), 65-70 (1994).
53. B. Singh, D. Velázquez, J. Terry and M. R. Linford, Journal of Electron Spectroscopy and Related Phenomena **197**, 56-63 (2014).
54. B. Singh, D. Velázquez, J. Terry and M. R. Linford, Journal of Electron Spectroscopy and Related Phenomena **197**, 112-117 (2014).
55. T. d. O. Zuppa Neto, T. G. Avval, P. A. d. O. Morais, W. C. Ellis, S. C. Chapman, A. E. de Oliveira, M. R. Linford, P. B. Farnsworth and N. R. Antoniosi Filho, Journal of the American Society for Mass Spectrometry **31** (7), 1525-1535 (2020).
56. S. Chatterjee, G. H. Major, B. Paull, E. S. Rodriguez, M. Kaykhaii and M. R. Linford, Journal of Chromatography A **1558**, 21-28 (2018).
57. S. Chatterjee, S. C. Chapman, B. M. Lunt and M. R. Linford, Bulletin of the Chemical Society of Japan **91** (12), 1775-1780 (2018).
58. R. L. McLaren, G. R. Owen and D. J. Morgan, Results in Surfaces and Interfaces **6**, 100032 (2022).
59. L. Edwards, P. Mack and D. J. Morgan, Surface and Interface Analysis **51** (9), 925-933 (2019).
60. R. N. Bracewell, Scientific American **260** (6), 86-95 (1989).
61. R. Kramer, *Chemometric techniques for quantitative analysis*. (CRC Press, 1998).