ON THE METHOD OF LEAST SQUARES. II

WILLIAM H. JEFFERYS
Department of Astronomy, University of Texas at Austin, Austin, Texas 78712
Received 30 July 1980; revised 22 September 1980

ABSTRACT

The general method of least squares described in the first paper of this series is expanded in two ways. Certain useful covariance matrices related to the solution are calculated and their interpretation is given. In addition, improved methods of solving the normal equations related to those of Marquardt (1963) and Fletcher and Powell (1963) are developed for this approach. They may converge, but with greater speed, when the method of the first paper either diverges or converges slowly.

I. INTRODUCTION

In an earlier paper (Jefferys 1980, hereafter referred to as Paper I), a method was proposed for solving a very general class of problems in least squares (see also Britt and Leucke 1973, a paper which came to light after the publication of Paper I, in which a more restricted algorithm, which my method generalizes, was presented). In this paper, I accomplish two things: First, I calculate a number of covariance matrices related to the adjustment which are of importance in interpreting results obtained by this method. Second, I recognize the fact (alluded to at the end of Paper I) that the method proposed there for solving the normal equations—namely, Newton's method—may not always converge, and I propose generalizations of the method of steepest descent, of Marquardt's algorithm (Marquardt 1963), and of the Fletcher–Powell algorithm (Fletcher and Powell 1963) appropriate to this algorithm, which may converge in cases where Newton's method diverges or converges slowly.

II. SUMMARY OF PAPER I

In Paper I it was shown that if a vector of observations $\mathring{\mathbf{x}}$ and a vector of parameters $\mathbf{a}$ are given which are supposed to satisfy the vector equations of condition

$$\begin{aligned} \mathbf{f}(\mathring{\mathbf{x}} + \mathbf{v},\mathbf{a}) &= \mathbf{0}, \\ \mathbf{g}(\mathbf{a}) &= \mathbf{0}, \end{aligned} \tag{1}$$

where $\mathbf{x} = \mathring{\mathbf{x}} + \mathbf{v}$ would have been the *actual* observations if there were no measurement errors, and therefore $\mathbf{v}$ is the vector of the *actual* residuals, and if the covariance matrix

$$\boldsymbol{\sigma} = \langle \mathbf{v}\mathbf{v}^t \rangle \tag{2}$$

of the observations is known, then the correct least-squares equations of condition are given by Eqs. (3) [Eqs. (10a)–(10d) of Paper I]:

$$\boldsymbol{\sigma}^{-1}\hat{\mathbf{v}} + \mathbf{f}_{\hat{\mathbf{x}}}{}^t(\hat{\mathbf{x}},\hat{\mathbf{a}})\hat{\boldsymbol{\mu}} = \mathbf{0}, \tag{3a}$$

$$\mathbf{f}_{\hat{\mathbf{a}}}{}^t(\hat{\mathbf{x}},\hat{\mathbf{a}})\hat{\boldsymbol{\mu}} + \mathbf{g}_{\hat{\mathbf{a}}}{}^t(\hat{\mathbf{a}})\hat{\boldsymbol{\lambda}} = \mathbf{0}, \tag{3b}$$

$$\mathbf{f}(\hat{\mathbf{x}},\hat{\mathbf{a}}) = \mathbf{0}, \tag{3c}$$

$$\mathbf{g}(\hat{\mathbf{a}}) = \mathbf{0}, \tag{3d}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\lambda}}$ are Lagrange multipliers, the caret denotes the least-squares estimates of the indicated vectors, and the subscripts denote the Jacobian matrix of partial derivatives with respect to that variable. Equations (3) are to be solved by any method, for example, Newton's method (as given in Paper I) or one of the methods developed in this paper in Secs. V–VII.

III. COVARIANCE MATRICES

Once the solution of Eqs. (3) is known, a number of covariance matrices related to the solution can be calculated. Some of these are of particular importance. (For the importance and use of the covariance matrix, the reader is referred to standard texts, e.g., Meyer 1975, pp. 416–417.)

To calculate them, it is necessary to expand Eqs. (3) about the *actual* (and hence unknown) values of $\mathbf{a}$ and $\mathbf{x}$ in powers of

$$\hat{\mathbf{d}} = \hat{\mathbf{a}} - \mathbf{a}$$

and

$$\hat{\mathbf{v}} - \mathbf{v} = (\hat{\mathbf{x}} - \mathring{\mathbf{x}}) + (\mathring{\mathbf{x}} - \mathbf{x}) = (\hat{\mathbf{x}} - \mathbf{x}). \tag{4}$$

One obtains Eqs. (5):

$$\boldsymbol{\sigma}^{-1}\hat{\mathbf{v}} + \mathbf{f}_{\mathbf{x}}{}^t(\mathbf{x},\mathbf{a})\hat{\boldsymbol{\mu}} = \mathbf{0}, \tag{5a}$$

$$\mathbf{f}_{\mathbf{a}}{}^t\hat{\boldsymbol{\mu}} + \mathbf{g}_{\mathbf{a}}{}^t\hat{\boldsymbol{\lambda}} = \mathbf{0}, \tag{5b}$$

$$\mathbf{f}_{\mathbf{x}}(\hat{\mathbf{v}} - \mathbf{v}) + \mathbf{f}_{\mathbf{a}}\hat{\mathbf{d}} = \mathbf{0}, \tag{5c}$$

$$\mathbf{g}_{\mathbf{a}}\hat{\mathbf{d}} = \mathbf{0}, \tag{5d}$$

where terms of second order have been discarded and use has been made of Eqs. (1).

Manipulations similar to those of Paper I (and which

 0004-6256/81/010149-07$00.90

need not be repeated here) lead to the basic relationships

$$\begin{pmatrix} \mathbf{f_a}^t \mathbf{W} \mathbf{f_a} & \mathbf{g_a}^t \\ \mathbf{g_a} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \mathbf{v} \\ \mathbf{0} \end{pmatrix}, \qquad (6)$$

$$\hat{\mathbf{v}} = \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \mathbf{v} - \mathbf{f_a} \hat{\mathbf{d}}),$$

where

$$\mathbf{W}^{-1} = \mathbf{f_x} \sigma \mathbf{f_x}^t. \qquad (7)$$

Solving Eqs. (6) for $\hat{\mathbf{d}}$, I obtain

$$\hat{\mathbf{d}} = \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \mathbf{v}, \qquad (8)$$

where the inverse of the matrix in Eq. (6) is given by

$$\begin{pmatrix} \alpha & \beta \\ \beta^t & \gamma \end{pmatrix} \qquad (9)$$

Then the covariance matrix of the parameters is given by

$$\begin{aligned} \sigma_{\hat{\mathbf{d}}\hat{\mathbf{d}}^t} &= \langle \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \mathbf{v} \mathbf{v}^t \mathbf{f_x}^t \mathbf{W} \mathbf{f_a} \alpha \rangle \\ &= \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \langle \mathbf{v} \mathbf{v}^t \rangle \mathbf{f_x}^t \mathbf{W} \mathbf{f_a} \alpha \\ &= \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \sigma \mathbf{f_x}^t \mathbf{W} \mathbf{f_a} \alpha \\ &= \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{W}^{-1} \mathbf{W} \mathbf{f_a} \alpha \\ &= \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_a} \alpha = \alpha. \end{aligned} \qquad (10)$$

This last step was demonstrated in Paper I. (Note that the proof given in Paper I of this formula used $\langle \hat{\mathbf{v}} \hat{\mathbf{v}}^t \rangle$ instead of $\langle \mathbf{v} \mathbf{v}^t \rangle$ and is therefore not quite correct. The final formula given in Paper I for the covariance matrix of the parameters is correct, however.)

Next, I have

$$\begin{aligned} \sigma_{\hat{\mathbf{d}}\mathbf{v}^t} &= \langle \hat{\mathbf{d}} \mathbf{v}^t \rangle = \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \langle \mathbf{v} \mathbf{v}^t \rangle \\ &= \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \sigma, \end{aligned} \qquad (11)$$

which makes it possible to calculate the important result

$$\begin{aligned} \sigma_{\hat{\mathbf{d}}\hat{\mathbf{v}}^t} &= \langle \hat{\mathbf{d}} \hat{\mathbf{v}}^t \rangle = (\langle \hat{\mathbf{d}} \mathbf{v}^t \rangle \mathbf{f_x}^t - \langle \hat{\mathbf{d}} \hat{\mathbf{d}}^t \rangle \mathbf{f_a}^t) \mathbf{W} \mathbf{f_x} \sigma \\ &= (\alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \sigma \mathbf{f_x}^t - \alpha \mathbf{f_a}^t) \mathbf{W} \mathbf{f_x} \sigma \\ &= \mathbf{0}, \end{aligned} \qquad (12)$$

that is, the *residuals and the parameter corrections are uncorrelated.* In deriving the last step of Eq. (12), use has been made of Eq. (7).

Next,

$$\begin{aligned} \sigma_{\hat{\mathbf{v}}\mathbf{v}^t} &= \langle \hat{\mathbf{v}} \mathbf{v}^t \rangle \\ &= \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \langle \mathbf{v} \mathbf{v}^t \rangle - \mathbf{f_a} \langle \hat{\mathbf{d}} \mathbf{v}^t \rangle) \\ &= \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \sigma - \mathbf{f_a} \alpha \mathbf{f_a}^t \mathbf{W} \mathbf{f_x} \sigma) \\ &= \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \sigma \mathbf{f_x}^t - \mathbf{f_a} \alpha \mathbf{f_a}^t) \mathbf{W} \mathbf{f_x} \sigma, \end{aligned} \qquad (13)$$

again using Eq. (7). We also have

$$\begin{aligned} \sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t} &= \langle \hat{\mathbf{v}} \hat{\mathbf{v}}^t \rangle \\ &= \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \langle \mathbf{v} \hat{\mathbf{v}}^t \rangle - \mathbf{f_a} \langle \hat{\mathbf{d}} \hat{\mathbf{v}}^t \rangle) \\ &= \sigma \mathbf{f_x}^t \mathbf{W} \mathbf{f_x} [\sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \sigma \mathbf{f_x}^t - \mathbf{f_a} \alpha \mathbf{f_a}^t) \mathbf{W} \mathbf{f_x} \sigma] \\ &= \sigma \mathbf{f_x}^t \mathbf{W} (\mathbf{f_x} \sigma \mathbf{f_x}^t - \mathbf{f_a} \alpha \mathbf{f_a}^t) \mathbf{W} \mathbf{f_x} \sigma \\ &= \sigma_{\hat{\mathbf{v}}\mathbf{v}^t} = \sigma_{\mathbf{v}\hat{\mathbf{v}}^t}, \end{aligned} \qquad (14)$$

where I have used Eqs. (7) and (12). This is a very important result which I discuss in more detail below.

Finally, the covariance matrix of the *adjusted observations* $\hat{\mathbf{x}}$ is

$$\begin{aligned} \sigma_{\hat{\mathbf{x}}\hat{\mathbf{x}}^t} &= \langle (\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^t \rangle \\ &= \langle (\hat{\mathbf{v}} - \mathbf{v})(\hat{\mathbf{v}} - \mathbf{v})^t \rangle \\ &= \sigma - 2\sigma_{\hat{\mathbf{v}}\mathbf{v}^t} + \sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t} \\ &= \sigma - \sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t}. \end{aligned} \qquad (15)$$

This last result allows one to estimate the improvement that has been accomplished by the adjustment. Brown (1955) has calculated analogs to all of these expressions for his approximate treatment of least squares, under more restrictive assumptions than those which I have adopted (in particular, he assumes that the matrix analogous to $\mathbf{f_a}^t \mathbf{W} \mathbf{f_a}$ in his scheme is nonsingular, which is not true in general).

The result for $\sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t}$ given by Eq. (14) is particularly interesting. It turns out that the second term in the parentheses is dominated by the first, being $O(1/N)$ relative to the first, where N is the number of observations. If we neglect it, we obtain the simplified expression

$$\sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t} \cong \sigma \mathbf{f_x}^t \mathbf{W} \mathbf{f_x} \sigma, \qquad (16)$$

which is somewhat easier to evaluate. The reason for the dominance of the first term is that it is proportional to the natural scatter of the observations, which is independent of N, whereas the second term is proportional to the variance of the estimated parameters, which naturally approaches zero with $1/N$ as the number of observations becomes arbitrarily large and the parameters become better determined.

The covariance matrices I have computed depend upon the actual values of the vectors $\mathbf{x}$ and $\mathbf{a}$, which, of course, are unavailable. In practice, they must be calculated using the estimated values $\hat{\mathbf{x}}$ and $\hat{\mathbf{a}}$. The error committed is $O(\mathbf{v}^3)$.

IV. REMARKS ON THE COVARIANCE MATRICES

The importance of understanding the covariance matrices of Sec. III, particularly $\sigma_{\hat{\mathbf{v}}\hat{\mathbf{v}}^t}$, was brought home to the author when two people working independently to apply the method of Paper I communicated some puzzling results (Leach 1980; Peters 1980). They had tested the method on artificial data and found that the calculated residuals did not behave as they expected. To understand what happened, let us consider the example of fitting a straight line, which was discussed in Paper I. To simplify it further, we assume that the observations of the quantities in the ordinate t and abscissa y are

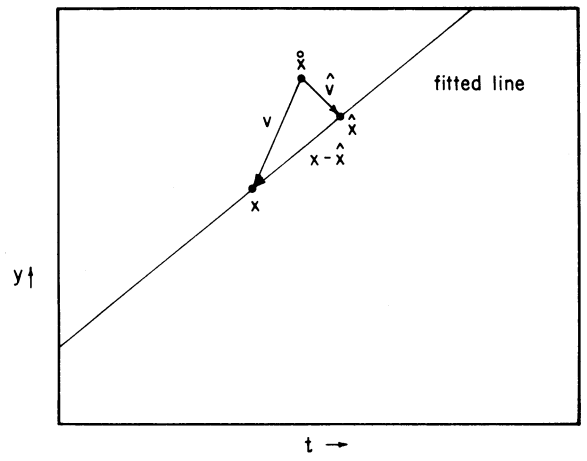

FIG. 1. The true ($\mathbf{x}$), observed ($\mathring{\mathbf{x}}$), and estimated ($\hat{\mathbf{x}}$) data points for a single observation. The component $\mathbf{x} - \hat{\mathbf{x}}$ of the vector $\mathbf{v}$ cannot be estimated, so the average residual is reduced, and a correlation between y and t residuals is induced. The point $\mathbf{x}$ actually lies off the fitted line by a small amount, owing to the second term of Eq. (14), but this effect is minimal.

uncorrelated. Then [using Eq. (16)] one calculates the covariance matrix of the residuals between the y's and t's as

$$\left\langle \begin{pmatrix} v_y \\ v_t \end{pmatrix} \begin{pmatrix} v_y \\ v_t \end{pmatrix}^t \right\rangle = \frac{1}{\bar{\sigma}} \begin{pmatrix} -\sigma_{yy}^2 & -\beta\sigma_{yy}\sigma_{tt} \\ -\beta\sigma_{yy}\sigma_{tt} & \beta^2\sigma_{tt}^2 \end{pmatrix} \tag{17}$$

for each observation of a point (t,y), in an obvious notation. σ_{yy} and σ_{tt} are the variances of the observations of y and t (as in Paper I), β is the slope of the line, and

$$\bar{\sigma} = \sigma_{yy} + \beta^2\sigma_{tt}. \tag{18}$$

Leach considered (among others) the case $\beta = 1$, $\sigma_{yy} = \sigma_{tt} = \sigma$, and found that (a) the y and t *residuals* were *correlated,* even though the (artificial) elemental observations were not, and (b) the *variances* of the calculated residuals were only half of the variances which the elemental observations possessed. Equations (17) and (18) allow us to understand this, for under Leach's assumptions, the covariance matrix of the residuals of an observational point is just

$$\frac{\sigma}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{19}$$

That this is intuitively correct can be seen from Fig. 1, which shows the plotted points $\mathbf{x} = (t,y)$, $\mathring{\mathbf{x}} = (\mathring{t},\mathring{y})$, and $\hat{\mathbf{x}} = (\hat{t},\hat{y})$ (respectively, the actual, observed, and adjusted points). In the special case discussed, the least-squares algorithm minimizes (for any slope) the mean-square distance from the fitted line to the observational points. Thus the adjusted point is obtained by dropping a perpendicular from the observed point to the fitted line. This distance is necessarily less than (or at most equal to) the *actual* residual $(\mathring{t} - t, \mathring{y} - y)$, and on average is less by the factor $1/\sqrt{2}$. Furthermore, for unity slope one sees that the calculated t residual is equal in magnitude and opposite in sign to the calculated y residual, and so the origin of the correlation is also easily understood.

In the more general case of least squares, it can be shown [see Eqs. (26a) and (26b), in which $\mathbf{f}(\hat{\mathbf{x}},\hat{\mathbf{a}})$ should be set equal to $\mathbf{0}$] that the final vector $\hat{\mathbf{v}}$ satisfies the condition $\hat{\mathbf{v}} = \mathbf{P}\hat{\mathbf{v}}$, where the matrix $\mathbf{P} = \boldsymbol{\sigma}\mathbf{f}_{\hat{\mathbf{x}}}{}^t \mathbf{W}\mathbf{f}_{\hat{\mathbf{x}}}$ is a *projection operator.* It has the effect of removing a component from $\mathbf{v}$ which lies in the surface $\mathbf{f}(\hat{\mathbf{x}},\hat{\mathbf{a}}) = \mathbf{0}$. This component is essentially "unknowable," because the constraints would continue to be satisfied if a vector satisfying $\mathbf{P}\hat{\mathbf{v}} = \mathbf{0}$ were to be added to $\hat{\mathbf{x}}$. Our inability to estimate this vector stems from our ignorance of the *true* error vector $\mathbf{v}$. Obviously, if we could estimate $\mathbf{v}$ there would be no need to resort to least squares! Basically, then, the method of least squares simply ignores this component, since it has no means of estimating it. This in turn leads both to the observed correlations among the residuals and to the reduction in the size of the residuals by an amount due to the unobservable component (Fig. 1).

V. THE METHOD OF STEEPEST DESCENT

Although rapidly convergent if the initial guess is good, Newton's method often fails to yield a solution in least-squares problems, particularly if the initial estimate $\mathring{\mathbf{a}}$ of the vector $\mathbf{a}$ is greatly in error. In such cases, the method of steepest descent may be used, although it too has its drawbacks, principally in its typically slow convergence. Marquardt's algorithm, to be discussed in Sec. VI, combines the best of the two methods by "interpolating" between them. It combines the certain improvement of the steepest descent method when far from the solution, with the rapid convergence of Newton's method when close to it. The Fletcher–Powell algorithm (Fletcher and Powell 1963) is another approach which I will also discuss. To adapt these sophisticated algorithms to my approach, we shall first have to consider the method of steepest descent as it applies to the new algorithm.

The basic idea of the method of steepest descent is to step to the next point in a sequence of better approximations to the solution by moving in the direction of the negative of the gradient of

$$S_0 = \frac{1}{2}\hat{\mathbf{v}}^t \boldsymbol{\sigma}^{-1}\hat{\mathbf{v}}.$$

If the step is not too large, the value of S_0 must necessarily decrease from one iteration to the next.

In our case, where S_0 depends *implicitly* rather than explicitly on the parameters, it is not immediately obvious how to compute S_0 or its gradient. (Britt and Luecke, in fact, declare that it cannot be done even for their less general method.) Nevertheless, it is possible to make an excellent approximation, as the following considerations show.

Suppose we were to pretend that the present value of

$\hat{\mathbf{a}}$ at any iteration is the true value of $\hat{\mathbf{a}}$. Even though this assumption is false, in general, it makes it possible to estimate the vector $\hat{\mathbf{v}}$ as a function of $\hat{\mathbf{a}}$. Suppose therefore we have an estimate $\hat{\mathbf{a}}$ of $\mathbf{a}$ and $\hat{\mathbf{v}}_0$ of $\mathbf{v}$ and we wish to obtain a better estimate of $\hat{\mathbf{v}}$, *assuming* $\hat{\mathbf{a}}$ *to be correct.* For the moment we assume that $\mathbf{g}(\hat{\mathbf{a}}) = \mathbf{0}$ has been satisfied. [I show below how to eliminate this assumption; see Eq. (29).] Then minimizing

$$S = \frac{1}{2}\hat{\mathbf{v}}^t \boldsymbol{\sigma}^{-1}\hat{\mathbf{v}} + \mathbf{f}^t\hat{\boldsymbol{\mu}}$$

relative to the remaining variables $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\mu}}$ yields

$$\mathbf{f}(\mathring{\mathbf{x}} + \hat{\mathbf{v}},\hat{\mathbf{a}}) = \mathbf{0}, \tag{20a}$$

$$\boldsymbol{\sigma}^{-1}\hat{\mathbf{v}} + \mathbf{f}_{\hat{\mathbf{x}}}{}^t\hat{\boldsymbol{\mu}} = \mathbf{0}. \tag{20b}$$

Letting $\hat{\mathbf{v}} = \hat{\mathbf{v}}_0 + \hat{\boldsymbol{\epsilon}}$ and expanding (20a) in powers of $\hat{\boldsymbol{\epsilon}}$, one arrives, after linearization and some manipulation, at the expression

$$\hat{\mathbf{v}} = -\boldsymbol{\sigma}\mathbf{f}_{\hat{\mathbf{x}}_0}{}^t\mathbf{W}[\mathbf{f}(\hat{\mathbf{x}}_0,\hat{\mathbf{a}}) - \mathbf{f}_{\hat{\mathbf{x}}_0}\hat{\mathbf{v}}_0], \tag{21}$$

where $\hat{\mathbf{x}}_0 = \mathring{\mathbf{x}} + \hat{\mathbf{v}}_0$. This equation may be iterated, if desired, to obtain a definitive value of $\hat{\mathbf{v}}$. Thus to each choice of parameters $\hat{\mathbf{a}}$, there is a corresponding $\hat{\mathbf{v}}$, and we can write $\hat{\mathbf{v}} = \hat{\mathbf{v}}(\hat{\mathbf{a}})$.

I note in passing that while it appears at first sight that Eq. (21) involves solving simultaneously for many unknown components of the vector $\hat{\mathbf{v}}$, this is only apparent. Because of the fact that in typical least-squares problems, the matrices $\boldsymbol{\sigma}$, $\mathbf{f}_{\hat{\mathbf{x}}}$, and $\mathbf{W}$ are sparse, reflecting the fact that typically only a few observations enter into each equation of condition, and observations are not highly correlated, Eq. (21) normally consists of a number of independent equations, each involving only a few of the components of $\hat{\mathbf{v}}$.

Now the interesting thing about Eq. (21) is that, although formally dependent on $\hat{\mathbf{v}}_0$ in first order, it is actually rather insensitive to the actual value of $\hat{\mathbf{v}}_0$ used, since it actually is *quadratic* in $\hat{\mathbf{v}}_0$. To see this, note that

$$\begin{aligned}&\mathbf{f}(\hat{\mathbf{x}}_0,\hat{\mathbf{a}}) - \mathbf{f}_{\hat{\mathbf{x}}_0}(\hat{\mathbf{x}}_0 - \mathring{\mathbf{x}})\\ &\quad = \mathbf{f}(\mathring{\mathbf{x}},\hat{\mathbf{a}}) + O(\hat{\mathbf{v}}_0^2),\end{aligned} \tag{22}$$

where $\mathring{\mathbf{x}}$ is the (constant) observation vector.

Hence the expression

$$\begin{aligned}S_0 &= \frac{1}{2}\hat{\mathbf{v}}^t\boldsymbol{\sigma}^{-1}\hat{\mathbf{v}}\\ &= \frac{1}{2}\hat{\boldsymbol{\phi}}^t\mathbf{W}\mathbf{f}_{\hat{\mathbf{x}}}\boldsymbol{\sigma}\boldsymbol{\sigma}^{-1}\boldsymbol{\sigma}\mathbf{f}_{\hat{\mathbf{x}}}{}^t\mathbf{W}\hat{\boldsymbol{\phi}}\\ &= \frac{1}{2}\hat{\boldsymbol{\phi}}^t\mathbf{W}\hat{\boldsymbol{\phi}},\end{aligned} \tag{23}$$

where

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}(\hat{\mathbf{x}}(\hat{\mathbf{a}}),\hat{\mathbf{a}}) = \mathbf{f}(\hat{\mathbf{x}},\hat{\mathbf{a}}) - \mathbf{f}_{\hat{\mathbf{x}}}\hat{\mathbf{v}} \cong \mathbf{f}(\mathring{\mathbf{x}},\hat{\mathbf{a}}) \tag{24}$$

was defined in Paper I, depends only insensitively on the value of $\hat{\mathbf{v}}$ chosen to calculate it. An approximate value of $\hat{\mathbf{v}}$ calculated from Eq. (21) will suffice. We can therefore use S_0 from Eq. (23) to estimate the negative gradient $\hat{\boldsymbol{\delta}}_g$ of S_0 with respect to $\hat{\mathbf{a}}$:

$$\begin{aligned}-\hat{\boldsymbol{\delta}}_g &= +\frac{\partial S_0}{\partial\hat{\mathbf{a}}} = \frac{\partial\hat{\boldsymbol{\phi}}^t}{\partial\hat{\mathbf{a}}}\mathbf{W}\hat{\boldsymbol{\phi}} + \frac{1}{2}\hat{\boldsymbol{\phi}}^t\frac{\partial\mathbf{W}}{\partial\hat{\mathbf{a}}}\hat{\boldsymbol{\phi}}\\ &= \left(\mathbf{f}_{\hat{\mathbf{x}}}\frac{\partial\hat{\mathbf{v}}}{\partial\hat{\mathbf{a}}} + \mathbf{f}_{\hat{\mathbf{a}}} - \mathbf{f}_{\hat{\mathbf{x}}}\frac{\partial\hat{\mathbf{v}}}{\partial\hat{\mathbf{a}}} - O(\hat{\mathbf{v}})\right)\mathbf{W}\hat{\boldsymbol{\phi}} + O(\hat{\mathbf{v}}^2) \cong \mathbf{f}_{\hat{\mathbf{a}}}\mathbf{W}\hat{\boldsymbol{\phi}},\end{aligned} \tag{25}$$

an expression that depends insensitively on $\hat{\mathbf{v}}$. I have used the symmetry of $\mathbf{W}$ to combine terms in deriving Eq. (25), and have finally dropped terms $O(\hat{\mathbf{v}}^2)$. The vector $\hat{\boldsymbol{\delta}}_g$, which points along the negative gradient of S_0, i.e., the direction of steepest descent, can now be used to give a modified method of steepest descent as follows:

At each step compute (from the best current values of $\hat{\mathbf{v}}$ and $\hat{\mathbf{a}}$)

$$\hat{\boldsymbol{\phi}} = \mathbf{f}(\hat{\mathbf{x}},\hat{\mathbf{a}}) - \mathbf{f}_{\hat{\mathbf{x}}}\hat{\mathbf{v}}, \tag{26a}$$

$$\hat{\mathbf{v}}_{\text{new}} = -\boldsymbol{\sigma}\mathbf{f}_{\hat{\mathbf{x}}}{}^t\mathbf{W}\hat{\boldsymbol{\phi}}, \tag{26b}$$

$$\boldsymbol{\delta}_g = -\nu\mathbf{f}_{\hat{\mathbf{a}}}{}^t\mathbf{W}\hat{\boldsymbol{\phi}}, \tag{26c}$$

$$\hat{\mathbf{a}}_{\text{new}} = \hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_g, \tag{26d}$$

where the proportionality constant ν in (26c) must be chosen small enough that the new value of S_0 is less than the old one. Since I am trying to incorporate these results into other algorithms, which provide criteria for estimating this constant, I do not pursue this point at this time.

Unfortunately, an important point has been neglected, for Eqs. (26) do not enforce the constraint that $\mathbf{g}(\hat{\mathbf{a}}) = \mathbf{0}$ either before or after the step is taken. It may well be that $\mathbf{g}(\hat{\mathbf{a}}) \neq \mathbf{0}$ initially, and steps must be taken to satisfy this condition (if it is not initially satisfied) and to maintain it once it has been satisfied. Therefore in the constrained case, Eqs. (26) need modification, in particular Eq. (26c), which is the only one affecting the constraint $\mathbf{g}(\hat{\mathbf{a}}) = \mathbf{0}$.

To see how to do this, consider the linearization of Eq. (3d):

$$\mathbf{g}(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_g) = \mathbf{g}_{\hat{\mathbf{a}}}\hat{\boldsymbol{\delta}}_g + \hat{\mathbf{g}} = 0. \tag{27}$$

To modify Eq. (26c) it will be necessary to do two things: First I must remove from the $\hat{\boldsymbol{\delta}}_g$ of Eq. (26c) any component which, when premultiplied by the matrix $\hat{\mathbf{g}}_{\mathbf{a}}$, would yield a nonzero result. Second, I must add to this a term which causes the (modified) $\hat{\boldsymbol{\delta}}_g$ to satisfy Eq. (27).

The first of these conditions can be satisfied by considering the set of all vectors $\hat{\boldsymbol{\delta}}_g$ such that $\mathbf{g}_{\hat{\mathbf{a}}}\hat{\boldsymbol{\delta}}_g = \mathbf{0}$. This set is the *null space* of $\mathbf{g}_{\hat{\mathbf{a}}}$, and it is a linear subspace of the set of all $\hat{\boldsymbol{\delta}}_g$. I can construct a pair of projection operators $P^{\perp}$ and P, which decompose each vector $\hat{\boldsymbol{\delta}}_g$ into a component in the null space of $\mathbf{g}_{\hat{\mathbf{a}}}$ and a component in its range.

They can be written

$$\mathbf{P} = \mathbf{B}^{-1}\mathbf{g}_{\hat{\mathbf{a}}}{}^{t}(\mathbf{g}_{\hat{\mathbf{a}}}\mathbf{B}^{-1}\mathbf{g}_{\hat{\mathbf{a}}}{}^{t})^{-1}\mathbf{g}_{\hat{\mathbf{a}}},$$
$$\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P},$$

so that

$$\mathbf{P}\cdot\mathbf{P} = \mathbf{P},$$
$$\mathbf{P}^{\perp}\cdot\mathbf{P}^{\perp} = \mathbf{P}^{\perp},$$

and (28)

$$\mathbf{g}_{\hat{\mathbf{a}}}\mathbf{P} = \mathbf{g}_{\hat{\mathbf{a}}},$$
$$\mathbf{g}_{\hat{\mathbf{a}}}\mathbf{P}^{\perp} = \mathbf{0},$$

where $\mathbf{B}$ is any nonsingular square matrix of order equal to the number of parameters. Note that $\mathbf{g}_{\hat{\mathbf{a}}}$ is of maximal rank, since by assumption the constraints $\mathbf{g}(\hat{\mathbf{a}}) = \mathbf{0}$ must be independent. Now multiply the $\hat{\boldsymbol{\delta}}_g$ of Eq. (26c) by $\mathbf{P}^{\perp}$ to obtain a vector which, when premultiplied by $\mathbf{g}_{\hat{\mathbf{a}}}$, gives zero.

To satisfy the second requirement, add to the projected vector $\mathbf{P}^{\perp}\hat{\boldsymbol{\delta}}_g$ the correction

$$\hat{\boldsymbol{\delta}}_c = -\mathbf{B}^{-1}\mathbf{g}_{\hat{\mathbf{a}}}{}^{t}(\mathbf{g}_{\hat{\mathbf{a}}}\mathbf{B}^{-1}\mathbf{g}_{\hat{\mathbf{a}}}{}^{t})^{-1}\hat{\mathbf{g}}, \tag{29}$$

which has the desired properties that $\mathbf{g}_{\hat{\mathbf{a}}}\hat{\boldsymbol{\delta}}_c = -\hat{\mathbf{g}}$, $\mathbf{P}\hat{\boldsymbol{\delta}}_c = \hat{\boldsymbol{\delta}}_c$, and $\mathbf{P}^{\perp}\hat{\boldsymbol{\delta}}_c = \mathbf{0}$, so that Eq. (26c) becomes modified to

$$\hat{\boldsymbol{\delta}}_g' = \mathbf{P}^{\perp}\hat{\boldsymbol{\delta}}_g + \hat{\boldsymbol{\delta}}_c, \tag{26c'}$$

which has the desired properties of satisfying Eq. (27) without altering the component of $\hat{\boldsymbol{\delta}}_g$ that lies in the null space of $\mathbf{g}_{\hat{\mathbf{a}}}$.

Note at this point that Eq. (26c′) can be duplicated by solving the linear system

$$\begin{pmatrix} \frac{1}{\nu}\mathbf{B} & \mathbf{g}_{\hat{\mathbf{a}}}{}^{t} \\ \mathbf{g}_{\hat{\mathbf{a}}} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\delta}}_g' \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = -\begin{pmatrix} \mathbf{f}_{\hat{\mathbf{a}}}{}^{t}\mathbf{W}\hat{\boldsymbol{\phi}} \\ \hat{\mathbf{g}} \end{pmatrix} \tag{30}$$

for $\hat{\boldsymbol{\delta}}_g$. Equation (30) should be compared with Eq. (17) of Paper I. The only difference is the substitution of the matrix $(1/\nu)\,\mathbf{B}$ in Eq. (30) for the matrix $\mathbf{A} = \mathbf{f}_{\hat{\mathbf{a}}}{}^{t}\mathbf{W}\mathbf{f}_{\hat{\mathbf{a}}}$ of Eq. (17).

The strict steepest descent solution would be given by setting $\mathbf{B} = \mathbf{I}$ (a unit matrix) in Eq. (30). However this turns out to be undesirable in general, and I shall retain the flexibility to choose $\mathbf{B}$ more freely. This will allow me to improve certain characteristics of the solution.

VI. MARQUARDT'S ALGORITHM

Section V has the primary purpose of presenting a treatment of the method of steepest descent that, as adapted to Eqs. (3), is most useful for the adaptation of Marquardt's algorithm (Marquardt 1963). The basic idea behind Marquardt's algorithm is, in some sense, to *interpolate* between Newton's method and the method of steepest descent in such a way that initially, far from the solution, steepest descent dominates, and that as the solution is approached, the more efficient Newton's method is chosen. To do this, a parameter ν (Marquardt's $1/\lambda$) is varied so that when it is small, the steepest descent solution dominates, and when it is large, Newton's method dominates. The size of ν is adjusted by monitoring the variation of S_0 (in my notation) so as to maximize its rate of decrease.

Comparison of Marquardt's paper with Eq. (17) of Paper I and Eq. (30) of this paper suggests that the natural generalization of his fundamental equation is

$$\begin{pmatrix} \mathbf{A} + \frac{1}{\nu}\mathbf{B} & \mathbf{g}_{\hat{\mathbf{a}}}{}^{t} \\ \mathbf{g}_{\hat{\mathbf{a}}} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = -\begin{pmatrix} \mathbf{f}_{\hat{\mathbf{a}}}{}^{t}\mathbf{W}\hat{\boldsymbol{\phi}} \\ \hat{\mathbf{g}} \end{pmatrix}, \tag{31}$$

where $\mathbf{B}$ is to be chosen appropriately. Because the gradient methods are not scale invariant, it is desirable to choose $\mathbf{B}$ so that it has the form

$$\mathbf{B} = \mathrm{diag}(A_{11}, A_{22}, \ldots, A_{nn}), \tag{32}$$

where A_{ij} are the elements of the matrix $\mathbf{A}$.

The effect of this is to scale the gradient along each coordinate axis by the factor $A_{ii}^{-1/2}$. Marquardt accomplishes the same thing by scaling the rows and columns of his system by the factors $A_{ii}^{-1/2}$.

Marquardt's algorithm would then take the following form:

Step 1: Initially, $\hat{\mathbf{a}} = \mathring{\mathbf{a}}$ is assumed and $\hat{\mathbf{v}}$ is unknown, so assume $\hat{\mathbf{v}} = \mathbf{0}$. If desired, Eqs. (26a) and (26b) may be used iteratively to improve the estimate of $\hat{\mathbf{v}}$. Since $\hat{\boldsymbol{\phi}}$ is insensitive to the value of $\hat{\mathbf{v}}$, this should converge rapidly.

Step 2: Adopt a starting value of ν (a value in the range 100 to 1000 is typical).

Step 3: Compute $S(\hat{\mathbf{a}}) = \frac{1}{2}\hat{\boldsymbol{\phi}}^{t}\mathbf{W}\hat{\boldsymbol{\phi}}$.

Step 4: Using Eqs. (31) compute $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{a}}_{\mathrm{new}} = \hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}$. Compute $\hat{\boldsymbol{\phi}}(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}})$ from Eqs. (26a) and $S(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}) = \frac{1}{2}\hat{\boldsymbol{\phi}}^{t}(\hat{\mathbf{a}} + \delta)\mathbf{W}\hat{\boldsymbol{\phi}}(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}})$.

Step 5: If $S(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}) \geq S(\hat{\mathbf{a}})$, set $\nu \leftarrow \nu/10$ and return to step 4.

Step 6: If $S(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}) < S(\hat{\mathbf{a}})$, set $\nu \leftarrow 10\nu$.

Step 7: Set $\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}} + \hat{\boldsymbol{\delta}} = \hat{\mathbf{a}}_{\mathrm{new}}$. Compute $\hat{\boldsymbol{\delta}}$ and $\hat{\mathbf{v}}$ using Eqs. (26a) and (26b). Set $S(\hat{\mathbf{a}})$ to its new value.

Step 8: Test for convergence. If not done, return to step 4. (Marquardt suggests testing the size of each component of $\hat{\boldsymbol{\delta}}$ against the corresponding component of $\hat{\mathbf{a}}$. When the change is sufficiently small for all components, the process may be said to have converged.)

VII. A MODIFICATION OF THE FLETCHER–POWELL ALGORITHM

Fletcher and Powell (1963), following Davidon (1959), have proposed an algorithm for minimizing functions of several variables that has great merits. This method is described by Acton (1970, pp. 467–469) and by Daniels (1978, pp. 226–232). The method has two basic features. First, at each iteration, once the *direction*

of the correction vector (which I have called $\hat{\boldsymbol{\delta}}$) has been chosen, an estimate is made (based on values at several points of the function S_0 to be minimized) of exactly *how far* in that direction one needs to go in order to attain the minimum in that direction, and the next iteration proceeds from *that point*. Second, although the method starts out using the steepest descent direction, in subsequent iterations an estimate is made of the quadratic terms in the Taylor's series expansion of S_0 to improve the estimate of the direction of the correction vector (i.e., to make it point more closely to the *actual* minimum of the function).

The Fletcher–Powell algorithm is designed to solve very general minimization problems, not necessarily those arising from least squares. It estimates the quadratic terms that a Newton's method calculation would require from information contained in the linear terms evaluated at two successive points. These terms are computed without the need of inverting a matrix, and approach the correct matrix in the limit of many iterations. Broyden (1967) calls such methods "Quasi-Newton."

In problems derived from least squares, the quadratic terms are readily available, for they are given by the inverse of the normal equation matrix. This suggests (if one is willing to calculate the inverse) that one should use the usual normal equations to estimate the *direction* of the vector at each step, and follow Fletcher and Powell only in the method of estimating its *length*. This also has the advantage of avoiding certain instabilities that have been noted with the Fletcher–Powell technique (Bard 1968). To do this, set $\mathbf{B} = \mathbf{A}$ in Eq. (30) to obtain

$$\begin{pmatrix} \frac{1}{\nu}\mathbf{A} & \mathbf{g}_{\hat{\mathbf{a}}}{}^t \\ \mathbf{g}_{\hat{\mathbf{a}}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\delta}}_\nu \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = - \begin{pmatrix} \mathbf{f}_{\hat{\mathbf{a}}}{}^t \mathbf{W} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{g}} \end{pmatrix} \tag{33}$$

and choose ν so that $S_0 = \frac{1}{2}\hat{\boldsymbol{\phi}}^t \mathbf{W} \hat{\boldsymbol{\phi}}$, evaluated *as a function of* ν, is a minimum. This procedure, when iterated, would closely approximate the Fletcher–Powell algorithm. This approach amounts to using Newton's method to estimate the direction to next value of $\hat{\mathbf{a}}$, but choosing the *length* of the step $\hat{\boldsymbol{\delta}}$ so as to minimize S_0 in the chosen direction.

As an aside, note that this choice of $\mathbf{B}$ is not far from the choice of $\mathbf{B}$ as given by Eq. (31), which ignores the correlations between variables. It is interesting to speculate that another interesting choice of $\mathbf{B}$ in Eq. (31) would be to set $\mathbf{B} = \mathbf{A}$. As shown below, this would make it possible to evaluate the inverse matrix of Eq. (31) only once per Marquardt iteration.

The family of vectors $\hat{\boldsymbol{\delta}}_\nu$ is easily calculated once the inverse of the matrix of Eq. (33) is known for $\nu = 1$. If that inverse is

$$\begin{pmatrix} \alpha & \beta \\ \beta^t & \gamma \end{pmatrix}, \tag{34}$$

then

$$\hat{\boldsymbol{\delta}}_\nu = \nu \hat{\boldsymbol{\delta}}_1 - (1 - \nu)\beta \hat{\mathbf{g}}, \tag{35}$$

where the second term on the right-hand side is the correction required to ensure that the constraints are satisfied.

The optimal value of ν can be estimated by evaluating $S_0 = \frac{1}{2}\hat{\boldsymbol{\phi}}^t \mathbf{W} \hat{\boldsymbol{\phi}}$ at three points, using Eqs. (21), (23), and (24). One can first evaluate S_0 at $\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_0 = \hat{\mathbf{a}} - \beta \hat{\mathbf{g}}$. For the second value, let $\nu_k = 2^{-k}$, $k = 0, 1, 2, \ldots$ successively in Eq. (35) until a value of ν_k is found such that $S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu_k}) \leq S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_0)$. Finally, for the third value choose either $S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu_{k+1}})$ or $S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu_{k-1}})$; the former may be more advantageous, but the latter may already have been calculated.

Then a parabola is passed through the three values, and the minimum found by differentiation. The value ν^* of ν found in this manner is the value to be used and I set

$$\begin{aligned} \hat{\mathbf{a}}_{\text{new}} &= \hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu^*}, \\ \hat{\boldsymbol{\phi}}_{\text{new}} &= \mathbf{f}(\hat{\mathbf{x}}, \hat{\mathbf{a}}_{\text{new}}) - \mathbf{f}_{\hat{\mathbf{x}}} \hat{\mathbf{v}}, \\ \hat{\mathbf{v}}_{\text{new}} &= -\sigma \mathbf{f}_{\hat{\mathbf{x}}}{}^t \mathbf{W} \hat{\boldsymbol{\phi}}_{\text{new}}. \end{aligned} \tag{36}$$

I am then ready for the next iteration. If the three values of S_0 that are obtained are $E_0 = S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_0)$, $E_1 = S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu_k})$, $E_2 = S_0(\hat{\mathbf{a}} + \hat{\boldsymbol{\delta}}_{\nu_{k\pm1}})$, and if $\alpha_1 = \nu_k$, $\alpha_2 = \nu_{k\pm1}$, then Daniels gives the useful formula

$$\nu^* = \frac{1}{2} \frac{(\alpha_1^2 - \alpha_2^2)E_0 + \alpha_2^2 E_1 - \alpha_1^2 E_2}{(\alpha_1 - \alpha_2)E_0 + \alpha_2 E_1 - \alpha_1 E_2}. \tag{37}$$

VIII. CONCLUSIONS

The solution methods for Eqs. (3) given in this paper provide several approaches for improving on the convergence of the method given in Paper I. Nevertheless there may be circumstances which arise in practice, particularly where the data are very noisy, when none of these methods will yield a solution. In such cases it may be necessary to resort to direct minimization methods. Such methods are slow, but they can succeed where other methods fail. One very useful method of this type is the "simplex algorithm" (which is not to be confused with Dantzig's simplex method of linear programming). This algorithm, which is described in Daniels (1978, pp. 183–202), requires only the existence of an objective function $S(\mathbf{a})$ to be minimized. In this paper, I have provided a method for computing the objective function through Eqs. (23), (26a), and (26b), and there seems to be no reason in principal why the simplex method could not be applied if needed, at least in the unconstrained case. The constrained case [when $\mathbf{g}(\mathbf{a}) = \mathbf{0}$ is stipulated] is a little trickier, since the constraints reduce the dimensionality of the parameter space. Still, there does not appear to be a fundamental obstacle to the application of methods such as the simplex method to this case.

I thank Dr. W. Peters and Dr. R. Leach for their interest and stimulating conversations, and Dr. R. Branham and Dr. R. H. Miller for drawing my attention to the Fletcher–Powell–Davidon work and related papers. I also am grateful to an anonymous referee, at whose suggestion the last section of the paper was added, and who was responsible for several other improvements in the final draft. The support of the National Aeronautics and Space Administration, under contract NAS 8-32906, is gratefully acknowledged.

REFERENCES

Acton, F. S. (1970). *Numerical Methods That Work* (Harper and Row, New York).
Bard, Y. (1968). Math. Comp. **22,** 665.
Britt, H. I., and Luecke, R. H. (1973). Technometrics **15,** 233.
Brown, D. (1955). "A Matrix Treatment of the General Problem of Least Squares Considering Correlated Observations," Ballistic Research Laboratories Report No. 937.
Broyden, C. G. (1967). Math. Comp. **21,** 368.
Daniels, R. W. (1978). *An Introduction to Numerical Methods and Optimization Techniques* (North-Holland, New York).
Davidon, W. L. (1959). "Variable Metric Methods for Minimization," AEC Research and Development Report ANL-5990 (Rev. TID-4500, 14th ed.).
Fletcher, W., and Powell, M. J. D. (1963). Comput. J. **6,** 163.
Jefferys, W. (1980). Astron. J. **85,** 177 (Paper I).
Leach, R. (1980). Private communication.
Marquardt, D. W. (1963). J. Soc. Ind. Appl. Math. **11,** 431.
Meyer, S. L. (1975). *Data Analysis for Scientists and Engineers* (Wiley, New York).
Peters, W. (1980). Private communication.