This is method IV in Pike & Jones.

J. Soc. Indust. Appl. Math.
Vol. 11, No. 2, June, 1963
Printed in U.S.A.

AN ALGORITHM FOR LEAST-SQUARES ESTIMATION OF NONLINEAR PARAMETERS*

DONALD W. MARQUARDT†

Introduction. Most algorithms for the least-squares estimation of nonlinear parameters have centered about either of two approaches. On the one hand, the model may be expanded as a Taylor series and corrections to the several parameters calculated at each iteration on the assumption of local linearity. On the other hand, various modifications of the method of steepest-descent have been used. Both methods not infrequently run aground, the Taylor series method because of divergence of the successive iterates, the steepest-descent (or gradient) methods because of agonizingly slow convergence after the first few iterations.

In this paper a *maximum neighborhood* method is developed which, in effect, performs an optimum interpolation between the Taylor series method and the gradient method, the interpolation being based upon the maximum neighborhood in which the truncated Taylor series gives an adequate representation of the nonlinear model.

The results are extended to the problem of solving a set of nonlinear algebraic equations.

Statement of problem. Let the model to be fitted to the data be

$$\begin{aligned} E(y) &= f(x_1, x_2, \cdots, x_m; \beta_1, \beta_2, \cdots, \beta_k) \\ &= f(\mathrm{x}, \beta), \end{aligned} \tag{1}$$

where $x_1, x_2, \cdots, x_m$ are independent variables, $\beta_1, \beta_2, \cdots, \beta_k$ are the population values of k parameters, and $E(y)$ is the expected value of the dependent variable y. Let the data points be denoted by

$$(Y_i, X_{1i}, X_{2i}, \cdots, X_{mi}), \qquad i = 1, 2, \cdots, n. \tag{2}$$

The problem is to compute those estimates of the parameters which will minimize

$$\begin{aligned} \Phi &= \sum_{i=1}^{n} [Y_i - \hat{Y}_i]^2 \\ &= \| \mathrm{Y} - \hat{\mathrm{Y}} \|^2, \end{aligned} \tag{3}$$

$(\|x\| = x^T x$ (Euclidean norm))

where $\hat{Y}_i$ is the value of y predicted by (1) at the ith data point.

It is well known that when f is linear in the β's, the contours of constant

* Received by the editors June 13, 1962, and in revised form December 3, 1962.

† Engineering Department, E. I. du Pont de Nemours & Company, Inc., Wilmington 98, Delaware.

Φ are ellipsoids, while if f is nonlinear, the contours are distorted, according to the severity of the nonlinearity. Even with nonlinear models, however, the contours are nearly elliptical in the immediate vicinity of the minimum of Φ. Typically the contour surface of Φ is greatly attenuated in some directions and elongated in others so that the minimum lies at the bottom of a long curving trough.

In this paper attention is confined to the algorithm for obtaining least-squares estimates. For discussion of the broader statistical aspects of nonlinear estimation and application to specific examples the reader is referred to items [1], [2], [4], [5], [6] in the References. Other references are given in the papers cited.

Methods in current use. The method based upon expanding f in a Taylor series (sometimes referred to as the Gauss method [1], [7], or the Gauss-Newton method [4]) is as follows.

Writing the Taylor series through the linear terms

$$\langle Y(\mathbf{X}_i, \mathbf{b} + \boldsymbol{\delta}_t)\rangle = f(\mathbf{X}_i, \mathbf{b}) + \sum_{j=1}^{k} \left(\frac{\partial f_i}{\partial b_j}\right)(\delta_t)_j \tag{4}$$

or

$$\langle \mathbf{Y}\rangle = \mathbf{f}_0 + P\boldsymbol{\delta}_t \tag{4a}$$

In (4), β is replaced notationally by $\mathbf{b}$, the converged value of $\mathbf{b}$ being the least-squares estimate of β. The vector $\boldsymbol{\delta}_t$ is a small correction to $\mathbf{b}$, with the subscript t used to designate $\boldsymbol{\delta}$ as calculated by this *Taylor series* method. The brackets $\langle\ \rangle$ are used to distinguish predictions based upon the linearized model from those based upon the actual nonlinear model. Thus, the value of Φ predicted by (4) is

$$\langle\Phi\rangle = \sum_{i=1}^{n} [Y_i - \langle Y_i\rangle]^2. \tag{5}$$

Now, $\boldsymbol{\delta}_t$ appears linearly in (4), and can therefore be found by the standard least-squares method of setting $\partial\langle\Phi\rangle/\partial\delta_j = 0$, for all j. Thus $\boldsymbol{\delta}_t$ is found by solving

$$A\boldsymbol{\delta}_t = \mathbf{g}, \tag{6}$$

where[1]

$$A^{[k\times k]} = P^T P, \tag{7}$$

[1] The superscript T denotes matrix transposition.

$$P^{[n\times k]} = \left(\frac{\partial f_j}{\partial b_j}\right), \qquad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, k, \tag{8}$$

$$\begin{aligned} \mathbf{g}^{[k\times 1]} &= \left(\sum_{i=1}^{n} (Y_i - f_i)\,\frac{\partial f_i}{\partial b_j}\right), \qquad j = 1, 2, \cdots, k, \\ &= P^T(\mathrm{Y} - \mathrm{f}_0). \end{aligned} \tag{9}$$

In practice it is found helpful to correct b by only a fraction of δ_t; otherwise the extrapolation may be beyond the region where f can be adequately represented by (4), and would cause divergence of the iterates. Various methods have, therefore, been used [1], [4], [6] to determine an appropriate step size $K\delta_t$, $0 < K \leqq 1$, once the direction has been specified by δ_t. Even so, failure to converge is not uncommon.

The *gradient methods* by contrast simply step off from the current trial value in the direction of the negative gradient of Φ. Thus

$$\delta_g = -\left(\frac{\partial \Phi}{\partial b_1}, \frac{\partial \Phi}{\partial b_2}, \cdots, \frac{\partial \Phi}{\partial b_k}\right)^T \tag{9}$$

Various modified steepest-descent methods have been employed [5] to compensate partially for the typically poor conditioning of the Φ surface which leads to very slow convergence of the gradient methods. With these gradient methods, as with the Taylor series methods, it is necessary to control the step size carefully once the direction of the correction vector has been established. Even so, slow convergence is the rule rather than the exception.

A number of convergence proofs, valid under various assumptions about the properties of f, have been derived for the Taylor series method [4], [7]. Convergence proofs have also been derived for various gradient type methods; e.g., [3]. However, mathematical proof of the monotonicity of Φ from iteration to iteration, assuming a well behaved function f, and assuming infinitely precise arithmetic processes, is at best a necessary condition for convergence in practice. Such proofs, in themselves, give no adequate indication of the *rate* of convergence. Among the class of theoretically convergent methods one must seek to find methods which actually do converge promptly using finite arithmetic, for the large majority of problems.

Qualitative analysis of the problem. In view of the inadequacies of the Taylor series and gradient methods, it is well to review the undergirding principles involved. First, any proper method must result in a correction vector whose direction is within 90° of the negative gradient of Φ. Otherwise the values of Φ can be expected to be larger rather than smaller at points along the correction vector. Second, because of the severe elongation

of the Φ surface in most problems, δ_t is usually almost 90° away from δ_g. (Indeed, we have monitored this angle, γ, for a variety of problems and have found that γ usually falls in the range $80° < \gamma < 90°$.) From these considerations it would seem reasonable that any improved method will in some sense interpolate between δ_t and δ_g.

Implicit in both of the procedures outlined is the choice of direction of the correction vector prior to the determination of an acceptable step size. In the algorithm to be described, the direction and step size are determined *simultaneously*.

Theoretical basis of algorithm. The theoretical basis of the algorithm is contained in several theorems which follow. Theorems 1 and 2 are due to Morrison [7]. The proof given here for Theorem 1 is different from that given by Morrison. The present proof affords perhaps a greater insight into the geometric relationships involved.

THEOREM 1. *Let* $\lambda \geqq 0$ *be arbitrary and let* δ_0 *satisfy the equation*

$$(A + \lambda I)\delta_0 = g. \tag{10}$$

Then δ_0 *minimizes* $\langle \Phi \rangle$ *on the sphere whose radius* $\| \delta \|$ *satisfies*

$$\| \delta \|^2 = \| \delta_0 \|^2.$$

Proof. In order to find δ which will minimize

$$\langle \Phi(\delta) \rangle = \| Y - f_0 - P\delta \|^2, \tag{11}$$

under the constraint

$$\| \delta \|^2 = \| \delta_0 \|^2, \tag{12}$$

a necessary condition for a stationary point is, by the method of Lagrange,

$$\frac{\partial u}{\partial \delta_1} = \frac{\partial u}{\partial \delta_2} = \cdots = \frac{\partial u}{\partial \delta_k} = 0, \qquad \frac{\partial u}{\partial \lambda} = 0, \tag{13}$$

where

$$u(\delta, \lambda) = \| Y - f_0 - P\delta \|^2 + \lambda(\| \delta \|^2 - \| \delta_0 \|^2) \tag{14}$$

and λ is a Lagrange multiplier.

Thus, taking the indicated derivatives

$$0 = -[P^T(Y - f_0) - P^T P\delta] + \lambda\delta, \tag{15}$$

$$0 = \| \delta \|^2 - \| \delta_0 \|^2. \tag{16}$$

For a given λ, (15) is satisfied by the solution δ of the equation

$$(P^T P + \lambda I)\delta = P^T(Y - f_0) \tag{17}$$

as can easily be verified by premultiplying (17) by $(P^TP)^{-1}$, and writing in the form

$$\delta = (P^TP)^{-1}P^T(Y - f_0) - (P^TP)^{-1}\lambda\delta, \tag{18}$$

and then substituting into (15). (10) and (17) are identical. That this stationary point is actually a minimum, is clear from the fact that $(A + \lambda I)$ is positive definite.

THEOREM 2. *Let* $\delta(\lambda)$ *be the solution of* (10) *for a given value of* λ. *Then* $\| \delta(\lambda) \|^2$ *is a continuous decreasing function of* λ, *such that as* $\lambda \to \infty$, $\| \delta(\lambda) \|^2 \to 0$.

Proof. Since the symmetric matrix A is positive definite, it may be transformed by an orthonormal rotation of axes into a diagonal matrix, D, without altering the distances between points. Let the transformation be denoted $S^TAS = D$, where $S^TS = I$, and all diagonal elements of D are positive. Thus, (10) takes the form $(D + \lambda I)S^{-1}\delta_0 = S^Tg$, so that

$$\delta_0 = S(D + \lambda I)^{-1}S^Tg. \tag{19}$$

Then, defining $v = S^Tg$

$$\begin{aligned} \| \delta_0(\lambda) \|^2 &= g^TS(D + \lambda I)^{-1}S^TS(D + \lambda I)^{-1}S^Tg \\ &= v^T[(D + \lambda I)^2]^{-1}v \\ &= \sum_{j=1}^{k} \frac{v_j^2}{(D_j + \lambda)^2} \end{aligned} \tag{20}$$

which is clearly a decreasing function of λ, $(\lambda \geqq 0)$, such that as $\lambda \to \infty$, $\| \delta_0(\lambda) \|^2 \to 0$.

The orthonormal transformation to a diagonal matrix has here been explicitly exhibited in order to facilitate the proof of the following theorem.

THEOREM 3. *Let* γ *be the angle between* δ_0 *and* δ_g. *Then* γ *is a continuous monotone decreasing function of* λ *such that as* $\lambda \to \infty$, $\gamma \to 0$. *Since* δ_g *is independent of* λ, *it follows that* δ_0 *rotates toward* δ_g *as* $\lambda \to \infty$.

Proof. We first observe that

$$\delta_g = \left(\sum_{i=1}^{n} (Y_i - f_i) \frac{\partial f_i}{\partial b_j} \right)^T, \qquad j = 1, 2, \cdots, k. \tag{21}$$

Thus (except for a scale factor which is irrelevant since only the orientation of the vectors is pertinent)

$$\delta_g = g. \tag{22}$$

By definition

$$\cos\gamma = \frac{\delta^T g}{(\|\delta\|)(\|g\|)} = \frac{v^T(D+\lambda I)^{-1}v}{(v^T[(D+\lambda I)^2]^{-1}v)^{1/2}(g^Tg)^{1/2}} = \frac{\sum_{j=1}^{k}\frac{v_j^2}{D_j+\lambda}}{\left[\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{1/2}(g^Tg)^{1/2}}. \tag{23}$$

Differentiating and simplifying

$$\frac{d}{d\lambda}\cos\gamma = \frac{\left[\sum_{j=1}^{k}\frac{v_j^2}{D_j+\lambda}\right]\left[\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^3}\right] - \left[\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^2}{\left[\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{3/2}(g^Tg)^{1/2}} \tag{24}$$

$$= \frac{[\sum_{j=1}^{k} v_j^2\prod_{1j}][\sum_{j=1}^{k} v_j^2\prod_{3j}] - [\sum_{j=1}^{k} v_j^2\prod_{2j}]^2}{\left[\sum_{j=1}^{k}\frac{v_j^2}{(D_j+\lambda)^2}\right]^{3/2}[\prod_{j=1}^{k}(D_j+\lambda)^2]^2(g^Tg)^{1/2}} \tag{25}$$

where $\prod_{1j} = \prod_{\substack{j'=1\\ j'\neq j}}^{k}(D_{j'}+\lambda)$, $\prod_{2j} = \prod_{\substack{j'=1\\ j'\neq j}}^{k}(D_{j'}+\lambda)^2$, $\prod_{3j} = \prod_{\substack{j'=1\\ j'\neq j}}^{k}(D_{j'}+\lambda)^3$. The denominator of (25) is positive, since each factor is positive. Hence the sign of $d\cos\gamma/d\lambda$ is the sign of the numerator. Noting that $\prod_{1j}\prod_{3j} = (\prod_{2j})^2$, the numerator can be written

$$\left[\sum_{j=1}^{k}(v_j\prod_{1j}^{1/2})^2\right]\left[\sum_{j=1}^{k}(v_j\prod_{3j}^{1/2})^2\right] - \left[\sum_{j=1}^{k}(v_j\prod_{1j}^{1/2})(v_j\prod_{3j}^{1/2})\right]^2. \tag{26}$$

By Schwarz's Inequality, (26) is positive. Thus $d\cos\gamma/d\lambda$ is always positive ($\lambda > 0$). Consequently, γ is a monotone decreasing function of λ.

For very large values of λ, the matrix $(A+\lambda I)$ is dominated by the diagonal λI. Thus it is seen from (10) that as $\lambda \to \infty$, $\delta_0 \to g/\lambda$, whence δ_0 and g become proportional in the limit, so that the angle between them approaches zero. On the other hand, if $\lambda = 0$ in (10), then (except for the trivial case where A is diagonal) the vectors δ_0 and g meet at some finite angle $0 < \gamma < \pi/2$. It follows that γ is a continuous monotone decreasing function of λ, such that as $\lambda \to \infty$, $\gamma \to 0$.

Scale of measurement. The relevant properties of the solution, δ_t, of (6) are invariant under linear transformations of the b-space. However, it is well known [3] that the properties of the gradient methods are not scale

invariant. It becomes necessary, then, to scale the b-space in some convenient manner. We shall choose to scale the b-space in units of the standard deviations of the derivatives $\partial f_i/\partial b_j$, taken over the sample points $i = 1, 2, \cdots, n$. Since these derivatives depend, in general, on the b_j themselves, the current trial values of the b_j are used as necessary in the evaluation of the derivatives. This choice of scale causes the A matrix to be transformed into the matrix of simple correlation coefficients among the $\partial f_i/\partial b_j$. This choice of scale has, in fact, been widely used in linear least-squares problems as a device for improving the numerical aspects of computing procedures.

Thus, we define a scaled matrix A^*, and a scaled vector $\mathbf{g}^*$:

$$A^* = (a^*_{jj'}) = \left(\frac{a_{jj'}}{\sqrt{a_{jj}}\sqrt{a_{j'j'}}}\right), \tag{27}$$

$$\mathbf{g}^* = (g_j^*) = \left(\frac{g_j}{\sqrt{a_{jj}}}\right) \tag{28}$$

and solve for the Taylor series correction using

$$A^*\delta_i^* = \mathbf{g}^*. \tag{29}$$

Then

$$\delta_j = \delta_j^*/\sqrt{a_{jj}}\,. \tag{30}$$

See attached note at end about this scaling.

Construction of the algorithm. The broad outline of the appropriate algorithm is now clear. Specifically, at the rth iteration the equation

$$(A^{*(r)} + \lambda^{(r)}I)\delta^{*(r)} = \mathbf{g}^{*(r)} \tag{31}$$

is constructed. This equation is then solved for $\delta^{*(r)}$. Then (30) is used to obtain $\delta^{(r)}$. The new trial vector

$$\mathbf{b}^{(r+1)} = \mathbf{b}^{(r)} + \delta^{(r)} \tag{32}$$

will lead to a new sum of squares $\Phi^{(r+1)}$. It is essential to select $\lambda^{(r)}$ such that

$$\Phi^{(r+1)} < \Phi^{(r)}. \tag{33}$$

It is clear from the foregoing theory that a sufficiently large $\lambda^{(r)}$ always exists such that (33) will be satisfied, unless $\mathbf{b}^{(r)}$ is already at a minimum of Φ. Some form of trial and error is required to find a value $\lambda^{(r)}$ which will lead to satisfaction of (33) and will produce rapid convergence of the algorithm to the least-squares values.

At each iteration we desire to minimize Φ in the (approximately) maximum neighborhood over which the linearized function will give adequate

representation of the nonlinear function. Accordingly, the strategy for choosing $\lambda^{(r)}$ must seek to use a small value of $\lambda^{(r)}$ whenever conditions are such that the unmodified Taylor series method would converge nicely. This is especially pertinent in the later stages of the convergence procedure, when the guesses are in the immediate vicinity of the minimum, where the contours of Φ are asymptotically elliptical, and the linear expansion of the model needs to be a good approximation over only a very small region. Large values of $\lambda^{(r)}$ should therefore be used only when necessary to satisfy (33). While it is true that $\Phi^{(r+1)}$ as a function of λ has a minimum, and choice of this value of λ at the rth iteration would maximize $(\Phi^{(r)} - \Phi^{(r+1)})$, such a locally optimum choice would be poor global strategy, since it typically requires a substantially larger value of λ than is necessary to satisfy (33). Such a strategy would inherit many of the properties of steepest-descent; e.g., rapid initial progress followed by progressively slower progress.

We shall therefore define our strategy as follows:

Let $\nu > 1$.

Let $\lambda^{(r-1)}$ denote the value of λ from the previous iteration. Initially let $\lambda^{(0)} = 10^{-2}$, say.

Compute $\Phi(\lambda^{(r-1)})$ and[2] $\Phi(\lambda^{(r-1)}/\nu)$.

i. If $\Phi(\lambda^{(r-1)}/\nu) \leqq \Phi^{(r)}$, let $\lambda^{(r)} = \lambda^{(r-1)}/\nu$.

ii. If $\Phi(\lambda^{(r-1)}/\nu) > \Phi^{(r)}$, and $\Phi(\lambda^{(r-1)}) \leqq \Phi^{(r)}$, let $\lambda^{(r)} = \lambda^{(r-1)}$.

iii. If $\Phi(\lambda^{(r-1)}/\nu) > \Phi^{(r)}$, and $\Phi(\lambda^{(r-1)}) > \Phi^{(r)}$, increase λ by successive multiplication[3] by ν until for some smallest w, $\Phi(\lambda^{(r-1)}\nu^{w}) \leqq \Phi^{(r)}$. Let $\lambda^{(r)} = \lambda^{(r-1)}\nu^{w}$.

---

[2] If $\lambda^{(r-1)}$ is already negligible by comparison with 1.0 to the number of significant figures carried, then go to test ii. or iii. immediately without computing $\Phi(\lambda^{(r-1)}/\nu)$, and ignore comparisons involving $\Phi(\lambda^{(r-1)}/\nu)$.

[3] On occasion in problems where the correlations among the parameter-estimates are extremely high (>0.99) it can happen that λ will be increased to unreasonably large values. It has been found helpful for these instances to alter test iii. The revised test is:

Let $$\mathbf{b}^{(r+1)} = \mathbf{b}^{(r)} + K^{(r)}\mathbf{\delta}^{(r)}, \qquad K^{(r)} \leqq 1.$$

Noting that the angle $\gamma^{(r)}$ is a decreasing function of $\lambda^{(r)}$, select a criterion angle $\gamma_0 < \pi/2$ and take

$$K^{(r)} = 1 \quad \text{if} \quad \gamma^{(r)} \geqq \gamma_0 .$$

However, if test iii. is not passed even though $\lambda^{(r)}$ has been increased until $\gamma^{(r)} < \gamma_0$, then do not increase $\lambda^{(r)}$ further, but take $K^{(r)}$ sufficiently small so that $\Phi^{(r+1)} < \Phi^{(r)}$. This can always be done since $\gamma^{(r)} < \gamma_0 < \pi/2$. A suitable choice for the criterion angle is $\gamma_0 = \pi/4$.

It may be noted that positivity of $\cos\gamma$ when $\lambda = 0$ is guaranteed only when A is

By this algorithm we always obtain a feasible neighborhood. Further, we almost always obtain, within a factor determined by ν, the maximum neighborhood in which the Taylor series gives an adequate representation for our purposes. The iteration is converged when $\frac{|\delta_j^{(r)}|}{\tau + |b_j^{(r)}|} < \epsilon$, for all j, for some suitably small $\epsilon > 0$, say 10^{-5} and some suitable τ, say 10^{-3}. The choice of ν is arbitrary; $\nu = 10$ has been found in practice to be a good choice.

Typically, condition iii. is met only rarely. Thus, it is most often required that (31) be solved for two values of $\lambda^{(r)}$ at each iteration. One such solution is required for the standard Taylor series method. The extra linear equation solution is generally much less computational effort than the evaluation of the A^* matrix, so that the small proportional increase in computation per iteration is more than offset by the gain in the power of an iteration.

It may be remarked that when using floating point arithmetic it is often desirable to accumulate the sum of squares of (3) (for use in tests i. and ii.) in double precision, after computing $\hat{Y}_i$ to single precision. Failure to do this can lead to erratic behavior of Φ near the minimum, due only to rounding error. In extreme cases it may also be necessary to compute $\hat{Y}_i$ to double precision.

A corollary numerical benefit associated with adding λ to the diagonal of the A^* matrix is that the composite matrix is always better conditioned than A^* itself.

Application to other problems. Clearly the method described can be applied to other types of problems. For example, it will enable interpolation between the gradient method and the Newton-Raphson method for solving systems of nonlinear algebraic equations, such as the simultaneous nonlinear difference equations associated with the solution of nonlinear boundary value problems by implicit methods. In a sense, such problems are a particularization of the least-squares problem to the case where $n = k$. Let the system of equations to be solved be written

$$0 = f_i(\mathbf{x}), \qquad i = 1, 2, \cdots, k, \tag{34}$$

where x is a k-dimensional vector of unknowns, x_j.

---

positive definite. In the presence of very high correlations, the positive definiteness of A can break down due to rounding error alone. In such instances the Taylor series method can diverge, no matter what value of $K^{(r)}$ is used. The present algorithm, by its use of $(A + \lambda I)$ will guarantee positive $\cos \gamma$ even though A may not quite be positive definite. While it is not always possible to forsee high correlations among the parameter-estimates when taking data for a nonlinear model, better experimental design can often substantially reduce such extreme correlations when they do occur.

For the gradient methods it is customary to define an error criterion

$$\Phi = \sum_{i=1}^{k} [f_i(\mathbf{x})]^2 \tag{35}$$

and to make corrections to trial vectors $\mathbf{x}^{(r)}$ by moving in the direction of the negative gradient of Φ, defined by

$$-\left(\frac{\partial \Phi}{\partial x_j}\right) = -\left(2 \sum_{i=1}^{k} f_i \frac{\partial f_i}{\partial x_j}\right). \tag{36}$$

So that

$$x_j^{(r+1)} = x_j^{(r)} + \delta_j^{(r)}, \qquad j = 1, 2, \cdots, k, \tag{37}$$

where

$$\delta_j^{(r)} = -a \frac{\partial \Phi}{\partial x_j}, \qquad j = 1, 2, \cdots, k, \tag{38}$$

and a is a suitably defined constant.

In matrix notation

$$\left(\frac{1}{a} I\right) \delta_g^{(r)} = \mathbf{g}^{(r)}. \tag{39}$$

On the other hand, the Newton-Raphson method, which involves expansion of the f_i in Taylor series through the linear terms, leads to the equations

$$\sum_{j=1}^{k} \left(\frac{\partial f_i}{\partial x_j}\right)^{(r)} \delta_j^{(r)} = -f_i^{(r)}, \qquad i = 1, 2, \cdots, k. \tag{40}$$

In matrix notation

$$B\delta_t = \mathbf{f} \tag{41}$$

where the matrix $B = -(\partial f_i / \partial x_j)$ is not, in general, symmetric. Premultiplying (41) by B^T we then form $B^T B = A$ and $B^T \mathbf{f} = \mathbf{g}$. The matrix A is symmetric. Matrix A and vector $\mathbf{g}$ are then normalized to A^*, $\mathbf{g}^*$.

Application of the new algorithm gives the formulation

$$(A^* + \lambda^{(r)} I) \delta^{*(r)} = \mathbf{g}^* \tag{42}$$

which is identical to the formulation of the nonlinear least-squares problem. The computational method for obtaining the solution is also identical. In this case, however, the minimum value of Φ is known to be zero, within rounding error.

Conclusion. The algorithm described shares with the gradient methods their ability to converge from an initial guess which may be outside the region of convergence of other methods. The algorithm shares with the Taylor series method the ability to close in on the converged values rapidly after the vicinity of the converged values has been reached. Thus, the method combines the best features of its predecessors while avoiding their most serious limitations.

Acknowledgments. The author is indebted to a referee for detecting an error in the proof of Theorem 3 in the original manuscript. The author is also grateful to Professor H. O. Hartley who has drawn his attention, since submission of this paper, to the related work of Levenberg [8]. From somewhat different reasoning Levenberg was also led to add a quantity to the diagonal of A. Levenberg's recommendation to minimize Φ locally as a function of the quantity added to the diagonal would give steepest-descent-like convergence for the reasons detailed earlier in this paper.

Computer Program. A FORTRAN program, "Least-Squares Estimation of Nonlinear Parameters", embodying the algorithm described in this paper is available as IBM Share Program No. 1428.

REFERENCES

[1] G. W. Booth, G. E. P. Box, M. E. Muller, and T. I. Peterson, *Forecasting by Generalized Regression Methods, Nonlinear Estimation* (Princeton — IBM), International Business Machines Corp., Mimeo. (IBM Share Program No. 687 WL NL1.), 1959.

[2] G. E. P. Box, and G. A. Coutie, Proc. Inst. Elec. Engrs. 103, Part B, Suppl. No. 1 (1956), Paper No. 2138 M.

[3] H. B. Curry, Quart. Appl. Math., 2 (1944), pp. 258–261.

[4] H. O. Hartley, Technometrics, 3 (1961), No. 2, pp. 269–280.

[5] D. W. Marquardt, Chem. Eng. Progr., 55 (1959), No. 6, pp. 65–70.

[6] D. W. Marquardt, R. G. Bennett, and E. J. Burrell, Jour. Molec. Spectroscopy, 7 (1961), No. 4, pp. 269–279.

[7] D. D. Morrison, *Methods for nonlinear least squares problems and convergence proofs, Tracking Programs and Orbit Determination*, Proc. Jet Propulsion Laboratory Seminar, (1960), pp. 1–9.

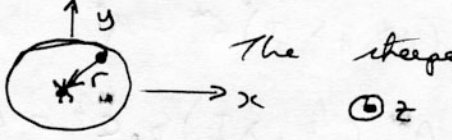
[8] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quart. Appl. Math., 2 (1944), pp. 164–168.

Scaling of the b space:.

Imagine an n dimensional space, say for simplicity $n = 3$. (ie 2 parameters).

If we are at a point on the 3D surface, then the steepest descent, and the gradient, are meaningless unless both parameters are in the same units.

Assume the ellipsoidal surface is spherical:



The steepest descent is towards the minimum, and the length is r. Now scale change x.

The steepest descent is not towards the minimum, and the optimum length (given the correct direction) is $\neq r$.

Scaling is ∴ necessary.

See R.A. Dankoehler, J. Biol. Chem. 241, 1955 (1966)

Marquardt's scaling:

$$A^* = (a^*_{ij}) = \left(\frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}} \right) = \left(\frac{\partial^2 f_i}{\partial p_i \partial p_j} \Big/ \sqrt{\frac{\partial^2 f_i}{\partial^2 p_i} \times \frac{\partial^2 f_j}{\partial^2 p_j}} \right)$$

Clearly if $i = j$, $a^*_{ij} = 1$. ie perfect correlation.

This scaling is already done if we change the identity matrix, I, for C where $C = (a_{ii})$. Hence the equivalence (algebraically) of eqn's [20] and [21] in Pitha & Jones, Can J. Chem, 44, 3031, (1966).

If $C = (a_{ii})$ then $C^{-1/2} = \left(\frac{1}{\sqrt{a_{ii}}}\right)$

and $(A^* + \lambda I)\delta^* = g^*$ becomes

$$(C^{-1/2} \cdot A^* \cdot C^{-1/2} + \lambda I)\delta \cdot C^{1/2} = g \cdot C^{-1/2}.$$

[note dot product.

Jones & Pitha point out that this is = to $(A + \lambda C)\delta = g$, which is ∴ scaled,