# Appendix VIII. Least-Squares Fitting

Revised July 16, 2004

## A. Introduction

When we make repeated measurements of the same variable, we calculate the mean to find our best estimate of the true value of the variable and we calculate the standard deviation to estimate the width of the experimental distribution of the data. See *Appendix V, Uncertainties and Error Propagation* for more information on the mean and standard deviation.

If we measure a two-dimensional distribution, *i.e.*, two variables $x$ and $y$, where $y$ is a function of $x$, then clearly the mean is not a useful parameter. We generally wish to find how the dependent variable $y$ depends on the independent variable $x$.

## B. Method of Least Squares

Suppose that we measure $N$ coordinate pairs, $x$ and $y$, and that we expect our data to fall on a curve represented by the function

$$y = f(x; A, B, ...) \qquad (1)$$

where $x$ is the independent variable and $A$, $B$, ... are unknown parameters. If we estimate trial values of the parameters, we can calculate from this function values $y_c$ at each measured value of $x$, and compare these values to the corresponding measured values $y_m$. Our object is to find values for the parameters that minimize the disagreement between $y_m$ and $y_c$.

For comparison, we define a *"goodness-of-fit"* parameter called $\chi^2$ (*chi-square*) as

$$\chi^2 = \sqrt{\sum_{i=1}^{N}\left(\frac{y_{ic} - y_{im}}{\sigma_i}\right)^2} \; . \qquad (2)$$

To assure that well-measured data have more influence on the fit than do poorly-measured data, we weight the data by the inverse squares of their uncertainties, the *weighting factor* $1/\sigma_i^2$. (*If we set $\sigma$ = 1 in Eq. 2, then all data points are treated equally. This is the simplest form of the least-squares method which is pre-programmed in calculators and business spreadsheets.*) We define the *maximum-likelihood* values of the parameters as those values that minimize $\chi^2$.

## C. Fit to a Straight Line

In order to solve for the parameters of the fit, $A$, $B$, ..., we could vary the trial values in a regular fashion and recalculate $\chi^2$ until we discover the values which minimize it. However, for functions $f(x;A,B, ...)$ that are linear in the parameters, the problem can be solved analytically. A straight line

$$y = f(x;A,B) = A + Bx \qquad (3)$$

is such as function (*as is any polynomial in x.*)

To find the maximum-likelihood values of $A$ and $B$, we first substitute $y$ from Equation 3 for $y_c$ in Equation 2, yielding

$$\chi^2 = \sqrt{\sum_{i=1}^{N}\left(\frac{A + Bx_i - y_{im}}{\sigma_i}\right)^2} \; . \qquad (4)$$

To minimize $\chi^2$, we take partial derivatives with respect to the two parameters $A$ and $B$ and set them to zero. This gives two coupled linear equations that can be solved directly for $A$ and $B$. We can find the uncertainties in the fitted values of the parameters $A$ and $B$ by applying the error propagation relations discussed in Appendix V, *Uncertainties and Error Propagation.*

The method can be readily expanded to find the parameters for fitting any function of the form

$f(x; a_0,a_1,...,a_m)=a_0f_0(x)+a_1f_1(x)+...a_mf_m(x)$,

or

$$f(x;a_0,a_1,\ldots,a_m)=\sum_{j=1}^{m} a_j f_j(x) \quad (5)$$

where the functions $f_j(x)$ can be any functions of the independent variable $x$, but must not be functions of any of the parameters. For example, the functions $f_j(x)$ might be terms in a quadratic function of $x$, such that

$$f_j(x;A,B,C) = A + Bx + Cx^2 . \quad (6)$$

For *non-linear fits*, *i.e.*, problems in which the fitting function is not linear in the parameters, more complicated, non-analytic or semi-analytic methods involving searches of parameter space are required.

## D. Chi-Square $\chi^2$

Once we have found the best-fit values of the parameters, we can use them to calculate $\chi^2$ from Eq. 2. It is convenient to define *chi-square per degree of freedom* as

$$\chi^2_{DOF} = \chi^2 / (N-m) \quad (7)$$

where $N$ is the number of data points, and $m$ is the number of parameters determined from the fit.

The expectation value of $\chi^2$ is

$$<\chi^2> = N-m. \quad (8)$$

so the expectation value of the reduced chi-square is

$$\left\langle \chi^2_{DOF} \right\rangle = 1$$

In introductory laboratory experiments, values between 0.5 and 1.5 are generally acceptable. Values that are too large may indicate poor measurements, an incorrect choice of theoretical function, or an underestimate of the uncertainties. Values that are too low almost always result from an overestimate of the uncertainties.

Sometimes you can spot a poorly measured quantity by comparing the calculated and measured values of $y$. It may be appropriate to re-measure or discard such a number if this can be reasonably justified. If it appears that you have under- or overestimated the uncertainties in the dependent variable, you may choose to make an adjustment. Note that, in general, you must scale all the uncertainties in the $y$-values and refit in order to correct a problem of this sort. Of course, after making such a correction, you will not be able to use $\chi^2$ as an indicator of the quality of the fit. Any adjustments to data must be explained in your report.

## E. Fitting with *Origin*

Fortunately, we don't have to work out the details of a fitting problem for ourselves. There are many computer programs available. In the laboratory, we use *MPLI*, *Logger Pro*, or *Origin* for all such calculations. (*See Appendix IV on Origin.*) **Students who use a standard spreadsheet, instead of *Origin*, for their calculations should program the correct equations to find the uncertainties in the parameters.**