Chapter 3

F. C. Hawthorne and G. A. Waychunas

SPECTRUM-FITTING METHODS

INTRODUCTION

When one records a spectrum, in the simplest case one gets an analogue representation of signal intensity as a function of energy. From this, one can approximately measure band positions and estimate qualitative intensities. However, a major thrust in the spectroscopy of minerals is to measure site-occupancies and/or chemical composition. For this one needs derivative results more precise (and hopefully more accurate) than those obtained by qualitative processing of analogue data. One needs digitized data and an objective method of deriving quantitative information from that data. Details of data acquisition are technique dependent, but much of the data reduction necessary to obtain quantitative results is common to many spectroscopic techniques. This chapter will provide a general background in the numerical methods and general philosophy of "curve-fitting" techniques.

There is often a tendency to treat curve-fitting as a "black-box" procedure. This is extremely dangerous. It is very easy to make mistakes, particularly when one lacks a basic understanding of the general principles. This has happened with distressing frequency in the mineralogical literature and has damaged the credibility of the technique, when it is the user rather than the technique that is at fault. Consequently, all spectroscopic practitioners should be aware of certain general principles and technique associated with this aspect of spectroscopy. For this, some knowledge of statistics is essential; you may refresh your memory at the end of this chapter, where Appendix A gives definitions of the common statistical quantities used here.

GENERAL PHILOSOPHY

One may summarize the general principles of curve-fitting very briefly:

(i) From one's physical/chemical knowledge of the experiment, one sets up a mathematical model that will describe the raw data of the experiment;

(ii) One then changes the variable parameters of the model to minimize the deviations between the calculated "data" and the observed data:

(iii) If the "fit" or agreement between the calculated data and the observed data is statistically acceptable, then the (mathematical) model is taken as being a possible description of the experimental situation.

Although this sounds very straightforward, it is necessary to define very exactly what we mean by many of the terms used in this description; this will be done in the following sections. Before we do this, there is a little matter of terminology. This procedure of curve-fitting is frequently referred to by the term deconvolution. This is wrong; deconvolution is a specific mathematical operation in Fourier analysis (that is discussed in detail later in this chapter). Our spectrum consists of a summation of a series of separate curves

(spectral bands), which together with a random noise component constitutes our observed envelope of curves. Thus we resolve the observed spectrum into its constituent bands, and refer to the process as spectrum resolution.

SETTING UP THE MODEL

In this section, we use the word model in the most general sense to mean the mathematical model we set up to describe the data. This type of model can be divided into three parts:

(i) data reduction procedures to correct for various experimental factors that are independent of the other two parts of the model; this converts the raw data into a form convenient for the ensuing calculations;

(ii) deciding on an algorithm to model the background, that is the response of the equipment when no spectroscopic signal is being observed;

(iii) deciding on a function that adequately models the digitized spectrum signal, that is the band shape.

Data reduction

This is normally quite a simple procedure in most spectroscopic techniques, often being restricted to assigning weights to each observation (this will be dealt with later on). For some methods, a "split-beam" technique is used, in which the incoming radiation is split into two beams, only one of which passes through the sample, the other passing through a similar path but without undergoing sample absorption. By suitable subtraction or ratioing methods, complex background profiles may be removed easily.

Background modeling

This varies tremendously from one spectroscopic method to another. In some cases, we have a good idea what the background function is, and this can be modeled as part of the fitting procedure. In other cases, there is no a priori ideal shape for the background and a variety of ad hoc methods are used, often at the data reduction stage. We will examine these in increasing order of sophistication.

Linear interpolation: this is diagrammatically illustrated in Figure 1(a). On either side of the band of interest, the background intensity is counted at specific points B(1) and B(2). A straight line is drawn between these two points, and the intensity below this line is taken as the background intensity. This is subtracted from the total integrated intensity between points B(1) and B(2) to get the intensity (≡ area) under the peak.

This method assumes that the background is a linear function of energy, which is often not the case. The effect of this is shown in Figure 1(b). A concave background can lead to significant underestimation of the peak intensity. Obviously, different nonlinear type backgrounds can lead to different sorts of error.

Non-linear interpolation: the principal problem here is to find a suitable analytical function with which to model background behavior.

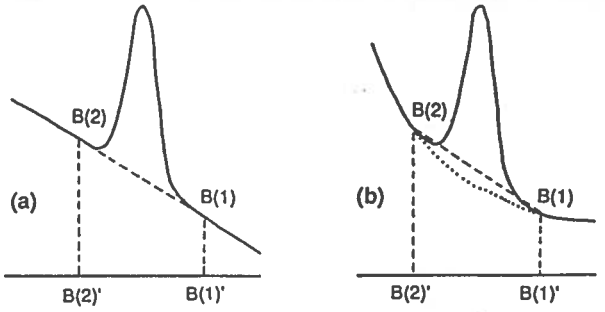

Figure 1. Background subtraction. For linear background (a), a linear interpolation gives the correct background; for a non-linear background (b), linear interpolation gives the wrong background and hence the wrong integrated peak intensity.

In some cases, there is insufficient data to warrant such a procedure, and backgrounds must be "drawn by eye". Obviously, any method that requires such a procedure cannot be considered as completely quantitative. On the other hand, a semi-quantitative method is better than no method at all.

Most analytical background functions involve either polynominals or circular functions (sines and/or cosines), and may be implemented in two different ways:

(i) the spectroscopic bands are removed from the data, and the remaining background points are then fit to the background function; this background function is then considered fixed. The background intensity for each data point in the complete data set is calculated from the background function and then subtracted from the total intensity at that point. The area under each band then represents the intensity and energy of the spectroscopic response to the incident radiation.

(ii) an approximate background function is derived as in (i). The subtraction of the background intensity is done in the actual curve-fitting process, with certain of the background terms considered as variable parameters. Thus the background fitting is iteratively optimized throughout the fitting procedure. This is the most satisfactory method, provided that there are enough background data points to constrain a good fit by the background function. This latter point is of considerable importance as there may be high correlation between background and band function variables if there are insufficient background data.

Modeling the band shape

Choosing an approximate function to accurately model the band shape is one of the more difficult aspects of generalized curve-fitting procedures. What one is essentially doing is looking at the distribution of (absorption) events as a function of energy. Consequently, band shapes are normally described by one of the usual distribution functions. We will take a look at four examples, but it is the Lorentzian and Gaussian functions that are of most use in spectroscopic applications.

The binomial distribution describes the probability P(x,n,p) c observing x positive responses from n tries where p is the probabilit of a positive response for an individual try:

$$P(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad , \qquad (1$$

where $\binom{n}{x} = n!/x!(n-x)!$. Although this is the most basic distributic function, it is not very amenable for general use as n and p ar usually not known; consequently various approximations to it are used

The Poisson distribution approximates the binomial distributic for $p << 1$, where $n \dashrightarrow \infty$ when $np = \mu$ = mean value = constant. Son algebraic juggling gives

$$P(x, \mu) = \frac{\mu^x}{x!} e^{-\mu} \quad . \qquad (2$$

The variance may be evaluated thus:

$$\sigma^2 = \sum_{x=0}^{\infty} \left[(x-\mu)^2 \frac{\mu^2}{x!} e^{-\mu}\right] = \mu \quad . \qquad (3$$

This result is of considerable importance to us with regard t the weighting of observations (see later section). When we make single observation, consisting of a number of positive responses (i.e. counts), the distribution of possible results should follow a Poisso distribution if the number of counts is small (i.e., $p << 1$). Thu the variance of the mean value μ is μ (i.e., equation (3)). If x is on value taken from this distribution (i.e., a single experimenta result), we can make the approximation

$$x = \mu \quad , \qquad (4$$

and thus the variance of the single observation x is x (and thus th standard derivation of x is $\sqrt{x}$).

The Gaussian distribution approximates the binomial distributic for $n \dashrightarrow \infty$ and $np >> 1$. It is probably the most useful distributio function, and many applications have shown that it is an ap description of the distribution of random observations for th conditions stated ($n \dashrightarrow \infty$, $np >> 1$):

$$P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad . \qquad (5$$

One extremely useful characteristic of this function is that th most probable estimate of the mean value, μ, is the average of th observations x:

$$\mu = \bar{x} \quad , \qquad (6$$

and the standard deviation, σ, is a variable in the function Another important characteristic of this curve is its half-width, Γ the full-width at half-maximum height; this has very importan spectroscopic implications. The half-width is defined as the range o x within which the probability $P(x,\mu,\sigma)$ is half its maximum value. little algebra shows that

$$\Gamma = 2.354\ \sigma \quad . \qquad (7$$

Another important distribution function is the Lorentzian (or Cauchy) distribution:

$$P(x, \mu, \Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(x-\mu)^2 + (\Gamma/2)^2} \quad . \qquad (8)$$

This function is unrelated to the binomial distribution but has been found appropriate for describing resonance data; hence it is perhaps the most important function from a spectroscopic viewpoint. For purely mathematical reasons, we cannot define a standard deviation for the Lorentzian distribution; instead we can characterize its dispersion by its half-width, Γ, which is defined such that $P(x,\mu,\Gamma) = \frac{1}{2}P(\mu,\mu,\Gamma)$ for $x-\mu = \Gamma/2$; that is, when the deviation from the mean is equal to one-half of the half-width, then the probability function is half of its maximum value (which occurs at the mean value $x = \mu$).

A comparison of the Lorentzian and Gaussian curves is shown in Figure 2, in which both curves have the same half-width. Note that the Gaussian curve has a higher maximum value, whereas the Lorentzian curve has a wider tail. Very frequently one finds that none of the ideal distribution functions adequately models the experimental data. In this case, "mixtures" of curves can be used, whereby one's model curve consists, for example, of A Lorentzian character and (1-A) Gaussian character.

Table 1 summarizes the details of the various distribution functions we have considered. Although they are the most common in spectroscopic applications, they are not the only sorts of curves one can use; and several other functions are also shown.

CRITERION OF "BEST FIT"

Having set up a mathematical model, the next step is to "fit" the model to the experimental data. This normally involves varying the parameters of the model until the model shows the best agreement with the observed data. This raises the question of what do we mean by the best agreement or best fit. Intuitively one expects the best agreement when some function of the deviations between the observed values and the corresponding values calculated from the model is minimized. However, what is this function?

Consider that our model (Fig. 3) is of the form

$$y = mx + c \quad . \qquad (9)$$

Intuitively one might expect that the best fit occurs when the sum of the deviations from the observed data are minimized. However, this is not a good measure, as positive and negative deviations tend to cancel each other out. This can be overcome by minimizing the sum of the magnitudes of the deviations, but this produces numerous practical problems with regard to the minimization procedure.

Least-squares method

There is no unique method for defining a correct criterion of best fit. However, if we assume a Gaussian distribution of

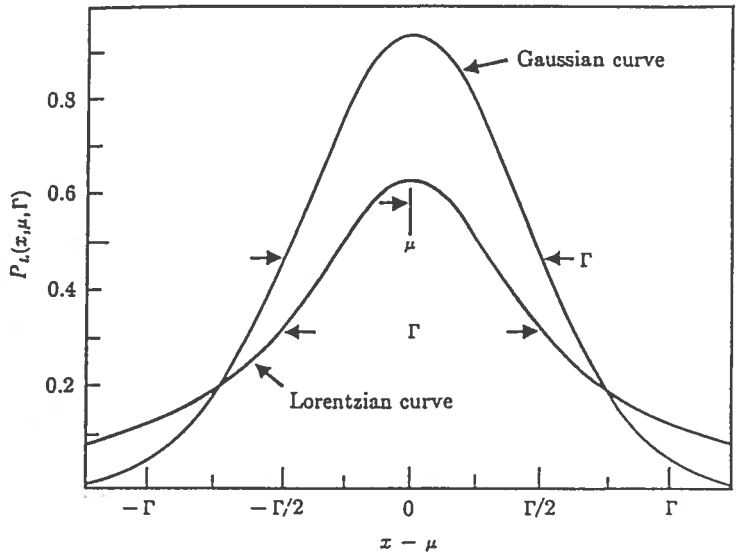

Figure 2. Comparison of Gaussian and Lorentzian curves of equal half-width. Note that the Gaussian curve has a higher maximum whereas the Lorentzian curve has a wider tail; after Bevington (1969).

Table 1. Various distribution functions

(1) Binominal distribution:	$P(x,n,p) = \binom{n}{x} p^x(1-p)^{n-x}$
(2) Poisson distribution:	$P(x,\mu) = \frac{\mu^x}{x!} e^{-\mu}$
(3) Gaussian distribution:	$P(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
(4) Lorentzian distribution:	$P(x,\mu,\Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(x-\mu)^2 + (\Gamma/2)^2}$

(1) $P(x,n,p)$ = probability of observing x positive responses from n tries, where p = probability of response for an individual try
(2) approximates (1) for $p \ll 1$, where $n \to \infty$ when $np = \mu$ = mean value = constant
(3) approximates (1) for $n \to \infty$ and $np \gg 1$; σ = standard deviation
(4) unrelated to (1) Γ = half-width, a measure of the dispersion of the distribution

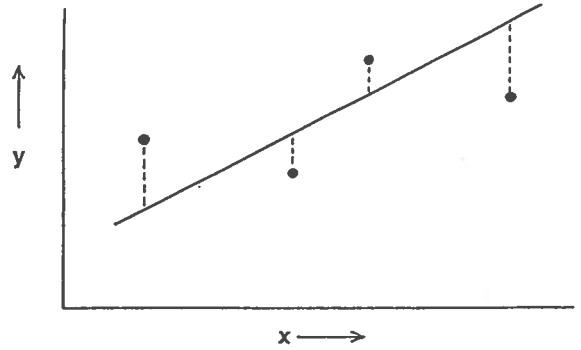

Figure 3. A linear model (y = mx + c) to the four data points shown in the graph. We wish to minimize some function of the deviations of the points from the line for it to be the "best fit".

probabilities (see equation (5)), we can derive a useful and robust method: the method of least-squares.

Suppose we wish to fit our experimental observations thus:

$$y = f(x) \quad . \tag{10}$$

Where our experimental observations are (x_i, y_i), the discrepancy δy_i between the observed and calculated value is given by

$$\delta y_i = y_i - f(x_i) \quad . \tag{11}$$

Let us write the actual relationship (which we cannot know) as

$$y(x) = f_o(x) \quad . \tag{12}$$

For any value x_i of x, the probability of making the observed measurement y_i is given by

$$P_i = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left\{\frac{y_i - y(x_i)}{\sigma_i}\right\}^2\right] \quad . \tag{13}$$

The probability of making the observed set of N observations of y_i is the product of the N probabilities of equation (3):

$$P(f_o) = \prod_{i=1}^{N} P_i = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left\{\frac{y_i - y(x_i)}{\sigma_i}\right\}^2\right] \quad . \tag{14}$$

For any estimated function f, the probability that we will make the observed set of measurements is

$$P(f) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{\delta y_i}{\sigma_i}\right\}^2\right] \quad . \tag{15}$$

The important point about equations (12) to (14) is that we do not know $f_o(x)$ and hence cannot evaluate the probabilities in

equations (13) and (14). However, what we can do is use the method o maximum likelihood to allow us to overcome our ignorance. We make th assumption that the observed set of measurements is more likely t come from the actual parent distribution (equation (12)) rather tha any other distribution with different numerical parameters for f(x) Thus the probability of equation (12) is the maximum probabilit possible with equation (14), and the best estimates for the parameter of f(x) are the values which maximize the probablity of equation (15)

In equation (15), maximizing P(f) requires minimizing th summation term inside the exponential, which is usually designated χ^2 and sometimes referred to as the residual:

$$\chi^2 \equiv \sum_{i=1}^{N}\left\{\frac{\delta y_i}{\sigma_i}\right\}^2 = \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left[y_i - f(x_i)\right]^2 \quad . \tag{16}$$

Thus the "best fit" to the data for a specific model is the on that minimizes the weighted sum of the squares of the deviation between the observed and 'calculated' data. The method by which thi fit is found is called the least-squares method, and will be discusse later.

MINIMIZATION METHODS

We have seen how we decide what is the criterion of minimizatio (optimal fit); now we need to examine the methods by which suc minimization is done. These fall into three broad groups, each o which uses a different general philosophy:

(i) Pattern search methods,
(ii) Gradient methods,
(iii) Analytical solution methods.

Pattern search methods are generally rather crude, and ar usually used only in conjunction with one of the other two methods However, it is very instructive to work through some very simpl pattern search examples, as they give one a feeling for th convergence process in general (and what can go wrong with it) something one does not easily get from the less visual methods of (ii and (iii).

Pattern search

The general idea of these methods is very simple, one just make a search throughout parameter space such that one always moves t lessen the residual, without using any of the properties of th algorithm used to model the spectrum.

The simplest method is to divide up parameter space into network and calculate the residual at nodes of this network. On picks an arbitrary starting point and determines by inspection of th surrounding grid points, which is the best way to move. Having don this, the process is iterated until one cannot move any more, at whic time one is, hopefully, at or near the minimum value of the residual.

A very simple 2-dimensional example is shown in Figure 4. W start at point (1,1) and move down the first column until we can mov

Figure 4. χ^2-values at network nodes in parameter space; lines connecting nodes show paths for various starting points.

	1	2	3	4	5	6	7	8	9	10
1	547	527	483	432	403	317	286	253	201	233
2	486	517	446	397	352	333	277	221	156	201
3	444	483	443	373	306	386	261	153	83	163
4	367	399	376	351	256	401	274	143	72	136
5	329	365	333	321	243	302	255	216	117	200
6	283	313	290	262	196	171	125	153	137	222
7	241	234	228	200	132	84	4	73	165	247
8	276	222	168	143	161	96	38	113	193	264
9	331	296	199	127	101	74	57	186	261	307
10	374	342	247	168	193	111	88	203	290	356

no further at row 7 in Figure 4. Then we inspect one point each way along row 7, and move in the direction defined by the lesser value until we can move no further along this row. The process is repeated until one cannot move at all [point (7,7)]; at this stage, it is (again, hopefully) at or close to the minimum (that is, the best fit). If one wishes to be more precise, then one can construct another such network on a larger scale close to the end-point of the first search, and repeat the procedure.

Let us repeat this example, but start at point (1,2) instead. This time we move along row 1 and follow a very different path; we arrive at the minimum point much more quickly. This illustrates two important points concerning fitting procedures in general:

(i) the path taken to a minimum can be dependent on one's starting point;

(ii) some paths to a minimum are much shorter than others.

Consequently, it is of importance to try and optimize one's path to the minimum to reduce computation time.

Let us repeat the example again, starting at point (1,5). We move along row 1, eventually converging at point (9,4). Note that this is a different end-point from the previous two paths. We are still at a minimum, that is we cannot move according to our criterion of moving along a row or column to a lower residual. However, the value of the residual is greater than that on the previous paths; we have converged to a false minimum, it is only a local minimum and not a global minimum. Thus

(iii) convergence at a false (local) minimum can be a major problem in all fitting procedures,

and there is no intrinsic way of identifying whether or not one has converged only at a local minimum.

The example we have been considering is only a 2-variable problem. It serves to indicate some of the features of convergence problems, but is not a practical method. Each of the parameters is varied independently (i.e., along a row or down a column); consequently, with

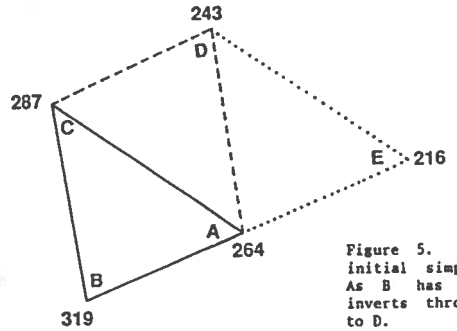

Figure 5. A two-variable simplex: ABC is the initial simplex with the χ^2 values indicated. As B has the largest χ^2 value, this vertex inverts through the center of the opposite edge to D.

a large number of variable parameters, the path taken through variab space is extremely convoluted and inefficient.

Simplex method

A more sophisticated pattern search approach is the simpl method, which varies all parameters simultaneously, exploiting t geometric properties of a simplex. A simplex is a geometrical figu with (n+1) vertices, where n is the number of variable parameter Figure 5 shows the analogous simplex for a 2-dimensional proble Optimization proceeds in the following manner. The values of χ^2 a calculated at the vertices of the simplex (points A,B and C in Fi 5). The vertex with the largest χ^2 value (B in Fig. 5) is invert through the centre of the opposite edge (to point D in Fig. 5). Nex χ^2 is evaluated at point D, and the process is repeated to form AD Iteration of this procedure moves the simplex around parameter spa until it arrives in the vicinity of the minimum, where it cannot move

There are various modifications of the simplex method, where the amount of movement during inversion is variable; this allows t simplex to expand or contract depending on how close it is to t minimum, and can greatly improve the rate of convergence. Simpl methods can be very useful when inadequate starting models a available for analytical least-squares methods.

Gradient method

Continuing with our 2-dimensional example developed in t previous section, a more efficient path down to the minimum would down the steepest slope (the analogy of a mountain stream is a go one), and the method is often referred to as the method of steepe descent.

The gradient $\nabla\chi^2$ is that vector the components of which are equ to the rate at which χ^2 increases in that direction:

$$\nabla\chi^2 = \sum_{j=1}^{N} \left(\frac{\partial \chi^2}{\partial a_j} \hat{a}_j \right) \quad , \qquad (1$$

where $\hat{a}_j$ is a unit vector. The gradient may be obtained exact by evaluation the partial derivatives of the χ^2 function,

approximated from the observed variation of χ^2 for small incremental changes in the variables:

$$(\nabla\chi^2)_j = \frac{\partial\chi^2}{\partial a_j} \sim \frac{\chi^2(a_j + w\delta a_j) - \chi^2 a_j}{w\delta a_j} \quad . \qquad (18)$$

where δa_j is an incremental step in a_j and w is a weight ($\approx$ 0.1). A combination of the two-gradient evaluation methods is optimal, as the analytical calculation of the partial derivative is slow, but the approximation is imprecise. Most efficient is to initially calculate the derivatives exactly, and then increment all variables in the optimum direction, either approximating χ^2 every (few) increments, or continuing until χ^2 begins to rise, whereupon χ^2 is calculated exactly again. This method does have problems with irregular surfaces (those containing curving valleys) and is not good close to the minimum. Consequently, it has tended to be replaced by analytical solutions.

Analytical solution methods

All of the previous methods miminize χ^2 and hence are classed as least-squares methods. However, common usage often reserves the use of this term to analytical solution methods.

Linear functions. Let us first consider the case in which our dependent variable, y(x) is a linear function of the coefficients, a_j, of the fitting function:

$$y(x) = \sum_{j=0}^{n} a_j f_j(x) \quad . \qquad (19)$$

Remember that here we are talking about the function being linear in the coefficients; thus the equation $y = ae^{bx}$ is linear in a, but not in b; similarly $y = ax + bx^2$ is linear in both a and b.

At the minimum point of equation (19), the derivative of χ^2 with respect to each variable is zero:

$$\frac{\partial\chi^2}{\partial a_i} = 0 \qquad (i = 1,n) \quad . \qquad (20)$$

Writing χ^2 as

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \left(y_i - y(x_i)\right)^2 \right] = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \left(y_i - \sum_{j=0}^{n} a_j f_j(x_i)\right)^2 \right] \qquad (21)$$

and applying the conditions of equation (20) we get

$$\frac{\partial\chi^2}{\partial a_k} = \frac{\partial}{\partial a_k} \left[\sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \left(y_i - \sum_{j=0}^{n} a_j f_j(x_i)\right)^2 \right] \right] \quad (k = 1,n) \quad . \qquad (22)$$

There are n of these equations, one for each coefficient a_k in the function of equation (22). Note that the subscript k must be used in the derivative, as this subscript is independent of j in the function expression itself; however, they are the same coeficients.

Taking the derivative and using the relation

$$\frac{\partial}{\partial a_k} \sum_{j=0}^{n} a_j f_j(x_i) = f_k(x_i)$$

together with a little algebraic manipulation gives

$$\sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} y_i f_k(x_i) \right] = \sum_{j=0}^{n} \left[a_j \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} f_j(x_i) f_k(x_i) \right] \right] \quad (k = 0,n) \quad . \qquad (23)$$

The (n+1) unknowns in this series of equations are a_j (j=0,n); if there are (n+1) equations of this form (i.e., (k=0,n) in equation (23)), then we may solve these equations for a_j, the (fitting parameters of interest in the model.

These equations could be solved by successive subtraction (elimination of variables) as one does with simple simultaneous equations. Of course, this is far too clumsy for most uses, and matrix methods are normally used. Equation (23) can be much more compactly represented in matrix form:

$$R_k = \sum_{j=0}^{n} a_j M_{jk} \quad \text{or} \quad R = a\,M \quad . \qquad (24)$$

where Rk is a column vector (the normal vector) given by

$$R_k = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} y_i f_k(x_i) \right] \quad . \qquad (25)$$

and Mjk is a symmetric n x n matrix (the curvature matrix) given by

$$M_{jk} = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} f_j(x_i) f_k(x_i) \right] \qquad (26)$$

The set of equations (24) are often called the normal equations. Multiplying both sides of equation (24) by M^{-1}, the inverse of the curvature matrix M is

$$aMM^{-1} = RM^{-1} = a \quad . \qquad (27)$$

Hence one may write the solution to the normal equations as

$$a_j = \sum_{k=0}^{n} R_k M_{jk}^{-1} \quad . \qquad (28)$$

and the principle labor involves the inversion of the curvature matrix; hence this is sometimes known as the matrix inversion method.

It is instructive to briefly consider the evaluation of the inverse matrix, as it is here that the least-squares method can give problems. The inverse of a matrix is defined as the adjoint of the matrix divided by its determinant:

$$M^{-1} = \frac{M^{*}}{|M|} \qquad (29)$$

If the determinant is zero (i.e., $|M| = 0$), the inverse of the matrix is indeterminate and the matrix is singular. When this is the case, one cannot find a solution to one's problem, the fitting process failing (diverging) when this happens. This occurs generally because one of the constituent equations of equation (23) can be expressed (either exactly or approximately) as a linear combination of some of the other equations.

Linearization of non-linear functions. Here we will consider the case in which there is a non-linear relationship between our dependent variable y(x), and the coefficients of our fitting function.

We may recognize two cases, one trivial and one not. Consider for example, the function

$$y(x) = ae^{bx} \quad . \tag{30}$$

This is linear in a, but not linear in b. However, we may take logarithms

$$\ln y(x) = \ln a + bx \quad , \tag{31}$$

whereupon $\ell n\ y(x)$ is linear in ℓn a and b, and we may proceed with our least-squares method as above.

In the second case, linearization is not achieved so easily. For example, in the case of fitting a Lorentzian curve plus a quadratic background, we can write

$$y(x) = \frac{1}{\pi} \frac{\frac{a_1}{2}}{(x-a_2)^2 + (\frac{a_1}{2})^2} + a_3 + a_4x + a_5x^2 \quad . \tag{32}$$

and we cannot separate out all of the a_j parameters to form an expression involving simple summation. In this case, we may use Taylor's expansion, which approximates a function y(x) around a point x = a by the following expression:

$$y(x) = y(a) + y'(a)\,(x-a) + \ldots + \frac{y^{n'}(a)}{n!}(x-a)^n + \ldots \quad . \tag{33}$$

If (x-a) is small, then we may ignore terms involving $(x-a)^n$, n>1 and writing (x-a) as δa, we get

$$y(x) = y_o(x) + \sum_{j=1}^{n} \left[\frac{\partial y_o(x)}{\partial a_j} \delta a_j \right] \tag{34}$$

This function is now linear in the parameter increments δa_j, and we can write χ^2 directly as a function of δa_j:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \left(y_i - y_o(x_i) - \sum_{j=1}^{n} \left[\frac{\partial y_o(x_i)}{\partial a_j} \delta a_j \right] \right)^2 \right] \quad . \tag{35}$$

As before, taking the derivative and setting it equal to zero (tl minimization criterion), we get

$$\sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} (y_i - y_o(x_i)) \right] = \sum_{j=1}^{n} \left[\delta a_j \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \frac{\partial y_o(x_i)}{\partial a_j} \cdot \frac{\partial y_o(x_i)}{\partial a_k} \right] \right] \quad (k=1,n) \quad . \tag{36}$$

Writing this in matrix form

$$R = \delta a M \quad , \tag{37}$$

where

$$R_k = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \left(y_i - y_o(x_i) \right) \right] \tag{38}$$

and

$$M_{jk} \approx \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \frac{\partial y_o(x_i)}{\partial a_j} \cdot \frac{\partial y_o(x_i)}{\partial a_k} \right] \quad . \tag{39}$$

Thus we now have a set of n linear equations by which we ca calculate δa_j, the parameter shifts necessary to move the initia parameter values to their values for the minimum χ^2. This set o equations may be solved by matrix inversion methods as outlined above.

SOME ASPECTS OF SPECTRUM REFINEMENT

Tactics

Here we will consider some of the tactics involved i least-squares refinement, as judicious use of these can greatl increase refinement efficiency (and decrease frustration for th spectroscopist).

One starts with a model function to represent the spectrum. Thi usually consists of a background function and several line-shap functions. Generally the model function is linear in some parameter and nonlinear in other parameters. The linear parameters wil converge to the neighborhood of their correct values whatever thei starting values, and thus should be refined first, with th "nonlinear" parameters held constant; of course, it is sensible t use starting values close to the final values just for efficiency o computation. The nonlinear parameters in one's model have normall been linearized by Taylor expansion of the initial nonlinea equations. We assume that δa_j (= x-a_j) is small and hence th nonlinear terms in the Taylor expansion can be ignored. There ar two important points with regard to this:

(i) if δa_j are too large (that is we are far from our minimun value of χ^2), then the approximation breaks down and th process probably will not converge.
(ii) the values we get for a_j are also only an approximation, an we have to iterate through the procedure until $\delta a_j \approx 0$.

Thus it is important to have a reasonably accurate starting estimate for the parameters. Generally one's model function will be linear in some parameters but not in others. It is usually good practice to first refine the important linear parameters (background, perhaps a scaling constant), and then start to refine the nonlinear parameters. If one's starting parameters are close to the true values, then one can immediately refine all variable parameters simultaneously. More often, one is not that close and a more cautious approach may be necessary, gradually increasing the number of variable parameters as χ^2 decreases. When the starting model is not good, then immediate refinement of all variables can cause oscillation and slow convergence or even divergence. In spectroscopic applications, a common tactic is to initially fix peak positions and half-widths, and refine peak areas (or even peak area ratios).

Convergence is attained when the least-squares refinement procedure calculates parameter shifts that are much less than the standard deviations of the variable parameters. At this stage, it is IMPERATIVE that all variable parameters be refined simultaneously. If this is not done, covariance terms will be missing from the dispersion (variance-covariance) matrix, and derivative parameter standard deviations will be wrongly calculated (i.e., underestimated by omission of covariance terms in equation A(12), see Appendix A).

Correlation and constraints

A spectrum contains a specific amount of information, and it is not possible to get out of the spectrum refinement procedure more information than the spectrum contains. However, what we can do is input external information into the fitting process such that we can extract the information in the spectrum in a form that we want. This procedure is very common in spectrum fitting, and involves the use of linear constraints in the fitting process. In the simplest case, a variable, a_k, is constrained to be related to a set of other variables:

$$a_k = \sum_{i=1}^{n} b_i f(a_i) \qquad (i=j) \qquad (40)$$

The set of functions $b_i f(a_i)$, $(i=1,n;\ i\neq j)$ is the external information put into the fitting process. Usually what this does is to greatly decrease variable correlation in the fitting process; this leads to more precisely determined variables, but it should be realized that their accuracy depends on the correctness of the constraint equations used. An example of this is shown in Figure 6. Mossbauer spectra of a series of minerals was used to determine Fe^{3+}/Fe^{2+} ratios, using the constraint that all the half-widths of the Fe^{2+} peaks were equal. Figure 6 shows that the half-widths of the Fe^{3+} doublet is strongly correlated with the Fe^{3+}/Fe^{2+} ratio, with weak Fe^{3+} doublets showing very wide half-widths. It seems probable that the half-width constraint for the Fe^{2+} peaks is not exact, and that the Fe^{3+} doublet is absorbing the error associated with this; the weaker and less well defined the Fe^{3+} doublet, the more it can widen and absorb error from the Fe^{2+} doublets. Thus one has to be very careful in the use of such constraints.

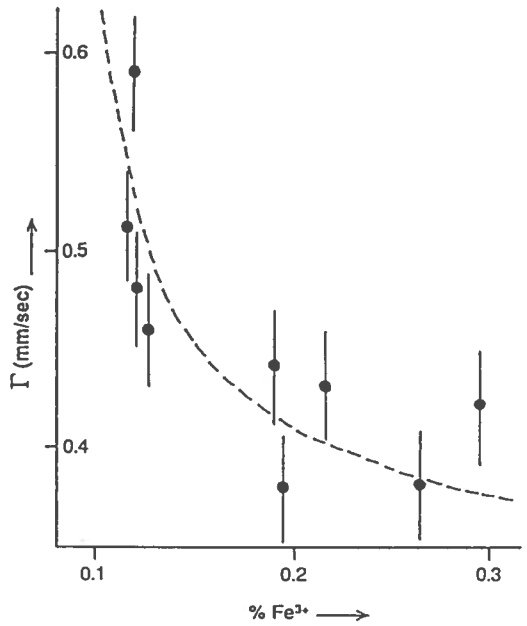

Figure 6. Half-width of Fe^{3+} doublet as a function of $Fe^{3+}/(Fe^{3+},Fe^{2+})$ for a series of isostructural minerals examined by Mössbauer spectroscopy.

In spectrum fitting procedures, constraints can involve

(i) constraining peak half-widths,
(ii) constraining peak intensities,
(iii) constraining peak positions,
(iv) constraining peak shapes,
(v) constraining peak splittings.

Often, the first two types involve constraining parameters to be equal (equal half-widths is a very common constraint); however, it is sometimes advantageous to constrain such parameters to have constant ratios.

The use of constraints in spectrum fitting is often essential to get any result at all. However, one must be very careful about the results; if the constraints are not appropriate, then the results will be wrong despite (possibly) having high precision.

False minima and model testing in least-squares fitting

It was noted earlier that there is no mathematical way to test least-squares minima for global validity. Rather, the general method is to attempt convergence from a variety of starting points in parameter space. This is frequently not an option if many constraints must be used for convergence, or if the fit is extremely sensitive to the values of particular parameters.

Another strategy involves making assumptions about the nature of false minima. In least-squares refinement, false minima occur because

of the existence of a relative minimum for a particular subset of the data set. If these data could be given zero weight in the refinement procedure, the true global minimim would be obtained. Although the subset responsible cannot be determined by analytic tests, it's effect can be overwhelmed or circumvented. The former is achieved by increasing the number of independent data. This effectively dilutes the possibility of finding a relative minimum in parameter space. The latter can be done (sometimes) by separating the data at random into multiple sets and refining each set separately. It would be most unlikely that each data subset would lead to the same minimum in the same region of parameter space unless the minimum were truly global.

Another consideration in the detection of false mimina is the reasonableness of the refined parameter values. If these can be subjected to model tests that independently affirm likelihood, or if constraints can be used, the possibility of false minima are reduced. It is also possible to put a restriction on refined parameters such that they will not adopt unreasonable values, (e.g., negative bond lengths). In the latter case, the refinement is said to be restrained (Prince, 1982).

Signal-to-noise effects in fitting

Up to now, we have considered only one goodnes-of-fit parameter, χ^2. This allows us to determine the probability that the model we fit is more (or less) representative of the data than some other fit. This is done by reference to $P_x(\chi^2,v)$, the integral of the χ^2 probability distribution function (Bevington, 1969). However, it will not allow us to compare fits to data of differing signal/noise ratios. This is because χ^2 is sensitive to the absolute size of the residuals. Fits of the same model to indentical data sets with different ratios of signal/noise will show a strong dependence of χ^2 on the signal/noise ratio (Waychunas, 1986). The easiest way around the problem is to attempt all trial models on all data sets and compare the χ^2 values obtained only for a given data set. Because this is time consuming or impractical, one must either make allowances in χ^2 or develop a new goodness-of-fit parameter insensitive to signal/noise variations. One such parameter was derived by Ruby (1973) and is called MISFIT.

Variations on χ^2 goodness-of-fit parameter

The MISFIT goodness-of-fit parameter is defined as the ratio of two values calculated from a test fit, M = D/S, where D is the discrepancy or distance of the fit from the actual data, and S is the signal or total spectrum above a baseline. MISFIT is thus a fractional assessment of the fit quality, and in an ideal case, should be independent of the magnitude of S.

The quantity D is derived from χ^2, and has the formulation

$$D = \sum_{I=1}^{N} \left[\left(\frac{(Y_c(I) - Y_d(I))}{\sqrt{Y_d(I)}} \right)^2 - 1 \right] \qquad (41)$$

where $Y_d(I)$ are the experimental data points,
$Y_c(I)$ are the calculated model data points,
N is the total number of data points.

S is defined as

$$S = \sum_{I=1}^{N} \left[\left(\frac{(Y_o(I) - Y_d(I))}{\sqrt{Y_d(I)}} \right)^2 - 1 \right] \qquad (4.$$

where Y_o is the baseline value.

Unlike χ^2, MISFIT is not compared with a distribution functio but with its own uncertainty, ΔMISFIT. This is a simple comparis which has obvious meaning. If ΔMISFIT is small relative to MISFI then we know the percentage error in the fit. We then seek minimize MISFIT by adopting improved models. If ΔMISFIT significant relative to MISFIT, then we are not testing the model ve well (i.e. the data are too poor). This use of ΔMISFIT is critic for MISFIT to have any utility over χ^2. ΔMISFIT is defined as

$$\Delta M = \left(\frac{1}{S}\right) \sqrt{\left[n(1 + M^2) + 4D(1 + M)\right]} \qquad (4$$

Some confusion in the use of MISFIT may occur, as Ruby (1973 actually devised two formulations based on differing assumptions Either form works fairly well, but both may diverge to large values a very low S values and large data point variances. These deviation have been discussed by Waychunas (1986).

Another function which measures goodness-of-fit is th crystallographic R-factor and similar functions. The R-factor i defined (Hamilton, 1964) as

$$R = \left[\frac{\sum_{I=1}^{N} w_i \left(|F|_i^{obs} - |F|_i^{calc} \right)^2}{\sum_{I=1}^{N} w_i \left(F_i^{obs} \right)^2} \right] \qquad (44$$

where $|F|_i^{obs}$ and $|F|_i^{calc}$ refer to the observed and calculated value of some function |F| at data point i, having weight w_i.

In the limit of a perfect fit, R goes to zero. The R-factor i scaled by the size of the observed points, so that it is relativel insensitive to the magnitude of differing data sets. R-factor calculated for fit models can be tested for significance by evaluatin the ratio R

$$R = \frac{R_1}{R_o} = \left[\frac{p}{n-p} * F_{p,n-p,\Omega} + 1 \right]^{\frac{1}{2}} \qquad (45$$

where p is the number of fitted parameters,
n is the number of data points,
F is the well-known probability distribution,
Ω is the significance level (1-confidence level to which th R-factors are to be tested.

For example, suppose one has 70 data points and must fit 10 parameter to a confidence level of 95%. From tables of the F distribution, $F_{10,\,60,\,0.05} = 1.9926$. R is thus found to be 1.1541. If satisfactory R value obtained from fitting procedures is 10.50%,

then a value of 12.12% represents a poorer fit in 95% of the possible cases. Similarly, an R value of 9.10% represents a better fit in 95% of all cases. A similar application of the F distribution can be used with X^2 values (Bevington, 1969).

It is also possible to devise weighting schemes for goodness-of-fit parameters, such that some data points do not contribute significantly to the fit parameter regardless of their true variance. This effectively results in selective fitting of some partition of the spectral data set. Such action would damage the utility of a statistical analysis of the goodness-of-fit parameter magnitude, but may aid least-squares fitting by improving chances for minimization of the fit parameter.

Variations in spectral line shape

When treating any aspect of mineralogical spectroscopy, we will frequently be concerned with the spectral line shape. We will need to understand the significance of broader lines in the spectrum of one material over another, of asymmetry in line shapes, and the types of line shape variations that are the result of particular physical processes in the sample or in the spectrometer. The ability to assume a particular line shape allows us to constrain fitting procedures and extract parameters that are much simpler to manipulate and compare. The basic experimental line shapes were noted earlier, viz. Gaussian and Lorentzian or Cauchy. These lines are symmetric about their centroids, and represent the distributions of energy emitted or absorbed by a particular atom or atomic system in the spectroscopic process. It is usually possible to calculate the minimum possible line width. Such a line width is never actually observed because of the effects of the spectrometer itself, or of non-ideal conditions in the sample. The line width and shape are affected by:

Excited-state lifetime. The natural line width for a spectroscopic line is determined by the lifetime of the excited state via the Heisenberg uncertainty principle: $T*\delta E \geq h/2\pi$, where T is the lifetime of the excited state, h is Planck's constant and δE is the energy uncertainty. The longer lived the excited state, the more sharply defined is the transition energy.

Döppler and collision effects. These effects are primarily relevant to the spectroscopy of independent molecules rather than atoms and molecules in solids. However, Gaussian line-broadening can occur due to the Döppler shifts created by rapid thermal vibrations at high temperatures.

Saturation effects. The line shape can be altered if excitation of a system is so powerful that the excited state is saturated (i.e., the population of excited atoms or electrons is equal to that in the ground (unexcited) state). Under such conditions, the absorption coefficient of the sample is dependent on the intensity of the incident radiation, instead of being an inherent characteristic of the sample.

Relaxation effects. The excited states of a system will decay with a characteristic rate that depends on the physical processes affecting the particular state. Optical spectra usually record

electronic states with lifetimes of about 10^{-6} s, so that a fa relaxation process occurs. In this process, energy from the excit state is removed and delivered to the crystal structure as vibration or to other electronic centers such as excited electrons or exciton Much slower relaxation rates may be found in NMR and certain far- excited states. Saturation broadening is increasingly more likely f slower relaxation rates.

Besides leading to a change in the line shape of a given excit state transition, relaxation processes can create "mixing" of sever excited (or ground) states, or averaging of physical parameters. Tl latter effects can create very complex line shapes, such as the parti evolution of multiline spectral features into singlets. In Figure the effect of averaging nuclear magnetic hyperfine spin states is sho in the Mössbauer spectrum of FeF_3. The low-temperature spectr consists of a six line pattern representing allowed transitions to t ground state from each of the hyperfine magnetic states in a stat magnetic field (see Chapter 7). This static field is due to magneti ordering of the Fe^{3+} electronic spins. As the temperature is increas toward the magnetic disordering point (the Néel point), the electron spins begin to decouple and spin fluctuations occur. Tl time averaged field seen by the Mössbauer hyperfine states is weak and varies in direction. This reduces the splitting of these state and the spectrum begins to collapse. Above the Néel point, t electronic spins are not coupled and there is no separation of t various hyperfine states. Temperature changes near the Néel point hav dramatic effects on the observed Mössbauer spectra, but there is littl effect over other temperature ranges.

In very dilute Fe^{3+} minerals and compounds, a related effec occurs but usually over a larger temperature range. In such materials there is no definite magnetic ordering point, but at low temperature the electronis spins on each individual Fe^{3+} will couple to produc hyperfine state splitting. As the individual ions are well separate in a dilute system, there is little relaxation due to spin-spi exchange and the relaxation rate is slow. Raising the temperatur increases fluctuation of the Fe^{3+} spins, but there is no sudden los of coupling as would occur with long-range magnetic order. Henc there is a gradual change in the observed Mössbauer spectrum (Fig. 8).

Relaxation due to fluctuations in chemical shifts can affect bot NMR and Mössbauer spectra. One example of this in the latter i electron hopping. If near-neighbor Fe^{2+} and Fe^{3+} ions exchange th 6^{th} 3d electron between them, the oscillating valence is manifested o the Mössbauer nuclei as a fluctuating chemical shift (isomer shift) At slow oscillation rates, discrete spectral lines due to each valenc state are observed, but these smear into one another as the rat increases. At hopping rates faster than the lifetime of the Mössbaue excited state, only one valence-averaged set of features remains (se Chapter 8). The NMR analog could be an atom whose chemical shif fluctuates as it diffuses through a crystal or liquid. The relaxatio spectra for these hopping and diffusion cases might look like th idealized spectra in Figure 9.

A very complete treatment of relaxation effects in spectroscop can be found in Poole and Farach (1971).

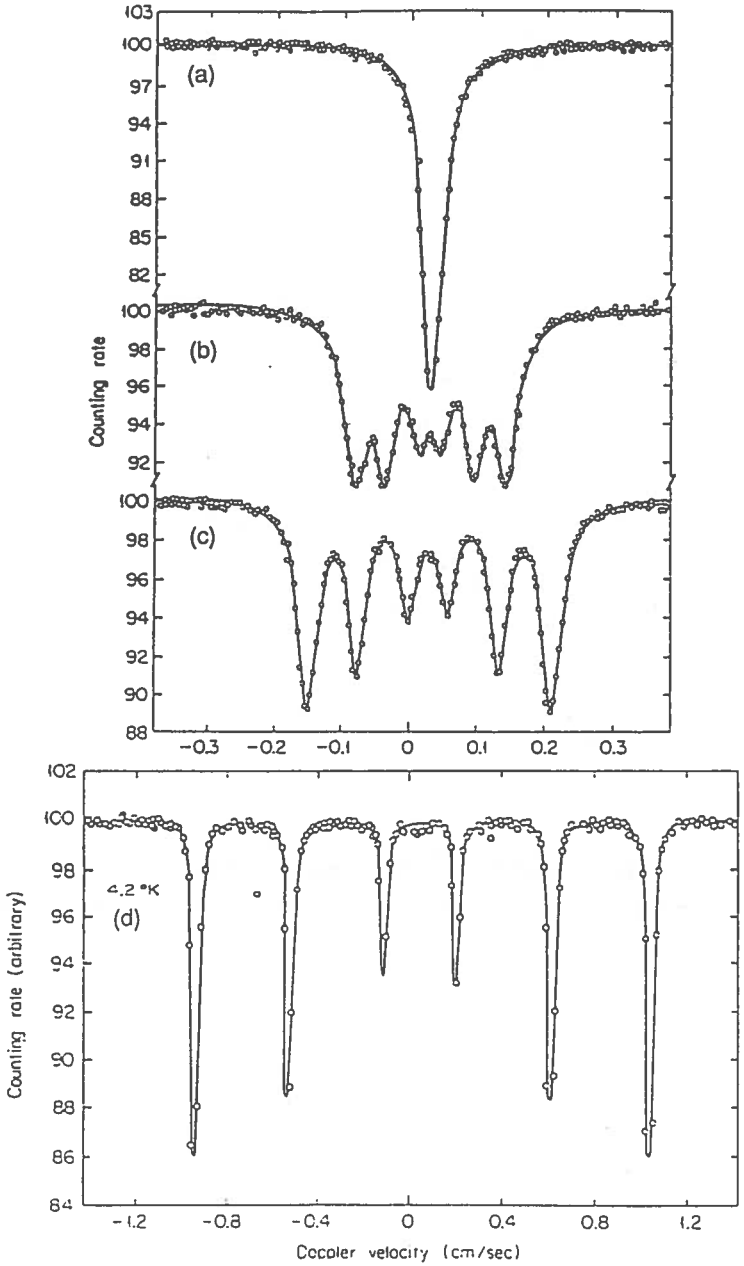

Figure 7. Mössbauer spectrum of FeF_3 at (a) 363.4 K, (b) 362.7 K, (c) 361.5 K, (d) 4.2 K. There is a tremendous change in the spectrum near the Néel temperature of 363.1 K. Note in particular how the spacing of the outmost lines changes; after Wertheim et al. (1968).

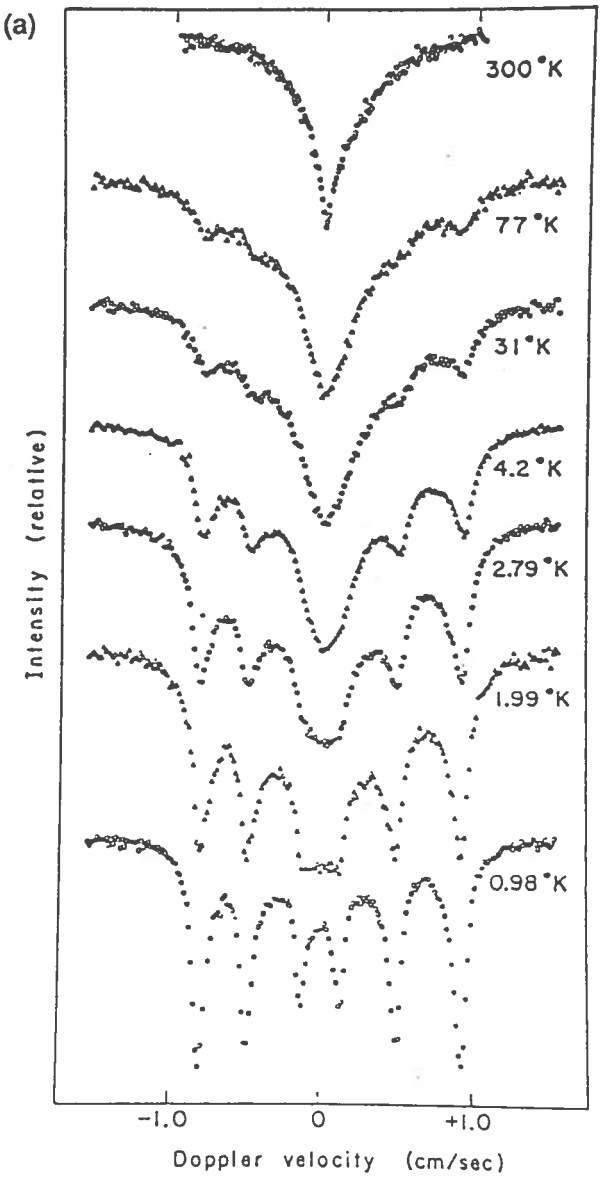

Figure 8. (a) Mössbauer spectra of dilute Fe^{3+} in a nonmagnetic material, Ferrichrome A. (b) [following page] Calculations of Fe^{3+} Mössbauer relaxation spectra for differing values of the relaxation rate t. The values for t are (top to bottom) 10^{-12} s, 10^{-9} s, 2.5×10^{-9} s, 5.0×10^{-9} s, 7.5×10^{-9} s, 2.5×10^{-8} s, 7.5×10^{-8} s and 10^{-6}s. The lifetime of the ^{57}Fe excited state is about 10^{-8} by comparison. Relaxation rates faster than the lifetime result in averaged spectra; after Wickman (1966).

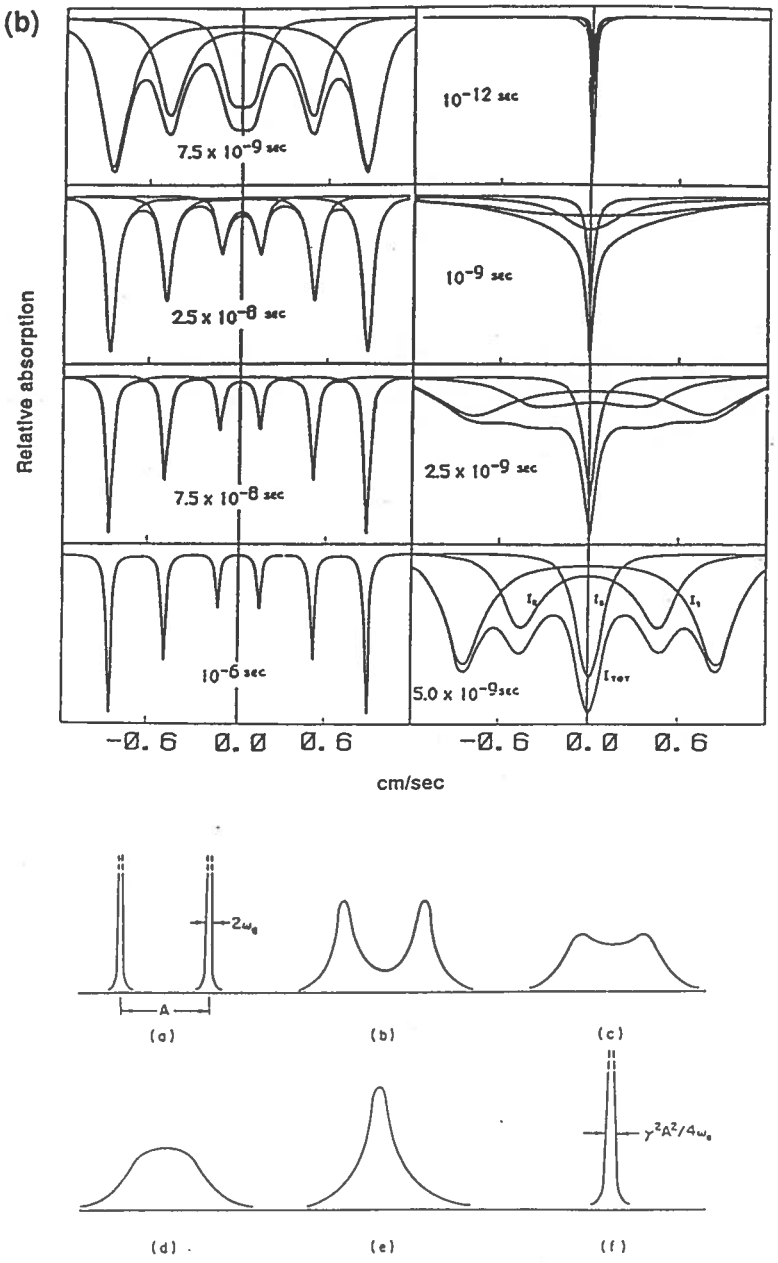

Figure 9. Collapse of a doublet spectrum due to increasing rate of fluctuations in the abcissa parameter: (a) through (f) represent an increase in the rate of 500 fold. Note that for individual doublet lines, the width is proportional to the fluctuation rate, ω_e, but for the averaged singlet (f), the width is inversely proportional to ω_e. The latter effect is sometimes termed "motional narrowing"; after Poole and Farach (1971).

Distribution of physical states. Samples are inherently inhomogeneous due to chemical variations, grain boundaries, dislocations and other defects. Depending on the type of spectroscopy attempted, these features will contribute to spectral line broadening or asymmetry. In general, the broadening is of the Gaussian type, suggesting essentially a random distribution of line-shifting perturbations. However, it is easy to imagine a context in which broadening and asymmetry are correlated. Suppose a spectroscopic measurement reveals a spectral line position sensitive to a metal-oxygen bond distance. If the potential well characterizing this bond has a hard repulsive potential but very soft attractive potential, we would expect random strains to create a non-symmetric distribution of bond distances, and hence an asymmetric line shape. The probability of a transition may also vary with bond distance or site geometry, and possibly in a non-linear fashion. Thus the observed line shape is actually a convolution over all of the "micro" states of the system.

Spectrometer resolution. Even the best spectrometer system has resolution limits. In many cases, these may derive from diffraction criteria, or they may be due to electronic considerations. An example of the former occurs in optical and optical-analog spectrometers, in which the actual spectrum being analyzed is convoluted by various slit functions. If the natural line-width and homogenous broadening effects are small, then the spectrometer output will be spectral lines whose width is determined by the slit functions. If the opposite is the case, the output spectrum will closely resemble the true sample spectrum. An example of electronic resolution limits can be seen in Si(Li) energy dispersive detector spectra, in which the output spectrum has lines hundreds of times broader than the actual X-ray emission spectrum.

Spectrometer aberrations. Most spectrometers introduce variations in line shapes that are asymmetric, in addition to the broadening effects already noted. In modern optical systems, these effects are minimized by a variety of strategies. In X-ray spectrometers, the possibilities for optical element design are severely limited relative to visible and IR optics. Thus aberrations can be more significant in X-ray spectra. The separation of spectrometer aberrations from physical state distributions may be quite difficult.

The manner in which these effects operate on a basic delta function line spectrum to produce the observed spectrum is one example of the convolution process. This and other Fourier integral operations are described in the next section.

Fourier processing of spectral data

Fourier methods are an integral part of the spectroscopic art. The action of any spectrometer system convolutes the initial spectrum with the slit functions and aberration functions of the spectrometer. The output spectrum is then frequently subjected to smoothing procedures which are actually convolution operations. Noise can be removed from the spectrum with autocorrelation or Fourier filtering procedures.

Attempts can also be made to enhance the apparent resolution of a spectrum via deconvolution and maximum entropy Fourier procedures. In the former case, we attempt to correct (i.e., remove) the spectrometer or other convolution effects, and in the latter case, we make approximations to get around the limited data set (much smaller than from negative to positive infinity) observed.

Fourier transforms and integrals

The Fourier transform of the function $f(x)$ is usually defined as

$$F(s) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi xs} dx \quad . \tag{46}$$

Substitution of $F(s)$ for $f(x)$ into the same integral will result in the original function. This is the cyclical nature of the Fourier transform. If $F(s)$ is the Fourier transform of $f(x)$, then $f(x)$ is the Fourier transform of $F(s)$. The key relation between the transform and the original function is the change of variable, from x to s. These variables have reciprocal dimensions. In harmonic analysis and electronics, the two variables are time and frequency; in vibrational systems, distance and wavelength; in diffraction analysis, they can be vectors in real and reciprocal space; in quantum mechanical operations, position and momentum. The usual formulae for Fourier transformation are written slightly differently:

$$F(s) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi xs} dx \quad , \tag{47}$$

$$f(x) = \int_{-\infty}^{\infty} F(s)\, e^{i2\pi xs} ds \quad . \tag{48}$$

These forms differ in the exponent, so that they are not identical and may operate differently on particular functions. To keep them straight, we may refer to the first as a forward transform and the second as a back transform (but the names can be reversed). Combining the two results in the Fourier integral theorem,

$$f(x) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x)\, e^{-i2\pi xs} dx \right] e^{i2\pi xs} ds \quad . \tag{49}$$

For conditions under which this integral can be evaluated, the reader should consult a text on Fourier methods, (e.g., Bracewell, 1986). In particular, discontinous functions create problems in evaluation. Some examples of $f(x)$ and $F(s)$ pairs are shown in Figure 10.

Properties of Fourier transforms

A few of the theorems important to Fourier analysis are stated here without proof; these can readily be found in most references on Fourier transforms.

Similarity theorem. If $f(x)$ has the Fourier transform $F(s)$, then $f(ax)$ has the transform $|a|^{-1} F(s/a)$. This theorem allows scaling of the area under transforms.

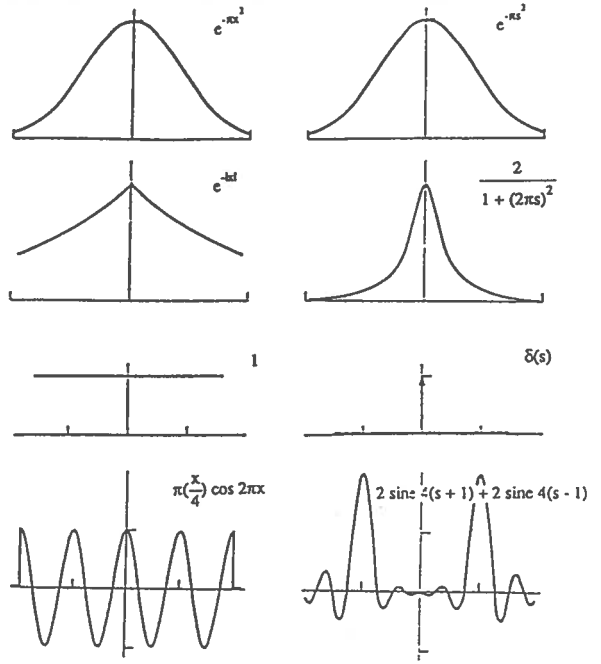

Figure 10. Some examples of Fourier transform pairs. From top to bottom: Gaussian --> Gaussian, exponential --> Lorentzian, constant --> delta function at position s, truncated cosine function --> peaks with finite widths and side lobes at position s; after Bracewell (1986).

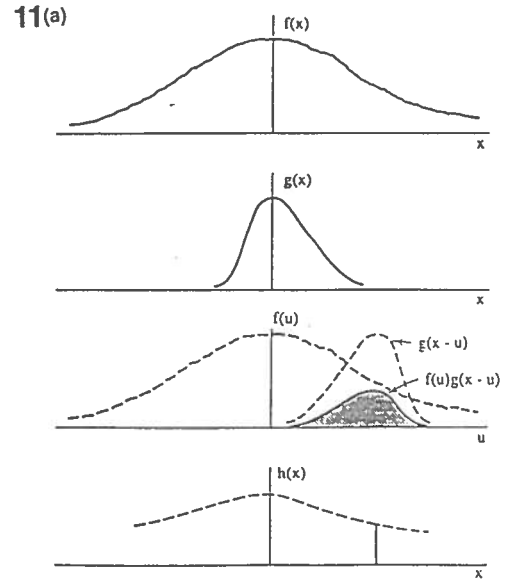

Figure 11. Graphical depiction of convolution: (a) the value of $h(x) = f(x)*g(x)$ calculated at one point from $f(u)g(x-u)$, and the form of the integral $h(x)$; (b) decomposition of small segments of $f(x)$ into the characteristic shape of $g(x)$, then superimposed to obtain part of the convolution integral; after Bracewell (1986).

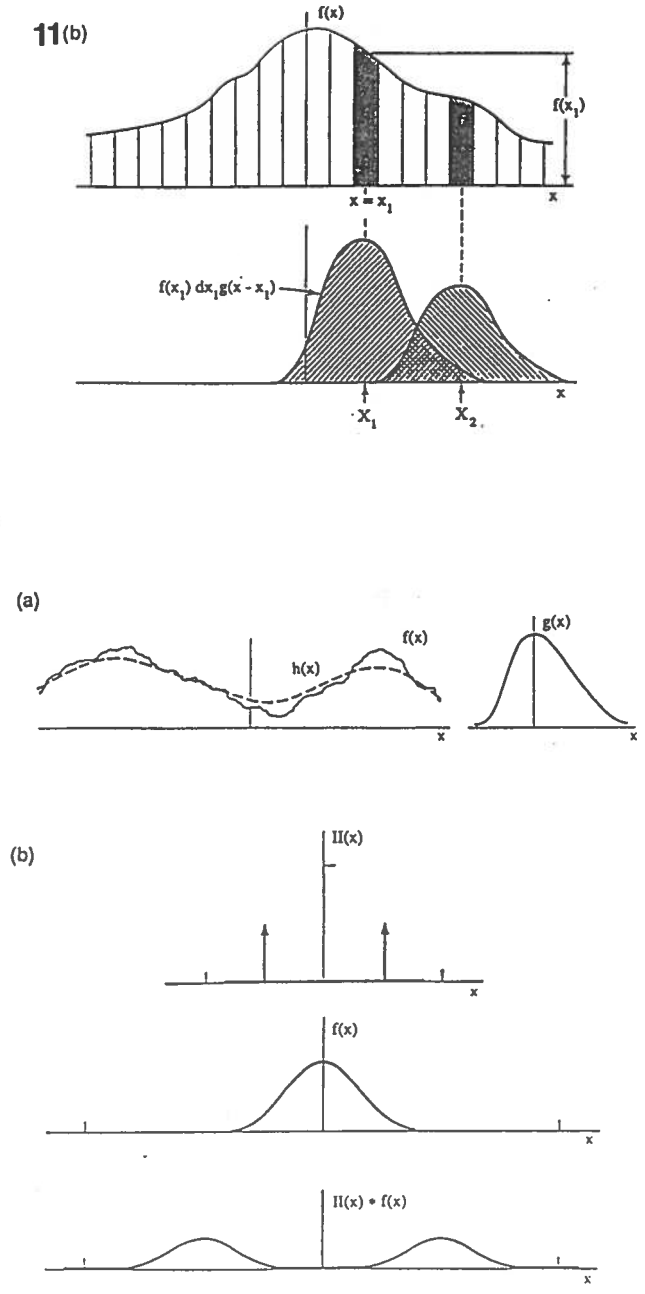

Figure 12. Examples of convolution; (a) smoothing effect of convolution; (b) convolution of delta function spectrum with Gaussian to obtain Gaussian line spectrum, after Bracewell (1986).

Addition theorem. If $f(x)$ and $g(x)$ have the transforms $F(s)$ and $G(s)$, respectively, then $f(x) + g(x)$ has the transform $F(s) + G(s)$.

Shift theorem. If $f(x)$ has the transform $F(s)$, then $f(x-a)$ has the transform $e^{-i2\pi as}F(s)$.

Convolution theorem. If $f(x)$ and $g(x)$ have transforms $F(s)$ and $G(s)$ respectively, then we define the convolution of the functions as $f(x)*g(x)$ having the Fourier transform $F(s)G(s)$. This theorem indicates that we can obtain the convolution of one function with another by first determining their Fourier transforms separately, then multiplying the transforms and back transforming the product. As a corollary, we can imagine a process to deconvolute two convoluted functions by division of the Fourier transform of the convoluted function by the transform of one of the original functions.

Autocorrelation theorem. If $f(x)$ has the Fourier transform $F(s)$ then its autocorrelation function is defined as

$$\int_{-\infty}^{\infty} f^*(u)\, f(u + x)\, du \quad , \tag{50}$$

which has the Fourier Transform $|F(s)|^2$.

Rayleigh's theorem. If $f(x)$ has the Fourier transform $F(s)$, then

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = \int_{-\infty}^{\infty} |F(s)|^2\, ds \quad . \tag{51}$$

This theorem is important, as it implies that least-squares minimization in frequency (or reciprocal) space corresponds to the same type of operation in time (or real) space. Hence we can operate in either variable space, whichever is more convenient.

Convolution-deconvolution. The most common application of the Fourier transform in spectroscopy is in convolution and deconvolution procedures. The convolution of functions $f(x)$ and $g(x)$ is defined as

$$f(x)*g(x) = \int_{-\infty}^{\infty} f(u)\, g(x-u)\, du \quad . \tag{52}$$

Here it is important to realize that the two functions being convoluted are within the integral, so that a simple multiplication of the functions does not reproduce the form of the convolution. Secondly, the second function is reversed and shifted. A graphical interpretation of convolution is depicted in Figure 11a, in which each small area segment of $f(x)$ is replaced by an equal area element having the shape of $g(x)$ and centered on the position of the original segment. The convolution of the two functions, $h(x)$, is then equal to the sum of all such contributions at the given position. Figure 11b gives another interpretation of the convolution, and Figure 12 gives several examples of convolutions.

One well-known type of convolution is a smoothing operation. It consists of replacing each element of the function $f(x)$ with a Gaussian, rectangular (boxcar), or other function of the same area. Various types of polynomial smoothing functions are frequently used to approximate Gaussian smoothing.

Deconvolution procedures are most often used to reduce spectral line width due to spectrometer broadening. However, when line shape analysis is crucial, such as in Mössbauer spectroscopy, the source line shape can be removed by deconvolution (Vincze, 1982; Lin and Preston, 1974). If all broadening functions and the natural line width can be removed from a spectrum, what remains is the distribution of physical parameters in the sample. As an example, the distribution of isomer shift and quadrupole splitting can be obtained in this way for Mössbauer spectra (Window, 1971). Various numerical approximations to the Fourier transform method have also been evaluated (Wivel and Morup, 1981). An example of a deconvoluted Mössbauer spectrum in which sample thickness effects have been removed (Ure and Flinn, 1971) is shown in Figure 13.

Fourier filtering

In most spectra, the background consists of a variety of differing noise signals. The usual noise is due to counting statistics (i.e., noise with a random of Gaussian distribution). Electronic effects may add periodic noise. Incorrect values in computer memory due to data transfer or transcription problems may introduce data spikes. Fourier filtering is a type of deconvolution in which the Fourier transform of the original spectrum is multiplied by a window function. Typically, the window function has non-zero values only in the vicinity of frequency (or reciprocal) space in which the signals of interest occur. The other frequencies are thus set to zero in the product of the transforms. Back Fourier transformation then yields a filtered spectrum; most statistical noise, as it contains a wide distribution of frequencies, will be removed along with any other periodic noise. Spikes in the raw spectrum will also be removed. An example of Fourier filtering is shown in Figure 14.

The advantage of filtering procedures is that often some sense can be made out of spectral data that would otherwise be too noisy for further analysis. An additional advantage, used in the analysis of EXAFS spectra and seismic frequency spectra, is that certain frequency components can be separated for detailed analysis procedures without handling the full spectrum. This is of particular value when one frequency range in the spectrum is subject to less uncertainty than others.

A major disadvantage of Fourier filtering is the inability to use normal goodness-of-fit parameters when least-squares fitting a filtered spectrum. The filtering removes the statistical variations in the spectrum, so that variances become small or zero. The way around this problem is to calculate some estimate of fit-quality, such as a crystallographic R-factor type parameter, in order to compare fitting models. Then these models can be applied to the original unfiltered spectrum with the usual X^2 parameter.

Correlation functions

The autocorrelation function has been defined above. Another type of correlation function is the cross correlation function defined as

$$\int_{-\infty}^{\infty} f^*(u)\, g(u+x)\, du \qquad (53)$$

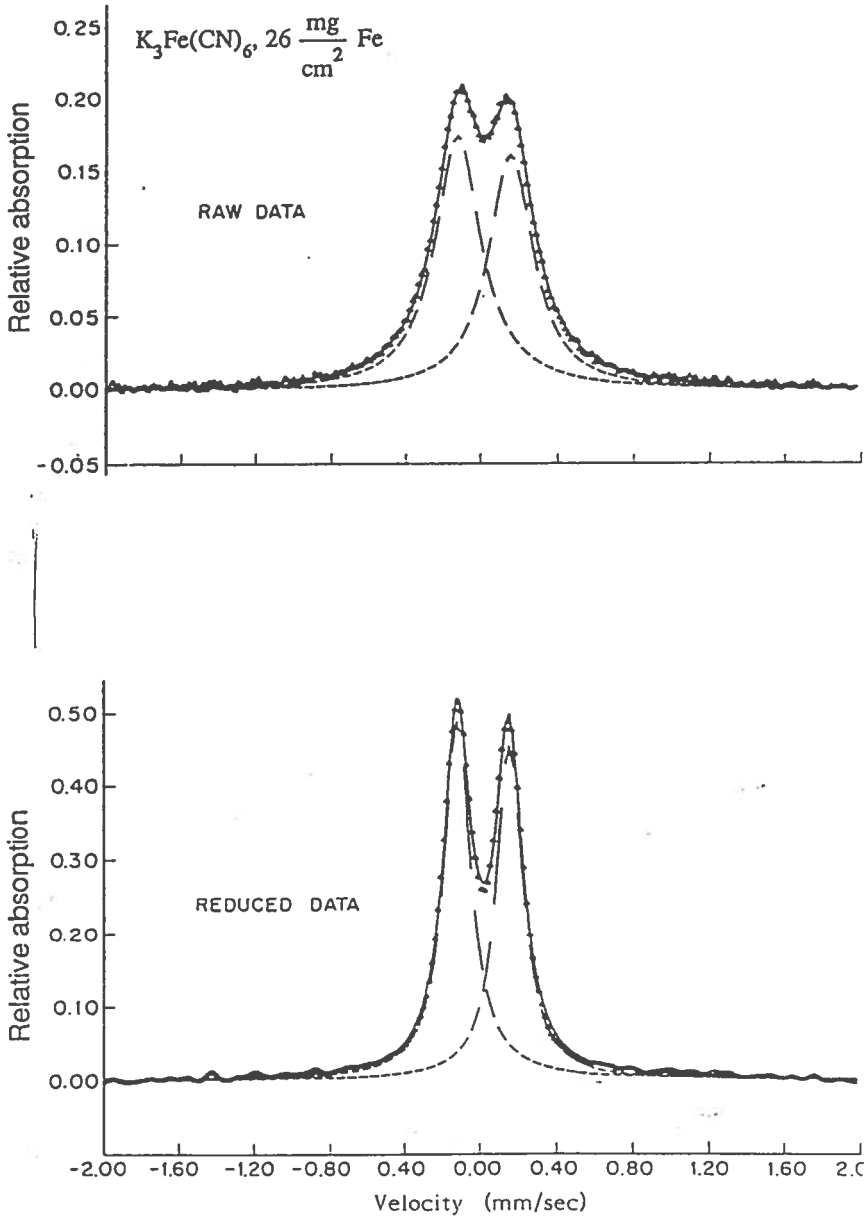

Figure 13. Deconvolution of blackness broadening in the Mössbauer spectrum of potassium ferricyanide. Top: raw data taken with a ^{57}Co:Cr source having a line width of 0.13 mm/s. Overall spectral line width is about 0.32 mm/s. Bottom: spectrum after deconvolution process. Spectral line width is now about 0.18 mm/s; after Ure and Flinn (1971).

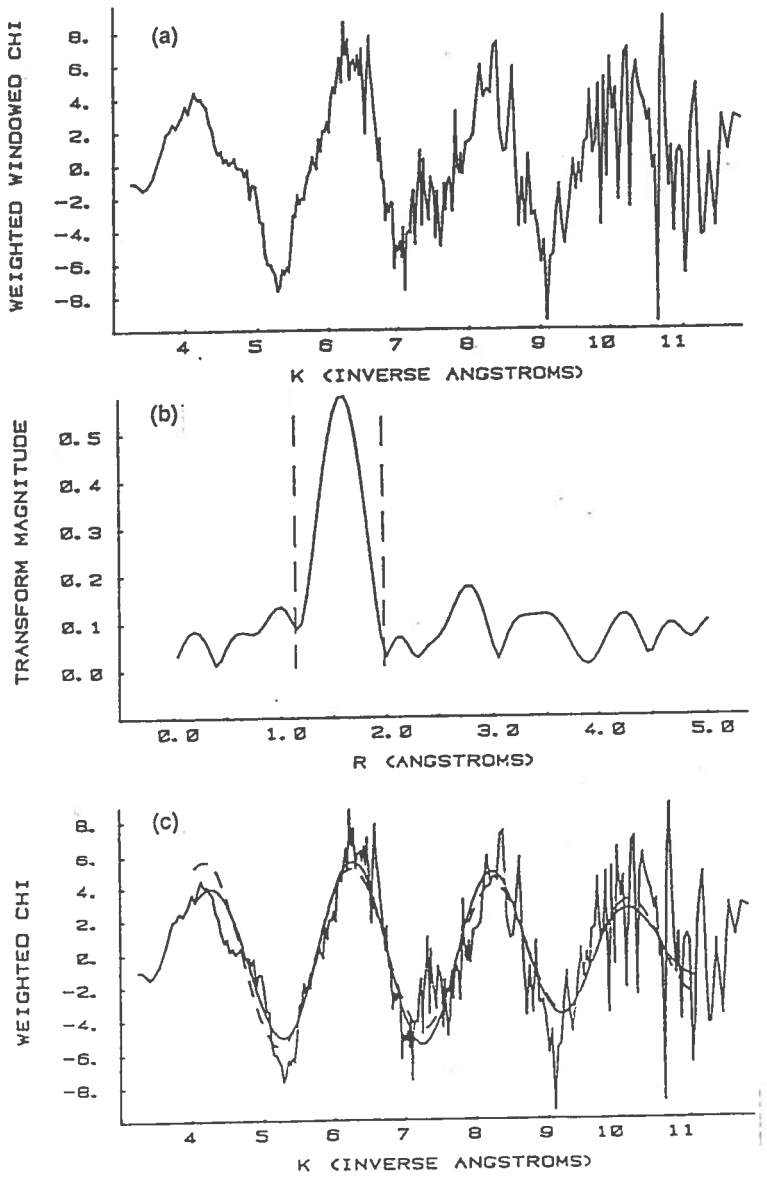

Figure 14. Fourier filtering of EXAFS spectrum; (a) original weighted EXAFS spectrum as as function of k ($Å^{-1}$); (b) Fourier transform of (a) to R space (Å); (c) Back transform of (b) over the range 1.2<R<2.0 Å compared to original data (solid lines). Calculated fit to back transform is given by the dashed line. Note evidence of higher frequency components in the raw spectrum not present in the filtered spectrum, (e.g. oscillations between 4 and 6 $Å^{-1}$ are due to the second small peak in R space at about 2.8 Å). After Waychunas et al. (1986).

These functions are extremely important in the analysis of diffraction patterns. For example, the autocorrelation function of the electronic charge density in a crystal is equivalent to the generalized Patterson function (for a complete derivation, see Cowley, 1981). Note that from the definition of the autocorrelation function, its Fourier transform is the square of the absolute value of the Fourier transform, F(s), of the original function. This means that the autocorrelation transform can carry no information about the phase of F(s). This is one description of the so-called phase problem in X-ray crystallography.

The autocorrelation function is a measure of the correlation between the values of a function evaluated at points differing by x. This aspect can be used to remove noise from periodic functions or spectra. The autocorrelation function of a periodic function with random noise is equivalent to the sum of the separate autocorrelation functions of the function and the noise. The periodic function will give rise to a periodicity in the autocorrelation from which it can be evaluated, but the noise will produce only a smooth background curve.

Fourier "ripples" -- limitations of finite data sets

Fourier transforms are defined over the range $-\infty$ to $+\infty$; as a result, the use of data sets any shorter causes perturbations in the transform. The endpoints of the data set act as rapid impulse functions and generate sinusoidal ripples in the transform. Transforms of single lines in spectra will thus not produce infinite oscillations in s space. Similarly, a short wave train in frequency space will transform to a broad peak with side lobes in x space. Thus for finite data sets in either x or s space, the forward Fourier transform - back Fourier transform process will introduce complicating oscillations. These are often called Fourier "ripples". An example of such ripples can be seen in the EXAFS Fourier transform (called a pair-correlation function) in Figure 14b. The peak at 1.6 Å has side lobes at 1.0 Å and 2.1 Å. These annoying features may interact in phase with side lobes from adjacent peaks in a spectrum, creating strange non-physical beats. Because of this, every effort should be made to extend the range of data in x and s for Fourier operations. The existance of severe ripples for small data sets may preclude use of convolution-deconvolution operations. However, it may be possible to extend the effective data range by making particular assumptions about the uncollected data. One way of doing this is described in the next section.

Maximum entropy methods

A type of spectral analysis much used with geophysical frequency spectra is the maximum entropy method (MEM). The name derives from the fact that by making reasonable assumptions, it is possible to reconstruct some of the frequency space that is not part of the original data set. The new enhanced data set spans a large region in frequency space, and thus its Fourier transform will have narrower lines and generally enhanced resolution over an original spectrum.

The assumptions in the MEM are mainly that in any limited data set, the unmeasured frequency components are unlikely to be all zero, or all have perfect periodicity to infinity. The actual case lies

somewhere in between, and can be estimated by generating a spectral distribution function with the maximum entropy. This function must be constrained to agree with that part of the spectral distribution function which can be measured. The MEM analysis has been applied to various forms of spectra with very short data ranges, with mixed results. Misapplication of the technique can result in spectra with high resolution but spurious features. A good review of progress in MEM with geophysical applications is that of Ulrych and Bishop (1975).

REFERENCES

Bevington, P.R. (1969) Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, New York, 336 p.

Bracewell, R. (1986) The Fourier Transform and its Applications. Second revised edition. McGraw-Hill, New York, 474 p.

Champeney, D.C. (1973) Fourier Transforms and their Physical Applications. Academic Press, London, 256 p.

Cowley, J. (1981) Diffraction Physics. Elsevier (North-Holland), Amsterdam, 430 p.

Hamilton, W.C. (1964) Statistics in Physical Science. Ronald Press, New York, 230 p.

Lin, T.M. and Preston, R.S. (1974) Comparison of techniques for folding and unfolding Mössbauer spectra and data analysis. In Gruverman, I.J., Seidel, C.W. and Dieterov, D.K., eds., Mössbauer Effect Methodology 9, 205-233.

Poole, C.P. and Farach, H.A. (1971) Relaxation in Magnetic Resonance. Academic Press, New York, 392 p.

Prince, E. (1982) Mathematical Techniques in Crystallography and Materials Science. Springer-Verlag, New York, 192 p.

Ruby, S.L. (1982) Why MISFIT when you already have X^2? In Gruverman, I.J. and Seidel C.W., eds., Mössbauer Effect Methodology 8, 263-276.

Ulrych, T.J. and Bishop, T.N. (1975) Maximum entropy spectral analysis and autoregressive decomposition. Rev. Geophys. Space Phys. 13, 183-200.

Ure, M.C.D. and Flinn, P.A. (1971) A technique for removal of the "blackness" distortion of Mössbauer spectra. In Gruverman, I.J., ed., Mösbauer Effect Methodology 7, 245-262.

Vincze, I. (1982) Fourier evaluation of broad Mössbauer spectra. Nucl. Spectra. Nucl. Instr. Methods 199, 247-262.

Waychunas, G.A. (1986) Performance and use of Mössbauer goodness-of-fit parameters to spectra of varying signal/noise ratio and possible misinterpretations. Am. Mineral. 71, 1261-1265.

-----, Brown, G.E.,Jr. and Apted, M.J. (1986) X-ray K-edge absorption spectra of Fe minerals and model compounds: II. EXAFS. Phys. Chem. Minerals 13, 31-47.

Wertheim, G.K., Guggenheim, H.J. and Buchanan, D.N.E. (1958) Sublattice magnetization in FeF_3 near the critical point. Phys. Rev. 169, 465-470.

Wickman, H.H. (1966) Mössbauer paramagnetic hyperfine structure. In Gruverman, I.J., ed., Mössbauer Effect Methodology 2, 39-66.

Window, B. (1971) Hyperfine field distributions from Mössbauer spectra. J. Phys. E: Sci. Instrum. 4, 401-402.

Wivel, C. and Morup, S. (1981) Improved computational procedure for evaluation of overlapping hyperfine parameter distributions in Mössbauer spectra. J. Phys. E: Sci. Instrum. 14, 605-610.

APPENDIX A: SOME STATISTICAL DEFINITIONS

ACCURACY: a measure of how close the experimental result is to the "true" value.

PRECISION: the measure of how exactly the result is determined (i.e reproducibility) without any reference to a "true" value.

RANDOM ERROR: indefiniteness of result due to finite precision of the experiment.

SYSTEMATIC ERROR: reproducible inaccuracy caused by faulty experimental technique or a faulty model.

PARENT POPULATION: hypothetical infinite set of "data" points of which the experimental data points are assumed to be a random sample.

PARENT DISTRIBUTION: probability distribution controlling the (random) sample data assumed to be drawn from the parent population.

EXPECTATION VALUE: denoted by $\langle\ \rangle$, it is the weighted average of a function f(x) over all values of x:

$$\langle f(x)\rangle = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N} f(x_i)\right] = \sum_{j=1}^{n}\left[f(x_i)P(x_j)\right] = \int_{-\infty}^{\infty} f(x)P(x)dx \quad , \quad A(1)$$

where P(x) is the probability function that defines the probability of obtaining a specific value of x in any random experiment.

MEAN VALUE: for a series of N observations, the sample mean value is the average of the observations, x; for the parent distribution (see above), the parent mean value, μ, is the limit of x as N -> ∞; thus

$$\mu \approx \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad . \qquad A(2)$$

MEDIAN VALUE: for the parent population, the median $\mu_{\frac{1}{2}}$ is that value of xi for which the probability of any observation being less than the median is equal to the probability of it being greater than the median.

$$P\{x_i \le \mu_{\frac{1}{2}}\} = P\{x_i \ge \mu_{\frac{1}{2}}\} = \frac{1}{2} \qquad A(3)$$

MOST PROBABLE VALUE: for the parent population, the most probable value μ_{max} is that value of x for which the parent distribution has its greatest value:

$$P\left(\mu_{max}\right) = P\left(x = \mu_{max}\right) \qquad A(4)$$

Note that for a symmetrical parent distribution function, all of these parent values are coincident.

AVERAGE DEVIATION: this is defined as the average of the magnitudes of the deviations from the mean of the parent distribution (cf., equation A(2)):

$$\alpha \equiv \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} |x_i-\mu| \right] \quad . \qquad A(5)$$

about the mean value.

VARIANCE: like the average deviation, the variance, σ^2, is a measure of the dispersion of the observations, defined as

$$\sigma^2 \equiv \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} (x_i-\mu)^2 \right] = \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right] - \mu^2 \quad . \qquad A(6)$$

This is a very convenient measure (more so than the average deviation) as the expression $(x_i-\mu)^2$ occurs in several distribution functions (e.g., Gaussian function). We can rewrite it as

$$\sigma^2 = \langle (x_i - \mu)^2 \rangle = \langle x^2 \rangle - \mu \quad . \qquad A(7)$$

For a finite set of observations, the sample variance is defined as

$$\sigma^2 \approx \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \quad . \qquad A(8)$$

This is normally modified (Bevington, 1969, p. 19) to the form

$$\sigma^2 \approx s^2 \equiv \frac{1}{N-1} \sum_{x=1}^{N} (x_i - \bar{x})^2 \quad , \qquad A(9)$$

where s^2 is called the sample variance.

COVARIANCE: this is defined by analogy with variance as

$$\sigma_{ij}^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left[(x_i - \bar{x})(y_i - \bar{y}) \right] \quad , \qquad A(10)$$

where x and y are different variables. If the deviations in x and y are random, then $\sigma_{ij}^2 = 0$. However, if the deviations are correlated, then $\sigma_{ij}^2 \neq 0$ and is a measure of the degree of correlation between x and y.

STANDARD DEVIATION: the standard deviation is defined as the square root of the variance

$$\sigma = \sqrt{\sigma^2} \quad . \qquad A(11)$$

Thus it is the root-mean-square of the deviations, and is a measure of the uncertainty of a result assuming random error only.

The standard deviation is a very important quantity because it allows us to perform hypothesis tests to determine the significance (or otherwise) of a result.

PROPAGATION OF ERROR: let $x_i (i=1,n)$ be a set of experimental results, each with an associated variance σ_i^2, and let y be some parameter that is related to the experimental results by the function

$$y = f(x_1, x_2 \ldots x_n) \quad . \qquad A(12)$$

When calculating y, we must assign a standard deviation or we cannot assess the significance of the result. If σ_y^2 is the variance of y,

$$\sigma_y^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \quad , \qquad A(13)$$

where σ_{ii}^2 and σ_{jj}^2 are the variances of x_i and x_j, and σ_{ij}^2 is the covariance of x_i and x_j.