MAXIMIZATION BY QUADRATIC HILL-CLIMBING

by

Stephen M. Goldfeld, Richard E. Quandt, Hale F. Trotter

35-48-65
19x400

Econometric Research Program
Research Memorandum No. 72
January 19, 1965

The research described in this paper was supported by National Science Foundation Grants NSF-GS 551 and NSF G 24462. The computer facilities used are supported by National Science Foundation Grant NSF-GP 579.

Princeton University
Econometric Research Program
92-A Nassau Street
Princeton, N. J.

ABSTRACT

The purpose of this paper is to describe a new gradient method for maximizing general functions. After a brief discussion of various known gradient methods the mathematical foundation is laid for the new algorithm which rests on maximizing a quadratic approximation to the function on a suitably chosen spherical region. The method requires no assumptions about the concavity of the function to be maximized and automatically modifies the step size in the light of the success of the quadratic approximation to the function. The paper further discusses some practical problems of implementing the algorithm and presents recent computational experience with it.

MAXIMIZATION BY QUADRATIC HILL-CLIMBING

1. Introduction

A variety of problems, when formulated mathematically, reduce to maximizing or minimizing functions of several variables. An important problem of this variety in econometrics is the computation of full-information maximum-likelihood estimates of the coefficients of a simultaneous-equations system which requires one to maximize the appropriate likelihood function (or more typically its logarithm).

All general computational techniques for maximization[1] take the form of an iterative procedure; given a point in n-dimensional space corresponding to a set of values for the independent variables, a new point at which the function is larger is computed. Repetition of this process leads to a sequence of points which, if the method is successful, converges more or less rapidly to the location of a maximum. Convergence proofs for these procedures generally require the assumption that the function to be maximized is strictly concave, at least in a region containing the sequence of computed points. Except in certain degenerate cases a function will be concave throughout some neighborhood of a

---

1. Since, given a function $H(x)$, $\max H(x) = \min(-H(x))$, all our remarks apply, mutatis mutandis, to minimization as well.

maximum and convergence can be guaranteed provided the initial value is sufficiently close to the maximum.[2]

In the absence of a priori knowledge of the behavior of the function, however, there is no way to ensure that the starting point will satisfy any such condition. The method proposed in this paper is specifically designed to work for functions which are not everywhere concave and for starting points which are not necessarily near a maximum.

Section 2 briefly discusses various gradient methods and Section 3 presents the mathematical background for our method. Section 4 discusses some more practical and computational aspects of the method while in Section 5 we present several examples of its performance. The Appendix summarizes some relevant standard results of matrix theory.

2. Alternative Gradient Methods

We consider a function $H(x_1,\ldots,x_n)$, denoted briefly by $H(x)$, of n variables which we wish to maximize. Let us denote by x the column vector of variables $(x_1,\ldots,x_n)$, by F_x the vector of first partial derivatives evaluated at x and by S_x the symmetric $(n \times n)$ matrix of second partial derivatives evaluated at x. There are many methods for maximizing $H(x)$. Usually one chooses a starting point $x^o = (x_1^o,\ldots,x_n^o)$ and iterates according to

$$x^{p+1} = x^p + h^p D^p \tag{2-1}$$

where h^p is a positive constant and D^p is an n-dimensional direction vector. In some methods the vectors D^p are chosen in some cyclic

2. See [8].

pattern.[3] In the so-called gradient methods the choice of D^p is given by

$$D^p = B^{-1}F_{x^p} \tag{2-2}$$

where B is a positive definite weighting matrix and F_{x^p} is the gradient F evaluated at x^p.

The Method of Steepest Ascent. A simple choice of B, suggested by Cauchy, is given by $B = I$ where I is the identity matrix. With this choice of B the procedure is known as the method of steepest ascent. The rationale behind this choice of B and hence of D^p rests on the result that the gradient points in the direction of the maximum increase of the best local linear approximation to $H(x)$.

Assume that $H(x)$ admits of a second-order Taylor series expansion around a point $a = (a_1, a_2, \ldots, a_n)$:

$$H(x) \approx H(a) + (x - a)'F_a + \frac{1}{2}(x - a)'S_a(x - a) \tag{2-3}$$

where the subscripts indicate the point of evaluation. Corresponding to (2-3) we have a first-order expansion for the first partials obtained by differentiating (2-3) with respect to x

$$F_x \approx F_a + S_a(x - a) \ . \tag{2-4}$$

The method of steepest ascent implies

---

3. In the simplest of these the vectors D^p are taken successively parallel to the coordinate axes, i.e., the variables are changed one at a time. See [9]. More sophisticated and efficient methods of this type are described in [6] and [7].

$$x^{p+1} = x^p + h^p F_{x^p} \tag{2-5}$$

and if we substitute (2-5) into (2-3), replacing a by x^p we obtain

$$H(x^{p+1}) - H(x^p) = h^p F'F + \frac{1}{2}(h^p)^2(F'SF) \tag{2-6}$$

where the subscripts have been omitted from F and S. One possible approach is to choose h^p so as to maximize (2-6). This, in fact, is a special case of what is known as the "optimum" gradient method.[4] Treating (2-6) as a function of h^p, say $G(h^p)$ we have

$$\frac{dG}{dh^p} = F'F + h^p(F'SF) = 0$$

or

$$h^p = -(F'SF)^{-1}F'F \ .$$

In order that this value of h^p yield a maximum we further require

$$\frac{d^2G}{d(h^p)^2} = F'SF < 0$$

which is necessarily so if S is negative definite. If x^p is not sufficiently close to the maximum to assure that S_{x^p} is negative definite this procedure may fail. In practice the "optimum" gradient choice of h^p has not worked well and alternatives have been used. If these alternatives do not involve S, the method of steepest ascent is computationally simple, but in general it involves other difficulties. In particular, (i) the sequence of points may converge to a saddle-point

---

4. This maximizes, one step at a time, the increase in the function given by the quadratic approximation as we move along the gradient. Other possibilities include maximization of the increase over the next r iterations. See [8].

rather than a true maximum and (ii) if the maximum lies on a narrow ridge, there is a tendency for successive steps to oscillate back and forth across the ridge, so that convergence to the maximum is very slow.

Newton's Method. The difficulties associated with steepest ascent methods lead to the second and most common version of the gradient method, known as Newton's method, which is obtained by maximizing (2-3) with respect to x. Setting $x^p = a$ and (2-4) equal to zero we obtain the iterative scheme

$$x^{p+1} = x^p - S^{-1}_{x^p} F_{x^p} . \qquad (2\text{-}7)$$

In other words this is a gradient method with $h^p \equiv 1$ and $B = -S^{-1}_{x^p}$. This may be regarded as a steepest ascent method with a different metric.[5]

If (2-3) is exact, i.e., if H is actually a quadratic polynomial, Newton's method yields the maximum in one step. If H is not quadratic but one has an approximation to the maximum which is sufficiently close to it, then (2-3) may be expected to be a very good approximation, and convergence is rapid.[6]

---

5. Distance from x to y in the Euclidean matrix is $[(x-y)'(x-y)]^{1/2}$ and if B is a positive definite symmetric matrix we can define a new distance measure by $[(x-y)'B(x-y)]^{1/2}$. The locus of points a distance k from x^o in the new metric is given by the hypersphere $(x-x^o)'B(x-x^o) = k^2$ with center x^o. This, of course, is an ellipsoid in the Euclidean metric and the direction of steepest ascent can be defined as the direction from x^o to the point on the ellipsoid where H is the greatest. In [2] it is shown that as $k \to 0$ this direction approaches $B^{-1}F_{x^o}$.

6. See [2].

In general, however, it may happen that (2-7) calls for taking a step so large that the quadratic approximation based on S_{x^p} and F_{x^p}, i.e., on the behavior of the function at x^p, has no validity at x^{p+1}. In addition, S_{x^p} may not be negative definite, in which case the quadratic approximation does not have a maximum. In either case the use of (2-7) is clearly inappropriate. While this has been observed before and attempts have been made to solve the problem of the non-negative-definiteness of S, no solution with a completely satisfactory rationale seems to have been proposed.[7]

We propose a new method which uses the same quadratic approximation, but includes a parameter which limits the size of the step taken. This parameter is altered according to the apparent accuracy of the quadratic approximation so that the step size is increased in regions where the approximation is good and cut down in regions where it is bad.

3. Restricted Maximization of a Quadratic Function

In this section we consider a quadratic function $Q(x)$. The matrix S is then constant and the expansions (2-3) and (2-4) are exact.

---

7. Chernoff and Divinsky [1] suggest a construction which yields a positive definite matrix for B and then suggest switching back to S^{-1} at some point. The decision when to switch is, however, arbitrary. Eisenpress [4] has also suggested a procedure for obtaining a positive definite B but offers little theoretical justification for it. Finally, Davidon [3] has suggested a procedure for modifying B at each iteration but assumes an arbitrary B to be given for the first iteration.

If S is non-singular it follows from (2-4) that

$$c = a - S^{-1}F_a \tag{3-1}$$

is the unique point where $F_x = 0$. If S is negative definite, Q has a unique global maximum at c; otherwise if S is non-singular Q is not bounded above. (If S is singular Q may have a whole linear subspace of maxima.)

Definition. $\| x \|$ denotes the length of the vector x which is defined to be $(x'x)^{1/2}$. Thus $\| x - y \|$ is the distance between x and y.

Lemma 1. Let α be any number such that $S - \alpha I$ is negative definite, and define

$$b_\alpha = a - (S - \alpha I)^{-1}F_a \tag{3-2}$$

$$r_\alpha = \| b_\alpha - a \| \ . \tag{3-3}$$

Then $Q(b_\alpha) \geqq Q(x)$ for all x such that $\| x - a \| = r_\alpha$.

Proof. Consider the quadratic function

$$\begin{aligned} R(x) &= Q(a) + (x - a)'F_a + \tfrac{1}{2}(x - a)'(S - \alpha I)(x - a) \\ &= Q(a) + (x - a)'F_a + \tfrac{1}{2}(x - a)'S(x - a) - \tfrac{1}{2}\alpha(x - a)'(x - a) \\ &= Q(x) - \tfrac{1}{2}\alpha \| x - a \|^2 . \end{aligned}$$

Since $S - \alpha I$ is negative definite, (3-1) with S replaced by $S - \alpha I$ applies to show that $R(x)$ has a global maximum at b_α. Thus for all x

$$\begin{aligned} Q(b_\alpha) - \tfrac{1}{2}\alpha \| b_\alpha - a \|^2 &= R(b_\alpha) \\ &\geqq R(x) = Q(x) - \tfrac{1}{2}\alpha \| x - a \|^2 \end{aligned}$$

and if $\|x - a\| = r_\alpha = \|b_\alpha - a\|$ then

$$Q(b_\alpha) \geqq Q(x)$$

as asserted.

Lemma 2. If $F_a \neq 0$ then the r_α defined by (3-2) and (3-3) is a strictly decreasing function of α on the interval (λ_1, ∞) where λ_1 is the maximum eigenvalue of S.[8]

Proof. From (3-2) and (3-3) we have

$$r_\alpha = \| (S - \alpha I)^{-1} F_a \| \tag{3-5}$$

and application of (A-4) in the Appendix with S replaced by $(S - \alpha I)^{-1}$ yields

$$r_\alpha^2 = \sum_{i=1}^{n} c_i^2 (\lambda_i - \alpha)^{-2} \tag{3-6}$$

where $c_1, \ldots, c_n$ are certain constants and $\lambda_1 \geqq \lambda_2 \geqq, \ldots, \geqq \lambda_n$ are the eigenvalues of S; this is a consequence of the easily verified fact that $(S - \alpha I)^{-1}$ has the same eigenvectors as S and has eigenvalues $(\lambda_i - \alpha)^{-1}$. For $\alpha > \lambda_1$, so that $\lambda_i < \alpha$ for all i, each of the coefficients $(\lambda_i - \alpha)^{-2}$ in (3-6) is a decreasing function of α and the stated conclusion follows.

---

8. Note that as stated in the last paragraph of the appendix, $S - \alpha I$ is negative definite precisely when α is in the interval (λ_1, ∞). It is easy to see that $r_\alpha \to 0$ as $\alpha \to \infty$. In general, $r_\alpha \to \infty$ as $\alpha \to \lambda_1$, but r_α tends to a finite limit if F_a has component 0 in the space of eigenvectors corresponding to λ_1.

Theorem. Let α, b_α and r_α be as in Lemma 1, let B_α be the region consisting of all x such that $\|x - a\| \leq r_\alpha$, and suppose $F_a \neq 0$. Then the maximum value of $Q(x)$ on B_α is attained at b_α if $\alpha \geq 0$, and is attained at b_o if $\alpha < 0$. (In this case b_o is interior to the region B_α.)

Proof. If S is not negative definite then (see Appendix) λ_1 and hence α are non-negative. Q can have no local maximum, and hence the maximum on a region such as B_α must occur on the boundary for some x with $\|x - a\| = r_\alpha$. By Lemma 1 the maximum is attained at b_α.

If S is negative definite then Q has an absolute maximum at b_o. Since $\lambda_1 < 0$ (see remark following (A-6) in the Appendix) both 0 and α are in the interval (λ_1, ∞) and by Lemma 2, $\|b_o - a\| \leq \|b_\alpha - a\|$ if and only if $\alpha < 0$. Thus if $\alpha < 0$, b_o is in the interior of B_α and the maximum of Q on B_α occurs at b_o. On the other hand, if $\alpha \geq 0$, b_o is not in the interior of B_α and there is no local maximum of Q in the interior. Hence just as in the case when S is not negative definite the maximum on B_α must occur at b_α.

If $F_a = 0$ the theorem does not apply (note that $b_\alpha = a$ and $r_\alpha = 0$ for all values of α in this case). The relevant statement is

Lemma 3. If $F_a = 0$ then the maximum value of Q on the region B_r consisting of all x with $\|x - a\| \leq r$ occurs at $a \pm ru_1$ if λ_1 (the maximum eigenvalue of S) is positive, and at a otherwise. (Here u_1 is a unit eigenvector associated with λ_1.)

Proof. Since in this case Q reduces to

$$Q(x) = Q(a) + \frac{1}{2}(x - a)'S(x - a)$$

the result follows almost immediately from (A-3).

It should be noted that if a metric $\| \ \|_A$ is defined by $\|x\|_A = (x'Ax)^{1/2}$ for a fixed positive definite matrix A, then all results of this section hold if $\| \ \|$ is replaced by $\| \ \|_A$ and $S - \alpha I$ is replaced by $S - \alpha A$ throughout. The regions B_α determined by such a metric would be ellipsoidal rather than spherical. We have made no attempt to exploit this generalization.

4. Implementation of the Algorithm

The iterative procedure we propose for finding the maximum of a general function is, given point x^p at which S_{x^p} and F_{x^p} are evaluated, to define the next point, x^{p+1}, as the maximum of the quadratic approximation (2-3) on a spherical region centered at x^p. Ideally, the region should be taken as large as possible provided that it is small enough that in the region the quadratic approximation is a satisfactory guide to the actual behavior of the function. The following procedure attempts to approximate this ideal.

Two distinct cases arise:

(a) F_{x^p} significantly different from 0.

In this event we choose a number

$$\alpha = \lambda_1 + R\| F_{x^p}\| \tag{4-1}$$

where λ_1 is, as before, the largest eigenvalue of S_{x^p}, and R is a

positive parameter determined by a rule to be described. We now take

$$x^{p+1} = x^p - (S_{x^p} - \alpha I)^{-1} F_{x^p} \tag{4-2}$$

or

$$x^{p+1} = x^p - S_{x^p}^{-1} F_{x^p}$$

according to whether α is positive or not. By the theorem of the previous section, x^{p+1} is the maximum of the quadratic approximation to the function on a region B_α of radius $\|(S_{x^p} - \alpha I)^{-1} F_{x^p}\|$ with center at x^p.

Lemma 2 shows that the larger the value of α, and hence the larger the value of R, the smaller the size of the region B_α. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of S_{x^p}, then the eigenvalues of $(S_{x^p} - \alpha I)^{-1}$ are $\mu_i = -1/(\lambda_1 + \|F_{x^p}\| R - \lambda_i)$, $i = 1, 2, \ldots, n$. The one with largest absolute value is μ_1, with $|\mu_1| = (\|F_{x^p}\| R)^{-1}$. Hence the radius of B_α

$$\|(S_{x^p} - \alpha I)^{-1} F\| \leq (\|F_{x^p}\| R)^{-1} \|F_{x^p}\| = R^{-1}$$

by (A-3). Equality holds only in exceptional cases, and it is possible for $\|(S_{x^p} - \alpha I)^{-1} F_{x^p}\|$ to be much smaller than R^{-1}, but it is reasonable to expect that the two quantities will in general be of the same order of magnitude.

In actual practice an initial value of R which appears reasonable is given to the algorithm and then R is automatically modified at each iteration; so that the step size tends to increase when the quadratic approximation appears to be satisfactory and tends to decrease when it

appears poor.[9] Given the value of α one computes a new iteration and accepts the step if the actual change in the function is positive. In the event the function deteriorates one can increase R so as to take a smaller step and repeat thus until an improvement is obtained.[10]

(b) F_{x^p} is so near 0 that the length of the step taken is within a preset tolerance of 0. Then, if S_{x^p} is negative definite, the process is terminated and x^p is accepted as the location of the maximum. If S_{x^p} is not negative definite, we are at a saddle point or at the bottom

---

9. More explicitly, we modify R as follows. Let ΔH be the actual change in the function due to the proposed Δx and let ΔQ be the corresponding change in the quadratic approximation. Let $z = \Delta H/\Delta Q$. If $z \leq 0$, the proposed Δx implies overshooting; it is therefore not accepted, R is increased by a factor of 4 and a new $(S - \alpha I)^{-1}$ is calculated. If $z > 0$ and close to unity (in practice, if z is between .7 and 1.3) R is decreased by multiplying R by a factor of .4. If $z > 2$, R is again increased by a factor of 4. For other values of z ($0 \leq z \leq .7$ and $1.3 \leq z \leq 2$) the magnitude of the factor multiplying R is determined by linear interpolation between .4 and 4.0. The form of the above rule was chosen as the simplest such rule with the correct qualitative behavior; the numerical values incorporated in the rule are those which seemed to be most successful in several experimental runs.

10. If $\alpha = 0$ at this point, we generally directly computed the step size necessary to produce a positive α. This typically saved a number of iterations.

of a cylindrical valley in which case Lemma 3 is applied. A step is taken along the eigenvector corresponding to λ_1 and the algorithm recycles in the usual manner.

One final feature, incorporated for reasons of computational efficiency rather than theoretical elegance, was the introduction of a scalar h^p into (4-1), writing it as

$$x^{p+1} = x^p - h^p(S_{x^p} - \alpha I)^{-1} F_{x^p} .$$

At each step the computation is first performed with $h^p = 1$. If this gives an improvement in $H(x)$, h^p is multiplied by a constant[11] and the function is examined at the new point so obtained. This process is repeated until the function declines in which event the last step is accepted. It should be noted that these attempts at stretching the step are relatively cheap since they require only an evaluation of the function. This is in contrast to changes in α within each iteration which require reinversion of $(S_{x^p} - \alpha I)$.

5. Some Computational Experience

Computational experience with the present algorithm is limited. The method appears to be reasonably successful in that on most examples it converges in a fairly small number of steps. It is clearly a relatively expensive method since each step involves calculating first and

---

11. The magnitude of the constant is a decreasing function of the absolute value of the angle between the current step and the immediately preceding step. Crockett and Chernoff, [2], suggest why even for the standard Newton method one might want h^p different from unity.

second derivatives, inverting a matrix, and finding its eigenvalues. An often suggested procedure for reducing computational effort is to modify S only after a number of iterations have occurred, in the belief that the consequent increase in the number of iterations will be offset by the reduced time per iteration. This may well be the case but we have not yet incorporated this feature in our computer programs. More generally, while we have tried to make the program efficient (in terms of total elapsed computer time) we have obviously not exploited all possible time-saving devices. Direct comparisons of the efficiency of various methods are very difficult, since the notion of an iteration may not be well-defined and variations in computer capabilities may render time comparisons meaningless.

In the remainder of this section we present computational experience based upon a limited number of functions, some of which have been reported in the literature to be appropriate test functions since they often cause difficulties.

The first of these is Rosenbrock's function given by[12]

$$z = 100(y - x^2)^2 + (1 - x)^2 \ .$$

This function has a minimum at (1,1) and is noted for resembling a U-shaped valley with very steep walls. The course of iterations from the starting point (-1.2, 1.0) suggested by Fletcher and Reeves is shown in Table 1.

---

12. See [6].

TABLE 1. Convergence for Rosenbrock's Function

Iteration	x	y	z
1	-1.2000	1.0000	24.2000
2	-1.1071	1.1850	4.6048
3	- .8762	.7143	3.8043
4	- .5820	.2964	2.6819
5	- .2867	.0275	1.9552
6	.2797	- .0095	1.2891
7	.3185	.0999	.4646
8	.4272	.1682	.3484
9	.6342	.3832	.1702
10	.7005	.4987	.0961
11	.8373	.6806	.0684
12	.8402	.7051	.0256
13	.9539	.8969	.0192
14	.9490	.9008	.0026
15	.9920	.9818	.0006
16	.9934	.9868	$.0004 \times 10^{-1}$
17	1.0000	.9999	$.0001 \times 10^{-3}$
18	1.0000	1.0000	$.0001 \times 10^{-9}$

A second function used for test purposes was

$$z = e^{-x^2-y^2}(2x^2 + 3y^2)$$

which has maxima at (1,0) and (-1,0), saddlepoints at (0,1) and (0,-1) and a minimum at (0,0). This function was maximized and the course of iterations from the two starting points (5,5) and (0,4) are shown in Tables 2 and 3.

TABLE 2. The Function $z = e^{-x^2-y^2}(2x^2 + 3y^2)$

Iteration	x	y	z
1	5.0000	5.0000	2.0000×10^{-20}
2	2.1833	2.1599	.0019
3	- .5395	- .6726	.8454
4	-1.0891	- .5290	.9506
5	- .9533	- .1924	1.0876
6	-1.0209	.0421	1.1020
7	-1.0002	.0030	1.1036
8	-1.0000	.0000	1.1036

TABLE 3. The Function $z = e^{-x^2-y^2}(2x^2 + 3y^2)$

Iteration	x	y	z
1	0.	4.0000	.0000
2	0.	2.0000	.1465
3	0.	1.0000	.7358
4	- .5000	1.0000	.7879
5	-1.0145	.5837	.9804
6	-1.0204	.0512	1.1017
7	-1.0004	.0037	1.1036
8	-1.0000	.0000	1.1036

It may be noted that in the second of these examples the algorithm takes us first straight to the saddlepoint (at iteration 3) from which we move counterclockwise, substantially along the rim of the crater.

As a final example we consider the function $z = e^{-\sum_{i=1}^{5} x_i^2}(3.0x_1^2 + 2.0x_2^2 + 3.5x_3^2 + 4.0x_4^2 + 2.7x_5^2)$. The course of convergence from the point (3.0,3.0,3.0,3.0,3.0) to (0,0,0,-1.0,0) is shown in Table 4.

TABLE 4. The Function $z = e^{-\sum_{i=1}^{5} x_i^2}(3.0x_1^2 + 2.0x_2^2 + 3.5x_3^2 + 4.0x_4^2 + 2.7x_5^2)$

Iteration	x_1	x_2	x_3	x_4	x_5	z
1	3.0000	3.0000	3.0000	3.0000	3.0000	$.3916 \times 10^{-17}$
2	1.2106	1.1971	1.2174	1.2241	1.2065	.0146
3	- .5297	- .6198	- .4836	- .4367	- .5574	.9940
4	- .4789	- .4704	- .4801	- .4786	- .4771	1.1116
5	.3131	- .0899	- .5219	- .8977	- .2273	1.3442
6	- .0142	.0197	- .1847	-1.0368	.0143	1.4572
7	.0019	- .0064	- .0208	- .9847	- .0031	1.4707
8	- .0002	.0002	.0059	-1.0002	.0002	1.4715
9	- .0000	- .0000	.0000	-1.0000	.0000	1.4715

Several other functions, including a function of ten variables, have been used for test purposes. With the exception of a four variable

function reported in [7] for which the matrix of second partial derivatives has rank equal to 2 at the minimum point, computational experience has been substantially similar to the above cases.

APPENDIX

This appendix contains a summary of various standard results of matrix theory in a form adapted to the requirements of this paper.

The principal fact used is that for any real symmetric $n \times n$ matrix S there exist n constants $\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_n$ called the <u>eigenvalues</u> of S and n mutually orthogonal unit column vectors $u_1,\ldots,u_n$ called the (normalized) eigenvectors of S such that

$$Su_i = \lambda_i u_i \qquad i = 1,2,\ldots,n \; . \tag{A-1}$$

Any $n \times 1$ vector x may be written in the form $x = \sum_{i=1}^{n} c_i u_i$ for a unique set of scalars $c_1,\ldots,c_n$ and then the quadratic function $x'Sx$ satisfies

$$x'Sx = \sum_{1}^{n} \lambda_i c_i^2 \tag{A-2}$$

For the squared norm of x one has

$$\| x \|^2 = x'x = \sum_{1}^{n} c_i^2$$

so that x is a unit vector if and only if $\sum_{1}^{n} c_i^2 = 1$. Under this restriction the maximum value of $\sum_{1}^{n} \lambda_i c_i^2$ is λ_1, achieved when $c_1 = \pm 1$ and all the other $c_i = 0$. Hence

$$\max_{\|x\|=1} x'Sx = \lambda_1 \tag{A-3}$$

and is achieved for $x = \pm u_1$.

The squared norm of Sx is given by

$$\|Sx\|^2 = x'S^2x = \sum_1^n \lambda_i^2 c_i^2 \qquad \text{(A-4)}$$

since S^2 has the same eigenvectors as S and eigenvalues $\lambda_1^2, \ldots, \lambda_n^2$. Thus the maximum value of $\|Sx\|^2$ for x a unit vector is $\max \lambda_i^2$. We then have, for all x ,

$$\|Sx\| \leq \|x\| \max |\lambda_i| \qquad \text{(A-5)}$$

Equality is attained if x is a multiple of the eigenvector corresponding to the eigenvalue of S of greatest absolute value.

S is said to be <u>negative definite</u> if

$$x'Sx < 0 \quad \text{for all} \quad x \neq 0 . \qquad \text{(A-6)}$$

In view of (A-2) this is equivalent to having $\lambda_1 < 0$ (and hence all $\lambda_i < 0$). Since the eigenvalues of $S - \alpha I$ are $\lambda_i - \alpha$, $S - \alpha I$ is negative definite if and only if $\lambda_1 - \alpha < 0$, i.e., $\alpha > \lambda_1$.

BIBLIOGRAPHY

[1] Chernoff, H. and N. Divinsky, "The Computation of Maximum-Likelihood Estimates of Linear Structural Equations," in Studies in Econometric Method, ed. W. C. Hood and T. C. Koopmans, New York, 1953, pp. 236-302.

[2] Crockett, Jean B. and Herman Chernoff, "Gradient Methods of Maximization," Pacific Jour. of Math., Vol. 5 (1955), pp. 33-59.

[3] Davidon, W. C., "Variable Metric Method for Minimization," Argonne National Lab., Report No. ANL-5990 Revised. (TID-4500, 14th ed.)

[4] Eisenpress, Harry, "Experiments in Convergence in Full-Information Estimation," mimeographed.

[5] Eisenpress, Harry, "Note on the Computation of Full-Information Maximum-Likehood Estimates of Coefficients of a Simultaneous System," Econometrica, Vol. 30, No. 2 (April 1962), pp. 343-349.

[6] Fletcher, R. and C. M. Reeves, "Function Minimization by Conjugate Gradients," The Computer Journal, Vol. 7, No. 2 (July 1964), pp. 149-153.

[7] Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," The Computer Journal, Vol. 7, No. 2 (July 1964), pp. 155-162.

[8] Saaty, Thomas L. and Joseph Bram, Nonlinear Mathematics, McGraw-Hill, New York, 1964.

[9] Spang, H. A., III, "A Review of Minimization Techniques for Non-Linear Functions," SIAM Review, Vol. 4, No. 4 (October 1962), pp. 343-365.