Data Analysis with Least-Squares Fitting

Matthew Newville

Center for Advanced Radiation Sources
The University of Chicago

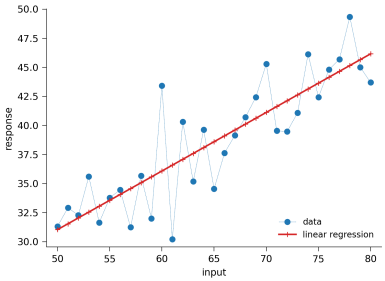
2022-July-27

2022 XAFS School

Fitting Scientific Data

Analyzing scientific data often involves finding and then refining a *Model* to best matches a *Measurement* or set of *Observations*.

A simple example of a Model would be a linear response: $y = mx + b$:



and the *Parameters* m and b would be adjusted to best match the data.

The most common way to decide which Parameters best match the measured data is the χ^2 or "least squares":

$$\chi^2 = \sum_i^N \frac{[y_i^{\mathrm{measured}} - y_i^{\mathrm{model}}(p)]^2}{\epsilon^2}$$

where

- N: number of data pointsto fit.
- ϵ: estimated noise in the data.
- p: the set of parameters to be optimized.

The Best Fit is the one with lowest χ^2.

$$\chi^2 = \sum_i^N \frac{[y_i^{\mathrm{measured}} - y_i^{\mathrm{model}}(p)]^2}{\epsilon^2}$$

Questions:

1. How can we compare different models?
2. What parameters in that model should be varied?
3. What is ϵ for the data?
4. What are the independent measurements in N?

These are general data fitting questions, with a huge literature.

Note 1: This approaches uses "a series of measurements", and there is no "independent variable" (x, or for XANES E and for EXAFS maybe k or R) in this desciption. Of course, we'll need that to model the data, but it is truly treated as independent.

Note 2: some people do things other than "squaring (data - model)". We'll leave that aside for this discussion, though there are ways to deviate from pure least-squares.

There are two Important cases for modeling data and least-squares analysis.

Linear Least Squares the model for y is linear *in the parameters*:

$$y = a + b * x + c * x^2$$

These cases can be solved with a single-step "regression". This is easy , robust, and very fast.

Non-Linear Least Squares the model for y is not linear in the parameters:

$$y(x, A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} e^{[-(x-\mu)^2/2\sigma^2]}$$

.

Such cases are solved iteratively, and are less easy, less robust, and less fast.

For some XANES analysis methods, we'll use Linear Least-Squares. For peak-fitting and EXAFS analysis, we'll use Non-Linear Least-Squares.

Fitting Statistics

To decide how well a *Model* matches a *Measurement*, χ^2 (don't confuse with EXAFS χ!!) is the main statistic used:

$$\chi^2 = \sum_i^N \frac{[y_i^{\rm measured} - y_i^{\rm model}(p)]^2}{\epsilon^2}$$

where

- N: number of data pointsto fit.
- ϵ: estimated noise in the data.
- p: the set of parameters to be optimized.

The Best Fit is the one with lowest χ^2.

Ideally – if ϵ is set correctly **and** the model is correct – then χ^2 should be around N:
the difference between model and data should be "at the noise level".

Other Fitting Statistics

χ^2 is useful, but other "goodness-of-fit statistics" are useful for comparing two models:

reduced chi-square: scale χ^2 by the "degrees of freedom" :

$$\chi^2_\nu = \chi^2/(N - N_{\rm varys})$$

where $N_{\rm varys}$ is the number of variable parameters in the fit.

For a "Good Fit", χ^2_ν should be ~ 1. Realistically, $\chi^2_\nu < 10$ is doing pretty good.

R-factor: $\mathcal{R}$ gives a "fractional misfit" (and so is not scaled by the data uncertainty ϵ):

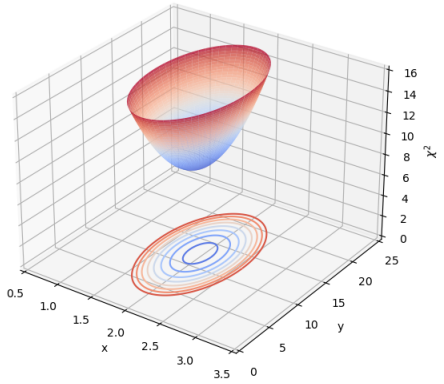
$$\mathcal{R} = \frac{\sum_i^N [y_i^{\rm measured} - y_i^{\rm model}(p)]^2}{\sum_i^N [y_i^{\rm measured}]^2}$$

Akaike Information Criterion: Also weights to account for degrees of freedom in fit:

$$\mathrm{AIC} = N \log(\chi^2/N) + 2N_{\rm varys}$$

Optimizing Variables: Finding the Best Fit

For Non-Linear Least-Squares fits, the fit will start at *Initial Values* and Refine those values to minimize the difference between model and data (χ^2).



Reminder: x and y here are parameters in the Model: the values to be optimized.

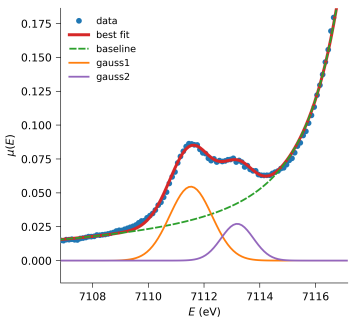
The refinement is iterative, trying to find the best set of parameter values.

Initial values are always important: a fit can get lost with poor initial values.

For XANES and EXAFS, these methods can be pretty robust.

Reigning in Variables: Bounds and Constraints

The fits allow for Parameters to be *unconstrained*: they can each take any continuous value.



Fit to XANES pre-edge peaks using a Model of 2 Gaussians, 1 Lorentzian (main edge) and 1 Line:

$$y(E) = \frac{A_1}{\sigma_1\sqrt{2\pi}} e^{-(E-\mu_1)^2/2{\sigma_1}^2} + \frac{A_2}{\sigma_2\sqrt{2\pi}} e^{-(E-\mu_2)^2/2{\sigma_2}^2} + \frac{A_3\sigma_3}{(E-\mu_3)^2 + {\sigma_3}^2} + mE + b$$

That makes 11 variable parameters.

We may want to:

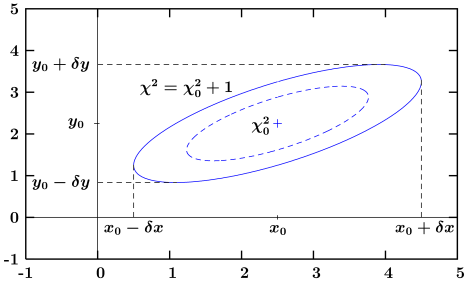
- set upper/lower bounds on Parameters: $sigma_1 > 0$, $7110 < \mu_1 < 7114$.
- Fix some Parameters - don't allow those to vary at all.
- *Constrain* or tie some values to each other: $\sigma_2 = \sigma_1$ or $\mu_2 = \mu_1 + 3$.

All of these are available in the libraries used.

Error Bars: the uncertainties in the fit variables

A fit finds the "best-fit" set of values for the variables $\{x, y, \ldots\}$: these give the lowest $\chi^2 = \chi^2_0$.

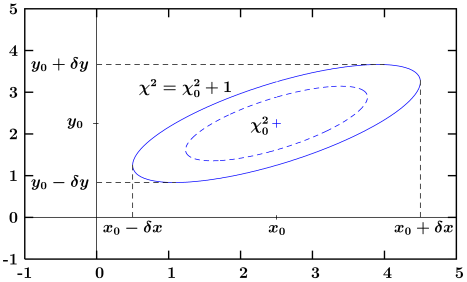
Uncertainties in Parameters x are estimated by increasing the χ^2 by 1:



Error Bars: the uncertainties in the fit variables

A fit finds the "best-fit" set of values for the variables $\{x, y, \ldots\}$: these give the lowest $\chi^2 = \chi^2_0$.

Uncertainties in Parameters x are estimated by increasing the χ^2 by 1:



Some Parameters are *Correlated*:

Changing the value for parameter x away from its best value will change the best value for another parameter, y.

For EXAFS, (R, E_0) and (N, σ^2) are usually very highly correlated (> 0.85).

Increasing χ^2 by 1 assumes we have a "Good Fit", with $\chi^2_\nu \approx 1$.

For EXAFS, we typically have $\chi^2_\nu \sim 10$, so we increase the best χ^2 by χ^2_ν to estimate error bars.

The reported uncertainties do take the correlation into account!

More rigorous methods for uncertainty analysis is available from the LARCH Python code.

Data Analysis and Least Squares Fitting

All data analysis needs to account for uncertainties in the data and should do some comparison of different models.

Using the normal fitting statistics such as χ^2 and reduced χ^2_ν are the starting points for comparing models . . . this includes whether "adding 1 more variable parameter" is needed to improve a fit.

Please remember to report uncertainties as well as best-fit values. For EXAFS, reporting "typical" uncertainties (say, 0.01Å for R) mayn be OK, but uncertainties from each fit should be reported.

More information on X-rays and X-ray Absorption Spectroscopy:

https://xafs.xrayabsorption.org/

Fundamentals of XAFS M. Newville, Reviews in Mineralogy & Geochemistry **78**, 2014.

Introduction to XAFS G. Bunker, Cambridge Univ Press, 2010.

XAFS for Everyone S. Calvin, CRC Press, 2013.

Elements of Modern X-ray Physics J. Als-Nielsen & D. McMorrow, John Wiley & Sons. 2001