Automatic estimation of unknown chemical components in a mixed material by XPS analysis using a genetic algorithm

Ryo Murakami, Hideki Yoshikawa, Kenji Nagata, Hiroshi Shinotsuka, Hiromi Tanaka, Takeshi Lizuka, and Hayaru Shouno

ABSTRACT
There is an urgent need to develop automatic analysis methods for the large number of X-ray photoelectron spectroscopy (XPS) spectra being obtained by methods such as 3D chemical analysis and operand analysis. In a previous study, we developed an automatic analysis method for mixed materials that can decompose the XPS spectra and estimate the compositional ratios by comparison with XPS reference spectra of many candidate single-phase compounds. This method needs access to the XPS reference spectrum of every possible compound in the sample. However, in many practical cases, the necessary XPS reference spectra are not always available. In this study, we developed an automatic analysis method to estimate the compositional ratios, even when all necessary XPS reference spectra are not available, i.e., the measured XPS spectra contain unknown peak structures. In particular, the new method can automatically estimate the number of unknown peaks by the combination of a genetic algorithm and the Bayesian information criterion. We applied the method to analyze the depth-resolved XPS spectra of a Pb(Zr, Ti)O3 (PZT) piezoelectric film and successfully identified the change in the chemical states of the components in the film without ambiguity.

1. Introduction
X-ray photoelectron spectroscopy (XPS) is used to analyze the elemental species and chemical states of the surface of a material. It is widely used in materials science and quality control in industry. XPS can analyze the chemical state distribution by depth profiling and in-plane mapping, and can evaluate the evolution of reactions by operand analysis. However, there are few automatic analysis methods for XPS spectra, and it takes an enormous amount of time to analyze a set of spectra. In particular, XPS spectra consisting of multiple chemical state peaks can only be decomposed precisely when all necessary reference spectra for the possible chemical states are available. Therefore, there is an urgent need to develop automatic analysis methods to estimate compositional ratios, even when the measured XPS spectra contain unknown peak structures not found in known reference spectra.

Recently, some automatic analysis methods for various spectra have been proposed [1–7]. For example, the method proposed by M. Shiga [1] automatically identifies components from spectral data by matrix factorization with sparsity. The method proposed by
T. Matsumura [2] can analyze spectra by expectation-maximization algorithm for high-throughput analysis. The method proposed by K. Nagata [4] are able to estimate the peak parameters and the number of peaks by Bayes free energy automatically. It has the problem that the separated peaks were not based on chemical states. During the analysis of XPS spectra, the spectra are usually decomposed into spectra of known single-phase compounds (hereafter, reference spectra) to estimate the compound ratio of chemical states manually. In our previous study, we developed an automatic analysis method to estimate compositional ratios using the peak structures (e.g. energy position, peak width, and peak area) of reference spectra [8]. In particular, our previous method automatically adjusts the peak width (i.e. energy resolution) originating from each XPS instrument and the energy shifts (i.e. charging) caused by the sample condition and the X-ray flux density, to fit the measured XPS spectra. Therefore, the previous method can use reference spectra from the literature, reference spectral databases, and measurement data from other institutions. It automatically estimates the ratio of chemical states by combining reference spectra from different sources. In addition, the previous method estimates compound ratios and selects compound species automatically. The development of such methods is being enabled by the expansion of reference databases [9].

Researchers often analyze samples containing unexpected compounds originating from very rare materials, or materials with unstable chemical bonding states. Therefore, unknown peak structures not found in known reference spectra are often present in experimental XPS spectra. Therefore, it is common that only some of the reference spectra necessary for a precise decomposition of the XPS spectra are available. To achieve fully automated spectral analysis, it is important to develop automated analysis methods that can decompose XPS spectra, even when lacking the necessary reference spectra. To this end, we developed a new analysis method that can automatically decompose XPS spectra into both expected peak structures identified by known reference spectra, and unknown peak structures. In this study, we expanded our previous analytical model [8] that can estimate compound ratios for a composite sample automatically, to the proposed method that can estimate the structures and the number of unknown peaks not found in reference spectra. Specifically, we added independent peaks to the previous analytical model composed of only known reference spectra. The proposed method can estimate the parameters of the known reference spectra and independent peaks. We aimed to develop an automatic method that can analyze the XPS spectra even if all necessary reference spectra are not prepared.

2. Methods

Figure 1 shows a conceptual diagram of the proposed method, which can estimate both the compositional ratio of the known compounds from reference spectra and the unknown peak structures without reference spectra by introducing a term describing the e.g. energy position, peak width, and peak area of unknown peaks structures. In particular, the new method automatically estimates the number of unknown peaks.

![Figure 1. Conceptual diagram of the new method. Our method simultaneously analyzes multiple core-level XPS spectra for automatic estimation of the compound ratio, the proposed method expresses measured XPS spectra as the sum of known reference spectra and unknown peak structures (without reference spectra), the proposed method can estimate the parameters and the number of unknown peaks automatically.](image-url)
2.1. Analysis model of a single core-level

Given a spectral data point \( \{ (y_n, x_n) \} \), the observation intensity

\[
y_n = f(x_n; \Theta, \mathcal{W}),
\]

where the observation energy \( x_n \in \mathbb{R} \) is given,

\[
S(x_n; c, \mathcal{D}, \theta^i) = T_\theta(x_n; c, \mathcal{D}) + U(x_n; \mathcal{D}, \theta^i),
\]

where \( c = \{ c_k \}^{K_1}_{k=1} \) is the compound ratio,

\[
T_\theta(x_n; c, \mathcal{D}) = h \sum_{k=1}^{K_2} c_k R_\theta(x_n; \mathcal{D}),
\]

where \( h \) is intensity dependent, and

\[
R_\theta(x_n; \mu, \sigma) = A_k \sum_{m=1}^{M_k} a_{km} V(x_n; p_{km} + \mu_k, \sigma_k, y_{km}),
\]

\[
V(x_n; p, \sigma, \gamma) = \int_{0}^{\infty} G(x'; p, \sigma) L(x_n - x'; p, \gamma) dx',
\]

where \( \mu = \{ \mu_k \}^{K_1}_{k=1} \) is the shift of a peak structure,

\[
U(x_n; \mathcal{D}, \theta^i) = \sum_{k=1}^{K_2} g_k a' k V(x_n; p' k, \sigma, \gamma' k),
\]

where \( K_2 \) is the upper value of the number of unknown peaks, and the unknown peak parameter set is \( \theta^i = \{ (g_k, a' k, p' k, \gamma' k) \}^{K_2}_{k=1} \), where \( g_k \in \{ 0, 1 \} \) is the indicator vector. This means that the number of unknown peaks is the sum of the indicator vector \( \sum_{k=1}^{K_2} g_k \). Here, \( a' k, p' k \) and \( \gamma' k \) are the area intensity, peak position, and Lorentz natural width of an unknown peak, respectively.

2.2. Analysis model of multiple core levels

This section extends the analytical model of a single core level described in the above section to unify the analytical models of multiple core levels. We assumed that the compound ratios \( c \) are uniform in depth of a sample. Therefore, the compound ratios

\[
y_n^{(l)} = f(x_n^{(l)}; \Theta^{(l)}, \mathcal{W}^{(l)})
\]

\[
= S(x_n^{(l)}; \Theta^{(l)}) + B(x_n^{(l)}; \Theta^{(l)}, \mathcal{W}^{(l)}),
\]

where the observation energy \( x_n^{(l)} \) of core level

\[
S(x_n^{(l)}; \Theta^{(l)}) = T_\theta(x_n^{(l)}; c, \mathcal{D}) + U(x_n^{(l)}; \mathcal{D}, \theta^{(l)}),
\]

\[
= h^{(l)} \sum_{k=1}^{K_2} c_k R_\theta(x_n^{(l)}; \mathcal{D}, \theta^{(l)}).
\]

Here, the general parameters are \( \beta = \{ h^{(l)}, \alpha, \sigma, \theta^{(l)} \} \).

2.3. Design of loss functions

We used the Bayesian information criterion (BIC)[33] as the loss function \( L_0(\beta) \) for determining \( \beta \) when reference parameter sets \( \theta \) is given. The BIC gives arsult with a low error and a small number of unknown peaks. Here, \( L_0(\beta) \) is expressed using the negativelog-likelihood \( E(\beta) \), as follows:

\[
L_0(\beta) = 2E_0(\beta) + \lambda \log(N),
\]

\[
\hat{N} = \sum_{l=1}^{L} N^{(l)},
\]

\[
\lambda = b \sum_{l=1}^{L} \sum_{k=1}^{K_2} g_k^{(l)} + d,
\]

where \( \hat{N} \) is the number of data points in all spectral data, \( N^{(l)} \) is the number of data points in the spectral data of core level \( l \), \( \lambda \) is the overall parameter size, \( b \) is the number of parameters of an unknown peak, \( d \) is the parameter size of the reference spectrum, and \( \{ b, d \} \) are given and fixed. The number of parameters of an unknown peak is \( b = 3 \times \{ (a^l, p^l, \gamma^l) \} \).

The negative log-likelihood \( E_0(\beta) \) denotes an error function assuming Poisson noise.

\[
E_0(\beta) = \sum_{l=1}^{L} \sum_{n=1}^{N^{(l)}} [f(x_n^{(l)}) - y_n^{(l)} \ln \{ f(x_n^{(l)}) \} + y_n^{(l)}].
\]

The details of the formulation are given in Appendix.

Therefore, the estimated parameters \( \hat{\beta} \) are expressed as follows:

\[
\hat{\beta} = \arg \min_{\beta} \{ L_0(\beta) \}.
\]

In this study, we optimized the overall parameters \( \beta \) by a distributed genetic algorithm, which is a multipoint search method [10–13,34,35]. The genetic algorithm is an optimization method.

2.4. Genetic algorithm

The genetic algorithm is a multipoint optimization method with search heuristics inspired by natural evolutionary theory. The genetic algorithm has been used in materials analysis and its effectiveness has been shown [14–16]. The genetic algorithm has the following advantages over the gradient method used for the conventional analysis of XPS spectra;
• search the global solution without initial value dependence
• no need to calculate the gradients of the loss function
• easy to implement because of the simplicity of the algorithm

The genetic algorithm consists of five processes. In the following explanation, an individual and a population mean a parameter set \( \beta_i \) and solution set \( \{ \beta_i \}^n \), respectively. \( n \) is the number of individuals. The processes are described as follows:

(1) initial population generation
This process generates the parameter set from the generating distributions for each parameter. This operation is carried out for individuals to generate an initial population.

(2) error calculation
The error of each individual is calculated according to \( E_0(\beta_i) \). In this study, we use the BIC as the error function.

(3) selection
This process takes a duplicate sampling of the population according to the error values. In this study, the ranking selection is adopted. This method converts the error values to a probability \( P_i \) and performs sampling with overlap according to the probability

\[
e_i = E_0(\beta_i),
\]

\[
P_i = \frac{e_i^{-1}}{\sum_j e_j},
\]

where \( n \) is the number of individuals.

(4) crossover
This process selects two individuals from a population and generates a new individual by the linear combination of them. The linear combination is performed for each parameter, and the combination weights are determined by random numbers.

(5) mutation
This process takes the operation of replacing the sampled parameter values from the generating distribution with the mutation probability.

The genetic algorithm provides the optimal parameter solution by repeating the process (2)–(5). The genetic algorithm has hyperparameters that control each process: the number of individuals, the number of generations, the probability of doing the crossover, and the probability of doing the mutation. Here, the number of generations is the number of iterations of the process (2)-(5).

3. Experiments
In this study, we applied the proposed method to analyze both artificial and measured spectral data. The following subsections discuss the results of these two experiments. In this study, the PC specification of the experimental environment is AMD Ryzen Thread ripper 3990 x 64-Core Processor, 128 core, 4 GHz, 256GB, Ubuntu 18.04.5 LTS. We have developed the proposed method using C++ and GNU Scientific Library.

3.1. Procedure for generating artificial spectra
We generated XPS spectra assuming that the sample contained four compounds, SrTiO\(_3\), SrO, SrCO\(_3\), and TiO\(_2\) with a compound ratio of .40 : .15 : .20 : .25, respectively. To generate the artificial XPS spectra, we used the binding energy of peaks from several previous studies [17–22]. Table 1 shows reference information for the peak parameters used to generate the artificial spectral data. In the generation of artificial data, we used the binding energy positions of XPS peaks based on several previous studies [17–22]. We set Gaussian widths \( \sigma \) and Lorentz natural widths \( \gamma \) to .4 and .1, respectively. The area ratios \( a \) of the orbital splitting peaks(e.g. Sr3d\(_{5/2}\) and Sr3d\(_{3/2}\)) were taken from the theoretical values.

Figure 2 shows the generated artificial XPS spectra of (a) O1s, (b) Sr3d, and (c) Ti2p. Each spectrum contains the contributions from the four compounds based on the ratio shown above. We generated the background following the generative process of Shirley method. We add artificial spectra to Poisson noise. The signal intensity was set to 5000 counts. The analysis results of the XPS spectra without unknown peaks have been shown in the paper of the previous study [8].

3.2. Procedure for collecting measured XPS spectra
Here, we describe the results of applying the proposed method to measured XPS spectra. We analyzed depth-resolved XPS spectra of piezoelectric films

| Table 1. Peak parameter values used to generate artificial XPS spectra. Energy position \( \rho \) is from the literature [17–22]. |
|---------------------------------|--------------------|-----------------|-------------------------------|-----------------|
| Orbital | Compound | Atomic | Energy | Gauss | Lorentz |
| species | species | fraction \( A \) | position \( \rho \) | width \( \sigma \) | width \( \gamma \) |
| O1s  | SrCO\(_3\) | 0.60 | 531.1 | .40 | .10 |
| O1s  | SrO | 0.50 | 530.0 | .40 | .10 |
| O1s  | SrTiO\(_3\) | 0.60 | 529.1 | .40 | .10 |
| O1s  | TiO\(_2\) | 0.67 | 530.0 | .40 | .10 |
| Sr3d\(_{5/2}\) | SrCO\(_3\) | 0.20 | 133.6 | .40 | .10 |
| Sr3d\(_{3/2}\) | SrCO\(_3\) | 0.20 | 135.4 | .40 | .10 |
| Sr3d\(_{5/2}\) | SrO | 0.50 | 132.3 | .40 | .10 |
| Sr3d\(_{3/2}\) | SrO | 0.50 | 134.2 | .40 | .10 |
| Sr3d\(_{5/2}\) | SrTiO\(_3\) | 0.20 | 133.2 | .40 | .10 |
| Sr3d\(_{3/2}\) | SrTiO\(_3\) | 0.20 | 134.9 | .40 | .10 |
| Ti2p\(_{3/2}\) | SrTiO\(_3\) | 0.20 | 458.2 | .40 | .10 |
| Ti2p\(_{1/2}\) | SrTiO\(_3\) | 0.20 | 463.8 | .40 | .10 |
| Ti2p\(_{3/2}\) | TiO\(_2\) | 0.33 | 459.3 | .40 | .10 |
| Ti2p\(_{1/2}\) | TiO\(_2\) | 0.33 | 464.9 | .40 | .10 |
(Pb/Zr, Ti)O$_3$; PZT) covered with a platinum electrode. Depth-resolved XPS spectra were obtained using in situ Ar ion-beam etching. The film formation temperature of PZT was 700 degrees. It is firstly etched by Ar$^+$ mono-atomic ion beam at 5 keV energy for 1400 seconds. Then it is also etched by Ar400$^+$ cluster ions at 5 keV energy for 6000 seconds. For measurements, we used a high-performance X-ray photoelectron spectrometer (AXIS-ULTRA DLD, SHIMADZU/KRATOS). We used the Al-Ka (1486.6 eV) monochromatic X-rays for the X-ray source. We show the resolution of the measurement system. The width of Ag3d$_{5/2}$ peak was 0.80 eV when the pass energy was 40.0 eV. The vacuum was about $3.0 \times 10^{-8}$ Torr at the time of the measurement. It is known that during obtaining the depth-profile of the components in the sample, this etching process results in shifts and/or broadening of the XPS peaks due to charging effects and/or ion-beam-induced chemical degradation of the sample. Because of these complicated phenomena during ion-beam etching, depth-resolved XPS chemical analysis cannot usually be fully automated. The proposed method was able to detect the depth-resolved changes in the chemical states caused by the etching process under the condition that XPS peak shifts/broadening due to variations in charging were automatically tuned. We analyzed the depth-resolved XPS spectra after etching using the proposed method by referring to the peak structures of the spectra before etching.

4. Results and discussion

4.1. Evaluation of the proposed method for artificial spectra

We applied the proposed method to the generated artificial XPS spectra assuming that SrCO$_3$ was an unknown compound without a registered reference spectrum. Therefore, we tried to reproduce the peak structures of the SrCO$_3$ reference spectrum and the compositional ratio of known compounds, i.e. SrTiO$_3$, SrO, and TiO$_2$. In this experiment, we set maximum peak numbers of $K_{ul}^{(i)} = 3$ ($l$ = 01s, Sr3d, and Ti2p) for the unknown peak structures. The $y$ values of the reference spectra are often unknown because they are difficult to measure accurately. The $y$ values sometimes are not known in advance. We have also evaluated the proposed method in the case of estimating the $y$ values. Therefore, we applied our method to artificial XPS spectra under two different conditions, where $y$ was either estimated as a parameter or was given as a fixed value.

We describe the setting of optimizing by the genetic algorithm. The settings of the genetic algorithm were as follows: individual number of 30, island number of 60, probability of cross-over of .8, probability of mutation of .1, and generation number of 6000. In the optimization by the genetic algorithms, the generation distribution of parameters needs to be set. The settings used to generate the distributions of peak parameters used in the analysis of artificial spectral data were as follows:
\[ h \sim \text{Gam}(90, 50), \]
\[ y \sim \text{Gam}(1.0, 0.15), \]
\[ \sigma \sim \text{Gam}(7.0, 0.05), \]
\[ \mu \sim \text{Nor}(1.0, 0.1), \]
\[ a' \sim \text{Gam}(4.0, 1000.0), \]
\[ p' \sim \text{Uni}(\min(x), \max(x)), \]
\[ y' \sim \text{Gam}(1.0, 0.15), \]

where \( \text{Gam}(\cdot) \) is the gamma distribution, \( \text{Nor}(\cdot) \) is the normal distribution, \( \text{Uni}(\cdot) \) is the uniform distribution, and \( x \) is the binding energy of input spectral data. To generate the indicator vector \( g \), we first used the peak number \( \hat{K} = \sum_{j=1}^{K_0} g_j \) by sampling from the categorical distribution \( \text{Cat}((K_0^{-1}, \cdots, K_0^{-1})) \). Then, we set elements of size \( \hat{K} \) randomly sampled to 1, and other elements were set to 0.

Figure 2 (d-f) show the peak structures estimated by the proposed method. In these spectra, the peaks shaded in gray are the predicted unknown peaks, which correspond well to the artificial peaks of SrCO\(_3\) (Figure 2 (a,b)). Additionally, the number of unknown peaks estimated by our method matched the true value for SrCO\(_3\). The estimated ratio of known compounds (SrTiO\(_3\) : SrO : TiO\(_2\) = .412 : .141 : .247) corresponded well to the compositional ratio of the artificial peak structures (SrTiO\(_3\) : SrO : TiO\(_2\) = .40 : .15 : .25). The proposed method was able to automatically estimate the ratio of the compounds, even when some necessary reference spectra were not available. Table 2 showed the estimated peak parameter values by the proposed method. The calculation cost was about 10 minutes.

We consider the analysis result of applying the previous method [8], which estimates a compound ratio using the reference spectra automatically, to spectral data containing unknown components. The compound ratios estimated by the previous method differ significantly from the true values. This is because the previous method is unable to detect and estimate unknown components that are not referenced. The previous method has to force the unknown component to be represented by using the referenced spectra. As a result, the previous method provides incorrect analysis results in spectral data containing unknown components. In contrast, the proposed method automatically estimates unknown peaks other than the referenced ones. The proposed method can estimate the compound ratio close to the true value in spectral data containing unknown components.

The results of spectral analysis are affected by the noise pattern of spectral data. Therefore, we applied the proposed method to 50 artificial XPS spectra with different noise patterns, where the random number seed that generates the noise was varied to verify the accuracy of the proposed method. Figure 3 shows the frequency of the estimated number of unknown peaks of each core level: (a) O1s spectrum, (b) Sr3d spectrum, and (c) Ti2p spectrum. The blue bars indicate the case where the \( y \) values were parametrically estimated, and the red bars are for the case where \( y \) values were given. In Figure 3, the gray hatching shows the true number of unknown peaks, which is the number of peaks of SrCO\(_3\).

| Table 2. Estimated parameter values by the proposed method. This table shows the peak parameters of known reference spectra and unknown components. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Orbital | Compound | intensity | compound | Energy | Gauss | Lorentz |
| species | species | \( h/l \) | \( a' \) | ratio | position \( p \) | width \( \sigma \) | width \( y \) |
|-------|--------|---------|--------|-------|--------|--------|--------|
| O1s | SrO | 5003.72 | .244 | 530.05 | .375 | .106 |
| O1s | SrTiO\(_3\) | 5003.72 | .474 | 529.12 | .375 | .134 |
| O1s | TiO\(_2\) | 5003.72 | .282 | 529.99 | .375 | .244 |
| O1s | Unknown | 2224.71 | — | 531.12 | .375 | .109 |
| O1s | Unknown | .00 | — | — | — | — |
| O1s | Unknown | .00 | — | — | — | — |
| Sr3d | SrO | 3845.23 | .244 | 132.35 | .375 | .203 |
| Sr3d | SrO | 3845.23 | .244 | 134.25 | .375 | .162 |
| Sr3d | SrO | 3845.23 | .474 | 133.22 | .375 | .121 |
| Sr3d | SrO | 3845.23 | .474 | 134.92 | .375 | .106 |
| Sr3d | Unknown | 1930.46 | — | 133.58 | .375 | .144 |
| Sr3d | Unknown | 1332.32 | — | 133.41 | .375 | .124 |
| Sr3d | Unknown | .00 | — | — | — | — |
| Ti2p | SrTiO\(_3\) | 4945.91 | .474 | 458.25 | .375 | .151 |
| Ti2p | SrO | 4945.91 | .474 | 463.85 | .375 | .158 |
| Ti2p | TiO\(_2\) | 4945.91 | .282 | 459.23 | .375 | .116 |
| Ti2p | TiO\(_2\) | 4945.91 | .282 | 464.23 | .375 | .118 |
| Ti2p | Unknown | .00 | — | — | — | — |
| Ti2p | Unknown | .00 | — | — | — | — |
As shown in Figure 3, in the case of known $\gamma$, the proposed method estimated the true number of peaks with a probability of about 90%. Furthermore, in the case of estimating $\gamma$, the proposed method estimated the true number of peaks with a probability of about 80%. In particular, the proposed method accurately estimated the number of peaks in the Sr3d and Ti2p XPS spectra. Therefore, it is concluded that the proposed method is effective for detecting unknown peak structures. In estimating the number of unknown peaks, the accuracy rate of the O1s spectrum was lower than that of the Sr3d and Ti2p spectra. This is because the O1s spectrum has a more complex structure than the Sr3d and Ti2p spectra as it contains four compound species.

Figure 4 shows the fluctuations in the binding energy of the estimated peaks with unknown peak structure compared to the SrCO$_3$ reference values (black lines). Here, we show the case where the number of unknown peaks was estimated correctly in Figure 3. Since the Ti2p spectrum did not contain unknown peak structures (as SrCO$_3$ does not contain Ti), Figure 4 shows the results for the O1s and Sr3d spectra. The vertical axis in Figure 4 shows the index of the data depending on the random number seed. As shown in Figure 4, the proposed method shows the estimated binding energy of unknown XPS peaks accurately. Accurate estimation of the binding energy can enable correct identification of the compound species, chemical bond states, and electronic structure of the compound.

**Figure 3.** Frequency of estimated unknown number of peaks for each core-level. [(a) O1s, (b) Sr3d, and (c) Ti2p spectra], the blue and red bars are for the cases where $\gamma$ values were parametrically estimated or fixed, respectively. The gray area indicates the true number of unknown peaks.

**Figure 4.** Estimated binding energy position of unknown peaks without reference spectra. [(a) O1s, estimated $\gamma$, (b) Sr3d, estimated $\gamma$, (c) O1s, fixed $\gamma$, (d) Sr3d, fixed $\gamma$], the black lines are the true binding-energy positions. The vertical axis shows the index of the data depending on the random number seed.
unknown peaks. In the estimation of the unknown peak positions, there was no significant difference between the cases where γ values were estimated or fixed.

Figure 5 shows the estimated area ratio of the unknown compound (SrCO$_3$) compared to sum of all other the compounds, as determined from the peak areas. In Figure 5, the black dashed line is the true compound ratio, while the vertical axis shows the index of the data depending on the random number seed. Since the Ti2p spectrum did not contain unknown peaks from SrCO$_3$, only the results for the O1s and Sr3d spectra are shown. The estimated compound ratio of the unknown peak structures was close

![Figure 5](image_url)

Figure 5. Box-Plot of the estimated area ratio of unknown peak structures with reference spectra. [(a) O1s and (b) Sr3d spectra]. the black line shows the true area ratio. $K_t$ is the true number of unknown peaks, and $K$ is the number of estimated unknown peaks. the blue dots show the case where the true number of unknown peaks could be estimated. the vertical axis shows the index of the data depending on the random number seed. the box extends from the lower to upper quartile values of the data, with a line at the median. the whiskers extend from the box to show the range of the data. the length of the whiskers is set to 1.5 boxes. We showed the sample points in a jitter plot.

![Figure 6](image_url)

Figure 6. Triangular diagram of known compounds ratio of SrTiO$_3$, SrO, and TiO$_2$ for the cases where γ was (a) parametrically estimated or (b) fixed. (c) Estimation result of a previous method[8] using the entire reference spectrum, where the green point is the true compound ratio. the axes are in units of percent.
to that of the true ratio. As shown in Figure 5 (a), the estimation of the compound ratio is more accurate when the true number of peaks can be estimated (blue dots) than when the estimated number of peaks deviated by 1 (red dots) from the true number. As shown in Figure 5 (b), the estimated result for the Sr3d peaks had a higher variance than the estimated result for O1s. This is because the Sr3d spectrum for unknown compounds closely overlapped with the other Sr3d spectra for known compounds, as shown in Figure 2 (e).

Figure 6 show triangular diagrams of the known compound ratio of SrTiO3, SrO, and TiO2 estimated by the proposed new method using three known reference spectra. Figure 6 (a) and (b) show results for the cases where γ values were parametrically estimated and fixed, respectively. Figure 6 (c) shows the compound ratio estimated using all necessary reference spectra for SrCO3, SrTiO3, SrO, and TiO2 by the previous method [8]. In Figure 6, the green points indicate the true compound ratios. As shown in Figure 6, the estimated compound ratios were distributed around the true values, while satisfying the condition that the compound ratio of TiO2 was almost constant. In particular, the accuracy of the compound ratio was lower in the case of estimating γ compared to using a fixed γ. Therefore, it is important to use fixed γ values from the reference spectra to estimate compound ratios accurately, and these values need to be accurately extracted from the spectral analysis.

The advantage of our method is that it can estimate a ratio of known compounds and unknown components simultaneously. The conventional method has not been able to achieve the above. Our method contributes to the development of materials that handle often samples containing unknown components.

4.2. Limitations and robustness on the practical

In this section, we show the limitations and the robustness on the practical of the proposed method. The proposed method calculated the error assuming Poisson noise. When applying the proposed method to the spectrum with a count per second or arbitrary unit, it cannot detect unknown peaks normally. Therefore, the proposed method has a limitation that it can only be used for count data.

The proposed method used Shirley method as the background model. When the proposed method is applied to the spectrum with a complex background, it provides incorrect results that appears the pseudo-negative intensity. Therefore, the proposed method has a limitation that it can only be used for the spectrum with a simple background.

Our method cannot assign the unknown peaks to the chemical state automatically. Analysts have to query the chemical state of unknown peaks from the database on the practical. This method gives the analysts the energy positions of the unknown peaks needed to query the chemical states.

We show the analysis results when we set the maximum number of unknown peaks KU to 6, which is twice the value of the setting in the main text. Figure 7 shows the frequency of the estimated unknown number of peaks for each core-level. As shown in Figure 7, the unknown peaks were accurately detected even when KU was doubled. The proposed method is robust to increases in KU. However, if KU is excessively large, it is expected that the proposed method will not work due to the large search space of the parameters.

We show the noise robustness of the proposed method. Figure 8 -(d,f) show the results of the analysis of artificial data with the signal intensity set to 500 counts. The settings for data generation and analysis are the same as in the main text. In these spectra, the peaks shaded in gray are the predicted unknown peaks, which correspond well to the artificial peaks of SrCO3. Figure 8 -(a-c) show the true peaks. As shown in Figure 8 (b) and (e), the unknown peaks (SrCO3) with large overlap cannot be detected if the noise level is high.

![Figure 7](image-url)  
**Figure 7.** Frequency of estimated unknown number of peaks for each core-level. This result is for the case where the maximum unknown peak numbers are KU = 6 (a) O1s, (b) Sr3d, and (c) Ti2p spectral. The blue and red bars are for the cases where γ values were parametrically estimated or fixed, respectively. The gray area indicates the true number of unknown peaks.
Figure 9 shows the relationship between the signal intensity (noise level) and the detection of the unknown peak. The solid line shows the true number of peaks in each core-level. The sample points are the estimated number of unknown peaks. In this artificial data, the signal intensity of about 5000 counts is required to estimate the true number of unknown peaks. Additionally, the Sr3d spectrum is the most difficult to estimate the number of unknown peaks.

The proposed method works to not detect unknown peaks when applied to noisy spectral data. Therefore, the proposed method requires sufficient counting (measurement time) to accurately detect the unknown peak structure.

4.3. Applying the proposed method to measured XPS spectra

First, we show the settings for the generation distribution of peak parameters in the analysis of measured XPS spectral data. The settings are as follows:

\[ h \sim \text{Gam}(1.0, \max(y) - \min(y)), \]
\[ \sigma \sim \text{Gam}(1.0, 0.4), \]
\[ \mu \sim \text{Nor}(1.0, 0.01), \]
\[ a' \sim \text{Gam}(1.0, \max(y) - \min(y)), \]
\[ p' \sim \text{Uni}(\max(x), \min(x)), \]

Figure 9. Relationship between the signal intensity (noise level) and the detection of the unknown peak. The solid line shows the true number of peaks in each core-level. The sample points are the estimated number of unknown peaks.
where $\text{Gam}(\cdot)$ is the gamma distribution, $\text{Nor}(\cdot)$ is the normal distribution, $\text{Uni}(\cdot)$ is the uniform distribution, and $\{x,y\}$ are the binding energy and intensity of the input spectral data, respectively. In this experiment, we set the maximum peaks numbers of the unknown peak structures without reference spectra to $K_0^c = 4$ ($l = \text{O1s, Pb4f, and Pt4f}$).

Figure 10 (a-c) shows the measured XPS spectra of (a) O1s, (b) Pb4f, and (c) Pt4f before etching, along with the deconvoluted peaks using an independent Voigt function as the basis function. In addition, we automatically estimated the number of peaks $K$ using the Bayesian information criterion [33]. This experiment focused on the inner shell orbitals of three elements (O1s, Pb4f, and Pt4f) identified from the survey spectrum before etching. As shown in Figure 10 (b) the Pb4f XPS peak was detected even though a platinum electrode was deposited on the PZT film. The Pb component other than Pt, that is the electrode, was indicated. The results suggested that the thickness of the Pt electrode was not uniform, as shown in Figure 11 (a). Additionally, we detected only Pb (and not Zr and Ti) on the surface of the measured sample from XPS spectra before etching. It was assumed that a Pb oxide (PbO$_x$) layer may be generated by heat treatment and/or atmospheric exposure. Therefore, the decomposed peaks can be assigned as follows. In Figure 10 (a), the peak at around 530 eV was attributed to the oxidation of lead (PbO$_x$). The two peaks at around 520 eV were attributed to Pt4f$_{3/2}$. In Figure 10 (b), the peaks may be the oxidation of lead (PbO$_x$). In the main peak (Pb4f$_{3/2}$), the high and low binding energy peaks were located at 138.2 and 137.2 eV, respectively. These peaks may be PbO and PbO$_2$ because it was in agreement with the energy position of the literature [23–26]. The energy positions of PbO and PbO$_2$ in the literature are 137.9–138.2 eV [23,24] and 136.8–137.4 eV [25,26], respectively. In Figure 10 (c), the peaks indicate Pt metal derived from the platinum electrode. Figure 11 (a) shows a model of the cross-sectional morphology of the sample before etching based on these results.

Figure 10 (d-f) shows the results of applying the proposed method to XPS spectra measured after etching the sample by referring to the peak structures of the XPS spectra before etching. In this figure, the orange peaks are reference peak structures that were estimated from XPS spectra before etching, and the gray shaded peaks are unknown peak structures detected by the proposed method. Table 3 shows the

**Figure 10.** Analysis result of measured XPS spectra using the proposed method. (a)-(c) XPS spectra before etching and peak deconvolution for (a) O1s, (b) Pb4f, and (c) Pt4f. $K$ is the number of peaks estimated using Bayesian information criterion automatically. (d)-(f) Analysis results of XPS spectra after etching and peak deconvolution for (d) O1s, (e) Pb4f, and (f) Pt4f. The orange peaks are reference components detected before etching, the peaks filled with gray are unknown components detected by the proposed method. In this figure (e), the red and blue dot lines are peaks positions that are considered metal Pb and PZT, respectively.
estimated parameter values by the proposed method. The calculation cost was about 11 minutes. As shown in Figure 10 (e), four unknown peaks were detected in the Pb4f spectrum. The two unknown peaks (red dotted lines) were located around 136.6 eV and 141.5 eV. The first peak position was identical to that of Pb metal (136.60 eV) listed in reference databases [27–29]. As shown in Figure 10 -(d,f), the unknown peaks were derived from Pb-containing components because there were no large unknown peaks in the O1s and Pt4f energy regions. The unknown low-binding-energy peaks showed that the etching process reduced the PbOx to Pb metal. As shown in Figure 10 (a) and (d), the intensity of the O1s peak around 530 eV was decreased by etching compared to the Pt4p3/2 peak, consistent with the hypothesis that reduction of PbOx occurred.

The Pb4f results in Figure 10 (e) shows that the two unknown peaks around 144.0 eV and 139.2 eV (blue dotted lines) were detected at higher binding energies than those of PbOx (orange peaks). As shown in Figure 10 -(d,f), these unknown peaks can originate from PZT, which was not detected in the XPS spectrum before etching. Furthermore, it is consistent with the appearance of Zr and Ti components indicated by the Zr3d and Ti2p peaks after etching. Figure 12 (a,b)
shows narrow XPS spectra after and before etching. Figure 12 (a) and (b) show Ti2p and Zr3d spectra, respectively. As shown in Figure 12, it was confirmed that the etching process removed the deposited platinum, and Zr and Ti (i.e. PZT) were observed by XPS. Figure 12 (c-d) show the fitting result of (c) Ti2p and (d) Zr3d by using the automatic peak separation method. As shown in Figure 12, these spectra may be mostly dominated by the PZT component. Additionally, the Zr3d spectrum has a small amount of sub-oxides derived from damage to the sample caused by etching [30]. However, the position of the peak (139.19 eV) attributed to the Pb4f contribution of PZT estimated by the proposed method was about 1.4 eV higher than the peak position (137.80 eV) from the literature [31]. This implies that charging probably occurred in the PZT film during XPS measurements, which is likely considering that PZT is a ferroelectric insulating material with a perovskite structure. Also, it is possible that the pyrochlore phase or another decomposed phase is present in the measured sample after etching.

Since large unknown peaks were not detected in the Pt4f region in Figure 10 (f), it was concluded that the chemical states of Pt4f components were almost unchanged after etching. As shown in Figure 10 (d) and (f), other tiny unknown peaks were detected at the binding energy positions indicated by gray arrows in the Pt4f spectrum. These peaks were located around 5 eV above the main peaks (indicated by dotted lines for both core levels). Therefore, these detected unknown peaks might not be photoelectron peaks but energy-loss peaks.

This analysis provided insights into the phenomena before and after etching, as schematically shown in Figure 11 (a) and (b). The assumed cross-sectional morphology model in Figure 11 (a,b) is one of the sample models only from the analysis result of XPS spectra. The result of unknown component analysis by the proposed method suggested that the etching process reduced the PbO to Pb metal and the substrate remained as the PZT film. In conclusion, the proposed method is effective for automatic depth-resolved XPS analysis under complex conditions where the chemical states of the constituent components change.

5. Conclusion

In this study, we aimed to develop an XPS analysis method to automatically estimate compositional ratios in a mixed material under the conditions where the measured XPS spectra contains unknown peak structures not found in known reference spectra. The proposed method can automatically decompose complicated measured XPS spectra by adjusting the peak parameters (binding energy, width, intensity, and shape) of the known peak structure of XPS reference spectra of candidate compounds and the peak parameters of unknown peak structures simultaneously. In particular, the method combining a genetic algorithm and Bayesian information criterion was able to reasonably estimate the number of unknown peaks, and accurately estimate the energy position of unknown peak structures, which are important for identifying compound species or chemical states.

We applied the method to effectively identify the complex change of chemical states from depth-resolved XPS spectra of a PZT piezoelectric film. The proposed method contributes to automatically and reasonably estimating unknown peak structures in practical situations where both known candidate compounds and unknown compounds coexist in a sample. Such methodologies are essential for future materials development.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by EXT KAKENHI under grant [number 18K05191]; and JSPS KAKENHI under grant [number 19K12154].

ORCID

Ryo Murakami http://orcid.org/0000-0001-8585-9268
Kenji Nagata http://orcid.org/0000-0001-9894-4461
Hiroshi Shinotsuka http://orcid.org/0000-0001-5147-1396
Hayaru Shouno http://orcid.org/0000-0002-2412-0184
References

[1] Shiga M, Tatsumi K, Muto S, et al. Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization. Ultramicroscopy. 2016;170:43–59.

[2] Matsumura T, Nagamura N, Akaho S, et al. Spectrum adapted the expectation-maximization algorithm for high-throughput peak shift analysis. Sci Tech Adv Mater. 2019;20:733–745.

[3] Murakami R, Kagayama H, Nakamura K, et al. Background estimation in X-ray photoelectron spectroscopy data using an active Shirley method with automated selection of the analytical range. E-J Surf Nanotechnol. 2019;17:61–68.

[4] Nagata K, Sugita S, Okada M. Bayesian spectral deconvolution with the exchange Monte Carlo method. Neural Netw. 2012;28:82–89.

[5] Matsumoto R, Nishiwaya Y, Katoa N, et al. Reproducibility of XPS analysis for film thickness of SiO2/Si by active Shirley method. J Electron Spectrosc Relat Phenom. 2016;207:55.

[6] Murakami R, Nakamura K, Tanaka H, et al. Efficient removal of noise-derived components for automatic XPS spectral decomposition using hierarchical clustering. E-J Surf Nanotechnol. 2020;18:201–207.

[7] Shinotsuka H, Yoshikawa H, Murakami R, et al. Automated information compression of XPS spectrum using information criteria. J Electron Spectrosc Relat Phenom. 2020;239:146903.

[8] Murakami R, Tanaka H, Shinotsuka H, et al. Development of multiple core-level XPS spectra decomposition method based on the Bayesian information criterion. J Electron Spectrosc Relat Phenom. 2020;245:147003.

[9] Suzuki M, Nagao H, Hanada Y, et al. Raw-to-repository characterization data conversion for repeatable, replicable, and reproducible measurements. J Vac Sci Technol A. 2020;38:023204.

[10] Proctor A, Sherwood PMA. Data analysis techniques in X-ray photoelectron spectroscopy. Anal Chem. 1982;54:13–19.

[11] Holland JH. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan Press; 1975.

[12] Shirley DA. High-resolution X-ray photoemission spectrum of the valence bands of gold. Phys Rev B. 1972;5:4709.

[13] Herrera-Gomez A, Bravo-Sanchez M, Cebllos-Sanchez O, et al. Practical methods for background subtraction in photoemission spectra. Surf. Interface Anal. 2016;46:897–905. DOI:10.1002/sia.5453

[14] Wierzbanowski K, Tarasiku J, Lodini A. Mater sci. 2010;6:652.

[15] Jennings PC, Lysgaard S, Hummelshøj JS, et al. Genetic algorithms for computational materials discovery accelerated by machine learning. npj Comput Mater. 2019;5:46. DOI:10.1038/s41524-019-0181-4

[16] Roberts J, Bursten JRS, Risko C. Genetic Algorithms and Machine Learning for Predicting Surface Composition, Structure, and Chemistry: A Historical Perspective and Assessment. Chem Mater. 2021;33(17):6589–6615.

[17] Pilleux ME, Grahm arg CR, Fuenzalida VM. Hydrothermal Strontium Titanate Films on Titanium: An XPS and AES Depth-Profiling Study. J Am Ceram Soc. 1994;77:1601–1604.

[18] Sosunnikov MI, Teterin YA, Doklady AN. X-ray photoelectron studies of Ca, Sr and Ba and their oxides and carbonates. J Electron Spectrosc Relat Phenom. 1991;59(2):111–126.

[19] Aas N, Pringle TJ, Bowker M, et al Adsorption and decomposition of methanol on TiO2, SrTiO3 and SrO. Soc Faraday Trans. 1994;90:1015–1022.

[20] Chaschechnikova IT, Vorotyntsev VM, Borovik VV, et al Strong metal-carrier interaction in cobalt- and nickel-titanium dioxide co-hydrogenation catalysts. Theor Eksp Khim. 1992;28(216):174–176.

[21] Vasquez RP. X-ray photoelectron spectroscopy study of Sr and Ba compounds. J Electron Spectrosc Relat Phenom. 1991;56(3):217–240.

[22] Lange F, Schmelz H, Knozinger H. An X-ray photoelectron spectroscopy study of oxides of arsenic supported on TiO2. J Electron Spectrosc Relat Phenom. 1991;57(3):307–315.

[23] Morgan WE, Van Wazer JR. Binding energy shifts in the x-ray photoelectron spectra of a series of related Group IVa compounds. J Phys Chem. 1973;77:964.

[24] Baker JM, Johnson RW, Pollak RA. Surface analysis of rf plasma oxidized In and PbmAu films using ESCA. J Vac Sci Technol. 1979;16:1534.

[25] Taylor JA, Perry DL. An x-ray photoelectron and electron energy loss study of the oxidation of lead. J Vac Sci Technol A. 1984;2:771.

[26] Pederson LR. Two-dimensional chemical-state plot for lead using XPS. J Electron Spectrosc Relat Phenom. 1982;28:203.

[27] McGilp JF, Weightman P, McGuire EJ. The N6,7O4,504.5 Auger spectra of thallium, lead and bismuth. J Phys C. 1977;10:3445.

[28] Bertrand PA, Fleischauer PD. X-ray photoelectron spectroscopy study of the surface adsorption of lead naphthenate. J Vac Sci Technol. 1980;17:1309.

[29] Kim KS, O’Leary TJ, Winograd N. X-ray photoelectron spectra of lead oxides. Chem Phys Lett. 1973;19:209.

[30] Sugiyama O, Murakami K, Kaneko S. XPS analysis of surface layer of sol-gel-derived PZT thin films. J Eur Ceram Soc. 2004;24:1157–1160.

[31] Change WL, He JL. XPS Studies of PZT Films Deposited by Metallic Lead and Ceramic PZT Dual Target Co-Sputtering. J Electroceramics. 2004;13:47–50.

[32] Bishop C Pattern Recognition and Machine New York (USA): Learning: Springer; 2006.

[33] Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–464.

[34] Goldberg DE. Genetic Algorithm in Search, Optimization and Machine Learning. Boston: Addison-Wesley; 1989.

[35] Gordon VS, Whitley D, Serial and parallel genetic algorithms as function optimizers. Proceedings of the Fifth International Conference on Genetic Algorithms, San Francisco;1993: 177–183.
Appendices

Detailed formulation in the Bayesian estimation

In this section, we show a supplementary explanation of the formulation for the loss function. In the photoelectron spectrum, the measured intensities $y$ are count data. The data

$$P(y|\beta, K) = \prod_{n=1}^{N} \frac{e^{-f_{n} y_{n}}}{y_{n}!},$$

(19)

where $\beta$ is parameter set,

$$E(\beta, K) = -\ln P(y|\beta, K),$$

(20)

$$= \sum_{n=1}^{N} [f_{n} - y_{n} \ln f_{n} + y_{n}].$$

(21)

In the Bayesian estimation, the posterior distribution $P(K|y)$ is estimated by the Bayes Theorem and the marginalization to get the optimal number of peaks

$$P(K|y) = \frac{P(K, y)}{P(y)} \propto P(K, y)$$

(22)

$$= \int P(K, \beta, y)d\beta$$

(23)

$$= \int P(y|K, \beta)P(\beta|K)P(K)d\beta$$

(24)

$$\propto \int P(y|K, \beta)d\beta$$

(25)

$$= \int \exp[-E(\beta, K)]d\beta.$$ 

(26)

We can estimate the number of peaks $K$ using the Bayes free energy $F = -\ln P(K|y)$.

The Bayesian information criterion (BIC) is the criterion value obtained by the Laplace approximation of the posterior distribution $P(K|y)$ [32]. The BIC requires that the posterior distribution can be approximated by a Gaussian distribution. This is related to the fact that we are dealing with the Poisson noise. According to the central limit theorem, we can assume that the statistical noise is generated from a Gaussian distribution if the counts of the spectral data are sufficiently large. We consider the BIC to work without the problem for the spectral data with a sufficient counting. It is important to investigate the behavior when the posterior distribution cannot be approximated by a Gaussian distribution, and that is a future work.

In this study, the loss function $L(\beta)$ was defined using the BIC. The BIC can select a model, such as the order or the number of bases[7,8]. In this study, the BIC is used to select the number of unknown peaks. The

$$F_{BIC} = 2E(\beta, K) + \lambda \ln N,$$

(27)

where $\lambda$ is the number of parameters.