Curve Fitting

Introduction:

The objective of curve fitting is to theoretically describe experimental data with *a* model (function or equation) and to find the parameters associated with this model. Models of primary importance to us are *mechanistic models*. *Mechanistic models* are specifically formulated to provide insight into a chemical, biological, or physical process that is thought to govern the phenomenon under study. Parameters derived from mechanistic models are quantitative estimates of real system properties (rate constants, dissociation constants, catalytic velocities etc.). It is important to distinguish *mechanistic* models from *empirical* models that are mathematical functions formulated to fit a particular curve but whose parameters do not necessarily correspond to a biological, chemical or physical property. An example of an empirical fit is a polynomial fit to the baseline of a NMR spectrum with the goal to baseline-correct the spectrum. The final coefficients are physically meaningless and also of no interest. The objective of curve fitting is different: one is just trying to draw a curve through the baseline. Other examples of empirical fitting include interpolations such as splines and smoothing. We will not be concerned with such empirical curve fitting methods in this discussion.

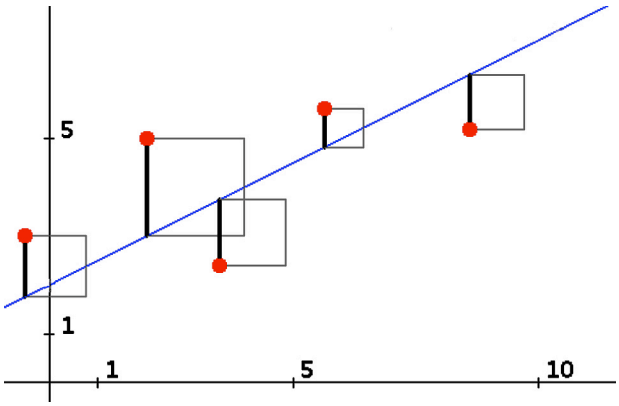
Curve Fitting: The Least-Squares method:

Curve fitting finds the values of the coefficients (parameters) which make a function match the data as closely as possible. The *best* values of the coefficients are the ones that minimize the value of Chi-square. Chi-square is defined as:

$$\chi^2 = \sum_i \left(\frac{y - y_i}{\sigma_i} \right)^2$$

where y is a fitted value for a given point, y_i is the measured data value for the point and σ_i is an estimate of the standard deviation for y_i (discussed under weighting). In other words, the best values of the coefficients are obtained when the sum of the squared distances $\sum(y\text{-}y_i)^2$ of the fit to the data is minimized while giving data points more weight during this minimization process that have smaller errors (σ_i).

This principle is illustrated easiest with the case of fitting data to a straight line: $y = ax + b$. We presume that there is a theoretical reason for our data to fall on a straight line ($\rightarrow$mechanistic model).



The sum-of-squares $\sum(y\text{-}y_i)^2$ is, like the name implies, the sum of the squares illustrated in the figure above. The *Least-Squares* method during curve-fitting aims to minimize this sum.

The method of minimization is therefore called least-squares fitting (or regression). The reason that the squares are minimized as opposed to just the vertical distances is related to the squared term in the Gaussian error distribution. No further details are provided here.

$P(x) \sim e^{-x^2}$

We distinguish *linear* least-squares from *nonlinear* least-squares fitting. If the dependent variable y is linearly dependent on the parameter the linear least-squares method of minimization applies (e.g. straight line, polynomial). This proceeds noniteratively in one step. Nonlinear least-squares fitting is computationally more intensive as the fit iteratively tries various values to yield the minimum value of chi-square.
Note that in the case of linear least squares fitting, the function need not be linear in the argument x, but only in the parameters that are determined to give the best fit. Therefore, $y=ae^x$ is an example of a linear least squares problem (y is linearly dependent on a) whereas $y=ae^{bx}$ is an example of a nonlinear least-squares problem due to the nonlinear dependence of y on the parameter b.

Nonlinear least-squares fitting employs the Levenberg-Marquardt algorithm to search for the coefficients that minimize chi-square. It has become the industry standard in nonlinear regression. As the fit proceeds and better values are found, the chi-square value decreases. The fit is finished when the *rate* at which chi-square decreases is small enough (other convergence criteria may apply).

Assumption of the nonlinear least-squares method:
(1) x is known precisely, all the error is in y.
(2) variation in y follows a Gaussian distribution
(3) scatter is the same all the way along the curve

Implications of this assumption on data processing prior to fitting:
(1) Shifting x-value by a constant: ok
(2) divide/ multiply all x-values by constant: ok
(3) smooth data prior to curve fitting (interpolation, spline, smoothing)
→ error distribution is no longer Gaussian resulting in misleading fits.
(4) Applying nonlinear transforms: $1/y$, $\sqrt{y}$, $\log(y)$ etc.
→again, error distribution is no longer Gaussian leading to less accurate or inaccurate estimates of parameters and skewed/biased error bars.

Weighting:
In general, a weight w_i must be assigned for each measurement y_i. This weighting term is used in the calculation of chi-square:

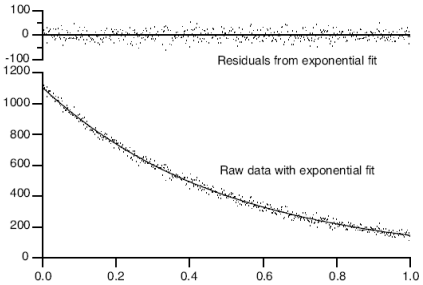
$$\chi^2 = \sum_i \left(\frac{y - y_i}{w_i} \right)^2$$

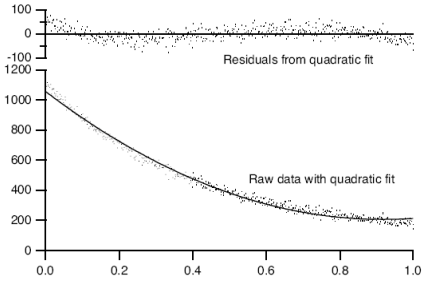
If estimates of the standard deviation σ_i are available, the true weights can be used ($w_i = \sigma_i$). One can provide this *a priori* by measuring the standard deviation of the experimental noise. Otherwise, σ_i is first set to unity and estimated *posteriori* from the residuals after the fit. This has no implications on the fitting process but rather is necessary to calculate valid error bars of the fit. One may provide user-specific weighting to assign greater or lesser importance to certain data points. This is based on experience or on common-sense criteria. This may introduce bias. However, there exist physical models that require such weighting terms. Examples include:
$w_i = 1/y_i$ and,
$w_i = \sqrt{y_i}$.

Analysis of the Fit

(1)Residuals - A good visual cue:

A residual is what is left when one subtracts the fit from the raw data. Ideally, the raw data is equal to some known function plus random noise. If one subtracts the function from the data, what is left therefore should be noise. If this is not the case, then the function does not properly fit the raw data. The residuals illustrated in the left-hand figure are the result of a good fit whereas the residuals in right-hand figure reflect a deviance.





95% confidence limit

-Assessing error of parameter estimates

Reporting the parameter estimates as '*best estimate* ±δ' means that the estimate of the parameter lies between '*best estimate* -δ' and '*best estimate* +δ' with a 95% probability. Note that valid error estimates are only valid if all the data points have roughly equal errors and if the fit function is, in fact, appropriate to the data. It is therefore important to inspect the residuals first before interpreting 95% confidence limits.

Correlation Matrix

-Assessing uniqueness of parameter estimates

A correlation matrix is a normalized form of the covariance matrix (covariance matrix is an array of covariances between the parameters. It is the natural generalization to higher dimensions of the concept of the variance). Each element of the correlation matrix shows the correlation between two fit coefficients as a number between *-1* and *1*. The correlation between two coefficients is perfect if the corresponding element is *1*, it is a perfect inverse correlation if the element is *-1*, and there is no correlation if it is *0*.

	k_1	k_2	k_{-2}	k_3	k_4	k_{-4}
k_1	1.00	-0.76	-0.74	0.11	-0.87	-0.63
k_2	-0.76	1.00	0.99	-0.41	0.58	0.41
k_{-2}	-0.74	0.99	1.00	-0.37	0.55	0.41
k_3	0.11	-0.41	-0.37	1.00	-0.20	-0.15
k_4	-0.87	0.58	0.55	-0.20	1.00	0.82
k_{-4}	-0.63	0.41	0.41	-0.15	0.82	1.00

Note that the matrix is symmetric with respect to the diagonal. One should be suspicious of fits in which an element of the correlation matrix is very close to *1* or *-1*. This may signal "identifiability" problems. That is, the fit does not distinguish between two of the parameters very well, and so the fit is not very well constrained.