
Munich Personal RePEc Archive

Hessian and approximated Hessian matrices in maximum likelihood estimation: a Monte Carlo study

Calzolari, Giorgio and Panattoni, Lorenzo
IBM Scientific Center, Pisa, Italy.

August 1983

Online at http://mpra.ub.uni-muenchen.de/28847/
MPRA Paper No. 28847, posted 17. February 2011 / 13:08

HESSIAN AND APPROXIMATED HESSIAN MATRICES IN MAXIMUM LIKELIHOOD ESTIMATION: A MONTE CARLO STUDY

by Giorgio CALZOLARI and Lorenzo PANATTONI

Centro Scientifico IBM, Pisa

ABSTRACT

Full information maximum likelihood estimation of econometric models, linear and nonlinear in variables, is performed by means of two gradient algorithms, using either the Hessian matrix or a computationally simpler approximation. In the first part of the paper, the behavior of the two methods in getting the optimum is investigated with Monte Carlo experimentation on some models of small and medium size. In the second part of the paper, the behavior of the two matrices in producing estimates of the asymptotic covariance matrix of coefficients is analyzed and, again, experimented with Monte Carlo on the same models. Some systematic differences are evidenced.

CONTENTS

The authors have benefitted from discussions with D.A.Belsley, E.M.Cleur, A.Holly, G.Laroque, P.Mazodier, T.J.Rothenberg, P.Schmidt and A Trognon. Responsibility, however, remains entirely with the authors.

1. MOTIVATION AND INTRODUCTION

The motivation underlying this Monte Carlo study on full information maximum likelihood is twofold. On the one side, this work was motivated by the optimization techniques which have been proposed in the last few years and experimented on linear and nonlinear models of increasing size. While some techniques are search algorithms which do not make use of information on first and second derivatives (e.g. Parke, 1982), it is generally acknowledged that gradient methods, and more specifically Newton-like methods, which make use of such information, should be superior to the others, at least near the optimum. The drawback of Newton-like methods, as well pointed out in Belsley (1980, p.222), lies in the excessive cost required in the calculation of the Hessian matrix. Therefore, methods have been proposed in the literature which replace the Hessian matrix with some approximations, like those adopted in Berndt, Hall, Hall and Hausman (1974), or in Dagenais (1978).

Belsley's findings, after comparing the computational optimization performances of different approximations, placed the algorithm which uses the "exact" Hessian in a dominant position for optimization of the FIML objective function. On the other hand Dagenais' experiments showed that a gradient method in which the Hessian is replaced by a suitable approximation can be computationally more efficient than a Newton-like algorithm, at least as long as the robustness with respect to the initial guess of the coefficients is concerned. This motivated the first part of the Monte Carlo study, described in sections 3 and 4 of this paper. The performances of the Newton-like method are compared with the performances of a gradient algorithm in which a more easily

obtainable matrix is used: in the linear case this matrix is the straightforward estimate of matrix R in Rothenberg and Leenders (1964, p.67), in the nonlinear case it is the matrix proposed in Amemiya (1977,p.963) and experimented with in Dagenais (1978). In both cases, from now on it will be referred to as R (or $\hat{R}$) matrix.

An almost systematic indication, in agreement with the results both of Belsley and Dagenais, is derived from the Monte Carlo experiments. The convergence with the Hessian is, in fact, usually faster near the optimum, while the $\hat{R}$ matrix works better far from it. However from our experiments it came out that "near the optimum" must be interpreted in a much more restrictive sense than usually believed. For models with annual data (and therefore short sample period), whichever "good" starting point of the iterative process was adopted, such as the point obtained from single equation estimation, only when more than 99% of the distance between the initial point and the optimum has been covered, the convergence becomes faster using the Hessian. This point seemed worthwhile a comment, since it might be useful for computationally efficient FIML procedures.

On the other hand, the Hessian matrix and its approximations are also used for tests of hypotheses. This study was motivated by some contraddictory results obtained when testing hypotheses on coefficients of some macroeconometric models by means of FIML estimates. According to which estimate was used for the asymptotic covariance matrix of coefficients, it was not too rare to have a zero-coefficients hypothesis rejected in one case and not rejected in another. This was particularly evident for the Klein-I model when using the coefficients covariance matrix obtained from the Hessian (as in Chernoff and Divinsky, 1953), or from the $\hat{R}$ matrix (as in Hendry, 1971, Rothenberg, 1973 and Hausman, 1974), or,

finally, from the outer product of first derivatives of the log-likelihoods (currently used with the Hessian in pseudo-maximum likelihood estimation, see Gourieroux, Monfort and Trognon, 1982, or White, 1982).

Differences were also encountered in other small-medium size models of the standard type; for almost all coefficients, the asymptotic variances obtained from the outer product of first derivatives were larger than those obtained from the Hessian, and the latter, in turn, were larger than those obtained from the $\hat{R}$ matrix.

As far as the first two groups of results are concerned, the inequality is of the same sign as that encountered in Artus, Laroque and Michel (1982, p.21) for a quite different type of model (with discontinuities).

These systematic differences among asymptotically equivalent estimates of variances seemed worthwhile a deeper investigation.

A difficulty, however, was encountered as soon as the estimation of the asymptotic variances was undertaken using the outer product of first derivatives. As pointed out by Hatanaka (1978, pp.332 and 345), consistency of these estimates is ensured only if derivatives are related to the "unconcentrated" log-likelihood, but not when the concentrated likelihood is considered. Dealing with the unconcentrated likelihood on the one side involves considerable increase of computational complexity, and on the other restricts the class of models which can be processed. This is due to the singularity of the so estimated information matrix in the case of undersized samples (see Hatanaka, 1978, p.333).

For these reasons we have temporarily abandoned the analysis concerning the estimation of the coefficients variances based on the outer product of first order derivatives, and confined the

comparison to the other two estimation methods. Moreover, in the case of nonlinear models, the $\hat{R}$ matrix would not produce a consistent estimate of the covariance matrix of coefficients; its expression should be adjusted and its computation would become considerably more complicated. For this reason, in the second part of the paper, we have confined comparisons to linear models; experiments on nonlinear models have been performed on linearized versions of the models.

It will be shown in sections 5 and 6 that the difference between the $\hat{R}$ matrix and the Hessian matrix (with minus sign), at the point which maximizes the likelihood, is equal to the sum of a positive semi-definite matrix and a matrix made up of blocks with rank zero or one (in particular all diagonal blocks are zero). These two last matrices, although vanishing in the limit, are responsible for the numerical differences encountered in practice. The positive semi-definite matrix, by itself, would imply the Hessian (with minus sign) to be always smaller (in matrix sense) than $\hat{R}$ and, therefore, after inversion, the variances obtained from the Hessian would always be greater than the others. The presence of the other matrix prevents this inequality from being systematic and from occurring in all cases. However, the Monte Carlo experiments suggest that the numerical magnitude of the off diagonal blocks of this matrix is usually small enough so as to leave the inequality (at least the inequality of the variances) unaffected in most cases. In other words, it is proved that, even if the inequality does not always hold, it should be expected in a fairly high percentage of cases.

2. NOTATIONS

In deriving the analytic expression of the Hessian matrix and of the $\hat{R}$ matrix we follow Amemiya (1977); reference to the same paper should also be done for detail on the underlying assumptions. Let the nonlinear simultaneous equation model be represented as

$$(1) \qquad f_i(y_t, x_t, a_i) = u_{it} \qquad i=1,2,\dots,m;\ t=1,2,\dots,T$$

where y_t is the $(m\times 1)$ vector of endogenous variables at time t, x_t is the vector of exogenous variables at time t and a_i is the vector of unknown structural coefficients in the i-th equation. The $(m\times 1)$ vector of random error terms at time t, $u_t=(u_{1t}, u_{2t}, \dots, u_{mt})'$, is assumed to be independently and identically distributed as $N(0,\Sigma)$, with Σ completely unknown, apart from being symmetric and positive definite. The complete vector of unknown structural coefficients of the system will be indicated as $a=(a_1', a_2', \dots, a_m')'$.

Under standard assumptions, by equating to zero the first order derivatives of the unconcentrated log-likelihood with respect to Σ, we get the concentrated log-likelihood function

$$(2) \qquad L = \sum_t \log\left|\frac{\partial f_t}{\partial y_t'}\right| - \frac{T}{2}\log\left|T^{-1}\sum_t f_t f_t'\right|$$

where $f_t=(f_{1t}, f_{2t}, \dots, f_{mt})'=u_t$, and the Jacobian determinant $\left|\partial f_t/\partial y_t'\right|$ is taken in absolute value. We define, for the i-th equation, $g_{it}=\partial f_{it}/\partial a_i$, which is a column vector with the same length as a_i; we define also, for any i and j, the matrix $g_{ijt}=\partial^2 f_{it}/\partial a_i \partial a_j'$. If $i\neq j$, g_{ijt} is zero; it is also always zero if the model is linear in the coefficients (even if nonlinear in the variables). We note, now, that g_{it} and g_{ijt} may be regarded as functions of u_t, x_t and a, under the standard assumption of a

one-to-one correspondence between u_t and y_t.

Differentiating L with respect to the coefficients of the i-th equation, and using $\partial g_{it}/\partial u_{jt} = \partial g_{it}/\partial y_t' \cdot (\partial f_t/\partial y_t')_j^{-1}$, we have

$$(3) \qquad \frac{\partial L}{\partial a_i} = \sum_t \frac{\partial g_{it}}{\partial u_{jt}} - T\left(\sum_t g_{it} f_t'\right)\left(\sum_t f_t f_t'\right)_j^{-1}.$$

Further differentiation of (3), with respect to the structural coefficients of equation j, gives the i,j-th block of the Hessian matrix

$$(4) \qquad \frac{\partial^2 L}{\partial a_i \partial a_j'} = \sum_t \frac{\partial g_{ijt}}{\partial u_{it}} - T\left(\sum_t g_{ijt} f_t'\right)\left(\sum_t f_t f_t'\right)_i^{-1} - \left(\sum_t \frac{\partial g_{it}}{\partial u_{jt}} \frac{\partial g_{jt}'}{\partial u_{it}}\right)$$
$$- T\left(\sum_t f_t f_t'\right)_{ij}^{-1}\left(\sum_t g_{it} g_{jt}'\right) + T\left(\sum_t g_{it} f_t'\right)\left(\sum_t f_t f_t'\right)_i^{-1}\left(\sum_t f_t f_t'\right)_i^{-1}\left(\sum_t f_t g_{jt}'\right)$$
$$+ T\left(\sum_t f_t f_t'\right)_{ij}^{-1}\left(\sum_t g_{it} f_t'\right)\left(\sum_t f_t f_t'\right)^{-1}\left(\sum_t f_t g_{jt}'\right)$$

where use has been made of $\partial a_{ijt}/\partial u_{it} = \partial g_{ijt}/\partial y_t' \cdot (\partial f_t/\partial y_t')_i^{-1}$; the inversion sign and a single subscript after a closed parenthesis indicate a column of the inverse of the matrix, while the double subscript indicates the i,j-th element of the inverse of the matrix.

If we confine the analysis to models which are linear in the coefficients (even if nonlinear in the variables), g_{ijt} and its derivatives are zero, so that the first two terms on the right hand side of equation (4) vanish. Moreover, g_{it} is nothing but the vector of values, at time t, of the explanatory variables of the i-th equation. Therefore, the numerical evaluation of equations (3) and (4) requires only one order of differentiation, that is the computation of derivatives of the explanatory endogenous variables appearing in the i-th and j-th equations with respect to the error terms of the same equations; furthermore, since $\partial g_{it}/\partial u_{jt} = \partial g_{it}/\partial y_t' \cdot (\partial f_t/\partial y_t')_j^{-1}$, this differentiation could even be

performed analytically without any particular difficulty. The use of equation (4) is, therefore, a sufficiently manageable matter even for medium-large models and, as far as our computational experience is concerned, its use with numerical calculation of the first derivatives of the g_{jt}'s always ensured quite accurate results, while the rough second order numerical differentiation of L (which, on its turn, involves a further differentiation to calculate the Jacobian determinant) is well known to produce inaccurate results at higher computational costs (see, for example, Eisenpress and Greenstadt, 1966, p.860 and also the discussion in Parke, 1982, p.94 on the difficulty of obtaining a positive definite matrix from calculating the Hessian with numerical differentiation).

If the model is also linear in the variables, equations (3) and (4) further simplify, since $\partial g_{it}/\partial u_{jt}$ is no more time varying; if the model is

$$(5) \qquad Ay_t + Bx_t = u_t$$

then the vector $\partial g_{jt}/\partial u_{jt}$, for any t, is made up of zeros (corresponding to the exogenous components of g_{jt}) and of elements of A^{-1} (corresponding to the endogenous components of g_{jt}).

Introducing the (T×m) matrix F, whose t,i-th element is $f_i(y_t, x_t, a_i)=u_{it}$, and the matrix G_i, whose t-th row is g_{it}', then the vector of first derivatives (3) can be rewritten as

$$(6) \qquad \left[T^{-1}\sum_t \frac{\partial g_{it}}{\partial u_t'}F' - G_i'\right] F(T^{-1}F'F)_i^{-1}.$$

We define, now,

$$(7) \qquad \hat{G}_i = G_i - T^{-1}F\sum_t \frac{\partial g_{it}'}{\partial u_t}$$

and build the block diagonal matrices G and $\hat{G}$, whose m diagonal blocks are G_i and $\hat{G}_i$, respectively. Moreover, evaluating all terms at $\hat{a}$, we have

$$(8) \qquad T^{-1}F'F = \hat{\Sigma}.$$

Equating (5) to zero, and combining all equations for i=1,2,...,m, we get

$$(9) \qquad \hat{G}'(\hat{\Sigma}^{-1}\otimes I)vecF = 0.$$

The left hand side of equation (9) is a computationally simple expression of the gradient of the concentrated log-likelihood.

In Amemiya (1977), an iterative procedure to get the maximum likelihood estimate of a is obtained from a Taylor expansion of vecF as a function of the coefficients vector, a. The iterative method which results is

$$(10) \qquad \hat{a}^{(r)} = \hat{a}^{(r-1)} - \left[\hat{G}'(\hat{\Sigma}^{-1}\otimes I)G\right]^{-1}\hat{G}'(\hat{\Sigma}^{-1}\otimes I)vecF.$$

For linear models, Hausman (1974) derives the analogous of equation (10) and shows that it corresponds to one iteration of Brundy and Jorgenson's (1971) instrumental variables method (full information case).

In both cases (for linear and nonlinear models) the square matrix which appears in brackets on the right hand side of (10) can be replaced by an asymptotically equivalent matrix

$$(11) \qquad \hat{\mathfrak{I}} = \left[\hat{G}'(\hat{\Sigma}^{-1}\otimes I)\hat{G}\right]$$

which has the advantage of being symmetric and positive definite. However, a distinction must be made between the linear and the nonlinear case. In case of linear models, in fact, we have

$$(12)\qquad -\operatorname*{plim}_{T\to\infty} T^{-1}\frac{\partial^2 L}{\partial a\,\partial a'}\bigg|_{a_0} = \operatorname*{plim}_{T\to\infty} T^{-1}\hat{G}'(\hat{\Sigma}^{-1}\otimes I)\hat{G},$$

whereas the equality does not hold for nonlinear models (the right hand side should be replaced by a more complicated expression, see Amemiya, 1977, eqs. (3.14) and (4.10)). This must not be forgotten if we want to compare the behavior of matrix $\hat{R}$ (11) and of the Hessian. Both matrices can be used, even for nonlinear models, in gradient procedures to maximize the likelihood. On the contrary, for the test of hypothesis, the Hessian can be used in all cases, while $\hat{R}$, as stated in (11), can be used only for linear models. For this reason, while experimental comparisons of the gradient algorithms using the two matrices will be performed either on linear or on nonlinear models (but linear in the parameters), comparisons of the estimated asymptotic variances of coefficients, calculated in the two ways, will be confined to linear models (or to the linearized version of a nonlinear model).

3. TWO GRADIENT PROCEDURES.

The optimization procedures used in this paper are two applications of the well known gradient algorithm. A gradient iterative procedure can be represented by the formula:

$$(13)\qquad \hat{a}^{(k)} = \hat{a}^{(k-1)} + \lambda Q\, p^{(k-1)}$$

where $\hat{a}$ is an estimate of the vector of the coefficients, p is the gradient of the log-likelihood function with respect to the vector a , Q is some matrix and λ is a real number.

Gradient methods differ in the way in which the matrix Q and the number λ are selected at each iteration. The selection of the matrix Q determines the choice of the direction along which the search for the maximization of the log-likelihood function will be made. The choice of λ determines the step size in this direction to obtain the new values of the coefficients.

As long as the choice of Q is concerned, two different approaches have been tried:

1) the matrix Q is given by the inverse of the Hessian of the log-likelihood function (with minus sign); in this case the optimization algorithm becomes a Newton-like algorithm;
2) the matrix Q is given by the inverse of the matrix $\hat{R}$ (see above), which can be considered as an approximation of the Hessian (asymptotically exact or not exact, according to the model).

The choice of the step size λ has been performed following an optimality criterion, i.e. trying to maximize the log-likelihood function by means of an univariate search in the selected direction (see also Eisenpress and Greenstadt, 1966, or Dagenais, 1978). Of course, the procedure is only based on heuristic considerations and there is no assurance that such a strategy for the selection of the value of λ is an optimal one; however, it appeared in practice to accelerate the calculations and to assure the convergence in most cases, and, therefore, it gave a good common basis for performing comparisons of the algorithms using the two matrices.

For the univariate search we used a part of Powell's algorithm, as described in Pierre (1969, pp.277-280), which does not involve the use of derivatives, but is quadratic convergent all the same. Particular care had to be used in the choice of the tolerance for the convergence in this univariate search because, although the maximization process improved the computational efficiency of the whole algorithm, this implied the evaluation of several values of the log-likelihood function. These computations, for medium and

large size models, are rather time consuming and it can happen that with a too tight tolerance the algorithm requires a high number of such computations without a corresponding improvement in the efficiency of the whole algorithm. For the experimented models we found that values 0.01-0.001 of the tolerance on λ are usually good values for the overall computational efficiency of the maximization algorithm.

4. EXPERIMENTAL COMPARISON

Monte Carlo experiments have been performed on five models of small and medium size. Three models are linear, while two are nonlinear in variables.

1) A multiplier-accelerator model, with three linear equations, two of which stochastic, and 6 unknown structural coefficients; the equations and empirical data can be found in Dhrymes (1970, pp.533-534).
2) Klein's model-I, that consists of six linear equations, three of which stochastic, and 12 unknown structural coefficients; the equations and empirical data can be found in Rothenberg (1973, ch.5).
3) A model for the Italian economy proposed in Sitzia and Tivegna (1975), consisting of 7 linear equations, 5 of which stochastic, and 19 unknown structural coefficients.
4) A mildly nonlinear version of Klein-I model, obtained by replacing the linear equation for consumption with a log-linear equation (see Belsley, 1980, model 3B).
5) The Klein-Goldberger model (Klein, 1969), which is nonlinear in variables and consists of 20 equations, 16 of which stochastic,

with 54 unknown structural coefficients.

Monte Carlo experiments on all models are based on a few hundred replications, each of which has been performed as follows. Starting from the model with a given set of parameters ("true" coefficients and covariance matrix of the structural disturbances, held fixed in all replications), random values of the endogenous variables over the sample period are generated by means of stochastic simulation and are used for FIML estimation with the two methods.

To reproduce as much as possible the conditions under which FIML estimation is performed in practice, we choose a "good" starting point for each estimation by getting a preliminary single equation estimate (least squares or instrumental variables).

Several convergence criteria (on coefficients, on the likelihood and on the gradient) have been experimented with. While some differences have been encountered in several cases, the overall behavior did not change very much with the different criteria, apart from the obvious lengthening of convergence "tails" when adopting a very tight tolerance. The same can be said about the choice of the predetermined variables in the sample period, (they have been either kept fixed in all experiments, or partially randomly generated using dynamic simulation), and about the choice of the "true" parameters of the model, on which Monte Carlo generations are based.

The results displayed in Table 1 are related to a convergence criterion with relative tolerance 0.0001 on coefficients. For each of the two methods the number of cases (as percentages) in which convergence has occurred after a given number of iterations is shown. The results in the table suggest that the use of the Hessian usually ensures faster convergence than the use of the $\hat{R}$ matrix, and in particular never requires very long tails for the convergence. However, apart from the increasing computational burden, a problem

Table 1

Number of experiments (percentage) in which convergence has been reached after k iterations using $\hat{R}$ and Hessian

	Multiplier-accelerator model		Klein-I model		Log-linear Klein-I model	
k	Matrix $\hat{R}$	Hessian	Matrix $\hat{R}$	Hessian	Matrix $\hat{R}$	Hessian
1	.	.	.	.	.	.
2	.	.	.	.	.	.
3	3.5	11.0	1.0	.	.	.
4	34.5	72.0	17.5	12.0	.	.5
5	37.0	15.0	28.7	30.5	3.5	3.0
6	19.0	1.0	25.0	46.7	13.0	28.5
7	2.5	.6	13.8	7.5	25.5	38.5
8	1.5	.3	6.5	3.2	21.5	20.5
9	1.0	.1	3.2	.1	11.0	6.0
10	.6	.	1.8	.	9.0	2.5
11	.3	.	1.2	.	6.0	.5
12	.1	.	.8	.	3.5	.
13	.	.	.4	.	2.5	.
14	.	.	.1	.	1.8	.
15	.	.	.	.	1.1	.
16	.	.	.	.	.8	.
17	.	.	.	.	.5	.
18	.	.	.	.	.2	.
19	.	.	.	.	.1	.
20	.	.	.	.	.	.

	Linear Italian model		Klein-Goldberger model	
k	Matrix $\hat{R}$	Hessian	Matrix $\hat{R}$	Hessian
1-3	.	.	.	.
4-6	2.5	78.4	.	1.1
7-9	32.0	18.5	.	27.2
10-12	33.0	1.5	3.0	31.5
13-15	14.0	.8	4.0	19.6
16-18	6.0	.5	10.5	11.9
19-21	4.5	.2	20.3	5.4
22-24	3.0	.1	13.2	2.2
25-27	2.0	.	13.2	1.0
28-30	1.3	.	7.9	.1
31-33	.8	.	7.0	.
34-36	.5	.	6.1	.
37-39	.3	.	3.5	.
40-42	.1	.	2.6	.
43-45	.	.	2.6	.
>45	.	.	6.1	.

arises when the Hessian is used for the estimation of rather complex models. In fact, in the experiments made with the Klein-Goldberger model, in about one out five cases it gave a convergence to a local maximum of the log-likelihood function (i.e. to a value remarkably smaller then the one reached with the matrix $\hat{R}$).

These considerations suggested that it could be interesting to have a better insight in the convergence process. For each Monte Carlo replication, we measure that fraction of the distance between the starting point and the optimum covered at each iteration, with the two methods. The distance is measured both on the values of the log-likelihood and as length of the difference between the current and the optimal coefficient vectors. As before, in some cases the two measures give different results, but the overall behavior is practically the same. In Table 2 only results related to the distances measured on the values of the log-likelihood function are displayed on a log-scale. If we call D(k) the distance which, after k iterations, still must be covered to get the optimum, the value which is calculated is

$$(14) \qquad d(k) = -\left[\text{Log } D(k)/D(0)\right]$$

The value of this variable is equal to zero at the starting point, increases at any new (k-th) iteration, as we move monotonically "uphill", and would be infinite at the optimum (in practice it assumes a value of a few units, depending on the choice of the tolerance in the convergence criterion).

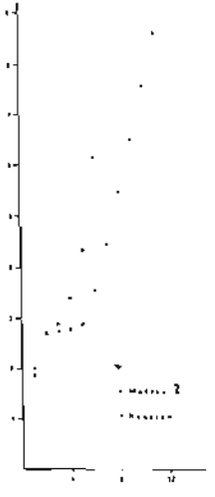
For each model, and for each iteration number (k), the value which is displayed in Table 2 is the average value of all d(k), across Monte Carlo replications, obtained from using $\hat{R}$ or the Hessian.

An interesting systematic behavior of the two methods can be

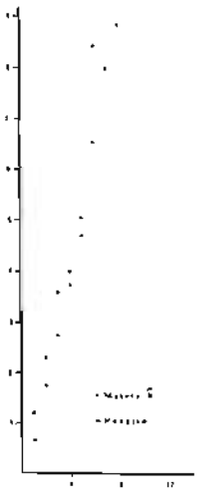
Table 2

Average rate of convergence of the two gradient algorithms

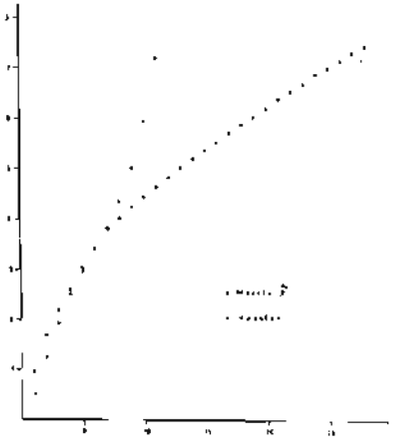
Multiplier-accelerator model



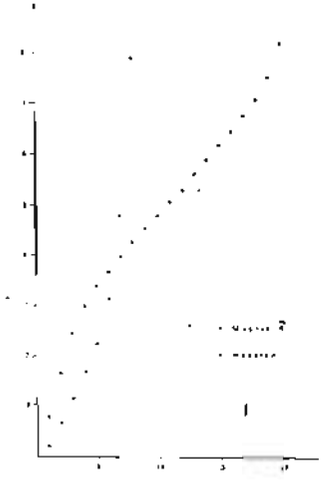
Klein-I model



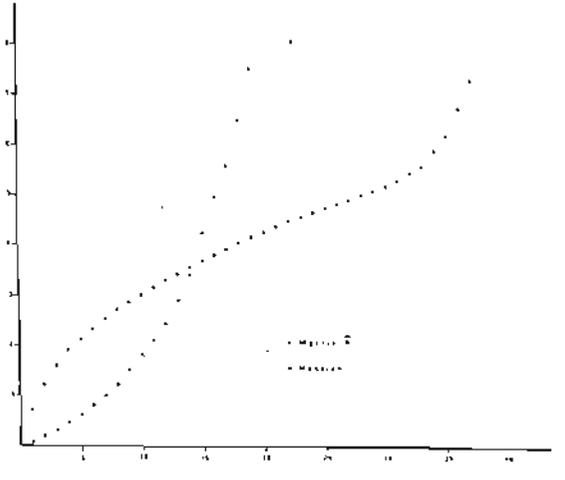
Linear Italian model



Log-linear Klein-I model



Klein-Goldberger model



observed for the models in Table 2. The gradient algorithm, which makes use of the $\hat{R}$ matrix, is considerably faster in the first iterations and, on average, it allows to cover up to 99.99% of the distance, from a "good" starting point up to the optimum, in a smaller number of iterations than the same algorithm which makes use of the Hessian matrix. The dominance of the Hessian matrix becomes effective only in a very tight neighborhood of the optimum, where it allows a considerable reduction of the number of iterations.

This systematic behavior might be interesting for improving the computational efficiency of FIML algorithms. The use of matrix $\hat{R}$ seems recommendable instead of the Hessian in the first iterations, even if the costs to compute the two matrices were the same. It becomes even more recommendable when considering that the computation of matrix $\hat{R}$ is rather simple and fast even for medium-large size models and is, in any case, considerably simpler and faster than computation of the Hessian. Obviously, how much simpler and faster the computation of the matrix $\hat{R}$ is depends on the size of the model; for example for the Klein-Goldberger model the computation of the Hessian took, on the average, about 2.3 times more CPU time than the computation of the $\hat{R}$ matrix, while the time required by the univariate search optimization was about one tenth of this.

5. TWO ESTIMATES OF THE ASYMPTOTIC COVARIANCE MATRIX

As anticipated in section 2, we confine comparison of the two different estimates of the asymptotic covariance matrix of coefficients to the case of linear models.

The first estimate is obtained from the inversion of the matrix $\hat{R}$

of equation (11), calculated upon convergence of the maximization process (so that the gradient (9) is zero). This way of estimating the asymptotic covariance matrix of the structural coefficients is often adopted in practice (see, for example, Hausman, 1974, Hendry, 1971, or Rothenberg, 1973).

In case of the linear system (5), let

$$(15)\quad \begin{cases} \Pi = -A^{-1}B \quad \text{and } \hat{\Pi} = -\hat{A}^{-1}\hat{B} \text{ its FIML estimate} \\ \hat{M} = T^{-1}\sum_t x_t x_t' \text{ and } M \text{ its probability limit} \\ \hat{N} = T^{-1}\sum_t x_t \hat{u}_t' \text{ with zero probability limit} \\ \hat{\Sigma} = T^{-1}\sum_t \hat{u}_t \hat{u}_t' \text{ and } \Sigma \text{ its probability limit.} \end{cases}$$

The matrix $\hat{G}_i$, which consists of T rows and as many columns as the number of explanatory variables in the i-th equation, can be obtained by properly selecting columns of the matrix

$$(16)\quad \begin{bmatrix} y_1' & x_1' \\ y_2' & x_2' \\ \cdot & \cdot \\ \cdot & \cdot \\ y_T' & x_T' \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \cdot \\ \cdot \\ x_T' \end{bmatrix} [\Pi' \;;\; I] + \begin{bmatrix} u_1' \\ u_2' \\ \cdot \\ \cdot \\ u_T' \end{bmatrix} [A'^{-1} \;;\; 0]$$

and the matrix $\hat{G}_i$, of equation (7), can be obtained from properly selecting columns of the matrix

$$(17)\quad \begin{bmatrix} x_1' \\ x_2' \\ \cdot \\ \cdot \\ x_T' \end{bmatrix} [\hat{\Pi}' \;;\; I]$$

The i,j-th block of the matrix $\hat{R}$ can be obtained from properly selecting rows (corresponding to the explanatory variables of equation i) and columns (corresponding to the explanatory variables of equation j) of the matrix

$$(18)\quad T \begin{bmatrix} \hat{\Pi}\hat{M}\hat{\Pi}' & \hat{\Pi}\hat{M} \\ \hat{M}\hat{\Pi}' & \hat{M} \end{bmatrix} \hat{\sigma}^{ij}.$$

The formulas which follow become simpler if we avoid to represent matrices like (18) in partitioned form. This can be simply accomplished by properly augmenting the vector of endogenous variables of the system with the inclusion of all variables which multiply structural coefficients in any stochastic equation and that are either predetermined or functions of endogenous variables. Such a representation of model (5) simply needs the addition of definitional equations, and all the unknown structural coefficients become elements of the matrix A, while no unknown coefficient appears any more in the matrix B. Since all explanatory variables on the right hand side of any stochastic equation are, now, formally represented as endogenous, we can drop the rightmost part of the partitioned matrices (16) and (17), while the whole matrix (18) can be represented as

$$(19)\quad T \left[\hat{\Pi}\hat{M}\hat{\Pi}'\right] \hat{\sigma}^{ij}.$$

As above, we must select rows and columns of matrix (19) in order to build the i,j-th block of the matrix $\hat{R}$.

The second estimate of the asymptotic covariance matrix of structural coefficients is obtained by inverting the Hessian, with minus sign, upon convergence of the maximization process. The i,j-th block of the Hessian (4), $-\partial^2 L/\partial a_i \partial a_j'$, calculated at the FIML estimate of a, can be built starting from the last four terms on the right hand side of equation (4) (the first two terms are zero in our case). As above, for any i and j, we must select the same rows and columns of the matrices which will be given below.

- 3rd term of equation (4):

$$(20)\quad T \left[\hat{A}^{-1} i_j \; i_i' \hat{A}'^{-1}\right]$$

where i_i and i_j are the i-th and j-th columns of the $(m \times m)$ unit matrix.

- 4th term of equation (4):

$$(21) \quad T\left[\hat{\Pi}\hat{M}\hat{\Pi}' + \hat{\Pi}\hat{N}\hat{A}^{-1} + \hat{A}^{-1}\hat{N}'\hat{\Pi}' + \hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}\right]\hat{\sigma}^{ij}.$$

- 5th term of equation (4):

$$(22) \quad -T\left[\hat{\Pi}\hat{N}\hat{\sigma}^{j}\hat{\sigma}^{i}\hat{N}'\hat{\Pi}' + \hat{\Pi}\hat{N}\hat{\sigma}^{j}i_{i}'\hat{A}^{-1} + \hat{A}^{-1}i_{j}\hat{\sigma}^{i}\hat{N}'\hat{\Pi}' + \hat{A}^{-1}i_{j}i_{i}'\hat{A}^{-1}\right]$$

where use has been done of $\hat{\Sigma}\hat{\sigma}^{j}=i_j$ and $\hat{\sigma}^{i}\hat{\Sigma}=i_i'$.

- 6th term of equation (4):

$$(23) \quad -T\left[\hat{\Pi}\hat{N}\hat{\Sigma}^{-1}\hat{N}'\hat{\Pi}' + \hat{\Pi}\hat{N}\hat{A}^{-1} + \hat{A}^{-1}\hat{N}'\hat{\Pi}' + \hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}\right]\hat{\sigma}^{ij}.$$

Several terms cancel, when summing the matrices (20), (21), (22) and (23); we get

$$(24) \quad T\hat{\Pi}\hat{M}\hat{\Pi}'\hat{\sigma}^{ij} - T\hat{\Pi}\hat{N}\hat{\Sigma}^{-1}\hat{N}'\hat{\Pi}'\hat{\sigma}^{ij} - T[\hat{\Pi}\hat{N}\hat{\sigma}^{j}\hat{\sigma}^{i}\hat{N}'\hat{\Pi}' + \hat{\Pi}\hat{N}\hat{\sigma}^{j}i_{i}'\hat{A}^{-1} + \hat{A}^{-1}i_{j}\hat{\sigma}^{i}\hat{N}'\hat{\Pi}'].$$

The first term of (24) is the matrix (19). The other terms, given the presence in each of them of the matrix $\hat{N}$, whose probability limit is zero, asymptotically vanish as expected. However, when calculating the Hessian of the likelihood, they contribute to produce results which numerically differ from those obtained from $\hat{R}$. In particular, as far as the second term of (24) is concerned, we must observe that its overall contribution is to make the resulting matrix smaller (in the usual matrix sense); in fact, if we build the entire matrix whose i,j-th block is obtained from selecting rows and columns of the second term of (24), we would get, by defining an appropriate matrix $\hat{P}$, a matrix of the form $\hat{P}'(\hat{\Sigma}^{-1}\otimes I)\hat{P}$, that is positive semidefinite, and this matrix should be subtracted from $\hat{R}$.

The last term of (24), in brackets, has, however, a quite different behavior. Each block has, in fact, a maximum rank equal to one, since the block is obtained as the product of a column vector with a row vector. Moreover, several elements of the matrix, made up of these blocks, would be zero if calculated upon convergence of the FIML estimation and in particular all the elements of the diagonal blocks would be zero. This follows from considering that, from equation (9), the subvector of the gradient, corresponding to the coefficients of the i-th equation, could be obtained from properly selecting elements from the vector $\hat{\Pi}\hat{N}\hat{\sigma}^{i}$; since this subvector is zero at the optimum, the matrix in brackets on the right hand side of equation (24) is zero when i=j.

Inversion of $\hat{R}$ and of $-\partial^2 L/\partial a \partial a'$, calculated upon convergence of the optimization process, produces two estimates of the asymptotic covariance matrix of structural coefficients. If the Hessian were derived only from the first two terms of (24), the inverted Hessian (with minus sign) would be always greater than $\hat{R}$. This inequality, however, is not exact, given the presence of the last term in (24). Experimentally, however, the contribution of such a term seems to be rather small, and the inequality holds in most cases, at least for the diagonal terms of the inverted matrices (variances).

6. EXPERIMENTAL COMPARISON

Experiments have been performed on the same models of section 4, but a linearized version of the two nonlinear models has been adopted. To be more precise, the two nonlinear models have been maintained in their original form, but the Jacobian determinant in the likelihood has been calculated, as for linear models, only in

Table 3

Klein-I model

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	Monte Carlo
18.34	2.49	4.62	82%
-.2324	.312	.581	70%
.3857	.217	.302	76%
.8018	.036	.044	94%
27.26	7.94	9.53	94%
-.2010	.491	.840	70%
1.052	.352	.424	74%
-.1481	.029	.047	95%
5.794	1.80	3.24	99%
.2341	.049	.095	90%
.2947	.045	.063	92%
.2348	.035	.057	97%

Klein-I

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	M C
1.4[illegible]	[illegible]	.076	[illegible]
.04[illegible]	[illegible]	.015	[illegible]
.03[illegible]	[illegible]	.016	[illegible]
.62[illegible]	[illegible]	.024	9[illegible]
34.[illegible]	[illegible]	10.3	99%
-.317[illegible]	.139	.245	[illegible]
.932[illegible]	.167	.203	[illegible]
-.215[illegible]	.041	.048	[illegible]
2.91[illegible]	1.[illegible]	1.36	[illegible]
.3371[illegible]	.030	.035	9[illegible]
.232[illegible]	[illegible]	.032	[illegible]
.161[illegible]	[illegible]	.029	[illegible]

Multiplier-accelerator model

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	Monte Carlo
.0528	.071	.076	70%
.9580	.108	.117	70%
-1.459	5.86	5.90	86%
-.1033	.253	.248	70%
.8850	.155	.155	84%
12.50	11.0	11.1	85%

Linear Canadian model (consump[illegible] invest. eqs.)

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	M C
.224[illegible]	[illegible]	.726	9
.72[illegible]	[illegible]	.102	[illegible]
.467[illegible]	[illegible]	.124	
.06[illegible]	[illegible]	.224	[illegible]
-9500	[illegible]	1604.	
.154[illegible]	[illegible]	.034	
107.[illegible]	1.[illegible]	17.7	[illegible]
.560[illegible]	.050	.104	[illegible]

Some equations of the Klein-Goldberger model

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	Monte Carlo
Consumption of durables			
.2637	.022	.044	99%
-.1364	.040	.093	99%
-6.042	.787	1.13	99%
Stock of inventories			
.2099	.009	.050	99%
.0428	.045	.137	99%
-38.33	1.99	5.68	99%

Consumption of non-durables

Estim. coeff.	Std.err. with $\hat{R}$	Std.err. Hessian	M C
.124[illegible]	[illegible]	.050	
.87[illegible]	[illegible]	.078	
-1.2[illegible]	[illegible]	1.76	
[illegible] inventories			
.97[illegible]	[illegible]	.025	
-1.0[illegible]	[illegible]	.051	
.39[illegible]	[illegible]	.264	

one year (middle of the sample period and, of course, the other matrices have been calculated accordingly); this is roughly equivalent to a linearization of the models.

For each model, in Table 3 FIML coefficients, estimated from historical data, are first displayed, followed by the two estimates of asymptotic standard errors obtained from the $\hat{R}$ matrix and from the Hessian (both calculated at the same point). Then, for each of the Monte Carlo replications (the same as in section 4), we have computed the asymptotic standard errors of coefficients with the two matrices, at the convergence point (the same, of course, for both). Across the replications, we check how many times, for each coefficient, the standard error computed with the Hessian is greater than the one computed with matrix $\hat{R}$. The number of times that such inequality holds is displayed, in percentage form, in the last column of each model.

7. SUMMARY

The results displayed in the first part of this paper (sections 3 and 4) may be interesting for increasing computational efficiency of full information maximum likelihood algorithms. It seems reasonable to conclude, from the empirical results, that the dominance of the Hessian should be confined only to a very tight neighborhood of the optimum.

The results derived in the second part of the paper (sections 5 and 6) may be interesting when using FIML estimates for the test of hypotheses. Between the two sets of estimates of asymptotic standard errors, those obtained with the Hessian are, in practice, almost systematically larger than the others. Which of them is a

more accurate estimate, however, is a problem to be further investigated.

REFERENCES

Amemiya,T. (1977), "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares in the General Nonlinear Simultaneous Equation Model", Econometrica 45, 955-968.

Artus,P., G.Laroque and G.Michel (1982), "Estimation of a Quarterly Model with Quantity Rationing". Paris: INSEE, discussion paper No.8209.

Belsley,D.A. (1980), "On the Efficient Computation of the Nonlinear Full-Information Maximum-Likelihood Estimator", Journal of Econometrics 14, 203-225.

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, 653-665.

Brundy,J.M. and D.W.Jorgenson (1971), "Efficient Estimation of Simultaneous Equations by Instrumental Variables", The Review of Economics and Statistics 53, 207-224.

Chernoff,H. and N.Divinsky (1953), "The Computation of Maximum-Likelihood Estimates of Linear Structural Equations", in Studies in Econometric Method, ed. by W.C.Hood and T.C.Koopmans. New York: John Wiley & Sons, Cowles Commission Monograph No.14, 236-302.

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.

Dhrymes,P.J. (1970), Econometrics: Statistical Foundations and Applications. New York: Harper & Row.

Eisenpress,H. and J.Greenstadt (1966), "The Estimation of Nonlinear Econometric Systems", Econometrica 34, 851-861.

Gourieroux,C., A.Monfort and A.Trognon (1982), "Pseudo Maximum Likelihood Methods: Theory". Paris: CEPREMAP, Discussion paper No.8129.

Hatanaka,M. (1978), "On the Efficient Estimation Methods for the Macro-Economic Models Nonlinear in Variables", Journal of Econometrics 8, 323-356.

Hausman,J.A. (1974), "Full Information Instrumental Variables Estimation of Simultaneous Equations Systems", Annals of Economic and Social Measurement 3, 641-652.

Hendry,D.F. (1971), "Maximum Likelihood Estimation of Systems of Simultaneous Regression Equations with Errors Generated by a Vector Autoregressive Process", International Economic Review 12, 257-272.

Klein,L.R. (1969), "Estimation of Interdependent Systems in Macroeconometrics", Econometrica 37, 171-192.

Parke,W.R. (1982), "An Algorithm for FIML and 3SLS Estimation of Large Nonlinear Models", Econometrica 50, 81-95.

Pierre,D.A. (1969), Optimization Theory with Applications. New York: John Wiley & Sons.

Rothenberg,T.J. (1973), Efficient Estimation with A Priori Information. New Haven: Yale University Press, Cowles Foundation Monograph 23.

Rothenberg,T.J. and C.T.Leenders (1964), "Efficient Estimation of Simultaneous Equation Systems", Econometrica 32, 57-76.

Sitzia,B. and M.Tivegna (1975), "Un Modello Aggregato dell'Economia Italiana 1952-1971", in Contributi alla Ricerca Economica No.4. Roma: Banca d'Italia, 195-223.

White,H. (1982), "Maximum Likelihood Estimation of Misspecified Models", Econometrica 50, 1-25.