CHAPTER 4
XAFS DATA ANALYSIS

In practice, the XAFS data is recorded as a function of energy. In most XAFS experiments, the exact concentration of the absorbing atom and the sample thickness is not precisely known. Besides, a variety of materials such as kapton windows and the sample matrix enter into the beam path before reaching the detector. This causes the observed signals to be multiplied by energy dependent absorption factors. To compensate for this, in transmission experiments, a logarithm of the ratio of the incident and transmitted intensities is taken so that the multiplicative factor is converted into an additive background which slowly varies with energy. However, in fluorescence experiments, no logarithm is taken and the energy dependent factors are carried through the entire analysis. The elastically and inelastically scattered x-ray photons from the incident beam contribute an additive background to the fluorescence data. Therefore,

$$[\mu(E)]_{Transmission} = \ln\left(\frac{I_0}{I_t}\right); \tag{4.1}$$

$$[\mu(E)]_{Fluorescence} = \frac{I_f}{I_0} \tag{4.2}$$

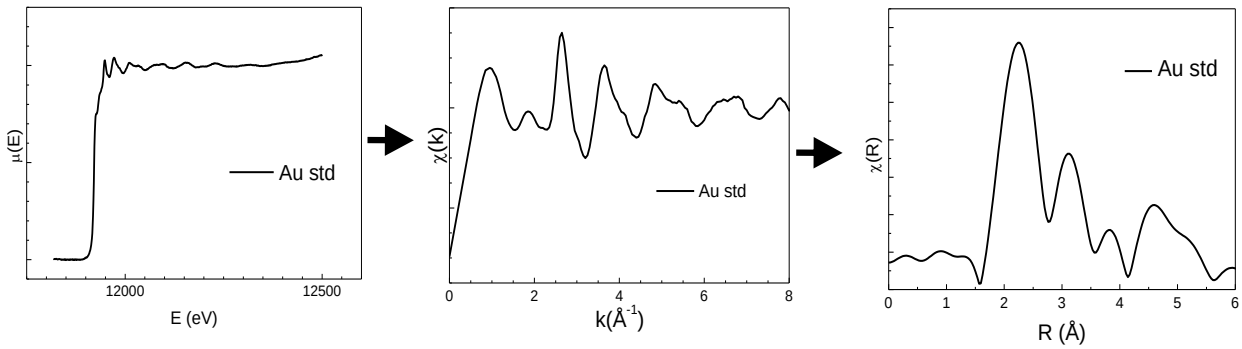

Fig. 4.1: Schematic of XAFS data analysis

The collected data is then converted to wave-number (k), which is Fourier Transformed to yield radial distribution about the absorbing atom. Fig. 4.1 shows the schematic of XAFS data analysis.

XAFS data analysis is non-trivial and proceeds via two main steps: (1) Data Processing, (2) Data Modeling and Fitting. (Fig. 4.2)

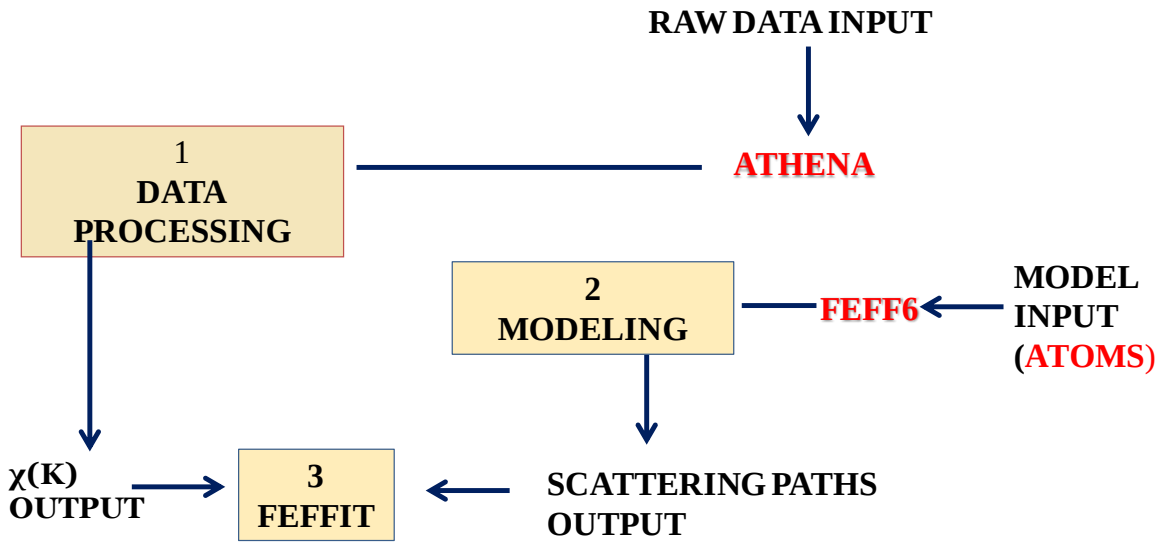

Fig. 4.2: Data Analysis Flowchart

4.1. Data Processing

Data processing is the most challenging step in XAFS data analysis. The most popular software for data processing is ATHENA.[83] Data processing involves several steps:

(i) Merging of scans: XAFS oscillations are very weak in nature and hence very high signal-to-noise ratio ($S/N \sim 10^3$) is required to get meaningful information from the data. Due to the effect of radiation on the sample and due to oxidation of the sample, the XAFS spectrum may vary from scan to scan. Moreover, there could be noise arising from instrumental errors. Therefore, in

practice, several data scans are taken to check data reproducibility. The scans are merged to reduce statistical noise.

(ii) Determining of threshold energy E_0: *k*, the photoelectron wave number in the EXAFS equation, is defined with respect to threshold energy E_0 (binding energy of the electron) as $k = \sqrt{\frac{2m(E - E_0)}{\hbar^2}}$. E_0 for the onset of absorption depends on the chemical state of the central atom and the structural symmetry of its surroundings. So, for the same absorbing atom in different samples, E_0 may vary upto a few eV. The purpose of E_0 determination is to correctly and quantitatively account for any edge shift observed in the different samples. In practice, E_0 is chosen to be the energy at half of the edge step. Only relative shift in E_0 between standard and unknown samples is very important: if both are shifted by the same amount, the answer will be the same to a good approximation. Relative E_0 shifts primarily affect the data at low values of k which are distinguishable from changes in other structural parameters. Thus, ambiguities in absolute E_0 position, and small (~3 eV) differences in relative E_0 position, do not introduce corresponding ambiguities in structure determination by EXAFS. The choice of E_0 does pose significant uncertainties for k-space analysis in the XANES region, however.

After the edge energy E_0 is fixed, the whole data gets divided into the pre-edge and the post-edge regions.

(iii) Deglitching of the data: Absorption and fluorescence data often include sharp points, called glitches, which may be many times larger than the EXAFS signal. Although they have a small energy width, because of the finite data content, upon Fourier transformation, they may be magnified manifold and

have a small effect on the data. The background subtraction may also be incorrect because of the presence of these spiky glitches and can change the chi-space spectrum. So these must be removed in the initial stages of data processing. In order to distinguish a glitch from a data point, one must look at the plot for I_0. A glitch in the μ(E) plot corresponds to a dip in energy in the I_0 plot.

(iv) Pre-edge background subtraction: The removal of pre-edge background involves fitting of a functional form to the data before the edge and extrapolation of this function to the XAFS data region. Since the pre-edge background is a slowly varying function of energy, it can be fitted to a linear function between 200 eV and 50 eV below the edge and then extrapolated to the post-edge region.

(v) Post-edge background subtraction: Extracting the post-edge background function, μ_0, is the most critical step of background removal, as this function can affect the final conclusions for the structural information. Incorrect background removal leads to leakage of the background peak into the nearest neighbor peak of interest and distorts the latter's structural information (particularly coordination number).

The XAFS oscillations should be even about the background line.

The background is approximated by a single cubic least-squares-spline polynomial[84] over the energy region of interest, with the first and last data points respectively at the beginning and end of the energy range defined. The polynomial function is solved using these data points. If the background subtraction using this method does not look good, then the background may be approximated by a series of cubic polynomial functions that are joined at

"knots" where the functions and their first derivative are continuous. It must be ensured that the data is not distorted by background removal process. Too few knots might not approximate the background oscillation well enough and might show as a low-*r* peak in the Fourier transform. On the other hand, too many knots might tend to follow the true XAFS oscillations and distort the analysis. The position of the knots can be varied and the background may also be forced to pass through a certain point in cases of a complicated background fitting. Once the fitting parameters are fixed, the fitting program tries to adjust the ordinates for the minimization of the mean square error between the data and the polynomial.

Alternately, a program called AUTOBK[85] is popularly used to remove the "background function" from a measured XAFS spectrum, using an automated algorithm based on the Fourier content of the XAFS spectrum. AUTOBK considers the background function to be that part of the measured absorption which does not contain any structural information. Salient features of this method of background subtraction are (1) No point on the absorption curve is assumed to be on the background curve, (2) AUTOBK chooses a smooth spline to best fit only the low-frequency components of $\mu(E)$, (3) The number and the abscissas of knots in the AUTOBK method are explicitly determined because of the restrictions put on the background spline from the information content of $\tilde{\chi}(r)$.

$$N_{knots} = \frac{2r_{bkg}\Delta k}{\pi}, \tag{4.3}$$

where Δk is the k-range of useful data and r_{bkg} is the upper limit of the low-R region over which the background is to be fit. r_{bkg} is typically about half the

distance of the first shell peak. The knots are set to be equally spaced in *k*-space while the ordinates of the knots are determined by minimizing the difference between data and standard $\tilde{\chi}(r)$ in the low-*r* region.

Difficulty in using AUTOBK arises when leakage from the first shell causes substantial low-r component of $\tilde{\chi}(r)$. In such cases, a suitable standard may be measured to give an estimate of the leakage into the low-*r* region from the first shell.

For very noisy data, the first background subtraction method tends to do a better job. Mostly, background subtractions from the two methods are comparable. However, AUTOBK tends to give slightly more even oscillations about the background line in *k*-space.

(vi) Transformation to k-space: This involves extracting the oscillations χ(k) and interpolating them to an equally spaced grid in photoelectron wave number k. The extraction to k-space (from energy space) is essential to derive the structural parameters since (k,R) are the conjugate variables, not (E,R).

(vii) Fourier Transform to r-space: Fourier transform of the data is taken over a finite range from k_{min} to k_{max}.

$$\tilde{\chi}(r) = \frac{1}{2\pi} \int_{k_{min}}^{k_{max}} \chi(k) e^{2ikr} dk \qquad (4.4)$$

In order to increase the data content, *k* is chosen upto a long range (at least upto 12 $Å^{-1}$). The choice of *k* is such that good quality data exists upto the value of $k_{max.}$ k_{min} is chosen in the region where the background doesn't change rapidly.[86]

A practical problem associated with the Fourier transforming technique here is due to the finite range. Truncation ripples often interfere with physical features

of interest. For a limited range of k as in our case, $\chi(k)$ may be replaced by

$W(k)\chi(k)$ where $W(k) = \begin{cases} 1, & k_{\min} < k < k_{\max} \\ 0, & otherwise \end{cases}$

The transform $\tilde{A}(r) = W(r) * A(r - R)$,

where $\tilde{W}(r) = \exp(ik_0 r)\dfrac{\sin \Delta k r}{r}$,

k_0 is the center of transform range and Δk is the width. This is a damped sine function with a peak at $r = 0$. The convolution of this $\tilde{A}(r)$ with $\tilde{W}(r)$ results in peak broadening as well as truncation ripples on the sides of the main transform peak. These ripples can contaminate neighboring peaks.

It is possible to reduce the amplitude of truncation ripple at the expense of further peak broadening by using a tapered windowing function of the form

$$W(k) = \cos^2\left(\frac{k - k_0}{\Delta k}\right), for \left|k - k_0\right| < \Delta k \qquad (4.5)$$

where k_0 and Δk are as defined before. The $\cos^2$ window is often referred to as Hanning function. The transform of this function is given by

$$\tilde{W}(r) = \exp(2ik_0 r)\frac{\sin(\Delta k r)}{r}\frac{\pi^2}{2(\pi^2 - 2\Delta k^2 r^2)} \qquad (4.6)$$

Comparing this with the result of that without window, it is apparent that the truncation ripples die out much faster at large $(r - r_0)$ using the smooth window, but the central peak is significantly broader. It is often useful to combine a flat region over the middle of the data range with a Hanning function at the ends, that is:

$$W(k) = \begin{cases} \sin^2\left(\dfrac{\pi(k - k_1)}{2(k_2 - k_1)}\right), & k_1 < k < k_2 \\ 1 & , k_2 < k < k_3 \end{cases}$$

$$= \cos^2\left(\frac{\pi(k-k_3)}{2(k_4-k_3)}\right), k_3 < k < k_4$$

where k_1, k_2, k_3, k_4 delimit three different ranges corresponding to the two end regions and the flat central region.

To sum up, window functions are used to weigh different portions of the data, truncate the data smoothly at the ends of the transform range and make data from two samples appear as close as possible.

After selecting appropriate k-range, the data is exported as a chi (k) file. Further analysis is performed on this exported data as described in the following section.

4.2. Data Modeling and Fitting

This step involves the following sub-steps:

4.2.1 Construction of the appropriate structural model using FEFF6[87]

Suitable structural model is simply based on physical intuition. The scattering phase shifts and amplitudes corresponding to the input model are generated *ab initio* by program FEFF6. FEFF6 calculates wave phase shifts, effective scattering amplitudes and single and multiple scattering curved XAFS spectra, including polarization dependence, for clusters of atoms (Z<95). FEFF6 requires atomic coordinates for the input model and physical information about the system such as absorbing atom and excited core level for its calculation. For crystalline systems, the program ATOMS[88] is used to generate list of atomic coordinates. ATOMS generates the required coordinates starting from a crystallographic description of the system, for e.g., lattice type, lattice parameters and space group. For a disordered system, the model bond lengths are assumed from the peak positions of the Fourier transformed data.

FEFF6 calculates EXAFS using four internal modules:

1. POTPH (Potentials and Phases): This module calculates the scattering potentials for each type of atom by overlapping their wave functions based on their geometric distribution. To calculate free atomic potentials, the inputs required by this module are the atomic numbers of all the atoms and the type of core-hole being considered.
2. PATHS (Path enumeration): This module identifies all significant single and multiple scattering paths for an arbitrary cluster of atoms, in order of increasing path length, based on a list of atomic coordinates. Different paths are weighed on the basis of scattering amplitudes and subsequently the low amplitude paths are neglected.
3. GENFMT (Scattering amplitude and other XAFS parameters calculation): For each path the code calculates the effective scattering amplitude (F_{eff} from which FEFF gets its name), total scattering phase shift and other XAFS parameters and writes the result to a corresponding file feffnnnn.dat.
4. FF2CHI (XAFS spectrum calculation): This module constructs the XAFS spectrum chi (k) using the XAFS parameters from one or more paths. Single and multiple scattering Debye-Waller factors are calculated using a correlated Debye model. Output from this module is the total XAFS spectrum and optionally, the contribution to the XAFS from each path individually.

4.2.2 Fitting the peaks in *r*-space using least squares fitting method and FEFFIT[89]

FEFFIT compares the experimental XAFS data to theoretical calculations. It fits the calculations from FEFF to XAFS chi (k) data, hence aiding in the determination of local structure around an atom. It can determine inter-atomic distances, co-ordination numbers and atomic species of neighboring atoms. Finer details of atomic configurations, including detailed descriptions of two-body distribution functions and certain aspects of three-body correlations, are also measurable with FEFFIT.

FEFFIT fits the XAFS data using FEFFnnnn.dat files output from FEFF as the basis. The XAFS contribution of each scattering path (read from each feffnnnn.dat file) is adjusted by applying standard XAFS parameters such as co-ordination number, change in distance, Debye Waller Factor, and shift in energy origin) until the best-fit to the data is found. Standard numerical techniques are used to find the set of variables that minimizes the sum of squares of the difference between model and data and to estimate the uncertainties in the variables. Fitting can be done in either *r*-space or back-transformed k-space, with Fourier Transforms done by FEFFIT on both the XAFS data and the calculated model during the fitting process. However, *r*-space is preferred because in that case, any difference between the fit and the data is a direct measure of the quality of fit. On the other hand, in *k*-space there is no limit to the number of multiple scattering paths required, which increases the number of variable parameters. Because of the finite content of an XAFS signal, so many parameters cannot all be independently measured as indicated from the equation

$$N_{idp} \approx \frac{\Delta k \Delta r}{\pi} \tag{4.7}$$

Therefore, if k-space fitting is considered, difference between data and fit has two contributions – one from the accuracy of the model used and the other from the neglect of higher order multiple scattering paths; there is no reliable way to separate the two. Thus, in order to limit the number of paths used, *r*-space fitting is considered.

Conversion to *r*-space is undertaken via equation 4.4. *k*-range should be selected such that data of good quality exists in *k*-space upto the maximum value of k_{max} and k_{min} is beyond 2 $Å^{-1}$, in order to avoid multiple scattering and other effects. The *r*-range is defined by the peaks (bonds) of interest.

4.3 Error Analysis

4.3.1 χ^2 as an estimate of goodness of fit

The best set of variables in FEFFIT will minimize the sum of the squares of the difference of model and data XAFS. The statistic called chi-square (χ^2), which is a scaled measure of the sum of squares of a function, is generally considered the best figure of merit to judge the quality of the fit. By definition,

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{f_i}{\varepsilon_i} \right)^2 , \tag{4.8}$$

where f_i is the function to be minimized, N is the number of function evaluations and ε_i are the uncertainties in f_i approximated by a constant value due to random fluctuations.

FEFFIT[90] allows the fit to be performed in *r*- or *k*-space, but there is no conceptual difference in the way the fit is done. When fitting in *r*-space, the function to minimize is

$$f(r_i) = \tilde{\chi}_{data}(r_i) - \tilde{\chi}_{model}(r_i), r_{\min} \le r_i \le r_{\max} , \tag{4.9}$$

and when fitting in *k*-space, the function to minimize is

$$f(k_i) = \tilde{\chi}_{data}(k_i) - \tilde{\chi}_{model}(h_i), k_{\min} \le k_i \le k_{\max} \tag{4.10}$$

There is one real and one imaginary evaluation for each data point. Therefore the number of evaluations is

$$N = \frac{2(r_{\max} - r_{\min})}{\delta r} \tag{4.11}$$

when fitting in *r*-space. Likewise, when fitting in *k*-space,

$$N = \frac{2(k_{\max} - k_{\min})}{\delta k} \tag{4.12}.$$

δr and δk are the grid spacings in *r* and *k* space respectively. N has no physical significance, since δr and δk are chosen arbitrarily. The more important number is the number of relevant independent measurements:

$$N_{idp} = \frac{2(k_{\max} - k_{\min})(r_{\max} - r_{\min})}{\pi} + 2 \tag{4.13}$$

The explanation for the above equation is as follows: (a) the conjugate variables considered are *k* and 2*r*, (b) since we are measuring real and imaginary parts of $\tilde{\chi}(r)$ the information must be an even number of points, (c) even for an infinitesimally small *r*-range, we must have atleast one pair of points.

This implies that $\chi^2 = \sum_{i=1}^{N_{idp}} \left(\frac{f_i}{\varepsilon_i} \right)^2 = \frac{N_{idp}}{N} \sum_{i=1}^{N} \left(\frac{f_i}{\varepsilon_i} \right)^2$ (4.14)

FEFIT uses a single value of ε for all values of ε_i for the sake of simplicity. Assuming a reasonable estimate for ε,

$$\chi^2 = \frac{N_{idp}}{N\varepsilon^2} \sum_{i=1}^{N} \{[\mathrm{Re}(f_i^{\,2})] + [\mathrm{Im}(f_i^{\,2})]\}. \tag{4.15}$$

This is the definition of χ^2 used by FEFFIT, and it is the primary figure of merit to characterize the goodness of fit. There is a related figure of merit, χ_v^2 or reduced chi-square.

$$\chi_v^2 = \frac{\chi^2}{v}, \tag{4.16}$$

where *v* = no. of degrees of freedom in the fit = $N_{idp} - N_{\mathrm{var}}$, N_{var} being the number of variables in the fit.

$\chi^2$and χ_v^2 are useful quantities for comparing the quality of different fits. The general rule is that the fit with lower χ_v^2 is the best. χ_v^2 is lowered for useful variables. If addition of a variable causes decrease in χ^2 but increase in χ_v^2 then the fit does not improve.

In principle, a good fit should have χ_v^2~1. However, practically, χ_v^2>>1! Specifically, it is unclear from the value of χ_v^2 alone if hard-to-estimate systematic errors are drowning out the random measurement errors or if the fit is truly bad. To distinguish

between these two very different conclusions, R-factor is introduced which is scaled to the magnitude of the data itself.

$$R = \frac{\sum_{i=1}^{N}\{[\mathrm{Re}(f_i)]^2 + [\mathrm{Im}(f_i)]^2\}}{\sum_{i=1}^{N}\{[\mathrm{Re}(\chi_{data})_i]^2 + [\mathrm{Im}((\chi_{data})_i]^2} . \qquad (4.17)$$

This number is directly proportional to χ^2, and gives a sum-of-squares measure of the fractional misfit. As long as the measurement uncertainty is not a significant fraction of the measurement itself, any fit with R-factor greater than a few percent is not a good fit. General acceptable values are R≤0.02, $\chi_v^2 > 10$. Such fits are considered as good fits as the theory and data agree within a percent.

4.3.2 Measurement Uncertainty

Estimating systematic errors in $\tilde{\chi}(r)$ is the main difficulty in error analysis in FEFFIT. ε contains both random fluctuations and systematic error in the data.

4.3.2.1 Systematic Uncertainty

Systematic errors are dominated by (i) leakage of imperfect background into the first few shells, (ii) systematic errors in the measurement of μ(E). It is possible to estimate the size of the systematic errors by trying different reasonable background removals. Analyzing different data scans taken under different experimental conditions could give an estimate of (ii). A third source of systematic errors in the fit comes from the FEFF calculation itself. Such errors are important as they contribute to the small amount of misfit expected in a good fit.

4.3.2.2 Statistical uncertainty

The scale of ε depends on the Fourier Transform parameters used. Assuming that the noise is dominated by random fluctuations, the measurement uncertainty in k-space data is linearly related to ε_k, i.e. the unfiltered fluctuations in χ(k) data:

$$\varepsilon_k = \varepsilon_R \sqrt{\frac{\pi(2\omega+1)}{\delta k(k_{\max}^{2\omega+1} - k_{\min}^{2\omega+1})}}, \tag{4.18}$$

where ω is the k-weighting.

4.3.3 Error estimation for the variables

FEFFIT estimates the uncertainties in the variables once the best-fit values of the variables are found. The uncertainties are estimated using standard χ^2 minimization technique. The goal of the fit is to minimize χ^2 in each of its N_{var} dimensions. Generally the Levenberg-Marquardt non-linear least squares fitting algorithm[91] is used (For details about the algorithm, see Appendix A). The first and second derivatives of χ^2 are found *wrt* each of the variables. These are used for finding the next estimate of the best variables, and are found to be useful for estimating the uncertainties in the variables after the best fit has been found.

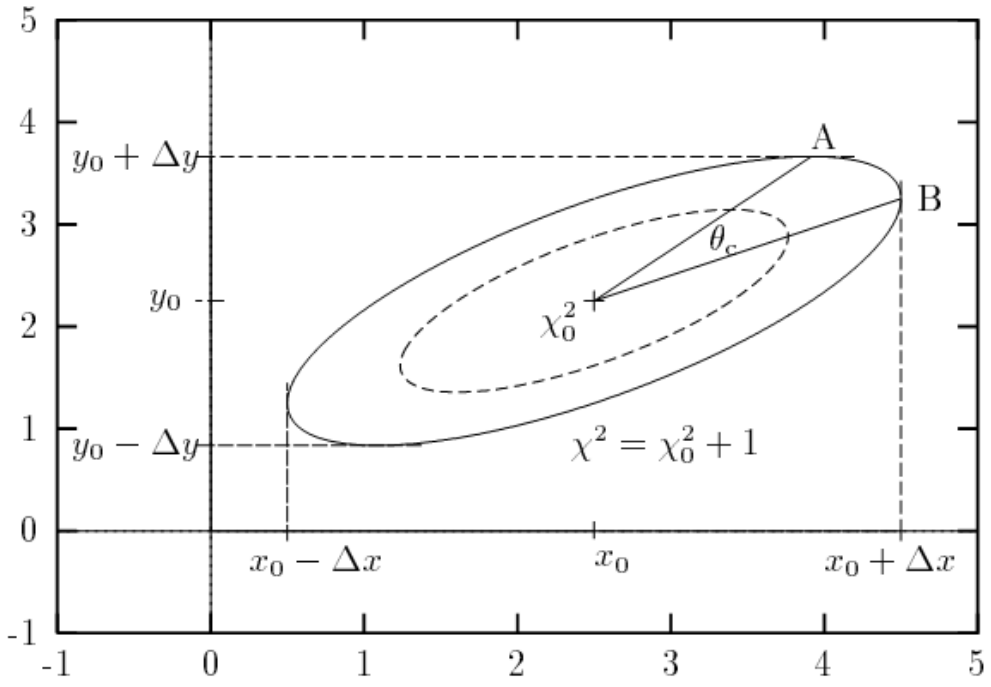

Fig. 4.3: Contour map of χ^2 as a function of two variables x and y

At the best fit solution, χ^2 is roughly parabolic in each of its N_{var} dimensions. The curvature of the χ^2 is given by the ($N_{\text{var}} \times N_{\text{var}}$) matrix of the second derivatives around the solution. Fig. 4.3 shows a contour plot of the χ^2 surface for a two variable problem. The surface is elliptical since the errors in the variables are not equal. At the solution, the variables x and y have values x_0 and y_0 respectively and $\chi^2 = \chi_0^2$. As x or y move away from their best fit solution, χ^2 increases. For normally distributed uncertainties, the contour of χ^2 is elliptical for two dimensions. The uncertainty in the value of a variable is the amount by which it can be increased and still have χ^2 below some limit. FEFFIT uses the $\chi_0^2 + 1$ criterion (same as the one commonly used for randomly distributed errors). The uncertainties in x and y are Δx and Δy respectively, and these ensure that χ^2 is increased by 1 from its best value.

While evaluating the uncertainty in a variable, all the other variables are allowed to vary, so that the correlations between variables can be taken in to account. The correlation is a measure of how much the best-fit value of one of the variables changes in response to changing another variable away from its best-fit value. In Fig. 4.3, the correlation between the variables x and y is equivalent to the projection of x on y, i.e. $\cos\theta_c$. If x and y had been completely correlated, the ellipse in Fig. 4.3 would have had its major and minor axes parallel to the x and y axes. If we ignore the correlations and hold y constant, then the uncertainty in x would be estimated to be $\Delta x'$. This is considerably smaller than Δx, and is a worse estimate of the uncertainty in x

Algebraically, the uncertainties in the variables are given by the inverse of the curvature matrix (the matrix of the second derivatives), called the correlation matrix. The uncertainties in the variables are the square roots of the diagonal terms, and the correlations between pairs of variables are given by the off-diagonal terms of this matrix.

The uncertainties of the variables and the correlations between variables are estimated at the best fit condition. Needless to say, each analysis requires its own assessment of uncertainties, including estimation of the relative importance of systematic and random errors; therefore it is not possible to state universal values for these uncertainties. Since the fits involve simultaneously varying several variables, it is possible that a local minimum may be found instead of a global minimum. Another possibility is that the initial structural model considered is very far away from the correct model, and the model does not have the freedom required to allow the necessary variation for reaching the correct model. Yet another possibility is that there exists some correlation between the variables employed and with the available data, it is not possible to unambiguously determine these correlated variables.

Before stating conclusive results, several models must be considered and accounted for by applying appropriate error bars. One particularly useful way to reduce error bars is by simultaneously fitting different datasets. Sometimes, choice of appropriate k-weight for the data is also important for obtaining low error bars.

Therefore, XAFS data analysis is largely non-trivial and rigorous, and arriving at the correct conclusion significantly depends on the ability of the analyst to judge the quality of fit.