Supplementary material for

Estimating the number of pure chemical components in a mixture by X-ray absorption spectroscopy

Authors: Alain Manceau, Matthew A. Marcus, Thomas Lenoir

This document provides information about the workings of the NSS analysis referred to in the main text and included in the `PCA_Estimator.exe` program supplied in the zipped folder. It is not the user manual for the program, which is a separate document (`PCA_Estimator_manual.pdf`) also included into the zipped folder. Although EXAFS is referred to below, the methods described here are for XANES, including extended XANES data, not for $k^n\chi(k)$-type EXAFS data.

1. Filtering problem definition

In order to use the NSS method for evaluating the number of PCA components required to fit a XANES dataset, it is necessary to de-noise the data, that is to remove that which is "obviously" noise, leaving what can be signal. One approach to doing this filtering is to use filters applied over different regions of energy, with interpolation where they connect, and adjust the parameters of the filters 'by eye'. However, this procedure introduces a number of arbitrary parameters.

The need for all these parameters comes from not using what is known about the physics of the data. We know, for instance, that no actual feature in the XANES can be narrower than the broadening induced by the instrument response and the core-hole lifetime. We also know that EXAFS features in the XANES data, even strong multiple-scattering peaks, tend to die off past some cutoff distance R_{max}, which can be reasonably well estimated based on the crystallinity of the material.

Suppose, before we do the filtering, we transform the energy abscissa into a new variable q such that these minimum-width features are all about the same width wherever they occur, from pre-edge peak to multiple-scattering EXAFS wiggle. Then, one filter applied uniformly should de-noise the data. Ideally, the parameters of this filter should be valid for all data taken at a particular X-ray absorption edge.

2. Mathematical maneuvers

Let us say that we want a step of 1 unit of q to represent the width of a minimum-width feature at any energy E. Make the approximation of replacing differences with derivatives, so that we get

$$\frac{dE}{dq} = \Delta E_{\min}(E) \qquad 1$$

where ΔE_{min} is the width of a sharpest-possible feature, taken as the FWHM of a XANES peak (i.e. the broadening Γ, including the instrument broadening), or as the energy difference between successive zeros of an EXAFS wiggle at R_{max}. Now, the EXAFS coordinate equation may be written as

$$k = c\sqrt{E - E_0} \qquad 2$$

where c is the conversion factor of $0.5132 Å^{-1} eV^{-1/2}$. To get a phase shift of π at a distance of R_{max} implies a step in k of $\Delta k = \frac{\pi}{4R_{\max}}$, or an energy step of

$$\Delta E_{EXAFS} = \Delta k \frac{dE}{dk} = \frac{\pi\sqrt{E - E_0}}{2cR_{\max}} \qquad 3$$

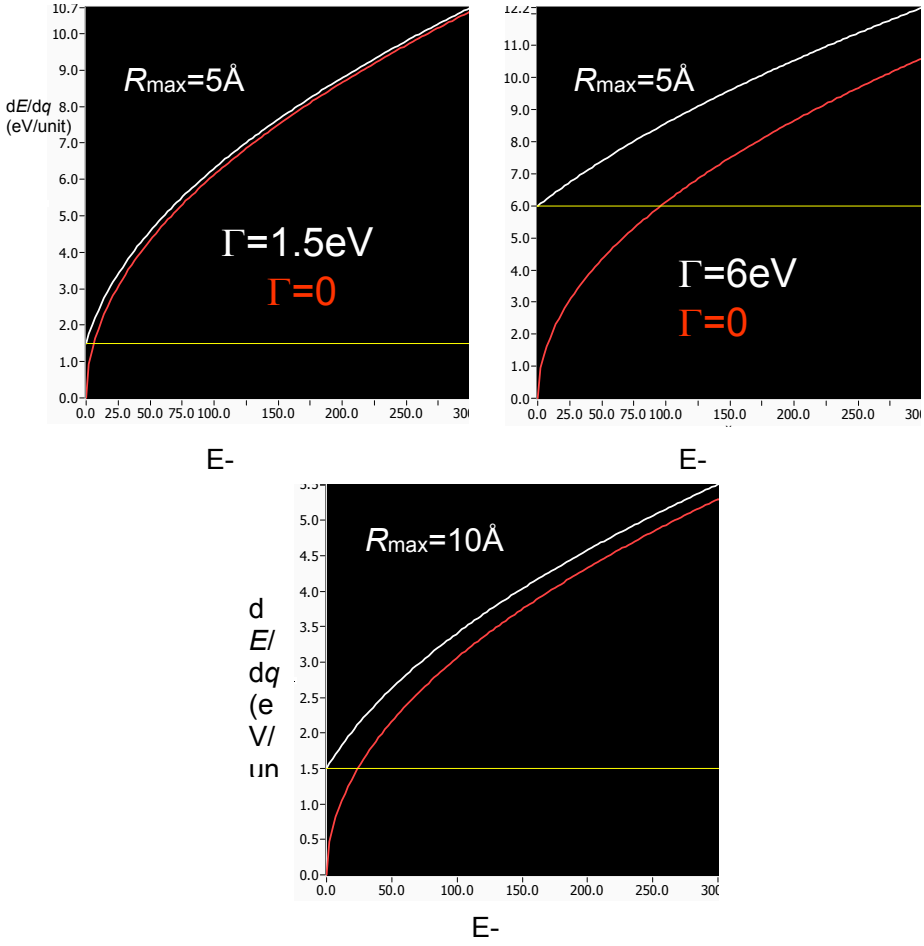
Therefore, we want the slope of the conversion curve of E to q to interpolate between these two limits:

$$\frac{dE}{dq} \to \begin{cases} \frac{\pi\sqrt{E-E_0}}{2cR_{\max}} & E-E_0 \gg \Gamma \\ \Gamma & E \le E_0 \end{cases} \qquad 4$$

which we can make happen in a single formula valid for $E \geq E_0$:

$$\frac{dE}{dq} = \sqrt{\Gamma^2 + \left(\frac{\pi}{2cR_{\max}}\right)^2 (E-E_0)} \qquad 5$$

Three examples of this formula are shown in the following figure, in which the white and red curves represent non-zero and zero Γ, respectively, to illustrate the transition between broadening-dominated and EXAFS-limited regions:



Now, what should happen for $E < E_0$? On the one hand, it could be argued that the pre-edge gets quickly flat as E decrease, but on the other hand, there are pre-edge features which are diagnostic in XANES analysis and must not be taken for noise. These two possibilities may be handled by two choices of the form of dE/dq for $E < E_0$:

$$\frac{dE}{dq} = \begin{cases} \sqrt{\Gamma^2 + \left(\frac{\pi}{2cR_{\max}}\right)^2 |E-E_0|} \\ \Gamma \end{cases} \qquad 6$$

with the upper form (square root) representing the flat pre-edge case.

The differential equation 5 may be solved by bringing the dq to the right hand side, and the square root to the left hand side and integrating:

$$q = \frac{2\left(\sqrt{\Gamma^2 + \varepsilon(E - E_0)} - \Gamma\right)}{\varepsilon}$$
$$\varepsilon = \left(\frac{\pi}{2cR_{\max}}\right)^2 \qquad 7$$

with ε a quantity with dimensions of energy. The inverse equation is

$$E = E_0 + \left(\left(\frac{q\varepsilon}{2} + \Gamma\right)^2 - \Gamma^2\right)/\varepsilon \qquad 8$$

These equations hold for $E \geq E_0$. For $E < E_0$, the two possibilities discussed above yield

$$q = -\frac{2\left(\sqrt{\Gamma^2 + \varepsilon(E_0 - E)} - \Gamma\right)}{\varepsilon}$$
$$E = E_0 - \left(\left(\Gamma - \frac{q\varepsilon}{2}\right)^2 - \Gamma^2\right)/\varepsilon \qquad 9$$

and

$$q = (E - E_0)/\Gamma$$
$$E = E_0 + q\Gamma \qquad 10$$

for the choice that preserves pre-edge features. This is the choice used by default in the `PCA_Estimator` program.

3. Implementation notes

For filtering, we used a second-order Butterworth filter with adjustable cutoff frequency ("*Fcut*" parameter). Our program also makes available elliptic, Bessel and Chebychev filters. Since the filtering algorithm requires uniformly-spaced points, we resampled the data to 2048 points by cubic spline interpolation after transformation to *q*-space, filtered, then re-interpolated back onto the original grid. In order to reduce the possible effects of this back-and-forth resampling, we constructed the final, filtered spectrum by resampling the filter residuals (filtered - original) onto the original grid and added these to the original, non-resampled data. That way, if the filter did nothing, the resulting filtered, resampled data would be identical to the original.

The filters available in our program are all causal filters, which implies a shift between the input and output. What we want is a filter whose impulse response is symmetric about zero lag. Thus, we reverse a copy of the data, run it through the same filter as the data, reverse it again, then add the two filtered versions together, with an abscissa offset. This abscissa offset is chosen so that the first moment of the impulse response of the direct and reversed filters are equal. With this procedure, features such as peaks are broadened by the filter but not shifted, so the difference between filtered and unfiltered resembles the second derivative of the signal, for high filter cutoff frequency. All of this manipulation is, of course, in the *q* domain

How does one define what E_0 should be? It is not reasonable for the user of this method to have to type in separate values for each spectrum, so a possible choice is to use a fixed value (user input) for all spectra. Another choice is to use some sort of automatic detection. In our case, we use the energy at which the normalized XANES first reaches a value of 0.5. This sort of automated detection takes into account the fact that the different spectra will often represent different oxidation states, so should get different E_0. In the supplied program, entering a zero value for E_0 enables automatic setting.

The broadening Γ should be the quadrature sum of the core-hole broadening (assuming that a high-resolution method like partial-fluorescence yield is not used) and the instrumental broadening. For most hard X-ray applications, the core-hole lifetime dominates. Approximate values may be obtained from M.O. Krause

and J. H. Oliver "Natural widths of atomic K and L levels, Kα X-ray lines and several KLL Auger lines", *J. Phys.Chem. Ref. Data* **8**, 329-338 (1979) and O. Keski-Rahkonen and M. O.. Krause, "Total and partial Atomic-level widths", *Atomic Data and Nuclear Data Tables* **14**, 139-146 (1974).

In Figure 5 of the main text, we used Γ = 2eV, R_{max} = 5 Å, Butterworth (2nd order) cutoff frequency *Fcut* = 0.7. All these parameters may be adjusted interactively in the program, thus allowing one to see the region of parameter space over which the results (i.e., NSS-stat) are robust.

4. Signal/noise measures

We have implemented two measures of signal/noise ratio (*s*/*n*) for spectra, both based on the approximation that the filtered spectrum is de-noised, so that noise is defined as the difference between filtered and unfiltered data. The simpler, more obvious measure is

$$s/n = \sqrt{\frac{\langle S^2 \rangle}{\mathrm{Var}(S-F)}} \qquad 11$$

where *S* and *F* are the original signal and filtered data, respectively, $\langle \cdots \rangle$ is a mean over the energy range, and Var(*x*) is the variance of *x* sampled over the energy range. This is essentially the intuitive definition of signal/noise. For XANES, we usually want to consider only the post-edge region in evaluating *s*/*n*, so all sums and means are carried out in a user-selected energy region.

The International X-ray Absorption Society's standard for reporting data (`http://ixs.iit.edu/subcommittee_reports/sc/err-rep.pdf`) implies a somewhat different definition:

$$s/n_{IXAS} = \sqrt{\frac{\langle S \rangle}{\mathrm{Var}((S-F)/\sqrt{S})}} \qquad 12$$

These two measures turn out to yield consistent results. The program reports both.