Propagation of statistical errors across EXAFS extraction and Fourier filtering

Emmanuel Curis[a]* and Simone Bénazeth[a,b]

[a]LURE, Bâtiment 209D, Centre Universitaire Paris-Sud, 91405 Orsay CEDEX, France, and [b]Laboratoire de Biomathématique, Faculté de Pharmacie, Université René Descartes, 4 Rue de l'Observatoire, 75005 Paris, France. E-mail: curis@lure.u-psud.fr

A statistical model for estimating the error bars on the EXAFS signal is developed. The average can be calculated at the beginning of the analysis, with the estimation of the error bars made directly on the individual measurements; the error bars are then propagated across the EXAFS extraction. Furthermore, this model also allows the estimation of the error bars on the Fourier transformation of the EXAFS signal, and on the filtered EXAFS spectrum after the inverse Fourier transform.

1. Introduction

A necessary step in EXAFS analysis is the estimation of the experimental error bars for the considered spectrum, as without these error bars it is impossible to estimate the errors on the structural parameters fitted by EXAFS. The experimental errors may originate either from statistical errors ('noise') or from systematic errors. In this paper we only deal with the statistical errors.

In order to estimate these error bars, the most reliable approach is the statistical approach, which requires the same spectrum to be recorded many times. Nevertheless, if only one spectrum is available, various methods have been proposed to provide an estimation of an average error bar, thus making the assumption that the error is constant with energy (Newville *et al.*, 1999); all these methods do not use any statistical tool. In addition, if these methods are applied to an averaged spectrum, obtained by many experiments, they provide an estimation of the error bars of the experimental *averaged* distribution and not of the experimental individual distribution, these two values being related by a factor $N^{1/2}$, where N is the number of averaged spectra. In the present paper we only deal with spectral recorded many times.

If many spectra are available, the proposed method is to extract the EXAFS signal for each spectrum and then to calculate the averages. While this averaging is performed, the error bars are estimated for each point and then averaged to estimate a global error bar for the spectrum, assumed to be constant with k (Vlaic *et al.*, 1999).

In this paper we use standard statistical tools to improve this error treatment; the model we develop allows us to calculate the averages at the very first step of the treatment (which helps in the EXAFS extraction, as the signal is less noisy) without losing any statistical information: we keep the error bars on each point at any step of the treatment.

Furthermore, this model also allows us to estimate the error bars for the Fourier transform itself, which is very interesting for fitting in *R*-space.

All the statistical tools presented in this paper have been implemented in our new EXAFS analysis software, *LASE* (Logiciel d'analyse des spectres EXAFS). It runs on any Unix/X-Window computer and includes the classical tools for EXAFS analysis, adapted to our model of the error propagation.

2. Statistical model

We only consider spectra recorded for a finite number of points, *i.e.* by a discrete method (for instance by a step-by-step method). An experimental spectrum is then a set of n points $(E_i, \mu_i)_{i=1...n}$, where E_i is the energy for the ith point and μ_i is the corresponding absorption coefficient.

We choose a very general model, the basic model used in statistics: the ith experimental point is modelled by a random variable X_i; we assume that the esperance $E(X_i)$ and the variance $V(X_i)$ exist. In that case each experimental measurement at energy E_i is a realization of the random variable X_i. With these hypotheses, the great number law shows that

$$m = (\mu_1 + \mu_2 + \ldots + \mu_n)/n$$

almost certainly converges towards $E(X_i)$ and

$$\sigma^2 = \left[(\mu_1 - m)^2 + (\mu_2 - m)^2 + \ldots + (\mu_n - m)^2\right]/(n-1)$$

almost certainly converges towards $V(X_i)$.

Furthermore, we can consider the set of random variables $(X_i)_{i=1...n}$ as an n-dimension real random vector $\mathbf{X}$, characterized by its mean vector $\mathbf{E}$ and its covariance matrix Σ. Each experimental spectrum is then a realization of this random vector; the estimated mean vector is

equivalent to the usual experimental spectrum after averaging, and all the statistical information is included in the estimated covariance matrix. In that model the following treatment of the experimental data to extract the EXAFS signal is applied to the random vector **X** instead of applying it only to the averaged experimental spectrum **E**, as is the usual practice. The propagation of the statistical errors is then accomplished by transforming the covariance matrix Σ according to the applied transformation. Hence, we use the information from all the individual measurements, not just the average experiment.

In order to simplify our model we also suppose that the experimental points are statistically uncorrelated, which means that $\mathrm{Cov}(X_i, X_j) = 0$ if $i \neq j$. In that case the covariance matrix Σ is strictly diagonal. This hypothesis is not required, since all that follows is still correct with a non-diagonal covariance matrix, but it makes computations quicker and it is physically meaningful.

3. Propagation across EXAFS extraction

Practically, the value of interest is not μ_i but the EXAFS oscillations $\chi(k_i)$. The relation between these two values is

$$\chi(k_i) = [\mu(k_i) - \mu_0(k_i) - \mu_m(k_i)]/\mu_0(k_i),$$

where $\mu_m(k_i)$ is the absorption coefficient of all atoms except the absorbing atom and $\mu_0(k_i)$ is the atomic absorption of the isolated absorber. In our model all these quantities are random variables; since $\mu_m(k_i)$ and $\mu_0(k_i)$ are estimated from the experimental data, these random variables are *a priori* correlated with $\mu(k_i)$. In addition, the relation defining $\chi(k_i)$ is not linear for $\mu_0(k_i)$. Owing to these two facts there is no simple relation between $\mathrm{E}[\chi(k_i)]$ and $\mathrm{E}[\mu(k_i)]$, nor between $\mathrm{V}[\chi(k_i)]$ and $\mathrm{V}[\mu(k_i)]$.

As a first approximation we considered that the random variables $\mu_m(k_i)$ and $\mu_0(k_i)$ are almost certainly constant. In that case the relation becomes very simple,

$$\mathrm{E}\big[\chi(k_i)\big] = \big\{\mathrm{E}[\mu(k_i)] - \mu_0(k_i) - \mu_m(k_i)\big\}/\mu_0(k_i),$$

$$\mathrm{V}\big[\chi(k_i)\big] = \mathrm{V}\big[\mu(k_i)\big]/\mu_0^2(k_i).$$

To check this approximation we compared these results with those obtained by averaging directly on the $\chi(k_i)$ values. Fig. 1 shows that the values for $\chi(k_i)$ obtained by the two methods cannot be distinguished. In Fig. 2 we compare the error bars obtained by the two methods, and we notice that they are very close, except at very low k. These results confirm the hypothesis that $\mu_m(k_i)$ and $\mu_0(k_i)$ are almost certainly constant. We interpret the difference at low k as a consequence of the uncertainties on the monochromator displacements, which are functions of the Bragg angle θ and not of k nor E. This leads to important variations in the measured intensities in the edge region, where μ varies quickly with E.

Except for this very low k region, we observed that the error bars are generally almost constant with k, as is usually assumed (Newville *et al.*, 1999; Vlaic *et al.*, 1999), but our model still works when the error bars are not constant with k.

We also observed that, by increasing the number of averaged spectra, the error bars are not really modified: the average is better estimated (according to the central limit theorem) and the constant behaviour of the error bars is more visible.

4. Propagation across Fourier filtering

4.1. *Error propagation across integration*

A classical way of estimating $I = \int_a^b f(x)\,\mathrm{d}x$, if only n points $[x_i, y_i = f(x_i)]_{i=1\ldots n}$ are known, is the trapeze method,

$$I = \tfrac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)(y_{i+1} + y_i)$$

$$= (x_1 - x_0)y_0/2 + \tfrac{1}{2}\sum_{i=2}^{n-1}(x_{i+1} - x_{i-1})y_i + (x_n - x_{n-1})y_n/2.$$

As we can see, I appears as a linear combination of y_i, whose coefficients are functions of x_i. Hence, it is possible to write $I = A\mathbf{Y}$ where $\mathbf{Y} = (y_i)_i$ is the column vector of the values and A is a $(1 \times n)$ matrix. Hence, the numerical integration is seen as a linear transformation.

If we now consider a random vector $\mathbf{X} = (X_i)_i$ with a covariance matrix Σ, we can define its 'integral', which is a
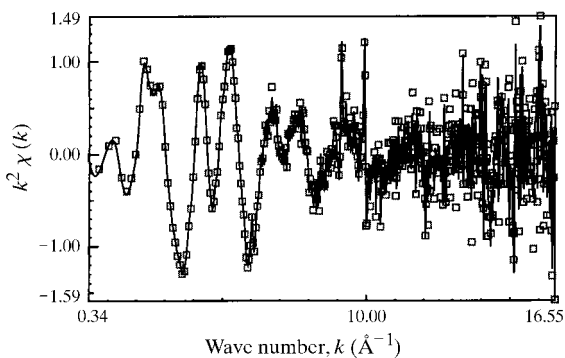

Figure 1
Comparison of the EXAFS signal extracted by the two methods. Continuous line: extractions, then average (method 1); squares: average, then extraction (method 2). 14 spectra are averaged.
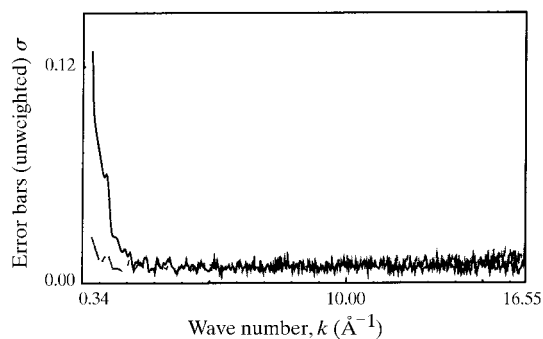

Figure 2
Comparison of the error bars on the EXAFS signal extracted by the two methods. Dashed line: method 1; continuous line: method 2.

random variable I, by $I = A\mathbf{X}$. The probability theory then shows that $E(I) = AE(\mathbf{X})$ and $V(I) = A\Sigma^t A$; we can propagate errors across integration.

With the help of the random function theory it may be proved that, when the number of points increases (hence $\mathbf{X}$ converges towards a random function f), the 'integral' defined above converges towards the real integral of f (Brard, 1966/1967).

4.2. *Application to the Fourier transformation*

By convention, in EXAFS the Fourier transform of the signal is defined by

$$\check{\chi}(R) = (2/\pi)^{1/2} \int_{-\infty}^{+\infty} \chi(k) \exp(-2ikR)\,\mathrm{d}k.$$

Since we only know $\chi(k)$ for a finite number of values, we use the trapeze method to evaluate the Fourier transform for the points $(R_j)_{j=1\ldots N}$. Use of the classical FFT algorithm (which corresponds to an estimation of the Fourier transform by the rectangle method) is not possible, since it requires a smoothing of the data to obtain a constant k-step, and such a smoothing is not compatible with a statistical use of the data. According to the model we developed above, if we model the experimental spectrum
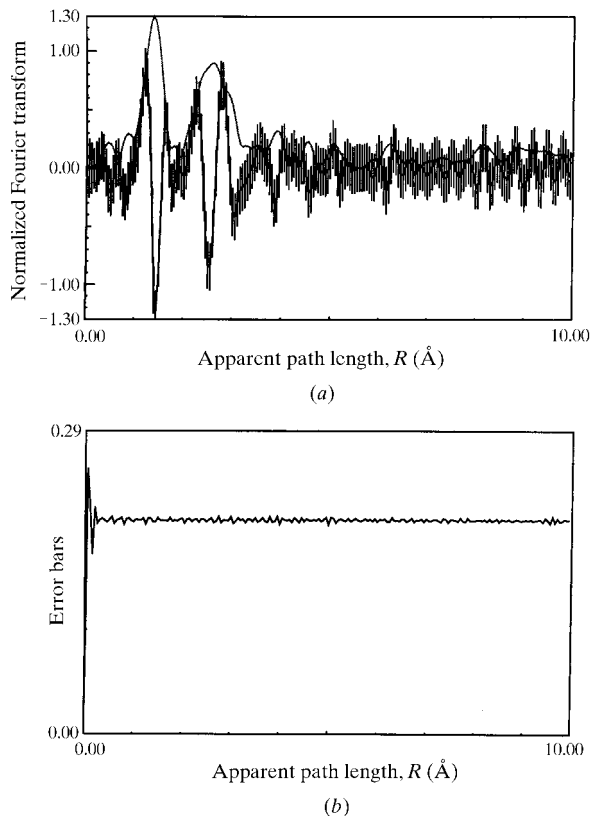

1.30
1.00
0.00
−1.00
−1.30

Normalized Fourier transform

0.00   10.00

Apparent path length, R (Å)

(*a*)

0.29

Error bars

0.00

0.00   10.00

Apparent path length, R (Å)

(*b*)

Figure 3
(*a*) The Fourier transform (FT) with its error bars on each point, estimated directly from the error bars on the experimental EXAFS signal. Only the error bars on the imaginary part of the FT are shown; the modulus of FT is also drawn. (*b*) The error bars on the real part of this FT.

by an n-dimensional vector χ, for each point R_j we can write $\Re[\check{\chi}(R_j)] = A_{2j}\chi$ and $\Im[\check{\chi}(R_j)] = A_{2j+1}\,\chi$, where

$$A_{2j} = (2\pi)^{-1/2}\big[(k_1 - k_0)\cos(2k_0 R_j)$$
$$\ldots (k_{i+1} - k_{i-1})\cos(2k_i R_j)$$
$$\ldots (k_n - k_{n-1})\cos(2k_n R_j)\big]$$

and

$$A_{2j+1} = (2\pi)^{1/2}\big[(k_1 - k_0)\sin(2k_0 R_j)$$
$$\ldots (k_{i+1} - k_{i-1})\sin(2k_i R_j)$$
$$\ldots (k_n - k_{n-1})\sin(2k_n R_j)\big]$$

are the matrices associated with the Fourier transformation.

If we now consider the estimated Fourier transform as a vector $\check{\chi} = \{\Re[\check{\chi}(R_j)], \Im[\check{\chi}(R_j)]\}_{j=1\ldots N}$, we can define a global matrix A for the complete Fourier transformation, and we have $\check{\chi} = A\chi$.

In our model, χ is a random vector with a covariance matrix Σ. Hence, $\check{\chi}$ is also a random vector whose covariance matrix is $\Sigma' = A\Sigma^t A$. The diagonal part of Σ' gives the error bars for each Fourier transform point (for their real and imaginary parts).

It is possible to develop exactly the same model for the inverse Fourier transformation: if we call B the associated matrix, χ_f the filtered EXAFS spectrum and Σ_f its covariance matrix, then we obtain $\chi_f = B\check{\chi} = BA\chi$ and $\Sigma_f = B\Sigma'^t B = AB\Sigma^t(AB)$. We then obtain the error bars for each point of the filtered spectrum, and all the correlation information between its points.

4.3. *Inclusion of window and filter effects*

To limit the distortions of the Fourier transform, one commonly multiplies the experimental EXAFS by a window function $w(k)$. With our model, this window function is modelled by a diagonal matrix W with $W_{ii} = w(k_i)$. We can also define a filter matrix F, related to the filter function used in R-space to isolate one or more peaks of the Fourier transform; this matrix is also a diagonal matrix.

With these two effects included, the global transformation equations are $\chi_f = W^{-1}BFAW\chi$ for the average and $\Sigma_f = W^{-1}BFAW\Sigma^t(W^{-1}BFAW)$ for the covariance matrix.

4.4. *Results*

4.4.1. *Direct Fourier transformation.* We may note (Fig. 3) that the error bars in R-space are almost constant (whatever k-weighting is used), with a residual oscillating part, especially for low R values.

We may also note that the variation of the error bars for different filtering methods follows the expected trend: if we increase the k-weighting of the EXAFS signal, the error bars on the Fourier transform are increased. The error bars on the Fourier transform increase in two ways depending on the choice of the window: they increase either with the rectangular tendency of the window, or with its wideness. The most important effect is the wideness of the window,

that is the k-range of the signal kept for the Fourier transform. Hence, when the EXAFS signal becomes too noisy (*i.e.* if one goes too far in k so, as the signal is attenuated whereas the noise is constant, one just adds noise to the Fourier transform), it leads to more noisy Fourier transforms. The kind of window used for the analysis seems to have a much lower effect on the noise of the Fourier transform. For instance, if we compare a Hanning window and a Kaiser–Bessel window over the same k-range, the Hanning window gives some higher error bars, but the Fourier transform intensity also increases, which gives a similar signal-to-noise ratio for the two windows.

We may note that an estimation of the error based on the high R values of the average Fourier transform (Newville *et al.*, 1999) gives a smaller value than the error bars in R-space. This is not surprising, because the first method estimates the error bar on the 'average' distribution and not on the experimental individual distribution. As expected by the theory, we find a factor of approximately $N^{1/2}$ between the two values, where N is the number of averaged spectra.
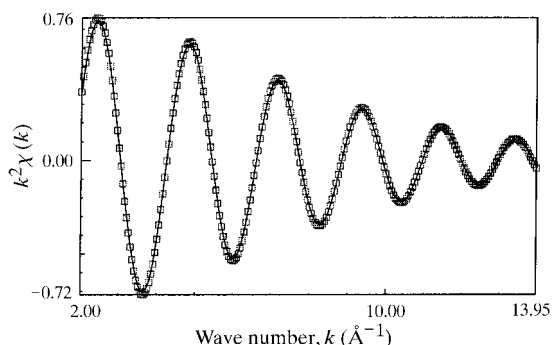

Figure 4
Comparison of the filtered EXAFS signal obtained by the two methods. Continuous line: EXAFS obtained by extracting the signal, Fourier filtering and then averaging (method 1). Squares: EXAFS obtained by averaging, extracting the signal and then Fourier filtering (method 2). The two spectra are similar. The signal corresponds to the first peak of the Fourier transform shown in Fig. 3.
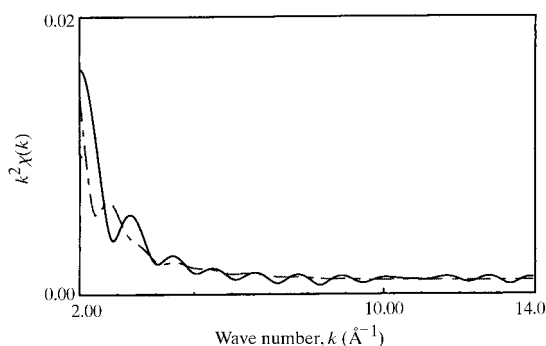
4.4.2. Inverse Fourier transformation. To check the results of our method for propagating errors across Fourier filtering, we compared the results obtained by this method with those obtained by the other sequence: extracting the EXAFS signal, Fourier filtering and then averaging. Fig. 4 provides graphical evidence that the filtered EXAFS spectra obtained by the two methods are identical.

Fig. 5 shows that the error bars obtained by both methods are roughly of the same magnitude, especially for high k values. The differences between the two methods may arise from inaccuracies in Fourier transformation evaluation for each individual spectrum, leading to larger error bars for that case. One may also note that, after filtering, the error bars are no longer constant with k, although they were before filtering. On all the spectra that we studied, we observed some error bars which varied with k after filtering; this variation is not similar from one spectrum to another and seems to be, for an important part, related to the window used for the Fourier transformation.

In Fig. 6 we observe that the filtering operation really has an effect on the uncertainties: the error bars after filtering are much smaller than the error bars before filtering. One notices the same results for the filter effect and for the window effect in the direct transformation (which is not surprising, since the mathematical functions are the same): given a Fourier transform, the filtered spectrum is less noisy if the R range is smaller. However, since the R range is more or less imposed by the peaks one wants to keep, the main effect arises from the Fourier transform itself: the more noisy the Fourier transform, the more noisy the filtered spectrum. Hence, the choice of the k range for the window of the Fourier transformation seems to be the most important step to achieving good filtered spectra. It seems that there is no simple relation between the error before and after filtering; in any case, even if such an expression exists, making use of it would mean losing important information, since it would not give the covariance matrix nor the k-evolution of the error bars after filtering.

We also noticed, by studying the covariance matrix after filtering, that there are no strictly independent points after
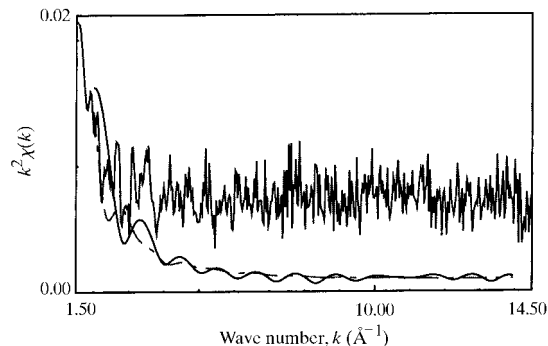

Figure 5
Comparison of the error bars on the filtered EXAFS signal obtained by the two methods. Continuous line: method 1. Dashed line: method 2. The error bars are similar; they are not constant with k.

Figure 6
Comparison of the error bars on the EXAFS signal after and before the Fourier filtration. Continuous line: method 1. Dashed line: method 2. Thick line: error bars before filtration (as in Fig. 2). The error bars are a lot smaller after filtration.

filtering (in the statistical meaning); nevertheless, the correlation coefficient is usually between 0.1 and 0.3, which indicates a weak linear correlation. If we assume that this can be neglected, this means it is possible to fit it with all 'experimental' points even after filtering; if not, this means a new statistical estimator needs to be developed which accounts for the covariance matrix.

5. In the case of a Gaussian distribution

In all the previous sections we did not make any special assumption about the experimental distribution function of μ_i, except that it had an esperance and a variance.

If we assume now that this distribution is Gaussian, we can go further: the initial experimental vector is a Gaussian vector. This means that every linear transformation of this vector leads to another Gaussian vector; in our model, all the transformations applied to the experimental data are linear, so at each step the result is a Gaussian vector. In particular, the crude EXAFS signal, its Fourier transform and the filtered EXAFS signal are Gaussian vectors. In addition, with our model the crude EXAFS signal has a diagonal covariance matrix, so the maximum-likelihood estimator is the classical estimator used for fitting,

$$\chi^2 = \sum_i \left\{ \left[\chi_{\exp}(k_i) - \chi_{\mathrm{th}}(k_i) \right] / \sigma_i \right\}^2 .$$

For the Fourier transform and the filtered signal, this estimator may still be used, but it is not the maximum-likelihood estimator, since the covariance matrix is not diagonal for these quantities.

6. Conclusion

We have developed a simple statistical model for EXAFS analysis that allows all statistical information extracted from the experimental data to be kept, and propagated across all steps of the EXAFS analysis. In particular, this model allows the averages to be calculated at the very beginning of the treatment – which permits a better EXAFS extraction since the less noisy average spectrum eases the background removal and saves time in the case where a great number of spectra are to be analysed – and keeps all the statistical information needed for the fitting step of the analysis procedure.

This model allowed us to confirm the usual results used in the analysis: there is no difference between 'average, then extraction' and 'extraction, then average', except that, previously, only the second method lead to statistical information. It also confirmed that, as far as we could see, the error bars are constant with k before filtering.

Use of this model also provides answers to the problem of the propagation of the error across Fourier filtering, and especially the conservation of the distribution law and the estimation of the error in R-space, which is necessary for some statistical treatments when fitting in R-space.

The next step of our work is to check the Gaussian distribution of the experimental points – is that model a theoretically plausible model, knowing that the measured quantities follow some Poissonian laws? – and the use of this model to estimate the distribution functions of the parameters fitted during the analysis: in which conditions is it valid to use the classical results, proved only for the linear case, knowing that the EXAFS equation is not linear?

References

Brard, R. (1966/1967). In *Cours de Calcul des Probabilités*, Part 2, *Fonctions Aléatoires*.
Newville, M., Boyanov, B. I. & Sayers, D. E. (1999). *J. Synchrotron Rad.* **6**, 264–265.
Vlaic, G., Andreatta, D., Cepparo, A., Colavita, P. E., Fonda, E. & Michalowicz, A. (1999). *J. Synchrotron Rad.* **6**, 225–227.