A new model for statistical error analysis in XAS: about the distribution function of the absorption coefficient

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*

Under some general hypothesis, this paper proposes a theoretical model, showing that a gaussian distribution is generally a good approximation of the experimental distribution of the absorption coefficient. This result is confirmed experimentally by usage of appropriate statistical tests.

Keywords: EXAFS; **statistical analysis**; **statistical errors**; **distribution function**; **absorption coefficient.**

1. Introduction

In the statistical analysis of EXAFS results, as for any statistical study, the estimation of numerical quantities from experimental results (for instance, the distance between the absorbing atom and the atoms of its first coordination shell) relies on the hypothesis that the different experimental values are derived from random variables; the knowledge of the distribution function of these random variables is required to construct some reliable estimators.

Usually, it is assumed in EXAFS analysis that these random variables are normal (gaussian) (see, for instance, Vlaic *et al.*, 1999; Filiponi, 1995; Krappe & Rossner, 1999). With this hypothesis, the natural estimator to use is then the least-squares estimator, either weighted or unweighted by the error bars (the former being preferable, as it corresponds to the maximum likelihood estimator). But, as far as we could see, no experimental verification or theoretical justification of this hypothesis has yet been proposed.

We already showed that, if the absorption coefficient follows a normal distribution, then any subsequent quantity used in EXAFS analysis can also be modelled by a normal distributed random variable (Curis & Bénazeth, 2000). Hence, in this paper, we only verify that, both theoretically and experimentally, this hypothesis of a normal distribution of the absorption coefficient μ is acceptable.

2. Theoretical model

As biological samples measurements are performed by transmission or by fluorescence technics, we developed our model only for that two methods – but it may extend for other methods. In both cases, the quotient of two experimental values is used, so we will begin with a generic model for this quotient. After that, we will introduce the model for fluorescence experiments, and then for transmission experiments.

2.1. Preamble: the exact law of the quotient of two independant normal random variables

Let X_1 and X_2 be two independant random variables, following a gaussian law, such that $E(X_1) = \mu_1, V(X_1) = \sigma_1^2$ and $E(X_2) = \mu_2, V(X_2) = \sigma_2^2$, where $E(X)$ is the expectation of X and $V(X$ its variance. We now consider the random variable Z defined by $Z = \frac{X}{Y}$. The problem is to determine the law of Z. Since it is possible for the denominator to cancel, it is necessary to use a slightly extended concept of real random variables, which take their values not only in R but in $\overline{R} = R \cup \{-\infty, +\infty\}$.

As presented in Brard (1966), in this conception the classical results extend in a very simple, intuitive way; the main difference is that, if F_X is the cumulated distribution function of the random variable X, $\lim_{-\infty} F_X = p(X = -\infty)$ and $\lim_{+\infty} F_X = 1 - p(X = +\infty)$. In the case of the quotient Z considered, $p(Z = -\infty) = p(Z = +\infty) = p(Y = 0) = 0$, so in practice there is no difference with the classical case.

A classical result (Saporta, 1990) then gives the distribution function f_Z of Z knowing the distribution function $g(x, y)$ of (X, Y) (so that $g(zy, y)$ quantifies the probability that $(X, Y) = (zy, y)$, that is $p(Z = \frac{X}{Y} = z)$).

$$f_Z(z) = \int_{-\infty}^{+\infty} |y| g(zy, y)\, dy$$

If X and Y are independant, one can write $g(zt, t) = f_X(zt)\, f_Y(t)$. In the case of a normal distribution, the definition of f_Z then becomes:

$$f_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2 z^2} \int_{-\infty}^{+\infty} |t| \exp\left(-\frac{(t-\mu_1)^2}{2\sigma_1^2} - \frac{(t - z\mu_2)^2}{2z^2\sigma_2^2}\right) dt$$

This integral may be exactly computed, giving the exact distribution function of the quotient Z; the resulting expression being quite long and complex, we do not present it here. But the knowledge of the expression allows a comparison between the exact law and the law estimated from the error propagation theorem, which predicts that Z may be modelled by a random variable following a normal law of mean $\frac{\mu_1}{\mu_2}$ and of variance $\frac{\sigma_1^2}{\mu_2^2} + \frac{\mu_1^2\sigma_2^2}{\mu_2^4}$. The main result is that the variable Z does not have any mean nor variance – this result is particularly evident when $\mu_1 = \mu_2 = 0$, since in that case Z follows a Cauchy law. Some graphical comparisons (see fig. 1) show that the approximation is valid only if the probability that both X and Y change of sign is negligible – that is, for instance, that $p(Y < 0)$ is negligible if $E(Y) = \mu_2 > 0$, which can be converted into a condition on the quotients $\frac{|\mu_1|}{\sigma_1}$ and $\frac{|\mu_2|}{\sigma_2}$ that must be both greater than 3, for classical usage. The exact value will of course depend on the accuracy level desired; the presented condition corresponds to the classical 3σ interval that contains 99,6% of the values and that must not contain 0. If one can estimate the values of μ and σ, this result can be used as a rule-of-thumb to determine if the experimental quotient follows, or not, a gaussian distribution.

2.2. Distribution for fluorescence experiments

In fluorescence experiments, the absorption coefficient μ is derived from the intensity before the sample I_0 and the intensity of the fluoresced beam I_f by the relation $\mu = \frac{I_f}{I_0}$.

The general model for the statistical distribution of the number of photons detected is a Poisson law, with very large hypothesis. With usual intensities, the average number of photons detected is very high – this means that the Poisson law may be accurately approximated by a gaussian law. In that case, we are in the scope of the model developed in the preamble (assuming that I_f and I_0

are independant random variables – note that this hypothesis is not in contradiction with the fact that $E(I_f) = f(E(I_0))$, where $E(X)$ denotes the expectation of X).

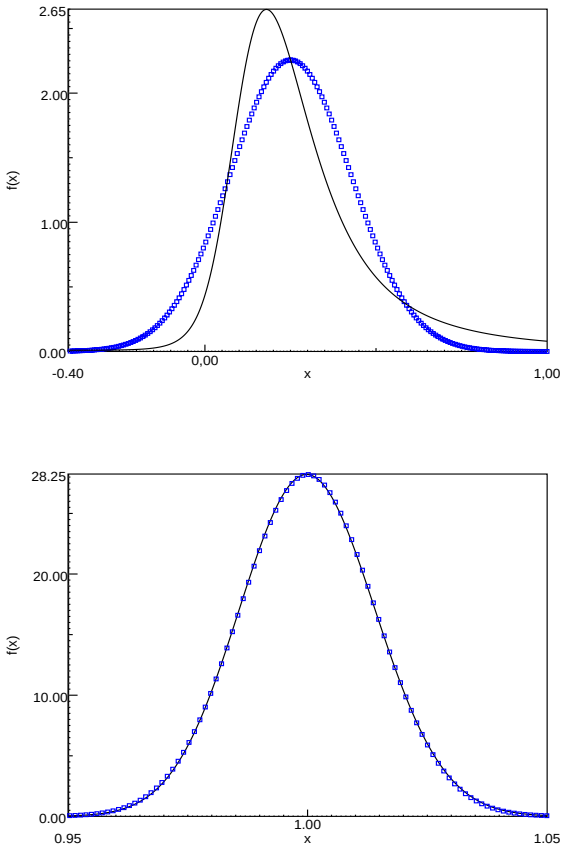

Figure 1
Comparison of the exact distribution function of a quotient of two independant gaussian random variables (continuous line) and of the distribution function of the corresponding gaussian estimated by the error propagation theorem (squares). Upper: numerator can take positive or negative values with significant probabilities. The error propagation model does not apply. Lower: numerator and denominator have negligible probabilities to take negative values, the error propagation model applies.

Also because of the high counting rate, the probability for I_f and I_0 to be negative (or, in the Poisson law model, to vanish) is negligible. Consequently, the results presented above are applicable: one can approximate safely the distribution of μ by a gaussian distribution.

Note that this model does not apply correctly for the preedge region, in which the counting rate for I_f is very small (ideally null): the approximation of the Poisson law by a gaussian law is not correct, and even if it were correct, the probability to vanish is not negligible: the approximation of the real distribution of the quotient by a gaussian law is not correct. Anyway, the preedge region is useless in EXAFS analysis, so this is not really a problem for daily analysis.

2.3. Distribution for transmission experiments

In transmission experiments, the absorption coefficient μ is derived from the intensity before the sample I_0 and the intensity after the sample I_1 by the relation $\mu = \ln \frac{I_0}{I_1}$.

As for fluorescence, the high counting rate in detectors allows the modelization of the Poisson law by a gaussian distribution, with a negligible probability to be non-positive. In that case, the quotient $\frac{I_0}{I_1}$ can be approximated by a gaussian law, as stated previously. The problem is now "what is the exact distribution of $Y = \ln X$ when X is a random variable following a gaussian law?"

The fact that X can take some negative values makes the correct definition of Y difficult. One can imagine three ways to handle that problem:

– using $Y = \ln |Y|$;

– using $Y = 0$ (arbitrarly) if $X \leq 0$;

– using the generalized real random variables with $Y = -\infty$ if $X \leq 0$.

As the introduced gaussian variables are models for non-negative Poisson laws, and since the probability of counting no photon is negligible, one can expect the three methods to give the same results. For mathematical convenience, we choosed the third model to do computations. In that case, the density function of Y is given by

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(y - \frac{(e^y - m)^2}{2\sigma^2}\right)$$

where m is the expectation of X and σ^2 its variance.

As in the quotient case, the comparison of the exact distribution above and of the gaussian distribution obtained by the error propagation theorem shows that the approximation by a gaussian law is usable only if the probability for X to be negative is negligible. Since it is the case in X-rays absorption experiments, it is theoretically funded to use the gaussian distribution as a model for the probability distribution of the absorption coefficient μ.

3. Experimental verification

The theoretical model presented in the previous section gives hints that a model of gaussian distribution for the experimental values used in XAS is reasonable. Anyway, many experimental effects may arise that make questionable the above hypotheses: if the Poisson law for the photons couting rate is generally accepted, the electronic treatments of the signal after detection may transform the distribution. Even if classical statistical theorems, like the central limit theorem, state that one can expect gaussian laws for the result of such processes (and, in that case, the above model is still correct), the complexity of these effects conducted us to experimentally test the normality of the distribution of the absorption coefficient.

Many statistical tests exist to check the hypothesis of the normal distribution for a set of experimental values (Csorgo *et al.*, 1973). We chose, to perform our tests, the Kolmogorov-Smirnov test, with the tabulated values computed to take into account the fact that both mean and variance are unknown (Saporta, 1990; Lilliefors, 1967), since it is one of the most powerful tests.

We used two kinds of experiments to check the normality of the experimental values. First, we fixed the energy of the incident beam and recorded a set of about 100 values of the absorption coefficient. Second, to be closer of the real experimental conditions, we recorded a set of 98 spectra of the same sample (a zinc complex, studied at K-edge on the D44 beamline of DCI, LURE (France)), in the same conditions. To limit the time of acquisition, we just

recorded the XANES part of the spectrum. We then selected all points at the same energy and applied to them the Kolmogorov-Smirnov test.

In both cases, the result of the test is that the distribution is gaussian to the 5% signification level.

4. What about systematic errors?

To evaluate the effect of systematic errors on the statistical model, we can consider a simple model of the measure. Let $\mu_{\exp}(E,t)$ be the experimental measured value and $\mu(E)$ the real value. One can always write, for a fixed energy value, that $\mu_{\exp}(E,t) = \mu(E) + \delta\mu(E,t) + \varepsilon(E,t)$, where $\delta\mu(E,t)$ accounts for the systematic effects and $\varepsilon(E,t)$ is a stationary random process that accounts for statistical errors. After experiments, we know a set of values of $\mu_{\exp}(E,t)$ for certain values of t. One can imagine three cases, depending on the behaviour of $\delta\mu(E,t)$ with time. First, $\delta\mu(E,t)$ does not depend of time. In that case, all the statistical model developed above and in Curis & Bénazeth (2000) applies – but the average values are not the expected value, there is a bias. Second, $\delta\mu(E,t)$ varies quickly with time, but does not vary in average. This case is similar to the previous one. Third, $\delta\mu(E,t)$ varies with time. In that case, it is not possible to do any statistics without correction for that effect, because $\mu_{\exp}(E,t)$ is no longer a stationary random process, and so average is not defined: no statistical model can apply.

5. Conclusion

Both the theoretical model presented above and the usage of classical statistical tests on experimental data justify the model of a gaussian distribution for the absorption coefficient at a given, fixed, energy. With the error propagation model we introduced in Curis & Bénazeth (2000), it follows that each point of the experimental spectrum at any stage of the treatment follows a normal distribution (crude EXAFS spectrum, filtered EXAFS spectrum, Fourier transform real or imaginary part – but *not* the modulus).

This result validates all the error analysis procedures that relies, more or less implicitly, on this hypothesis - and, in particular, the use of estimators like the least-squares, as maximum likelihood estimators (the exact kind of estimator depending on the statistical dependances between the experimental points), to estimate the parameters' values by fitting.

It also provides the distribution model indispensable to develop tools, like the Monte-Carlo methods, to estimate the uncertainties on fitted parameters with no other hypothesis than the experimental distribution function. The use of the Monte-Carlo methods needs much less hypotheses than the usual methods, so we are currently working on this method to check its suitability in EXAFS. Since this work is still in development, we will present its results elsewhere.

Acknowledgements

We want to thank Bruno Blanchet for his help in the mathematical part of this work and Alain Michalowicz for the discussions about statistics in EXAFS.

References

Brard, R. (1966). *Calcul des probabilités. Première partie : variables aléatoires*. École polytechnique.
Csorgo, M., Seshadri, V. & Yalovsky, M. (1973). *J. R. Stat. Soc. Ser. B*, **35**(3), 507–522.
Curis, E. & Bénazeth, S. (2000). *J. Synchrotron Rad.* **7**, 262–266.
Filipponi, A. (1995). *J. Phys. Condens. Matter*, **7**, 9343–9356.
Krappe, H. J. & Rossner, H. H. (1999). *J. Synchrotron Rad.* **6**, 302–303.
Lilliefors, H. W. (1967). *J. Am. Stat. Assoc.* **62**, 399–402.
Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Technip.
Vlaic, G., Andreatta, D., Cepparo, A., Colavita, P. E., Fonda, E. & Michalowicz, A. (1999). *J. Synchrotron Rad.* **6**, 225-227.