User's Manual

Logistic Function Profile Fit

Software for Estimating the Position, Width, and Asymmetry of the Interface between Dissimilar Materials

Version 1.3

William H. Kirchhoff

Surface and Microanalysis Science Division, National Institute of Standards and Technology

Gaithersburg, Maryland 20899-8370

October 2013

Preface

This document describes the use of the extended logistic function and the associated software for conveniently estimating the position, width, and asymmetry of interfaces between dissimilar materials (such as might be measured in depth profiles and linescans in surface analysis) in an unbiased fashion from a set of discrete measurements.

Section 1 provides a brief explanation of this function and a rationale for its use in describing interface profiles. Section 2 provides instructions on the use of the program LFPF (Logistic Function Profile Fit) for fitting profiles to the extended logistic function. The results of tests to investigate the performance of the algorithm using synthetic profile data are given in Section 3. Section 4 gives a detailed account of the development of the algorithm to perform iterative least-squares fits efficiently with this function for those who wish to know more about how the program LFPF works or wish to develop their own algorithm for performing similar analyses.

Those wishing to use and evaluate the LFPF software should read Section 2 which describes the functions of the program and its options. This software was developed by William H. Kirchhoff who also prepared the documentation. Any questions or comments on the software or the documentation should be sent to lfpf@nist.gov.

The Logistic Function Profile Fitting program is based on a Fortran program written for DOS and originally issued under the name LOGIT. This program was successfully used to fit Auger sputter-depth-profile data [W. H. Kirchhoff, G. P. Chambers, and J. Fine, J. Vac. Sci. Tech. A **4**, 1666 (1986)]. This approach and the associated software were applied in a number of laboratories, and formed the basis for an ASTM standard [E 1636-10: Standard Practice for Analytically Describing Sputter-Depth-Profile Data by an Extended Logistic Function]. The logistic function (although not the specific LOGIT or LFPF software) has also been used to describe Auger linescans [S. A. Wight and C. J. Powell, J. Vac. Sci. Tech. A **24**, 1024 (2006)]. An updated description of LFPF and its application to depth profile and line scan measurements has been recently published to serve as a reference when using this program [W. H. Kirchhoff, J. Vac. Sci.Tech. A, **30,** 051101 (2012).]

The name Logistic Function Profile Fit (LFPF) has been adopted because (1) LOGIT has come to signify a statistical package for analyzing logistic distributions, and (2) LFPF more directly relates to its intended use in profile analyses.

A compact disc is available that contains the LFPF software and documentation. The CD contains the software as an executable file LFPF.exe, the documentation (this manual, LFPFdoc.pdf), and a help file, LFPFHelp.chm. Various text files with test data, described in this documentation are included on the CD so that the user can test the software and compare results with those in the documentation. It is suggested that these files be copied to an appropriate directory on the user's personal computer.

If this program is installed from LFPF Setup.msi no further installation is required.

LFPF.exe, Version 1.3, requires Version 2.0 or higher of the .NET framework. All versions of the .NET framework can be installed by running the appropriate versions of dotnetfx.exe which can be downloaded without charge from Microsoft. All versions can also be downloaded directly.

If version 2.0 or higher of the .NET framework is not installed, attempting to run LFPF setup.msi to install LFPF or attempting to run LFPF.exe will result in an error message along the lines of:

> **To run this application, you first must install one of the following versions of the .NET Framework:**
>
> **v2.0.50727**
>
> **Contact your application publisher for instructions about obtaining the appropriate version of the .NET Framework.**

The LFPF software can be started simply by double clicking LFPF.exe in Windows Explorer.

"All models are wrong; some models are useful" George E.P. Box

Table of Contents

1 The use of an extended logistic function for systematically analyzing interface profiles

This document describes the use of an extended logistic function for systematically estimating the width and asymmetry of interfaces between dissimilar materials as measured, for example, by depth profile analyses. Specifically, it describes the rationale for the choice of this particular function as an empirical description of an interface profile, how to use the function in a least squares fit of the function's parameters to a measured profile, and how to interpret the statistics associated with the least squares fit.

1.1 Depth Profiles

The logistic function in its simplest form is given by $Y = \frac{1}{1+e^{X}}$. As X varies from -∞ to +∞, Y varies from 1 to 0 with a sigmoidal shape.

That the logistic function might provide a reasonable representation of an interface is suggested by the following argument. If we represent an interface between spheres labeled A and B as in Figure 1-1 to the right., the probability that an exchange of two neighboring spheres in a horizontal direction will result in the interchange of an A sphere and a B sphere is

$$P_{AB} = kf_A f'_B = kf_A(1-f'_A),$$

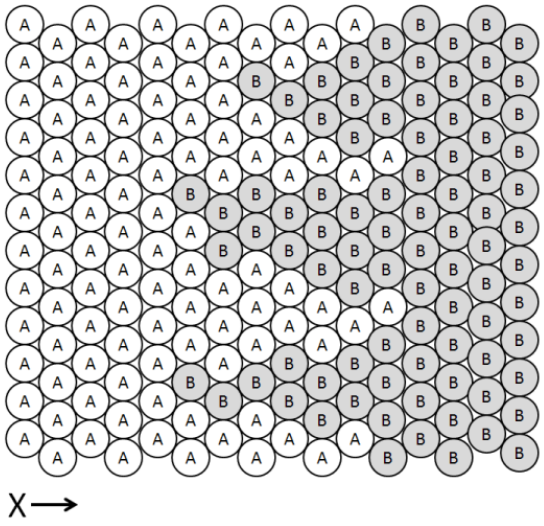


Figure 1-1 Cartoon of an imagined interface

where f_A is the fraction of A in a particular layer at X, f'_A and f'_B are the fractions of A and B in the neighboring layer X + δX, and k is some measure of the propensity for exchange.

This, plus the fact that at some distance from the interface the material is either pure A or pure B, *suggests* that the change in f_A as a function of X can be expressed as

$$\frac{df_A}{dX} = kf_A(1-f_A) \tag{1-1}$$

which, upon integration, gives

$$f_A = \frac{1}{1+e^{-kX}} \tag{1-2}$$

Since k will have the units of 1/X, we can replace k by 1/D. Furthermore, if Y is an instrument response to a measurement of species A so that Y is proportional to the fraction f_A, then Y will be given by

$$Y = \frac{A}{1+e^{(X-X_0)/D}} \tag{1-3}$$

where X_0 is the midpoint of the interface where Y = A/2. Y varies from A to 0 through the interface. The parameter D is seen to be a scaling parameter that defines the width of the interface. As D→0, the profile of Y approaches a step function.

The scaling parameter itself may not be constant. If the spheres in the cartoon above were, for example, of a different size, the rate of change in f_A with X might well vary with X. If we allow D to vary logistically with its own scaling factor, for example,

$$D = \frac{2D_0}{1+e^{Q(X-X_0)}} \tag{1-4}$$

the sigmoidal shape will be sharper at one side of the interface than the other.

Equation (1-3) and (1-4) can be further generalized to

$$Y = \frac{A+A'(X-X_0)+A''(X-X_0)^2}{1+e^{(X-X_0)/D}} + \frac{B+B'(X-X_0)+B''(X-X_0)^2}{1+e^{-(X-X_0)/D}}, \tag{1-5}$$

where the instrument response of the species of interest is allowed to vary with time (and therefore with X) as $A+A'(X-X_0)+A''(X-X_0)^2$ and, where a background signal remains when the species of interest is depleted, as $B+B'(X-X_0)+B''(X-X_0)^2$. The value of Y thus varies from $A+A'(X-X_0)+A''(X-X_0)^2$ to $B+B'(X-X_0)+B''(X-X_0)^2$. In practice, A'' and B'' are almost never included, A' or B' is occasionally included, and the baseline A or B can often be held fixed at 0.

Substitution of Equation (1-4) into Equation (1-5) results in the extended logistic function.

In addition to the three parameters that define the interface region, X_0, D_0, and Q, the interface can also be characterized *independently of an assumed functional form* by a width W and an asymmetry η (to be distinguished from the asymmetry *parameter* Q) in the following way. We define the width as beginning where the interface is some fraction, f, of the distance between the pre-interface asymptote and the post-interface asymptote, and ending where the interface is the fraction $(1 - f)$ of the distance between the two asymptotes. This is particularly useful when the beginning and ending points of the interface are ambiguous or difficult to determine. If we designate the corresponding values of X as X_f and $X_{(1-f)}$ then

$$W = X_{1-f} - X_f . \tag{1-6}$$

The asymmetry η (as opposed to the asymmetry *parameter* Q – repetition added for emphasis) is defined as the skewing of X_f and $X_{(1-f)}$ about the center of the interface X_0, namely,

$$\eta = \frac{2X_0 - (X_{1-f} + X_f)}{X_{1-f} - X_f} . \tag{1-7}$$

(With this definition of η, η and Q have the same sign.) Clearly, if X_f and $X_{(1-f)}$ are equally spaced about X_0, $\eta = 0$. To emphasize the point that W and η are functions of the choice of f, they can be designated as W_f and η_f. W_f and η_f can be calculated graphically from the

profile using rulers or can be related to the interface parameters of the extended logistic function through

$$X_f = X_0 - \frac{2D_0}{1+e^{Q(X_f - X_0)}} \ln\left(\frac{1-f}{f}\right) \text{ and } X_{1-f} = X_0 + \frac{2D_0}{1+e^{Q(X_{1-f} - X_0)}} \ln\left(\frac{1-f}{f}\right) \quad (1\text{-}8)$$

Depth profile measurements are complex processes (see, for example, S. Hoffman, Rep. Prog. Phys. 61 (1998) 827–888 and references quoted therein.) This introduction in no way should suggest that the extended logistic function is being advocated as an atomic scale model for describing depth profile analyses. It merely provides a rationale for the use of the logistic function as a convenient and reasonable means for estimating the position, width and asymmetry of the interfacial profile in a systematic fashion from a set of discrete measurements.

1.2 Line Scans

In the case of line scans of materials deposited on surfaces or of surfaces with discrete regions of differing composition, Fig. 1-1 could be thought to represent an interfacial region between two surface areas with different compositions. However, one important purpose of performing surface line scans is the determination of the lateral resolution of the measuring instrument when the transition region between the two materials is smaller than the expected lateral resolution. For particle or photon beams incident on a suitable test surface, the lateral resolution can be usefully determined from a line scan over a sharply defined interface. In a so-called knife-edge measurement, the beam intensity is measured as a function of the position of a knife edge partially or fully occluding the beam and thus represents the intensity distribution from zero to maximum intensity. Such a measurement is equivalent to measuring the response of a beam (or for that matter an Atomic Force Microscope tip) scanned over a sufficiently sharp interface between two materials. The measured response is a convolution of the point spread function of the beam and the object being measured (M. Senoner, T. Wirth and W. E. S. Unger, J. Anal. At. Spectrom.., 2010, 25, 1440–1452.) The line scan, sometimes referred to as the line spread function, has a sigmoidal shape representative of the underlying point spread function. The width of the line spread function can be taken as a measure of lateral resolution. Current practice, as recommended in ISO 18516: 2006, Surface chemical analysis - Auger electron spectroscopy and X-ray photoelectron spectroscopy - Determination of lateral resolution (International Organization for Standardization, Geneva, 2006), calls for the width to be defined as various percentages (12% to 88%, 16% to 84%, 20% to 80%, or 25% to 75%) of the maximum intensity in the line spread function. The choice of percentages for determining lateral resolution from the line scan reflects assumptions about the underlying point spread function as well as effects from the test sample (e.g., the role of backscattered electrons in Auger electron spectroscopy). Most prominent among the functions used to describe point spread functions are the Gaussian (12% to 88% for the width at half height and 16% to 84% for the 1σ width) and Lorentzian, (25% to 75% for the width at half height) though sums of Gaussians and pseudo-Voigt profiles have been considered along with the so-called top hat model (uniform beam intensity across a circular profile).

The point spread function is the derivative, dY/dX of line scan measurement across the interface and the width at half height is a measure of the lateral resolution. From the logistic

function parameters derived from the fit of the line scan measurements, the derivative, dY/dX, can be calculated from which the value of X where dY/dX is a maximum, $X_{\max}$,can be determined as can the two values of X, X_- and X_+, where dY/dX is half its maximum value. From these we can define a width and asymmetry analogous to the definitions of W_f and η_f in Eq. (1-6) and Eq. (1-7) above, namely

$$W_{hh} = W_+ - W_- \tag{1-9}$$

$$\eta_{hh} = \frac{2X_{\max} - (X_+ + X_-)}{(X_+ - X_-)} \tag{1-10}$$

If $Q = 0$, then $X_{\max} - X_- = X_+ - X_{\max} = 2\ln\left(3+\sqrt{8}\right)D_0$ and $X_{\max} = X_0$ for the symmetric profile. If $Q \neq 0$, $X_{\max}$, X_- and X_+ must be solved numerically. For symmetric profiles where $Q = 0$, the width at half height is 14.64% to 85.36% and these are the default values used in LFPF.

An unusual feature of the point distribution function based on the extended logistic function is that η_{hh} has a maximum value of 0.2589 at $QD_0 = 0.8346$ and $X_{\max} - X_0$ a maximum value of $0.7587\,D_0$ at $QD_0 = 0.829$. Similarly, η_{hh} has a minimum value of -0.2589 at $QD_0 =$ -0.8346 and $X_{\max} - X_0$ a minimum value of $-0.7587\,D_0$ at $QD_0 = -0.829$. In effect this limits the asymmetry of the point distribution function and the uncertainties returned by the least squares fit for these parameters should take these limits into account.

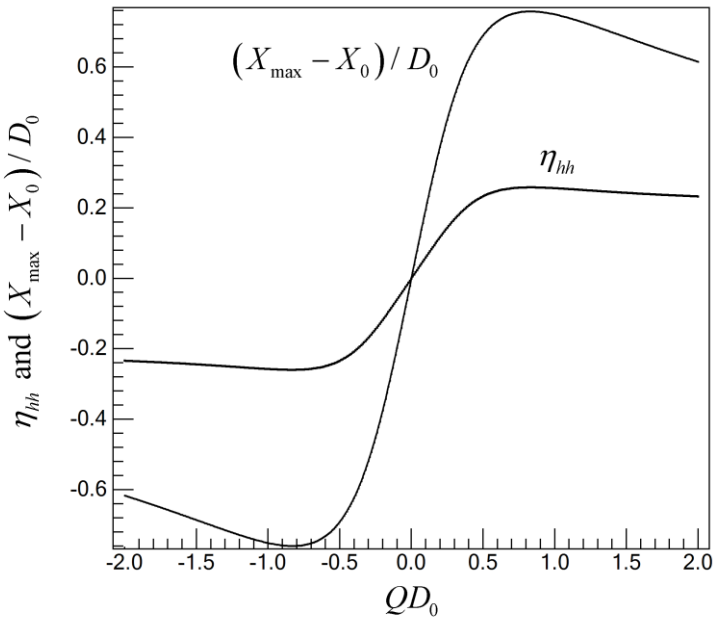


Figure 1-2 Asymmetry and position of the point distribution function as a function of QD_0

In practice, the line spread function is the result of unknown or incompletely known factors and can be expected to deviate from any particular mathematical function such as a Gaussian.

The extended logistic function is thus no more than another possible function to use in fitting a measured line scan to provide a suitable measure of lateral resolution. A measured line scan typically consists of intensity measurements at discrete steps (e.g., of a beam across a knife edge or a sharp compositional interface), and use of a suitable fitting function is convenient for obtaining an objective measure of lateral resolution given the finite steps and the presence of statistical noise. The logistic function has slightly longer tails than a Gaussian function with the same width (whose integral is the error function) though not nearly as long as the

Lorentzian (whose integral is an arctangent function). When integrals of Gaussian, Lorentzian, or pseudo-Voigt functions are fitted with the extended logistic function, the residual standard deviation is less than 1.5% of the maximum intensity. Therefore, the choice of which function to use is at present mostly one of ease and convenience. The inclusion of an asymmetry factor in the extended logistic function is a further indicator of the shape of the point spread function.

2 A program for obtaining a least squares fit of an extended logistic function to a measured profile

A computer program for fitting an extended logistic function to depth profile measurements has been written in Microsoft Visual Basic.Net. The following discussion serves as a "user manual" for that program. Many features have been added to the program beyond fitting interfacial profiles, mostly for the benefit of program development and the testing and interpretation of the profile fits. While some of these features may be of limited interest to the average user, we have decided retain them so that those who may be more concerned about details of the fitting process can do their own testing. While this decision may leave the analysis options more extensive than necessary for many users, the display has been designed to be as simple and intuitive as possible. In short, the name of a data file is entered, an "OK" button is clicked after the data are listed in the LFPF window, and when the graph of the data is displayed, a button labeled "Fit (Converge)" is clicked and that is it. The following is a description of the operation of all the program features.

An extensive Help file (LFPFHelp.chm) accompanies the program which contains most of the information contained in this documentation, albeit in very abbreviated form. Several data files, most notably BgtAQp 25.txt (*B* greater than *A*, *Q* positive with 25 data points), are included with the program and the analyses of these data are described in this documentation.

2.1 Program Startup

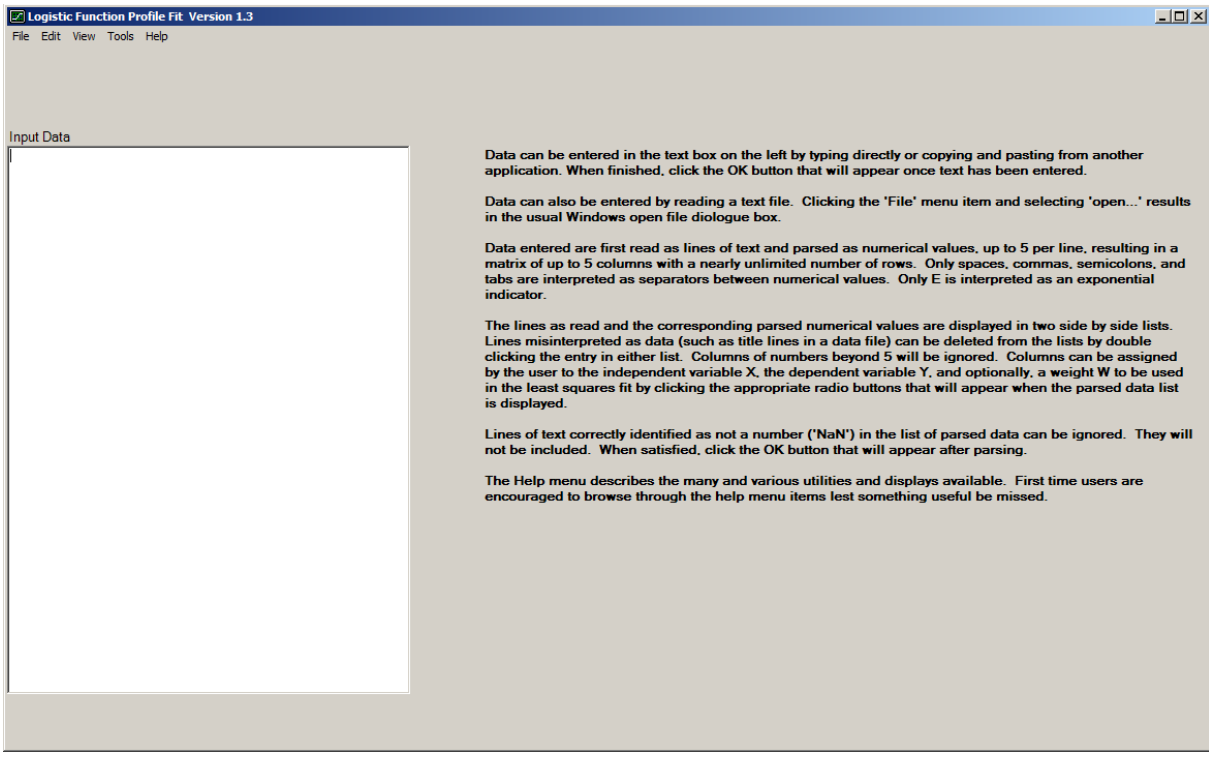


Figure 2-1 Logistic Function Profile Fit (LFPF) opening display

The program is run in the usual Windows manner, either by double clicking the file name

"LPFP.EXE" in Windows Explorer, or Clicking Start, Run, and entering "LPFP.exe" with its full path name, or clicking a shortcut icon to LPFP.exe on the desktop.

Note: The first time LFPF is run, two directories are created by the program. The first, \NIST\LFPF is created in the user's APPS directory (an often hidden directory), either C:\Users\<*username*>\AppData\Roaming\NIST\LFPF\<*version number*> or C:\Documents and Settings\<*username*>\Application Data\NIST\LFPF\<*version number*> depending on the Windows operating system. This directory is used to contain a text file with the list of the five most recent data files opened. The second directory created by LFPF is in the user's \Documents directory and is used as a fall back default directory for various program outputs as described later in this document. If either of these directories is erased, it is re-created the next time LFPF is run.

The program begins with a window that displays instructions and a text box, as in Figure 2-1, into which data from another application can be copied and pasted. Alternatively data could be entered directly into the text box from the keyboard. Up to five entries per line can be accepted which can be assigned, once entered, to the independent variable X, the dependent variable Y and optionally a weighting factor W. The data entries in each line can be separated by tabs, spaces, commas, or semicolons. Separators used in combination such as a comma followed by a space, or several spaces together, are considered as a single separator. Spaces are always interpreted as separators except when used in exponential notation such as nnn.nn Emm where the space before the E is ignored. As soon as any entry is made an OK button and a Cancel button appear to the right of the values entered as in Figure 2-2 below.

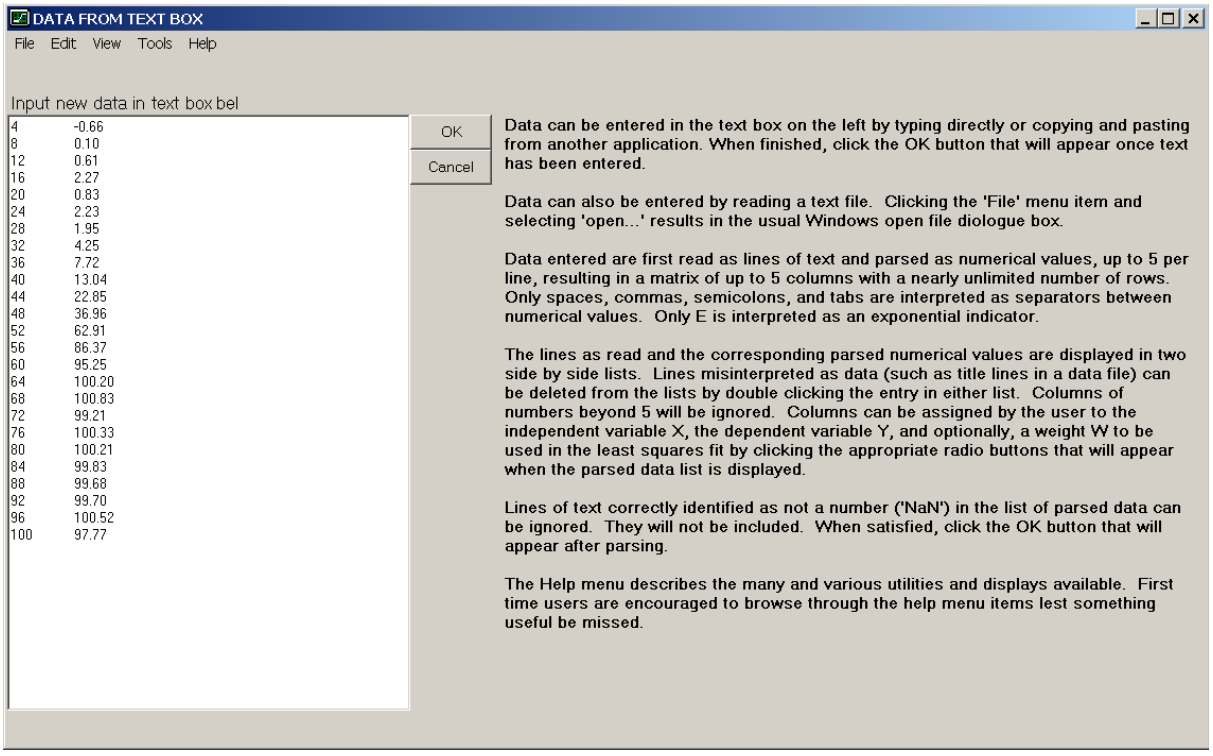
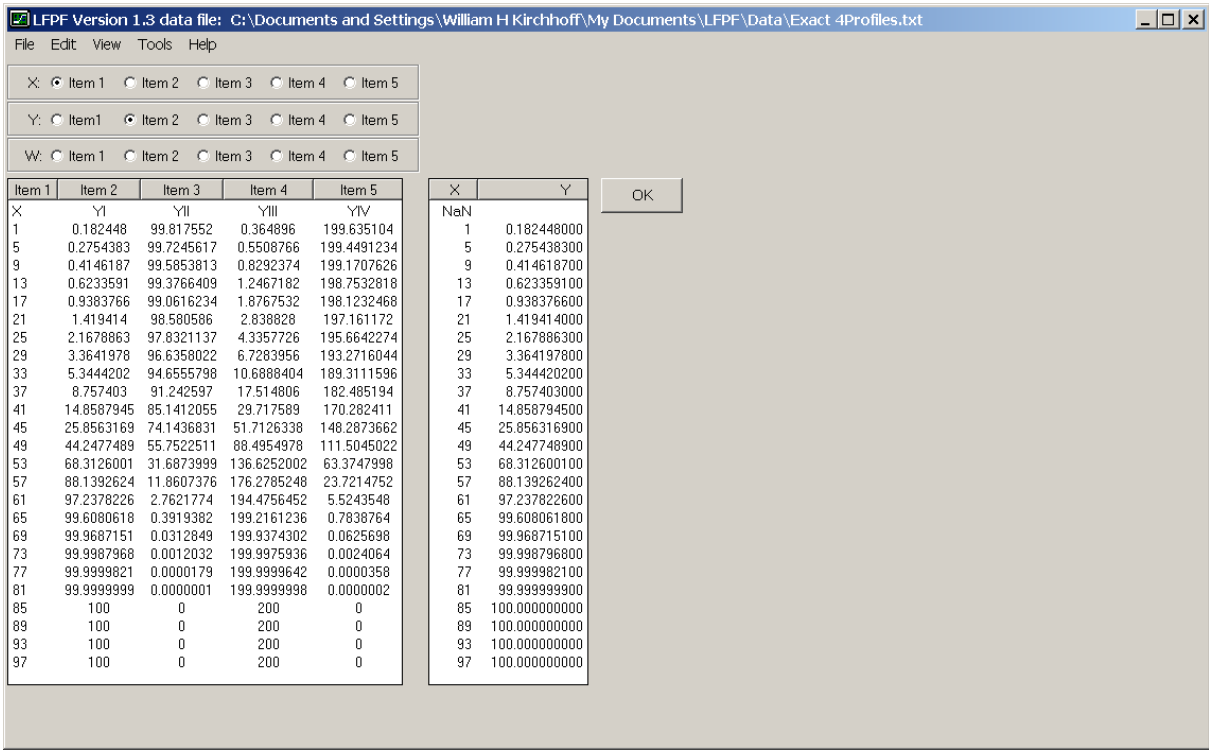


Figure 2-2 Entering data in the text box of the opening display

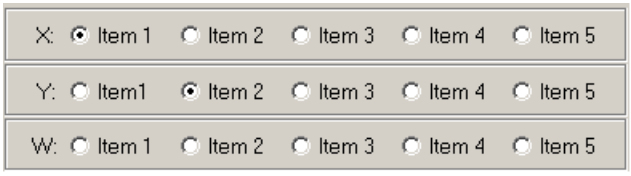


Item 1	Item 2	Item 3	Item 4	Item 5
X	YI	YII	YIII	YIV
1	0.182448	99.817552	0.364896	199.635104
5	0.2754383	99.7245617	0.5508766	199.4491234
9	0.4146187	99.5853813	0.8292374	199.1707626
13	0.6233591	99.3766409	1.2467182	198.7532818
17	0.9383766	99.0616234	1.8767532	198.1232468
21	1.419414	98.580586	2.838828	197.161172
25	2.1678863	97.8321137	4.3357726	195.6642274
29	3.3641978	96.6358022	6.7283956	193.2716044
33	5.3444202	94.6555798	10.6888404	189.3111596
37	8.757403	91.242597	17.514806	182.485194
41	14.8587945	85.1412055	29.717589	170.282411
45	25.8563169	74.1436831	51.7126338	148.2873662
49	44.2477489	55.7522511	88.4954978	111.5045022
53	68.3126001	31.6873999	136.6252002	63.3747998
57	88.1392624	11.8607376	176.2785248	23.7214752
61	97.2378226	2.7621774	194.4756452	5.5243548
65	99.6080618	0.3919382	199.2161236	0.7838764
69	99.9687151	0.0312849	199.9374302	0.0625698
73	99.9987968	0.0012032	199.9975936	0.0024064
77	99.9999821	0.0000179	199.9999642	0.0000358
81	99.9999999	0.0000001	199.9999998	0.0000002
85	100	0	200	0
89	100	0	200	0
93	100	0	200	0
97	100	0	200	0

X	Y
NaN	
1	0.182448000
5	0.275438300
9	0.414618700
13	0.623359100
17	0.938376600
21	1.419414000
25	2.167886300
29	3.364197800
33	5.344420200
37	8.757403000
41	14.858794500
45	25.856316900
49	44.247748900
53	68.312600100
57	88.139262400
61	97.237822600
65	99.608061800
69	99.968715100
73	99.998796800
77	99.999982100
81	99.999999900
85	100.000000000
89	100.000000000
93	100.000000000
97	100.000000000

Figure 2-4 Comparing data as read with data as parsed

The data in the text box can be edited as with any text editor. Clicking the Cancel button erases the entries in the text box. Clicking the OK button clears the window and displays two lists as in Figure 2-4 above.



Instead of copying and pasting, data contained in text files can be read by clicking "open…" in the file menu, whereupon the usual Windows open file dialog box appears (Figure 2-3). Once the file is opened, the text box is replaced with two lists as in Figure 2-4. Note that the title bar of the window now contains the name of the data file.

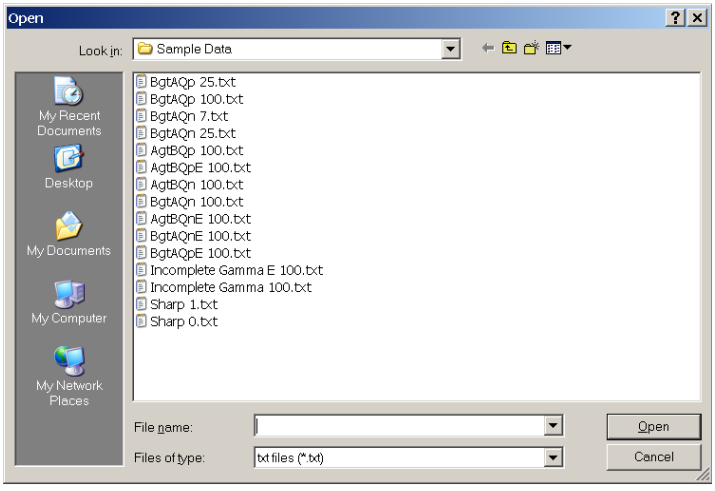


Figure 2-3 Open File dialog box

The list on the left in Figure 2-4 contains the unparsed data as entered. The list on the right contains the data as parsed and assigned to X and Y based on the most recent assignment used.

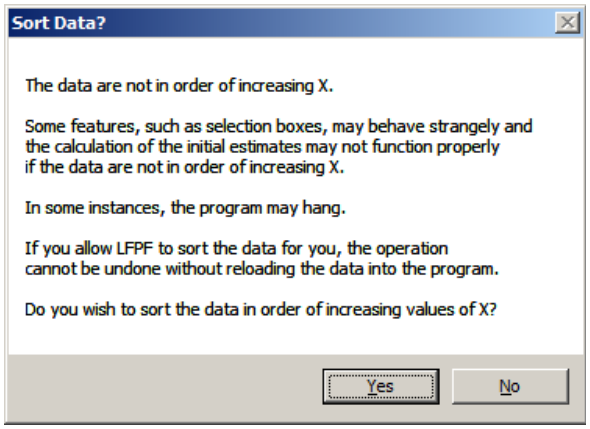
The radio buttons above the unparsed list indicate which item is X and which is Y (and, optionally, which is the weighting factor W.) Up to five items per line of text can be accepted and interpreted. If the data file contains only two items of data per line of text, only the option

of identifying which entry is X is given. At this point, individual entries, such as titles, as in the top line of the two lists in Figure 2-4 can be deleted by double clicking the corresponding entry in either list. If an entry is identified as "NaN", i.e., not a number, as in the parsed, assignment list in Figure 2-4, it will be eliminated from the data table automatically when the OK button is pressed.

It is important to emphasize the role of data separators when entering data or when reading data as lines of text in a text file. Data separators can be spaces, commas, and semicolons and are always interpreted as such. Combined separators such as a comma or semicolon followed by a space or several spaces together are considered to be a single separator. In addition, in text files, tabs are considered to be separators and will appear as tabs in the unparsed list on the left of Figure 2-4. The only time a space is not interpreted as a separator is when it precedes an E in numbers using exponential notation nnnnn.nn Emm. Commas appearing as thousands markers will be ignored so that if they are used as data separators they should be followed by a space.

Once the data as interpreted are deemed correct, clicking the OK button clears the window and replaces it with the data analysis display which includes a graph of the data, the list of data, a list of the extended logistic function parameters, buttons to initiate the least squares fit, and additional parameters associated with the fit as shown in Figure 2-5 on the following page.

Note: For the routine that calculates the initial estimates of the parameters to work correctly, the data may have to be in order of increasing X. Consequently, the data are tested and if not in order of increasing X a warning message is printed and the option of sorting the data is offered before the analysis display appears.



In the list of the logistic function parameters in the lower left hand side of the window shown in Figure 2-5 , only those parameters whose boxes are checked will be evaluated by the least squares fit. Any parameter can be held at a fixed value entered by the user. The default parameters to be evaluated are A, B, X_0, D_0, and Q. Unless the check box below the list of parameters, labeled "Permit A´ and/or B´" is checked, the routine determining the initial estimates does not attempt to give them values other than 0. *Additionally, the parameters accommodating curvature in the pre and post interface asymptotes, A˝ and B˝, cannot be varied until at least one least squares fit of the data has been performed in which they are not varied*. The reason for doing this is to discourage the addition of parameters merely to improve the residual standard deviation.

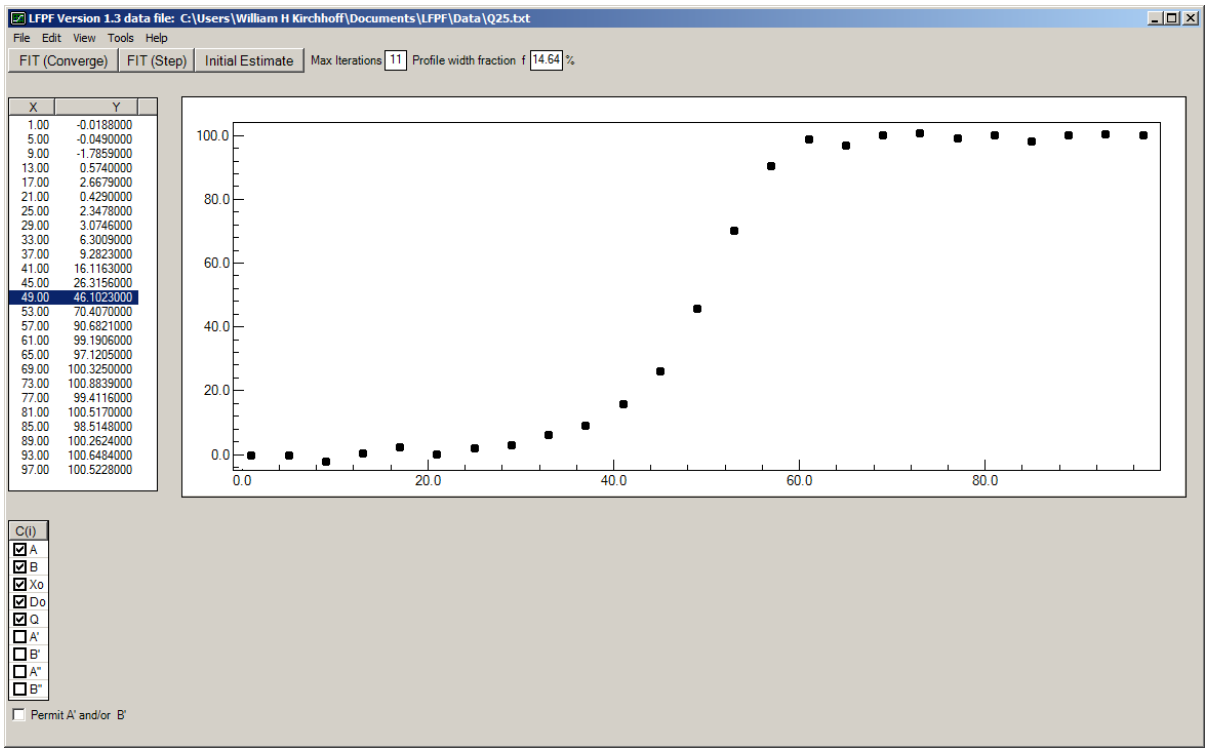


Figure 2-5 Initial graphical display of data along with the analysis options and parameter list

2.2 Data Selection and Identification

Data can be identified by clicking the individual data points in the graph. When a point on the graph is clicked with either mouse button, a crosshair appears at that point and the values of X and Y are printed at the top of the graph. If clicked near a data point, the crosshair is moved to that data point which is then highlighted in red while the corresponding values of X and Y are highlighted in the data list. Similarly, if an entry in the data list is clicked, it is highlighted and the corresponding point on the graph is marked with a crosshair. If the data list is active (the selected item in the list is highlighted) the cursor keys (up, down, left, right) move the selected entry up and down the list and the crosshairs to the previous or next point in the graph. **If a data point on the graph is double clicked, its display changes (dimmed, replaced by a single screen pixel, or replaced by an X, See Section 2.6.18 View > Ignored Data) and it is ignored in the least squares fit of the data.** If an ignored data point is double clicked, its display returns to normal and it is subsequently included in the least squares fit of the data. The entries for ignored data in the displayed list are dimmed. The delete and insert keys also mark data as ignored or not. **A range of data may be selected by employing the data selection box (See Section 2.6.10 below, View > Data Selection Box.)** If a Data Selection Box is displayed on the screen, the delete key will mark all the data in the box as ignored. The insert key will restore them.

If a least squares fit of the data to the extended logistic function has been performed, the calculated value of the function is drawn on the graph. When the calculated values are displayed, clicking on a data point displays the value of Y(observed) – Y(calculated) along with

the confidence interval (not the standard deviation) of that *difference*. If any point other than a measured point is clicked, the calculated value of Y for the selected value of X is printed on the top line along with its confidence interval (not its standard deviation.) The confidence interval is the standard deviation multiplied by the t distribution confidence limit for the selected confidence level (95% default.)

2.3 The Least Squares Fit

The least squares fit minimizes the sum $\sum_{i=1}^{n} W_i (Y_i^{obs} - Y_i^{calc})^2$ where Y_i^{obs} are the measured values of the profile and Y_i^{calc} are the values calculated from Equation (1-5). If the weights, W_i, are proportional to the inverse square of the standard deviation of the measured values, then the sum of the squares should follow a chi square distribution. $s^2 = \sum_{i=1}^{n} W_i (Y_i^{obs} - Y_i^{calc})^2 / (n-m)$ is the estimate of the variance (square of the standard deviation) of the normally distributed errors in *Y*. The variance of a particular measurement, Y_i is s^2 / W_i. s^2 is the variance of a measurement with unit weight.

The three buttons shown in Figure 2-5, FIT (Converge), FIT (Step) and Initial Estimate, control the least squares fit of the data to the extended logistic function. The function is non-linear in the parameters and the least squares fit is based on an iterative Newton-Raphson linearization of the function, that is, a Taylor series approximation cutting off at the linear term as described in Section 4 of this document. Each iteration calculates corrections to the parameter values. The rapidity of convergence, indeed whether the procedure converges at all, depends on the quality of the initial estimates of the parameters. The calculation of initial estimates is also discussed in Section 4 of this report. Briefly, the curvature parameters for the asymptotes, A″ and B″, are always assumed to be 0 and are not varied in the analysis unless explicitly requested by checking their boxes in the parameter list, and only after a least squares analysis of the data has been performed at least once. Initial values of the slopes of the asymptotes, A′ and B′, will be calculated only if the box labeled "Permit *A′* and/or *B′*" is checked and only if it appears that their values differ significantly from 0. If the fit is unstable, preference is always given to evaluating Q over evaluating the slopes of the asymptotes with which the value of Q is usually highly correlated. The remaining parameters are given initial estimates by examination of the data, identifying the asymptotic regions and the interface region. If the data are well structured, the initial estimates routine is reasonably robust. By well structured is meant at least 7 data points for which each asymptote has at least two values within 5% of its limiting value, and at least three values within the interface region that lie more than N standard deviations away from each asymptote, where N is the normal distribution confidence limit. The confidence limits for both the normal and the student's t distribution are calculated by the program.

When Initial Estimate is clicked, the starting values of the parameters are estimated as described in Section 4.1 of this documentation and are reported and graphed on the analysis display as seen in Figure 2-6 below. Next to the parameter names are the initial estimates of their values. The "Data Scatter" is a model independent measure of the noise in the data (See Section 2.6.3, View > Data Scatter) and is estimated for the purpose of comparison with the standard deviation returned by the least squares fit of the data to the extended logistic function. A standard deviation significantly greater than the data scatter indicates the likely influence of model errors. Since the extended logistic function is an empirical representation of the interface, it should always be assumed that model errors will likely be present along with random measurement errors.

The Residual Standard Deviation is that calculated from Equation (4-12), namely,

$s = \sqrt{\sum_{i=1}^{n} W_i (Y_i^{obs} - Y_i^{calc})^2 / (n-m)}$, where n is the number of data points, m is the number of fitting parameters, W_i are the weights of each datum (= 1 if weights are not included in the data file) and Y_i^{calc} are calculated from the extended logistic function using, in this case, the initial estimates of the parameters. In the case of the initial estimates, model errors are expected from the approximate nature of the initial, estimated values of the parameters and, as seen in Figure 2-5, the data scatter was 1.389 compared with the calculated standard deviation of 1.759. This comparison is the only indication that the parameter values have not yet been optimized. The fit of the initial values of the parameters to the data appears to be quite good. Inspection of the residuals, $Y_i^{obs} - Y_i^{calc}$, shows no systematic trend. (Residuals can be displayed by clicking the

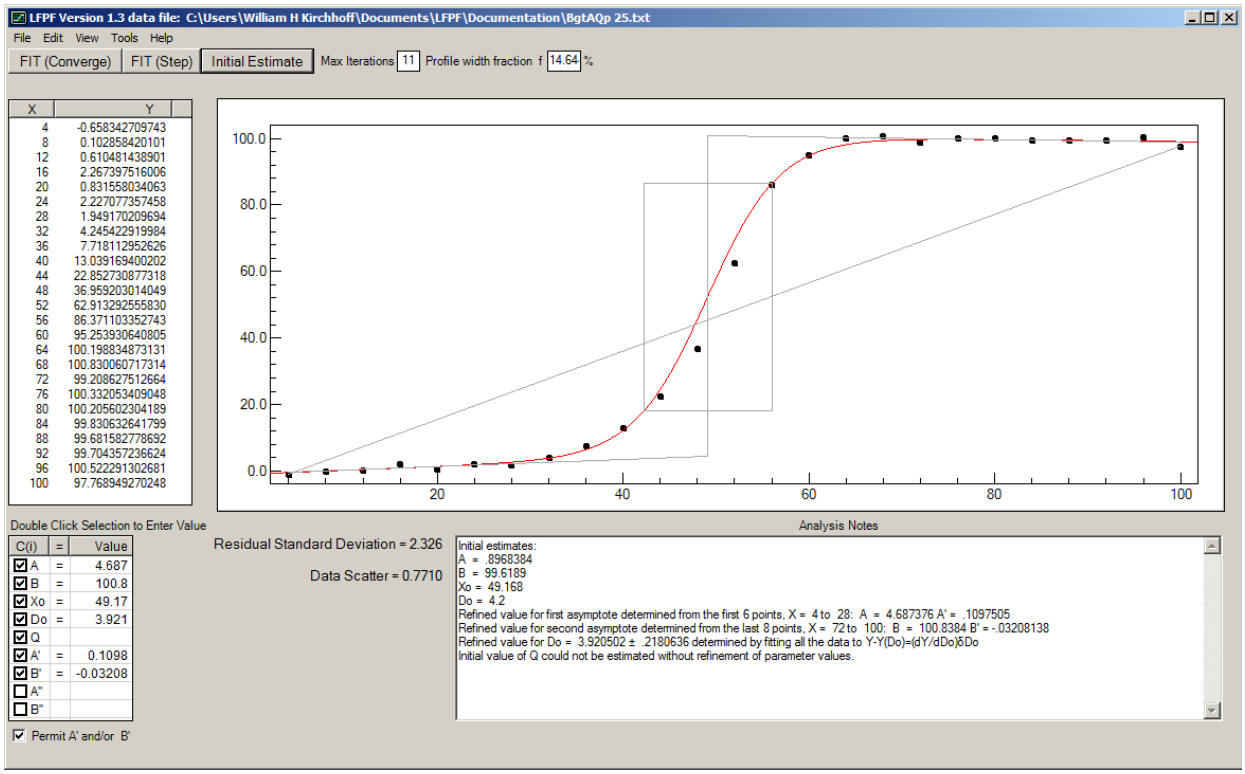


Figure 2-6 Display of initial estimates of the extended logistic function parameters

Residuals item in the View menu as described below.) The initial estimates are arrived at by using a variety of techniques depending on the structure of the data as described in Section 4.1. (See also Section 2.6.10 View > Data Selection Box.) Notes on the calculation of the initial estimates appear in the text box labeled "Analysis Notes" in the lower right of the window such as the following:

```
Initial estimates:
 A  =  .3028667
B  =  99.9795
Xo =  49.144
Do =  4.2
Refined value for first asymptote determined from the first 7 points, X =  1 to  29:  A  =  5.2112 A' =  .1261202
Refined value for second asymptote determined from the last 8 points, X =  69 to  97:  B  =  100.0594 B' =  .002253869
Refined value for Do = 3.894566 ± .1834962 determined by fitting all the data to Y-Y(Do)=(dY/dDo)δDo
Initial value of Q = .01103739 ± .0105855 determined by fitting all the data to Y = (dY/dQ)Q
```

Because the check box labeled Permit A′ and/or B′ below the list of parameters and their values was checked, the slope of the initial baselines, A′and B′, were determined and found to be significantly different from 0.

The central box on the graph in Figure 2-6 defines the interface region used for the initial estimate of D_0 as determined by the program. For poorly structured data, this region can be controlled by the user by clicking the "Select Data Box" item in the View menu. This will be discussed below in Section 2.6.10.

Two small text boxes labeled "Max Iterations" and "Profile Percentage limit" appear on the display:

Max Iterations: The iterative fit is curtailed at the specified maximum number of iterations if it has not yet converged. A prime number for the maximum number of iterations is desirable to identify when the least squares fit is oscillating between two neighboring minima. A value of 11 seems to be adequate for most cases tested and is the default value. If convergence is not reached, re-clicking the FIT (Converge) button repeats another round of iterations beginning with the current values.

Profile width fraction f: The profile width fraction f defines the reported width and asymmetry of the interface. Because of the exponential nature of the extended logistic function, the asymptotes are never reached. The reported width of the interface is therefore taken as the spread in X from the value at which the interface is f percent complete to the value at which the interface is (1-f) percent complete. Sigmoidal depth profiles were originally fit to error functions as a way of parameterizing their width so that the values of X corresponding to $x = \pm \sigma$ in the normal probability function were used as a measure of the width. Other measures have included 12% and 88% for the width at half height of a Gaussian function or 25% and 75% for the width at half height of a Lorentzian function (the integral of which is the arctangent function.) The distribution function underlying the logistic function ($dY / dX = \left[\left(1+e^{z}\right)\left(1+e^{-z}\right)\right]^{-1}$) has a width at half-height of 14.64% $\left(1/\left(4+\sqrt{8}\right)\right)$ to 85.36% and this is taken to be the default value for the profile width fraction f. In LFPF any value between 0 and 50 can be entered into the box labeled "Profile width fraction f." If 0 is entered (the width would be infinite), the default value of 14.64 is restored. This measure of the width, i.e. between f and 1-f, is somewhat insensitive to, though not completely independent of, the functional form used to represent the interface profile. The

width and asymmetry values along with their confidence limits are printed in the Analysis Notes following the least squares fit.

When the FIT (Converge) button is clicked, the values of the parameters are iteratively refined until convergence is achieved or the maximum number of iterations is reached, following which the display will resemble Figure 2-7. The tests for convergence are based on the changes in the values of the parameters compared with their standard deviations and on changes in the standard deviation of the fit. Convergence is declared when the following occurs: (1) the corrections to the parameters are all less than 1 percent of the values of their standard deviations and (2) the standard deviation does not change from one iteration to the next by greater than 1 part in a thousand. In some instances, most notably when exact data are being fit, the convergence limit may never be reached because of round off errors.

Clicking the FIT (Converge) button always begins the iterative procedure starting with the *current* values of the parameters. If no initial values have been estimated, they are first estimated as if the Initial Estimate button had been clicked. To restart from scratch, the Initial Estimate button must first be clicked.

Clicking the FIT (Step) button performs one iteration of the least squares fit beginning with the current estimates of the parameters, which could be the initial estimates. Repeated clicking of the FIT (Step) button differs from the Fit (Converge) button in that the slope parameters, A´ and B´, and the curvature parameters, A″ and B″, are evaluated (if their boxes are checked) even if their values do not differ significantly from 0. The Fit (Converge) button will set these parameters to 0 if their confidence limits include 0 as in Figure 2-7.

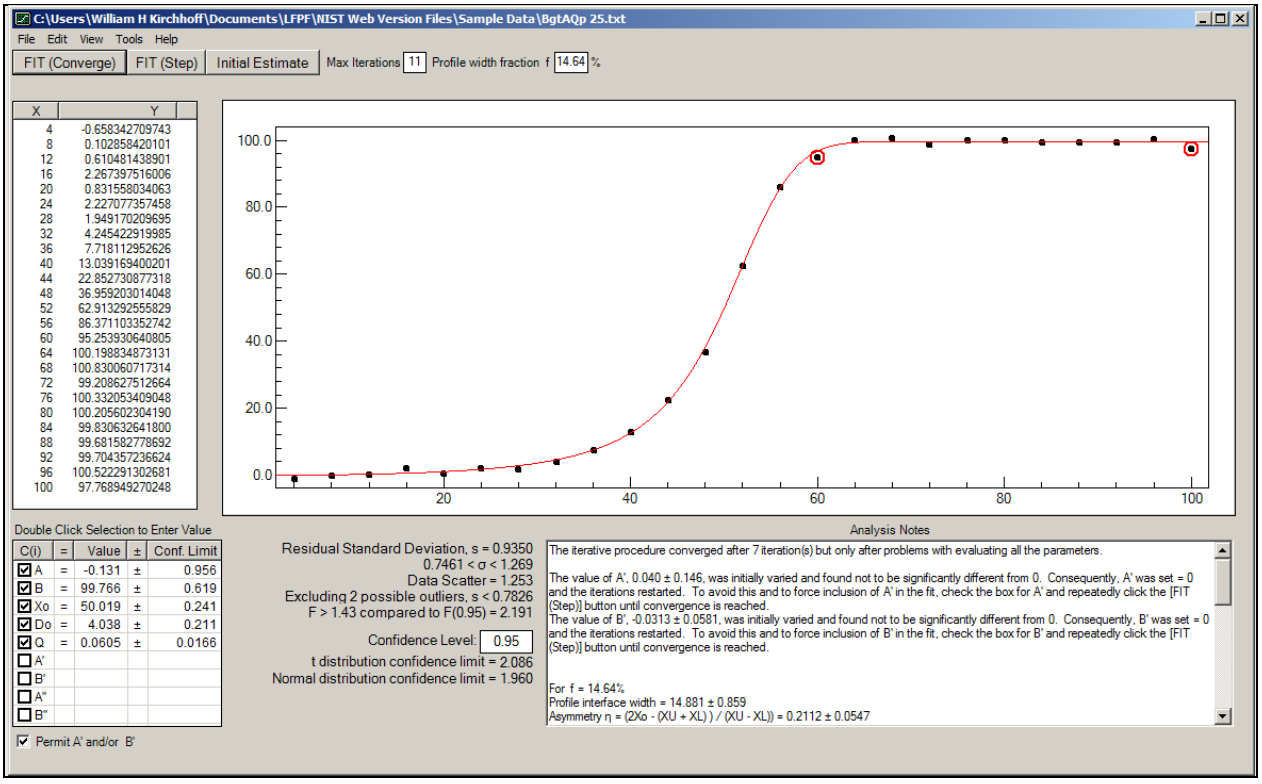


Figure 2-7 Results of a least squares fit of the profile data to the extended logistic function

2.3.1 Wild Excursions, Divergences and Instabilities

Primarily because FIT (Converge) and FIT (Step) always proceed with the *current* values of the parameters, instabilities that occur with poorly structured data (see Section 3.2 below) can get out of hand and drive the parameter values so far from their least squares values that the program more or less freezes on values far from the convergent solution in order to avoid crashing. The many procedures incorporated into the program to deal with unstable or diverging situations do not work for all eventualities. In such cases it will usually be necessary to click the Reset All item in the Tools menu and start afresh. It would then be advisable to proceed by clicking Initial Estimate and continue step by step to determine where the instability appears and which parameters must be held at some fixed value selected by the user.

2.4 Parameter Values, Associated Statistical Statements, and Analysis Notes

Following Fit (Converge) or FIT (Step), the graph of the extended logistic function

$$Y = \frac{A + A'(X - X_0) + A''(X - X_0)^2}{1 + e^{(X - X_0)/D}} + \frac{B + B'(X - X_0) + B''(X - X_0)^2}{1 + e^{-(X - X_0)/D}} \quad \text{where} \quad D = \frac{2D_0}{1 + e^{Q(X - X_0)}}$$

is drawn on the graph of the data as in Figure 2-7.

The values of the parameters are printed with their confidence limits. The confidence limits are based on the confidence level which is under the control of the user. (Default value = 0.95) If the measurement errors in the values of Y are normally distributed, the values of the determined parameters should follow a student's t distribution. The confidence limits reported for the parameters are calculated by multiplying their standard deviations returned by the least squares fit by the value of t, labeled on the display "t distribution confidence limit," that satisfies the stated confidence level entered in the box labeled "Confidence Level" (see Equation (4-33) and accompanying discussion). **Note: The values for the normal distribution confidence limit and the t distribution limit, as reported in the LFPF program are both two-tailed limits. If the confidence level is 95% then 2.5% of the distribution fall above the confidence limit and 2.5% fall below the negative value of the confidence limit.** The value of t will depend only on the number of degrees of freedom (number of data being fit minus the number of parameters varied) and the confidence level. The confidence limit for the normal distribution does not depend on the number of degrees of freedom and it is the limiting value for t as the number of degrees of freedom approaches infinity and the student's t distribution approaches the normal distribution. Note that the errors in the parameters are correlated, the values and correlations being contained in the so-called variance-covariance matrix (Equation (4-15).) **It cannot be stressed often enough that the confidence limits are based not only on the assumption of normally distributed errors in Y, but also on the assumption that the values of X are error free.**

As mentioned above, the reported Residual Standard Deviation is that calculated from $s = \sqrt{\sum_{i=1}^{n} W_i (Y_i^{obs} - Y_i^{calc})^2 / (n-m)}$ where Y^{calc} is calculated using the parameters returned by the least squares fit and reported in the parameter table.

If the measurement errors follow a normal distribution, the estimate of the variance (square of the residual standard deviation) of a sample of the data will follow a so-called chi-square distribution. **In contrast to the normal distribution confidence limits and the t distribution confidence limits, the chi-square distribution confidence limits used by LFPF to calculate the confidence limits of the residual standard deviation are one-tailed. 5% of the time, the residual standard deviation will fall below the lower confidence limit and 5% of the time above the upper confidence limit when the confidence level is 0.95.** The true value of the population standard deviation will fall between the values of the 95% confidence limits, determined from the sample variance and reported below the Residual Standard Deviation, 90% of the time. (See Equation (4-18) and its accompanying discussion.) In Figure 2-7 those limits are $0.8348 < \sigma < 1.420$. The data scatter, being based only on an estimate of the noise in the data, remains the same as it was in the display for the initial estimates.

Because the "Identify Outliers" item was checked in the View Menu when the FIT(Converge) button was clicked, the expression "Excluding 2 possible outliers, s < 0.7826," appearing in Figure 2-7 below the Data Scatter, is the value of the standard deviation obtained using the most recent values of the parameters but excluding all those data identified as possible outliers from the calculation of the standard deviation though not from the fit itself. If they were excluded from the fit, the standard deviation would be less than the figure quoted because the exclusion would lead to a slightly lower minimum, hence the < sign in the expression. Below that on the display is the F test result comparing the standard deviations with and without the outliers: F > 1.43 compared to F(0.95) = 2.191. In this case, because F < F(0.95), the exclusion of outliers does not lead to a statistically significant drop in the standard deviation. For more discussion on outliers, see Section 2.6.7 View > Identify Outliers

The Analysis Notes give additional information on the analysis, such as:

The iterative procedure converged after 7 iteration(s) but only after problems with evaluating all the parameters.

The value of A', 0.0121 ± 0.0140, was initially varied and found not to be significantly different from 0. Consequently, A' was set = 0 and the iterations restarted. To avoid this and to force inclusion of A' in the fit, check the box for A' and repeatedly click the [FIT (Step)] button until convergence is reached.

For f = 14.64%
Profile interface width = 14.5816 ± 0.0813
Asymmetry η = (2Xo - (XU + XL)) / (XU - XL)) = 0.17475 ± 0.00542
Dimensionless QD0 = 0.20066 ± 0.00640

At XL = 41.43037 the interface region is 14.64% complete
At XU = 56.01193 the interface region is 85.36% complete

The maximum magnitude of dY/dX = 6.481 at X = 51.5013 ± 0.0539
The full width of the derivative dY/dX at half height = 13.2159 ± 0.0853
Corresponding asymmetry η = 0.11788 ± 0.00352

dY/dX = 3.241 at X = 44.11433880 where Y = 21.79325000 (21.77% complete)
dY/dX = 3.241 at X = 57.33025649 where Y = 90.37413506 (90.32% complete)

The number of data included in the fit was 25
The number of parameters varied in the fit was 5 giving 20 degrees of freedom

Corr Coef	A	B	Xo	Do	Q
A	1.0000	0.0622	0.1640	-0.4667	-0.5044
B	0.0622	1.0000	0.1798	0.1999	-0.2512
Xo	0.1640	0.1798	1.0000	-0.2296	0.3340
Do	-0.4667	0.1999	-0.2296	1.0000	-0.1099
Q	-0.5044	-0.2512	0.3340	-0.1099	1.0000

Minimum precision of X = 1 and of Y = 0.000000000001
(Standard Deviation)/(Minimum precision of Y) = 9.45E+10
At the initial point, X = 4, the interface is 0.18% complete.
At the final point, X = 100, the interface is 100.00% complete.

The Analysis Notes, along with the values of the parameters and the statistics of the fit are automatically copied to the Windows clipboard for pasting in other applications.

The analysis notes include the number of iterations performed, various warning messages, information on the interface width and asymmetry, the number of data included in the fit, the number of parameters varied, the number of degrees of freedom used in calculating confidence limits for the selected confidence level, additional information on the distribution of residuals and scatter, and the correlation coefficients among the parameters varied in the fit. The notes conclude (not shown in the sample) with the minimum precision for X and Y determined from the data appearing in the unparsed list, the ratio of standard deviation to minimum precision which indicates whether the precision of the data is limiting the accuracy, and finally, a statement on the completeness of the interface at the beginning and the end of the data. If incomplete to an extent greater than 5% at either end, a warning is included in the analysis notes to be careful in interpreting the confidence limits for D_0 and Q.

As noted in the discussion of initial estimates and depicted in Figure 2-7, A' and B' were determined to be possibly significant, assigned starting values, and varied in the fit. After completion of the least squares fit, the value of A' and B' were found to be less than their confidence limits, whereupon their values were set equal to 0, and the analysis continued. The analysis notes mention this and give the values that were obtained for A' and/or B'. This is one difference between Fit (Converge) and repeatedly clicking FIT(Step) until convergence is reached. In the latter case, A' and B' would have continued to have been included in the fit even though their values were not statistically significant. Other warning messages can be quite

lengthy, reflecting difficulties encountered in the analysis. If A' and/or B' were varied and found to be significantly different from 0, then the graphs of the asymptotes would have been drawn on the screen.

Along with the values of the interface width and asymmetry, the interface percent values for the values of *X* at the width limits are given as a check on the calculation. The dimensionless asymmetry parameter, η, and its uncertainty are discussed in Section 4.5, Equation (4-43). The dimensionless quantity QD_0 which, if less in magnitude than 1, is comparable in magnitude to η is also given. In general, a value of *QDo* much greater than 1 indicates an unrealistic asymmetry and possibly a runaway value for *Q*. The interface width can be displayed on the graph in the form of a box when the interface item in the View menu is checked.

As mentioned above in the discussion of the use of LFPF in determining lateral resolution from surface line-scan measurements, another measure of the width of the interface is the width at half-height of the derivative *dY/dX*. Following the report of W_f and η_f in the analysis notes, the value of *X* where *dY/dX* is a maximum, $X_{\max}$, is given along with the values of *X* where *dY/dX* is half its maximum value. These two values of *X*, X_- and X_+, can then be used to provide a width W_{hh} and η_{hh} with Eqs. (1-9) and (1-10)

2.4.1 Statistically Significant Interface Region

Because the interface is essentially infinite, owing to the exponential nature of the logistic function, all of the data correspondingly fall within the interface. The statistically significant interface region is that range in X, X_{lower} to X_{upper}, where the **calculated** value of Y lies more than an exaggerated confidence interval (*eci)* away from either asymptote. The *eci* begins with the estimate of the maximum value for the standard deviation based on the residual variance and the upper percentile of the χ^2 distribution for the selected confidence level. We multiply the upper confidence limit for σ by the normal distribution confidence limit to give the *eci*. We note those data whose measured values of *X* lie between X_{lower} and X_{upper} and whose measured values of *Y* lie between the two asymptotes and more than *eci* from each asymptote. **We describe these data as lying in the "statistically significant interface region."** The statistically significant interface region is thus defined by range in *X*, X_{lower} to X_{upper} for which *Y* lies between *f* and (1-*f*) of completion where $f = eci/|B-A|$. The number of such data will inform the program whether the interface parameters X_0, D_0 and Q, are likely to be reliable or even determinable.

2.4.2 Warning Messages in the Analysis Notes

The analysis notes may contain additional warning messages if problems are encountered or if the structure of the data might indicate concern about the interpretability of the parameter confidence limits. These messages are described below.

If the measurement errors follow a normal distribution, the ratio of the variances (squares of the standard deviations) of two independent samples should follow an F distribution. If we assume that the residual standard deviation and the data scatter (see Section 4.2.2 below for a discussion of the data scatter) are two such independent samples (in the sense that one depends on the extended logistic model and the other does not) then the value of F for the ratio of the square of

the standard deviation of the fit over the square of the scatter in the data should be less than the value of F for the number of degrees of freedom for each (see Equation (4-30) and the preceding discussion.) If it is **greater**, a message to this effect appears in the Analysis Notes, such as,

> F(std/scatter) = 2.39 compared to F(0.99, ndf1 = 105, ndf2 = 107) = 1.575

This is an indication that, at the 0.99 confidence level, model errors may dominate random measurement errors limiting the interpretability of the confidence intervals for the parameters.

Consider now the separation of the data into three regions, the statistically significant interface and the pre- and post-interface regions. The region prior to the statistically significant interface is dependent almost solely on the parameters *A*, *A′*, and *A″*. Similarly, the region following the statistically significant interface is dependent almost solely on the parameters *B*, *B′*, and *B″*. While the statistically significant interface depends on all the parameters, it is this region that is most sensitive to X_0, D_0, and *Q*. *Since the asymptotic regions are virtually model-independent*, the variance of those regions will not be sensitive to model errors whereas the statistically significant interface will be. The variances of the three regions are calculated from:

$$s_A^2 = \sum_{i=1}^{n_A} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_A - p_A}, \quad s_I^2 = \sum_{i=1}^{n_I} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_I - p_I}, \quad s_B^2 = \sum_{i=1}^{n_B} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_B - p_B} \qquad (2\text{-}1)$$

n_{A}, n_{I}, and n_{B} are the numbers of data in the pre-interface asymptotic region, the statistically significant interface region, and the post-interface asymptotic region respectively and p_{A}, p_{I}, and p_{B} are the number of varied parameters on which each of the regions is dependent so that the three regions have, respectively, ν_A, ν_I, and ν_B degrees of freedom where $\nu_A = n_A - p_A$, etc. Typically p_A and p_B will each be 1 and p_I will be 2 or 3 depending on whether Q is varied. We can perform the F test on the ratio of the interface variance with each of the asymptotic variances to test for systematic errors. If $s_I^2 / s_A^2 > F(\nu_I, \nu_A, \alpha)$ or $s_I^2 / s_B^2 > F(\nu_I, \nu_B, \alpha)$ where α is the confidence level for the F distribution, we may have reason to suspect model errors. As usual, the more data available for the three regions, the more likely the effect will be noticed.

If both interface/asymptote F tests fail, a warning message similar to:

> The F test for the ratios of the interface/asymptote variances suggests the possibility of systematic error
> F(interface/pre-interface) = 13.598 compared to F(0.95) = 1.732
> F(interface/post-interface) = 2.636 compared to F(0.95) = 2.185

will appear in the Analysis Notes.

The spacing of the values of *X* is noted. If the values of *X* are not uniformly spaced and if the standard deviation of the spacing in *X* is more than 1% of the average spacing, a warning comment will be printed in the Analysis Notes such as:

> NOTE!!! The values of X are not uniformly spaced.
> The average spacing is 3.971 with a standard deviation of 10.93%
> If the values of X are not error free the parameter confidence limits may be underestimated.

The spacing need not be uniform for the statistical interpretation of the confidence limits for the parameter values, but they *must be error free*. If they are not, they will likely fail the

interface/asymptote F test. See the discussion in Section 3.2.5 Errors in the Independent Variable X.

If the data are poorly structured, the least squares analysis can become unstable and diverge. When this occurs, the program attempts to fit the data by holding Q, D_0, and/or X_0 at some predetermined value. The three typical cases are 1) those situations where the interface is very sharp and less than 3 data fall in the statistically significant interface region, 2) those situations where the interface is not complete and one of the two asymptotes is not reached and 3) the noise level is on the order of 10% or more of the separations between asymptotes or greater. The instability of the least squares analysis is also noted when numerical overflows or underflows occur or when the corrections to X_0 or D_0 or the confidence limits for X_0 or D_0 on any iteration indicate that either is poorly determined. See the discussion in Section 3.2 Difficult Data and Analysis Instabilities.

The program notes the number of data falling in the statistically significant interface region and if less than five, a message to this effect appears in the Analysis Notes similar to the following:

> 3 data between 42.0 and 58.0 with | Y-Asymptote | > 1.90 appeared to fall in the statistically significant interface region
> 7 possible interface values from X = 38.0 to X = 62.0, were tested
> Based on the statistics of the fit, the upper limit for Do was 1.88

In this message, | Y-Asymptote | > 1.90 defines the confidence limit for deciding if a datum differs significantly from an asymptote. If two or fewer data fall in the statistically significant interface region and the least squares fit becomes unstable, Q is first set equal to 0. If the least squares fit continues to be unstable, X_0 or D_0 is held fixed at a value determined from the distribution of the data surrounding the interface. The details of how D_0 and X_0 are handled when fewer than three data are found in the statistically significant interface region are further discussed in Section 3.2. The analysis notes report on which of the parameters is held fixed and why.

If either asymptote appears to be incomplete, a corresponding warning message appears in the Analysis Notes:

> Warning!! At the final point, X = 55, the interface is only 89.38% complete. The final asymptote is not reached and the confidence limits for Xo, Do, Q, and B may be underestimated.

If the noise in the data becomes significant compared with the separation between asymptotes, a warning similar to the following appears in the Analysis Notes:

> The ratio of the upper limit of the standard deviation from the chi squared distribution to the value of A-B, 19.2%, may make the determination of Xo, Do, and Q problematic and possibly result in false, local minima..

Both of these situations are discussed further in 3.2.

2.5 Setting the values of parameters

If an entry in the parameter table is double clicked (or the key combination <ctrl>-Enter is

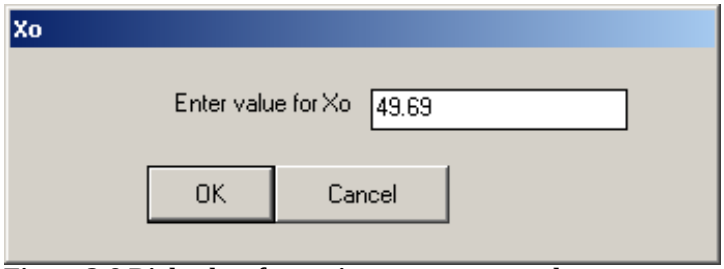


Figure 2-8 Dialog box for setting a parameter value

pressed), as noted in the label above the parameter table, a dialog box appears on the screen as in Figure 2-8 where X_0 has been double-clicked.

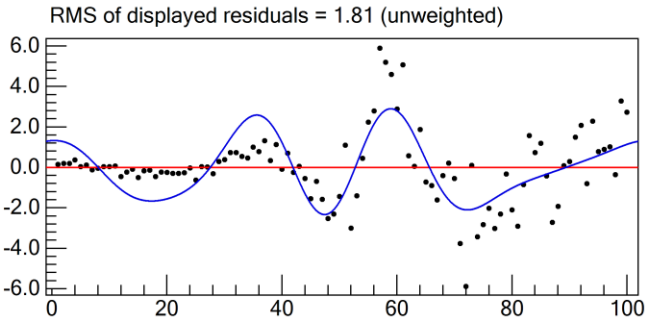


Figure 2-9 Display of residuals with a trend line

Entering a value for X_0 in the text box assigns that value to X_0. When the OK button is clicked (or the enter key pressed), the dialog box disappears and the graph of the calculated value of Y is redrawn using the new value of X_0. Unchecking the check box for, in this case X_0, will cause X_0 to be held fixed at this value in subsequent analyses while the other parameters are varied. To vary X_0 starting with the entered value, just make sure its check box is checked.

2.6 Additional Displays and the View Menu

The View Menu allows display of additional information concerning the analysis.

2.6.1 View > Residuals

The examination of the residuals, that is, the values of the observed data, Y^{obs}, minus the calculated values, Y^{calc}, as in Figure 2-9, is **by far** the best way to detect systematic errors inherent in a semi-empirical model. The eye can quickly detect trends in what should be a random scatter of points. Clicking Residuals on the View Menu marks it with a check mark and replaces the graph of the data with the graph of the residuals. To return to a display of the data, just re-click and uncheck Residuals. When the residuals are displayed, the root mean square, rms, of $Y_{obs} - Y_{calc}$, $rms = \sqrt{\sum_{i=1}^{n} \left(Y_{obs} - Y_{calc}\right)^2 / n}$, is printed above the graph. Note that the rms is *unweighted* while the standard deviation incorporates the weight factors. The *weighted* (standardized) residuals, $W\left(Y_{obs} - Y_{calc}\right)$, are tested in the trend analysis.

Discussion of the residuals is more obvious if we take as our example a data set with the same parameter values as those we have been using but with 100 data instead of 25. Analyzing these data while holding Q fixed at zero results in confidence limits for the standard deviation of $1.702 < \sigma < 2.160$. The lower limit is almost twice the estimate of the standard deviation from the data scatter, 0.9972, suggesting a possible model error or a systematic trend in the residuals. This can be further tested by clicking the Residuals item in the View menu to display the graph of the residuals as in Figure 2-9, where the residuals appear to indicate an oscillating trend suggesting model errors

2.6.2 View > Trends

In a multi parameter fit of say, n parameters, to data with only systematic errors, the residuals will typically cross the Y^{obs} - Y^{calc} axis n or n+1 times resulting in a seemingly oscillatory pattern. This suggests the use of a Fourier analysis of the residuals to test for trends. Clicking the Trends item in the View menu fits the *weighted* residuals to a Fourier series:

$$\left(Y_{obs} - Y_{calc}\right)W = \sum_i U_i \sin(iz) + V_i \cos(iz) \quad \text{where } z = 2\pi \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2-2}$$

Figure 2-9 shows the resulting trend line as well as the *unweighted* residuals themselves. Each of the values of U_i and V_i returned by the least squares fit is compared to its confidence limits at the confidence level selected and if the ratio of the absolute value to its standard deviation is greater than the value of $t_{\nu,\alpha}$ the ratio is printed. The Analysis Notes accompanying the trend line in Figure 2-9 demonstrate this:

> Sequentially fitting the weighted residuals to Ui sin (iz) + Vi cos (iz) for i = 1 to 6, where z = 2π (X - Xmin) / (Xmax – Xmin) . For each value of Ui and Vi calculate T = value / (standard deviation of value). The following values of T were significant beyond the 95% confidence level where T > 1.985. The percentage in parentheses is the confidence level that Ui or Vi is non-zero.
>
> T(cos(2x) term) = 3.141 (99.8315%)
> T(sin(3x) term) = -3.298 (99.9026%) T(cos(3x) term) = 4.857 (99.9999%)
> T(sin(4x) term) = 6.439 (100%) T(cos(4x) term) = -4.012 (99.994%)
> T(sin(5x) term) = -3.024 (99.7504%) T(cos(5x) term) = 3.556 (99.9623%)
>
> This analysis suggests a standard deviation in the range 0.843 < σ < 1.14 is due to random error at the 95% confidence level. The graph represents the sum of all Ui sin (iz) + Vi cos (iz) whether or not individual coefficients

We note that the "improved" residual standard deviation is more in line with the standard deviation estimated by the data scatter. Even if no single term appears to be statistically different from 0 because no value of *t* is greater than $t_{\nu,\alpha}$, some linear combination of the coefficients may be significant and if the upper value of the standard deviation of this fit is less than the standard deviation of the residuals, a warning message along the lines of the following is printed in the Analysis Notes:

> Sequentially fitting the weighted residuals to Ui sin (iz) + Vi cos (iz) for i = 1 to 5, where z = 2π (X - Xmin) / (Xmax – Xmin) , showed no single term to be significantly non-zero at the 99% confidence level. This analysis suggests a standard deviation in the range 0.807 < σ < 1.19 is due to random error at the 99% confidence level. The graph represents the sum of all Ui sin (iz) + Vi cos (iz) whether or not individual coefficients are significantly non-zero.

This analysis reinforces the conclusion reached from visual inspection of the residuals that the analysis suffers from slight systematic errors. The, in this case obvious, source of error is the constraint $Q = 0$. This would have been less obvious if only 25 data had been included in the analysis. Note that the Trends item on the View menu is enabled only when the residuals are displayed. .

It should be noted that the value of $D_0 = 3.872 \pm 0.182$ corresponding to the residuals in Figure 2-9 falls at the low end of the 95% confidence limits given for D_0 obtained for 25 data when *Q* was varied. See Figure 2-7 above, where $D_0 = 4.038 \pm 0.211$. The presence of systematic errors can seriously cloud and may even negate the statistical interpretation of confidence limits, which can, in extreme cases, be underestimated by as much as an order of magnitude.

2.6.3 View > Data Scatter

A model-independent estimate of the standard deviation of the data can be obtained from third differences in the data. Given a set of measurements Y_i, the first differences are defined as

$Y_i^{(1)} = Y_i - Y_{i-1}$, second differences as $Y_i^{(2)} = Y_i^{(1)} - Y_{i-1}^{(1)} = Y_i - 2Y_{i-1} + Y_{i-2}$, and third differences as $Y_i^{(3)} = Y_i^{(2)} - Y_{i-1}^{(2)} = Y_i - 3Y_{i-1} + 3Y_{i-2} - Y_{i-3}$. If Y is a slowly varying function of X so that the change in Y between neighboring data is less than the variability in the point to point scatter of Y, the third differences, which magnify the point to point scatter but minimize the systematic variation in Y, can provide a model-independent estimate of the standard deviation of the measurements. (Indeed, if Y were a linear or quadratic function of X and the values of X were evenly spaced, the contribution from the systematic variation in Y would vanish identically.) In the presence of non-uniform weighting, we use a modified form of the third differences:

$\bar{Y}_i^{(3)} = \frac{1}{\sqrt{20}}\left(Y_{i+3}\sqrt{W_{i+3}} - 3Y_{i+2}\sqrt{W_{i+2}} + 3Y_{i+1}\sqrt{W_{i+1}} - Y_i\sqrt{W_i}\right)$, from which we calculate

$\varepsilon_{3d}^2 = \frac{1}{(n-3)}\sum_{i=1}^{n-3}\bar{Y}_i^{(3)}$ as a model independent estimate of the variance against which to compare the variance from the fit of the data to the logistic function.

In order to avoid confusion between ε_{3d} and *s* we refer to ε_{3d} as the scatter in the data. In the calculation of ε_{3d}^2 those values of $\bar{Y}_i^{(3d)}$ suspected of having a pronounced contribution from the underlying sigmoidal function are excluded. (See Section 4.2.2 for a discussion of third difference estimates of the variance, their calculation and their interpretation.)

Clicking **View > Data Scatter** displays the values of $Y_i^{(3)} / \sqrt{20}$ (not $\bar{Y}_i^{(3)}$ though for unit weighting, the two are the same) as shown in Figure 2-10 Note that the *displayed* values of $Y_i^{(3)}$ are unweighted as are the residuals even though the calculation of ε_{3d}^2 is weighted as is *s* (when weights are used.)

In Figure 2-10 the values strongly affected by the sigmoidal profile interval have nearly all been identified by red x's. Only the two points prior to the identified values have been missed. The header reads "*Third differences with a rms value (excluding 5 values) = 0.253 (0.741).*" The number in parentheses is the (weighted) root mean square of the $\bar{Y}_i^{(3)}$ with all values included. If the data have been fit to a logistic function, the scatter is recalculated from the residuals, not the data, to remove the systematic contribution from the logistic function itself and to minimize the contribution of any remaining systematic errors. See Section 4-6 below.

Figure 2-11 shows a comparison of the residuals of Figure 2-9 with the scatter from the same data. The graph of the data scatter has been displaced downward for ease of comparison. The pattern of the data scatter does not echo the pattern of the residuals because each point in the scatter is the result of a combination of four adjacent data values in addition to the presence of systematic errors in the residuals.

2.6.4 View > Connect

Connects the displayed Data points with a straight line as in Figure 2-11.

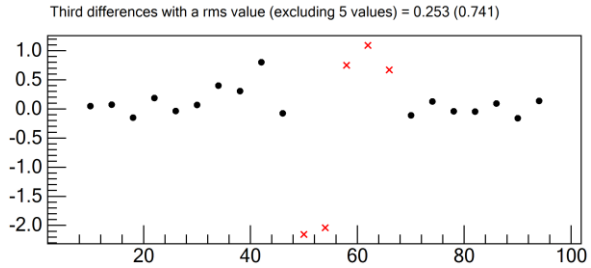


Figure 2-10 Scatter with values most strongly affected by systematic error and excluded from the calculation displayed as red x's

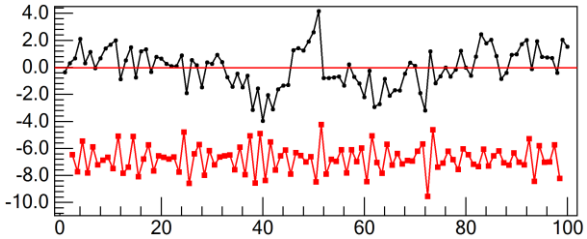


Figure 2-11 Comparison of Data Scatter (lower trace) with the Residuals (upper trace) from the least squares fit

2.6.5 View > View Memory (Data)

If data (or a range of data) have been saved in memory by clicking Remember on the Tools menu, this menu item becomes visible. When it is checked by clicking it, the memorized data are drawn on the display. If the memorized data fall off the scale of the display, they are shifted to be superimposed on the existing display. If a least squares fit was active when the graph was saved to memory, and a least squares fit is currently active, the memorized graph is shifted to match the midpoints of the interfaces. *This item is not visible if no graph has been "remembered."* The remembered graph can be shifted vertically by dragging it up or down while the Ctrl key is depressed. Figure 2-11was created in this manner by saving the Data Scatter graph (See Section 2.6.3 above.)

2.6.6 View > View Memory (Calc)

If a least squares fit is active when the "Remember" item on the Tools menu is clicked, this menu item becomes visible. When it is checked by clicking it, the memorized line calculated from the least squares fit of the corresponding data in memory is drawn on the display. If the memorized line falls off the scale of the display, it is shifted to be superimposed on the existing display. (If a least squares fit is currently active, the shift matches the midpoints of the interfaces.) *This item is not visible if no graph has been "remembered."* The remembered graph can be shifted vertically by dragging it up or down while the Ctrl key is depressed.

2.6.7 View > Identify Outliers

When the Identify Outliers item of the View menu is clicked, the item is checked and those data for which $Y^{obs} - Y^{calc}$ fall outside the confidence limits for a normal distribution (see Equation (4-38)) are identified by circling the points in red. The default confidence limits are based on the two-tailed 95% confidence level for a student's t distribution.

If any outlier is found, the statistics associated with excluding outliers from the calculation are displayed below the Data Scatter. For example, as in Figure 2-7:

```
Excluding 2 possible outliers, s < 0.7826
F > 1.43 compared to F(0.95) = 2.191
```

If the two data identified as outliers were excluded from the fit, the standard deviation would be less than the figure quoted because the exclusion would lead to a slightly lower minimum, hence the < sign in the expression. (When this was done, the resulting residual standard deviation was 0.7120.) The F test result compares the ratios of the standard deviations with and without the outliers (> 1.43) with the value of $F_{\alpha}(n_1, n_2) = 2.191$, where α is the confidence level and n_1 and n_2 are the number of degrees of freedom for the data including and the data excluding the outliers respectively. In this case, performing the least squares fit excluding the two data identified as outliers gave a value of F=1.724 which, because it was less than $F_{0.95} = 2.191$ (which is also greater than 1.43) indicated that the two standard deviations were not statistically distinguishable and the two outliers are simply two data on the wings of a normal distribution.

Upon clicking a data point, *when the calculated function is displayed on the graph*, the values of X, $Y^{obs} - Y^{calc}$, and the confidence limits of $Y^{obs} - Y^{calc}$ are printed above the graph. When a data point identified as an outlier is clicked on the graph, a message similar to:

X = 65.000, Y(obs) − Y(calc) = −2.660 ± 0.994 (0.743%)

is printed on the top of the graph. Here, the difference Y(obs)-Y(calc) = −2.660 is seen to be well beyond its standard deviation, 0.994 and $t = \left(Y^{obs} - Y^{calc}\right) / s_{\left(Y^{obs} - Y^{calc}\right)}$ = −2.69 while the 95% confidence limit for the normal distribution is 1.96. The value in parentheses, 0.743%, is the confidence level of this difference. That is, we would expect only 0.743% of the data whose difference, $Y^{obs} - Y^{calc}$, has an uncertainly of 0.994 to fall 2.660 or more from its expected value of 0. See also Eqs. (4-37) and (4-38) for the difference in the standard deviation of $Y^{obs} - Y^{calc}$ when Y^{obs} is included in the fit or not.

Information on the data identified as outliers is included in the extended analysis notes and can be displayed *and at the same time copied to the windows clipboard for pasting in another application* by clicking **View > Analysis Notes** (see Section 2.6.19 below.)

Data identified as outliers are still included in the least squares fit. **Any point, whether identified as an outlier or not, can be excluded from future fits by double clicking that point**. Double clicking a point will change the display of that point by dimming it, replacing it with a single pixel, or replacing it with an **x**. In subsequent least squares fits, data marked as ignored are not included in the fit. Double clicking an ignored point restores that point to its original status.

It should be stressed that the identification of a particular point as an outlier should not suggest that point be necessarily excluded from the fit. The identification of a point as an outlier only means that its value falls beyond the selected confidence limits. Exclusion of a point as an outlier should be justified by considerations of why the point may be suspect beyond the fact that its value falls in the tail of the distribution. Out of 100 data, we would expect 5 to fall beyond the 95% confidence limits based on the standard deviation of the fit and a normal distribution of errors. As already mentioned, the uncertainty in $Y^{obs} - Y^{calc}$ that is serving as the basis for considering a point as an outlier must take into account that the value of Y^{obs} was used in the calculation of Y^{calc} so that their errors are correlated. This correlation must be included explicitly as described in the discussion accompanying Equation (4-38)

2.6.8 View > Error Bars

When the Error Bars item on the View menu is checked by clicking it, error bars are drawn on each of the displayed data equal to $Ns/\sqrt{W_i}$ where W_i is equal to the weight of the ith datum (usually equal to 1), *s* is the standard deviation of the fit and *N* is the two-tailed normal distribution confidence limit (=1.96 for the default 0.95 confidence level). The error bars are displayed on both the data and the residuals graphs. Note, the error bars are confidence limits, not standard deviations.

2.6.9 View > Confidence Limits

When the Confidence Limits item on the View menu is checked by clicking it, confidence bands are drawn for the *calculated* values of Y equal to $t_{\nu,\alpha}s(Y_i^{calc})$ where $t_{\nu,\alpha}$ is the two-tailed *t* distribution confidence limit for ν degrees of freedom at the α confidence level and $s(Y_i^{calc})$ is the standard deviation of the calculated value of Y from Equation (4-36). The confidence limits are displayed for both the data and for the residuals and can be displayed simultaneously with the error bars. The confidence limits take into account the correlation of errors among the parameters from the least squares fit.

2.6.10 View > Data Selection Box

Data can be "selected" for special treatment by drawing a box around the selected data.

Clicking the Select Data Box item on the View menu draws a box on the screen as the starting data selection box such as that seen in Figure 2-12. The Analysis Notes give guidance on how to size and position the selection box.

The Data Selection Box can also be invoked with a combination of left and right mouse clicks. Clicking a point on the graph with the RIGHT mouse button when a crosshair is already displayed displays the data selection box as defined by the crosshair and the right mouse click.

To resize and reposition the box, drag either side of the box to its new position, or click anywhere on the area of the graph and the side nearest to the spot clicked will be moved to that spot. Only the *X* values of the selection box are significant. The top and bottom of the box are set to include the minimum and maximum values of *Y* in the box.

Clicking anywhere in the LFPF window outside the graph erases the box. Unchecking the Select Data Box item on the View menu erases the selection box. If the selection box is erased, clicking the Select Data Box item on the View menu again, or clicking the right mouse button when the cursor is on the graph *and* a crosshair is *not* displayed, redisplays the selection box as it was when it was erased.

When a selection box is displayed, the Zoom in item on the View menu is enabled and clicking "Zoom in" redraws the graph of only those data inside the selection box. *Alternatively, clicking the right mouse button alone at any point inside the selection box performs the same function as checking the "Zoom in" item in the View menu.* When zoomed, the axes are correspondingly rescaled. If a zoomed graph is displayed, the "Restore" item on the View menu is enabled and clicking "Restore" returns the display to the full range of data. Alternatively, clicking the right mouse button anywhere on the zoomed graph returns the display to the full range of data.

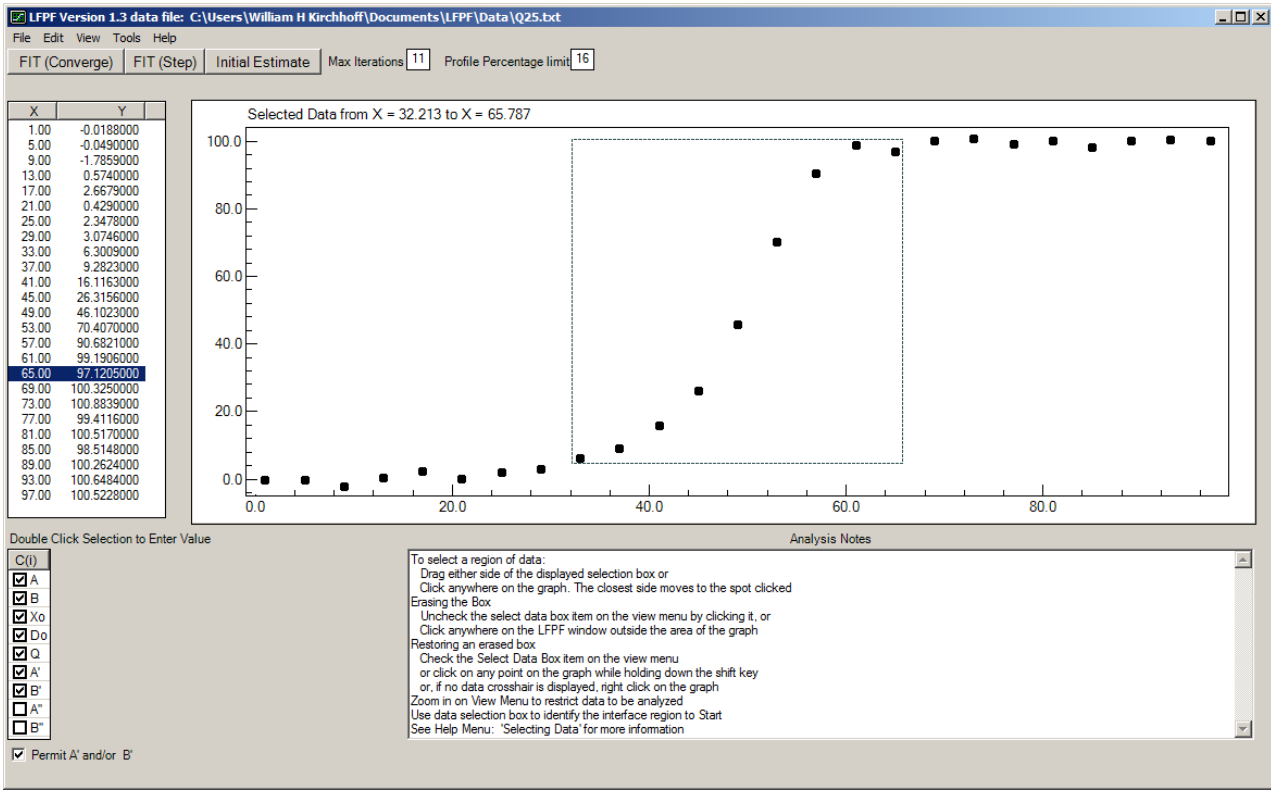


Figure 2-12 Data Selection Box

Note: The range in *X* for the selected data printed on the top of the graph is calculated from the pixel values of *X* for the graph. Typically, there are less than 1000 pixels in the width of a displayed graph and this limits the precision with which the values of *X* can be calculated. Slight differences may be noted when redisplaying a previously displayed selection box.

When doing a least squares analysis, only the displayed data are used for the analysis. This is one of the major uses for the data selection procedure. The data on a zoomed graph can be further zoomed by displaying a data selection box on the zoomed graph and proceeding as above. If the "Zoom out" item on the View menu is clicked, the range of data for the display is increased by 20%, 10% in each direction. This can be used for fine tuning of the data selection or for extrapolating the calculated graph of the interface profile. When the "Restore" item on the View menu is clicked, the graph returns to the full range of data no matter how many times the zoomed graphs were nested.

2.6.10.1 Using the Data Selection Box for calculating initial estimates

If a selection box is displayed when Initial Estimate (or Fit (Converge) or FIT (Step) for a new data set) **is clicked, the initial estimates of the parameters are based on straight lines through the three regions identified by the selection box.** The pre-interface baseline is calculated from the data to the left of the selection box and the post-interface baseline is calculated from the data to the right of the selection box. The values of X_0 and D_0 are calculated from a straight line passing through the five points (if there are more than five) nearest the center

(in the *Y* direction) of the interface region inside the box. This line is interpreted as a tangent line to the logistic function. X_0 corresponds to the midpoint of the "tangent" line where $Y = (A+B)/2$. D_0 is determined from the slope, $dY/dX = (B-A)/4D_0$. The value of *Q* is initially set equal to 0 and the slopes of the asymptotes are allowed to vary. This alternative method for making initial estimates is provided primarily for the analysis of poorly structured data where the algorithms for making the initial estimates from the structure of the data fail.

2.6.11 View > Zoom in

If a selection box is displayed, the "Zoom in" item on the View menu is enabled and clicking it redraws the graph of only those data that were inside the selection box. The axes are correspondingly rescaled. This can also be accomplished by clicking the right mouse button with the cursor positioned in the selection box. *A least squares fit, if performed, is based only on the data displayed.*

2.6.11.1 View > Zoom out (Ctrl-Z)

Clicking the "Zoom out" item on the View menu expands the scale of the *displayed* X axis symmetrically by 20%, 10% in each direction. The purpose of this is primarily to look at extrapolation of the calculated logistic function. This can be repeated as often as one wants. Ctrl-Z performs the same function (overriding the usual undo or delete windows function for Ctrl-Z)

2.6.11.2 View > Restore

If a graph is a zoomed graph, whether zoomed in or zoomed out, the "Restore" item on the View menu is enabled and clicking it restores the graph of the data to its original scale. This can also be accomplished by clicking the right mouse button while no crosshair is displayed *and* the cursor is positioned anywhere on the graph.

2.6.12 View > Interface

Clicking the "Interface" item on the View menu draws a box around the width of the interface as defined by the profile width fraction f (default values 14.64% and 85.36%) with opposing corners at the exact points, *X* and *Y*, given by the calculation. If the interface box is displayed, checking "Select Data Box" on the view menu will erase the interface box and replace it with a data selection box of the same size and the graph can be zoomed to display only those data within the defined width of the interface. (Clicking the right mouse button does the same thing.)

2.6.13 View > Statistical Interface

When this item is clicked, a box is drawn representing the range of the *calculated* interface for which the *calculated* values of Y differ from the two limiting asymptotes by more than the confidence limits of the residual standard deviation (See 2.4.1.) Data falling in this box are those that contribute most significantly to the determination of X_0, D_0, and Q. The **statistically significant interface region** is invoked to estimate the maximum value for D_0 when less than three data fall in this region. In *determining* the number of data falling in the statistically significant interface, only the deviations of the observed data from the asymptotes are used. In *testing* for data falling in the statistically significant interface, a range approximately equal to the statistically significant interface above and below X_0 is used. The test region is therefore

approximately twice the width of the statistically significant interface. The statistically significant interface drawn on the graph is calculated from the values of X_0, D_0, Q, and the confidence limits of the residual standard deviations. The statistically significant interface region is also used to test for the possible influence of errors in the independent variable X (see Section 4.2.3 below)

2.6.14 View > Asymptotes

Clicking the "Asymptotes" item on the View menu draws the two asymptotes, $A+A'(X-X_0)+A''(X-X_0)^2$ and $B+B'(X-X_0)+B''(X-X_0)^2$ connected by a vertical line at X_0. Inclusion of any of the parameters A', B', A'', or B'' in the least squares fit will always improve the residual standard deviation but the asymptotes themselves may not be physically reasonable. For this reason, whenever they are varied in the fit the "Asymptotes" item in the View menu is checked by the program and the asymptotes are displayed, prompting the analyst to consider whether these asymptotes are physically reasonable or not. If any of A', B', A'', or B'' is required *only* for an improved fit to the data, very likely these terms indicate that the extended logistic function is not an accurate representation of the data. Moreover, their inclusion in the fit may shift the width and asymmetry well beyond their confidence limits.

2.6.15 View > Draw dY/dX

Clicking the "Draw dY/dX" item on the View menu draws the derivative *dY/dX* on the graph, scaled to fit the graph with the half height marked by a horizontal line and with vertical lines marking the maximum value of *dY/dX* and connecting the half-height points to the logistic function line. On the top of the graph, the values of *X* where *dY/dX* is a maximum as well as the two values of *X* at half-height, X_- and X_+ are printed (see Eqs. (1-9) and (1-10))

2.6.16 View > ΔX/ΔY from Data

Replaces the graph of the data with the derivative of the data approximated by the ratio of differences from neighboring points: $dY/dX \approx \Delta Y/\Delta X = (Y_{i+1}-Y_i)/(X_{i+1}-X_i)$ for $i=1 \rightarrow n-1$ with corresponding values of $X=(X_i+X_{i+1})/2$. Any random noise in the data will be magnified but the resulting display can be compared with dY/dX from the logistic function by also clicking **View > Draw dY/dX**. The comparison is strictly visual but may help in the interpretation of dY/dX, When *ΔY/ΔX* and *dY/dX* are simultaneously displayed they are displayed with the same, correct scale. That is, *dY/dX* is not scaled as it is when displayed with the original data.

2.6.17 View > Parameter Derivatives

If a parameter is selected by clicking its entry in the parameter list, for example *Q*, that entry is highlighted in the list. The View menu then displays the additional item, in the present example, "Draw dY/dQ." Clicking "Draw dYdQ" draws *dY/dQ* on the current graphical display, scaled to fit the display as in Figure 2-13 below. This provides some insight into which parameters are

sensitive to which regions of the data. Note that it is not necessary for any parameter to be varied (have its entry checked) in order to draw its derivatives.

If Q is non-vanishing and the parameter D_0 is selected then an additional item, "Draw D," is offered on the View menu to display the variation in D, scaled from 0 to $2D_0$, over the range of data as shown in Figure 2-14

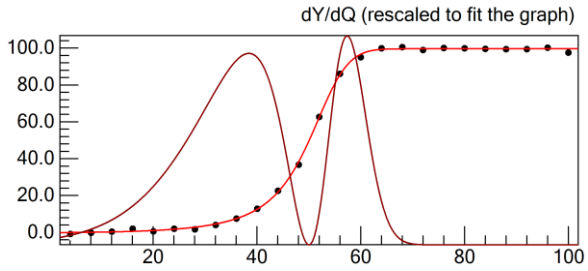


Figure 2-13 Displayin dY/dQ

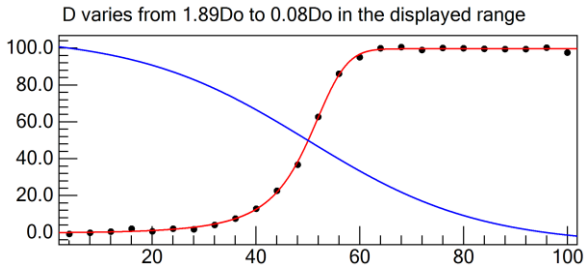


Figure 2-14 Displaying D for non-zero Q

2.6.18 View > Ignored Data

This item allows the user to choose how data which have been designated to be ignored in the least squares fit are displayed. Data can be designated as ignored by double clicking a data point, hitting the delete key when a data point is highlighted or double clicking the corresponding entry on the data list. The display options include a dimmed point, a single screen pixel, or an x.

2.6.19 View > Analysis Notes

Displays extended analysis notes as a message box on the screen and copies the same to the windows clipboard. The extended analysis notes include the values of the parameters the statistics of the least squares fit and scatter, and outliers (if **Menu > View > Outliers** is checked.) Along with a copy of the graph (**Menu > Edit > Copy Graph**) it provides a complete record of the analysis which can then be pasted into another application and archived.

2.7 Edit Menu: Editing and Copying Data, Results and Graphs

The "Edit" item on the Menu Bar contains several items to allow the user to paste data into the starting window for subsequent analysis, edit or reassign X and Y, or copy the results of the analysis, the currently active data, and the graphical display.

2.7.1 Edit > Paste

Pastes the contents of the Windows clipbord into the data entry text box on the starting screen. Ctrl-V works as well. This is enabled only when the data entry text box is displayed on the screen as in Figure 2-2

2.7.2 Edit > Edit Data

Returns to the initial display with the currently active data in the text box (see Figure 2-2) where it can be subsequently edited as with any text editor.

2.7.3 Edit > Interchange X, Y

Exchanges X for Y and Y for X and reorders the data in order of increasing X for the new values of X.

2.7.4 Edit > Reassign X, Y

If the current data file being analyzed contains more than two "columns" of data, this menu item becomes visible and when clicked, the screen displays the original columns of data as in Figure 2-4 above where the columns corresponding to X and Y (and/or W) can be reassigned.

2.7.5 Edit > Normalize Y

Shift and rescale the Y axis so that the maximum value of Y is 1.0 and the minimum is 0.0.

2.7.6 Edit > Copy Data

Copy the table of displayed data onto the clipboard for subsequent pasting into a word processor or spread sheet program or the input text box of this program. If a zoomed graph is displayed, *only the data in the zoomed graph* is copied. This is useful for generating various test data sets for further testing or intercomparisons of computational approaches. Note also that the displayed data can be saved to a file by clicking the save item on the File menu.

2.7.7 Edit > Copy Results

Copies a summary of the results of the least squares analysis to the Windows Clipboard. The summary includes the values of the parameters and their confidence limits, the values of the standard deviation and data scatter, all of the information in the Analysis Notes as they appear on the screen, and the data included in the analysis. (Note that following an analysis, the contents of the Analysis Notes are automatically copied to the clipboard.)

2.7.8 Edit > Copy Graph

Clicking Edit > Copy Graph brings up a dialog box, Figure 2-15, providing a number of options for copying a displayed graph including such items as the interface box, outliers, various derivatives, etc.

The first option copies the displayed graph to the Windows clipboard for subsequent pasting into a word processing document. Figure 2 13 and Figure 2 14 above were generated in this fashion. Alternatively, the graphical image can be saved in a variety of graphical file formats including bit mapped (BMP), Graphics Interchange Format (GIF), Joint Photographic Expert Group format (JPEG), Portable Network Graphics (PNG), or Tagged Image File Format (TIFF) with Lempel-Ziv-Welch (LZW) compression. The graphics file is stored, if possible and as a default, in the same folder as the data file with the same file name but with the file extension replaced by bmp, gif, jpg, png, or tif. The usual windows save file dialog box provides the user with the ability to

store the file anywhere with any name. The folder (directory) \NIST\LFPF created by the program in the users' \Documents folder when it is first run is the default, fall back folder for saving graphical images. The graph can be saved with or without its original colors. X and Y axis labels can be added as well as the file name and date as a caption. If the graph contains header information it can be included or ignored or replaced with a caption entered by the user. If a header caption is added by the user, it is automatically copied unless the "Retain Header" box is checked, whereupon the header displayed on the graph will be copied. If no header is displayed on the graph and the Retain Header is checked, no header will be copied, even if one is entered in the Copy Options form. The font size for the axis labels can also be set with the default value being based on the number of pixels displayed on the graph. Font size is specified in points, one point being 1/72 of an inch. The resolution (dots per inch, dpi) can be specified. Screen resolution is typically 96 dpi and printed graphics typically have to be at resolutions of 300 or 600 dpi. The width of the image must also be specified. All of the graphics routines are pixel or dot centered so the working scale for the program is determined from the specified size and resolution. The screen size, on the other hand, is specified in terms of pixels. As the window is resized, the pixel values for the height and width of the graph are printed on the top line of the graph. When saving a graphics file, only the width of the saved graph is specified. LFPF maintains the aspect ratio of the graph when saving a graphics scale regardless of the resolution and width.

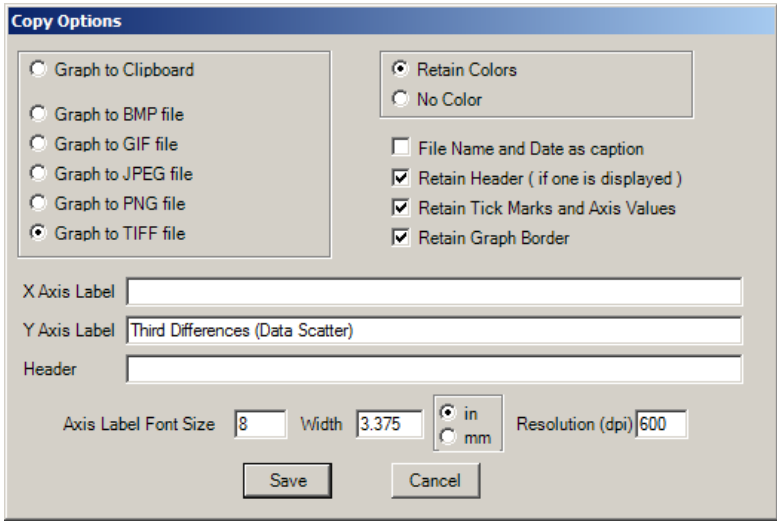


Figure 2-15 Dialog Box for copying or saving a graph

2.8 Tools

The Tools menu includes routines that were used during the development of LFPF and have been retained because of possible general interest.

2.8.1 Tools > Reset All (Ctrl-R)

Clicking Reset simply redraws the graph of the data, omitting any other lines or diagrams that may have been displayed. The parameters are reset to 0 and the default parameters are checked in the parameter table.

2.8.2 Tools > Remember

Stores the currently displayed graph and, if a least squares fit has been performed, the current parameter values in memory to allow redrawing for comparison. This would allow, for example, display of the residuals on the same graph with the original data, or comparison of the data

scatter with the residuals (as in Figure 2-11), or two sets of data. When graphs of data are stored in memory, the item View Memory (Data) on the View Menu is visible and can be checked to draw the memorized data on the display. If a least squares fit is active when Remember is clicked, the item View Memory (Calc) on the View Menu is visible as well and can be checked to draw the memorized calculated graph on the display. The data in memory remains until the program is terminated or a new graph is saved to memory.

2.8.3 Tools > Log Results

When this item is checked the results of intermediate calculations as well as the analysis notes on the final calculation are printed in a file with the same name as the data file being analyzed, but with a **.**log extension, and in the same folder as the data currently being analyzed. If that is not allowed by the operating system, the log file is saved in the default fall back subfolder \NIST\LFPF in the user's \Documents folder created by the program when it is first run. The log file may be of some use in interpreting situations that lead to unexpected or bizarre results or program crashes.

2.8.4 Tools > Statistics

The program provides a means for performing multiple calculations on synthetic sets of data that differ only in the distribution of errors. Initially these calculations were incorporated into the program for the purpose of program development and testing. However, the results may be of more general interest and therefore the capability has been retained and can be accessed through clicking the "Statistics" item on the "Tools" menu. Using this feature, the reliability of the reported confidence intervals can be tested. For example, slight model errors might be introduced by the linearization process of a truncated Taylor's series (no such errors have yet been detected). Were this so, the validity of the confidence limits calculated from the linearized function would be called into question.

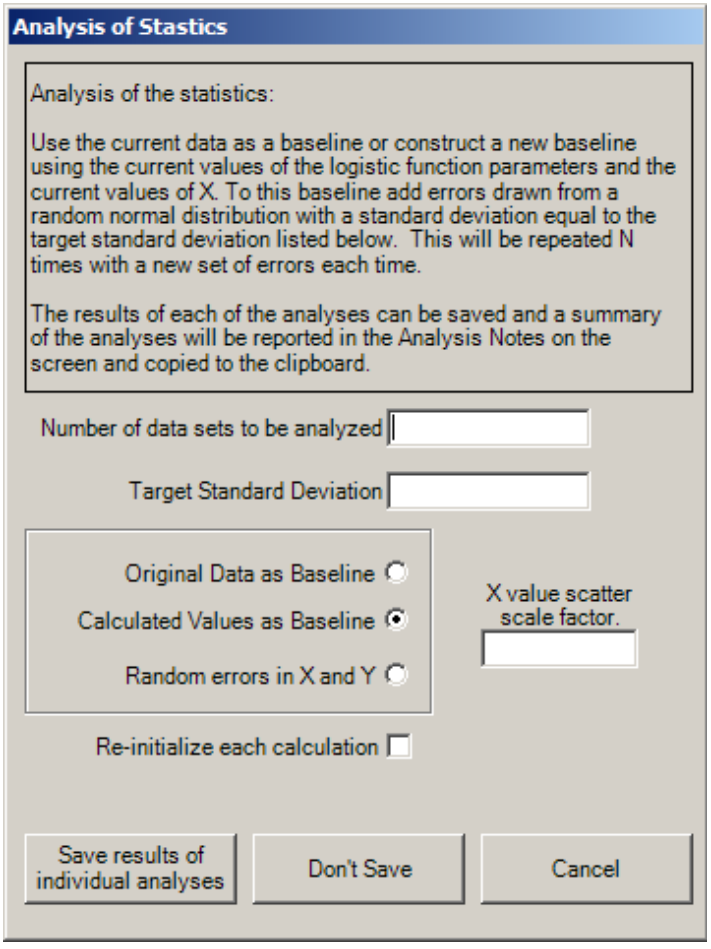


Figure 2-16 Analysis of Statistics Dialog Box

The calculation proceeds in the following way. Baseline data are defined. A least squares fit is performed on the baseline data providing the true values of the logistic function parameters. Following this, random normal deviates with a standard deviation of unity are generated,

multiplied by a target standard deviation entered by the user, and added to the Y values of the baseline data. This new set of data is then fit by least squares to the logistic function. This can be repeated for any number of times under the control of the user.

When the "Statistics" item on the Tools menu is clicked, a new dialog box appears as in Figure 2-16**Error! Reference source not found.**. Several options are provided. Enter the number of data sets to be analyzed and the target standard deviation. The baseline for the data sets can consist of the original *X* and *Y* values themselves, i.e., the original data, or the original *X* values with *Y* values calculated from the current values of the parameters. If the calculated values of *Y* are to be used as the baseline to which errors are added, the user has the additional option of non-uniform spacing in *X*. If $\bar{\Delta}_X$ is the average separation between adjacent values of *X*, r_i is a random, normal deviate with unit standard deviation, and s_f is the "X value scatter scale factor" entered by the user, each value of *X* is shifted by $s_f \bar{\Delta}_X r_i$ giving $X_i' = X_i + s_f \bar{\Delta}_X r_i$ as *X* values for the baseline data. The corresponding baseline *Y* values, Y_i', are calculated from X_i'. If the "X value scatter scale factor" is left blank or set to zero, the baseline *X* values are the original values of *X*. If, on the other hand, "Random errors in X and Y" is checked, then the values of *Y* are the values of *Y* calculated from the original values of *X*. The baseline values of *Y*, Y_i', no longer correspond to the values of X_i'. The *X* values are therefore no longer error free. In this case, the effect of errors in the X values on the confidence limits of the parameter values can be evaluated.

If "Re-initialize each calculation" is checked, then each calculation begins with the calculation of initial estimates. If not, each calculation begins with the original values of the parameters. The latter results in a somewhat more rapid calculation because convergence is reached more rapidly. If a parameter box is unchecked, its value will be held fixed at its current value in the calculations of all the data sets. If this is not the first time the statistics have been tested, an additional option to re-initiate the random normal deviates is offered. If unchecked, the generation of random errors continues from where the previous calculation left off.

If desired, the resulting parameter values from each of the data sets analyzed can be saved by clicking the Save results of individual analyses button. The results are saved to a text file named "Results.txt" residing in the same folder (directory) as the original data (or if not allowed by the operating system in the fall back directory \Documents\NIST\LFPF created by the LFPF program). If the button Don't Save is clicked, only a summary of all the data sets will be printed in the Analysis Notes.

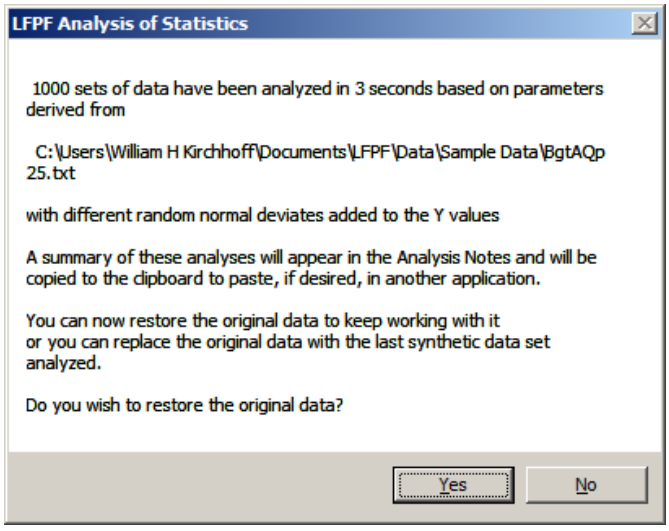


Figure 2-17 Dialog Box following statistical analysis

Depending on the number of data sets being analyzed, the calculation can last for a few seconds to many minutes. An estimate of the time required is printed in the Analysis Notes based on the time

taken to do the first 50 datasets.

When the analysis is complete, a message is printed on the screen similar to that in Figure 2-17**.** The option is given to continue working with the original data or to replace the original data with the data from the last data set analyzed.

(*Note: If the statistics test is performed with one data set, a target standard deviation of 0, and the calculated values as a baseline, the last (and only) data set analyzed will consist of values calculated from the extended logistic function with no errors. Opting NOT to restore original data will replace the data with data calculated from the displayed parameters and will be exact. Furthermore, the calculated profile can then be copied to the clipboard by clicking the "Copy Data" item of the Edit menu or saved as a new data file from the File menu.)*

After selecting yes or no, the screen will appear something like Figure 2-18 below.

The summary of the statistics appearing in the Analysis Notes will be difficult to read but these results are also copied to the Windows clipboard and can be pasted into another application. For the analysis appearing in Figure 2-18, the contents of the Analysis Notes were pasted into this document, formatted, and appear following Figure 2-18.

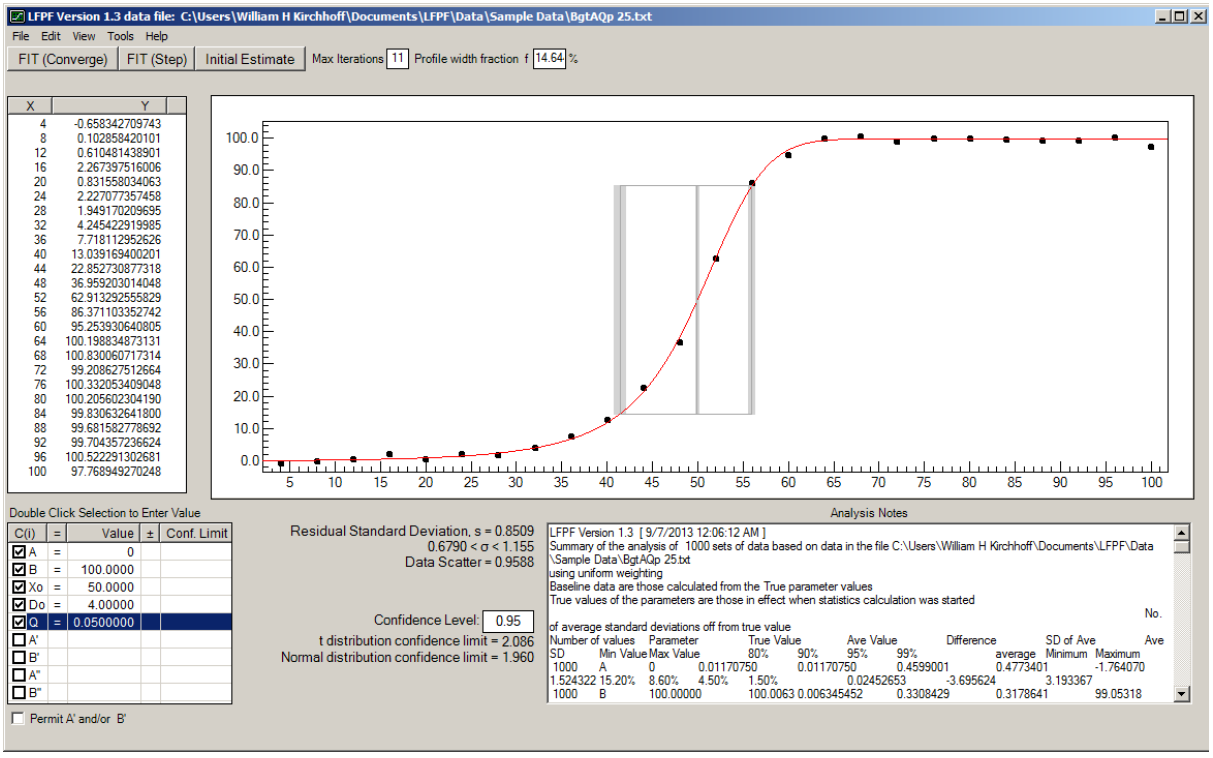


Figure 2-18 Screen shot following completion of a statistics analysis

LFPF Version 1.3 2/27/2013 2:07:09 PM
Summary of the analysis of 1000 sets of data based on data in the file C:\Users\William H Kirchhoff\Documents\NIST\LFPF\Sample Data\BgtAQp 25.txt
using uniform weighting

Number of values	Parameter	True Value	Ave Value	Difference	SD of Ave	Ave SD	Min Value	Max Value	80%	90%	95%	99%
1000	A	0	0.01	0.01	0.46	0.48	-1.76	1.52	15.2%	8.6%	4.5%	1.5%
1000	B	100	100.01	0.01	0.33	0.32	99.05	100.91	21.8%	11.7%	5.7%	1.1%

1000	Xo	50	50.004	0.004	0.125	0.120	49.642	50.345	20.8%	10.8%	5.1%	1.0%
1000	Do	4	4.002	0.002	0.105	0.104	3.596	4.351	20.3%	10.8%	5.1%	0.7%
1000	Width (14.6%)	14.53	14.55	0.02	0.41	0.41	13.07	15.79	18.4%	9.3%	4.8%	0.9%
1000	dY/dX W	13.18	13.17	-0.01	0.43	0.43	11.68	14.85	19.8%	9.7%	5.0%	1.0%
1000	Q	0.05	0.0502	0.0002	0.0080	0.0081	0.0279	0.0779	18.8%	8.9%	4.5%	0.5%
1000	ETA	0.1742	0.1745	0.0003	0.0266	0.0273	0.0938	0.2670	17.9%	8.6%	4.6%	0.8%
1000	QD0	0.2	0.2006	0.0006	0.0315	0.0323	0.1067	0.3135	18.1%	8.3%	4.6%	0.7%
1000	dY/dX ETA	0.1175	0.1175	0.0000	0.0172	0.0176	0.0639	0.1741	18.2%	8.9%	4.3%	1.1%
1000	X at max dY/dX	51.497	51.499	0.002	0.264	0.270	50.712	52.466	17.9%	8.1%	3.6%	0.9%
	Std. Dev.	1.004	0.992	-0.012	0.158		0.520	1.505	20.5%	9.6%	5.1%	0.8%
	True SDev.	1.004	0.994	-0.010	0.140		0.628	1.474	19.0%	10.2%	4.9%	0.8%
	Scatter	1	0.978	-0.022	0.226		0.297	2.053	40.3%	28.2%	19.7%	8.2%
	True Sctr	1	0.982	-0.018	0.225		0.296	2.054	38.2%	26.8%	18.5%	6.2%
	Ftest								16.8%	8.0%	3.2%	0.8%
	Outliers								20.8%	10.0%	4.9%	0.6%
	True Outliers								20.2%	10.3%	5.2%	1.0%

As a check on the calculation of standard deviation of the asymmetry calculated from the half heights of dY/dX using numerical derivatives d(dY/dX)/dCi, the ratio of the average standard deviation of η to the standard deviation of the average values of η = 1.027

Corr Coef

	A	B	Xo	Do	Q
A	1	0.0617	0.1651	-0.4659	-0.4975
B	0.0617	1	0.1787	0.2004	-0.2518
Xo	0.1651	0.1787	1	-0.2288	0.3377
Do	-0.4659	0.2004	-0.2288	1	-0.1123
Q	-0.4975	-0.2518	0.3377	-0.1123	1

Each data set contained 25 data

For the distribution of the scatter and the standard deviation in the 1000 data sets:

Moment	Scatter	True Scatter	Std. Dev.	True Std. Dev.
Average	0.977687112	0.982028895	0.992319106	0.994393382
Std. Dev	0.225670465	0.224759501	0.158337107	0.139654575
Skewness	0.493256933	0.493528129	0.227271229	0.239418033
Kurtosis	0.729293315	0.768131034	-0.193102754	-0.07486338

For the individual data sets:

The average skewness for the scatter was -0.001874 and for the residuals was 0.009358

The average kurtosis for the scatter was -0.4818 and for the residuals was -0.2052

The average scatter from the logistic function was 0.1663

For these 1000 data sets, the scatter = -0.0231 (± 0.0629 at the 95% confidence level) + 1.0086 X the standard deviation

The ratio of (Scatter RMS)/(Standard Deviation) ranged from 0.4419 to 1.4366

The ratio was, on average, 0.9847 (± 0.203 0.260 0.311 0.408 at the 80%, 90%, 95%, 99% confidence level)

The analyses required, on average, 3.5 iterations to reach convergence

Tests for divergence were based on the 95% confidence level

For each data set, the initial values of the parameters were estimated from scratch

1000 data sets used the Trial & error method for initial estimates.

In the parameter value statistics above:

Only those parameters whose values could be evaluated in the least squares fit were included

The confidence limits are based on the "True" values of the parameters

Outliers were identified only after convergence was reached which can affect the count of data in the interface

The minimum number of data in the statistically significant interface region was 6

The average number of data in the statistically significant interface region was 8.39

The statistically significant interface region, based on the target standard deviation, was between X = 24 and X = 64

The maximum interface width is 34.60 and the average data spacing is 4.000

The ratio is 8.649858

The maximum limit for Do averaged 4.940 ± 0.3277

At the minimum value of X, 4.00, the fractional completeness ranged from 0.0% to 0.3% complete (True = 0.2%)

At the maximum value of X, 100.0, the fractional completeness ranged from 100.0% to 100.0% complete (True = 100.0%)

Initial error seed = -1 and last error value added = -0.213317558169 and the calculation took 3 seconds.

33 sets failed the pre-interface/interface F test

39 sets failed the post-interface/interface F test

4 sets failed both.

The following are the root mean square values of Y(obs)-Y(calc) for each datum averaged over all data sets. Note that they are all less than the root mean square of the errors added in the column labeled 'True' because the values of Y(calc) also contain errors that are correlated with the errors in Y(obs). The 'adjusted' values are scaled by the square root of $s^2/(s^2-s^2(Ycalc))$ and these should agree with the 'True' values. If not, the data should be considered ill structured and the confidence limits of the affected parameters cannot

be trusted. The sums of the squares of the rms values of Y(obs)-Y(calc) divided by the square of the rms value of the standard deviation is given for the pre-interface region, the interface region, and the post interface region. These should reflect the number of degrees of freedom for each of these regions.

X	rms(Yo-Yc)*W		adjusted	True*W	
4	0.887349	0.993011	0.985487		
8	0.932877	1.03455	1.00878		
12	0.923358	1.01273	0.992343		
16	0.951506	1.03013	1.04012		
20	0.95019	1.01541	1.02482		
24	0.941412	0.997099	0.975993		$\Sigma W(Yo-Yc)^2/s^2 = 5.15$, n = 6
28	0.966969	1.02643	1.01935		
32	0.867965	0.942428	0.95065		
36	0.910529	1.03606	1.00863		
40	0.881504	1.05318	1.0681		
44	0.833066	1.01687	0.984773		
48	0.714806	0.987283	1.02864		
52	0.6561	0.963585	1.0067		
56	0.629513	0.972434	0.978541		
60	0.791358	0.968196	0.969662		$\Sigma W(Yo-Yc)^2/s^2 = 5.89$, n = 9
64	0.986982	1.04176	1.05249		
68	0.968811	1.01919	1.0212		
72	0.926975	0.978093	0.981731		
76	0.941887	0.994225	0.992789		
80	0.948788	1.00161	1.00478		
84	0.950368	1.00317	0.994723		
88	0.94707	0.999745	1.00371		
92	0.95627	1.00953	1.0237		
96	0.961496	1.01495	0.996269		
100	0.916382	0.967367	0.981157		$\Sigma W(Yo-Yc)^2/s^2 = 8.95$, n = 10

Although in this example every parameter reported was determined for every data set analyzed, this is not always so and for this reason, the values of the number of analyses that led to the values reported are given under the heading "Number of Values". Regardless of whether the original data or calculated values were selected as the baseline, the "True Values" reported are the values of the parameters when the "Statistics" item on the Tools Menu was clicked. Note, these values can be entered into the parameter list by the user just before clicking the statistic menu item and need not have any relationship to the data being analyzed before clicking the statistics menu item.

The "Average Value" reported is the average over the N values that were determined. The "Average std dev" is the average over the N values of the standard deviation for the parameter returned by the least squares fits while the standard deviation of the average is just that. The confidence limits for the parameter values, based as they are on the standard deviation of the fit which is an estimate of the standard deviation of the population of all errors, should follow a student's distribution . The percentages given in the table above are the percentage of those values that fall beyond the 80, 90, 95, and 99 two-tailed percentile values for a student's t distribution. Because the true standard deviation is known, the standard deviations of the parameters can by scaled by the ration of the standard deviation of the population divided by the standard deviation of the fit and those values should follow a normal distribution. This has been tested and found to be so.

In the above example, good agreement with a student's t distribution resulted, as well they should have, including the statistics associated with the normal distribution of those data whose value of $Y_i(obs) - Y_i(calc)$ lie outside of the 80%, 90%, 95% and 99% confidence limits and are identified as "outliers" in the analysis. These should follow the statistical distribution dictated by:

$$s^2(Y_i^{obs} - Y_i^{calc}) = s^2 - s^2(Y_i^{calc}) \quad (2\text{-}3)$$

where s^2 is the standard deviation of the data (which should follow a χ^2 distribution) and

$$s^2(Y_i^{calc}) = \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{\partial Y_i^{calc}}{\partial C_j}\frac{\partial Y_i^{calc}}{\partial C_k}\mathbf{V}_{jk}^{cv} \quad (2\text{-}4)$$

(See Equation (4-16) and accompanying discussion.) The – sign in equation (2-3) arises from the fact that the error in Y_i^{calc} is correlated with the overall standard deviation through the variance-covariance matrix, $\mathbf{V}_{jk}^{cv}$, which is a multiple of s^2. Whereas the uncertainties in Y_i should follow a normal distribution, the uncertainties in $s^2(Y_i^{calc})$ should follow a student's t distribution. In counting outliers, we use the confidence limits of the normal distribution as if our sample size were infinite. For higher values of the confidence limit and lower values for the number of degrees of freedom, we will underestimate the number of data falling outside the confidence limits. This, fortunately, corresponds to the guidance that outliers should not be identified as outliers based solely on their falling in the tails of the normal distribution but on factors related to the manner in which their values were measured.

The F test percentages are those for which the square of the standard deviation divided by the square of the data scatter from third differences is greater than the value of $F_{\nu_1,\nu_2,1-\alpha}$ (see Equation (4-30) and its accompanying discussion.) While fewer than expected in the current example fail the F test, at all confidence levels, 16.8%, 8.0%, 3.2%, 0.8% compared to 20%, 10%, 5%, 1%, they are close enough to be useful in signaling the possible presence of model errors. Of interest, this underestimation remains more or less constant as the number of data in each set increases from 25 to 1000 (see Table 3-4 in the following chapter.)

Much of the remaining report from the statistics analysis deals with tests of the consequences of the possible influence of systematic error. While most of the information is straightforward, the items "33 sets failed the pre-interface/interface F test, 39 sets failed the post-interface/interface F test, 4 sets failed both." requires additional explanation.

When the calculation of the 1000 data sets was repeated with errors in the X values as well as the Y values, the resulting statistics did not follow the normal and student's t distributions because the effect of introducing errors into the X values causes those errors to be interpreted as errors in Y. Those errors will be greater in the transition region where small changes in X result in large changes in Y whereas in the asymptotic regions, errors in X will have no effect on the perceived errors in Y. Hence the errors are not randomly distributed and the statistics resulting from the calculation are wrong. The effect of errors in the X values can only be seen in the display of residuals in Figure 2-19 below which also includes the statistically significant interface box.

If we were to perform the F test on those data in the pre-interface region and those data in the interface region, that is comparing the ratio of the variance of the data in the interface region over the variance in the pre-interface region with the value of F from the F distribution with the appropriate number of degrees of freedom of the two populations and the stated confidence level, we would see that the two sets of residual standard deviations represented different populations. A similar statement can be made for the ratio of the variance from the post-interface region to that of the interface region. In the example for which only the Y values contained errors, the ratios of the variances were less than the value of F for the confidence level and the number of

degrees of freedom. For the extreme case data, represented by Figure 2-19, the F test values were F(interface/pre-interface) = 68.876 compared to F(0.95) = 6.094 and F(interface/post-interface) = 125.138 compared to F(0.95) = 3.293. This will be discussed further in Section 4.2.3 below.

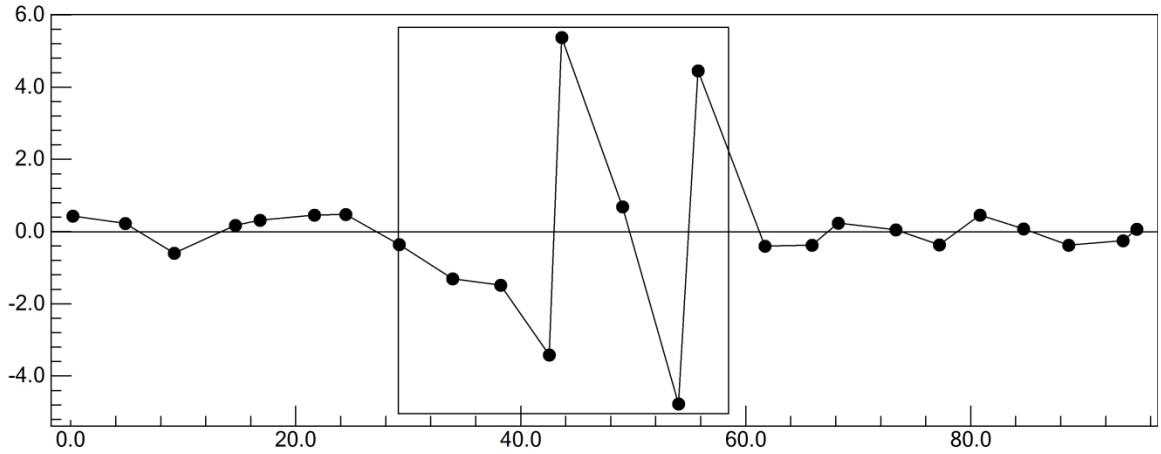


Figure 2-19 Residuals when all errors are in the values of X. The box represents the statistically significant interface region used for the F test.

2.8.5 Tools > Smooth Data

When Smooth Data on the Tools menu is selected, an n (where n is the square root of the number of data) point quadratic fit of the data is performed, replacing the center datum with its calculated value from the fit. The purpose of this is to examine the effects of one particular type of data smoothing on the statistics associated with the least squares fit. In particular, examining the residuals shows the effects of the averaging. This "smoothing"can be performed as often as desired. Once data has been smoothed at least once, the option of restoring the original data is offered on the Tools menu.

Smoothing alters the functional dependence of the resulting values of Y on X. If the original data obey an extended logistic function, the smoothed data will not and will exhibit systematic errors characteristic of model errors. It is recommended *never* to smooth the data, or, if doing so, to use methods that do not introduce correlation between the data. The fit of the raw data to the extended logistic function is itself a more rational kind of smoothing for the purpose of estimating the interface properties of position, width, and asymmetry.

2.8.6 Tools > Straight Line

Performs a least squares fit of the displayed data to a straight line, displays the resulting line on the graph, and prints the equation for the line on the top of the graph. If a selection box is displayed, the line is estimated for those points in the selection box only.

2.9 Help

Clicking the help menu displays an abbreviated version in html format of the information in this section.

2.10 Conclusion

While the preceding description of the details of the working of the LFPF program may seem overly complicated, it should nevertheless prove to be intuitive and the user is encouraged to simply try it, using this documentation as a reference. If the program crashes because of some circumstance unanticipated by the author, simply restart the program. The authors would greatly appreciate any feedback on problems encountered it using this application.

3 Results of analyses of synthetic interface data using the extended logistic function.

The easiest and most useful approach to assessing the performance of a computer program written to implement the analysis of interfacial data has been to construct data sets calculated using the extended logistic function to which random, normally distributed errors have been added. The random, normal deviates were generated using a Basic version of the Fortran function codes GASDEV and RAN1 found in Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterline, W.T., "Numerical Recipes, The Art of Scientific Computing", Cambridge University Press, 1989, 191-203. 100,000 random, normal deviates generated by these algorithms were compared with the table of 100,000 random normal deviates originally published in 1955 by the Rand Corporation ("A Million Random Digits with 100,000 Normal Deviates," Rand Corporation monograph no. MR-1418-RC, Rand Corpopation, Santa Monica, CA 2001.) The two sets of numbers had comparable first, second, third, and fourth moments.

Ten thousand data sets consisting of 100 X,Y data pairs, 25 X,Y data pairs, and 7 X,Y data pairs were analyzed. The values of X in each of the data sets were evenly spaced. Random errors with a standard deviation (square root of the variance) equal to 1% of the separation between *A* and *B* were added to the Y values only. The results of these analyses are presented in Tables 1 – 3.

Table 3-1 Extended Logistic Function fit to 1000 data sets of 100 data each

Parameter	True Value	Average Value	Std Dev of Ave	Average Std Dev	Minimum Value	Maximum Value	% Beyond Confidence Limits			
							80%	90%	95%	99%
A	0	0.002	0.23	0.23	-1.15	0.82	20.2%	10.2%	5.4%	1.2%
B	100	100.00	0.16	0.16	99.33	100.63	19.4%	9.8%	4.9%	0.9%
X_0	50	50.00	0.06	0.06	49.75	50.21	19.8%	10.0%	5.1%	1.0%
D_0	4	4.00	0.05	0.052	3.81	4.24	20.2%	10.0%	4.8%	1.1%
Width (14.5%)	14.53	14.53	0.20	0.20	13.79	15.53	20.2%	10.2%	5.2%	0.9%
Width (dY/dX)	13.18	13.17	0.21	0.22	12.34	14.05	19.3%	9.6%	4.9%	1.0%
Q	0.05	0.0500	0.0041	0.0040	0.0353	0.0669	19.8%	10.2%	5.1%	1.1%
η	0.174	0.174	0.014	0.014	0.123	0.233	19.8%	10.4%	5.0%	1.2%
QD_0	0.2	0.200	0.016	0.016	0.140	0.271	19.7%	10.4%	5.1%	1.2%
η *(dY/dX)*	0.118	0.118	0.009	0.009	0.084	0.154	19.9%	10.5%	5.1%	1.2%
X at $(dY/dX)_{max}$	51.50	51.50	0.14	0.14	50.97	52.05	19.7%	10.0%	5.3%	1.2%
Std. Dev.	1	0.997	0.073		0.755	1.285	20.2%	10.0%	5.1%	1.1%
True Std. Dev.	1	0.997	0.071		0.749	1.267	20.3%	10.2%	5.3%	1.0%
Scatter	1	0.988	0.112		0.590	1.458	36.2%	24.1%	16.5%	7.1%
True Scatter	1	0.994	0.109		0.610	1.458	39.7%	27.6%	19.3%	8.8%
F Test							17.5%	7.9%	4.0%	1.0%
Outliers							20.2%	10.0%	5.0%	0.9%
True Outliers							20.0%	10.0%	5.0%	1.0%

Table 3-2 Extended Logistic Function fit to 1000 data sets of 25 data each

Parameter	True Value	Average Value	Std Dev of Ave	Average Std Dev	Minimum Value	Maximum Value	% Beyond Confidence Limits			
							80%	90%	95%	99%
A	0	0.00	0.48	0.48	-1.82	1.67	20.2%	10.1%	5.0%	1.1%
B	100	100.01	0.32	0.32	98.76	101.15	20.0%	9.8%	4.9%	1.0%
X_0	50	50.00	0.12	0.12	49.54	50.43	20.0%	10.4%	5.1%	0.9%
D_0	4	4.00	0.10	0.10	3.60	4.46	19.9%	9.9%	5.2%	1.0%
Width (14.5%)	14.53	14.55	0.41	0.41	13.07	16.29	19.6%	9.7%	5.1%	1.1%
Width (dY/dX)	13.18	13.16	0.43	0.43	11.43	14.85	19.7%	9.5%	4.8%	1.0%
Q	0.05	0.050	0.008	0.008	0.021	0.081	19.8%	9.8%	4.7%	0.9%
η	0.174	0.175	0.027	0.027	0.076	0.274	19.7%	10.0%	4.9%	0.9%
QD_0	0.200	0.201	0.032	0.032	0.086	0.322	19.6%	9.9%	4.8%	0.8%
η *(dY/dX)*	0.118	0.118	0.018	0.018	0.052	0.178	20.0%	10.4%	5.3%	1.0%
X at $(dY/dX)_{max}$	51.50	51.50	0.27	0.27	50.40	52.50	19.9%	10.2%	5.1%	1.0%
Std. Dev.	1	0.988	0.158		0.449	1.647	20.1%	10.1%	5.3%	1.0%
True Std. Dev.	1	0.991	0.141		0.515	1.589	19.9%	10.0%	4.9%	1.0%
Scatter	1	1.134	0.280		0.290	2.157	58.3%	43.4%	32.2%	15.0%
F Test							14.0%	9.1%	5.4%	1.6%
Outliers							20.8%	10.0%	4.7%	0.6%
True Outliers							20.1%	10.1%	5.0%	1.0%

Table 3-3 Extended Logistic Function fit to 1000 data sets with 7 data each

Parameter	True Value	Average Value	Std Dev of Ave	Average Std. Dev	Minimum value	Maximum value				
							80%	90%	95%	99%
A	0	0.70	1.12	0.98	-4.75	3.78	30.1%	14.8%	7.2%	1.5%
B	100	100.17	0,63	0.63	97.85	102.83	17.6%	8,7%	4.5%	1.0%
X_0	50	49.39	0.70	0.66	46.10	51.33	43.7%	25.2%	10.8%	2.2%
D_0	4	3.999	0.163	0.188	3.372	4.653	13.9%	6.9%	3.4%	0.9%
Width (14.5%)	14.53	14.36	0.68	0.65	12.26	18.01	18.6%	8.7%	4.4%	0.8%
Width (dY/dX)	13.18	13.55	0.93	1.26	7.95	16.13	20.8%	9.8%	5.3%	1.7%
Q	0.05	0.030	0.040	0.040	-0.168	0.114	24.4%	12.9%	6.6%	1.3%
η	0.174	0.099	0.134	0.128	-0.447	0.365	24.8%	13.1%	6.6%	1.3%
QD_0	0.2	0.115	0.156	0.154	-0.606	0.452	24.7%	13.1%	6.7%	1.3%
η *(dY/dX)*	0.118	0.066	0.088	0.081	-0.249	0.222	25.0%	13.3%	6.8%	1.4%
X at $dY/dX)_{max}$	51.50	50.39	1.85	1.72	43.64	53.59	15.6%	8.4%	4.2%	0.8%
Std. Dev.	1.003	1.020	0.523		0.009	3.054	30.2%	18.1%	10.7%	3.2%
True Std. Dev.	1.003	0.967	0.266		0.223	2.235	20.9%	10.2%	5.1%	1.1%
Scatter	"1"	14,2	0.7		11.5	16.9	100%	100%	100%	100%
F Test							0.0%	0.0%	0.0%	0.0%
Outliers							28.3%	8.9%	6.0%	5.3%
True Outliers							20.1%	10.2%	5.2%	1.0%

The distribution of results appearing in Tables 1 - 3 are consistent with what would be expected from random errors, normally distributed. The number of parameter values falling outside their 80%, 90%, 95%, and 99% confidence limits is consistent with a normal distribution. The square root dependence of the parameter standard deviations on sample size is also obvious in comparing the tables. Since, in each data set, errors were drawn from a population of random normal deviates with a standard deviation of 1.000, one can compare the "true" standard deviation for each data set with that returned by the least squares fit of an exact extended logistic

function to which those errors had been added. These are given in Tables 1 - 3 with the heading "True Std. Dev." It can be seen that the least squares fit of the extended logistic function returns consistent estimates of the standard deviations when compared to the values added to the calculated data. This is also borne out in the comparison of the average of the standard deviations of each parameter returned by the fits with the standard deviation of the average of the parameter values returned by the fits. As the number of data in the sets increases, these numbers become closer. When the number of data in each set was 7, however, the beginnings of the breakdown in the statistics can be seen with lower or higher than expected values falling outside the confidence limits. In fact, the value of Q (and associated asymmetries) could be determined for only 6,400 of the 10,000 data sets. With only 2 or 3 degrees of freedom, it is surprising that the confidence limits are obeyed as well as they appear to be. In fact, it is fortuitous. In these data sets, the separation between data was equal to 2.5 D_0 with, on average, 2.3 data values falling in the statistically significant interface region. When the spacing was increased to 3.5 D_0 with, on average, 2 data values falling in the statistically significant interface region, the fits were not nearly as good with larger ranges for the returned values, particularly X_0, and larger than expected numbers of parameter values falling outside their confidence regions. If the number of data in each set is increased from 7 to 10, all sets returned values for Q and the distribution of parameter errors followed the student's t distribution more closely. Similarly, when the standard deviation of the errors added to the data sets (of 7 data) is decreased to 0.2 % of the separation between *A* and *B*, the distribution of errors again follows a student's t distribution.

In Table 3-4, the data scatter estimated from third differences is compared with the standard deviation of the fit as a function of the number of data. One thousand data sets were analyzed. The sets contained varying number of data from 7 to 4000.

The F test compares the value of $\left(\varepsilon_{3d} / s\right)^2$ with the value of F for the F distribution for the corresponding degrees of freedom (See Section 4.2.3 below.) For the data sets consisting of 100 data, 20% of the sets should give values of $\left(\varepsilon_{3d} / s\right)^2 < (1.00 - 0.10)^2$ or $> (1.00+0.10)^2$ and out of 1000 data sets, 15.6% did. This reflects the effect of the correlation between neighboring third difference values, showing a broader distribution for the data scatter than for a χ^2 distribution. Except only for the very small number of data per set, the distribution of data scatter was sufficiently close to a normal distribution that it might prove

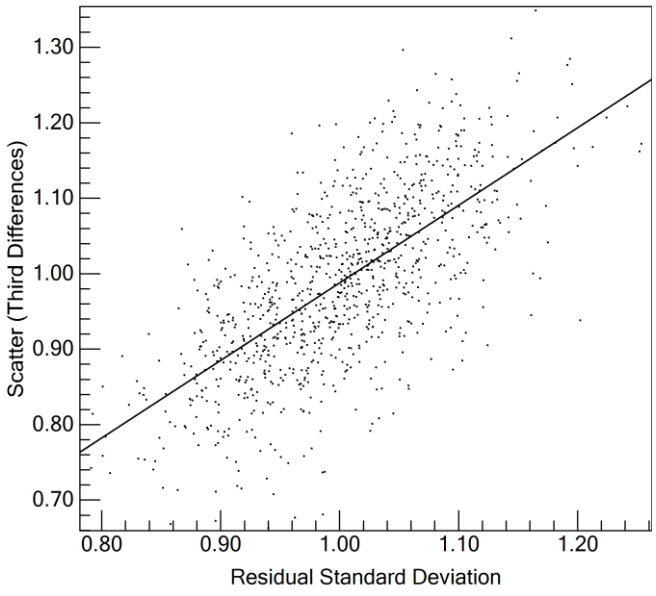


Figure 3-1 Scatter as a function of the Residual Standard Deviation for 1000 analyses of data sets containing 100 data.

useful for signaling the possible presence of systematic errors. However, in the following section dealing with systematic errors, it proved for the case tested, only marginally useful.

Table 3-4 Comparison of Scatter and Standard Deviation for 1000 sets of data with varying numbers of data in the set

Number of Data	Average Ratio of ε_{3d} / s	Confidence Limits of Ratio				Min Value	Max Value	% Failing the F test beyond F distribution confidence limits			
		80%	90%	95%	99%			80%	90%	95%	99%
4000	1.00	0.016	0.021	0.025	0.033	0.959	1.039	13.7%	5.9%	2.2%	0.2%
1000	1.00	0.033	0.042	0.051	0.067	0.895	1.087	16.7%	6.5%	1.9%	0.3%
100	1.00	0.10	0.13	0.16	0.21	0.70	1.24	15.6%	6.0%	2.7%	0.7%
25	0.98	0.20	0.26	0.31	0.41	0.44	1.44	16.8%	8.0%	3.2%	0.8%
10	0.78	0.28	0.36	0.42	0.56	0.12	1.22	40%	21%	11%	3%
7	0.59	0.23	0.29	0.35	0.46	0.19	1.04	56%	25%	14%	4%

It is interesting to note that as the number of data increase, the spread in the ratio of scatter to standard deviation decreases, but the number of values falling beyond the χ^2 distribution confidence limits remains stable. This reflects the parallelism of the scatter and standard deviations apparent in Figure 3-1exhibiting the full set of values of the scatter, ε_{3d} as a function of the residual standard deviation, s, for 1000 analyses of 100 data.

3.1 Notes on Systematic (Model) Errors

As pointed out in this documentation on several occasions, the uncertainties presented by the least squares fit of the logistic function have little significance beyond what to expect for the parameter values were the measurement of the profile to be repeated because of systematic deviations from a pure logistic function. The results of these tests do demonstrate that the extended logistic function can be used to provide a width and asymmetry of a profile in a systematic and reliable manner.

The above tests were repeated with randomly spaced values of X and with different random spacing patterns for each data set. No change in the values of the parameters or the distribution of the reported confidence limits was detected.

The tests were repeated yet again, but this time with random *errors* added to the X values. That is, for each value of X a value of Y was calculated *after which* a random error was added to the X value. In this case, the uncertainties for the interface parameters X_0, D_0, and Q, were slightly underestimated but not by much as long as the errors in X produced *apparent* errors in Y comparable to the *true* errors in Y. On the other hand, the uncertainties for the asymptotes, A and B were overestimated. As anticipated, when errors were added to the values of X alone, the confidence limits on the values of the parameters no longer reflected the student's t distribution. This is not surprising since placing all the errors in X gives rise to effective errors in Y that are larger in the interface (where Y is rapidly varying) and vanishing in the wings of the distribution which can be seen in examining the residuals shown in Figure 3 2. The student's t distribution in this case is based on the incorrect assumption that the errors in the wings of the profile are the same as the errors in the interface region; hence the confidence limits for A and B are overestimated while those for X_0, D_0 and Q are underestimated as judged by the overall standard deviation for all the data. This is a form of a model error, the model being a normal distribution of errors in Y. In this case the F test indicated model errors in all 1000 sets.

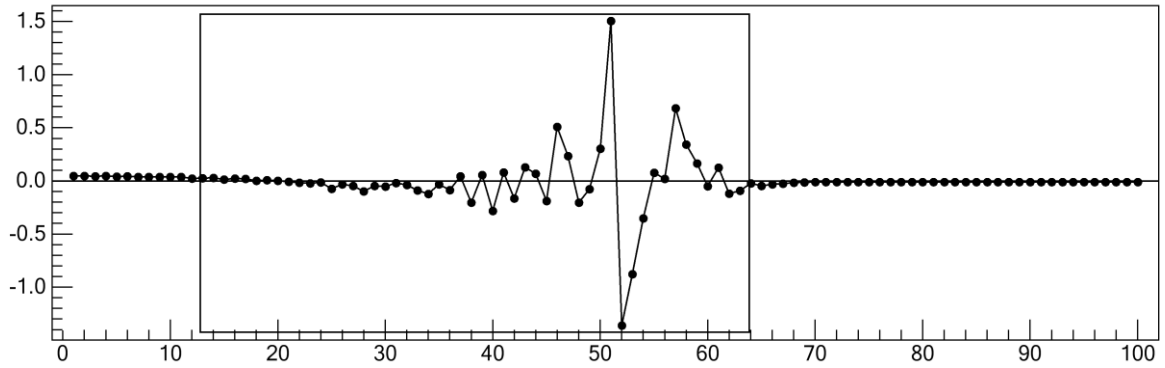


Figure 3-2 Residuals when all errors reside in X values but are assumed to reside in Y values. The rectangle represents the statistical interface region

When, instead of an exact extended logistic function, an incomplete gamma function, $\Gamma(s,x) = 1/\Gamma(s)\int_0^x e^{-t}t^{s-1}dt$, is used as the basis for generating test data, systematic, i.e., model, errors will be present which can affect the interpretation of the results. The incomplete gamma function, unlike the error function, presents an asymmetric profile. A fit of the extended logistic function to an incomplete gamma function is shown in Figure 3-3.

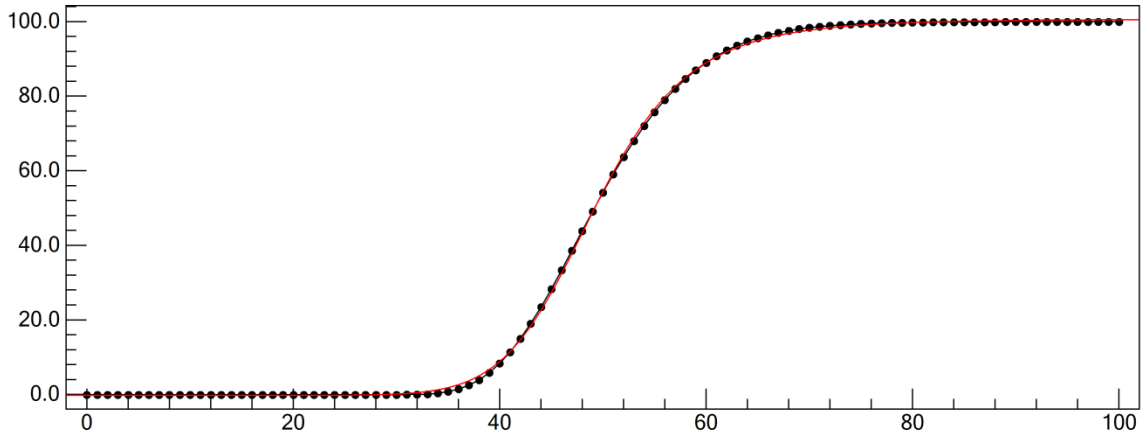


Figure 3-3 Analysis of data representing an incomplete gamma function

While the agreement between the calculated fit and data appears quite good in Figure 3-3, the residuals, the values of the incomplete gamma function minus the fitted logistic function, clearly show the presence of model errors in Figure 3-4. The standard deviation of the fit is 0.5199, half a percent of the range in Y, compared to the estimate of the standard deviation from third differences of 0.0039 (with the systematic error minimized though not eliminated.) For comparison, the data scatter (third differences) are also displayed in Figure 3-4 and seen to be barely discernible. The standard deviation estimated by third differences is due exclusively to the functional difference between neighboring *residuals*. (The value of $F = (0.52)^2/(0.0039)^2 = 17{,}777$ was far in excess of $F_{95,99,0.95} = 1.40$ though such a comparison is inappropriate for completely non-statistical errors.)

When random normal deviates are added to the incomplete gamma function and then analyzed with LFPF the presence of systematic errors is detectable but much less so.

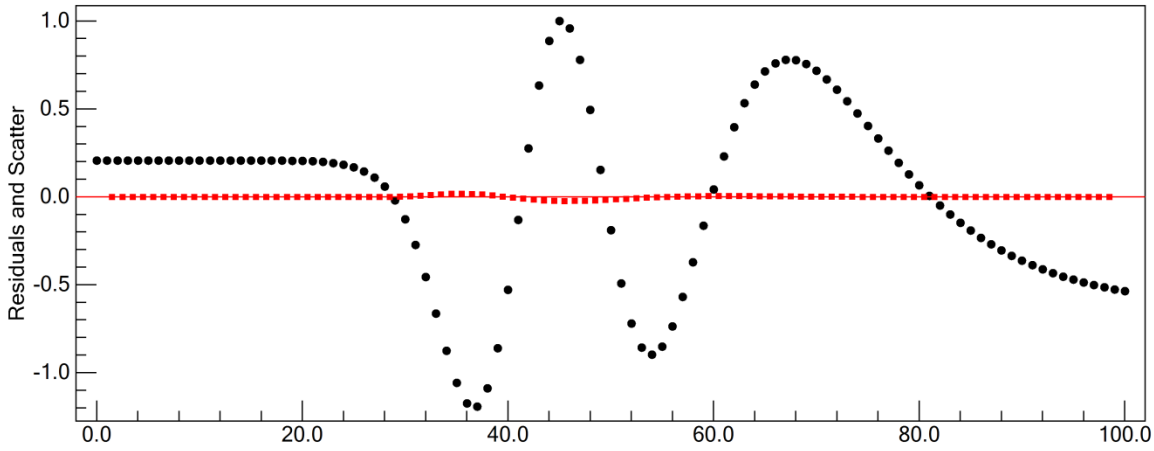


Figure 3-4 Residuals from a fit of the extended logistic function to an incomplete gamma function along with the data scatter (red points, deviation from 0 barely discernible.)

Two series of numerical experiments were performed. In both cases, 1000 data sets, based on the 100 data values depicted in Figure 3-3 and Figure 3-4, were analyzed. To each data set, random normal deviates were added. In the first series, the standard deviation of the random errors, 0.5, was comparable to the standard deviation of the fit of the exact incomplete gamma function of Figure 3-3. The results are summarized in Table 3 5.

Table 3-5 Summary of 1000 fits of the extended logistic function to an incomplete gamma function to which random errors *comparable to the systematic error* have been added.

Parameter		True Value	Average	Difference	Average Standard Deviation	Per Cent Beyond Confidence Intervals			
						80.0%	90.0%	95.0%	99.0%
A		0	-0.20	-0.20	0.13	68.4%	47.8%	28.9%	7.2%
B		100	100.64	0.64	0.17	99.6%	99.6%	99.2%	95.2%
X_0		49.17	49.23	0.057	0.05	47.2%	28.1%	15.8%	2.5%
X at $(dY/dX)_{max}$		47.5	48.08	0.58	0.11	99.9%	100.0%	100.0%	100.0%
Width	12 to 88%	18.29	18.61	0.32	0.17	79.5%	60.6%	44.0%	14.3%
	14.6 to 88.4%	16.36	16.41	0.05	0.15	9.3%	2.8%	0.8%	0.0%
	16 to 84%	15.46	15.41	-0.05	0.14	10.5%	4.0%	1.2%	0.0%
	20 to 80%	13.07	12.84	-0.23	0.11	86.2%	70.6%	53.5%	22.4%
	25 to 75%	10.47	10.15	-0.32	0.09	99.9%	99.8%	99.2%	92.2%
	dY/dX half hgt	17.74	15.73	-2.00	0.15	99.9%	100.0%	100.0%	100.0%
η	12 to 88%	-0.125	-0.128	-0.003	0.010	7.2%	1.4%	0.4%	0.0%
	14.6 to 88.4%	-0.112	-0.113	-0.001	0.009	5.7%	1.2%	0.3%	0.0%
	16 to 84%	-0.106	-0.107	-0.001	0.009	5.1%	0.9%	0.2%	0.0%
	20 to 80%	-0.090	-0.089	0.001	0.007	5.7%	1.3%	0.3%	0.0%
	25 to 75%	-0.072	-0.071	0.001	0.006	7.4%	2.0%	0.3%	0.1%
	dY/dX half hgt	-0.130	-0.077	0.053	0.006	99.9%	100.0%	100.0%	100.0%
Residual Std. Dev.		0.5017	0.7215	0.2198		100.0%	100.0%	100.0%	100.0%
"True" Var.		0.5017	0.5004	-0.013		21.5%	10.1%	5.3%	1.3%
Data Scatter		0.4984	0.4946	-0.0038		38.2%	27.6%	17.7%	7.1%
Outliers						20.6%	10.2%	5.0%	0.8%
$\mathcal{E}_{3d}$ / S F test						99.8%	98.9%	96.3%	87.1%

The width, W_f, and asymmetry, η_f, are given for commonly used values for f. The true values of the parameters are those of the exact incomplete gamma function itself. Because the parameters D_o and Q have no relevance to the incomplete gamma function, there are no "true" values for these quantities and they are not included in the table. The average value is the average over the 1000 data sets. The confidence limits for the parameters are derived from the values of the parameters determined by the fit minus the "true" value divided by the standard deviation of the parameter from the fit. These ratios should follow a student's t distribution and the percentage of values falling beyond the student's confidence limits for 95 degrees of freedom appear in the table.

In the second series, random normal errors from a population with a standard deviation twice that of the first series, i.e. twice the systematic error, were added. The results are summarized in Table 3-6.

Table 3-6 Summary of 1000 fits of the extended logistic function to the incomplete gamma function to which random errors *twice the systematic error* have been added.

Parameter		True Value	Average	Average Standard Deviation	Per Cent Beyond Confidence Intervals			
					80.0%	90.0%	95.0%	99.0%
A		0.00	-0.20	0.20	37.0%	21.8%	11.3%	3.2%
B		100.00	100.64	0.26	89.3%	81.5%	69.0%	41.5%
X0		49.17	49.23	0.07	30.5%	17.8%	10.2%	2.2%
X dY/dX max		47.50	48.08	0.17	98.7%	97.4%	92.6%	78.8%
Width	12 to 88%	18.29	18.61	0.27	47.8%	31.2%	18.9%	5.1%
	14.6 to 88.4%	16.36	16.41	0.23	17.0%	7.1%	2.3%	0.3%
	16 to 84%	15.46	15.41	0.22	15.8%	7.8%	3.3%	0.2%
	20 to 80%	13.07	12.84	0.16	55.6%	43.8%	32.1%	15.2%
	25 to 75%	10.47	10.15	0.14	86.5%	76.5%	63.3%	37.0%
	dY/dX width at half hgt	17.74	15.74	0.23	99.6%	99.7%	99.7%	99.7%
η	12 to 88%	-0.125	-0.128	0.016	13.8%	5.3%	1.7%	0.2%
	14.6 to 88.4%	-0.112	-0.113	0.014	13.4%	5.1%	2.0%	0.1%
	16 to 84%	-0.106	-0.106	0.014	13.6%	4.7%	2.1%	0.1%
	20 to 80%	-0.090	-0.089	0.010	20.2%	10.7%	5.1%	1.0%
	25 to 75%	-0.072	-0.071	0.009	13.5%	5.9%	2.2%	0.1%
	dY/dX width at half hgt	-0.130	-0.077	0.010	99.6%	99.7%	99.7%	99.7%
Variance		1.00	1.1287		20.9%	10.9%	5.2%	1.4%
"True" Var.		1.00	1.0007		22.4%	10.8%	5.7%	1.8%
Data Scatter			0.9968		43.9%	32.4%	22.2%	10.6%
ε_{3d} / S F test					67.2%	47.6%	30.3%	11.8%

The extended logistic fit gave a residual standard deviation of s = 1.13 ± 0.08 in good agreement with the true value of 1.0. For the incomplete gamma function, the width and asymmetry as measured in the vicinity of f = 15% gave good agreement with the true value. On the other hand, the agreement between the maximum of dY/dX from the logistic function and that from the incomplete gamma function as well as the width at half height and the corresponding value of η were quite poor compared with the standard deviation for these quantities returned by the fit. Agreement is much better the smaller the asymmetry. Different systematic differences may give better agreement at different percentage points.

If, instead of comparing the values of the parameters with their "true" values, we made the comparison with the average of the 1000 values returned from the individual fit, the number of

values falling beyond the confidence intervals proved to be less than the expected number. This is to be expected because each set has the same systematic error which the average over all sets takes into account. But this also shows that as long as the systematic error is constant, repeated measurements will result in differences from set to set less than the statistics associated with a normal distribution of errors would indicate.

The F test, invoked to compare the standard deviation of the fit with the data scatter estimated from third differences, proved to be a moderately strong test for the second series with 30% of the sets failing at the 95% level. Had a smaller number of data been used, the distribution of F test failures would be just that expected for two samples from the same population of errors.

The usual means for estimating the presence of systematic errors is in examination of the residuals shown in Figure 3-5 to Figure 3-7 below.

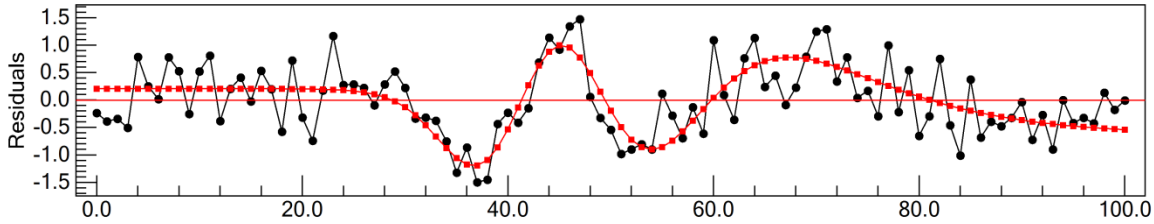


Figure 3-5 Residuals from a fit of a logistic function to an incomplete gamma function with random normal *errors comparable to the systematic error*. The residuals from the fit of the exact incomplete gamma function are shown in red. The systematic error can be easily discerned in the residuals from the data with random errors plus systematic error.

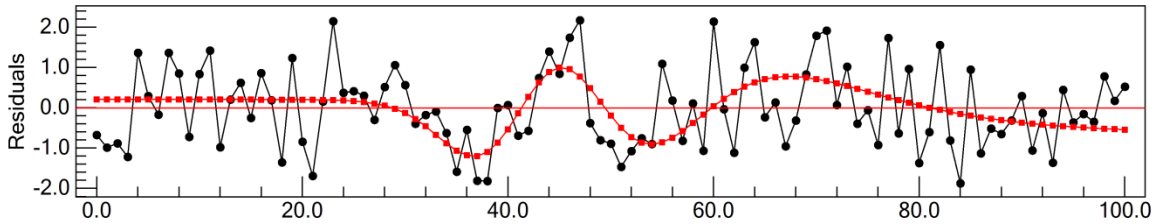


Figure 3-6 Residuals from a fit of a logistic function to an incomplete gamma function with random normal *errors twice that of the systematic error*. The residuals from the fit of the exact incomplete gamma function are shown in red. The systematic error can still be discerned but is aided by the superposition with the systematic error.

Figure 3-5 corresponds to one of the data sets contributing to Table 3-5 with random errors comparable to the systematic error. The systematic error is obvious. Figure 3-6 corresponds to one of the data sets contributing to Table 3-6 with random errors twice that of the systematic error. In this case, the systematic error is less obvious but easily seen when superimposed with the systematic error (from a fit of the exact incomplete gamma function data.) If we look at the same residuals without the random error superimposed, Figure 3-7, it is not at all obvious that a systematic error would have been detected. If the residuals in Figure 3-5 are fit with a Fourier series five terms are found to have t values (parameter value divided by its standard deviation) greater than the 95% confidence limit. Moreover, the inclusion of these five terms reduced the standard deviation of the fit below its lower chi-square confidence limit and comparable to the added random errors. If the residuals in Figure 3-6 are fit with a Fourier series, no terms are found to be significant.

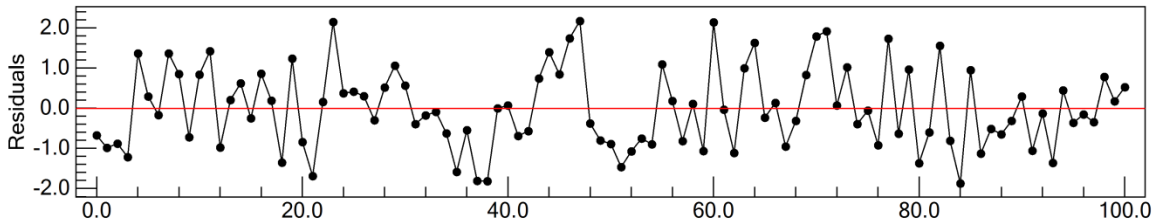


Figure 3-7 Residuals from a fit of a logistic function to an incomplete gamma function with random normal errors twice that of the systematic error. This is the same graph as that appearing in Figure 3-6 but *without* the systematic error superimposed, the systematic error in the noisy data is less obvious.

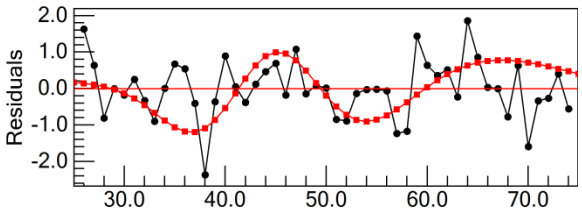


Figure 3-8 Residuals of a fit of the central 49 data points from data consisting of an incomplete gamma function with errors on the order of twice the systematic error

On the other hand, had we performed the same numerical experiments with data sets consisting of only 25 data, it would not have been possible to spot the systematic error.

In line with the observation made above that the asymptotic data appear to absorb the systematic errors, data from the central half of the data were analyzed, throwing out the upper and lower quartiles. The number of parameters exceeding their confidence levels increased, particularly for the interface parameters X_0, W_f, and η_f. When the average values are substituted for the true values, the values falling outside the confidence levels were about 3/4 as many as expected. The residuals of the set giving the largest value of $\left(W_f\left(obs\right)-W_f\left(true\right)\right)/s\left(W_f\right)$, where $s\left(W_f\right)$ is the standard deviation of W_f from the least squares fit, show no hint of a systematic error as can be seen in Figure 3-8.

Does any of this really matter? Probably not since model errors are generally unknown so that the measured profile, even though it may contain systematic differences from a logistic function profile, will generally be close enough that the model errors will be small or even undetectable. The analysis in terms of the logistic function gives a central position, a width, and an asymmetry to a measured profile that can serve as a description of a depth profile or the lateral resolution of a surface line scan.

3.2 Difficult Data and Analysis Instabilities

In the preceding discussion, all of the data sets discussed included values in the asymptotic regions and at least three data with values lying between the asymptotes and more than a confidence limit away from each asymptote (which we call here the statistically significant interface region, see Section 2.4.1 above). In all cases, the standard deviations were one percent of the separation between the asymptotes. Data this well behaved may be encountered often, but not always. We therefore discuss briefly three cases of difficult data, namely, incomplete data for which one of the asymptotes is not reached, i.e. where the values at one end or the other are more than 5% of the asymptotic separation away from the corresponding asymptote; very sharp interfaces in which only one datum or none falls in the statistically significant interface region; and data with errors on the order of 10% or greater of the separation between asymptotes, so-called noisy data.

The extended logistic function is continuous and well behaved (with the exception of a singularity when $D \rightarrow 0$, with analytic first derivatives of Y with X and first derivatives of Y with each of the function's parameters. However each parameter is sensitive only to certain regions of the profile. A (as well as A' and A'') is sensitive to the pre-interface and the early stages of the interface and B (as well as B' and B'') to the late stages of the interface and the post-interface region. The interface parameters X_0, D_0, and Q are sensitive only to data in the interface, significantly distant from either asymptote, i.e. to data in the statistically significant interface region. In addition, correlation between parameters (such as Q and A' if Q is positive or Q and B' if Q is negative) can cause indeterminacy in the parameters and the iterative process can not only converge slowly, it can also diverge.

Generally, a fit of all the desired parameters is first attempted and if it appears the iterative procedure may be diverging, certain parameters are held fixed at predetermined values depending on the parameter and the structure of the data as revealed by the nature of the divergence. The divergence tests include the following:

- A 10% increase in the standard deviation of the fit from one iteration to the next when the number of parameters being varied has not changed
- An increase by a multiplicative factor on the order of 2 (the value of the normal distribution confidence limit for the selected confidence level is the actual, somewhat arbitrary, factor chosen) in the standard deviation of any parameter value when the number of parameters being varied has not changed.
- A correction to D_0 that would make its value negative or confidence limits for D_0 that are larger than the magnitude of D_0, i.e., include 0 in the range.
- A correction to X_0 that would move its value beyond the interface width (by default the 14.6% and 85.4% limits) or confidence limits for X_0 that are greater than the interface width

These tests are performed before the corrections are added to the parameters being varied. When any of these tests indicates divergence, one of the parameters being fit is fixed at its current value or some predetermined value, depending on the situation, and corrections to the remaining parameters are ignored for that one iteration where the divergence was noted. Descriptions of the actions taken are included in the analysis notes, examples of which will appear in the following sections.

3.2.1 Incomplete Profiles

If the data being fit do not reach one or another of the asymptotes, it becomes problematic to form an initial estimate of the asymmetry parameter Q because the value of Q depends primarily on the difference in curvature in the profile near the pre- and post-interface asymptotes. It may still be possible to determine Q if the level of noise in the existing data is small, but it is generally not possible to make an initial estimate for the value of Q. Hence Q is initially set to zero but may still be varied in the iterative analysis. For the same reason, the slope of the undeveloped asymptote is held fixed at 0. A typical result appears in Figure 3-9 below where only the data in the displayed selection box have been included in the fit in order to demonstrate effect of an incomplete profile.

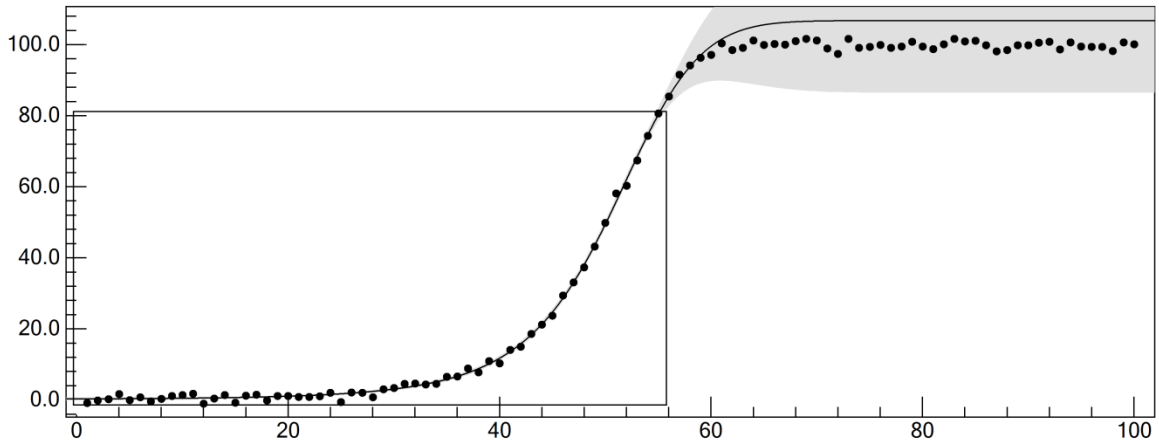


Figure 3-9 Fit to incomplete data. Only data in the box were included in the fit. The gray area represents the confidence limits for the value of Y calculated from the extended logistic function

The Analysis Notes will contain a message similar to:

> At the final point, X = 55, the interface is only 75.12% complete. The final asymptote is not reached and the confidence limits for Xo, Do, Q, and B may be underestimated.

Fitting 1000 data sets, based on the logistic function with random normal errors added but with the upper, in this case more rapidly converging, portion of the profile incomplete resulted in analyses similar to the one pictured in Figure 3-9. All 1000 analyses produced values for all parameters including *Q* and *B*. The estimated degree of completion from the analyses varied from 52% to 94% (true value = 81%) with corresponding values of *B* ranging from 160 down to 85 (true value = 100), width from 16.23 down to 9.8 (true value 11.3) and the asymmetry from -.01 to +0.26 (true value = 0.138.) Figure 3-10 shows the fits of the data set with the most negative value of $t(B)$ $(=(B-B_{calc})/s(B))$ and most positive value of $t(B)$. For all the parameters the distribution of parameter values was symmetric above and below their true values but on the most negative side, the uncertainties on *B* were typically underestimated.

When the lower portion, the more slowly converging portion, of the profile was incomplete and *A* was poorly determined, only about 97% of the data produced a stable value for *A*. In this case the distribution of values of $t(A)=(A-A_{calc})/s(A)$) gave fewer values at the 80% confidence level than expected and more values at the 99% confidence level than expected.

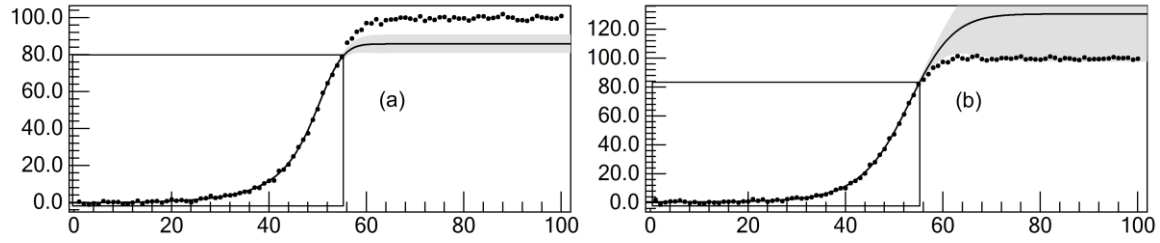


Figure 3-10 Fit to incomplete incomplete data: (a) data set with the most negative value for t(B) (b) data set with the most positive value for t(B). Both data sets based on the same underlying function.

If one were limited to the incomplete data alone, then there would be nothing more one could do other than holding *B* fixed at its true value supposing that true value were known. This situation could occur, for example, if the spectroscopic values for pure species *A* and *B* were known and those spectroscopic properties were the basis for the profile measurement. In the example in Figure 3-9, holding *B* fixed at its true value of 100 resulted in accurate and reliable values for X_0, D_0, and *Q* with confidence limits consistent with a student's t distribution. For the case of the data with systematic error, the analyses were less stable and the variability from one data set to the next was less but the width, W_f, for example, varied by 16% of its value.

If the divergence test for the incomplete asymptote parameter fails, i.e., if the confidence limit for *A* or *B* increases by a multiple of the confidence limit for a normal distribution from one iteration to the next, that asymptotic parameter is held fixed at its most recent stable value, which may be its initial estimate. A sentence in the Analysis Notes similar to the following warns of the action taken:

> Procedure began to diverge on iteration 1. Because B appears to be ill determined (confidence limits = ± 156.0). it has been held fixed at its most recent stable value.

In addition, the value of *Q* may not be determinable because, as already mentioned, it depends on the difference in curvature between the approaches to the two asymptotes. If the iterative procedure continues to diverge, it may be necessary to hold *Q* fixed at 0 and a message similar to the following appears in the Analysis Notes:

> Procedure began to diverge on iteration 4. Assume Xo, Do, and Q could not be determined simultaneously. The confidence Limits for Do, 11.79077, were greater than Do (= 6.354834) on iteration 1 The iterative procedure was continued with Q fixed at 0. The value of Q, determined from the data by varying only Q and holding the remaining parameters fixed at their current values, = 0.0142 ± 0.0247

Note that while the confidence limits for D_0 should have caused D_0 to be held constant at its most recent stable value, setting the value of *B* removed the instability for D_0 and the iterative procedure continued until it became necessary to hold *Q* fixed at 0. *Q* will always be held fixed at 0 and another iteration attempted before setting the values of X_0 or D_0. Note also that *Q* can always be held fixed at any value set by the user.

3.2.2 Sharp Interface Regions

When only one or no datum falls within the interface region, as in **Error! Reference source not ound.** below, it is not possible to determine *Q* and D_0 and/or X_0 though limits may be placed on their values based on the standard deviation of the data.

Defining A as the point just before the interval (because of its proximity to the asymptote A) and B as the point just after the interval, the value of X_0 will lie somewhere between X_A and X_B. Similarly, the upper limit for the value of D_0 will be less than that which would cause Y_A to differ from its asymptote, A, by more than the confidence limit for the value of Y_A and Y_B from its asymptote, B, by more than the confidence limit for the value of Y_B. Therefore, from Equation (4-6),

$$X_A < X_0 - D_0 \ln\left(\frac{1-f_A}{f_A}\right) \quad \textit{where} \quad f_A = \frac{CL(Y_A)}{|B-A|} \qquad (2\text{-}5)$$

and $X_B > X_0 - D_0 \ln\left(\frac{1-f_B}{f_B}\right)$ *where* $f_B = \frac{CL(Y_B)}{|B-A|}$ (2-6)

and where CL(Y_A) and CL(Y_B) are the confidence limits on the values of Y_A and Y_B.

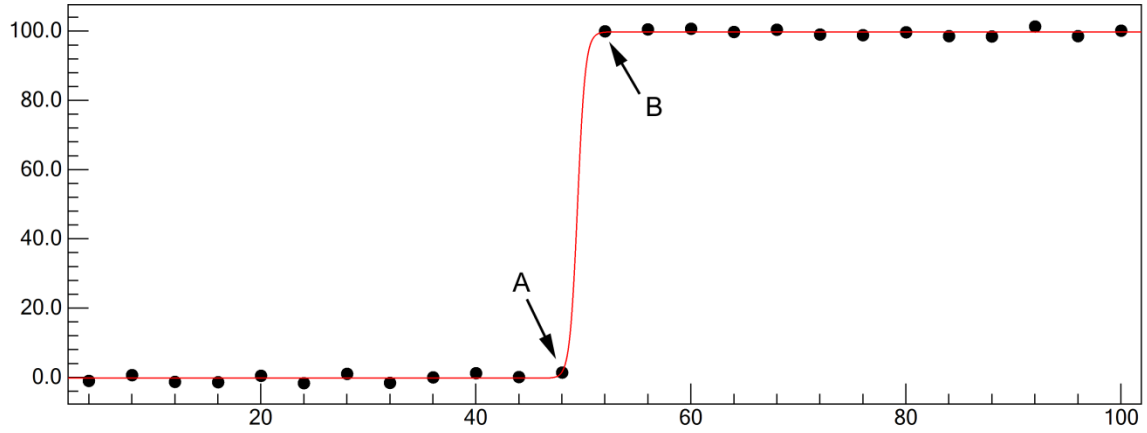


Figure 3-11Analysis of a profile with no data in the interface

Generally, the confidence limits on both are the same and equal to the standard deviation multiplied by the cumulative factor for a normal distribution of unit variance, namely, 1.96 for the 95% confidence limit. This gives

$$D_0 < \frac{X_B - X_A}{2\ln\left(\frac{1-f}{f}\right)} \quad \textit{where} \quad f = f_A = f_B \tag{2-7}$$

The confidence limits CL(Y_A) and CL(Y_B) are calculated from the one tailed χ^2 distribution for the stated confidence limit and the standard deviation of the fit. This gives the maximum value for the standard deviation and from this, the one tailed probability that a measured point is not a random error for a normal distribution at the stated confidence limit, but in fact lies within the statistically significant interface region. The upper limit for D_0 from Equation (2-7) is referred to in the analysis notes as the statistical upper limit of D_0.

If any of the divergence tests for D_0, X_0, or Q fails, Q is first set equal to 0 and the iterative step is repeated without any changes in the remaining parameters.

If any of the divergence tests for D_0 fails (whether or not any of the divergence tests for X_0 fails), and Q is not being varied, the value of D_0 is held fixed at its most recent stable value (which could be its initial value) *unless* it is less than half the statistical upper limit for D_0 or greater than the statistical upper limit for D_0 whereupon it is set equal to D_0(upper limit)/2 or D_0(upper limit) respectively. The iterative step is repeated without adding the corrections to the remaining parameters from the iteration where the divergence was noted.

If any of the divergence tests for X_0 fails, the value of X_0 is set equal to the average of the two values of X bordering the interface. Generally this occurs only when no datum falls in the statistically significant interface region. The corrections to the remaining parameters that were

calculated in the iteration where the divergence was noted are ignored and the iterative procedure is then resumed.

In effect this approach gives the values of X_0 and D_0 that one could have obtained by inspection with the exception of the factor $2\ln\left((1-f)/f\right)$ appearing in Equation (2-7) and the averaging of the asymptotic limits.

Reports on all of the above actions appear in the Analysis Notes following the iterative least squares fit along the lines of the following:

> The iterative procedure began to diverge on iteration 0.
> Assume Xo, Do, and Q could not be determined simultaneously.
> The confidence Limits for Do, 35.6, were greater than Do (=0.424) on iteration 0
> The change in Xo, 400.0, was greater than the interface halfwidth, 4.00, on iteration 0
> The iterative procedure was continued with Q fixed at 0.
> The value of Q, determined from the data by varying only Q and holding the remaining parameters fixed at their final, converged values = 0.007 ± 0.386

In this case, several divergence tests failed and because the values of X_0 and D_0 could have been displaced far from their probable values, the iterative procedure was halted and begun again with the initial estimates of the parameters, holding Q fixed at 0. Even so, holding Q fixed at 0 was not enough to force convergence and the standard deviation of D_0 was still too large forcing D_0 to be held at its most likely value with an additional message in the Analysis Notes:

> The confidence Limits for Do, 0.683, were greater than Do (=0.374) on iteration 2
>
> The standard deviation(s) for Xo, Do increased by more than a factor of 1.960 on iteration number 2
> Consequently, Do was held fixed at 0.362, its value determined by varying Do alone at iteration 2.
>
> 0 data with | Y-Asymptote | > 2.20 appeared to fall in the statistically significant interface region
> 4 possible interface values from X = 44.0 to X = 56.0, were tested
> Based on the statistics of the fit, the upper limit for Do was 0.489

While this may give the impression that D_0 can be determined, its value is strongly correlated with the value of X_0 and hence the confidence limits are underestimated by an unknown amount.

Four ensembles of 1000 data sets each were generated using a logistic function as the basis but with a very sharp interface equal to 1/8 the separation between adjacent data. In each ensemble, the data were shifted relative to X_0 so that one or no point fell in the interval. For the ensembles with one point in the interval, X_0 was near one or the other asymptote or in the center. For the ensemble with X_0 equally spaced between X_A and X_B, those two points differed from their asymptotes by 3.4% of the separation between A and B. With a true value of $D_0 = (X_A\text{-}X_B)/8$, and a standard deviation equal to 1% of the separation between A and B, only data falling more than 2.6% from either asymptote would be counted as falling in the interval. Figure 3-12 below represents one of those 1000 data sets in which only one value of Y_A or Y_B (before the addition of random errors) is further from its nearest asymptote by more than 2.6% (Note: the errors added to the data represented by Figure 3-11 and Figure 3-12 were the same.)

It should be noted that with no datum falling in the interface, the interface will be defined by the separation between the two embracing data above and below the apparent midpoint of the interface. Therefore, D_0 will typically have tighter limits placed on it (assuming X_0 is determined by the least squares fit) if no point falls in the interval than when one point does even though the two data sets represent the same values of D_0 and X_0. On the other hand, X_0 will be more

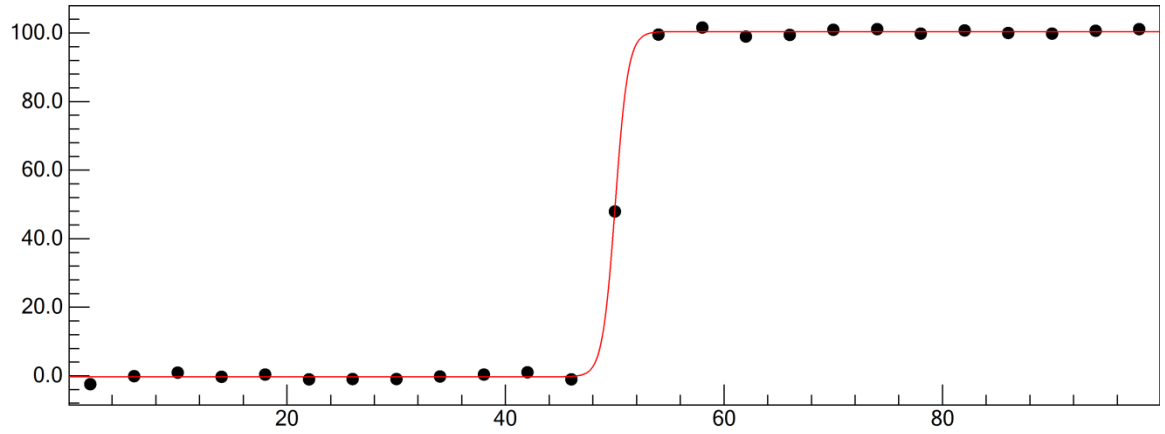


Figure 3-12Analysis of a profile with 1 datum in the interface

difficult to evaluate for no point falling in the interval but will be well determined if one point falls in the interval. In the analysis of the four ensembles of 1000 data each, this was borne out.

For an ensemble with no point in the interval (really two equally spaced points barely in the interval,) of which Figure 3-11is one example, the analyses of 1000 synthetic data sets determined that 444 data sets had one datum in the interval and 154 sets had two. Since the difference between the true values of Y and the corresponding asymptotes are 1.8% (for the parameter values used in this example) we would expect a certain number of the data sets to have values within the 2.2% boundary for being considered within the interval with 95% confidence. In fact we would expect 33% of the data to have at least one point falling between A+0.018(B-A) and B-0.018(B-A) (for a standard deviation of unity) and we observe 44%. We would only expect 11% of the data to have at least two points falling in the interval and we observe, for the sample studied, 15%. The uncertainties returned by the analysis for X_0 and D_0 did not follow a student's t distribution nor were they expected to since the values for the interface parameters, X_0, D_0, and Q were underdetermined. Nevertheless, the uncertainties proved to be conservative.

Since the parameters D_0 and X_0 cannot be independently calculated from profiles with fewer than two data in the interface region, one may wonder why one would try to include their evaluation in this analysis. The answer is to allow for a reasonable determination of D_0 and X_0 without knowing *a priori* that the data would present a sharp interface. The idea is to allow the analysis to proceed without the need for operator intervention and to present values of X_0 and D_0 in a manner close to what one would use without the logistic model and to take advantage of the statistics from the remaining data to place limits on the determination of D_0.

In general, the number of data in the statistically significant interface region must be greater than the number of interface parameters, X_0, D_0, and Q, being varied. See also Equation (4-38) and the discussion following Equation (4-38) on the idea of localized degrees of freedom.

3.2.3 Highly Asymmetric Profiles: Runaway Q

Closely related to the problem of narrow interface regions with few data in the region is the situation where the magnitude of Q becomes quite large (whether positive or negative.) D then varies from its minimum value of 0 to its maximum value of $2D_0$ over a narrower range of X

than the interface itself. The interface region is skewed heavily so that one edge of the region is almost coincident with X_0. When this happens, no data falls between its closest edge and X_0 and the data on the opposite side of X_0 depend solely upon D_0. Q then becomes poorly determined and can increase in magnitude dramatically from one iteration to the next until the least squares fit becomes completely unstable and, frequently, the program aborts.

The program constantly monitors the value of $|QD_0|$ and when its value becomes 10 (the logistic range of D is 1/10 the range of the interface itself) and the correction to Q is greater than Q itself, the value of Q is frozen at its value from the previous iteration and the analysis continues. This places a lower bound on the magnitude of Q. The value of Q is immaterial at this point and only its sign is important. Note also that at this point, the asymmetry η is approaching its limit of ± 1. Because this allows, in effect, the remaining parameters to catch up to the value of Q, it is occasionally possible to achieve convergence by clicking Fit (Converge) a second time.

3.2.4 Noisy Data

We have already mentioned from time to time situations where the standard deviation of the data becomes an appreciable fraction of the spacing between the asymptotes. As long as the data are normally distributed, this does not seem to present too much of a problem. However, as the standard deviation approaches 10% or higher of the separation between A and B, the initial estimates, which depend on individual data rather than the full range of data, can be sufficiently in error to lead to false minima. As a warning, a message something like the following will appear:

> The ratio of the upper limit of the standard deviation from the chi squared distribution to the value of A-B, 19.2%, may make the determination of Xo, Do, and Q problematic and possibly result in false, local minima..

It is sometimes noted that the minimum value for the standard deviation occurred on an iteration before the final one and a note to this effect is included. This can most often occur when one of the parameters cannot be determined and is set to some otherwise determined value.

If an analysis of noisy data begins with initial estimates sufficiently far from their proper values, the procedure may converge very slowly or even diverge. If the procedure is not diverging but has not yet converged, repeated clicking of Fit (Converge) may eventually lead to convergence. Alternatively, or for a procedure that has begun to diverge, it may be possible to obtain convergence by using the data selection box to identify the interface region and restart the analysis by clicking the Initial Estimate button which may give a better estimate of the starting values of the parameters.

The limit on the value of the standard deviation relative to the separation between asymptotes to allow fitting the data depends on the number of data being analyzed. For few data, the statistical advantage of the averaging inherent in the least squares fit is lost. Data sets with larger numbers of data can support larger variations, the $1/\sqrt{n}$ advantage.

3.2.5 Errors in the Independent Variable X

Comment has been made earlier in this manual of the fact that the analysis being described and used in the program LFPF presumes all the statistical error is contained in the values of Y; that

the X values are precisely and accurately known. The statistical behavior described by the program assumes a normal distribution of errors in Y, and numerical experiments with synthetic data are in agreement with behavior expected for a normal distribution. This is true even when the values of X are not evenly spaced but still accurately known.

If the values of X themselves have errors, the analysis described herein and realized in the program LFPF will still ascribe all scatter to a normal distribution in Y. Whereas errors in X in the asymptotic regions will not contribute significantly to the perceived errors in Y, they will in the interface region where small changes in X are accompanied by large changes in Y. In this case, the error distribution is not uniform but, in fact, is larger in the interface region. It may well be that this problem can be overcome with appropriate weighting of the data to reflect this, but in this version of the program, this has not been done. If the errors in X are small compared to the errors in Y, they will not present much of a problem as has been seen to date in analyses of various series of synthetic data.

In examining the residuals, the residuals in the interface region are expected to be slightly smaller than in the asymptotic region because more parameters are affected by their values. (See the discussion following Equation (4-38)) If the residuals in the interface appear larger as exaggerated in Figure 3-2, the effects of errors in the values of X can be easily deduced. If not, it may still be possible to use the F test to compare the standard deviations of the statistical interface region (to which the interface parameters X_0, D_0, and Q are sensitive) with the standard deviations of each of the asymptotic regions.

Consider the separation of the data into three regions, the statistically significant interface and the pre- and post-interface regions. The region prior to the statistically significant interface is dependent almost solely on the parameters A, A', and A''. Similarly, the region following the statistically significant interface is dependent almost solely on the parameters B, B', and B''. While the statistically significant interface depends on all the parameters, it is most sensitive to X_0, D_0, and Q. Since the asymptotic regions are virtually model independent, the variance of those regions will not be sensitive to model errors whereas the statistically significant interface will be. The variances of the three regions are calculated from:

$$s_A^2 = \sum_{i=1}^{n_A} \frac{(Y_i^{obs} - Y_i^{calc})^2}{n_A - p_A}, \quad s_I^2 = \sum_{i=1}^{n_I} \frac{(Y_i^{obs} - Y_i^{calc})^2}{n_I - p_I}, \quad \textit{and} \quad s_B^2 = \sum_{i=1}^{n_B} \frac{(Y_i^{obs} - Y_i^{calc})^2}{n_B - p_B} \qquad (2\text{-}8)$$

n_A, n_I, and n_B are the numbers of data in the pre-interface asymptotic region, the statistically significant interface region, and the post-interface asymptotic region respectively and p_A, p_I, and p_B are the number of varied parameters on which each of the regions is dependent so that the three regions have, respectively, ν_A, ν_I, and ν_B degrees of freedom where $\nu_A = n_A - p_A$, etc. Typically p_A and p_B will each be 1 and p_I will be 2 or 3 depending on whether Q is varied. If $s_I^2 / s_A^2 > F(\nu_I, \nu_A, \alpha)$ or $s_I^2 / s_B^2 > F(\nu_I, \nu_B, \alpha)$ where α is the confidence level for the F distribution, we may have reason to suspect model errors and, particularly for the case under discussion, errors in the values of X. Similar to other situations already mentioned, the more data available for the three regions, the more likely the effect will be noticed.

The values of X are tested and if they are not evenly spaced, a message to that effect appears in the analysis notes much like the following:

NOTE!!! The values of X are not uniformly spaced.
The average spacing is 3.971 with a standard deviation of 10.93%
If the values of X are not error free the parameter confidence limits may be underestimated.
F(interface/pre-interface) = 4.287 compared to F(0.95) = 6.041
F(interface/post-interface) = 7.928 compared to F(0.95) = 3.438.

In this case, the interface/post-interface F test failed indicating the possibility of a model error and more specifically, of possible errors in the X values.

For many data sets encountered in practice where the number of data through an interface is on the order of 25 or less, it will be difficult to extract much from the data in way of determining the aptness of the logistic function model. Inspection of the residuals will give some idea of the importance, if any, of systematic errors compared to random errors. Even so, the analysis will still give systematic and reasonable measures of the position, width and asymmetry of the interface, whether or not the statistics of the least squares fit can be exploited.

4 Detailed discussion of the least squares fit of an extended logistic function to a measured profile

The least squares fit of the measured values of Y to Equation (1-5) is a nonlinear one and is approximated by fitting the linear form,

$$Y_{obs}(X) - Y_{calc}(X, \{C^{|k}\}) = \sum_{i=1}^{m} \frac{\partial Y}{\partial C_i^{|n}} \delta_i^{|k+1}, \qquad (4\text{-}1)$$

i.e., as a Taylor series expansion of Y about the values of $Y_{calc}(X, \{C^{|k}\})$ calculated using the values of the parameters $\{C^{|k}\}$ (= A, A', A'', B, B', B'', X_0, D_0, Q) following the k[th] iteration and where the derivatives $\frac{\partial Y}{\partial C_i}$ are all evaluated using the values of the parameters from the k[th] iteration, namely $C_i^{|k}$. Y_{obs} are the measured values of Y being fit.

The corrections to the parameters $\delta_i^{|k+1}$ are obtained from the linear least squares fit of $Y - Y\{C^{|k}\}$ and the corrected values of the parameters, $C_i^{|k+1} = C_i^{|k} + \delta_i^{|k+1}$, are used for the next iteration. The procedure continues until convergence when the corrections to the parameters are insignificant compared to the uncertainties in the parameters returned by the least squares fit.

A particularly convenient feature of the logistic function is that all of the derivatives of Y with respect to the parameters can be evaluated analytically from a set of current values $\{C_i\}$. In particular,

defining $z = (X - X_0) / D$

$\partial Y / \partial A = 1/\left(1 + e^{z}\right)$

$\partial Y / \partial A' = (X - X_0)(\partial Y / \partial A)$

$\partial Y / \partial A'' = (X - X_0)(\partial Y / \partial A')$

$\partial Y / \partial B = 1/\left(1 + e^{-z}\right)$

$\partial Y / \partial B' = (X - X_0)(\partial Y / \partial B)$

$\partial Y / \partial B'' = (X - X_0)(\partial Y / \partial B')$

The remaining derivatives are more complicated and it helps to further define:

$E_p = 1 + e^{z}$ *and* $E_m = 1 + e^{-z}$

$A^* = (A + A'(X - X_0) + A''(X - X_0)^2) / E_p$

$B^* = (B + B'(X - X_0) + B''(X - X_0)^2) / E_m$

so that

$Y = A^* + B^*$

$$\frac{\partial Y}{\partial z} = \frac{B^*}{E_p} - \frac{A^*}{E_m}$$

$$\frac{\partial Y}{\partial D_0} = \left(\frac{\partial Y}{\partial z}\right)\left(\frac{\partial z}{\partial D_0}\right) = -\left(\frac{z}{D_0}\right)\left(\frac{\partial Y}{\partial z}\right)$$

$$\frac{\partial Y}{\partial Q} = \left(\frac{\partial Y}{\partial z}\right)\left(\frac{\partial z}{\partial Q}\right) = \left(\frac{(X - X_0)^2 e^{Q(X - X_0)}}{2D_0}\right)\left(\frac{\partial Y}{\partial z}\right)$$

$$\frac{\partial Y}{\partial X_0} = \left(\frac{\partial Y}{\partial z}\right)\left(\frac{\partial z}{\partial X_0}\right) - \frac{A' + 2A''(X - X_0)}{E_p} - \frac{B' + 2B''(X - X_0)}{E_m} \qquad (4\text{-}2)$$

$$\textit{where} \quad \left(\frac{\partial z}{\partial X_0}\right) = -\frac{(1 + (1 + Q(X - X_0))e^{Q(X - X_0)}}{2D_0}$$

Of importance in the discussion that follows is the calculation of the fraction of completeness of the interface. The fraction *f* of completeness at $X = X_f$ where $Y = Y_f$ is given by:

$$f = \frac{Y_f - (A + A'(X_f - X_0))}{(B + B'(X_f - X_0)) - (A + A'(X_f - X_0))} \qquad (4\text{-}3)$$

$$= \frac{\dfrac{(A + A'(X_f - X_0))}{1 + e^{z_f}} + \dfrac{(B + B'(X_f - X_0))}{1 + e^{-z_f}} - (A + A'(X_f - X_0))}{(B + B'(X_f - X_0)) - (A + A'(X_f - X_0))}$$

which simplifies to

$$f = \frac{1}{1 + e^{-z_f}} \quad \textit{where} \quad z_f = \frac{X_f - X_0}{D} \quad \textit{and} \quad D = \frac{2D_0}{1 + e^{Q(X_f - X_0)}} \qquad (4\text{-}4)$$

so that $X_f = X_0 - \dfrac{2D_0}{1 + e^{Q(X_f - X_0)}} \ln\left(\dfrac{1 - f}{f}\right)$ and $X_{1-f} = X_0 + \dfrac{2D_0}{1 + e^{Q(X_{1-f} - X_0)}} \ln\left(\dfrac{1 - f}{f}\right)$ (4-5)

If the profile is symmetric, i.e., if Q=0, then

$$X_f = X_0 - D_0 \ln\left(\frac{1 - f}{f}\right) \quad \textit{and} \quad X_{1-f} = X_0 + D_0 \ln\left(\frac{1 - f}{f}\right) \qquad (4\text{-}6)$$

4.1 Initial Estimates of the Parameters

The rapidity of convergence, if the iterative process does indeed converge, depends on the quality of the initial estimates of the parameters. **When the program for whatever reason is unable to calculate reasonable initial estimates of the parameters, the user can define, by inspection of the graph of the data, the interface region and the program can then draw straight lines through the data in each of the three regions. The straight line through the data in the region identified as the interface is then interpreted as a tangent to the logistic function from which D_0 can be determined from the slope and X_0 can be**

determined as the value of X where the tangent is midway between the pre-interface and post-interface lines. *Q* in such cases is initially assumed to be 0.

While this works very well for all kinds of data, well behaved or not, it is desirable to find algorithms that can yield initial estimates automatically without requiring the user to define the interface. Several approaches have been evaluated and all work well for well structured data with small random errors. A trial and error approach that appears to work well with incomplete profiles, high levels of random noise, and very sharp profiles with few if any data in the interface region, is one of assigning values for the asymptotic parameters from the first and last data values (in the sense of increasing values of X), a width parameter D_0 equal to the average separation between X values, evaluating the root mean square (rms) deviation for various values of X_0, and selecting the value that gives the lowest rms deviation. This is accomplished by dividing the data range into 10, testing the midpoint of each section, selecting the section containing the value of X_0 with the lowest rms deviation, dividing that section into 10 sections and repeating the process and continuing until the separation between trial values of X_0 is equal to 0.1% of the range of X values. Following this, the starting estimate of D_0 is obtained by first setting D_0 equal to ¼ the range of X values, calculating the rms deviation, dividing the value of D_0 by 2 and continuing until the minimum value of D_0 is reached. D_0 is then determined by sampling the region around this minimum value of D_0.

Finally, D_0 is further refined by fitting the linear form, Equation (4-1), varying only δD_0. The initial value of Q is estimated by fitting the linear form varying only δQ.

This procedure for determining the initial estimates of the parameters must be modified if the data do not encompass the entire interface region as in Figure 4-1 below:

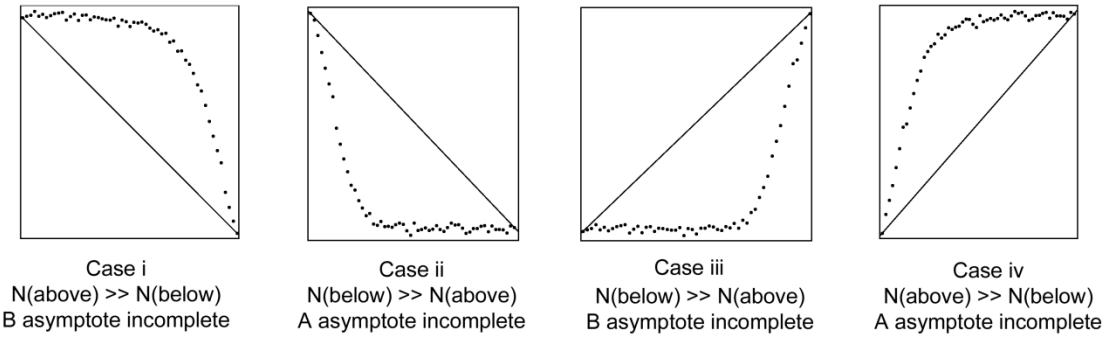


Figure 4-1 Initial estimates from an incomplete interface

A straight line is drawn connecting the first and last data values and the number of points falling above and below that line is calculated. If the number is greater than 80% (or less than 20%) of the data, the interface is considered to be incomplete. In this case the initial values of X_0 and D_0 are obtained by locating short line segments (five points) where the maximum slope is observed. The center of the segment is taken to be X_0 and the slope is taken to be $1/4D_0$.

If these methods fail to give reasonable initial estimates, then the user can identify the interface region graphically using a data selection box as described in Section 2.6.10.1 above.

4.2 Review of linear regression and confidence limits

Linear regressions, i.e., linear least squares fits, are the subjects of numerous textbooks on statistics and the interpretation of the quality of the fit, i.e., the measure of the agreement between Y_{obs} and Y_{calc} can be as complicated as desired. Most of the conclusions that can be drawn from an analysis of residuals (Y_{obs}-Y_{calc}) rely on the assumption that all the variability resides in the values of Y_{obs} and arises from a normally distributed population of errors. This is seldom the case. However, this is mostly of concern when attempting to determine a true value for some quantity derived from the data, a true value that can be compared with a fundamental calculation such as an average atomic separation in a crystal. In the case of the measurement of interfaces, there is no such true value for the width, the center and the asymmetry. The best that can be said about the estimates returned by the least squares analysis is that a repetition of the same measurements on the same material will return the same values within the stated uncertainties. This will now be discussed in some detail. In so doing, it will be necessary for establishing a frame of reference to review briefly the least squares fit calculation.

For simplicity we rewrite (4-1) as

$$y_i = \sum_{j=1}^{m} c_j x_{ji} \quad where \quad y_i = Y_{obs}(X_i) - Y_{calc}(X_i, \{C^{|k}\}),$$
$$x_{ji} = \frac{\partial Y_{calc}(X_i, \{C^{|k}\})}{\partial C_j}, \quad and \quad c_j = \delta_j^{|k+1} \tag{4-7}$$

for the i = 1 to n (> m) measured values of Y_i. In a least squares fit, we are seeking those values of the parameters $\{c_j\}$ for which sum of the squares $\sum_{i=1}^{n} W_i \left(y_i - \sum_{k=1}^{m} c_k x_{ki} \right)^2$ is a minimum, that is, those values of $\{c_j\}$ which satisfy $\frac{\partial}{\partial c_j} \left(\sum_{i=1}^{n} W_i \left(y_i - \sum_{j=1}^{m} c_j x_{ji} \right)^2 \right) = 0$, or,

$$\sum_{i=1}^{n} y_i W_i x_{ji} = \sum_{k=1}^{m} c_k \sum_{i=1}^{n} W_i x_{ji} x_{ki} \quad for \quad j = 1 \, to \, m \tag{4-8}$$

If we adopt a matrix notation, $\mathbf{y} = (y_1\ y_2\ \ldots\ y_n)$, $\mathbf{c} = (c_1\ c_2\ \ldots\ c_m)$,

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{m1} & x_{m2} & \cdot & \cdot & \cdot & x_{mn} \end{pmatrix} \quad where \quad x_{ji} = \frac{\partial Y_{calc}(X_i, \{C^{|k}\})}{\partial C_j^{|k}} \tag{4-9}$$

and **W** is an n dimensional, diagonal matrix (all measurement errors are uncorrelated) whose elements are the weights W_i, then the least squares equations (4-8) become

$$\mathbf{yWx^T = cxWx^T} \tag{4-10}$$

and solving for **c**,

$$\mathbf{c} = \left(\mathbf{xWx^T}\right)^{-1}\left(\mathbf{yWx^T}\right) \tag{4-11}$$

The matrix **x** is often referred to as the design matrix. The quality of the fit is determined by the residual standard deviation of the fit,

$$s = \sqrt{\frac{\sum_{i=1}^{n} W_i\left(y_i - \sum_{j=1}^{m} c_j x_{ji}\right)^2}{n-m}} = \sqrt{\frac{\sum_{i=1}^{n} W_i (Y_i^{obs} - Y_i^{calc})^2}{n-m}}, \tag{4-12}$$

where the second equality holds at convergence of the iterative, linear process. The standard deviations of the values of the parameters are obtained through the usual propagation of errors formula:

$$s^2(c_k) = \sum_{i=1}^{n}\left(\frac{\partial c_k}{\partial y_i}\right)^2 s_i^2 \tag{4-13}$$

where s_i is the standard deviation of the i^{th} measurement. Note that this assumes that the individual measurement errors are uncorrelated. If not, and if that correlation is known, then Eq. (4-13) can be suitably modified to carry along this correlation. Generally the s_i are not known and s_i is set equal to the standard deviation of the fit, s, as determined by Eq. (4-12). Substituting Eq. (4-11) into Eq. (4-13) and performing some algebra, the standard deviation of the k^{th} parameter is easily seen to be

$$s(c_k) = \sqrt{\mathbf{V}_{kk}^{cv}} \tag{4-14}$$

where $\mathbf{V}^{cv}$, known as the variance-covariance matrix, is given by

$$\mathbf{V}^{cv} = \left(\mathbf{xWx^T}\right)^{-1} s^2 \tag{4-15}$$

$\mathbf{V}^{cv}$ carries not only the errors in the determined parameters c_k but also the correlation of errors among the parameters c_k so that the variance of any function of the parameters, $f\{C\}$, can be obtained from

$$s^2(f\{C\}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial f}{\partial C_i}\frac{\partial f}{\partial C_j}\mathbf{V}_{ij}^{cv} \tag{4-16}$$

4.2.1 Variance and the Chi-Square distribution.

The confidence levels for reporting uncertainties are strictly valid only if the errors in Y_i are normally distributed. Even if this condition is not met, they can provide a guide for determining whether a second measurement of an interfacial profile is different from the first.

The number of degrees of freedom in a least squares analysis, often designated as ν, is equal to the number of measurements included in the fit minus the number of parameters varied, ν=n-m. For ν degrees of freedom, the standard deviation of a set of measurements, s, taken from a normal population having a standard deviation of σ will follow a chi square distribution:

$$\int_0^{\chi^2_{\nu,\alpha}} \varphi(\chi^2)d\chi^2 = \int_0^{\chi^2_{\nu,\alpha}} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} e^{-\chi^2/2}(\chi^2)^{\nu/2-1}d\chi^2 = 1-\alpha \tag{4-17}$$

where $\chi^2 \geq 0$ and α is the probability

$$\alpha = P\left(\frac{\nu s^2}{\chi^2_{\nu,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\nu s^2}{\chi^2_{\nu,\frac{\alpha}{2}}} \right) \tag{4-18}$$

That is, the variance of the population will have a probability α of falling between $\frac{\nu s^2}{\chi^2_{\nu,1-\frac{\alpha}{2}}}$ and $\frac{\nu s^2}{\chi^2_{\nu,\frac{\alpha}{2}}}$.

The values of χ^2 can be found in tables or calculated using readily available algorithms. In determining whether model errors may be dominating the least squares fit of the extended logistic function to a measured profile, Equation (4-18) can be quite helpful if we have some independent estimate of the standard deviation of the measurement population.

4.2.2 Third Differences as an estimate of the variance

A model independent estimate of the variance which has proved useful is that obtained from so-called "third differences" of the observed data. Given a set of measurements Y_i, the first differences are defined as $Y_i^{(1)} = Y_{i+1} - Y_i$, second differences as
$Y_i^{(2)} = Y_{i+1}^{(1)} - Y_i^{(1)} = Y_{i+2} - 2Y_{i+1} + Y_i$, and third differences as
$Y_i^{(3)} = Y_{i+1}^{(2)} - Y_i^{(2)} = Y_{i+3} - 3Y_{i+2} + 3Y_{i+1} - Y_i$.

If $Y_i = \hat{Y}_i + \varepsilon_i$ where $\hat{Y}_i$ is the arbitrarily accurate value of Y_i from the logistic function and ε_i is the random error in Y_i and if the variation in $\hat{Y}_i$ from one value to the next is negligible compared to the values of ε_i, so that $\hat{Y}_{i+3} - 3\hat{Y}_{i+2} + 3\hat{Y}_{i+1} - \hat{Y}_i \approx 0$ and
$Y_i^{(3)} \approx \varepsilon_i^{(3)} = \varepsilon_{i+3} - 3\varepsilon_{i+2} + 3\varepsilon_{i+1} - \varepsilon_i$ then
$\sum_{i=1}^{n-3}\left(Y_i^{(3)}\right)^2 \approx \varepsilon_1^2 + 10\varepsilon_2^2 + 19\varepsilon_3^2 + \left[\sum_{i=4}^{n-3}\varepsilon_i^2\right] + 19\varepsilon_{n-2}^2 + 10\varepsilon_{n-1}^2 + \varepsilon_n^2$ where all the cross terms $\varepsilon_i\varepsilon_j$ for $i \neq j$ have been ignored because they will average will tend to vanish. Substituting the average value $\langle \varepsilon^2 \rangle$ for each value of ε_i^2 gives

$$\sum_{i=1}^{n-3}\left(Y_i^{(3)}\right)^2 \approx 20(n-4)\langle \varepsilon^2 \rangle \tag{4-19}$$

The third differences, which magnify the point to point random variations but minimize the systematic variation in Y, can provide a model-independent measure of the standard deviation of the measurements. (Indeed, if Y were a linear or quadratic function of X and the

values of X were evenly spaced, the contribution from the systematic variation in Y would vanish identically.) If s^2 is the variance of the values of Y attributable to measurement error, then the variance in $Y_j^{(3)}$ is

$$Var(Y_i^{(3)}) = \sum_{j=i}^{j=i+3}\left(\frac{\partial Y_i^{(3)}}{\partial Y_j}\right)^2 s_j^2 = s^2\sum_{j=i}^{j=i+3}\left(\frac{\partial Y_i^{(3)}}{\partial Y_j}\right)^2 = (1+9+9+1)s^2 = 20s^2 = 20\left\langle \varepsilon^2 \right\rangle \quad (4\text{-}20)$$

so that $\sum_{i=1}^{n-3} Var(Y_i^{(3)}) = s_1^2 + 10s_2^2 + 19s_3^2 + \sum_{i=4}^{n-3} 20s_i^2 + 19s_{n-2}^2 + 10s_{n-1}^2 + s_n^2$. (4-21)

Assuming all s_i have equal variance *s*, substituting *s* for s_i in Eq. (4-21) yields

$$\sum_{i=1}^{n-3} Var\left(Y_i^{(3)}\right) = 20(n-4)s^2 = 20(n-3)\left\langle \varepsilon \right\rangle^2 \quad (4\text{-}22)$$

Comparison of Eq. (4-19) with Eq.(4-22) gives us the estimate of the standard deviation from the data scatter:

$$s^2 \approx Var\left(Y^{(3)}\right) = \varepsilon_{3d}^2 = \frac{1}{20(n-3)}\sum_{i=1}^{n-3}\left(Y_i^{(3)}\right)^2 \quad (4\text{-}23)$$

The assumption that the systematic contribution to Y_i^3 can be ignored can be checked once we have the least squares fit of the logistic function to the data by calculating $\sum_{i=1}^{n-3}\left(\hat{Y}_i^{(3)}\right)^2$ and comparing it with the value of s^2 calculated from Eq. (4-23).

In the presence of weights, where the weights are proportional to the inverse square of the measurement uncertainty, $s_i^2 = \frac{s^2}{W_i}$, and $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left\langle \varepsilon_i^2 \right\rangle = \frac{1}{n}\sum_{i=1}^{n} W_i s_i^2 = \frac{1}{n}\sum_{i=1}^{n} W_i \frac{s^2}{W_i} = s^2$ and

$s^2 = \frac{1}{n-m}\sum_{i=1}^{n} W_i\left(Y_i - \hat{Y}_i\right)^2$ where m is the number of parameters varied. If we then define a weighted third difference as

$$\bar{Y}_i^{(3d)} = Y_{i+3}\sqrt{W_{i+3}} - 3Y_{i+2}\sqrt{W_{i+2}} + 3Y_{i+1}\sqrt{W_{i+1}} - Y_i\sqrt{W_i} \quad (4\text{-}24)$$

and assume the systematic contribution to the weighted third difference $\bar{Y}_i^{(3)}$ is small, i.e.,

$$\hat{Y}_{i+3}\sqrt{W_{i+3}} - 3\hat{Y}_{i+2}\sqrt{W_{i+2}} + 3\hat{Y}_{i+1}\sqrt{W_{i+1}} - \hat{Y}_i\sqrt{W_i} \ll \varepsilon_{i+3}\sqrt{W_{i+3}} - 3\varepsilon_{i+2}\sqrt{W_{i+2}} + 3\varepsilon_{i+1}\sqrt{W_{i+1}} - \varepsilon_i\sqrt{W_i}$$

then we can let

$$\left(\bar{Y}_i^{(3)}\right)^2 \approx W_{i+3}\varepsilon_{i+3}^2 + 9W_{i+2}\varepsilon_{i+2}^2 + 9W_{i+1}\varepsilon_{i+1}^2 + W_i\varepsilon_i^2 \quad (4\text{-}25)$$

where we have ignored cross terms $\left(\sqrt{W_i W_j}\right)\varepsilon_i\varepsilon_j$ for $i \neq j$ because they will tend to vanish when summed over all the data. Then summing over all n-3 values of $\bar{Y}_i^{(3)}$ gives

$$\sum_{i=1}^{n-3}\left(\bar{Y}_i^{(3)}\right)^2 \approx W_1\varepsilon_1^2 + 10W_2\varepsilon_2^2 + 19W_3\varepsilon_3^2 + \sum_{i=4}^{n-3} 20W_i\varepsilon_i^2 + 19W_{n-2}\varepsilon_{n-2}^2 + 10W_{n-1}\varepsilon_{n-1}^2 + W_n\varepsilon_n^2$$

$$\approx 20(n-3)s^2 \qquad (4\text{-}26)$$

$$\text{or } s_1^2 \approx \varepsilon_{3d}^2 = \frac{1}{20(n-3)}\sum_{i=1}^{n-3}\left(\bar{Y}_i^{(3)}\right)^2 \qquad (4\text{-}27)$$

where we have defined ε_{3d} as the estimate of s from third differences. By ignoring the cross terms in Eq. (4-27), we cannot expect ε_{3d}^2 from either weighted or unweighted third differences to be distributed as a χ^2 distribution as in Equation(4-18).

As already mentioned, a problem with the use of third differences to estimate the random scatter in the data is the influence of systematic change from the underlying sigmoidal function that we approximate with the logistic function. For data following a logistic function, the scatter might look like that shown in Figure 4-2. The large excursions represent the functional change of the logistic function in the middle of the profile. In order to remove these contributions and arrive at a better measure of the random noise in the data, the values surrounding the third difference with the largest magnitude are subtracted one by one from a calculation of the root mean square value of all the third differences. If the third difference of a particular value of $Y_i^{(3)}$ is more than the confidence level for that value, (assuming it follows approximately a student's *t* distribution and using the remaining third differences to estimate the variance) that value is discarded from the calculation and the process is repeated until no more systematic differences are discovered. When the data scatter is then displayed, (see section 2.6.3) the discarded third differences are displayed with red x's as in **Error! Reference source not ound.**

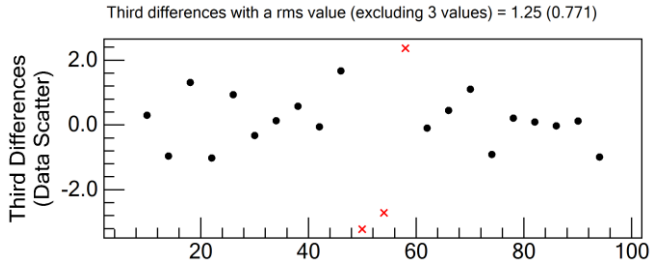


Figure 4-2 Data Scatter from Third Differences showing values excluded from the calculation of the estimated standard deviation.

As noted earlier, the systematic contribution to ε_{3d}^2 can be estimated by calculating $\hat{Y}_i^{(3)} = \hat{Y}_{i+3}\sqrt{W_{i+3}} - 3\hat{Y}_{i+2}\sqrt{W_{i+2}} + 3\hat{Y}_{i+1}\sqrt{W_{i+1}} - \hat{Y}_i\sqrt{W_i}$ once a logistic function has been fit to the data. Prior to having such a logistic function, the systematic contributions are identified statistically (assuming sufficient data). But after we have the values of $\hat{Y}_i^{(3)}$ we can subtract the $\hat{Y}_i^{(3)}$ from the $Y_i^{(3)}$ to give values of ε_{3d}^2 that are more free of systematic error than those values for which the data statistically demonstrating systematic contributions have been subtracted.

$$s_2^2 \approx \varepsilon_{3d}^2 = \frac{1}{20(n-3)}\sum_{i=1}^{n-3}\left(\bar{Y}_i^{(3)} - \hat{Y}_i^{(3)}\right)^2 \qquad (4\text{-}28)$$

The resulting scatter, based on the same data as Figure 4-2, is shown in Figure 4-3 where the logistic function itself is drawn

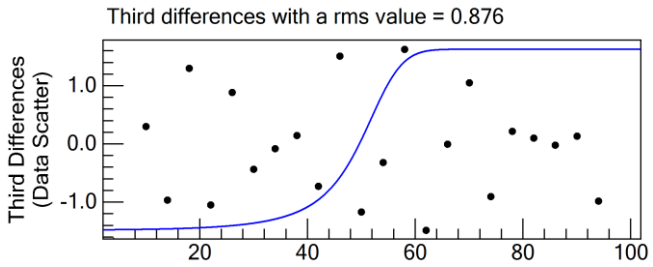


Figure 4-3 Scatter with systematic contribution from the extended logistic function removed

A comparison between the two can be seen, as was done here, by displaying the scatter before and after a least squares fit of the data have been performed. At first, it might be thought that making this subtraction merely reproduces the residuals, but it actually it reproduces the third differences of the residuals themselves in which any unresolved systematic trends in the residuals have been minimized. The scatter as estimated by ε_{3d}^2 from Eq. (4-27) and/or Eq. (4-28) has proven to provide a convenient measure against which to compare s^2 to signal the possible presence of systematic error.

In the analysis of 1000 sets of synthetic data generated from the logistic functions with added random, normal errors and consisting of 100 data in each set (as summarized in Table 3-1), the ratio of scatter was, on average, 0.997 (± 0.164 at the 95% confidence level) and ranged from 0.6921 to 1.2442. (See Table 3-4 and its accompanying discussion beginning on page 3-1)

If the value of s^2 from the least squares fit of the extended logistic function is significantly larger than ε_{3d}^2 for a particular class of data, then the extended logistic function may not be a good model of the measured data.

4.2.3 F tests for the comparison of variance

A common statistical test for comparing sample variances is derived from the ratio of the squares of two sample standard deviations. If two independent samples of data have ν_1 and ν_2 degrees of freedom and standard deviations of s_1 and s_2, then the ratio of the squares of s_1 and s_2 should follow the so-called F distribution:

$$1-\alpha = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\int_0^{F_{\nu_1,\nu_2,\alpha}} F^{\left(\frac{\nu_1}{2}-1\right)}\left(1+\frac{\nu_1}{\nu_2}F\right)^{-\left(\frac{\nu_1+\nu_2}{2}\right)} dF \qquad (4\text{-}29)$$

If the ratio, F, of the variances of two samples with degrees of freedom ν_1 and ν_2 has a value greater than $F_{\nu_1,\nu_2,1-\alpha}$, i.e., if

$$F = \frac{s_1^2}{s_2^2} > F_{\nu_1,\nu_2,1-\alpha} \qquad (4\text{-}30)$$

then the probability that sample 1 arises from a population with a greater standard deviation than sample 2 will be 1-α.. Taking 1-α = 0.95, we calculate s^2 / ε_{3d}^2 and compare it to $F_{\nu_1,\nu_2,0.05}$ and if it is greater there will be a greater than 95% probability that we have model errors. The application of the F test here is not strictly appropriate since the test is based on the independence of the sample standard deviations s^2 and ε_{3d}^2 (See Table 3-1 to Table 3-3.) However it can serve as a convenient suggestion of systematic, i.e., model errors

Another comparison is the variance for different regions of the data sensitive to different parameters of the logistic function. Consider the separation of the data into three regions, the statistically significant interface and the pre- and post-interface regions. The region prior to the statistically significant interface is dependent almost solely on the parameters A, A′, and A″. Similarly, the region following the statistically significant interface is dependent almost solely on the parameters B, B′, and B″. While the statistically significant interface depends on all the parameters, it is most sensitive to X_0, D_0, and Q. Since the asymptotic regions are virtually model independent, the variance of those regions will not be sensitive to model errors whereas the statistically significant interface will be. The variances of the three regions are calculated from:

$$s_A^2 = \sum_{i=1}^{n_A} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_A - p_A}, \quad s_I^2 = \sum_{i=1}^{n_I} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_I - p_I}, \quad s_B^2 = \sum_{i=1}^{n_B} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n_B - p_B} \qquad (4\text{-}31)$$

n_A, n_I, and n_B are the numbers of data in the pre-interface asymptotic region, the interface region, and the post-interface asymptotic region respectively and p_A, p_I, and p_B are the number of varied parameters on which each of the regions is dependent so that three regions have ν_A, ν_I, and ν_B degrees of freedom where $\nu_A = n_A - p_A$, etc. Typically p_A and p_B will each be 1 and p_I will be 2 or 3 depending on whether Q is varied. If $s_I^2/s_A^2 > F(\nu_I,\nu_A,\alpha)$ or $s_I^2/s_B^2 > F(\nu_I,\nu_B,\alpha)$ where α is the confidence level for the F distribution, we may have reason to suspect model errors.

4.2.4 Parameter Confidence Limits

The parameters obtained from the least squares fit, again assuming a normal distribution of measurement errors, will follow the so-called student's t distribution.

$$\phi(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad (4\text{-}32)$$

so that the probability that the value of a parameter C_k determined from the least squares fit lies in the range

$$P\left(C_k - t_{\nu,\frac{\alpha}{2}}\sqrt{\mathbf{V}_{kk}^{cv}} < C_k < C_k + t_{\nu,\frac{\alpha}{2}}\sqrt{\mathbf{V}_{kk}^{cv}}\right) = 1-\alpha \qquad (4\text{-}33)$$

where the quantities $\pm t_{\nu,\alpha}$ are the limits of the integral of the student's t distribution that satisfy

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu}}\int_{-t}^{t}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}dx=1-\alpha \tag{4-34}$$

for $0 < \alpha < 0.5$. (Note some references use single tailed integrals, i.e. from $-\infty$ to $t_{\nu,\alpha/2}$ and from $-t_{\nu,\alpha/2}$ to ∞ so that $\alpha/2$ values must be used for a probability of $1-\alpha$.) The limits of Equation (4-33) are called the $100(1-\alpha)\%$ confidence limits.

It also follows from Equation (4-33) that for any arbitrary function of the parameters, $f\{C\}$,

$$P\left(\left(f\{C\}-t_{\nu,\frac{\alpha}{2}}s(f\{C\})\right) < f\{C\} < \left(f\{C\}+t_{\nu,\frac{\alpha}{2}}s(f\{C\})\right)\right)=1-\alpha \tag{4-35}$$

where $s(f\{C\})$ is obtained from Equation (4-16).

One particular function $f\{C\}$ is important and that is the calculated value of Y itself, namely Y_i^{calc}, corresponding to the i[th] observation, and the difference between the calculated and observed value of Y, namely, $Y_i^{obs}-Y_i^{calc}$. The variance of Y_i^{calc} is given by Equation (4-16) as

$$s^2(Y_i^{calc})=\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{\partial Y_i^{calc}}{\partial C_j}\frac{\partial Y_i^{calc}}{\partial C_k}\mathbf{V}_{jk}^{cv} \tag{4-36}$$

where the derivatives, $\frac{\partial Y_i^{calc}}{\partial C_k}$, are those listed collectively as Equations (4-2). The variance of $Y_i^{obs}-Y_i^{calc}$ is given by

$$s^2(Y_i^{obs}-Y_i^{calc})=s^2+s^2(Y_i^{calc}) \tag{4-37}$$

if Y_i^{obs} was ***not*** included in the least squares fit and by

$$s^2(Y_i^{obs}-Y_i^{calc})=s^2-s^2(Y_i^{calc}) \tag{4-38}$$

if Y_i^{obs} ***was*** included in the fit.

The minus sign in (4-38) arises because the variance of Y_i^{calc} and the variance of Y_i^{obs} are correlated since Y_i^{obs} was used, through the least squares fit, to calculate Y_i^{calc}. This can be shown using the usual propagation of error formulas and a few pages of algebra.

That this is significant can be seen from an analysis of 25 synthetic data with a standard deviation of unity, a value of $|B-A| = 100$ and a value of D_0 such that on average between 2 and 4 values fall in the statistically significant interface region (3 values in the 16% to 84% region.) One hundred thousand data sets with different random errors drawn from a normal population with $\sigma = 1$ were analyzed and the root mean square values of $(Y_{obs}-Y_{calc})^2$ were calculated for each value of X for all 100,000 data sets. The results are summarized in Table

4-1 below. The values of the root mean squares of $(Y_{obs} - Y_{calc})^2$ are also displayed in Figure 4-4 below to the right of the table.

From the definition of the standard deviation the sum of the squares, $(Y_{obs} - Y_{calc})^2$, is equal to $(n-m)s^2$. The large dip in the graph of the root mean square deviations represents the fewer effective degrees of freedom associated with the interface region where only the few data in the interface region are sensitive primarily to the 2 interface parameters (Q was held fixed at 0) whereas each asymptotic region is sensitive to only one asymptotic parameter each. The statistically significant interface region for this example included the 4 data from X=45 to X=57. The sum $(Y_{obs} - Y_{calc})^2/s^2$ for the first 11 data comprising the pre-interface region is 9.86, approximating 10 degrees of freedom for 11 data and one adjustable parameter. Similarly the sum $(Y_{obs} - Y_{calc})^2 /s^2$ for the last 10 data comprising the post-interface region is 8.98, approximating 9 degrees of freedom for 10 data and one adjustable pararameter. The sum $(Y_{obs} - Y_{calc})^2/s^2$ for the 4 data in the interface region is 2.16, approximating 2 degrees of freedom for 4 data and 2 adjustable parameters.

In the column labeled "Adjusted" in Table 4-1, the influence of the uncertainty in Y_{calc} has been taken into account and the column labeled "True" is the root mean square of the 100,000 deviations added to each point.

Table 4-1 Distribution of Errors

X	rms(Yo-Yc)	Adjusted	TRUE
1	0.949	1.000	1.001
5	0.946	0.997	0.997
9	0.948	0.999	0.999
13	0.950	1.002	1.002
17	0.949	1.001	1.000
21	0.949	1.000	1.000
25	0.951	1.002	1.002
29	0.948	0.997	0.997
33	0.953	1.000	0.999
37	0.955	1.000	1.000
41	0.920	1.003	1.002
45	0.735	0.999	1.000
49	0.686	0.996	0.996
53	0.660	0.998	0.998
57	0.844	0.998	0.997
61	0.944	0.999	1.001
65	0.957	1.004	1.005
69	0.953	1.004	1.003
73	0.948	1.001	1.000
77	0.948	1.002	1.002
81	0.946	1.000	0.999
85	0.943	0.997	0.997
89	0.947	1.001	1.001
93	0.947	1.001	1.000
97	0.945	0.999	0.999

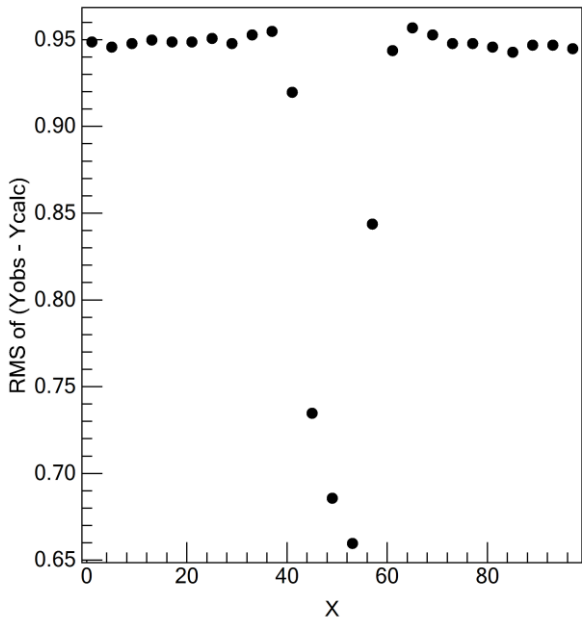


Figure 4-4 Graphical representation of the first two columns of Table 4-1

4.2.5 Skewness and Kurtosis

Two other measures of the distribution of the residuals are the skewness and kurtosis. If the standardized residuals $r_i = \sqrt{W_i}\left(Y_i(obs) - Y_i(calc)\right)$ and the moments of the standardized

residuals are defined as $\mu_k = \left(\sum_{i=1}^{N} r_i^k \right) / N$ then the skewness and kurtosis of the distribution are given by

$$skewness = \mu_3 / \mu_2^{3/2} \text{ and } kurtosis = \mu_4 / \mu_2^2 - 3 \quad (4\text{-}39)$$

For a normal distribution the skewness and kurtosis are both vanishing, but convergence to zero is very slow as N increases so that they have only limited usefulness for data sets with less than, say, 100 data.

4.3 Algorithm for the Linear Least Squares Fit

Various programs exist for conducting linear regressions and the one used here is a program called ORTHO, originally written in Algol by Walsh (P. J. Walsh, *Commun. Assoc. Comput. Mach*. **5**, 511(1962)), which is based on a Gramm-Schmidt orthonormalization of the design matrix **x** , following which the solution of the least squares equations becomes trivially simple. In this procedure, the inversion of the design matrix implicit in Equation (4-11) is avoided. Using double precision arithmetic, and re-orthogonalization after normalization, the program has been found to be simple and robust.

4.4 Poorly Structured Data

Poorly structured data are those for which the least squares fit becomes unstable because some parameter or linear combination of parameters cannot be determined from the data. One such example has already been discussed in the section on initial estimates, namely where the interface is not complete. In general, the interface should reach within 5% of completion at both ends of the interface to obtain reliable confidence limits for the values of the width and asymmetry parameters D_0 and Q. For any value of X, the fractional completion of the transition (strictly speaking never exactly 0 nor 1) is calculated from Equation (4-3) or Equation (4-4).

The extended logistic function is a continuous function all of whose derivatives exist which makes calculation easy within a Taylor's Series linearization. But it does have a singularity, namely where $D_0 \to 0$. When this occurs, the number of measurements falling in the interface region approaches 0 and it is not possible to determine D_0 though an upper limit may be placed on its value based on the standard deviation of the data as determined by the least squares fit and the separation between neighboring data.

The idea has been presented earlier in this documentation of a statistically significant interface region where a measurement is considered to be in the interface region when it falls between the two asymptotes and it's deviation from each of the asymptotes, *A* and *B*, is statistically significant. If s^2 is the variance of the measured data, calculated from $s^2 = \sum_{i=1}^{n} \frac{W_i (Y_i^{obs} - Y_i^{calc})^2}{n-m}$, and Y^{obs} - Y^{calc} follows a normal distribution so that s^2 follows a χ^2 distribution, the probability that the true variance is less than $s_\alpha^2 = \frac{(n-m)s^2}{\chi^2_{n-m,\alpha}}$ is α where

$0<\alpha<1$. If we have a normal distribution of errors $\phi(x)$ so that $N_\alpha = \int_{-\infty}^{\alpha} \phi(x)$ then the probability that a single measurement Y_i, drawn from a population with a standard deviation of *s* will differ from its true value by more than $N_\alpha s$ will be 1- α. Therefore a measurement Y_i will be considered to be statistically different from A or B when

$$|A-Y_i| > N_\alpha s, \quad |B-Y_i| > N_\alpha s \quad and \quad A<Y_i<B \quad or \quad B<Y_i<A \tag{4-40}$$

Before performing each iteration in the analysis, the number of data falling in the interface region with values of Y satisfying Equation (4-40) is counted. If one or none falls in the statistically significant interval and if the least squares fit appears to be diverging, then the value of *Q* is held fixed at 0, the upper limit of D_0 is estimated from the separation between the points bordering the interval, and the value of X_0 is the average of the two values of X bordering the interval. Exactly which parameter is held fixed at what value depends on the manner in which the iterative process is diverging. This was discussed earlier in the beginning of Section 3.2 Difficult Data and Analysis Instabilities. By varying the value of the confidence limit α, the test for significance can be made more or less stringent.

If one or no point falls in the interval and X_b,Y_b is the measurement just before the interface region and X_a,Y_a is the measurement just after the interface region, then the interface width $W_{ts} < X_a - X_b$ and, from Equation (4-5),

$$D_0 < \frac{1}{2} \frac{X_a - X_b}{\ln\left(\frac{1-f}{f}\right)} \tag{4-41}$$

Where $f = N_\alpha s / |A-B|$. In the course of the subsequent iterations, D_0 is held at a value between one half of this upper limit and its upper limit. If the last stable value of D_0 is between these two limits, D_0 will retain that value in subsequent iterations.

As a check on the reasonableness of setting values of *Q* and D_0, a least squares fit is performed varying only *Q* or D_0, holding the remaining parameters fixed at their current values. The thus determined values of *Q* and D_0 should not change significantly from the values assigned to them in the full analysis. That their values sometimes do not change at all is a result of the effect that the derivatives of Y with respect to *Q* or to D_0 nearly vanish for all but one or two values, as can be seen by displaying their derivatives on a graph of the data.

4.5 Calculation of the interface width and asymmetry

The width, W_f, of an interface is taken to be the range in X in which Y varies from a fraction f of completion to a fraction $(1-f)$ of completion, where completion is represented by the second asymptote, *B*. W_f is calculated from *Q* and D_0 using Equation(4-5). If *Q* is non-zero, W_f must be calculated by successive approximations since X_f appears on both sides of Equation (4-5). Using a Newton-Raphson approach and taking X_f^i as the value of following the i^{th} iteration,

$$X_f^{i+1} = X_f^i - \frac{2\ln D_0\left(\frac{1-f}{f}\right) + (X_f^i - X_0)(1+e_f^i)}{1+e_f^i + Q(X_f^i - X_0)e_f^i}$$

where $e_f^i = e^{Q(X_f^i - X_0)}$ (4-42)

The initial value for $X_f^i = X_f^0$ is given by

$$X_f^0 = X_0 - \frac{2\ln\left(\frac{1-f}{f}\right)D_0}{1+e^{-QD_0\ln\left(\frac{1-f}{f}\right)}}$$

Typically the procedure converges in less than five iterations. However, this approach does not converge and even diverges if $|QD_0| >> 1$. The LFPF program therefore takes the safe and sure route of successive range bisecting to find the value of X_f. For $f < 0.5$, X_f will lie below X_0. The midpoint between X_0 and the lowest value of X, X_{test}, is tested. If it yields a value of f less than target value, then the desired value of X_f lies between X_{test} and X_0 and a new test point X_{test} between the two is tested. The region containing X_f is again bisected and tested and the procedure continues until the desired precision is achieved. Taking the data range times 10^{-8} gives more precision than necessary and takes 27 successive bisections. The value of X_{1-f} is found in the same way.

When Q=0, the calculation of the width reduces to Equation (4-6). By convention, the values selected for *f* and 1-*f* are 16% and 84%.. The reason for this choice is that the 16% and 84% completion points for an error function correspond to the x = -σ to x = +σ width of the normal distribution function, the integrand of the error function which was first used for characterizing depth profiles. The width at half height of the derivative dY/dX of a symmetric logistic function is ± 1.762D_0 corresponding to the 14.64% and 85.36% completion points of the logistic function whereas the width at half height of the Normal Distribution function would be 1.1762σ corresponding to the 12% and 88% completion points of the error function. The logistic function has slightly longer tails than the error function.

While the *parameter Q* describes the asymmetry of the extended logistic function profile, the asymmetry η (as distinguished from the asymmetry *parameter Q* of the extended logistic function) of the interface is in practice described by the skewing of X_f and X_{1-f} about X_0.

$$\eta = \frac{2X_0 - (X_{1-f} + X_f)}{X_{1-f} - X_f} \text{ for } 0 < f \le 0.5 \quad (4\text{-}43)$$

As defined, *Q* and η will have the same sign. When both are negative the interface is sharper before the midpoint than after. When both are positive, the interface is sharper after the midpoint than before. The dimensionless quantity, QD_0 is similar in magnitude to η but η has the advantage of being defined independently of the function being used to fit the data.

The confidence limits of the width of the interface, $W_{f,1-f} = X_{1-f} - X_f$ (f<0.5), and the asymmetry parameter η are calculated from Equations (4-35), (4-16), (4-5), and (4-43). In

calculating the derivative of, for example, X_f with respect to D_0 or Q, one must keep in mind the appearance of X_f in the exponential in the denominator of equation (4-5) so that

$$\frac{\partial X_f}{\partial D_0} = \frac{\partial}{\partial D_0}\left(X_0 - \frac{2D_0F}{1+e^{Q(X_f - X_0)}} \right) = -\frac{2F}{1+e^{Q(X_f - X_0)}} + \frac{2D_0FQe^{Q(X_f - X_0)}}{(1+e^{Q(X_f - X_0)})^2}\frac{\partial X_f}{\partial D_0}$$

where $F = \ln\left(\frac{1-f}{f}\right)$

Rearranging gives:

$$\frac{\partial X_f}{\partial D_0} = \left(\frac{X_f - X_0}{D_0}\right)\left(1 + \frac{Q(X_f - X_0)e^{Q(X_f - X_0)}}{1+e^{Q(X_f - X_0)}}\right)^{-1} \tag{4-44}$$

Similarly,

$$\frac{\partial X_f}{\partial Q} = \frac{2D_0Fe^{Q(X_f - X_0)}}{(1+e^{Q(X_f - X_0)})^2}\left(X_f - X_0 + Q\frac{\partial X_f}{\partial Q}\right) \tag{4-45}$$

which, on rearrangement, results in

$$\frac{\partial X_f}{\partial Q} = -\frac{(X_f - X_0)^2}{e^{-Q(X_f - X_0)} + 1 + Q(X_f - X_0)} \tag{4-46}$$

Finally

$$\frac{\partial \eta}{\partial D_0, Q} = \frac{2(X_f - X_0)\left(\frac{\partial X_{1-f}}{\partial D_0, Q}\right) - 2(X_{1-f} - X_0)\left(\frac{\partial X_f}{\partial D_0, Q}\right)}{(X_{1-f} - X_f)^2} \tag{4-47}$$

Where $\frac{\partial \eta}{\partial D_0, Q}$ represents either $\frac{\partial \eta}{\partial D_0}$ or $\frac{\partial \eta}{\partial Q}$, etc.

This completes the description of how interface data can be analyzed by a linearized, least squares fit to an extended logistic function. This approach is the basis for the computer program described in this manual.

One other measure of the width mentioned in the discussion of surface line scans is the width at half height of the derivative, dY/dX, of the measured profile. The derivative, which is taken to be the point spread function, can be evaluated once the logistic function defining the line spread function has been determined. While the derivative can be evaluated analytically, the location of the maximum of dY/dX as well as the half-height points of dY/dX must be determined numerically. This is done with a simple brute force method of locating the upper and lower bounds of each value and then narrowing the range until the desired level of precision is obtained. For a symmetric profile with $Q = 0$, the maximum of dY/dX occurs at $X = X_0$ and the half-height values occur at $X = X_0 + \ln\left(3 \pm \sqrt{8}\right)D_0 = X_0 \pm 1.7627D_0$

which correspond to the $100/\left(4+\sqrt{8}\right)=14.64\%$ and $100/\left(4-\sqrt{8}\right)=85.36\%$ points on the extended logistic function profile. For this reason, 14.64% to 85.36% is taken as the default width in the LFPF program though this can be set to any value such as the commonly used ranges of 12% to 88% (width at half height of a Gaussian function) 16% to 84%, (1σ width for a Gaussian function) or 25% to 75% (width at half height of a Lorentzian function).

Acknowledgements

I wish to acknowledge the contributions of Cedric Powell and David Simons for their encouragement to start this project when I entered retirement and for their continuing assistance and discussion. It has been a pleasure to return to programming and a little algebra. I would also like to dedicate this program to the memory of Joseph Fine, a long time friend who involved me many years ago in applying what little I had learned about statistics to the analysis of sigmoidal profiles in depth profile analyses. I am sorry that he is no longer around to see this come to fruition after 25 years while I was away doing other things.