JVST A
Journal of Vacuum Science & Technology A

Practical guide on chemometrics/informatics in x-ray photoelectron spectroscopy (XPS). II. Example applications of multiple methods to the degradation of cellulose and tartaric acid

F

Cite as: J. Vac. Sci. Technol. A **40**, 063205 (2022); https://doi.org/10.1116/6.0001969
Submitted: 17 May 2022 • Accepted: 09 September 2022 • Published Online: 22 November 2022

Tahereh G. Avval, Hyrum Haack, Neal Gallagher, et al.

COLLECTIONS

Paper published as part of the special topic on Reproducibility Challenges and Solutions II with a Focus on Surface and Interface Analysis

F This paper was selected as Featured

View Online Export Citation CrossMark

Practical guide on chemometrics/informatics in x-ray photoelectron spectroscopy (XPS). II. Example applications of multiple methods to the degradation of cellulose and tartaric acid

Cite as: J. Vac. Sci. Technol. A **40**, 063205 (2022); doi: 10.1116/6.0001969
Submitted: 17 May 2022 · Accepted: 9 September 2022 ·
Published Online: 22 November 2022

View Online Export Citation CrossMark

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

AFFILIATIONS

[1] Department of Chemistry and Biochemistry, Brigham Young University, C100 BNSN, Provo, Utah 84602
[2] Eigenvector Research, Inc., Manson, Washington, DC 98831
[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, United Kingdom
[4] HarwellXPS—EPSRC National Facility for Photoelectron Spectroscopy, RCaH, Didcot, Oxon OX11 0FA, United Kingdom
[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 TQ14 8NE, United Kingdom
[7] Nantes Université, CNRS, Institut des Matériaux de Nantes Jean Rouxel, IMN, F-44000 Nantes, France

Note: This paper is part of the Special Topic Collection: Reproducibility Challenges and Solutions II with a Focus on Surface and Interface Analysis.

[a)] **Author to whom correspondence should be addressed:** mrlinford@chem.byu.edu

ABSTRACT

Chemometrics/informatics, and data analysis in general, are increasingly important in x-ray photoelectron spectroscopy (XPS) because of the large amount of information (spectra/data) that is often collected in degradation, depth profiling, operando, and imaging studies. In this guide, we present chemometrics/informatics analyses of XPS data using a summary statistic (pattern recognition entropy), principal component analysis, multivariate curve resolution (MCR), and cluster analysis. These analyses were performed on C 1s, O 1s, and concatenated (combined) C 1s and O 1s narrow scans obtained by repeatedly analyzing samples of cellulose and tartaric acid, which led to their degradation. We discuss the following steps, principles, and methods in these analyses: gathering/using all of the information about samples, performing an initial evaluation of the raw data, including plotting it, knowing which chemometrics/informatics analyses to choose, data preprocessing, knowing where to start the chemometrics/informatics analysis, including the initial identification of outliers and unexpected features in data sets, returning to the original data after an informatics analysis to confirm findings, determining the number of abstract factors to keep in a model, MCR, including peak fitting MCR factors, more complicated MCR factors, and the presence of intermediates revealed through MCR, and cluster analysis. Some of the findings of this work are as follows. The various chemometrics/informatics methods showed a break/abrupt change in the cellulose data set (and in some cases an outlier). For the first time, MCR components were peak fit. Peak fitting of MCR components revealed the presence of intermediates in the decomposition of tartaric acid. Cluster analysis grouped the data in the order in which they were collected, leading to a series of average spectra that represent the changes in the spectra. This paper is a companion to a guide that focuses on the more theoretical aspects of the themes touched on here.

I. INTRODUCTION

In this guide, we show the analysis of two rather large x-ray photoelectron spectroscopy (XPS) data sets using various exploratory data analysis (EDA) methods. In particular, we analyze two XPS data sets obtained from the repeated analyses of filter paper (cellulose, a natural polymer containing C, O, and H) and tartaric acid (a small, symmetric molecule that also contains only C, O, and H). We focus here on carbon and oxygen containing materials because the C 1s and O 1s XPS narrow scans are the most commonly shown and analyzed in the scientific literature. Both data sets reveal significant degradation of the materials during XPS analyses that appears to lead to graphitization. These data sets were analyzed/probed with a series of EDA chemometrics/informatics methods that include a summary statistic (pattern recognition entropy, PRE), principal component analysis (PCA), multivariate curve resolution (MCR), and cluster analysis. This work also presents an examination of the raw spectra, identifies anomalies in the data sets, covers methods for determining the number of abstract factors to keep (that best describe a data set), discusses data preprocessing, shows XPS peak fitting of MCR components (to the best of our knowledge this is the first time this has been done), identifies intermediates revealed in an MCR analysis (to the best of our knowledge this is also the first time this has been done), shows the evolution of XPS data using cluster analysis, and compares the results from multiple EDA methods. This guide has Paper I that focuses on more general and theoretical aspects of the techniques and analyses shown here. The chemometrics/informatics methods employed in this study have been reviewed and discussed multiple times in the literature.[1–8]

XPS is the most widely used and important method for chemically analyzing surfaces.[9–12] 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 a sample surface. 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.[13] While the x rays used in XPS can penetrate ca. 1 μm into a material, the photoelectrons they generate can only escape in an unattenuated fashion from the upper ca. 5–10 nm of it. Accordingly, XPS is a surface sensitive spectroscopy. Furthermore, while little or no sample damage occurs in many XPS analyses, e.g., for many inorganic materials, it does occur in some cases. This damage is often caused more by photoelectrons than the x rays themselves. 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,[14,15] a group of experts has recently produced a series of guides that cover multiple aspects of XPS.[12,16–25] 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.

Materials containing carbon and oxygen (and hydrogen) have been extensively analyzed by XPS. Indeed, Beamson and Briggs' classic work on organic polymers suggests that a large subset of the organic polymers of interest in XPS are those that contain only carbon, hydrogen, and oxygen.[26] Such materials include the acrylates, methacrylates, polyethylene glycol/oxide, polypropylene glycol/oxide, polyethyleneterephthalate, polyether ether ketone, and the naturally occurring polymers lignin and cellulose. All are of practical and theoretical importance, and there are multiple examples of their characterization in the literature by XPS,[27–29] including by near ambient pressure (NAP)-XPS.[30–34] These polymers are dominated by a series of functional groups that contain increasing numbers of carbon—oxygen bonds, including reduced carbon with no carbon—oxygen bonds (C—C/C—H, where carbon is usually sp^2 or sp^3 hybridized), C—O (alcohols, ethers, and epoxides), C═O (carbonyls) and O—C—O (acetals), C(O)O (carboxyls and esters), and O—C(O)O (carbonates).[13] While both the C 1s and O 1s narrow scans are important for understanding these polymers, the C 1s narrow scan is usually more informative because (i) the chemical shifts exhibited by carbon in its different oxidation states occur over ca. 10 eV, which is quite a bit more than for oxygen (ii) organic polymers generally contain more carbon atoms than oxygen atoms, i.e., the C 1s narrow scan often represents a larger fraction of the atoms in the material; and (iii) the XPS of carbon is quite strongly determined by initial state effects, i.e., the state of the atom as influenced by those it is bonded to. As a result of this first point, the C 1s spectrum is often easier to fit/interpret. The large spread in binding energies for carbon is, no doubt, a reflection of its lower electronegativity compared to oxygen. That is, carbon may be bonded to elements that are more electronegative than it is, e.g., nitrogen, oxygen, chlorine, and fluorine, to those that have roughly the same electronegativity, e.g., hydrogen and sulfur, and to those that are more electropositive, e.g., silicon and germanium, while there is only one element (fluorine) that is more electronegative than oxygen. Sulfur, which has about the same electronegativity as carbon, also shows a wide range of chemical shifts. In addition to polymers, some small organic molecules with sufficiently low volatilities can be analyzed by conventional XPS. Such molecules are often held together by multiple hydrogen bonds. More volatile organic molecules may be analyzed by NAP-XPS.[30]

While XPS causes little or no sample damage in many cases, organic materials sometimes degrade during XPS analyses. This damage usually occurs gradually, over multiple scans. Damage can be identified by comparing different scans in an analysis, e.g., by ratioing spectra.[25] In describing the damage caused by x rays and photoelectrons during XPS, Baer *et al.* noted that, in general, sample damage takes place in an approximately linear fashion at the beginning of an analysis but nonlinearly at later times.[35] Because it undergoes rapid damage during XPS analysis, polyvinyl chloride (PVC) is often used as a standard in damage studies.[35–38] Even though clean cellulosic filter paper is damaged during XPS, it has been proposed as an *in situ* reference for analysis of organics and polymeric materials.[39] Cellulosic filter paper stays relatively clean in and out of vacuum, and, more importantly, its C 1s envelope is different from adventitious carbon contamination. Related studies have shown damage to polymers when they are irradiated with energetic electrons or photons, where these conditions appear to lead to an increase in sp^2 carbon/sample graphitization and cross-linking.[40,41] Large numbers of spectra may be needed to understand sample damage. These large data sets may be difficult to interpret and visualize by conventional methods.

Chemometrics/informatics methods have been used for years to analyze large and complex data sets. However, in spite of previous work in this area,[1,5,42–44] this capability has been overlooked by much of the 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 techniques. The first extensive use of chemometrics algorithms, such as PCA, MCR, and image classification, in degradation studies was done on a PVC/polymethymethacrylate (PMMA) blend.[41,45] 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.[42,45–51] Chemometrics/informatics techniques can be used as an alternative to or in combination with conventional peak fitting because they reduce the dimensionality of large and complex data sets and may extract hidden features in the data. Fundamentally, these multivariate methods work in XPS data analysis because of the high degree of correlation between the spectra in many data sets. Chemometrics/informatics methods are particularly relevant to XPS today because of the trend to collect increasingly large data sets in degradation, depth profiling, *operando*, and imaging studies. Thus, methods are increasingly needed to more efficiently analyze and visualize these data sets. In addition to providing a wide variety of analysis methodologies, chemometrics/informatics can guide experimental design to ensure maximal interpretability of experimental results. Finally, while the particular EDA methods employed herein are, for the most part, widely used and effective, we have not covered all possible EDA methods in this guide—there are many more than may be considered.

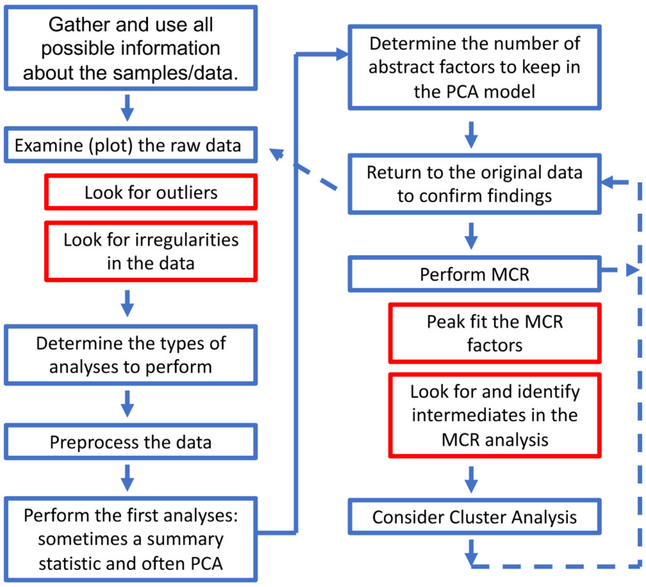
Gather and use all possible information about the samples/data.
Examine (plot) the raw data
Look for outliers
Look for irregularities in the data
Determine the types of analyses to perform
Preprocess the data
Perform the first analyses: sometimes a summary statistic and often PCA
Determine the number of abstract factors to keep in the PCA model
Return to the original data to confirm findings
Perform MCR
Peak fit the MCR factors
Look for and identify intermediates in the MCR analysis
Consider Cluster Analysis

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

Section III of this paper is organized into sections that cover multiple aspects of the chemometrics/informatics analysis of XPS data. To help the reader understand the connections between these sections/concepts, they have been organized into a flowchart (Fig. 1). This diagram teaches that one should first gather (and then use) all the information available about one's samples and data. The raw data should then be plotted and its general structure should be analyzed, where one should look for any outliers or irregularities in it. At this point, one should determine which chemometrics/informatics analyses to perform. The data preprocessing necessary for these analyses should then be undertaken. Because of its widespread use and power, we recommend PCA first be performed. The reader may also wish to consider a summary statistic analysis. One then determines the appropriate number of abstract factors to keep in the PCA model. After obtaining these initial results, one should return to the original data to confirm them. We then recommend that MCR be performed. Peak fitting of the MCR factors can help reveal the chemical evolution of a data set. Chemical intermediates may even appear in this analysis. Finally, one may wish to consider cluster analysis to obtain another mathematical perspective of one's data. As suggested by the dashed lines in the flowchart, we believe that chemometrics/informatics analyses should always point one back to the original data. At that point, initial findings 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.

Finally, someone new to chemometrics/informatics may want to apply these methods in their work, but be put off by all the new vocabulary, concepts, and techniques in this paper and the previous 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/mathematical 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) Secs. III A and III B and then skipping to the sections on MCR. The other sections of this document and the information in the previous 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 factors themselves to see where they no longer show meaningful structure,
(iii) creating models with successively larger numbers of factors in them, evaluating the chemical reasonableness of the models (this approach is shown in Sec. III),
(iv) perhaps reconstructing spectra from one's data set with MCR factors as is done in Figs. 13–15 for PCA, and
(v) using what one knows about one's sample to determine the appropriate number of MCR factors to keep/expect.

It would probably be best to use a combination of these approaches. While our view may not be shared by all chemometricians, we believe that MCR is the most powerful and relevant tool for analyzing the types of data sets considered in this work, and that if one were to learn and apply only one of these techniques, it should be MCR. However, in the long run, if one is to be effective in this space—if one wishes to be able to apply chemometrics/informatics methods to a wide variety of data sets, one should become familiar with at least PCA, and, in time, with other chemometrics/informatics methods as well.

II. EXPERIMENT

The impact of repeated XPS analyses (x-ray irradiation of the samples) on two organic materials (cellulose and tartaric acid) was examined using the data sets from two different instruments as described below.

A. Materials analyzed

1. Cellulose

Sixty C 1s and O 1s XPS narrow scans of a cellulose sample (filter paper) were collected with a Kratos AXIS Ultra instrument with an Al K-alpha monochromatic source at 300 W. In our experience, the charge compensation system for this instrument is superb—the data were not shifted or otherwise corrected after they were collected. The pass energy for these measurements was 10 eV. The instrumental resolution determined from the Fermi edge of silver yielded a resolution of 0.5 eV with a step size of 0.1 eV. The region analyzed was about $150 \times 350\,\mu m^2$ (FOV2 slot). Acquisition of each spectrum took about 10 min. The total analysis time was 36 h.

2. Tartaric acid

One hundred and one C 1s and O 1s XPS narrow scans of a tartaric acid sample were collected with a Thermo Fisher Scientific K-alpha+ spectrometer. Samples were mounted by pressing them into a well on the Thermo K-Alpha copper powder sample exchangeable top plate. Data were recorded using a microfocused monochromatic Al K-alpha x-ray source ($6\,mA \times 12\,kV = 72\,W$) using the 400-$\mu$m spot option which forms an ellipse of approximately $600 \times 400\,\mu m^2$. Data were recorded at pass energies of 150 eV for survey scans and 40 eV for high resolution/narrow scans with 1 and 0.1 eV step sizes, respectively. A total of two scans each for the C1s and O1s regions were acquired, totaling approximately 50 s per iteration. Charge compensation was achieved using a combination of both low energy electrons and argon ions with the flood source operating at the following conditions: beam = 0.2 V, emission = 100 mA, and extractor = 40 V. However, in spite of the reasonable efforts undertaken to provide adequate charge compensation for the sample, the less expensive K-alpha+ instrument probably does not have the capabilities of the higher end instrument used to analyze the cellulose sample. Accordingly, in this work, these data were handled in both corrected and uncorrected ways. That is, all of the analyses in this work, except those in Figs. 13–15, were performed with uncorrected/unshifted data. Appendix describes the approach taken to shift the peak positions of the O 1s peaks to a common value.

B. Data organization

The spectra analyzed herein were organized row-wise into data matrices, where each row of the data set contained one spectrum/scan. The concatenated data set consisted of C 1s and O 1s narrow scans joined/linked together into a single spectrum.

C. Computer/software

The computer programs used to analyze the data sets with summary statistics were written in the MATLAB computing environment (Version R2018b, Release No. 8.6.0.267246, The MathWorks Inc., 1 Apple Hill Drive, Natick, MA). The computer used for this work was an Intel Corei7-4770 CPU@3.40 GHz with 8.0 GB of RAM on a 64-bit Windows 10 Enterprise Edition operating system. PCA and MCR were performed using the PLS Toolbox, version 8.7, and MIA Toolbox, version 3.0.9 from Eigenvector Research, Inc., Wenatchee, WA, in the MATLAB programming environment. Curve fitting was performed in CasaXPS 2.3.25. The PCA abstract factors used in Figs. 13–15 were computed using Iterative SVD (Ref. 52) implemented in CasaXPS.

III. RESULTS AND DISCUSSION

We now show the chemometrics/informatics analyses of two moderately large XPS data sets, as presented in a series of subsections. These subsections cover important concepts/steps that should be considered in performing chemometrics/informatics analyses including gathering all the information possible about the samples, examining/plotting the raw data, determining the types of analyses to perform, preprocessing the data, knowing where to begin the chemometrics/informatics analysis, identifying outliers or other unexpected features in data sets, returning to the original data to confirm chemometrics/informatics results, determining the number of factors to keep in a model, MCR, peak fitting of MCR factors, more complicated MCR factors and the presence of intermediates, and cluster analysis, including using the average spectra from clusters to follow an analysis. We again refer the reader to the flowchart in Fig. 1, which shows the logical sequence of and connections between these topics.

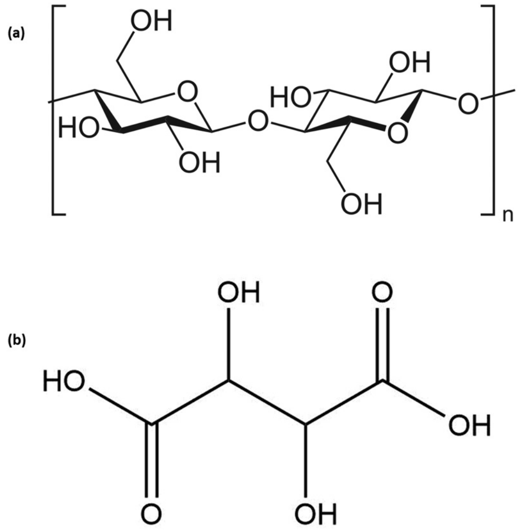

FIG. 2. Chemical structures of cellulose (top) and tartaric acid (bottom).

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

As discussed in Paper I,[53] all of the information that is available about a sample, including its chemical/structure information, should be considered in a chemometrics/informatics analysis of it. Most of this paper is about analyses of two data sets obtained from samples of cellulose and tartaric acid. The structures of these polymers/molecules are shown in Fig. 2. In both cases, these structures suggest that two types of chemically different carbon are in these materials: for cellulose, we expect carbon in +1 (C—O) and +2 (O—C—O) oxidation states, while for tartaric acid, we expect carbon in +1 (C—O) and +3 [C(O)O] oxidation states.[13] Thus, if additional (more than two) signals are present in the XPS spectra of these materials, they must come from impurities or (in the case of cellulose) additives.

B. Examine (plot) the raw data

As discussed in Paper I,[53] an early step in a chemometrics/informatics analysis should be to visually examine/plot the data. Figure 3 presents overlay plots of the C 1s and O 1s narrow scans from the cellulose and tartaric acid data sets, i.e., all the spectra are plotted on top of each other. These plots show significant changes in the data, where such changes in XPS spectra may indicate sample degradation or charging. These plots also suggest that there is a rather significant break or discontinuity in the cellulose O 1s data set. These plots provide good motivation for the chemometrics/informatics analyses of these spectra.

Waterfall plots show spectra in a side-by-side, sequential fashion. The waterfall plots in Figs. 4(a)–4(c) for cellulose show a decrease in the C—O peak, an increase in the reduced carbon peak, and a decrease in the O 1s signal. Because of the more three-dimensional nature of waterfall plots, it can be advantageous to view them from different angles. Figure 4 shows "high binding energy" and "low binding energy" views of the cellulose and tartaric acid C 1s data sets. These plots again suggest that there is a break/discontinuity in the cellulose data, which will be discussed below. Like the plots of the cellulose data, the waterfall plots of the tartaric acid data show an increase in the reduced carbon peak and a decrease in the O 1s signal [see Figs. 4(d)–4(f)].

Another possible way to view spectra is by plotting their derivative. Figure 5 shows the first and last (60th) C 1s and O 1s narrow scans of the cellulose data set and their derivatives. These plots reveal considerable differences between the 1st and 60th C 1s narrow scans in both their differentiated and undifferentiated forms. However, the changes in the O 1s spectra are more subtle—the most obvious change in them is that the O 1s peaks decrease in size. However, the O 1s peak position and peak shape do change to some degree, where the shift in this peak position is nicely revealed by the change in the zero-crossings of the corresponding derivative curves. There is generally more complexity/"wiggles" in derivative spectra than undifferentiated spectra.

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 Fig. 1. Of course, there are other chemometrics/informatics analyses and approaches that the analyst will learn in time and be able to consider. However, one new to this area may wish to follow the approach outlined in Fig. 1 because (i) PCA, MCR, and cluster analysis are very well accepted and established and (ii) they have been shown to be effective on many types of data sets. We have also found summary statistics to be helpful in the initial evaluation of our data. Of course, those who are more experienced with chemometrics/informatics may see more tailored/focused approaches for analyzing particular data sets.

D. Preprocess the data

"Data preprocessing" or just "preprocessing" refers to any mathematical treatment of a data set prior to a chemometrics/informatics analysis. The objectives of data preprocessing are to suppress signal that is not of interest, bring signal of interest to the forefront, and make the data mathematically consistent with the analyses that are to be performed on it, e.g., one may add an extremely small number to zero values in a data set to prevent an algorithm from dividing by zero. Paper I[53] describes multiple ways of preprocessing data for chemometrics/informatics analyses that include no preprocessing at all, normalization with the 1-norm, baselining, variable selection, mean centering, autoscaling, Poisson scaling, concatenation, differentiation, smoothing, and the use of multiple preprocessing methods. Some of these methods are important for XPS data analysis, and some of them are discouraged. An advantage of pattern recognition entropy (PRE), which will be described in Sec. III E, is that it requires little or no preprocessing.

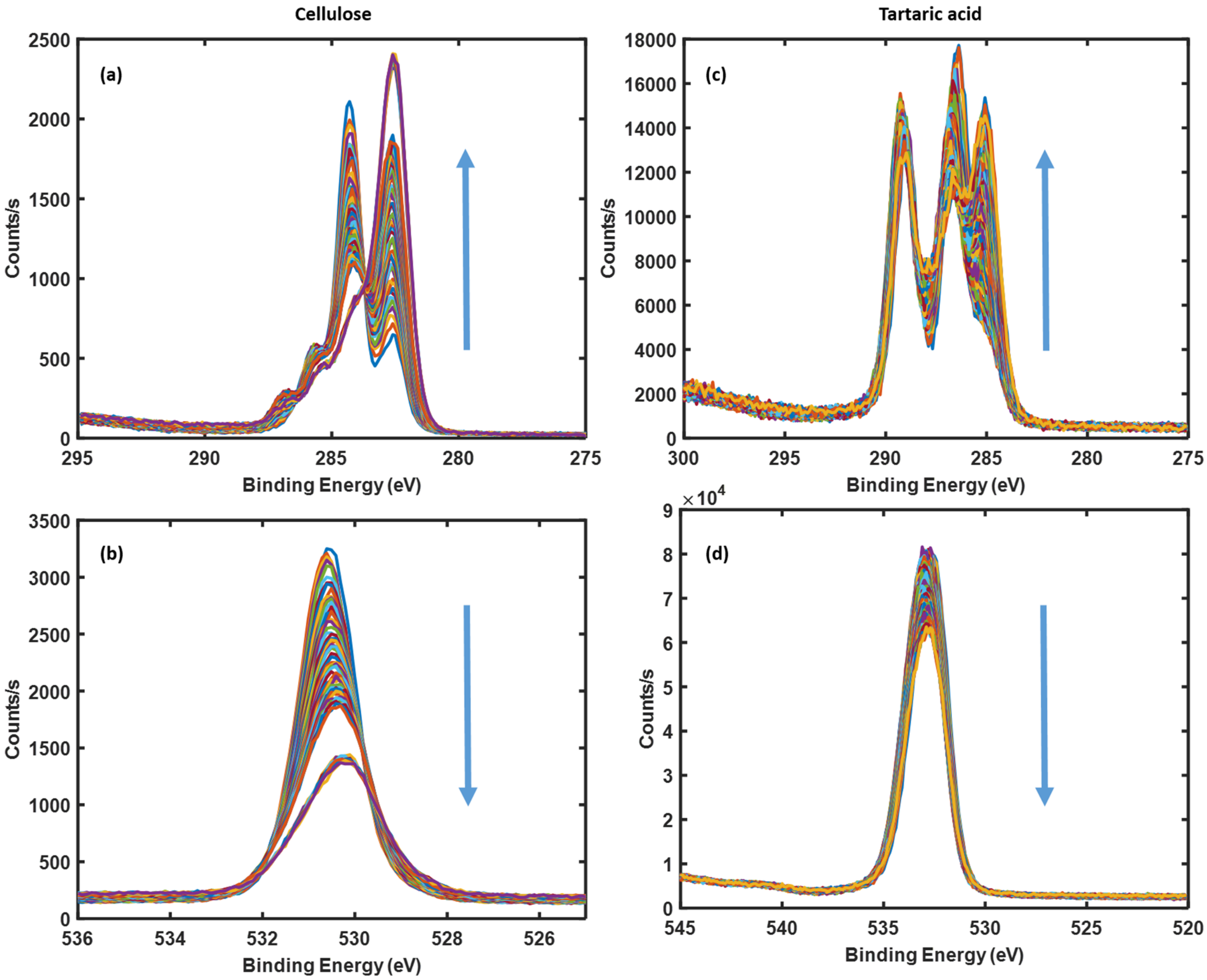

FIG. 3. Overlay plots of 60 (a) C 1s and (b) O 1s narrow scans from an XPS analysis of cellulose and 101 (c) C 1s and (d) O 1s narrow scans from an XPS analysis of tartaric acid. Arrows show the general direction of early time to later time in data collection. See Appendix for an approach used to energy shift the O 1s and C 1s tartaric acid peaks such that the O 1s signals would align.

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

A summary statistic is a single number that characterizes a spectrum. Summary statistical analyses are quite easy to perform and can be helpful in identifying trends in data/spectra. Accordingly, we suggest that a summary statistic be applied early in a data analysis. PRE,[1,5,6,59–61] which is based on Shannon's entropy, clusters and reveals trends in data, where its results are often similar to those in PCA scores plots. PRE is particularly useful in image analysis. Figure 6 shows the PRE analyses of the C 1s and O 1s spectra from the cellulose and tartaric acid data sets. First, Fig. 6 simply reveals that the PRE values change, which suggests that the spectra are changing (in three of the four subplots in Fig. 6 these changes are basically monotonic). Of course, this is not surprising because the original data/underlying spectra are also changing (see Figs. 3 and 4). Second, PRE reveals an abrupt change in the cellulose C 1s and O 1s spectra, where this discontinuity occurs between spectra 51 and 52. No evidence for a gap or jump is present in either the raw spectra (Figs. 3 and 4) or in the PRE analyses [Figs. 6(c) and 6(d)] of the tartaric acid spectra. Figure S1 in the supplementary material[65] shows other summary statistical

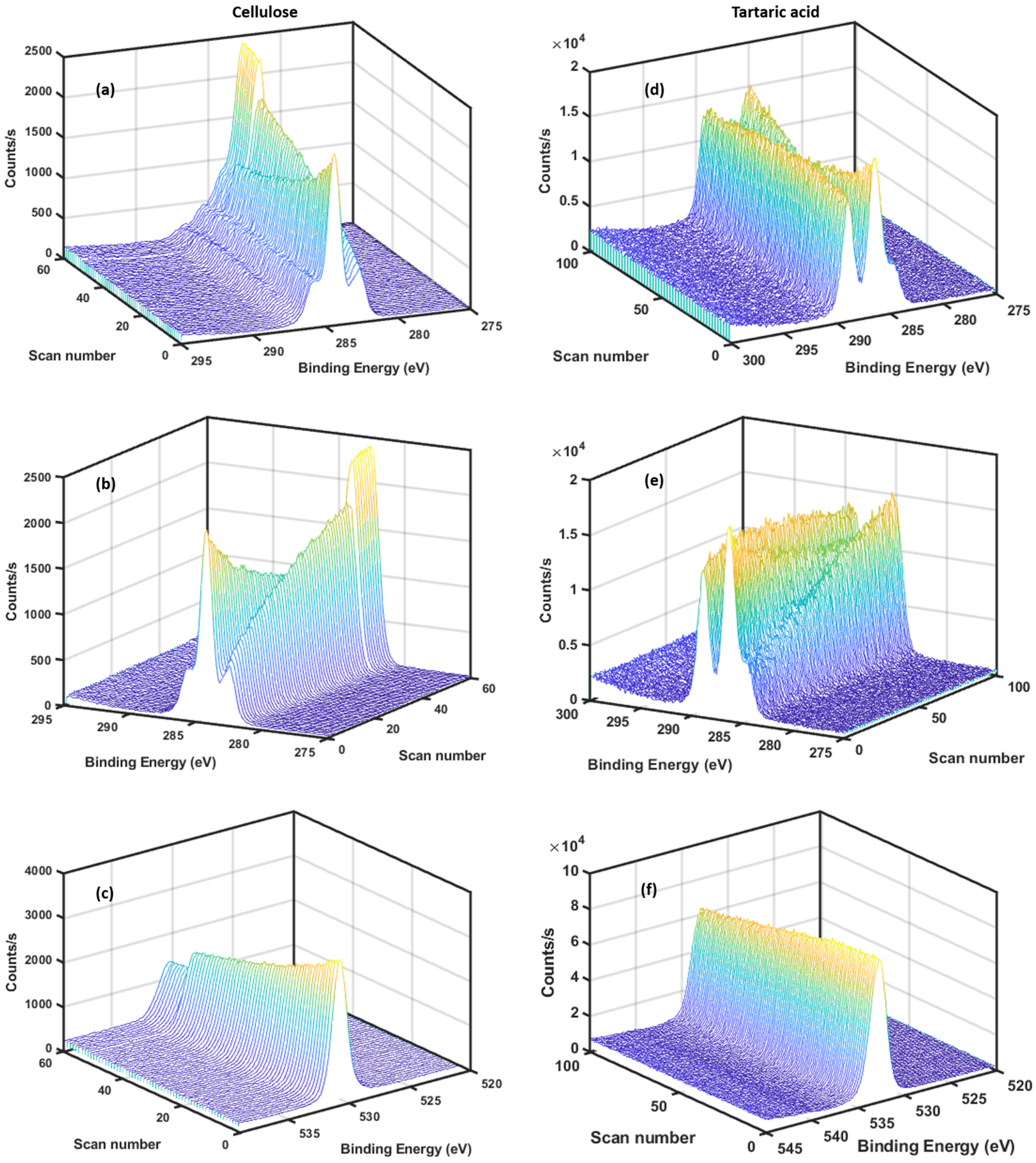

FIG. 4. Waterfall plots of the C 1s and O 1s narrow scans in the cellulose [(a)–(c)] and tartaric acid [(d)–(f)] data sets. Two different views of the C 1s data sets [(a) and (b) and (d) and (e)] and one view of the O 1s data sets [(c) and (f)] are presented. The cellulose and tartaric acid data sets here contain 60 and 101 spectra, respectively.

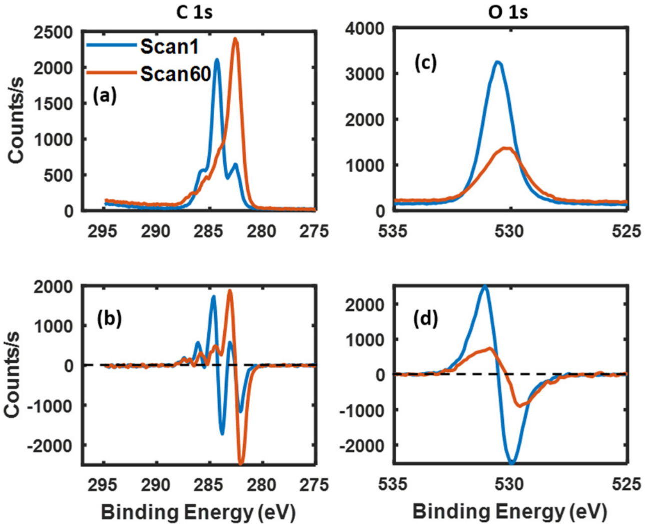

FIG. 5. First and the last (60th) undifferentiated [(a) and (c)] and differentiated/derivative [(b) and (d)] C 1s and O 1s narrow scans of the cellulose data set. A smoothing/differentiating Savitzky–Golay filter[54–58] was used to produce the results in (b) and (d).

analyses of these data. PRE is a rather new chemometrics/informatics tool. It was developed by some of the authors on this paper.

As single numbers, summary statistics are limited in the amount of insight they can provide about spectra. Accordingly, we next recommend that a whole-spectrum analysis be performed. The most common, and arguably important, of these EDA methods is PCA. Figure 7 shows the two-dimensional PCA scores plots of the C 1s, O 1s, and concatenated C 1s and O 1s spectra of the cellulose and tartaric acid data sets, which were preprocessed by mean centering. In all cases, the data points/spectra fall along trajectories, which suggest steady changes in the spectra. These types of trajectories are often observed in PCA analyses. As can be seen in the x- and y-labels of these plots, most of the variance in these data sets is captured by these first two PCs. In Paper I,[53] and also below, we discuss methods for determining the number of abstract factors to keep in a chemometrics/informatics analysis. The breaks in the trajectories of the data points/spectra in the PCA of the cellulose data set [Figs. 7(a)–7(c)] take place at the same point as the breaks in the PRE analysis in Figs. 6(a) and 6(b). This discontinuity in the data was confusing to us. The scientist who took the data informed us that after scan 51, a different analysis was performed on this cellulose sample, after which the remainder of the data for this analysis were collected. That is, the cellulose received additional irradiation between scans 51 and 52. As in its PRE analysis, no break or discontinuity is present in the tartaric acid data set [Figs. 7(d)–7(f)]. Rather, continuous trajectories are observed, which, again, suggest steady changes in the spectra.

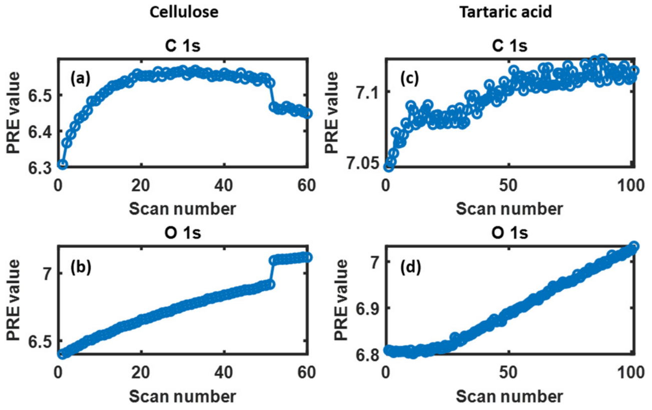

FIG. 6. PRE analysis of 60 C 1s (a) and O 1s (b) narrow scans from an XPS analysis of cellulose and 101 C 1s (c) and O 1s (d) narrow scans from an XPS analysis of tartaric acid. No preprocessing was performed on the data before these analyses. PRE is a summary statistic based on Shannon's entropy. That is, PRE takes a spectrum and turns it into a single, characteristic number, which is plotted here.

Outlier identification is an early step in an informatics analysis. In the C 1s, O 1s, and concatenated C 1s and O 1s PCA scores plots of cellulose [Figs. 7(a)–7(c)], the first points (corresponding to the first narrow scans collected) are either fairly far from the next points and/or inconsistent with the trajectories of the points that follow them. These results suggest that the first C 1s and O 1s scans of the cellulose data set may be outliers. These effects are even more pronounced in the 3D PCA scores plot (on the first three PCs of these data) in Fig. 8. This result illustrates that even if most of the variation in a data set is be captured by a few PCs, the higher PCs sometimes contain useful, and even important, information about the data set. The same applies for MCR. Finally, additional information may be added to PCA scores plots. Figure 9 shows a plot of the PCA of the concatenated C 1s and O 1s narrow scans of the cellulose data set, where the elapsed time of the analysis has been added to the plot via the color of the data points. This type of plot allows additional information/another dimension to be rather easily added to a graph.

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 captures the relevant variance in a data set. 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 to this end, including scree plots, cross-validation, and reconstructing the data from increasing numbers of PCs. Figure 10 shows scree plots obtained from the PCA analysis of the cellulose data set after mean centering. The top row of plots in this figure, which show the cumulative variance captured by the PCs, reveals that for all three data sets (the C 1s, O 1s, and concatenated C 1s and O 1s data sets), the first two PCs capture more than 99% of the variance in the data. The bottom row of scree plots in Fig. 10 shows the log of the eigenvalues (a measure of the amount of variance captured per PC) as a function of the principal component number. In these types of plots,

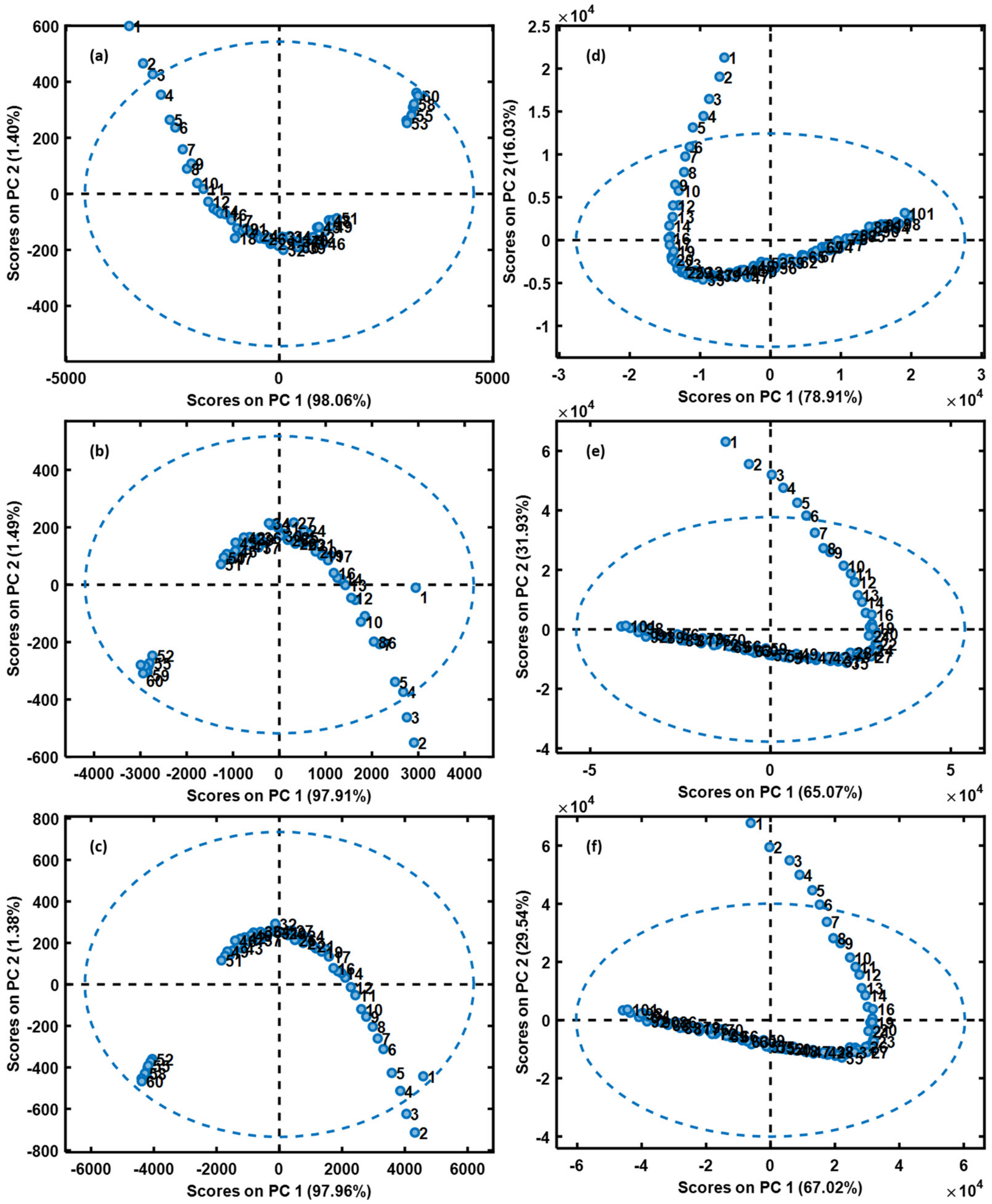

FIG. 7. Two-dimensional PCA scores plots of the C 1s [(a) and (d)], O 1s [(b) and (e)], and concatenated C 1s and O 1s [(c) and (f)] spectra of the cellulose (first column) and tartaric acid (second column) data sets after preprocessing by mean centering. Each point in these plots corresponds to a spectrum, where the "scores" here are the projections of these spectra on a new set of rotated, orthogonal axes, which are called principal components, or "PCs." The first two PCs here account for most of the variance in the data sets. For example, in panel (c), the "(97.96%)" value in the *x* axis label "Scores on PC1 (97.96%)" indicates that PC1 accounts for 97.96% of the variance in the data set.

one typically looks for a discontinuity in the plot (a "knee") where the slope of the results (as viewed from right to left) changes. This point in the plot is often taken as the number of PCs that describe a data set. Accordingly, these plots suggest that five PCs describe the cellulose data sets quite well. That is, while, in some cases, a two-PC (two-abstract factor) model may adequately describe the cellulose data sets because of the high amount of variance it captures, the higher PCs appear to contain some relevant (non-noise)

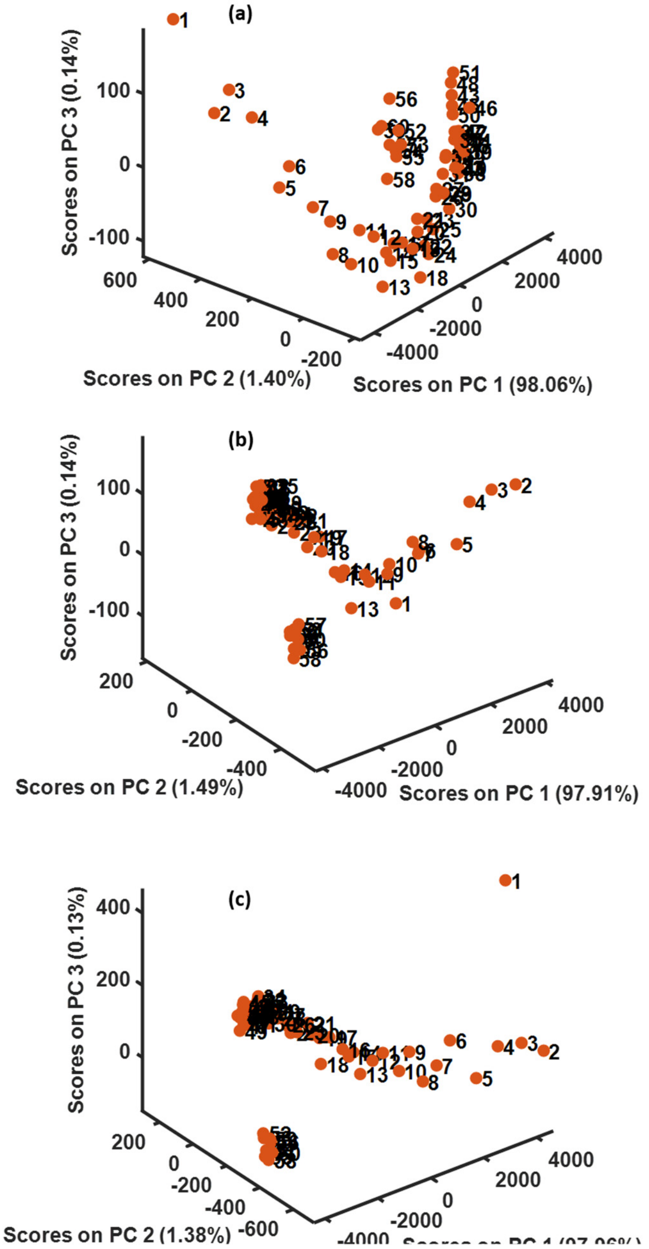

FIG. 8. Three-dimensional PCA scores plots of the (a) C 1s, (b) O 1s, and (c) concatenated C 1s and O 1s spectra of the cellulose data set after preprocessing by mean centering. PC 3 here only accounts for a small amount of the variance in these data sets.

information. These results are typical of the PCA of many data sets. The scree plots for the tartaric acid data sets in Fig. 11 show that the first two PCs account, on average, for a lower fraction of the variance in the data sets than for the cellulose data set, and that a total of four to five PCs probably describe these data sets.

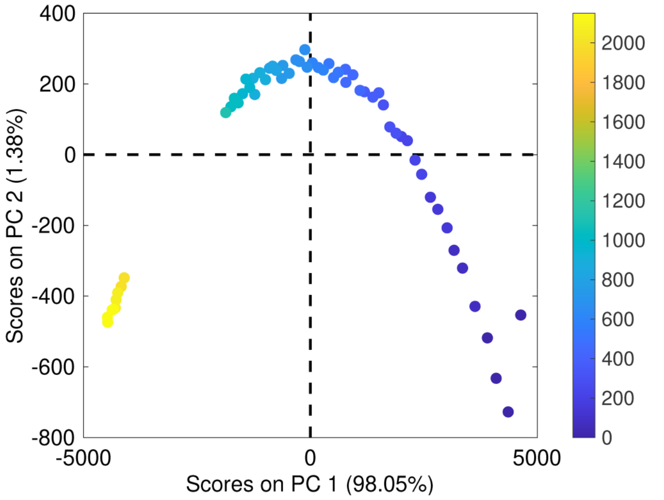

FIG. 9. Two-dimensional PCA scores plots of the concatenated C 1s and O 1s narrow scans of the cellulose data set with the elapsed time shown as the color of the data points. In PCA, the spectra are, in essence, plotted as single points in a hyperspace, where the axes of this coordinate system are rotated to align with the greatest amount possible 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."

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 examined here. It is often better *not* to preprocess the data prior to this type of reconstruction, and the data were not preprocessed in the analyses that are now described. Figure 12 shows reconstructions of the first spectrum of the tartaric acid C 1s data set from increasing numbers of abstract factors. The high RSD values and the presence of significant structure in the residuals of the reconstructions from one [Fig. 12(a)] and two [Fig. 12(b)] abstract factors suggest that the spectrum is inadequately described by one or two PCs. Reconstructing the spectrum from three or more abstract factors yields spectra that no longer change significantly. However, the residuals in Fig. 12(c) still show some structure, which mostly disappears when the spectrum is reconstructed from four abstract factors. The loadings of these factors in Fig. 13 suggest that abstract factors 1–4 have meaningful structure and that four abstract factors (PCs) adequately describe this data set. Nevertheless, like scree plots and cross-validation, this is an inexact approach. There appears to still be a small amount of structure/information in abstract factors 5 [Fig. 13(e)] and 6 [Fig. 13(f)]. Nevertheless, these factors are becoming noisier, which also suggests that they are contributing less useful information to the analysis. With the exception of the first abstract factor, the negative peaks in the abstract factors in Fig. 13 make them hard to interpret chemically. As noted in Paper I,[53] the approach of reconstructing data from abstract factors should be applied to different spectra in a data set. Figure 14 shows the reconstruction of the 50th C 1s spectrum in the tartaric acid

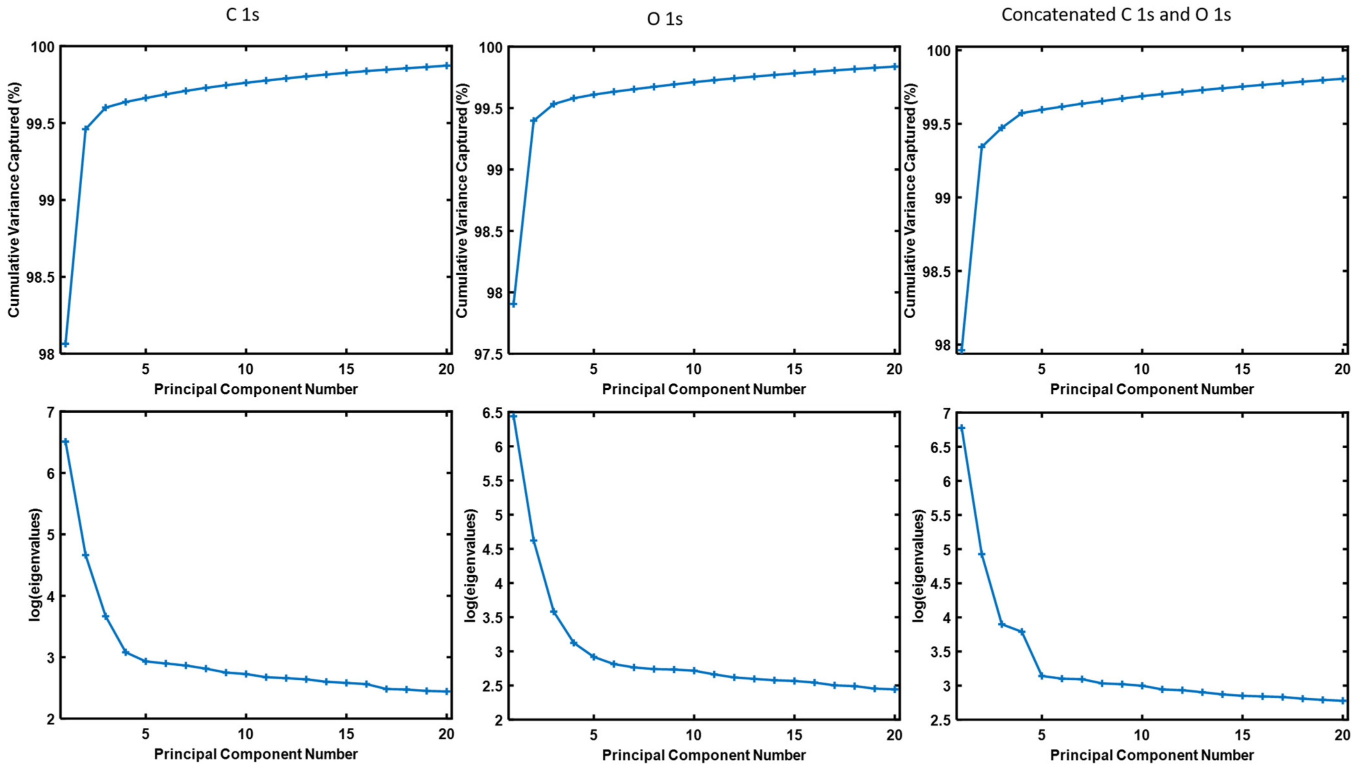

FIG. 10. Scree plots from the PCA analyses of the cellulose data sets after mean centering. Scree plots show the amount of variance captured by a PCA model vs the principal component number.

data set. Fortuitously, this spectrum is very well described by the first abstract factor. That is, if only this spectrum were reconstructed, one might conclude that only one, or perhaps two, abstract factors are necessary to describe this data set. Finally, one can denoise/smooth a spectrum by reconstructing it from a limited number of abstract factors. This removal of noise is illustrated in Fig. 12—compare the noise levels on the spectra reconstructed from three and four abstract factors to the original spectrum, i.e., the spectrum reconstructed from all the abstract factors.

G. Return to the original data after an informatics analysis to confirm findings

With modern chemometrics/informatics software, it is easy to perform many different analyses on data sets. However, the predictions and findings from these analyses should always be confirmed in the original data. We now follow this procedure for the outlier in the cellulose data set (spectrum 1) suggested in Figs. 7–9. Figure 15 shows the raw, concatenated C 1s and O 1s data for the first three narrow scans of this data set. Included in this plot are enlarged views of the tips of the peaks. While one might argue that these three scans are not terribly different from each other, Fig. 15 suggests that spectrum 1 is indeed different from spectra 2 and 3. These results underscore the ability of chemometrics/informatics methods to differentiate between spectra, even when the differences between them are fairly subtle. These differences might have been missed otherwise.

H. Multivariate curve resolution (MCR) (of the cellulose data set)

MCR has become popular among chemometricians as it offers various advantages over PCA. For example, because of the non-negativity constraints that are usually applied in MCR, MCR loadings have the appearance of real spectra, making them easier to interpret, while PCA loadings often have negative peaks (see, for example, Fig. 13). Figure 16 shows scores and loadings plots for two-component MCR models of the C 1s, O 1s, and concatenated C 1s and O 1s data of the cellulose data set. In all three cases, more than 99% of the variance in the data sets is captured by two components. In each case, the scores on one of the components rise monotonically, while the scores on the other fall monotonically. Accordingly, one might expect that the first and last scans of the data sets would basically be the same as the MCR components, which is confirmed in the last row of Fig. 16. In other words, MCR makes the interesting prediction that the cellulose spectra are essentially linear combinations of the first and last spectra in this data set. Chemically, this implies that there are two chemical states for

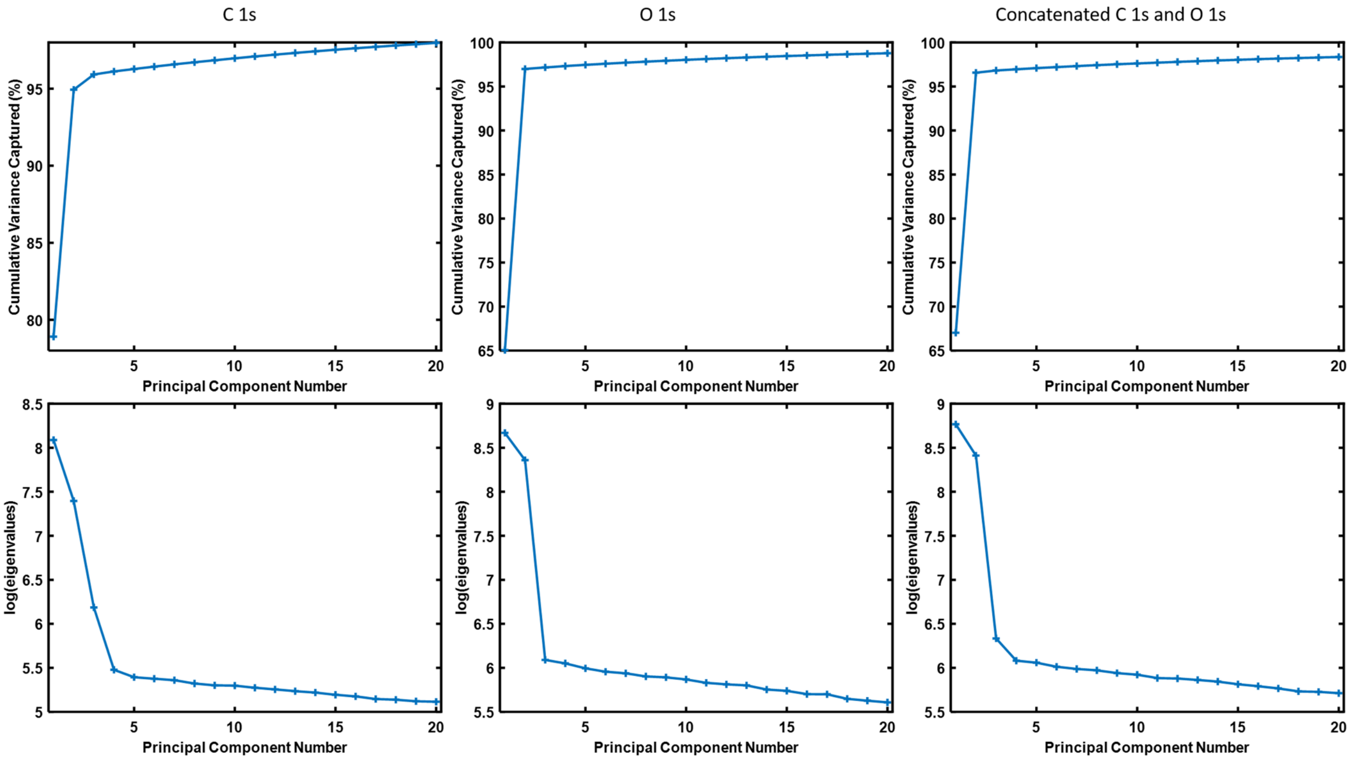

FIG. 11. Scree plots from the PCA analyses of the tartaric acid data sets after mean centering. Scree plots show the amount of variance captured by a PCA model vs the principal component number.

the material: an undamaged state and a damaged one. As we will see below, it is not always the case that series of spectra can be conveniently described with only two components.

The following are additional conclusions/considerations from the MCR analyses of the cellulose data set in Fig. 16.

(i) The break in the cellulose data between spectra 51 and 52, which was apparent in the summary statistic [Figs. 6(a) and 6 (b)] and PCA (Figs. 7–9) analyses, is also obvious in the MCR scores plots for all three models (the C 1s, O 1s, and concatenated C 1s and O 1s models) [see Figs. 16(a)–16(c)].

(ii) While the PCA analyses in Figs. 7(a)–7(c), 8 and 9 suggest that spectrum 1 is an outlier in the C 1s, O 1s, and concatenated data sets, this effect was only observed in the MCR model of the O 1s spectra [see Fig. 16(b)], where spectrum 2 has the highest and lowest scores on components one and two, respectively. It is not entirely clear why this effect is only apparent in Fig. 16(b). Nevertheless, these somewhat different results from PCA and MCR underscore the importance of using multiple informatics methods to analyze data sets. The different mathematics of these methods probe data sets differently.

(iii) Preprocessing usually affects informatics analyses. For example, the outlier in the cellulose data set became apparent when the data were mean centered prior to the PCA analysis (Figs. 7–9), but not when no preprocessing was applied (see Fig. S2 in the supplementary material).[65] No preprocessing was applied to the cellulose data set prior to MCR. However, it is incorrect to mean center (or autoscale) spectra prior to MCR because of its non-negativity constraints, unless special considerations/changes are applied in the analysis.

(iv) In Fig. 16, we obtain loadings with very similar shapes from the C 1s, O 1s, and concatenated data sets, which suggests that all of these analyses are revealing/exposing the same chemical variation/evolution in the data.

(v) The relative concentrations of the different chemical components of a material are often more obvious/better preserved in the loadings obtained from the concatenated data set. For example, in Fig. 16, the loadings of the concatenated data set are closer to the real spectra. MCR results from concatenated data can be easier to interpret—concatenation forces the relative areas of the peaks/different signals to be constant.

I. Peak fitting the MCR factors (of the cellulose data set)

We believe we now show for the first time that chemical information can be extracted from MCR factors of XPS data sets by peak fitting. Such fits can help us understand the chemical changes

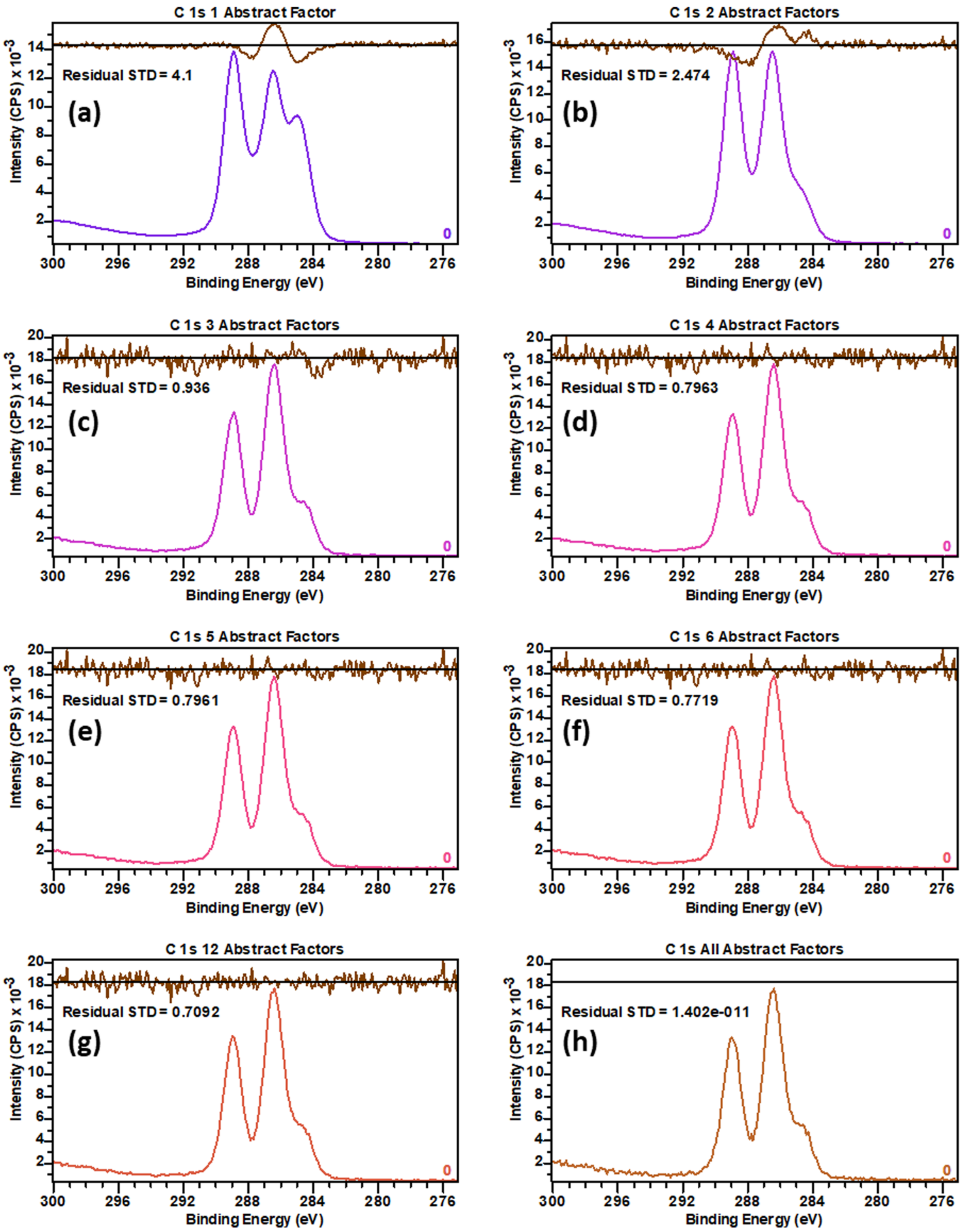

FIG. 12. Reconstructions of the first, unpreprocessed, C 1s spectrum from the tartaric acid data set using (a) one, (b) two, (c) three, (d) four, (e) five, (f) six, (g) twelve, and (h) all the PCs (abstract factors). The residuals of these reconstructions are shown above the spectra in each panel. As the number of PCs used to reconstruct the raw data increases, the residuals and residual standard deviations (residual STDs) decrease.

that may take place in a material. In this section, we focus on fitting MCR factors of C 1s narrow scans. Figure 17 shows both the fits of the two MCR components of the cellulose C 1s spectra shown in Fig. 16, and the fits of the first and last C 1s narrow scans in this data set. The protocol used in these fits was determined as follows. First, the spectra and MCR components were fit with four synthetic peaks (Voigt functions with a mixing parameter, m, that was allowed to vary from 0 to 100)[62] that represent the following

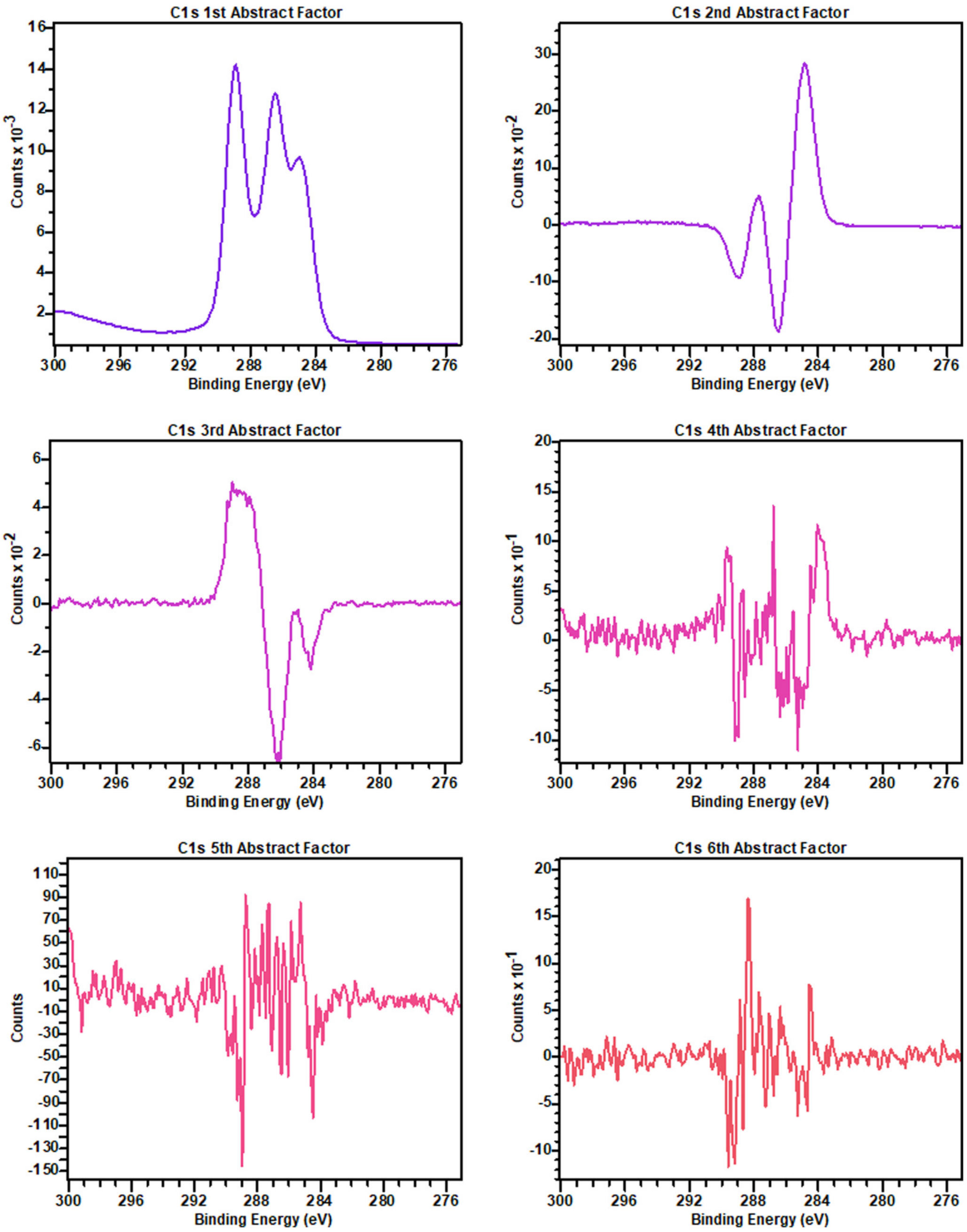

FIG. 13. First six loadings (abstract factors) from a PCA analysis of the C 1s tartaric acid data set in which no preprocessing was performed on the data. The "loadings" contain the contributions of the original axes to the new (rotated) axes in PCA. These abstract factors were used to regenerate the spectra in Figs. 12 and 14. For these analyses (in Figs. 12–14), the binding energy scale was adjusted to align the O 1s spectra peak maxima in all the spectra. Appendix contains more details of this adjustment.

chemical states: C—C/C—H (peak 1), C—O (peak 2), O—C—O (peak 3), and carboxyl carbon/C(O)O (peak 4),[31] where these peaks were constrained to have equal widths, their Gaussian contributions/contents were optimized such that all four peaks in a fit had the same value of the mixing parameter, and the position of the highest binding energy peak was constrained to be at least 1.2 eV above the previous peak. This last constraint was only necessary in the fits of the first two narrow scans in the data set. These

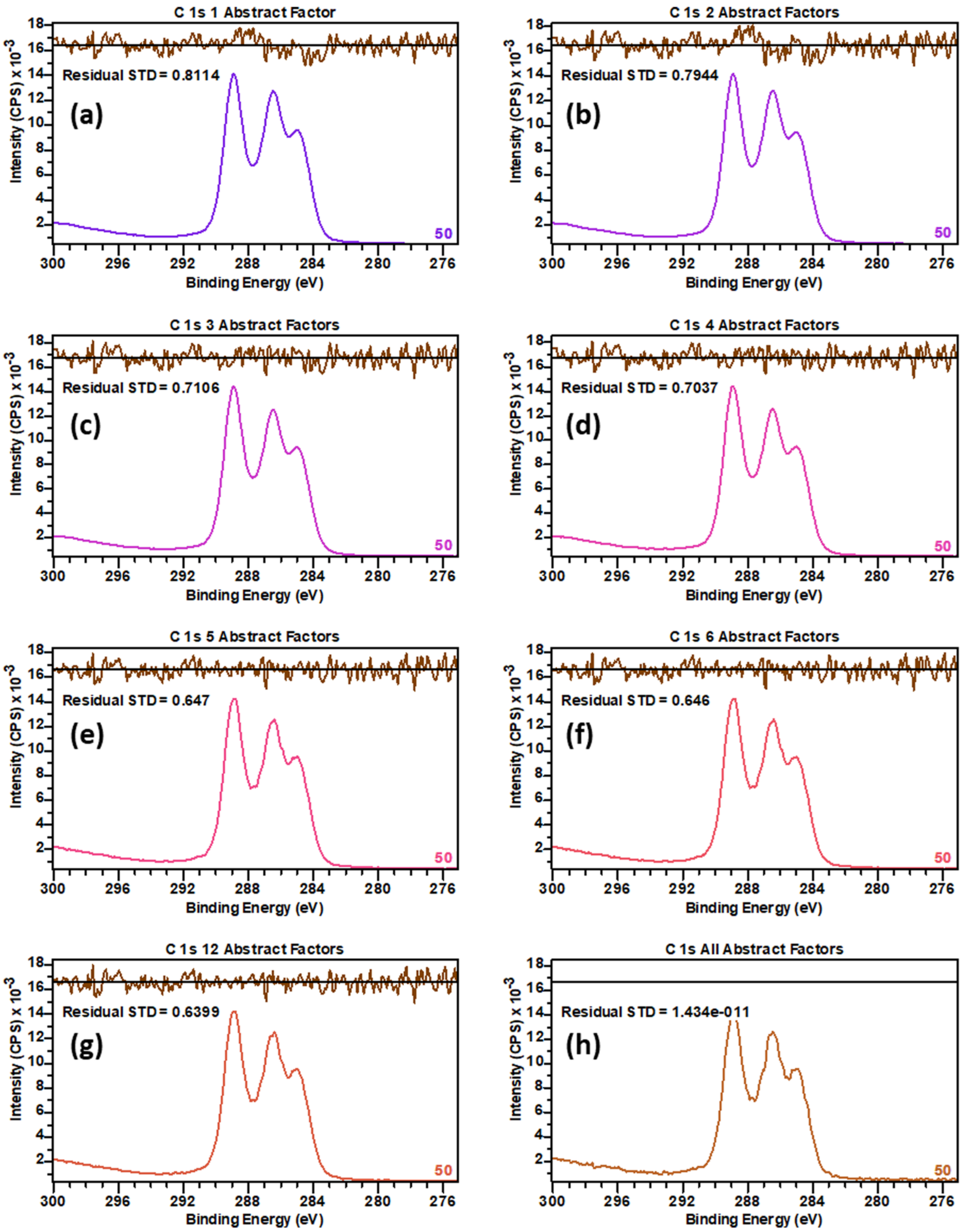

FIG. 14. Reconstruction of the 50th unpreprocessed C 1s spectrum from the tartaric acid data set using (a) one, (b) two, (c) three, (d) four, (e) five, (f) six, (g) twelve, and (h) all abstract factors. The residuals of these reconstructions are shown above the spectra in each panel. As the number of PCs used to reconstruct the raw data increases, the residuals and residual standard deviations (residual STDs) decrease.

same chemical states of carbon were used in a recent XPS study of cellulose.[39] A universal polymer Tougaard background was used for all the fits.[63] No other constraints were applied. These fits indicate that significant changes take place during the XPS analysis of cellulose, i.e., the first and last narrow scans (and also the two MCR components) are very different. As expected from the results in Fig. 16 for cellulose, the fits to MCR components 1 and 2 are quite similar to the fits of the first and last C 1s narrow scans in the data set, respectively. These results are also consistent with significant sample degradation during the analysis. For example, an obvious change in the spectra is the decrease in intensity of the C—O peak and the concomitant increase in intensity of the C—C/C—H peak, which suggests carbonization of the material.

The protocol used to fit the C 1s narrow scans and MCR components in Fig. 17 was applied to all the spectra in the cellulose data set. Figure 18 shows the percent areas of the four synthetic peaks used in these fits plotted as a function of sample irradiation time [not scan number as in Figs. 6 and 16(a)–16(c)]. This plot clearly shows the break in the data that is apparent in Figs. 6 and 16, indicating that the latter data points (after the break) are an extension of the earlier ones. Figure 18 also shows the total C/O area ratio for cellulose as a function of x-ray exposure. The increase in this ratio suggests sample damage, and it is also consistent with the increase in the area of peak 1 and the decrease in the areas of peaks 2 and 3 in Fig. 18. Over the course of this damage, peak 4 (the carboxyl signal) increases and then begins to decrease, suggesting it is an intermediate (there should not be any carboxyl functionality in pure cellulose, see Fig. 1). Sample damage is expected to randomize and/or introduce new chemical states into a material. Therefore, the best synthetic peaks for the fits to the data may change over the course of the analysis. In particular, a more random material is often better described by a more Gaussian fit component. We optimized the mixing parameter, m, in all the fits. However, there was no clear trend in the results, e.g., the average value of m for these scans was 10., the standard deviation here was 11, and, in general, for each fit, the plot of the error in the fit versus m was flat (at a minimum value) from m = 0 to m = 20–40. Even though m did not change/show a trend in these fits, we still believe it is a good idea to check for this possibility.

While MCR can be extremely useful in understanding series of spectra, MCR components may contain artifacts or anomalies. For example, component 1 [see Fig. 17(a)] contains a small carboxyl peak that is not in the first spectrum in the data set. A more subtle example of an artifact is on the right side of component 1. Here, as indicated in the residuals, component 1 is not precisely fit with the first synthetic peak. Nevertheless, in spite of these artifacts, MCR is an extremely powerful tool for understanding series of spectra. However, the possibility of artifacts in an MCR analysis 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—an artifact created by one chemometrics/informatics analysis may not be present in the results of another chemometrics/informatics tool.

J. Identification of intermediates in an MCR analysis

In addition to the methods mentioned in Sec. III F, another way to determine the number of abstract factors that describe a data set is to create models with successively larger numbers of factors in them, evaluating the chemical reasonableness of the models. As noted above, PCA of the mean-centered C 1s data set of tartaric acid suggested that a minimum of four abstract factors is necessary to describe the data set. Figure 19 shows MCR models of the tartaric acid data set with three, four, five, and six factors. The MCR models with three to six factors depicted in Fig. 19 all capture more than 99% of the variance in the data, which is definitely a positive sign. The loadings of the three-factor model [Fig. 19(b)] are smooth and appear to be chemically reasonable. The scores on these components [Fig, 19(a)] suggest that there is an initial state (described by component 3), an intermediate state (described by component 1, which grows in and then disappears), and a final state (described by component 2) for this material. However, the scores here are somewhat unreasonable because those corresponding to the initial and final states do not change in a monotonic fashion. These results suggest that a model with more components should be considered. Figures 19(c) and 19(d) show the four-component MCR model of the tartaric acid C 1s data set. Again, the loadings [Fig. 19(d)] are smooth and chemically reasonable. The corresponding scores plot [Fig. 19(c)] indicates that there are two initial states (components 3 and 4), one intermediate state (component 1), and one final state (component 2). However, as before, the initial and final states do not change in an entirely monotonic fashion. These results again prompted us to consider a model with more factors.

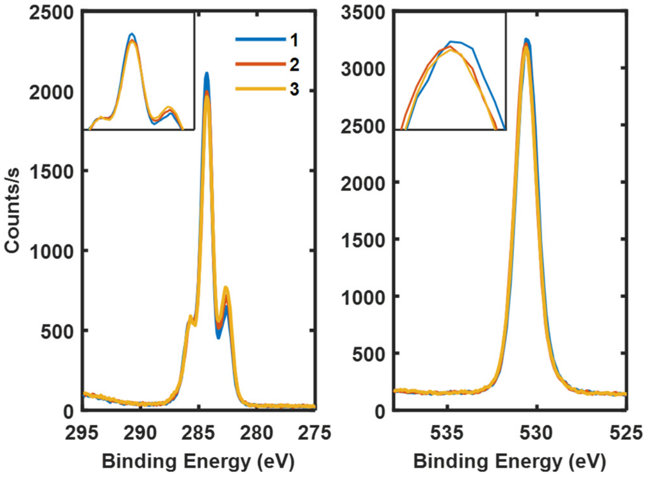

FIG. 15. First, second, and third concatenated C 1s (left) and O 1s (right) spectra of the cellulose data set. The insets of these panels show zoomed-in views of the data.

Both the five and six-component models of the tartaric acid data set are satisfactory in many ways. First, all of the initial and final states in the models change in a monotonic fashion. In addition, both models have scores and loadings that are not overly noisy, although we would be uncomfortable with any more noise in the results than that in the six-component model because noisy loadings suggest that we are fitting/adding noise in a model. The five-component model decomposes the spectra into two initial states (components 4 and 5), two intermediate states (components 1 and 2), and one final state (component 3). These results raise the interesting possibility that component 3 (the final state) is also an intermediate state, i.e., that the scores on this component will also eventually decrease. Obviously, more scans would be

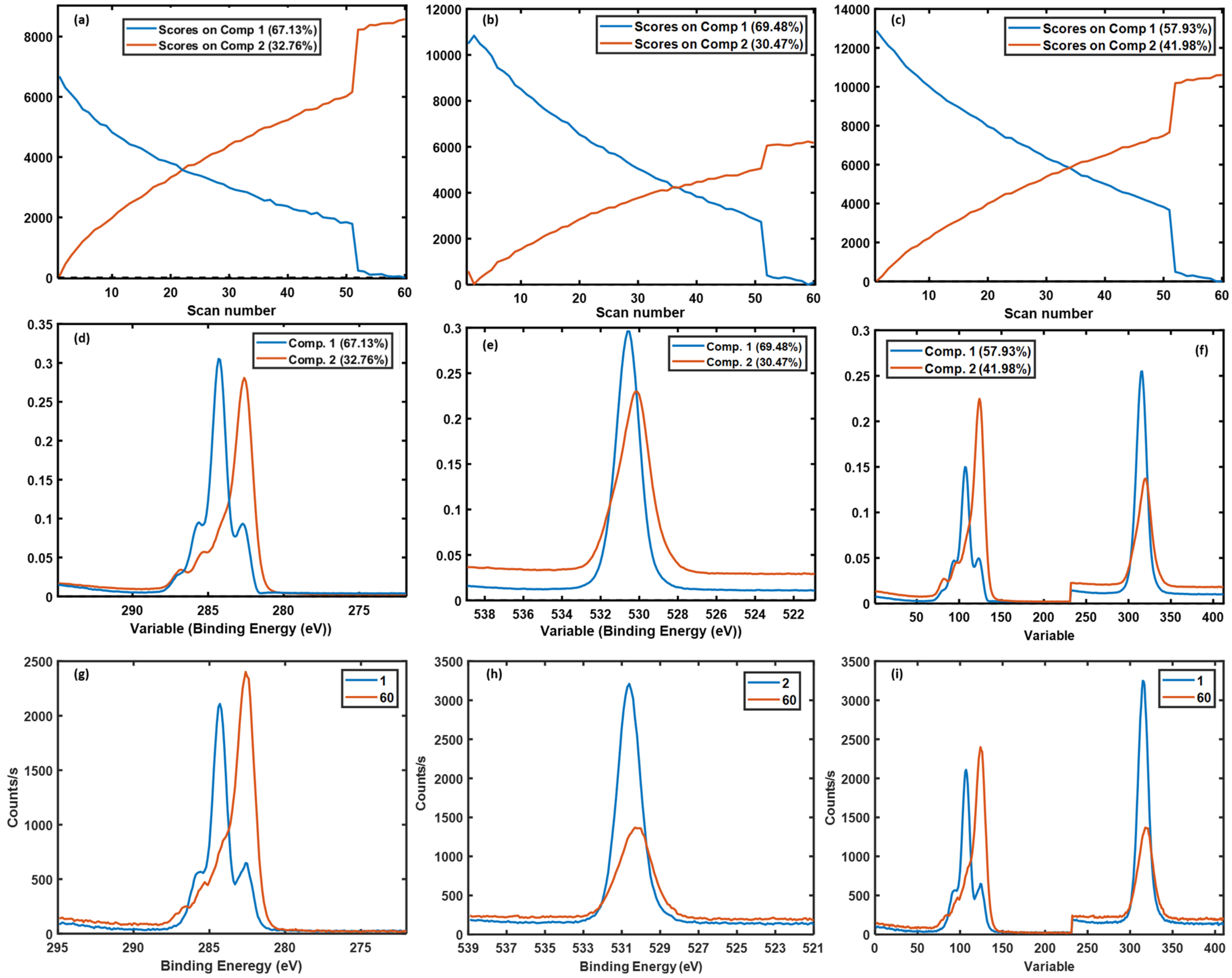

FIG. 16. MCR of the 60 XPS spectra in the cellulose data set. (a)–(c) Scores plots, (d)–(f) loadings on components 1 and 2, and (g)–(i) plot of the first and last scans from the C 1s [(a), (d), and (g)], O 1s [(b), (e), and (h)], and concatenated C 1s and O 1s [(c), (f), and (i)] data sets. The scores in (a)–(c) are the projections/contributions of the loadings (abstract factors) to the original spectra. That is, the spectra are represented as a linear combination of the two components shown in (d)–(f).

needed to confirm or reject this hypothesis. The reduced carbon (C—C/C—H) signal in the loadings increases from the initial states through the intermediate states to the final state, suggesting a carbonization of the material.

The six-component MCR model of the tartaric acid data set presents a particularly interesting view of the evolution and degradation of the material. The C 1s spectrum of pure tartaric acid should contain two equal-area, chemically shifted signals corresponding to the two chemically different carbons in the molecule. However, in addition to the two expected signals, the initial states in the six-component model (components 5 and 6) also show reduced carbon, and these components do not have the two main signals in exactly the expected 1:1 ratio. That is, these initial states suggest the presence of adventitious carbon contamination. Component 4 then grows in as the initial states (components 5 and 6) disappear. Interestingly, component 4 contains the two equal-area signals expected from tartaric acid, with little reduced carbon. These results suggest that the x-ray beam and photoelectrons “clean” the surface of adventitious carbon. Thereafter, two intermediate states (components 1 and 2) and a final one (component 3) appear. Again, the final state (component 3) may actually be an intermediate. We believe that this analysis is the first time these types of intermediate states have been shown/suggested in an XPS degradation study. MCR is a

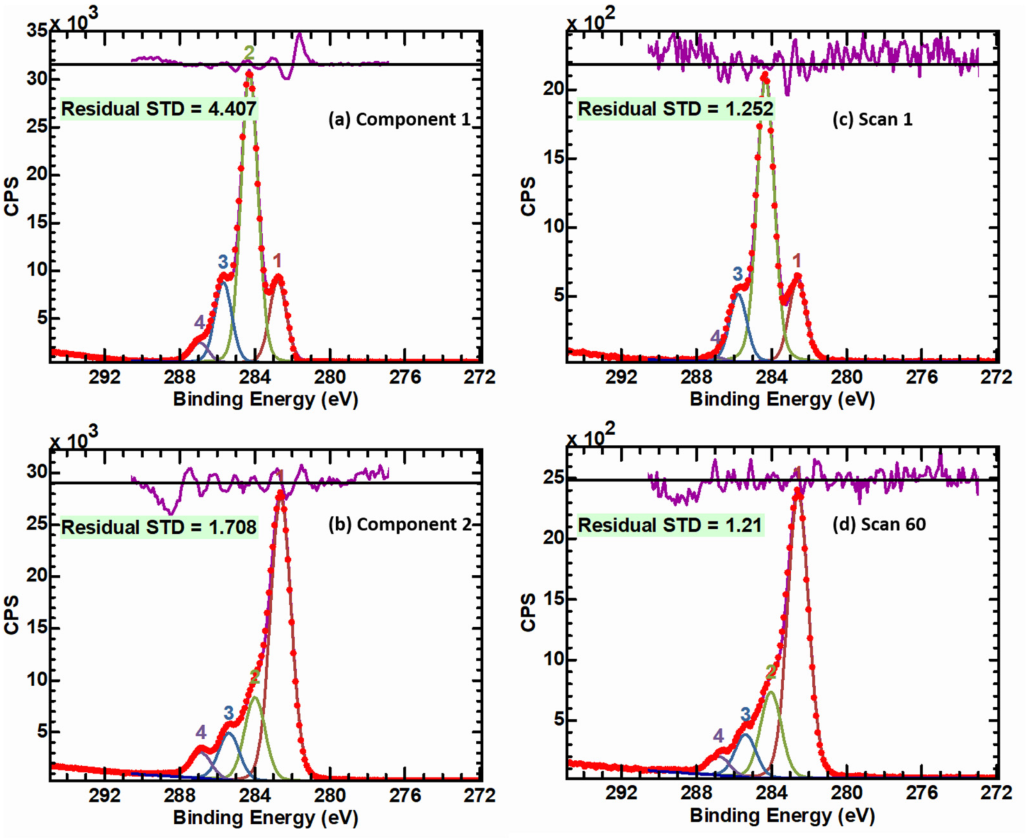

FIG. 17. Peak fitting of the two MCR components used to describe the cellulose data set and of the first and last spectra of this data set. See the text for the fitting protocol. The abstract factors (components) here were multiplied by a factor of 10^3.

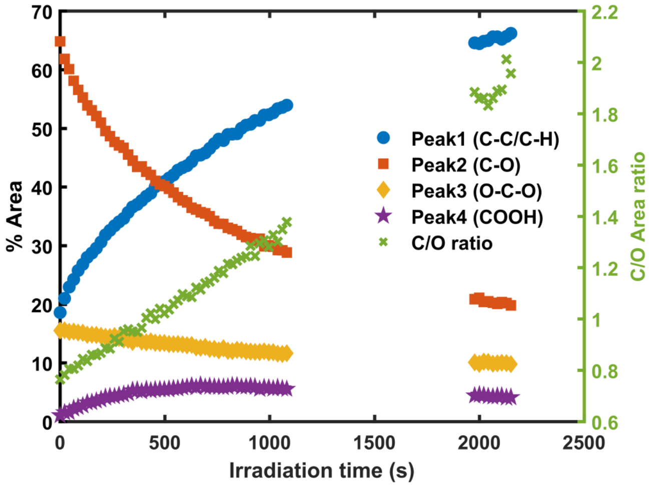

FIG. 18. Areas (as percentages of the total area) of the four synthetic peaks in the peak fits to the C 1s narrow scans in the cellulose data set (see the text for the fitting protocol).

powerful tool for these types of analyses. However, this all begs the question, how do we actually know that the degradation of this material involves multiple intermediates? First, intermediates are present in many complex chemical reactions, and the degradation of tartaric acid is probably complex. Second, the five-component MCR model, where component 1 has grown in (around scan 22), should mostly describe the data (the score on the other components is low at that point). As shown in Fig. S3 in the supplementary material,[65] component 1 and scan 22 are indeed very similar, i.e., the model appears to be representing the data at this point, which suggests it has some validity. In summary, these results suggest the interesting possibility that MCR can be used to uncover the underlying chemistry, including intermediates, in complex XPS data sets.

To better understand their chemistry, we peak fit the MCR factors in the six-component model in Fig. 19(h). To find an appropriate protocol for this fitting, the raw spectra in the data set were first fit. This protocol consisted of three synthetic peaks with equal widths representing the C—C/C—H (peak 1), C—OH (peak 2), and COOH (peak 3). No other constraints were applied to these fits. The optimal m values (mixing parameters) for the fit

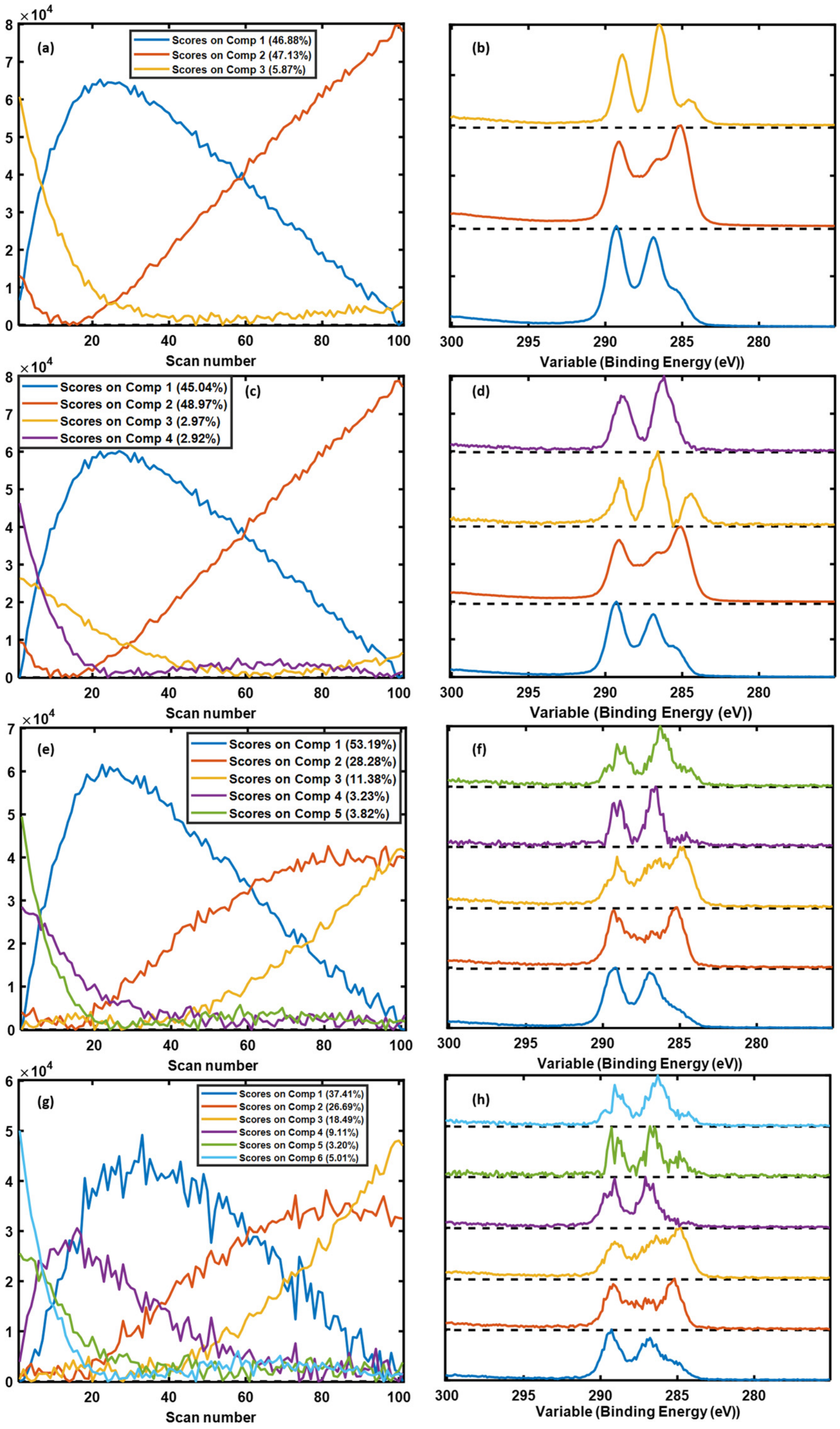

FIG. 19. MCR analyses with different numbers of components of the C 1s narrow scans of the tartaric acid data sets. MCR scores (left) and loadings (right) from models with three [(a) and (b), four [(c) and (d)], five [(e) and (f)], and six [(g) and (h)] components. The scores and loadings become noisier as the number of components in the models increase, i.e., increasing amounts of noise are being incorporated into the models.

components ranged from 0.6 to 0.8, and the average spacings between the first two peaks and the last two peaks were 1.72 ± 0.04 and 2.33 ± 0.09 eV, respectively. These results prompted us to fit the loadings in Fig. 19 with three synthetic peaks of equal widths, spacings of 1.72 and 2.33 eV, and an m value of 0.7. Very good fits were obtained with this protocol (see Fig. 20). There is little evidence of sample charging in these fits, i.e., the first peak stayed at a relatively constant position of 285.0 ± 0.2 eV, where the shifts in the

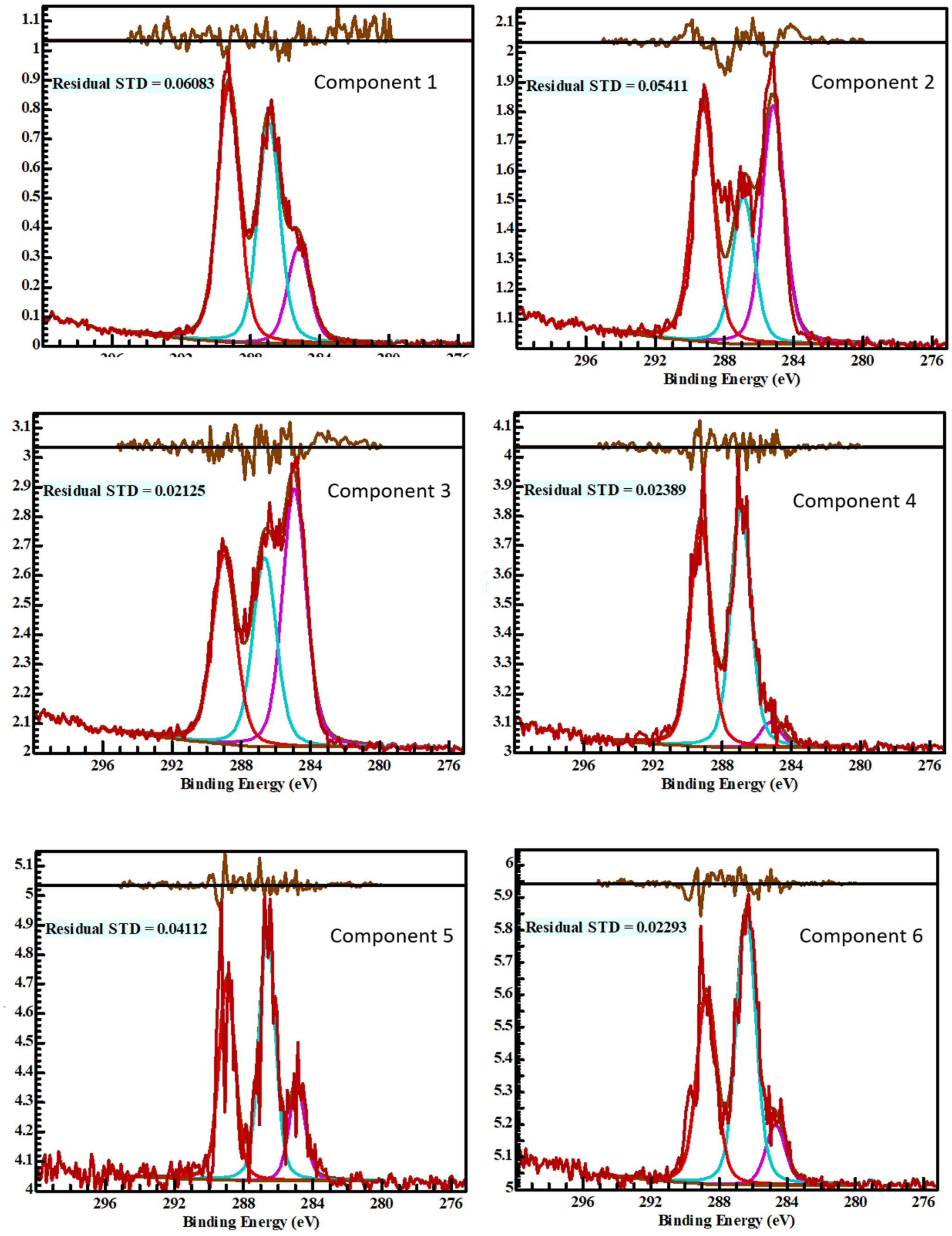

FIG. 20. Peak fits of the MCR loadings of the six-component MCR analysis of the tartaric acid C 1s data set in Fig. 19(h). See the text for the fitting protocols used here.

peak positions were not monotonic. Figure 21 is a plot of the areas of the three synthetic peaks used to fit the six MCR components. It shows that (after the initial, apparent cleaning of the material) the amount of reduced carbon increases monotonically from the earlier to the later components, while, overall, the areas of the two oxygen containing peaks decrease somewhat. These results suggest that, as was the case with cellulose, x-ray exposure and photoelectrons carbonize tartaric acid.

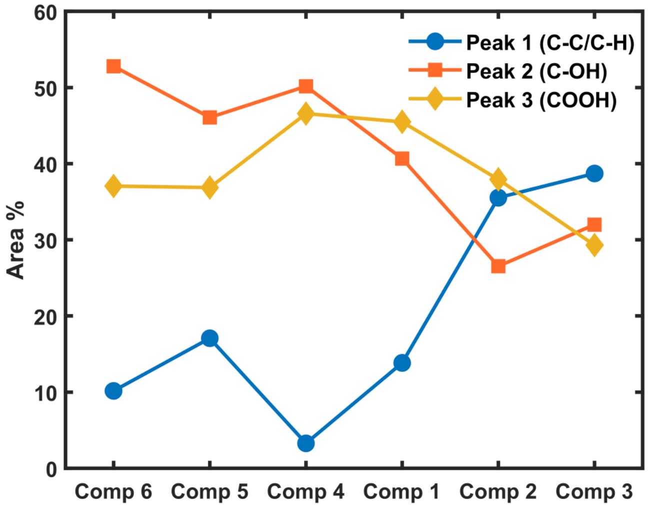

FIG. 21. Percent areas of the synthetic peaks used to fit the MCR components of the tartaric acid data set in Fig. 20.

K. Cluster analysis

Cluster analysis is another widely used EDA method. Cluster analysis groups similar samples/spectra according to their distances in a multidimensional space. The resulting groupings are typically shown as a dendrogram. Figure 22(a) shows the cluster analysis/dendrogram of the tartaric acid C 1s data set that grouped the data into five classes. (The number of clusters/classes in a cluster analysis can be chosen by the user.) Interestingly, the spectra naturally clustered in this analysis in the same consecutive order that they appear in in the data set. In other words, the series of clusters reflects the evolution/changes that are taking place in the data set. Figure 22(b) shows the average spectrum for each of the five clusters. As in the MCR analysis [Fig. 19(h)], these spectra indicate that the sample is carbonizing with x-ray exposure (the reduced carbon peak grows in). Thus, cluster analysis confirms the other results in this work—as previously noted, it is good to verify the trends/conclusions of one chemometrics/informatics method with others. Cluster analysis was also performed on the O 1s and concatenated data sets. As with the C 1s data, the clustering took place consecutively. However, different spectra appeared in the different clusters, i.e., the groupings were not the same. Cluster is relatively

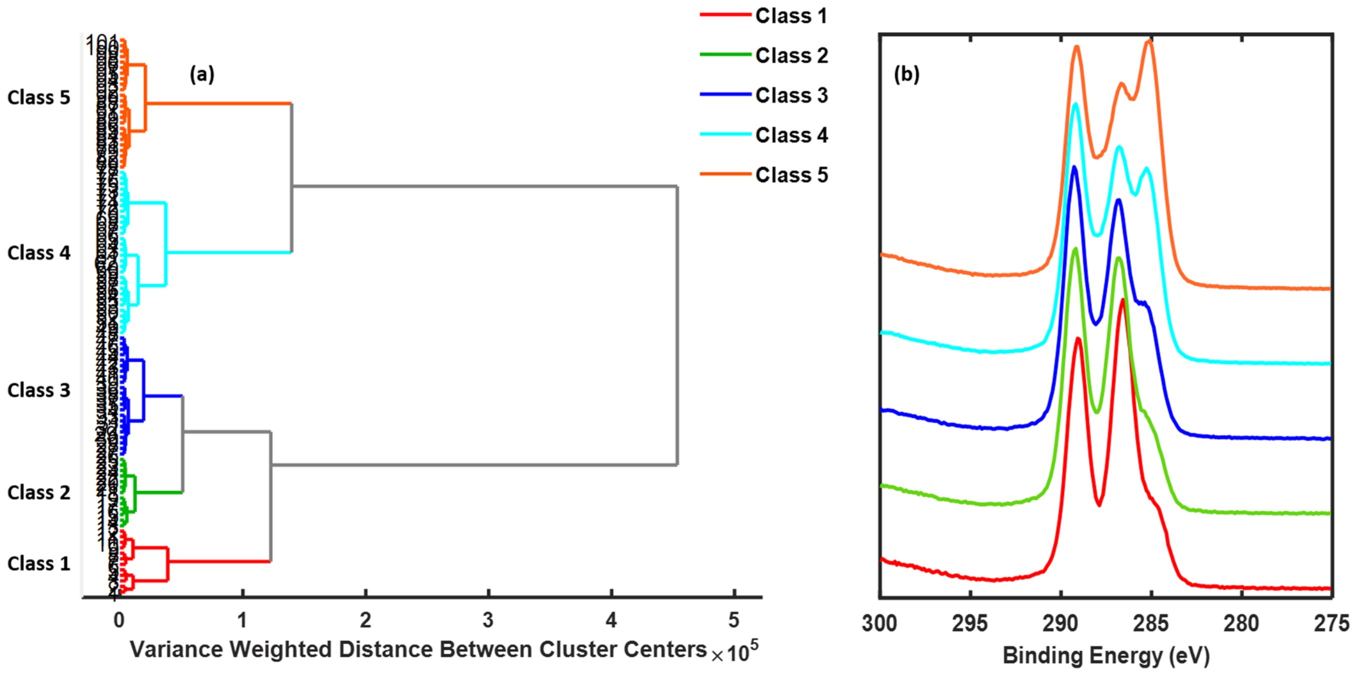

FIG. 22. (a) The dendrogram from the cluster analysis of the tartaric acid C 1s data set and (b) the average spectrum of each cluster identified in the dendrogram in (a). The numbers on the left side of the figure correspond to the spectra in the data set. The user can select the number of clusters in a cluster analysis. That is, imagine moving a vertical line back and forth across the dendrogram, e.g., if the line were at 3 × 10^5 on the *x* axis, two clusters would have been selected. Five clusters have been selected here, which are color coded and numbered. The distances between the data points (spectra) or clusters of data points in a dendrogram are given by the lengths of the lines parallel to the *x* axis.

easy to apply and conceptually simpler than some other chemometrics/informatics methods. However, cluster analysis does not generally provide as much insight or information as MCR or PCA. For example, although the cluster analysis in Fig. 22 groups the data in a reasonable way, it does not suggest or reveal the presence of intermediates. Cluster analysis 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 spectra in Fig. 22(b) could be peak fit.

L. Summary of the results

The following is a summary of the information provided by each of the chemometrics/informatics methods applied to the cellulose and tartaric acid data sets, including where the analyses agreed, any problem areas, and differences in the results.

(i) For the cellulose data set, in plotting (examining) the raw data, PRE, PCA, and MCR showed a break in the data set. This break would probably be harder to identify in a cluster analysis. Two- and three-dimensional PCA scores plots indicated that the first scan in the data set is an outlier, where the presence of this outlier was confirmed by returning to the original data. Neither PRE nor MCR produced this same result, although MCR of the O 1s data suggested that the second data point (spectrum) in the data set may be an outlier. Scree plots suggested that five PCs describe the cellulose data set, although most of the variance in the data sets was captured by only two PCs—for at least some applications, a two-abstract factor MCR model will be reasonable because it captures so much of the variance in the data set. However, as is often the case in factor-based analyses of data, even though higher abstract factors may account for quite small amounts of the variance in a data set, they may still contain useful information about it. Overall, MCR indicated that the cellulose data set could be quite well described by only two-abstract factors, which closely resembled the first and last spectra of the data set. In other words, MCR makes the interesting prediction that the spectra in the data set are essentially linear combinations of two-abstract factors (basically the first and last spectra of the data set), which represent two chemical states. In general, concatenated (combined C 1s and O 1s spectra) gave the most chemically meaningful results in the MCR analysis. MCR is unique in its ability to produce abstract factors that closely resemble real, underlying spectra. Peak fitting of the original C 1s data, and of the two MCR components that describe it, better revealed the significant chemical changes the material underwent as it carbonized. While PCA is, and will continue to be, extremely important in chemometrics/informatics, its orthogonality constraints do not allow it to produce the same type of intuitive information.

(ii) None of the chemometrics/informatics methods applied to the tartaric acid data set (PRE, PCA, MCR, and cluster analysis) suggested that there were any outliers or discontinuities in it. Rather, PRE and especially PCA, suggested quite smooth trajectories (changes) for the spectra. Scree plots suggested that four to five PCs describe the tartaric acid data set. Reconstruction of this data set from abstract factors suggested that even a 6th PC may contain meaningful information. As is common in MCR analyses, models of the data set with successively more abstract factors were considered. Models with three and four abstract factors were not favored because of the lack of chemical reasonableness in their scores. The scores in the models with five and six-abstract factors were both more chemically reasonable *and* they suggested the presence of intermediates in the decomposition of this material. In our opinion, this is an extremely important result. It is the first time something like this has been observed. It suggests that MCR can be used to identify intermediates/intermediate states in XPS data sets where decomposition and other chemical changes are occurring. The PCs from the PCA analysis were not fit in either this analysis or the previous one as their more abstract nature simply does not allow it. While cluster analysis provided a series of average spectra that seemed to reveal the chemical evolution of the tartaric acid as it degraded, it did not suggest intermediates in the decomposition of the material. Also, while cluster analysis is easier to apply, and we *do* recommend it for EDA analyses (years ago, some of us successfully analyzed hyperspectral ToF-SIMS images using cluster analysis that produced very interpretable and useful results[64]), we have not, in general, found it to be as powerful as MCR.

IV. SUMMARY AND CONCLUSIONS

This article shows the application of some of the more common EDA methods to the analysis of two XPS data sets. It is intended to be a guide to using these methods. The current trend in XPS is to collect increasingly large data sets in degradation, depth profiling, *operando*, and imagining studies, which should make chemometrics/informatics techniques increasingly relevant in the field. The first step in an informatics analysis is to gather and consider whatever information one has about one's material. A next logical step is to plot the raw data in different ways. One should then develop a strategy for the analysis of one's data. Next, the data are preprocessed, and chemometrics/informatics analyses are performed. Summary statistics are a quick method of analyzing data sets, where PRE is often sensitive to their underlying structure. PCA is another "first technique" that should be applied in chemometrics/informatics analyses. Considerations for PCA include the number of PCs (abstract factors) to keep in an analysis, different preprocessing methods, and different ways of plotting/representing the results, including the addition of extra information to scores plots. Scree plots, reconstruction of the data from abstract factors, and consideration of the chemical reasonableness of a model can be used to determine the number of abstract factors that describe a data set. One should return to the original data after an informatics analysis to confirm predicted data structures or outliers in the raw data. We strongly recommend MCR as an EDA method for uncovering the underlying structure of complex XPS data sets. For example, MCR analysis of the cellulose data suggested that two states, representing the damaged and undamaged material, describe the data well. These loadings closely resembled the first and last scans of the data set. Concatenation of data can be useful in MCR (and PCA) analysis—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. MCR factors of XPS narrow scans may be peak fit to better reveal their underlying chemistry. The protocol for peak fitting MCR factors may be based on fits of the raw data. The C/O area ratios from the C 1s and O 1s narrow scans in the cellulose data set correlated with the increase in reduced carbon in the material and were consistent with the proposed carbonization of the material. The degradation of tartaric acid appeared to be more complex. Models based on two to four abstract factors were not entirely chemically reasonable. Five or six-abstract factors appeared to better describe the data, where these models raised the possibility of a contaminated surface state, a cleaned surface state, and multiple intermediates. We believe this is the first time the evolution of an XPS data set has been revealed in this way. These data also show that the sample is carbonizing with x-ray exposure. The MCR loadings of the six-abstract factor model were peak fit. Finally, we showed cluster analysis of the tartaric acid data set. The average spectra of each cluster were also used to follow changes in this data set.

ACKNOWLEDGMENTS

Because of the similarity and close connection between this guide and Paper I,[53] some of the text and figures in these papers are either the same or very similar. This reuse of material is done with permission from the publisher.

AUTHOR DECLARATIONS

Conflict of Interest

Some of the authors work for organizations that sell the chemometrics software (Neal Gallagher, PLS_Toolbox from Eigenvector) and XPS data analysis software (Neal Fairley, CasaXPS) used in this study.

Author Contributions

Tahereh G. Avval: Formal analysis (lead); Investigation (lead); Writing – original draft (equal). **Hyrum Haack:** Formal analysis (supporting); Investigation (supporting). **Neal Gallagher:** Conceptualization (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **David Morgan:** Data curation (supporting). **Pascal Bargiela:** Data curation (supporting). **Neal Fairley:** Conceptualization (supporting); Formal analysis (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Vincent Fernandez:** Conceptualization (supporting); Formal analysis (supporting). **Matthew R. Linford:** Conceptualization (equal); Formal analysis (equal); Investigation (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The cellulose, tartaric acid, and sucrose data sets used in this work are provided in the supplemental material[65] as VAMAS files.

APPENDIX

This appendix focuses on small shifts observed in the peak positions of the O 1s tartaric acid spectra as they were collected by a Thermo Fisher Scientific K-alpha+ spectrometer [see these raw spectra in Fig. 3(d)]. As indicated in its structure (Fig. 2) and discussed herein, two well-separated C 1s signals of equal area are expected from tartaric acid. In contrast, O 1s signals typically undergo less chemical shifting—they are less sensitive to the chemical state of oxygen. The O 1s spectra in the tartaric acid data set are featureless and rather stable in shape—they can be well approximated as a single peak. Figure 23 shows the first 20 O 1s spectra in the tartaric acid data set. These spectra shift by a fraction of an eV to higher binding energy. We took two approaches to analyzing this (and its accompanying C 1s) data. In the first, the raw data were used as collected. This approach was taken in all the analyses shown in this work, except those in Figs. 12–14. In the second, the O 1s signals in Fig. 23 were aligned to a common value, where their accompanying C 1s spectra were shifted by the same amount. In all likelihood, the consequences of *not* shifting these O 1s spectra to a common binding energy value were that (i) more abstract factors were needed to describe the data set (the spectra were probably more spread out in the hyperspace they occupy), (ii) the resulting chemometrics/informatics analysis was somewhat more complicated because a larger number of abstract factors was probably needed to describe data, (iii) it was probably somewhat more challenging to determine the dimensionality of the data set (the number of abstract factors that best describe it), (iv) the abstract factors were probably a little harder to interpret because they had to account for both sample charging (peak shifts) and chemical effects, and (v) attempts that might be made to denoise the spectra would require a larger number of abstract factors than if the spectra were aligned. A negative effect of aligning the O 1s

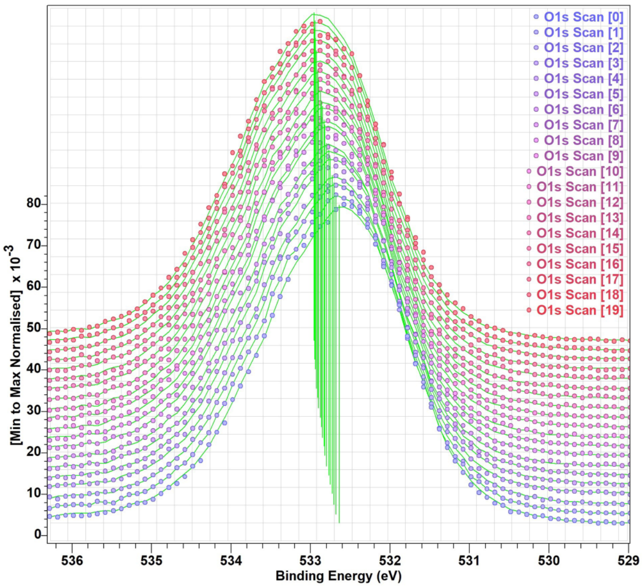

FIG. 23. First 20 raw O 1s spectra from the tartaric acid data set fitted using a single component to obtain the peak position/maximum.

spectra in this data set is that any real chemical shifts in these data, which may very well be present, are lost.

REFERENCES

[1] T. G. Avval *et al.*, J. Chem. Inf. Model. **61**, 4173 (2021).
[2] R. Bro and A. K. Smilde, Anal. Methods **6**, 2812 (2014).
[3] A. De Juan and R. Tauler, Crit. Rev. Anal. Chem. **36**, 163 (2006).
[4] N. B. Gallagher, J. M. Shaver, E. B. Martin, J. Morris, B. M. Wise, and W. Windig, Chemom. Intell. Lab. Syst. **73**, 105 (2004).
[5] S. Chatterjee, B. Singh, A. Diwan, Z. R. Lee, M. H. Engelhard, J. Terry, H. D. Tolley, N. B. Gallagher, and M. R. Linford, Appl. Surf. Sci. **433**, 994 (2018).
[6] S. Chatterjee and M. R. Linford, Bull. Chem. Soc. Jpn. **91**, 824 (2018).
[7] J. E. Jackson, *A User"s Guide to Principal Components* (Wiley, New York, 1991).
[8] B. M. Wise and N. B. Gallagher, J. Process Control **6**, 329 (1996).
[9] P. Van der Heide, *X-ray Photoelectron Spectroscopy: An Introduction to Principles and Practices* (Wiley, New York, 2011).
[10] S. Hofmann, *Auger-and X-ray Photoelectron Spectroscopy in Materials Science: A User-Oriented Guide* (Springer Science & Business Media, New York, 2012).
[11] F. A. Stevie and C. L. Donley, J. Vac. Sci. Technol. A **38**, 063204 (2020).
[12] D. R. Baer *et al.*, J. Vac. Sci. Technol. A **37**, 031401 (2019).
[13] V. Gupta, H. Ganegoda, M. H. Engelhard, J. Terry, and M. R. Linford, J. Chem. Educ. **91**, 232 (2014).
[14] D. R. Baer and I. S. Gilmore, J. Vac. Sci. Technol. A **36**, 068502 (2018).
[15] E. National Academies of Sciences and Medicine, *Reproducibility and Replicability in Science* (National Academies, 2019).
[16] D. R. Baer, J. Vac. Sci. Technol. A **38**, 031201 (2020).
[17] S. Tougaard, J. Vac. Sci. Technol. A **39**, 011201 (2021).
[18] C. J. Powell, J. Vac. Sci. Technol. A **38**, 023209 (2020).
[19] A. G. Shard, J. Vac. Sci. Technol. A **38**, 041201 (2020).
[20] T. R. Gengenbach, G. H. Major, M. R. Linford, and C. D. Easton, J. Vac. Sci. Technol. A **39**, 013204 (2021).
[21] M. J. Sweetman, S. M. Hickey, D. A. Brooks, J. D. Hayball, and S. E. Plush, Adv. Funct. Mater. **29**, 1808740 (2019).
[22] D. R. Baer, G. E. McGuire, K. Artyushkova, C. D. Easton, M. H. Engelhard, and A. G. Shard, J. Vac. Sci. Technol. A **39**, 021601 (2021).
[23] J. V. Macpherson, Phys. Chem. Chem. Phys. **17**, 2935 (2015).
[24] P. A. Navrátil, B. Westing, G. P. Johnson, A. Athalye, J. Carreno, and F. Rojas, paper presented at the Advances in Visual Computing, Berlin, Heidelberg, 2009.
[25] J. Wolstenholme, J. Vac. Sci. Technol. A **38**, 043206 (2020).
[26] G. Beamson and D. Briggs, *High Resolution XPS of Organic Polymers: The Scienta ESCA300 Database* (Wiley, Chichester, 1992).
[27] G. P. López, D. G. Castner, and B. D. Ratner, Surf. Interface Anal. **17**, 267 (1991).
[28] D. G. Castner and B. D. Ratner, Surf. Interface Anal. **15**, 479 (1990).
[29] H. E. Canavan, X. Cheng, D. J. Graham, B. D. Ratner, and D. G. Castner, Langmuir **21**, 1949 (2005).
[30] D. I. Patel *et al.*, Surf. Sci. Spectra **26**, 016801 (2019).
[31] T. G. Avval, G. T. Hodges, J. Wheeler, D. H. Ess, S. Bahr, P. Dietrich, M. Meyer, A. Thißen, and M. R. Linford, Surf. Sci. Spectra **27**, 014006 (2020).
[32] G. Jiang, G. A. Husseini, L. L. Baxter, and M. R. Linford, Surf. Sci. Spectra **11**, 91 (2004).
[33] D. Shah, S. Bahr, P. Dietrich, M. Meyer, A. Thißen, and M. R. Linford, Surf. Sci. Spectra **26**, 024009 (2019).
[34] S. C. Tahereh, G. Avval, S. Bahr, P. Dietrich, M. Meyer, A. Thißen, and M. R. Linford, Surf. Sci. Spectra **27**, 014006 (2020).
[35] D. R. Baer, M. H. Engelhard, and A. S. Lea, Surf. Sci. Spectra **10**, 47 (2003).
[36] M. D. Duca, C. L. Plosceanu, and T. Pop, J. Appl. Polym. Sci. **67**, 2125 (1998).
[37] L. A. Pesin, E. M. Baitinger, Y. P. Kudryavtsev, and S. E. Evsyukov, Appl. Phys. A **66**, 469 (1998).
[38] I. I. Vointseva, L. M. Gilman, Y. P. Kudryavtsev, S. E. Evsyukov, L. A. Pesin, I. V. Gribov, N. A. Moskvina, and V. V. Khvostov, Eur. Polym. J. **32**, 61 (1996).
[39] L.-S. Johansson, J. M. Campbell, and O. J. Rojas, Surf. Interface Anal. **52**, 1134 (2020).
[40] L.-L. Chua, M. Dipankar, S. Sivaramakrishnan, X. Gao, D. Qi, A. T. S. Wee, and P. K. H. Ho, Langmuir **22**, 8587 (2006).
[41] H. Ahn, D. W. Oblas, and J. E. Whitten, Macromolecules **37**, 3381 (2004).
[42] K. Artyushkova and J. E. Fulghum, J. Electron Spectrosc. Relat. Phenom. **121**, 33 (2001).
[43] S. Pylypenko, K. Artyushkova, and J. E. Fulghum, Appl. Surf. Sci. **256**, 3204 (2010).
[44] M. P. Felicissimo, J. L. S. Peixoto, R. Tomasi, A. Azioune, J.-J. Pireaux, L. Houssiau, and U. P. Rodrigues Filho, Philos. Mag. **84**, 3483 (2004).
[45] K. Artyushkova and J. E. Fulghum, Surf. Interface Anal. **31**, 352 (2001).
[46] C. Bittencourt, M. P. Felicissimo, J.-J. Pireaux, and L. Houssiau, J. Agric. Food Chem. **53**, 6195 (2005).
[47] B. J. Tyler, Appl. Surf. Sci. **252**, 6875 (2006).
[48] R. E. Peterson and B. J. Tyler, Appl. Surf. Sci. **203–204**, 751 (2003).
[49] L. Yang, Y.-Y. Lua, G. Jiang, B. J. Tyler, and M. R. Linford, Anal. Chem. **77**, 4654 (2005).
[50] D. J. Graham and D. G. Castner, Biointerphases **7**, 49 (2012).
[51] M. S. Wagner, D. J. Graham, and D. G. Castner, Appl. Surf. Sci. **252**, 6575 (2006).
[52] P. Bargiela, V. Fernandez, C. Cardinaud, J. Walton, M. Greiner, D. Morgan, N. Fairley, and J. Baltrusaitis, Appl. Surf. Sci. **566**, 150728 (2021).
[53] Tahereh G. Avval, Neal Gallagher, David Morgan, Pascal Bargiela, Neal Fairley, Vincent Fernandez, and M. R. Linford, "Practical guide on chemometrics/informatics in x-ray photoelectron spectroscopy (XPS). I. Introduction to methods useful for large or complex datasets," J. Vac. Sci. Technol. A (to be published).
[54] P. Kraniauskas, *Transforms in Signals and Systems (Modern Applications of Mathematics)* (Addison-Wesley Longman, Incorporated, Reading, MA, 1992).
[55] R. Schafer, IEEE Signal Process. Mag. **28**, 111 (2011).
[56] W. H. Press and S. A. Teukolsky, Comput. Phys. **4**, 669 (1990).
[57] J. Luo, K. Ying, and J. Bai, Signal Process. **85**, 1429 (2005).
[58] A. Savitzky and M. J. E. Golay, Anal. Chem. **36**, 1627 (1964).
[59] T. O. Zuppa Neto *et al.*, J. Am. Soc. Mass Spectrom. **31**, 1525 (2020).
[60] S. Chatterjee, G. H. Major, B. Paull, E. S. Rodriguez, M. Kaykhaii, and M. R. Linford, J. Chromatogr. A **1558**, 21 (2018).
[61] S. Chatterjee, S. C. Chapman, B. M. Lunt, and M. R. Linford, Bull. Chem. Soc. Jpn. **91**, 1775 (2018).
[62] G. H. Major, V. Fernandez, N. Fairley, and M. R. Linford, Surf. Interface Anal. **54**, 262 (2022).
[63] S. Tougaard, Surf. Interface Anal. **25**, 137 (1997).
[64] L. Pei, G. Jiang, R. C. Davis, J. M. Shaver, V. S. Smentkowski, M. C. Asplund, and M. R. Linford, Appl. Surf. Sci. **253**, 5375 (2007).
[65] See the supplementary material at https://www.scitation.org/doi/suppl/10.1116/6.0001969 for summary statistics analyses of the data, PCA of the unpreprocessed cellulose data, comparison of component 1 and scan 22 in the MCR analysis of tartaric acid C 1s narrow scans with six components, and MCR analyses with different numbers of components of the concatenated C 1s and O 1s narrow scans of the tartaric acid data sets.