
Department of Statistics

University of Wisconsin
Madison, WI

TECHNICAL REPORT NO. 321

November 1972

TOPICS IN MODEL BUILDING

PART II

ON NONLINEAR LEAST SQUARES

By

G. E. P. Box* and

H. Kanemasu**

Typist: Candy Smith

* This author was supported by the United States Air Force through the Air Force Office of Scientific Research under Grant AFOSR-72-2363.

** This author was supported by the United States Navy through the Office of Naval Research under Contract No. N00014-67-A-0128-0017.

TOPICS IN MODEL BUILDING

PART II ON NONLINEAR LEAST SQUARES

Gauss suggested that, when the model is a nonlinear function of parameters, least square parameter estimates might be obtained by iterative linearization. To prevent difficulties in convergence, Levenberg, and later Marquardt, proposed a constrained minimization procedure. On critically examining this method with a linearly invariant metric for the parameters, we find this to be equivalent to a simple modification of the Gauss method which had been proposed earlier. Procedures to decide how far one should go along the Gauss solution vector are introduced which utilize only quantities already computed.

2.1 Least squares

Suppose an observation y_u is described by the model

$$y_u = f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}) + \varepsilon_u \qquad u=1,2,\ldots,n \tag{2.1.1}$$

where $\underset{\sim}{\xi}_u = (\xi_{1u}, \xi_{2u}, \ldots, \xi_{ku})$ are the levels of k independent variables, $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ are p unknown parameters, $f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})$ is a known function of $\underset{\sim}{\xi}_u$ and $\underset{\sim}{\theta}$, and ε_u is a random error. Then the method of least squares obtains values $\hat{\underset{\sim}{\theta}}$ of the parameters $\underset{\sim}{\theta}$ which minimize the sum of squares

$$S(\underset{\sim}{\theta}) = \sum_{u=1}^{n} (y_u - f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}))^2 \tag{2.1.2}$$

Using vector notation (2.1.2) may be written

$$S(\underset{\sim}{\theta}) = (\underset{\sim}{y} - \underset{\sim}{f}_\theta)' (\underset{\sim}{y} - \underset{\sim}{f}_\theta) \tag{2.1.3}$$

where $\underset{\sim}{y}$ is the n×1 vector of y_u, u=1,2,...,n and $\underset{\sim}{f}_\theta$ is the n×1 vector whose u th element is $f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})$.

We shall say that the function $f(\underset{\sim}{\xi}, \underset{\sim}{\theta})$ is linear in the parameters if the derivatives $\partial f(\underset{\sim}{\xi}, \underset{\sim}{\theta})/\partial\theta_j$ are independent of $\underset{\sim}{\theta}$ for all j. Otherwise it is said to be nonlinear in the parameters.

In the linear case, it is easily shown that the least squares estimates $\hat{\underset{\sim}{\theta}}$ are the solutions of the normal equations

$$X'X\hat{\underset{\sim}{\theta}} = X'\underset{\sim}{y} \qquad (2.1.4)$$

where X is the n×p matrix of derivatives $\partial f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})/\partial\theta_j$. Also, if $X'X$ is non-singular

$$\hat{\underset{\sim}{\theta}} = (X'X)^{-1}X'\underset{\sim}{y} \; . \qquad (2.1.5)$$

If the random elements ε_u, u=1,2,...,n of the n×1 vector of errors $\underset{\sim}{\varepsilon}$ are such that $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $E(\underset{\sim\sim}{\varepsilon\varepsilon}') = I\sigma^2$, then the Gauss theorem tells us that $\underset{\sim}{c}'\hat{\underset{\sim}{\theta}}$ is a minimum variance unbiased linear estimator of $\underset{\sim}{c}'\underset{\sim}{\theta}$, where $\underset{\sim}{c}$ is any p dimensional vector of constants. Furthermore, if the errors ε's are normally distributed, $h(\hat{\underset{\sim}{\theta}})$ is the maximum likelihood estimator of any one to one function $h(\underset{\sim}{\theta})$ of $\underset{\sim}{\theta}$.

2.2 Gauss method and "overshooting"

More specifically local approximations to nonlinear functions $f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})$ (u=1,2,...,n) may be obtained by expanding around the current best estimates $\underset{\sim}{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$ of $\underset{\sim}{\theta}$ in a Taylor expansion and truncating after the first order terms. Then,

$$f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}) \cong f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}^{(0)}) + \sum_{i=1}^{p} \left[\frac{\partial f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})}{\partial \theta_i} \right]_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(0)}} (\theta_i - \theta_i^{(0)}) \quad u=1,2,\ldots,n, \qquad (2.2.1)$$

or, in vector notation,

$$\underset{\sim}{f}_{\theta} \cong \underset{\sim}{f}_{o} + X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}), \qquad (2.2.2)$$

where $\underset{\sim}{f}_{o}$ is the n×1 vector of $f(\underset{\sim}{\xi}_{u},\underset{\sim}{\theta}^{(0)})$; u=1,2,...,n and X_{o} is the n×p matrix whose (u,j) element is $[\partial f(\underset{\sim}{\xi}_{u},\underset{\sim}{\theta})/\partial\theta_{j}]_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(0)}}$.

Using this approximation, the approximate sum of squares corresponding to a particular choice $\underset{\sim}{\theta}$ of the parameters is

$$\begin{aligned}\bar{S}(\underset{\sim}{\theta}) &= (\underset{\sim}{y}-\underset{\sim}{f}_{o}-X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}))'(\underset{\sim}{y}-\underset{\sim}{f}_{o}-X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})) \\ &= \underset{\sim}{\varepsilon}_{o}'\underset{\sim}{\varepsilon}_{o}-2\underset{\sim}{\varepsilon}_{o}'X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) + (\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})'X_{o}'X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) \qquad (2.2.3)\end{aligned}$$

where $\underset{\sim}{\varepsilon}_{o} = \underset{\sim}{y}-\underset{\sim}{f}_{o}$. Setting the derivative $\partial\bar{S}(\underset{\sim}{\theta})/\partial\theta_{j}$ to zero gives the normal equations

$$X_{o}'X_{o}(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) = X_{o}'\underset{\sim}{\varepsilon}_{o}. \qquad (2.2.4)$$

Provided that $X_{o}'X_{o}$ is nonsingular, as is usually the case, new estimates $\underset{\sim}{\theta}^{(1)}$ of the parameters are given by

$$\underset{\sim}{\theta}^{(1)} - \underset{\sim}{\theta}^{(0)} = (X_{o}'X_{o})^{-1}X_{o}'\underset{\sim}{\varepsilon}_{o}. \qquad (2.2.5)$$

Figure 2.2.1 gives the parameter space representation of this procedure in case of two parameter model. True and

approximate sum of squares contours could in principle be obtained by plotting points (θ_1, θ_2) which satisfy

$$S(\underset{\sim}{\theta}) = c \text{ and } \bar{S}(\underset{\sim}{\theta}) = c \qquad (2.2.6)$$

respectively for various values of the constant c. The approximate contours are necessarily elliptical while the true ones will typically have the appearance of distorted ellipses. Thus we move from the initial point $\underset{\sim}{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$ to the new point $\underset{\sim}{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)})$, the minimum point of the approximate sum of squares. Around the new estimates $\theta^{(1)}$, the model is again linearized and new estimates $\theta^{(2)}$ are obtained, and so on.

This procedure, which we call the Gauss iteration procedure, does not always lead to convergence. In fact, when the initial estimates of parameters $\theta^{(0)}$ are poor and/or the model is severely nonlinear, it has sometimes been found that wild oscilation occurs from iteration to iteration. Figure 2.2.1 shows a typical case of divergence, where the Gauss solution vector (2.2.5) "overshoots" and the sum of squares given by the new estimates $\underset{\sim}{\theta}^{(1)}$ is larger than that given by the initial estimates $\underset{\sim}{\theta}^{(0)}$. Repetition of this process could lead further and further away from the minimum point. A frequent cause of divergence is that the adjustments $\underset{\sim}{\theta}_1 - \underset{\sim}{\theta}_1^{(0)}$ are too large and so invalidate the linear approximation (2.2.1).

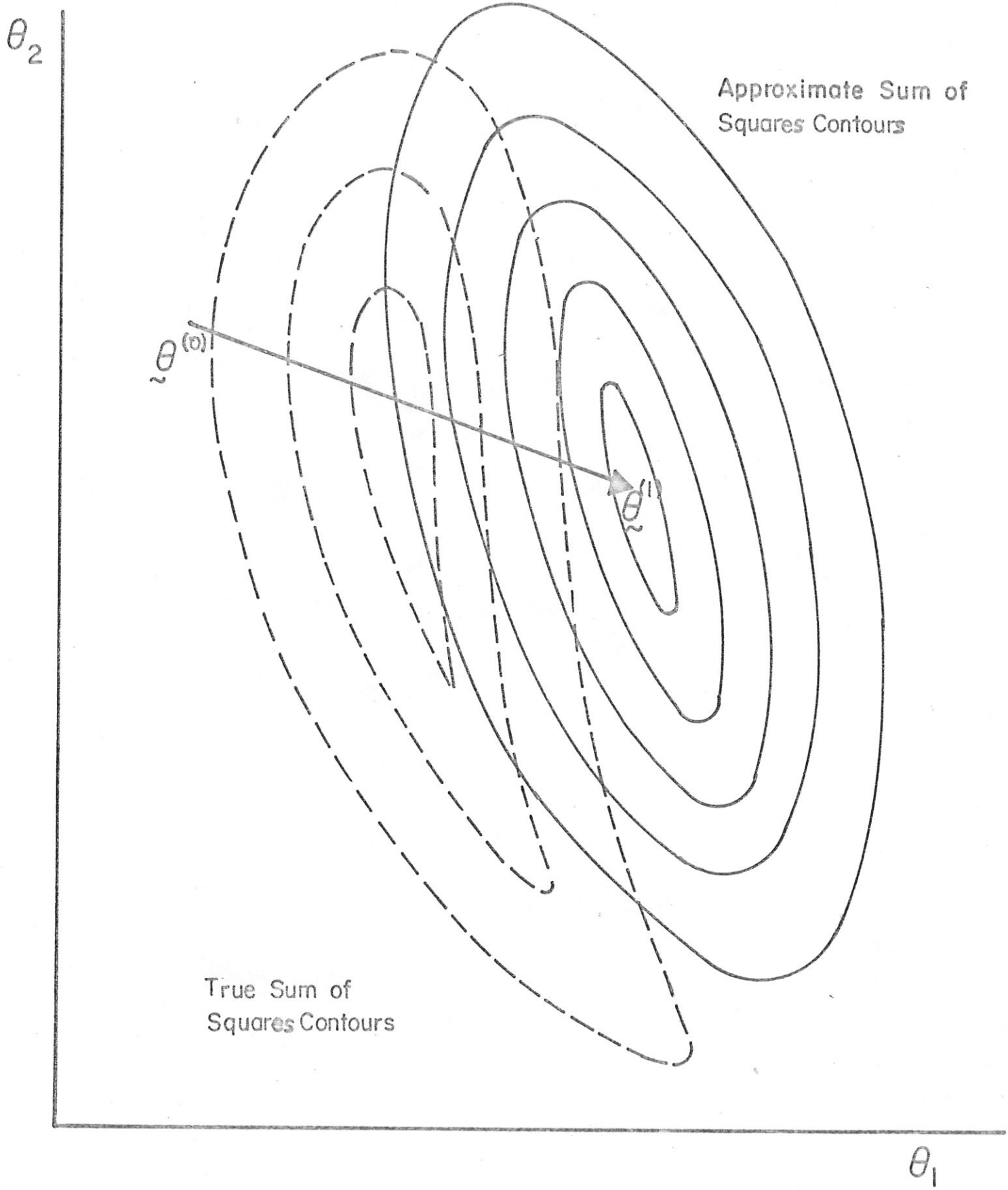


Figure 2.2.1 Parameter space representation of the Gauss iteration.

2.3 Modified Gauss Iteration

One way to overcome the difficulty of overshooting in the Gauss iteration is to go only part of the way along the Gauss solution vector $\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)} = (X_o'X_o)^{-1}X_o'\underset{\sim}{\varepsilon}_o$. Thus the adjustment vector $\underset{\sim}{\theta} - \underset{\sim}{\theta}^{(0)}$ is given by

$$\underset{\sim}{\theta} - \underset{\sim}{\theta}^{(0)} = v(X_o'X_o)^{-1}X_o'\underset{\sim}{\varepsilon}_o \qquad (2.3.1)$$

where v is a certain positive quantity less than unity. This modified Gauss iteration was suggested by Box [4] and incorporated into a computer program described by Booth, Box, Muller and Peterson [1]. In order to determine the value of v that approximately minimizes the sum of squares along the Gauss solution vector, they used what may be called the "halving and doubling" method in which, starting from v=1, the value of v is successively halved (or doubled) until the sum of squares finally starts to increase and then a quadratic curve is fitted to the last three points to locate an approximate local minimum. Hartley [10] later proved that, under a set of mild regularity conditions, the modified Gauss iteration as described above converges to the solution of $\partial S(\theta)/\partial\theta_j = 0$; $j = 1,2,\ldots,p$ and also proposed a similar method to determine the value of v. This method suffers from the disadvantage that extensive further calculations may be needed to decide the best point on the Gauss vector.

In particular the function $f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})$ must be evaluated for $u = 1,2,\ldots,n$ to calculate $S(\underset{\sim}{\theta})$ at each new "test point" in the parameter space.

2.4 Application of response surface methods

In an earlier paper, Box and Coutie [6] had suggested application of response surface techniques ([9], [2]) to the problem of nonlinear least squares. If the initial parameter values are remote from the minimum, the sum of squares surface $S(\underset{\sim}{\theta})$ could be locally approximated by a polynomial in $\underset{\sim}{\theta}$ of first degree. Sums of squares were therefore determined at a series of points in the parameter space which formed a first order design, making it possible to calculate a direction of steepest descent. This direction was followed until an increase in the sum of squares was encountered. The whole process was repeated until the need for a second degree approximation became manifest. Then, what we may call the "second order procedure" was used in which the sum of squares was determined at a series of points in the parameter space arranged in a second order design, from which the second degree polynomial could be fitted and the approximate minimum and a confidence region determined.

It is easy to show that this method which makes use only of the sums of squares of the residuals and not of the

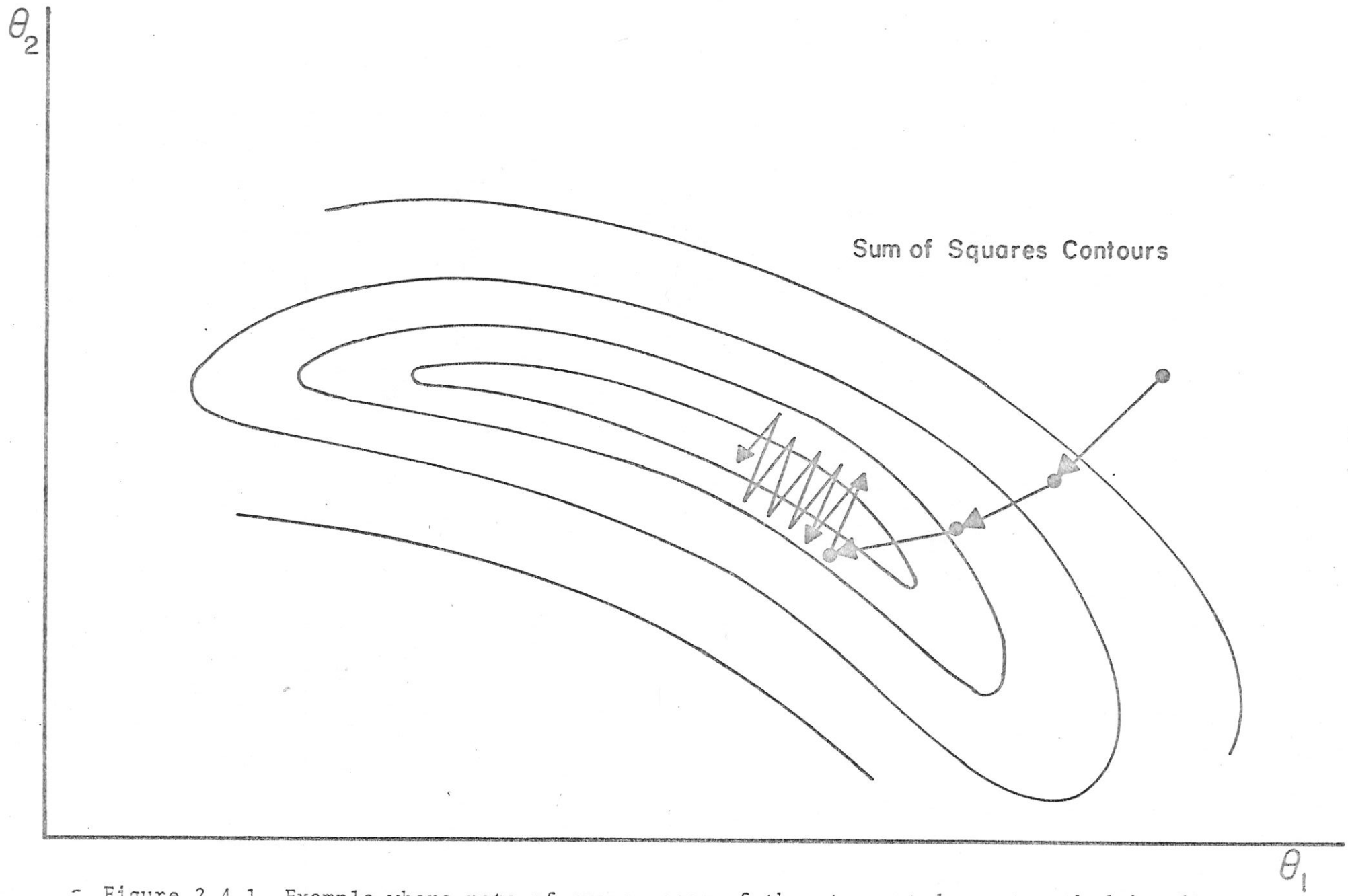


Figure 2.4.1 Example where rate of convergence of the steepest descent method is slow.

individual residuals themselves, makes only partial use of available information and, further, that when this missing information is included, we are brought back to the method suggested by Gauss. (See Appendix A2.1).

There are many versions of the steepest decent method which could be applied to the nonlinear least squares problem. However, they all suffer from the difficulties of this kind and the rate of convergence could be extremely slow for a ridgy minimum as is illustrated in Figure 2.4.1.

2.5 Levenberg-Marquardt's constrained iteration

2.5.1 Levenberg's damped least squares

Levenberg [11] tried to overcome the difficulty of "overshooting" in the Gauss iteration by introducing constraints into the minimization of the sum of squares. Instead of minimizing the approximate sum of squares $\bar{S}(\underset{\sim}{\theta})$ itself he proposed to minimize

$$F(\underset{\sim}{\theta}) = \bar{S}(\underset{\sim}{\theta}) + \lambda \sum_{i=1}^{p} \omega_i (\theta_i - \theta_i^{(0)})^2 \qquad (2.5.1)$$

where $\lambda\omega_1, \lambda\omega_2, \ldots, \lambda\omega_p$ are weighting factors expressing the relative importance of damping the different increments. Substituting (2.2.3) into (2.5.1),

$$F(\underset{\sim}{\theta}) = \underset{\sim}{\varepsilon}_o'\underset{\sim}{\varepsilon}_o - 2\underset{\sim}{\varepsilon}_o' X_o(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) + (\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})' X_o' X_o (\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})$$
$$+ \lambda(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})'\Omega(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}), \qquad (2.5.2)$$

where Ω is a p×p diagonal matrix whose i th diagonal element is ω_i. Setting the derivatives $\partial F(\underset{\sim}{\theta})/\partial\theta_i$ to zero

$$-X_o'\underset{\sim}{\varepsilon}_o + X_o' X_o(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) + \lambda\Omega(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) = 0 \; . \qquad (2.5.3)$$

Solving for $\underset{\sim}{\theta}$

$$\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)} = (X_o' X_o + \lambda\Omega)^{-1} X_o'\underset{\sim}{\varepsilon}_o \; . \qquad (2.5.4)$$

Geometrically, in the parameter space representation, this amounts to minimizing the approximate sum of squares $\bar{S}(\underset{\sim}{\theta})$ on the elliptical constraint whose principal axes are parallel to the axes of $\theta_1, \theta_2, \ldots, \theta_p$. This is illustrated in Figure 2.5.1. Levenberg proved that, provided the true sum of squares $S(\underset{\sim}{\theta})$ does not have stationary values at the current estimate $\underset{\sim}{\theta}^{(0)}$, the sum of squares initially decreases as we move off the initial point $\underset{\sim}{\theta}^{(0)}$ changing the value of λ. He also recommended using

$$\Omega = I_p \qquad (2.5.5)$$

which corresponds to the use of a spherical constraint.

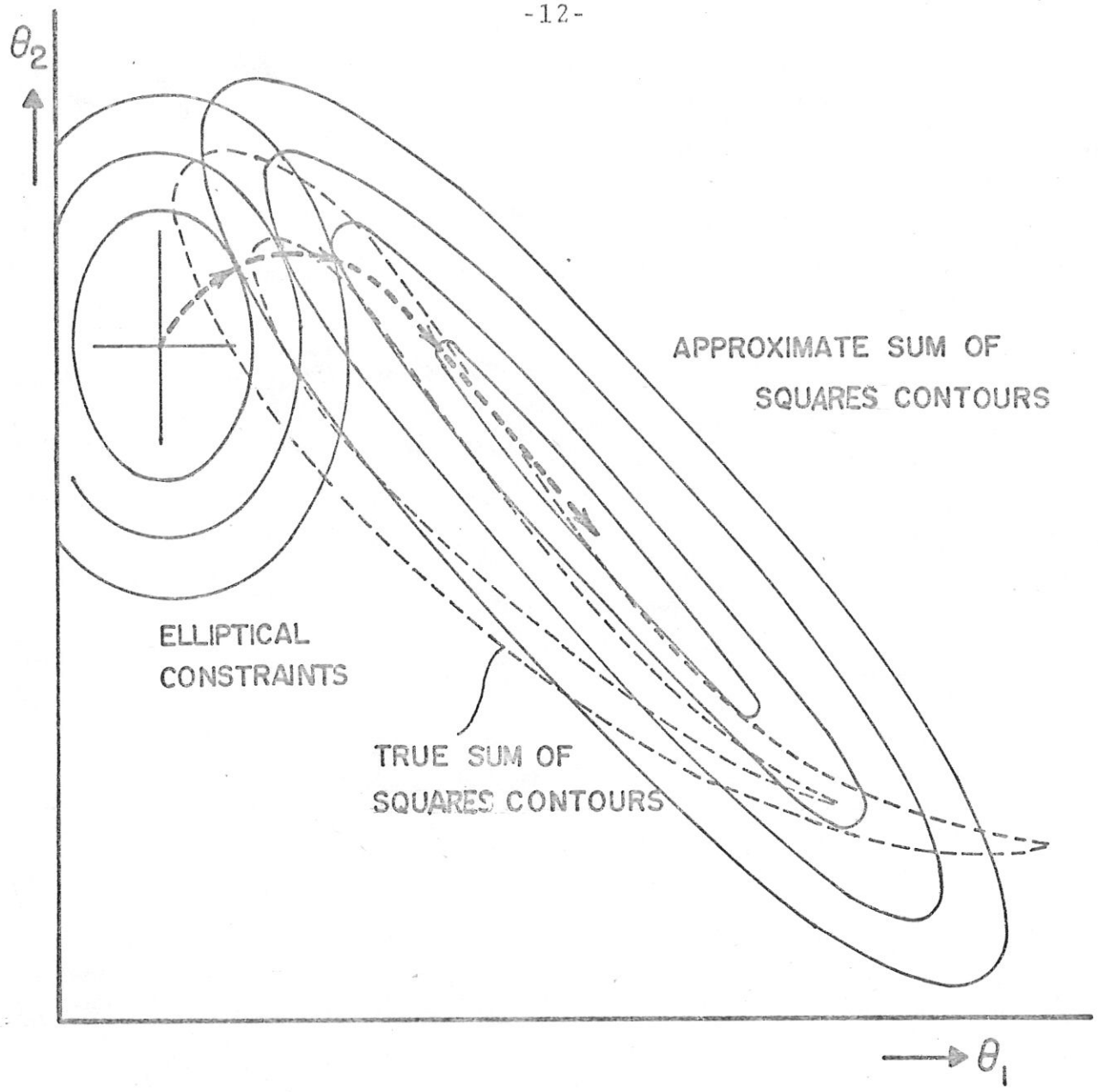


Figure 2.5.1 Parameter space representation of Levenberg's constrained minimization.

Fig 2.5.1

Another proposal Levenberg made was that Ω be set equal to a p×p diagonal matrix D whose (j,j) element was the j th diagonal element of $X_o'X_o$ so that

$$= D = \begin{bmatrix} [11] & & & \\ & [22] & & \\ & & \ddots & \\ & & & [jj] \end{bmatrix}$$

As will be seen in the next section, this has the effect of making the problem invariant under scale changes.

2.5.2 Constrained minimization in the transformed space

Now usually there are many ways in which a problem could be parameterized. Instead of considering $\theta_1,\theta_2,\ldots,\theta_p$ we could with equal reason consider, say, $\psi_1,\psi_2,\ldots,\psi_p$, where $\psi_j = \psi_j(\underset{\sim}{\theta})$ in some 1:1 transformation of $\underset{\sim}{\theta}$. Clearly the nature of constraints applied (for example in Levenberg's procedure) would differ depending on which parameterization is considered. In particular, consider a linear transformation

$$\underset{\sim}{\psi} = T\underset{\sim}{\theta} \qquad (2.5.7)$$

with T a p×p nonsingular matrix. The linearized model in the new parameters $\underset{\sim}{\psi}$ will be

$$f_\psi \cong f_o + Z_o(\underset{\sim}{\psi}-\underset{\sim}{\psi}^{(0)}) \qquad (2.5.8)$$

where $\underset{\sim}{f}_\psi$ is the n×1 vector of $f(\underset{\sim}{\xi}_u,\underset{\sim}{\theta}(\underset{\sim}{\psi}))$; u = 1,2,...,n,

$\underset{\sim}{\psi}^{(0)} = T\underset{\sim}{\theta}^{(0)}$, and Z_o is the n×p matrix whose (u,j) element is $[\partial f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}(\underset{\sim}{\psi}))/\partial\psi_j]_{\underset{\sim}{\psi}=\underset{\sim}{\psi}^{(0)}}$. Notice that Z_o is related to X_o by

$$Z_o = X_o T^{-1}. \tag{2.5.9}$$

If in the space of $\underset{\sim}{\psi}$, we minimize the approximate sum of squares on the spherical constraint, we will obtain

$$\underset{\sim}{\psi}-\underset{\sim}{\psi}^{(0)} = (Z_o'Z_o+\lambda I_p)^{-1}Z_o'\underset{\sim}{\varepsilon}_o, \tag{2.5.10}$$

in exactly the same manner as in Section 2.5.1. Transforming this result back into the original space of $\underset{\sim}{\theta}$ gives

$$\begin{aligned}\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)} &= T^{-1}(T'^{-1}X_o'X_oT^{-1}+\lambda I_p)^{-1}T'^{-1}X_o\underset{\sim}{\varepsilon}_o \\ &= (X_o'X_o+\lambda T'T)^{-1}X_o'\underset{\sim}{\varepsilon}_o.\end{aligned} \tag{2.5.11}$$

From this it follows that minimization of the approximate sum of squares with a spherical constraint in the new metric $\underset{\sim}{\psi}$ is equivalent to minimization in the original metric $\underset{\sim}{\theta}$ with some elliptical constraint

$$(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)})'T'T(\underset{\sim}{\theta}-\underset{\sim}{\theta}^{(0)}) = \text{constant}. \tag{2.5.12}$$

Marquardt [12] did not consider this general transformation but considered only a special case by which the problem was made invariant under scale changes. This amounted

to using the constraint matrix Ω=D in (2.5.4), or alternatively, as he described it, to using spherical constraints with scaled parameters

$$\psi_j = [jj]^{\frac{1}{2}}\theta_j \qquad j=1,2,\ldots,p \tag{2.5.13}$$

In this parametization then the parameters $\underset{\sim}{\psi}$ were scaled so that the contours of the approximate sum of squares were contained in squares, cubes or hypercubes, as illustrated in Figure 2.5.2 for the case p=2. Marquardt also pointed out that (i) when λ=0 the equation (2.5.10) gives the Gauss solution and (ii) when λ is very large

$$\underset{\sim}{\psi}-\underset{\sim}{\psi}^{(0)} \propto Z_o'\underset{\sim}{\varepsilon}_o \tag{2.5.14}$$

which is the steepest descent vector on the approximate sum of squares surface in the matrics $\underset{\sim}{\psi}$. Thus by choosing an intermediate value for λ, a compromise between the Gauss method and the steepest descent method is obtained. Based on this observation, Marquardt proposed what we call the (λ,ν) algorithm in which λ is decreased gradually by a factor ν from a relatively large initial valve λ_o as the iteration proceeds. Thus in successive iterations he would choose Levenberg's matrix $\lambda\Omega$ to be $\lambda_o D$, $\lambda_o\nu^{-1}D$, $\lambda_o\nu^{-2}D$ etc. This method would then possess the apparent virtue of each

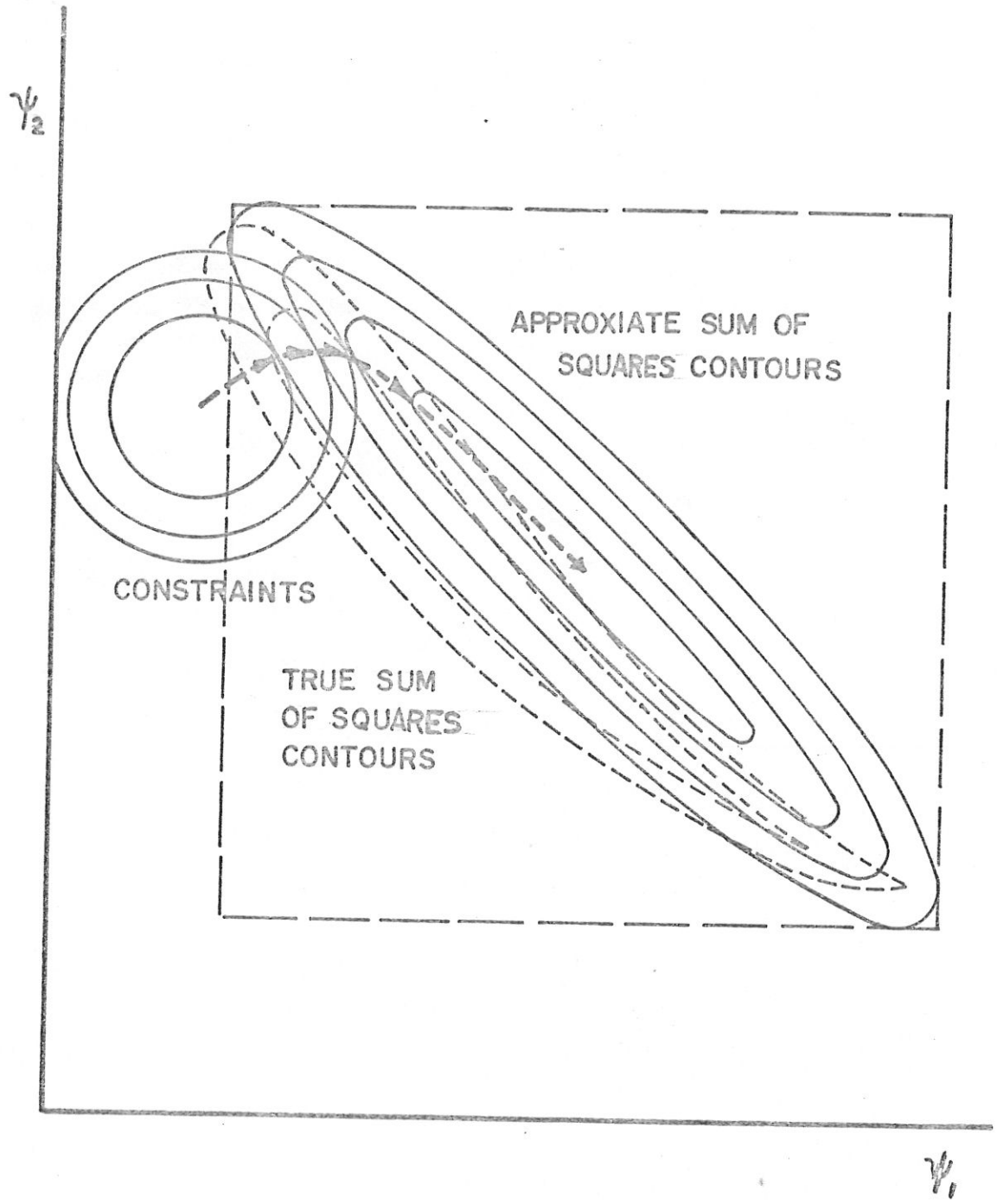


Figure 2.5.2 Constrained minimization of sum of squares in Marquardt's scale invariant metric.

method in the circumstances where it is most effective. Marquardt's recommendations were followed by Meeter [13] in writing the program GAUSHAUS (later UWHAUS) at the University of Wisconsin and similar programs have been available elsewhere.

In Marquardt's method then the constrained minimization is carried out in parameter metrics which are scale invariant. However, it is easy to see that we can go further and conduct the minimization in metrics which are not only scale invariant but which are also linearly invariant. Such a transformation $\underset{\sim}{\psi} = H\underset{\sim}{\theta}$ is provided by any H for which

$$H'H = X_o'X_o. \qquad (2.5.15)$$

To see this suppose we make the arbitrary non-singular transformation

$$\underset{\sim}{\tilde{\theta}} = L\underset{\sim}{\theta}; \qquad (2.5.16)$$

then X_o will be transformed into

$$\tilde{X}_o = X_o L^{-1} . \qquad (2.5.17)$$

The transformation of $\underset{\sim}{\tilde{\theta}}$ corresponding to the one given by the equation (2.5.15) will be

$$\tilde{\underset{\sim}{\psi}} = \tilde{H}\tilde{\underset{\sim}{\theta}} \text{ with } \tilde{H}'\tilde{H} = \tilde{X}_o'\tilde{X}_o \qquad (2.5.18)$$

However, the requirement on $\tilde{H}$ will be satisfied by $\tilde{H} = HL^{-1}$ since

$$\tilde{H}'\tilde{H} = (HL^{-1})'(HL^{-1}) = L^{-1}{}'X_o'X_oL^{-1} = \tilde{X}_o'\tilde{X}_o \quad . \qquad (2.5.19)$$

Therefore

$$\tilde{\underset{\sim}{\psi}} = \tilde{H}\tilde{\underset{\sim}{\theta}} = (HL^{-1})(L\underset{\sim}{\theta}) = H\underset{\sim}{\theta} = \underset{\sim}{\psi}, \qquad (2.5.20)$$

establishing that $\underset{\sim}{\psi}$ is invariant under linear transformation.

In such a metric $\underset{\sim}{\psi}$, the sum of squares contours for the linearized model

$$(\underset{\sim}{y}-\underset{\sim}{f}_o-Z_o(\underset{\sim}{\psi}-\underset{\sim}{\psi}^{(0)}))'(\underset{\sim}{y}-\underset{\sim}{f}_o-Z_o(\underset{\sim}{\psi}-\underset{\sim}{\psi}^{(0)})) = \text{constant} \qquad (2.5.21)$$

will be spherical because

$$Z_o = X_oT^{-1} = X_oH^{-1} \qquad (2.5.22)$$

and so

$$Z_o'Z_o = (X_oH^{-1})'(X_oH^{-1}) = H^{-1}{}'X_o'X_oH^{-1} \qquad (2.5.23)$$
$$= I_p.$$

The representation would now be that illustrated for p=2 in Figure 2.5.3 in which the constraining contours are circles, spheres, or hyperspheres, when the approximate sum of squares contours are circles, spheres or hyperplanes. However, if we carry out the constrained minimization in this linearly invariant metric, on transforming the result back into the original metric $\underset{\sim}{\theta}$ by setting $T'T = H'H = X_o'X_o$ in (2.5.11) we obtain

$$\underset{\sim}{\theta} - \underset{\sim}{\theta}^{(0)} = (X_o'X_o + \lambda X_o'X_o)^{-1} X_o' \underset{\sim}{\varepsilon}_o$$

$$= \frac{1}{1+\lambda} (X_o'X_o)^{-1} X_o' \underset{\sim}{\varepsilon}_o \quad . \qquad (2.5.24)$$

Thus by setting $v = 1/(1+\lambda)$ in (2.3.1) we see, somewhat surprisingly that the Levenberg-Marquardt constrained minimization in the more reasonable linearly invariant metric is exactly the modified Gauss method already discussed. It should be noted also that, in the linearly invariant metric there is no question of a compromise between the Gauss method and the steepest descent method. In fact, in this metric, because of the property of Z_o given in (2.5.23), the Gauss solution vector and the steepest descent vector are identical. It is noted that, in Figure 2.5.2, the solution given by the constrained minimization, as the constraint is relaxed, follows the curved path starting into the direction of

steepest descent and ending up with the Gauss solution while, in Figure 2.5.3, the path taken by the solution is a straight line which is the Gauss solution vector. In the latter figure, the path which the Marquardt method would take in these linearly invariant metrics is indicated by a bold connected curve.

So far as the problem of speedy convergence is concerned, there does not therefore seem to exist any concrete theoretical basis for Marquardt's interpolation between the two classical methods.

One incidental advantage of the Levenberg-Marquardt procedure is that the matrix $(X_o'X_o+\lambda\Omega)$ can be inverted even when the matrix $X_o'X_o$ is singular or nearly singular, thus always giving a "solution". Practical experience, however, leads us to believe that the possibility of not having a singularity or near-singularity brought to one's attention is a disadvantage rather than an advantage. It has often been pointed out (for example [2]) that the minimum is often better envisioned as being approximated by a line, plane or hyperplane rather than by a point. When this happens, it is important that it be brought to the investigator's notice. One method for doing this is by means of a canonical analysis as has been suggested, for example, in [5].

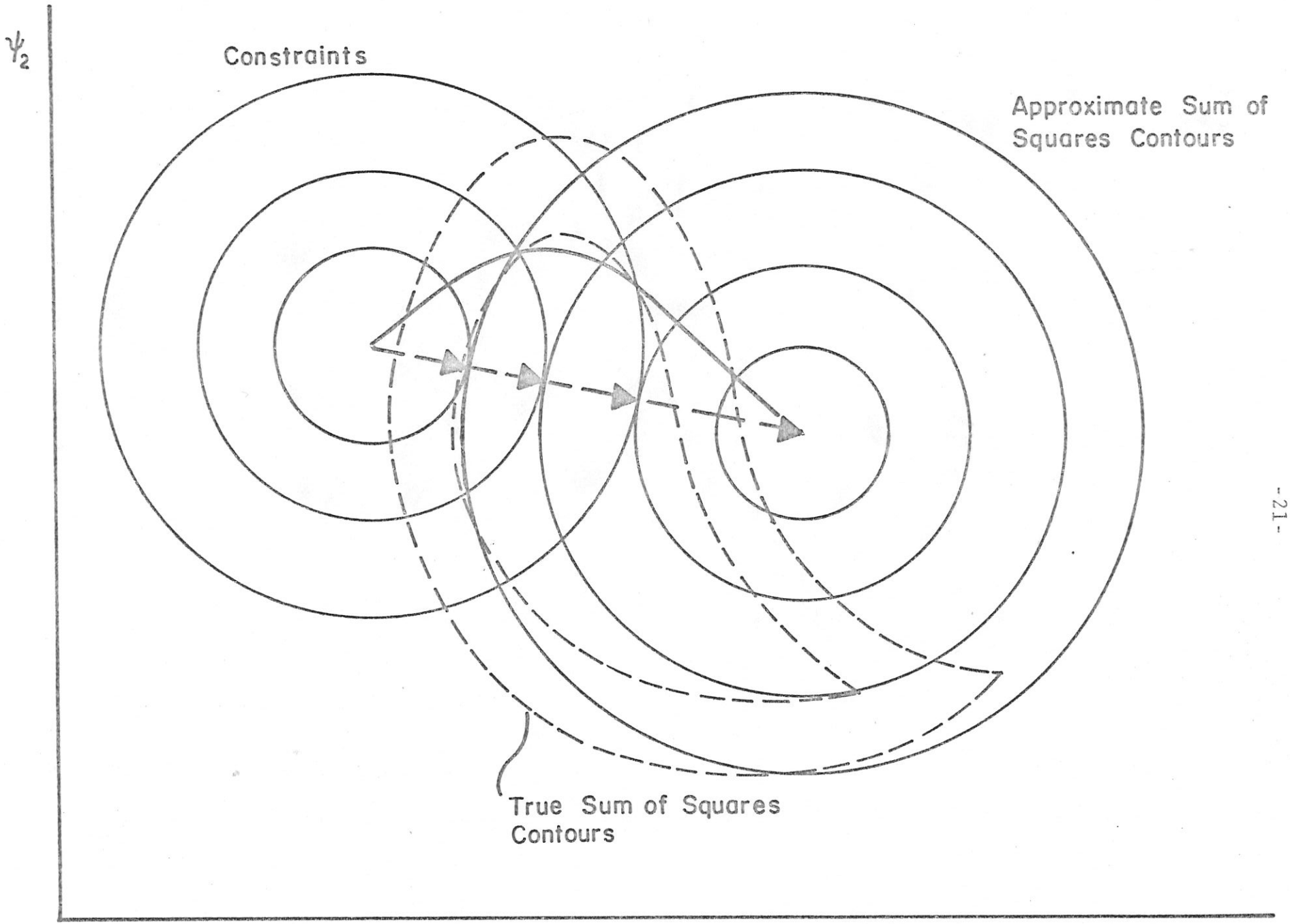


Figure 2.5.3 Constrained minimization of the sum of squares in the linearly invariant metric.

2.6 Methods to determine how far one should go along the Gauss solution vector

In the preceding section, we have given theoretical support to the idea that we should explore the straight line given by the Gauss vector itself rather than the curved path followed by Marquardt. We have done this by demonstrating that the Gauss vector may be arrived at by applying Levenberg-Marquardt constrained minimization in the linearly invariant parameter metrics.

An important fact to bear in mind is that, provided the current best estimate $\underset{\sim}{\theta}^{(0)}$ is not the stationary point, it follows from Levenberg's result mentioned in the section 2.5.1 that the true sum of squares initially decreases when we start off the point $\underset{\sim}{\theta}^{(0)}$ along the Gauss solution vector $\underset{\sim}{\theta}^{(g)}-\underset{\sim}{\theta}^{(0)}$ so that this direction is certainly worth investigating.

It still remains to be decided how far one should go along the Gauss vector. The "halving and doubling" method already mentioned could be used at the expense of further calculation. We also could use the (λ,ν) algorithm by Marquardt. Although no compromise between the original Gauss method and the steepest descent method is here involved, it might still make sense to gradually decrease λ so as to constrain the iteration less and less as the minimum is

approached in successive iterations. Again, however, because there is no way to know in advance how best to choose λ and ν, the method is inefficient, and does not make use of information already available from previous computation.

Modification A To obtain the Gauss solution vector it is necessary to compute the matrix X_o of the partial derivatives $[\partial f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta})/\partial\theta_j]_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(0)}}$. Using this matrix, it is clearly possible to obtain the initial rate of change of the true sum of squares along the Gauss solution vector. In fact,

$$\left[\frac{dS}{dv}\right]_{v=0} = \sum_{i=1}^{p} \left[\frac{\partial S}{\partial \theta_i}\right]_{v=0} \left[\frac{d\theta_i}{dv}\right]_{v=0}$$

$$= \left[\frac{\partial S}{\partial \underset{\sim}{\theta}}\right]'_{v=0} \left[\frac{d\underset{\sim}{\theta}}{dv}\right]_{v=0} \qquad (2.6.1)$$

where $\left[\frac{\partial S}{\partial \underset{\sim}{\theta}}\right]_{v=0}$ is the p×1 vector of $\left[\frac{\partial S}{\partial \theta_i}\right]_{v=0}$; i=1,2,...,p and $\left[\frac{d\underset{\sim}{\theta}}{dv}\right]_{v=0}$ is the p×1 vector of $\left[\frac{d\theta_i}{dv}\right]_{v=0}$; i=1,2,...,p.

However, since $S = (\underset{\sim}{y}-\underset{\sim}{f}_\theta)'(\underset{\sim}{y}-\underset{\sim}{f}_\theta)$ and $\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)} = (X_o'X_o)^{-1}X_o'\underset{\sim}{\varepsilon}_o$,

we obtain

$$\left[\frac{dS}{dv}\right]_{v=0} = -2\underset{\sim}{\varepsilon}_o' X_o (X_o' X_o)^{-1} X_o' \underset{\sim}{\varepsilon}_o$$

$$= -2(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)})' X_o' \underset{\sim}{\varepsilon}_o . \qquad (2.6.2)$$

Incidentally, equation (2.6.2) gives a direct proof of a corollary of Levenberg's result mentioned above that the true sum of squares initially decreases as we start moving from $\underset{\sim}{\theta}^{(0)}$ to $\underset{\sim}{\theta}^{(g)}$, because provided that $X_o' X_o$ is positive definite (as is usually the case), $\left[\frac{dS}{dv}\right]_{v=0}$ will be negative except when $X_o' \varepsilon_o = -\frac{1}{2}\left[\frac{\partial S}{\partial \theta}\right]_{v=0}$ is $\underset{\sim}{0}$.

To locate a point along the Gauss solution vector at which the true sum of squares is approximately minimized, we first suppose that the true sum of squares follows, along this vector, the form of a quadratic function

$$S = a + bv + cv^2 . \qquad (2.6.3)$$

We can determine the constants a, b and c in (2.6.3) by setting $S = S_o$ for $v=0$, $S = S_g$ for $v=1$ and

$$\left[\frac{dS}{dv}\right]_{v=0} = -2(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)})' X_o' \underset{\sim}{\varepsilon}_o . \qquad (2.6.4)$$

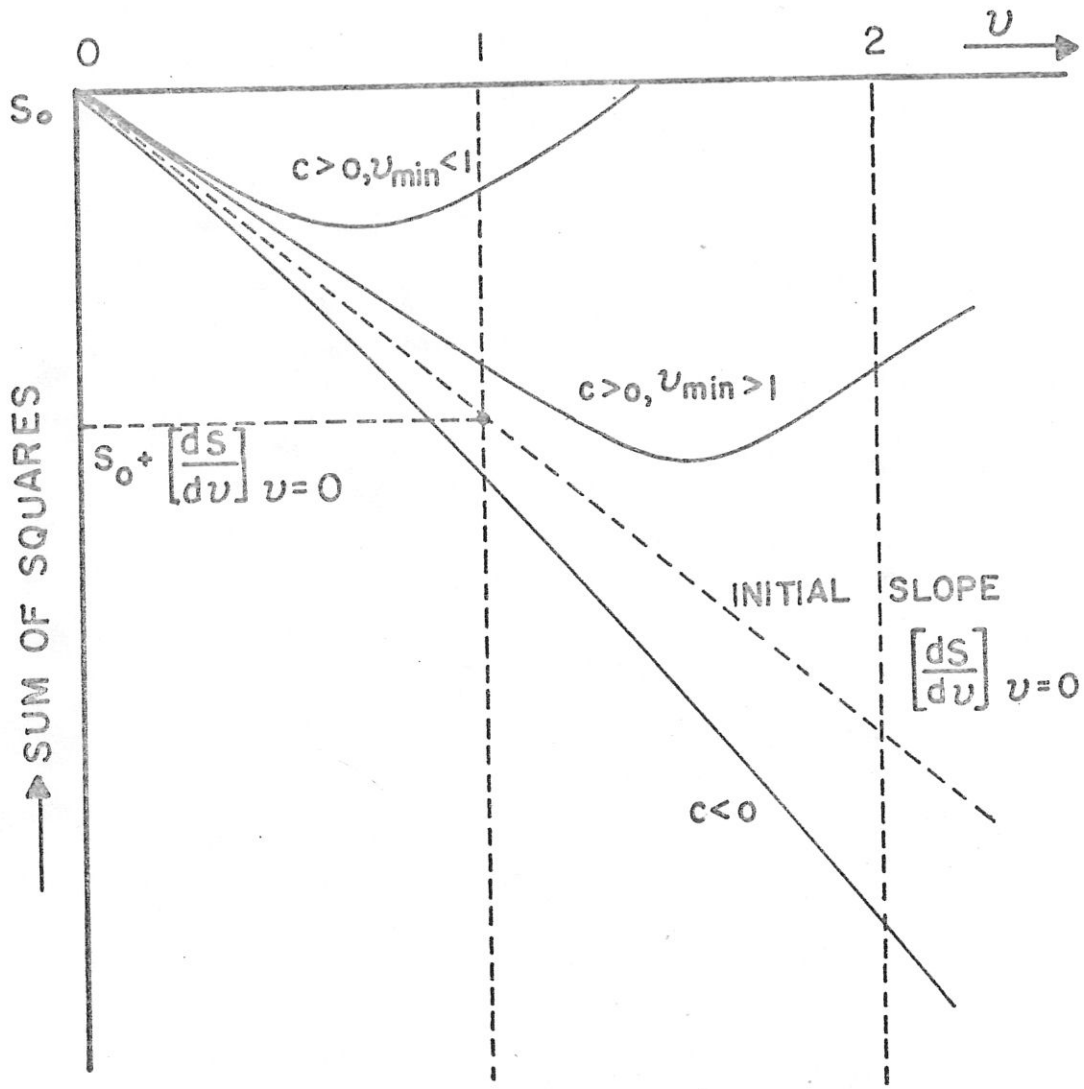


Figure 2.6.1 Various quadratic curves approximating the sum of squares along Gauss vector.

Consequently, provided $c = S_g - S_o + 2(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)})' X_o' \underset{\sim}{\varepsilon}_o > 0$, the value of v for which the true sum of squares is approximately minimized is given by

$$v_{min} = \frac{(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)})' X_o' \underset{\sim}{\varepsilon}_o}{S_g - S_o + 2(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)})' X_o' \underset{\sim}{\varepsilon}_o} \quad . \qquad (2.6.5)$$

Thus, we may take our next estimate as

$$\underset{\sim}{\theta}^{(1)} = \underset{\sim}{\theta}^{(0)} + v_{min}(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)}) \quad . \qquad (2.6.6)$$

If $c \leq 0$, we have that the actual sum of squares S_g at distance v=1 is already smaller than that predicted by the initial slope since in this case $S_g \leq S_o + [dS/dv]_{v=0}$. Therefore we may settle at v=1, or double, or redouble the distance, checking at each point to see if the decreasing trend is continuing. Figure 2.6.1 illustrates various situations that could occur.

Modification B A procedure developed following a suggestion by Jack Draffen[1] is another method that makes use of the existing information in an interesting manner. The quantities S_o and S_g are obtained by computing, squaring, and summing the elements of the two residual vectors

---

[1] Personal communication (1972).

$$\underset{\sim}{\varepsilon}_o' = (y_1 - f(\underset{\sim}{\xi}_1, \underset{\sim}{\theta}^{(0)}), y_2 - f(\underset{\sim}{\xi}_2, \underset{\sim}{\theta}^{(0)}), \ldots, y_n - f(\underset{\sim}{\xi}_n, \underset{\sim}{\theta}^{(0)}))$$

and

$$\underset{\sim}{\varepsilon}_g' = (y_1 - f(\underset{\sim}{\xi}_1, \underset{\sim}{\theta}^{(g)}), y_2 - f(\underset{\sim}{\xi}_2, \underset{\sim}{\theta}^{(g)}), \ldots, y_n - f(\underset{\sim}{\xi}_n, \underset{\sim}{\theta}^{(g)})) \tag{2.6.7}$$

Consider a point along the Gauss solution vector with the distance $v||\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)}||$ away from the origin $\underset{\sim}{\theta}^{(0)}$. By linear interpolation we can estimate the residuals at this point by

$$\underset{\sim}{\varepsilon} = (1-v)\underset{\sim}{\varepsilon}_o + v\underset{\sim}{\varepsilon}_g \tag{2.6.8}$$

which may be written

$$\underset{\sim}{\varepsilon}_o = v(\underset{\sim}{\varepsilon}_o - \underset{\sim}{\varepsilon}_g) + \underset{\sim}{\varepsilon} \quad . \tag{2.6.9}$$

Thus the value $\hat{v}$ of v for which the sum of squares of the estimated residuals will be as small as possible can be obtained by regressing $\underset{\sim}{\varepsilon}_o$ on $\underset{\sim}{\varepsilon}_o - \underset{\sim}{\varepsilon}_g$ so that

$$\hat{v} = \frac{(\underset{\sim}{\varepsilon}_o - \underset{\sim}{\varepsilon}_g)' \underset{\sim}{\varepsilon}_o}{(\underset{\sim}{\varepsilon}_o - \underset{\sim}{\varepsilon}_g)' (\underset{\sim}{\varepsilon}_o - \underset{\sim}{\varepsilon}_g)} , \tag{2.6.10}$$

whence our next estimate of parameters is obtained by

$$\underset{\sim}{\theta}^{(1)} = \underset{\sim}{\theta}^{(0)} + \hat{v}(\underset{\sim}{\theta}^{(g)} - \underset{\sim}{\theta}^{(0)}) \quad . \tag{2.6.11}$$

Again, computing $\hat{v}$ is very simple and makes use of information already available.

The relationship of this procedure to modification A can be seen as follows. Equation (2.6.9) can be written

$$\underset{\sim}{y} - \underset{\sim}{f}_o = v(\underset{\sim}{f}_g - \underset{\sim}{f}_o) + \underset{\sim}{\varepsilon}, \qquad (2.6.12)$$

where $\underset{\sim}{f}_g$ is the n×1 vector of $f(\underset{\sim}{\xi}_u, \underset{\sim}{\theta}^{(g)})$, u=1,2,...,n. The sum of squares of the estimated residuals is thus

$$\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = (\underset{\sim}{y}-\underset{\sim}{f}_o-v(\underset{\sim}{f}_g-\underset{\sim}{f}_o))'(\underset{\sim}{y}-\underset{\sim}{f}_o-v(\underset{\sim}{f}_g-\underset{\sim}{f}_o)) \qquad (2.6.13)$$

which is a quadratic in v and passing through the points $(0,S_o)$ and $(1,S_g)$, and $\hat{v}$ obtained by (2.6.10) minimizes $\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$. Furthermore, differentiating equation (2.6.13), the initial slope of the sum of squares of the estimated residuals is given by

$$\left[\frac{d\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}}{dv}\right]_{v=0} = -2(\underset{\sim}{f}_g-\underset{\sim}{f}_o)'(\underset{\sim}{y}-\underset{\sim}{f}_o)$$

$$= -2(\underset{\sim}{f}_g-\underset{\sim}{f}_o)'\underset{\sim}{\varepsilon}_o \qquad (2.6.14)$$

which is identical to the initial slope used in the previous method provided that $\underset{\sim}{f}_g = \underset{\sim}{f}_o + X_o(\underset{\sim}{\theta}^{(g)}-\underset{\sim}{\theta}^{(o)})$.

In the sample space, modification B amounts to dropping a perpendicular line from $\underset{\sim}{y}$ to $\underset{\sim}{f}_g-\underset{\sim}{f}_o$ whose foot gives

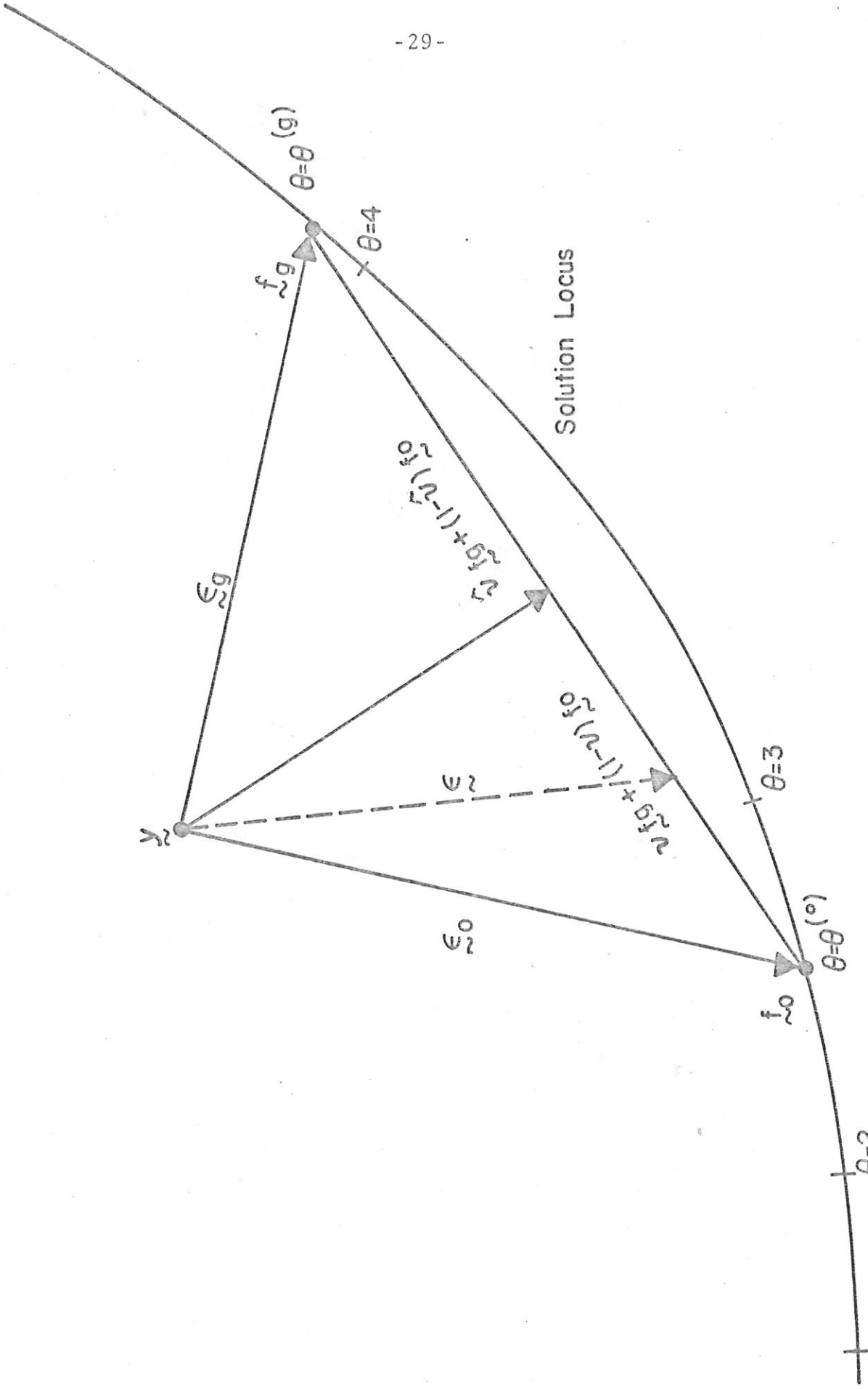


Figure 2.6.2. Sample space representation of modification B for the case of one parameter model.

the vector $(1-\hat{v})\underset{\sim}{f}_o + \hat{v}\underset{\sim}{f}_g$. This is illustrated in Figure 2.6.2 for a model with only one parameter.

2.7 Example

In testing numerical methods, little in the way of general conclusions can be based on the behavior of particular examples. Nevertheless it is worthwhile to illustrate the performance of the procedures we have described above. We do so using a simple example which Box and Hunter [7] employed previously. The results for this example are certainly not discouraging.

The model[2] is

$$f(\underset{\sim}{\xi},\underset{\sim}{\theta}) = \frac{\theta_2\theta_1\xi_1}{1+\theta_1\xi_1+5000\xi_2} \qquad (2.7.1)$$

and the data is

ξ_1	ξ_2	y
1	1	0.1165
2	1	0.2114
1	2	0.0684
2	2	0.1159

The sum of squares surface, which of course one would

[2] In the original problem there was the third parameter θ_3 which was estimated. We set it to 5000 for ease of illustration.

not usually know, is plotted in Figure 2.7.1 and is seen to be very curved and ridgy. The values $(\theta_1,\theta_2) = (300,6)$ corresponding to the point P_o in the figure were chosen for the initial estimates of parameters. The numbers of times, n_f, that the function $f(\underset{\sim}{\xi},\underset{\sim}{\theta})$ had to be evaluated before the iteration reached the minimum point P_m (at which the sum of squares was 3.82750×10^{-5}) was employed as a measure of the effectiveness of different methods.

The methods studied were:

(1) The Levenberg-Marquardt's constrained iteration with the (λ,ν) procedure.

(2) The modified Gauss iteration with the (λ,ν) procedure.

(3) The modified Gauss method with the "halving and doubling" procedure.

(4) The modified Gauss method with the modification A described in the section 2.6.

(5) The modified Gauss method with the modification B described in the section 2.6.

The values for λ_o and ν have to be specified in using the methods (1) and (2), so these two methods were compared for many different choices of λ_o and ν. Figure 2.7.2 presents the result. The performance of the methods (3), (4) and (5) is indicated by three horizontal lines since these methods are "unique", not depending on the choice of λ_o and ν.

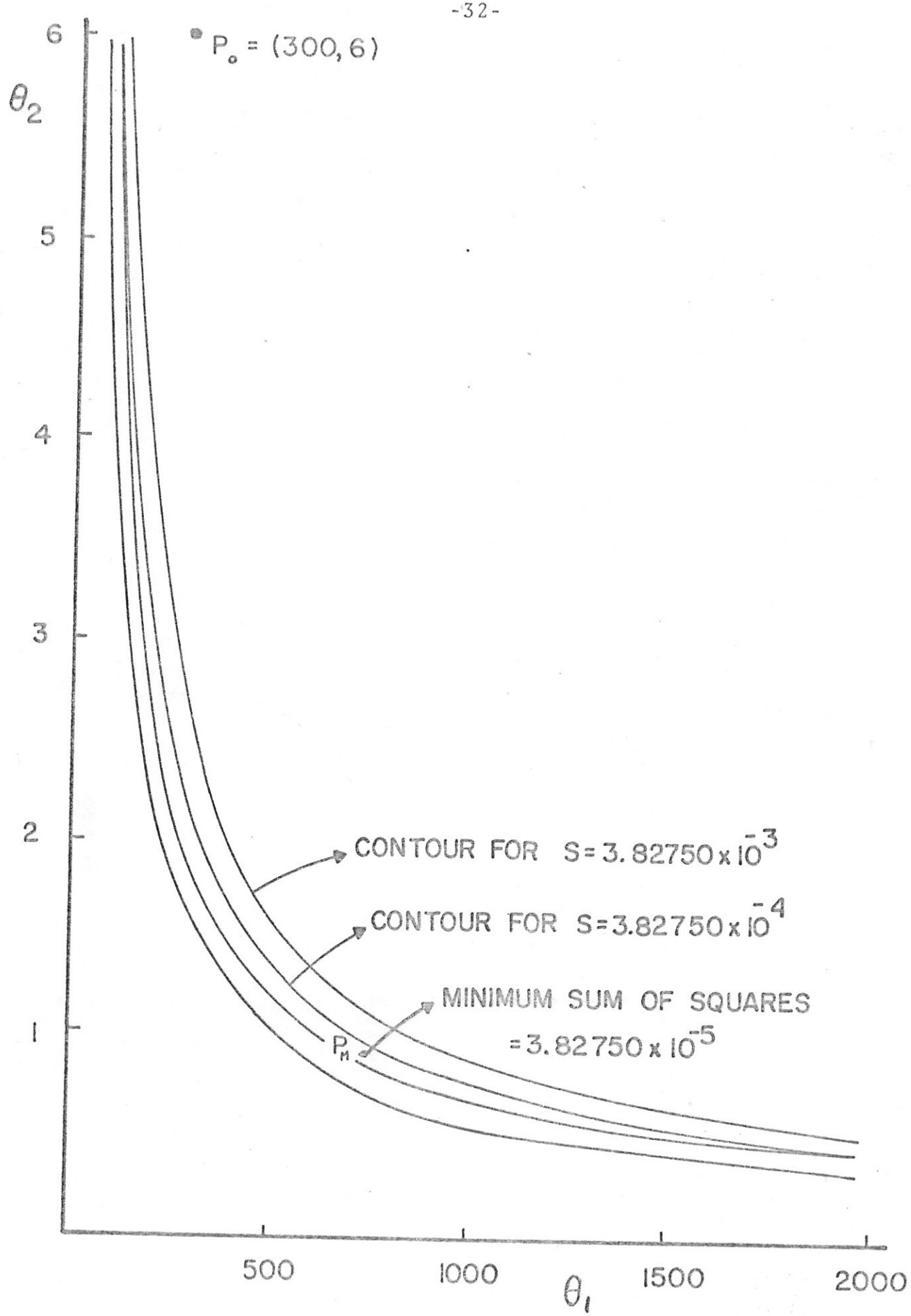


Figure 2.7.1 Sum of squares contours for the example.

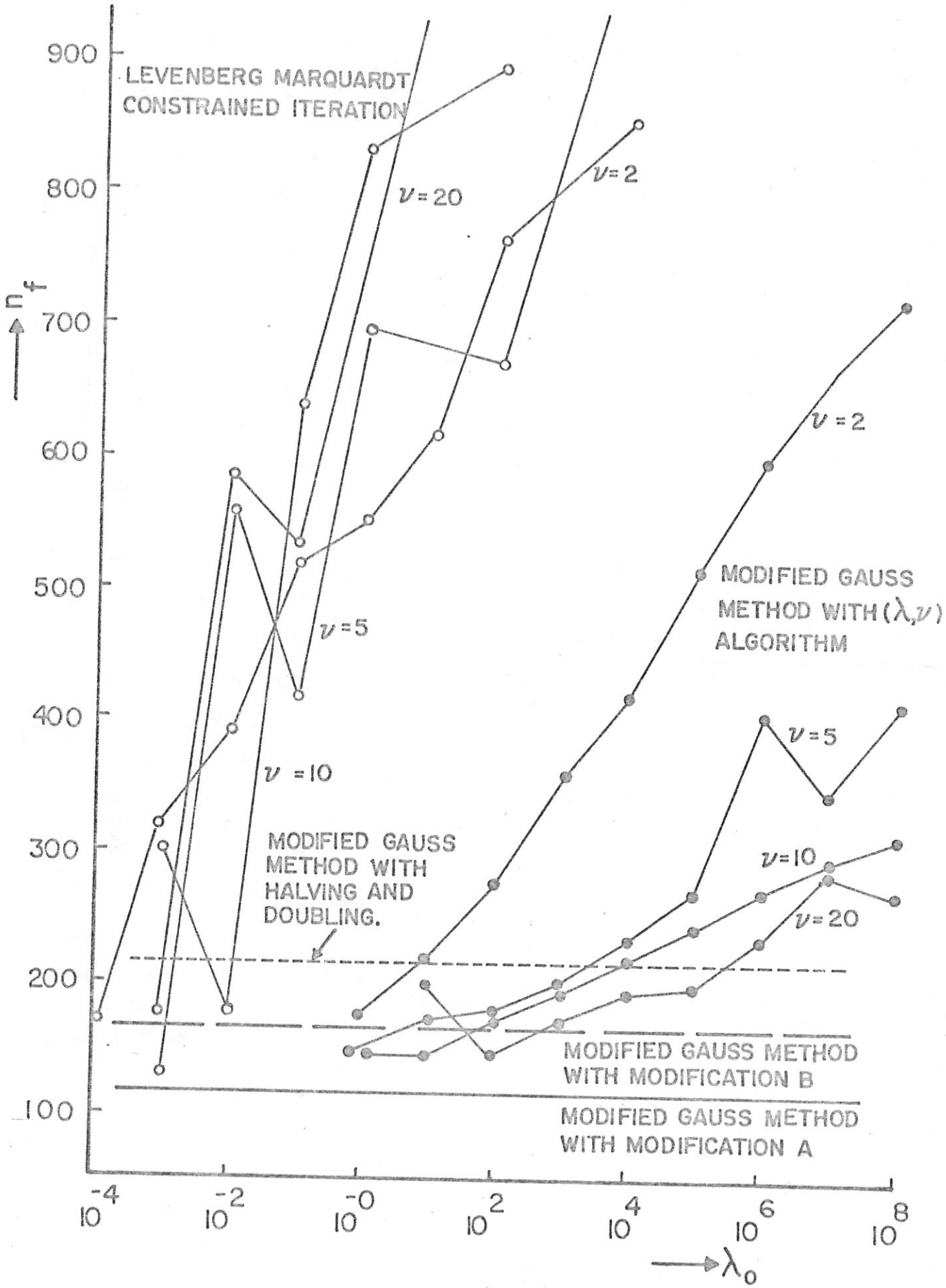


Figure 2.7.2 Comparison of the Levenberg-Marquardt constrained iteration and the modified Gauss methods.

For this particular example, the following conclusions can be drawn:

a) With the (λ,ν) procedure used in common, the modified Gauss method is far more stable than the Levenberg-Marquardt's constrained iteration in a sense that at each value of ν the range of λ_o for which the former converges with a reasonable amount of calculation is much wider (notice that the log scale is used for λ in Figure 2.7.2).

b) The modified Gauss method with the modification A turns out to be remarkably effective. The amount of necessary computation ($n_f = 116$) is even less than the best the (λ,ν) procedure can attain for the best choice of λ_o and ν($n_f = 148$ for $\lambda_o = 1$ and $\nu = 10$).

c) The modified Gauss method with the modification B also performs well, its n_f being 164 which is slightly greater than the best n_f for the methods (1) and (2).

2.8 Conclusion

Because of its practical effectiveness and apparent justification as a compromise between the steepest descent method and the Gauss method, the Levenberg-Marquardt's constrained iteration has been widely used since the publication of Marquardt's paper in 1963.

However, if the parameters are transformed into the linearly invariant metric, the steepest descent vector and the Gauss solution vector are found to be identical and thus there does not exist any need to compromise between directions given by the two different methods. We have also shown that the constrained minimization in this metric is merely equivalent to using the modified Gauss method in the original metrics which was proposed earlier.

Modifications A and B, to determine how far one should go along the Gauss solution vector, are also proposed. These have the virtue that their computations involve only information that already exists. A numerical illustration for a particular set of data illustrates the advantages of the new procedures.

References

1. Booth, G.W., Box, G.E.P., Muller, M.E. and Peterson, T.I. (1959), Forecasting by Generalized Regression Methods, Nonlinear Estimation (Princeton-IBM), International Business Machines Corp., Mimec. (IBM SHARE Program No. 687).

2. Box, G.E.P. (1954), The exploration and exploitation of response surface: some general considerations and examples, Biometrics, 10, 16-60.

3. Box, G.E.P. (1956), Some notes on nonlinear estimation, Technical Report, Statistical Techniques Research Group, Princeton University.

4. Box, G.E.P. (1958), Use of statistical methods in the elucidation of physical mechanism, Bulls. Inst. Intern De Statistique, 36, 215-225.

5. Box, G.E.P. (1960), Fitting empirical data, Proceedings N.Y. Academy of Sciences, 86, 792.

6. Box, G.E.P. and Coutie, G.A. (1956), Application of digital computers in the exploration of functional relationship, Proceedings of the Institute of Electrical Engineers, 103, Part B, Supplement No. 1, 100-107.

7. Box, G.E.P. and Hunter, W.G. (1965), A useful method for model building, Technometrics, 4, 301.

8. Box, G.E.P. and Hunter, W.G. (1965), Sequential design of experiments for nonlinear models, IBM Scienfitic Computing Symposium in Statistics, 113.

9. Box, G.E.P. and Wilson, K.B. (1951), On the experimental attainment of optimum conditions, JRSS, Series B, 13, 1.

10. Hartley, H.O. (1961), the modified Gauss-Newton method for fitting of nonlinear regression functions by least squares, Technometrics, 3, 269.

11. Levenberg, K. (1944), A method for the solution of certain, nonlinear problems in least squares, Quart. Appl. Math., 2, 1964.

12. Marquardt, D.W. (1963), An algorithm for least squares estimation of nonlinear parameters, SIAM J. Numer. Anal., 7, No. 1, 157.

13. Meeter, D.A. (1966), Nonlinear least squares (GAUSHAUS), University of Wisconsin Computing Center Users Manual, 4, Section 3, 22.

Appendix A2.1 Partial use of available information in the first and second order steepest descent methods.

It has been pointed out ([3]) that all the available information concerning the sum of squares surface will not be used when the first or second order steepest descent procedures are applied to the problem of least squares.

The first-order steepest descent procedure requires determination of at least (p+1) values of S to which a first degree polynomial

$$S = \alpha_o + \alpha_1\theta_1 + \alpha_2\theta_2 + \ldots + \alpha_p\theta_p \qquad \text{(A2.1.1)}$$

will be fitted and the direction of steepest descent will be given by $(-\alpha_1, -\alpha_2, \ldots, -\alpha_p)$. However this does not tell us how far one should move.

The second order procedure used near the stationery region determines the sum of squares at at least $\frac{1}{2}(p+1)(p+2)$ points. The fitted second degree polynomial

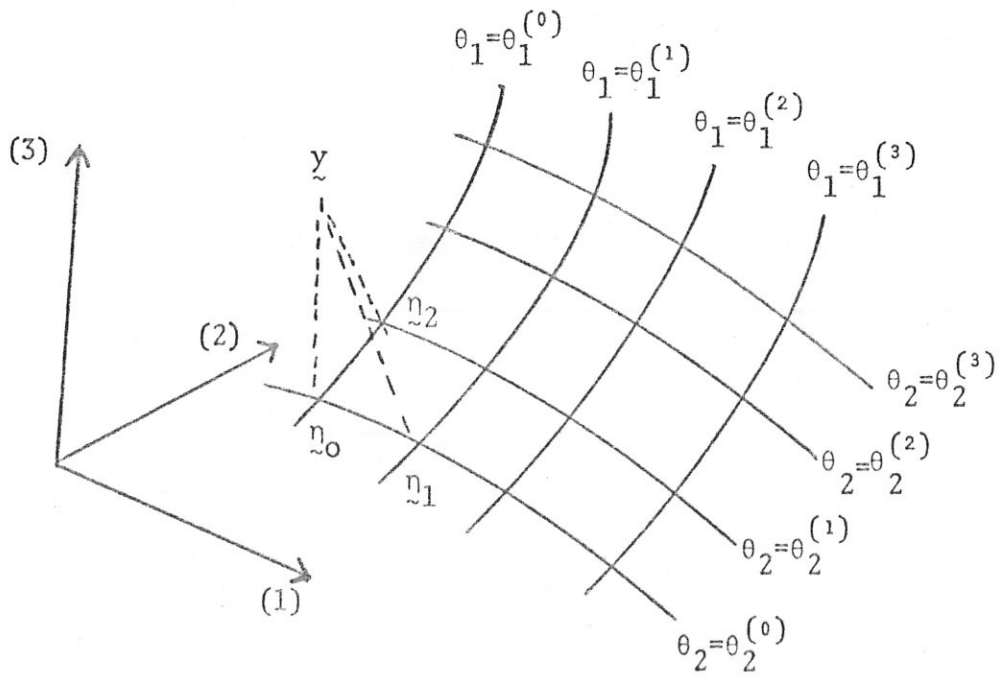
$$S = \alpha_o + \sum_{i=1}^{p} \alpha_i\theta_i + \sum_{i=1}^{p}\sum_{j=i}^{p} \alpha_{ij}\theta_i\theta_j \qquad \text{(A2.1.2)}$$

can then be used to locate approximately the parameter point yielding the minimum sum of squares.

In these procedures only the sums of squares are used. To examine whether there is any information that remains unused, we consider the problem in the sample space. Suppose, for example, we are fitting a functional relation

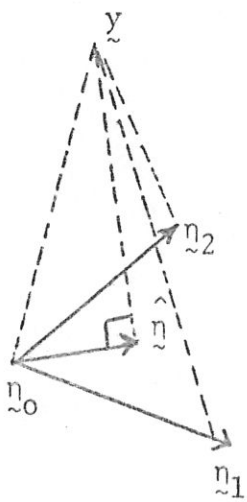
$$\eta_u = f(\xi_u;\ \theta_1,\theta_2) \qquad \text{(A2.1.3)}$$

where η_u is the expected value of the u th observation y_u. When there are n observations, the locus of the point $(\eta_1,\eta_2,\ldots,\eta_u)$ in the n dimensional space for all possible values of (θ_1,θ_2) will be defined by equation (A2.1.3) and generate a "parameter surface". For example, if n=3 the parameter surface in 3 dimensional sample space might look like that shown below.



Geometrically the least square estimates of parameters θ_1 and θ_2 are the values of θ_1 and θ_2 given by the point on the parameter surface that is closest to the point $\underset{\sim}{y}$.

Suppose that, in applying the steepest descent method, we have computed the sum of squares at three points $(\theta_1^{(0)},\theta_2^{(0)})$, $(\theta_1^{(1)},\theta_2^{(0)})$ and $(\theta_1^{(0)},\theta_2^{(1)})$ to determine the direction of steepest descent. These sums of squares are, in fact, the squared distances between the point $\underset{\sim}{y}$ and the points $\underset{\sim}{\eta}_k (k=0,1,2)$ on the locus. However, we also know how far apart the points $\underset{\sim}{\eta}$'s are from each other. Clearly if the locus can be taken to be locally planar, this knowledge is sufficient to determine the position of $\underset{\sim}{y}$ relative to the solution locus, and thus to locate the point $\hat{\underset{\sim}{\eta}}$ on the locus that is nearest to $\underset{\sim}{y}$. Such point can be located by projecting the vector $\underset{\sim}{y}-\underset{\sim}{\eta}$ onto the plane spanned by the vectors $\underset{\sim}{\eta}_1-\underset{\sim}{\eta}_o$ and $\underset{\sim}{\eta}_2-\underset{\sim}{\eta}_o$.



For the general p parameter case, this projection is a linear combination of $\underset{\sim}{\eta}_j - \underset{\sim}{\eta}_o$ (j=1,2,...,p)

$$\hat{\underset{\sim}{\eta}} - \underset{\sim}{\eta}_o = \hat{\phi}_1(\underset{\sim}{\eta}_1 - \underset{\sim}{\eta}_o) + \ldots + \hat{\phi}_p(\underset{\sim}{\eta}_p - \underset{\sim}{\eta}_o) \qquad \text{(A2.1.4)}$$

where $\hat{\phi}$'s are obtained from the "normal" equations

$$(\underset{\sim}{y} - \hat{\underset{\sim}{\eta}})'(\underset{\sim}{\eta}_j - \underset{\sim}{\eta}_o) = 0 \quad j=1,2,\ldots,p \qquad \text{(A2.1.5)}$$

or more specifically,

$$D'D\hat{\underset{\sim}{\phi}} = D'(\underset{\sim}{y} - \underset{\sim}{\eta}_o) \qquad \text{(A2.1.6)}$$

where $\hat{\underset{\sim}{\phi}}$ is the vector of $\hat{\phi}_j$; j=1,2,...,p and D is the n×p matrix whose j th column is $\underset{\sim}{\eta}_j - \underset{\sim}{\eta}_o$. The coordinate of the solution $\hat{\underset{\sim}{\theta}}$ will then be given by

$$\hat{\theta}_j = \theta_j^{(0)} + \hat{\phi}_j(\theta_j^{(1)} - \theta_j^{(0)}) \quad j=1,2,\ldots,p. \qquad \text{(A2.1.7)}$$

Thus, when all the available information is used, determination of the sums of squares at only (p+1) points is sufficient to locate the approximate minimum, while the sum of squares at $\frac{1}{2}(p+1)(p+2)$ points must be computed if we are using the sums of squares only.

Furthermore, let X be the n×p matrix whose j th column is $(\underset{\sim}{\eta}_j - \underset{\sim}{\eta}_o)/(\theta_j^{(1)} - \theta_j^{(0)})$. Then we have

$$D = XB \qquad \text{(A2.1.8)}$$

where B is a p×p diagonal matrix with the j th diagonal element $\theta_j^{(1)}-\theta_j^{(0)}$. Substituting (A2.1.8) into (A2.1.6) and noticing $B\underset{\sim}{\hat{\phi}} = \underset{\sim}{\hat{\theta}}-\underset{\sim}{\theta}^{(0)}$ from (A2.1.7), we obtain

$$X'X(\underset{\sim}{\hat{\theta}}-\underset{\sim}{\theta}^{(0)}) = X'(\underset{\sim}{y}-\underset{\sim}{\eta}_o) \qquad \text{(A2.1.9)}$$

It is then easy to see that the method described above is equivalent to the Gauss method assuming that the numerical estimates for derivatives are used

$$\left[\frac{\partial f(\underset{\sim}{\xi}_u,\underset{\sim}{\theta})}{\partial \theta_j}\right]_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(0)}} = \frac{\eta_{ju}-\eta_{ou}}{\theta_j^{(1)}-\theta_j^{(0)}} \quad . \qquad \text{(A2.1.10)}$$