Lawrence Berkeley National Laboratory

Title
Construction of Mathematical Software Part IV: Nonlinear Least Squares Codes

Permalink
https://escholarship.org/uc/item/0fv2v6nx

Author
Johnston, Brad

Publication Date
1973

DISCLAIMER

LAWRENCE LIVERMORE LABORATORY

University of California/Livermore, California

CONSTRUCTION OF MATHEMATICAL SOFTWARE

PART IV:

NONLINEAR LEAST SQUARES CODES

Brad Johnston

January 1973

Foreword

This is Part IV of the five-part report <u>Construction of Mathematical</u> <u>Software</u>.

The following outline of the complete report lists the topics covered in each of the five parts:

Contents of Part IV

Chapter 4 (Continued)

CONSTRUCTION OF MATHEMATICAL SOFTWARE
PART IV:
NONLINEAR LEAST SQUARES CODES

Introduction

The present report is mostly a commentary on three nonlinear least squares codes taken from a book by Bevington.[1] Bevington's book has four nonlinear least squares codes; but one of them (CHIFIT) was left out of this report because, so far, I have been able neither to understand it completely nor to make it run on the computer.

To really understand Bevington's book, one must have some background in statistics. Here I try to make it possible for one to use and understand the codes without such a need; one will end up knowing the meaning of the terms, but the interpretation of the real value of the results of particular least squares problems may still require the help of a statistician.

Chapter 1. Some Elementary Statistics

1. Basic Definitions

Consider N elementary events that make up all the possible outcomes of some experiment. (The word "elementary" is used here to mean that these events are mutually exclusive and equally likely.) Let some higher-level event A be defined by the occurrence of any one of some subset of N_A events out of the set of N elementary events. The probability $P(A)$ of A is then defined to be N_A/N. A set of all mutually exclusive outcomes of an experiment is called a sample space.

Consider two events A and B. The event "A and B" written $A \cap B$, means that both events A and B occur. We will show that

$$P(A \cap B) = P(A) \, P_A(B), \tag{1.1}$$

where $P_A(B)$ means the probability that B happens, given that A happened. Let N_A, N_B, and N_{AB} be the number of elementary events that constitute A, B, and A ∩ B, respectively. Then

$$P(A \cap B) = \frac{N_{AB}}{N} = \frac{N_A}{N} \frac{N_{AB}}{N_A} .$$

By definition, $N_{AB}/N = P(A \cap B)$. The meaning of $P_A(B)$ is that the sample space has been narrowed down to the elementary events in A. Therefore $P_A(B) = N_{AB}/N_A$, and the result (1.1) follows. The events A and B are said to be independent if the fact that A happened does not affect P(B). In terms of numbers of elementary events, the independence of A and B means that

$$\frac{N_{AB}}{N_A} = \frac{N_B}{N} , \qquad \text{or} \qquad P(B) = P_A(B) ,$$

so that

$$P(A \cap B) = P(A) \, P(B). \tag{1.2}$$

Equation (1.2) can obviously be extended to any number of independent events. One must not confuse mutually exclusive events with independent events. If A and B are mutually exclusive events, then $P(A \cap B) = 0$.

Any set of N observations y_i is a finite sampling of a possibly infinite set of observations. In terms of the infinite set of observations, one can define various statistical quantities, such as the mean and the variance. The mean μ is the limit of the average values of the y_i's. It is defined as

$$\mu = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} y_i .$$

For any real problem we have only a finite number N of observations, and we define the sample mean $\bar{y}$ as

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i .$$ (1.3)

We say that the sample mean is the best estimate E of the mean, and we write $E(\mu) = \bar{y}$. The deviation d_i of the observation y_i can be defined with respect to the mean or with respect to the sample mean. Usually, we take the deviation to be $d_i = y_i - \bar{y}$.

The variance of the observations has to do with the spread of the observations about the mean. Many sets of observations could have the same mean. Some could all be clustered very close to the mean, while others are spread out quite far from the mean. The more spread-out the observations, the less confidence one has in the sample mean being a good estimate of the true mean. The variance σ_y^2 is defined as

$$\sigma_y^2 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 .$$

The standard deviation σ_y of the observations is defined as the square root of the variance. As we will see later in Eq. (1.23), the uncertainty in determining the mean is proportional to the standard deviation.

For a finite number of observations we must define the sample variance. To begin with, if we knew the mean, we could define the sample variance as

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 .$$ (1.4)

But we do not know the mean μ; we know only the sample mean $\bar{y}$. The first thing that comes to mind is to replace μ in Eq. (1.4) by $\bar{y}$. However, we must look a little deeper to get the correct formula.

-3-

It turns out that, for N observations, y_i, the sum $S \equiv \sum_{i=1}^{N} (y_i - \xi)^2$ is minimized when $\xi = \bar{y}$. This can easily be seen by calculating $dS/d\xi$, setting it equal to zero, and solving the resulting equation for ξ. Therefore, we would expect that replacing μ in Eq. (1.4) by $\bar{y}$ would lead to an underestimate of the variance. To make up for this sum being too small, we must also reduce N in the factor $1/N$. This is done by replacing N by the number of "independent" observations (called the number of degrees of freedom), that is, the number of observations in excess of those required to determine an estimate for the unknown quantity μ. At least one observation must be made to get this estimate. Therefore, we replace N in Eq. (1.4) by N-1 and define the sample variance σ_y^2 by

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2. \tag{1.5}$$

To make this formula more plausible, consider what would happen if we had just one observation and used Eq. (1.4) with μ replaced by $\bar{y}$. We would take $\bar{y}$ to be equal to this observation, and get a zero deviation, which would give zero as the estimate of the variance; this would, in general, be a very poor estimate indeed. However, by having the factor $1/(N-1)$ in front of Eq. (1.5), the formula becomes indeterminate, as it should be in this case. As N becomes larger, $\bar{y} \to \mu$, and the sample variance approaches the variance.

Many observed physical quantities are more or less symmetrically distributed about their mean. Next we consider an important class of distributions of observations that are symmetrical. Some of the observations are too large and some are too small; but, for a large enough set of observations, there are about the same number of small ones as large ones.

2. The Gaussian Error Distribution

Suppose Δ is the observation spacing, in the sense that a plot made to the following prescription has a "reasonable" shape. There is a total of N observations.

N_1 of the y_i's are in the interval $(\overline{y}-\Delta, \overline{y}+\Delta)$

$$
\left\{
\begin{array}{l}
\dfrac{N_2}{2} \quad " \quad " \quad " \quad " \quad (\overline{y}-2\Delta, \overline{y}-\Delta) \\[2ex]
\dfrac{N_2}{2} \quad " \quad " \quad " \quad " \quad (\overline{y}+\Delta, \overline{y}+2\Delta)
\end{array}
\right.
$$

$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$

$$
\left\{
\begin{array}{l}
\dfrac{N_i}{2} \quad " \quad " \quad " \quad " \quad (\overline{y}-i\Delta, \overline{y}-(i-1)\Delta) \\[2ex]
\dfrac{N_i}{2} \quad " \quad " \quad " \quad " \quad (\overline{y}+(i-1)\Delta, \overline{y}+i\Delta)
\end{array}
\right.
$$

$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$

Then make a plot (histogram) as in Fig. 1.1, where

$$h_i = \frac{N_i}{2\Delta N} . \tag{1.6}$$

From the definition of N_i, it is clear that $\Sigma N_i = N$, where the sum is taken over the number of rectangles on either side of the mean. The area under the histogram is equal to unity, since

$$2\Delta \ \Sigma h_i = 1. \tag{1.7}$$

In the limit, as $N \to \infty$ and $\Delta \to 0$, a smooth bell-shaped curve results, as shown in Fig. 1.2. If there are numbers μ and σ such that the equation of the above curve is given by

$$h(y) = \frac{1}{\sigma\sqrt{2\pi}} \ e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \tag{1.8}$$

then the experimental points are said to follow the <u>Gaussian error distribution</u>.
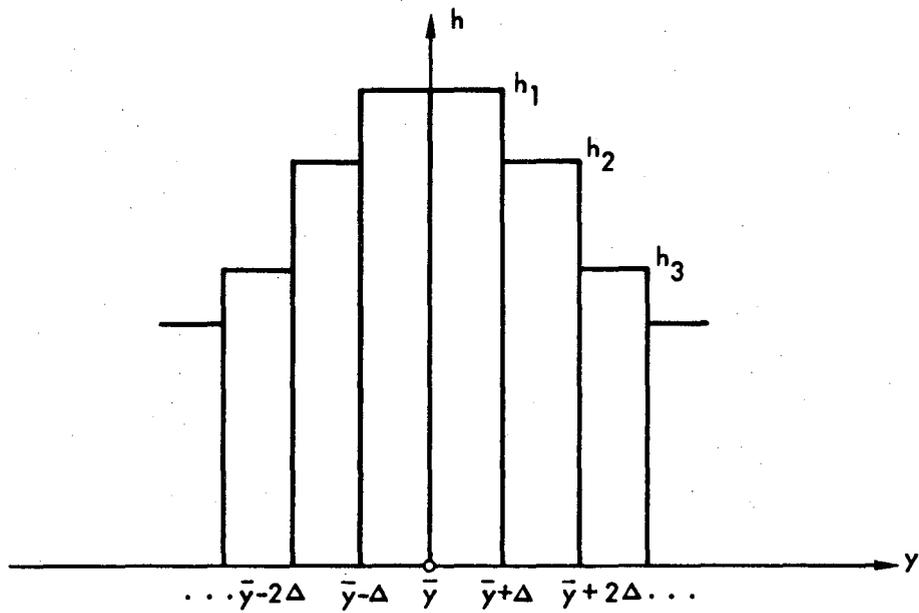
Fig. 1.1 Histogram of experimental observations.
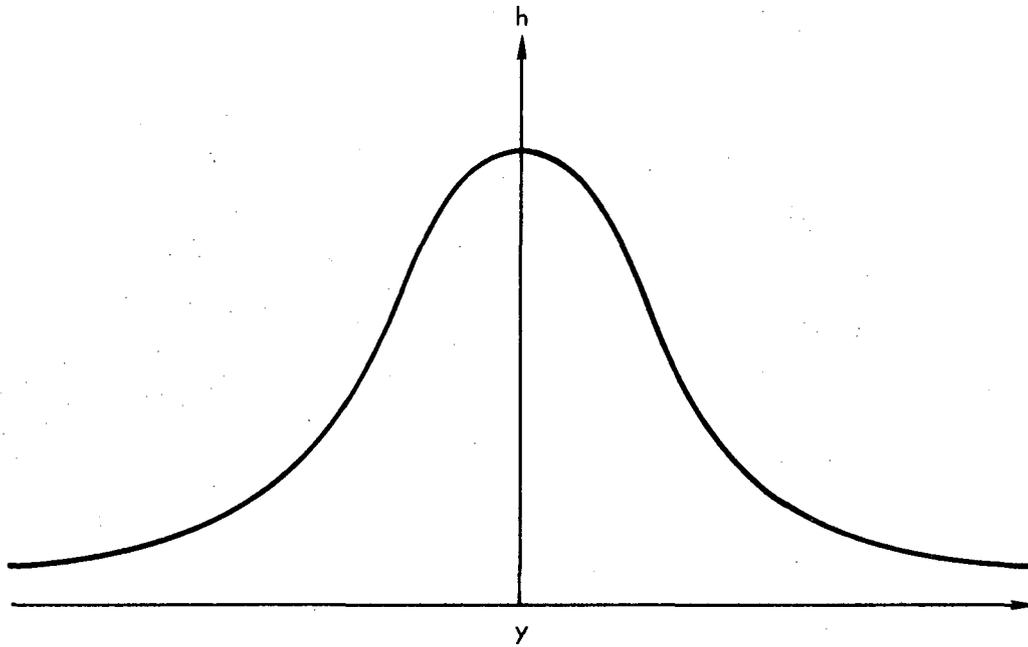
Fig. 1.2 Gaussian error distribution.

Return to the definition of the mean given by Eq. (1.3). Suppose the set of N observed values y_i contains only $\tilde{N} < N$ distinct y_i's. Let these distinct values be denoted by $\tilde{y}_j$, and let $\tilde{N}_j$ of the y_i's be equal to $\tilde{y}_j$. Then the mean $\bar{y}$ could be written

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{\tilde{N}} \tilde{y}_j \tilde{N}_j = \sum_{j=1}^{\tilde{N}} \left(\frac{\tilde{N}_i}{N}\right) \tilde{y}_j = \sum_{j=1}^{\tilde{N}} P(\tilde{y}_j) \tilde{y}_j, \qquad (1.9)$$

where the sums are taken over all the distinct values $\tilde{y}_j$, and $P(\tilde{y}_j) = \frac{\tilde{N}_j}{N}$ is an estimate for the probability that $\tilde{y}_j$ will be observed.

In the same way, the variance σ_y^2, defined by Eq. (1.4), can be written (we must here use μ in place of $\bar{y}$)

$$\sigma_y^2 = \sum_{j=1}^{\tilde{N}} (\tilde{y}_j - \mu)^2 \, P(\tilde{y}_j). \qquad (1.10)$$

If we had available to us an infinite number of observations of y, then these sums would be replaced by integrals, and our sample estimates of statistical quantities would be replaced by their true values. In this case, Eqs. (1.9) and (1.10) become

$$\bar{y} = \int_{-\infty}^{\infty} y \, f(y) \, dy \qquad (1.11)$$

$$\sigma_y^2 = \int_{-\infty}^{\infty} (y - \mu)^2 \, f(y) \, dy. \qquad (1.12)$$

In these formulas, $f(y) \, dy$ is the probability that y will have a value between y and $y + dy$, and $f(y)$ is called the <u>density function</u>.

From the way $h(y)$, given by Eq. (1.8), was constructed, it follows that, if the experimental points follow the Gaussian error distribution, then

$$f(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}. \qquad (1.13)$$

It turns out, as one would expect, that, for the Gaussian error distribution, the mean and variance are the μ and σ, respectively, that appear in Eq. (1.13). To prove this, one has only to put Eq. (1.13) into Eqs. (1.11) and (1.12) and perform the indicated integrations.

There are many other distribution functions, but for least squares theory, the Gaussian is the most important.

Suppose y is an observation from a "parent" (or "true") distribution that is Gaussian with mean μ and standard deviation σ. From Eq. (1.13) the probability P that the observed value of y will be between $\bar{y}$ - d and $\bar{y}$ + d, for any fixed d > 0, is given by

$$P = \int_{\mu-d}^{\mu+d} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = \frac{1}{\sigma\sqrt{2\pi}} \int_{-d}^{d} e^{-\frac{1}{2}\left(\frac{\xi}{\sigma}\right)^2} d\xi.$$

We have used the obvious change of variable $\xi = y - \mu$.

Therefore, the probability of y falling within one standard deviation of μ [this means y is the interval (μ-σ, μ+σ)] is given by

$$P = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sigma}^{\sigma} e^{-\frac{1}{2}\left(\frac{\xi}{d}\right)^2} d\xi = \frac{2}{\sigma\sqrt{2\pi}} \int_{0}^{\sigma} e^{-\frac{1}{2}\left(\frac{\xi}{d}\right)^2} d\xi.$$

If we make the change of variable $t = \frac{1}{\sqrt{2}} \frac{\xi}{d}$, we obtain

$$P = \frac{2}{\sqrt{\pi}} \int_{0}^{1/\sqrt{2}} e^{-t^2} dt \equiv erf\left(\frac{1}{\sqrt{2}}\right)^* \approx 0.68$$

3. Many Observations of One Thing vs One Observation of Many Things

Up to this point, we have been considering many observations of one thing. For example, one could make many measurements of the length

---

*See Ref. 2, p. 297.

of a desk top. Then one could calculate the mean by using Eq. (1.3) and the variance by using Eq. (1.5). Now we consider a different problem, where one variable depends on another. Suppose, for example, we measure some time-dependent quantity. We look at the clock, and read a dial on some instrument. Assume the time measurements are exact, but there are statistical fluctuations of the dial readings of the instrument. We end up with a set of pairs of quantities $\left(t_i, f(t_i)\right)$, where t_i is the time of the ith observation, and $f(t_i)$ is the corresponding dial reading. Because there are statistical fluctuations in the dial readings, we must have some measure of the fluctuations. For this we can use the variance of $f(t_i)$. But now we have a problem because, to compute the sample variance $\sigma^2_{f(t_i)}$ by using Eq. (1.5), we must have many dial readings for each t_i, not just one. Therefore, one must use some estimate for these variances, based on the known accuracy of the instrument, etc. The lack of good variances is one of the inherent weaknesses of least squares theory. Another problem with least squares is that the entire structure depends on the dial readings having a Gaussian distribution; and, more often than not, one does not know the true distribution. The method of least squares is often used, despite these shortcomings, to fit functions to observations and the value of the results must be judged with these limitations in mind.

4. The Propagation of Errors

Suppose the dependent variable y is a function of M independent variables $u_1, u_2, \ldots, u_M$, so that (for the ith observation)

$$y_i = y(u_1^i, u_2^i, \ldots u_M^i).$$ (1.14)

If $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_M$ are the means of the independent variables, then the most probable value of y is given by

-10-

$$\bar{y} = y(\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_M). \tag{1.15}$$

If more and more observations $u_1^i, u_2^i, \ldots, u_M^i$ were made, giving many different values of y_i, the mean of these y_i's would approach the mean given by Eq. (1.15).

The sample variance σ_y^2 of y can be found by using Eq. (1.5) where N sets of data, $u_1^i, u_2^i, \ldots, u_M^i$, $i = 1, 2, \ldots, N$, were taken. The deviations $y_i - \bar{y}$ can be expressed in terms of the deviations $u_1^i - \bar{u}_1$, $u_2^i - \bar{u}_2, \ldots,$ using the first part of Taylor's series:

$$y_i - \bar{y} \approx \sum_{k=1}^{M} \frac{\partial y}{\partial u_k}\bigg|_{u_1, u_2, \ldots} (u_k^i - \bar{u}_k). \tag{1.16}$$

Putting Eq. (1.16) into Eq. (1.5) gives (see Appendix A)

$$\sigma_y^2 \approx \sum_{k=1}^{M} \sum_{\ell=1}^{M} \left[\frac{1}{N-1} \sum_{i=1}^{N} (u_k^i - \bar{u}k)(u_\ell^i - \bar{u}_\ell) \right] \left(\frac{\partial y}{\partial u_k} \right) \left(\frac{\partial y}{\partial u_\ell} \right). \tag{1.17}$$

If $k \neq \ell$, we define the <u>sample covariance</u> $\sigma_{u_k u_\ell}^2$ between the variables u_k and u_ℓ as

$$\sigma_{u_k u_\ell}^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (u_k^i - \bar{u}_k)(u_\ell^i - \bar{u}_\ell) \tag{1.18}$$

When $\ell = k$, the variance $\sigma_{u_k}^2$ is given by $\sigma_{u_k u_k}^2$. Therefore, Eq. (1.17) can be written

$$\sigma_y^2 \approx \sum_{k=1}^{M} \sum_{\ell=1}^{M} \sigma_{u_k u_\ell} \frac{\partial y}{\partial u_k} \frac{\partial y}{\partial u_\ell}. \tag{1.19}$$

The covariance $\sigma_{u_k u_\ell}^2$ is the average of cross terms involving products of deviations of u_k and u_ℓ simultaneously. If the fluctuations in u_k and u_ℓ are uncorrelated, there would be about as many negative terms in Eq. (1.18) as positive ones, and we would expect the covariance to be small in comparison with the variances $\sigma_{u_k}^2$ and $\sigma_{u_\ell}^2$. This assumption is often

made when dealing with the independent variables of a least squares problem.

Therefore, if the variables u_1, u_2,... are uncorrelated, then

$$\sigma_y^2 \approx \sum_{k=1}^{M} \sigma_{u_k} \left(\frac{\partial y}{\partial u_k} \right)^2 . \qquad (1.20)$$

5. The Mean and Its Estimated Error

If the observed y_i's follow the Gaussian error distribution and each y_i has associated with it a standard deviation σ_i, one can show, using the method of maximum likelihood (see Ref. 1, pp. 69-70), that the mean $\bar{y}$ is given by

$$\bar{y} = \frac{\displaystyle\sum_{i=1}^{N} \left(\frac{y_i}{\sigma_i^2} \right)}{\displaystyle\sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2} \right)} . \qquad (1.21)$$

By using Eq. (1.20) we can find an expression for the variance of the mean. Since $\bar{y}$ is a function of the observed y_i's, Eq. (1.20) gives

$$\sigma_{\bar{y}}^2 \approx \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial \bar{y}}{\partial y_i} \right)^2 . \qquad (1.22)$$

If the uncertainties of the data points are equal, so that $\sigma_i = \sigma$, then by using Eq. (1.3) one gets

$$\sigma_{\bar{y}}^2 \approx \frac{1}{N} \sigma^2 . \qquad (1.23)$$

If the uncertainties are not equal, one uses Eq. (1.21), and the result is

$$\sigma_{\bar{y}}^2 \approx \frac{1}{\displaystyle\sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2} \right)} . \qquad (1.24)$$

Chapter 2. The Least Squares Problem

1. The Principle of Least Squares

Suppose we have made N observations y_i, each corresponding to some variable x_i. The x_i's are assumed to be exact, and let σ_i be the standard deviations of y_i. Let $y = f(x, \vec{a})$ be a mathematical model for a physical experiment, where $\vec{a}$ is a vector of parameters we wish to find. Assume that the observations follow the Gaussian error distribution. This means that if we took an infinite number of observations y_{ij}, $j = 1, 2, \ldots$ for each x_i, instead of just one, then these y_{ij}'s would be symmetrically scattered about their mean $\bar{y}_i$ and their histogram would become a bell-shaped curve as in Fig. 1.2. Suppose now that $\vec{a}$ is the parameter vector corresponding to these $\bar{y}_i$'s. This vector is called <u>parent</u> parameter vector. But in fact we have made only one observation y_i for each x_i, and we want to find a parameter vector $\vec{a}$ so that $f(x_i, \vec{a})$ will be the best estimate for the mean $\bar{y}_i$. The tool we use to accomplish this is the <u>method of maximum likelihood</u>, which is the assumption that the observations y_i are better approximated by $f(x_i, \vec{a})$, where $\vec{a}$ is the parent parameter vector, than by any other parameter vector. To use this idea, we calculate the probability of getting our actual observations y_i and then maximize this probability with respect to the parameter vector.

From Eq. (1.13), the probability P_i for getting the observed value y_i is given by[*]

$$P_i = K \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{y_i - f(x_i, \vec{a})}{\sigma_i}\right]^2}, \tag{2.1}$$

where we have replaced the mean $\bar{y}$ by $f(x_i, \vec{a})$, and K is the appropriate constant of proportionality. We assume that all our observed y_i's are

---

[*]Actually the probability of getting a y_i between y_i and $y_i + dy_i$.

independent, so that, by extending Eq. (1.2) to any number of independent events, the probability of getting the complete set of y_i's is equal to the product of the individual probabilities. We therefore get

$$P(\vec{a}) = K \prod_{j=1}^{N} \left(\frac{1}{\sigma_j \sqrt{2\pi}}\right) e^{-\frac{1}{2} \sum_{i=1}^{N} \left(\frac{y_i - f(x_i, \vec{a})}{\sigma_i}\right)^2}, \qquad (2.2)$$

where we have written $P(\vec{a})$ to indicate the dependence of the probability on the parameter vector, and the K in Eq. (2.2) is not the same as in Eq. (2.1). Using the method of maximum likelihood, we maximize $P(\vec{a})$ with respect to $\vec{a}$. The first term in Eq. (2.2) is a known constant. Therefore, to maximize $P(\vec{a})$, we minimize the sum in the exponential. We define this quantity (chi-squared) χ^2, as follows:

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} [y_i - f(x_i, \vec{a})]^2. \qquad (2.3)$$

We have shown therefore that, if the observed data y_i follows the Gaussian error distribution, then the best parameter vector $\vec{a}$ is the one that minimizes Eq. (2.3). This is called the principle of least squares.

2. Some Statements About the Results of a Least Squares Calculation

Given a function $f(x, \vec{a})$ where $\vec{a}$ is a vector of parameters, N data points (x_i, y_i), and standard deviations σ_i of the y_i's, we have seen that the least squares problem is to find $\vec{a}$ that minimizes χ^2, given by Eq. (2.3). If $\vec{a}$ has n parameters, then the number of data points (observations) minus the number of parameters, N - n, is called the number of degrees of freedom. It can be shown that under certain circumstances the expected value of χ^2, $E(\chi^2)$, is equal to the number of degrees of freedom.

$$E(\chi^2) = N - n \qquad (2.4)$$

(see Ref. 3, Section 12). However, we can make this seem plausible by noting that, on the average, $|y_i - f(x_i, \vec{a})| \approx \sigma_i$, so that $\sum_{i=1}^{N} \frac{1}{\sigma_i^2}[y_i - f(x_i \vec{a})]^2 \approx N$, rather than $N - n$. Suppose, however, that there were just as many parameters as observations. Then we would expect χ^2 to be very small. This is another way of saying that one can usually fit a function of n parameters through n points. This is one argument for putting $N - n$ in place of N in Eq. (2.4). If N > n, which is the usual case, then n of the observations would be taken up in determining the n parameters, and the remaining $N - n$ parameters would be available to allow some flexibility in this determination, in order to minimize χ^2. Thus, statistically one has only $N - n$ parameters to work with. This suggests the introduction of the reduced χ_r^2, defined as $\chi^2/(N-n)$. If the observations are Gaussian, and the standard deviations σ_i of the y_i correspond to experimental reality, then we would expect that, for a good fit, $\chi_r^2 \approx 1$.

Suppose, however, that for some least squares problem χ_r^2 is either much larger than 1 or much smaller than 1. What are we to conclude then?

Roughly, the implications are as follows.

(1) χ_r^2 too large:

(a) The model was wrong, which means that the functional relation $y = f(x, \vec{a})$ was not correct.

(b) The model was correct, but the estimates of the standard deviations were too small. This means that we overestimated the accuracy of the experiment.

(2) χ_r^2 too small:

(a) There were too few data points for the number of parameters to be found.

(b) The estimates of the standard deviations were too large. This means that we underestimated the accuracy of the experiment.

3. The Difference Between Linear and Nonlinear Least Squares Problems

If the function $f(x,\vec{a})$ is of the form $f(x,\vec{a}) = \sum_{i=1}^{n} a_i F_i(x)$, then we have a linear least squares problem. If $f(x,\vec{a})$ is not of this form, we have a nonlinear least squares problem. Some nonlinear least squares problems can be changed into linear ones, for example, by taking the logarithm of both sides of $y = f(x,\vec{a})$. The difficulty with doing this is that the resulting least squares problem is really a different problem, in the sense that the weighting is changed.

4. The Solution of the Linear Least Squares Problem

Given N data points (x_i, y_i), standard deviations σ_i of the y_i's, and the fitting function

$$f(x,\vec{a}) = \sum_{j=1}^{n} a_j F_j(x), \tag{2.5}$$

(the x_i's are assumed to be exact), we must find $\vec{a}$ so that

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} [y_i - f(x_i,\vec{a})]^2 \tag{2.6}$$

is minimized.

Minimizing χ^2 with respect to $\vec{a}$ requires $\partial \chi^2 / \partial a_j = 0$, for $j = 1, 2, \ldots, n$. Doing this gives the normal equation

$$\vec{\beta} = \alpha \vec{a}, \tag{2.7}$$

where $\vec{\beta}$ is the vector with

$$\beta_k = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} y_i F_k(x_i), \tag{2.8}$$

and α is the symmetric matrix with

$$\alpha_{jk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \; F_j(x_i) F_k(x_i). \tag{2.9}$$

One can also show that

$$\alpha_{jk} = \frac{1}{2} \; \frac{\partial^2 \chi^2}{\partial a_j \, \partial a_k} \; .^* \tag{2.10}$$

The matrix α, given by Eq. (2.10), is called the curvature matrix. The solution of Eq. (2.7) is

$$\vec{a} = \alpha^{-1} \vec{\beta}. \tag{2.11}$$

If we define the matrix ϵ to be the inverse of the matrix α, then Eq. (2.11) becomes

$$\vec{a} = \epsilon \vec{\beta}. \tag{2.12}$$

From Eqs. (2.12) and (2.8) one gets

$$a_j = \sum_{k=1}^{n} \epsilon_{jk} \beta_k = \sum_{k=1}^{n} \left\{ \epsilon_{jk} \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2} \; y_i F_k(x_i) \right] \right\}. \tag{2.13}$$

5. Weights and Errors in Linear Least Squares Problems

One can weight the data in many different ways for a least squares problem. By this we mean that χ^2 could be defined as

$$\chi^2 = \sum_{i=1}^{N} w_i [y_i - f(x_i, \vec{a})]^2, \tag{2.14}$$

where w_i is a given positive quantity and is called the _weight_ of y_i. We now distinguish three types of weights as follows.

Case I. $w_i = 1/\sigma_i^2$, where σ_i^2 is the variance of y_i.

---

*Note that, without the factor $\frac{1}{2}$, α would be the Hessian matrix. See Ref. 4, p. 39.

Case II. The weights w_i are chosen to be certain values for reasons
peculiar to some problem.

Case III. The weights are unknown, so one sets $w_i = 1$.

We now derive formulas for the statistical quantities appropriate
for these three cases, and it will make the discussion easier if we intro-
duce some additional notation. There are n parameters $\vec{a}$; they are dis-
tinguished by subscripts a_i. Different experimental determinations of
them are distinguished by superscripts. The same goes for the y's: each
x_i, assumed to be exact, corresponds to a y_i, and m^* different experimental
determinations of y_i will be denoted by y_i^j, $j = 1, 2, \ldots, m$.

If there were m experimental determinations of $\vec{a}$, then from Eq.
(1.18) the sample covariance $\sigma^2_{a_j a_k}$ between a_j and a_k is given by

$$\sigma^2_{a_j a_k} = \frac{1}{m-1} \sum_{i=1}^{m} \left(a_j^i - \bar{a}_j \right) \left(a_k^i - \bar{a}_k \right),$$ (2.15)

where $\bar{a}_j$ and $\bar{a}_k$ are the mean values of these parameters. The errors in
the parameters come from errors in the y's. Expanding in a Taylor
series, as we did in Eq. (1.16), we get

$$a_j^i - \bar{a}_j \approx \sum_{\ell=1}^{N} \frac{\partial a_j}{\partial y_\ell} \bigg|_{\bar{y}_1, \bar{y}_2, \ldots} (y_\ell^i - \bar{y}_\ell).$$ (2.16)

The notation on the right side of Eq. (2.16) means that the partial
derivatives are to be evaluated at the mean values of the y's. One gets
a better understanding of the notation above by knowing that, as the
number of experimental determinations of the y's gets larger and larger,

$$\bar{y}_i \rightarrow f(x_i, \bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n).$$ (2.17)

---

*One should get clearly in mind that there are
   N data points, (x_i, y_i)
   n parameters, and
   m different determinations of the N data points.
Therefore, m least squares problems could be solved.

Putting Eq. (2.16) into Eq. (2.15), one gets (see Appendix A)

$$\sigma^2_{a_j a_k} = \sum_{s=1}^{N} \sum_{\ell=1}^{N} \left[\frac{1}{m-1} \sum_{i=1}^{m} (y^i_\ell - \bar{y}_\ell)(y^i_s - \bar{y}_s) \right] \frac{\partial a_j}{\partial y_\ell} \frac{\partial a_k}{\partial y_s}$$

$$= \sum_{s=1}^{N} \sum_{\ell=1}^{N} \sigma^2_{\ell s} \frac{\partial a_j}{\partial y_\ell} \frac{\partial a_k}{\partial y_s} , \qquad (2.18)$$

where $\sigma^2_{\ell s}$ is the covariance between y_ℓ and y_s. If we assume that the fluctuations in y_ℓ and y_s are uncorrelated, then

$$\sigma^2_{\ell s} = \begin{cases} \sigma^2_{\ell \ell} = \sigma^2_\ell, & \text{the variance of } y_\ell, \text{ when } \ell = s \\ \\ 0, & \text{when } \ell \neq s. \end{cases}$$

Then Eq. (2.18) becomes

$$\sigma^2_{a_j a_k} \approx \sum_{\ell=1}^{N} \sigma^2_\ell \frac{\partial a_j}{\partial y_\ell} \frac{\partial a_k}{\partial y_\ell} . \qquad (2.19)$$

When we set $j = k$ in Eq. (2.19), we get the variance of a_j, given by

$$\sigma^2_{a_j} = \sigma^2_{a_j a_j} \approx \sum_{\ell=1}^{N} \sigma^2_\ell \left(\frac{\partial a_j}{\partial y_\ell} \right)^2 . \qquad (2.20)$$

Case I

The weights for Case I are given by $w_i = 1/\sigma^2_i$. From Eq. (2.13),

$$\frac{\partial a_j}{\partial y_\ell} = \sum_{s=1}^{n} \epsilon_{js} \frac{1}{\sigma^2_\ell} F_s(x_\ell), \qquad (2.21)$$

so that Eq. (2.19) becomes

$$\sigma^2_{a_j a_k} \approx \sum_{\ell=1}^{N} \sigma^2_\ell \left(\sum_{s=1}^{n} \epsilon_{js} \frac{1}{\sigma^2_\ell} F_s(x_\ell) \right) \left(\sum_{p=1}^{n} \epsilon_{kp} \frac{1}{\sigma^2_\ell} F_p(x_\ell) \right)$$

$$= \sum_{s=1}^{n} \sum_{p=1}^{n} \left(\sum_{\ell=1}^{N} \frac{1}{\sigma^2_\ell} F_s(x_\ell) F_p(x_\ell) \right) \epsilon_{js} \epsilon_{kp} .$$

-19-

From Eq. (2.9), this becomes

$$\sigma^2_{a_j a_k} \approx \sum_{s=1}^{n} \sum_{p=1}^{n} \alpha_{sp} \epsilon_{js} \epsilon_{kp} = \sum_{p=1}^{n} \left(\sum_{s=1}^{n} \epsilon_{js} \alpha_{sp} \right) \epsilon_{kp} = \sum_{p=1}^{n} \delta_{jp} \epsilon_{kp},$$

$$= \epsilon_{kj} = \epsilon_{jk}, \qquad (2.22)$$

where δ_{jp} is the Kronecker delta function. (We have used the fact that the inverse of a symmetric matrix is also symmetric.)

When $j = k$, one gets the variance of a^j,

$$\sigma^2_{a_j} = \sigma^2_{a_j a_j} \approx \epsilon_{jj}. \qquad (2.23)$$

Case II

In case II, the w_i's are given positive numbers, χ^2 is given by Eq. (2.14), and $f(x,a)$ is given by Eq. (2.5). The normal equation is $\vec{\beta} = \alpha \vec{a}$, where

$$\beta_k = \sum_{i=1}^{N} w_i y_i \, F_k(x_i) \qquad (2.24)$$

$$\alpha_{jk} = \sum_{i=1}^{N} w_i \, F_j(x_i) F_k(x_i). \qquad (2.25)$$

If ϵ is the inverse of the α matrix, then the solution to the normal equation is given by

$$a_j = \sum_{k=1}^{n} \epsilon_{jk} \sum_{i=1}^{N} w_i y_i \, F_k(x_i). \qquad (2.26)$$

From Eq. (2.26), one can get an expression for $\partial a_j / \partial y_k$, which is

$$\frac{\partial a_j}{\partial y_k} = \sum_{p=1}^{n} \epsilon_{jp} \omega_k \, F_p(x_k). \qquad (2.27)$$

-20-

We must put Eq. (2.27) into Eq. (2.19) to get an expression for $\sigma^2_{a_j a_k}$, but Eq. (2.19) requires the variances σ^2_i, of the y_i's, which we do not have. To make the necessary approximations for these variances, we will consider how the weights should be chosen.

For any number of reasons, weights are assigned certain values in least squares problems; but, if one wishes any statistical statements about the results, he should pick the weights so that the absolute values of the terms in the sum [Eq. (2.14)] are all more or less the same. Therefore one must pick the weights w_i so that, on the average,

$$\sqrt{w_i} \, |y_i - f(x_i, \vec{a})| \approx 1. \tag{2.28}$$

Define the weighted difference e_i as

$$e_i = \sqrt{w_i} \left(y_i - f(x_i, \vec{a}) \right). \tag{2.29}$$

Then Eq. (2.14) becomes $\chi^2 = \sum_{i=1}^{N} e_i^2$. From Eq. (2.29),

$$y_i - f(x_i, \vec{a}) = \frac{e_i}{\sqrt{w_i}}, \tag{2.30}$$

so that $E(y_i) = f(x_i, \vec{a})$ implies

$$E(e_i) = 0. \tag{2.31}$$

When the superscript notation, with m experimental determinations, is used, the variance of y_i, which we write as $\text{var}(y_i)$, is given by

$$\text{var}(y_i) = \lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \left(y_i^j - f(x_i \vec{a}) \right)^2 = \lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \left(\frac{e_i^j}{\sqrt{w_i}} \right)^2$$

$$= \frac{1}{w_i} \lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} \left(e_i^j - E(e_i) \right)^2 = \frac{1}{w_i} \operatorname{var}(e_i) . \qquad (2.32)$$

Note that we have used Eq. (2.31). We do not know the variances of the e_i's, so we use Eq. (2.28) and assume they are all equal to the variance of some single quantity e, and we get

$$\operatorname{var}(e_i) \approx \operatorname{var}(e) \approx \frac{1}{N-n} \sum_{i=1}^{N} [e_i - E(e_i)]^2 = \frac{1}{N-n} \sum_{i=1}^{N} e_i^2$$

$$= \frac{1}{N-n} \chi^2 \equiv s^2, \overset{*}{} \qquad (2.33)$$

where χ^2 is given by Eq. (2.14), and s^2 is defined by Eq. (2.33). From Eq. (2.32), one gets

$$\widehat{\operatorname{var}(y_i)} = \frac{1}{w_i} \widehat{\operatorname{var}(e_i)}, \qquad (2.34)$$

where the "hats" mean "an approximation to." Then, from Eqs. (2.33) and (2.34), one gets

$$\widehat{\sigma_i^2} = \widehat{\operatorname{var}(y_i)} = \frac{1}{w_i} s^2 . \qquad (2.35)$$

Returning to Eq. (2.27), we put this expression into Eq. (2.19) and replace σ_ℓ^2 in the latter formula by $\widehat{\sigma_\ell^2}$ given by Eq. (2.35):

$$\sigma_{a_j a_k}^2 \approx s^2 \sum_{\ell=1}^{N} \frac{1}{w_\ell} \left[\left(\sum_{p=1}^{n} \epsilon_{jp} w_\ell F_p(x_\ell) \right) \left(\sum_{q=1}^{n} \epsilon_{kq} w_\ell F_q(x_\ell) \right) \right]$$

$$= s^2 \sum_{p=1}^{n} \sum_{q=1}^{n} \left[\sum_{\ell=1}^{N} w_\ell F_p(x_\ell) F_q(x_\ell) \right] \epsilon_{jp} \epsilon_{kq} .$$

---

*In this formula, there are N-n degrees of freedom. It is similar to Eq. (1.5), where there were N-1 degrees of freedom. One should notice, however, that Eq. (1.5) comes from many observations of one thing, but in Eq. (2.33) there is one observation of many things. The complete justification of Eq. (2.33) requires more statistical background than is assumed in this report.

Using Eq. (2.25) and the fact that ε is the inverse of α, one gets

$$\sigma^2_{a_j a_k} \approx s^2 \varepsilon_{jk}. \tag{2.36}$$

Putting $j = k$, we get the variance of a_j

$$\sigma^2_{a_j} \approx s^2 \varepsilon_{jj}. \tag{2.37}$$

Case III

In Case III, $w_i = 1$ for all i, and the normal equation is $\vec{\beta} = \alpha \vec{a}$, where

$$\beta_k = \sum_{i=1}^{N} y_i F_k(x_i), \tag{2.38}$$

$$\alpha_{jk} = \sum_{i=1}^{N} F_j(x_i) F_k(x_i). \tag{2.39}$$

If ε is the inverse of the α matrix, then the parameter vector $\vec{a}$ is given by

$$a_j = \sum_{k=1}^{n} \varepsilon_{jk} \left(\sum_{i=1}^{N} y_i F_k(x_i) \right). \tag{2.40}$$

In this case, we do not have variances of the observed data, so that, following Case II, so set $\sigma_i^2 = \sigma^2 \approx s^2$, where s^2 is given by Eq. (2.33), with the w_i's $= 1$.

Using Eq. (2.19), replacing σ_i^2 by s^2, and using Eq. (2.40) to calculate the partial derivatives, we get

$$\sigma^2_{a_j a_k} \approx s^2 \varepsilon_{jk}, \tag{2.41}$$

-23-

so that the variance of a_j is

$$\sigma_{a_j}^2 = \sigma_{a_j a_j}^2 \approx s^2 \epsilon_{jj}. \tag{2.42}$$

[Note that Eqs. (2.37) and (2.42) are the same as in Ref. 5, p. 18, Eq. (5)].

6. The Variance of the Observations

In Cases II and III, the variances of the y's were not given, but they were approximated by $\sigma_i^2 = (1/w_i)s^2$ in Case II and $\sigma_i^2 = s^2$ (with the w_i's = 1) in Case III. Using these results, we got approximations for the covariances of the parameters. One can go back and get another approximation for the variance of the y_i's. Thinking of the y's as being functions of the a's, one uses Eq. (1.19) and gets

$$\sigma_i^2 = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \sigma_{a_k a_\ell}^2 \frac{\partial[f(x_i,\vec{a})]}{\partial a_k} \frac{\partial[f(x_i,\vec{a})]}{\partial a_\ell}. \tag{2.43}$$

For Cases II and III, one uses Eq. (2.41) for $\sigma_{a_k a_j}^2$ in formula (2.43), and the result is

$$\sigma_i^2 \approx s^2 \sum_{k=1}^{n} \sum_{\ell=1}^{n} \epsilon_{k\ell} \frac{\partial[f(x_i,\vec{a})]}{\partial a_k} \frac{\partial[f(x_i,\vec{a})]}{\partial a_\ell}.$$

From Eq. (2.5), we find that $\partial[f(x_i,\vec{a})]/\partial a_k = F_k(x_i)$, so that

$$\sigma_i^2 \approx s^2 \sum_{k=1}^{n} \sum_{\ell=1}^{n} \epsilon_{k\ell} F_k(x_i) F_\ell(x_i). \tag{2.44}$$

This is the same formula that is given by Varljen [Ref. 5. p. 18, Eq. (7)].

Chapter 3. The Univariate Algorithm*

1. Introduction

The univariate method of finding the least squares solution for the parameter vector $\vec{a}$ consists of changing one parameter at a time. There are many variations of the univariate method, depending on how the parameter changes are made. The algorithm described below takes a minimizing step in the sense that χ^2 is minimized with respect to each parameter separately.

2. The Basic Procedure

(1) First guesses for the parameters a_j and the absolute values of increments Δa_j are provided by the user. One parameter a_j at a time is incremented by an amount Δa_j, where the sign of Δa_j is chosen by the code so that χ^2 decreases.

(2) The parameter a_j is repeatedly incremented by Δa_j until χ^2 starts to increase.

(3) After the relative sizes of the three χ^2's corresponding to the a_j's are as in Fig. 3.1, one fits a parabola through the three points labeled 1, 2, and 3 on the curve. Point 3 must be higher than point 2, but it does not matter whether it is higher or lower than point 1. We then find the a_j corresponding to the minimum of the parabola. (The code does not test to see if the value of χ^2 at the parabolic minimum is less than at point 2, but this should be done.)

(4) χ^2 is minimized for each parameter in turn, until all the parameters have been modified once. In two dimensions (that is, for two parameters), this corresponds to one zig-zag. Control is then returned

---

*Bevington[1] calls this algorithm the grid search, and the name of his code for this method is GRIDLS.
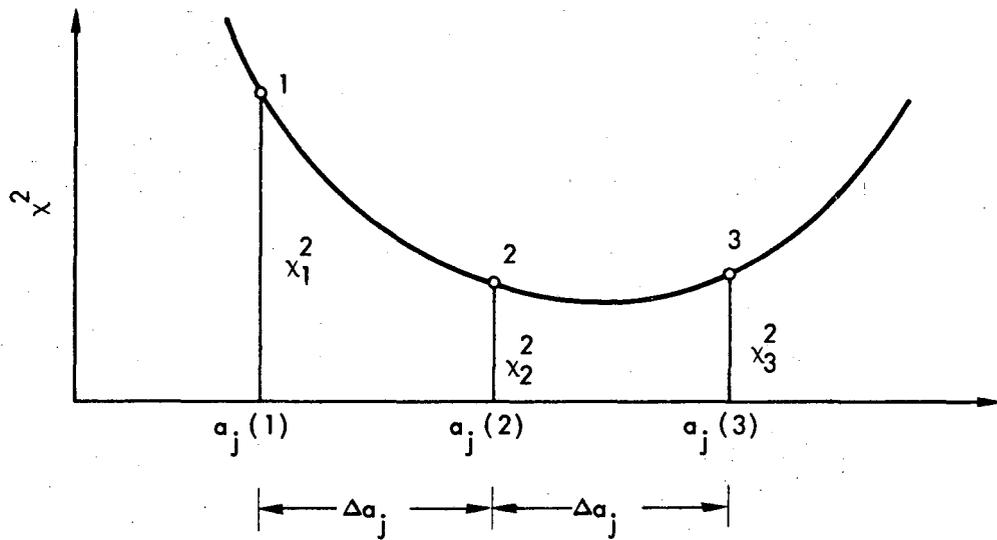
Fig. 3.1. χ^2 as a function of a_j. The remaining parameters are fixed.

to the main code, which must call GRIDLS as many times as needed for convergence.

A property of any algorithm that takes minimizing steps is that the minimum points occur where the X^2 contours are tangent to the direction of search. This is true regardless of the dimensionality of the space and the direction of search. For a proof of this, see Ref. 4, Section 2.7. This method is quite dependent on scale changes and rotation of axis. Consider Fig. 3.2, where contour curves for X^2 are shown for a least squares problem in two dimensions.

If the univariate search were started at point A, where neither of the directions of search is "downhill," the algorithm could not get started. If the axis were rotated through an angle α, then the minimum point M would be reached in one step. If the axis were not rotated, but the scale changed so that the contours were shortened in the direction along the ridge, as shown in Fig. 3.3, then the univariate method could get started. This is because the X^2 contour through point A in Fig. 3.2 lies within a right angle, and the contour in Fig. 3.3 goes outside a right angle. There-fore, a good univariate algorithm should have a scale-change option. The code described below does not have this option.

The univariate method would seem to be appropriate when the variation of X^2 with respect to each parameter were more or less independent of the values of the other parameters. This is almost never the case, but this independence is assumed to get a formula for the standard deviations of the parameters. Without doubt, this code usually gives a poor estimate of these standard deviations. The grid search method usually converges very slowly toward the minimum, but the simplicity of the calculations often compensates for this inefficiency.

3. Powell's Modification of the Univariate Search[6]

Before going on with the detailed discussion of the flow chart for Bevington's code GRIDLS, it is appropriate to point out that better methods
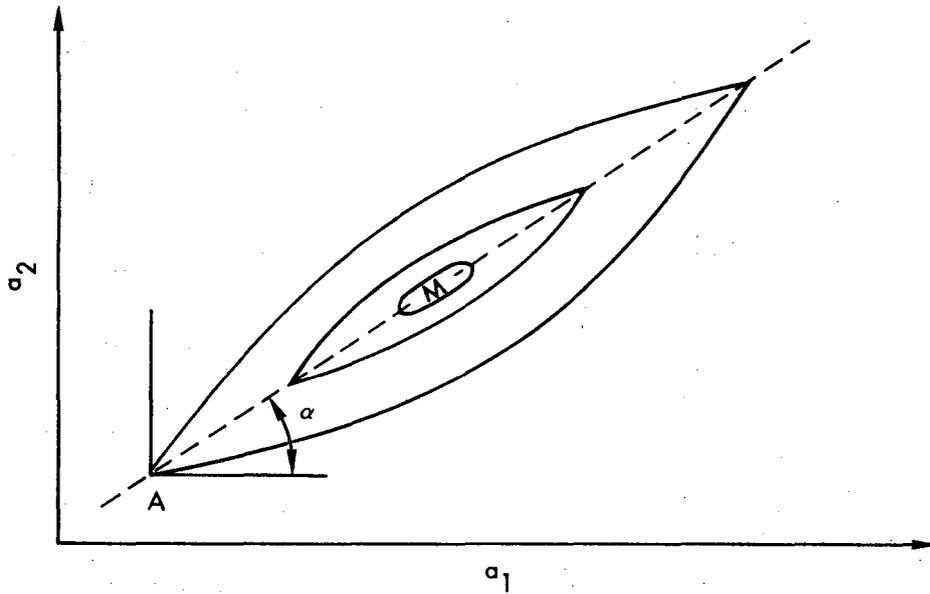
Fig. 3.2. Contour curves for χ^2 in two dimensions. The χ^2 surface has a ridge along the dotted line, where the direction of the contour curve is discontinuous. Point A is the starting point, and M is the minimum point we wish to find.
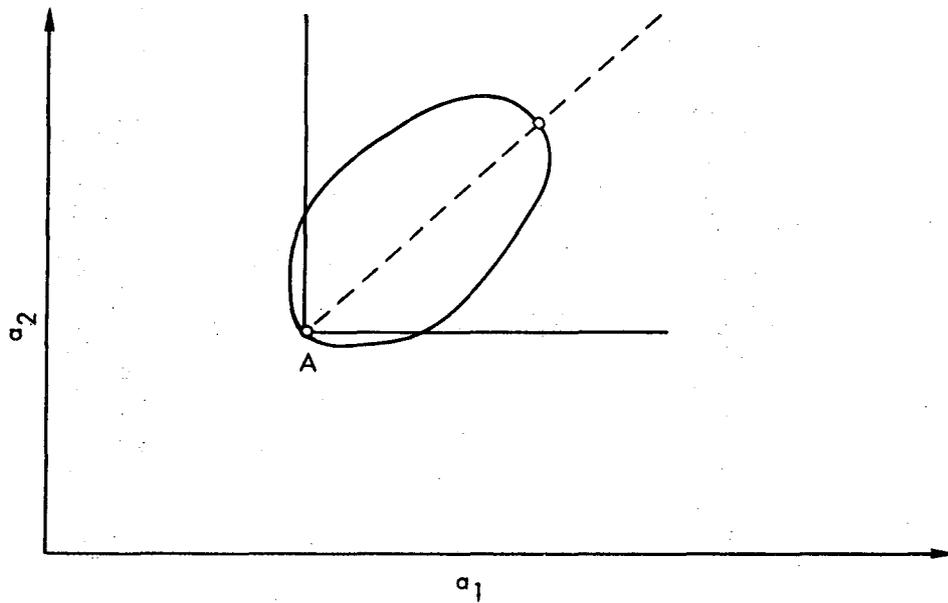

Fig. 3.3. The effect of a change of scale.

for solving least squares problems, when derivatives are not used, are extensions of the simple univariate method. The simple search described above often becomes bogged down, but it reveals a pattern of directions that point toward the minimum. For illustrations of this phenomenon, see Ref. 4, Section 2.5. One can move in the direction of this pattern and speed up the minimization process. Powell's method uses conjugate directions (Ref. 4, Section 2.6), the nature of which will not be defined here. By using conjugate directions one ensures that the minimum of a quadratic function will be reached in a predetermined number of steps, where this number is related to the number of parameters. Quadratic functions are important in minimization methods, not so much because they frequently occur, but because many functions are closely approximated by a quadratic near their minima.

The first iteration of Powell's algorithm is the same as the first iteration of the univariate algorithm (one iteration being to search on each parameter just once). In the parameter space, each parameter can be thought of as an independent direction. The direction of the net result (on all parameters) of the first iteration is made one of the new independent directions for the next iteration. The direction that is thrown away, in order to add the new direction, is the (old) direction that caused the largest decrease in χ^2. Since this is usually the major component of the new direction being added, doing this retards the build-up of linear dependence among the parameter directions.

4. Usage of Subroutine GRIDLS

The calling sequence for GRIDLS is as follows:

CALL GRIDLS (X,Y,SIGMAY,NPTS,NTERMS,MODE,A,DELTAA,SIGMAA,YFIT,CHISQR)

It should be emphasized that only one step is done per call to the subroutine. (Each parameter is modified just once.)

(1) Input Variables

X: array of independent variables x_i.

-29-

Y: **array** of dependent variables y_i.

SIGMAY: array of standard deviations of the y_i's.

NPTS: number of pairs of data points.

NTERMS: number of parameters.

MODE: determines the method of weighting the least squares fit.

$$
\left.\begin{array}{r} +1 \\ 0 \\ -1 \end{array}\right\} \rightarrow \left\{\begin{array}{l} w(i) = 1./SIGMAY(i)**2 \\ w(i) = 1 \\ w(i) = 1/y(i)^{*} \end{array}\right\} \quad \begin{array}{l} \text{for all} \\ i = 1, NPTS. \end{array}
$$

A: array of parameters. (Must start with approximate values.)

DELTAA: array of parameter increments.

(2) Output Variables

A: array of final values of parameters.

SIGMAA: array of standard deviations of parameters.

YFIT: array of calculated values of the dependent variables.

[YFIT(i) = $f(x_i, \vec{a})$, where $\vec{a}$ is the final parameter vector.]

CHISQR: χ^2_r for the fit.

5. The GRIDLS Flow Chart (Fig. 3.4)

The variable NFREE = N-n, which is the number of degrees of freedom in the fit, must be greater than zero. The user must supply a function subprogram FUNCTN(X,I,A) that calculates $f(x_i, \vec{a})$. The value of χ^2 is evaluated with the function subprogram FCHISQ (see Ref. 1, p. 194). FCHISQ calculates

$$
\chi^2_r = \frac{1}{N-n} \sum_{i=1}^{N} w(i) \left[y_i - f(x_i, \vec{a}) \right]^2 .
$$

[Note that $f(x_i, \vec{a})$, rather than x_i and $\vec{a}$, is input to FCHISQ.]

---

*There are physical processes where the number of successes is much smaller than the possible number of events. An example is the counting of radioactive decay rates. For these problems, one uses the Poisson distribution (see Ref. 1, p. 36) rather than the Gaussian distribution, and it is often true that the uncertainty in each measurement is related to the magnitude of the thing observed. But just how are we to relate σ_i and y_i? The answer is that the Poisson distribution automatically gives the standard deviation; it is given by $\sigma_i = \sqrt{y_i}$ (see Ref. 1, p. 109).
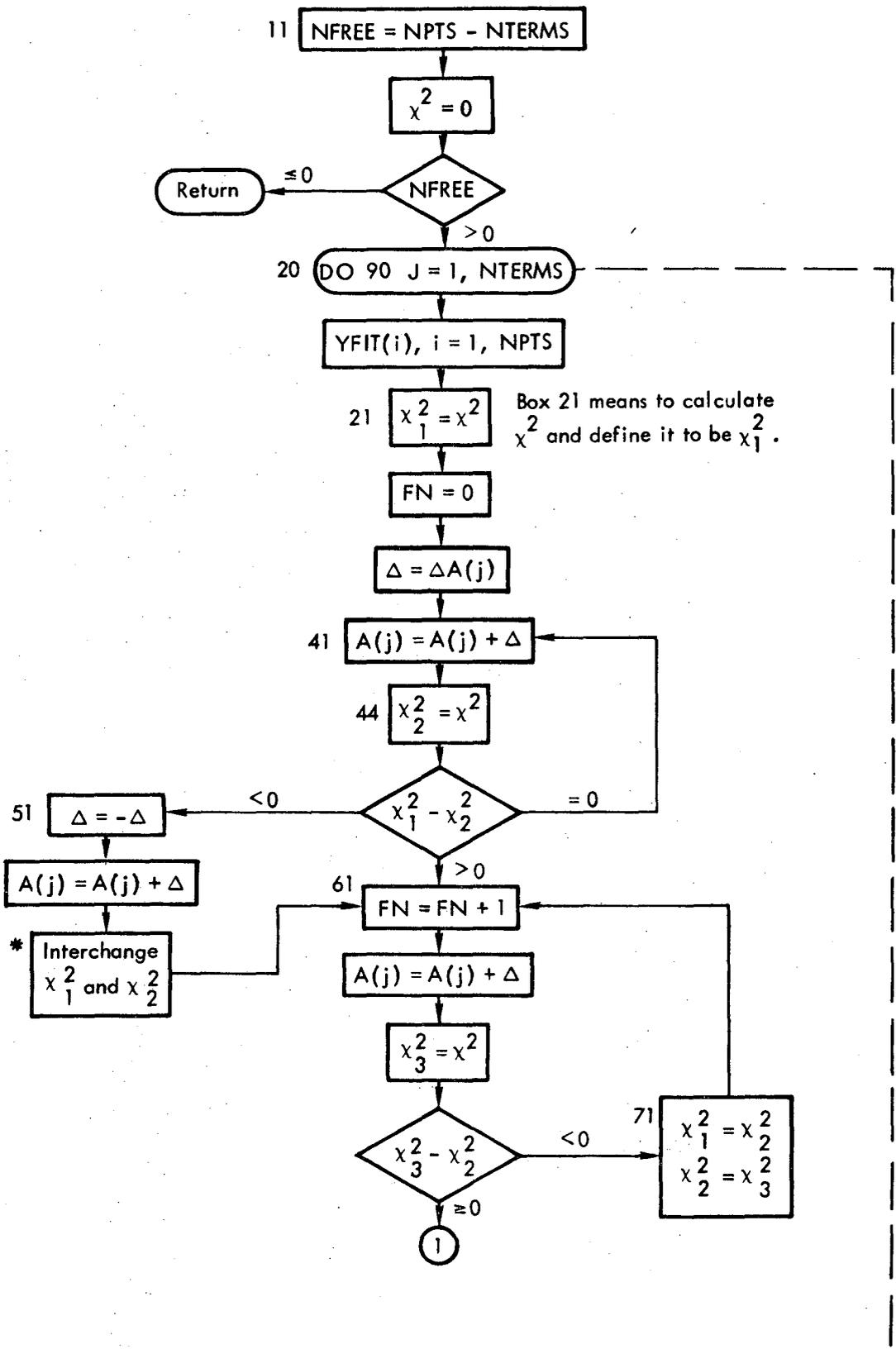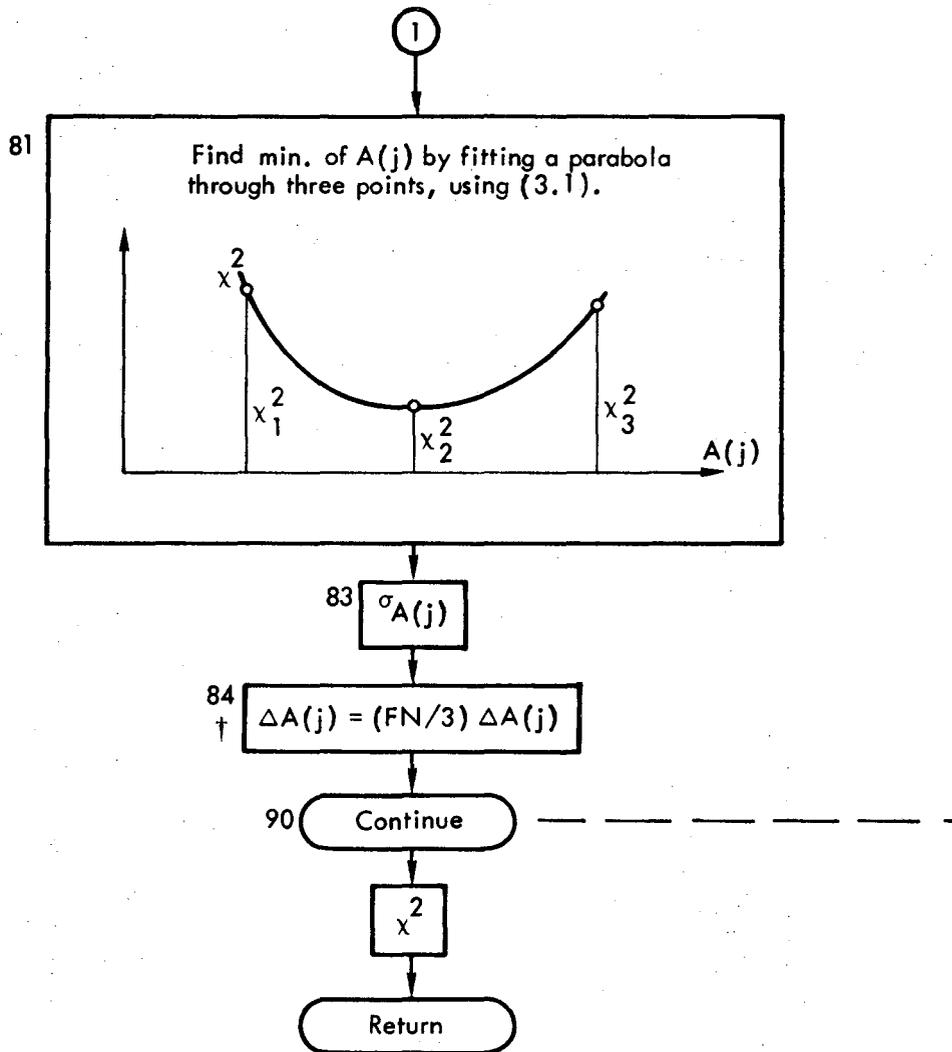
Fig. 3.4. Flow chart for the subroutine GRIDLS (see Ref. 1, p. 213).

Fig. 3.4. Continued.

In the flow chart, (Fig. 3.4), the parameter increment DELTAA(i) is written $\Delta A(i)$. The formula used to calculate the minimum of the parabola shown in Fig. 3.1 is

$$a_j(min) = a_j(3) - \Delta a_j \left[\frac{\chi^2(3) - \chi^2(2)}{\chi^2(3) - 2\chi^2(2) + \chi^2(1)} + \frac{1}{2} \right]^* . \qquad (3.1)$$

When discussing flow charts, either in the text or in footnotes on the flow charts themselves, we will, for example, refer to "box 21" and mean statement 21 in the code to which the flow chart refers. These statement numbers are also indicated on the flow chart and are taken there to refer to that box.

6. The Standard Deviation of the Parameters

For the linear least squares problem, the standard deviation of the parameter a_j is gotten from Eq. (2.23) and is

$$\sigma_{a_j} = \sqrt{\epsilon_{jj}} , \qquad (3.2)$$

where ϵ is the inverse of the curvature matrix α, given by Eq.(2.9). For the nonlinear least squares problem, one may still be able to use Eq. (3.2) for the standard deviations, where α is given by Eq. (2.10). To find out, we look a little farther into the linear problem and then extend the ideas in an obvious way to the nonlinear problem.

First solve the least squares problem for $\vec{a}$. Let any one of the parameters, say a_m, be changed by an amount Δa_m. Now solve the least squares problem again, with respect to all the parameters except a_m. Bevington[1] (p. 224) gives a proof that, if Δa_m is small enough, then

$$\Delta \chi^2 \approx \frac{\Delta a_m^2}{\epsilon_{mm}} . \qquad (3.3)$$

---

*See Fritsch[7] for a discussion of formulas like this.

In this formula, ΔX^2 is the increase in X^2 caused by the change Δa_m in the parameter a_m, and ε_{mm} is the mth diagonal element of the inverse of the curvature matrix. Equation (3.3) comes from a second-order Taylor's series expansion of X^2 around the minimum, where first derivatives with respect to the parameters are zero. Therefore Eq. (3.3) is correct for the linear least squares problem and, if we set $\Delta a_m = \sigma_{a_m}$, the standard deviation of a_m, then from Eq. (2.23), $\Delta a_m^2 / \varepsilon_{mm} = 1$, so that from Eq. (3.3), $\Delta X^2 = 1$. Therefore, we see that, in a linear least squares problem, a change of any one of the parameters by one standard deviation causes X^2 to increase by 1.

For a nonlinear least squares problem, Eq. (3.3) may be far from correct when Δa_m is set equal to σ_{a_m}. However, these ideas can be used in the nonlinear problem as follows:

1. Solve the nonlinear least squares problem for the correct parameter vector $\vec{a}$.

2. Find what change Δa_m is required when varying each parameter a_m separately, to increase X^2 by 1 (in the sense explained above).

3. If X^2 changes by about 1, whether we add Δa_m or $-\Delta a_m$ to a_m, then this is an indication that the change is a good estimate of the standard deviation of that parameter.

4. If one has an approximation to ε_{mm}, then compare $\sqrt{\varepsilon_{mm}}$ with the result of step 3. If they are approximately equal, then this is a further indication that this is a good estimate of the standard deviation of a_m.*

In all of Bevington's nonlinear least square codes, the approximation (3.2) is used without the further check on a unit change in X^2.

In the GRIDLS subroutine, the assumption is made that the variation of X^2 with respect to each parameter is independent of the other parameters (at least near the minimum of X^2). Therefore, Eq. (2.10) implies that the

_____

*These rough ideas on the problem of obtaining statistical quantities for nonlinear least squares problems will be clarified in a future report by N. Smiriga and J. Karush.

curvature matrix is diagonal, so that its inverse, ϵ, is also diagonal, and these diagonal elements are given by

$$\epsilon_{jj} = \frac{2}{(\partial^2 \chi^2 / \partial a_j^2)} \ .$$ (3.4)

An estimate for $\partial^2 \chi^2 / \partial a_j^2$ can be obtained from the last three points along the search (referring to Fig. 3.1), by using the well-known formula for the numerical approximation of the second derivative of a function of one variable:

$$\frac{\partial^2 \chi^2}{\partial a_j^2} \approx \frac{\chi_1^2 - 2\chi_2^2 + \chi_3^2}{(\Delta a_j)^2} \ .$$ (3.5)

The calculation of χ^2 is done with the subroutine FCHISQ. As we saw earlier, this subroutine calculates χ_r^2, rather than χ^2. The relation between them is

$$\chi_r^2 = \frac{\chi^2}{N-n} \ .$$ (3.6)

Using Eqs. (3.2), (3.4), (3.5), and (3.6), one gets

$$\sigma_{a_j} \approx \Delta a_j \sqrt{\frac{2}{(N-n)\left[\chi_{r,1}^2 - 2\chi_{r,2}^2 + \chi_{r,3}^2\right]}} \ ,$$

where $\chi_{r,i}^2$ means χ_r^2 at point i in Fig. 3.1, where i = 1,2, or 3. This last formula is used in statement 83 of GRIDLS (Fig. 3.4).

Chapter 4. The Gradient Search Algorithm

1. The Basic Procedure

A gradient algorithm does its search in the direction of the negative gradient of χ^2 with respect to $\vec{a}$. This is also called the method of steepest descent. The methods differ in how far they go in the direction

-35-

of the negative gradient. The algorithm discussed below approximately minimizes χ^2 along the direction of the negative gradient.

One could also introduce a scale change, since this method can be quite sensitive to scale changes. Bevington's gradient search code, called GRADLS (Ref. 1, pp. 215-222), is a method of steepest descent combined with a change of scale. A change of scale is a special type of change of variable, where the parameter vector $\vec{a}$ is replaced by a new parameter vector $\vec{b}$ such that

$$b_i = \xi_i a_i . \tag{4.1}$$

The ξ_i's are either provided by the user or built into the algorithm. In GRADLS, the relation is $b_i = a_i/\Delta a_i$ (so that $\xi_i = 1/\Delta a_i$), where the user picks the Δa_i's. In GRADLS, however, the user is not entirely free in his choice of the Δa_i's. The reason for this is that the gradients are calculated numerically according to

$$\frac{\partial \chi^2}{\partial a_j} \approx \frac{\chi^2(a_j + f\Delta a_j) - \chi^2(a_j)}{f\Delta a_j} , \tag{4.2}$$

where $f = 0.1$. Therefore, Δa_j must be small enough to give an accurate estimate for these partial derivatives, and yet big enough to allow for reasonable step sizes in $\vec{a}$. A true scale change gives one complete freedom in the choice of the ξ_i's in Eq. (4.1), so that, for example, one could possibly change the χ^2 contours from those shown in Fig. 3.2 to those in Fig. 3.3. In GRADLS, therefore, one is not really dealing with a change of scale, but with a change to dimensionless parameters. This is a good place to say a few things about a true change of scale in connection with the method of steepest descent .

2. Change of Scale [*]

A problem is "well scaled" when similar changes in the different parameters cause similar changes in χ^2. Given $f(x,\vec{a})$, and a <u>scale change vector</u> $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$, one defines a new parameter vector $\vec{b}$, given by Eq. (4.1) and a new function $g(x,\vec{b})$ defined by

$$g(x,\vec{b}) \equiv f\left(x, \frac{b_1}{\xi_1}, \frac{b_2}{\xi_2}, \ldots, \frac{b_n}{\xi_n}\right) = f(x,\vec{a}). \qquad (4.3)$$

The relation between the gradients is then

$$\frac{\partial g}{\partial b_i} = \frac{1}{\xi_i} \frac{\partial f}{\partial a_i}. \qquad (4.4)$$

Also from Eq. (4.1), the relation between increments in the parameters $\vec{a}$ and $\vec{b}$, denoted by $\vec{a}$ and $\vec{b}$, respectively, is

$$\delta a_i = \frac{1}{\xi_i} \delta b_i. \qquad (4.5)$$

Suppose λ is a positive number such that $\delta b_i = -\frac{1}{\lambda} \frac{\partial g}{\partial b_i}$ is a "small" step in the direction of the negative gradient in $\vec{b}$-space. Then the corresponding change in $\vec{a}$-space is

$$\delta a_i = -\frac{1}{\xi_i} \frac{1}{\lambda} \frac{\partial g}{\partial b_i} = -\frac{1}{\xi_i^2} \frac{1}{\lambda} \frac{\partial f}{\partial a_i}. \qquad (4.6)$$

Therefore, it is clear that $\delta\vec{a}$ is not in the direction of the negative gradient in $\vec{a}$-space, unless the ξ_i^2 are all equal.

The reason for the above development is the hope that one could pick a scale change vector $\vec{\xi}$, so that the problem would be well scaled in $\vec{b}$-space. Then one uses the method of steepest descents in $\vec{b}$-space. One could stay in $\vec{b}$-space until the "bottom of the hole," or the solution for $\vec{b}$ is found,

---

[*] See Appendix B for a discussion of the geometry of parameter changes.

and only then return to the corresponding $\vec{a}$ in $\vec{a}$-space. In GRADLS, how-
ever, control is returned to the user after each step, so it is necessary
to provide the corresponding δa_i's after each step. Another case for
which it would be necessary to calculate the δa_i's after each step is
when the scale-change vector $\vec{\xi}$ itself changes at each step. (The vector
$\vec{\xi}$ does change for each step in the subroutine CURFIT to be described
below.)

Consider a least squares problem with just two parameters. It would
be interesting to have corresponding plots of the χ^2 contours in $\vec{a}$-space
and $\vec{b}$-space, along with the paths from the corresponding starting points
to the minimum points. Such a study could not be done with GRADLS because
of the limitation, pointed out above, on the Δa's. One would need another
steepest-descent code, allowing a scale change, which would use exact
expressions for the partial derivatives.

There is a disadvantage to two-dimensional illustrations of the method
of steepest descents, because the path is made up of parallel and perpendi-
cular segments, so that the path looks very much like that of the univar-
iate method. This is true in two dimensions, but it does not generalize
to higher dimensions (see Ref. 4, pp. 75-76).

3. Powell's Modified Gradient Method

It is interesting to note that Powell[8] devised a strategy using gra-
dients that is much more effective than the simple gradient method described
above. The method finds the common center of nested similar ellipses.
Consider Fig. 4.1. The starting point is at P, and we wish to find the
minimum point C. Powell finds a point M such that the direction PM passes
near the point C. This is done in three steps:

(1) Proceed down the gradient from the starting point P to a point R,
where χ^2 is minimum.

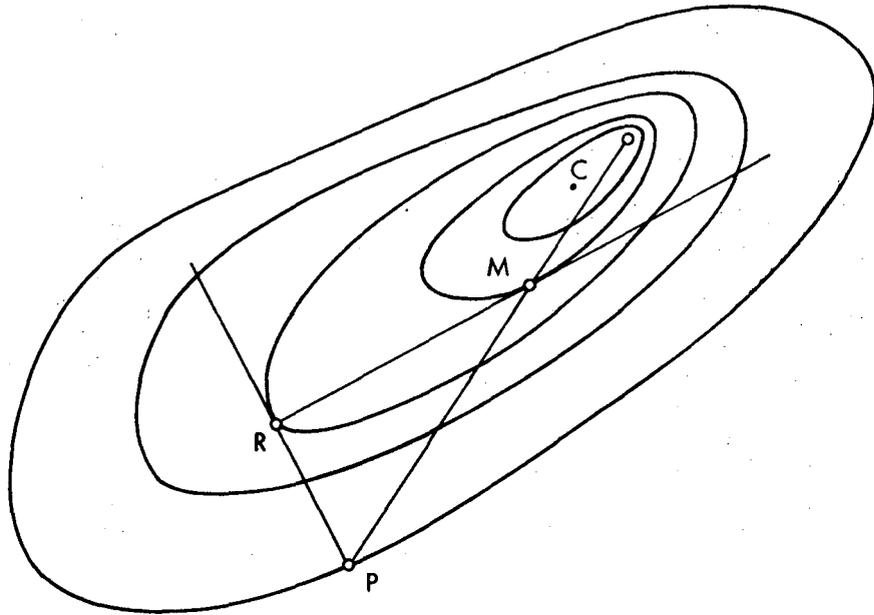(2) In the space perpendicular to PR, find the point M at which χ^2 is

Fig. 4.1. Powell's method of finding the minimum point.

minimum.

(3) Go from M in the direction PM until χ^2 is minimum near C.

This idea can be generalized to any number of dimensions (see Ref. 9, pp. 458-460).

4. The GRADLS Subroutine

The input and output quantities for GRADLS are the same as those for GRIDLS, except that the present code does not provide the standard deviations of the parameters, which is the array SIGMAA. The reason for this is that the second derivatives of χ^2 are never computed with respect to any individual parameter.

The calling sequence for GRADLS is as follows:

CALL GRADLS(X,Y,SIGMAY,NPTS,NTERMS,MODE,A,DELTAA,YFIT,CHISQR)

As in all of Bevington's codes, GRADLS does only one iteration, so that the user's code must call GRADLS over and over again, until some stopping condition is met.

The method of steepest descents is used in $\vec{b}$-space, so that one must have some step size in order to run a search in the direction of the negative gradient. In GRADLS this step size is given by the unit vector $\vec{\gamma}$ in the gradient direction. The jth component of this unit vector is given by

$$\gamma_j = \frac{\partial \chi^2/\partial b_j}{\sqrt{\sum_{k=1}^{n} \left(\frac{\partial \chi^2}{\partial b_k}\right)^2}} , \qquad (4.7)$$

so that one puts $\delta b_j = -\gamma_j$. From Eq. (4.5) and the fact that $\xi_i = 1/\Delta a_i$, one gets

$$\delta a_j = -\gamma_j \Delta a_j . \qquad (4.8)$$

One could question the wisdom of using $-\gamma_j$ for the b_j-step in Eq. (4.8). This seems to be a large step, since $\vec{\gamma}$ is a unit vector in

$\vec{b}$-space. The reason why this makes sense is that one usually has $\Delta a_j \ll a_j$, so that one unit of length along any vector in $\vec{b}$-space corresponds to moving a much smaller distance in $\vec{a}$-space.

It should be emphasized here that δa_j is the step size for the increments in a_j, used in the search for $\vec{a}$ that minimizes χ^2, while Δa_j is the (user supplied) increment for calculating the partial derivatives and for defining the dimensionless parameter b_j.

The formula used for finding the minimum of the parabola through three points is (Fig. 4.2)

$$a_j(\min) = a_j(3) - \delta a_j \left[\frac{1}{1 + \left(\dfrac{\chi^2(1) - \chi^2(2)}{\chi^2(3) - \chi^2(2)} \right)} + \frac{1}{2} \right]. \qquad (4.9)$$

Note that this formula is different from the one used in the GRIDLS subroutine, given by Eq. (3.1). The above formula is used in box 91 in the detailed flow chart, and its use there may be a little confusing, because we are changing all the variables at once. The method of finding the minimum of the parabola can easily be applied to n-dimensions, because we are always searching along a line anyway. Figure 4.2 illustrates the idea for the case of parameters a_1 and a_2. Along the line S in the $a_1 - a_2$ plane, the vectors $\vec{a}_1$, $\vec{a}_2$, and $\vec{a}_3$ are shown analogous to the values $a_j(1)$, $a_j(2)$, and $a_j(3)$ in Fig. 3.1.

5. The GRADLS Flow Charts

Figure 4.3 is a rough outline flow chart, and Fig. 4.4 is a detailed flow chart. The Roman numerals on the flow charts identify portions of the detailed flow chart that correspond with the rough flow chart.
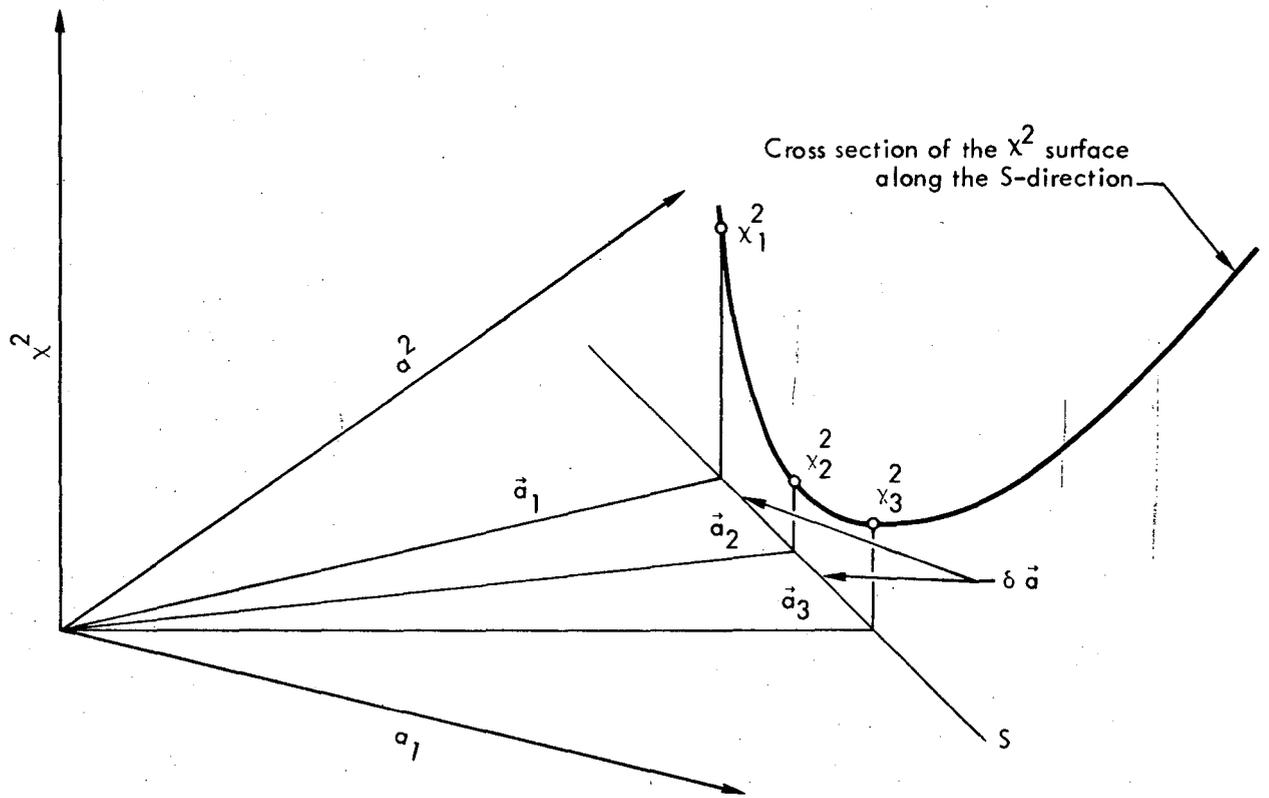
Fig. 4.2. Method of finding the minimum, given the parameters a_1 and a_2.

```
            (I)

        x²₁ = x² (a)

           (II)

Increment aⱼ according to δaⱼ = -λⱼ Δaⱼ, for j = 1,2, ..., n

          (III)

        x²₂ = x² (a⃗)

          (IV)

If x²₂ is greater than x²₁, go back and increment the original
aⱼ's by 1/2 δaⱼ.
If x²₂ is less than x²₁, go to the next box.

           (V)

Keep on incrementing the aⱼ's until x² starts to increase.

          (VI)

Pass a parabola through three points that include the minimum
and pick for a⃗ the minimum of the parabola.

          (VII)

If x²(a⃗) is less than x²₂, we are finished; otherwise, use a⃗₂

          Return
```

Box I: $x_1^2 = x^2(a)$

Box II: Increment a_j according to $\delta a_j = -\lambda_j \Delta a_j$, for $j = 1, 2, \ldots, n$

Box III: $x_2^2 = x^2(\vec{a})$

Box IV: If x_2^2 is greater than x_1^2, go back and increment the original a_j's by $1/2 \, \delta a_j$. If x_2^2 is less than x_1^2, go to the next box.

Box V: Keep on incrementing the a_j's until x^2 starts to increase.

Box VI: Pass a parabola through three points that include the minimum and pick for $\vec{a}$ the minimum of the parabola.

Box VII: If $x^2(\vec{a})$ is less than x_2^2, we are finished; otherwise, use $\vec{a}_2$

Return

Fig. 4.3. Outline flow chart for the subroutine GRADLS.

Detailed Flow Chart for the
Subroutine GRADLS

(I)

11 NFREE = NPTS - NTERMS

13 $\chi^2 = 0$

≤ 0

NFREE

> 0

Return

21 YFIT(i), i = 1, NPTS

$\chi_1^2 = \chi^2$

The statement $\chi_1^2 = \chi^2$
means to calculate
χ^2 and define it to be χ_1^2.

(II)

31 SUM = 0.

32 DO 39 J = 1, NTERMS

$\Delta = 0.1 \, \Delta A(j)$

$A(j) = A(j) + \Delta$

YFIT(i), i = 1, NPTS

$A(j) = A(j) - \Delta$

$GRAD(j) = \chi_1^2 - \chi^2$

39 $SUM = SUM + [\, GRAD(j)\,]^2$

41 $GRAD(j) = \Delta A(j) \, \dfrac{GRAD(j)}{\sqrt{SUM}}$, j = 1, NTERMS

(III)

51 A(j) = A(j) + GRAD(j), j = 1, NTERMS

53 YFIT(i), i = 1, NPTS

$\chi_2^2 = \chi^2$

(IV)
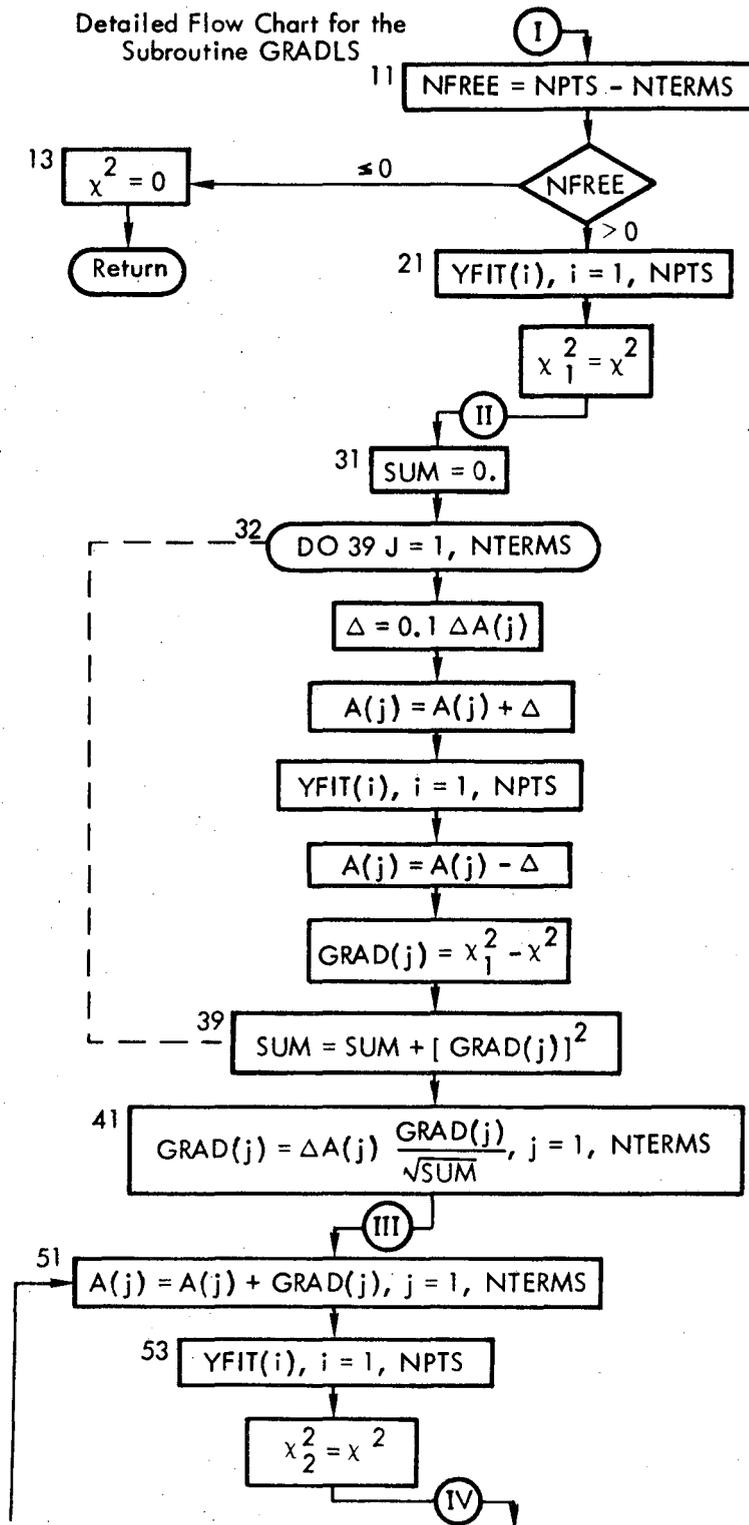
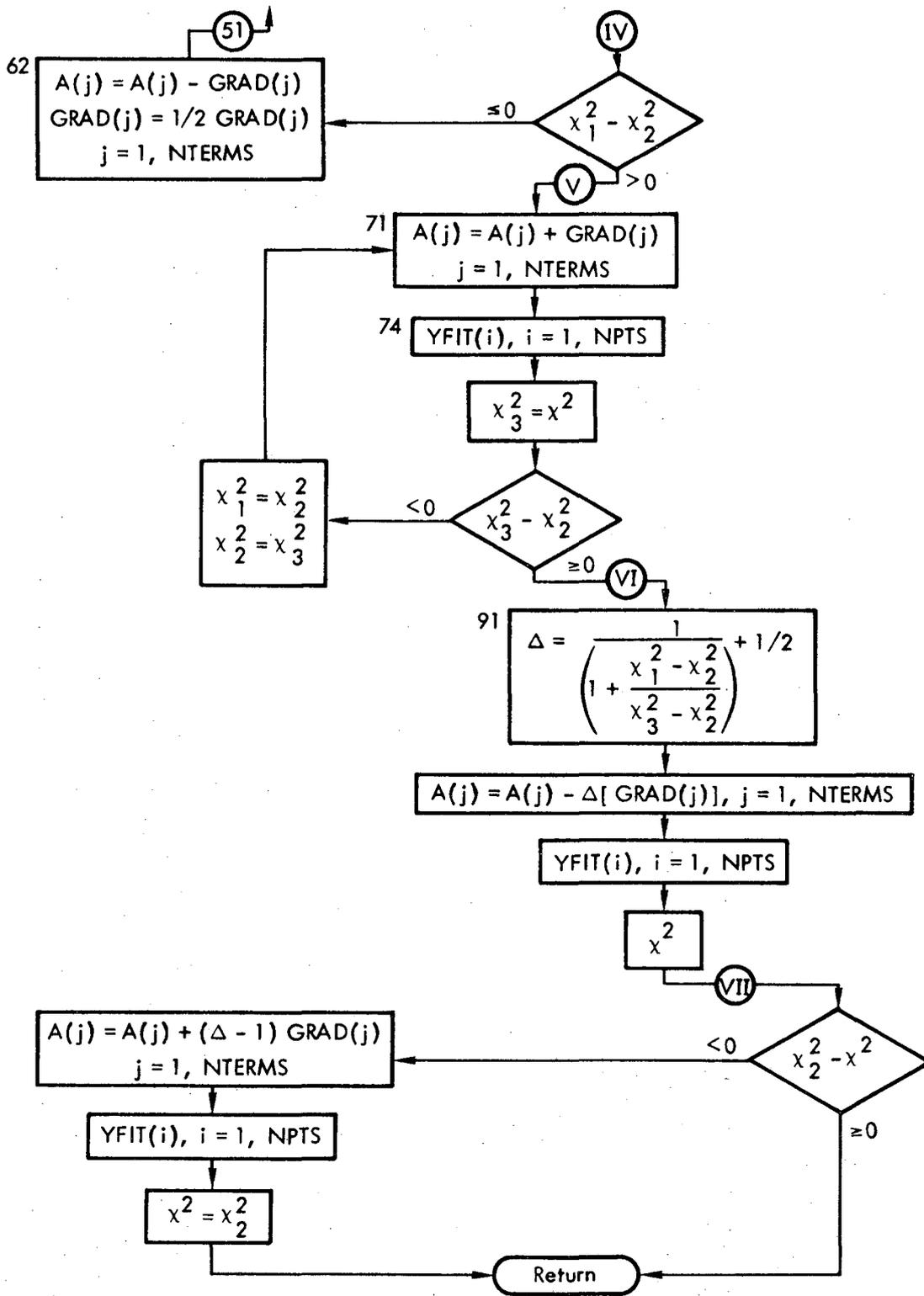Fig. 4.4. Detailed flow chart for the subroutine GRADLS.

Fig. 4.4. Continued.

Chapter 5. Gauss's Method

1. The Basic Procedure

The method of Gauss starts by expanding the fitting function $f(x,\vec{a})$ about $\vec{a}_0$ to first order in a Taylor series in the parameter increments $\delta\vec{a}$, where $\vec{a} = \vec{a}_0 + \delta\vec{a}$.

$$f(x,\vec{a}) \approx f(x,\vec{a}_0) + \sum_{j=1}^{n} \frac{\partial f(x,\vec{a})}{\partial a_j}\Bigg|_{\vec{a}_0} \delta a_j. \tag{5.1}$$

The right-hand side of Eq. (5.1) is a linear function in the parameter increments δa_j, to which one can apply the method of linear least squares. The partial derivatives on the right side of Eq. (5.1) are to be evaluated at the starting point $\vec{a}_0$. We will write $\frac{\partial f}{\partial a_j}(x,\vec{a}_0)$ for $\frac{\partial f}{\partial a_j}(x,\vec{a})\big|_{a_0}$.

With the above approximation for $f(x,\vec{a})$, χ^2 can be expressed in terms of the parameter increments:

$$\chi^2 \approx \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - f(x_i,\vec{a}_0) - \sum_{j=1}^{n} \frac{\partial f}{\partial a_j}(x_i,\vec{a}_0)\delta a_j \right]^2. \tag{5.2}$$

Forming the partial derivatives of χ^2 with respect to the parameter increments, one gets

$$\frac{\partial \chi^2}{\partial(\delta a_k)} \approx -2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - f(x_i,\vec{a}_0) \right.$$
$$\left. - \sum_{j=1}^{n} \frac{\partial f}{\partial a_j}(x_i,\vec{a}_0)\delta a_j \right] \frac{\partial f}{\partial a_k}(x_i,\vec{a}_0). \tag{5.3}$$

Setting $\frac{\partial \chi^2}{\partial(\delta a_k)} = 0$, for $k = 1,2,\ldots,n$, gives the equations

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - f(x_i, \vec{a}_0) \right] \frac{\partial f}{\partial a_k} (x_i, \vec{a}_0)$$

$$= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(\sum_{j=1}^{n} \frac{\partial f(x_i, \vec{a}_0)}{\partial a_j} \frac{\partial f(x_i, \vec{a}_0)}{\partial a_k} \right) \delta a_j .$$

Interchanging the i and j summations on the right gives

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - f(x_i, \vec{a}_0) \right] \frac{\partial f}{\partial a_k} (x_i, \vec{a}_0)$$

$$= \sum_{j=1}^{n} \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f}{\partial a_j} (x_i, \vec{a}_0) \frac{\partial f}{\partial a_k} (x_i, \vec{a}_0) \right) \delta a_j .$$

This is a set of n linear equations (the normal equations in $\delta \vec{a}$),

$$\vec{\beta} = \alpha \, \delta \vec{a}, \tag{5.4}$$

where the vector $\vec{\beta}$ and the matrix α are given by

$$\beta_k = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - f(x_i, \vec{a}_0) \right] \frac{\partial f}{\partial a_k} (x_i, \vec{a}_0), \tag{5.5}$$

$$\alpha_{jk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f}{\partial a_j} (x_i, \vec{a}_0) \frac{\partial f}{\partial a_k} (x_i, \vec{a}_0). \tag{5.6}$$

Solving Eq. (5.4) for $\delta \vec{a}$, one replaces $\vec{a}_0$ by $\vec{a}_0 + \delta \vec{a}$ and starts all over again until the process converges ($\delta \vec{a} \rightarrow 0$).

If one is going to use the Gauss method to solve the nonlinear least squares problem, it is often better not to form the normal equation (5.4), but to solve directly the overdetermined system

$$\frac{y_i}{\sigma_i} = \frac{1}{\sigma_i} f(x_i, \vec{a}), \qquad\qquad i = 1, 2, \ldots, N, \tag{5.7}$$

using MLR.[10] The procedure for doing this is given in Appendix C.

2. The Effect of a Scale Change on the Gauss Method

It is interesting to know and easy to show that the Gauss method is independent of the scale change $\vec{a} \rightarrow \vec{b}$, where $b_i = \xi_i a_i$. We have seen that a scale change can be very important in the method of steepest descents. To show this, first make the (induced) mapping Eq. (4.3) from $f(x,\vec{a})$ to $g(x,\vec{b})$. Now perform the Gauss process using $g(x,\vec{b})$, just as we did to get from Eq. (5.1) to Eq. (5.4). This results in the equation

$$\vec{\beta}' = \alpha' \, \delta\vec{b}, \tag{5.8}$$

where

$$\beta'_k = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - g(x_i,\vec{b}_0) \right] \frac{\partial g}{\partial b_k} (x_i,\vec{b}_0) \tag{5.9}$$

and

$$\alpha'_{jk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial g}{\partial b_j} (x_i,\vec{b}_0) \frac{\partial g}{\partial b_k} (x_i,\vec{b}_0). \tag{5.10}$$

Using Eq. (4.4), Eqs. (5.9) and (5.10) become

$$\beta'_k = \frac{1}{\xi_k} \beta_k, \tag{5.11}$$

$$\alpha'_{jk} = \frac{1}{\xi_j \xi_k} \alpha_{jk}, \tag{5.12}$$

where β_k and α_{jk} are given by Eqs. (5.5) and (5.6), respectively. When the diagonal (scale change) matrix D is introduced as

$$D = \begin{pmatrix} \xi_1 & & & & \\ & \xi_2 & & & \\ & & \cdot & & 0 \\ & & & \cdot & \\ 0 & & & & \cdot \\ & & & & \xi_n \end{pmatrix}, \tag{5.13}$$

Eqs. (5.11) and (5.12) become

$$\vec{\beta}' = D^{-1}\vec{\beta},^*$$ (5.14)

$$\alpha' = D^{-1}\alpha D^{-1}.$$ (5.15)

Therefore, Eq. (5.8) becomes

$$D^{-1}\vec{\beta} = (D^{-1}\alpha D^{-1})\delta\vec{b}.$$

But, from Eq. (4.5), $\delta\vec{b} = D\delta\vec{a}$, so that $D^{-1}\vec{\beta} = (D^{-1}\alpha D^{-1})(D\delta\vec{a})$, from which one gets Eq. (5.4). Therefore, mathematically the scale change has no effect on the Gauss method; but, of course, numerically it may have some effect.

Chapter 6. The Levenberg-Marquardt Algorithm[11,12]

1. The Basic Procedure

If the starting point $\vec{a}_0$ is too far from the final solution $\vec{a}$ of a nonlinear least squares problem, the Gauss method often gives corrections δa_j to the parameters that are much too large. To prevent this, the Levenberg-Marquardt algorithm minimizes the sum of χ^2 and a weighted sum of squares of the corrections. Therefore, S is minimized, where

$$S = \chi^2 + \lambda \sum_{j=1}^{n} (\delta a_j)^2.$$ (6.1)

---

*It should be pointed out that

$$D^{-1} = \begin{pmatrix} 1/\xi_1 & & & & \\ & 1/\xi_2 & & & \\ & & \cdot & & 0 \\ 0 & & & \cdot & \\ & & & & 1/\xi_n \end{pmatrix}.$$

The factor λ is a weighting factor for the second term in S.

To minimize S, we set the partial derivatives of S with respect to the δa_j's equal to zero, getting

$$\frac{\partial \chi^2}{\partial(\delta a_k)} + 2\lambda \, \delta a_k = 0, \qquad \text{for } k = 1,2,\ldots,n. \qquad (6.2)$$

Putting Eq. (5.3) into Eq. (6.2) gives

$$\beta_k = \sum_{j=1}^{n} \alpha_{jk} \, \delta a_j + \lambda \, \delta a_k, \qquad (6.3)$$

where β_k and α_{jk} are defined by Eqs. (5.5) and (5.6). If we define a new matrix α' such that

$$\alpha'_{ij} = \left\{ \begin{array}{l} \alpha_{ij} \\ \\ \alpha_{ii} + \lambda \end{array} \right\} \quad \text{when} \quad \left\{ \begin{array}{l} i \neq j \\ \\ i = j \end{array} \right\}, \qquad (6.4)$$

then the Levenberg-Marquardt equation corresponding to Eq. (5.4) is

$$\vec{\beta} = \alpha' \vec{\delta a}. \qquad (6.5)$$

It is clear that, when $\lambda \to 0$, this gives Eq. (5.4), which is the Gauss method. On the other hand, as $\lambda \to \infty$, then

$$\beta_j \approx \lambda \, \delta a_j, \qquad (6.6)$$

from which

$$\delta a_j \approx \frac{1}{\lambda} \beta_j. \qquad (6.7)$$

Return to the definition of χ^2 given by Eq. (2.3), from which it follows that

$$\frac{\partial \chi^2}{\partial a_j} = -2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - f(x_i, \vec{a}) \right] \frac{\partial f}{\partial a_j} (x_i, \vec{a}). \tag{6.8}$$

Therefore, from Eq. (5.5),

$$\beta_j = -\frac{1}{2} \frac{\partial \chi^2}{\partial a_j}, \tag{6.9}$$

so that from Eqs. (6.7) and (6.9) it is clear that $\lambda \to \infty$ gives a correction $\delta \vec{a}$, which is a small step in the direction of the negative gradient.

Thus the Levenberg-Marquardt idea of minimizing S given by Eq. (6.1) produces an algorithm that is a hybrid between the Gauss and gradient methods. At the start of a least squares problem, λ should be large enough to take advantage of the slower but more positive convergence of the gradient method. As λ is made larger, the length of the step gets smaller, and the direction of the step approaches that of the negative gradient. As λ is made smaller, the step size gets larger and tends in the direction given by the Gauss method.

Levenberg said it would be best to find the minimum of χ^2 as a function of λ on each iteration. This requires quite a bit of calculation. One objection to this is that it may be less efficient to spend time looking for a minimum of χ^2 with respect to λ than to start a new iteration with more up-to-date information for α' and $\vec{\beta}$. Marquardt suggested that Levenberg's scheme would lead to overemphasis toward the gradient solution. On the other hand, Fletcher does not agree with Marquardt on this point (see Ref. 13, p. 2). The algorithms for the Levenberg-Marquardt process differ in the way λ is selected at each iteration.

Marquardt[12] found that, on the average, there was an 80° to 90° difference in direction between the negative gradient direction and the direction given by the Gauss method. This can be seen in Fig. 6.1.
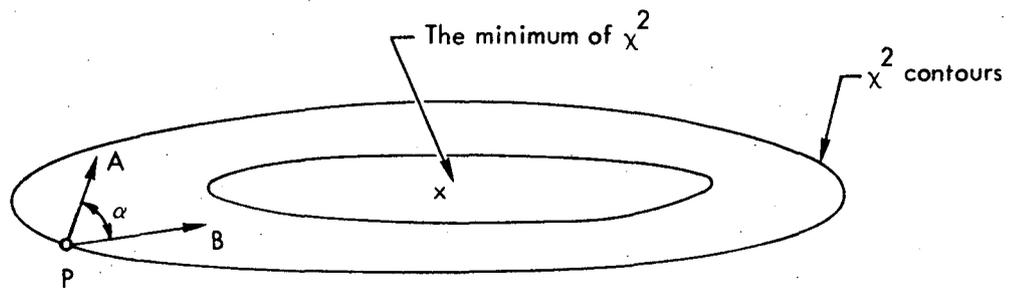
Fig. 6.1. The difference in directions at point P between the negative
gradient (arrow A) and the Gauss direction (arrow B).

2. The Levenberg-Marquardt Scale Change Algorithm

We have seen that, as $\lambda \to 0$, the Levenberg-Marquardt algorithm approaches the Gauss algorithm for which no scale change is necessary. A modification in the α' matrix of Eq. (6.4) makes it possible to introduce a scale change that becomes effective as $\lambda \to \infty$, where the gradient method is used. We have seen that the scale changes in the gradient method can be very important.

The new α' matrix is defined as

$$\alpha'_{ij} = \left\{ \begin{array}{l} \alpha_{ij} \\ \\ \alpha_{ii}(1+\lambda) \end{array} \right\} \quad , \text{ where } \quad \left\{ \begin{array}{l} i \neq j \\ \\ i = j \end{array} \right\} . \tag{6.10}$$

The equation to be solved for $\delta\vec{a}$ is still Eq. (6.5), where $\vec{\beta}$ is also the same as before, namely, Eq. (5.5). As before, when $\lambda \to 0$, this new equation approaches the Gauss equation. When $\to\infty$, Eq. (6.6) is replaced by

$$\beta_j \approx \lambda \, \alpha_{jj} \delta_j , \tag{6.11}$$

so that

$$\delta a_j \approx \frac{1}{\alpha_{jj}} \left(\frac{\beta_j}{\lambda} \right). \tag{6.12}$$

The difference between this result and the previous result, Eq. (6.7), is the presence of a factor α_{jj} in the denominator of Eq. (6.12). Using Eq. (6.9), Eq. (6.12) can be put in the form

$$\delta a_j \approx - \frac{1}{\left(\sqrt{2\alpha_{jj}} \right)^2} \left(\frac{1}{\lambda} \frac{\partial \chi^2}{\partial a_j} \right). \tag{6.13}$$

Referring to Eq. (4.6), one sees that Eq. (6.13) implies a scale change, such that

$$\xi_j = \sqrt{2\alpha_{jj}} . \tag{6.14}$$

From Eq. (5.6),

$$\alpha_{jj} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(\frac{\partial f}{\partial a_j} (x_i, \vec{a}) \right)^2 \geq 0, \tag{6.15}$$

so that we are safe in using Eq. (6.14).

Now we ask what is the effect of the scaling given by Eq. (6.14)? To answer this question, we must digress to another view of the scaling question.

3. More About Scaling

Differentiate Eq. (6.8) again with respect to a_j. One gets

$$\frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_j^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial f}{\partial a_j} (x_i, \vec{a}) \right]^2$$

$$- \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - f(x_i, \vec{a}) \right] \frac{\partial^2 f}{\partial a_j^2} (x_i, \vec{a}). \tag{6.16}$$

If the second sum in Eq. (6.16) is small with respect to the first, then

$$\frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_j^2} \approx \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(\frac{\partial f}{\partial a_j} (x_i, \vec{a}) \right)^2 . \tag{6.17}$$

Therefore, from Eq. (6.15), $\alpha_{jj} \approx \frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_j^2}$, so that, from Eq. (6.14),

$$\xi_j \approx \sqrt{\frac{\partial^2 \chi^2}{\partial a_j^2}} . \tag{6.18}$$

Contour curves of a quadratic function (in two dimensions) are made less eccentric by a simple scale change that makes the coefficients of the

-54-

squared terms equal. For example, consider the function

$$f(x_1, x_2) = 36x_1^2 + x_2^2 - 6x_1 x_2.$$

The scale change $\tilde{x}_1 = 6x_1$, $\tilde{x}_2 = x_2$ gives

$$f(x_1, x_2) = \tilde{x}_1 + \tilde{x}_2^2 - \tilde{x}_1 \tilde{x}_2 \equiv \tilde{f}(\tilde{x}_1, \tilde{x}_2).$$

Rough contour plots of f and $\tilde{f}$ are shown in Fig. 6.2.

In the least squares problem the natural extension of the above process would be a scale change that makes $\partial^2 \chi^2 / \partial a_k^2 \approx 1$, for all $k = 1, 2, \ldots, n$. The reason for this is that, near the solution $\vec{a}$ of a least squares problem, the first derivatives of χ^2, with respect to the parameters, are approximately zero. When χ^2 is expanded in a Taylor series about $\vec{a}$ in the quantities δa_i, the coefficients of the terms δa_i^2 are $\frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_i^2}$. Near the solution $\vec{a}$, the coefficients of these squared terms are the most important. Therefore, the χ^2 contours near $\vec{a}$ are approximated by the contours of a Taylor series expansion up to second-order terms. Consider the Taylor expansion of χ^2 around the least squares solution, at which $\chi^2 = \chi_0^2$:
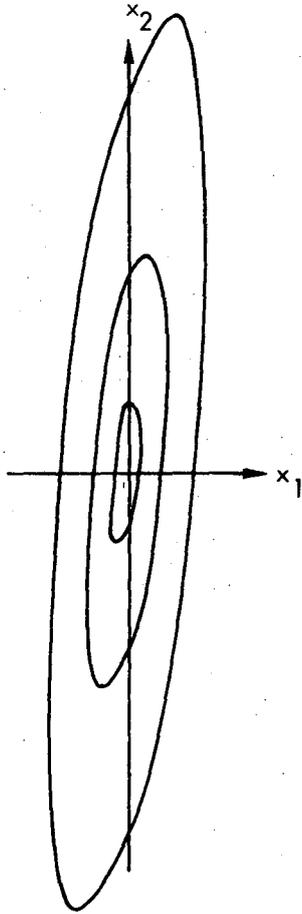
$$\chi^2 \approx \chi_0^2 + \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial a_1^2} \right|_0 \delta a_1^2 + \frac{1}{2} \left. \frac{\partial \chi^2}{\partial a_2^2} \right|_0 \delta a_2^2 + \ldots . \qquad (6.19)$$
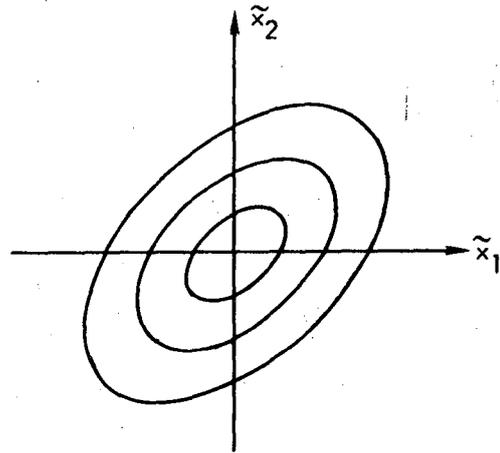
Define

$$\tilde{a}_i = \sqrt{\left. \frac{\partial^2 \chi^2}{\partial a_1^2} \right|_0}\ a_i, \ \text{ so that } \ \delta a_i^2 = \frac{1}{\left. \frac{\partial \chi^2}{\partial a_i^2} \right|_0} \delta a_i^2 ,$$

and

$$\chi^2 \approx \chi_0^2 + \frac{1}{2} \delta \tilde{a}_1^2 + \frac{1}{2} \delta \tilde{a}_2^2 + \ldots . \qquad (6.20)$$

(a)   (b)

Fig. 6.2. (a) Contour curves of $f(x_1 x_2) = 36x_1^2 + x_2^2 - 6x_1 x_2$. (b) Contour
curves of $\tilde{f}(\tilde{x}_1 \tilde{x}_2) = \tilde{x}_1^2 + \tilde{x}_2^2 - \tilde{x}_1 \tilde{x}_2$.

From Eq. (6.18), we see that the scale change induced by the definition of the matrix α', given by Eq. (6.10), could be expected to improve the scaling.

There is a theory of optimal scaling,[14] and the above scaling is not optimal (except in the two-variable quadratic problem); however, it is found in practice that the above type of "diagonal" scaling is of considerable benefit. For more information and examples on scaling, see Ref. 4, pp. 90-97.

4. The Solution of Eq. (6.5)

Given the diagonal matrices A and C,

$$
A = \begin{pmatrix} a_1 & & & & \\ & a_2 & & & \\ & & \cdot & & 0 \\ & & & \cdot & \\ 0 & & & & \cdot \\ & & & & a_n \end{pmatrix}, \qquad C = \begin{pmatrix} c_1 & & & & \\ & c_2 & & & \\ & & \cdot & & 0 \\ & & & \cdot & \\ 0 & & & & \cdot \\ & & & & c_n \end{pmatrix}, \quad (6.21)
$$

and any full matrix $B = (b_{ij})$, one can establish the following identities:

$$
\left.\begin{aligned}
(AB)_{ij} &= a_i\, b_{ij} \\
(BC)_{ij} &= b_{ij}\, c_j \\
(ABC)_{ij} &= a_i\, b_{ij}\, c_j
\end{aligned}\right\}. \qquad (6.22)
$$

The solution of Eq. (6.5) is

$$
\delta\vec{a} = (\alpha')^{-1}\vec{\beta}. \qquad (6.23)
$$

Define the diagonal matrix D such that

$$
D = \begin{pmatrix} \sqrt{\alpha_{11}} & & & & \\ & \sqrt{\alpha_{22}} & & & \\ & & \cdot & & 0 \\ & & & \cdot & \\ 0 & & & & \cdot \\ & & & & \sqrt{\alpha_{nn}} \end{pmatrix}, \qquad (6.24)
$$

where the α' matrix is given by Eq. (6.10). In terms of D and α', define a new matrix α'' such that

$$\alpha' = D\alpha''D. \qquad (6.25)$$

The matrix α'' is then given by

$$\alpha'' = \begin{pmatrix} 1 + \lambda & \dfrac{\alpha_{12}}{\sqrt{\alpha_{11}}\sqrt{\alpha_{22}}} & \cdots & \dfrac{\alpha_{1n}}{\sqrt{\alpha_{11}}\sqrt{\alpha_{nn}}} \\ \dfrac{\alpha_{21}}{\sqrt{\alpha_{22}}\sqrt{\alpha_{11}}} & 1 + \lambda & \cdots & \dfrac{\alpha_{2n}}{\sqrt{\alpha_{22}}\sqrt{\alpha_{nn}}} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\alpha_{n1}}{\sqrt{\alpha_{nn}}\sqrt{\alpha_{11}}} & \dfrac{\alpha_{n2}}{\sqrt{\alpha_{nn}}\sqrt{\alpha_{22}}} & \cdots & 1 + \lambda \end{pmatrix}. \qquad (6.26)$$

In Bevington's[1] subroutine CURFIT, the array α'' is called ARRAY, so that ARRAY$(i,j) = \alpha''_{ij}$.

The reason for forming α'' is the hope that it will be easier to find its inverse than the inverse of α', because the elements of α'' should be more or less the same size. This scaling of the matrix α' is necessary because Bevington uses the Gauss-Jordan matrix inversion method, which is far from the best way to invert matrices. The code is called MATINV (see Ref. 1, p. 302).

From Eq. (6.25),

$$(\alpha')^{-1} = (D\alpha''D)^{-1} = D^{-1}(\alpha'')^{-1}D^{-1}, \qquad (6.27)$$

where

$$D^{-1} = \begin{pmatrix} \dfrac{1}{\sqrt{\alpha_{11}}} & & & 0 \\ & \dfrac{1}{\sqrt{\alpha_{22}}} & & \\ & & \ddots & \\ 0 & & & \dfrac{1}{\sqrt{\alpha_{nn}}} \end{pmatrix}. \qquad (6.28)$$

Therefore,

$$\delta \vec{a} = \left[D^{-1} (\alpha'')^{-1} D^{-1} \right] \vec{\beta}. \qquad (6.29)$$

Let $\varepsilon = (\alpha'')^{-1}$, so that

$$\delta a = (D^{-1} \varepsilon D^{-1}) \vec{\beta}. \qquad (6.30)$$

[In CURFIT, ε is also called ARRAY, after the matrix inverter MATINV finds the inverse of the matrix α''.] Writing out Eq. (6.30) in component form,

$$\delta a_j = \sum_{k=1}^{n} \frac{\beta_k \ \varepsilon_{jk}}{\sqrt{\alpha_{jj}} \sqrt{\alpha_{kk}}}. \qquad (6.31)$$

This is the equation used in statements 81-84 in CURFIT.

Numerically it would be better to solve Eq. (6.5) for $\delta \vec{a}$ using some method other than by forming the inverse of the α' matrix, as is done in CURFIT. Other methods are Gaussian elimination and Householder reflections, as in MLR.[10]

5. The Variance of the Parameters

The variance of the parameter a_j can be approximated by [see Eq. (3.2)]

$$\sigma_{a_j}^2 \approx \varepsilon_{jj}, \qquad (6.32)$$

where ε is the inverse of the curvature matrix. We have seen in the discussion just before Eq. (6.18) that α is an approximation to the curvature matrix, not α' given by Eq. (6.10). Nevertheless, the code uses the inverse of α' rather than the inverse of α. One reason for doing this is that the subroutine has the inversion of α', but not that of α. Another reason is that λ is usually small when the solution of a least squares

-59-

problem is reached, so there is then very little difference between α and α'. Therefore, from Eq. (6.27),

$$\sigma^2_{a_j} \approx \left[(\alpha')^{-1} \right]_{jj} = \left[D^{-1} \epsilon D^{-1} \right]_{jj} = \frac{\epsilon_{jj}}{\alpha_{jj}} = \frac{ARRAY(j,j)}{\alpha(j,j)} , \qquad (6.33)$$

so that

$$\sigma_{a_j} \approx \sqrt{\frac{ARRAY(j,j)}{\alpha(j,j)}} . \qquad (6.34)$$

This last equation is used in statement 101 in CURFIT.

6. The Algorithm Given by Marquardt

(1) Compute $\chi^2(\vec{a})$.

(2) The user sets λ (usually $\lambda = 10^{-3}$).

(3) Compute $\delta\vec{a}$ and $\chi^2(\vec{a} + \delta\vec{a})$ with this choice of λ.

(4) If $\chi^2(\vec{a} + \delta\vec{a}) > \chi^2(\vec{a})$, increase λ by a factor of 10, and repeat Step 3.

(5) If $\chi^2(\vec{a} + \delta\vec{a}) < \chi^2(\vec{a})$ decrease λ by a factor of 10, take $\vec{a}' = \vec{a} + \delta\vec{a}$ to be the new starting point, and return to Step 3, substituting $\vec{a}'$ for $\vec{a}$.

It may be necessary to recompute $\delta\vec{a}$ several times in order to get a value of λ that works, but the matrix α and the vector $\vec{\beta}$ need be calculated just once per iteration (because the matrix α', given by Eq. (6.10), is gotten from the matrix α, and $\vec{\beta}$ does not depend on λ).

The input and output variables are the same as in the other codes, except for the extra input variable FLAMDA = λ. The calling sequence for CURFIT is as follows:

CALL CURFIT(X,Y,SIGMAY,NPTS,NTERMS,MODE,A,DELTAA,SIGMAA,FLAMDA,YFIT,CHISQR)

This subroutine does only one step of the Levenberg-Marquardt process, leaving the convergence criterion to the user. This subroutine has the special feature that it allows either analytical or numerical derivatives.

7. The Flow Chart for CURFIT

Figure 6.3 is an outline flow chart for CURFIT, and Fig. 6.4 is a detailed flow chart. The Roman numerals identify corresponding parts on the two flow charts.
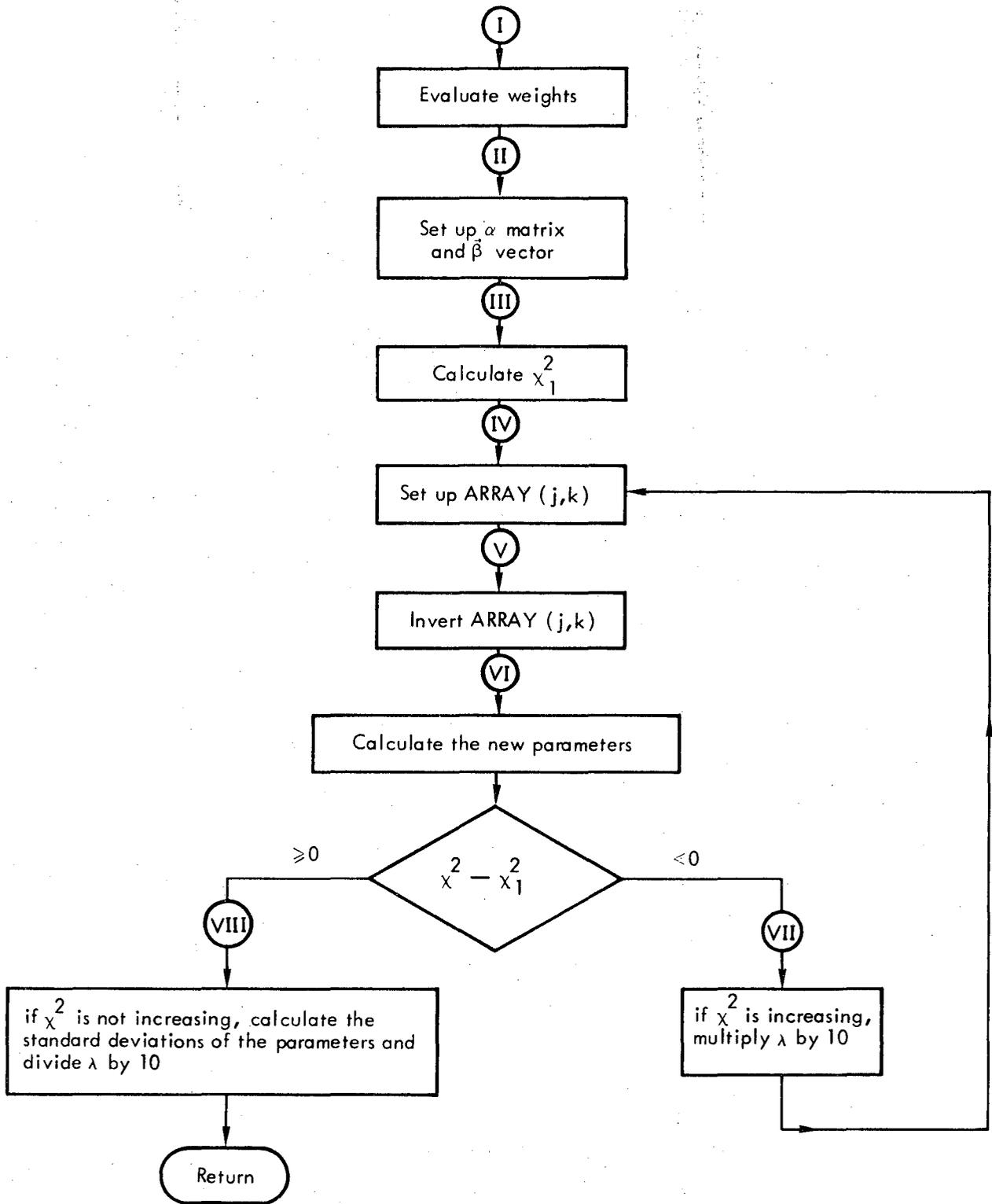
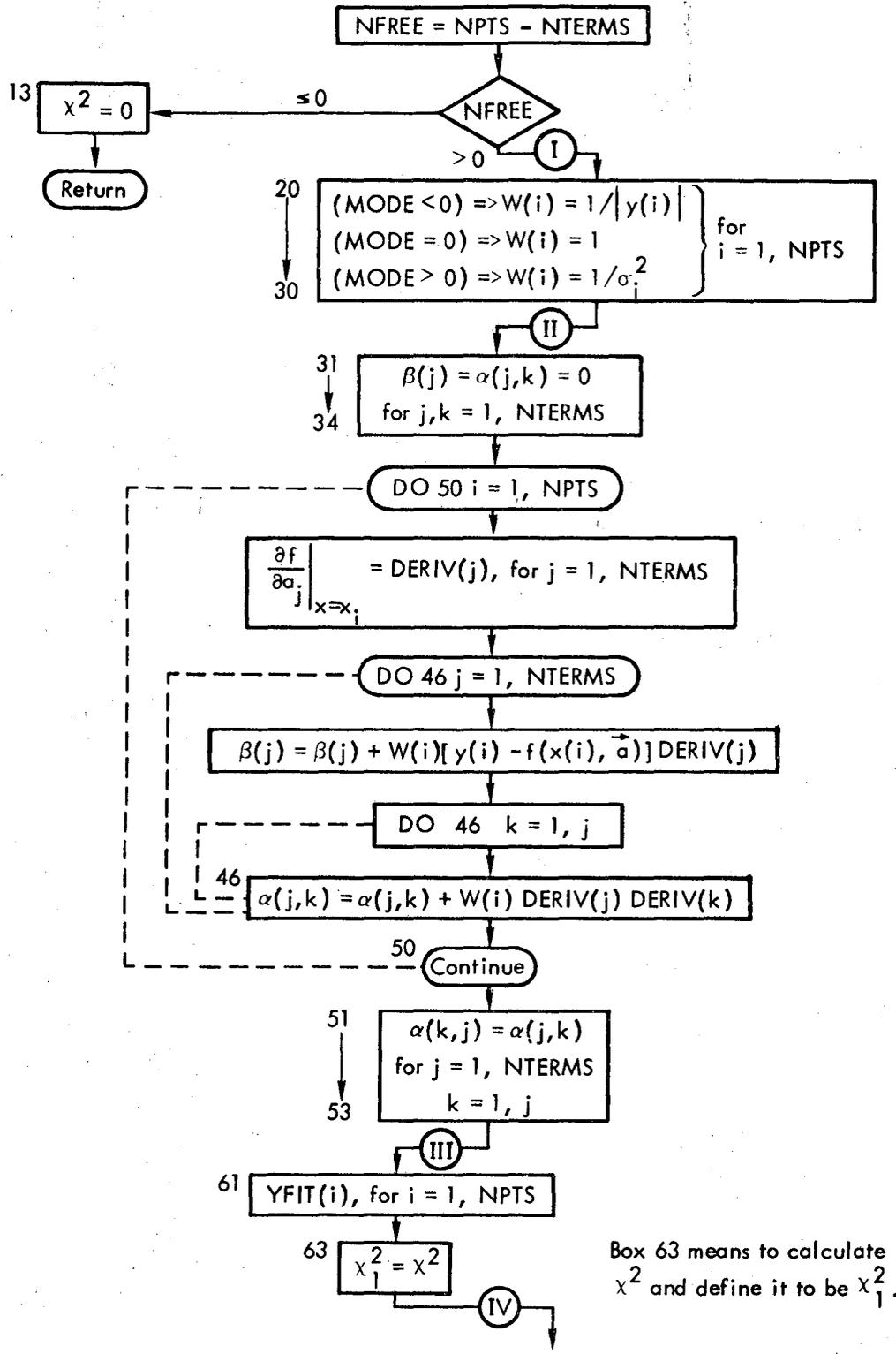Fig. 6.3. Outline flow chart for the subroutine CURFIT.

NFREE = NPTS − NTERMS

13 $\chi^2 = 0$ ⟵ ≤0 ⟵ NFREE

(Return)

> 0 (I)

20
30
$$(MODE < 0) \Rightarrow W(i) = 1/\left| y(i) \right|$$
$$(MODE = 0) \Rightarrow W(i) = 1$$
$$(MODE > 0) \Rightarrow W(i) = 1/\sigma_i^2$$
for i = 1, NPTS

(II)

31
34
$$\beta(j) = \alpha(j,k) = 0$$
for j,k = 1, NTERMS

DO 50 i = 1, NPTS

$$\left.\frac{\partial f}{\partial a_j}\right|_{x=x_i} = DERIV(j),\ \text{for } j = 1,\ NTERMS$$

DO 46 j = 1, NTERMS

$$\beta(j) = \beta(j) + W(i)[\, y(i) - f(x(i), \vec{a})\,]\, DERIV(j)$$

DO 46 k = 1, j

46 $$\alpha(j,k) = \alpha(j,k) + W(i)\, DERIV(j)\, DERIV(k)$$

50 (Continue)

51
53
$$\alpha(k,j) = \alpha(j,k)$$
for j = 1, NTERMS
k = 1, j

(III)

61 YFIT(i), for i = 1, NPTS

63 $$\chi_1^2 = \chi^2$$

(IV)

Box 63 means to calculate
χ^2 and define it to be χ_1^2.

Fig. 6.4. Detailed flow chart for the subroutine CURFIT (see Ref. 1, p. 237).
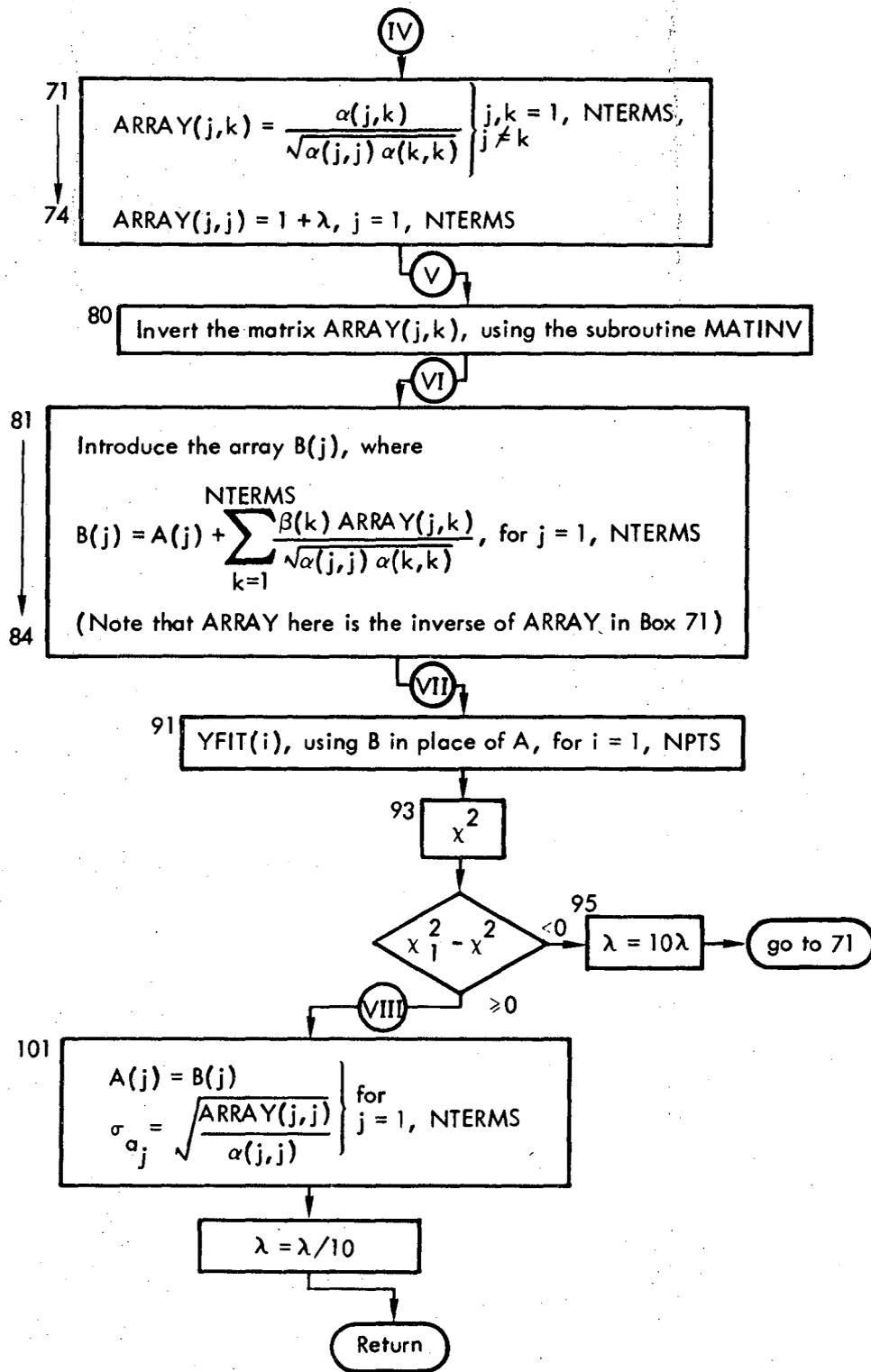
Fig. 6.4. Continued.

Acknowledgments

References

1. P. R. Bevington, <u>Data Reduction and Error Analysis for the Physical Sciences</u> (New York, McGraw-Hill, 1969).

2. M. Abramowitz and I. A. Stegun, <u>Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables</u> (New York, Dover, 1965).

3. E. M. Pugh and G. H. Winslow, <u>The Analysis of Physical Measurements</u> (Reading, Mass., Addison-Wesley, 1966).

4. R. L. Fox, <u>Optimization Methods for Engineering Design</u> (Reading, Mass., Addison-Wesley, 1971).

5. T. C. Varljen, Lawrence Livermore Laboratory, Internal Document, CIC Report E2.2-005 (1967). Readers outside the Laboratory who desire further information on LLL internal documents should address their inquiries to the Technical Information Department, Lawrence Livermore Laboratory, Livermore, California 94550.

6. M. J. D. Powell, "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," <u>Comput. J.</u> <u>7</u>, 155-162 (1964).

7. F. N. Fritsch, Lawrence Livermore Laboratory, Internal Document, LLL Numerical Analysis Note No. 24 (1971).

8. M. J. D. Powell, "An Iterative Method for Finding Stationary Values of a Function of Several Variables," <u>Comput. J. 5</u>, 147-151 (1962)

9. F. S. Acton, <u>Numerical Methods That Work</u> (New York, Harper and Row, 1970).

10. R. E. von Holdt, Lawrence Livermore Laboratory, Internal Document, CIC report F3-001 (1969).

11. K. Levenberg, " A Method for the Solution of Certain Non-Linear Problems in Least Squares," <u>Quart. Appl. Math. 2</u>, 164-168 (1944).

12. D. W. Marquardt, "An Algorithm for Least Squares Estimation of Nonlinear Parameters," <u>J. Soc. Indust. Appl. Math. 2</u>, 431-441 (1963).

13. R. Fletcher, A Modified Marquardt Subroutine for Non-Linear Least Squares AERE-R 6799 (Atomic Energy Research Establishment, Harwell, England, 1971).

14. F. L. Bauer, "Optimally Scaled Matrices," Num. Math. 5, 73-87 (1963).

Appendix A
Expressions Involving Squares of Sums

Writing out expressions involving squares of sums is illustrated by the following examples:

$$\text{I.} \quad \left(\sum_{i=1}^{n} b_i \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} b_j b_i . \qquad (A.1)$$

The right-hand side of Eq. (A.1) is the sum of all pairs of the b's.

$$\text{II.} \quad \sum_{j=1}^{m} \left(\sum_{i=1}^{n} a_i b_{ij} \right)^2 = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \sum_{k=1}^{n} a_i a_k b_{ij} b_{kj} \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n} a_i a_k \left(\sum_{j=1}^{m} b_{ij} b_{kj} \right). \qquad (A.2)$$

$$\text{III.} \quad \sum_{j=1}^{m} \left[\sum_{i=1}^{n} a_i (b_{ij} + c_i) \right]^2$$

$$\sum_{j=1}^{m} \left[\sum_{i=1}^{n} \sum_{k=1}^{n} a_i a_k (b_{ij} + c_i)(b_{kj} + c_k) \right]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n} a_i a_k \left[\sum_{j=1}^{m} (b_{ij} + c_i)(b_{kj} + c_k) \right]. \qquad (A.3)$$

Appendix B
The Geometry of Parameter Changes

The change of variable from $\vec{a}$ to $\vec{b}$ is a cause of some confusion because the gradient of χ^2 with respect to $\vec{b}$ points in a different direction than the gradient with respect to $\vec{a}$, and it is the latter that we are interested in.

The difficulty is more apparent than real. Starting with the fitting function $f(x,\vec{a})$, consider a continuous mapping ξ from $\vec{a}$-space to $\vec{b}$-space, so that $\vec{b} = \xi(\vec{a})$. This mapping must be such that for each $\vec{a}$ there corresponds one and only one $\vec{b}$ and the mapping goes both ways. This mapping need not be linear, as in Eq. (4.1). Then for any $\vec{a}$ there is a corresponding $\vec{b}$ such that

$$\vec{a} = \xi^{-1}(\vec{b}), \tag{B.1}$$

where ξ^{-1} is the inverse mapping. Replacing $\vec{a}$ in $f(x,\vec{a})$ by Eq. (B.1), one gets a function $g(x,\vec{b})$ such that

$$f(x,\vec{a}) = g(x,\vec{b}). \tag{B.2}$$

Denote by χ^2_a the function χ^2 with respect to $f(x,\vec{a})$, and denote by χ^2_b the function χ^2 with respect to $g(x,\vec{b})$.

Suppose the contour surface of χ^2_a (in two dimensions) is as in Fig. B.1. In $\vec{b}$-space, the contour surface χ^2_b might look like the contour curves of Fig. B.2. The point P in Fig. B.1 corresponds to point Q in Fig. B.2, in the sense that $\chi^2_a(P) = \chi^2_b(Q)$.

The gradient of χ^2_a at P will, in general, point in a different direction and be of a different length than the gradient of χ^2_b at Q. The
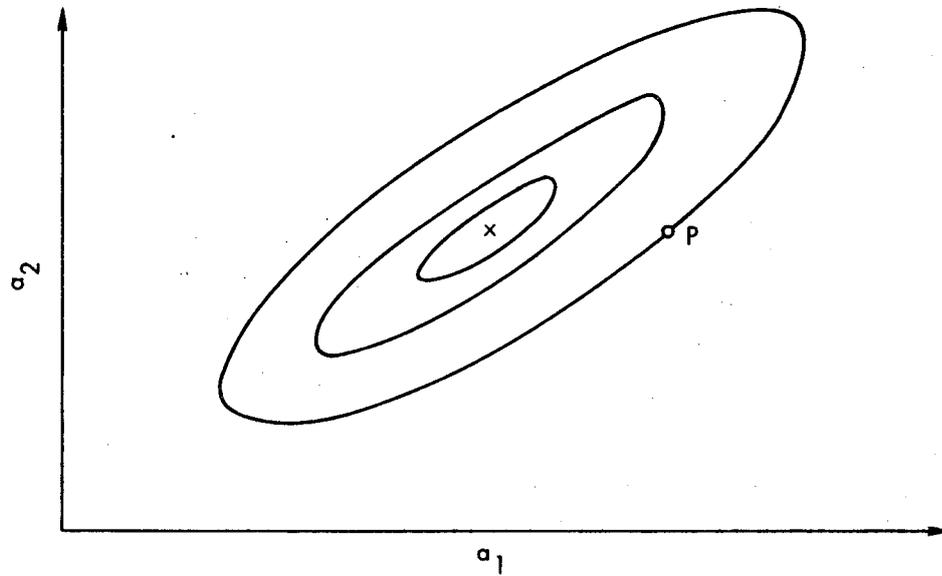
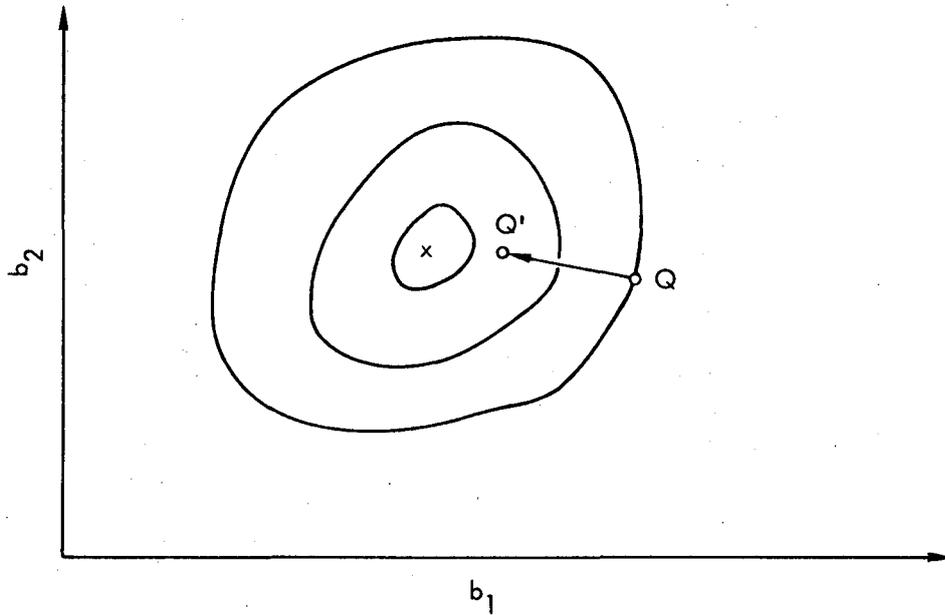Fig. B.1. The contour surface of χ^2_a.

Fig. B.2. The contour surface of χ_b^2.

corresponding minimum points in both figures are indicated by x's.

Suppose we start at point Q in Fig. B.2 and proceed some distance along the negative gradient to a point Q', so that

$$\chi_b^2(Q') \quad < \quad \chi_b^2(Q). \tag{B.3}$$

Corresponding to the point Q' in $\vec{b}$-space, there is a point P' in $\vec{a}$-space, such that

$$\chi_a^2(P') = \chi_b^2(Q'). \tag{B.4}$$

For the general nonlinear transformation ξ, lines in Fig. B.2 correspond to curves in Fig. B.1. For the linear transformation, Eq. (4.1), lines in Fig. B.2 correspond to lines in Fig. B.1. But, even in the linear case of (4.1), unless $\xi_i = 1$ for all i, P' will not lie on the line through the gradient at P. However, the important point to see is that, in any case,

$$\chi_a^2(P') < \chi_a^2(P) . \tag{B.5}$$

Appendix C

Nonlinear Least Squares, the Gauss Method, Using MLR

Given:

(1) N data points $(x_i y_i)$.

(2) Standard deviations σ_i, associated with the y_i's (we assume that the x_i's are correct).

(3) A function $f(x, \vec{a})$.

To solve:

The equation $y_i \approx f(x_i, \vec{a})$ in the least squares sense ($\vec{a}$ is a vector of parameters with n elements). We have seen that the Gauss method requires that one find $\delta \vec{a}$ that minimizes X^2, given by Eq. (5.2). When using MLR, one writes the following (overdetermined) equations that result from setting $\chi^2 = 0$ in Eq. (5.2):

$$\frac{y_i}{\sigma_i} - \frac{f(x_i, \vec{a}_0)}{\sigma_i} + \sum_{j=1}^{n} \frac{1}{\sigma_i} \frac{\partial f}{\partial a_j} (x_i \vec{a}_0) \delta a_j \, ,$$

or

$$\frac{y_i}{\sigma_i} - \frac{f(x_i, \vec{a}_0)}{\sigma_i} = \sum_{j=1}^{n} \frac{1}{\sigma_i} \frac{\partial f}{\partial a_j} (x_i, \vec{a}_0) \delta a_j \, .$$

This last equation can be put in the vector-matrix form

$$AX = Y, \tag{C.1}$$

where A is a matrix with N rows and n columns,

$$a_{ij} = \frac{1}{\sigma_i} \frac{\partial f}{\partial a_j} (x_i, \vec{a}_0). \tag{C.2}$$

X is the n-vector of unknowns:

$$X_i = \delta a_i. \tag{C.3}$$

Y is the N-vector:

$$Y_i = \frac{y_i}{\sigma_i} - \frac{1}{\sigma_i} \; f(x_i, \vec{a}_0). \tag{C.4}$$

When using MLR to solve for X, one gives MLR the quantities N, n, A, Y. Then, MLR gives back a vector X that is the solution of the overdetermined system, in the sense that the length of the vector

$$R = AX - Y \tag{C.5}$$

is minimized. It is easy to show that this gives the least squares solution for $\delta \vec{a} = X$.

The Normal Equation

We are trying to find the vector X that gives an approximate solution to Eq. (C.1). If we multiply both sides of this equation by A^T, the transpose of the matrix A (recall that $a^T_{ij} = a_{ji}$), we get the <u>normal</u> <u>equation</u>

$$A^T AX = A^T Y . \tag{C.6}$$

This equation is a set of n equations in n unknowns. If we define the matrix B and vector Z as

$$\left.\begin{array}{l} B \equiv A^T \\ Z \equiv A^T Y \end{array}\right\}, \tag{C.7}$$

then Eq. (C.6) becomes

$$BX = Z, \qquad\qquad\qquad\qquad (C.8)$$

and the solution for X is

$$X = B^{-1} Z. \qquad\qquad\qquad\qquad (C.9)$$

This is the sort of thing that is always done in Bevington's codes, but there are distinct numerical advantages of solving Eq. (C.1) in the sense of Eq. (C.5).

Recall from Eq. (2.22) that the variances and covariances come from B^{-1}. In Eq. (2.2) the symbol ε is the same as B^{-1} here. The matrix B^{-1} is (optionally) available from MLR.

Distribution

External Distribution Continued

B. K. Swartz
Los Alamos Scientific Laboratory
Los Alamos, New Mexico

C. W. Gear
University of Illinois
Urbana, Illinois

J. F. Traub
Carnegie-Mellon University
Pittsburgh, Pennsylvania

RE-RUN 6/18/74

TID Library   100

R. J. Thompson
D. A. Young
Sandia Laboratories
Albuquerque, New Mexico

J.N. Rogers
Sandia Laboratories
Livermore, California

Technical Information Center
Oak Ridge, Tennessee   2

LOF/eap

-77-