Principal Component Analysis

Introduction

XPS is a technique that provides chemical information about a sample that sets it apart from other analytical tools. However, the key information sought by the analyst is locked into a data envelope and as a consequence the need for powerful algorithms is paramount when reducing the data to chemically meaningful quantities. Two approaches have been employed on XPS data:

- Curve synthesis and fitting (see Section on Quantification).
- Techniques from multivariate statistical analysis of which Principal Component Analysis (PCA) is the most common form.

Curve synthesis is probably the most widely used method for data analysis practised by XPS researchers. Unfortunately, statistically good curve fits are not always physically meaningful and, in many cases, great care must be exercised when choosing the model to describe the data. Any assistance in understanding the model is therefore of great value and it is with this end that Principal Component Analysis is offered as a supplementary tool.

Factor analysis is a field that is as broad as it is deep. It is a mathematically challenging tool that requires knowledge of matrix algebra coupled with a feel for a statistical approach to data interpretation. A true understanding for the subject can only be obtained by studying the literature[1] and through practical experience. Therefore the material presented here is only an introduction rather than a complete set of works.

Theory Behind Principal Component Analysis

Factor analysis is a multivariate technique for reducing matrices of data to their lowest dimensionality by use of orthogonal factor space. The challenge is to identify the number of significant factors (principal components) and use this information to model the data using techniques such as Target Transformations or curve fitting.

In XPS the data matrix is composed of spectra where each acquisition channel is viewed as a co-ordinate in an r-dimensional space; r is equal to the number of acquisition channels per spectrum. The problem addressed by PCA is that of determining the number of distinct spectroscopic features present in a particular set of c spectra.

The following example tries to illustrate the nature of the problem. Consider a set of three spectra; each spectrum has three acquisition channels:

s1 = (4, 3, 6**) , s2 =** (2, 3, 2**) , s3 =** (2, 0, 4**)**

The data matrix is given by

$$D = \begin{pmatrix} 4 & 2 & 2 \\ 3 & 3 & 0 \\ 6 & 2 & 4 \end{pmatrix}$$

These three vectors belong to a 3-dimensional space, however they do not span 3-dimensional space for the following reason. If a linear combination of the vectors **s1**, **s2** and **s3** is used to construct a new vector **v**, then **v** always lies in a plane (a 2-dimensional sub-space of 3-dimensional space). The fact that **v** lies in a plane is a consequence of the following relationships between the three spectra.

s3 = **s1** – **s2**,

so **v** = a **s1** + b **s2** + c **s3**
= a **s1** + b **s2** + c (**s1** – **s2**)
= (a + c) **s1** + (b – c) **s2.**

Thus, two principal components exist for the set of three spectra.

The analysis of the data matrix in the above simple example has been performed by observation. Unfortunately real spectra are not so simple and spotting the linear relationships between the columns of the data matrix requires a more sophisticated approach.

PCA, also known as Eigenanalysis, provides a method for identifying the underlying spectra that form the building blocks for the entire set of spectra. The data matrix is transformed into a new set of r-dimensional vectors. These new vectors span the same subspace as the original columns of the data matrix, however they are now characterised by a set of eigenvalues and eigenvectors. The eigenvalues provide a measure for the significance of the abstract factors with respect to the original data. Various statistics can be computed from these values that aid in identifying the dimensionality of the subspace spanned by the spectra.

The procedure for calculating the abstract factors has its roots in linear least square theory. In fact the preferred method is to form a Singular Value Decomposition (SVD) for the data matrix.

$$D = USV'$$

Where *D* is the data matrix formed from c spectra, each containing r channels. *U* is the same dimension as *D*, while *S* and *V* are c by c matrices. *S* is a diagonal matrix; the diagonal elements are the square root of the eigenvalues of the correlation matrix

$$Z = D'D$$

The abstract factors are computed from *US*. The rows of *V* are the corresponding eigenvectors of *Z*; the co-ordinates of the eigenvectors represent the loading for the abstract factors and specify how linear combinations of these factors can be used to reproduce the original data. Including all of the abstract factors with the appropriate loading enables the data to be reproduced to an accuracy only limited by the precision characteristic of the Eigenanalysis procedure.

The essential feature of the SVD procedure is to compute the abstract factors so that the factor corresponding to the largest eigenvalue accounts for a maximum of the variation in the data. Subsequent abstract factors are generated such that 1) as much variance as possible is accounted for by each new factor and 2) the newest

axis is mutually orthogonal to the set of axes already located. The procedure therefore computes an orthogonal basis set for the subspace spanned by the original data matrix that is oriented with respect to the data in a linear least square sense.

In principle, the number of non-zero eigenvalues is equal to the number of linearly independent vectors in the original data matrix. This is true for well posed problems, but even the presence of errors due to numerical operations will result in small eigenvalues that theoretically should be zero. Numerical errors are an insignificant problem compared to the one presented by the inclusion of experimental error in the calculation. Noise in the data changes the underlying vectors so that almost every data matrix of c spectra with r acquisition channels, where c <= r, will span a c-dimensional subspace. This is true even though the underlying vectors should only span fewer than c dimensions.

Various statistics are available for identifying the mostly likely dimensionality of a data matrix. These statistics are designed to aid partitioning the abstract factors into primary and secondary factors. The primary factors are those corresponding to the largest n eigenvalues and represent the set of abstract factors that span the true subspace for the data. The secondary factors are those factors that can be associated with the noise and, in principle, can be omitted from subsequent calculations. It is not possible to completely disassociate the true data from the error within the measured data, however the statistics guide the analyst in choosing the most appropriate number of abstract factors that describe the data and therefore the “best guess” dimensionality for the data matrix.

In the case of XPS spectra the experimental error is known to be the square root of the number of counts in an acquisition channel. Under these circumstances where the experimental error is known, a number of statistics have been proposed for determining the size of the true factor space.

Residual Standard Deviation (RSD or Real Error RE)

An alternative name for the RSD (used by Malinowski) is the Real Error (RE).

The RSD is defined to be:

$$RSD_n = \sqrt{\frac{\sum_{j=n+1}^{c} E_j}{r(c-n)}}$$

where E_j is the j^{th} largest eigenvalue, n is the number of abstract factors used to reproduce the data; c spectra each with r channels are used to construct the data matrix.

RSD_n must be compared against the estimated experimental error. If the value computed for RSD_n is approximately equal to the estimated error then the first n abstract factors span the factor space. The dimensionality of the original data matrix is therefore n.

Two further statistics may be derived from RSD_n, namely, IE_n (Imbedded Error) and IND_n (Indicator Function) given by:

$$IE_n = RSD_n\left(\frac{n}{c}\right)^{1/2}$$

And

$$IND_n = \frac{RSD_n}{(c-n)^2}$$

IE_n and IND_n are statistics that should decrease as the number of primary abstract factors is increased. Once all the primary factors have been included, these statistics should begin to increase since at this point factors from the noise subspace start to interfere with the accuracy of the data description. This minimum is therefore an indicator of the dimensionality of the data subspace.

Chi-square

Bartlett[2] proposed using the chi-square criterion for situations similar to XPS data, where the standard deviation varies from one data channel to the next.

The procedure involves reproducing the data matrix using the abstract factors. Each abstract factor is progressively included in a linear combination in the order defined by the size of the eigenvalues and weighted by the co-ordinates of the corresponding eigenvectors. The chi-square value for a set of n abstract factors is computed using:

$$\chi_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(d_{ij} - \underline{d}_{ij})^2}{\sigma_{ij}^2}$$

where d_{ij} is an element of the data matrix, $\underline{d}_{ij}$ is the corresponding approximation to the data point constructed from the first n abstract factors with the largest eigenvalues. The standard deviation for XPS data s_{ij} is the square root of d_{ij}.

The expected value for each n is given by c_n^2 (expected) = (r-n)(c-n). A comparison between the expected value and the computed value is the basis for determining the number of principal components. Both c_n^2 and its expected value decrease as n increases. c_n^2 initially is larger than c_n^2 (expected) but as n increases a crossover occurs. The true dimensionality of the data matrix is chosen to be the value of n for which c_n^2 is closest to its expected value.

Note that smoothing the data will alter the characteristics of the noise. Performing such pre-processing therefore invalidates the c_n^2 statistic.

Target Factor Analysis

Principal Component Analysis provides a set of basis vectors that describe the original set of spectra. Although useful as a means of characterising the data, these abstract factors are in general not physically meaningful. Target Factor Analysis is concerned with identifying vectors that can also describe the data, but with the additional property that they are recognisable as spectra rather than simply abstract vectors in an r-dimensional space.

There are numerous methods for transforming the PCA abstract factors to provide vectors that are more open to chemical interpretation. These involve constructing abstract rotation transformations that map the abstract factors into one of the infinite number of alternative basis sets for the factor space. Fortunately there is a technique which when coupled with curve synthesis, lends itself to the analysis of XPS data, namely, Target Testing.

Target Testing

Once a Principal Component Analysis has been performed, the mathematical bridge between abstract and real solutions is Target Testing. Individual spectra can be evaluated to assess whether the corresponding vector lies in the subspace spanned by the chosen primary abstract factors. The essential feature of Target Testing is to form the projection of the target vector onto the subspace spanned by the primary factors, then compute the predicted target vector using this projection. Statistical tests applied to the predicted and test vectors determine whether these two vectors are one and the same. These tests serve as a mean of accepting or rejecting possible fundamental components of the sample.

Ultimately, the goal of target testing is to identify a set of spectra that span the same subspace as the primary abstract factors. Complete models of real factors are tested in the target-combination step. In the combination step the data matrix is reproduced from the real factors (spectra) rather than from abstract factors and by comparing the results for different sets of real factors, the best TFA solution to a problem can be determined.

Testing a target vector **x** with respect to the chosen set of primary abstract factors involves forming the projection **t** onto the subspace spanned by the PCA primary abstract factors. The predicted vector **x,** calculated using the co-ordinate values of **t** to load the corresponding abstract factors, is compared to the original target vector. A target vector that belongs to the subspace spanned by the primary abstract factors should result in a predicted vector that is identical to the initial target vector. Errors in the original data matrix and similar errors in the measured target vector mean that the predicted and target vector differ form each other as well as from the pure target vector $\mathbf{x}^*$ (**x** but without error). Estimates for these differences allow a comparison to be made between the predicted and target vector and a decision as to which targets to include in the target combination step.

The apparent error in the test vector (AET) measures the difference between the test and predicted vectors in a root mean square (RMS) sense. Similarly two other RMS quantities estimate the real error in the target vector (RET) and the real error in the predicted vector (REP)[3]. These error estimates form the basis for the SPOIL function defined to be approximately equal to the ratio RET/REP.

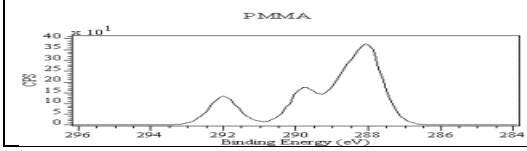
Principal Component Analysis by Example

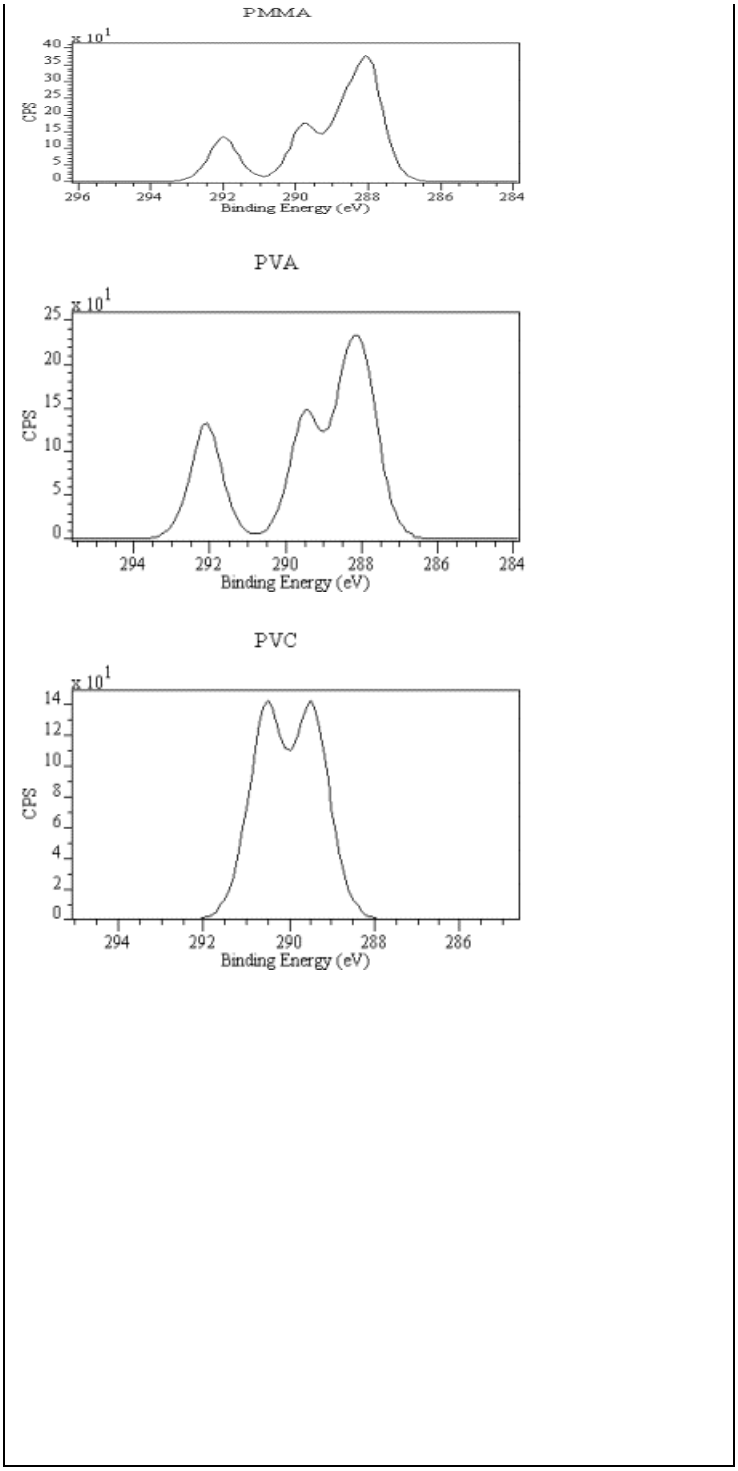
The first example illustrating the characteristics of PCA uses a set of artificial data.

Three sets of spectra prepared from synthetic components are used in the PCA. The structure of the artificial data derives from Carbon 1s states within three compounds, namely, PMMA, PVA and PVC (Figure 1.). The proportion of each compound varies throughout each set of ten VAMAS blocks. The data is located in the files c1stest1.vms, c1stest2.vms and c1stest3.vms. The underlying trends introduced into each file are as follows: peaks corresponding to PMMA and PVC obey quadratic adjustments in intensity over the set of ten spectra (PMMA decreases while PVC increases). The difference between the three files is the proportion of PVA in each data envelope. The first file (c1test1.vms) has a constant level of PVA (Figure 2); the second file (c1stest2.vms) varies linearly, first increasing then decreasing; the third file (c1stest3.vms) includes a linear increase in the level of PVA.

The objective is to show how the statistics used in PCA behave for a known problem. Data matrices constructed from the three sets of spectra should have a dimensionality of three.

Figure 1: Artificial C 1s Data

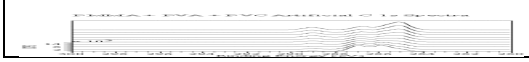


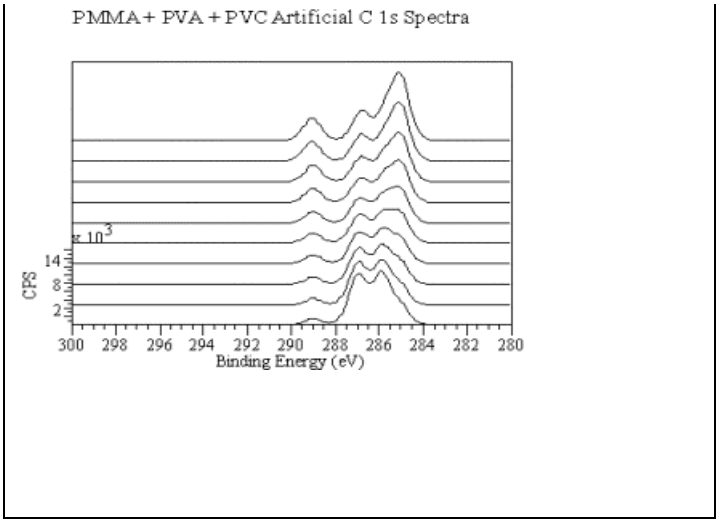


Note that, although each compound is constructed from a number of C 1s peaks (PMMA 4, PVA 4 and PVC 2), the stoichiometry of these compounds masks the true number of synthetic components actually present in the data. Hence the dimensionality of the data should be three not ten (4+4+2). An additional twist to this example is that two of the underlying envelopes are similar in shape to each other, though not identical (see Figure 1).

The trend throughout the first data set may be seen in Figure 2.

Figure 2: c1s_test1.vms C 1s Spectra





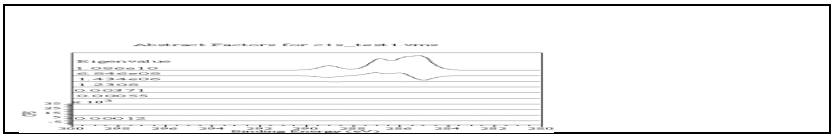
No noise is present in the data; therefore eigenvalues belonging to the primary set of three abstract factors should be non-zero, while the remaining seven eigenvalues should be zero. The results of applying PCA to these data sets (Table 1) illustrate the uncertainty associated in estimating the dimensionality of the data matrix from the statistics. The fourth largest eigenvalue in each case is small but non-zero. Also the statistics for IE and IND indicate a minimum at eigenvalues other than the expected result. Chi-square is not a valid statistic since no noise is present in the data, however it does show that three abstract factors are sufficient to reproduce the data to within reason.

Table 1: PCA report for file c1stest.vms

Factor	Eigenvalue	RMS	RE (RSD)	IE	IND * 1000	Chi-sq Calc.	Chi-sq Expected
C 1s/1	10861690000	266.7944	615.8053	194.7347	7602.534	29055.89	1800
C 1s/2	684568100	13.0963	29.86145	13.35445	466.5852	172.5176	1592
C 1s/3	1433862	0.01064209	0.02964045	0.01623474	0.6049071	0.000190564	1386
C 1s/4	1.230771	0.000391784	0.002107629	0.001332982	0.05854525	9.86356E-08	1182
C 1s/5	0.003712633	0.000385433	0.001279202	0.000904532	0.05116807	1.20367E-07	980
C 1s/6	0.000545838	0.000230365	0.001168993	0.000905498	0.07306205	4.86336E-08	780
C 1s/7	0.000473059	0.000240069	0.001018602	0.000852223	0.113178	5.18512E-08	582
C 1s/8	0.000306354	0.000155331	0.000891207	0.000797119	0.2228016	2.20338E-08	386
C 1s/9	0.000200465	0.00012725	0.00076887	0.000729414	0.7688698	1.91008E-08	192
C 1s/10	0.000118823	4.68061E-13	0	0	0	6.13259E-23	0

It is interesting to see how the eigenvalues change with respect to the three data sets (Figure 3 and Figure 4). The same spectra varied in different ways results in slightly different orientations for the principal component axes and hence different eigenvalues.

The PCA statistics IE and IND have implied a dimensionality other than three (Table 1). The clue to the correct dimensionality of the data lies in the relative size of the eigenvalues. The fourth eigenvalue is in two cases better than five orders of magnitude smaller than the third eigenvalue. This statement has been made with the benefit of a good understanding of what is present in the data. In real situations such statements are themselves suspect and so require support from other data reduction techniques. For example curve fitting using three sets of synthetic peaks all linked with the appropriate stoichiometric relationships would lend support to the hypothesis. Curve fitting such structures is not an exact science and such fits themselves should be supported by statistics gathered from the fitting parameters.



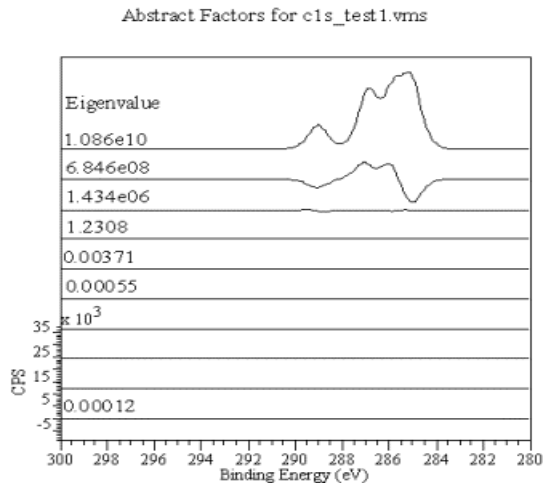


Figure 3: Abstract Factors and Eigenvalues.

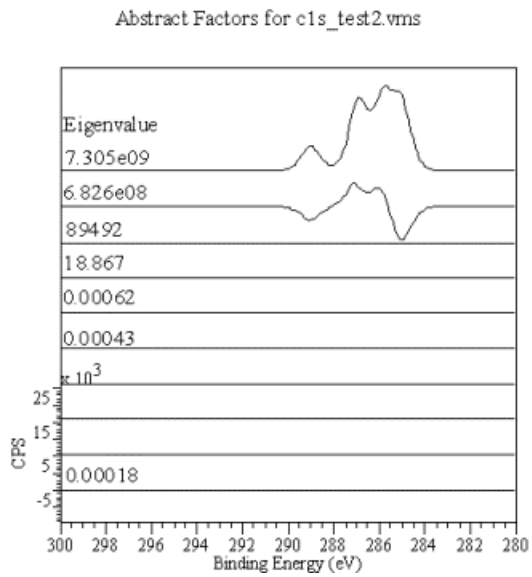


Figure 4. Eigenvalues for c1s_test2.vms Abstract Factors

The second example illustrates the effects of experimental error on a PCA calculation.

Table 2: PCA Applied to Artificial Data with Simulated Noise.

Factor	Eigenvalue	RMS	RE (RSD)	IE	IND * 1000	Chi-sq Calc.	Chi-sq Expected
C 1s/1	10883620000	267.626	617.9728	195.4202	7629.294	50489.63	1800
C 1s/2	687172800	18.93302	47.75373	21.35612	746.152	3572.687	1592
C 1s/3	2714766	6.90718	26.01386	14.24838	530.8951	483.7729	1386
C 1s/4	946794.7	0.1649431	2.106561	1.332306	58.51557	0.5688374	1182
C 1s/5	5095.574	0.0622169	0.5048702	0.3569971	20.19481	1.247441	980
C 1s/6	226.9936	0.01656361	0.1904914	0.147554	11.90571	0.002773735	780
C 1s/7	29.13143	0.002688086	0.00847462	0.007090376	0.9416245	5.0833E-05	582
C 1s/8	0.03943089	0.000339289	0.003105158	0.002777337	0.7762894	3.05649E-07	386
C 1s/9	0.003409306	0.000206314	0.001523904	0.001445703	1.523904	9.22657E-08	192
C 1s/10	0.000466779	1.26006E-12	0	0	0	7.4912E-22	0

Real data includes noise. The effect of noise on a PCA calculation can be see from Figure 5 together with the report in Table 2. The data in the file c1stest1.vms has been used together with a pseudorandom number generator to simulate noise that would typically be found in XPS data. The consequence of including a random element in the data is that the eigenvalues increase in size and lead to further uncertainty with regard to which eigenvalues belong to the set of primary abstract factors. Note that the abstract factors in Figure 5 are plotted in the reverse order to the ones in Figure 3 and Figure 4.

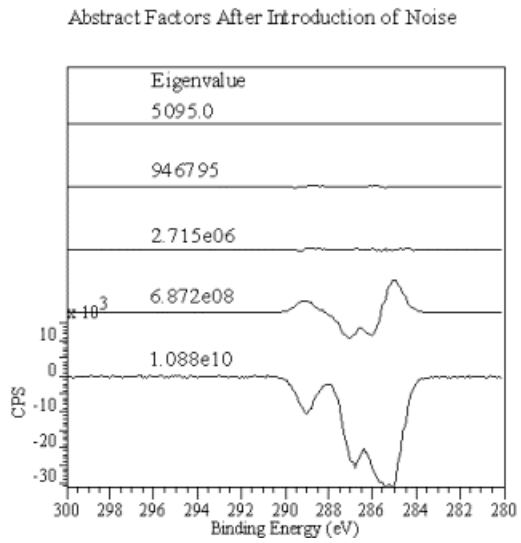


Figure 5: c1stest1.vms data plus artificial noise.

Fortunately, the chi-square statistic becomes more meaningful when noise is introduced into the problem. A comparison between the computed chi-square and its expected values do seem to point to a 3-dimensional subspace. The crossover between the two quantities suggests the need for three abstract factors when approximating the data matrix using the results of PCA.

Principal Component Analysis and Real Data

XPS depth profiles generate sets of spectra that are idea for examination via PCA. The spectra are produced by repeatedly performing etch cycles followed by measuring the count rate over an identical energy region. The resulting data set therefore varies in chemical composition with respect to etch time and the common acquisition conditions provide data in a form that is well suited to PCA.

An aluminium foil, when profiled using a Kratos Analytical Axis Ultra, provides a good example of a data set that can be analysed using some of the features on offer in CasaXPS. The data is not chemically or structurally interesting, but does show how trends can be identified and anomalies isolated.

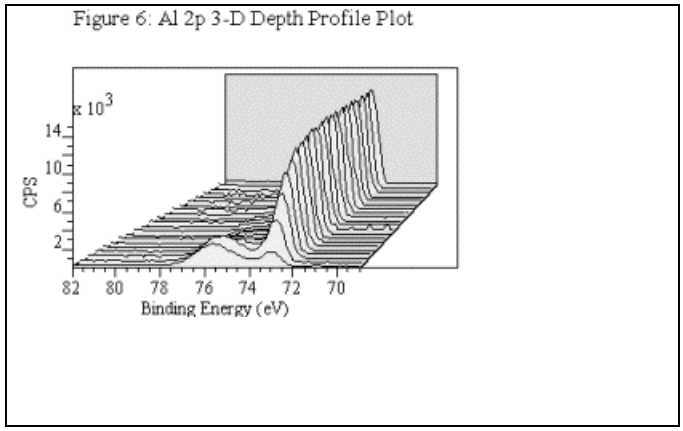


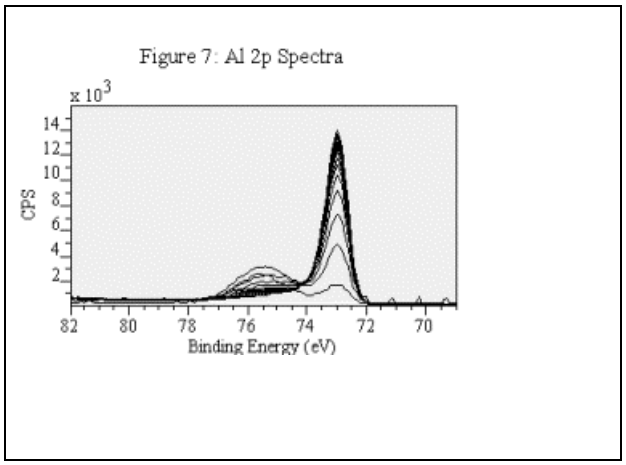
Figure 6 shows a plot of the Al 2p energy-region profiled against etch time. The data envelopes change in shape as the surface oxide layer is removed by the etch cycles to reveal the homogeneous bulk aluminium metal.

It should also be noted from Figure 6 that the data contains an imperfection. One of the energy scans includes data acquired during an instrumental event. Noise spikes are superimposed on the true data and these should be removed before techniques such as curve synthesis are applied. In this case the spikes appear in the background and are therefore obvious to the eye, however similar non-physical structure that occurs on the side of a peak is less obvious and could be missed.

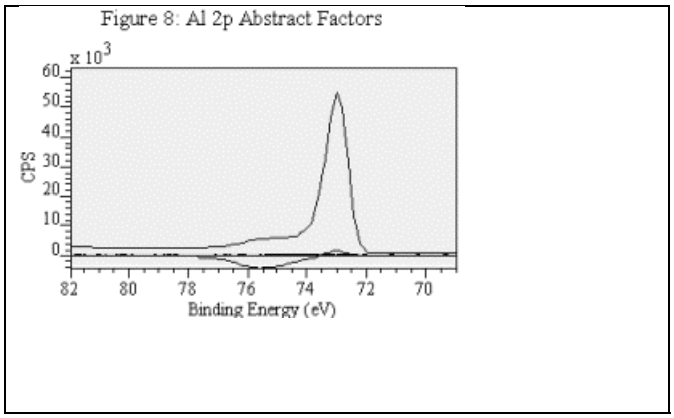
The first step in performing the Principal Component Analysis is to define a quantification region for each of the spectra to be used in the analysis. These regions specify the acquisition channels that will be used in the data matrix. Also any shifts in the data due to charging can be removed from the calculation by using an offset in the energy region for those spectra affected.

Next, select the set of spectra in the Browser View and display the data in the active tile. The processing property page labelled "PCA" offers a button labelled "PCA Apply". On pressing this button, those spectra displayed in the active tile are transformed into abstract factors. Figure 7 displays the spectra before the PCA transformation while Figure 8 shows the abstract factors generated from the eigenanalysis.

Note that the abstract factors are truly abstract. The first factor (Figure 8) looks like an Al 2p metal doublet, however this is because the Al 2p metal envelope dominates the data set and therefore a vector having a similar shape accounts for most of the variation in the overall set of spectra. A more even weighting between the underlying line-shapes would produce abstract factors that are less physically meaningful in appearance.



The only real use for the abstract factors is judging their significance with respect to the original data. Abstract vectors that derive from noise look like noise, factors that contribute to the description of the data contain structure. The dividing line between the primary and secondary abstract factors can sometimes be assessed based on the appearance of the abstract factors.



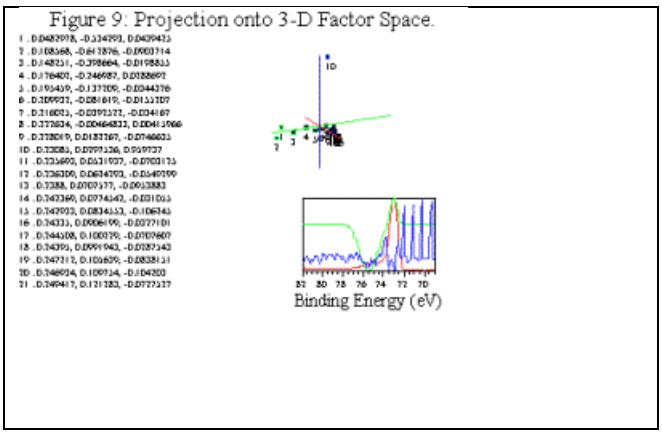
Analysing the Al 2p spectra generates abstract factors and eigenvalues that represent the PCA fingerprint for the data. Table 3 is a report of the Al 2p data set generated by CasaXPS and formatted using a spreadsheet program. Each row of the report is labelled by the VAMAS block name that contains the abstract factor corresponding to the listed eigenvalue.

Table 3: Report generated by a PCA for Al 2p Profile.

Factors	Eigenvalue	RMS	RE (RSD)	IE	IND * 1000	Chi-sq Calc.	Chi-sq Expected
Al 2p/3	4727872000	82.5236	159.3106	34.76442	398.2765	26032.99	2600
Al 2p/8	65086870	3.784555	23.78739	7.340948	65.89305	136.4193	2451
Al 2p/13	393178.7	5.875731	20.74929	7.842494	64.04102	133.7949	2304
Al 2p/18	305001.5	3.48073	17.85783	7.793798	61.79182	78.0949	2159
Al 2p/23	156694.8	3.249367	16.25038	7.929371	63.47803	59.6866	2016
Al 2p/28	98359.06	2.757161	15.2192	8.135007	67.64091	51.1035	1875
Al 2p/33	86168.29	2.48836	14.18397	8.189118	72.36718	41.82519	1736
Al 2p/38	65267.54	2.22333	13.35424	8.242415	79.01916	37.98182	1599
Al 2p/43	53613.14	2.247765	12.61316	8.257255	87.59142	42.46731	1464
Al 2p/48	43569.08	0.1744253	11.97161	8.261198	98.93895	0.1869893	1331
Al 2p/53	32387.23	1.710532	11.52946	8.344409	115.2946	29.01139	1200
Al 2p/58	28174.98	2.021671	11.12658	8.410907	137.3652	33.42906	1071
Al 2p/63	24742	1.261896	10.75487	8.461885	168.0448	15.17705	944
Al 2p/68	23980.27	0.40768	10.29759	8.407944	210.1548	1.649807	819
Al 2p/73	20710.42	1.113217	9.867347	8.33943	274.093	10.69842	696
Al 2p/78	18345.12	1.155456	9.424948	8.226769	376.9979	15.75993	575
Al 2p/83	16109.71	0.7358818	8.960655	8.062218	560.0409	5.605129	456
Al 2p/88	13003.76	0.7303461	8.600543	7.962556	955.6159	5.882052	339
Al 2p/93	12307.15	0.7049177	7.99876	7.608339	1999.69	6.193881	224
Al 2p/98	9285.948	0.000443667	7.554815	7.372745	7554.815	2.12747E-06	111
Al 2p/103	7476.855	1.90305E-13	0	0	0	2.03675E-25	0

The chi-square indicates that the data matrix can be reproduced to within experimental error using two abstract factors. This is a result that is consistent with the physical nature of the sample. It is also interesting (from a mathematical standpoint) to note that using all the abstract factors to reproduce the data matrix returns a chi-square of zero (allowing for round-off errors in the computation). This should always be the case and provides an easy check to see that the calculation has been performed correctly.

All the statistics expect the Indicator Function point to two abstract factors being sufficient to span the factor space for the data matrix.



It is worth examining the data set using a subset of the spectra and Target Testing the spectra not used in the PCA. This allows anomalies to be identified such as spikes in the data. Selecting a representative subset of spectra for the PCA then target testing the remainder is particularly useful for large sets of data.

Table 4: Target Test Report for a Subset of Al 2p Data Set.

Target	AET	REP	RET	SPOIL	Al 2p/3	Al 2p/8

Al 2p/23	20.93032	10.38374	18.17295	1.750135	0.39033	-0.04947
Al 2p/28	23.83028	11.05117	21.11288	1.910467	0.41839	0.017257
Al 2p/33	19.83736	11.47927	16.17861	1.409376	0.42997	0.065749
Al 2p/38	19.8507	12.01348	15.80274	1.315418	0.44268	0.10609
Al 2p/43	19.9069	12.46508	15.52116	1.245171	0.4531	0.133366
Al 2p/48	57.16561	12.70691	55.73546	4.386233	0.45854	0.14688
Al 2p/53	15.37333	13.18052	7.912861	0.600345	0.46791	0.174614
Al 2p/58	21.39836	13.30379	16.76004	1.259795	0.46901	0.184805
Al 2p/63	19.92528	13.5238	14.63296	1.082016	0.47386	0.195062
Al 2p/68	27.73522	13.78354	24.06775	1.746122	0.48087	0.203826
Al 2p/73	19.10189	13.88023	13.12332	0.945469	0.48192	0.210646
Al 2p/78	20.9575	13.98145	15.61204	1.116625	0.48264	0.218455
Al 2p/83	19.03813	14.15492	12.7314	0.899433	0.48483	0.229382
Al 2p/88	18.38591	14.11378	11.78317	0.83487	0.48374	0.228046

The SPOIL function and AET statistics (Table 4) show that Al 2p/48 differs in some respect from the other spectra in the list tested. The spectrum in question corresponds to the trace displaying the spikes seen in Figure 6. Also, another spectrum that could be looked at is Al 2p/68. The AET value is high compared to the other spectra. Such spectra may highlight interfaces where either new chemical states appear (either directly from features in the data or indirectly through changes in the background due features outside the acquisition region) or energy shifts due to sample charging have altered the characteristics of the data.

The PCA report in Table 3 includes the spectrum labelled Al 2p/48 in the data matrix. The consequence of not removing the spikes is apparent in the 3-D factor space shown in Figure 9, where the abstract factor with third largest eigenvalue clearly contains spikes and the projection point number 10 derived from the Al 2p/48 spectrum is obviously a statistical outlier.

PCA and CasaXPS

Principal Component Analysis is offered on the “processing” window. The options on the property page labelled “PCA” allow spectra to be transformed into abstract factors according to a number of regimes. These include covariance about the origin and correlation about the origin. Each of these pre-processing methods may be applied with and without background subtraction.

Quantification regions must be defined for each spectrum included in the factor analysis. In addition, each spectrum must have the same number of acquisition channels as the others in the set of spectra to be analysed. The first step in the calculation replaces the values in each spectrum by the result of interpolating the data within the defined quantification region for the spectrum. This is designed to allow energy shifts to be removed from the data used in the factor analysis.
The quantification region also provides the type of background to the spectrum. Performing the analysis on background subtracted data attempts to remove artifacts in the spectrum that derive from other peaks within the vicinity of the energy region. Background contributions can be significant in PCA. Additional primary abstract factors are often introduced as a consequence of changes in the background rather than the underlying peaks within the region of interest. The presence of such abstract factors can be viewed as information extracted from the data, although in many circumstances they can lead to incorrect synthetic models if background contributions are misunderstood.

A factor analysis is performed on the set of spectra displayed in the active tile. Although PCA is offered as a processing option, it is the only processing option that acts on a collection of spectra. Any other option from the processing window would only act upon the first VAMAS block in a selection when that selection is displayed in a single tile.

The principal component analysis is performed when the “Apply” button is pressed. Each spectrum displayed in the active tile is replaced by the computed abstract factors. The order of the VAMAS blocks containing the spectra is used as the order for the abstract factors. The factor corresponding to the largest eigenvalue is entered first. Subsequent blocks receive the abstract factors in descending order defined by the size of the corresponding eigenvalues. A report showing the statistics for understanding the dimensionality of the factor space appears in a dialog window.

A button labelled “PCA Report” allows the current PCA report to be re-displayed. Care should be exercised since the values are subject to any additional processing (including PCA) that may subsequently be applied to any of the spectra included in the original analysis.

The PCA property page includes a button to reset the processing operations for every spectrum displayed in the active tile. This allows a PCA calculation to be undone in one stroke. It will also undo any processing previously performed on the data. PCA is aimed at the raw data; the chi-square statistic is referenced to the raw data and has an undefined meaning when the data have been processed prior to performing factor analysis.

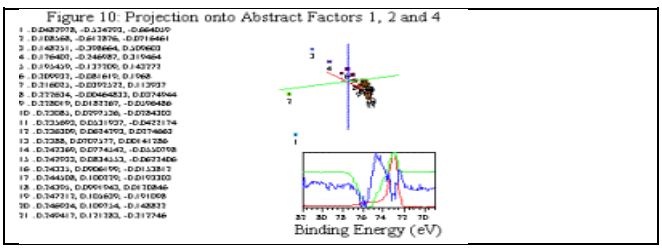
Target Factor Analysis in the form of target testing is also available on the PCA property page. Following a PCA, candidates for the physically meaningful components may be assessed individually or collectively. Choose an abstract factor from the PCA and entering this factor into the active tile. Then select the number of primary abstract factors for use in the target test procedure. A text field is offered on the PCA property page for this purpose and is found in the section headed "Target FA". Next, select the target test spectra in the Browser view and press the button labelled "TFA Apply". A report detailing the statistics calculated from the TFA procedure will appear in a dialog window.

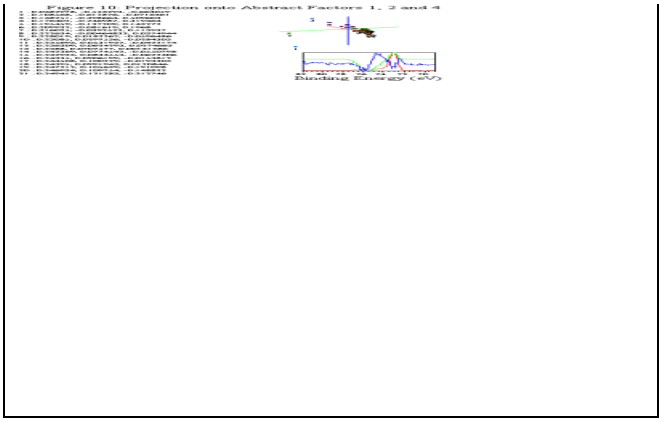
The TFA report may be written to file in an ASCII format with TAB separated columns. When pressed, any of the buttons above the columns on the report will display a file dialog window from which the output text-file can be specified. This method for saving a report to file is used by the PCA report (above) and the Linear Regression Report described below.

Once a set of target spectra has been identified, these spectra can be used to reproduce the original set of spectra through a linear regression step. Enter the set of target spectra into the active tile; then select the original spectra in the Browser view. Press the button labelled "Linear Regression". A report shows the RMS differences between each of the original spectra and the predicted spectra calculated from a linear combination of the set of target spectra displayed in the active tile. The loading used to compute the predicted spectra are listed in the report. The report may be written to file using a similar procedure to the TFA report described above.

Viewing the Data in Factor Space

CasaXPS offers an option on the “Geometry” property page on the “Tile Display” dialog window labelled “Factor Space”. If selected, the VAMAS blocks displayed in a tile are used to define the axes for a subspace and the original data are plotted, if possible, as a set of co-ordinates with respect to these axes. The plot represents a projection of the data space onto the subspace defined by a set of two or three abstract factors.





The abstract factors defining the axes are graphed together with a list of the co-ordinate values for each of the spectra projected onto the subspace spanned by the chosen abstract factors (Figure 9). A 3-dimensional plot provides a visual interpretation for the spectra. Patterns formed by the spectra highlight trends within the data set and the relative importance of the abstract factors can be examined. A plot in which the axes are defined by unimportant factors generally appear random, while factors that are significant when describing the data typically produce plots containing recognisable structure.

---

[1] E.R. Malinowski. Factor Analysis in Chemistry. Wiley. NY. 1991
M. Meloun, J. Militky and M.Forina. Chemometrics for Analytical Chemistry. Ellis Horwood NY. 1992
J.N. Fiedor, A. Proctor, M. Houalla and D.M. Hercules. Surface and Interface Analysis Vol. 20 1-9 (1993)
[2] M.S. Bartlett, Brit. J. Psychol. Stat. Sect. 3, 77 (1950).
[3] E.R. Malinowski. Factor Analysis in Chemistry. Wiley. NY. 1991

