Chapter 12

COMPUTATIONAL PROBLEMS AND METHODS

RICHARD E. QUANDT*

Princeton University

Contents

*I am indebted to David A. Belsley, Angus Deaton, Ray C. Fair, Stephen M. Goldfeld, Jerry A. Hausman, and Mark Plant for constructive comments.

Handbook of Econometrics, Volume I, Edited by Z. Griliches and M.D. Intriligator

1. Introduction

The very substantial growth in econometric and statistical theory in the last 30 years has been at least matched by the explosive growth of computer technology and computational methods and algorithms. For the average researcher 30 years ago it was a problem of some moment to need the inverse of a matrix of relatively small size, say 5×5. Many procedures that are routinely applied today were not even attempted, even if they had been thought of.

The impressive advances of hardware, software, and algorithmic technology since that time have significantly advanced the state of econometrics; they have, however, not been an unmixed blessing. On the one hand, new problems have emerged which can trap the unwary. On the other hand, there has occurred an increase in the capital/output ratio in research. It is difficult to escape the conclusion that, as a consequence, the average researcher today spends a higher fraction of his time in data management, computer-program writing and adaptation, in interpretation of masses of computed output and a lesser fraction of his time in reasoning about the underlying problem than did his predecessor.

The purpose of this chapter is to highlight some of the most important computational methods and problems of today. The emphasis is on algorithms and general procedures for solving problems and not on detailed implementation in concrete computer programs or systems. Hence, names familiar to many such as TSP, ESP, GREMLIN, TROLL, AUTOREG, SHAZAAM, etc. will not be discussed. For some classical approaches to numerical analysis the reader is referred to Hildebrand (1956). For detailed computer implementation see Carnahan, Luther and Wilkes (1969).

Section 2 is devoted to certain matrix methods involved in estimating the parameters of single and simultaneous equation models. Sections 3–7 cover various aspects of numerical optimization. These methods become relevant whenever the first-order conditions for a maximum are not linear in the parameters to be estimated. Section 3 gives a survey of the typical functions that are optimized. Section 4 discusses the basic theory of optimization. Section 5 covers special purpose algorithms and simplifications useful in econometrics; Section 6 considers some further aspects of algorithms. Section 7 deals with very particular difficulties encountered only in problems of certain types. Section 8 is devoted to numerical integration and Section 9 to random number generation.

The list is obviously incomplete and problems that are treated are covered only in broad outlines. An extensive bibliography refers the interested reader to many extensions.

2. Matrix methods

As is well known, many commonly used estimators of the coefficients of econometric equations are the solutions to equations of the form

$$A\hat{\beta} = c, \tag{2.1}$$

where $\hat{\beta}$ represents the k-element coefficient vector estimated, A is a $k \times k$ matrix (usually non-singular), c a k-element vector, and where A and c depend only on the data. Some examples are discussed below. For the pertinent econometric theory see Schmidt (1976), Theil (1971), and for computational aspects see Belsley (1974).

(1) Ordinary Least Squares. If the model is

$$Y = X\beta + u, \tag{2.2}$$

where Y and u are $n \times 1$ and X is $n \times k$ (and usually of rank k), then $A = X'X$ and $c = X'Y$.

If linear restrictions are imposed on β by

$$R\beta = r,$$

where R is $p \times k$ and of rank p, then $A = X'X$ as before and $c = X'Y + R'(R(X'X)^{-1}R')^{-1}(r - R(X'X)^{-1}X'Y)$. If the ridge estimator [Schmidt (1976)] is required instead, $c = X'Y$ as before but $A = X'X + sI$, where s is a constant.

(2) k-Class. Consider a full system of simultaneous equations

$$Y\Gamma + XB = U,$$

where Y and U are $n \times g$, Γ is $g \times g$ and non-singular, X is $n \times k$, and B is $k \times g$. To discuss single equations estimators consider the first equation of the system written as

$$y = Y_1\gamma + X_1\beta + u_{\cdot 1} = Z_1\delta + u_{\cdot 1},$$

where $Z_1 = [Y_1 \quad X_1]$, $\delta' = (\gamma' \quad \beta')$, and $u_{\cdot 1}$ is the first column of U. Then the following k-class estimators for δ are immediate from $A\hat{\delta} = c$. Let A be given by

$$A = \begin{bmatrix} Y_1'Y_1 - k_1V_1'V_1 & Y_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix},$$

where $V_1 = Y_1 - X(X'X)^{-1}X'Y_1$ and c by

$$c = \begin{bmatrix} (Y_1' - k_2 V_1')y \\ X_1'y \end{bmatrix}.$$

If $k_1 = k_2 = 1$, two-stage least squares results. If $k_1 = k_2 =$ the smallest eigenvalue λ of

$$|Y_1^{0\prime}Y_1^0 - Y_1^{0\prime}X_1(X_1'X_1)^{-1}X_1'Y_1^0 - \lambda\left(Y_1^{0\prime}Y_1^0 - Y_1^{0\prime}X(X'X)^{-1}X'Y_1^0\right)| = 0,$$

where $Y_1^0 = [y \quad Y_1]$, we obtain limited information maximum likelihood estimates. If $k_1 = k_2 = 1 + (k - k^* - g - 1)/n$, where k^* is the number of columns in X_1, we obtain Nagar's $O(n^{-1})$ unbiased estimator. Other estimators are obtained by choosing k_1 and k_2 to be unequal.

If W is a $(g_1 + k^*) \times n$ matrix of instruments uncorrelated with $u_{\cdot 1}$, instrumental variables estimators (as are the above) are given in general by setting $A = W'Z$ and $c = W'y$, which also includes the indirect least squares estimator.

(3) Three-stage least squares. Write the full system as

$$y_i = Z_i\delta_i + u_i, \qquad i = 1, \ldots, g,$$

and define $y' = (y_1', \ldots, y_g')$, $Z = \operatorname{diag}(Z_i)$ a block-diagonal matrix with Z_i in the ith position, $\hat{\delta}_i$ as the two-stage least squares estimate of δ_i, and S the square matrix with (ij)th element $S_{ij} = (y_i - Z_i\hat{\delta}_i)'(y_j - z_j\hat{\delta}_j)/n$. Then if $A = Z'(S^{-1} \otimes X(X'X)^{-1}X')Z$ and $c = Z'(S^{-1} \otimes X(X'X)^{-1}X')y$, we have the three-stage least squares estimator.

2.1. *Methods for solving $A\hat{\beta} = c$*

The computation of each of the above estimators, as well as of many others, requires the inverse of A. Error in the inversion process accumulates as a result of rounding error in each computation. Rounding error, in turn, is due to the fact that the representation of numbers in a computer occupies a fixed number of places. In a binary computer floating point numbers are of the form $(\cdot a)(2^b)$, where a, the mantissa, and b, the characteristic, are binary integers stored in the computer and where the binary point "$\cdot$" and "2" are implied. The extent to which rounding error may affect the results is indicated by the condition number κ, which is the ratio of absolute value of the largest eigenvalue of A to the absolute value of the smallest [Golub (1969) and Jennings (1980)].[1] Various

[1]Since the matrix A is positive definite in all our examples, we may dispense with the absolute value in the definition of the condition number.

illustrative examples are provided by Golub (1969) and Wampler (1980). Consider as an example a case of OLS in which $A = X'X$ and assume

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ \varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 0 \\ 0 & 0 & 0 & \varepsilon \end{bmatrix}.$$

The eigenvalues of A are $4+\varepsilon^2$, ε^2, ε^2, and ε^2. If $\varepsilon < 2^{-t/2}$, where t is the number of binary digits in the mantissa of a floating point number, A will be a matrix with unity for each element and hence of rank 1 and not invertible. In general, the bound for the relative or proportionate error in the solution of an OLS problem is $\eta\kappa$, where η measures machine precision (e.g. 10^{-6}). Some principal matrix methods for controlling rounding error are discussed briefly below; for detailed application to econometric estimators see Belsley (1974), Golub (1969), and Wampler (1980). We illustrate the methods with reference to the ordinary regression model.

(1) Scaling. If the model is given by (2.2), it can also be written as $Y = Z\alpha + u$, where $Z = XB$ and B is a suitable diagonal matrix. The estimate for α is $\hat{\alpha} = (Z'Z)^{-1}Z'Y$ and $\hat{\beta} = B\hat{\alpha}$. Choosing b_{jj} as $[1/\sum_{i=1}^{n} x_{ij}^2]^{1/2}$ generally improves the conditioning of $Z'Z$.

(2) Cholesky factorization [Golub (1969), Klema (1973)]. If A is a positive definite matrix of order k, A may be factored so that

$$A = R'R, \tag{2.3}$$

where R is upper triangular. Error bounds for the factorization can be computed. Replacing A by $R'R$:

$$R'R\hat{\beta} = c,$$

and the solution proceeds in two steps: we first solve $R'\xi = c$ which is a triangular system and is solved easily; we next solve $R\hat{\beta} = \xi$ which is another triangular system. Cholesky factorizations for the $k \times k$ matrix A can be obtained in two ways [Golub (1969)]:

(a) Define

$$r_{11} = a_{11}^{1/2},$$
$$r_{1j} = a_{1j}/r_{11}, \qquad j = 2,\ldots,k,$$

and then let

$$r_{ii} = \left(a_{ii} - \sum_{p=1}^{i-1} r_{pi}^2 \right)^{1/2}, \qquad i = 2,\ldots,k,$$

$$r_{ij} = \left(a_{ij} - \sum_{p=1}^{i-1} r_{pi} r_{pj} \right) \Big/ r_{ii}, \qquad \begin{matrix} j = i+1,\ldots,k, \\ i = 2,\ldots,k. \end{matrix}$$

(b) Define $a_{ij}^1 = a_{ij}$ for all i, j. Then set

$$r_{pp} = \left(a_{pp}^p\right)^{1/2}, \qquad p = 1,\ldots,k,$$

$$r_{pj} = a_{pj}^p / r_{pp}, \qquad p = 1,\ldots,k, \quad j > k,$$

$$a_{ij}^{p+1} = a_{ij}^p - \frac{a_{pi}^p a_{pj}^p}{a_{pp}^p}, \qquad p = 1,\ldots,k, \quad i = p+1,\ldots,k, \quad j \geqslant i.$$

The decompositions are themselves subject to rounding error and there is no guarantee that (b) can be completed even if A is positive definite.

(3) The QR decomposition [Belsley (1974), Golub (1969) and Jennings (1980)]. For all $n \times k$ matrices X there exists an $n \times n$ orthogonal matrix Q and a $k \times k$ upper triangular matrix R_1 such that

$$QX = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = R.$$

Partitioning $Q' = [Q_1 \quad Q_2]$ it follows that

$$X = Q_1 R_1 = Q'R.$$

The solution of $A\hat{\beta} = c$ in the ordinary least squares case is then particularly easy since $R_1'Q_1'Q_1R_1\hat{\beta} = R_1'Q_1'Y$ or $R_1\hat{\beta} = Q_1'Y$, which is triangular system. The relative error of its solution is small if the regression residuals are small and is given by $\eta_1\kappa^{1/2} + \eta_2\kappa(Y - X\hat{\beta})'(Y - X\hat{\beta})/\hat{\beta}'\hat{\beta}$, where η_1 and η_2 are functions of machine precision and κ is the condition number of $X'X$ [Jennings (1980)]. Moreover, $X'X = R'R$ and R is a Cholesky factorization of $X'X$. Two alternative methods are often employed to obtain the QR decomposition.

(a) The Householder transformation. Let $P = I - 2vv'$, where v is a column vector and where $v'v = 1$. Then P is a Householder transformation. Define $X^{(1)} = X$ and let $X^{(p+1)} = P^{(p)}X^{(p)}$, where $P^{(p)} = I - 2v_pv_p'$, $v_p'v_p = 1$, and v_p is

chosen to make $X_{jp}^{(p+1)} = 0$ for $j = p+1,\ldots,n$. Then $R = X^{(k+1)}$ and $Q = P^{(k)}P^{(k-1)}\ldots P^{(1)}$. For an application of Householder transformations to estimating regression coefficients subject to linear restrictions see Dent (1980).

(b) Gram–Schmidt orthogonalization. Two such procedures are in use: the classical and the modified methods. The former can be found in numerous algebra texts [Hoffman and Kunze (1961)]. The latter is preferred from the computational point of view, although in the absence of rounding errors they produce identical answers [Golub (1969)]. For the modified method replace $Q'R$ by PS, where S has unity on the diagonal and $P'P$ = diagonal. Now define

$$X^{(p)} = \left(p_1,\ldots,p_{p-1}, x_p^{(p)},\ldots,x_k^{(p)}\right),$$

where p_i is the ith column of P and $x^{(p)}$ are columns defined below. Then at the pth step we let $p_p = x_p^{(p)}$ and set $d_p = p_p'p_p$, $s_{pr} = p_p'x_r^{(p)}/d_p$, and $x_r^{(p+1)} = x_r^{(p)} - s_{pr}p_p$ for $p+1 \leqq r \leqq k$.

Some recent experimental results (Wampler (1980)) indicate that the QR method with either the Householder transformation or the modified Gram–Schmidt orthogonalization gives more accurate results than the Cholesky factorization. For application see Belsley (1974), Dent (1977), and Jennings (1980).

2.2. *Singular value decomposition* [Belsley (1974) and Chambers (1977)]

Any $n \times k$ matrix X can be decomposed as

$$X = U\Sigma V', \tag{2.4}$$

where the columns of U and V are orthonormal eigenvectors of XX' and of $X'X$, respectively, and where Σ is diagonal and contains the square roots (positive) of the eigenvalues of $X'X$ and XX'. If X has rank $r < k$, then (2.4) can be written with U as $n \times r$, Σ as $r \times r$, and V' as $r \times k$.

The singular value decomposition can be employed to compute the pseudoinverse of any matrix X, defined as X^+ satisfying (a) $XX^+X = X$, (b) $X^+XX^+ = X^+$, (c) $(XX^+)' = XX^+$, and (d) $(X^+X)' = X^+X$. By substituting in (a) through (c) it can be shown that $X^+ = V\Sigma^+U'$, where Σ^+ is the same as Σ except that its diagonal elements are the reciprocals of the non-zero diagonal elements of Σ.

Consider a regression model $Y = X\beta + u$ and the normal equations $X'X\hat{\beta} = X'Y$. Assume a case of exact multicollinearity so that the rank r of X satisfies $r < k$. Replacing X by its singular value decomposition leads to

$$VV'\hat{\beta} = X^+Y. \tag{2.5}$$

Substitution of $\hat{\beta} = X^+ Y$ in the transformed normal equations (2.5) shows that they remain satisfied and that $X^+ Y$ is a least squares estimate. It can be shown further that $\hat{\beta}$ has shortest length in the set of all least squares estimates. The singular value decomposition thus permits the computation of the shortest least squares coefficient vector in the presence of multicollinearity. It can also be employed for the computation, via the pseudoinverse, of least squares estimates subject to linear restrictions on the coefficients [Gallant and Gerig (1980)]. For the calculation of the singular value decomposition see Golub (1969) and Bussinger and Golub (1969).

2.3. *Sparse matrix methods*

In some applications, such as optimal control problems or in seemingly unrelated regression models, there may occur matrices in which the non-zero elements are a small fraction of the total number of elements. Computational efficiency can be gained by not storing and manipulating the matrices in their full size but only their non-zero elements and identification as to the location of these. The resulting techniques are called sparse matrix techniques [see Drud (1977/78) and Belsley (1980)]. Their use can result in dramatic reductions in computer time. Fair (1976) reports that the time required to evaluate the Jacobian in full-information maximum likelihood (see Section 3) was reduced by a factor of 28 when sparse methods were employed.

3. Common functions requiring optimization

The computation of econometric estimates characteristically requires the maximization or minimization of some function. Some of these possess first-order conditions that are linear in the parameters to be estimated, and the matrix techniques discussed in Section 2 have wide applicability in these cases. In many other instances, however, the first-order conditions for an optimum cannot be solved in closed form. In these cases one must either solve the equations representing the first-order conditions by numerical methods or apply numerical methods to the direct optimization of the function in question. The present section briefly outlines some of the principal types of objective functions.

3.1. *Likelihood functions*

Specific assumptions about the distribution of error terms characteristically permit the derivation of the likelihood function. Maximum likelihood estimates are desired because of their favorable asymptotic properties.

One of the most common models requiring numerical maximization for the attainment of maximum likelihood estimates is the linear simultaneous equations model

$$Y\Gamma + XB = U, \tag{3.1}$$

where Y is an $n \times g$ matrix of endogenous variables, X an $n \times k$ matrix of predetermined variables, U an $n \times g$ matrix of error terms, Γ a $g \times g$ non-singular matrix, and B a $k \times g$ matrix of coefficients. If it is assumed that the rows of U are distributed identically and independently as $N(0, \Sigma)$, where Σ is a $g \times g$ positive definite matrix, the likelihood function is

$$L = (2\pi)^{-gn/2}|\Sigma|^{-n/2}(\text{abs}|\Gamma|)^n \exp\{-\tfrac{1}{2}\text{tr}[\Sigma^{-1}(Y\Gamma + XB)'(Y\Gamma + XB)]\}, \tag{3.2}$$

where $|\cdot|$ denotes taking the determinant and where $|\Gamma|$ is the Jacobian of the transformation $U \to Y$ [Schmidt (1976)]. The logarithm of the condensed likelihood function is

$$\log L = \text{constant} - \frac{n}{2}\log|S| + \frac{n}{2}\log[|\Gamma|]^2, \tag{3.3}$$

where S has elements $s_{jk} = \sum_{i=1}^{n} \hat{u}_{ij}\hat{u}_{ik}$ and where $\hat{u}_{ij}$ is the ith residual in the jth equation. If the system is non-linear and is given by

$$f_j(y_i, x_i, \beta) = u_{ij}, \qquad i = 1, \ldots, n; \quad j = 1, \ldots, g, \tag{3.4}$$

eq. (3.3) becomes

$$\log L = \text{constant} - \frac{n}{2}\log|S| + \tfrac{1}{2}\sum_{i=1}^{n}\log[|J_i|]^2, \tag{3.5}$$

where J_i is the Jacobian matrix corresponding to the ith observation with typical element $J_{ikl} = \partial u_{ik}/\partial y_{il}$. For a modification of (3.5) to perform robust estimation, see Fair (1974a). It should be noted that most linear simultaneous equations estimators that superficially might not be thought to be related to the maximization of (3.5) are in fact approximate solutions to the first-order conditions corresponding to (3.5) [Hendry (1976)].

Another very common example is provided by the ordinary regression model with error terms that obey a first-order Markov process $u_i = \rho u_{i-1} + \varepsilon_i$, $\varepsilon \sim N(0, \sigma^2 I)$. The log likelihood is

$$\log L = \text{constant} - \frac{n}{2}\log\sigma^2 + \tfrac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2}(Y - X\beta)'R'R(Y - X\beta), \tag{3.6}$$

where R is the matrix

$$R = \begin{bmatrix} (1-\rho^2)^{1/2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}.$$

Additional specific likelihood functions are discussed in subsequent sections as necessary.

3.2. *Generalized distance functions*

A number of estimates are obtained by minimizing a suitable distance function. A simple example is the non-linear least squares estimator of the parameters of

$$y_i = f(x_i, \beta) + u_i, \tag{3.7}$$

obtained by minimizing

$$D = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2.$$

More complicated examples arise in simultaneous equation estimation.

If eqs. (3.4) are in reduced form,

$$y_j = g_j(x_j, \beta_j) + v_j, \qquad j = 1, \dots, g, \tag{3.8}$$

where $y_j' = (y_{1j}, \dots, y_{nj})$ and where x_j and β_j are the predetermined variables and coefficients in the jth equation, a non-linear two-stage estimator is given by minimizing

$$D = \left[Y_j - g_j(x_j, \beta_j)\right]' X (X'X)^{-1} X' \left[Y_j - g_j(x_j, \beta_j)\right].$$

Stacking the equations in (3.7) as

$$y = g(X, \beta) + V,$$

where $y' = (y_1', \dots, y_g')$, we obtain an analogue of three-stage least squares by

minimizing.

$$D = [Y - g(X,\beta)]'\left[\hat{\Omega}^{-1} \otimes X(X'X)^{-1}X'\right][Y - g(X,\beta)],$$

where $\hat{\Omega}$ is a consistent estimate of Ω in $E(VV') = \Omega \otimes I$. [See Jorgenson and Laffont (1974), Berndt, Hall, Hall and Hausman (1974), and Parke (1979).]

3.3. *Functions in optimal control*

Consider a set of structural equations

$$f_j(y_i, x_i, z_i, \beta) = u_{ij}, \qquad j = 1,\ldots,g; \quad i = 1,\ldots,n, \tag{3.9}$$

where the y_i are vectors of g endogenous variables to be controlled, x_i are vectors of exogenous variables, and z_i are vectors of control variables. Then the optimal control problem is to minimize some loss function $W(y_1,\ldots,y_n; x_1,\ldots,x_n; z_1,\ldots, z_n)$ subject to eqs. (3.9). A frequent assumption is that the loss function is quadratic as in

$$W = \sum_{i=1}^{n} (y_i - a_i)'K_i(y_i - a_i),$$

where the vectors a_i and matrices K_i are given [Fair (1974b), Chow (1975), and Chow and Megdal (1978)].

4. Algorithms for optimizing functions of many variables

4.1. *Introduction*

The present section deals with the fundamental ideas of optimization algorithms. Refinements and special problems encountered in individual cases are discussed in Sections 5, 6, and 7. For the sake of convenience we adopt the convention that functions are to be maximized; hence the problem is to

$$\text{maximize } F(x_1,\ldots,x_n) \tag{4.1}$$

with respect to the elements of the vector $x = (x_1,\ldots,x_n)$.[2] Under normal

[2] Obvious alterations of the algorithms to be discussed turn them into methods for minimizing functions. Equivalently, one may maximize $-F(x)$.

circumstances $F(x)$ is taken to be twice continuously differentiable; however, under some circumstances this assumption may be violated (see Section 7). Most often maximization is unconstrained and the present section is exclusively restricted to this case. Some techniques for dealing with constraints are discussed in Section 7. Since $\partial F/\partial x = 0$ is a necessary condition for maximizing $F(x)$, optimization methods can be adapted in a natural way to solving systems of equations.

Numerical methods of optimization characteristically assume that an initial value x^0 is given for vector of variables.[3] Algorithms are iterative procedures or sequences of steps with the k th step defined by

$$x^{k+1} = x^k + \lambda^k d^k, \tag{4.2}$$

where d^k is a direction vector and λ^k a suitable constant. Algorithms differ in the way in which they select λ^k and d^k.

The classification of algorithms could be based on numerous criteria. We adopt a simple classification according to whether the algorithm requires the evaluation of no derivatives, or of first partial derivatives, or of first as well as second partial derivatives.

Algorithms have many characteristics of interest and the choice of an algorithm represents a trade-off among these. Clearly, no "best" algorithm exists and the mix of characteristics possessed by an algorithm will vary from problem to problem to a greater or lesser extent. Two fundamental characteristics of algorithms are of interest here: (a) their robustness, i.e. the degree to which they are capable of providing an estimate $\hat{x}$ of the true maximum x^* such that $\|\hat{x} - x^*\| < \varepsilon$ for some prespecified positive ε, and (b) their cost. This latter measure is not uniquely given by the specification of the algorithm but is dependent on the actual charging scheme in effect for the various resources of a computer such as execution time, core, I/O requests, etc. Cost is frequently and heuristically taken to be proportional to the number of iterations (a concept not well defined when comparing different algorithms) or the number of function evaluations. In any event, the speed with which an algorithm can be expected to converge is a relevant consideration. An algorithm is said to be quadratically convergent if it attains the maximum of a quadratic function in a finite number of steps. Various criteria exist for defining the speed of convergence. One of these may be stated in terms of $c = \lim_k \sup |x^k - x^*|^{1/k}$. Convergence is sublinear, linear, or superlinear

[3] The choice of x^0 may itself be a non-trivial task. Clearly, even approximate information about the shape of the function is valuable in that convergence to the maximum is likely to be the faster the closer x^0 is to the location of the maximum. It is often asserted that in estimation problems x^0 must be a consistent estimate. This may well be essential for statistical reasons as in the computation of linearized maximum likelihood estimates [Rothenberg and Leenders (1964)], but is not necessary for computational reasons.

when x^k converges to x^* according to whether the asymptotic rate of convergence satisfies $c=1$, $0<c<1$, or $c=0$. Sublinear convergence to zero is provided by $1/k$, linear by 2^{-k}, and superlinear by k^{-k} [Brent (1973)]. The notion of quadratic convergence is important, for in the neighborhood of the maximum the function F is approximately quadratic in the following sense. Let the Hessian matrix of $F(x)$ be $G(x)=[\partial^2 F(x)/\partial x_i \partial x_j]$ and let G satisfy the Lipschitz condition

$$|G(x^1)-G(x^2)| \leqq M\|x^1-x^2\|$$

for all x^1, x^2 in some domain R of F containing x^* in its interior, where $\|x^1-x^2\|$ is the Euclidean norm and M is a matrix of constants and where $|G(x^1)-G(x^2)|$ denotes a matrix the elements of which are the absolute values of $\partial^2 F(x^1)/\partial x_i \partial x_j - \partial^2 F(x^2)/\partial x_i \partial x_j$. Then

$$F(x)=F(x^*)+\tfrac{1}{2}(x-x^*)'G(x^*)(x-x^*)+Q(x) \tag{4.3}$$

for $x \in R$, where $|Q(x)| \leqq M\|x-x^*\|^3$. For x sufficiently near x^* the first two terms on the right-hand side of (4.3) provide a good approximation to $F(x)$.

4.2. *Methods employing no derivatives*

In principle, such methods are appealing because the computation of derivatives is almost always computationally costly. Nevertheless, relatively few algorithms of this type are in frequent use, particularly on problems of more than moderate size.

One class of derivative-free algorithms employs the notion of searching on a suitable grid of lattice points. A simple procedure is to start at some point x^0 and evaluate the function at x^0 and at the $2n$ lattice points $x^0 \pm he_i$, where e_i $(i=1,\ldots,n)$ is a vector with unity in the ith position and zeros elsewhere and where h is the preassigned lattice width. A step is taken from x^0 to x^1, where x^1 is the value of $x^0 \pm he_i$ for which $F(x^1)=\sup F(x^0 \pm he_i)$. The procedure is repeated starting from x^1 until no improvement is found for the given value of h. The value of h is then reduced and the search renewed. When h is finally reduced to the preassigned level of accuracy, the search is terminated and the last value of x taken as the location of the maximum. An algorithm in this class is that of Berman (1969).

Although the above algorithm is guaranteed to converge to a local maximum, in practice it is prohibitively expensive to employ. A different and more efficient version of search algorithms is that of Hooke and Jeeves (1961). The Hooke and Jeeves algorithm employs exploratory moves which are parallel to the coordinate

axes and pattern moves which represent the average direction of several past moves together. If an exploratory move and a subsequent pattern move together result in function improvement, they are both accepted; otherwise only an exploratory move is made. Computation again begins with a prespecified value of h and ends when h has been reduced to the desired accuracy.

Search methods do have advantages over methods using (first and second) derivatives. These are the assurance of eventual convergence and their independence of the concavity or convexity of the function $F(x)$. Nevertheless, in practice they are not employed frequently. They tend to converge slowly even in the immediate vicinity of the location of a maximum and, as a rule, are computationally very expensive. An even more serious problem is that algorithms that change only one variable at a time may fail to converge altogether. Consider the simple algorithm that changes at each iteration one variable according to

$$F(x^{k+1}) = \max_x F(x_1^k, \ldots, x_{k-1}^k, x, x_{k+1}^k, \ldots, x_n^k).$$

Methods of this type are in common use; see for example the Cochrane–Orcutt iterations used to maximize (3.6). These methods frequently work well if precautions are taken to terminate iterations when function improvement becomes small. Nevertheless, the gradient may remain strictly positive over the path taken by an algorithm and Powell (1973) has given examples in which this algorithm could cycle indefinitely around the edges of a hypercube.

An alternative direct search method is the Simplex method of Nelder and Mead (1965).[4] The function is first evaluated at the $n+1$ vertices $x^0, \ldots, x^n$ of an (irregular) simplex in the space R^n of variables. The corresponding function values, denoted by F_i $(i = 0, \ldots, n)$, are assumed to be ordered $F_n > F_{n-1} > \ldots, > F_0$. Among the points thus examined, x^n is currently the best, x^0 the worst. Compute the centroid c of the points not including the worst: $c = \sum_{j=1}^n x^j/n$. The steps of the algorithm are as follows:

(1) Reflect the simplex about the subsimplex given by $x^1, \ldots, x^n$ by choosing a point $x^r = c + \alpha(c - x^0)$ where $\alpha > 0$ is a coefficient chosen for the algorithm. If F_r, the function value corresponding to x^r, is such that $F_1 < F_r < F_n$, then x^r replaces x^0 and we return to Step 1.

(2) If $F_r > F_n$, then the simplex may profitably be stretched in the direction of x^r and an x^s is defined by $x^s = c + \beta(x^r - c)$, where $\beta > 1$ is a coefficient chosen for the algorithm. If $F_s > F_n$, x^s replaces x^0. Otherwise x^r replaces x^0. In either event we return to Step 1.

[4] Not to be confused with the simplex method of linear programming. See also Swann (1972) and, for variants and computational experience, Parkinson and Hutchinson (1972).

(3) If $F_r < F_1$, then the simplex should be contracted. A positive $\gamma < 1$ is chosen and x^c set to $c + \gamma(x^0 - c)$ if $F_r < F_0$ and to $c + \gamma(x^r - c)$ if $F_r > F_0$. If $F_c > \max(F_r, F_0)$, x^c replaces x^0 and we return to Step 1. Otherwise the points other than the best point x^n are shrunk toward x^n by a preselected proportion and we return to Step 1.

The algorithm is useful because it does not require derivatives. Unfortunately, its performance depends on the values of the various (expansion, contraction, reflection) coefficients and it is not easy to develop sound intuition as to desirable values.

An even more useful algorithm is the conjugate gradient method of Powell (1964). The basic motivation of this method is its behavior in the case of quadratic functions and its application to more general functions rests on analogy, or at least the heuristic observation, that near a maximum well-behaved functions are approximately quadratic.[5]

Two direction vectors, p and q, are said to be conjugate relative to a symmetric matrix A if $p'Aq = 0$. The essence of the algorithm is a sequence of n linear searches of the function in n linearly independent, mutually conjugate directions. Assume that n such directions, $d_1^k, \ldots, d_n^k$, are given at the beginning of the kth iteration and that the most recent estimate of the location of the maximum is x^k. The steps of an iteration are as follows.

(1) Calculate values v_r $(r = 1, \ldots, n)$ sequentially such that $F(x^k + \sum_{j=1}^{r} v_j d_j^k)$ is a maximum.

(2) Replace d_r^k by d_{r+1}^k $(r = 1, \ldots, n-1)$.

(3) Replace d_n^k by $\sum_{j=1}^{n} v_j d_j^k$.

(4) Calculate v such that $F(x^k + \sum_{j=1}^{n} v_j d_j^k + v(\sum_{j=1}^{n} v_j d_j^k))$ is a maximum and let x^{k+1} be given by $x^{k+1} = x^k + \sum_{j=1}^{n} v_j d_j^k + v(\sum_{j=1}^{n} v_j d_j^k)$.

The justification of the algorithm rests upon its convergence in the case of quadratic functions $F(x) = x'Ax + b'x + c$ and is established by the following theorems due to Powell (1964).

Theorem 4.1

Let $d_1, \ldots, d_m$, $m \leqq n$, be mutually conjugate directions in a subspace of dimension m and let x^0 be the starting point in that subspace. Then the maximum of the quadratic function $F(x)$ in the subspace is found by searching along each direction only once.

[5] For details beyond those provided here see also Goldfeld and Quandt (1972), Brent (1973), Murray (1972), and Fletcher (1965).

Proof

The location of the maximum can be written $x^0 + \sum_{i=1}^{m} \nu_i d_i$ and parameters ν_i are chosen so as to maximize $F(x^0 + \sum_{i=1}^{m} \nu_i d_i)$. Substituting $x^0 + \sum_{i=1}^{m} \nu_i d_i$ into the quadratic it can be seen that terms involving $d_i' A d_j$ vanish by the assumption of conjugacy. Hence, the maximum with respect to ν_i does not depend on the value of any $\nu_j, j \neq i$, proving the assertion.

Theorem 4.2

Let x^0 and x^1 be the locations of the maxima when the function is searched twice in the direction d from two starting points. Then the direction $x^1 - x^0$ is conjugate to d.

Proof

Any point x is the location of the maximum in the direction d if

$$\frac{\partial}{\partial \nu} F(x + \nu d) = 0 \quad \text{at } \nu = 0.$$

Performing the differentiation and alternately substituting x^1 and x^0 yields

$$\left(2(x^1)'A + b'\right)d = \left(2(x^0)'A + b'\right)d = 0$$

or

$$(x^1 - x^0)'Ad = 0.$$

The convergence of the algorithm can then be proved by induction. Assume that on the kth iteration of the algorithm the last k directions searched were mutually conjugate. The x^k which is the starting point of the next iteration and the $x^k + \sum \nu_j d_j^k$ defined in it represent maxima involving the same search directions, hence their difference is also conjugate to the previously conjugate directions by the parallel property stated in Theorem 4.2. Thus, after two iterations two conjugate directions exist, and after n iterations n such directions will exist, each of which will have been searched once. Q.E.D.

The conjugate gradient method is usually initiated by taking the columns of an identity matrix as the search directions. In practice it is often a useful method, although it has been conjectured that for problems in excess of 10–15 variables it may not perform as well. The principal reason for this may be [see Zangwill (1967)] that at some iteration the optimal value of ν_i in the linear search may be zero. The resulting set of directions $d_1, \ldots, d_n$ then become linearly dependent and henceforth the maximum can be found only over a proper subspace of the

original n-space. Near linear dependence and slow convergence can occur if ν_i is approximately zero. There are at least three devices for coping with this, with no clear evidence as to which is preferable.

(1) If the search directions become nearly linearly dependent, we may reset them to the columns of the identity matrix.

(2) We may skip Step 3 of the algorithm and search again over the same n directions used previously.

(3) We may replace the matrix of direction vectors with a suitably chosen orthogonal matrix [Brent (1973)]. These vectors are computed on the assumption that $F(\cdot)$ is quadratic and negative definite as follows.

Let A be the matrix of the (approximating) quadratic function. A is generally unknown (although it could be obtained at significant cost by evaluating the Hessian of F). Let D be the matrix of direction vectors. Then, since the directions are mutually conjugate with respect to A,

$$D'AD = M, \tag{4.4}$$

where M is diagonal with negative diagonal elements. The linear search in each of the n directions may be accomplished by evaluating $F(x^k + \sum_{i=1}^{j} \nu_i d_i^k)$ at three points ν_j^1, ν_j^2, and ν_j^3 $(j = 1,\ldots,n)$ and fitting a parabola to the function values (see Section 6). This involves computing the second differences of the function values which are easily shown to be

$$d_i' A d_i = \mu_i, \tag{4.5}$$

where d_i is a column of D and μ_i is a diagonal element of M. Define $R = D(-M)^{1/2}$ and $H = A^{-1}$. Then $H = DM^{-1}D' = -RR'$. Since D and M are known from the iteration, H can be computed. It remains to compute Q such that

$$Q'HQ = M^{-1} \tag{4.6}$$

and the columns of Q are orthogonal eigenvectors of A. If the quadratic approximation is good, the resulting search directions are conjugate to a matrix that is approximately the true Hessian and, hence, convergence can be expected to be fast. In order to avoid bad rounding errors in the computation of eigenvectors for a badly conditioned matrix it may be desirable to find the singular value decomposition $Q'R'S$ of the matrix R', where Q is the matrix of directions sought.

4.3. Methods employing first and second derivatives

A reasonable starting point for very general methods is to approximate $F(x)$ by a second-order Taylor approximation about x^0:

$$F(x) \simeq F(x^0) + g(x^0)'(x - x^0) + \tfrac{1}{2}(x - x^0)'G(x^0)(x - x^0), \tag{4.7}$$

where $g(x^0)$ denotes the gradient of $F(x)$ evaluated at x^0. Maximizing F by setting its partial derivatives equal to zero yields

$$g(x^0) + G(x^0)(x - x^0) = 0, \tag{4.8}$$

or, replacing x^0 by the current value of x at the kth iteration and replacing x by x^{k+1}, the new value sought is

$$x^{k+1} = x^k - \left[G(x^k)\right]^{-1} g(x^k), \tag{4.9}$$

which forms the basis of iterating according to Newton-type algorithms.[6] A very general class of algorithms is obtained by writing (4.9) as

$$x^{k+1} = x^k - \lambda^k H^k g(x^k), \tag{4.10}$$

where λ^k is a suitable constant and H^k is a matrix. Eq. (4.10) is of the same general form as (4.2) with $-H^k g(x^k)$ being the search direction. It can be shown that search direction d^k guarantees an improvement in the function if and only if it can be written as $-H^k g(x^k)$, with the matrix H^k being negative definite [Bard (1974)]. Numerous choices are available for λ^k as well as H^k; $\lambda^k = 1$ and $H^k = [G(x^k)]^{-1}$ yields Newton's method. It is a method with the best asymptotic rate of convergence $c = 0$.[7] It is, however, clearly expensive since it requires the evaluation of n first and $n(n+1)/2$ second derivatives. Moreover, (4.8) corresponds to a maximum only if the second-order conditions are satisfied, i.e. if $G(x^k)$ is a negative definite matrix. Obviously this may be expected to be the case if x^k is near the maximum; if not, and if $G(x^k)$ is not negative definite, iterating according to (4.9) will move the search in the "wrong" direction. A much simpler alternative is to set $H^k = -I$. The resulting method may be called the steepest ascent method. It locally always improves the value of the function but tends to

[6] Chow (1968, 1973) recommended this method for maximizing the likelihood for systems of simultaneous linear equations. Instead of directly maximizing the likelihood, he suggested the method for solving the first-order condition. It is also called the Newton–Raphson method. See also Hendry (1977) for various applications.

[7] See Parke (1979). Parke also discusses the asymptotic rates of convergence of the steepest ascent and univariate search methods. See also Dennis and Moré (1977).

behave badly near the optimum in that it tends to overshoot (indeed, for arbitrary fixed λ it is not guaranteed to converge) and near ridges in that it induces motion that is orthogonal to the contours of the function; these directions may well be nearly orthogonal to the desirable direction of search. Newton's method is useful precisely where the steepest ascent method is likely to fail. If $-G$ is positive definite, we have the decompositions

$$G = \sum_{i=1}^{n} \lambda_i P_i P_i', \qquad G^{-1} = \sum_{i=1}^{n} P_i P_i'/\lambda_i, \tag{4.11}$$

where the λ_i are the eigenvalues and the P_i the orthogonal eigenvectors of G. The eigenvectors point in the direction of the principal axes of the ellipsoid defined by $-y'Gy = \text{constant}$ and the quantities $(-1/\lambda_i)^{1/2}$ give their lengths. Since the eigenvectors are linearly independent, we can write $g = \sum_{i=1}^{n} \beta_i P_i$. Hence, the step defined by (4.9) can be expressed as

$$-\left[\sum_{i=1}^{n} P_i P_i'/\lambda_i\right] \sum_{j=1}^{n} \beta_j P_j = -\sum_{j=1}^{n} \beta_j P_j/\lambda_j.$$

If one of the λ's, say the kth, is very small, i.e. if the quadratic approximation defines ellipsoids that are highly elongated in the direction P_k, then the component P_k receives a weight proportional to $1/\lambda_k$ and the step will be nearly parallel to the ridge.

Several modifications exist for coping with the possibility that G might not be negative definite.

(1) Greenstadt (1967) replaces G by $-\sum_{i=1}^{n} |\lambda_i| P_i P_i'$.

(2) Marquardt (1963) suggests replacing G by $G - \alpha A$, where α is a small positive constant and A is a diagonal matrix with $a_{ii} = |G_{ii}|$ if $G_{ii} \neq 0$ and $a_{ii} = 1$ otherwise.

(3) In maximum likelihood problems, in which $\log L$ is to be maximized, it may be possible to compute the value of $[E(\partial^2 \log L/\partial\theta\partial\theta')]^{-1}$, where θ is the vector of variables with respect to which one wishes to maximize. Setting H^k equal to this matrix yields the method of scoring [Rao (1973), and Aitcheson and Silvey (1960)].

(4) In non-linear least squares problems [see eq. (3.7)] the objective function is $D = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2$. The second derivative matrix is

$$\frac{\partial^2 D}{\partial\beta\partial\beta'} = 2\sum_{i=1}^{n}\left[\frac{\partial f(x_i,\beta)}{\partial\beta}\frac{\partial f(x_i,\beta)}{\partial\beta'} - (y_i - f(x_i,\beta))\frac{\partial^2 f(x_i,\beta)}{\partial\beta\partial\beta'}\right]. \tag{4.12}$$

If H^k is set equal to the first term of (4.12), it is guaranteed to be positive definite and the resulting method is known as the Gauss or Gauss–Newton method [Goldfeld and Quandt (1972), and Bard (1974)].

A quadratic hill-climbing algorithm due to Goldfeld, Quandt and Trotter (1966) attacks the non-negative-definiteness of G directly and replaces G by $G - \alpha I$, where α is chosen so that $G - \alpha I$ is negative definite. In practice $\alpha = 0$ when G is negative definite and $\alpha > \lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of G, when G is not negative definite. The justification for the algorithm is based on the behavior of quadratic functions and is contained in the following theorems.[8] Let $Q(x)$ be an arbitrary quadratic function of x, let $Q'(x)$ denote the vector of first partial derivatives, and $Q''(x)$ the matrix of second partial derivatives. Define the iteration

$$x^{k+1} = x^k - \left(Q''(x^k) - \alpha I\right)^{-1} Q'(x^k) \tag{4.13}$$

and

$$r(\alpha) = \|x^{k+1} - x^k\|.$$

Then the following are true:

Theorem 4.3

For any α such that $Q''(x^k) - \alpha I$ is negative definite and any x such that $\|x - x^k\| = r(\alpha)$, $Q(x^{k+1}) \geqq Q(x^k)$.

Theorem 4.4

For all $Q'(x^k) \neq 0$, the radius $r(\alpha)$ is a strictly decreasing function of α for all $\alpha > \lambda_{\max}$.

Theorem 4.5

Let $R_\alpha = \{x \| \|x - x^k\| \leqq r(\alpha)\}$, and assume $Q'(x^k) \neq 0$. Then the maximum of $Q(x)$ on R_α is at the boundary point x^{k+1} if $\alpha \geqq 0$ and at the interior point $x^k - [Q''(x^k)]^{-1} Q'(x^k)$ otherwise.

The algorithm thus basically works as follows: at each step $G(x^k)$ is examined for negative definiteness. If G is negative definite, a step equal to $-[G(x^k)]^{-1} g(x^k)$ is taken.[9] Otherwise the step taken is $-[G(x^k) - \alpha I]^{-1} g(x^k)$, where α is taken to be $\lambda_{\max} + \rho \|g(x^k)\|$. The quantity ρ is itself adjusted from

[8] For proof see Goldfeld, Quandt and Trotter (1966).

[9] In practice, the direction $-[G(x^k)]^{-1} g(x^k)$ is computed and a one-dimensional line search is performed since line searches are computationally efficient ways of improving the function value.

iteration to iteration since the radius $r(\alpha) \leqq \rho^{-1}$. At each step the actual improvement in the function is compared with the improvement in the quadratic Taylor series approximation to it; if the comparison is unfavorable, ρ is increased and the radius is shrunk. It should be noted that the resulting changes of α not only change the step size (which may be overridden anyway by a subsequent line search) but also change the direction of movement. In any event, the direction will tend to be intermediate between that of a Newton step ($\alpha = 0$) and that of a steepest ascent step ($\alpha \to \infty$). It also follows that if α is very large, convergence is certain, albeit slow since $x^{k+1} \simeq x^k + g(x^k)/\alpha$. The comparison of the present method with Greenstadt's suggests that the latter may make a non-optimal correction in the step if F has "wrong" curvature in some direction. Assume, for example, that $\lambda_1 = \sup \lambda_i > 0$. Using (4.11), the step according to the quadratic hill-climbing method is given by

$$x^{k+1} = x^k - \sum_i \left[\frac{P_i' g}{\lambda_i - \alpha} \right] P_i,$$

and according to Greenstadt's suggestion by

$$x^{k+1} = x^k - \sum_i \left[\frac{P_i' g}{|\lambda_i|} \right] P_i. \tag{4.14}$$

Since α is chosen so that $\lambda_1 - \alpha < 0$, we have $|1/(\lambda_1 - \alpha)| > |1/(\lambda_j - \alpha)|$ for all $j = 2, \ldots, n$ and hence the step will contain the direction P_1 with relatively largest weight, a fact that need not hold for (4.14). Thus, the step will be relatively closer to the direction in which the function is convex [Powell (1971)].

A further refinement of the quadratic hill-climbing algorithm rests on the observation that recently successful directions of search may well be worth further searches. Thus, if the step from x^{k-1} to x^k is given by $x^k - x^{k-1} = \xi^k$, then the decomposition of any vector into its projection on ξ and its orthogonal complement permits the component parallel to ξ to be emphasized. To distinguish an actual x^k from arbitrary members of the coordinate system prevailing at the jth iteration, we use the notation $x(j)$. Thus, the coordinate system prevailing at the jth iteration may be transformed into a system prevailing at the $(j+1)$th by

$$x(j+1) = B_j x(j),$$

where $B_j = I + (1-\beta)M_j$ and where $0 < \beta < 1$ and $M_j = \xi_j(\xi_j'\xi_j)^{-1}\xi_j'$. A sequence of such transformations allows the original coordinate system and the one prevailing at the jth iteration to be related by $x(j) = Bx(0)$. Applying the hill-climbing algorithm thus alters the original procedure from maximizing at

each step on a sphere to maximizing on a suitably oriented ellipsoid, since $x(j)'x(j) = x(0)'B'Bx(0)$. Writing the function in the jth coordinate system as $F(x(j))$ and differentiating, $\partial F(x(j))/\partial x(0) = B'\partial F(x(j))/\partial x(j)$. Hence, the gradient of $F(x(j))$ in terms of the original system is $g(x(j)) = (B^{-1})'\partial F(x(j))/\partial x(0)$. By similar reasoning the Hessian is $(B^{-1})'G(x(j))B^{-1}$. It follows that the step taken can be expressed in the $x(j)$-coordinate system as $-[(B^{-1})'G(x(j))B^{-1} - \alpha I]^{-1}(B^{-1})'g(x(j))$. Premultiplying by B^{-1} yields the step in the $x(0)$-coordinate system and is $-(G(x(j)) - \alpha B'B)^{-1}g(x(j))$ and is equivalent to replacing I in (4.13) by a positive definite matrix $B'B$.[10]

4.4. *Methods employing first derivatives*

A general theory of quadratically convergent algorithms has been given by Huang (1970).[11] The objective of Huang's theory is to derive a class of algorithms with the following properties: (a) searches at each iteration are one-dimensional; (b) the algorithms are quadratically convergent; (c) they calculate only function values and first derivatives; and (d) at the kth iteration they only employ information computed at the kth and $(k-1)$th iterations.

Requirement (a) states that at each iteration $k = 1, 2, \ldots$ a direction d^k be chosen and a scalar λ_k be determined such that

$$\partial F(x^k + \lambda_k d^k)/\partial \lambda_k = 0. \tag{4.15}$$

This determines a displacement $\Delta x^k = \lambda_k d^k$ or $x^{k+1} = x^k + \lambda_k d^k$. Restricting attention [by Property (b)] to quadratic functions $F(x) = x'Ax + b'x + c$, it follows that

$$F(x^{k+1}) = F(x^k) + g_k'\Delta x^k + \Delta x^{k\prime} A \Delta x^k, \tag{4.16}$$

where g_k denotes the gradient $g(x^k)$ at x^k. From the first order condition (4.15) it follows that $g_{k+1}'d^k = 0$. Since $g_{k+1} = g_k + 2A\Delta x^k$ and $\Delta x^k = \lambda_k d^k$, the optimal λ_k is given by

$$\lambda_k = -\frac{d^{k\prime} g_k}{2 d^{k\prime} A d^k}.$$

[10]A final modification pertains to the case when $g = 0$ to the required accuracy without having achieved a negative definite $G(x^k)$. This is the case in which x^k is a saddlepoint. Although such cases may be only rarely encountered, the following proposition ensures that the algorithm does not terminate prematurely. If $g(x^k) = 0$ and if $\lambda_{\max} > 0$, the maximum of F within the sphere $\|x - x^k\| \leqq r$ occurs at $x^k \pm rP_{\max}$, where $P_{\max}$ is the eigenvector corresponding to the $\lambda_{\max}$. Thus, in such instances a search direction is provided by the appropriate eigenvector.

[11]See also detailed discussion in Powell (1971), Broyden (1972), and Dennis and Moré (1977).

Substituting for Δx^k and λ_k in (4.16) yields

$$F(x^{k+1}) - F(x^k) = -\frac{(d^{k\prime} g_k)^2}{2 d^{k\prime} A d^k},$$

which is positive if A is negative definite, thus ensuring that the function is monotone increasing over successive iterations. If it is further required that the successive search directions be conjugate with respect to A, quadratic convergence can be proved in straightforward fashion. Taking the search direction d^k to be a matrix multiple of the gradient

$$d^k = H^k g_k$$

produces an algorithm of the general form of (4.10) and for that reason these algorithms may be called quasi-Newton algorithms. The conjugacy condition places restrictions on the matrix H^k; however, the restrictions will be observed if H^k is updated according to

$$H^{k+1} = H^k + \theta_1 \Delta x^k \Delta x^{k\prime} + \theta_2 H^k \Delta g_k \Delta g_k' H^k + \theta_3 \left[\Delta x^k \Delta g_k' H^k + H^k \Delta g_k \Delta x^{k\prime}\right], \tag{4.17}$$

where $\Delta g_k = g_{k+1} - g_k$. Different choices for θ_1, θ_2, and θ_3 yield different members of this class of algorithms. In any event, θ_2 and θ_3 must satisfy

$$1 + \theta_2 \Delta g_k' H^k \Delta g_k + \theta_3 \Delta x^{k\prime} \Delta g_k = 0. \tag{4.18}$$

At the start, H^1 is usually initialized to $-I$ (or I for minimization). Some of the alternatives are as follows.

(1) If $\theta_1 = 1/\Delta x^{k\prime} \Delta g_k$, $\theta_2 = -1/\Delta g_k' H^k \Delta g_k$, and $\theta_3 = 0$, the resulting algorithm is known as the Davidon–Fletcher–Powell (DFP) algorithm [Davidon (1959), and Fletcher and Powell (1963)]. In this case $F(x)$ is not required to be quadratic for convergence but very strict concavity conditions are required. If $F(x)$ is not concave, there is no assurance that convergence will take place. It should be noted that the quantity $g_k' H^k g_k$ increases monotonically over the iterations; since $g_k' H^k g_k$ is negative for concave functions, this implies that the search direction tends to become more nearly orthogonal to the gradient which can interfere with speedy convergence.

An important feature of DFP is contained in the following:

Theorem 4.6

If $F(x)$ is quadratic, then $H^n = G^{-1}$. The convergence of H to the inverse Hessian in the quadratic case is used in practice to obtain estimates of asymptotic variances and covariances in the presence of maximum likelihood estimation. However, care must exercised for if apparent convergence occurs in less than n iterations, H will not contain usable quantities. It is important, therefore, that computer implementations of DFP contain a restart facility by which computations can be (re)initiated not only with $H^1 = -I$, but $H^1 =$ a previously computed H^k.

(2) In order to make H^k approximate the inverse of the true Hessian $G(x^k)$, one may set

$$\theta_1 \Delta x^{k\prime} \Delta g_k + \theta_3 \Delta g_k' H^k \Delta g_k = 1 \tag{4.19}$$

and also require (4.18) to hold. Obviously, the DFP algorithm satisfies (4.19). In that case the required approximation holds because (4.18) and (4.19) imply $H^{k+1}\Delta g_k = \Delta x^k$, which is just the Taylor series approximation $g_{k+1} = g_k + G(x^k)\Delta x^k$. A special case is the rank-one correction formula, according to which

$$H^{k+1} = H^k + \frac{\left[\Delta x^k - H^k \Delta g_k\right]\left[\Delta x^k - H^k \Delta g^k\right]'}{\Delta x^{k\prime} \Delta g_k - \Delta g_k' H^k \Delta g_k}.$$

Several other algorithms are defined by Huang and Powell and the reader is referred to Huang (1970) and Powell (1971) for details.[12]

Computational experience with many members of this class of algorithms is lacking. The best documented member of the class is almost certainly DFP. Overall its performance is good enough to make it a reasonable first choice for many problems, although it is both generally less robust than some variants of Newton's method (particularly near the maximum) and less efficient computationally than algorithms especially tailored to a problem [see Parke's (1979) Algorithm A] or possessing special *ad hoc* features such as MINOPT [Belsley (1980)]. The latter in particular makes a choice at each iteration whether to employ a steepest ascent step or a Newton step, allowing the former to guide computations at an initial point far from the optimum and the latter near it. It can thus perform more efficiently than DFP. In general, however, DFP and other members of the class must continue to be regarded as viable alternatives for many

[12]A recent new member of the class is due to Davidon (1975) and is designed to work without any line searches.

problems. Two related reasons why members of the quasi-Newton class may fail or perform poorly in practice are: (a) H^{k+1} may become (nearly) singular and (b) H^{k+1} may fail to provide a good approximation to G^{-1}. The latter affects the speed of convergence as a result of the following:

Theorem 4.7

If $F(x)$ is a negative definite quadratic function and x^* the location of the maximum, then $F(x^*)-F(x^{k+1})) \leqq [(K(R_k)-1)/(K(R_k)+1)](F(x^*)-F(x^k))$, where $K(R_k)$ is the condition number of the matrix $R_k = G^{1/2}H^kG^{1/2}$.

Hence, $K(R_k)$ should be small and decreasing which will be the case if H^k increasingly approximates G^{-1}. Oren and Luenberger (1974) designed a "self-scaling" algorithm in this class which guarantees that $K(R_{k+1}) \leqq K(R_k)$, with updating formulae given by

$$v^k = \left(\Delta g_k' H^k \Delta g_k\right)^{1/2}\left[\frac{\Delta x^k}{\Delta g_k' \Delta x_k} - \frac{H^k \Delta g_k}{\Delta g_k' H^k \Delta g_k}\right],$$

$$H^{k+1} = \left[H^k - \frac{H^k \Delta g_k \Delta g_k' H^k}{\Delta g_k' H^k \Delta g_k} + \theta_{4k} v^k v^{k\prime}\right]\theta_{5k} + \frac{\Delta x^k \Delta x^{k\prime}}{\Delta x^{k\prime} \Delta g_k},$$

$$\theta_{5k} = \frac{\Delta x^{k\prime} \Delta g_k}{\Delta g_k' H^k \Delta g_k'}(1-\theta_{6k}) + \frac{g_k' \Delta x^k}{g_k' H^k \Delta g_k}\theta_{6k},$$

where θ_{4k} and θ_{6k} are parameters to be chosen. In recent experiments various self-scaling and other quasi-Newton algorithms gave satisfactory results [Van der Hoek and Dikshoorn (1979)].

5. Special purpose algorithms and simplifications

There is no hard-and-fast dividing line between general and special purpose algorithms. In the present section we discuss some algorithms that are either especially suited for problems with a particular structure or contain more or less *ad hoc* procedures that appear to be useful in particular contexts.

5.1. Jacobi and Gauss–Seidel methods

Both of these procedures are designed to solve systems of (linear or non-linear) equations. In the context of maximizing a likelihood function, they are applied to solving the first-order conditions, the likelihood equations. Both Jacobi's method

and the Gauss–Seidel method presuppose that the equation system can be solved in a particular manner. In the former case we require a solution

$$x = f(x), \tag{5.1}$$

where x is a vector of unknowns and $f(x)$ a vector-valued function. Jacobi's method iterates according to

$$x^{k+1} = f(x^k).$$

The Gauss–Seidel method is similar, except that the ith equation in (5.1), $i = 1,\ldots,n$, is assumed to have the structure $x_i = f_i(x_1,\ldots,x_{i-1}, x_{i+1},\ldots,x_n)$. A further distinction may be obtained depending on whether in a given iteration of the algorithm all n x's are computed and then used only in the next iteration, or whether an x_i computed in a given iteration is used immediately in the computation of other x's in that same iteration. There is some reason to think that the latter procedure is more efficient [Fromm and Klein (1969)].

Jacobi's method was applied to Klein's Model I by Chow (1968). As shown in Section 3, the condensed log-likelihood function for a system of simultaneous linear equations $Y\Gamma + XB = U$ can be written as

$$L = \text{constant} - \frac{n}{2}\log|S| + \frac{n}{2}\log\left[|\Gamma|^2\right], \tag{5.2}$$

where Γ is the matrix of coefficients associated with the jointly dependent variables and S is the estimated covariance matrix of residuals with typical element

$$S_{ij} = \frac{1}{n}\sum_{k=1}^{n} u_{ik}u_{jk}.$$

S itself is a function of the parameters in Γ and B and setting derivatives of L with respect to the non-zero elements of Γ and B equal to zero yields equations of the form of (5.1). Jacobi's method or the Gauss–Seidel method are also routinely applied to solving non-linear systems of simultaneous equations as is required for the solution of stochastic control problems [Chow and Megdal (1978)] or for simulating non-linear econometric models after estimation [Duesenberry et al. (1969), and Fair (1976)].

The objectives of simulation may be to assess the sources of uncertainty and the quality of the predictions over several models or to estimate the effects and the uncertainty of various policy variables [Fair (1980a, 1980b)]. Simulations are stochastic if repeated trials are made in which either the error terms, or the

coefficients employed in computing predictions, or exogenous variable values, or all of these are drawn from some appropriate distribution. Whatever simulation variant is chosen, the simulated endogenous variable values must be obtained by solving the system of econometric equations, which is typically non-linear.

A particularly interesting application is due to Fair (1979) in which models with rational expectations in bond and stock markets are simulated. In these models two layers of Gauss–Seidel alternate: for certain initial values of some variables, Gauss–Seidel is used to solve the system for the remaining ones. These solution values are used to obtain new values for the initial set of variables and the system is solved again for the remaining variables, etc.

Neither Jacobi's nor the Gauss–Seidel method can be expected to converge in general. A sufficient condition for convergence is that $f(x)$ be continuous and a contraction mapping; that is, given the distance function d over a compact region R, $f(x)$ is a contraction mapping if for $x \neq x^*$, $x, x^* \in R$, and $d(f(x), f(x^*)) < d(x, x^*)$. An example of such a contraction mapping is provided by Ito (1980) in connection with solving a two-market disequilibrium model with spillovers for values of the endogenous variables. The equations of such models are

$$
\begin{aligned}
y^{\mathrm{d}} &= \alpha_{01}' x_1 + \alpha_1 (l - \tilde{l}^{\mathrm{s}}) + \varepsilon_1, \\
y^{\mathrm{s}} &= \alpha_{02}' x_2 + \alpha_2 (l - \tilde{l}^{\mathrm{d}}) + \varepsilon_2, \\
y &= \min(y^{\mathrm{d}}, y^{\mathrm{s}}), \\
l^{\mathrm{d}} &= \beta_{01}' z_1 + \beta_1 (y - \tilde{y}^{\mathrm{s}}) + \varepsilon_3, \\
l^{\mathrm{s}} &= \beta_{01}' z_2 + \beta_2 (y - \tilde{y}^{\mathrm{d}}) + \varepsilon_4, \\
l &= \min(l^{\mathrm{d}}, l^{\mathrm{s}}),
\end{aligned}
$$

where $\tilde{y}^{\mathrm{d}} = \alpha_{01}' x_1 + \varepsilon_1$, $\tilde{y}^{\mathrm{s}} = \alpha_{02}' x_2 + \varepsilon_2$, $\tilde{l}^{\mathrm{d}} = \beta_{01}' z_1 + \varepsilon_3$, and $\tilde{l}^{\mathrm{s}} = \beta_{02}' z_2 + \varepsilon_4$. The x's and z's are exogenous variables and the ε's random errors. The y^{d}, y^{s}, l^{d}, and l^{s} are effective goods and labor demand and supply, respectively, the same symbols with a tilde (˜) represent notional demands and supplies, and y and l represent the actually transacted quantities. Ito shows that values of y and l may be calculated by Jacobi's method (for given values of x's, z's, and ε's) by starting with arbitrary y and l if $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ and if $1 - \alpha_i \beta_j > 0$ for all $i = 1, 2$ and $j = 1, 2$.

Some algorithms that appear to be principally gradient methods exploit the idea of a Jacobi or a Gauss–Seidel iteration. Thus, for estimating the parameters of systems of simultaneous equations, Dagenais (1978) first computes a Jacobi iteration of the parameter vector and then further displaces the parameter vector in the direction of the difference between the original value and the Jacobi iterate.

This yields an iteration of the form $x^{k+1} = x^k + \lambda H^k g_k$, where H^k is a positive definite matrix and λ a scalar.

5.2. *Parke's Algorithm A*

An algorithm particularly suited for estimating the coefficients of linear or non-linear simultaneous equations by full-information maximum likelihood or by three-stage least squares is Parke's (1979) Algorithm A. Algorithms that are especially useful for simultaneous equation estimation have been used before. A case in point is the procedure implemented by Chapman and Fair (1972) for systems with autocorrelations of the residuals: their algorithm is a sequence of pairs of Newton steps in which the first operates only on the coefficients of the equations and the second on the autocorrelation coefficients.

Algorithm A performs sequences of searches at each iteration in order to exploit two empirical generalizations about the structure of simultaneous equations models: (a) that the coefficients in any one equation are more closely related than those in separate equations, and (b) that change in the values of the residuals of the equations usually has a substantial effect on the objective function. The algorithm uses searches, no derivatives, and performs numerous searches at each iteration; these facts make it superficially resemble the Powell (1964) class of algorithms.

The sequence of searches in an iteration may be briefly summarized as follows.

(a) For each equation in turn the coefficients of the equation are perturbed one by one (and in a particular order) with the constant term being continually readjusted so as to stabilize the residuals in the sense of holding the mean residual constant. Finally, the constant term itself is perturbed and then the change in the full set of coefficients for that equation is used as a search direction.

(b) After (a) is complete, the change in the coefficients for the system as a whole is used as a search direction.

(c) The last (equation-by-equation) search directions in (a) and the direction in (b) are searched again.

Searches in (a) are linear for linear equations but non-linear otherwise, since the constant term is not, in general, a linear function of the other coefficients when mean residuals are kept constant. The algorithm also provides for the case in which there are constraints on the coefficients.

General theorems about the convergence properties of Algorithm A are difficult to come by. On a small number of test problems the convergence rate of Algorithm A compares favorably with a simple steepest ascent or a simple univariate relaxation algorithm that searches parallel to the coordinate axes. No

claim is made that Algorithm A's convergence rate can approximate that of Newton's method (although the latter is very much more expensive per iteration than the former), nor that Algorithm A will necessarily perform well on problems other than simultaneous equation estimation. Computational experience so far is fairly limited and appears to consist of estimates of two versions of the Fair (1976) model [see Fair and Parke (1980) and Parke (1979)]. In spite of the scant evidence the algorithm appears to be quite powerful in a rather sizeable model: in the model of Fair and Parke (1980), Algorithm A estimates 107 coefficients.

5.3. *The EM algorithm*

A particularly effective algorithm becomes possible in models involving incomplete data or latent or unobservable variables. The basic properties of the algorithm are given in Dempster, Laird and Rubin (1977); particular applications are treated in Hartley (1977a, 1977b) and Kiefer (1980).

The incomplete data problem may be stated as follows. Consider a random variable x with pdf $f(x|\theta)$ and assume the existence of a mapping from x to $y(x)$. It is assumed that x is not observed but is known to be in a set $X(y)$, where y represents the observed data. The y-data are incomplete in the sense that a y-observation does not unambiguously identify the corresponding x, but only $X(y)$. The y-data are generated by the density function

$$g(y|\theta) = \int_{X(y)} f(x|\theta)\,\mathrm{d}x. \tag{5.3}$$

A simple example is a multinomial model with k possible outcomes but with the restriction that for some pair of possible outcomes only their sum is observed. Another example is the switching regression model with the structure

$$\begin{aligned} y_i &= \beta_1' x_i + u_{1i} \quad \text{with probability } \lambda, \\ y_i &= \beta_2' x_i + u_{2i} \quad \text{with probability } 1-\lambda. \end{aligned} \tag{5.4}$$

In this model the x_i are exogenous variables, the β_j unknown parameters, the u_i the usual error terms, and the y_i the observed values of the dependent variables [see Hartley (1977a) and Kiefer (1980)]. The probability λ is unknown and we do not observe whether a particular y_i observation is generated by regime (5.4) or by (5.5). Other cases where the method is applicable are censored or truncated data, variance component estimation, estimation in disequilibrium models, etc.

The essential steps of the EM algorithm are the E-step and the M-step which are carried out at each iteration. At the kth iteration we have:

E-step: Given the current value θ^k of the parameter vector and the observed data y, calculate estimates for x^k as $\mathrm{E}(x|y, \theta^k)$.

M-step: Using the estimated values x^k, maximize the likelihood for the complete-data problem $\prod_{i=1}^{n} f(x_i^k|\theta)$ to determine θ^{k+1}.

The most important feature of the EM algorithm is that if it converges to a θ^*, then θ^* is a stationary point of the likelihood function $L(\theta) = \sum \log g(y_i|\theta)$. It has been suggested as a technique preferable to outright maximization of $L(\cdot)$ in instances (see Section 7) in which the likelihood function is unbounded in parameter space or, possibly, cases in which false apparent maxima exist. Whether these problems can typically be avoided by using the EM algorithm is not yet clear; nevertheless it is a powerful algorithm which may simplify as well as speed up convergence in the class of problems to which it is applicable. As an example we discuss the application to the switching regression model by Kiefer (1980).

Assume that n observations are generated by (5.4) with i.i.d. normal errors and the additional restriction that $\sigma_1^2 = \sigma_2^2$. Let W_1 be a diagonal matrix of order n where the ith diagonal element w_i represents the expected weight of the ith observation in the first regime and let $W_2 = I - W_1$. Then, maximizing the likelihood $\prod_{i=1}^{n} f(y_i|x_i, \theta)$, where $\theta = (\lambda, \beta_1, \beta_2, \sigma^2)$ yields

$$\hat{\beta}_j = (X'W_jX)^{-1}X'W_jY, \qquad j = 1,2,$$

$$\hat{\sigma}^2 = \sum_{j=1}^{2} (Y - X\beta_j)'W_j(Y - X\beta_j), \tag{5.5}$$

where X and Y are the matrices of observations on the x's and on y. Regarding regime choice as a Bernoulli experiment, λ is estimated as

$$\hat{\lambda} = \operatorname{tr}(W_1)/n.$$

Given these estimates for θ, representing the M-step, one can obtain new estimates for W_1 since for the ith observation

$$\begin{aligned} \mathrm{E}(w_i) &= (1)p(w_i = 1|y_i) + (0)p(w_i = 0|y_i) \\ &= p(w_i = 1|y_i) = \frac{\lambda g(y_i|w_i = 1)}{\lambda g(y_i|w_i = 1) + (1-\lambda)g(y_i|w_i = 0)}, \end{aligned} \tag{5.6}$$

by Bayes' Theorem. Since the right-hand side of (5.6) is easily computable, we can alternate between E and M steps as required.

5.4. *Simplified Jacobian computation*

If we are seeking FIML estimates for the coefficients of a system of simultaneous linear equations, the transformation from the pdf of the error terms to the pdf of

the jointly dependent variables involves the Jacobian $|\Gamma|$ of the transformation as in eq. (5.2). In the event that the equation system is non-linear the term $(n/2)\log[|\Gamma|]^2$ in (5.2) is replaced by

$$\sum_{i=1}^{n} \log |J_i|, \tag{5.7}$$

where J_i is the Jacobian corresponding to the ith observation. Clearly, the evaluation of (5.7) is likely to be much more expensive than the corresponding term in a linear system. Parke (1979), Fair and Parke (1980), and Belsley (1979) report good success with approximations that do not compute all n terms in the summation of (5.7). Various alternatives can be employed, such as approximating (5.7) by $(n/2)(\log|J_1| + \log|J_n|)$ or by computing a somewhat larger number of distinct Jacobians and interpolating for the missing ones. Fair and Parke report an example in which computations start with the simpler approximation and switch to a somewhat more expensive one with six Jacobians being computed for 98 data points. Belsley employs three Jacobian terms. All authors report that the approximations work quite well. The two- and six-term Jacobian approximations lead to substantially similar coefficient estimates and the corresponding objective functions rank the coefficient vectors consistently. The three-term approximation produces essentially the same results as the full Jacobian. It is difficult to predict how this type of approximation will perform in general. The acceptability of the approximation will surely depend on the degree of non-linearity:[13] the greater the non-linearity the worse the approximation may be expected to be. The time saving in computation may, however, be appreciable enough to recommend the procedure in most if not all instances of non-linear models.

6. Further aspects of algorithms

The previous two sections dealt with general as well as with special purpose optimization algorithms in rather broad terms, i.e. in terms that emphasized the general strategy and the key ideas of the algorithms in question. Most of these algorithms share certain detailed aspects which have been neglected up to now. The present section considers some of the salient aspects in this category. We specifically discuss (a) the computation of derivatives, (b) the techniques of linear searches, (c) stopping criteria, and (d) the problem of multiple optima.

[13] For some measures of non-linearity see Beale (1960) and Guttman and Meeter (1965).

6.1. Computation of derivatives

As shown above, many algorithms require that at least the first partial derivatives of the function be calculated; Newton-type methods also require the computation of second partial derivatives. Derivatives may be calculated analytically, i.e. by writing computer programs that evaluate the formulae that result from formal differentiation of the function in question or numerically by finite differencing. The evidence is clear that, other things equal, the former is vastly preferable. Not only do the various convergence properties presume the use of analytic derivatives, but in terms of the required computer time analytic derivatives clearly dominate their numerical counterparts, particularly for Newton-type methods [Belsley (1980)]. Unfortunately, for all but the smallest problems the calculations of analytic derivatives is highly labor intensive and in practice numerical derivatives are often employed, although some computer programs for symbolic differentiation exist (e.g. FORMAC). For numerical evaluation at least two choices have to be made: (a) Should derivatives be evaluated symmetrically or unsymmetrically? (b) How should one choose the length of the interval over which function differences are computed for arriving at a derivative approximation? First partial derivatives at x^0 are given by

$$\frac{\partial F(x^0)}{\partial x_i} = \frac{F(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_n^0) - F(x_1^0,\ldots,x_n^0)}{\varepsilon_i} \tag{6.1}$$

if evaluated unsymmetrically about x^0, and by

$$\frac{\partial F(x^0)}{\partial x_i} = \frac{F(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_n^0) - F(x_1^0,\ldots,x_i^0-\varepsilon_i,\ldots,x_n^0)}{2\varepsilon_i} \tag{6.2}$$

if evaluated symmetrically. If the value of $F(x^0)$ is already available (i.e. having already been computed by the algorithm), (6.1) requires n and (6.2) $2n$ additional function evaluations. Second direct partial derivatives are

$$\frac{\partial^2 F(x^0)}{\partial x_i^2} = \frac{F(x_1^0,\ldots,x_i^0-\varepsilon_i,\ldots,x_n^0) - 2F(x_1^0,\ldots,x_n^0) + F(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_n^0)}{\varepsilon_i^2}. \tag{6.3}$$

Second cross partial derivatives are

$$\frac{\partial^2 F(x^0)}{\partial x_i \partial x_j} = \Big[F\big(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_j^0+\varepsilon_j,\ldots,x_n^0\big) - F\big(x_1^0,\ldots,x_j^0+\varepsilon_j,\ldots,x_n^0\big) - F\big(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_n^0\big) + F\big(x_1^0,\ldots,x_n^0\big)\Big]/\varepsilon_i\varepsilon_j \tag{6.4}$$

or

$$\frac{\partial^2 F(x^0)}{\partial x_i \partial x_j} = \Big[F\big(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_j^0+\varepsilon_j,\ldots,x_n^0\big) - F\big(x_1^0,\ldots,x_i^0-\varepsilon_i,\ldots,x_j^0+\varepsilon_j,\ldots,x_n^0\big) - F\big(x_1^0,\ldots,x_i^0+\varepsilon_i,\ldots,x_j^0-\varepsilon_j,\ldots,x_n^0\big) + F\big(x_1^0,\ldots,x_i^0-\varepsilon_i,\ldots,x_j^0-\varepsilon_j,\ldots,x_n^0\big)\Big]/4\varepsilon_i\varepsilon_j. \tag{6.5}$$

The symmetric version (6.5) requires $(n-1)n/2$ more function evaluations than (6.4). The total number of function evaluations required for derivative calculations is:

	First derivatives	Second derivatives	Total
Unsymmetric	n	$(3n^2+n)/2$	$3n(n+1)/2$
Symmetric	$2n$	$2n^2$	$2n(n+1)$

Further compromises are clearly possible and often implemented, such as when first derivatives are calculated analytically and second derivatives numerically by differencing first derivatives.

An important question is how the values of ε_i and ε_j ought to be chosen. In practice they are chosen as small but arbitrary proportions of x_i^0 and x_j^0. For example, Chow and Megdal (1978) choose $\varepsilon_j = \max(\delta_1 x_j, \delta_2)$ where δ_1 and δ_2 are both 0.001. A procedure has been developed by Stewart (1967) for choosing the intervals optimally. Consider for simplicity a function $\phi(x)$ of a single variable and assume it to be the quadratic. Then, expanding $\phi(x+\varepsilon)$ about x and evaluating at $x=0$, the function can be written as

$$\phi(\varepsilon) = \alpha_0 + \alpha_1\varepsilon + \tfrac{1}{2}\alpha_2\varepsilon^2.$$

The first derivative can then be approximated by

$$\phi'(0) = \alpha_1 \simeq \frac{\phi(\varepsilon)-\alpha_0}{\varepsilon}. \tag{6.6}$$

The approximation (6.6) may be in error because it and all difference approximations to derivatives are Taylor series approximations and thus involve truncation error. The relative magnitude of this error is

$$\left|\frac{(\phi(\varepsilon)-\alpha_0)/\varepsilon-\alpha_1}{\alpha_1}\right| = \left|\frac{\alpha_2\varepsilon}{\alpha_1}\right|/2,$$

which clearly increases in the interval length ε. Another error is introduced if ε is small since in the numerator of (6.6) two numbers of comparable size are subtracted: in fixed word length computing serious rounding error may arise. We need to know an error bound $\bar{\eta}$ such that the computed value ϕ_c and true value ϕ are related by $\phi_c = \phi(1+\eta)$, where $|\eta| \leqq \bar{\eta}$. Then $\phi(\varepsilon)$ and α_0 can be computed as $\phi_c(\varepsilon) = \phi(\varepsilon)(1+\eta_1)$ and $\alpha_{0c} = \alpha_0(1+\eta_2)$, where $|\eta_1|$ and $|\eta_2| \leqq \bar{\eta}$. If ε is small so that $\alpha_0 \simeq \phi(\varepsilon)$, $\phi_c(\varepsilon)-\alpha_{0c} \simeq \phi(\varepsilon)-\alpha_0+\eta_3\alpha_0$, where $|\eta_3| \leqq 2\bar{\eta}$. It follows that the relative cancellation error is

$$\left|\frac{(\phi_c(\varepsilon)-\alpha_{0c})-(\phi(\varepsilon)-\alpha_0)}{\phi(\varepsilon)-\alpha_0}\right| \leqq 2\bar{\eta}\left|\frac{\alpha_0}{\phi(\varepsilon)-\alpha_0}\right|,$$

which is decreasing in ε if $\phi(\varepsilon)-\alpha_0$ is increasing in ε. Stewart suggests choosing ε so that the errors from the two sources are equal, which can be determined as the solution of one of the cubic equations below, where $\varepsilon_1 = |\varepsilon|$:

$$\begin{aligned}
&\tfrac{1}{2}\alpha_2^2\varepsilon_1^3 + |\alpha_2||\alpha_1|\varepsilon_1^2 - 4|\alpha_0||\alpha_1|\bar{\eta} = 0 && \text{if } \varepsilon > 0,\\
&-\tfrac{1}{2}\alpha_2^2\varepsilon_1^3 + |\alpha_2||\alpha_1|\varepsilon_1^2 - 4|\alpha_0||\alpha_1|\bar{\eta} = 0 && \text{if } -2|\alpha_1|/|\alpha_2| \leqq \varepsilon \leqq 0,\\
&\tfrac{1}{2}\alpha_2^2\varepsilon_1^3 - |\alpha_2||\alpha_1|\varepsilon_1^2 - 4|\alpha_0||\alpha_1|\bar{\eta} = 0 && \text{if } \varepsilon \leqq -2|\alpha_1|/|\alpha_2|.
\end{aligned} \tag{6.7}$$

Simplified versions of (6.7) as well as optimal solutions based on symmetric approximations exist and are generally desirable.

The computational cost of second derivative methods is significant enough to make them less than fully practical for large problems. Fair (1973) has applied such methods to problems with up to 40 variables, but problems not much larger than these may well represent the practical upper bound for using Newton-type methods for most researchers. How to economize on second derivative evaluations has been a problem of high priority. A simple solution that works in practice is to evaluate the matrix of second partial derivatives not at every iteration but at, say, every second or third iteration. The degradation of convergence that may occur is often more than made up by the savings in function evaluations.[14] An

[14] If the same Hessian is used for a number of iterations and happens to be a good estimate of $E[\partial^2 \log L/\partial\theta\partial\theta']$, the method is an approximation to the method of scoring.

important modification based on statistical theory is that of Berndt, Hall, Hall and Hausman (1974) and is applicable to the maximization of likelihood functions. The negative inverse of the matrix of second partial derivatives required for computation is a sample estimate of

$$-\left[\mathrm{E}\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]^{-1} = \left[\mathrm{E}\frac{\partial \log L}{\partial\theta}\frac{\partial \log L}{\partial\theta'}\right]^{-1}.$$

The (negative) expected value of the Hessian is thus, at the optimum, the covariance matrix of the gradient. It may therefore be inexpensively approximated by using the first derivatives of the likelihood function. Given, for example, the system of simultaneous non-linear equations

$$f_i(y_i, x_i, \theta) = u_i, \qquad i = 1, \ldots, n,$$

where f_i is a row vector with a component for each of g equations, y_i the vector of jointly dependent variables, and x_i the vector of predetermined variables, the log-likelihood function can be written analogously to (5.2) as

$$L = \text{const} - \tfrac{1}{2}\log|f'f| + \sum_{i=1}^{n} \log|J_i|,$$

where f is the $n \times g$ matrix containing as its rows f_i, and where $J_i = \partial f_i(y_i, x_i, \theta)/\partial y_i$. It is easy to show that the matrix of second partials can be approximated by

$$n\sum_{i=1}^{n}(p_i - q_i)(p_i - q_i)',$$

where $p_i = \partial \log|J_i|/\partial\theta$ and $q_i = (\partial f_i/\partial\theta)(\sum f_i' f_i)^{-1} f_i'$. The use of this approximation can be powerful by eliminating many of the function evaluations required for numerical derivatives. In addition, the approximation is positive definite and iterations will move uphill along the likelihood function. In practice it appears to work very well in many problems [Belsley (1980)], although it need not always provide a good approximation to the Hessian, particularly in small samples and at points not close to the optimum.

In spite of the several useful techniques discussed above, it may occur that a numerical approximation to the Hessian at the optimum is not negative definite. Although some algorithms may, in principle, converge to saddlepoints, this must generally be regarded as an unlikely event. The most plausible conclusion is that the numerical approximation to the Hessian has failed. Such an outcome is most frequent in cases where the function is extremely flat. It is clearly not an

acceptable outcome in any event, but particularly not in the case of maximum likelihood estimation for then the negative inverse of the Hessian is used as an estimate of the asymptotic covariance matrix. A heuristic technique that may occasionally be employed with success in such cases is as follows. Choose alternative values of the intervals ε_i over some fairly wide range and evaluate the Hessian for each. For large ε_i the truncation error, and for small ε_i the cancellation error, may predominate. For extreme values of the ε_i the estimates of the Hessian are likely to be unstable in the sense that the values of ε_i that are near to each other do not yield Hessians that are comparable. There may exist a (problem dependent) range of ε_i over which the estimates of the Hessian appear to be stable. If such a range exists, it is likely to be associated with numerically more accurate estimates of the Hessian.

6.2. *Linear searches*

Many algorithms require at each iteration, say the kth, the computation of λ_k such that $F(x^k + \lambda_k d^k)$ is maximized. It is clearly too expensive to require that (4.15) be satisfied exactly. In fact, normally it appears not worthwhile to calculate λ_k very accurately because of the excessive number of function evaluations this tends to require. Three procedures are discussed briefly.

(1) Fibonacci search [Spang (1962)]. Assume that the location of a unique maximum after p cycles of the linear search is known to be between $x + \lambda_a^p d$ and $x + \lambda_b^p d$. Then two more test values, λ_1^p and λ_2^p, are selected ($\lambda_a^p < \lambda_1^p < \lambda_2^p < \lambda_b^p$). If $F(x + \lambda_1^p d) > F(x + \lambda_2^p d)$, the maximum is between $x + \lambda_a^p d$ and $x + \lambda_2^p d$; the new lower and upper limits are $\lambda_a^{p+1} = \lambda_a^p$ and $\lambda_b^{p+1} = \lambda_2^p$, and the procedure is repeated with the new interval. (The corresponding actions are obvious if the inequality is reversed or if equality should be attained.) The values of λ_1 and λ_2 after the pth shrinking of the interval are obtained as

$$\lambda_1^p = \frac{U_{N-1-p}}{U_{N+1-p}}(\lambda_b^p - \lambda_a^p) + \lambda_a^p$$

and

$$\lambda_2^p = \frac{U_{N-p}}{U_{N+1-p}}(\lambda_b^p - \lambda_a^p) + \lambda_a^p,$$

where U_i denotes the ith Fibonacci number and N is a predetermined number of shrinkages. The rationale for Fibonacci search rests on the relationship between

the number of function evaluations and the size of the region covered when the error in finding the maximum satisfies a certain bound. Assume specifically that $F(x+\lambda d)$ has a unique maximum in the interval $[0, \Lambda_N]$, where Λ_N allows the maximum to be found with an error bounded by unity in no more than N function evaluations. Define λ_N as $\inf(\Lambda_N)$. Then one can prove the following:

Theorem 6.1.

The sequence λ_N is the Fibonacci sequence. Thus, for a given error bound of unity, the area searched increases with N fastest for the Fibonacci sequence; conversely, for a bounded area the error declines fastest. However, in practice this requires a prior determination of N. If the initial interval is large, the temptation may be to use a large value of N which will result in a large number of function evaluations at each line search. The method has been successful in applications [Daganzo, Bouthelier and Sheffi (1977)].

(2) Successive quadratic approximation [Powell (1964)]. Evaluate the function at $x+\lambda_1 d$, $x+\lambda_2 d$, and $x+\lambda_3 d$ and determine the stationary point of the quadratic function of λ fitted exactly to those points. The stationary point occurs at

$$\lambda_4 = \frac{1}{2}\frac{(\lambda_2^2-\lambda_3^2)F(x+\lambda_1 d)+(\lambda_3^2-\lambda_1^2)F(x+\lambda_2 d)+(\lambda_1^2-\lambda_2^2)F(x+\lambda_3 d)}{(\lambda_2-\lambda_3)F(x+\lambda_1 d)+(\lambda_3-\lambda_1)F(x+\lambda_2 d)+(\lambda_1-\lambda_2)F(x+\lambda_3 d)}.$$

If the stationary point is a maximum and if the proposed step is not greater than a preassigned tolerance, the λ corresponding to the smallest value of $F(x+\lambda d)$ is discarded and the computation repeated with the surviving three λ's. If λ_4 corresponds to a minimum or implies a step greater than the maximum allowed, λ_4 is chosen to correspond to the largest permitted step, the λ-value farthest from λ_4 is discarded and the computation repeated. If a computed λ_4 is within a preassigned ε-distance of λ_1, λ_2, and λ_3, it is accepted as the maximum and the line search is terminated. [A more elaborate cubic interpolation scheme is suggested by Davidon (1959). Dagenais (1978) first fits a quadratic from two function values and the gradient, adds a third point which is the maximum of the fitted quadratic, and then fits a cubic.]

(3) Powell (1971) and Berndt, Hall, Hall and Hausman (1974) recommend the following approximate procedure. First select an ε such that $0 < \varepsilon < \frac{1}{2}$. Then find λ^k such that

$$\varepsilon \leqq \frac{F(x^k+\lambda^k d^k)-F(x^k)}{\lambda^k g_k' d^k} \leqq 1-\varepsilon. \tag{6.8}$$

Since the fraction in (6.8) approaches 1 as $\lambda^k \to 0$ and approaches zero or a negative number as $\lambda^k \to \infty$, a suitable value of λ^k exists and can be found by successive evaluations; the first λ^k that satisfies (6.8) is then used.

6.3. *Stopping criteria*

An important part of every algorithm is the criterion it employs for terminating computations. The ideal of reaching a point x^k such that $\partial F(x^k)/\partial x_i = 0$ $(i = 1,\ldots,n)$ is not attainable in practice and the question is what compromise is most reasonable. In the neighborhood of the maximum any algorithm is likely to take small steps in the sense that $|x_i^{k+1} - x_i^k|$ is likely to be small for all values of i and in the sense that $|F(x^{k+1}) - F(x^k)|$ is likely to be small; accordingly both of these quantities have been employed as stopping criteria. In fact, it is theoretically possible for either difference to become small while the other is large; it is thus preferable to use both and continue iterating unless both criteria are satisfied. In addition, since at the exact maximum the gradient in zero, $g_k' g_k$ is also an obvious choice for judging closeness to the maximum. Perhaps the most common is to test the relative change in the variables and accordingly computations terminate if

$$\max \frac{|x_i^{k+1} - x_i^k|}{\max(\varepsilon_1, |x_i^k|)} \leqq \varepsilon_2,$$

where ε_1 and ε_2 are preassigned tolerances [see Powell (1964) and Berndt, Hall, Hall and Hausman (1974)]. A variant is to terminate if

$$2\left[\sum_{i=1}^{n} \left(x_i^{k+1} - x_i^k\right)^2\right]^{1/2} \leqq \varepsilon_1^{1/2}\left[\sum_{i=1}^{n} \left(x_i^{k+1}\right)^2\right]^{1/2} + \varepsilon_2,$$

where ε_1 is machine precision defined as $\beta^{1-\tau}$ (for truncated arithmetic) or $\beta^{1-\tau}/2$ (for rounded arithmetic), where β is the base of arithmetic and τ the number of floating point digits [Brent (1973)]. Some algorithms employ a combination of criteria and may terminate when any of the criteria are satisfied. Belsley (1980) criticizes the gradient criterion as not being scale independent and ignoring statistical considerations in the sense that a gradient component in the direction of an insignificant parameter has the same weight as one in the direction of a statistically significant one. He criticizes the relative variable-change criterion as treating all variables equally (although the criterion is scale invariant). The relative function-change criterion is not invariant with respect to the scaling of the function. Belsley suggests a weighted-gradient stopping criterion $-g_k' H^{-1} g_k < \varepsilon$, where H^{-1} is the inverse Hessian. A similar suggestion in a least squares context

is contained in Dennis, Gay and Welsch (1979). This criterion is clearly scale invariant and in maximum likelihood estimation it can be interpreted to weight parameter estimates according to their (asymptotic) significance. Moreover, the criterion may be recognized as the Lagrange multiplier test statistic and accordingly iterations continue until the test statistic confirms that the gradient is small. Computational experience with the criterion appears to be quite satisfactory and it results in terminations that are relatively model independent.[15]

6.4. *Multiple optima*

All algorithms discussed so far (with the possible exception of appropriately designed grid searches) locate only local maxima. There is no guarantee that only one maximum exists or, if more than one exists, that the maximum found is the global maximum. There are, of course, numerous instances in which the likelihood function can be proved to be globally concave, in which case a unique maximum exists. Cases in point are the likelihood function associated with the classical normal regression model or that of the probit or logit models. For the latter two, in the simplest case, the dependent variable value at the ith observation, y_i, is dichotomous:

$$y_i = \begin{cases} 0 & \text{with probability } 1 - \Phi(x_i'\beta), \\ 1 & \text{with probability } \Phi(x_i'\beta), \end{cases}$$

where x_i' is the (row) vector of independent variables, β a (column) vector of parameters, and Φ a distribution function, normal for the probit model and logistic for the logit model. The likelihood is

$$L = \prod_{i=1}^{n} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i},$$

and it is not difficult to show that

$$\frac{\partial^2 \log L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \phi(x_i'\beta) x_i x_i',$$

where $\phi(z) = d\Phi(z)/dz$ is the probability density function, is negative definite [Dhrymes (1978)].

[15]A termination criterion that is qualitatively different from all of the ones discussed is the criterion employed in the Nelder and Mead (1965) algorithm. According to this criterion computations stop if the sample variance of the function values at the vertices of the current simplex falls below a preset tolerance.

In some cases the existence or the possibility of multiple maxima can be shown analytically. Some illustrative examples are provided by Goldfeld and Quandt (1972); a realistic example pertaining to pooling cross-section and time-series data is discussed by Maddala (1971). Numerous instances exist in which a global maximum has not necessarily been found [see Fair (1974a)]. When several maxima exist, it is important to attempt to find the global maximum since it is (customarily if not uniformly) implicit in the asymptotic justification for maximum likelihood estimation that the global maximum is attained.

Unfortunately, mostly *ad hoc* methods are employed for locating multiple optima. (a) The most common method is to select several (or many, if cost considerations permit) starting values for the vector of unknowns and to reoptimize repeatedly using the algorithms discussed previously. If all starting points lead to the same local optimum, the tendency is to declare it to be the unique maximum with substantial confidence. (b) Assume that a local maximum has been found at x^0. Goldfeld and Quandt (1972) propose to find a solution to $F(x) = F(x^0) + \varepsilon$ for small positive ε; if there exists a solution to the equation, x^0 cannot correspond to a global maximum. (c) A deflation method is suggested by Brown and Gearhart (1971) and explored by Salmon (1978) with a view towards solving large-scale econometric models. It appears feasible to apply the method to solving the first-order conditions of a maximum problem. The first-order conditions are

$$g(x) = 0 \tag{6.9}$$

and assume that (6.9) has been written as

$$x = \phi(x), \tag{6.10}$$

as would be required to obtain a solution by the Jacobi or Gauss–Seidel method. Let x^0 be a solution obtained by one of these methods. Define

$$\|x - x^0\|_p = \left[\sum_{i=1}^{n} \left(x_i - x_i^0\right)^p \right]^{1/p},$$

p normally > 1, and consider solving

$$g^*(x) = \frac{g(x)}{\|x - x^0\|_p} = x - \phi^*(x) = 0.$$

This suggests iterations according to $x^{k+1} = \phi^*(x^k)$, where

$$\phi^*(x) = x - \frac{x - \phi(x)}{\|x - x^0\|_p},$$

leading to iterations defined by

$$x^{k+1} = \left(1 - \frac{1}{\|x^k - x^0\|_p}\right) x^k + \frac{1}{\|x^k - x^0\|_p} \phi(x^k), \tag{6.11}$$

which shields the algorithm from the previously obtained solution x^0. Additional solutions give rise to similar deflators of $g(x)$ until no further solutions are found. Experience with the method appears to be limited to test examples and to solving the Australian Treasury N1F7 macro model with 128 equations. Although the method will share the difficulties of the Gauss–Seidel method and need not converge after deflation even if a second solution exists, it appears to be an interesting candidate for further study.

7. Particular problems in optimization

In addition to the standard features of optimization algorithms and problems, there occur in practice several particular problems that may be difficult to cope with. Most of these are associated with special models; accordingly, their treatment will be relatively brief.

7.1. *Smoothing of non-differentiable functions*

These are several related contexts in which non-differentiable likelihood functions may arise. One is the switching regression model

$$\begin{aligned} y_i &= \beta_1' x_i + u_{1i} \quad \text{if } \pi' z_i \leqq 0, \\ y_i &= \beta_2' x_i + u_{2i} \quad \text{if } \pi' z_i > 0, \end{aligned} \tag{7.1}$$

where β_1, β_2, and π are unobserved parameter vectors and x_i and z_i are exogenous variable vectors. Define $D = D(\pi' z_i) = 0$ if $\pi' z_i \leqq 0$ and $D_i = 1$ otherwise. Eqs. (7.1) may then be rewritten as

$$y_i = (1 - D_i)\beta_1' x_i + D_i \beta_2' x_i + u_{1i}(1 - D_i) + u_{2i} D_i. \tag{7.2}$$

If u_{1i} and u_{2i} are distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, the likelihood function is

$$L = (2\pi)^{-n/2} \left(\sigma_i^2\right)^{-1/2} \exp\left\{ - \sum_{i=1}^{n} \left(y_i - (1 - D_i)\beta_1' x_i - D_i \beta_2' x_i\right)^2 \Big/ 2\sigma_i^2 \right\}, \tag{7.3}$$

where σ_i^2 is defined as $\sigma_1^2(1-D_i)^2+\sigma_2^2 D_i^2$. The unknowns in this problem are β_1, β_2, σ_1^2, σ_2^2, and the D_i which are discrete; hence, derivatives with respect to the D_i do not exist.

An alternative model, first suggested by Tishler and Zang (1977) and shown to be a special case of a general disequilibrium model by Goldfeld and Quandt (1978), is

$$\begin{aligned} y_{1i} &= \beta_1' x_{1i}, \\ y_{2i} &= \beta_2' x_{2i}, \\ y_i &= \min(y_{1i}, y_{2i}) + u_i, \end{aligned} \tag{7.4}$$

where y_{1i} and y_{2i} are not observed and y_i is observed. The likelihood function corresponding to (7.4) is

$$\begin{aligned} L = &\prod_{\beta_1' x_{1i} \leqq \beta_2' x_{2i}} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \beta_1' x_{1i})^2 \right\} \\ &\times \prod_{\beta_1' x_{1i} > \beta_2' x_{2i}} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \beta_2' x_{2i})^2 \right\}. \end{aligned} \tag{7.5}$$

Rewriting $\beta_1' x_{1i} - \beta_2' x_{2i} \leqq 0$ as $\pi' z_i \leqq 0$ shows that (7.5) is formally identical with (7.3) with the added restriction that $\sigma_1^2 = \sigma_2^2$. The function (7.3) is not differentiable everywhere because of the discrete D_i and function (7.5) exhibits the same problem whenever the β's pass through values at which the sorting of observations between the two types of products changes. Various types of smoothing have been suggested to cope with this problem. The technique employed by Goldfeld and Quandt (1972) replaces D_i with the approximation

$$D_i \simeq \int_{-\infty}^{\pi' z_i} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2}\frac{\xi^2}{\sigma^2} \right\} \mathrm{d}\xi, \tag{7.6}$$

where σ^2 is a new parameter that has been interpreted as a measure of the extent to which the approximation can reproduce the exact sorting of the data. Clearly, other distribution functions might be used in the place of (7.6). Tischler and Zang (1979) recommend several spline approximations of which the most promising appears to be the quintic given by

$$D_i = \begin{cases} 0 & \text{if } \pi' z_i \leqq -\alpha, \\ \dfrac{3}{16}\left(\dfrac{r}{\alpha}\right)^5 - \dfrac{5}{8}\left(\dfrac{r}{\alpha}\right)^3 + \dfrac{15}{16}\left(\dfrac{r}{\alpha}\right) + \dfrac{1}{2} & \text{if } -\alpha \leqq \pi' z_i \leqq \alpha, \\ 1 & \text{if } \alpha \leqq \pi' z_i, \end{cases} \tag{7.7}$$

where α is a new (positive) parameter to be determined. In practice both types of approximation appear to work well.

7.2. *Unbounded likelihood functions and other false optima*

Ordinarily the value of the likelihood function is bounded. In certain classes of models, usually dealing with unobserved or latent variables, the likelihood function may become unbounded. This is a serious difficulty in that any algorithm that happens to locate a neighborhood within which the function is unbounded is likely to break down; moreover, the point or points in parameter space at which unboundedness occur have no desirable statistical properties. Three examples are given.

The first is a case in which unboundedness may occur under special circumstances but is unlikely to do so in general. It is the well-known tobit model:

$$\begin{aligned} y_i &= \beta' x_i + u_i \quad \text{if } \beta' x_i + u_i > 0, \\ y_i &= 0 \qquad\quad \text{if } \beta' x_i + u_i \leqq 0. \end{aligned} \tag{7.8}$$

The likelihood function is

$$L = \prod_{i \in I} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta' x_i}{\sigma} \right)^2 \right\} \prod_{i \in \bar{I}} \left[1 - \Phi\left(\frac{\beta' x_i}{\sigma} \right) \right], \tag{7.9}$$

where I and $\bar{I}$ are the sets of indices for which the two inequalities in (7.8) hold, respectively, and where $\Phi(\cdot)$ is the standard cumulative normal distribution. Assume that β has k elements and that the number of elements in I is less than k. Then if the x_i $(i \in I)$ are linearly independent, there exists a value β^* such that $y_i - \beta^{*\prime} x_i \equiv 0$, $i \in I$. Assume further that $\beta^{*\prime} x_i \leqq 0$, $i \in \bar{I}$; i.e. that $\beta^{*\prime} x$ is a supporting hyperplane for the convex hull of the x_i, $i \in \bar{I}$. Then consider the sequence of points obtained by letting $\sigma \to 0$. The product of terms for $i \in I$ becomes unbounded, whereas the product of terms for $i \in \bar{I}$ remains bounded away from zero; hence $L \to \infty$.

The second example is the well-known switching regression model [Quandt (1972), and Quandt and Ramsey (1979)]:

$$y_i = \beta_1' x_{1i} + u_{1i} \quad \text{with probability } \lambda,$$

$$y_i = \beta_2' x_{1i} + u_{2i} \quad \text{with probability } 1 - \lambda.$$

The likelihood function with normal errors is

$$L = \prod_{i=1}^{n} \left[\frac{\lambda}{\sqrt{2\pi}\,\sigma_1} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta_1' x_i}{\sigma_1} \right)^2 \right\} + \frac{1-\lambda}{\sqrt{2\pi}\,\sigma_2} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta_2' x_i}{\sigma_2} \right)^2 \right\} \right]. \tag{7.10}$$

Choose a β_1^* such that the kth observation $y_k - \beta_1^{*\prime} x_k \equiv 0$ and let $\sigma_1^2 \to 0$. Then $\lambda \exp\{-(y_k - \beta^{*\prime} x_k)^2/2\sigma_1^2\}/\sqrt{2\pi}\,\sigma_1$ becomes unbounded, but $(1-\lambda)\exp\{-(y_i - \beta_2' x_i)^2/2\sigma_2^2\}/\sqrt{2\pi}\,\sigma_2$ remains bounded away from zero for all i; hence $L \to \infty$.

The third example is that of a simple disequilibrium model:

$$\begin{aligned} y_{1i} &= \beta_1' x_{1i} + u_{1i}, \\ y_{2i} &= \beta_2' x_{2i} + u_{2i}, \\ y_i &= \min(y_{1i}, y_{2i}). \end{aligned} \tag{7.11}$$

The likelihood function with normal errors and u_{1i} and u_{2i} independent is

$$L = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta_1' x_{1i}}{\sigma_1} \right)^2 \right\} \left(1 - \Phi\left(\frac{y_i - \beta_2' x_{2i}}{\sigma_2} \right) \right) + \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta_2' x_{2i}}{\sigma_2} \right)^2 \right\} \left(1 - \Phi\left(\frac{y_i - \beta_1' x_{1i}}{\sigma_1} \right) \right) \right]. \tag{7.12}$$

An argument very similar to that employed in the case of the switching regression model can be used to show that the simple disequilibrium model has an unbounded likelihood function [Quandt (1978a)]. In most instances there is an infinity of points at which the likelihood function becomes unbounded: thus, for example, β_1^* in (7.10) can be chosen in infinitely many ways so as to make one residual exactly equal to zero. No completely satisfactory methods are known to avoid the computational problems. A device employed frequently is to constrain the variances by $\sigma_1^2 = \alpha\sigma_2^2$, where α is selected *a priori*. This guarantees to solve the computational problem at the cost of introducing misspecification. Other methods [Hartley, (1977a, 1977b)] have suggested that adaptations of the EM algorithm may tend to avert the problem. Although the problem is not encountered very frequently in practice, it must be considered a difficult one.

Another problem with potentially equally bad computational consequences is the appearance of "false optima". A case in point is the likelihood function corresponding to (7.11) when $E(u_{1i}u_{2i}) = \sigma_{12} \neq 0$. Define $r_{ij} = (y_i - \beta_j' x_i)/\sigma_j$ as a

normalized residual. Then the following can be shown [Goldfeld and Quandt (1978)].

(a) If values of β_1, β_2, σ_1^2, and σ_2^2 have been selected such that $r_{i1} + r_{i2} < 0$ for all i, then the likelihood function increases as the correlation, ρ, between u_1 and u_2 approaches -1.

(b) If $r_{i2} - r_{i1} > 0$, when $\exp(-r_{i2}^2)/\sigma_2 > \exp(-r_{i1}^2)/\sigma_1$ and $r_{i2} - r_{i1} < 0$ otherwise, then the likelihood function increases as $\rho \to 1$.

The practical consequences of the result is that if values of parameters happen to be achieved that satisfy the above conditions during the iterations of an algorithm, then the algorithm may attempt to push the value of ρ arbitrarily close to ± 1. Since the likelihood function is not defined at $\rho = \pm 1$, computation normally fails at some point in such a neighborhood. Such a point is not a true local maximum since the likelihood function is defined only over the open interval $-1 < \rho < 1$, but it may computationally appear as one. There is as yet no obvious method to guard against the occurrence of this problem. One might wish to impose constraints so that the inequalities required by the difficulty do not hold. However, this may be unsatisfactory since the inequalites may be satisfied even if the true values of the β's and σ's are substituted.

7.3. *Constraints on the parameters*

Classical constrained optimization problems arise if there are equality or inequality constraints on the parameters arising from the intrinsic aspects of the problem. A case in point would be the requirement that the exponents of a Cobb–Douglas production function add to unity. Another example, discussed by MacKinnon (1979), is when economic considerations require that the Jacobian term in a likelihood function have a particular sign (even though it is the absolute value of the Jacobian that enters the likelihood function). Equality constraints, the more common case, may often be handled adequately by using the constraints to eliminate variables of optimization from the objective function.[16] In general, problems of this type need to be handled by the methods of non-linear programming. We briefly discuss only two classes of algorithms.

[16] This has the added advantage of reducing the number of variables in the optimization. It follows that whenever a subset of the first-order condition can be solved for a subset of variables, the (likelihood) function should be condensed. It is not hard to show that the negative inverse Hessian of the condensed log-likelihood function is the appropriate estimator of the asymptotic covariance matrix of the variables that have not been condensed out.

Consider the constrained optimization problem:

$$\begin{aligned} &\text{maximize} \quad F(x) \\ &\text{subject to} \quad \psi_i(x) \geqq 0, \qquad i = 1,\ldots,m. \end{aligned} \tag{7.13}$$

If none of the constraints in (7.13) is binding, then an iterative step may be taken as if one were dealing with an unconstrained problem. If, however, one or more constraints are binding, the question arises of how to choose a search direction. A particular approach to this problem is provided by the Rosen gradient projection method [Rosen (1960, 1961) and Walsh (1975)]. Assume that $m_1 \leqq m$ of the constraints are binding and consider the subspace spanned by the constraint gradients $\nabla\psi_i$, $i = 1,\ldots,m_1$. The key of Rosen's algorithm is to choose as the search direction the projection of the gradient of $F(x)$, g, on the orthogonal complement of the subspace spanned by the $\nabla\psi_i$. Denoting the matrix the columns of which are $\nabla\psi_i$ by Ψ, the search direction is $d = (I - \Psi(\Psi'\Psi)^{-1}\Psi')g$.

A different class of algorithms converts a constrained maximization problem into a sequence of unconstrained ones. This is usually accomplished by penalizing the objective function for (near) violations of the constraint by adding to it penalty or barrier functions [Fiacco and McCormick (1964), Osborne (1972), and Walsh (1975)]. Thus, one might define $\xi(\psi_1,\ldots,\psi_m) = \sum_{i=1}^{m} \log \psi_i(x)$ and consider the unconstrained maximization of

$$W = F(x) + \gamma\xi(\psi(x))$$

for some positive number γ. Solve this unconstrained problem repeatedly for a sequence of $\gamma \to 0$. It can be shown under general conditions that the corresponding sequence of solutions, x, converges to the solution of (7.13). The obvious advantage of this approach is that it converts the original problem into one which is generally easier to solve. The disadvantage is that the single constrained optimization problem has been replaced by a sequence of unconstrained problems, which can result in high computational cost.

An interesting variant of this uses an augmented Lagrangean expression [Pierre and Lowe (1975)]. Consider the standard Lagrangean:

$$L(x,\alpha) = F(x) + \alpha'\psi(x), \tag{7.14}$$

where $\psi(x)$ is the vector with elements $\psi_i(x)$ and α a vector of constants. Form the augmented Lagrangean

$$L(x,\alpha,\beta_1,\beta_2) = L(x,\alpha) - \beta_1 \sum_{i \in I_1} \psi_i(x)^2 - \beta_2 \sum_{i \in I_2} \psi_i(x)^2, \tag{7.15}$$

where β_1 and β_2 are preassigned weights and where $I_1 = \{i | \alpha_i > 0\}$, $I_2 = \{i | \alpha_i = 0$, and $\psi_i(x_i) \geqq 0\}$. The algorithm alternates between two steps: the maximization step in which an unconstrained maximum of (7.15) is found and an adjustment step in which the Lagrange multipliers are adjusted so as to equate the gradient of the augmented Lagrangean to that of the simple Lagrangean and in which β_1 and β_2 may be increased. The procedure obtains its justification from the following:

Theorem 7.1

If x^*, α^* solves the appropriate Kuhn–Tucker conditions, then x^* satisfies sufficient conditions for an unconstrained maximum for $L(x, \alpha^*, \beta)$ and sufficiently large β if and only if x^*, α^* solves the non-linear programming problem.

Constraints on the variables can also arise in a different manner. During the optimization the variables may stray into a region in which the function is not defined. This may occur in a number of different ways. In maximizing a function such as (7.10) it is not possible to condense out the variances and an algorithm may wish to take a trial step that would make a variance negative. Alternatively, an equation to be estimated may involve functional forms that are defined only for certain parameter values, say as in $y_i = \log(1 + \alpha x_i) + u_i$. Technically, another case is provided by a simultaneous linear equation model in which the number of endogenous and exogenous variables is greater than the number of observations, in which case the estimated covariance matrix of residuals will be singular and consequently the likelihood function undefined [Parke (1979)]; in this case, however, one would not actually attempt optimization.

At least three *ad hoc* remedies may be employed, although none of them is assured of success. First, in some cases it may be possible to reparameterize the variables. Thus, if in (7.10) σ_1^2 has a tendency to become negative, one may replace it by e^{ω}. This may exchange one problem for another: in the transformed space the likelihood function may become very flat. Alternatively, one may continually test whether the current values of the variables are attempting to enter a forbidden region and inhibit the algorithm from proceeding in such a direction, either by returning to the algorithm an extremely unfavorable (pseudo-) function value associated with the illegitimate point, or by shrinking the step size. The former technique gives seriously erroneous estimates of derivatives. The latter may slow down convergence considerably. In spite of these difficulties, these *ad hoc* techniques are often employed and often work reasonably well. What must be stressed, however, is that the latter two may be employed only to guard against attempts to evaluate the function at points at which it is not defined; they ought not be used to substitute for general non-linear programming in cases in which the constraints represent economic restrictions. There is no *a priori* reason to believe that just because an algorithm strays into a forbidden region in an economic sense, the location of the maximum will also be in a forbidden region. If

several local constrained optima exist, a casual use of the constraint within an otherwise unconstrained maximization algorithm may well jeopardize locating the desired point.

8. Numerical integration

Normally, there are two econometric contexts in which numerical integration becomes necessary. The first is Bayesian analysis in which, say, the moments of the posterior density are required. An example is provided by Kloek and Van Dijk (1978) who deal with the linear simultaneous equation model:

$$Y\Gamma + ZB = U. \tag{8.1}$$

Denoting those elements of Γ and B that are not constant terms by θ and the prior density of θ by $p(\theta)$ and assuming (a) that the constant terms have uniform prior, (b) that the prior of the covariance matrix Σ is of the form $|\Sigma|^{-(G+1)/2}$, where G is the number of equations, and (c) that the constant terms and covariance matrix elements have been integrated out, the marginal posterior of θ can be shown to be proportional to

$$\kappa(\theta | Y, Z)p(\theta), \tag{8.2}$$

where $\kappa(\theta | Y, Z)$ depends on the structural parameters other than constant terms and on the observations. The moments are functions of θ, say $g(\theta)$, and can be written as

$$E[g(\theta) | Y, Z] = \frac{\int g(\theta)\kappa(\theta | Y, Z)p(\theta)\,\mathrm{d}\theta}{\int \kappa(\theta | Y, Z)p(\theta)\,\mathrm{d}\theta}. \tag{8.3}$$

Kloek and Van Dijk (1978) consider various alternatives for $p(\theta)$ in a small model such as the normal and beta distributions for which the integrals in (8.3) are not available in closed form.

The second context in which numerical integration is required is in finding maximum likelihood estimates in models with qualitative dependent variables, i.e. variables that are endogenous and involve essential elements of discreteness. Models of this type invariably contain some features that are not observable. Simple cases in point are the following models, where greek letters denote parameters, x's exogenous variables, i indexes observations, and $\Phi(\cdot)$ denotes the standard cumulative normal integral.

(a) The Probit Model:

$$y_i = 1 \quad \text{if } \beta' x_i + u_i > 0,$$
$$y_i = 0 \quad \text{otherwise,}$$

where $u_i \sim N(0,1)$. The likelihood function is

$$L = \prod_{y_i=1} (1 - \Phi(-\beta' x_i)) \prod_{y_i=0} \Phi(-\beta' x_i).$$

(b) The Tobit Model which is stated in (7.8) with likelihood function (7.9).

(c) The Simple Disequilibrium Model which is stated in (7.11) with likelihood function (7.12).

In all of these likelihood functions $\Phi(\cdot)$ appears which is not available in closed form. Fortunately, simple and accurate approximations for the cumulative normal integral exist in the univariate case as given, for example, by the FORTRAN subroutines ERF and DERF. The problem becomes much more difficult in problems in which multiple integrals of multivariate densities are required [see Hausman and Wise (1978)]. An important example is provided by discrete choice models in which an individual must chose among m possibilities. Let the ith individual's utility from choosing alternative j be

$$U_{ij} = V(C_{ij}, \beta) + \varepsilon_{ij} = \overline{U}_{ij} + \varepsilon_{ij},$$

where $\overline{U}_{ij} = V(C_{ij}, \beta_i)$ represents the systematic part of utility and where C_{ij} are objective measures of the individual's and the alternative's characteristics, β are parameters, and ε_{ij} is a random variable. Then the probability that alternative k is chosen is

$$P_{ik} = \Pr\{\varepsilon_{ij} \leqq \overline{U}_{ik} - \overline{U}_{ij} + \varepsilon_{ik} \quad \text{for all } j \neq k\}.$$

If $h(\varepsilon_{i1}, \ldots, \varepsilon_{im})$ is the joint density of the ε_{ij}, P_{ik} is

$$P_{ik} = \int_{-\infty}^{\infty} \int_{-\infty}^{\overline{U}_{ik} - \overline{U}_{i1} + \varepsilon_{ik}} \cdots \int_{-\infty}^{\overline{U}_{ik} - \overline{U}_{im} + \varepsilon_{ik}} h(\varepsilon_{i1}, \ldots, \varepsilon_{im})\, d\varepsilon_{im} \ldots d\varepsilon_{i1}\, d\varepsilon_{ik} \tag{8.4}$$

or

$$P_{ik} = \int_{-\infty}^{\overline{U}_{ik} - \overline{U}_{i1}} \cdots \int_{-\infty}^{\overline{U}_{ik} - \overline{U}_{im}} h^k(\eta_{i1k}, \ldots, \eta_{imk})\, d\eta_{i1k} \ldots d\eta_{imk}, \tag{8.5}$$

where $\eta_{ijk} = \varepsilon_{ij} - \varepsilon_{ik}$ and where $h^k(\cdot)$ is the joint density of the $(m-1)$ η_{ijk}. Hence, with m alternatives, an $(m-1)$-fold integral must be evaluated. If there are n individuals, the likelihood of a sample of observations is

$$L = \prod_{i=1}^{n} P_{i1}^{y_{i1}} \dots P_{im}^{y_{im}}, \tag{8.6}$$

where $y_{ij} = 1$ if the ith individual chooses alternative j and zero otherwise and where the P_{ij} are replaced by the expressions in (8.4) or (8.5). Eq. (8.6) must then be maximized with respect to the parameter vector β; hence, every function evaluation requires the evaluation of multiple integrals. If the errors ε_{ij} are multivariate normally distributed with non-diagonal covariance matrix, as is assumed by Hausman and Wise (1978), the integrals must be obtained numerically. A similar situation arises in the multimarket disequilibrium model of the type given in Section 5: in general, the density function for the observable variables in an m-market disequilibrium model involves an m-fold integral of the m-variate normal distribution.

In general, one would expect to make greater demands for accuracy in the case of likelihood maximization for models with qualitative dependent variables than in the case of computing the moments of posterior densities. In the latter case it might be acceptable to have a 10 percent error from the point of view of providing economic interpretation for the results. In the case of likelihood maximization an average error of 10 percent in evaluating the likelihood function is likely to cause serious problems of convergence. Hence, methods that are suitable for one type of problem are not likely to be suitable for the other. In what follows we do not distinguish systematically between univariate and multivariate integration.

8.1. *Monte Carlo integration*

Assume that we require the integral

$$I = \int_a^b f(x)\,\mathrm{d}x. \tag{8.7}$$

Let $g(x)$ be a probability density defined over (a, b). Then

$$E(f(x)) = \int_a^b f(x)g(x)\,\mathrm{d}x,$$

and if $g(x)$ is uniform,

$$E(f(x)) = \frac{I}{b-a}.$$

If n points $x_1,\ldots,x_n$ are selected randomly, $\sum_{i=1}^{n} f(x_i)/n$ converges in probability to $E(f(x))$ and I can be approximated by I_a defined as

$$I_a = \frac{b-a}{n} \sum_{i=1}^{n} f(x_i). \tag{8.8}$$

Clearly, $E(I_a) = I$ and I_a is the sum of n i.i.d. random variables with mean I/n and variance $\omega^2 = (b-a)^2\sigma^2/n^2$, where $\sigma^2 = \operatorname{var} f(x_i)$ and may be estimated by the sample variance. By the Central Limit Theorem, with probability α,

$$|I_a - I| \leqq z_\alpha \frac{(b-a)\sigma}{\sqrt{n}},$$

where z_α satisfies $\Phi(z_\alpha) - \Phi(-z_\alpha) = \alpha$. The error decreases as $n^{-1/2}$ and is independent of the dimensionality of the integral [Hammersley and Handscomb (1964) and Shreider (1964)].

A variance reducing technique is importance sampling. Write (8.7) as

$$I = \int_a^b \frac{f(x)}{g(x)} g(x)\,\mathrm{d}x = E(f(x)/g(x)), \tag{8.9}$$

where $g(x)$ is a pdf over (a, b) as before. If points $x_1,\ldots,x_n$ are generated with pdf $g(x)$, I is now estimated by

$$I_a = \frac{1}{n} \sum \frac{f(x_i)}{g(x_i)}. \tag{8.10}$$

The variance of $f(x)/g(x)$ can be made to equal zero by setting $g(x) = |f(x)|/\int_a^b |f(x)|\,\mathrm{d}x$, which is not practical for it requires knowledge of the integral in question. As a practical matter, $g(x)$ is chosen so as to make the variation in $f(x)/g(x)$ small: the implication is that x will be sampled relatively more frequently in regions in which $f(x)$ is large or important. Examples of importance sampling are in Kloek and Van Dijk (1978) and Quandt (1978b).

8.2. *Polynomial approximations*

If a function $f(x)$ is approximated by a polynomial of degree n such that the approximating polynomial agrees with it at $n+1$ equally spaced points, it can be written as

$$f(x) \simeq \sum_{k=0}^{n} \lambda_k(x) f(x_k),$$

where the $\lambda_k(x)$ are the Lagrangean polynomial coefficients. A class of integration formulae employ the integral of the approximating polynomial and are called Newton–Cotes formulae. Simpson's Rule is a special case of Newton–Cotes integration. [For an application see Richard and Tompa (1980).]

Although simple to implement, Newton–Cotes formulae can be subject to serious error and there are cases where the approximation does not converge to the true integral as $n \to \infty$. More stable are Gaussian Quadrature formulae using n points of evaluation for which the approximation is exact if $f(x)$ is polynomial of degree $\leqq 2n-1$ [Hildebrand (1956) and Stroud and Secrest (1966)]. Gaussian formulae are obtained from approximating polynomials, the coefficients of which are derived by requiring $f(x)$ and the approximation to have the same values and derivatives at n points. In more than one dimension it is customary to take the Cartesian product of the points calculated by one-dimensional quadrature formulae. If n is the number of points used in one dimension, the integrand will have to be evaluated at n^k points, where k is the multiplicity of the integral; hence, multidimensional polynomial quadrature formulae tend to be too expensive in problems such as likelihood function maximization in which the integrals have to be computed many times. For a bivariate Simpson's Rule see Zellner (1971).

8.3. *Evaluation of multivariate normal integrals*

Possibly the most common multivariate integration problem is that of integrating the multivariate normal density. Let $N(x|0, \Sigma)$ be a multivariate normal density for the k-dimensional vector variable x with mean 0 and covariance matrix Σ. We require

$$I = \int_{l_1}^{\infty} \cdots \int_{l_k}^{\infty} N(x|0, \Sigma)\, \mathrm{d}x_1 \ldots \mathrm{d}x_k. \tag{8.11}$$

Many fundamental relations concerning the bivariate, trivariate, and multivariate normal integrals are contained in Johnson and Kotz (1972). The bivariate case can be handled effectively by using Hausman and Wise's (1978) modification of a technique due to Owen (1956). For purposes of the method, one expresses the bivariate integral as

$$I(h, k; \rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \times \int_{-\infty}^{h} \int_{-\infty}^{k} \exp\{-\tfrac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2)\}\, \mathrm{d}x\, \mathrm{d}y.$$

If this is differentiated with respect to ρ, it can be integrated with respect to x and y and the result reintegrated with respect to ρ. The result is

$$\tfrac{1}{2}\Phi(h)+\tfrac{1}{2}\Phi(k)-T\left(h,\frac{k-\rho h}{h(1-\rho^2)^{1/2}}\right)-T\left(k,\frac{h-\rho k}{k(1-\rho^2)^{1/2}}\right)$$

if $hk>0$ or if $hk=0$ and h or $k\geqq 0$;

$$I(h,k;\rho)=\tfrac{1}{2}\Phi(h)+\tfrac{1}{2}\Phi(k)-T\left(h,\frac{k-\rho h}{h(1-\rho^2)^{1/2}}\right)-T\left(k,\frac{h-\rho k}{k(1-\rho^2)^{1/2}}\right)-\frac{1}{2}$$

if $hk<0$ or if $hk=0$ and h or $k<0$,

where

$$T(u,v)=-\frac{\tan^{-1}a}{2\pi}-\frac{1}{2\pi}\sum_{j=0}^{\infty}c_j v^{2j+1}$$

and

$$c_j=(-1)^j\frac{1}{2j+1}\left[1-\exp\left(-\tfrac{1}{2}u^2\right)\sum_{i=0}^{j}\frac{u^{2i}}{2^i i!}\right].$$

The Hausman–Wise modification works very fast and is suitable for use in optimizing likelihood functions.

Dutt (1976) represents I by Kendall's tetrachoric series which leads to

$$I=\left(\frac{1}{2}\right)^k-\left(\frac{1}{2}\right)^{k-1}\sum_{i=1}^{k}D^*_{1,i}+\left(\frac{1}{2}\right)^{k-2}\sum_{i=1}^{k}\sum_{k=2}^{k}D^*_{2,ij}+\cdots+D^*_{k,ij\cdots k},\quad i<j,$$

where the individual terms are defined in Dutt (1976) and involve Gaussian quadratures. The technique is recommended for up to $k=4$, beyond which it is likely to become expensive. Rules of thumb are suggested for selecting the degree N of the Hermite polynomials needed in the computation of the D^*'s. In practice it appears that poor choices of N can result either in costly computations or in poor approximations. In any event, for $k\geqq 3$ it is likely to be too expensive a method if integrals are required repetitively, as when a likelihood function is being maximized.

The multivariate normal integral (8.5) gives the probability that $\max_j[U_{ij}-U_{ik}]\leqq 0$. This observation allows Daganzo, Bouthelier and Sheffi (1977) to exploit an

approximation due to Clark (1961). Consider random variables $x_1, x_2, \ldots, x_k$ distributed as $N(\mu, \Sigma)$. It is not difficult to obtain exact formulae for $E(\max(x_1, x_2))$, $\text{var}(\max(x_1, x_2))$, and $\text{cov}(\max(x_1, x_2), x_3)$. If one were to assume that $\max(x_1, x_2)$ is normally distributed, which it is not, then one can recursively calculate the moments of $\max(x_1, \ldots, x_k)$ and, by the (incorrect) distributional assumption, the univariate normal distribution which proxies that of $\max(x_1, \ldots, x_k)$, from which the required probability is easily obtained. The behavior of the Clark probabilities is tested by Manski and Lerman (1981) in choice problems involving three or five alternatives. They compare the computation of choice probabilities according to this method with a particular Monte Carlo procedure. The Clark probabilities agree remarkably well with the Monte Carlo results and are obtained with substantially smaller computational cost. It is estimated that for comparable accuracy, the Monte Carlo approach may be as much as 100 times more expensive. However, the Clark approximation tends to be unsatisfactory when the variances of the x's are relatively unequal [Danganzo (1979)]. Why the Clark probabilities are as accurate as they are in other cases, given that they are derived from a false assumption, is not known.

8.4. *Special cases of the multivariate normal integral*

The normal integrals occurring in discrete choice models may have simpler representations than (8.4) or (8.5) under certain circumstances. Two of these will be examined briefly [Hausman (1980)]. As before, the subscript i denotes the ith individual.

Case 1. If all ε_{ij}, $j = 1, \ldots, m$, in (8.4) are i.i.d. normal with mean zero and unit variance, (8.4) becomes

$$P_{ik} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-\varepsilon_{ik}^2}{2}\right\} \prod_{j \neq k} \Phi\left(\overline{U}_{ik} - \overline{U}_{ij} + \varepsilon_{ik}\right) d\varepsilon_{ik}, \tag{8.12}$$

which is substantially easier to evaluate than (8.4) without the simplifying assumption, since it requires only a single-dimensional numerical integral [with $\Phi(\cdot)$ being efficiently evaluated by a partial fraction expansion routinely available in program libraries].

Case 2. Consider the special case in which

$$\overline{U}_{ij} = z_{ij}'\beta, \tag{8.13}$$

where z_{ij}' is a vector of p observable variables and β a coefficient vector. Assume that β is a random vector distributed normally with mean β_μ and covariance matrix Σ_β. Let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})$ be normal with mean zero and covariance matrix Σ_ε and independent of β. The probability that the kth alternative is chosen is then

$$\begin{aligned} P_{ik} &= \Pr\{(z_{ik} - z_{ij})'\beta \geqq \eta_{ijk} \quad \forall j \neq k\} \\ &= \Pr\{(z_{ik} - z_{ij})'\beta_\mu \geqq \eta_{ijk} + (z_{ij} - z_{ik})'(\beta - \beta_\mu) \quad \forall j \neq k\}. \end{aligned} \tag{8.14}$$

Define the right-hand side in the last probability in (8.14) as ξ_{ijk}. The random vector $\xi_{ik}' = (\xi_{i1k}, \ldots, \xi_{imk})$ is normal with mean zero and covariance matrix $\Sigma_\xi = A\Sigma_\varepsilon A' + Z\Sigma_\beta Z'$, where

$$A = \begin{bmatrix} 1 & & & -1 & & & \\ & \ddots & & \vdots & & & \\ & & \ddots & \vdots & & & \\ & & 1 & -1 & & & \\ & & & -1 & 1 & & \\ & & & -1 & \ddots & & \\ & & & \vdots & & \ddots & \\ & & & -1 & & & 1 \end{bmatrix}$$

is $(m-1)\times m$ and has zeros except on the main diagonal and the kth column and where

$$z = \begin{bmatrix} (z_{i1} - z_{ik})' \\ \vdots \\ (z_{im} - z_{ik})' \end{bmatrix}$$

and is $(m-1)\times p$. If we assume that Σ_β and Σ_ε are both diagonal, it is easy to verify that Σ_ξ can be written as

$$\Sigma_\xi = QQ' + \Sigma_{\varepsilon(k)},$$

where $\Sigma_{\varepsilon(k)}$ is the matrix Σ_ε from which its kth row and column have been deleted and $Q = (z\Sigma_\beta^{1/2} \vdots \sigma_{\varepsilon_k} i)$, $i' = (1 \quad 1 \ldots 1)$. Let $(v, w)' = (v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_m, w_1, \ldots, w_{p+1})$ be a vector of $m + p$ elements independently and normally distributed with mean zero and unit variance. Then

$$\xi_{ik} = \Sigma_{\varepsilon(k)}^{1/2} v - Qw,$$

and substituting on the right, (8.14) becomes

$$P_{ik} = \Pr\{(z_{ik} - z_{ij})'\beta_\mu \geqq \xi_{ijk} \quad \forall j \neq k\}$$
$$= \Pr\left\{ v_j \leqq \left[(z_{ik} - z_{ij})'^{-}\beta_\mu + \sum_{l=1}^{p+1} q_{jl} w_l \right] \Big/ \sigma_{\varepsilon_j} \quad \forall j \neq k \right\}. \tag{8.15}$$

If F and G are the distribution functions of scalars v and w, respectively, the following convolution formula holds [Marsaglia (1963)]:

$$\Pr\{v < c - w\} = \int \Pr\{v < c - w | w\} \, dG = E[F(c - w)],$$

where c is a constant vector and the expectation is taken with respect to G. It follows that (8.15) is

$$P_{ik} = \frac{1}{(2\pi)^{(p+1)/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j \neq k}$$
$$\times \Phi\left[\left((z_{ik} - z_{ij})'\beta_\mu + \sum_{l=1}^{p+1} q_{jl} w_l \right) \Big/ \sigma_{\varepsilon_j} \right] \times \exp\left\{ \frac{-t't}{2} \right\} dt_1 \ldots dt_{p+1}. \tag{8.16}$$

Hence, an $(m-1)$-fold numerical integral has been transformed into a $(p+1)$-fold numerical integral which is computationally advantageous if m is relatively large and p is small. [See Webster (1970) and particularly Hausman (1980) for details and extensions.]

It is an open question as to which method of integration of the multivariate normal density is best in several dimensions from the combined points of view of accuracy and cost. The issue is even less settled when multivariate integrals of other density functions are required.

9. The generation of random numbers

Monte Carlo or sampling experiments are the most common instances in econometrics in which it is necessary to generate (pseudo-) random numbers. An example is the following problem. Assume that an estimator $\hat{\theta}(x_1, \ldots, x_n)$ can be calculated from a sample $\{x_1, \ldots, x_n\}$, where the x_i are i.i.d. with pdf $f(x)$. If obtaining the sampling distribution of $\hat{\theta}$ analytically is an intractable problem, one may prefer to obtain experimental evidence about its behavior by repeated simulated drawings of samples of x's and by examination of the resulting $\hat{\theta}$'s. In

order to enable one to perform such experiments, one must be able to sample from arbitrary distributions $f(x)$ [Cragg (1968)].

The present section is devoted to the principal features of generating random numbers and sampling from various distributions. It deals with the computational features of various techniques, but not with the principles of the design of sampling experiments [see, for example, Naylor (1971)]. Among the computational features of greatest interest are the accuracy with which $f(x)$ is sampled and the computational cost of the methods.

9.1. *The generation of uniformly distributed variables*

A fundamental step in generating a variable, x, with pdf $f(x)$, is first to generate u distributed uniformly on (0,1). A common method of generating x is based on the observation that the quantity $y = F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ is distributed as $U(0,1)$ for any continuous $f(x)$. Letting x, y denote random variables and $\bar{x}$, $\bar{y}$ particular values, the assertion follows from

$$\Pr\{y \leqq \bar{y}\} = \Pr\{F(x) \leqq \bar{y}\} = \Pr\{x \leqq F^{-1}(\bar{y})\} = F(F^{-1}(\bar{y})) = \bar{y}.$$

Given a sample of $u_1,\ldots,u_n$ from $U(0,1)$, a corresponding sample $x_1,\ldots,x_n$ for $f(x)$ is obtained by solving

$$u_i = F(x_i), \qquad i = 1,\ldots,n. \tag{9.1}$$

Other uses of uniform deviates occur when some other function of uniform variables has the required distribution (see next subsection for generating normally distributed variables) or when Monte Carlo integration is to be performed (see Section 8).

The most commonly employed generators of $U(0,1)$ variables are based on the recurrence relation

$$R_{i+1} = \lambda R_i + \mu \pmod{M}, \tag{9.2}$$

where λ, μ, M and R_0 are integers chosen by the user. Uniform variables are obtained by calculating R_{i+1}/M, R_{i+2}/M, etc. Generators of the form of (9.2) are congruential generators; they are called mixed or linear if $\mu \neq 0$ and simple or multiplicative in the reverse case. All generators of type (9.2) have finite periods p such that $R_{i+p} = R_i$; the magnitude of p depends on λ, μ, and M. It is obviously desirable to choose these in such a manner as to make p large. The maximal p is easy to find if M is of the form 2^m. For $\mu \neq 0$, the maximal p is 2^m and is obtained if $\lambda = 1 \pmod 4$ and μ is odd; for $\mu = 0$ the maximal period is 2^{m-1} if $\lambda = 3$ or 5

(mod 8) and R_0 is odd [Newman and Odell (1971), Chambers (1977), and Atkinson (1980)]. In practice, for computers with a word-length of 32 bits, M is frequently chosen to be 2^{31} or $2^{31}-1$. The latter is particularly attractive since it is prime and has period $2^{31}-2$ which can be attained if λ is a primitive root of M [Hoaglin (1976)].

The extent to which numbers R_i/M can be thought to have been drawn from $U(0,1)$ can be tested by numerous statistical techniques, such as run tests, tests based on serial correlation properties including computation of the spectral density function, χ^2-tests, Kolmogorov–Smirnov tests, lattice tests, and many others. None of these may reveal adequately the following number-theoretic fact: that all n-tuples $(R_{i+1},\ldots,R_{i+n})$ lie on at most a certain number of parallel hyperplanes in n-dimensional space. The extent to which the generated numbers approximate the uniform distribution will depend on the separation between the hyperplanes which can either be calculated explicitly or approximated from the n-dimensional spectrum of the R_i. Computations by Hoaglin (1976) indicate that for $M=2^{31}-1$ and $\mu=0$, suitable values of λ are 764261123, 1323257245, 1078318381, 1203248318, 397204094, 2027812808. Comparable information for $M=2^{31}$ does not yet appear to be available.

Several ways exist to improve the quality of random number generators. A simple device is to shuffle blocks of k successive random numbers $R=(R_{i+1},\ldots,R_{i+k})'$ into a set $R^*=(R^*_{i+1},\ldots,R^*_{i+k})'$ by $R^*=PR$, where P is a permutation matrix and may itself be altered from time to time. A somewhat more time-consuming but desirable technique is shuffling with replacement [Hill and Holland (1977)]. Two random number sequences $R_1,\ldots,R_n$ and $S_1,\ldots,S_n$ are generated from two different congruential generators. Select an integer k and set $m=2^k$. In the present application k will determine storage requirements and normally one would set $k\leqq 8$. Set $T_1=R_1,\ldots,T_m=R_m$. Now examine a predetermined set of k bits in S_1 which form an integer j' such that for $j=j'+1$, $1\leqq j\leqq m$. The first random number to be selected is then T_j. The content of T_j is replaced by R_{m+1} and the bits of S_2 are examined to select a new member of the array $T_1,\ldots,T_m$. Thus, elements of the sequence S provide index values for selecting from among the elements of T which are then replenished by the next unused element of the sequence R.

9.2. *The generation of normally distributed variables*

Equation (9.1) is cumbersome to solve if $F(x_i)$ is the cumulative normal distribution and various alternative techniques are employed. A method for generating variables distributed approximately as $N(0,1)$ is by appealing to the Central Limit Theorem. The required variable might thus be computed as $\sum_{i=1}^{n}u_i/n-$

$0.5)/(1/\sqrt{12n})$. This method is currently regarded as not being sufficiently accurate.

Much more popular are transformations of pairs of uniform variables. One of the most frequently used of these is the Box–Müller transformation yielding two independent $N(0,1)$ variables x_1 and x_2 according to the transformation [Box and Müller (1958)]:

$$\begin{aligned} x_1 &= (-2\log u_1)^{1/2}\sin(2\pi u_2), \\ x_2 &= (-2\log u_1)^{1/2}\cos(2\pi u_2), \end{aligned} \tag{9.3}$$

where u_1 and u_2 are independent $U(0,1)$. The joint pdf of u_1 and u_2 is $f(u_1, u_2) = 1$ and the pdf of x_1 and x_2 is $g(x_1, x_2) = |J^{-1}| f(u_1, u_2)$, where J is the Jacobian $\partial(x)/\partial(u)$. Its absolute value is easily shown to be $2\pi/u_1$, and it follows that $g(x_1, x_2) = \exp\{-(x_1^2 + x_2^2)/2\}/2\pi$.

Since (9.3) requires the evaluation of trigonometric functions in addition to logarithms, the Box–Müller method is fairly slow. What is more serious is that the exact distribution of variables generated from (9.3) with the aid of uniform variables obtained from a congruential generator such as (9.2) is not normal and contains 2λ discontinuities [Neave (1973)]. A more accurate and faster alternative to (9.3) is to generate x_1, x_2 according to the Marsaglia–Bray transformation:

$$\begin{aligned} x_1 &= v_1\left[-2\log\left(v_1^2 + v_2^2\right)/\left(v_1^2 + v_2^2\right)\right]^{1/2}, \\ x_2 &= v_2\left[-2\log\left(v_1^2 + v_2^2\right)/\left(v_1^2 + v_2^2\right)\right]^{1/2}, \end{aligned} \tag{9.4}$$

where v_1 and v_2 are $U(-1,1)$ and are conditioned by $v_1^2 + v_2^2 < 1$. An argument similar to that used above verifies that x_1 and x_2 are independent $N(0,1)$ variables [Marsaglia and Bray (1964)].

By far the most effective techniques are the decomposition and the acceptance–rejection methods, often used in combination. The basic rationale of these techniques are described by Newman and Odell (1971); various details and computational experience with different methods of this as well as of other types are given by Ahrens and Dieter (1972) and Kinderman and Ramage (1976).

Both techniques allow one to exploit the fact that it may be much easier to generate random variables from one distribution than from another. Assume we wish to sample from $f(x)$ over some interval (a, b) and that $\phi(x)$ is some other pdf over the same interval (from which it may be particularly easy to sample). Determine a coefficient α such that $f(x) \leqq \alpha\phi(x)$ over (a, b). Then we may sample from $f(x)$ by the following procedure. (1) Draw an x from $\phi(x)$, say $\bar{x}$. (2) Draw a uniform $U(0,1)$ variate u. If $u \leqq f(\bar{x})/\alpha\phi(\bar{x})$, then we accept $\bar{x}$. Otherwise we return to the first step. It is clear that the probabilities $\Pr\{x \leqq x^0\}$ will be

proportional to $\int_a^{x^0} f(x)\,dx$ for all x^0 with the factor of proportionality simply measuring the frequency with which step (2) leads to acceptance of x; hence, an easy-to-sample distribution is employed, instead of the possibly difficult-to-sample $f(x)$, to yield the same result.

An interesting variant is the ratio-of-uniforms method [Robertson and Walls (1980)]. To sample from pdf $f(x)$, define the region $R = \{(u, v) | 0 \leqq u \leqq f^{1/2}(v/u)\}$. Generate u and v uniformly over R. Defining $u = y$ and $v = xy$, it is easy to show that v/u is distributed with pdf $f(x)$. The required procedure then is (1) to generate u and v, (2) reject (u, v) if (u, v) does not fall in the region R, or (3) accept (u, v) and form $x = v/u$ otherwise. Efficiency requires that step (2) not be encountered too often. Robertson and Walls (1980) consider in detail sampling from the normal, Cauchy, t-, and gamma distributions.

A similar idea is exploited by decomposition methods in that they replace a difficult-to-sample distribution with a finite mixture of relatively easy-to-sample distributions. Consider $f(x)$ and $\phi_1(x)$, with $\phi_1(x)$ being easy-to-sample. Then choose $\alpha > 0$ such that $f(x) - \alpha\phi_1(x) \geqq 0$. Unless $f(x)$ and $\phi_1(x)$ coincide,

$$\int_{-\infty}^{\infty} (f(x) - \alpha\phi_1(x))\,dx = 1 - \alpha > 0.$$

Hence, defining $\phi_2(x) = (f(x) - \alpha\phi_1(x))/(1 - \alpha)$, it is clear that $f(x) = \alpha\phi_1(x) + (1 - \alpha)\phi_2(x)$, an $(\alpha, 1 - \alpha)$ weighted mixture of ϕ_1, ϕ_2. If $\phi_1(x)$ is easily sampled and if α can be chosen to be relatively large, we may sample from $f(x)$ efficiently by generating x from $\phi_1(x)$ with probability α and from $\phi_2(x)$ with probability $1 - \alpha$. Clearly, the same type of decomposition may be applied to $\phi_2(x)$ and a decomposition scheme may set $f(x) = \sum_{j=1}^{m} \alpha_j \phi_j(x)$, $0 < \alpha_j < 1$, $j = 1, \ldots, m$. A simple algorithm for sampling from $N(0, 1)$ is given by Newman and Odell (1971). Let u_1, u_2, and u_3 denote independent $U(0, 1)$ variates. The normal density is decomposed into four components with probabilities

$$\alpha_1 = 0.8638554642,$$
$$\alpha_2 = 0.110817965,$$
$$\alpha_3 = 0.002699796063,$$
$$\alpha_4 = 0.02262677245.$$

For each of the four possibilities we generate x as follows:

(1) $x = 2(u_1 + u_2 + u_3 - 1.5)$.

(2) $x = 1.5(u_1 + u_2 - 1)$.

(3) $x =$ the first normal variate from repeated Box–Müller transformations for which $|x| > 3$.

(4) Generate $x = 6u_1 - 3$ and $y = 0.3181471173u_2$ repeatedly if necessary until $y \leqq \psi(x)$, and then accept x, where

$$\psi(x) = \begin{cases} ae^{-x^2/2} - b(3-x^2) - c(1.5-|x|), & |x| < 1, \\ ae^{-x^2/2} - d(3-|x|)^2 - c(1.5-|x|)^2, & 1 \leqq |x| < 1.5, \\ ae^{-x^2/2} - d(3-|x|)^2, & 1.5 \leqq |x| < 3, \\ 0, & \text{otherwise}, \end{cases}$$

where $a = 15.75192787$, $b = 4.263583239$, $c = 1.944694161$, and $d = 2.1317916185$. More complicated algorithms are described by Kinderman and Ramage (1976) and by Peterson and Kronmal (1982).

References

Ahrens, J. H. and U. Dieter (1972) "Computer Methods for Sampling from the Exponential and Normal Distributions", *Communications of the ACM*, 15, 873–882.

Aitcheson, J. and S. D. Silvey (1960) "Maximum Likelihood Estimation Procedures and Associated Tests of Significance", *Journal of the Royal Statistical Society*, Ser. B, 154–171.

Atkinson, A. C. (1980) "Tests of Pseudo-Random-Numbers", *Applied Statistics*, 29, 164–171.

Bard, Y. (1974) *Nonlinear Parameter Estimation*. New York: Academic Press.

Beale, E. M. L. (1960) "Confidence Regions in Non-linear Estimation", *Journal of the Royal Statistical Society*, Ser. B, 22, 41–76.

Belsley, D. A. (1974) "Estimation of Systems of Simultaneous Equations and Computational Specifications of GREMLIN," *Annals of Economic and Social Measurement*, 3, 551–614.

Belsley, D. A. (1979) "On the Computational Competitiveness of Full Information Maximum Likelihood and Three Stage Least Squares in the Estimation of Nonlinear, Simultaneous-Equations Models", *Journal of Econometrics*, 9, 315–342.

Belsley, D. A. (1980) "On the Efficient Computation of the Nonlinear Full-Information Maximum Likelihood Estimator", Technical Report no. 5, Center for Computational Research in Economics and Management Science, Vol. II, Cambridge, Mass.

Berman, G. (1979) "Lattice Approximations to the Minima of Functions of Several Variables", *Journal of the Association of Computing Machinery*, 16, 286–294.

Berndt, E. K., B. H. Hall, R. E. Hall and J. A. Hausman (1974) "Estimation and Inference in Nonlinear Structural Models", *Annals of Economic and Social Measurement*, 3, 653–666.

Box, G. E. P. and M. E. Müller (1958) "A Note on the Generation of Random Normal Deviates", *Annals of Mathematical Statistics*, 26, 610–611.

Brent, R. P. (1973) *Algorithms for Minimization without Derivatives*. Englewood Cliffs, N.J.: Prentice-Hall.

Brown, K. M. and W. B. Gearhart (1971) "Deflation Techniques for the Calculation of Further Solutions of a Nonlinear System", *Numerische Mathematik*, 16, 334–342.

Broyden, C. G. (1972) "Quasi-Newton Methods", in: W. Murray (ed.), *Numerical Methods for Unconstrained Optimization*. New York: Academic Press.

Bussinger, P. A. and G. H. Golub (1969) "Singular Value Decomposition of a Complex Matrix", *Communications of the ACM*, 12, 564–565.

Carnahan, B., H. A. Luther and J. D. Wilkes (1969) *Applied Numerical Methods*. New York: John Wiley & Sons.

Chambers, J. M. (1977) *Computational Methods for Data Analysis*. New York: John Wiley & Sons.
Chapman, D. R. and R. C. Fair (1972) "Full-Information Maximum Likelihood Program: User's Guide", Research Memo. no. 137, Econometric Research Program, Princeton University.
Chow, G. C. (1968) "Two Methods of Computing Full-Information Maximum Likelihood Estimates in Simultaneous Stochastic Equations", *International Economic Review*, 9, 100–112.
Chow, G. C. (1973) "On the Computation of Full Information Maximum Likelihood Estimates for Nonlinear Equation Systems", *The Review of Economics and Statistics*, LV, 104–109.
Chow, G. C. (1975) *Analysis and Control of Dynamic Economic Systems*. New York: John Wiley & Sons.
Chow, G. C. and S. B. Megdal (1978) "The Control of Large-Scale Nonlinear Econometric Systems", *IEEE Transactions on Automatic Control*, AC-23, 344–349.
Clark, C. E. (1961) "The Greatest of a Finite Set of Random Variables", *Operations Research*, 9, 145–162.
Cragg, J. C. (1968) "Some Effects of Incorrect Specifications on the Small Sample Properties of Several Simultaneous-Equation Estimators", *International Economic Review*, 9, 63–86.
Dagenais, M. G. (1978) "The Computation of FIML Estimates as Iterative Generalized Least Squares Estimates in Linear and Nonlinear Simultaneous Equations Models", *Econometrica*, 46, 1351–1362.
Daganzo, C. F. (1979) *Multinomial Probit*. New York: Academic Press.
Daganzo, C. F., F. Bouthelier and Y. Sheffi (1977) "Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm", *Transportation Science*, II, 339–358.
Davidon, W. C. (1959) "Variable Metric Method for Minimization", AEC Research and Development Report ANL-5990 (Rev.).
Davidon, W. C. (1975) "Optimally Conditioned Optimization Algorithms Without Line Searches", *Mathematical Programming*, 9, 1–30.
Dempster, A. P., N. M. Laird and D. B. Rubin (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.
Dennis, J. E. Jr., D. M. Gay and R. E. Welsch (1979) "An Adaptive Nonlinear Least-Squares Algorithm", Technical Report TR-1, Alfred T. Sloan School of Management, Massachusetts Institute of Technology.
Dennis, J. E. and J. J. Moré (1977) "Quasi-Newton Methods, Motivation and Theory", *SIAM Review*, 9, 46–89.
Dent, W. T. (1976) "Information and Computation in Simultaneous Equation Systems", *Journal of Econometrics*, 4, 89–95.
Dent, W. T. (1977) "On Numerical Computation in Simultaneous Equation Systems", *Annals of Economic and Social Measurement*, 6, 123–125.
Dent, W. T. (1980) "On Restricted Estimation in Linear Models", *Journal of Econometrics*, 12, 49–58.
Dhrymes, P. J. (1978) *Introductory Econometrics*. New York: Springer-Verlag.
Drud, A. (1977/78) "An Optimization Code for Nonlinear Econometric Models Based on Sparse Matrix Techniques and Reduced Gradients", *Annals of Economic and Social Measurement*, 6, 563–580.
Duesenberry, J. S., G. Fromm, L. R. Klein and E. Kuh (eds.) (1969) *The Brookings Model: Some Further Results*. Chicago: Rand McNally.
Dutt, J. E. (1976) "Numerical Aspects of Multivariate Normal Probabilities in Econometric Models", *Annals of Economic and Social Measurement*, 5, 547–561.
Fair, R. C. (1974a) "On the Robust Estimation of Econometric Models", *Annals of Economic and Social Measurement*, 3, 667–678.
Fair, R. C. (1974b) "On the Solution of Optimal Control Problems as Maximization Problems", *Annals of Economic and Social Measurement*, 3, 135–154.
Fair, R. C. (1976) *A Model of Macroeconomic Activity, Volume II: The Empirical Model*. Cambridge: Ballinger.
Fair, R. C. (1979) "An Analysis of a Macro-economic Model with Rational Expectations in the Bond and Stock Markets", *American Economic Review*, 69, 539–552.
Fair, R. C. (1980a) "Estimating the Expected Predictive Accuracy of Econometric Models", *International Economic Review*, 21, 701–724.
Fair, R. C. (1980b) "Estimating the Uncertainty of Policy Effects in Nonlinear Models", *Econometrica*, 48, 1381–1391.

Fair, R. C. and W. R. Parke (1980) "Full Information Estimates of a Nonlinear Macroeconomic Model", *Journal of Econometrics*, 13, 269–292.
Fiacco, A. V. and G. P. McCormick (1964) "Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method", *Management Science*, 10, 361–366.
Fletcher, R. (1965) "Function Minimization Without Evaluating Derivatives—A Review", *Computer Journal*, 8, 33–41.
Fletcher, R. and M. J. D. Powell (1963) "A Rapidly Convergent Descent Method for Minimization", *Computer Journal*, 6, 163–168.
Fromm, G. and L. R. Klein (1969) "Solutions of the Complete System," in: J. S. Duesenberry, G. Fromm, L. R. Klein and E. Kuh (eds.), *The Brookings Model: Some Further Results*. Chicago: Rand McNally.
Gallant, R. A. and T. M. Gerig (1980) "Computations for Constrained Linear Models", *Journal of Econometrics*, 12, 59–84.
Goldfeld, S. M. and R. E. Quandt (1972) *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland Publishing Co.
Goldfeld, S. M. and R. E. Quandt (1978) "Some Properties of the Simple Disequilibrium Model with Covariance", *Economics Letters*, 1, 341–346.
Goldfeld, S. M., R. E. Quandt and H. F. Trotter (1966) "Maximization by Quadratic Hill-Climbing", *Econometrica*, 34, 541–551.
Goldfeld, S. M., R. E. Quandt and H. F. Trotter (1968) "Maximization by Improved Quadratic Hill-Climbing and Other Methods", Research Memo no. 95, Econometric Research Program, Princeton University.
Goldfeld, S. M. and R. E. Quandt (1979) "Recent Problems and Advances in Estimating Disequilibrium Models", Paper given at the Western Economic Association Meeting, Las Vegas.
Golub, G. H. (1969) "Matrix Decompositions and Statistical Calculations", in: *Statistical Computation*. New York: Academic Press, pp. 365–397.
Greenstadt, J. (1967) "On the Relative Efficiencies of Gradient Methods", *Mathematics of Computation*, 21, 360–367.
Guttman, I. and D. A. Meeter (1965) "On Beale's Measures of Non-Linearity", *Technometrics*, 7, 623–637.
Hammersley, J. M. and D. C. Handscomb (1964) *Monte Carlo Methods*. London: Methuen.
Hartley, M. J. (1977a) "On the Estimation of a General Switching Regression Model Via Maximum Likelihood Methods", Discussion Paper 415, Dept. of Economics, State University of New York at Buffalo.
Hartley, M. J. (1977b) "On the Calculation of the Maximum Likelihood Estimator for a Model of Markets in Disequilibrium", Discussion Paper 409, Dept. of Economics, State University of New York at Buffalo.
Hausman, J. A. (1980) "Les Modèles Probit de Choix Qualitatifs", *Cahiers du Seminaire d'Econométrie*, 21, 11–31.
Hausman, J. A. and D. A. Wise (1978) "A Conditional Probit Model For Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences", *Econometrica*, 46, 403–426.
Hendry, D. F. (1976) "The Structure of Simultaneous Equation Estimators", *Journal of Econometrics*, 4, 51–88.
Hendry, D. F. (1977) "Numerical Optimization Methods", London School of Economics, mimeo.
Hildebrand, F. B. (1956) *Introduction to Numerical Analysis*. New York: McGraw-Hill.
Hill, R. W. and P. W. Holland (1977) "Two Robust Alternatives to Least-Squares Regression", *Journal of the American Statistical Association*, 72, 828–833.
Hoaglin, D. C. (1976) "Theoretical Properties of Congruential Random Number Generators: An Empirical View", Department of Statistics, Memo. NS-340, Harvard University.
Hoffman, K. and R. Kunze (1961) *Linear Algebra*. Englewood Cliffs, N.J.: Prentice-Hall.
Huang, H. Y. (1970) "Unified approach to Quadratically Convergent Algorithms for Function Minimization", *Journal of Optimization Theory and Applications*, 5, 405–423.
Ito, T., (1980) "Methods of Estimation for Multimarket Disequilibrium Models", *Econometrica*, 48, 97–126.
Jennings, L. S. (1980) "Simultaneous Equations Estimation", *Journal of Econometrics*, 12, 23–39.

Johnson, N. L. and S. Kotz (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley & Sons.
Jorgenson, D. W. and J. J. Laffont (1974) "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances", *Annals of Economic and Social Measurement*, 3, 615–641.
Kiefer, N. M. (1980) "A Note on Switching Regressions and Logistic Discrimination", *Econometrica*, 48, 1065–1069.
Kinderman, A. J. and J. G. Ramage (1976) "Computer Generation of Normal Random Variables", *Journal of the American Statistical Association*, 71, 893–896.
Klema, V. (1973) "A Note on Matrix Factorization", *Annals of Economic and Social Measurement*, 2/3, 317–321.
Kloek, T. and H. K. van Dijk (1978) "Bayesian Estimates of Equation Systems Parameters: An Application of Integration by Monte Carlo", *Econometrica*, 46, 1–19.
Lootsma, F. A. (ed.) (1972) *Numerical Methods for Non-Linear Optimization*. New York: Academic Press.
MacKinnon, J. G. (1979) "Convenient Singularities and Maximum Likelihood Estimation", *Economics Letters*, 3, 41–44.
Maddala, G. S. (1971) "The Use of Variance Components Models in Pooling Cross Section and Time Series Data", *Econometrica*, 39, 341–358.
Manski, C. F. and S. R. Lerman (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities", in: C. F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data (with Econometric Applications)*. Cambridge, MA: MIT Press.
Marsaglia, G. (1963) "Expressing the Normal Distribution with Covariance Matrix $A + B$ in Terms of One with Covariance Matrix A", *Biometrika*, 50, 535–538.
Marsaglia, G. and T. A. Bray (1964) "A Convenient Method for Generating Normal Variables", *SIAM Review*, 6, 260–264.
Marquardt, D. W. (1963) "An Algorithm for Least Squares Estimation of Nonlinear Parameters", *SIAM Journal*, 11, 431–441.
McCarthy, M. D. and C. J. Palash (1977) "The Use of Almon- and Other Dummy-Variable Procedures to Increase the Efficiency of Maximization Algorithms in Economic Control", *Annals of Economic and Social Measurement*, 6, 225–230.
Murray, W. (1972) *Numerical Methods for Unconstrained Optimization*. New York: Academic Press.
Naylor, T. H. (1971) *Computer Simulation Experiments with Models of Economic Systems*. New York: John Wiley & Sons.
Neave, H. R. (1973) "On Using the Box–Müller Transformation with Multiplicative Congruential Pseudo-random Number Generators", *Applied Statistics*, 22, 92–97.
Nelder, J. A. and R. Mead (1965) "A Simplex Method for Function Minimization", *Computer Journal*, 7, 308–313.
Newman, T. G. and P. L. Odell (1971) *The Generation of Random Variates*. New York: Hafner.
Oren, S. S. and D. G. Luenberger (1974) "Self Scaling Variable Metric Algorithms, Part I", *Management Science*, 20, 845–862.
Osborne, M. R. (1972) "On Penalty and Barrier Function Methods in Mathematical Programming", in: S. Andersen, L. S. Jennings and D. M. Ryan (eds.), *Optimization*. St. Lucia: University of Queensland Press.
Owen, D. B. (1956) "Tables for Computing Bivariate Normal Probabilities", *Annals of Mathematical Statistics*, 27, 1075–1090.
Parke, W. R. (1979) "An Algorithm for Full Information Estimation", Ph.D. Dissertation, Yale University.
Parkinson, J. M. and D. Hutchinson (1972) "An Investigation into the Efficiency of Variants on the Simplex Method," in: F. A. Lootsman (ed.), *Numerical Methods for Non-Linear Optimization*. New York: Academic Press.
Peterson, A. V. and R. A. Kronmal (1982) "On Mixture Methods for the Computer Generation of Random Variables", *The American Statistician*, 36, 184–191.
Pierre, D. A. and M. J. Lowe (1975) *Mathematical Programming Via Augmented Lagrangians*. Reading, MA: Addison–Wesley.
Powell, M. J. D. (1964) "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives", *Computer Journal*, 7, 155–162.

Powell, M. J. D. (1971) "Recent Advances in Unconstrained Optimization", *Mathematical Programming*, 1, 26–57.
Powell, M. J. D. (1973) "On Search Directions for Minimization Algorithms", *Mathematical Programming*, 4, 193–201.
Powell, M. J. D. (1976) "Some Convergence Properties of the Conjugate Gradient Method", *Mathematical Programming*, 11, 42–49.
Quandt, R. E. (1972) "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association*, 67, 306–310.
Quandt, R. E. (1978a) "Maximum Likelihood Estimation of Disequilibrium Models", in: *Pioneering Economics*. Padova: Cedam.
Quandt, R. E. (1978b) "Tests of the Equilibrium vs. Disequilibrium Hypotheses", *International Economics Review*, 19, 435–452.
Quandt, R. E. and J. B. Ramsey (1978) "Estimating Mixtures of Normal Distributions and Switching Regressions", *Journal of the American Statistical Association*, 73, 730–752.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications* (2nd edn.). New York: John Wiley & Sons.
Richard, J. F. and H. Tompa (1980) "On the Evaluation of Poly-*t* Density Functions", *Journal of Econometrics*, 12, 335–352.
Riddell, W. C. (1975) "Recursive Estimation Algorithms for Economic Research", *Annals of Economic and Social Measurement*, 4, 397–406.
Robertson, I. and L. A. Walls (1980) "Random Number Generators for the Normal and Gamma Distributions Using the Ratio of Uniforms Method", AERE-R 10032, Computer Science and Systems Division, AERE Harwell.
Rosen, J. B. (1960) "The Gradient Projection Method for Nonlinear Programming, Part I. Linear Constraints", *Journal of the Society of Industrial and Applied Mathematics*, 8, 181–217.
Rosen, J. B. (1961) "The Gradient Projection Method for Nonlinear Programming, Part II. Nonlinear Constraints", *Journal of the Society of Industrial and Applied Mathematics*, 9, 514–532.
Rothenberg, T. J. and C. T. Leenders (1964) "Efficient Estimation of Simultaneous Equation Systems", *Econometrica*, 32, 57–76.
Salmon, M. (1978) "The Detection of Successive Solutions to Nonlinear Econometric Models", CRES Working Paper R/WP 30, ISSN 0313 7414.
Schmidt, P. (1976) *Econometrics*. New York: Marcel Dekker.
Shreider, Y. A. (1964) *Method of Statistical Testing*. Amsterdam: Elsevier.
Spang, H. A., III (1962) "A Review of Minimization Techniques for Nonlinear Functions", *SIAM Review*, 4, 343–365.
Stewart, G. W. III (1967) "A Modification of Davidon's Minimization Method to Accept Difference Approximation of Derivatives" *Journal of the Association of Computing Machinery*, 14, 72–83.
Stroud, A. H. and D. Secrest (1966) *Gaussian Quadrature Formulas*. Englewood Cliffs: Prentice-Hall.
Swann, W. H. (1972) "Direct Search Methods", in: W. Murray (ed.), *Numerical Methods for Unconstrained Optimization*. New York: Academic Press, pp. 13–28.
Theil, H. (1971) *Principles of Econometrics*. New York: John Wiley & Sons.
Tischler, A. and I. Zang (1977) "Maximum Likelihood Method for Switching Regression Models Without A Priori Conditions", Tel Aviv University.
Tischler, A. and I. Zang (1979) "A Switching Regression Method Using Inequality Conditions", *Journal of Econometrics*, 11, 259–274.
Van der Hoek, G. and M. W. Dijkshoorn (1979) "A Numerical Comparison of Self Scaling Variable Metric Algorithms", Report 7910/0, Erasmus University, Rotterdam.
Walsh, G. R. (1975) *Methods of Optimization*. New York: John Wiley & Sons.
Wampler, R. H. (1980) "Test Procedures and Test Problems for Least Squares Algorithms", *Journal of Econometrics*, 12, 3–22.
Webster, J. T. (1970) "On the Application of the Method of Das in Evaluating a Multivariate Normal Integral", *Biometrika*, 57, 657–660.
Zangwill, W. I. (1967) "Minimizing a Function without Calculating Derivatives", *Computer Journal*, 10, 293–296.
Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.