Uncertainty Estimates in XPS

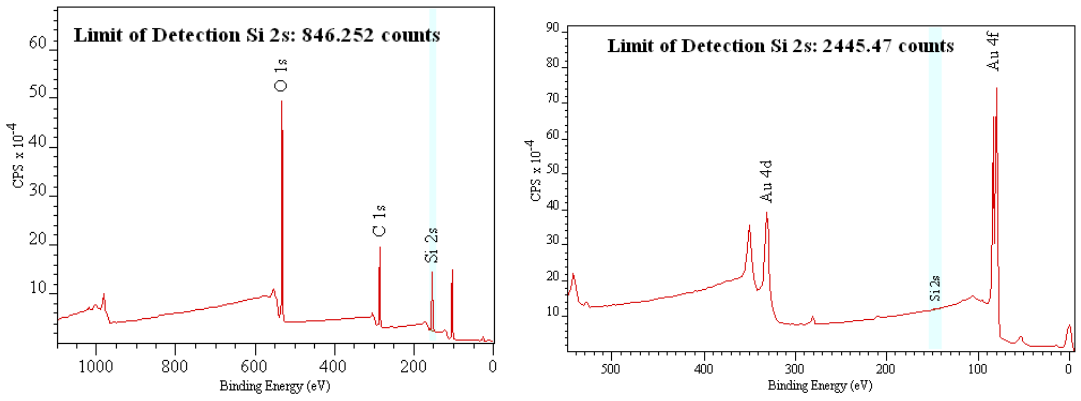
Limit of Detection

When acquiring XPS data a number of variables determine the quality of a spectrum in terms of signal-to-noise. Examples of variables are:

1. X-ray flux
2. Analyser pass energy
3. Aperture settings
4. Dwell time per data channel
5. Area of sample available for analysis

When measuring the signal for a transition from a given element, the absence or presence of a peak must be assessed in the context of the sample and the acquisition conditions. While a peak may be visible for data acquired at PE160, the same peak may be missing from a measurement of the same sample at PE20 simply because the time taken to measure the PE20 spectrum was insufficient to distinguish a peak from the background and noise when using the reduced signal on offer at the lower pass energy. The question therefore arises: for a given set of sample and acquisition conditions, at what point can it be concluded an element is present in the surface?

It should be emphasised that the limit of detection is a value specific to a given measurement. The value for the detection limit depends both on the sample composition and the localised signal intensity, which may change due to instrumental measurement conditions and the proximity of peaks in the spectrum. For example, the detection limit for silicon in the presence of gold is different than for silicon in a carbon polymer.



The reason for the difference in detection limit for these two samples is the position of the Si 2s peak and the background intensity beneath the Si 2s peak resulting from the Si 2p for the polymer and the Au 4f peak of gold. The noise for pulse counted intensity is assumed to be Poisson in nature and therefore the expected noise level is characterised by the square root of the counts per bin. For a conservative detection limit, three times the standard deviation for the given data is the measure indicated in the spectra above.

The limit of detection for the Si 2s peak in the polymer sample would change if the amount of silicon decreased or the silicon was buried or only located near the sample surface, since all these scenarios

would alter the intensity of the background beneath the Si 2s peak. The limit of detection is therefore a value appropriate for a given measurement and is not transferrable between samples.

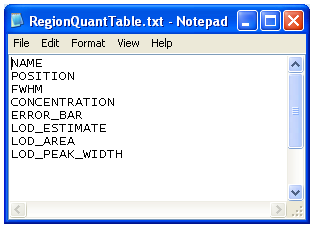
Two methods for calculating the limit of detection are offered in CasaXPS 2.3.16:

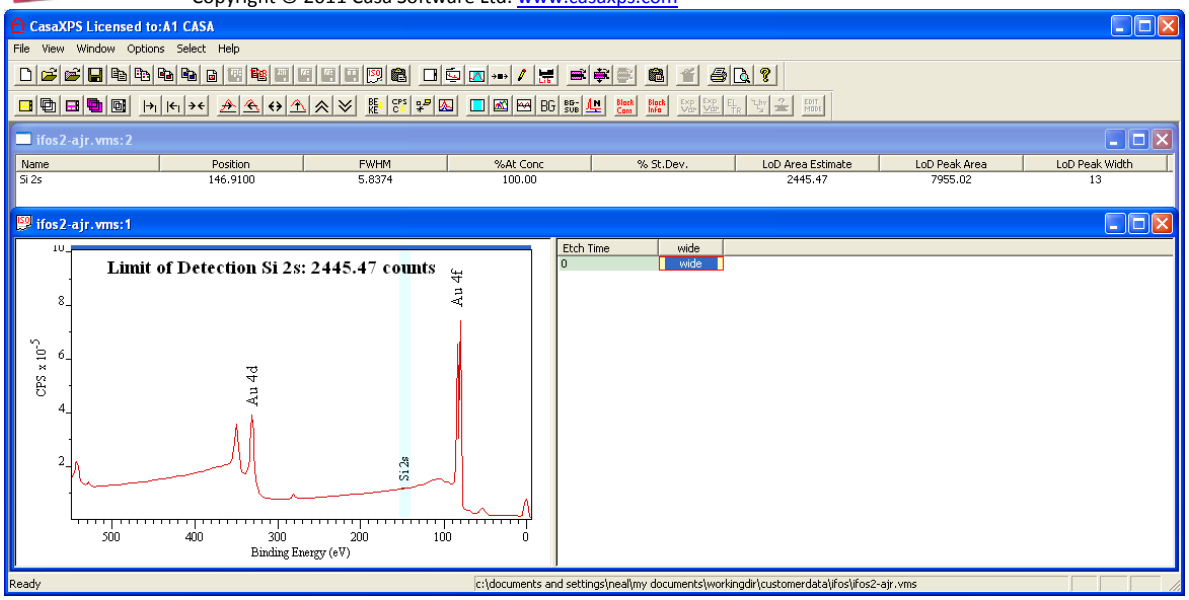
1. Assuming a peak can be measured within a quantification region, the peak characteristics in terms of background intensity and the energy interval over which a peak is spread are calculated from the region and used to estimate the limit of detection.
2. No peak is evident in a region, the LOD or “Limit of Detection” background type is used to supplement the quantification region and allows the user to specify the interval over which the peak counts are spread. The LOD background type does not calculate a background but is equivalent to the SKIP background type and therefore must be defined within an existing region already defined over the same energy interval.

Both methods for calculating the limit of detection populate reporting information such that the Region standard report on the Report Spec property page can be configured to include:

1. The estimated limit of detection for a peak (LOD_ESTIMATE)
2. The calculated peak intensity in terms of counts integrated over the width established for the peak (LOD_AREA)
3. The peak width in terms of data bins used to calculate the limit of detection (LOD_PEAK_WIDTH)
4. The background intensity used in the calculation of limit of detection estimate (LOD_BG_VALUE).

These keywords can be entered into an appropriate configuration file for the Standard Report





When assessing if a peak is detectable, the LOD_ESTIMATE should be compared against the LOAD_AREA. If the LOD_AREA is greater than a chosen multiple (see below) of the LOD_ESTIMATE, then a peak can be detected.

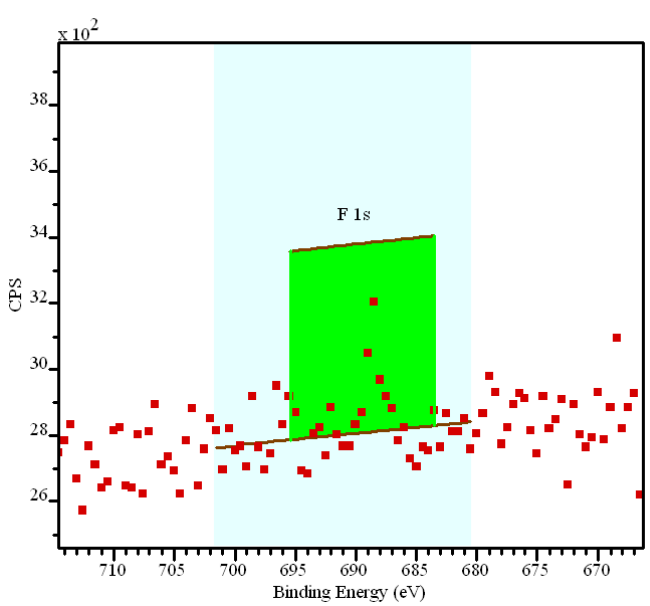
Limit of Detection Estimate

Method adopted for measuring the limit of detection is based on a method developed by Alex Shard (National Physical Laboratory).

Pulse counted intensities measured over a time interval are assumed to obey Poisson statistics and therefore if I counts are recorded for a given electron energy interval and time interval, the standard deviation for a sequence of identical measurements in theory is $\sqrt{I}$.

For small peaks the intensity of the peak and the intensity of the background are approximately the same and therefore for small peaks one standard deviation in the measured intensities, for a background intensity B, can be approximated by $\sqrt{B}$.

Since a spectra may be measured using a variety of energy step-sizes as well as dwell-times, rather than using the intensity per acquisition bin, the intensity for the signal is defined as the intensity falling within a defined quantification region and above the defined background.



For an energy interval and acquisition time corresponding to the quantification region specified with background type Limit of Detection (abbreviation LOD) one standard deviation is given by $\sqrt{BN}$, where N is the number of data channels falling within the region limits. The corresponding intensity against which this standard deviation can be compared is the sum of the positive background subtracted counts falling within the region limits LOD_{AREA}.

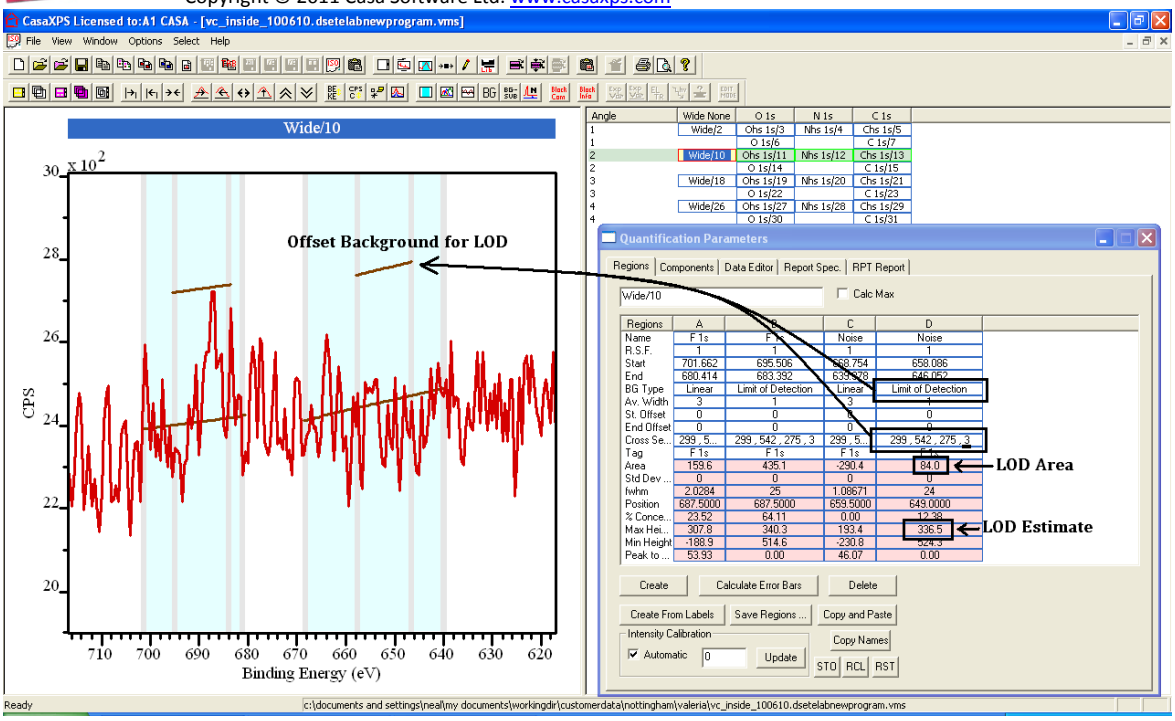
Assuming Poisson behaviour, the test for a peak is of the form

$$if\ (LOD_{AREA} > factor * \sqrt{BN})\ then\ "peak\ detected"$$

For a 99% confidence interval and Poisson statistics $factor = 3.0$.

A complication is introduced by modern instruments with multiple detector systems. Pulse counted data measured using a single channeltron detector does obey Poisson statistics to a high degree, however multiple data streams merged to form a spectrum results in smoothing of the data as recorded. The factor used in the test for a peak varies depending on the detector system and must be determined appropriately for a spectrum and the application.

To assist the selection of an appropriate factor, the Limit of Detection background type uses the 4th parameter (T_0) in the cross-section field to draw an offset curve to the background located $T_0\sqrt{B}$ above the background. The following data illustrates a small F 1s peak and a region defined on an energy interval without a peak.



The above data was collected with a multiple detector instrument and as a result using a factor of 3 in the test is likely to be more insensitive to peaks than might be expected. Adjusting the value of the 4th cross-section parameter causes the LOD Estimate to be scaled by the factor as specified by the cross-section parameter and so the LOD test can be made more sensitive to changes in the data around a peak.

Note: the LOD estimate depends on the choice of width for the region. A narrow peak is easier to distinguish from the background at low intensities compared to the same number of counts spread out over many acquisition bins due to a broad peak structure. Selecting the correct width for the LOD region is therefore an important decision when establishing the LOD test.

Limit of Detection Calculated from Quantification Regions

While the "Limit of Detection" background type offers more control over the calculated LOD estimate value, all region background types calculate an estimate for the LOD. For all regions with background types other than "Limit of Detection" the LOD estimate is one standard deviation and can only be seen if configured using the Standard Report. The application of a factor as shown in the LOD test above must be introduced using a spreadsheet program based on the exported Standard Report.

The Poisson distribution and Pulse Counted Data

For pulse counted data it can be assumed there exists a count rate ν such that

1) The probability of a single counting event occurring in a small time interval of length δt is approximately equal to $\nu\delta t$.
2) The probability of more than one counting event occurring in a small time interval δt is negligible when compared to a single counting event occurring in the same time interval.
3) The numbers of counting events in non-overlapping time intervals are independent.

Given these assumptions it can be shown that the number of counting events occurring in a period of time t has a Poisson distribution with parameter $\lambda = \nu t$. If the random variable $X(t)$ denotes the number of counting events in the time interval t then $P[X(t) = r] = \frac{e^{-\nu t}(\nu t)^r}{r!}$ for $r = 0,1,2, \ldots$

Given that $X$is a Poisson distributed random variable; the expected value and variance for $\lambda = \nu t$ are as follows

$$\mathcal{E}[X] = \lambda$$

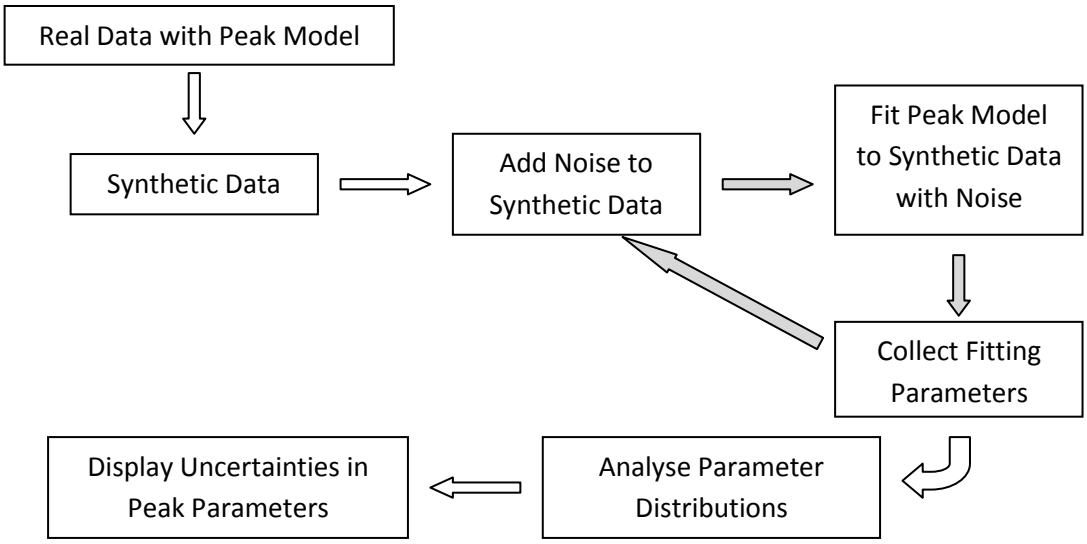
and

$$var[X] = \lambda$$

Since for pulse counted XPS data λ corresponds to the counts per bin, the standard deviation in the counts per bin is $\sigma = \sqrt{var[X]} = \sqrt{\lambda}$.

Provided an XPS spectrum can be expressed as counts per bin, assuming Poisson behaviour for the noise in the data allows error estimates for peak fitting parameters to be calculated using a Monte Carlo approach.

Monte Carlo Simulation



A Monte Carlo procedure involves a simple sequence of steps aimed at estimating the precision error in output quantities from a calculation. By synthesising the problem before adding noise back to the synthetic problem followed by repeating the calculation, an understanding of how noise perturbs the current set of peak parameters is established. After iterating through these steps collecting the output parameters from each iteration, distributions for the output parameters are gather where the variation in the output values are due to the influence of noise on the calculation in question. For the problem of fitting peaks to data, the calculation is that of optimising a set of peak parameters so as to reproduce the data envelope in a least squares sense.

Random Variables, Expectation and Variance in XPS

A random variable in statistics is a function defined in terms of events resulting in a numerical value. For the purposes of XPS quantification, the numerical values might be the intensity of a peak as

measured by a quantification region, and while the values for the, so called, random variable are specific for a given set of conditions, the random element of the measurement process leads to slightly different experimental conditions and therefore many possible values for the intensity for a peak. That is, by repeating an experiment and recalculating the peak intensity different peak intensities are obtained and these variations are due to random noise in the measurement process. The analysis must therefore choose a value that best represents the peak intensity, which more often than not is the value obtained from a single measurement. However, if a measurement is performed multiple times, using the assumption the results obtained for each experiment is equally likely, the value for the peak intensity is typically obtained by calculating the mean average for the set of measurements. The mean average provides as estimate of the expected value for the random variable corresponding to the counts per seconds for electrons allocated to a peak.

The term expected value for a random variable X with discrete values corresponding to a set of peak intensity measurements $\{A_1, A_2, A_3, \cdots, A_n\}$ with probability $P[X = A_i]$ is given by

$$\mu_X = \mathcal{E}[X] = \sum_{i=1}^{n} A_i P[X = A_i]$$

The uncertainty for such a measurement is obtained from the square root of the variance for the random variable as follows.

$$var[X] = \sum_{i=1}^{n} (A_i - \mu_X)^2 P[X = A_i]$$

The standard deviation for the random variable from the expected value is

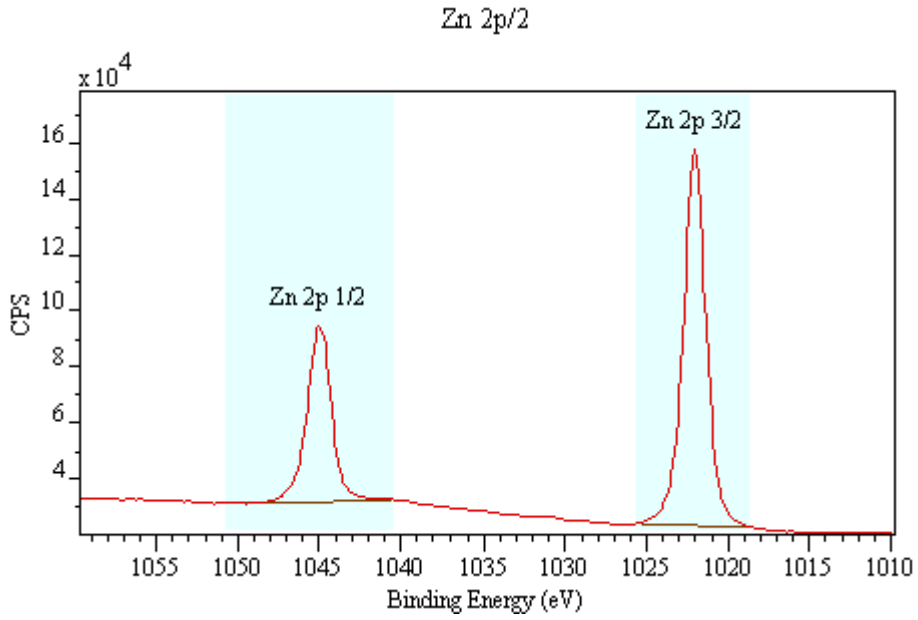
$$\sigma_X = +\sqrt{var[X]}$$

and is a measure for the spread from the expected value in units of the expected value.

When a single measurement is used to estimate the intensity for a peak, the expected value is approximated by the peak intensity integrated from the data envelope. An estimate for the uncertainty is obtained by applying Monte Carlo methods to the intensity calculation. Each peak used in XPS quantification represents a distinct random variable for which expectation and variance are calculated. While in general random variables may potentially be dependent, random variables measured from XPS quantification regions are assumed to be independent.

Intensity and Uncertainty for Multiple Peaks from a Single Transition

The intensity for a transition may be split between a pair of doublet peaks, such as is the case for Zn 2p, where a single integration region provides a poor definition of the background and therefore the intensity of the transition is measured using two regions defined independently for these widely separated Zn $2p_{1/2}$ and Zn $2p_{3/2}$ peaks.



Provided the peak area is scaled using the combined RSF appropriate for adjusting the integrated area from both peaks in the doublet, the total intensity is obtained by summing the intensity from the two quantification regions. This statement is equivalent to the statement that the expected value for the sum of two random variables $X_1$and X_2, $\mathcal{E}[X_1 + X_2]$ is the sum of the individual expected values

$$\mathcal{E}[X_1 + X_2] = \mathcal{E}[X_1] + \mathcal{E}[X_2]$$

The random variables $X_1$and X_2 represent the peaks intensities for the Zn $2p_{1/2}$ and Zn $2p_{3/2}$ peaks.

When the uncertainty for these two peaks is calculated based on two quantification regions, the uncertainty for the summed peak intensity is obtained by considering the variance for each random variable and the covariance between the random variables. The variance is the square of the standard deviation therefore the standard deviation for the summed peak intensity is obtained using

$$var[X_1 + X_2] = var[X_1] + var[X_2] + 2cov[X_1, X_2]$$

For independent random variables corresponding to region areas corrected for transmission, escape depth and relative sensitivity, $cov[X_1, X_2] = 0$ therefore

$$var[X_1 + X_2] = var[X_1] + var[X_2]$$

$$\Longrightarrow \sigma_{X_1+X_2} = \sqrt{{\sigma_{X_1}}^2 + {\sigma_{X_2}}^2}$$

	Zn $2p_{3/2}$	Zn $2p_{1/2}$	Zn 2p
Corrected Area	37932.6	18051.9	55984.5
St.Dev.	76.1778	120.328	142.4145

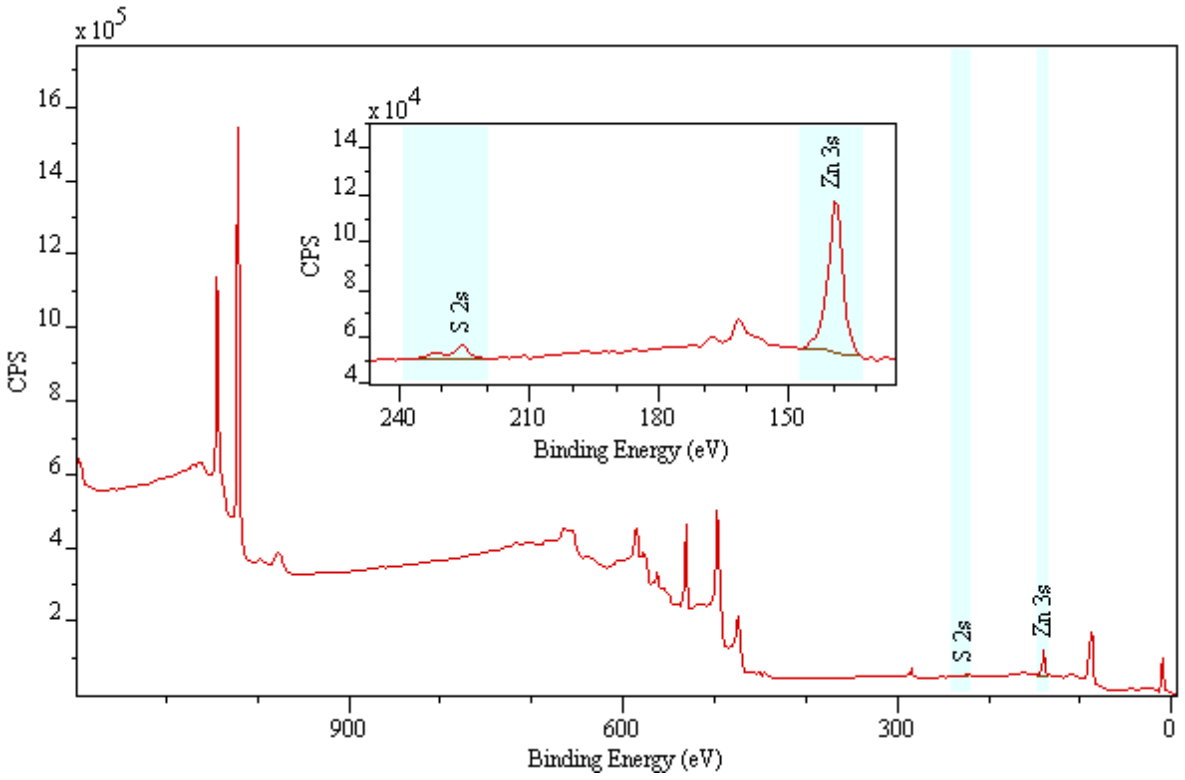
$$\mathcal{E}[X_1 + X_2] = 37932.6 + 18051.9 = 55984.5$$

$$\sigma_{X_1+X_2} = \sqrt{{\sigma_{X_1}}^2 + {\sigma_{X_2}}^2} = \sqrt{76.1778^2 + 120.328^2} = 142.4145$$

Note: intensities measured from peak fitting typically are not independent, therefore $cov[X_1, X_2] \neq 0$ for overlapping peaks.

Calculating the Ratio of Peak Intensities and Uncertainties

Consider, by way of example, the ratio of sulphur to zinc measured from the following survey spectrum. The regions selected for the ratio are not the most intense peaks from either sulphur or zinc, however the peaks are close in energy and therefore both the transmission and sampling depth for these transitions are well matched. The low intensity for these peaks is also useful to illustrate the uncertainty associated with calculating the ratio for these elements from XPS data.



In general, the variance for a quotient of two random variables X and Y with expected values μ_X and μ_Y, and variances $var[X]$ and $var[Y]$ is approximated by

$$var\left[\frac{X}{Y}\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{var[X]}{{\mu_X}^2} + \frac{var[Y]}{{\mu_Y}^2} - \frac{2cov[X,Y]}{\mu_X \mu_Y}\right)$$

Since the random variables corresponding to the peak intensities measured from distinct energy intervals are independent measurements, $cov[X,Y] = 0$ therefore for peak intensities measured using regions

$$var\left[\frac{X}{Y}\right] \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{var[X]}{{\mu_X}^2} + \frac{var[Y]}{{\mu_Y}^2}\right)$$

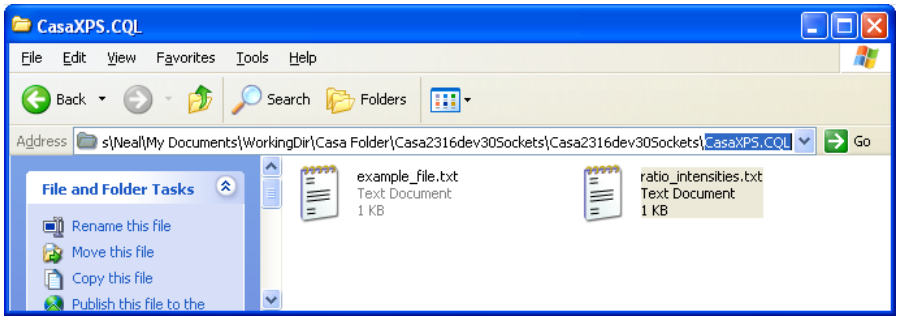
The uncertainty in the ratio of two peak intensities measured from regions is obtained from

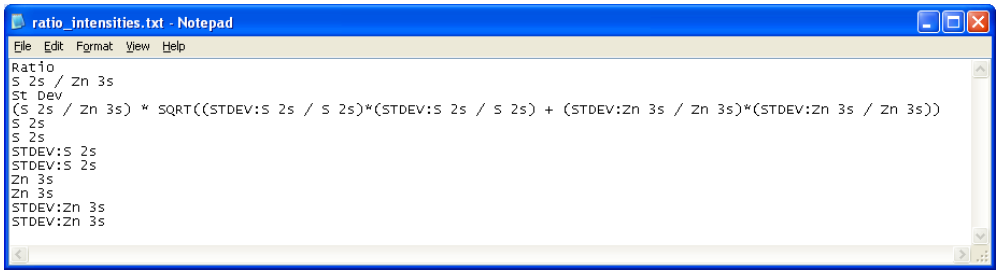
$$\sigma_{\frac{A_1}{A_2}} \approx \left(\frac{A_1}{A_2}\right)\sqrt{\left(\frac{{\sigma_{A_1}}^2}{{A_1}^2} + \frac{{\sigma_{A_2}}^2}{{A_2}^2}\right)}$$

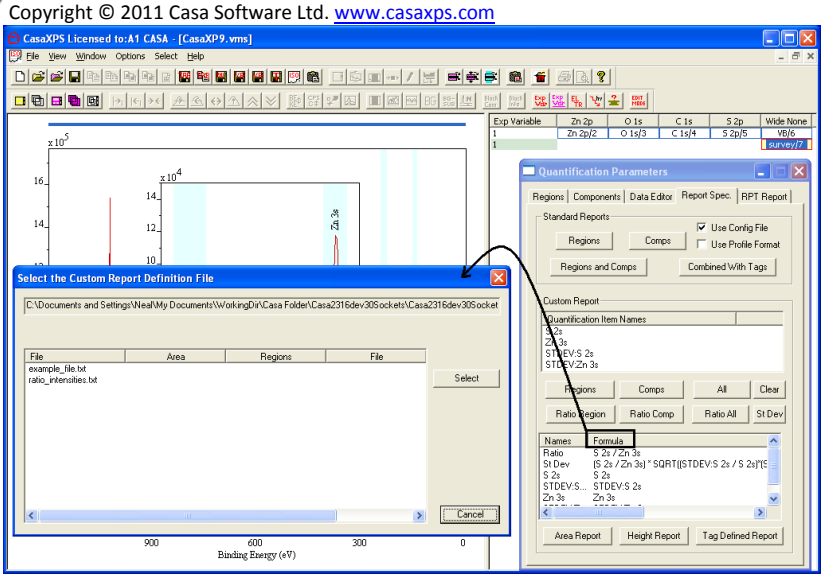
where A_i and σ_{A_i} are RSF, transmission and escape depth corrected intensity and uncertainty quantities. In CasaXPS these are calculated using a Monte Carlo procedure assuming the counts per bin obey Poisson behaviour.

Thus, using the Custom Report on the Report Spec property page of the Quantification Parameters dialog window, the relevant information can be gathered as follows.

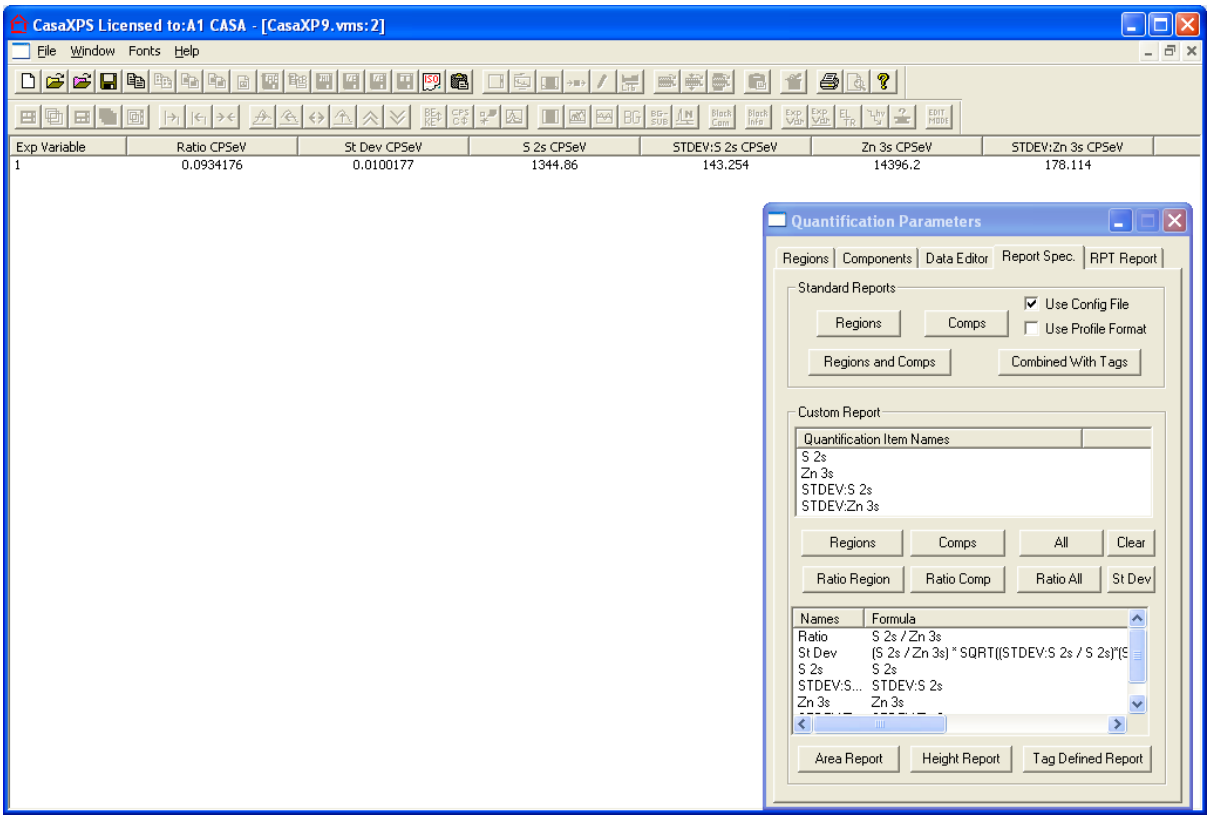
1. Select the VAMAS block containing the survey spectrum in the right-hand pane of the experiment frame.
2. On the Regions property page, press the Calculate Error bars button. This will ensure the St Dev for the peak areas are available to the Custom Report.
3. Use the StDev button on the Custom Report to create entries suitable for calculating the ratio and the uncertainty in the ratio using a spreadsheet program. Alternatively create a Custom Report configuration file in the CasaXPS/CasaXPS.CQL directory already to generate the ratio and the uncertainty for the ratio. Load the file using the Formula column header button.







4. Press the Area Report button on the Custom Report to generate the ratio together with the uncertainty in the ratio.



Uncertainty in Percentage Area

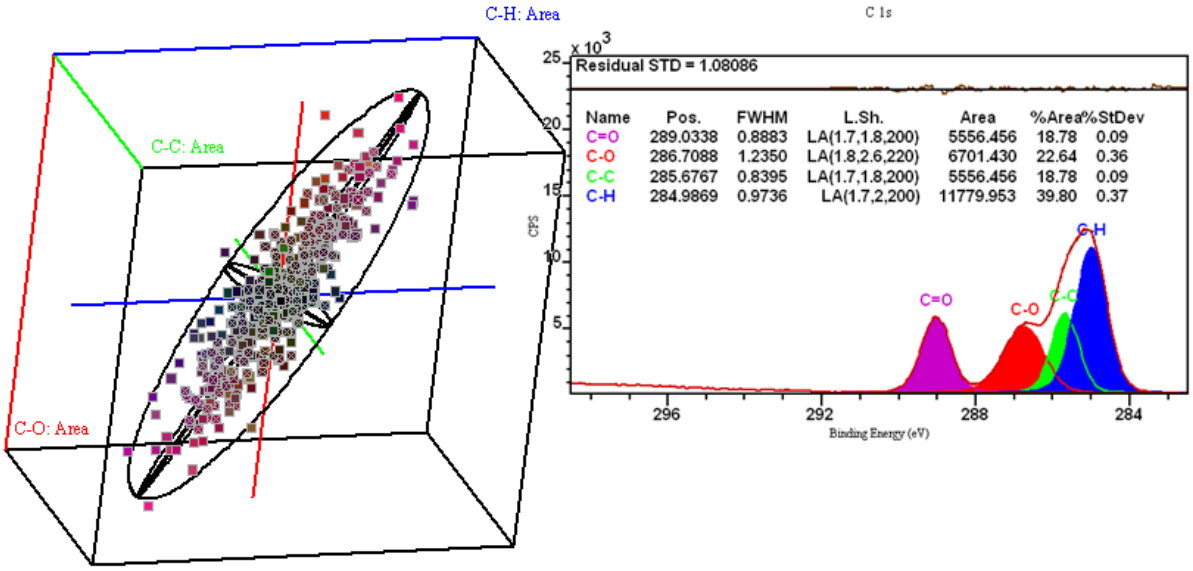
The problem of determining the uncertainty in atomic concentration measured using quantification regions provides an insight into the problems involved when area measurements are made using overlapping peaks. Atomic concentration for an element measured from regions is obtained by a ratio of two quantities, namely, the intensity of the element and the total intensity summed over all regions corresponding to each element used in the measurement. The concentration measured from region i out of a set of n quantification regions is given by

$$B_i = 100 \times \frac{A_i}{\sum_{j=1}^{n} A_j}$$

Since the measurement for each peak intensity A_i is influenced by noise, the uncertainty in B_i must be determined by considering two random variables X_i and Y corresponding to the possible outcomes for the values A_i and $\sum_{j=1}^{n} A_j$, respectively. While the intensities measured from quantification regions are assumed to be independent, the random variables as stated X_i and Y are not independent since $\sum_{j=1}^{n} A_j$ includes A_i. To correctly combine error estimates for each of the A_i in order to obtain an error estimate for B_i requires a determination based on the random variable $Z_i = \left(\sum_{j=1}^{n} A_j\right) - A_i$, where the functional form for the random variable W_i corresponding to the atomic concentration measurements B_i is of the form

$$W_i = \frac{X_i}{X_i + Z_i}$$

When expressed in this form, W_i is defined in terms of two independent random variables X_i and Z_i. Since for independent random variables X_i and $Z_i . cov[X_i, Z_i] = 0$, it is therefore possible to express the standard deviation of W_i in terms of the variance of X_i and Z_i only. This is in contrast to the random variables X_i and Y as, in general, $cov[X_i, Y] \neq 0$ and so the $var\left[\frac{X_i}{Y}\right]$ cannot be expresses in terms of $var[X_i]$ and $var[Y]$ alone.



The analogous problem, involving intensities measured using overlapping synthetic components optimised to fit a data envelope, the percent area calculation yields numerous random variables with varying degrees of dependency with one another. For data in which multiple overlapping peaks are used to measured intensity that cannot be reduced to a set of independent random variables, determining the uncertainty in the percentage area from the uncertainties in the peak areas is far from trivial. Fortunately this problem is similar in nature to the initial problem of determining the uncertainties in the peak fitting parameters. Monte Carlo simulation can be similarly applied to the percentage area calculation; these Monte Carlo results are the uncertainties available for display via

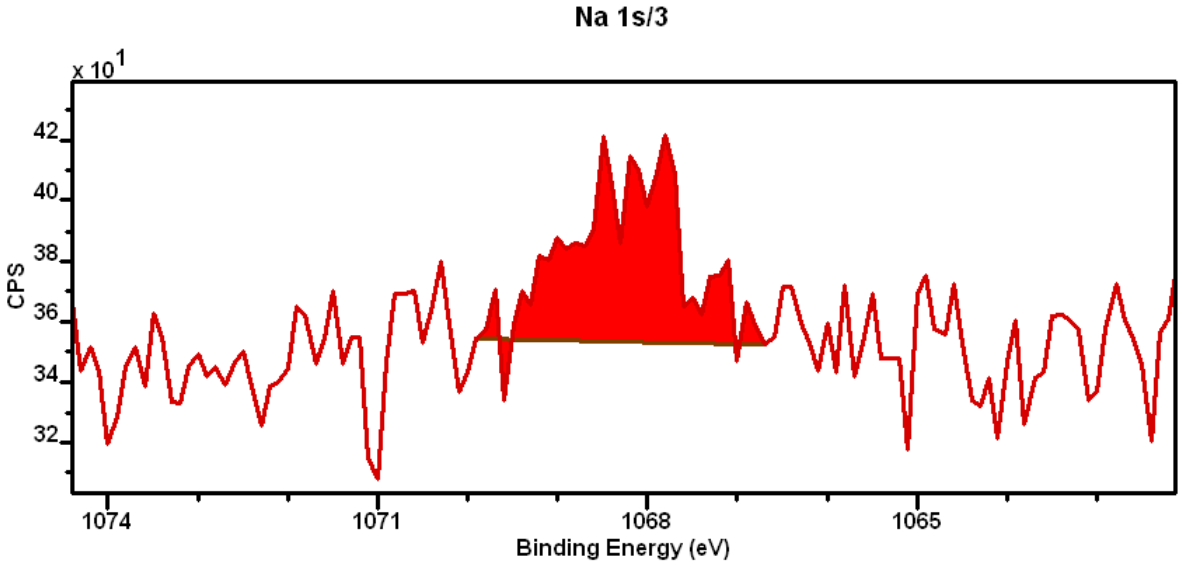
the Components annotation window and inclusion in quantification reports on the Report Spec property page of the Quantification Parameters dialog windows.

Uncertainties are typically reported in terms of confidence intervals. For normally distributed measurements about a mean, 68.3% of the values lie within one standard deviation of the mean. Other than adopting the 68.3% criterion the availability of value distributions from a Monte Carlo simulation means standard formulae for variance and covariance are not required to establish the confidence interval for a given value. It is sufficient to collect the set of parameter values from these Monte Carlos distributions which lie within an error ellipsoid, the dimensions of which ensure 68.3% of the points belong to this set. The dimensions of the ellipsoid provides the relationship between the different distributions in terms of correlation, and the projection of the extreme value for each parameter within the ellipsoid defines the ranges for which the assertion that the expected values are precise with a probability of 0.683 is supported.

Quantification Regions and Statistical Independence

Noise and Quantification Regions

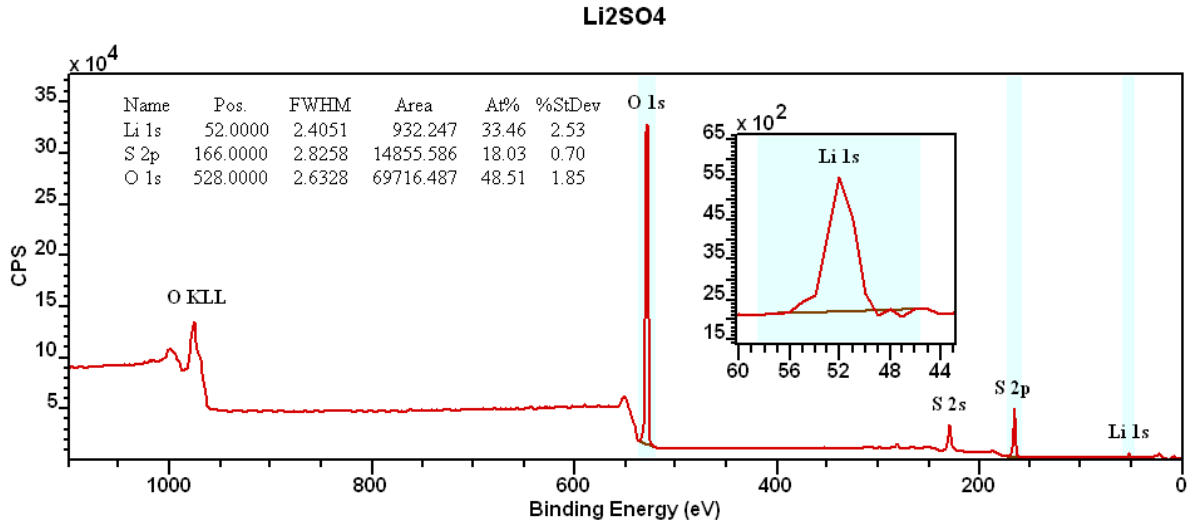
Peak intensity measured using quantification regions involve integrating the background-subtracted counts per second with respect to energy yielding an area which is corrected for transition probability and instrumental intensity variations. These corrected area values are used to calculate the relative intensity for an element in a sample. Intensity adjustment is achieved using a scaling factor computed from the relative sensitivity of the transition, the transmission function for the instrument response, and a correction for escape depth dependency on kinetic energy of the recorded electrons.



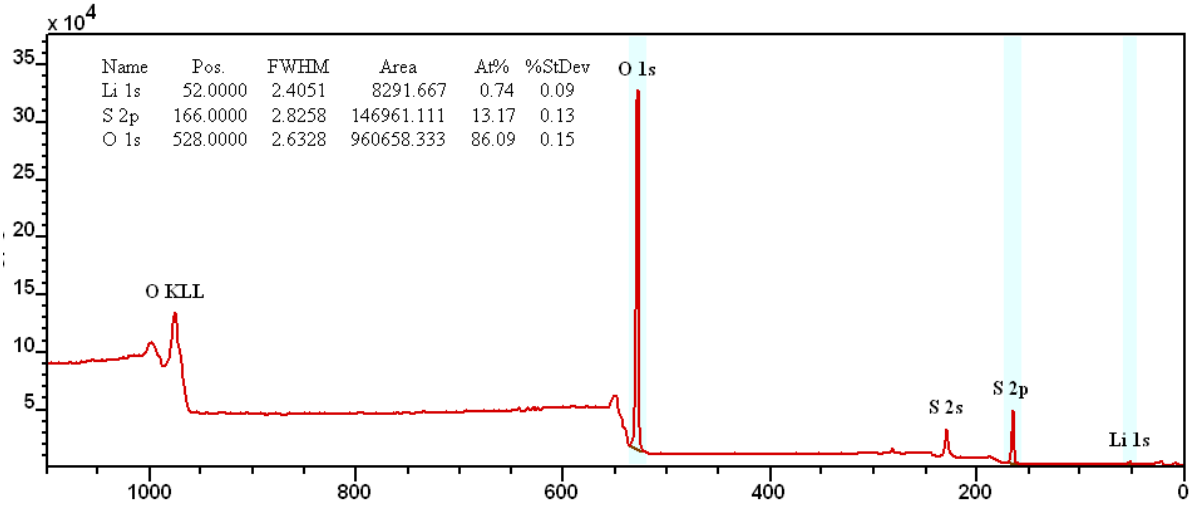
The reported intensity of a peak is an estimate, and as an estimate requires additional information to convey the level of confidence which can be placed in the intensity. Noise associated with the measured data influences the value computed for the peak intensity and is the source for precision errors in peak intensity. However, predominantly noise feeds into the measured intensity through the background calculation particularly where a background depends heavily on the intensity of data at the limits of the energy interval over which the background is defined.

The level of confidence is not simply a characteristic of the signal-to-noise in the data, but is also influenced by the statistic calculated from the data. The following two statistics measured from the same data illustrates this point.

By way of example, consider a Li_2SO_4 sample where the lithium peak, representing a significant proportion of the surface material, is nevertheless small compared to the S and O peaks. The first statistic measured from the data is the atomic concentration. Atomic concentration must be calculated from scaled peak areas. The act of scaling the peak areas also scales the effect of noise and the resulting percentage concentrations with 68.3% confidence intervals are displayed in the table overlaying the spectrum below.



A similarly calculated statistic, computed without any intensity scaling, offers a different table. By omitting intensity scaling the uncertainty in the value computed for oxygen has dramatically improved compared to the uncertainty for the atomic concentration calculation above.



These two calculations provide an example of how an error associated with the measurement of a small peak, in this case Li, is transformed into a large error in what would otherwise appear to be a precise measurement, namely O 1s, as a consequence of simply applying relative sensitivity

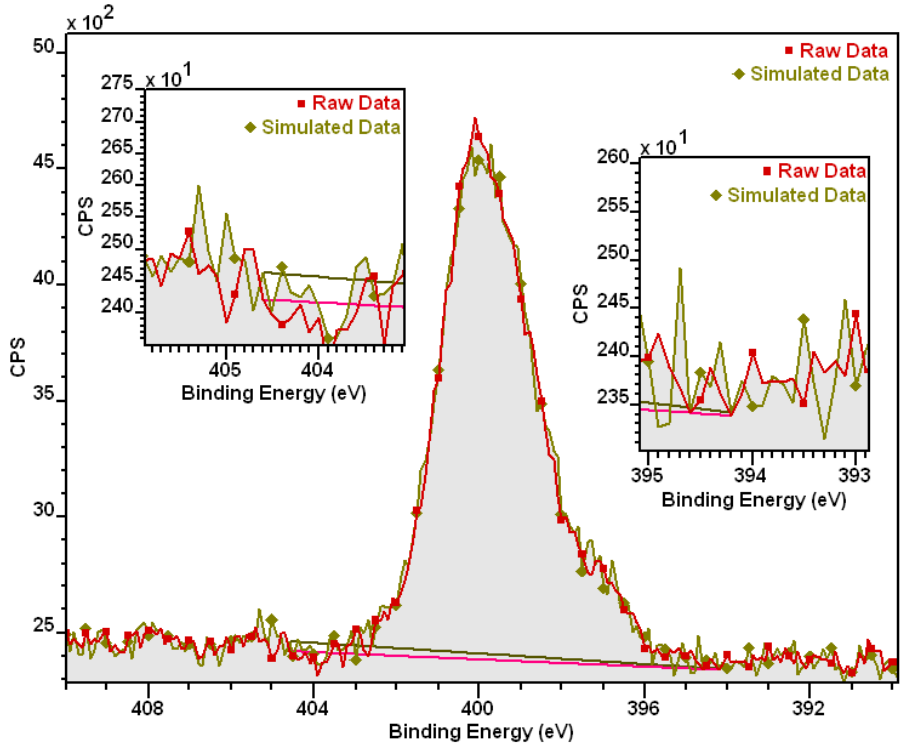
correction to the peak intensities. The lesson to be learnt from these calculations is it is worth investing the time to improve the signal-to-noise in the lithium peak, which will be the limiting factor in determining the atomic concentration for the sample in question.

Ideally, data should be acquired to achieve a consistent signal-to-noise for each peak used to assess the composition of a sample. A typical measurement aimed at quantifying the surface composition is performed using a survey spectrum where a fixed step-size and acquisition time is used to measure all peaks within a wide energy range. The signal-to-noise ratio varies depending on factors such as the background and the relative sensitivity of the peaks used in the quantification, so peak area measurements from a single survey spectrum tend to vary in precision. Narrow scan spectra with differing dwell-time are a means of focusing acquisition time on peaks such as the Li 1s peak for which an improved quantification would result from increasing the dwell-time compared to the O 1s data.

Calculating Uncertainty for Peak Intensity measured from Regions

The method used to calculate uncertainties for peak intensity measured using a quantification region assumes noise in the spectral data obeys a Poisson distribution. A random number generator is used to add noise to the data consistent with a noise distribution having a mean and variance equal to the counts per bin. A data envelope is synthesised from the raw spectrum to which simulated noise is added before calculating the peak intensity from the synthetic data. The intention is to simulate an identical experiment to the one yielding the original spectrum.

To illustrate the consequences of simulating a fresh measurement an example using a linear background type highlights how changes to the calculated background occur and therefore alter peak intensity measured from differing but essentially identical data. Note how the intensity at the end points used to define the background change as a result of noise.

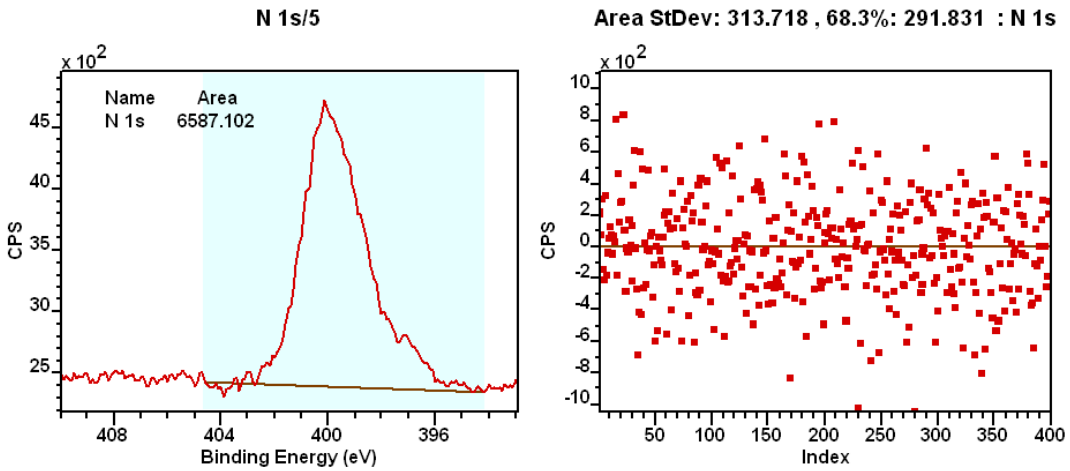


Repeating the simulation step and collecting the set of peak intensities calculated from these data provides a distribution of possible values from which a sample variance can be calculated.

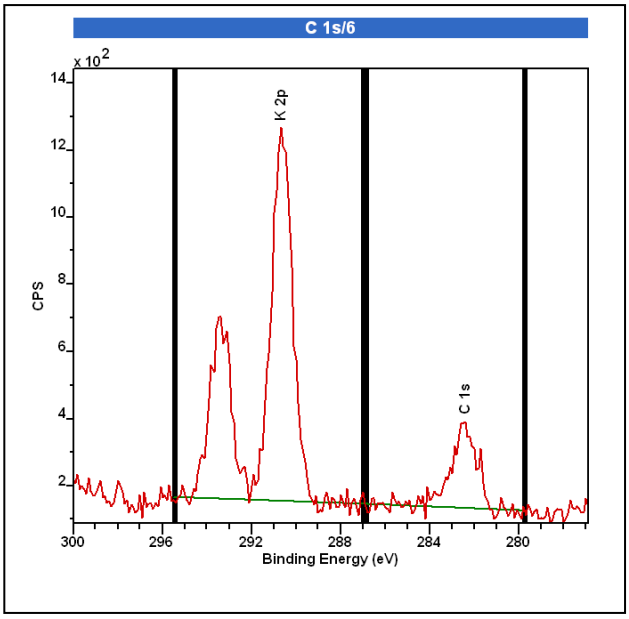
Two methods can be used to measure the standard deviation from the simulated distribution containing n simulations:

1. The standard deviation is calculated using the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \bar{A})^2$
2. Alternatively the interval is computed to ensure number of distribution points closest to the mean for the distribution, $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i$, includes 68.3% of the distribution points.

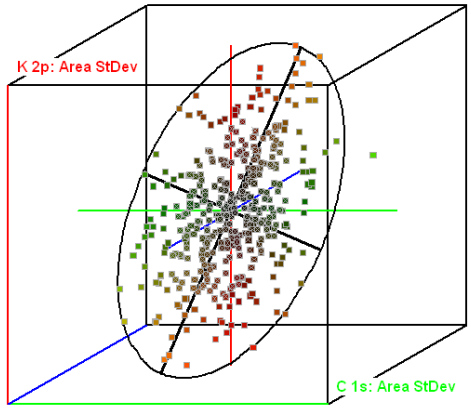
The error estimates based on both methods are reported above the distribution plotted below for the variation in peak area of a nitrogen 1s spectrum.



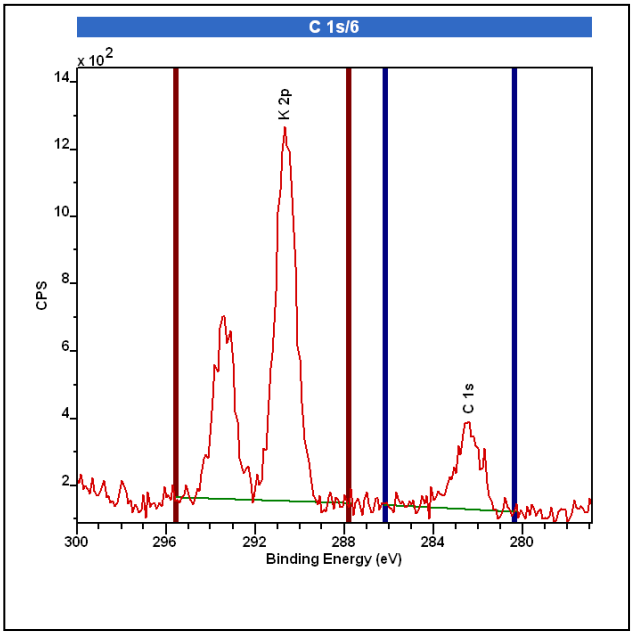
Estimating the uncertainty for a peak based on a region assumes the peak intensity is independent of any other measurements similarly made using quantification regions. While the assumption of independence is generally acceptable for peak intensity measured from regions, there are circumstances in which the assumption is not strictly true.



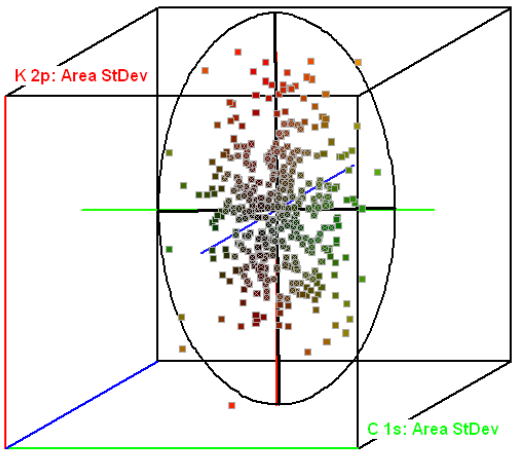
Correlated peak intensities due to the juxtaposition of quantification regions.



If two quantification regions are defined for K 2p and C 1s transitions, firstly locating the regions one directly next to the other, then secondly adjusting the regions to ensure a significant separation between the two regions, plotting the deviation from the mean peak intensities for these two cases illustrates potential correlation caused by quantification regions. The reason a dependency between the two regions is introduced in the former case, where the regions connect, is the common end and start limit for the K 2p and C 1s regions allows the same noise to influence both backgrounds beneath the K 2p and C 1s data. The dependency is due to the common data channels used in the calculation. Preventing the regions from using the same data to calculate the backgrounds removes the dependency between the two peak intensities.



Uncorrelated peak intensities due to the distinct quantification regions.



If peak intensities become correlated, the uncertainty in derived statistics which assume independence will potentially under estimate the uncertainty in these derived statistics. For example, if the statistic desired is the sum of the K 2p and C 1s peak intensities, for the correlated definition for the quantification regions the uncertainty in the summed intensity would require knowledge of the variance for both peaks and the covariance calculated from both distributions. The following table represent an error matrix calculated from the two distributions generated from the peak intensities for the K 2p and C 1s quantification regions.

Error Matrix

	1:Area:K 2p	2:Area:C 1s
1:Area:K 2p	4816.71	1393.72
2:Area:C 1s	1393.72	3005.49

To estimate the uncertainty in the statistic $Area{:}\,K\ 2p\ +\ Area{:}\,C\ 1s$ the formula

$$var[X_1 + X_2] = var[X_1] + var[X_2] + 2cov[X_1, X_2]$$

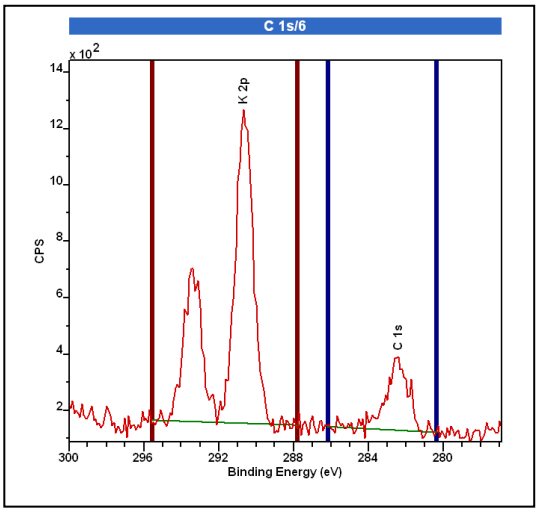
must be used where random variables $X_1$and X_2 represent the peak intensity values. The uncertainty in the summed peaks is therefore

$$\sigma_{X_1+X_2} = \sqrt{{\sigma_{X_1}}^2 + {\sigma_{X_2}}^2 + 2cov[X_1, X_2]}$$

$$\Longrightarrow \sigma_{X_1+X_2} = \sqrt{4816.71 + 3005.49 + (2)(1393.72)} = 103$$

An alternative method for calculating the same uncertainty is to introduce a third region spanning both the K 2p and C 1s peaks with background type ***SKIP***. A background type ***SKIP*** causes the quantification region to use the background calculated by other regions defined on the data. The peak area calculated from the third region is simply the sum of the two regions defined for K 2p and C 1s data. Calculating the distribution generated by a Monte Carlo simulation for the region measuring the sum of the K 2p and C 1s peaks yields an uncertainly of 102.98. If the covariance between the K 2p and C 1s distributions is assumed to be zero, that is, these distributions are assumed to be independent the uncertainty calculated by summing the variances alone would be 88.4, a value which underestimates the uncertainty in the summed intensity.

Provided quantification regions are defined to be independent, the uncertainty can correctly be calculated by summing the variances. Performing the equivalent calculation as above but for regions defined with distinct start and end limits, the outcome is as follows.



Error Matrix

	1:Area:K 2p	2:Area:C 1s
1:Area: K 2p	4036.95	-85.4642
2:Area: C 1s	-85.4642	1863.94

$$\sigma_{X_1+X_2} = \sqrt{{\sigma_{X_1}}^2 + {\sigma_{X_2}}^2}$$

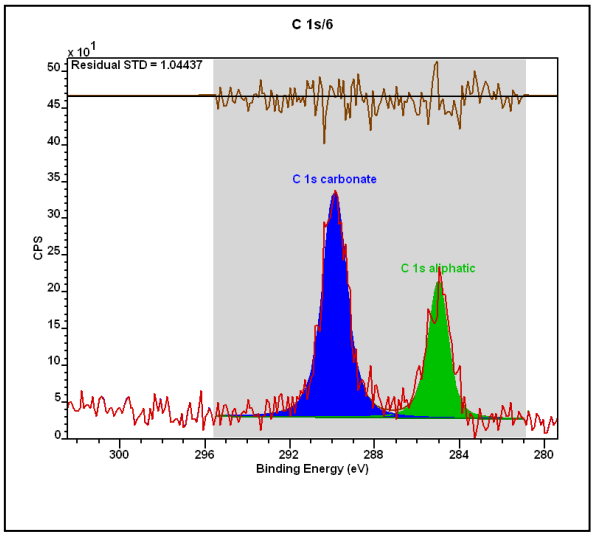
$$\Longrightarrow \sigma_{X_1+X_2} = \sqrt{4036.95 + 1863.94} = 76.8$$

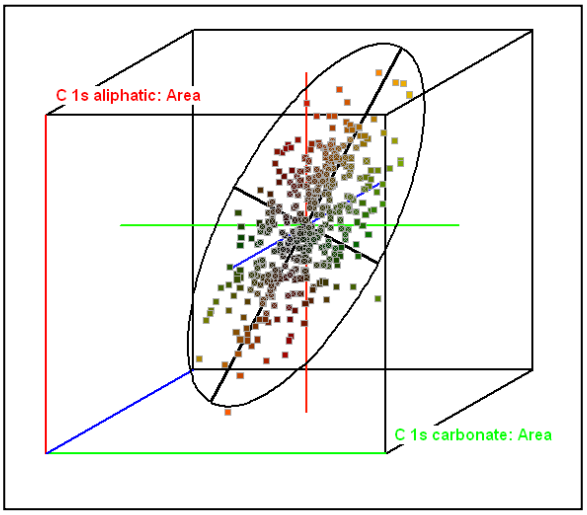
Again using a third region to span both K 2p and C 1s peaks from which an estimate for the uncertainty in the sum of the two regions can be made yields a value for the uncertainty of 75.7. Thus provided an element of care is used to ensure quantification regions are indeed independent of one another, uncertainties in derived statistics such as ratios and atomic concentrations can be calculated using only the variance obtained from distributions calculated by Monte Carlo simulation.

The example presented here where quantification regions are defined to be dependent is of greater relevance to uncertainties in statistics derived from peak intensities measured using synthetic component peaks in a peak model. If a common background is used beneath a set of apparently independent peaks, including only the variance for the uncertainty in the peak parameters will potentially underestimate the true uncertainty in derived statistics. For this reason it is necessary to perform the analysis for peak parameter uncertainties in the context of joint distributions rather than each distribution in isolation.

Uncertainties Calculated from Peak Model Parameters

A simple problem of identifying peak intensity for a pair of C 1s transitions separated sufficiently to expect a degree of independence for intensities illustrates the danger of assuming independent for these types of measurements when considering uncertainties in atomic concentration calculations.





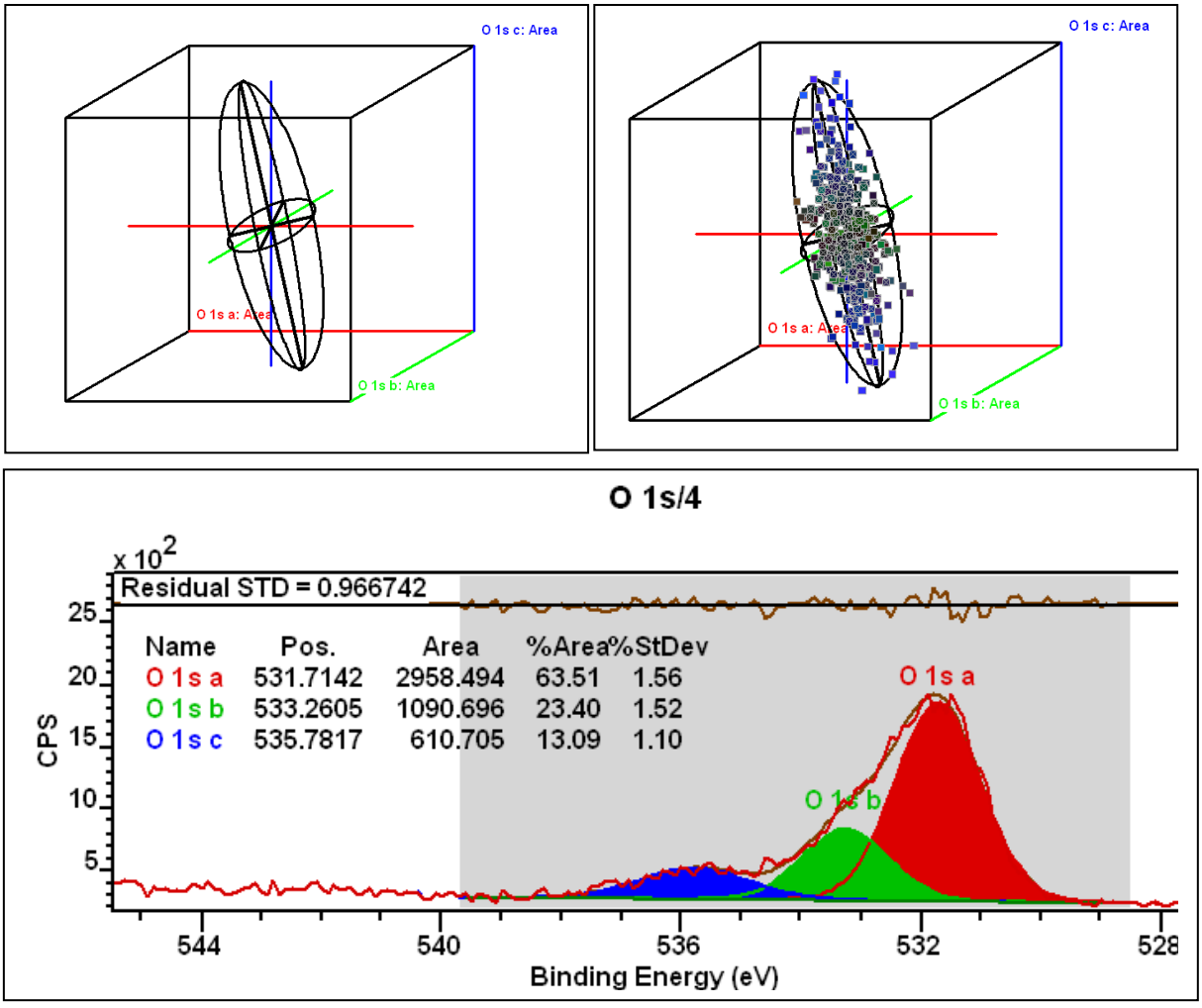
A scatter plot for the peak intensity measured using these two synthetic components reveals a correlation between these two intensity values not dissimilar to the dependent regions above. The background calculation which spans both peaks is the source for the connection between these two peak intensities. Any statistic calculated from these two intensity values would require a non-trivial uncertainty calculation before a confidence interval could be established for the statistic.

If two error distributions exhibit independence, the probability for a quantity lying within one standard deviation of the mean is presumed to be 0.683. The measured quantity and corresponding uncertainty cares little for other independent measurements. When two or more quantities change depending on the values measured within the set of quantities of interest the problem changes as follows. The set of acceptable parameter sets changes from simply deviation from the mean in each distribution to considering the deviation from a centre of mass for the joint distributions. For a point within the joint distribution to be accepted as a valid set of parameters, all parameters representing the point must lie within a volume of space containing 68.3 % of the joint distribution. Since this criterion for an acceptable parameter set is more exacting than the one used for independent parameters, some values which would have been accepted as within the 68.3 % limit for a parameter taken in isolation will be rejected by the collective perspective requiring all parameters are within the defined limit, and therefore uncertainties in the individual parameters increase when dependences are involved.

When dealing with multiple parameter distributions the term “lie within a volume of space” is somewhat imprecise. Precision is given to this statement by choosing the region of space bounded by an ellipsoid taking on the proportions defined by the variation in the distribution of parameter sets about the centre of mass. These dimensions for the ellipsoid are obtained from the principal axes calculated for the distributions. In using this criterion to partition the set of parameter sets into

those lying within a volume containing 68.3 % of the parameter sets and those outside the volume as prescribed, a definition for the confidence region is made specific.

An example based upon an O 1s data envelope fitted with three component peaks provides an illustration for the concept of an ellipsoid region of interest. The parameters calculated from the three component peaks when fitted to the O 1s data are analysed via Monte Carlo methods to produce the 3D distribution for the errors in the area parameters. The concept of an error ellipsoid is used in CasaXPS to provide the measure of how close a particular set of fitting parameters are to the distribution centroid. Parameter sets closest to the centroid, in an elliptical sense, such that 68.3 % of the parameter sets are bounded by a common ellipsoid are marked with a cross.



Once a list of parameter sets are ordered according to proximity to the centre of mass, an uncertainty for each parameter is obtained by determining the extreme points from the list containing 68.3% of the distribution points. The method for determining the range of possible acceptable outcomes is therefore analogous to method 2) above for calculating the standard deviation from a single Monte Carlo distribution obtained for a quantification region.

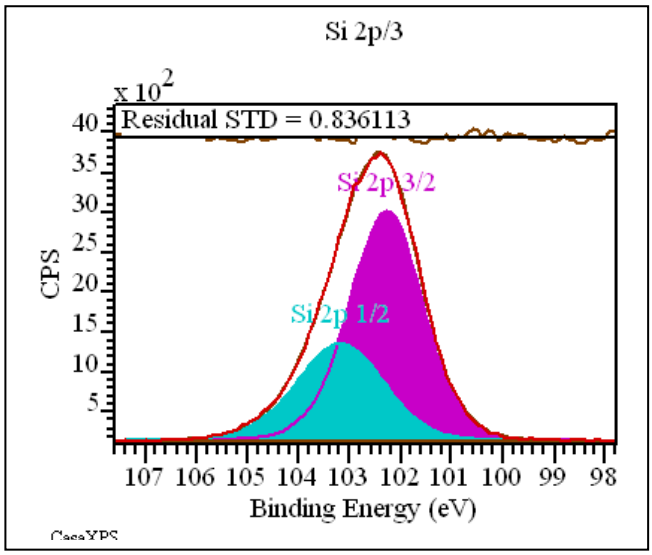
Confidence intervals are offered to indicate how precise a value is computed from the data given a peak model. The uncertainty estimate is exactly as stated, an estimate, with the merit of indicating when a parameter value ***cannot*** be relied upon. The values calculated by CasaXPS are only as good

as the method used and the applicability of the definition for the confidence region to the distribution data. Calculating the confidence intervals for the parameters from the Monte Carlo data directly simply provides an error estimate for the fitting parameters which should be used to compare uncertainties obtained by one peak model against another. If a different method or definition of the confidence region were adopted the values for the uncertainties would be different from those reported by CasaXPS. It is therefore important to appreciate uncertainties are context sensitive, where the context is both the method used to calculate the uncertainty from a set of outcomes to an experiment and the criterion used to decide if a result is within the set of acceptable values.

Also, as with all numerical methods, there are circumstances where any given method will fail. The result of the Monte Carlo simulation provided in VAMAS format and the ability to plot these distributions provides a means of assessing the uncertainty calculation itself. For example, a given peak model may be unstable resulting in two very different outcomes depending on the noise in the data. Plotting the distributions might show two or more clusters of parameter sets with distinct and separate centres. Under these conditions assuming the centre of mass as being representative of the distribution is possibly wrong and an ellipsoidal confidence region only has context for a distribution with a single centre. Such an outcome for the error distributions is indicative of a poorly defined peak model. Adjusting constraints will usually alter the result to one consistent with the uncertainty calculation and often with the beneficial consequence of improving the stability and precision of the peak model too.

Peak Fitting and Error Estimates

Quantification of a sample using XPS is typically presented as a set of atomic concentrations for the elements evident in the data. Evidence of an element in the sample consists of a set of peaks in the spectra and the ability to measure the contribution from each peak to an atomic concentration calculation is dependent on separating peaks arising from different elements. The separation of overlapping peak intensities is achieved by constructing a peak model from known lineshapes and fitting these component peaks to the data envelope. The following data envelope is from a silicon dioxide sample. The measured data envelope is a simple example of two component peaks.



The usual objective for modelling a data envelope is to estimate the relative intensity of elements using peak area. XPS often presents situations where peaks can be identified from the same element with differing position and FWHM as well as from different elements with coincidently overlapping peaks. When peaks overlap there are two problems involved in the calculation: the first, and instrumental in the atomic concentration calculation, involves determining the peak area for each peak underlying a peak structure; the second problem is to estimate the precision associated with these area measurements. The subject addressed in this section is the latter, namely, assuming a peak model is correctly defined, estimate the uncertainty in the peak area values determined from the model.

Peak Parameters

Consider the case in which a component peak is defined by a Gaussian lineshape:

$$G(E) = Ae^{-\left[\frac{(E-P)}{F}\right]^2} \quad \cdots \quad (1)$$

The functional form in Equation (1) contains three parameters P, A and F which alter the mean position for the functional form, the area between the abscissa and the function, and the spread of peak area over the energy axis, respectively. Given a data envelope which can be well approximated by a single Gaussian as defined by Equation (1), the problem is to choose values for these parameters P, A and F which minimise the chi square

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{d_i - Ae^{-\left[\frac{(E_i-P)}{F}\right]^2}}{\sigma_i} \right)^2$$

where $\{d_1, d_2, \cdots, d_n\}$ are the measured data intensities corresponding to energies $\{E_1, E_2, \cdots, E_n\}$ with individual standard deviations $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$.

There are many methods for minimising the χ^2. Essentially these methods iteratively adjusting the current parameter set P, A and F until the χ^2 function appears to be at a minimum, and the method chosen to minimise the χ^2 is of no importance to the problem provided the method yields a reliable minimum in a timely fashion.

Monte Carlo Simulation

The area for the Gaussian in Equation (1) can be calculated from the three fitting parameters once a minimum is achieved. The uncertainty for the measured area may be estimated by taking a set of identical samples, repeating the measurement several times and fitting the same model Gaussian to these independent measures. Since the variable element in each determination of the peak area is the noise contribution to the measured signal, the fitting parameters will only vary from the first set determined due to the instrumental noise contribution in the data.

The principle behind Monte Carlo error estimation is analogous to the experimental method just described. The only difference between repeating the experimental measurement and the Monte Carlo approach is noise is introduced into the results from a single measurement using a random number generator rather than allowing the noise inherent in the measurement process for a sequence of measurements to alter the initial conditions to the fitting procedure. The output from

both approaches is a list of fitting parameters differing from the original values only due to the influence of noise.

The uncertainties for the adopted fitting parameters and therefore the peak area are calculated from these variations in the parameter sets resulting from noise. One difference between an empirical approach and Monte Carlo is the measurement process will necessarily introduce noise characteristic of the instrument and sample while the Monte Carlo method requires a theoretical specification of the noise.

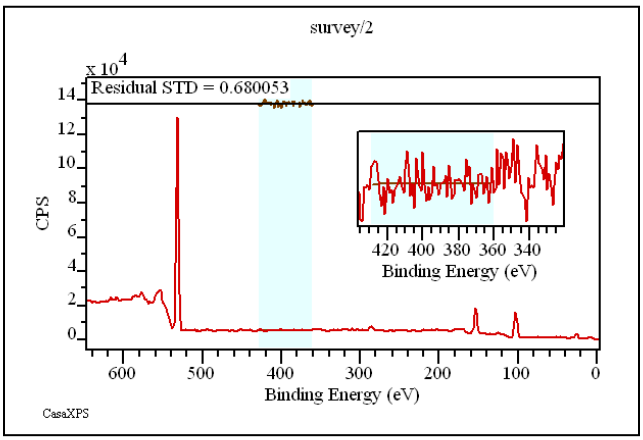
For pulse counted data measured from a large sample of random events, in this case induced by photoionisation, the noise distribution is modelled using a Poisson distribution where the standard deviation in the recorded intensity in a data channel is the square root of the counts per bin. The validity of assuming a Poisson noise distribution for the counts per bin is perhaps the weak link in estimating the errors using Monte Carlo or any other theoretical methods reliant on knowledge of the noise distribution. Problems with such an assumption exist for instruments for which data are collected using multiple detectors or for detecting systems which are not strictly reporting raw intensities. The consequence of multiple detector systems is the *raw* data appears smoothed by the averaging procedures typically adopted when combining spectral information from multiple data streams.

A simple procedure for testing the validity of assuming Poisson statistics for the spectral data bins is to measure an energy range without any peaks in the data. Using the Regions property page, add a region to the spectrum and select the regression background type for the region. A linear background is added to the data chosen to minimise

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{d_i - \left[a\left(E_i - P_{region}\right) + b \right]}{\sigma_i} \right)^2$$

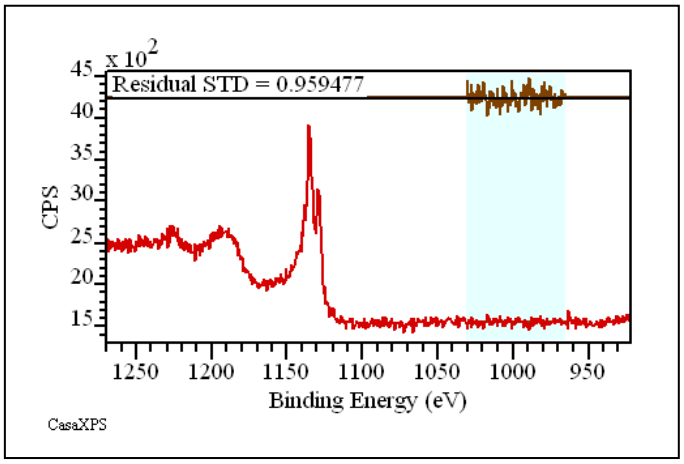
by calculating the linear parameters a and b over the interval defined by the region. The standard deviation reported for the residual will be close to unity suggests the noise in the data channels obey Poisson statistics.

The following data are from a multiple detector XPS instrument. The residual standard deviation is too good and is the result of merging multiple data streams to produce the spectrum.



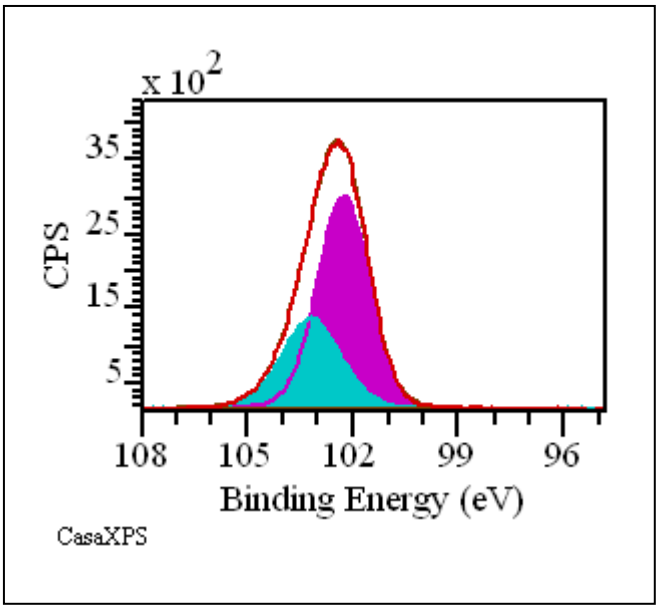
Error estimates assuming Poisson distributed noise will be conservative for cases where the data have a better than expected residual standard deviation.

Data collected using a single channeltron electron detector typically results in the expected Poisson behaviour. The following data are collected with such a detection system which is typical of older instruments. However, it is also possible that an instrument can introduce more than the expected noise, in which case the uncertainties will be under estimated by the Monte Carlo approach.

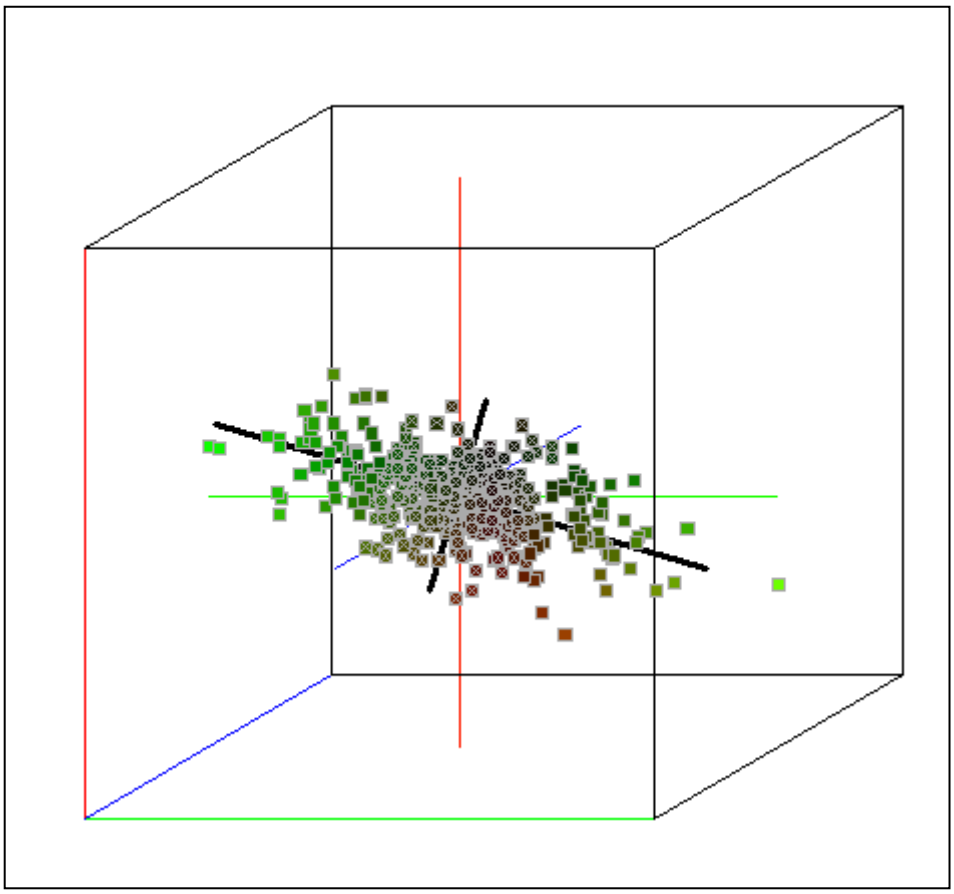


Estimating the Errors in Peak Parameters

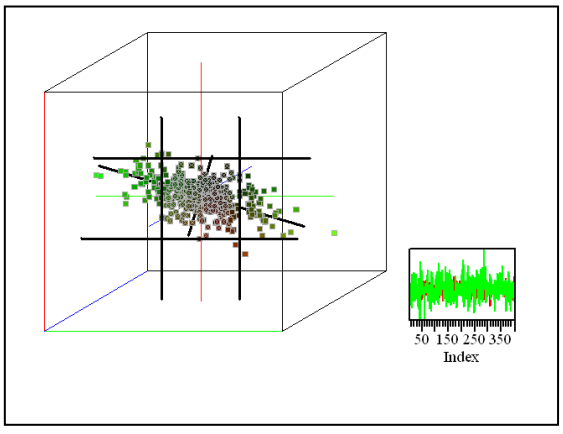
For uncorrelated fitting parameters, one standard deviation in the distribution for the individual fitting parameters offers the uncertainty interval with a 68.3% confidence. Unfortunately the only time peak fitting is performed is precisely when the peaks are correlated and therefore the fitting parameters such as those in Equation (1) are correlated too. Instead of considering the range of variation for each individual parameter is becomes necessary to consider a multi-dimensional distribution from which a region containing 68.3% of the parameter sets must be determined and by projecting the extent of the region within which 68.3% of the possible parameter sets lie from the mean, the uncertainty for each of the fitting parameters can be established.



To illustrate the procedure consider the case where two peaks are defined which model a data envelope and for which the peaks are assumed to have fixed position and FWHM but the area for each of the peaks is allowed to vary. After performing a Monte Carlo simulation for such a peak model, the two free fitting parameters produce two distributions from which the uncertainties for the areas can be estimated.



A scatter plot of these two parameter distributions illustrates anti-correlation between these two fitting parameters and also highlights the set of points belonging to the region which contains 68.3% of the simulations outcomes. The scatter plot represents the outcome for the peak areas determined for each optimisation step in the Monte Carlo procedure by positioning a marker at the coordinate $\left(area_{peak1}, area_{peak2}\right)$. These markers are filled with a colour determined by the size of the coordinates relative to the mean for the individual distributions, where each coordinate axis is assigned to a colour intensity red, green, blue (RGB) as a right-handed coordinate system. Markers within the set of peak areas lying within an elliptical (in the case of a 2D plot) region containing 68.3% of these coordinates are additionally marked with a cross. An estimate for the uncertainty in these two peak area parameters is obtained by projecting the extreme values from this 2D confidence region onto the coordinate axes as illustrated below.



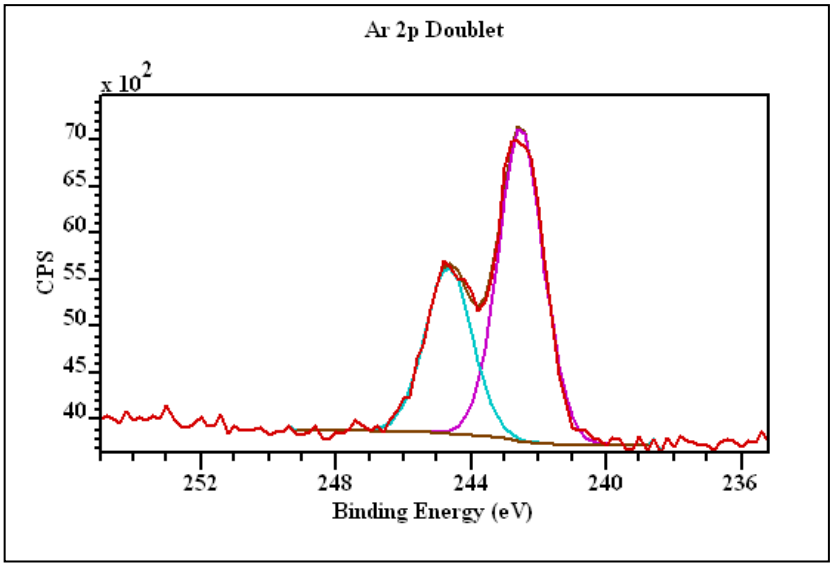
The principal axes determined from the parameter distribution are also shown. The length of these axes are three standard deviations in the distributions obtained from the coordinates of the area parameters after transformation to coordinates with respect to the principal coordinate axes.

While discussed in terms of two fitting parameters, the principles illustrated with 2D scatter diagrams can be extended to m-dimensional problems involving m fitting parameters. The visualisation of these m dimensional distributions is limited to a projection of these distributions onto at most 3D scatter plots, but the mathematics for determining the confidence region in m-dimensional space remains the same and can be used to obtain error estimates for fitting problems requiring multiple component peaks with parameters equivalent to those illustrated by Equation (1).

Principal Axes and Peak Fitting Uncertainties

When peaks are fitted to data the answer returned by the fitting procedure for the peak areas is just one of many possible answers. The source for these alternative outcomes to the fitting algorithm is random variations in the data due to noise which differs with each measurement. The influence noise on the outcome is dependent on the peak model used to approximate the data envelope and the nature of the envelope in terms of underlying peak proximity.

The following data from a sample measured after bombardment with argon ions by XPS pulse counted signal will be used to facilitate a discussion of the issues associated with estimating errors in the parameters adjusted to fit the synthetic peaks to the data. An energy range appropriate for electrons emitted from the Ar 2p core level provides an example of a doublet electron state yielding two overlapping peaks. The discussion will focus on the determination of the peak areas for each of the Ar $2p_{1/2}$ and Ar $2p_{3/2}$ component peaks.



A Monte Carlo procedure is used to simulate repeating the measurement many times and with each simulated data set the peak area is recalculated. The result of the Monte Carlo simulation is a table of variations from the initial peak parameters calculated from the Ar 2p spectrum. In this current example of fitting two synthetic peaks the number of adjustable parameters is six, namely, two sets of area, position and FWHM parameters, one set per synthetic peak. While these six parameters

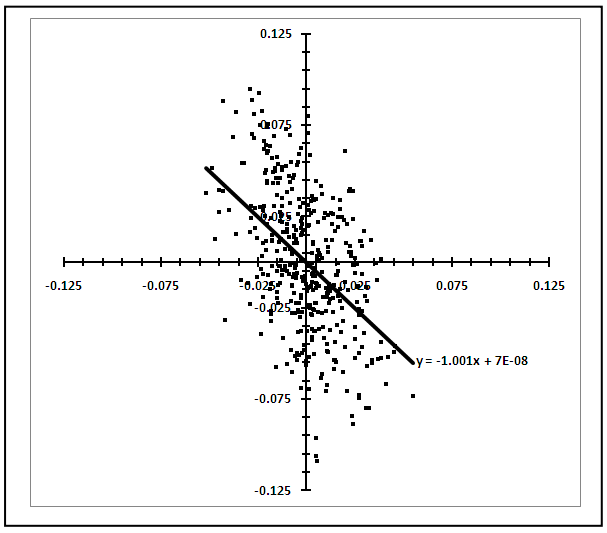
yield six error distributions, the discussion will proceed by focusing only on the two area distributions in isolation. This is only to highlight the nature of the analysis involved in understanding error distributions. The determination of the uncertainties for these area parameters in CasaXPS includes all fitting parameters and not just the two distributions now discussed.

Methods for Characterising Trends in Scatter Plots

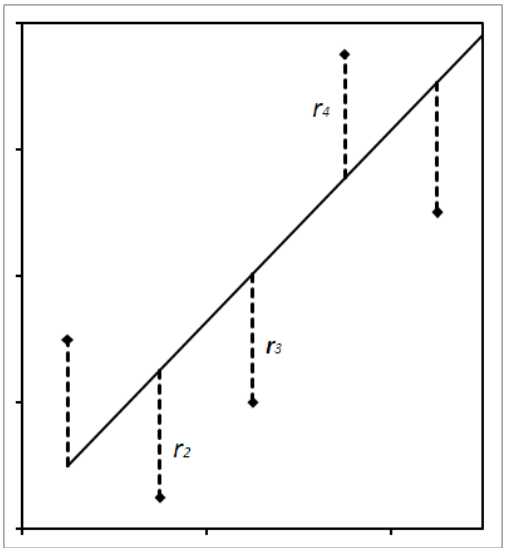
The peak areas, measured relative to the area parameter calculated from the data, are plotted for the two peaks fitted to the Ar 2p spectrum as a set of points on Cartesian axes. Since the two peaks overlap it is logical that if one peak increases in area, to fit the same data envelope, the second peak must reduce in area. It is therefore reasonable to believe the scatter plot for these area parameters is anti-correlated. A regression line calculated for the set of coordinates

$$\left(\frac{area_{simulated\ peak_1} - area_{peak_1}}{area_{peak_1}}, \frac{area_{simulated\ peak_2} - area_{peak_2}}{area_{peak_2}}\right)$$

supports this theory.



A Scatter Diagram in which the y-axis is the area calculated for the Ar $2p_{1/2}$ peak and the x-axis is the area for the Ar $2p_{3/2}$ peak. The peak areas are plotted centred with respect to the initial values for the areas determined from fitting the model to the data and normalised to these initial areas.

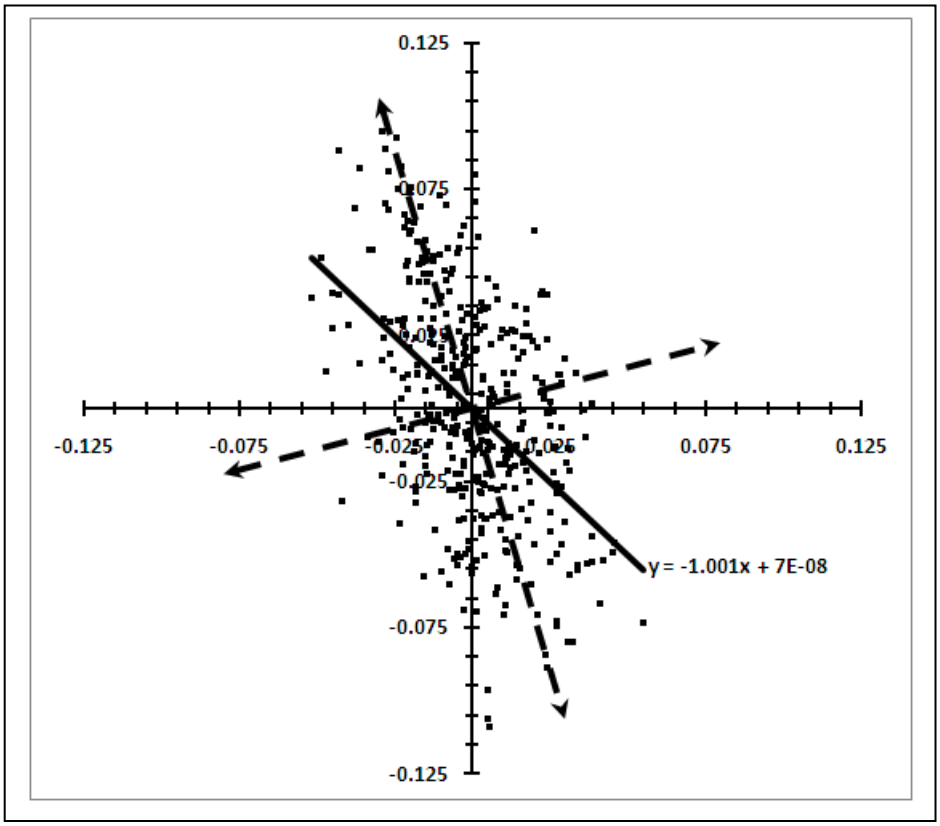


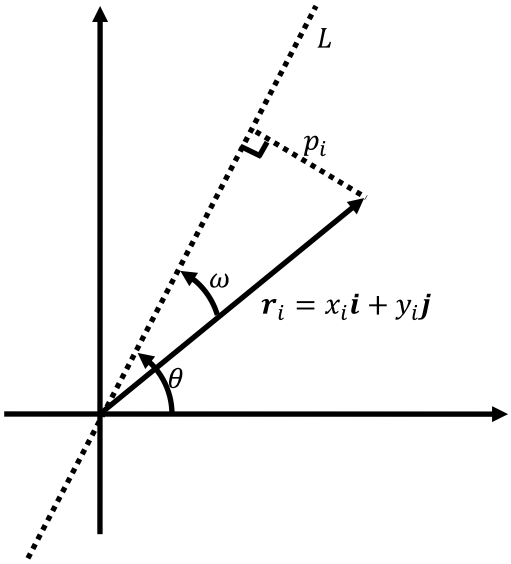
The trend in the above scatter diagram is illustrated using a regression line. The values $r_i = y_i - (a + bx_i)$ are called the residuals and are depicted graphically as vertical lines between the data points and the line of best fit.

The line of best fit is therefore considered to be the line which minimises the sum of the squares of the residuals.

$$\chi^2 = \sum_{i=1}^{n} {r_i}^2$$

While a regression line is often used to determine the line of best fit, the error distribution is visually different in nature from the regression line. The regression line is limited to minimising the residuals with respect to one distribution only. An alternative approach for lines passing through the origin is to consider the shortest distance from each point on the scatter plot to a line of best fit. If the set of points are considered to be the positions for a set of unit mass particles relative to the centre of mass located at the origin, then the problem of calculating the lines of best fit for these two distributions is equivalent to finding the principal axes for the moment of inertia for a collection of particles.





Consider a set of particles of unit mass with position vectors $\boldsymbol{r}_i = x_i\boldsymbol{i} + y_i\boldsymbol{j}$ and a line L making an angle θ with the positive x direction. Let the shortest distance between the point with position vector $\boldsymbol{r}_i$ and the line L be p_i.

A line of best fit through the origin may be obtained by calculating the minimum for

$$I = \sum_{i=1}^{n} {p_i}^2$$

Let the unit vector in the direction of the line L be $\hat{\boldsymbol{l}} = \alpha\boldsymbol{i} + \beta\boldsymbol{j}$ the $p_i = |\boldsymbol{r}_i| \sin\omega = |\boldsymbol{r}_i \times \hat{\boldsymbol{l}}|$

$$\boldsymbol{r}_i \times \hat{\boldsymbol{l}} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ x_i & y_i & 0 \\ \alpha & \beta & 0 \end{vmatrix} = \boldsymbol{k}(x_i\beta - y_i\alpha)$$

Therefore ${p_i}^2 = (x_i\beta - y_i\alpha)^2$ and so

$$I = \sum_{i=1}^{n} {p_i}^2 = \sum_{i=1}^{n} (x_i\beta - y_i\alpha)^2$$

$$\Rightarrow I = \sum_{i=1}^{n} ({x_i}^2\beta^2 - 2x_iy_i\alpha\beta + {y_i}^2\alpha^2)$$

$$\Rightarrow I = \beta^2 \sum_{i=1}^{n} {x_i}^2 + \alpha^2 \sum_{i=1}^{n} {y_i}^2 - 2\alpha\beta \sum_{i=1}^{n} x_iy_i$$

Let

$$A = \sum_{i=1}^{n} {x_i}^2, B = \sum_{i=1}^{n} {y_i}^2 \; and \; C = \sum_{i=1}^{n} x_iy_i$$

$$\Rightarrow I = A\beta^2 + B\alpha^2 - 2C\alpha\beta$$

Since $\hat{\boldsymbol{l}} = \alpha\boldsymbol{i} + \beta\boldsymbol{j}$ can be expressed in terms of the angle θ between the line L and the x-axis, the expression for I can also be expressed in terms of the angle θ; the values A, B and C are all calculated from the data and are therefore known.

$$\alpha = \cos\theta \; and \; \beta = \sin\theta$$

Therefore

$$I = A\sin^2\theta + B\cos^2\theta - 2C\sin\theta\cos\theta$$

The line of best fit can be obtained by minimising I with respect to θ.

$$\frac{dI}{d\theta} = A2\sin\theta\cos\theta - B2\sin\theta\cos\theta - 2C(\cos^2\theta - \sin^2\theta)$$

Since $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\sin\theta\cos\theta$

$$\frac{dI}{d\theta} = (A - B)\sin 2\theta - 2C\cos 2\theta$$

The minimum in I occurs when $\frac{dI}{d\theta} = 0$ therefore

$$(A - B)\sin 2\theta - 2C\cos 2\theta = 0$$

$$\tan 2\theta = \frac{2C}{(A - B)}$$

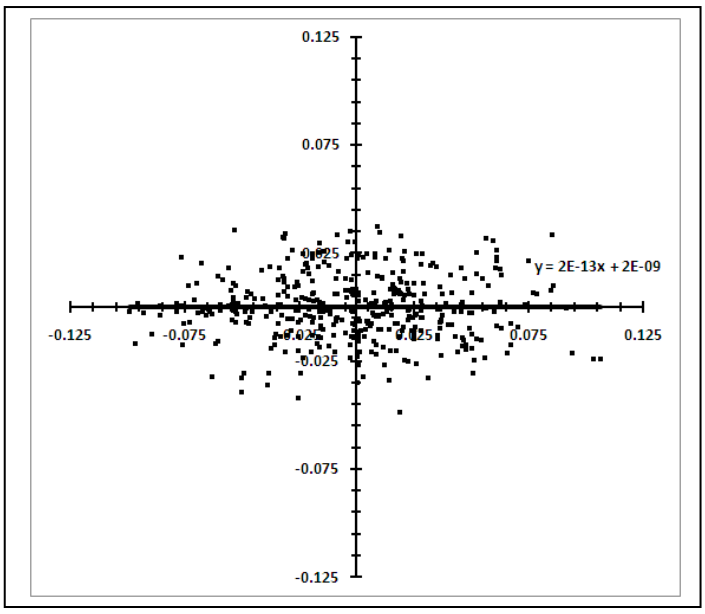
Applying the result to the two distributions for the area parameters from the Ar 2p spectrum the minimum for I occurs for ($A = \sum_{i=1}^{400} {x_i}^2 = 0.132165691$, $B = \sum_{i=1}^{400} {y_i}^2 = 0.563587216$ and $C = \sum_{i=1}^{400} x_i y_i = -0.132312195$

$$\tan 2\theta = \frac{2(-0.132312195)}{(0.132165691 - 0.563587216)} \Longrightarrow 2\theta = \tan^{-1} 0.613377809$$

$$\Longrightarrow \theta = \frac{0.550198098}{2} + k\frac{\pi}{2}\ radians$$

Therefore two extrema occur for lines at $\theta = 15.76°$ and $\theta = 15.76° + 90° = 105.76°$ with respect to the x-axis.

If the set of data points are transformed by rotation by $-105.76°$ the image for the distributions appears as follows.



An alternative means of finding the principal axes which generalises to multidimensional distributions is to minimise the function $I = A\beta^2 + B\alpha^2 - 2C\alpha\beta$ subject to the constraints $\alpha^2 + \beta^2 = 1$.

Applying the method of Lagrange multipliers to this constrained optimisation problem involves minimising

$$\Psi(\alpha, \beta) = I(\alpha, \beta) - \lambda(\alpha^2 + \beta^2 - 1)$$

The extrema are obtained by the condition $\frac{\partial \Psi}{\partial \alpha} = 0$ and $\frac{\partial \Psi}{\partial \beta} = 0$

$$\Longrightarrow \frac{\partial \Psi}{\partial \alpha} = 2B\alpha - 2\,\beta C - 2\lambda\alpha = 0 \text{ and } \frac{\partial \Psi}{\partial \beta} = 2A\beta - 2\,\alpha C - 2\lambda\beta = 0$$

Resulting in the simultaneous equations

$$2B\alpha - 2\,\beta C - 2\lambda\alpha = 0$$
$$2A\beta - 2\,\alpha C - 2\lambda\beta = 0$$

$$\Rightarrow \begin{matrix}(B-\lambda)\alpha - \beta C = 0\\ -\alpha C + (A-\lambda)\beta = 0\end{matrix}$$

or in matrix notation

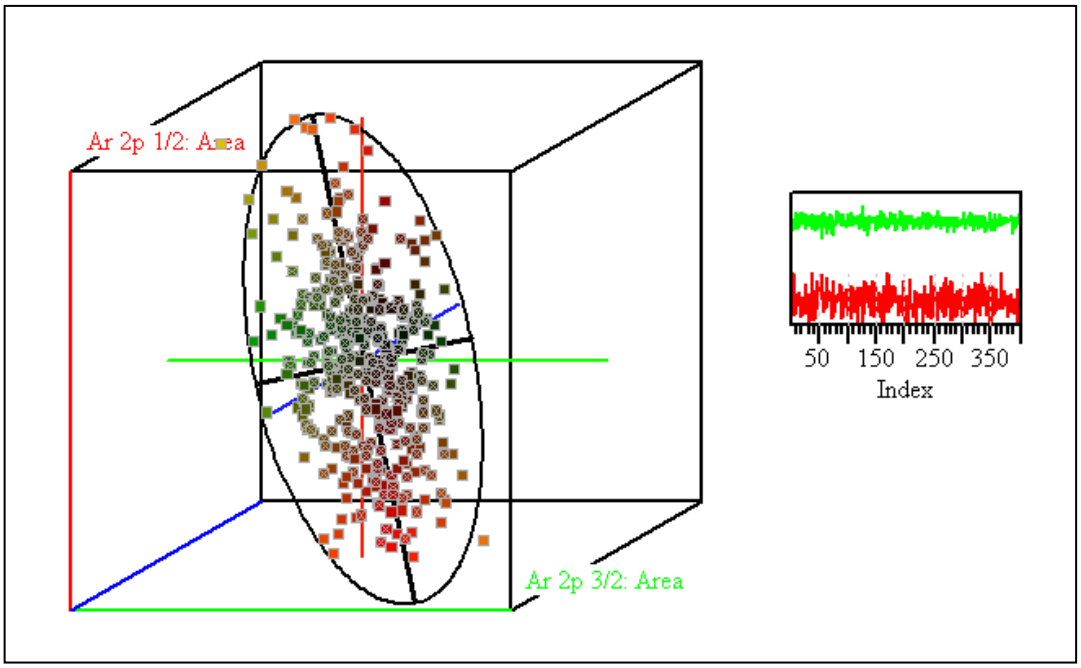
$$\begin{bmatrix}(B-\lambda) & -C\\ -C & (A-\lambda)\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \mathbf{0}$$

A non-trivial solution is only possible if these two lines are parallel which means mathematically λ is an eigenvalue of the matrix

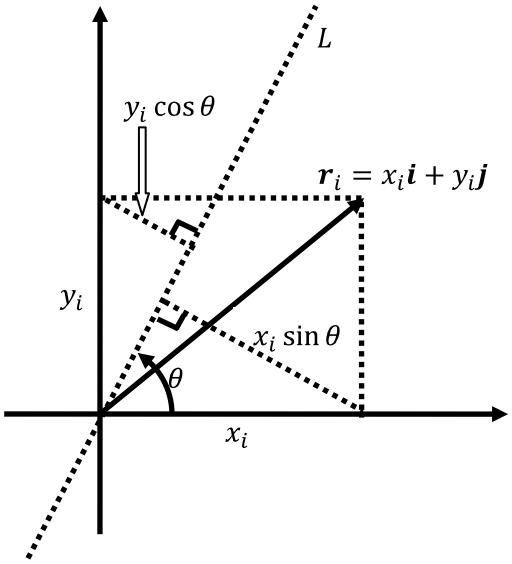
$$\boldsymbol{Z} = \begin{bmatrix}B & -C\\ -C & A\end{bmatrix}$$

The eigenvalues and eigenvectors for $\boldsymbol{Z}$ are the principal axes shown on the scatter diagram.

A geometric interpretation, at least for the 2D problem, is the principal axes are the directions about which the variation in the data points plotted in the plane are a minimum in one direction and a maximum in the orthogonal direction. Principal axes are calculated for multi-dimensional distributions by performing an eigenanalysis. The residual option on the toolbar of CasaXPS turns these principal axes on and off when data are displayed as a scatter plot.



Principal Component Analysis and Principal Axes



Consider a set of particles of unit mass with position vectors $\boldsymbol{r}_i = x_i\boldsymbol{i} + y_i\boldsymbol{j}$ and a line L making an angle θ with the positive x direction. The direction cosines for the line L are $\alpha\boldsymbol{i} + \beta\boldsymbol{j} = \cos\theta\,\boldsymbol{i} + \sin\theta\,\boldsymbol{j}$.

A line of best fit through the origin may be obtained by calculating the minimum for

$$J = \sum_{i=1}^{n} (x_i\beta + y_i\alpha)^2$$

subject to the constraint $\alpha^2 + \beta^2 = 1$

The analysis for I can be applied for J resulting in the following optimisation problem.

$$\text{Minimise } J = A\beta^2 + B\alpha^2 + 2C\alpha\beta \text{ subject to the constraint } \alpha^2 + \beta^2 = 1$$

Where again

$$A = \sum_{i=1}^{n} {x_i}^2, B = \sum_{i=1}^{n} {y_i}^2 \; and \; C = \sum_{i=1}^{n} x_i y_i$$

So applying the method of Lagrange Multipliers leads to finding the eigenvectors and eigenvalues of

$$\boldsymbol{W} = \begin{bmatrix} B & C \\ C & A \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}.\boldsymbol{y} & \boldsymbol{x}.\boldsymbol{y} \\ \boldsymbol{x}.\boldsymbol{y} & \boldsymbol{x}.\boldsymbol{x} \end{bmatrix}$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$. The matrix $\boldsymbol{W}$ is the covariance matrix associated with a Principal Component determination. The eigenvalues for $\boldsymbol{W}$ and $\boldsymbol{Z}$ are identical since both are calculated from the same quadratic

$$(B - \lambda)(A - \lambda) - C^2 = 0$$

The eigenvectors for $\boldsymbol{W}$ and $\boldsymbol{Z}$ are related by a reflection in the x axis. The principal axes can therefore be obtained using either $\boldsymbol{W}$ or $\boldsymbol{Z}$.

Line of Best Fit using Z

As an alternative to linear regression, in general, a line of best fit in the moment of inertia sense, where the distance between a point and the line is measured using the shortest distance from the point to the line, is obtained by calculating the gradient for the line of best fit passing through the mean coordinate for the two data distributions, with gradient obtained from the principal axis vector determined from the matrix $\boldsymbol{Z}$.

Applying the notation used for linear regression, the mean centred distributions are summarised as follows:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where

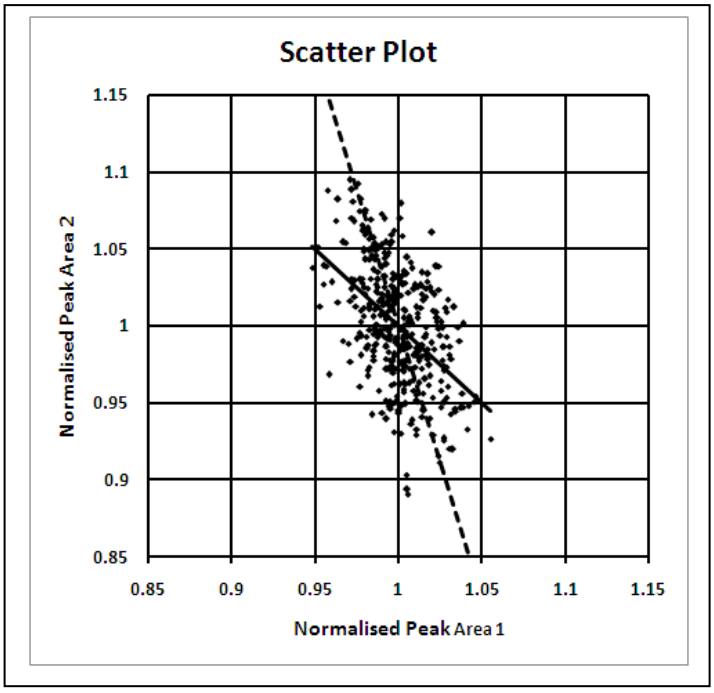
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The equation of best fit $y = a + bx$ is given by

$$b = \frac{(S_{yy} - S_{xx}) + \sqrt{(S_{yy} - S_{xx})^2 + 4(S_{xy})^2}}{2S_{xy}}$$

and

$$a = \bar{y} - b\bar{x}$$



For the scatter plot of the peak area for each of the two Ar 2p peaks relative to the initial values for the peak areas in the Monte Carlo simulation, the two possible lines of best fit have gradient -1 in the case of linear regression, while the line of best fit based on principal axis has gradient -3.54. Both lines pass through the mean coordinate for the data $(\bar{x}, \bar{y}) = (1,1)$.