Combining PCA and nonlinear fitting of peak models to re-evalute cellulose C 1s region

Vincent Fernandez,[1] David Morgan,[2] Pascal Bargiela,[3] Neal Fairley[4] and Jonas Baltrusaitis[5,*]

[1]Université de Nantes, CNRS, Institut des Matériaux Jean Rouxel, IMN, F-44000 Nantes, France
[2]School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff CF10 3AT, United Kingdom
and
HarwellXPS – EPSRC National Facility for photo-electron Spectroscopy, Research Complex at Harwell (RCaH), Didcot, Oxon, OX11 0FA
[3]The Institute for Research on Catalysis and the Environment of Lyon (*IRCELYON)*, 2 Avenue Albert Einstein, 69626 Villeurbanne, France
[4]Casa Software Ltd, Bay House, 5 Grosvenor Terrace, Teignmouth, Devon TQ14 8NE, UK
[5]Department of Chemical and Biomolecular Engineering, Lehigh University, 111 Research Drive, Bethlehem, PA 18015, USA

Abstract

The practice of fitting a peak model constructed from bell-shaped components by nonlinear least squares to multiple spectra in unison is investigated and compared to similar concepts fundamental to the analysis of sets of spectra using linear algebra. A mathematical analysis of the least squares optimisation is presented which demonstrates the relationship between optimisation and the covariance matrix used in PCA. The act of fitting a peak model formed from bell-shaped curves similarly applies the least squares criterion to compute a curve from multiple spectra. A comparison between the results of PCA and nonlinear fitting of peak models to data is made to demonstrate that, while similar curves are created by PCA and nonlinear curve fitting, the results are different. X-ray Photoelectron Spectroscopy (XPS) of cellulose was used to create a set of spectra that evolve in shape upon sample surface exposure to X-ray beam which were analyzed by both nonlinear least squares applied to multiple spectra and linear least squares fitting of spectral forms calculated directly from data. These examples of fitting curves to data are used to demonstrate how both data analysis approaches combine to help in the understanding of chemistry at the surface using XPS.

*Corresponding author: job314@lehigh.edu; +1-610-758 6836

Keywords: XPS; PCA; cellulose; peak model; spectra

Introduction

Cellulose fibers provide a useful model system for spectroscopic polymer studies as the X-ray Photoelectron Spectroscopy (XPS) data obtained is usually subject to contamination, spectral artifacts, impurities as well as degradation due to the energetic X-ray beam [1,2]. Of particular interest is the ability to accurately interpret cellulose XPS spectra obtained from wide range of natural and modified cellulose sample surfaces. The cellulose sample surface often is subjected to an oxidative treatment [3] due to the necessity to modify it both chemically (oxidation and hydrolysis) and structurally (type of crystal and degree of crystallinity) to change their reactivity [4] to tailor a wide range of applications [5]. A typical four-peak structure in the XPS C 1s region has been reported for cellulose surface subjected to a varying degree of oxidation, assigned to C-C, C-O, O-C-O or C=O and O=C-O bonds with chemical shifts of 1.7±0.1 eV, 3.1 ± 0.1 eV and 4.4 ± 0.2 eV, respectively [4,6]. However, the chemistry of the unmodified cellulose only warrants two peaks, e.g. C-O and O-C-O [2]. The monitoring of cellulose oxidation as a model system using XPS has been shown to generate a series of spectra whereby the evolution of the four-peak structure can be used to monitor the rate of oxidation [1] or evolution of other, non-cellulosic components [2]. Finally, charging artifacts can have a significant effect on the four-peak intensity distribution [1]. Since modern XPS instruments can acquire large amounts of data at speeds that tailor to combinatorial screening methods [7], large dataset processing methods need to be developed that alleviate the data analysis and provide for consistent peak fitting procedures.

The technique of fitting simultaneously a single peak model to multiple spectra can be used to estimate peak full-width-half-maximum (FWHM) and binding energy for sets of bell-shaped components in a peak model that can be consistently applied to all spectra in a data set. The approach is appropriate for spectra measured from samples for which the chemistry present in each sample is assumed to be identical, typical for cellulose modified using oxidative treatment, but for which the proportions of chemical state may change. Spectra used in the analysis also need to be calibrated in energy to align variation in signal in each spectrum with corresponding energy. These conditions necessary for fitting multiple spectra by a single peak model are also the conditions, for the most part, required when performing linear analysis of such a data set and both approaches only perform well provided energy calibration is precise. To demonstrate how a common FWHM and energy are achieved by nonlinear optimization, a set of spectra measured from cellulose comprised coffee filter paper is analyzed in this work. Repetitions of the same C 1s spectra are measured from the same location on a sample. Repeating measurements by XPS results in evolution in spectral shapes yielding spectra that are well aligned in energy for which signal intensity (over a range of energies) alters systematically throughout the experiment due to the exposure to X-rays. It can be assumed, when calculating FWHM and binding energy, that these spectra are simply altering in relative proportions in four different chemical states [3,4]. This assumption of four chemical states will be called into doubt by performing further analysis of the entire data set, the result of which suggests fitting cellulose spectra with four bell-shaped components is a crude model for chemical changes in the sample modified using X-rays. In particular, analysis of the data set by PCA and linear optimisation [8–10] by constructing spectral-forms from data suggest at least six chemical states contribute to spectra as-measured from cellulose.

Experimental and theoretical methods

XPS data acquisition

The experiment was performed on a Kratos Axis Ultra, using a pass energy of 10 eV (PE10) and field of view 1 (FoV1). Charge compensation via low-energy electrons was active throughout the experiment and provided a stable uniform potential on the surface of the sample over the 700x300 μm^2 analysis area. As-measured photoemission peaks are offset in energy (as a consequence of charge compensation) to lower binding energy from the expected binding energy. Nevertheless, all spectra appear to be stable in terms of the apparent binding energy for photoemission. No energy calibration was performed and spectra are displayed and processed using the apparent binding energy corresponding to the measured spectra. Only C 1s spectra are used in the following work, but the role of O 1s in the experiment is important since switching between energy intervals while acquiring data, represents a perturbation to the state of the sample with respect to charge compensation that may result is shifts in energy between measurement cycles. Careful investigations aimed at eliminating concerns of energy shifts between cycles were performed. All data processing was performed using CasaXPS [11].

Theoretical methods

Two approaches to fitting curves to data are typically used for the analysis of XPS data. The first approach most commonly used is that of nonlinear optimization. The second approach is fitting curves to data using linear least squares. Both methods of fitting curves to data make use of a least squares criterion, however, nonlinear least squares are more flexible but less robust than linear least squares, since nonlinear least squares optimization does not provide a unique solution to an optimization problem for algorithmic reasons, whereas if a solution is possible then linear least squares does provide a unique solution. Nevertheless, both approaches indirectly or directly make use of least squares to solve fitting curves to data. The challenge for linear least squares fitting is to identify spectral forms of physical meaning that also allow a fit to data. PCA is central to establishing spectral forms that allow the use of linear least squares fitting, therefore the following provides the basis for least squares optimization, then mathematically shows how PCA is formulated in terms of least squares and how optimization is equivalent to solving an eigen analysis of the covariance matrix formed from a data set.

Least Squares and Fitting Curves to Data

Nonlinear optimisation fitting of a peak model constructed from bell shaped curves to data involves the use of the so-called least squares criterion. Therefore, before describing the analysis of spectra, concepts central to the use of least squares in data analysis are presented. When a curve is fitted to data, adjustments are made to parameters that alter the characteristics of the curve until, by some means, it is decided the curve fits all data. The decision to accept a curve as the best fit to data is performed by calculating a value (figure-of-merit) which indicates a positive outcome has been achieved by some optimization algorithm. In the case of nonlinear optimization, the best set of parameters is obtained by performing a search in which a sequence of steps in parameters defining the curve is tested against a figure-of-merit. The figure-of-merit does not directly influence the next step toward a solution, but rather is a measure that indicates if a step moved the curve closer to a best fit to data. For XPS, the figure-of-merit is computed from two measures based on the difference between the curve and data for each data bin. To obtain a single number capable of guiding optimization algorithms to the best fit these differences are squared and then summed. Two variations on this theme for a figure-of-merit are now described.

Given a data vector $\boldsymbol{d} = (d_1, \quad d_2, \quad d_3 \quad \dots \quad d_n)$ and the vector of expected values for the data vector $\boldsymbol{e} = (e_1, \quad e_2, \quad e_3 \quad \cdots \quad e_n)$, a measure for the deviation of the data vector $\boldsymbol{d}$ from the expected vector $\boldsymbol{e}$ is given by (1)

$$\chi^2 = \sum_{i=1}^{n} \frac{(d_i - e_i)^2}{e_i} \qquad (1).$$

If data are measured using pulse counting, then the standard deviation expected for d_i is modeled making use of Poisson statistics which implies uncertainty in counts d_i for each data bin i is $\sqrt{e_i}$. When attempting to fit a vector $\boldsymbol{y} = (y_1, \quad y_2, \quad y_3 \quad \cdots \quad y_n)$ constructed with parameters that are determined to minimize a figure of merit (Equation (2)), then the fit of $\boldsymbol{y}$ to $\boldsymbol{d}$ for pulse counted data is achieved through scaling each data vector coordinate to allow for the increased variance for high-intensity signals, such as peak maxima, compared to the low-intensity signal

$$figure\ of\ merit = \sum_{i=1}^{n} \frac{(d_i - y_i)^2}{y_i} \cong \sum_{i=1}^{n} \left(\sqrt{d_i} - \frac{y_i}{\sqrt{d_i}}\right)^2 \qquad (2).$$

In the context of XPS data measured by counting electrons emitted with given energy over a specific time, optimization using the figure-of-merit Equation (2) has the advantage that fitting parameters are determined with equal weight for both background signal and photoemission signal superimposed on background signal. By contrast, due to variation in the magnitude of random noise with signal intensity in pulse counted data, optimization based on minimizing a direct sum of squares (Equation (3)) yields solutions that favor fitting data vector coordinates with high count rates, such as photoemission peak maxima, rather than uniformly treating all data vector coordinates equally

$$sum\ of\ squares = \sum_{i=1}^{n} (d_i - y_i)^2 \qquad (3).$$

Fitting curves to data based on minimizing Equation (2) or Equation (3) may create different outcomes, but both are reasonably described as the figure-of-merit for least squares optimization. When these figure-of-merits are applied to a peak model and an individual spectrum the result is a curve that approximates the spectrum.

Covariance Matrix and Least Squares Optimisation

Linear least squares fitting of data is achieved by making use of the covariance matrix formed from the set of spectral forms selected to represent distinct chemical states of a sample. The covariance matrix form taken from a set of spectra represents part of the construction of a PCA used to characterize and count the number of spectral shapes within a set of spectra. Therefore, a geometric interpretation of how a least squares criterion is transformed by the calculus of variation into an eigenproblem of fundamental importance to data analysis is presented in this section. The mathematical reasoning described below makes use of scatter plots constructed from spectra. The reason for doing so is because plotting points in a 3-dimensional diagram, where the coordinates for each point are the intensities for the same data-bin taken from three spectra, links together these three intensities from spectra for which a line within the 3-dimensional space can be determined that minimizes the distances between all such points formed from intensities and the line. While a line within a 3-dimensional space may seem abstract from spectra, a line in 3-dimensions is linked to a specific curve corresponding to spectra by a transformation constructed by the mathematics of singular value decomposition (SVD). Therefore, before showing the mathematical relationship between least squares optimization and determining the eigenvalues of a covariance matrix $\boldsymbol{Z}$, the concept of singular valued decomposition for a set of spectra will be presented.

Given three spectra with intensities measured at a set of n corresponding energies, neglecting the absolute values for these energies at which intensities are measured and simply listing intensities, a spectrum can be written mathematically as a vector $\boldsymbol{d_1} = (x_1, \quad x_2, \quad x_3 \quad \cdots \quad x_n)$. Simply because a graphical interpretation is possible in 3-dimensions, three spectra are considered rather than being forced to work in higher dimensions by using more than three spectra. The mathematics as described generalizes to higher dimensions, but for the sake of clarity the following uses only three spectra represented as vectors: $\boldsymbol{d_1} = (x_1, \quad x_2, \quad x_3 \quad \cdots \quad x_n)$, $\boldsymbol{d_2} = (y_1, \quad y_2, \quad y_3 \quad \cdots \quad y_n)$ and $\boldsymbol{d_3} = (z_1, \quad z_2, \quad z_3 \quad \cdots \quad z_n)$. The use of x, y and z in these vectors is intended to provide a narrative link to the coordinate axes in the scatter plot formed from these vectors. From these three vectors, a matrix $\boldsymbol{D}$ is formed of dimension $n \times 3$, that is an array of three columns where each row contains three entries $(x_i \quad y_i \quad z_i)$. Thus, the columns of $\boldsymbol{D}$ are the vectors formed from spectra. The rows of $\boldsymbol{D}$ are 3-dimensional coordinates for points that can be plotted as a scatter diagram. The mathematics of SVD [12] prescribes a method for expressing $\boldsymbol{D}$ in terms of three matrices $\boldsymbol{U}$, $\boldsymbol{W}$ and $\boldsymbol{V}$ as follows in (4)

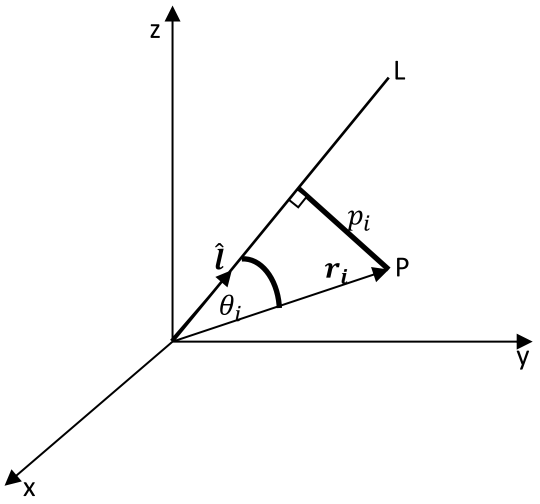
$$\boldsymbol{D} = \boldsymbol{UWV}^T \qquad (4).$$

The matrix $\boldsymbol{U}$ has the same dimensions as $\boldsymbol{D}$. The matrices $\boldsymbol{W}$ and $\boldsymbol{V}$ are square matrices of dimension equal to the number of columns in $\boldsymbol{D}$. These matrices $\boldsymbol{U}$, $\boldsymbol{W}$ and $\boldsymbol{V}$ are not arbitrary but have specific mathematical properties. However, the critical point derived from Equation [4] for the following discussion is that assuming inverse matrices exist for matrices $\boldsymbol{U}$, $\boldsymbol{W}$ and $\boldsymbol{V}$, Equation (4) provides a means of converting 3-dimensional vectors into equivalent n-dimensional vectors. Thus, a line in three dimensions passing through the origin for a scatter plot may be converted to a spectral shape. On this basis, constructing a scatter plot from three spectra and fitting a line to these scatter plot points in the least squares sense results in a curve that can be interpreted as a spectral shape, namely the first PCA AF. The following now considers the mathematics of fitting a line to points in a 3-dimensional scatter plot.

The common practice in PCA of working directly with the covariance matrix $\boldsymbol{Z} = \boldsymbol{D^T D}$ can be understood by following through the logic of minimizing the sum of squares of the perpendicular distances from each point in a scatter plot to the principal axis line. Note that Equation (3) is not the same figure of merit used in the following steps. The following makes use of Pythagoras to determine the distance from a point in the scatter plot to a line through the origin with direction cosines $\hat{\boldsymbol{l}} = (\alpha, \beta, \gamma)$, the minimization problem can be expressed as follows. That is, a vector is computed that minimizes I subject to the constraint that the vector has unit length, where I and the constraint enforcing a non-zero vector $\hat{\boldsymbol{l}}$ is defined as follows in (5)

$$I = \sum_{i=1}^{n} {p_i}^2 \quad \text{subject to the constraint } \alpha^2 + \beta^2 + \gamma^2 = 1 \qquad (5).$$

The mathematical logic, now described for the case of three data vectors, shows how a least squares figure-of-merit is formulated such that calculus, when used to determine turning-points for the figure-of-merit in terms of coordinates of a unit vector, results in systems of linear equations expressed as an eigenvalue/eigenvector problem, the solution of which determines the position vector line of best-fit for all points in the scatter plot (Figure 1).



Applying Pythagoras theorem to determine the distance p_i

$$p_i{}^2 = |\boldsymbol{r_i}|^2 - (|\boldsymbol{r_i}|cos\theta_i)^2$$

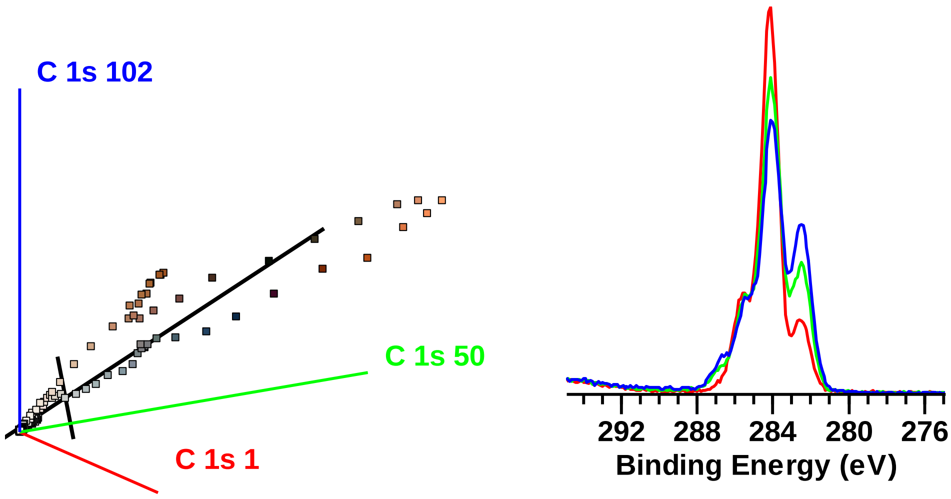
$$cos\theta_i = \frac{\boldsymbol{r_i}.\hat{\boldsymbol{l}}}{|\boldsymbol{r_i}||\hat{\boldsymbol{l}}|}$$

Since

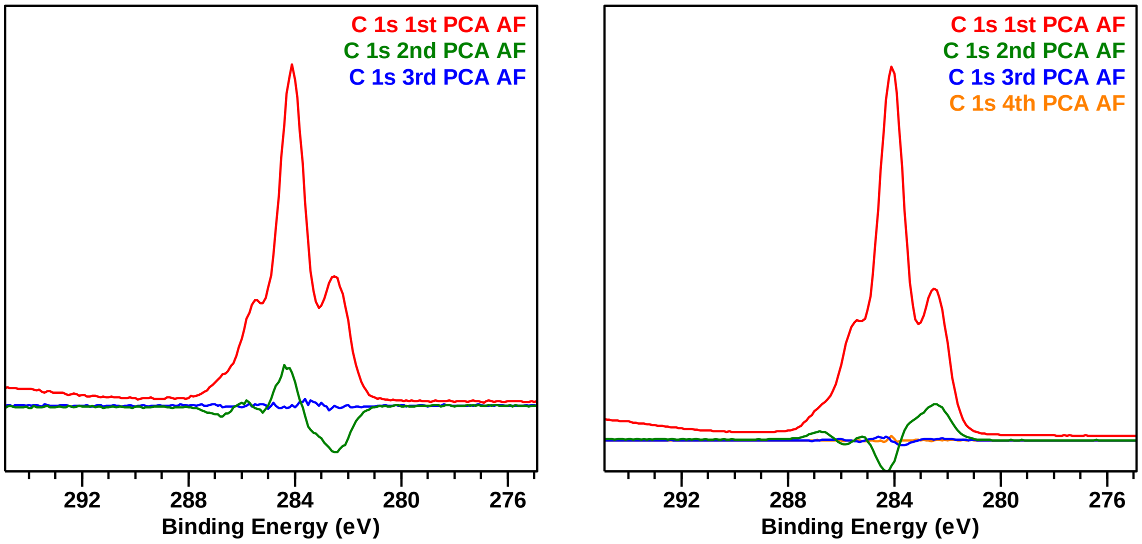
$$|\hat{\boldsymbol{l}}|^2 = \alpha^2 + \beta^2 + \gamma^2 = 1$$

$$p_i{}^2 = |\boldsymbol{r_i}|^2 - (\boldsymbol{r_i}.\hat{\boldsymbol{l}})^2$$

(a)



(b)



(c)

(d)

Figure 1. (a) Geometrical interpretation of a point within a scatter plot formed from three spectra. (b) Three dimensional scatter plot was formed from three spectra selected from the data set measured from cellulose. Two principal axes corresponding to the first and second most significant PCA AFs are plotted over the scatter points formed from the three spectra display to accompany the scatter plot. (c) PCA AFs calculated from the 3 spectra shown in (b). The PCA AF labeled C 1s 1st PCA AF corresponds to the principal axes of greatest length displayed in (b). (d) PCA AFs calculated from all 102 spectra measured from cellulose. Figure 2b is constructed by making use of the first two PCA AFs shown in (d).

Each point in the scatter plot has coordinates $\boldsymbol{r_i} = (x_i, y_i, z_i)$, therefore using the vector algebra in Figure 1 the perpendicular distance of a point to the line is defined by the position vector $\hat{\boldsymbol{l}}$ is given by the following by (6)

$$ {p_i}^2 = {x_i}^2 + {y_i}^2 + {z_i}^2 - (\alpha x_i + \beta y_i + \gamma z_i)^2 \qquad (6). $$

The sum of squares in Equation [5] can be written as follows in (7)

$$ I = \sum_{i=1}^{n}({x_i}^2 + {y_i}^2 + {z_i}^2 - (\alpha x_i + \beta y_i + \gamma z_i)^2) \qquad (7). $$

The covariance matrix is derived by applying the method of Lagrange multipliers to include the constraint $\alpha^2 + \beta^2 + \gamma^2 = 1$, namely the optimization of the parameters $\alpha, \beta\ and\ \gamma$ subject to the constraint that $\hat{\boldsymbol{l}}$ is a unit vector that best fits all points in a scatter plot and is performed by computing extrema for the function (8)

$$ \Psi = \lambda(\alpha^2 + \beta^2 + \gamma^2 - 1) + I \qquad (8). $$

Partial differentiating Ψ with respect to α yields (9)

$$\frac{\partial\Psi}{\partial\alpha} = 2\alpha\lambda + \sum_{i=1}^{n}(-2(\alpha x_i + \beta y_i + \gamma z_i)x_i) = 2\alpha\lambda - 2\alpha\sum_{i=1}^{n} x_i{}^2 - 2\beta\sum_{i=1}^{n} x_i y_i - 2\gamma\sum_{i=1}^{n} x_i\, z_i$$

(9).

Equating to zero yields (10)

$$\alpha\sum_{i=1}^{n} x_i{}^2 + \beta\sum_{i=1}^{n} x_i y_i + \gamma\sum_{i=1}^{n} x_i\, z_i = \alpha\lambda \qquad (10).$$

Similarly, $\frac{\partial\Psi}{\partial\beta} = 0$ and $\frac{\partial\Psi}{\partial\gamma} = 0$ provides two more equations as (11) and (12)

$$\alpha\sum_{i=1}^{n} x_i y_i + \beta\sum_{i=1}^{n} y_i{}^2 + \gamma\sum_{i=1}^{n} y_i\, z_i = \beta\lambda \qquad (11),$$

$$\alpha\sum_{i=1}^{n} x_i z_i + \beta\sum_{i=1}^{n} y_i z_i + \gamma\sum_{i=1}^{n} z_i{}^2 = \gamma\lambda \qquad (12).$$

Using $\boldsymbol{d}_1.\boldsymbol{d}_1 = \sum_{i=1}^{n} x_i{}^2$, $\boldsymbol{d}_2.\boldsymbol{d}_2 = \sum_{i=1}^{n} y_i{}^2$, $\boldsymbol{d}_3.\boldsymbol{d}_3 = \sum_{i=1}^{n} z_i{}^2$, $\boldsymbol{d}_1.\boldsymbol{d}_2 = \sum_{i=1}^{n} x_i y_i$, $\boldsymbol{d}_1.\boldsymbol{d}_3 = \sum_{i=1}^{n} x_i z_i$ and $\boldsymbol{d}_2.\boldsymbol{d}_3 = \sum_{i=1}^{n} y_i z_i$ and expressing the above simultaneous equations in matrix form the eigenvector problem expressed in terms of the original data vectors reduces as follows in (13)

$$\begin{bmatrix} \boldsymbol{d}_1.\boldsymbol{d}_1 & \boldsymbol{d}_1.\boldsymbol{d}_2 & \boldsymbol{d}_1.\boldsymbol{d}_3 \\ \boldsymbol{d}_1.\boldsymbol{d}_2 & \boldsymbol{d}_2.\boldsymbol{d}_2 & \boldsymbol{d}_2.\boldsymbol{d}_3 \\ \boldsymbol{d}_1.\boldsymbol{d}_3 & \boldsymbol{d}_2.\boldsymbol{d}_3 & \boldsymbol{d}_3.\boldsymbol{d}_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \lambda \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \qquad (13).$$

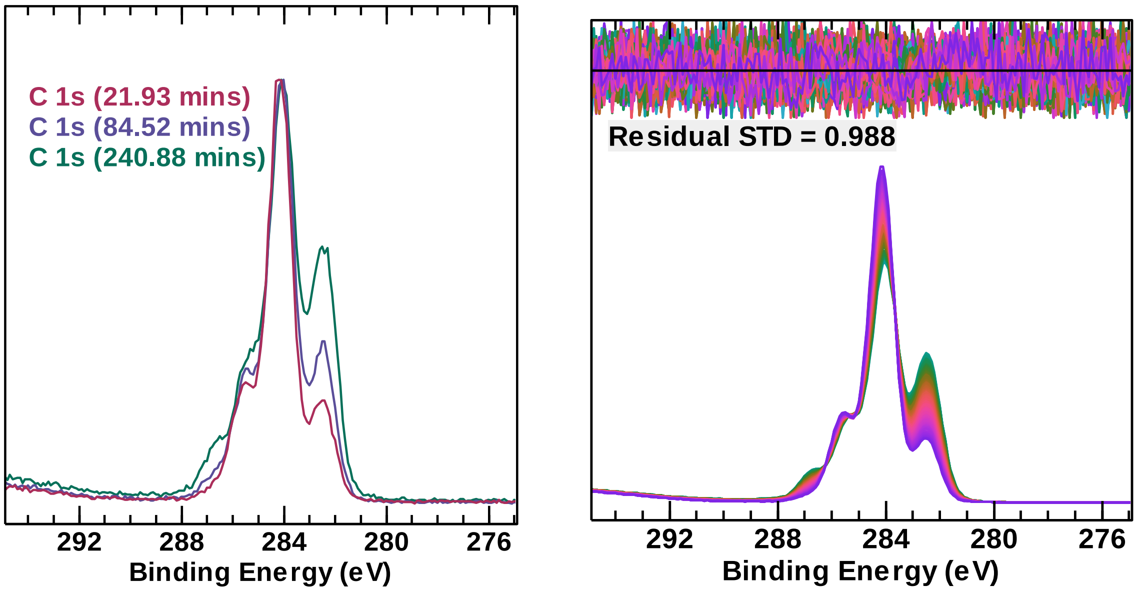
Thus, the covariance matrix $\boldsymbol{Z}$ is recovered from the optimization problem by making use of the least squares figure-of-merit in Equation (5).

Results

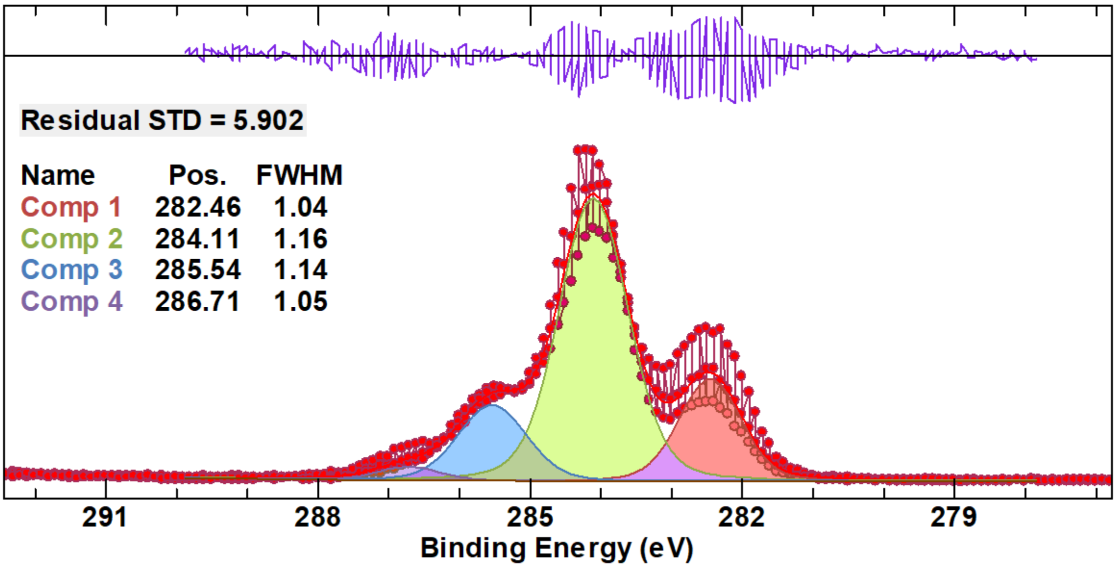
Cellulose C 1s Peak Model Design

An average spectrum is a curve that influences the selection of a common FWHM and energy determined by the nonlinear fitting of a peak model to data. An average spectrum is readily computed from the set of spectra by summation of spectra; however, the average spectrum does not provide any indication of how well it represents spectra within the data set. PCA AFs, on the other hand, provide feedback in the form of the less significant PCA AFs. The scatter plot in Figure 1b and corresponding PCA AFs shown in Figure 1c demonstrate that, even based on three spectra, PCA (by the relative size for the second PCA AF shown in Figure 1c) highlights that an average spectrum is not entirely characteristic of all C 1s spectra in Figure 1. The outstanding question that justifies the use of an average spectrum to test for common FWHM and energy not answered by PCA is *to what extent the second PCA AF exists because of changes in proportions in common components in a peak model*. The obvious similarity between the average spectrum (an example of which is shown in Figure 2c) for these spectra measured from cellulose and the most significant PCA AF (first PCA AF) requires some explanation that may help to make an informed decision as to the number of components to use in a peak model. In particular, the first PCA AF is computed based on the least squares principle in Equation (7), which is not the same as Equation (3) and also involves the use of a constraint stated in Equation (5).

Figure 2a presents three raw spectra sub-sampled from the full set of 102 spectra measured as a function of time of cellulose surface. Principal Component Analysis for the full set of 102 spectra indicates these spectra lie within a plane defined by the first two PCA AFs (as already shown in Figure 1d). In support of this assertion, approximations to raw spectra are computed by projecting raw spectra onto a 2-dimensional subspace defined by the two most significant PCA AFs yielding processed spectra that fit the original spectra with precision indicated by the residual plots (residual STD close to unity and uniform residual plots) in Figure 2b. Systematic changes in spectral shapes shown in Figure 2b and PCA both support the assumption that these data are well behaved in terms of the response of the sample to charge compensation.



(a) (b)

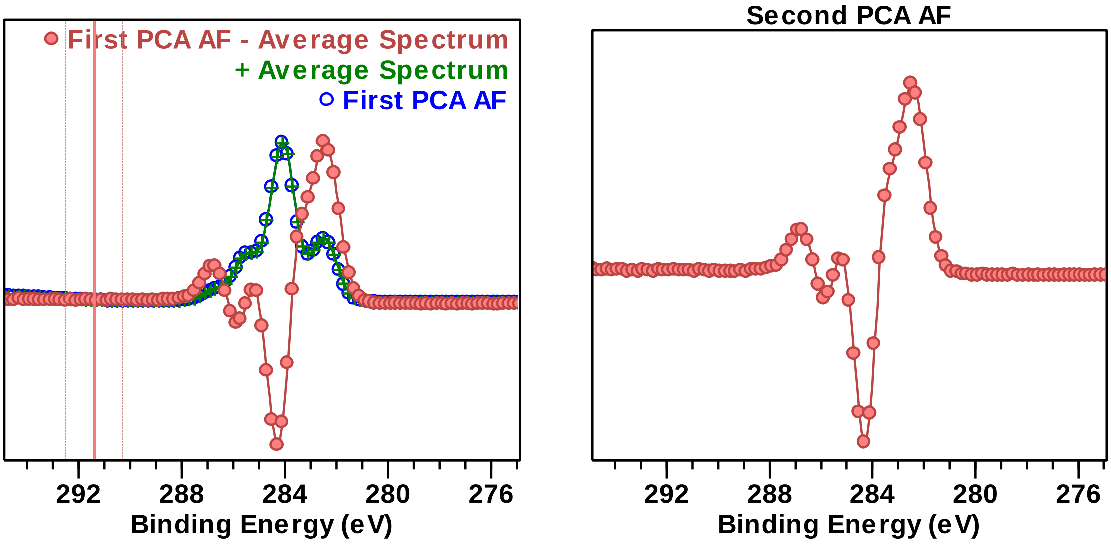


(c)

Figure 2. (a) Three raw spectra sub-sampled from the data set corresponding to the processed spectra in (b) displayed normalized for each spectrum to a common display area. (b) PCA enhanced C 1s spectra for data measured from cellulose surface. The residual plots represent the normalized difference between raw spectra and PCA enhanced spectra, where these PCA enhanced spectra are computed by fitting the two most significant PCA AFs to raw spectra in the least squares sense. (c) Average spectrum constructed from all three spectra in (a). For each abscissa, there are three ordinates. Therefore, optimization for the peak model with four bell-shaped components yields a fit to the average spectrum rather than to individual spectra.

A peak model designed to fit all three spectra in Figure 2a, when applied to each spectrum individually might yield anomalous FWHM. Nonphysical FWHM obtained after optimization for certain spectra is a consequence of insufficient signal such as signal corresponding to the high binding energy component in Figure 2c. When the peak model in Figure 2c is applied to the as-received cellulose spectrum in Figure 2a results in an FWHM for one component close to 0.5 eV, compared to an expectation (on the grounds of physics) of FWHM for all components close to unity. However, when presented with three spectra of the form shown in Figure 2a, the curve computed from a peak model such as that shown in Figure 2c is a fit that moderates non-physical FWHM by requiring the fit to be valid, in some sense for all three spectra. In terms of component area in the peak model, the optimum solution does not fit any spectrum in the data set, but rather fits the average spectrum for the data as shown in Figure 2a and the alternative merged form for these same three spectra in Figure 2c. The objective of fitting data using the merged data from three spectra in Figure 2c is to estimate the FWHM and energy for components in the peak model that are common to all three spectra. The advantage of so doing lies in following the physics of the photoemission process, which implies electrons involved in the chemical state of atoms, on being excited by photons of a given energy, emit electrons with intensity and energy that match the component shapes and energy in the peak model. Once the best estimates for FWHM and energy are established, uncertainty in intensity is greatly reduced when determined by fitting a peak model to individual spectra with known FWHM and energy.

Differences between the average spectrum and the first PCA AF become more apparent when PCA is performed on larger data sets with greater variation in spectral shapes than the example data set used here. Figure 3 is constructed from the data set of 102 C 1s spectra used in Figure 2, where PCA is performed on all 102 spectra and the average spectrum for all 102 spectra is calculated. While remarkably similar, forming the difference between the normalized first PCA AF and the normalized average spectrum (Figure 3a) reveals a shape that is very similar to the second PCA AF (Figure 3b). The implication is that within the average spectrum there is a shape that is removed from the first PCA AF by the steps performed by SVD that underpins PCA calculations.

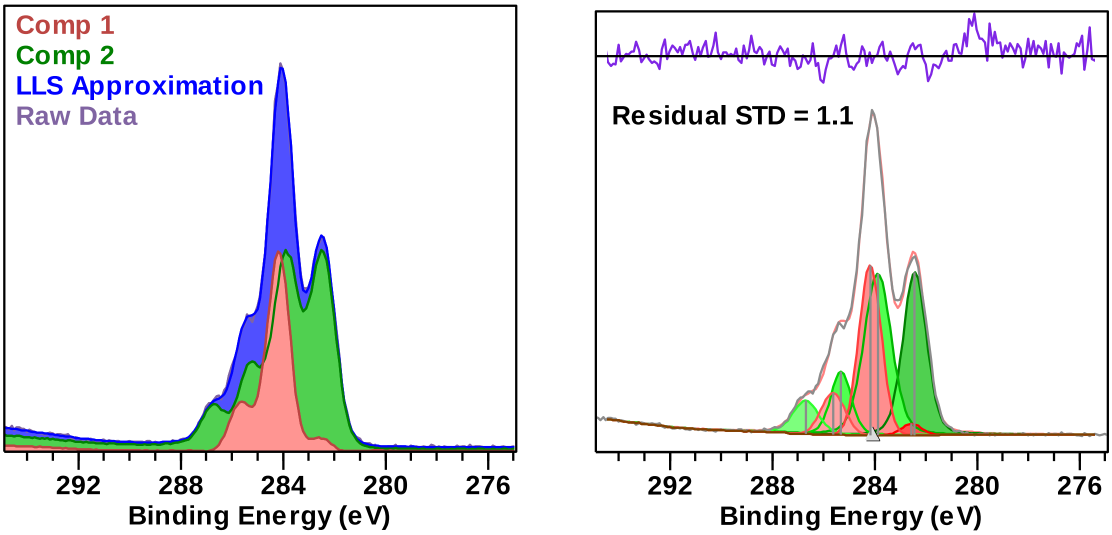


(a)

(b)

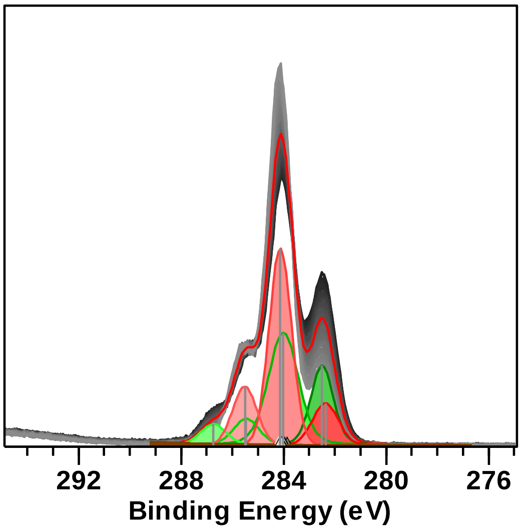
Figure 3. (a) Overlay of the normalized first PCA AF, the normalized average spectrum and the difference between the normalized first PCA AF and the normalized average spectrum. Data are scaled by averaging intensities over the interval marker by vertical lines. The difference curve (red) is small for this case constructed from 102 C 1s spectra illustrated in Figure 2, but mathematically the difference between the first PCA AF and the average spectrum is significant in the sense the shape includes characteristics of the second PCA AF. (b) Second PCA AF was computed for all 102 spectra used in (a).

The peak model used to compute an estimate for the average spectrum for the spectra in Figure 2c is shown to be an oversimplification of the processes responsible for spectra measured from cellulose sample surface, by virtue of the analysis leading to the results in Figure 4. While the peak model offers FWHM and energy for the components as defined in Figure 2c, the analysis of all 102 spectra by means of considering spectral forms that emerge when difference spectra are calculated (Figure 4a), demonstrate an alternative interpretation for this data set. The analysis shown in Figure 4a and Figure 4b makes use of two spectral forms to fit one spectrum from the data set of 102 C 1s spectra. These two spectral forms (Comp 1 and Comp 2) are capable of fitting all spectra in the data set with similar precision.



(a)

(b)



(c)

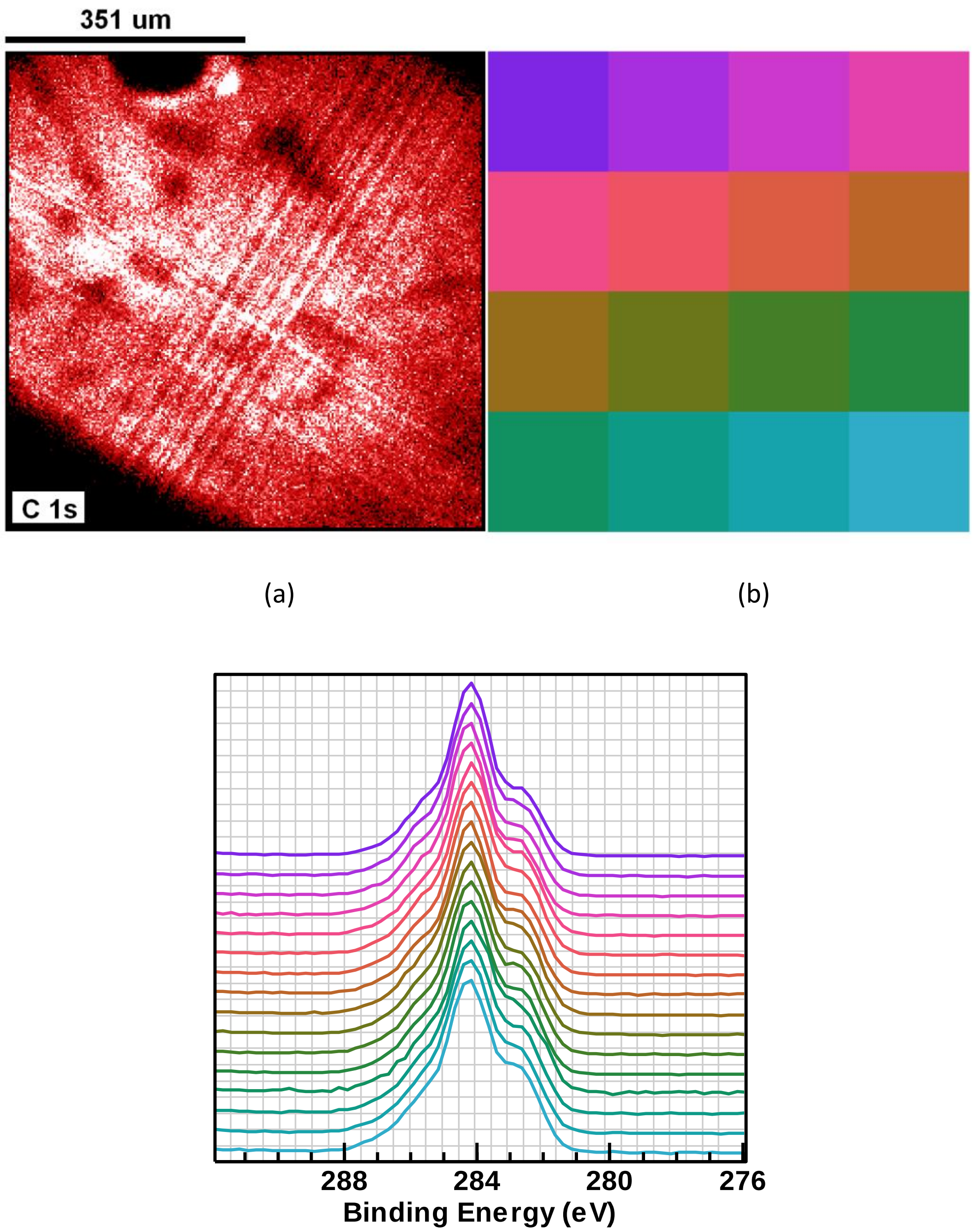
Figure 4. (a) Analysis of data set in Figure 2b into two component spectral forms labelled Comp 1 and Comp 2. The selected spectral forms, Comp 1 and Comp 2 are capable of fitting all 102 spectra with residual standard deviations consistent with pulse counted data. Comp 1 is a curve with spectral shapes compatible with cellulose. Comp 2 is a shape of unknown nature, but which is sufficient to illustrate that these cellulose spectra are constructed from at least six bell-shaped components to a peak model. (b) Data in both (a) and (b) are identical, however (b) illustrates a peak model determined via the fit of seven bell-shaped components to data in (a). The peak model was constructed with seven bell shaped components corresponding to two sets of bell-shaped components defined for each of the spectral forms labeled Comp 1 (three bell-shaped components) and Comp 2 (four bell-shaped components). The peak model in (b) is determined by using two separate nonlinear optimisation steps in which a fit of three bell-shaped components to Comp 1 is performed and then an independent fit to Conp 2 involving four bell-shaped components. The fit shown in (b) is the result of copying these two sets of bell-shaped components in the proportions calculated in (a) to (b). (c) The fit of the seven-component peak model in (b) to the entire data set of C 1s spectra in unison. Note how the separation in the energy of components achieved in (b) is not maintained in (c), but rather components appear to align in energy suggesting fewer components can be used in a peak model for this data set.

The method used to compute these component spectra Comp 1 and Comp 2 in Figure 4a is described in detail in Garland *et al*. [8]. While the solution offered in Figure 4a is not unique, similar analyses reveal similar peak structures, namely, one spectral-component peak model contains three bell-shaped components while the other spectral-component peak model contains four bell-shaped components. The significance of being able to identify two component-spectra in the form Comp1 and Comp 2 is that from these two component-spectra it is clear there are at least six bell-shaped curves that might underlie all spectra in the data set. The implication is that the fit in Figure 2c using four bell-shaped components is offering some information about the material, but is potentially hiding the true number of chemical states for these data. The peak model in Figure 4b highlights a problem that can limit the ability to fit a peak model to multiple spectra in unison in an attempt to determine the FWHM and energy for components in a peak model, namely, if it is assumed the peak model in Figure 4b is correct, then seven highly correlated bell-shaped components with no parameter constraints are unlikely to return appropriate FWHM and energy based on nonlinear optimization. The fit of two component-spectra Comp 1 and Comp 2 to the spectrum in Figure 4 is possible only because the shapes for these components derive from the data set and therefore represent any spectrum from the data set adequately by design. Linear analysis of a spectrum in terms of component-spectra calculates relative intensities for these seven bell-shaped components guided by the constraint of the specific shapes chosen for Comp 1 and Comp 2. By contrast, the peak model consisting of seven bell-shaped components when fitted to all 102 spectra results in the outcome shown in Figure 4c. No relational parameter constraints are applied to these bell-shaped components before the optimization is performed. The outcome of using nonlinear least squares to fit these seven bell-shaped components to all spectra without constraints is seldom repeatable. That is, the outcome for nonlinear optimization can be perturbed from the result shown in Figure 4c by changing the starting parameters, for example, so the conclusions drawn from the result in Figure 4c are subject to a degree of doubt. Nevertheless, Figure 4c does provide evidence that seven bell-shaped components when optimized to fit all spectra simultaneously, move components within the peak model so that bell-shaped components are less well defined in terms of energy, which suggests fewer than seven components are required to model these data in Figure 4c.

The results from applying these two methods to the cellulose sample surface should be considered in the context of the sample and the measurement process. Charge compensation is performed for these data and the apparent binding energy for spectra indicated the sample is measured whilst a net negative charge establishes the potential experienced by emitted electrons. When making use of an analysis area defined by FoV1, there is the possibility that different zones within the analysis area achieve different potentials and as the sample is measured the relative proportions of these zones change. The counter to this line of thinking for the coffee filter paper is the non-cellulose spectral component in Figure 4a (Comp 2) includes bell-shaped components that move to both lower and higher binding energy than the spectral component Comp 1. Further, an experiment in which the sample is imaged by XPS creating a data set of spectra over the analysis area shows little variation in C 1s spectra. Thus, at the micron scale, charge compensation appears to be uniform over the analysis area used to measure these data in Figure 4. The fact that both analyses Figure 2 and Figure 4 yield symmetrical bell-shaped components with FWHM close to unity and that data were measured using PE10 suggests the energy resolution achieved for these bell-shaped components is limited by the sample as measured with no obvious indication that charge compensation has distorted photoemission peak shapes. These observations about spectra used in the analysis shown in Figure 4 are compatible with results obtained by repeating the experiment using a range of cellulose samples measured by different researchers using different instruments. While the rate of change in these different cellulose samples differed depending on these variables in the analysis, the essential results shown in Figure 4 are reproducible. An ability to perform an analysis of data based on the approach in Figure 2 and also an analysis resulting in Figure 4a combine tools that help move towards a better understanding of data and therefore of a sample. In this sense, both approaches provide insight into spectra that increase knowledge about a sample without necessarily providing the full solution.

Testing the Stability of Charge Compensation

Stable charge compensation for photoemission, when measuring spectra from an insulating material such as cellulose, is important to ensure sets of spectra measured over an extended period, as shown in Figure 2, correctly correlate counts with energy. The analyses of spectra in Figure 2 and Figure 4 rely heavily on spectra that are free from variations that might result if the analysis area includes zones of differing potentials. In an attempt to support the assertion that data used in these analyses are without significant artifacts caused by issues with charge compensation, imaging XPS was used to examine lateral changes to C 1s spectra. The image in Figure 5a is constructed for the C 1s peak area from a data set that contains image data measured sequentially, where each image binding energy is incremented over an energy interval expected to include C 1s photoemission. The image for C 1s intensity (CPSeV) shows evidence that intensity variations over the analysis area do occur. Such variations may be a result of shadowing effects for X-ray flux due to topographical characteristics of the X-ray spot on the sample. The dark triangular zone in the bottom left-hand corner of the C 1s image in Figure 5a is the edge of an aperture, however other measurement artifacts in the form of striations are due to the delay-line detector used to collect spatially resolved intensity. Despite the non-uniform appearance of the C 1s map in Figure 5a, by summing spectra-at-pixels using the color mask shown in Figure 5b, spectra calculated from the image data set (Figure 5c) are generally stable concerning energy and the most intense photoemission signal in Figure 5c align for all sixteen spectra. The spectra in Figure 5c do vary to some extent in shape. However, a sample that changes chemistry during measurement by XPS may show lateral variation due to X-ray-induced changes at different points on the sample owing to variation in flux with the position. Another factor may be the sample itself started with differing levels of chemistry over the analysis area. A characteristic of cellulose is that damage or contamination of the pure chemistry for cellulose is a constant feature of spectra measured by XPS. It should also be noted that these spectra in Figure 5c are influenced by changes to sample chemistry observed in Figure 2. These changes in spectral shapes demonstrated in Figure 2 have an impact on spectra measured by imaging due to the length of time between the first image measured and the last. The period over which an imaging data set is measured compared to the time required to measure a spectrum in Figure 2 means spectral shapes in Figure 5c are not directly comparable to spectra in Figure 2. For these reasons and perhaps others, it is not surprising to observe minor differences in spectra such as those shown in Figure 5c. Nonetheless, the level of consistency of spectral shapes in Figure 5c suggests charge compensation applied to coffee filter paper is providing the necessary stability in energy required for the processing of C 1s spectra as performed in the paper.



(a)

(b)

(c)

Figure 5. (a) C 1s map of the area (CPSeV) measured over the analysis area from which spectra in Figure 2 are collected. (b) Colour zones defined for the C 1s map are used to identify locations suitable for accumulating spectra-at-pixels to form C 1s spectra with signal-to-noise appropriate for assessing the alignment of signal with binding energy. (c) C 1s spectra accumulated from spectra-at-pixels (color-coded to correlate with the color zones in (b)) computed from the XPS image data set used to construct the image in (a).

Conclusions

A commonly offered method for estimating FWHM and energy for bell-shaped components in a peak model obtained by fitting a peak model to multiple spectra in unison has been discussed. Concepts from the linear analysis are used to enhance an appreciation of this approach in which a fit of a peak model to multiple spectra identifies common trends in data sets capable of offering guidance when selecting FWHM and energy for components. Theory relating to least squares fitting of curves to data is developed which allows a comparison of results from PCA to the outcome for fitting a single peak model to multiple spectra. An example based on XPS of the cellulose sample surface is used to show that basic assumptions about the number of components in a peak model may yield plausible results without highlighting the need for further research into the possible chemistry of a sample. In conclusion, it is recommended to approach data treatment using a range of options, none of which individually provide a full picture but collectively can develop a more coherent picture for a sample composition.

Acknowledgments

This work by J.B. was supported as part of UNCAGE-ME, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award No. DE-SC0012577. The CNRS is acknowledged for financial support to the Thematic Workshop (N° 1317144) held at the Station Biologique, Roscoff, France.

Author contributions

Vincent Fernandez: Investigation (equal); Writing – review and editing (equal). David Morgan: Investigation (equal); Writing – review and editing (equal). Pascal Bargiela: Investigation (equal); Writing – review and editing (equal). Neal Fairley: Investigation (supporting); Methodology (lead); Writing – review and editing (equal). Jonas Baltrusaitis: Methodology (supporting); Supervision (lead); Writing – review and editing (equal).

Supporting information

A tutorial on the spectral manipulation of the data described in this work using CasaXPS [11] software is presented in the form of the tutorial videos at https://www.youtube.com/watch?v=2rJHQBUx8HE and https://www.youtube.com/watch?v=YVuTmJu1F4E.

References

[1] L.-S. Johansson, J.M. Campbell, Reproducible XPS on biopolymers: cellulose studies, Surf. Interface Anal. 36 (2004) 1018–1022. https://doi.org/10.1002/sia.1827.

[2] L. Johansson, J.M. Campbell, O.J. Rojas, Cellulose as the in situ reference for organic XPS. Why? Because it works, Surf. Interface Anal. 52 (2020) 1134–1138. https://doi.org/10.1002/sia.6759.

[3] M.N. Belgacem, G. Czeremuszkin, S. Sapieha, A. Gandini, Surface characterization of cellulose fibres by XPS and inverse gas chromatography, Cellulose. 2 (1995) 145–157. https://doi.org/10.1007/BF00813015.

[4] L. Fras, L.-S. Johansson, P. Stenius, J. Laine, K. Stana-Kleinschek, V. Ribitsch, Analysis of the oxidation of cellulose fibres by titration and XPS, Colloids Surfaces A Physicochem. Eng. Asp. 260 (2005) 101–108. https://doi.org/10.1016/j.colsurfa.2005.01.035.

[5] F. Khili, J. Borges, P.L. Almeida, R. Boukherroub, A.D. Omrani, Extraction of Cellulose Nanocrystals with Structure I and II and Their Applications for Reduction of Graphene Oxide and Nanocomposite Elaboration, Waste and Biomass Valorization. 10 (2019) 1913–1927. https://doi.org/10.1007/s12649-018-0202-4.

[6] L.-S. Johansson, J.. Campbell, K. Koljonen, P. Stenius, Evaluation of surface lignin on cellulose fibers with XPS, Appl. Surf. Sci. 144–145 (1999) 92–95. https://doi.org/10.1016/S0169-4332(98)00920-9.

[7] L.C.W. Bodenstein-Dresler, A. Kama, J. Frisch, C. Hartmann, A. Itzhak, R.G. Wilks, D. Cahen, M. Bär, Prospect of making XPS a high-throughput analytical method illustrated for a Cu x Ni 1− x O y combinatorial material library, RSC Adv. 12 (2022) 7996–8002. https://doi.org/10.1039/D1RA09208A.

[8] B.M. Garland, N. Fairley, N.C. Strandwitz, R. Thorpe, P. Bargiela, J. Baltrusaitis, A study of in situ reduction of MoO3 to MoO2 by X-ray Photoelectron Spectroscopy, Appl. Surf. Sci. 598 (2022) 153827. https://doi.org/10.1016/j.apsusc.2022.153827.

[9] V. Fernandez, D. Kiani, N. Fairley, F.-X. Felpin, J. Baltrusaitis, Curve fitting complex X-ray photoelectron spectra of graphite-supported copper nanoparticles using informed

line shapes, Appl. Surf. Sci. 505 (2020) 143841. https://doi.org/10.1016/j.apsusc.2019.143841.

[10] J. Baltrusaitis, B. Mendoza-Sanchez, V. Fernandez, R. Veenstra, N. Dukstiene, A. Roberts, N. Fairley, Generalized molybdenum oxide surface chemical state XPS determination via informed amorphous sample model, Appl. Surf. Sci. 326 (2015) 151–161. https://doi.org/10.1016/j.apsusc.2014.11.077.

[11] N. Fairley, V. Fernandez, M. Richard-Plouet, C. Guillot-Deudon, J. Walton, E. Smith, D. Flahaut, M. Greiner, M. Biesinger, S. Tougaard, D. Morgan, J. Baltrusaitis, Systematic and collaborative approach to problem solving using X-ray photoelectron spectroscopy, Appl. Surf. Sci. Adv. 5 (2021) 100112. https://doi.org/10.1016/j.apsadv.2021.100112.

[12] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420. https://doi.org/10.1007/BF02163027.