Data Fitting

In many circumstances, one has a theory prediction of the relationship between two quantities. For example, in Phy335, we may have a circuit theory prediction for $|V_{out}/V_{in}|$ as a function of the frequency $\omega = 2\pi f$. We measure V_{out} at some different ω values, and we'd like to compare our measurements with the theory. This handout gives a (very basic) introduction to this topic which is usually called "curve fitting" or "parameter estimation". This short note describes the standard method, called χ^2 fitting or χ^2 minimization. The result is presented without going into the (well established) formal derivation using the method of maximum likelihood.

Suppose one has a set of N measurements $\{y_i\},\ i = 1,..,N$ taken at specific values of an indpendent variable $\{x_i\}$, the uncertainty on each measurement σ_i, and a theory prediction for $y = f(x; p_j)$ in which $p_j,\ j = 1,..,M$ are the parameters needed to define the function. In the above circuit example, the y values are output voltages, and the x values are the frequencies dialed up on the signal generator. The goal of the fitting procedure is to to determine the best values of the parameters p_j and to determine the quality with which the function describes the data.

General Procedure

Step 1: The first step is to define the χ^2

$$\chi^2 \equiv \sum_{i=1}^{N} \Big(\frac{y_i - f(x_i; p_j)}{\sigma_i}\Big)^2 \tag{1}$$

In this definition only the parameters p_j are not known. The values $x_i,\ y_i$, and σ_i are the numerical values from the experiment. Thus, for this purpose, the χ^2 is a function of the parameters p_j.

Step 2: TThe method of maximum likelihood says that the parameter values which give the best match to the specific data set are found by minimizing the χ^2, treating as a functino of the parameters. Any method which does the minimization works. In terms of simple calculus, a function is minimized (or maximized) by setting its first derivatives to zero,

$$\frac{\partial \chi^2}{\partial p_j} = 0 \qquad j = 1, 2, 3, ..., M \tag{2}$$

This is actually M equations, one for each parameter. So, one must solve a system of M equations in M unknowns p_j. *Once the system has been solved, the best fit parameters p_{j0} have been determined.*

Step 3: The third step is to determine the uncertainties σ_{p_j} on the parameters, p_{j0}. This is done by forming the curvature matrix, $\mathbf{C}$. Each term in the curvature matrix C_{ij} is defined as a second partial derivative of the χ^2,

$$C_{ij} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial p_i \partial p_j} \tag{3}$$

The curvature matrix is inverted to get the "error" or "covariance" matrix

$$\epsilon = \mathbf{C^{-1}} \tag{4}$$

The uncertainties σ_{p_j} on the fitted parameters are determined from the terms of the error matrix as

$$\sigma_{p_j} = \sqrt{\epsilon_{jj}}. \tag{5}$$

In addition, each term ρ_{ij} in the correlation matrix ρ describing the impact each parameter has on another is given by

$$\rho_{ij} = \frac{\epsilon_{ij}}{\sigma_{p_i}\sigma_{p_j}} \tag{6}$$

Step 4: For the last step, now that we have determined the best values of p_j for our data, the method allows us to ask if the function really matches our data. To do this compute the χ^2 numerically, using the values for p_j determined in step two. That is, let

$$\chi_0^2 = \sum_{i=1}^{N} \Big(\frac{y_i - f(x_i; p_j)}{\sigma_i}\Big)^2 \tag{7}$$

with the subscript "0" denoting this is the particular value found using our best fit values p_{j0}. The value

$$\chi_0^2/(N - M) \tag{8}$$

is a goodness–of–fit and $N - M$ is called the *number of degrees of freedom*. On average, if the function is the right one, then $< \chi_0^2 >= (N-M)$ and the RMS of the χ_0^2 value is $\sqrt{2(N-M)}$. One can also look up the probability $P(\chi_0^2, N - M)$ in standard places like Excel, Root, Mathematica or the web[1].

Example 1: Fitting data to a constant

Suppose we want to know the single value which best describes our data. That is, if we expect (e.g. have a theory relation) that our data is a constant c, or in the notation above

$$y = f(x; c) = c$$

with c a constant, what value of c is our data most consistent with? The prescription says make χ^2 and find its minimum. So

$$\chi^2 = \sum_{i=1}^{N} \Big[\frac{y_i - c}{\sigma_i}\Big]^2$$

and we find the minimum by differentiating and setting the result to zero:

$$0 \;\; = \;\; \frac{\partial \chi^2}{\partial c} = -2 \sum \Big[\frac{y_i - c}{\sigma_i^2}\Big]$$

Solving this for c gives

$$c = \frac{\sum(y_i/\sigma_i^2)}{\sum(1/\sigma_i^2)}$$

This is just the familiar weighted mean, and if all values of σ_i are the same, then

$$c = \frac{1}{N} \sum y_i$$

which is just the usual mean. So, we've derived from first principles the statement that the mean is the choice if one wants a single number to characterize a set of data values!

The prescription also says we can determine the uncertainty of c by taking the inverse of the curvature matrix. For one parameter, the matrix is 1×1, or just a number. The curvature is

$$C = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial c^2} = \sum \frac{1}{\sigma_i^2}$$

[1]Look for chi square probability.

and then the uncertainty on c is just

$$\sigma_c = \sqrt{1/C} = \sqrt{\frac{1}{\Sigma \frac{1}{\sigma_i^2}}}.$$

If all values of $\sigma_i = \sigma_0$ (they're the same), then this becomes

$$\sigma_c = \sigma_0/\sqrt{N}.$$

Finally, the prescription tells us that we can determine the goodness-of-fit by recomputing χ^2 after plugging in the value of c determined above. Rather than derive a separate expression, just plug back in the value and determine the number for χ^2.

Example 2: Linear Function
For the case of a linear relationship,

$$y = mx + b,$$

the χ^2 becomes

$$\chi^2 = \sum_{i=1}^{N}\Big[\frac{y_i - f(x_i;m,b)}{\sigma_i}\Big]^2 = \sum_{i=1}^{N}\Big[\frac{y_i - (mx_i + b)}{\sigma_i}\Big]^2$$

with $p_1 = m$ and $p_2 = b$ the (undetermined) parameter values. The prescription says minimize χ^2. In this case its easiest to do by taking the partial derivative of χ^2 with respect to each parameter and setting the results to zero, so

$$\begin{aligned} 0 &= \tfrac{\partial\chi^2}{\partial m} &= -2\sum\Big[\frac{y_i - (mx_i + b)}{\sigma_i^2}\Big]x_i \\ 0 &= \tfrac{\partial\chi^2}{\partial b} &= -2\sum\Big[\frac{y_i - (mx_i + b)}{\sigma_i^2}\Big] \end{aligned}$$

These are just two simultaneous equations in two unknowns m and b because all of the other things are numbers from the experiment. Solving these two equations gives the **best fit parameter values**

$$b = \frac{1}{\Delta}\Big[\Big(\sum\frac{x_i^2}{\sigma_i^2}\Big)\Big(\sum\frac{y_i}{\sigma_i^2}\Big) - \Big(\sum\frac{x_i}{\sigma_i^2}\Big)\Big(\sum\frac{y_i x_i}{\sigma_i^2}\Big)\Big] \tag{9}$$

$$m = \frac{1}{\Delta}\Big[\Big(\sum\frac{1}{\sigma_i^2}\Big)\Big(\sum\frac{x_i y_i}{\sigma_i^2}\Big) - \Big(\sum\frac{x_i}{\sigma_i^2}\Big)\Big(\sum\frac{y_i}{\sigma_i^2}\Big)\Big] \tag{10}$$

with

$$\Delta = \Big(\sum\frac{1}{\sigma_i^2}\Big)\Big(\sum\frac{x_i^2}{\sigma_i^2}\Big) - \Big(\sum\frac{x_i}{\sigma_i^2}\Big)^2.$$

Then to determine the uncertainties on m and b, we first make the curvature matrix

$$\mathbf{C} = \begin{pmatrix} \frac{1}{2}\frac{\partial^2\chi^2}{\partial b^2} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial b\,\partial m} \\ \frac{1}{2}\frac{\partial^2\chi^2}{\partial b\,\partial m} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial m^2} \end{pmatrix} = \begin{pmatrix} \sum\frac{1}{\sigma_i^2} & \sum\frac{x_i}{\sigma_i^2} \\ \sum\frac{x_i}{\sigma_i^2} & \sum\frac{x_i^2}{\sigma_i^2} \end{pmatrix} \tag{11}$$

and then get the uncertainties σ_b and σ_m by inverting $\mathbf{C}$

$$\epsilon = \mathbf{C^{-1}} = \begin{pmatrix} \sigma_b^2 & \sigma_{bm}^2 \\ \sigma_{bm}^2 & \sigma_m^2 \end{pmatrix} = \frac{\mathbf{1}}{\mathbf{\Delta}}\begin{pmatrix} \sum\frac{x_i^2}{\sigma_i^2} & -\sum\frac{x_i}{\sigma_i^2} \\ -\sum\frac{x_i}{\sigma_i^2} & \sum\frac{1}{\sigma_i^2} \end{pmatrix} \tag{12}$$

and remembering that the uncertainties are the square roots of the diagonal elements

$$\sigma_b = \sqrt{\frac{1}{\Delta}\sum\frac{x_i^2}{\sigma_i^2}} \tag{13}$$

$$\sigma_m = \sqrt{\frac{1}{\Delta}\sum\frac{1}{\sigma_i^2}} \tag{14}$$

and the correlation is

$$\rho_{bm} = \frac{\sigma_{mb}^2}{\sigma_b\sigma_m}. \tag{15}$$

Finally, determine the goodnes–of–fit by evaluating the χ^2 using the best fit b and m.