CURVE FITTING

S. Rajasekar
School of Physics
Bharathidasan University
Tiruchirapalli - 620 024
email: rajasekar@cnld.bdu.ac.in
srj.bdu@gmail.com

What is a curve fitting?

The general problem of construction of an appropriate mathematical equation which fit a set of given data is called curve fitting. Let us assume that a set of data (x_k, y_k), $k = 1, 2, ..., n$ containing two variables is given. The goal of the curve fitting is to determine a formula $y = f(x)$. How does one choose an appropriate form of $f(x)$? Usually, by inspecting the given data and the physical situation a particular form of $f(x)$ is chosen. For example, the data points in Fig. 1a appear to be varying linearly with x. Therefore, the linear relation $y = ax + b$ can be tried.

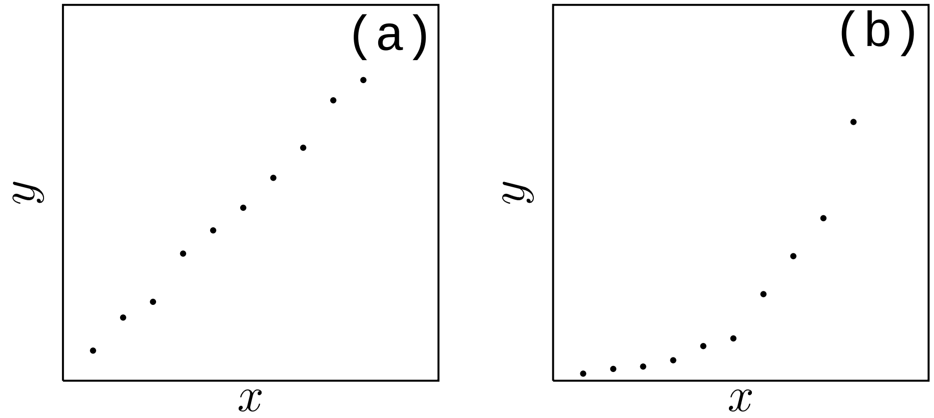

Figure 1: Data points y varying roughly (a) linearly and (b) nonlinearly with x.

For the data in Fig. 1b an exponential function $y = b\mathrm{e}^{ax}$ is suitable. Methods are available to determine the constants a and b.

Method of Least-Squares

Let (x_k, y_k), $k = 1, 2, ..., n$ be n sets of observations and it is required to make a fit to the function

$$y = F\left(x, c_1, c_2, ..., c_m\right) \;, \tag{1}$$

where F is some function of x which depends linearly on the parameters $\{c_i\}$. That is,

$$F = c_1\phi_1(x) + c_2\phi_2(x) + ... + c_m\phi_m(x) \;, \tag{2}$$

where $\{\phi_i(x)\}$ are a priory selected set of functions and $\{c_i\}$ are to be determined. Normally, m is small compared with the number, n, of the data points. The functions $\{\phi_i(x)\}$ may, for example, be simply x or x^k or $\cos(x)$ and so on.

Basic Idea

The basic idea of curve fitting is to choose the parameters $\vec{c} = \{c_i\}$ in such a way that the deviation error or the residuals e_k given by

$$e_k = y_k - F\left(x_k, \vec{c}\right), \qquad k = 1, 2, ..., n \tag{3}$$

are minimum. In other words, the difference between the observed or given y and the value of y calculated from the relation (1) to be determined is made as small as possible. Here e_k's can take both positive and negative values. In order to give equal weighage to positive and negative e_k's consider the root-mean-square (rms) error

$$E_{\text{rms}} = \left[\frac{1}{n}\sum_{k=1}^{n} e_k^2\right]^{1/2} . \tag{4}$$

A best fit is then obtained by minimizing E_{rms}.

From Eq. (4) it is clear that the E_{rms} will be minimum if and only if the quantity

$$E(\vec{c}) = \sum_{k=1}^{n} e_k^2 \tag{5}$$

is a minimum. Hence, *the best curve fit to a set of data is that for which the sum of the squares of the residuals is a minimum*. This is known as least-squares criterion and the resulting approximation $F(x, \vec{c})$ is called a least-squares approximation to the given data.

Derivation of Equations for the Unknowns $\{c_i\}$

It is desired to find $\{c_i\}$ for which $E(\vec{c})$ given by Eq. (5) is a minimum. Recall that the slope $\mathrm{d}f(x)/\mathrm{d}x = f'(x)$ of a single variable function $f(x)$ is zero at the minimum value of f. A function $f(\vec{X})$ of several variables is minimum when $\nabla f(\vec{X}) = 0$ where ∇ is the gradient operator. Thus the necessary condition for the function E to be minimum is

$$\nabla E(\vec{c}) = 0\,, \tag{6}$$

where the gradient operator ∇ here takes the form

$$\nabla = \vec{i}_1 \frac{\partial}{\partial c_1} + \vec{i}_2 \frac{\partial}{\partial c_2} + \ldots + \vec{i}_m \frac{\partial}{\partial c_m} \tag{7}$$

with $\vec{i}_j$'s being unit vectors. Equation (6) is satisfied if and only if all the components of ∇E are identically zero: $\partial E/\partial c_i = 0$. That is,

$$\frac{\partial E}{\partial c_i} = \frac{\partial}{\partial c_i} \sum_{k=1}^{n} \left[y_k - F\left(x_k, \vec{c}\right)\right]^2 = 0, \qquad i = 1, 2, \ldots, m. \tag{8}$$

From Eq. (2) one has $\partial F/\partial c_i = \phi_i$. Therefore, Eqs. (8) become

$$-2 \sum_{k=1}^{n} \left[y_k - F\left(x_k, \vec{c}\right)\right] \phi_i\left(x_k\right) = 0\,. \tag{9}$$

Derivation of Equations for $\{c_i\}$

Then using Eq.(3) in (9) one gets

$$\sum_{k=1}^{n} e_k \phi_i\left(x_k\right) = 0\,, \quad i = 1, 2, ..., m. \tag{10}$$

That is,

$$\vec{e} \cdot \vec{\Phi}_i = 0\,, \quad i = 1, 2, ..., m \tag{11}$$

where

$$\vec{e} = \left[e_1, e_2, ..., e_n\right]^{\mathrm{T}}\,, \quad \vec{\Phi}_i = \left[\phi_i(x_1), \phi_i(x_2), ..., \phi_i(x_m)\right]\,. \tag{12}$$

Equation (11) implies that the error vector $\vec{e}$, for all n, should be normal or orthogonal to each of the n vectors $\vec{\Phi}_i$. Because of this, the m equations, namely Eqs.(9) are called *normal equations*.

Application to Straight-Line Fit

Now, consider how to make a linear or straight-line fit to a given set of n data. Essentially the problem is to determine the values of the constants a and b in the function

$$y = f(x, a, b) = ax + b\,. \tag{13}$$

Simple formulas for a and b can be obtained by solving the normal equations of a and b. From Eqs.(8) and (13) the normal equations are written as

$$\frac{\partial E}{\partial a} = \frac{\partial}{\partial a}\sum_{k=1}^{n}(y_k - ax_k - b)^2 = 0\,, \tag{14a}$$

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b}\sum_{k=1}^{n}(y_k - ax_k - b)^2 = 0\,. \tag{14b}$$

Performing the partial derivatives in Eqs. (14) one has

$$-2\sum(y_k - ax_k - b)\,x_k = 0\,, \tag{15a}$$

$$-2\sum(y_k - ax_k - b) = 0\,, \tag{15b}$$

where the suffices in the summations are dropped for simplicity.

Application to Straight-Line Fit

Equations (15) can be rewritten as

$$a \sum x_k^2 + b \sum x_k = \sum x_k y_k \,, \tag{16a}$$

$$a \sum x_k + nb = \sum y_k \,. \tag{16b}$$

The above set of equations are linear in the unknowns a and b and can be easily solved. The equation for a is obtained by multiplying Eq. (16a) by n, Eq. (16b) by $\sum x_k$ and then subtracting one from another. Similarly, equation for b is obtained by multiplying Eq. (16a) by $\sum x_k$, Eq. (16b) by $\sum x_k^2$ and then subtracting one from another. The equations for a and b are obtained as

$$a = \frac{n \sum x_k y_k - \sum x_k \sum y_k}{n \sum x_k^2 - \left(\sum x_k\right)^2} \,, \tag{17a}$$

$$b = \frac{\sum x_k^2 \sum y_k - \sum x_k \sum x_k y_k}{n \sum x_k^2 - \left(\sum x_k\right)^2} \,. \tag{17b}$$

Example

Find the least-squares straight-line fit for the data given in the following table.

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7
y	0.16	0.21	0.23	0.30	0.36	0.39	0.46

For the above data the values of x_k^2, $x_k y_k$, $\sum x_k$, $\sum y_k$, $\sum x_k^2$ and $\sum x_k y_k$ are given in the following table:

	x_k	y_k	x_k^2	$x_k y_k$
	0.1	0.16	0.01	0.016
	0.2	0.21	0.04	0.042
	0.3	0.23	0.09	0.069
	0.4	0.30	0.16	0.120
	0.5	0.36	0.25	0.180
	0.6	0.39	0.36	0.234
	0.7	0.46	0.49	0.322
$\sum$	2.8	2.11	1.40	0.983

Example

Now, from Eqs. (17) a and b are calculated as

$$\begin{aligned} a &= \frac{7 \times 0.983 - 2.8 \times 2.11}{7 \times 1.4 - 2.8 \times 2.8} &= 0.49643\,, \\ b &= \frac{1.4 \times 2.11 - 2.8 \times 0.983}{7 \times 1.4 - 2.8 \times 2.8} &= 0.10286\,. \end{aligned}$$

Thus the least-squares straight-line fit is

$$y = 0.49643x + 0.10286\,. \tag{18}$$